ECONSTOR
Make Your Publications Visible.

A Service of
ZBW
Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Koller, Wolfgang

Book

Prognose makroökonomischer Zeitreihen: Ein Vergleich linearer Modelle mit neuronalen Netzen

Forschungsergebnisse der Wirtschaftsuniversität Wien, No. 63

Provided in Cooperation with:
Peter Lang International Academic Publishers

Suggested Citation: Koller, Wolfgang (2013) : Prognose makroökonomischer Zeitreihen: Ein Vergleich linearer Modelle mit neuronalen Netzen, Forschungsergebnisse der Wirtschaftsuniversität Wien, No. 63, ISBN 978-3-653-03344-1, Peter Lang International Academic Publishers, Frankfurt a. M., https://doi.org/10.3726/978-3-653-03344-1

This Version is available at:
https://hdl.handle.net/10419/178471

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

https://creativecommons.org/licenses/by/4.0/

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

Prognose makroökonomischer Zeitreihen: Ein Vergleich linearer Modelle mit neuronalen Netzen

WOLFGANG KOLLER

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

PL ACADEMIC RESEARCH

Prognose makroökonomischer Zeitreihen:
Ein Vergleich linearer Modelle mit neuronalen Netzten

Forschungsergebnisse der WU Wirtschaftsuniversität Wien

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Band 63

PL ACADEMIC RESEARCH

WOLFGANG KOLLER

Prognose makroökonomischer Zeitreihen: Ein Vergleich linearer Modelle mit neuronalen Netzen

PL ACADEMIC RESEARCH

Bibliografische Information der Deutschen Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the internet at http://dnb.d-nb.de.

Open Access: Die Online-Version dieser Publikation ist unter der internationalen Creative Commons Lizenz CC-BY 4.0 auf www.peterlang.com und www.econstor.eu veröffentlicht.
Erfahren Sie mehr dazu, wie Sie dieses Werk nutzen können:
http://creativecommons.org/licenses/by/4.0.

Dieses Buch ist Open Access verfügbar aufgrund der freundlichen Unterstützung der ZBW – Leibniz-Informationszentrum Wirtschaft.

Gefördert durch die WU Wirtschaftsuniversität Wien.

Umschlaggestaltung:
Atelier Platen, nach einem Entwurf
von Werner Weißhappl.

Universitätslogo der WU Wirtschaftsuniversität Wien:
Abdruck mit freundlicher Genehmigung
der WU Wirtschaftsuniversität Wien.

ISSN 1613-3056
ISBN 978-3-631-64336-5 (Print)
E-ISBN 978-3-653-03344-1 (E-Book)
DOI 10.3726/978-3-653-03344-1

© Peter Lang GmbH
Internationaler Verlag der Wissenschaften
Frankfurt am Main 2014

PL Academic Research ist ein Imprint der Peter Lang GmbH.

Peter Lang – Frankfurt am Main · Bern · Bruxelles · New York · Oxford · Warszawa · Wien

Dieses Buch erscheint in einer Herausgeberreihe bei
PL Academic Research und wurde vor Erscheinen peer reviewed.

www.peterlang.com

Vorwort

Diese Publikation basiert auf meiner Dissertation, die ich an der Wirtschaftsuniversität Wien geschrieben habe und die im Oktober 2012 approbiert wurde.

Die vorliegende Arbeit weist sowohl theoretische als auch angewandte Zielsetzungen auf. Einerseits soll ein Überblick sowohl über die univariate lineare und nicht-lineare Zeitreihenanalyse und -prognose als auch über die neuronalen Netze, im konkreten die sogenannten autoregressiven Neuronale-Netz-Modelle geboten werden, wobei bestimmte theoretische Probleme besonders fokussiert werden. Andererseits ist der Blick immer auf das Anwendungsgebiet der Modellierung und Prognose makroökonomischer Zeitreihen gerichtet. Alle Analysen und entwickelten Methoden werden auf die beiden Beispielzeitreihen der monatlichen österreichischen Arbeitslosenrate und des monatlichen österreichischen Industrieproduktionsindex angewandt. Auf diese Weise können sowohl Aussagen über das Instrumentarium als auch über die untersuchten Zeitreihen gemacht werden.

Aufgrund meiner Arbeit soll ein interessierter Praktiker der linearen und nichtlinearen Zeitreihenanalyse eine umfassende Beurteilung der Eignung von neuronalen Netzen zur Modellierung und Prognose von ökonomischen Zeitreihen zur Verfügung haben. Dies soll die Grundlage bieten für die Entscheidung, ob und mit welchen zur Verfügung stehenden Modellierungszugängen man das Instrumentarium der neuronalen Netze einsetzt.

Gegenüber der Fassung der Dissertation wurden nur wenige inhaltliche Änderungen vorgenommen, wobei es sich meist um Kürzungen handelt. Insbesondere wurde die Dokumentation der beiden selbst entwickelten R-Softwarepakete `NNUTS` und `seasonal` weggelassen, die in die Dissertation als eigener Anhang aufgenommen worden war. Diese Software ist vom Autor bzw. vom Comprehensive R Archive Network (CRAN) unter `http://CRAN.R-project.org` erhältlich.

Danksagung

Mein erster Dank gilt dem Erstbetreuer meiner Dissertation, o.Univ.Prof. Dr. Manfred Fischer. Er ist mir immer mit kritischem Rat zur Seite gestanden und hat mich in meinem Vorhaben bestärkt. Auch dem Institut für Wirtschaftsgeographie und Geoinformatik insgesamt, seinen Mitarbeitern und Mitarbeiterinnen, seinen Studenten und Studentinnen möchte ich meinen Dank aussprechen. Sie haben mir unter anderem im Rahmen des Privatissimums wertvolles Feedback gegeben. Meinem Zweitbetreuer a.o.Univ.Prof. Dr. Alfred Stiassny sage ich Dank für zahlreiche Anregungen, von denen die Arbeit ebenfalls profitiert hat.

Mein besonderer Dank gebührt meinem Freund und früheren Kollegen Dr. Adrian Trapletti, mit dem ich viele Aspekte der Zeitreihenanalyse und -prognose mit neuronalen Netzen oftmals diskutiert habe und auf dessen Vorarbeiten, u.a. im Bereich der Programmierung in R, meine Arbeit in einigen Bereichen aufbaut. Auch der Gemeinschaft aller Personen, die in verschiedenster Weise zum R-Projekt beigetragen haben, bin ich dankbar. Auch mein Freund und Kollege Mag. Philipp Brunner ist mir insbesondere in der Schlussphase der Arbeit mit wervollen Ratschlägen und als Gesprächspartner zur Verfügung gestanden, wofür ich dankbar bin.

Zu größtem Dank bin ich meiner Familie verpflichtet, vor allem meinen Eltern, die mir alles ermöglicht haben, und meiner Frau, Agnes, die mich auch in schwierigen Phasen unterstützt hat und mit Geduld die Fertigstellung erwartet hat. Ihr und meinem Sohn Moritz sei die Arbeit gewidmet.

Inhaltsverzeichnis

Abbildungsverzeichnis **xi**

Tabellenverzeichnis **xiii**

1 Einleitung **1**
1.1 Zielsetzung und Ausrichtung der Arbeit 2
1.2 Vorschau 3
1.3 Mathematischen Notation und Sprachgebrauch 4

2 Lineare Modellierung von Zeitreihen **7**
2.1 Grundlagen der linearen Zeitreihenanalyse 7
2.2 Beispielzeitreihen: Arbeitslosenrate und Industrieproduktionsindex . 14
2.3 Nicht-Stationarität und Herbeiführung der Stationarität 18
2.4 AR-, ARIMA-, SARMA- und ARDS-Modellierung von Zeitreihen . 30
2.5 Schlussfolgerungen 39

3 Tests auf Nicht-Linearität **41**
3.1 Grundlagen und Bedeutung von Nicht-Linearitäten 41
3.2 Teststrategien und eine Auswahl von Testverfahren 48
3.3 Ergebnisse für die beiden Beispielzeitreihen 56
3.4 Strukturbrüche und scheinbare Nicht-Linearitäten 64
3.5 Schlussfolgerungen 66

4 Neuronale Netze und Zeitreihenanalyse **69**
4.1 Grundlagen und Terminologie von ARNN-Modellen 70
4.2 Generalisierungsproblem und Modellierungsstrategien 78
4.3 Datentransformation und Parameterinitialisierung 84
4.4 Lokale Optimierung 89
4.5 Globale Optimierung 97
4.6 Statistisch-parametrischer Ansatz 102
4.7 Klassischer Ansatz mit Early-Stopping 123
4.8 Ansatz mit Regularisierung 129
4.9 Bayesianischer Ansatz 139
4.10 Schlussfolgerungen 147

5 Evaluierung der Prognosen **151**
5.1 Grundlagen 151
5.2 Mehr-Schritt-Prognosen 153
5.3 Anordnung der Prognoseerstellung zum Zweck der Evaluierung . . . 163
5.4 Ergebnisse 170
5.5 Schlussfolgerungen 191

6 Schlussfolgerungen und Ausblick **193**
6.1 Erkenntnisse zum Instrumentarium 193
6.2 Erkenntnisse zu den Zeitreihen 195
6.3 Empfehlungen für den Praktiker 196
6.4 Ausblick auf mögliche weitere Forschung 199

A Ableitungen der Fehlerfunktion im ARNN-Modell **201**
A.1 Fall A: ARNN-Modell 201
A.2 Fall B: ARNNDS-Modell 202

B Ergänzende Ergebnisse zu den Sensitivitätsanalysen **205**

Verzeichnis der Abkürzungen **241**

Verzeichnis der mathematischen Notation **243**

Literatur **247**

Abbildungsverzeichnis

1 Österreichische Arbeitslosenrate (ALR) und Industrieproduktionsindex (IPI), Originalzeitreihe und transformierte Zeitreihe 15
2 Buys-Ballot-Plots und saisonale Zeitreihenplots, ALR und IPI . . . 18
3 Zeitreihenplots für verschieden differenzierte Zeitreihen, ALR und IPI 26
4 Akaike-Informationskriterium in Abhängigkeit von der Modellordnung eines AR-Prozesses, ALR und IPI 32
5 Autokorrelationsfunktion und Partielle Autokorrelationsfunktion, ALR und IPI 35
6 Autokorrelationsfunktion der quadrierten Residuen eines ausführlich spezifizierten AR-Modells auf der saisonal differenzierten Zeitreihe, ALR und IPI 57
7 Ergebnisse des Hsieh-Tests für die Residuen eines ausführlich spezifizierten AR-Modells auf der saisonal differenzierten Zeitreihe, ALR und IPI 59
8 Ergebnisse des TLG-Tests bei Beschränkung des nicht-linearen Teils des Tests auf nur zwei Lags, ALR und IPI 63
9 Simulierte Zeitreihe (SIM), Zeitreihenplot einer Realisation und Quantilsplot der Residuen eines geschätzten AR(10)-Modells 84
10 Exploration der lokalen Minima, ALR und IPI 101
11 Entwicklung der Prognosegüte während des Pruningprozesses, SIM und IPI 119
12 Lernkurven sowie Verteilung der Ergebnisse mit dem Early-Stopping-Ansatz bei wiederholtem Lernen, SIM, ALR und IPI 127
13 Ergebnisse mit dem Regularisierungsansatz in Abhängigkeit von der Wahl der Regularisierungsparameter, ALR und IPI 137
14 Ergebnisse mit dem BEA bei überparametrisierten Modellen im Vergleich mit nicht-bayesianischen Methoden auf einem AR(1)-Prozess und auf SIM 146
15 Iterierte Prognosen auf Basis eines ARNN-Modells: IPI, saisonale Differenzen und SIM 162
16 Vergleich der realisierten Zeitreihe mit Mehr-Schritt-Prognosen auf Basis der getesteten linearen Modellierungsansätze, ALR und IPI . 172

Tabellenverzeichnis

1 Methode der variaten Differenzen, ALR und IPI 27
2 Ergebnisse der Tests auf Unit-Root bzw. Stationarität, ALR und IPI 28
3 Schätzergebnisse für ein AR-Modell, ALR 33
4 Schätzergebnisse für ein AR-Modell, IPI 33
5 Schätzergebnisse für ein ARMA-Modell, ALR 36
6 Schätzergebnisse für ein ARMA-Modell, IPI 36
7 Schätzergebnisse für ein ARDS-Modell, ALR 38
8 Schätzergebnisse für ein ARDS-Modell, IPI 39
9 Ergebnisse des McLeod-Li-Tests auf Residuen eines AR-, ARMA- bzw. ARDS-Modells, ALR und IPI 57
10 Struktur der auf Residuen geschätzten ARCH-Modelle, ALR und IPI 58
11 Ergebnisse des BDS-Tests auf ARCH-bereinigten Residuen eines ausführlich spezifizierten AR-Modells, ALR 60
12 Ergebnisse des BDS-Tests auf ARCH-bereinigten Residuen eines ausführlich spezifizierten AR-Modells, IPI 60
13 Ergebnisse des modifizierten TLG-Tests, ALR und IPI 62
14 Ergebnisse der Tests auf Strukturbruch, ARL und IPI 65
15 Ergebnisse des TLG-Tests auf Signifikanz von zusätzlichen verdeckten Neuronen, SIM 114
16 Ergebnisse des TLG-Tests auf Signifikanz von zusätzlichen verdeckten Neuronen, IPI 115
17 Schätzergebnisse für ein ARNN-Modell, ALR 122
18 Schätzergebnisse für ein ARNN-Modell, IPI 123
19 Koeffizienten der für Mehr-Schritt-Prognosen benötigten Lag-Polnome in Abhängigkeit von Differenzenfilter und Prognosehorizont . 159
20 Modellordnung und Anzahl der enthaltenen Lags in den Modellen der Evaluierungsanordnung, ALR und IPI 166
21 Out-of-Sample-Prognosegüte mit linearen Modellen, ALR und IPI . 174
22 Out-of-Sample-Prognosegüte mit dem statistisch-parametrischen Ansatz, ALR und IPI 175
23 Out-of-Sample-Prognosegüte mit dem klassischen Ansatz, ALR und IPI 177
24 Out-of-Sample-Prognosegüte mit dem Ansatz mit Regularisierung, ALR und IPI 179
25 Out-of-Sample-Prognosegüte mit dem Bayesianischen Ansatz, ALR und IPI 180
26 Hauptergebnisse zur Out-of-Sample-Prognosegüte von linearen und ARNN-Modellen, ALR 186
27 Ergebnisse des Diebold-Mariano-Tests für den Vergleich der Prognosegüte von linearen und ARNN-Modellen, ALR 187

28 Hauptergebnisse zur Out-of-Sample-Prognosegüte von linearen und ARNN-Modellen, IPI 189
29 Ergebnisse des Diebold-Mariano-Tests für den Vergleich der Prognosegüte von linearen und ARNN-Modellen, IPI 190
30 Out-of-Sample-Prognosegüte mit dem statistisch-parametrischen Ansatz auf Basis von alternativen Differenzenfiltern, ALR 206
31 Out-of-Sample-Prognosegüte mit dem klassischen Ansatz auf Basis von alternativen Differenzenfiltern, ALR 207
32 Out-of-Sample-Prognosegüte mit dem Ansatz mit Regularisierung auf Basis von alternativen Differenzenfiltern, ALR 208
33 Out-of-Sample-Prognosegüte mit dem Bayesianischen Ansatz auf Basis von alternativen Differenzenfiltern, ALR 209
34 Out-of-Sample-Prognosegüte mit dem statistisch-parametrischen Ansatz auf Basis von alternativen Differenzenfiltern, IPI 210
35 Out-of-Sample-Prognosegüte mit dem klassischen Ansatz auf Basis von alternativen Differenzenfiltern, IPI 211
36 Out-of-Sample-Prognosegüte mit dem Ansatz mit Regularisierung auf Basis von alternativen Differenzenfiltern, IPI 212
37 Out-of-Sample-Prognosegüte mit dem Bayesianischen Ansatz auf Basis von alternativen Differenzenfiltern, IPI 213
38 Out-of-Sample-Prognosegüte mit dem statistisch-parametrischen Ansatz bei Verwendung der naiv-iterierten Prognosemethode, im Vergleich zur simulativ iterierten Prognosemethode, ALR 214
39 Out-of-Sample-Prognosegüte mit dem klassischen Ansatz bei Verwendung der naiv-iterierten Prognosemethode, im Vergleich zur simulativ iterierten Prognosemethode, ALR 215
40 Out-of-Sample-Prognosegüte mit dem Ansatz mit Regularisierung bei Verwendung der naiv-iterierten Prognosemethode, im Vergleich zur simulativ iterierten Prognosemethode, ALR 216
41 Out-of-Sample-Prognosegüte mit dem Bayesianischen Ansatz bei Verwendung der naiv-iterierten Prognosemethode, im Vergleich zur simulativ iterierten Prognosemethode, ALR 217
42 Out-of-Sample-Prognosegüte mit dem statistisch-parametrischen Ansatz bei Verwendung der naiv-iterierten Prognosemethode, im Vergleich zur simulativ iterierten Prognosemethode, IPI 218
43 Out-of-Sample-Prognosegüte mit dem klassischen Ansatz bei Verwendung der naiv-iterierten Prognosemethode, im Vergleich zur simulativ iterierten Prognosemethode, IPI 219
44 Out-of-Sample-Prognosegüte mit dem Ansatz mit Regularisierung bei Verwendung der naiv-iterierten Prognosemethode, im Vergleich zur simulativ iterierten Prognosemethode, IPI 220

45 Out-of-Sample-Prognosegüte mit dem Bayesianischen Ansatz bei Verwendung der naiv-iterierten Prognosemethode, im Vergleich zur simulativ iterierten Prognosemethode, IPI 221
46 Out-of-Sample-Prognosegüte mit dem klassischen Ansatz unter Weglassung von Direktverbindungen, ALR und IPI 222
47 Out-of-Sample-Prognosegüte mit dem Ansatz mit Regularisierung unter Weglassung von Direktverbindungen, ALR und IPI 223
48 Out-of-Sample-Prognosegüte mit dem Bayesianischen Ansatz unter Weglassung von Direktverbindungen, ALR und IPI 224
49 Out-of-Sample-Prognosegüte mit linearen Modellen mit Jan1975 als Start des In-Samples, ALR 225
50 Out-of-Sample-Prognosegüte des statistisch-parametrischen Ansatzes mit Jan1975 als Start des In-Samples, ALR 226
51 Out-of-Sample-Prognosegüte des klassischen Ansatzes, iterierte Prognosen, mit Jan1975 als Start des In-Samples, ALR 227
52 Out-of-Sample-Prognosegüte des klassischen Ansatzes, direkte Prognosen, mit Jan1975 als Start des In-Samples, ALR 228
53 Out-of-Sample-Prognosegüte des Ansatzes mit Regularisierung, iterierte Prognosen, mit Jan1975 als Start des In-Samples, ALR 229
54 Out-of-Sample-Prognosegüte des Ansatzes mit Regularisierung, direkte Prognosen, mit Jan1975 als Start des In-Samples, ALR 230
55 Out-of-Sample-Prognosegüte des Bayesianischen Ansatzes, iterierte Prognosen, mit Jan1975 als Start des In-Samples, ALR 231
56 Out-of-Sample-Prognosegüte des Bayesianischen Ansatzes, direkte Prognosen, mit Jan1975 als Start des In-Samples, ALR 232
57 Out-of-Sample-Prognosegüte mit linearen Modellen mit Jan1975 als Start des In-Samples, IPI 233
58 Out-of-Sample-Prognosegüte des statistisch-parametrischen Ansatzes mit Jan1975 als Start des In-Samples, IPI 234
59 Out-of-Sample-Prognosegüte des klassischen Ansatzes, iterierte Prognosen, mit Jan1975 als Start des In-Samples, IPI 235
60 Out-of-Sample-Prognosegüte des klassischen Ansatzes, direkte Prognosen, mit Jan1975 als Start des In-Samples, IPI 236
61 Out-of-Sample-Prognosegüte des Ansatzes mit Regularisierung, iterierte Prognosen, mit Jan1975 als Start des In-Samples, IPI 237
62 Out-of-Sample-Prognosegüte des Ansatzes mit Regularisierung, direkte Prognosen, mit Jan1975 als Start des In-Samples, IPI 238
63 Out-of-Sample-Prognosegüte des Bayesianischen Ansatzes, iterierte Prognosen, mit Jan1975 als Start des In-Samples, IPI 239
64 Out-of-Sample-Prognosegüte des Bayesianischen Ansatzes, direkte Prognosen, mit Jan1975 als Start des In-Samples, IPI 240

1 Einleitung

Seit der Wiederentdeckung der neuronalen Netze (NN) etwa Anfang der 80er Jahre des vorigen Jahrhunderts erlebte dieses Forschungsgebiet ein spektakuläres Anwachsen von Forschungsaktivitäten und -ergebnissen. Inspiriert durch die ursprüngliche Idee, mit Hilfe einfacher rechnerischer Einheiten die Funktionsweise von Neuronen und des Gehirns nachzubilden, wurde die Theorie immer weiter ausgebaut und breite Anwendungsmöglichkeiten gefunden, zu denen neben Mustererkennung, Signalverarbeitung und Prozesssteuerung bald auch die Zeitreihenanalyse und -prognose gehörte. Für dieses Anwendungsgebiet kommen meist die auch in der vorliegenden Arbeit untersuchten Feedforward-Netze zum Einsatz, für die besonders viele und gut abgesicherte Ergebnisse vorliegen. Wenn in dieser Arbeit von NN die Rede ist, bezieht sich dies ab nun auf die Feedforward-Netze.

NN-Modelle werden heute als semi-parametrische oder parametrische nicht-lineare Modelle interpretiert, für die die entsprechende statistische Theorie zur Anwendung kommen sollte. Viele der ursprünglichen Innovationen der NN-Forschung haben eine statistisch-theoretische Absicherung erfahren bzw. wurden in die bestehende statistische Theorie integriert.

Frühe Arbeiten im Bereich der Prognose von Zeitreihen ließen die Zeitdimension der Daten oft außer Acht und bauten kaum auf den verfügbaren Grundlagen der Zeitreihenökonometrie und angewandten Statistik auf. Inzwischen hat sich die Sichtweise durchgesetzt, dass in diesem Anwendungsgebiet NN als nicht-lineare Zeitreihenmodelle eingesetzt werden können und im Vergleich mit etablierten linearen und nicht-linearen Zeitreihenmodellen bewertet werden sollten. Insbesondere für die autoregressiven Neuronale-Netz-Modelle (ARNN), die als Generalisierung der autoregressiven Modelle (AR) gelten, wurden seit Mitte der 90er Jahre wichtige theoretische und praktische Ergebnisse erzielt. Diese Entwicklungen erfolgten parallel und in wachsender Verzahnung mit der Forschung zu anderen nicht-linearen Zeitreihenmodellen und zu anderen benachbarten Gebieten der Zeitreihenanalyse wie der Testung auf Nicht-Linearität von Zeitreihen und der Messung und Testung der Prognosegüte.

Der Einsatz von NN für die Modellierung und Prognose ökonomischer Zeitreihen stellt eine Herausforderung dar, da diese eine Reihe von Eigenschaften aufweisen, die besondere Aufmerksamkeit und teilweise eine Erweiterung des Instrumentariums der NN erforderlich machen. Ökonomische Zeitreihen zeichnen sich durch vergleichsweise geringe Länge und einen hohen Anteil des nicht-prognostizierbaren Fehlers aus. Beides sind Faktoren, die gerade den Einsatz von NN erschweren. Zudem spielen nicht-stationäre Komponenten (stochastischer bzw. deterministischer Trend und Saisonalität) eine wichtige Rolle, deren richtige Handhabung auf die Prognosequalität große Auswirkungen haben kann. Schließlich kann a priori nicht mit Sicherheit gesagt werden, ob in einer ökonomischen Zeitreihe Nicht-Linearität jener Art vorhanden ist, die mit neuronalen Netzen modelliert werden kann.

1.1 Zielsetzung und Ausrichtung der Arbeit

Das Ziel der Arbeit ist es, die Eignung des Instrumentariums der NN, im konkreten der ARNN-Modelle, zur Modellierung und Prognose von ökonomischen Zeitreihen zu untersuchen und mit jener der AR- und ARMA-Modelle (autoregressive Moving-Average-Modelle) als den wichtigsten Vertretern der univariaten linearen Zeitreihenmodelle zu vergleichen. Als beispielhaftes Anwendungsgebiet werden die beiden monatlichen Zeitreihen der österreichischen Arbeitslosenrate und des österreichischen Industrieproduktionsindex herangezogen. Die Arbeit beinhaltet die Enwicklung einer Reihe von Erweiterungen am Instrumentarium der ARNN-Modelle, die durch die besonderen Anforderungen des Anwendungsgebiets motiviert sind. Somit umfasst die Zielsetzung der Arbeit sowohl die Gewinnung von Aussagen über das eingesetzte Instrumentarium als auch über die untersuchten Zeitreihen:

Erkenntnisse zum Instrumentarium Wie wird die Familie der ARNN-Modelle definiert und welche Erweiterungen an der Spezifikation sind vorzusehen? Welche Techniken und Vorgangsweisen, die für die Prognose mit ARNN-Modellen, aber auch mit linearen Modellen entwickelt werden, haben eine gute Prognosegüte hinsichtlich Effizienz und Verlässlichkeit?

Erkenntnisse zu den Zeitreihen Was sind die besonderen Eigenschaften der beiden Beispielzeitreihen, die ihre Vorhersagbarkeit durch lineare und nichtlineare Methoden möglich machen? Dies sind z.B. Trend, Saisonalität, Strukturbrüche, vor allem aber die Frage nach nicht-linearen Strukturen, die den Einsatz von nicht-linearen Modellen erst aussichtsreich erscheinen lassen.

Es werden sowohl die theoretischen Grundlagen als auch die angewandten Aspekte des Einsatzes von NN und linearen Modelle erläutert. Die hauptsächliche Ausrichtung ist eine angewandte und kann mit jener von wichtigen Referenzarbeiten wie Swanson und White (1997b), Teräsvirta et al. (2005), und Medeiros et al. (2006) verglichen werden, die ARNN-Modelle in ähnlicher Weise wie die vorliegende Arbeit spezifizieren und ebenso auf einer Auswahl von Zeitreihen deren Eignung im Vergleich zu alternativen Modellen untersuchen. Während in diesen Arbeiten die Auswahl der Beispielzeitreihen breiter ist, wird in der vorliegenden Arbeit auf nur zwei realweltliche Zeitreihen eingeschränkt, dafür jedoch die Anzahl der verschiedenen Modellierungsoptionen vertieft betrachtet. Ein systematischer, umfangreicher Prognosevergleich streng nach dem „Out-of-Sample“-Prinzip erlaubt einen fairen Vergleich der Prognosegüte der verschiedenen Ansätze und Modellierungsstrategien, wobei auch statistische Tests auf Prognosegüte zum Einsatz kommen.

Die in dieser Arbeit eingeführten Neuheiten am Instrumentarium der ARNN-Modellierung sind meist naheliegende Erweiterungen bestehender Methoden und nicht von großer theoretischer Reichweite, haben jedoch, wie sich zeigen wird, großen potentiellen praktischen Nutzen auf dem Anwendungsgebiet der ökonomischen Zeitreihen. Besonders zentral ist die Erweiterung der Spezifikation der ARNN-Modelle

zur Berücksichtigung von sparsamer Modellspezifikation, Einschluss von deterministischer Saisonalität und Trendkomponente und schließlich zur Ermöglichung von direkten Mehr-Schritt-Prognosen.

Weiters wird ein Repertoire von verschiedenen konkreten Vorgangsweisen für die Modellierung und Prognose von ökonomischen Zeitreihen entwickelt, das auf verschiedenen Ansätzen beruht: dem statistisch-parametrischen Ansatz, dem klassischen Ansatz mit Early-Stopping, dem Ansatz mit Regulierung und dem Bayesianischen Evidenzansatz. Für die Auswahl dieser Methoden sowie deren Weiterentwicklung im Rahmen der vorliegenden Arbeit sind als besonders wichtige Ausgangspunkte die Monographien von Bishop (1995, 2006) und Anders (1997) hervorzuheben, letztere insbesondere für die zentrale Idee von Modellierungsstrategien.

Die Analysen und Modelle werden mit Hilfe der frei verfügbaren Statistik-Software R implementiert, wobei die selbst entwickelten Techniken und Methoden in Form von zwei R-Stoftwarepaketen zusammengefasst und dokumentiert werden. Es ist dies einerseits das Paket `NNUTS`, welches Funktionen und Algorithmen im Zusammenhang mit ARNN-Modellen enthält und andererseits `seasonal`, das verschiedene Funktionen im Zusammenhang mit der linearen Modellierung von saisonalen Zeitreihen bereitstellt. Diese Software und ihre Dokumentation sind vom Autor bzw. vom Comprehensive R Archive Network (CRAN) unter `http://CRAN.R-project.org` erhältlich. Es ist zu hoffen, dass auch auf dem Weg der Bereitstellung von freier wissenschaftlicher Software ein Beitrag zur weiteren Forschung auf dem Gebiet der Modellierung und Prognose von ökonomischen Zeitreihen mit linearen und Neuronale-Netz-Methoden geleistet werden kann.

Aufgrund der Arbeit soll ein interessierter Praktiker der Zeitreihenanalyse eine umfassende Beurteilung der Eignung von neuronalen Netzen zur Modellierung und Prognose von ökonomischen Zeitreihen zur Verfügung haben. Dies soll die Grundlage bieten können sowohl für die Entscheidung, ob man überhaupt das Instrumentarium der neuronalen Netze einsetzt, als auch für die Auswahl eines bestimmten Modellierungszugangs.

1.2 Vorschau

Den Zielsetzungen der Arbeit folgend, sowohl Aussagen über das lineare und NN-Instrumentarium als auch über die Zeitreihen zu generieren, weist jedes Kapitel sowohl theoretische bzw. methodische Abschnitte als auch Anwendungen anhand der beiden Zeitreihen auf. Die Arbeit ist in vier Hauptkapitel (Kap. 2–4) gegliedert. Kapitel 1 führt in die Grundlagen der Zeitreihenanalyse und der linearen Modellierung ein und hat in diesem Sinne eher vorbereitenden Charakter. Doch auch hier werden sowohl wichtige Aussagen zu den beiden untersuchten Zeitreihen als auch zu den verschiedenen Techniken der Modellierung gewonnen.

Kapitel 3 bringt Grundlagen zu nicht-linearen Zeitreihenmodellen und stellt die Frage nach dem Vorhandensein von Nicht-Linearität aus der Perspektive von stati-

stischen Hypothesentests. Nur wenn in einer Zeitreihe ausreichend Nicht-Linearität von jenem Typ vorhanden ist, der mit NN behandelt werden kann, darf man erwarten, mit NN die Prognosegüte von linearen Modellen zu übertreffen. Eine Abfolge von Nicht-Linearitätstests, die in Hinblick auf die Erfordernisse von ökonomischen Zeitreihen zusammengestellt und teilweise angepasst wird, wird diskutiert und auf die beiden Beispielzeitreihen angewendet.

In Kapitel 4 werden die ARNN-Modelle entwickelt und Methoden für ihre Schätzung und Spezifikation, d.h. für die Modellierung insgesamt, vorgestellt und weiterentwickelt. Insbesondere wurde der Schwerpunkt auf vier verschiedene Modellierungsansätze gelegt, mit denen das für NN besonders prekäre Problem des Overfitting bzw. der Generalisierungsfähigkeit gelöst werden kann. Die Funktionsweise dieser Techniken wird sowohl anhand einer einfachen simulierten nicht-linearen Zeitreihe als auch anhand der beiden Beispielzeitreihen demonstriert. Dieses Kapitel richtet zwar das Hauptinteresse auf Fragen zum Instrumentarium, aber auch hier können aus den Problemen, die bei der Anwendung des Instrumentariums auf die realweltlichen Zeitreihen auftreten, wichtige Schlüsse über die Eigenschaften derselben gezogen werden.

Das letzte Hauptkapitel (Kap. 5) ist schließlich einer systematisch angelegten Prognose- und Evaluierungsanordnung gewidmet, in der auf einem Evaluierungsset („Out-of-Sample“), d.h. einem zur Seite gelegten Abschnitt der Zeitreihe, die Prognosegüte der linearen und NN-Methoden getestet wird. In diesem Kapitel liegt der Schwerpunkt auf Erkenntnissen zu den Daten. Sind die Zeitreihen mit einer der verwendeten Varianten der ARNN-Modelle besser, d.h. deutlich bzw. in statistisch signifikanter Weise besser, als mit linearen Modellen zu prognostizieren, so ist dies ein Nachweis von nicht-linearen Strukturen in diesen beiden konkreten Zeitreihen. In diesem Sinne liefert der systematische Prognosevergleich ein aussagekräftigeres Ergebnis zum Vorhandensein nicht-linearer Strukturen als die Nicht-Linearitätstests des Kapitels 3. Außerdem erfordert die Zusammenstellung der Prognose- und Evaluierungsanordnung die Entwicklung neuen und zusätzlichen Instrumentariums, insbesondere zur Ermöglichung von Mehr-Schritt-Prognosen.

Zwei Anhänge enthalten die Herleitung der Ableitungen der Fehlerfunktion eines ARNN-Modells nach den einzelnen Koeffizienten des Modells, die für die numerische Umsetzung der verschiedenen Lernverfahren notwendig sind (Anhang A), sowie zusätzliche Tabellen mit Ergebnissen zu verschiedenen Sensitivitätsanalysen im Zusammenhang mit dem systemantischen Vergleich der Prognosegüte (Anhang B).

1.3 Mathematischen Notation und Sprachgebrauch

In dieser Arbeit wird versucht, so weit wie möglich eine einheitliche mathematische Notation zugrunde zu legen. Aufgrund der großen Anzahl und Vielfalt der formelmäßig darzustellenden Konzepte, die zudem aus unterschiedlichen Fachrichtungen stammen, ist dies nicht immer möglich bzw. sinnvoll, da neben das Ziel der

einheitlichen Notation auch jenes der Verwendung der aus der jeweiligen Fachliteratur vertrauten Notation tritt. Es müssen daher in einzelnen Fällen mathematische Symbole in verschiedener Bedeutung verwendet werden, wofür der Leser um sein Verständnis gebeten wird. In jedem dieser Fälle ist die gewählte Definition des Symbols im unmittelbaren Zusammenhang angegeben und aus dem Kontext unmissverständlich. Für die Arbeit zentrale Konzepte werden weitgehend in einheitlicher Notation dargestellt. Manche Symbole, die fest im ökonometrischen Sprachgebrauch verankert sind, wie t-Wert, F-Statistik, Q-Statistik u.ä. werden ohne besonderen Hinweis verwendet, auch wenn diese Symbole an anderer Stelle in anderer Definition eingesetzt werden.

Mathematische Symbole werden im allgemeinen kursiv gesetzt. Davon ausgenommen sind verschiedene besondere statistische und mathematische Funktionen wie Erwartungswert, Varianz, Kovarianz, Indikatorfunktion usw., die aufrecht gesetzt werden. Ebenso aufrecht gesetzt werden Vektoren und Matrizen, die zusätzlich fettgeschrieben werden. Für Mengen wird kalligraphischer Schriftsatz verwendet. Griechische Symbole werden vorwiegend für Koeffizienten in Modellen sowie für Hyperparameter und Heuristiken im Rahmen von Algorithmen und Modellierungsstrategien verwendet. Durch die über einem Symbol gesetzten Superskripte $\tilde{\cdot}$, $\bar{\cdot}$, $\hat{\cdot}$, $\dot{\cdot}$ wird eine Modifikation des ursprünglichen Terms, Durchschnitts- oder Erwartungswertbildung, Schätzung sowie Vorläufigkeit oder Ziehung aus einer Simulation zum Ausdruck gebracht (in dieser Reihenfolge). Durch ein hochgestelltes Sternchen bei einer Variable, $\cdot^*$, wird angezeigt, dass diese einen besonderen Wert annimmt, z.B. im Sinne einer Restriktion oder Optimierung. Ein hochgestelltes $\cdot^{(\tau)}$ bezeichnet die Abhängigkeit vom Iterationszeitpunkt. Für die Transposition eines Vektors oder einer Matrix, ebenso wie für die Ableitung einer Funktion wird das Apostroph verwendet, $\cdot'$. Bei der Schreibweise von Dezimalzahlen wird, der englischsprachigen Konvention folgend, nicht das Komma sondern der Punkt als Dezimaltrennzeichen verwendet. Eine umfassende Aufstellung der verwendeten mathematischen Symbole befindet sich am Ende der Arbeit vor dem Literaturverzeichnis.

Auch zum Sprachgebrauch sind einige Vorbemerkungen angebracht. Englischsprachige Fachausdrücke werden soweit als möglich ins Deutsche übertragen. Meist gibt es eindeutige deutsche Übersetzungen. Doch gelegentlich ist eine Übersetzung nur schwer möglich oder verwirrend und wird in diesen Fällen nicht vorgenommen. Die Einführung von neuer deutschsprachiger Fachsprache ist keine wesentliche Aufgabe dieser Arbeit. Wird in Einzelfällen dennoch der Versuch unternommen, so wird immer auf den englischen Ausdruck hingewiesen. Teilweise kann das Problem durch die Verwendung von Abkürzungen umgangen werden. Abkürzungen werden bei ihrer erstmaligen Verwendung definiert und sind in einem Abkürzungsverzeichnis am Ende der Arbeit aufgelistet.

2 Lineare Modellierung von Zeitreihen

Lineare univariate Zeitreihenanalyse wird in der Ökonomie so oft und in so vielen verschiedenen Gebieten angewendet, dass eine Rechtfertigung der Verwendung dieses Instrumentariums nicht notwendig erscheint. Die Alternativen zu linearen Modellen bestehen einerseits im gänzlichen bzw. teilweisen Verzicht auf quantitative Methoden (d.h. im Rückgriff auf reine Heuristiken, Expertenurteile und sogenannte „Judgement-Methoden“) andererseits in der Anwendung nicht-linearer Modelle, die jedoch als Weiterentwicklungen und Verallgemeinerungen der linearen Modelle gesehen werden können und daher auf den Begriffen und der Theorie der linearen Zeitreihenanalyse aufbauen.

In diesem Sinne liefert dieses Kapitel einen Überblick nicht nur über die theoretischen und praktischen Aspekte der linearen Zeitreihenanalyse, sondern greift auch die Grundlagen der Analyse und insbesondere der Prognose von Zeitreihen insgesamt auf. Diese Fragen nehmen mitunter einen ausgesprochen wissenschaftstheoretischen Charakter an („Was ist überhaupt Prognose?“). Andererseits werden die in diesem Kapitel gelegten theoretischen Grundlagen und praktischen Aspekte auch in den darauffolgenden Kapiteln aufgegriffen und vorausgesetzt.

Abschnitt 2.1 führt in die statistischen Grundlagen der Zeitreihenanalyse ein und stellt jene linearen Zeitreihenmodelle vor, die im Rahmen dieser Arbeit untersucht und mit nicht-linearen Modellen verglichen werden sollen. Abschnitt 2.2 bringt einen Überblick über die Anwendung linearer Zeitreihenanalyse für ökonomische Zeitreihen und stellt jene beiden Zeitreihen vor, die in dieser Arbeit durchgehend als Anwendungsbeispiele herangezogen werden. Abschnitt 2.3 widmet sich der Frage, wie die Stationarität einer Zeitreihe festgestellt werden kann und was geeignete Wege zur Herbeiführung von Stationarität sind. Abschnitt 2.4 dient der Entwicklung einer konkreten Vorgangsweise bei der linearen Modellierung und wendet diese auf die beiden Beispielzeitreihen an. Abschnitt 2.5 fasst einige wichtige Erkenntnisse zusammen, die sich aus der linearen Modellierung der beiden Beispielzeitreihen ergeben.

2.1 Grundlagen der linearen Zeitreihenanalyse

Dieser Abschnitt enthält eine Einführung in Denkweisen, Grundbegriffe und die wichtigsten Modelle der linearen Zeitreihenanalyse. Da insbesondere die Details in einer Vielzahl von ausgezeichneten Lehrbüchern der Zeitreihenanalyse nachgelesen werden können (z.B. Mills, 1990; Brockwell und Davis, 1991; Harvey, 1993; Enders, 1995; Schlittgen und Streitberg, 1997), werden einige Aspekte nur knapp behandelt, während andere im Zusammenhang mit dem Ziel der Arbeit stehende Aspekte hingegen stärker hervorgehoben werden. Zudem wird die Behandlung einiger wichtiger Konzepte, die auch für die lineare Zeitreihenanalyse wichtig sind, auf spätere Kapitel verschoben, da sie, wie etwa das Generalisierungsproblem (vgl. Abschnitt 4.2)

oder Modellselektionskriterien (vgl. Abschnitt 4.6) im Zusammenhang mit der Verwendung von neuronalen Netzen eine besonders ausgeprägte Bedeutung haben.

Denkweisen und Grundbegriffe

Eine Zeitreihe (x_t) ist eine zeitlich geordnete Folge von Beobachtungen einer interessierenden Größe x_t, wobei $t \in \mathbb{N}$ der Zeitindex ist.[1] Kennt man von einer Zeitreihe nur vergangene Beobachtungen $x_t, t \leq t^*$, wobei t^* der gegenwärtige Zeitpunkt ist, so ist in vielen Anwendungsgebieten die Prognose zukünftiger $x_t, t > t^*$ eine interessante Aufgabenstellung.

Prognose beruht auf der Annahme, dass die Zukunft – zumindest teilweise – wie die Vergangenheit ist. In der Sprache der Zeitreihenanalyse manifestiert sich diese Denkweise im Begriff des „datengenerierenden Prozesses“ (DGP): es wird angenommen, dass eine im Zeitverlauf unveränderliche Gesetzmäßigkeit existiert, die der beobachteten Zeitreihe zugrundeliegt. Auf der Basis dieser Annahme kann man es wagen, von der Vergangenheit auf die Zukunft zu schließen.[2] Im Rahmen der Zeitreihenanalyse wird versucht, den DGP durch ein Modell abzubilden und dieses gemeinsam mit den bekannten Werten der Zeitreihe zur Prognose der zukünftigen Ausprägungen der Zeitreihe zu verwenden. Die vorliegende Arbeit beschränkt sich auf die univariate Zeitreihenanalyse, die zur Prognose zukünftiger $x_t, t > t^*$ ausschließlich die Informationen verwendet, die in den bisherigen $x_t, t \leq t^*$ enthalten ist.

Für ökonomische Zeitreihen ist es eine weithin akzeptierte Hypothese, dass im Rahmen des DGP auch der Zufall eine wesentliche Rolle spielt, man spricht daher von „stochastischen Prozessen“.[3] Jede einzelne Beobachtung x_t ist hierbei die Realisation einer Zufallsvariable. Die gegenteilige Annahme, ökonomische Zeitreihen könnten das Ergebnis eines rein deterministischen Prozesses sein, wird in Abschnitt 3.1 aufgegriffen. Die Unterscheidung zwischen der Zeitreihe und dem sie generierenden stochastischen Prozess ist essentiell. Die beobachtete Zeitreihe ist nur eine der möglichen Realisationen des stochastischen Prozesses, genauso wie eine einzelne Beobachtung im Rahmen eines Zufallsexperiments nur eine Realisation einer Zufallsvariablen ist. Da in der Wirtschaft wie auch in den meisten anderen

[1] Da in dieser Arbeit immer aus dem Zusammenhang klar ist, wann von der Zeitreihe und wann von der einzelnen Beobachtung die Rede ist, wird diese Unterscheidung in der Notation im weiteren fallen gelassen.

[2] An der Annahme eines DGP ist vor allem in der Ökonomie vielfach Kritik geäußert worden. Oberflächliche Kritik argumentiert, dass es in ökonomischen Zeitreihen häufig zu Strukturbrüchen kommt bzw. dass die Gesetzmäßigkeiten einem stetigen Wandel unterworfen sind. Diese Kritik kann leicht entkräftet werden, indem man zu einem entsprechend allgemeiner formulierten Modell übergeht (vgl. Abschnitt 3.4). Hingegen stellt Keuzenkamp (1995) das Konzept eines DGP als eine „Reifikation“ eines Denkkonstrukts grundsätzlich in Frage.

[3] Für eine grundlegende und aktuelle Einführung in stochastische Prozesse bieten sich unter anderem an: Brockwell und Davis (1991), Mills (1990), Harvey (1993) und Schlittgen und Streitberg (1997)

Anwendungsgebieten wiederholbare Experimente selten anzutreffen sind, steht die Zeitreihenanalyse vor dem Problem, aus nur *einer* Realisation auf den stochastischen Prozess schließen zu müssen. Dieses Problem ist nur mit Hilfe bestimmter Annahmen bezüglich des stochastischen Prozesses lösbar. Es sind dies die beiden Annahmen der Ergodizität und der Stationariät.

Ergodizität eines stochastischen Prozesses ist, grob gesprochen, dann gegeben, wenn die Stichprobenmomente für endlich lange Realisationen des Prozesses gegen die wahren Momente konvergieren, lässt man die Länge der Realisationen gegen unendlich gehen. Da diese Annahme naturgemäß nicht getestet werden kann, wird sie bei der Anwendung zeitreihenanalytischer Instrumente implizit vorausgesetzt.

Die Annahme der *Stationarität* wird in zwei verschieden starken Ausprägungen verwendet. Streng stationär ist ein Prozess, wenn seine Eigenschaften unabhängig von einer Verschiebung des Zeit-Ursprungs $t = 0$ sind. Das heißt, die gemeinsame Verteilung für jede beliebige Indexmenge $\{t_1, t_2, \ldots, t_m\}$ muss identisch mit der gemeinsamen Verteilung für $\{t_1 + k, t_2 + k, \ldots, t_m + k\}$ sein, wobei k eine beliebige Verschiebung entlang der Zeitachse ist. Die Annahme der schwachen Stationarität hingegen verlangt nur, dass die ersten und zweiten Momente unabhängig von einer Verschiebung entlang der Zeitachse sind. Es muss also für alle t und k gelten:

(i) $\mathrm{E}(x_1) = \mathrm{E}(x_2) = \ldots = \mathrm{E}(x_t) = \mu$,

(ii) $\mathrm{Cov}(x_1, x_{1+k}) = \mathrm{Cov}(x_2, x_{2+k}) = \ldots = \mathrm{Cov}(x_t, x_{t+k})$.

Es ist klar, dass der zweite Teil dieser Bedingung auch die Unabhängigkeit der Varianz von Verschiebungen entlang der Zeitachse einschließt: $\mathrm{Var}(x_1) = \mathrm{Var}(x_2) = \ldots = \mathrm{Var}(x_t) = \sigma^2$. Strenge Stationarität bringt immer schwache Stationarität mit sich. Hingegen kann von schwacher Stationarität nur dann auf starke Stationarität geschlossen werden, wenn zusätzlich der Prozess ein Gauß'scher Prozess ist, da eine Normalverteilung durch die ersten beiden Momente bereits vollständig definiert ist.

Aus dem in der Definition des Stationaritätsbegriffes verwendeten Begriff der Autokovarianz

$$\mathrm{Cov}(x_t, x_{t-k}) = \mathrm{E}((x_t - \mu)(x_{t-k} - \mu)) \tag{2.1}$$

leitet sich der Begriff der Autokorrelation ab:

$$\rho_k = \frac{\mathrm{Cov}(x_t, x_{t-k})}{(\mathrm{Var}(x_t)\mathrm{Var}(x_{t-k}))^{0.5}}. \tag{2.2}$$

Die Annahme der Stationarität impliziert, dass die Autokorrelation nur vom *Lag* (d.h. der zeitlichen Verzögerung) k abhängt. Jene Funktion, die jedem Wert k die Autokorrelation ρ_k zuordnet, heißt Autokorrelationsfunktion (ACF). Ein eng mit der Autokorrelationsfunktion verbundenes Konzept ist die Partielle Autokorrelationsfunktion (PACF), die jedem k die partielle Autokorrelation ϕ_{kk} zuordnet. Diese wird mit Hilfe der Autoregression der Ordnung k definiert:

$$x_t = \phi_{k1}x_{t-1} + \phi_{k2}x_{t-2} + \ldots + \phi_{kk}x_{t-k} + u_t, \tag{2.3}$$

wobei u_t ein stationärer Residualprozess ist. Auf diese Weise misst ϕ_{kk} die verbleibende Korrelation zwischen x_t und x_{t-k}, nachdem um die Korrelation zwischen x_t und den dazwischen liegenden Variablen $x_{t-1}, \ldots, x_{t-k+1}$ bereinigt worden ist. ACF und PACF sind für die Identifikation linearer Modelle ein hilfreiches Instrument, spielen aber auch im Rahmen nicht-linearer Modellierung eine gewisse Rolle.

Ein grundlegender Baustein komplizierterer stochastischer Prozesse ist der als „Weißes Rauschen“ oder als (unkorrelierter) Zufallsprozess bezeichnete stochastische Prozess ε_t mit den Eigenschaften

(i) ε_t identisch verteilt,

(ii) $\mathrm{E}(\varepsilon_t) = 0$

(iii) $\mathrm{Cov}(\varepsilon_t, \varepsilon_{t-k}) = 0$ für alle $k \neq 0$.

Ein Zufallsprozess, der zusätzlich die Bedingung erfüllt, dass ε_t für alle t unabhängig verteilt ist, ist ein Reiner Zufallsprozess („Reines Weißes Rauschen“). Ein normalverteilter Zufallsprozess heißt Normalprozess oder Gaußprozess. Realweltliche ökonomische Zeitreihen, die einem (Reinen) Weißen Rauschen entsprechen, sind selten. Üblicherweise wird jedoch angenommen, dass der Fehlerprozess (auch: Innovationsprozess) in einem linearen Modell ein weißes Rauschen ist. Die strengere Annahme eines Reinen Weißen Rauschens oder eines Gaußprozesses als Fehlerprozess ist für die theoretische Herleitung der Eigenschaften der linearen Modelle und der praktischen Schätzung im allgemeinen nicht notwendig.

Schließlich gehören der Backshift-Operator (auch: Lag-Operator) B^k, der eine Verzögerung einer Zeitreihe x_t um k Zeiteinheiten, $B^k x_t = x_{t-k}$, bewirkt, und der Differenzenoperator $\nabla_k^m = (1 - B^k)^m$ zu den Konzepten, von denen im folgenden ständig Gebrauch gemacht wird.

Lineare stochastische Prozesse

Generell wird ein linearer Prozess dadurch definiert, dass er als unendlicher linearer Filter eines unabhängig und identisch verteilten Zufallsprozesses ε_t angeschrieben werden kann:

$$x_t - \mu_t = \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \ldots = \sum_{j=0}^{\infty} \phi_j \varepsilon_{t-j}, \ \phi_0 = 1. \tag{2.4}$$

Hierbei bezeichnet μ_t linear deterministische Komponenten, etwa einen konstanten Mittelwert, $\mu_t = \mu$, saisonale Mittelwerte oder Trendkomponenten. Die Annahme der Unabhängigkeit von ε_t ist in diesem Zusammenhang wesentlich. In der folgenden Diskussion der linearen Modelle reicht jedoch wieder die Annahme der Unkorreliertheit von ε_t aus.[4] Wold'sche Dekompositionstheorem besagt, dass jeder schwach

[4] Harvey (1993, S. 15) verwendet eine eigentümliche Trennung zwischen den Begriffen (linearer) Prozess und (lineares) Modell. Er nennt einen Prozess linear, wenn er nach Gleichung (2.4)

stationäre, rein nicht-deterministische Prozess $(x_t - \mu_t)$ in Form von Gleichung (2.4) ausgedrückt werden kann, wenn für ε_t die schwächere Annahme der bloßen Unkorreliertheit getroffen wird. Bei der Diskussion der linearen Prozesse wird im folgenden zwecks besserer Übersichtlichkeit der Notation angenommen, dass die Zeitreihe x_t keine lineare deterministische Komponenten μ_t enthält.

Das autoregressive Modell der Ordnung p, AR(p), ist gegeben als

$$x_t - \phi_1 x_{t-1} - \ldots - \phi_p x_{t-p} = \varepsilon_t \tag{2.5}$$

bzw.

$$(1 - \phi_1 B - \ldots - \phi_p B^p) x_t = \phi(B) x_t = \varepsilon_t. \tag{2.6}$$

Hierbei bezeichnet $\phi(B)$ ein Polynom im Lag-Operator. Die Bedingung für die Stationarität des AR(p)-Modells ist, dass alle Wurzeln g_i der charakteristischen Gleichung

$$\phi(B) = (1 - g_1 B)(1 - g_2 B) \ldots (1 - g_p B) = 0 \tag{2.7}$$

innerhalb des Einheitskreises liegen. Für den einfachsten Fall eines AR(1)-Modells bedeutet das $-1 < \phi_1 < 1$.

Das Konzept des autoregressive Modells kann auch herangezogen werden, um den wichtigen Begriff der „Unit Root“ (der englische Ausdruck wird im folgenden in der Schreibweise Unit-Root der deutschen Bezeichnung „Einheitswurzel“ vorgezogen) einzuführen. Allgemein gesprochen ist eine Unit-Root dann gegeben, wenn eine der Wurzeln der charakteristischen Gleichung (2.7) genau auf dem Einheitskreis der Ebene der komplexen Zahlen liegt. In der Ökonomie hat sich jedoch die Sprachregelung durchgesetzt, dass eine Unit-Root nur den Fall bezeichnet, dass eine der Wurzeln der charakteristischen Gleichung genau den Wert Eins annimmt. Dann ist eine Repräsentation der Form

$$x_t - x_{t-1} = \nabla x_t = u_t \tag{2.8}$$

gegeben, wobei u_t ein stationärer, linearer Prozess ist, der eventuell als AR-Prozess dargestellt werden kann. Ein Prozess nach Gleichung (2.8) heißt integrierter Prozess.

Das Moving-Average-Modell der Ordnung q, MA(q), ist gegeben als

$$x_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q} \tag{2.9}$$

angeschrieben werden kann, ohne Aussagen über die Eigenschaften des Fehler- bzw. Innovationsprozesses ε_t zu machen. Von einem Modell spricht er, wenn zusätzlich Aussagen über die Eigenschaften von ε_t getroffen werden. Ein lineares Modell nach der Konvention von Harvey entspricht daher genau einem linearen Prozess nach der hier verwendeten Definition. Diese Terminologie von Harvey wird in dieser Arbeit nicht übernommen, da sie nicht der allgemeinen Sprachregelung zu entsprechen scheint. Es ist praktischer, von einer Analogie zwischen Prozess und Modell in dem Sinne auszugehen, dass lineare Prozesse am besten durch rein lineare Modelle behandelt werden, während nicht-lineare Prozesse die Verwendung nicht-linearer Modelle erfordern, vgl. dazu auch Abschnitt 3.1.

bzw.

$$x_t = (1 - \theta_1 B - \ldots - \theta_p B^p)\varepsilon_t = \theta(B)\varepsilon_t. \tag{2.10}$$

MA-Prozesse sind immer stationär. Allerdings ist im Zusammenhang mit MA-Prozessen eine wichtige Eigenschaft die Invertierbarkeit, die dann gegeben ist, wenn alle Wurzeln g_i der charakteristischen Gleichung

$$\theta(B) = (1 - g_1 B)(1 - g_2 B)\ldots(1 - g_p B) = 0 \tag{2.11}$$

außerhalb des Einheitskreises liegen. Nur unter der Annahme der Invertierbarkeit kann aus der ACF eines Prozesses eindeutig auf die Koeffizienten des MA-Modells geschlossen werden.

Zwischen AR-Prozessen und MA-Prozessen bestehen enge Zusammenhänge. Jeder stationäre AR(p)-Prozess kann als MA(∞)-Prozess geschrieben werden, und jeder invertierbare MA(q)-Prozess kann als AR(∞)-Prozess geschrieben werden. Außerdem kann ein stationärer AR(p)-Prozess (ein invertierbarer MA(q)-Prozess) durch einen MA-Prozess (AR-Prozess) genügend großer Ordnung $\tilde{q} > p$ (bzw. $\tilde{p} > q$) beliebig genau approximiert werden. Dies legt die Verbindung beider Modelle im Rahmen des ARMA(p, q)-Modells nahe:

$$x_t - \phi_1 x_{t-1} - \ldots - \phi_p x_{t-p} = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q}, \tag{2.12}$$

bzw. in knapper Notation:

$$\phi(B)x_t = \theta(B)\varepsilon_t. \tag{2.13}$$

In diesem hängt die Stationarität nur vom AR-Teil und die Invertierbarkeit nur vom MA-Teil ab. Ein stationäres ARMA-Modell kann auch folgendermaßen geschrieben werden:

$$x_t = \frac{\theta(B)}{\phi(B)}\varepsilon_t, \tag{2.14}$$

während ein invertierbares ARMA-Modell auch durch

$$\frac{\phi(B)}{\theta(B)}x_t = \varepsilon_t, \tag{2.15}$$

ausgedrückt werden kann.

Eine weitere Generalisierung sind ARIMA(p, d, q)-Modelle,

$$\phi(B)\nabla^d x_t = \theta(B)\varepsilon_t. \tag{2.16}$$

In dieser Schreibweise wird zum Ausdruck gebracht, dass die Ordnung des Differenzierens, d, als Teil der Definition eines Stochastischen Prozesses, bzw. der Modellierungsentscheidung gesehen werden kann.

Ein wichtiger Aspekt, vor allem in makroökonomischen Zeitreihen, ist Saisonalität. Saisonalität ist ein systematischer, wenn auch nicht unbedingt regelmäßiger,

unterjähriger Verlauf einer Zeitreihe, der im Zusammenhang mit Wetter- und Kalendereffekten, und zeitlicher Einteilung von Konsum- und Produktionsentscheidungen zu erklären ist (Hylleberg, 1992). Sie kann auf verschiedene Weise zustande kommen, z.B.:

(i) durch das Zusammenwirken von Parametern in einem ARMA-Modell

(ii) durch eine saisonale Unit-Root

(iii) durch deterministische saisonale Muster.

Im Fall (i) handelt es sich um sogenannte Pseudozyklen, die in fast jedem ARMA-Modell, das mehrere Lags enthält, auftreten, insbesondere wenn saisonale Lags, d.h. Lags zur saisonalen Frequenz, im Modell eine Rolle spielen. Die Stationarität der Zeitreihe ist dadurch nicht berührt. Pseudozyklen werden die auftretenden Muster genannt, weil sie unterschiedlicher Länge und von zufälligen Schocks leicht beeinflussbar sind.

Der Fall (ii) wird unter Verwendung des saisonalen Differenzenoperators ∇_s, wobei hier s die saisonale Frequenz der Zeitreihe bezeichnet, angeschrieben:

$$\nabla_s x_t = u_t,\ u_t \text{ ein stationärer, linearer Prozess} \tag{2.17}$$

Hierbei kann die saisonal differenzierte Zeitreihe $u_t = \nabla_s x_t$ eventuell einem ARMA-Prozess folgen. Ein Prozess nach Gleichung (2.17) heißt saisonal integrierter Prozess. Eine allgemeinere Formulierung berücksichtigt, dass in der Zeitreihe eventuell mehrere Unit-Roots, saisonaler oder nicht saisonaler Natur, präsent sein können. Unter Berücksichtigung dieser Möglichkeit ist ein saisonales ARIMA-Modell:

$$\phi(B)\nabla^d\nabla_s^D x_t = \theta(B)\varepsilon_t. \tag{2.18}$$

In einem multiplikativen saisonalen ARIMA-Modell wird außerdem von der nützlichen Möglichkeit der multiplikative Verknüpfung eines Lag-Polynoms, das konventionelle Lags enthält, mit einem Lag-Polynom, das nur saisonale Lags enthält, Gebrauch gemacht:

$$\phi(B)\Phi(B^s)\nabla^d\nabla_s^D x_t = \theta(B)\Theta(B^s)\varepsilon_t. \tag{2.19}$$

Aus der Sichtweise der linearen stochastischen Prozesse hat das multiplikative gegenüber dem konventionellen saisonalen ARIMA-Modell keine eigenständige Bedeutung, da das Produkt zweier Lag-Polynome wieder nur ein Lag-Polynom ist, also $\phi(B)\Phi(B^s) = \tilde{\phi}(B)$ und $\theta(B)\Theta(B^s) = \tilde{\theta}(B)$. Allerdings erlaubt es eine knappere Formulierung und eine Schätzung, die im Vergleich zu nicht-multiplikativen Modellen Parameterrestriktionen einführt. In der Tradition von Box-Jenkins wird das multiplikative saisonale ARIMA-Modell als SARIMA$(p,d,q)(P,D,Q)_s$ bezeichnet, wobei (p,d,q) die Ordnung des nicht saisonalen Teils und $(P,D,Q)_s$ die Ordnung des saisonalen Teils bezeichnet.

Von hoher Relevanz ist auch der Fall (iii) der obigen Aufstellung. Deterministische saisonale Muster kommen durch saisonal unterschiedliche Mittelwerte bzw. Trends zustande. Entsprechende Modelle können in allgemeiner Form folgendermaßen angeschrieben werden:

$$\begin{aligned} y_t &= \mu_t + x_t \\ \mu_t &= \mu_1 + \sum_{i=2}^{s} \mu_i s_{it} + \mu_1^t t + \sum_{i=2}^{s} \mu_i^t s_{it} t \\ x_t &\sim ARMA. \end{aligned} \tag{2.20}$$

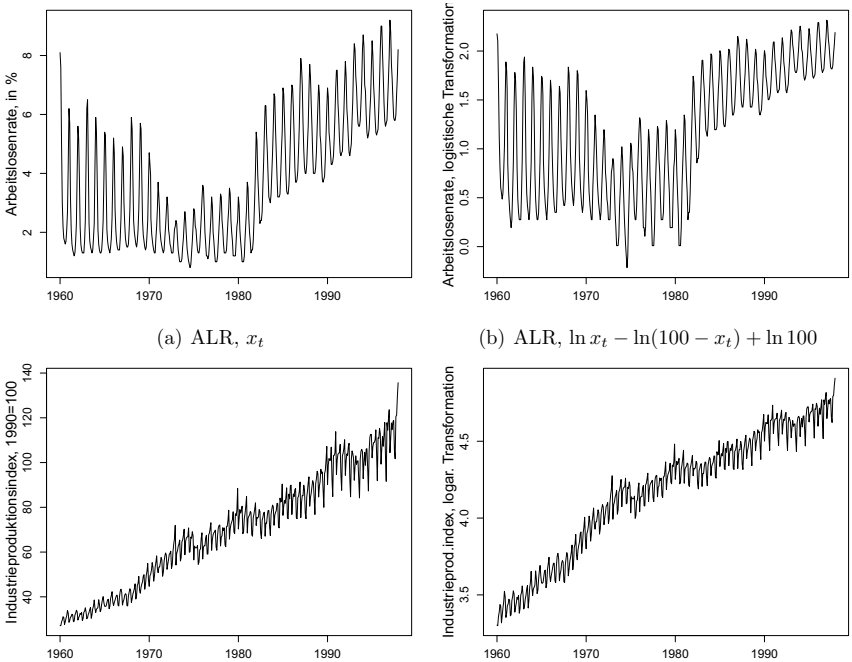
Hierbei ist s_{it} eine saisonale Dummy-Variable: $s_{it} = 1$ wenn $t \mod s = i$ und $s_{it} = 0$ sonst. In der weiteren Folge wird ein solches Modell als ARMADS-Modell bezeichnet. Es soll auch kurz darauf hingewiesen werden, dass dieses Modell auch ein Modell mit einem konventionellen (d.h. nicht-saisonalen) Trend einschließt, wenn $\mu_i^t = 0$ für $1 > i \leq s$. Eine naheliegende Verallgemeinerung dieses Modells, das sowohl saisonale Unit-Roots als auch saisonale Trends berücksichtigt ist:

$$\begin{aligned} \nabla^d \nabla_s^D y_t &= \mu_t + x_t \\ \mu_t &= \mu_1 + \sum_{i=2}^{s} \mu_i s_{it} + \mu_1^t t + \sum_{i=2}^{s} \mu_i^t s_{it} t \\ x_t &\sim ARMA. \end{aligned} \tag{2.21}$$

Die Unterscheidung zwischen saisonalen Unit-Roots und deterministischen saisonalen Mustern, wie auch die Unterscheidung zwischen konventionellen Unit-Roots und deterministischen Trends ist von hoher praktischer Relevanz. Zwar haben die oben genannten Möglichkeiten (ii) und (iii) nicht-stationäres Verhalten zur Folge, erfordern jedoch unterschiedliche Methoden der Herbeiführung der Stationarität, bzw. der Modellierung. Der Frage der Herbeiführung der Stationarität widmet sich Abschnitt 2.3. Die praktische Modellierung von Zeitreihen anhand der soeben vorgestellten linearen Modelle wird anhand von zwei Beispielzeitreihen in Abschnitt 2.4 erläutert. Zunächst werden jedoch im folgenden Abschnitt die beiden Beispielzeitreihen vorgestellt und die praktischen Hintergründe geliefert.

2.2 Beispielzeitreihen: Arbeitslosenrate und Industrieproduktionsindex

In der vorliegenden Arbeit werden die Instrumente der linearen (AR-, ARMA- und SARMA- und ARMADS-Modelle) und der nicht-linearen Zeitreihenanalyse (neuronale Netze bzw. ARNN-Modelle, siehe Kap. 4) auf zwei ausgewählte Zeitreihen angewendet: Arbeitslosenrate und Industrieproduktionsindex. Beide Zeitreihen können als typische Beispiele für univariate Zeitreihenanalyse angesehen werden. Sie wurden nicht nur in zahlreichen angewandten Arbeiten untersucht, sondern auch in



(a) ALR, x_t

(b) ALR, $\ln x_t - \ln(100 - x_t) + \ln 100$

(c) IPI, x_t

(d) IPI, $\ln x_t$

Abb. 1: Österreichische Arbeitslosenrate und Industrieproduktionsindex (1960/1-1997/12), jeweils Originalzeitreihe (in Prozent bzw. mit 1990=100) und logarithmisch bzw. logistisch transformierte Zeitreihe

Lehrbüchern als Beispiele herangezogen (z.B. Mills, 1990; Brockwell und Davis, 1991; Franses, 1996a).

Konkret handelt es sich um die österreichische Arbeitslosenrate (in der Folge kurz ALR) und den österreichischen Industrieproduktionsindex (IPI), jeweils in monatlichen Werten von 1960/1 – 1997/12 (Quelle: OECD). Die Zeitreihen wurden in saisonal nicht angepasster Form verwendet, da es prinzipiell von Interesse ist, wie die angewandten Prognoseinstrumente, insbesondere die neuronalen Netze, mit Saisonalitäten in den Daten zurecht kommen. Außerdem wurden in der Forschung der letzten Jahre vermehrt Hinweise gefunden, dass die von der amtlichen Statistik angewandten Verfahren zur Saisonanpassung (X11-Verfahren) bestehende lineare und vor allem nicht-lineare Muster rausglätten könnten. So liefern etwa Franses und De Bruin (2000) empirische Hinweise, dass das saisonale Muster vom Konjunkturzyklus abhängen könnte.

In Abb. 1 sind für beide Zeitreihen jeweils die Zeitreihenplots für die Originalzeitreihe und die logistisch bzw. logarithmisch transformierte Zeitreihe wiedergegeben. Die Wahl einer geeigneten Transformation steht an erster Stelle in der Modellierung von Zeitreihen. Das Ziel ist, zusammen mit anderen Verfahren wie Trendbereinigung und Differenzieren, die Stationarität der Zeitreihe herbeizuführen. Für die Wahl der geeigneten Transformation sollte sowohl a priori-Wissen über die Zeitreihe als auch eine empirische Auswertung herangezogen werden, wobei üblicherweise eine rein visuelle Auswertung der Zeitreihen- bzw. Residuenplots als genügend angesehen wird.

Im Falle der Arbeitslosenrate wurde eine logistische Transformation, $y = \ln x - \ln(100 - x) + \ln 100$, herangezogen. Diese ist deswegen angebracht, weil die Arbeitslosenrate auf das Intervall $[0, 100]$ beschränkt ist (vgl. Wallis, 1987). Außerdem bewirkt diese Transformation eine fast vollständige Bereinigung der Zeitreihe um saisonale Heteroskedastizität, wie ein Vergleich der Diagramme (a) und (b) in Abb. 1 zeigt. Während für die Originalzeitreihe die Werte für die saisonalen Täler (die Sommermonate) kaum statistische Variation aufweisen, treten in den saisonalen Spitzen (Wintermonate) beträchtliche Schwankungen auf. Wie sich bei Modellierungsversuchen anhand der Originalzeitreihe gezeigt hat, bewirkt das, dass Modelle vor allem die Varianz in den Spitzenmonaten erklären bzw. vorhersagen werden.[5] Allerdings hat der Gebrauch der logistischen Transformation auch Nachteile, wie etwa, dass Heteroskedastizität trivialen Typs eingeführt wird. Die Varianz der Fehlers linearer Modelle nimmt mit dem Zeitverlauf in geringem aber kontinuierlichem Maße ab. In der vorliegenden Arbeit wurde dieser „Nachteil“ in Kauf genommen und vor allem nicht als Indiz für eine unrichtige Transformation der Daten gewertet. Im Gegenteil, es erscheint ökonomisch sehr plausibel, dass der Arbeitsmarkt heute im Vergleich zu früher besser funktioniert und sich weniger sprunghaft entwickelt (wenn auch in Verbindung mit einer höheren Arbeitslosenrate). Entsprechendes gilt auch für die im Zeitverlauf leicht, aber kontinuierlich abnehmende Stärke der saisonalen Bewegungen. Die Originalzeitreihe, weist im Gegensatz dazu eine sehr unregelmäßige Entwicklung der Stärke saisonaler Muster auf. Zu ergänzen ist, dass die logistische Transformation numerisch sehr nahe an die logarithmische Transformation herankommt, da die österreichische Arbeitslosenrate kaum Werte über 9 % erreicht.[6]

Im Falle des Industrieproduktionsindex wurde die Zeitreihe einer logarithmischen Transformation unterzogen. Diese Modellierungsentscheidung ist für vergleichbare Daten allgemein üblich und ökonomisch sinnvoll, da makroökonomische Modelle üblicherweise die Wachstumsrate als stationär annehmen (die Differenzen der

[5] Für eine Untersuchung der nicht-linearen Abhängigkeiten in den Residuen von auf der Originalzeitreihe ALR-Zeitreihe geschätzten linearen Modellen vgl. Koller und Fischer (2001). Diese Arbeit kommt zu sehr ähnlichen Ergebnissen wie die in Kap. 2 und 3 präsentierten, insbesondere kann das Vorhandensein additiver Nicht-Linearität nachgewiesen werden.

[6] Die Addition von $\ln 100$ im Rahmen der Transformation soll die numerische Approximation der ln-Transformation sicherstellen und den Vergleich der Koeffizienten der geschätzten Modelle mit denen aus Modellen, welche die ln-Transformation verwenden, erleichtern.

logarithmierten Zeitreihe stellen eine sehr gute Näherung der Wachstumsrate dar). Die logarithmische Transformation bewirkt eine Bereinigung um die Heteroskedastizität und die von der Zeit abhängige Breite der saisonalen Muster in IPI.

Im Rahmen dieser Arbeit beziehen sich ab nun Aussagen, insbesondere Analysen und Modelle, immer auf die logistisch bzw. logarithmisch transformierten Variablen. Anderenfalls wird ausdrücklich darauf hingewiesen.

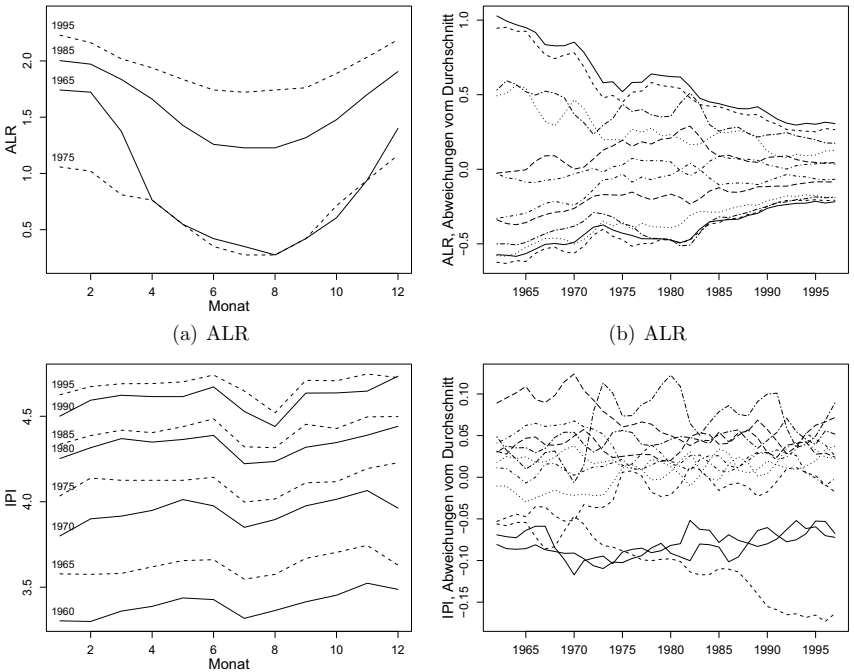
Im Vergleich der beiden Zeitreihen zeigen sich bedeutende Unterschiede im nichtstationären Verhalten. Während in der ALR kein Trend erkennbar ist, weist der IPI offenbar einen linearen Trend auf. Auch die Art der saisonalen Muster wirkt visuell sehr verschieden. Die Saisonalität als wichtiger Aspekt der beiden Zeitreihen kann mithilfe besonderer Analysemethoden untersucht werden. Es sind dies Buys-Ballot-Plots und saisonale Zeitreihenplots, die für die beiden Beispielzeitreihen in Abb. 2 zusammengefasst wurden.

Buys-Ballot-Plots sind eine Darstellungsform einer Zeitreihe, bei der die Zeitreihe gegen die Saison, z.B. Monat, in einem Diagramm aufgetragen werden, wobei jeweils die Datenpunkte einer Periode. z.B. eines Jahres, durch eine Linie verbunden werden. Diese Darstellungsform verdeutlicht das saisonale Muster einer Zeitreihe, das ja aus den gewöhnlichen Zeitreihenplots oft nur sehr schwer herauszulesen ist. In umgekehrter Weise wird in saisonalen Zeitreihenplots die Zeitreihe für jede Saison getrennt in ein Zeitreihendiagramm eingetragen. Von saisonalen Zeitreihenplots existieren verschiedene Varianten, je nachdem ob um Periodenmittelwerte und Trends bereinigt wird und ob gleitende Durchschnitte verwendet werden. In der vorliegenden Arbeit wurde ein Verfahren verwendet, das zunächst die Zeitreihe um einen linearen Trend bereinigt, dann für jede Monatszeitreihe getrennt gleitende 3-Jahres-Durchschnitte (also Durchschnitte von jeweils drei Werten) bildet und dann für jedes Jahr die Werte des Zeitreihenbündels auf den Durchschnitt 0 normiert. Gemeinsam erlauben Buys-Ballots-Plots und saisonale Zeitreihenplots die Beurteilung der Frage nach der Stabilität saisonaler Muster.[7]

Im Falle der ALR scheint das saisonale Muster sehr stabil zu sein, da die Arbeitslosigkeit immer in den Sommermonaten ein Tal und in den Wintermonaten eine Spitze aufweist und auch die Rolle der anderen Monate innerhalb des saisonalen Zyklus gleich bleibt. Allerdings wird, wie bereits zuvor bemerkt, die Stärke des saisonalen Musters im Zeitverlauf geringer.

Auch die IPI hat ein stabiles Muster, wie vor allem der entsprechende Buys-Ballot-Plot zeigt. Bei der Interpretation des saisonalen Zeitreihenplots des IPI muss man berücksichtigen, dass saisonale Schwankungen im Vergleich zu anderen Komponenten der Zeitreihe im IPI eine geringere Rolle spielen, weshalb es zu "zufälligen"

[7] Die Bezeichnung Buys-Ballot-Plot geht auf den holländischen Meteorologen Christoph H. D. Buys-Ballot (1817 – 1890) zurück. Ökonometrische Arbeiten, die Buys-Ballot-Plots verwenden sind beispielsweise Hylleberg (1992) und Sorensen (1999). Saisonale Zeitreihenplots finden sich unter anderem bei Hylleberg (1992), Franses (1996b), Clements und Hendry (1997), und Sorensen (1999).



(a) ALR (b) ALR

(c) IPI (d) IPI

Abb. 2: Buys-Ballot-Plots und saisonale Zeitreihenplots, Arbeitslosenrate (ALR) und Industrieproduktionsindex (IPI)

Überkreuzungen der Zeitreihen der einzelnen Monate kommt. Allerdings ist das Ausscheren einzelner Monate klar zu erkennen, etwa des Monats August, für den im Vergleich zur Entwicklung in anderen Monaten die Industrieproduktion im betrachteten Zeitraum abgenommen hat.

Die Frage nach der Stabiltität saisonaler Muster hat auch für die Frage nach der saisonalen Unit-Root Bedeutung, die unter anderem im nächsten Abschnitt ausführlich behandelt wird.

2.3 Nicht-Stationarität und Herbeiführung der Stationarität

Die Bedeutung der Frage nach der Stationarität bzw. den Unit-Roots von Zeitreihen wurde bereits unter 2.1 besprochen. Ob eine konkrete Zeitreihe stationär ist und auf welche Weise man eine nicht-stationäre Zeitreihe in eine stationäre Zeitreihe

überführt, ist von hoher praktischer Relevanz und wurde insbesondere für ökonomische Zeitreihen intensiv untersucht.

In der folgenden Diskussion wird zunächst das Problem der Unterscheidung eines stochastischen von einem deterministischen Trend besprochen. Danach geht es um den Aspekt der Saisonalität und der Berücksichtigung der Möglichkeit einer saisonalen Unit-Root. Um die Frage der korrekten Überführung in eine stationäre Zeitreihe für den konkreten Anwendungsfall der beiden Zeitreihen ALR und IPI zu beantworten, werden verschiedene Testverfahren angewendet, deren Funktionsprinzipien sowie Vor- und Nachteile in aller Kürze erläutert werden. Schließlich werden die Ergebnisse der Analyse zusammengefasst und in Hinblick auf die weitere Vorgehensweise bei der linearen und auch nicht-linearen Modellierung interpretiert.

Differenz-Stationarität versus Trend-Stationarität

Box und Jenkins (1970) schlagen die Anwendung des Differenzenoperators als Weg vor, um von einer nicht-stationären zu einer stationären Zeitreihe zu gelangen. Diese Idee gewann in der Ökonomie weite Anerkennung, nachdem Nelson und Plosser (1982) in ihrer Pionierarbeit starke Hinweise dafür präsentieren konnten, dass viele der wichtigsten makroökonomischen Zeitreihen einen stochastischen Trend (Unit-Root) aufweisen, den man durch einmalige Anwendung des Differenzenoperators beseitigen kann. Solche Prozesse werden als „differenz-stationär“ bezeichnet:

$$x_t = x_{t-1} + \beta + u_t \quad \text{bzw.} \quad \nabla x_t = \beta + u_t, \tag{2.22}$$

wobei u_t ein stationärer Prozess mit Mittelwert Null ist, der gegebenenfalls durch ein ARMA-Modell modelliert werden kann. Die Alternative hierzu ist, dass die Zeitreihe einen deterministischen Trend aufweist, der am besten durch Regression auf t geschätzt und um den durch einfache Subtraktion bereinigt werden kann. Der Trend wird im einfachsten Fall als linear angenommen. Es ergibt sich ein „trend-stationärer“ Prozess:

$$x_t = \alpha + \beta t + u_t, \tag{2.23}$$

Für $\beta = 0$ ergibt sich Stationarität als Spezialfall der Trend-Stationarität. Bis zur Veröffentlichung der Arbeiten von Nelson und Plosser galt im wesentlichen die Trend-Stationarität als allgemein akzeptierte Hypothese für makroökonomische Zeitreihen.

Im Falle der Trend-Stationarität verschwindet der Einfluss eines gegenwärtigen Schocks ε_t auf weit in der Zukunft liegende Werte der Zeitreihe, x_{t+k}, da die Zeitreihe zum deterministischen Trend zurückzukehren tendiert. Im Gegensatz dazu bleibt bei Differenzstationarität der Einfluss von ε_t auf zukünftige x_{t+k} bestehen und konvergiert bei $k \to \infty$ zu einem konstanten Faktor $c_\infty > 0$ (bei Trend-Stationarität ist $c_\infty = 0$). Aus diesem Umstand erklärt sich auch die große wirtschaftspolitische Bedeutung der Unterscheidung zwischen (2.22) und (2.23).

Für Prognosezwecke, die im Mittelpunkt der vorliegenden Arbeit stehen, ist eine korrekte Zuordnung einer konkreten Zeitreihe zu (2.22) oder (2.23) zwar weniger wichtig, da für einen kurzfristigen Zeithorizont differenz- und trendstationäre Prozesse sehr ähnliches Verhalten zeigen können. Jedoch konnte Rudebusch (1993) am Beispiel einer zentralen makroökonomischen Zeitreihe (reales US-amerikanisches BIP, Jahreswerte) zeigen, dass für mittelfristige Prognoseziele (drei bis fünf Jahre) die plausibelsten geschätzten Modelle nach (2.22) und (2.23) ökonomisch sehr wohl relevante Unterschiede liefern.

Es treten noch weitere Probleme auf, wenn ein trend-stationärer Prozess als differenz-stationärer Prozess modelliert wird oder umgekehrt (für einen Überblick vgl. Mills, 1990, S. 200ff). An dieser Stelle soll besonders auf die Gefahr des Über-Differenzierens hingewiesen werden. Wird auf eine Zeitreihe x_t, die bereits stationär ist, der Differenzenoperator angewandt, hat das unangenehme Auswirkungen auf die Eigenschaften der resultierenden Zeitreihe $u_t = \nabla x_t$. Einerseits ist die Varianz von u_t im Vergleich zu x_t höher, andererseits werden zusätzliche MA-Terme eingeführt, die außerdem die Nicht-Invertierbarkeit des MA-Teils zur Folge haben. Für den einfachen Fall eines AR(1)-Prozesses $x_t = \phi x_{t-1} + \varepsilon_t$ kann das sehr leicht gezeigt werden: $u_t = x_t - x_{t-1} = \phi(x_{t-1} - x_{t-2}) + \varepsilon_t - \varepsilon_{t-1} = \phi u_{t-1} + \varepsilon_t - \varepsilon_{t-1}$. Wird eine trend-stationäre Zeitreihe nach (2.23), $\beta > 0$, differenziert, ist zwar die Varianz der differenzierten Zeitreihe geringer, da um den Trend bereinigt wurde, doch das Problem der Einführung zusätzlicher MA-Terme ist das gleiche. Insbesondere wenn man sich bei den verwendeten linearen Modellen auf reine AR-Modelle beschränken will, sollte man daher mit der Diagnose einer Unit-Root vorsichtig sein. Aber auch im Falle der ARMA-Modellierung gestaltet sich für eine überdifferenzierte Zeitreihe die Modellselektion und -schätzung erheblich schwieriger.

Aus diesen Gründen wird den Möglichkeiten, die Stationarität bzw. das Vorhandensein einer Unit-Root einer Zeitreihe festzustellen, hier mehr Raum als üblich gewidmet und von in den Lehrbüchern oft vorzufindenden vereinfachenden Empfehlungen, wie z.B. zu differenzieren, wenn die ACF „langsam“ abnimmt, Abstand genommen.

Saisonale Unit-Roots

Es wurde bereits darauf hingewiesen, dass ökonomische Zeitreihen, die zu unterjährigen Zeitabständen erhoben werden, fast ausnahmslos saisonale Muster aufweisen und Saisonalitäten auch in den Beispielzeitreihen deutlich erkennbar sind. Die Behandlung von Saisonalitäten ist im Rahmen der Zeitreihen-Modellierung ein unerlässlicher Bestandteil. In Abschnitt 2.1 wurde die Anwendung des saisonalen Differenzenoperators ∇_s als eine Möglichkeit eingeführt, die saisonalen Muster zu berücksichtigen. Diese Vorgangsweise ist streng genommen jedoch nur dann zulässig, wenn die Zeitreihe eine saisonale Unit-Root aufweist:

$$x_t = x_{t-s} + \beta + u_t \quad \text{bzw.} \quad \nabla_s x_t = \beta + u_t, \tag{2.24}$$

wobei s die saisonale Frequenz ist (z.B. $s = 12$ für monatliche Daten) und u_t, wie zuvor in Gleichung (2.22), ein stationärer Prozess mit Mittelwert Null ist, der gegebenenfalls durch ein ARMA-Modell modelliert werden kann.

Saisonale Muster können durchaus auch auftreten, ohne dass Gleichung (2.24) gegeben ist. In diesem Falle müssten andere Verfahren der Saisonbereinigung angewandt werden, z.B. Einführung saisonaler Dummy-Variablen. Durch saisonales Differenzieren wird nicht nur eine eventuell vorhandene saisonale Unit-Root entfernt, sondern auch deterministische saisonale Muster. Dieser Effekt vereinfacht die Modellierung von saisonalen Zeitreihen, was die in angewandten Arbeiten bestehende Tendenz erklärt, saisonal zu differenzieren, auch wenn nicht klar ist, ob Gleichung (2.24) zutrifft. Verschiedene Arbeiten weisen jedoch in diesem Zusammenhang auf die Gefahr des Überdifferenzierens hin (z.B. Osborn, 1990).

Außerdem ist die Unterscheidung zwischen einer konventionellen und einer saisonalen Unit-Root ein wichtiges Thema, das in einigen Arbeiten untersucht wurde (z.B. Hylleberg et al., 1993; Franses, 1996b). Am einfachsten Beispiel eines I(1)-Prozesses, dem Random Walk $x_t = x_{t-1} + \varepsilon_t$, kann leicht gezeigt werden, dass durch saisonales Differenzieren mit Hilfe von ∇_s ein nicht invertierbarer MA($s-1$)-Prozess entsteht: $\nabla_s x_t = x_t - x_{t-s} = \varepsilon_t + \varepsilon_{t-1} + \ldots + \varepsilon_{t-s+1}$.

Umgekehrt wird eine saisonal integrierte Zeitreihe durch die Anwendung des normalen Differenzenoperators nur von der konventionellen Unit-Root befreit, nicht jedoch von den Unit-Roots zu den saisonalen Frequenzen. Dies wird ersichtlich, wenn man das dem saisonalen Differenzenoperator entsprechende Lag-Polynom faktorisiert, z.B. für den Fall einer monatlich beobachteten Zeitreihe:

$$\begin{aligned}
\nabla_s = 1 - B^{12} &= (1-B)(1+B)(1+B^2)(1+B^4+B^8) \qquad (2.25)\\
&= (1-B)(1+B)(1-iB)(1+iB)\\
&\quad \times[1+(\sqrt{3}+i)B/2][1+(\sqrt{3}-i)B/2]\\
&\quad \times[1-(\sqrt{3}+i)B/2][1-(\sqrt{3}-i)B/2]\\
&\quad \times[1+(i\sqrt{3}+1)B/2][1-(i\sqrt{3}-1)B/2]\\
&\quad \times[1-(i\sqrt{3}+1)B/2][1+(i\sqrt{3}-1)B/2],
\end{aligned}$$

wobei in dieser Gleichung i die imaginäre Zahl i bezeichnet. Von den Wurzeln dieser Gleichung ist die erste, Eins, die gewöhnliche nicht-saisonale Unit-Root, während die anderen elf Unit-Roots zu verschiedenen saisonalen Frequenzen darstellen und unterjährigen Zyklen unterschiedlicher Länge entsprechen.

Die Feststellung einer saisonalen Unit-Root ist gerade für makroökonomische Zeitreihen von hoher praktischer Relevanz. Osborn (1990) kommt anhand einer Untersuchung von 30 britischen makroökonomischen Zeitreihen zum Schluss, dass die Mehrzahl *keine* saisonale Unit-Root aufweist. Daher soll die Hypothese, dass die beiden Beispielzeitreihen ALR und IPI eine saisonale Unit-Root aufweisen, später einer Überprüfung unterzogen werden.

Tests auf Unit-Root bzw. Stationarität

Zur Feststellung der Stationarität einer Zeitreihe bzw. einer konventionellen oder saisonalen Unit-Root gibt es verschiedene Verfahren. Bei der Feststellung der richtigen Ordnung des Differenzierens kombiniert man diese Verfahren und wendet sie nicht nur auf die Originalzeitreihe, sondern auch auf die differenzierten Zeitreihen an. Die visuelle Analyse des Zeitreihen-Plots oder der ACF kann als erste Orientierungshilfe herangezogen werden.

Eine weitere brauchbare Heuristik steht in der Methode der Variaten Differenzen zur Verfügung, die auf dem Umstand beruht, dass sowohl eine integrierte Zeitreihe bzw. eine Zeitreihe mit deterministischen Komponenten als auch eine überdifferenzierte Zeitreihe eine höhere Varianz aufweisen, als die „richtig“ differenzierte, stationäre Zeitreihe. Eine Erweiterung der Methode der Variaten Differenzen erlaubt auf dem gleichen Prinzip basierend auch die Unterscheidung zwischen saisonalen und konventionellen Unit-Roots, wobei bestimmte Einschränkungen angebracht sind. Nach der Methode der Variaten Differenzen ist jene Kombination (d, D) der Anwendung des Differenzenoperators $\nabla^d \nabla_s^D$ die richtige, für die die Zeitreihe die kleinste Varianz hat (vgl. Schlittgen und Streitberg, 1997, S. 297f).

Am sichersten ist die richtige Art der Herbeiführung der Stationarität jedoch nur durch die systematische Anwendung mehrerer geeigneter statistischer Tests festzustellen. In den letzten Jahren wurden eine Vielzahl von Testverfahren vorgeschlagen, und es liegt nunmehr eine umfangreiche ökonometrische Literatur zu den Eigenschaften der Tests sowie zu angewandten Untersuchungen vor (für einen Überblick siehe Phillips und Xiao, 1998).

Die vorliegende Untersuchung basiert auf einer Auswahl von fünf mittlerweile als klassisch anzusehenden Tests:

- den Augmented-Dickey-Fuller-Test (ADF-Test),
- den Phillips-Perron-Test (PP-Test),
- den Kwiatkowski-Phillips-Schmidt-Shin-Test (KPSS-Test)
- den Dickey-Hasza-Fuller-Test (DHF-Test)
- und den Hylleberg-Engle-Granger-Yoo-Test (HEGY-Test).

ADF-Test und PP-Test haben die Annahme einer Unit-Root als Null-Hypothese, der KPSS-Test die Annahme der Stationarität und DHF-Test und HEGY-Test die Annahme einer saisonalen Unit-Root.

Sowohl der ADF-Test wie auch der PP-Test stellen Erweiterungen bzw. Modifikationen des Dickey-Fuller-Tests (DF-Test) dar, der in einer seiner Varianten auf der Schätzgleichung

$$x_t = \alpha_0 + \alpha_1 x_{t-1} + \delta t + \varepsilon_t \tag{2.26}$$

beruht (Dickey und Fuller, 1979). Im Rahmen des Testverfahrens wird die sehr restriktive Annahme getroffen, dass die Residuen ε_t in Gleichung (2.26) i.i.d $(0, \sigma^2)$ sind. Der ADF-Test lockert diese Restriktion auf parametrischem Wege, der PP-Test auf nicht-parametrischem Wege.

Der ADF-Test (Said und Dickey, 1984) basiert auf der Schätzgleichung

$$\nabla x_t = \alpha_0 + \alpha_1 x_{t-1} + \sum_{j=1}^{p} \beta_j \nabla x_{t-j} + \delta t + \varepsilon_t \tag{2.27}$$

Unter der Null-Hypothese einer Unit-Root gilt $\alpha_1 = 0$. Die Alternativ-Hypothese ist $\alpha_1 < 0$. Die Teststatistik ist der „t-Wert“ des geschätzten Koeffizienten α_1. Anstatt einer t-Verteilung folgt dieser der Dickey-Fuller-Verteilung, deren kritische Werte z.B. bei Cromwell et al. (1994) tabelliert sind. Die Aufnahme von p verzögerten abhängigen Variablen, ∇x_{t-j}, als Regressoren in die Gleichung (2.27) dient der Berücksichtigung der Autokorrelation der ε_t der Gleichung (2.26). Die Größe p sollte empirisch gewählt werden, wobei als Richtlinie $p = N^{1/3}$ gilt (N ist die Länge der Zeitreihe).

Im Gegensatz dazu wendet der PP-Test (Phillips, 1987; Phillips und Perron, 1988; Perron, 1988) eine nicht-parametrische Strategie an, um die Korrelation der ε_t zu berücksichtigen. Die zugrunde liegende Schätzgleichung ist Gleichung (2.26). Unter der Null-Hypothese einer Unit-Root gilt $\alpha_1 = 1$. Um zur Teststatistik zu gelangen wird eine hier nicht näher beschriebene Transformation des t-Werts von $(\hat{\alpha_1} - 1)$ vorgenommen, die die Korrelationsstruktur der Residuen berücksichtigt. Die Teststatistik folgt der Dickey-Fuller-Verteilung. Der PP-Test hat gegenüber dem ADF-Test den Vorteil, dass für die gemeinsame Verteilung der ϵ_t bedeutend weniger strenge Annahmen getroffen werden müssen. Sie müssen etwa einem stark mischendem Prozess folgen, was für die meisten stationären Prozesse zutrifft. Für die vorliegende Arbeit ist insbesondere wichtig, dass gezeigt werden kann, dass der Test auch angewendet werden kann, wenn die ε_t einem ARNN-Prozess folgen.

In der praktischen Anwendung hat der PP-Test in Bezug auf viele Alternativen eine hohe Macht.[8] Allerdings hat der PP-Test im Vergleich zum ADF-Test auch Schwächen. Beispielsweise kann das Testverfahren nur ungenügend um Saisonalitäten in den Residuen korrigieren, wodurch für saisonale Zeitreihen oft die Null-Hypothese abgelehnt wird, auch wenn sie zutrifft. Beide Tests ergänzen einander insoferne.

Statistische Tests sind so angelegt, dass sie die Null-Hypothese bei einem vorgegebenen Signifikanzniveau nur dann ablehnen, wenn starke Beweise zu ihrer Ablehnung vorliegen. An diesem Umstand schloss die Kritik an, mit Tests vom ADF- oder PP-Typ würde man für ökonomische Zeitreihen zu oft die Hypothese einer Unit-Root akzeptieren und zu selten ablehnen (z.B. Rudebusch, 1993). Eine interessante

[8] Der Zusammenhang zwischen Macht und Größe von Tests wird unter anderem in Abschnitt 3.2 genauer erläutert.

Ergänzung zu ADF-Test und PP-Test ist daher ein Test, der als Null-Hypothese die Annahme der Stationarität und als Alternativ-Hypothese die Annahme einer Unit-Root hat. Ein solcher Test, der KPSS-Test, wurde von Kwiatkowski et al. (1992) entwickelt.

Der KPSS-Test geht von folgendem Modell aus:

$$x_t = \delta t + z_t + \varepsilon_t, \quad \text{wobei} \quad z_t = z_{t-1} + u_t, \tag{2.28}$$

Man beachte, dass hierbei z_t ein Random-Walk ist. Unter der Null-Hypothese des Tests verschwindet die Varianz von u_t. Um diese Hypothese zu testen wird die Hilfsregression

$$x_t = \alpha + \delta t + \varepsilon_t \tag{2.29}$$

durchgeführt. Die Teststatistik ist als $\sum_{t=1}^{T} S_t^2/\hat{\sigma}_\varepsilon^2$ definiert, wobei $S_t = \sum_{i=1}^{t} \hat{\varepsilon}_i$ und $\hat{\sigma}_\varepsilon^2 = \mathrm{Var}(\varepsilon_t)$ ist. Diese Berechnungsformel ist jedoch nur für iid-verteilte ε_t zu verwenden. Trifft diese Annahme nicht zu, insbesondere wenn Autokorrelation vorliegt, wird eine Korrektur vorgenommen, die ähnlich funktioniert wie beim PP-Test.

Der DHF-Test und der HEGY-Test, die zur Überprüfung der Hypothese der saisonalen Unit-Root-Hypothese ausgewählt werden, sind beide einfache Modifikationen des ADF-Test. Der DHF-Test (vgl. Dickey et al., 1984, wo auch die kritischen Werte der Verteilung der Teststatistik tabelliert sind) testet die Hypothese einer saisonalen Unit-Root gegen die Alternativhypothese, dass keine Unit-Root vorhanden ist. In einer der üblichen Versionen des Tests lautet die Schätzgleichung:

$$\nabla_s x_t = \alpha_1 x_{t-s} + \sum_{j=1}^{p} \beta_j \nabla_s x_{t-j} + \sum_{j=1}^{s} \gamma_j s_{jt} + \delta t + \varepsilon_t, \tag{2.30}$$

wobei die saisonalen Dummyvariablen s_{jt} den Wert Eins annehmen, wenn die Beobachtung t der Saison j zu zuordnen ist, ansonsten den Wert Null. Anhand des „t-Werts“ von α wird die Null-Hypothese $\alpha = 0$ gegen die Alternativ-Hypothese $\alpha < 0$ getestet.

Der DHF-Test erlaubt nicht den Test der Hypothese einer saisonalen Unit-Root gegen die Alternativ-Hypothese einer konventionellen Unit-Root. Diese Lücke füllt der HEGY-Test (Hylleberg et al., 1990).[9] Streng genommen ist nur dann, wenn sämtliche Wurzeln des charakteristischen Polynoms in Gleichung (2.26) auf dem Einheitskreis liegen, die Anwendung des saisonalen Differenzenfilters berechtigt. Die

[9] Hylleberg et al. (1990) entwickelten ein Testverfahren für vierteljährliche Daten. Dieses Konzept wurde von Franses (1991) bzw. von Beaulieu und Miron (1993) auf monatliche Daten übertragen. In der vorliegenden Arbeit wird der HEGY-Test für monatliche Zeitreihen nach dem Verfahren von Beaulieu und Miron (1993) angewandt, das sich leicht von dem von Franses (1991) unterscheidet.

Schätzgleichung, die dem HEGY-Test für monatliche Zeitreihen zugrundeliegt, lautet

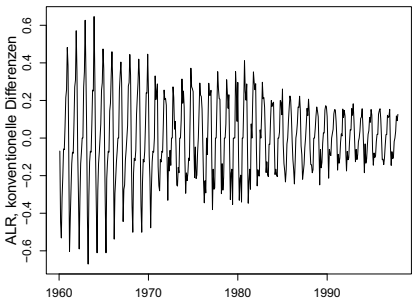
$$\nabla_s x_t = \sum_{j=1}^{12} \alpha_j y_{j,t-1} + \sum_{j=1}^{p} \beta_j \nabla_s x_{t-j} + \sum_{j=1}^{12} \gamma_j s_{jt} + \delta t + \varepsilon_t, \tag{2.31}$$

wobei die Variablen $y_{j,t}, j = 1, \ldots, 12$ gebildet werden, indem man ein Lag-Polynom, das bestimmte, aber nicht alle Faktoren der Faktorisierung von $(1-B^{12})$ gemäß Gleichung (2.26) enthält, auf x_t anwendet. Beaulieu und Miron (1993) führen zusätzlich durch die Konstruktion der $y_{j,t}$ deren paarweise Orthogonalität herbei und geben eine detaillierte Aufstellung für die Definition der $y_{j,t}$. Für die Interpretation des Tests ist es wichtig, dass jedes $y_{j,t}$ mit einer Frequenz assoziert werden kann. So ist $y_{j,t}$ mit j als ein Element aus den Mengen $\{1\}, \{2\}, \{3,4\}, \{5,6\}, \{7,8\}, \{9,10\}$ und $\{11,12\}$ den Frequenzen 0, π, $\pi/2$, $2\pi/3$, $\pi/3$, $5\pi/6$, bzw. $\pi/6$ zuzuordnen. Um zu zeigen, dass bei keiner der saisonalen Frequenzen eine Unit-Root vorhanden ist, muss α_j für $j = 2$ und für jeweils mindestens ein Element der Mengen $\{3,4\}, \{5,6\}, \{7,8\}, \{9,10\}$ von Null verschieden sein. Dementsprechend liefert der HEGY-Test fünf Teststatistiken, die t-Statistik für den Koeffizienten α_2 und vier F-Statistiken für die gemeinsame Signifikanz der Koeffizienten α_3 und α_4, bzw. α_5 und α_6 usw. Außerdem wird üblicherweise auch der t-Wert für α_1 angegeben, der als Test für die konventionelle Unit-Root angesehen werden kann.

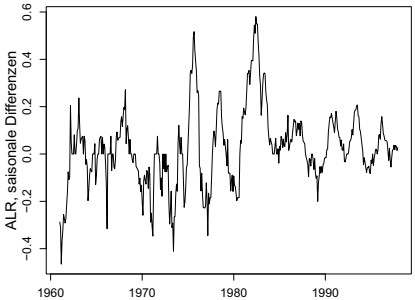
Ergebnisse für die beiden Beispielzeitreihen

Um zunächst eine visuelle Auswertung der Zeitreihen zu ermöglichen sind Zeitreihenplots für ∇x_t, $\nabla_s x_t$ und $\nabla_s \nabla x_t$ jeweils für die ALR und den IPI in Abb. 3 zusammengefasst. Für weitere Versionen differenzierter Zeitreihen wurden aus Platzgründen keine Abbildungen aufgenommen, da – wie die spätere Analyse bestätigt – öfteres Differenzieren bereits überdifferenzierte Zeitreihen ergeben würde. Eine visuelle Auswertung bezieht auch die ACF mit ein. Diese ist weiter unten in Abb. 5 enthalten.

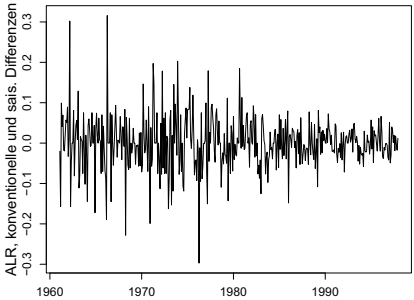
Tabelle 1 enthält die Ergebnisse der Methode der variaten Differenzen. Im Rahmen dieser Methode wird die Varianz der Ausgangszeitreihe x_t mit der Varianz der differenzierten Zeitreihen $\nabla_s^D \nabla^d x_t$ verglichen, indem für ein „Gitter“ von Kombinationen (d, D) die Verhältniszahlen $\mathrm{Var}(\nabla_s^D \nabla^d x_t)/\mathrm{Var}(x_t)$ gebildet werden. Als Ausgangszeitreihe wird hierbei nicht $x_t = \mathrm{ALR}$ bzw. $x_t = \mathrm{IPI}$ genommen sondern eine um Trend und saisonale Mittelwerte bereinigte Zeitreihe, $x_t - \sum_{j=1}^{12} \hat{\gamma}_j s_{jt} - \hat{\delta} t$, sowie eine um saisonale Trends und saisonale Mittelwerte bereinigte Zeitreihe, $x_t - \sum_{j=1}^{12} \hat{\gamma}_j s_{jt} - \sum_{j=1}^{12} \hat{\delta}_j s_{jt} t$, wobei mit $\hat{\gamma}$ und $\hat{\delta}$ OLS-Schätzer sowie mit s_{jt} saisonale Dummyvariable bezeichnet sind. Diese vorgeschaltete Bereinigung erhöht die Aussagekraft der Methode der variaten Differenzen, weil anderenfalls mindestens einmaliges konventionelles und einmaliges saisonales Differenzieren als empfehlenswert angezeigt werden würde, um deterministische Saisonalitäten und Trends zu entfernen.



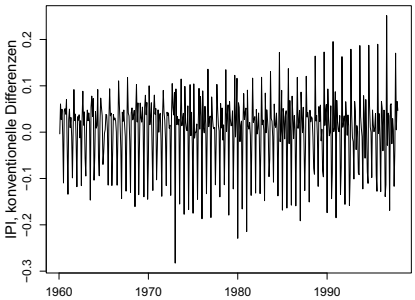
(a) ALR, ∇x_t



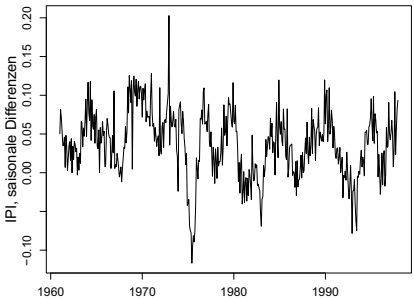
(b) ALR, $\nabla_s x_t$



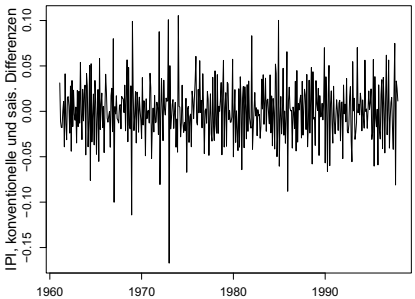
(c) ALR, $\nabla_s \nabla x_t$



(d) IPI, ∇x_t



(e) IPI, $\nabla_s x_t$



(f) IPI, $\nabla_s \nabla x_t$

Abb. 3: Zeitreihenplots für verschieden differenzierte Zeitreihen, Arbeitslosenrate (ALR) und Industrieproduktionsindex (IPI)

Tab. 1: Methode der variaten Differenzen, Arbeitslosenrate (ALR) und Industrieproduktionsindex (IPI)

	ALR				IPI			
Um Trend und saisonale Mittelwerte bereinigte Zeitreihe:								
	$D=0$	$D=1$	$D=2$	$D=3$	$D=0$	$D=1$	$D=2$	$D=3$
$d=0$	1.0000	0.2062	0.3516	0.9200	1.0000	0.2461	0.4656	1.3117
$d=1$	0.0720	0.0327	0.0871	0.2729	0.1766	0.1435	0.3431	1.0170
$d=2$	0.0842	0.0732	0.2052	0.6531	0.5368	0.4147	0.9723	2.8438
$d=3$	0.1999	0.2282	0.6418	2.0327	1.7820	1.3496	3.1496	9.1703
Um saisonale Trends und saisonale Mittelwerte bereinigte Zeitreihe:								
	$D=0$	$D=1$	$D=2$	$D=3$	$D=0$	$D=1$	$D=2$	$D=3$
$d=0$	1.0000	0.2349	0.4025	1.0532	1.0000	0.2535	0.4798	1.3517
$d=1$	0.0357	0.0371	0.0997	0.3123	0.1210	0.1476	0.3536	1.0480
$d=2$	0.0681	0.0837	0.2348	0.7477	0.3633	0.4266	1.0020	2.9306
$d=3$	0.1942	0.2611	0.7347	2.3268	1.2033	1.3884	3.2457	9.4503

Die Ergebnisse sind für beide Zeitreihen, ALR und IPI, ähnlich. In jedem Fall ist das Minimum in der zweiten Zeile zu finden, was der Empfehlung entspricht, genau einmal den konventionellen Differenzenfilter anzuwenden, $d = 1$. Was die Anwendung des saisonalen Differenzenfilters betrifft, so scheint die Methode der variaten Differenzen seine einmalige Anwendung, $D = 1$, zu favorisieren. Allerdings hängt dieses Ergebnis von der Art der vorgeschalteteten Bereinigung ab. Werden durch zusätzliche Berücksichtigung von saisonalen Trends die deterministischen Komponenten vollständiger herausgefiltert, wird für die IPI die Anwendung des saisonalen Differenzenfilters nicht mehr als notwendig angezeigt. Die Entscheidung für oder gegen die Anwendung des saisonalen Differenzenfilters fällt insgesamt recht knapp aus.

Tabelle 2 enthält die Ergebnisse der verschiedenen statistischen Testverfahren für die Zeitreihen ALR und IPI. Es wurden jeweils die nicht differenzierte Zeitreihe x_t, die konventionell differenzierte Zeitreihe, ∇x_t, die saisonal differenzierte Zeitreihe, $\nabla_s x_t$, sowie die saisonal und konventionell differenzierte Zeitreihe, $\nabla_s \nabla x_t$, untersucht. Die in der Tabelle angegebenen Lag-Parameter wurden unter der Berücksichtigung der Literaturempfehlungen gewählt, sämtliche Ergebnisse und Schlussfolgerungen sind in Hinblick auf andere plausible Lag-Parameter robust.

Die Ergebnisse in Tab. 2 sind sehr komplex und erfordern eine detaillierte Diskussion. Es können allerdings die Ergebnisse für ALR und IPI gemeinsam besprochen werden, da sie sich kaum unterscheiden.

Für die nicht differenzierte Zeitreihe, x_t, akzeptiert der ADF-Test die Hypothese einer Unit-Root, und damit übereinstimmend verwirft der KPSS-Test die Hypothese der Stationarität. Die Ablehnung der Unit-Root-Hypothese anhand des PP-Tests ist irreführend und kommt durch eine Schwäche des PP-Tests bei saisonalen Zeitreihen zustande. Die Hypothese einer saisonalen Unit-Root wird für x_t sowohl vom

Tab. 2: Ergebnisse der Tests auf Unit-Root bzw. Stationarität für ALR und IPI

		ALR		
Testverfahren	x_t	∇x_t	$\nabla_{12} x_t$	$\nabla_{12} \nabla x_t$
ADF, Lag=24	-1.77 (0.6763)	-4.29 (< 0.01)**	-2.98 (0.1619)	-6.67 (< 0.01)**
PP, Lag=17	-5.38 (< 0.01)**	-6.55 (< 0.01)**	-4.75 (< 0.01)**	-23.71 (< 0.01)**
KPSS, Lag=13	0.62 (< 0.01)**	0.06 (0.5133)	0.11 (0.1514)	0.02 (0.9727)
DHF, Lag=24	-6.93 (< 0.01)**	-6.14 (< 0.01)**	-16.61 (< 0.01)**	-15.74 (< 0.01)**
HEGY, Lag=13:				
$t(1)$	-1.75 (0.6874)	-4.28 (< 0.01)**	-2.88 (0.1525)	-6.82 (< 0.01)**
$t(2)$	-3.56 (< 0.01)**	-3.81 (< 0.01)**	-6.84 (< 0.01)**	-7.01 (< 0.01)**
$F(3,4)$	8.41 (0.0112)*	8.42 (0.0111)*	46.51 (< 0.01)**	44.49 (< 0.01)**
$F(5,6)$	4.89 (0.1578)	5.18 (0.1255)	28.98 (< 0.01)**	30.78 (< 0.01)**
$F(7,8)$	3.20 (0.3803)	2.99 (0.4256)	33.67 (< 0.01)**	28.71 (< 0.01)**
$F(9,10)$	10.50 (< 0.01)**	11.70 (< 0.01)**	41.56 (< 0.01)**	41.31 (< 0.01)**
$F(11,12)$	1.74 (0.6919)	1.49 (0.7423)	28.96 (< 0.01)**	26.06 (< 0.01)**
		IPI		
Testverfahren	x_t	∇x_t	$\nabla_{12} x_t$	$\nabla_{12} \nabla x_t$
ADF, Lag=24	-1.80 (0.6607)	-4.24 (< 0.01)**	-4.56 (< 0.01)**	-5.59 (< 0.01)**
PP, Lag=17	-11.16 (< 0.01)**	-57.53 (< 0.01)**	-10.70 (< 0.01)**	-41.99 (< 0.01)**
KPSS, Lag=13	0.66 (< 0.01)**	0.03 (0.8192)	0.07 (0.3325)	0.02 (0.9395)
DHF, Lag=24	-7.17 (< 0.01)**	-7.33 (< 0.01)**	-16.98 (< 0.01)**	-18.55 (< 0.01)**
HEGY, Lag=13:				
$t(1)$	-1.86 (0.6268)	-4.17 (< 0.01)**	-4.50 (< 0.01)**	-5.54 (< 0.01)**
$t(2)$	-2.24 (0.1944)	-2.37 (0.1513)	-5.25 (< 0.01)**	-4.29 (< 0.01)**
$F(3,4)$	8.66 (< 0.01)**	7.45 (0.0234)*	29.93 (< 0.01)**	36.52 (< 0.01)**
$F(5,6)$	7.25 (0.0272)*	6.29 (0.0562)	36.19 (< 0.01)**	43.24 (< 0.01)**
$F(7,8)$	13.05 (< 0.01)**	11.02 (< 0.01)**	34.09 (< 0.01)**	39.13 (< 0.01)**
$F(9,10)$	5.21 (0.1232)	5.49 (0.0963)	45.10 (< 0.01)**	35.90 (< 0.01)**
$F(11,12)$	6.36 (0.0527)	5.53 (0.0944)	32.41 (< 0.01)**	35.91 (< 0.01)**

** und * bezeichnen Werte, die bei einem Konfidenzniveau von 1 % bzw. 5 % signifikant sind. Die Berechnung der Wahrscheinlichkeitswerte beruht auf den Tabellen für kritische Werte, die sich bei Banerjee et al. (1993), Kwiatkowski et al. (1992), Dickey et al. (1984) und Beaulieu und Miron (1993) finden und durch eigenene Monte-Carlo-Simulationen ergänzt wurden. Da für dazwischenliegende Werte interpoliert wurde, sind genaue Wahrscheinlichkeitswerte unterhalb von 0.01 und oberhalb von 0.99 nicht verfügbar.

DHF-Test als auch vom HEGY-Test verworfen. Das Ergebnis des DHF-Tests verwundert ein wenig, unter der Alternativ-Hypothese des Tests gar keine Unit-Root existiert, was sich mit den Ergebnissen des ADF-Tests und des KPSS-Tests nicht gut vereinbaren lässt.[10] Der HEGY-Test verwirft seine Null-Hypothese, indem die Hypothese von Unit-Roots für bestimmte saisonale Frequenzen verworfen wird, wobei diese für ALR und IPI verschieden sind. Allerdings liefert der HEGY-Tests (vgl. die $t(1)$-Statistik des HEGY-Tests) auch zusätzliche Unterstützung der Hypothese einer gewöhnlichen Unit-Root in x_t.

Für die konventionell differenzierte Zeitreihe, ∇x_t, verwerfen alle Unit-Root-Test ihre Null-Hypothese, und damit übereinstimmend akzeptiert der KPSS-Test die Hypothese der Stationarität. Das Ergebnis mit dem HEGY-Test unterscheidet sich bis auf die $t(1)$-Statistik, die nun signifikant verworfen wird, kaum von dem für die nicht differenzierte Zeitreihe erhaltenem.

Auch die saisonal differenzierte Zeitreihe, $\nabla_s x_t$, hat den Tests gemäß (mit Ausnahme des ADF-Test-Ergebnisses für ALR, das eventuell wegen mangelnder Macht des ADF-Tests zustandekommt) keine konventionelle oder saisonale Unit-Root bzw. ist stationär. Der HEGY-Test kann hier die Null-Hypothese nicht nur einer saisonalen Unit-Root insgesamt, sondern auch die Unit-Root-Hypothese für jede einzelne saisonale Frequenz ablehnen.

Auf der saisonal und konventionell differenzierten Zeitreihe, $\nabla_s \nabla x_t$, liefern sämtliche Tests das erwartete Ergebnis. Für die Ablehnung der Hypothese einer konventionellen oder saisonalen Unit-Root erlauben die Test-Statistiken eine sehr hohe Konfidenz.

Für die Beantwortung der Frage nach der richtigen Ordnung des Differenzierens, (d, D), erlauben die vorliegenden Resultate keine eindeutigen Schlussfolgerungen, insbesondere wenn man auch die Ergebnisse der Methode der variaten Differenzen hinzuzieht. Es besteht kein Zweifel, dass die Zeitreihe mindestens einmal differenziert werden muss. Wenn man sich dafür entscheidet, nur einmal zu differenzieren, ist unklar, ob dem konventionellen oder saisonalen Differenzenfilter der Vorzug gegeben werden soll. Da in beiden Zeitreihen, ALR und IPI, Unit-Roots zu bestimmten aber nicht allen saisonalen Frequenzen vorhanden sind, impliziert die Anwendung des ∇-Operators, dass Unit-Roots zu einigen saisonalen Frequenzen nicht aus den Daten gefiltert werden, während die Anwendung des ∇_s-Operators einen MA-Prozess einführt, der für bestimmte andere saisonale Frequenzen nicht invertierbar ist.[11] Auch für die Anwendung des $\nabla_s \nabla$-Operators spricht einiges, insbesondere im Falle der ALR.

[10] Die Hypothese einer Unit-Root, sei diese konventionell oder saisonal, kann auch geprüft werden, indem die Jahreszeitreihe auf Unit-Root getestet wird. Nimmt man von ALR bzw. IPI nur die Jännerwerte, so lässt sich für die so erzeugte Zeitreihe klar eine Unit-Root diagnostizieren. Dieses Ergebnis entkräftet die unerwarteten Ergebnisse mit dem PP-Test und DHF-Test und stützt zusätzlich die Hypothese der Unit-Root in der nicht differenzierten Zeitreihe.

[11] Die Anwendung eines Differenzenfilters, der nur um Unit-Roots zu bestimmten saisonalen Frequenzen bereinigt, wäre hier möglicherweise eine interessante Neuerung.

In dieser Situation erscheint es als das sinnvollste, sowohl für ∇x_t als auch für $\nabla_s x_t$ geeignete lineare Modelle zu entwickeln. Für $\nabla_s \nabla x_t$ werden hier jedoch aus Platzgründen keine Modelle präsentiert. Dennoch kommen sowohl ∇x_t, $\nabla_s x_t$, als auch $\nabla_s \nabla x_t$ für die Modellbildung in Frage. Insbesondere wenn es um die Erstellung von Prognosen geht, kann erst die Auswertung der Prognosegüte anhand eines beiseite gehaltenen Teils der Zeitreihe („Out-of-Sample“) ein verlässliches Urteil sprechen. Beim systematischen Vergleich, welcher in Kapitel 5 durchgeführt wird, werden sowohl NN-Modelle als auch lineare Modelle auf allen drei auf verschiedene Art differenzierten Zeitreihen berücksichtigt und die Unterschiede in einer Sensitivitätsanalyse analysiert werden.

2.4 AR-, ARIMA-, SARMA- und ARDS-Modellierung von Zeitreihen

Die Modellbildung von Zeitreihen umfasst die Wahl einer bestimmten Modellklasse, die Modellselektion (bzw. -spezifikation), die Schätzung und die Residuenanalyse, wobei in dieser Abfolge Rückkoppelung vorgesehen ist. Für lineare Modellklassen können Schätzung und Residuenanalyse als weitgehend problemlos angesehen werden, nicht zuletzt aufgrund der leichten Verfügbarkeit und Verwendbarkeit von ökonometrischen Softwarepaketen.[12] Die Frage der Modellselektion ist hingegen anspruchsvoller und erfordert auch in rein angewandten Arbeiten ein großes Maß an Wissen und Erfahrung.

Als Teil der Modellselektion kann auch die Wahl einer Transformation der Zeitreihe (vgl. Abschnitt 2.2) gesehen werden. Die Wahl der Ordnung d im Rahmen des ARIMA(p, d, q)-Modells (vgl. Abschnitt 2.3) zählt ebenso zur Modellselektion, obwohl in diesem Abschnitt nicht mehr darauf eingegangen wird. In der Folge wird daher einfachkeitshalber von ARMA-Modellen (bzw. SARMA-Modellen usw.) anstatt von ARIMA-Modellen (bzw. SARIMA-Modellen) gesprochen.

Im Rahmen der Modellselektion ist zunächst die Entscheidung zwischen einem ausführlich und einem sparsam spezifizierten Modell zu treffen. Ein ausführlich spezifiziertes Modell bezieht alle Parameter eines Modells bis zur Ordnung des Modells in die Schätzung ein, während ein sparsames Modell mit möglichst wenig zu schätzenden Parametern auszukommen trachtet und die restlichen Parameter auf den Wert Null fixiert. Mit der Schätzung eines ausführlichen Modells ist die Gefahr verbunden, dass das Modell auch zufällige Muster in den Daten anpasst, was bei Prognosen zum Problem des „Overfitting“ führen kann (vgl. Abschnitt 4.2 für

[12] Leider stimmt diese Einschätzung nur eingeschränkt, da auch die am meisten verbreiteten Softwarepakete durchaus für das gleiche zu schätzende Modell unterschiedliche Ergebnisse liefern können. Unterschiede in den Details der Schätzverfahren können bei komplexen Modellen zu erheblichen Unterschieden in den Schätzergebnissen führen. Newbold et al. (1994) haben einen Artikel über die „Überraschungen“ verfasst, die man mit der üblichen Software zur Schätzung von ARIMA-Modellen erleben kann.

eine ausführlichere Diskussion des Problems). In dieser Arbeit werden prinzipiell sparsam spezifizierte Modelle untersucht. Für bestimmte Problemstellungen, z.B. die Erstellung einer um lineare Strukturen bereinigten Zeitreihe zur Durchführung von Tests auf Nicht-Linearität, wird jedoch ein ausführlich spezifiziertes Modell verwendet.

Jener Aspekt der Modellselektion, der im folgenden behandelt wird, betrifft die Wahl der Modellordnung und die Auswahl der zu schätzenden Parameter. Für AR-, ARMA-, SARMA- und ARMADS-Modelle sind unterschiedliche Strategien bei der Modellselektion möglich. Als Entscheidungshilfen werden üblicherweise die ACF und PACF, die Signifikanzniveaus der einzelnen geschätzten Koeffizienten sowie Informationskriterien (insbesondere das Akaike Informationskriterium, AIC, und das Bayes'sche Informationskriterium, BIC) herangezogen. Da verschiedene Informationskriterien auch im Zusammenhang mit neuronalen Netzen angewendet werden können und hierbei zusätzliche theoretische Aspekte wichtig sind, werden sie im Kapitel 4 genauer erläutert.

AR-Modellbildung

Für die Modellierung von Zeitreihen mit Hilfe eines reinen autoregressiven Modells empfiehlt sich eine stufenweise Vorgehensweise, die darauf beruht, zuerst ein ausführliches Modell der Ordnung n zu schätzen und von diesem Modell ausgehend durch Zurückschneiden („Pruning“) zu einem sparsamen Modell zu kommen, das man – unter der Annahme, dass die Zeitreihe tatsächlich von einem AR-Prozess generiert wurde, – als das „wahre“ Modell ansehen kann. Im konkreten wird in dieser Arbeit die folgende Vorgehensweise verwendet (dieser Ansatz weist große Ähnlichkeit mit dem von Brockwell und Davis (1991, S. 287ff) vorgeschlagenen Ansatz auf):

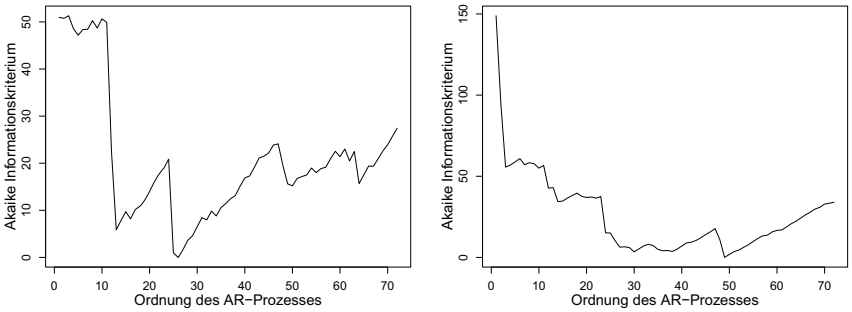
(i) Schätzung von N AR(n)-Modellen mit $n = 1, \ldots, N$, wobei N genügend groß gewählt wird

(ii) Auswahl jenes n^*, für das das AR(n^*)-Modell das kleinste AIC liefert

(iii) Elimination eines beliebigen Lags, dessen Koeffizient nicht signifikant von Null verschieden ist (z.B. bei einem Konfidenzniveau von 5 %) und neue Schätzung des vereinfachten Modells

(iv) Wiederholung des Schritts 3, bis das Modell nur mehr signifikante Koeffizienten enthält

Diese Vorgehensweise lässt noch Raum für Modifikationen im Einzelfall, z.B. bei der Festlegung der Reihenfolge der zu eliminierenden Lags. Bei den konkreten untersuchten Beispielzeitreihen ergeben sich jedoch keine Unklarheiten bei der Modellwahl. Genausowenig wären andere Modelle selektiert worden, hätte man statt



(a) ALR (b) IPI

Abb. 4: Akaike-Informationskriterium in Abhängigkeit von der Modellordnung eines AR-Prozesses: (a) Arbeitslosenrate (ALR) und (b) Industrieproduktionsindex (IPI)

der Signifikanz der Koeffizienten eine Verbesserung des BIC als Entscheidungskriterium in Schritt 3 der obigen Vorgangsweise gewählt.

Es soll außerdem an dieser Stelle darauf hingewiesen werden, dass die t-Werte, die man für das auf diese Weise optimierte Modell erhält, nach oben verzerrt sind, da nicht nur die Schätzung des AR-Modells selbst, sondern *die oben beschriebene Vorgehensweise insgesamt* von einem Fehler betroffen ist (Pötscher, 1991). Da eine Korrektur der Signifikanzen nur sehr schwer möglich und bisher in der Zeitreihenanalyseliteratur unüblich ist, werden hier die nicht korrigierten Schätzergebnisse berichtet. Die Problematik ist zudem in erster Linie für die Frage nach dem „wahren Modell“ und weniger für die Erstellung von Prognosen relevant.

Abbildung 4 illustriert die Veränderung des AIC mit wachsender Ordnung eines ausführlich spezifizierten AR-Modells. Es zeigt sich, dass für beide untersuchten Zeitreihen anhand des AIC eine relativ große Modellordnung gewählt wird, 26 für die Arbeitslosenrate und 49 für den Industrieproduktionsindex.

Tab. 3 und 4 enthalten die Schätzergebnisse des sparsam spezifizierten AR-Modells. Beide Prozesse sind – typisch für makroökonomische Zeitreihen – knapp an der Unit-Root: die größte Lösung des AR-Polynoms beträgt 0.94 für die Zeitreihe ALR und 0.97 für die Zeitreihe IPI. Die Residuenanalyse mit Hilfe der Q-Statistik weist in keinem der beiden Modelle auf Fehlspezifikation hin.

Die Ergebnisse sind auch inhaltlich plausibel. Die signifikanten Lags liegen jeweils nahe an den saisonalen Lags. Die hohe Ordnung des Modells deutet darauf hin, dass die Zeitreihen besser durch ein ARMA-Modell geschätzt werden sollten, da ein ARMA- oder MA-Prozess in der Praxis nur unzureichend durch einen reinen AR-Prozess angenähert werden kann.

Tab. 3: Schätzergebnisse für ein AR-Modell, ALR

Zeitreihe: $\nabla_{12} \ln ALR - \ln(100 - ALR) + \ln 100$
Zeitraum: 1961/1 – 1997/12 ($N = 444$)
Schätzmethode: Least Squares

Koeff.	Wert	Standardf.	Koeff.	Wert	Standardf.
ϕ_1	0.90876	0.04470 ***	ϕ_{12}	-0.46628	0.05177 ***
ϕ_2	0.14100	0.05144 **	ϕ_{13}	0.34362	0.05139 ***
ϕ_4	-0.10361	0.04234 *	ϕ_{16}	-0.06187	0.02897 *
ϕ_7	-0.10762	0.04163 **	ϕ_{24}	-0.19885	0.04373 ***
ϕ_9	0.10434	0.05050 *	ϕ_{25}	0.31554	0.05476 ***
ϕ_{10}	0.14095	0.05118 **	ϕ_{26}	-0.09889	0.04103 *

R^2adj.: 0.8779
F-Stat.: 262.9701*** (mit 12 und 406 Freiheitsgraden)
$\hat{\sigma}^2$: 0.00295
Q-Stat.(Lag=24): 21.3245
AIC: -1302.93

***, ** und * bezeichnen Werte, die bei einem Konfidenzniveau von 0.1 %, 1 % bzw. 5 % signifikant sind.

Tab. 4: Schätzergebnisse für ein AR-Modell, IPI

Zeitreihe: $\nabla_{12} \ln IPI$
Zeitraum: 1961/1 – 1997/12 ($N = 444$)
Schätzmethode: Least Squares

Koeff.	Wert	Standardf.	Koeff.	Wert	Standardf.
Interzept	0.00768	0.00246 **	ϕ_{24}	-0.37714	0.04529 ***
ϕ_1	0.37283	0.04027 ***	ϕ_{26}	0.15608	0.04306 ***
ϕ_2	0.25915	0.04269 ***	ϕ_{27}	0.18071	0.04637 ***
ϕ_3	0.33204	0.04426 ***	ϕ_{36}	-0.08613	0.02896 **
ϕ_{12}	-0.42558	0.04398 ***	ϕ_{48}	-0.17875	0.03821 ***
ϕ_{14}	0.18665	0.04196 ***	ϕ_{49}	0.18474	0.03810 ***
ϕ_{15}	0.18271	0.04680 ***			

R^2adj.: 0.6976
F-Stat.: 76.7574*** (mit 12 und 382 Freiheitsgraden)
$\hat{\sigma}^2$: 0.000572
Q-Stat.(Lag=36): 28.7625
AIC: -2029.20

***, ** und * bezeichnen Werte, die bei einem Konfidenzniveau von 0.1 %, 1 % bzw. 5 % signifikant sind.

ARMA-Modellbildung

Für die Identifikation von ARMA-Modellen ist keine ähnlich einfache Vorgangsweise wie die zuvor für die reinen AR-Modelle beschriebene verfügbar. Insbesondere erscheint eine gänzlich automatisierte Vorgangsweise inpraktikabel, obwohl Versuche in diese Richtung unternommen wurden. Beispielsweise schlagen Brockwell und Davis (1991, S. 296) vor, in einem ersten Schritt mit Hilfe von Informationskriterien die Ordnung p eines ausführlich spezifizierten ARMA(p, p)-Modells zu bestimmen und von hier, geleitet durch die Standardfehler der geschätzten Koeffizienten, durch Beschneiden des Modells jenes (sparsam spezifizierte) Modell zu finden, das das Informationskriterium minimiert. Die Inpraktikabilität dieser Vorgangsweise ist insbesondere für Prozesse, die einerseits eine hohe Ordnung p und andererseits eine relativ geringe Anzahl Parameter aufweisen, unmittelbar einzusehen. Der von Brockwell und Davis vorgeschlagene Ansatz verwendet außerdem das Informationskriterium als bloße Entscheidungsheuristik (vgl. die Ausführungen in Kapitel 4). Weiters weicht diese Vorgangsweise der korrekten Bestimmung der Modellordnung (p, q) aus und wird so der Komplexität der ARMA-Modelle grundsätzlich nicht gerecht. Hinzu kommt, dass die Fehlerfunktion eines ARMA-Modells eine stark nichtlineare Funktion der Parameter ist, sodass die Eigenschaften des Schätzverfahrens (Startwerte und Optimierungsalgorithmus) eine Rolle spielen und erschwerend auf die Modellselektion wirken.

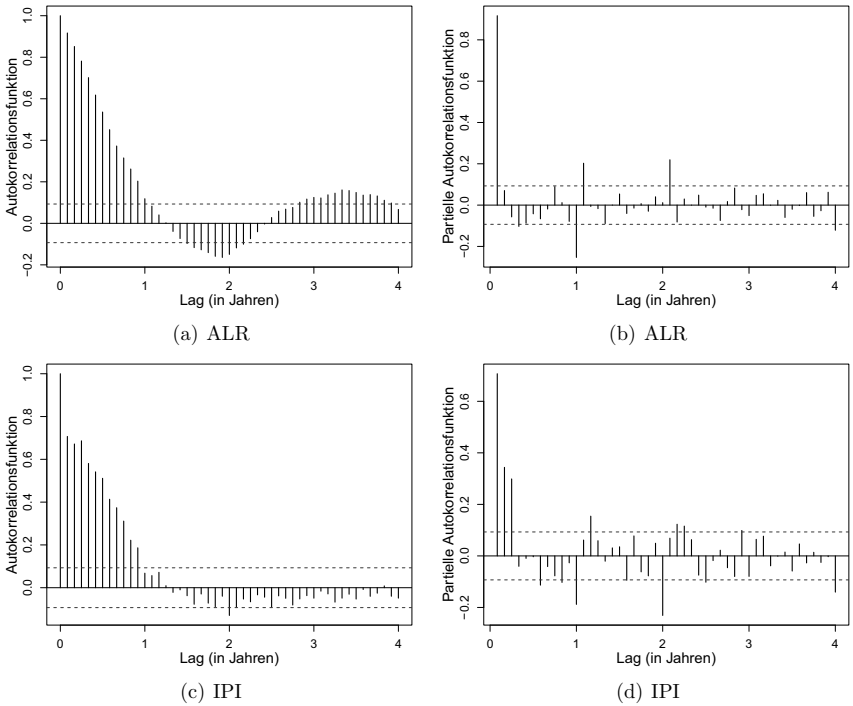
Aus diesen Gründen basiert die Modellselektion, den ursprünglichen Vorschlägen von Box und Jenkins (1970) folgend, vor allem auf einer visuellen Auswertung der ACF und PACF der Zeitreihe. Erst bei der Entscheidung zwischen einigen wenigen in die nähere Auswahl kommenden Modellen sollten Informationskriterien bzw. die Standardfehler der geschätzten Koeffizienten berücksichtigt werden. Ein ARMA-Modell sollte zudem solange als vorläufig angesehen werden, bis die Residuenanalyse ohne Hinweis auf Missspezifikationen durchgeführt werden konnte.

Die visuelle Auswertung von ACF und PACF basiert auf dem Erkennen von abbrechenden, abschwingenden oder sinusoidal schwingenden Mustern. Eine Systematik findet sich in den meisten Lehrbüchern (z.B. Mills, 1990, S. 130). Nach diesen Mustern kann auch getrennt bei den saisonalen Lags gesucht werden. Außerdem sollten einzelne signifikante ACF-Werte berücksichtigt werden.

Abbildung 5 enthält die ACF und PACF[13] der beiden Beispielzeitreihen. Es werden zuerst die ACF und PACF sowie das gefundene Modell für die Arbeitslosenrate besprochen.

Die ACF zeigt ein langsames Abschwingen. Da die ACF vom hohen Wert zum Lag 1 dominiert wird, ist es schwer, andere Muster zu erkennen. Die langen Schwin-

[13] Korrekterweise müsste hier von der Sample-Autokorrelationsfunktion (SACF) bzw. Sample-Partielle-Autokorrelationsfunktion (SPACF) gesprochen werden, da es sich um Schätzungen der „wahren“ ACF bzw. PACF aufgrund der vorliegenden Realisation des Prozesses handelt. Da jedoch im Folgenden aus dem Zusammenhang klar ist, welches Konzept gemeint ist, wird die korrekte Bezeichnung zugunsten der allgemein üblichen fallengelassen.



(a) ALR (b) ALR

(c) IPI (d) IPI

Abb. 5: Autokorrelationsfunktion und Partielle Autokorrelationsfunktion: (a) und (b) Arbeitslosenrate (ALR) sowie (c) und (d) Industrieproduktionsindex (IPI)

gungen deuten auf die Signifikanz saisonaler Lags. Die PACF enthält zu den gewöhnlichen Lags ein deutliches bei Lag 1 abbrechendes Muster, das auf ein signifikantes ϕ_1 hinweist. Weiters ist in der PACF ein (sinusoidal) abschwingendes Muster bei den saisonalen Lags zu erkennen, was auf entsprechende MA-Parameter θ_{12} und θ_{24} schließen lässt. Das "Basismodell„ $x_t = \phi_1 x_{t-1} + \phi_{12} x_{t-12} + \theta_{12} x_{t-12} + \theta_{24} x_{t-24}$, wobei $x = \nabla_{12} ALR$, wurde anhand der Residuenanalyse und mit Hilfe des BIC noch um die Parameter ϕ_7, ϕ_9, ϕ_{10}, ϕ_{12}, θ_3, θ_{10} und θ_{13} ergänzt (vgl. Tab. 5). Das gefundene Modell enthält außerdem kein Interzept, was die Tatsache widerspiegelt, dass in der Arbeitslosenrate kein Trend vorhanden ist. Die Q-Statistik des geschätzten Modells zeigt keine Autokorrelation in den Residuen an.

Die ACF der Industrieproduktion schwingt langsam ab, ohne Schwingungen nach dem Lag 12 aufzuweisen. Die PACF bricht (zunächst) bei Lag 3 ab. Das Modell enthält daher jedenfalls die Parameter ϕ_1, ϕ_2 und ϕ_3. Die PACF ist weiters geprägt von signifikanten Werten zu den saisonalen Lags. Der Umstand, dass der

Tab. 5: Schätzergebnisse für ein ARMA-Modell, ALR

Zeitreihe: $\nabla_{12} \ln ALR - \ln(100 - ALR) + \ln 100$
Zeitraum: 1961/1 – 1997/12 ($N = 444$)
Schätzmethode: Least Squares

Koeff.	Wert	Standardf.	Koeff.	Wert	Standardf.
ϕ_1	0.95284	0.02294 ***	θ_3	0.10003	0.04559 *
ϕ_7	-0.11282	0.03624 **	θ_{10}	-0.18365	0.05866 **
ϕ_9	0.11678	0.05253 *	θ_{12}	-0.33400	0.05060 ***
ϕ_{10}	0.19506	0.05654 ***	θ_{13}	0.19000	0.04818 ***
ϕ_{12}	-0.23002	0.03848 ***	θ_{24}	-0.10485	0.04732 *

R^2adj.: 0.8892
F-Stat.: 338.2209*** (mit 10 und 410 Freiheitsgraden)
$\hat{\sigma}^2$: 0.00296
Q-Stat.(Lag=24): 12.3109
AIC: -1304.67

***, ** und * bezeichnen Werte, die bei einem Konfidenzniveau von 0.1 %, 1 % bzw. 5 % signifikant sind.

Tab. 6: Schätzergebnisse für ein ARMA-Modell, IPI

Zeitreihe: $\nabla_{12} \ln IPI$
Zeitraum: 1961/1 – 1997/12 ($N = 444$)
Schätzmethode: Least Squares

Koeff.	Wert	Standardf.	Koeff.	Wert	Standardf.
Interzept	0.00250	0.00097 *	ϕ_{10}	-0.07003	0.03257 *
ϕ_1	0.36010	0.04464 ***	θ_{12}	-0.46686	0.05042 ***
ϕ_2	0.26380	0.04684 ***	θ_{24}	-0.22812	0.04974 ***
ϕ_3	0.37981	0.04600 ***			

R^2adj.: 0.7069
F-Stat.: 169.4116*** (mit 6 und 413 Freiheitsgraden)
$\hat{\sigma}^2$: 0.000587
Q-Stat.(Lag=36): 34.2026
AIC: -2029.53

***, ** und * bezeichnen Werte, die bei einem Konfidenzniveau von 0.1 %, 1 % bzw. 5 % signifikant sind.

Wert zum Lag 24 größer ist als zum Lag 12, bedeutet, dass das Modell Ordnung 24 (d.h. ϕ_{24} oder θ_{24}) aufweist. Nach Ergänzung um weitere signifikante Parameter ergibt sich das in Abb. 6 zusammengefasste Modell. Das Modell hat ein Interzept, was der Tatsache entspricht, dass die Industrieproduktion in logarithmierten Werten einen positiven linearen Trend aufweist. Das Modell erscheint aufgrund der Residuenanalyse wohlspezifiziert. Es ist im Vergleich zum entsprechenden AR-Modell bedeutend sparsamer spezifiziert.

SARMA-Modellbildung

Für die SARMA-Modellbildung gilt im Prinzip das gleiche wie für die gewöhnlichen ARMA-Modelle. Aufgrund der in SARMA-Modellen unterstellten Multiplikativität des konventionellen und des saisonalen Lag-Polynoms ist die Unterscheidung zwischen saisonalen und nicht-saisonalen Mustern in ACF und PACF besonders wichtig.

Es ist für keine der beiden untersuchten Zeitreihen gelungen, mit Hilfe von SARMA-Modellen ein wohlspezifiziertes Modell zu finden. Offenbar sind die durch die unterstellte Multiplikativität eingeführten Restriktionen zu stark. Für die Arbeitslosenrate ist es vor allem der stark signifikante Koeffizient θ_{10}, der mit einem SARMA-Modell unvereinbar zu sein scheint. Bei der Zeitreihe IPI zeigt sich, dass das geschätzte Modell ohnehin weitgehend einem SARMA$(3,0)(0,2)$-Modell entspricht, allerdings ohne von der Multiplikativität Gebrauch zu machen.

Aufgrund dieses Ergebnisses werden die SARMA-Modelle auch im Zuge des systematischen Vergleichs der Prognosegüte von linearen Modellen und neuronalen Netzen in Kapitel 5 außer Acht gelassen und nur AR-, ARDS-, ARMA- sowie ARMADS-Modelle verwendet.

ARDS-Modellbildung

Im Rahmen der ARDS-Modellbildung wurde im Gegensatz zu den zuvor präsentierten AR- und ARMA-Modellen die bloß einmal differenzierte Zeitreihe herangezogen. Die beträchtlichen saisonalen Schwankungen der Zeitreihen (vgl. Abb. 3 (a) und (d)) können in diesem Modell, zusätzlich zur Erklärung durch den AR-Teil, durch saisonale Dummy-Variablen erklärt werden. Für naheliegende Erweiterungen, wie etwa zusätzliche Berücksichtigung eines MA-Teils (ARMADS-Modell) werden aus Platzgründen keine Ergebnisse präsentiert, sie wurden jedoch im Rahmen des systematischen Vergleichs in Kapitel 5 berücksichtigt.

Die ARDS-Modellspezifizierung funktioniert nach dem gleichen Prinzip wie die AR-Modellspezifikation. Zunächst wird mithilfe der Minimierung des AIC die Ordnung eines ausführlich spezifizierten Modells bestimmt. Es ergibt sich eine Modellordnung von 36 für ALR und von 14 für IPI. Dieses Ergebnis überrascht, weil die selektierte Ordnung für die ALR länger und für die IPI kürzer ist als die Modellordnung des entsprechenden AR-Modells. Ausgehend vom ausführlichen ARDS-Modell

Tab. 7: Schätzergebnisse für ein ARDS-Modell, ALR

Zeitreihe: $\nabla_1 \ln ALR - \ln(100 - ALR) + \ln 100$
Zeitraum: 1960/2 – 1997/12 ($N = 455$)
Schätzmethode: Least Squares

Koeff.	Wert	Standardf.	Koeff.	Wert	Standardf.
ϕ_3	0.10261	0.04446 *	Interzept	0.15387	0.05717 **
ϕ_7	-0.09775	0.03653 **	μ_2	-0.18173	0.04475 ***
ϕ_8	-0.13681	0.03580 ***	μ_3	-0.33603	0.07057 ***
ϕ_{11}	0.18684	0.04393 ***	μ_4	-0.29551	0.08542 ***
ϕ_{12}	0.46746	0.03900 ***	μ_5	-0.29486	0.09636 **
ϕ_{13}	0.10046	0.03236 **	μ_6	-0.27936	0.09975 **
ϕ_{14}	-0.09394	0.03387 **	μ_7	-0.14183	0.09891
ϕ_{15}	-0.12773	0.04359 **	μ_8	-0.12694	0.10047
ϕ_{16}	-0.12515	0.03468 ***	μ_9	-0.08171	0.09875
ϕ_{22}	-0.08375	0.03186 **	μ_{10}	0.01489	0.08859
ϕ_{23}	-0.16845	0.04669 ***	μ_{11}	-0.06151	0.07237
ϕ_{35}	0.09277	0.04099 *	μ_{12}	-0.03044	0.04509
ϕ_{36}	0.12210	0.03674 ***			

R^2adj.: 0.9307
F-Stat.: 234.8375*** (mit 24 und 394 Freiheitsgraden)
$\hat{\sigma}^2$: 0.00259
Q-Stat.(Lag=36): 23.8362
AIC: -1368.52

***, ** und * bezeichnen Werte, die bei einem Konfidenzniveau von 0.1 %, 1 % bzw. 5 % signifikant sind.

gelangt man durch schrittweises Weglassen nicht-signifikanter Lags („Pruning“) zu einem sparsamen Modell. Hierbei betrifft das Pruning nur die AR-Variablen, nicht die saisonalen Dummies. In den im folgenden präsentierten Modellen könnte man auch nicht-signifikante Dummy-Variablen weglassen (bzw. solche, die sich nicht von einander unterscheiden, zusammenfassen), ohne dass die restlichen Koeffizienten wesentlich verändert würden.

Die Tabellen 7 und 8 enthalten die Schätzergebnisse für die beiden Zeitreihen. Beide Modelle scheinen aufgrund der Residuenanalyse (vgl. Q-Statistik) richtig spezifiziert zu sein. Wie auch bei den auf der saisonal differenzierten Zeitreihe geschätzten AR- und ARMA-Modellen sind die Modelle nahe der Unit-Root. Im Falle der ALR ist die grösste Wurzel der charakteristischen Gleichung 0.99 im Falle des IPI 0.93. Aufgrund der niedrigen Modellordnung und des Umstands, dass bereits ein sehr sparsam spezifiziertes Modell ausreicht, ergibt sich der Eindruck, dass insbesondere für IPI ein ARMADS-Modell eine wertvolle Option ist.

Tab. 8: Schätzergebnisse für ein ARDS-Modell, IPI

Zeitreihe: $\nabla_1 \ln IPI$					
Zeitraum: 1960/2 – 1997/12 ($N = 455$)					
Schätzmethode: Least Squares					
Koeff.	Wert	Standardf.	Koeff.	Wert	Standardf.
ϕ_1	-0.63964	0.04378 ***	μ_5	0.16903	0.01622 ***
ϕ_2	-0.36521	0.04343 ***	μ_6	0.15244	0.01463 ***
ϕ_{12}	0.50283	0.03991 ***	μ_7	0.01409	0.01514
ϕ_{13}	0.21304	0.05172 ***	μ_8	0.11967	0.01466 ***
ϕ_{14}	0.16088	0.04598 ***	μ_9	0.27446	0.01618 ***
Interzept	-0.14136	0.01071 ***	μ_{10}	0.15909	0.01397 ***
μ_2	0.20762	0.01926 ***	μ_{11}	0.17831	0.01425 ***
μ_3	0.16577	0.01432 ***	μ_{12}	0.14264	0.01910 ***
μ_4	0.15143	0.01401 ***			
R^2adj.: 0.9063					
F-Stat.: 266.9942*** (mit 16 und 424 Freiheitsgraden)					
$\hat{\sigma}^2$: 0.000563					
Q-Stat.(Lag=36): 15.909					
AIC: -2079.11					

***, ** und * bezeichnen Werte, die bei einem Konfidenzniveau von 0.1 %, 1 % bzw. 5 % signifikant sind.

2.5 Schlussfolgerungen

In diesem Kapitel wurde in die Grundlagen der linearen Zeitreihenanalyse eingeführt. Es wurden die wichtigsten Probleme diskutiert, die sich einem Praktiker, der lineare Modelle auf ökonomischen Zeitreihen schätzen möchte, unweigerlich stellen. Diese Probleme sind teilweise unmittelbar auf die Aufgabenstellung nichtlinearer Modellbildung zu übertragen oder tauchen zumindest nur in einem anderen Gewand wieder auf.

Relativ klar ist die Bestimmung der richtigen Transformation der Zeitreihe, die im Zusammenhang mit der allgemeinen Vorstellung der Beispielzeitreihen, Arbeitslosenrate (ALR) und Industrieproduktionsindex (IPI) diskutiert wurde. Es erscheint klar, dass für einen Vergleich von linearen Modellen mit ARNN-Methoden die Zeitreihen in einer identischen Transformation zugrundegelegt werden müssen.

Ein wichtiger Problemkreis ist die Bestimmung der anzuwendenden Differenzenfilters. Die Forschung hat sich von der auf Box und Jenkins (1970) zurückgehenden Empfehlung, unterjährige Zeitreihen prinzipiell sowohl konventionell als auch saisonal zu differenzieren, distanziert und statt dessen große Bemühungen darin gesetzt, Testverfahren zu entwickeln, die die richtige Ordnung des Differenzierens bestimmen helfen. Diesem Ansatz folgend wurden diese Verfahren auf die beiden Beispielzeitreihen angewandt, allerdings mit uneindeutigen Ergebnissen. Wie wohl es klar ist, dass in beiden Zeitreihen eine Unit-Root vorhanden ist, lässt sich nicht leicht sagen ob es eine konventionelle oder saisonale Unit-Root ist, bzw. ob nicht trotz der

Ablehnung einer saisonalen Unit-Root dennoch aus pragmatischen Gründen saisonal differenziert werden sollte.

In dieser Situation erscheint es das beste, als bevorzugten Differenzenfilter ∇_s zu betrachten, zusätzlich für die lineare Modellierung aber auch anders differenzierte Zeitreihen heranzuziehen. Es wurden AR- und ARMA-Modelle auf $\nabla_s x_t$ und ARMADS-Modelle auf ∇x_t geschätzt, zusätzlich wird beim systematischen Prognosevergleich in Kapitel 5 auch $\nabla_s \nabla x_t$ berücksichtigt werden. Es erscheint vielversprechend auch ARNN-Modelle auf genau diesen differenzierten Zeitreihen zu schätzen und einem systematischen Vergleich zu unterziehen. Ein Vergleich der in diesem Kapitel präsentierten linearen Modelle untereinander („In-Sample“) ist nur sehr eingeschränkt möglich, insbesondere aufgrund der unterschiedlichen Modellordnung und der sich daraus ergebenden unterschiedlichen Länge der Residuenzeitreihe. Dennoch bietet sich ein Vergleich anhand des AIC an, wobei das AIC als bloße Heuristik angesehen werden muss. Hierbei schneidet das ARMADS für beide Zeitreihen am besten ab, und zwischen dem AR und dem ARMA-Modell bestehen kaum Unterschiede.

Ein weiteres Ziel dieses Kapitels war es, Modellselektionsstrategien zu diskutieren. Es zeigte sich, dass bereits bei linearen Modellen die Modellselektion sehr komplex werden kann, sodass ein automatisiertes Vorgehen, wie es für einen systematischen Vergleich notwendig ist, nur schwer zu bewerkstelligen ist. Insbesondere die Hinzunahme von MA-Termen in linearen Modellen verkompliziert die Modellspezifikation, da hier Erschwernisse bei der Modellschätzung von Prozessen, die nahe der Unit-Root sind (wie für ökonomische Zeitreihen typisch), auf die Modellspezifikation zurückwirken.

Generell kann jedoch festgestellt werden, dass die geschätzten linearen Modelle wohlspezifiziert und robust sind, und dass keine Gründe dagegen sprechen, lineare Modelle zur Modellierung und Prognose einzusetzen.

3 Tests auf Nicht-Linearität

Nicht-lineare Modellierung, im speziellen mit Hilfe neuronaler Netze, ist im Vergleich zu linearer Modellierung, wie sie in Kapitel 2 besprochen wurde, um einiges anspruchsvoller. Nicht-lineare Modellierung ist vor allem mit der Gefahr des „Overfittings“ verbunden, wenn die Nicht-Linearitäten, die in der Zeitreihe vermutet werden, nicht vorhanden, nicht stark genug ausgeprägt sind oder von einem Typ sind, wie er von der verwendeten Modellklasse nicht behandelt werden kann. Um dieser Gefahr aus dem Weg zu gehen, sollte man zunächst verschiedene Tests auf Linearität durchführen und die Anwendung nicht-linearer Modelle von der Ablehnung der Null-Hypothese der Linearität abhängig machen.

Der Zeitreihenanalytiker sollte sich auch darüber Gedanken machen, ob prinzipiell in der untersuchten Zeitreihe Nicht-Linearitäten zu erwarten sind. Auf dem Anwendungsgebiet der ökonomischen Zeitreihen bedeutet das, nach theoretischen Gründen zu fragen, warum in dem der Zeitreihe zugrundeliegenden datengenerierenden Prozess (DGP) nicht-lineare Elemente eine Rolle spielen könnten.

In den folgenden Abschnitten werden die Grundlagen und Begriffe im Zusammenhang mit Nicht-Linearitäten in univariaten Zeitreihen, ökonomische Interpretationen nicht-linearer Modelle sowie die möglichen Teststrategien und die im Rahmen der vorliegenden Arbeit verwendete Abfolge von Hypothesentests erläutert (Abschnitt 3.1 und 3.2). Es werden weiters die Ergebnisse für die beiden Beispielzeitreihen (Abschnitt 3.3) präsentiert. Schließlich wird noch darauf eingegangen, inwiefern Strukturbrüche in Zeitreihen die Diagnose von nicht-linearen Effekten in Zeitreihen erschweren und auch für die beiden Beispielszeitreihen einige Aspekte der Ergebnisse erklären könnten (Abschnitt 3.4). Abschnitt 3.5 fasst die Schlussfolgerungen in Bezug auf eventuell in den beiden Beispielzeitreihen vorhandene Nicht-Linearitäten zusammen.

3.1 Grundlagen und Bedeutung von Nicht-Linearitäten

Dieser Abschnitt führt in die wichtigsten Grundkonzepte und ökomischen Hintergründe im Zusammenhang mit Nicht-Linearität ein. Zu den Literaturquellen, die eine Gesamtsicht für dieses Gebiet zu vermitteln versuchen, gehören Tong (1990), Brock et al. (1991), Granger und Teräsvirta (1993), Cromwell et al. (1994) und Franses und van Dijk (2000). Auf diese Literatur und auf einige weitere Übersichtsartikel (Granger, 1991; De Gooijer und Kumar, 1992; Barnett et al., 1997) stützt sich dieser Abschnitt in erster Linie. Die statistisch-theoretische und ökonomische Literatur ist in den letzten Jahren jedoch so gewaltig angewachsen, dass eine einheitliche Sichtweise nur bedingt herbeizuführen ist. Die hohe Relevanz nicht-linearer Zusammenhänge ist natürlich schon sehr früh und in verschiedenen Teilgebieten der Ökonomie und in anderen Disziplinen erkannt worden, und die Wichtigkeit der entwickelten Modelle und Instrumente wurde von den jeweils anderen Disziplinen

nur schrittweise entdeckt. Diesem Umstand wird im folgenden durch entsprechende Querverweise und eine Diskussion der ökonomischen Relevanz der Konzepte Rechnung getragen.

Definitionen und Begriffsabgrenzungen

Nicht-Linearität wird unter Rückgriff auf die in Abschnitt 2.1 durch Gleichung (2.4) gegebene und hier wiederholte Definition eines linearen Prozesses

$$x_t - \mu_t = \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \ldots = \sum_{j=0}^{\infty} \phi_j \varepsilon_{t-j},\ \phi_0 = 1, \tag{3.1}$$

definiert, wobei ε_t ein unabhängig und identisch verteilter Innovationsprozess, μ_t eine etwaige determinististische Komponente ist und ϕ_j Koeffizienten des Prozesses sind. Ein schwach stationärer Prozess ist dann nicht-linear, wenn er *nicht* nach Gleichung (3.1) dargestellt werden kann. Nach dem Wold'schen Dekompositionstheorem können auch nicht-lineare Prozesse, so sie schwach stationär sind, nach Gleichung (3.1) dargestellt werden, wenn sich die Annahmen bezüglich ε_t auf Unkorreliertheit beschränken – die Nicht-Linearität manifestiert sich in dieser Darstellungsform in den höheren Momenten von ε_t.

Dieser Zusammenhang ist nicht nur von theoretischer Relevanz. Zwar ist man bei der praktischen Modellierung von Zeitreihen an einer knappen Spezifizierung des Modells interessiert ist, also keinesfalls an einer solchen mit unendlicher Modellordnung. Doch auch für endliche Filter gilt entsprechendes. Nicht-lineare Prozesse haben eine Darstellungsform als endliche lineare Filter eines Innovationsprozesses, für den zumindest bis zu einer endlichen Ordnung p die ACF(p) verschwindet. Dieses Ergebnis folgt unmittelbar aus dem Wold'schen Dekompositionstheorem, das ja diese Aussage für $p = \infty$ macht. Ein auf einer nicht-linearen Zeitreihe angepasstes AR(p)-Modell ist ein offensichtliches Beispiel. Durch Schätzung eines AR(p)-Modells wird die Nicht-Linearität in die Residuenzeitreihe überführt und kommt dort in den höheren Momenten zu tragen.

Nicht-Linearität tritt in zwei verschiedenen Formen auf.[14] Additive Nicht-Linearität (auch: Nicht-Linearität im bedingten Mittelwert) in einer schwach stationären, rein nicht-deterministischen Zeitreihe x_t liegt vor, wenn die Prognose zukünftiger Werte (unter Kenntnis des DGP, d.h. des wahren Modells) eine nicht-lineare Funktion gegenwärtiger und vergangener Werte ist:

$$\mathrm{E}(x_{t+1}|x_t, \ldots, x_{t-k}) = g(x_t, \ldots, x_{t-k}),\ \ g \text{ nicht-linear.} \tag{3.2}$$

Hierbei bezeichnet $\mathrm{E}(x_{t+1}|x_t, \ldots, x_{t-k})$ den bedingten Erwartungswert von x_{t+1} unter Kenntnis von $x_t, \ldots, x_{t-k}$.

[14] Im folgenden werden nur univariate Konzepte berücksichtigt. Die Verallgemeinerung für den multivariaten Fall ist aus der Sicht der statistischen Theorie problemlos, erhöht jedoch enorm die Komplexität.

Im Gegensatz dazu handelt es sich um multiplikative Nicht-Linearität (auch: Nicht-Linearität in der bedingten Varianz), wenn unter Kenntnis des DGP die Varianz der Prognosen zukünftiger Werte eine beliebige Funktion gegenwärtiger und vergangener Werte ist:

$$\mathrm{Var}(x_{t+1}|x_t, \ldots, x_{t-k}) = g(x_t, \ldots, x_{t-k}), \;\; g \text{ beliebig.} \tag{3.3}$$

Geeignete nicht-lineare Modelle erlauben daher im Vergleich zu bloß linearen Prognoseinstrumenten bei additiver Nicht-Linearität bessere Punktprognosen und bei multiplikativer Nicht-Linearität bessere Konfidenzintervalle. Für die obigen Begriffe existieren auch anderslautende Definitionen, z.B. Hsieh (1989, S. 354), Granger und Teräsvirta (1993, S. 8f) und Lee et al. (1993, S. 270), die jedoch nichts wesentliches ändern bzw. Verallgemeinerungen darstellen. So kann man etwa in den Gleichungen (3.2) und (3.3) als Argumente der Funkion g zusätzlich auch die Fehler (Innovationen) $u_t, \ldots, u_{t-l}$ aufnehmen, was jedoch redundant ist, da man bei genügend großem k die Innovationen $u_t, \ldots, u_{t-l}$ beliebig genau aus $x_t, \ldots, x_{t-k}$ wiedergewinnen kann.

Des weiteren ist der Zusammenhang des Begriffs der additiven Nicht-Linearität mit jenem des Chaos klarzustellen. Charaktaristisch für chaotische Zeitreihen ist, dass sie rein deterministisch sind, jedoch stochastisch aussehen und dass kleine Veränderungen in den Anfangsbedingungen der Zeitreihe große Veränderungen in einer genügend weit entfernten Zukunft zur Folge haben. Die Sensitivität der Zeitreihe für Anfangsbedingungen ist das wesentliche Merkmal praktisch aller in der Literatur existierender Definitionen von Chaos und wird mit Hilfe des größten Lyapunov-Exponenten, Q^L, formalisiert, wobei im Falle von Chaos $Q^L > 0$ gilt (vgl. Brock et al., 1991, S. 8ff). Stark vereinfacht ausgedrückt misst Q^L die Geschwindigkeit der Divergenz von Trajektorien, die sich in den Anfangsbedingungen nur gering unterscheiden. Chaos ist ein Spezialfall von Gleichung (3.2), da nur nicht-lineare Abhängigkeit zu $Q^L > 0$ führen kann. Ein offensichtliches Beispiel für Chaos sind Zufallszahlengeneratoren, wie sie in statistischen Computerpaketen eingesetzt werden. Hierbei handelt es sich um hochdimensionales Chaos, d.h. die nicht-lineare Abhängigkeit der Zeitreihe wird erst sichtbar, wenn man die Zeitreihe in einen hochdimensionalen Raum einbettet und Trajektorien sehr großer Länge betrachtet. Viel interessanter ist jedoch niedrigdimensionales Chaos, das durch ein Modell geringer Ordnung zustandekommt. Ein oft angeführtes Beispiel ist die „Tent-Map“, die durch die Iteration

$$x_t = \begin{cases} 2x_{t-1} & \text{wenn } x_{t-1} \leq 0.5, \\ 2 - 2x_{t-1} & \text{sonst} \end{cases} \tag{3.4}$$

definiert ist.

Entgegen der anfänglichen Euphorie haben sich Chaos-Konzepte in der Ökonomie als wenig nützlich erwiesen, einfach weil ökonomische Systeme inhärent stochastisch sind. Gerade für makroökonomische Systeme ist die Existenz von unerwarte-

ten Schocks und Innovationen offensichtlich und bewirkt zusammen mit Messfehlern die stochastische Natur der Ökonomie.[15] Granger (1999) geht sogar soweit, deterministisches Chaos in der Ökonomie als „empty box“, als Konzept ohne jeden Erklärungsgehalt für die Realität, zu bezeichnen.

Einige nicht-lineare Modelle

Zu einfachen Modellen mit additiver Nicht-Linearität gehören das bilineare Modell (BL) (Granger und Andersen, 1978), z.B.

$$x_t = \varepsilon_t + \alpha x_{t-1}\varepsilon_{t-1} \tag{3.5}$$

und das Threshold-Autoregressive Modell (TAR) (Tong und Lim, 1980), z.B.

$$x_t = \begin{cases} \alpha x_{t-1} + \varepsilon_t & \text{wenn } x_{t-1} \leq \gamma, \\ \beta x_{t-1} + \varepsilon_t & \text{sonst} \end{cases} \tag{3.6}$$

Es existieren zahlreiche Verallgemeinerungen und Erweiterungen der Modelle (3.5) und (3.6), wobei am naheliegendsten Verallgemeinerungen für mehrere Lags sind. Ein erstaunliches Merkmal dieser Modelle ist, dass sie oft Zeitreihen ohne Autokorrelation generieren, z.B. das Modell (3.5). Dennoch ist im Regelfall zu erwarten, dass nicht-lineare Modelle auch eine lineare Struktur in der Zeitreihe erzeugen, z.B. das Modell (3.6).

Eine besondere Rolle unter den additiv nicht-linearen Zeitreihen-Modellen spielen die in Kapitel 4 ausführlich behandelten Autoregressiven Neural-Network-Modelle (ARNN), bei dem die nicht-lineare Funktion, die die Vergangenheit einer Zeitreihe $x_{t-1}, \ldots, x_{t-j}$ auf den gegenwärtigen Wert abbildet, mit Hilfe eines neuronalen Netzes (NN) realisiert wird:

$$x_t = \delta + \sum_{k=1}^{K} \gamma_k \psi(\beta_k + \sum_{j=1}^{J} \alpha_{kj} x_{t-j}) + \varepsilon_t, \tag{3.7}$$

wobei α_{kj}, β_k, γ_k, und δ die Koeffizienten des Modells (im Sprachgebrauch der NN-Forschung die Gewichte des neuronalen Netzes) sind und ψ eine geeignete Aktivierungsfunktion, z.B. die tanh-Funktion ist. Für neuronale Netze liegen Theoreme vor, die sie als universale Funktionsapproximatoren ausweisen (Hornik et al., 1989, 1990). Das bedeutet, dass mit Hilfe von Modell (3.7) alle denkbaren nicht-linearen Modelle beliebig genau angenähert werden können. Daher sind ARNNs besonders in jenen Fällen attraktiv, wenn die ökonomische Theorie keine Präferenzen für ein bestimmtes nicht-lineares Modell liefert.

[15] Granger und Teräsvirta (1993, S. 2) betonen, dass Messfehler in der Ökonomie nicht mit den Messfehlern in der Physik zu vergleichen sind, da sie von den ökonomischen Agenten in ihre Entscheidungen miteinbezogen werden und auf diese Weise dynamisch in das System eingebunden sind.

Für Modelle, die multiplikative Nicht-Linearität generieren, ist vor allem an Autoregressive-Conditional-Heteroskedasticity-Modelle (ARCH) (Engle, 1982),

$$\varepsilon_t \sim \mathrm{N}(0, z_t),\ z_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2\,, \tag{3.8}$$

und Generalized-Autoregressive-Conditional-Heteroskedasticity-Modelle (GARCH) (Bollerslev, 1986),

$$\varepsilon_t \sim \mathrm{N}(0, z_t),\ z_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i z_{t-i}\,, \tag{3.9}$$

zu denken. In diesen Modellen ist die bedingte Varianz zum Zeitpunkt t einer Zeitreihe eine lineare Funktion der vergangenen quadrierten Innovationen ε_t^2 und der vergangenen bedingten Varianz z_t. Es existieren zahlreiche Erweiterungen und Verallgemeinerungen, die stattdessen eine nicht-lineare Funktion verwenden oder eine nicht-lineare Transformation von z_t implizieren. Weiters gibt es Erweiterungen der Modelle (3.8) und (3.9) um ein lineares Modell des Mittelwerts, z.B. ARMA-GARCH,

$$\phi(B)x_t = \theta(B)\varepsilon_t,\ \ \varepsilon_t \sim \mathrm{GARCH}\,, \tag{3.10}$$

sowie solche, die einen Einfluss der bedingten Varianz auf den Mittelwert der Zeitreihe zulassen, z.B. das „ARCH-in-mean“-Modell (ARCH-M) (Engle et al., 1987). Das ARCH-M-Modell beinhaltet sowohl additive als auch multiplikative Nicht-Linearität.

Wie für lineare Modelle sind auch für nicht-lineare Modelle Stationarität und Invertierbarkeit wichtige Eigenschaften. Der Begriff der Invertierbarkeit wurde in Abschnitt 2.1 nur im Zusammenhang mit MA- und ARMA-Modellen definiert. Im Zusammenhang mit nicht-linearen Modellen ist eine allgemeinere Definition angebracht. Wenn für ein lineares oder nicht-lineares Modell

$$x_t = g(x_{t-1}, \ldots, x_{t-p}, \varepsilon_{t-1}, \ldots, \varepsilon_{t-q}) + \varepsilon_t \tag{3.11}$$

eine bestimmte Realisation bekannt ist, so stellt sich mitunter die Frage, ob mit dieser Information auch eine Schätzung der (nicht unmittelbar beobachtbaren) ε_t möglich ist. Diese Frage ist besonders für die Erstellung von Prognosen wichtig. Eine Schätzung der Zeitreihe der Innovationen ε_t, $t = 1, \ldots$, setzt die Annahme oder Schätzung von bestimmten Start-up-Werten $\varepsilon_{-k}, \ldots, \varepsilon_0$ voraus. Ein Modell heißt invertierbar, wenn für jede vernünftige Wahl von Start-up-Werten die Schätzungen $\hat{\varepsilon}_t$ zu den wahren ε_t konvergieren: $\mathrm{E}([\hat{\varepsilon}_t - \varepsilon_t]^2) \to_t 0$.

Die Bedingungen für Stationarität und Invertierbarkeit sind mitunter auch in sehr einfach gehaltenen nicht-linearen Modellen nur sehr schwer zu ermitteln, oder sie sind zu kompliziert um in der Praxis verwendet zu werden (vgl. Tong, 1990,

Kap. 4). Z.B. lautet die Bedingung für die Stationarität des TAR-Modells nach Gleichung (3.6) $|\alpha| < 1$, $|\beta| < 1$ und die Bedingung für Invertierbarkeit des BL-Modells nach Gleichung (3.5) $|\alpha\sigma_\varepsilon| < 0.606$. Ein ARNN-Modell nach Gleichung (3.7) ist jedoch immer stationär und invertierbar. Die Stationariät ist intuitiv wegen der Beschränktheit der Aktivierungsfunktion eines neuronalen Netzes, in Gleichung (3.7) die tanh-Funktion, plausibel, und die Invertierbarkeit ist unmittelbar klar, weil in ein ARNN-Modell keine ε_t als Regressoren eingehen.

Relevanz nicht-linearer Zeitreihenmodelle in der Ökonomie

Als besonders wichtig für die Entstehung nicht-linearen Zeitreihenverhaltens von wichtigen wirtschaftlichen Indikatoren wie Arbeitslosenrate und Industrieproduktion wird oft angenommen, dass Schocks von einem ökonomischen System in den unterschiedlichen Phasen des Konjunkturzyklus verschieden propagiert werden oder dass der Propagationsmechanismus vom Vorzeichen des Schocks abhängt, vgl. Granger und Teräsvirta (1993), Potter (1994) und mehrere Beiträge in Semmler (1994). Dieser Umstand führt zu Asymmetrien im Konjunkturzyklus, engl. „Business Cycle Asymmetries“ (BCA). Eine informale Definition von BCA basiert auf der Distanz von einem Tal zu einer Spitze bzw. von einer Spitze zu einem Tal im Zeitverlauf des Indikators. Wenn diese Distanzen im Durchschnitt nicht gleich lang sind, ist der Konjunkturzyklus asymmetrisch. Bereits sehr früh befasste sich die Wirtschaftszyklentheorie mit diesem Phänomen (Mitchell, 1927; Burns und Mitchell, 1946). Es ist klar, dass lineare Modelle BCA nicht berücksichtigen können.

Die traditionelle Wirtschaftszyklusforschung ist oft informal und geht nicht darauf ein, durch welches Entscheidungsverhalten der ökonomischen Akteure BCAs zustandekommen. Konkrete ökonomisch-theoretische Modelle, die nicht-lineares Zeitreihenverhalten erzeugen, sind zahlreich, wobei der Aspekt der Nicht-Linearität erst unter dem Eindruck der Neuentwicklungen auf dem Gebiet der nicht-linearen Zeitreihenanalyse bewusst in den Vordergrund gestellt wurde. Es folgt eine (unvollständige) Auswahl solcher theoretischer Ansätze:

- Asymmetrische Anpassungskosten: Studien haben gezeigt, dass die mit der Anstellung neuer Arbeitskräfte einerseits und der Entlassung von Arbeitskräften andererseits verbundenen Kosten nicht symmetrisch sind (Hamermesh und Pfann, 1996). Auch für andere Produktionsfaktoren existieren ähnliche Ansätze. Spezifisch arbeitsmarktökonomisch ist das Argument, dass die Kosten für Suche und Besetzung sowohl auf Seite der Unternehmen als auch der Arbeitnehmer von der Angespanntheit des Arbeitsmarktes abhängen (Burgess, 1992).

- Insider-Outsider-Modelle: Erwerbstätige und Arbeitslose (also Insider und Outsider) haben unterschiedliche Präferenzen und Interessen und werden von Interessensgruppen, etwa Gewerkschaften, unterschiedlich vertreten, da sich

mit dem Konjunkturzyklus die relative Stärke der Insider und Outsider ändert, verändert sich auch das Verhalten der ökonomischen Akteure in Abhängigkeit vom Konjunkturzyklus. Ähnliche Argumente können auch für eine Unterscheidung zwischen der Gruppe der Kurzzeit- und Langzeitarbeitslosen vorgebracht werden (Layard et al., 1991, Kap.3).

- Rezessionen als Reinigungsphasen der Wirtschaft: Auf Ideen der alten liquidationistischen Sichtweise, die vor der Keynesianischen Revolution weit verbreitet war, bauen Ansätze auf, die Rezessionen als Phasen der Selbstreinigung der Ökonomie interpretieren (Caballero und Hammour, 1994). Der Prozess der (kreativen) Zerstörung von Arbeitsplätzen bzw. Kapital und der Prozess der Neuschaffung von Arbeitsplätzen bzw. Kapital (also Investition) sind asymmetrisch.

Einige der nicht-linearen Modelle sind auf der Mikro-Ebene angesiedelt, und es stellt sich daher die Frage, ob die nicht-linearen Struktur die Aggregation über eine große Zahl von ökonomischen Agenten „übersteht“ (Brock, 2000). Aus der statistisch-theoretischen Perspektive kann diese Frage mit Hilfe von Simulationsstudien beantwortet werden, indem für bestimmte nicht-lineare Prozesse Cross-Section-Aggregate gebildet werden. Granger und Teräsvirta (1993, Abschn. 8.2) fassen die Ergebnisse einer solchen Simulation zusammen und stellen unter anderem fest, dass wesentlich für den Erhalt der nicht-linearen Struktur nach der Aggregation das Ausmaß der Korrelation des Innovationsprozesses ε_t über die Cross-Section-Einheiten ist, über die aggregiert wird.

Die empirische Literatur zu Asymmetrien in makroökonomischen Aggregaten ist sehr umfangreich. (Neftci, 1984; Brock und Sayers, 1988; Acemoglu und Scott, 1994; Gallegati und Mignacca, 1995; Peat und Stevenson, 1996; Peel und Speight, 1998b; Stanca, 1999; Peel und Speight, 2000). Während für die Arbeitslosenrate das Vorhandensein nicht-linearer Struktur unbestritten ist, liegen für die Industrieproduktion gemischte Ergebnisse vor. Am besten untersucht ist die US-amerikanische Arbeitslosenrate, für die auch eine rein visuelle Analyse das Vorhandensein von BCAs verdeutlicht (Montgomery et al., 1998; Rothman, 1998). In den oben zitierten Arbeiten wurden meist nicht-lineare Modelle (TAR, SETAR, BL, usw.) oder Tests auf Nicht-Linearität verwendet, wie sie im folgenden Abschnitt behandelt werden. Einige dieser Arbeiten hatten explizit das Ziel, die Möglichkeit chaotischer Zeitreihen zu untersuchen, konnten jedoch durchwegs nur gewöhnliche nicht-lineare Struktur nachweisen. Es existieren auch Arbeiten, die nicht-lineare Modelle u.a. auf österreichischen makroökonomischen Zeitreihen anpassen (Skalin und Teräsvirta, 1999).

Die bisherige Diskussion betraf additive Nicht-Linearität und ihre mögliche theoretische Begründung bzw. empirischen Nachweis. Was multiplikative Nicht-Linearität in makroökonomischen Zeitreihen betrifft, gibt es keine ökonomisch-theoretischen Modelle, die sie explizit vorhersagen. Das Vorhandensein von multi-

plikativer Nicht-Linearität in makroökonomischen Zeitreihen (z.B. vom ARCH-Typ oder verwander Modelle) scheint jedoch anhand reiner ad-hoc-Theorien plausibel und aus empirischer Sicht unbestritten zu sein (Fornari und Mele, 1997; Peel und Speight, 1998a).

3.2 Teststrategien und eine Auswahl von Testverfahren

Dieser Abschnitt dient der Beschreibung der ausgewählten Testverfahren auf Nicht-Linearität und der Vorgangsweise ihrer praktischen Implementierung. Zuvor werden einige zentrale Aspekte der Theorie statistischer Tests diskutiert und mögliche Strategien für Testverfahren auf Nicht-Linearität aufgezeigt. Dadurch soll die getroffene Auswahl an Testverfahren transparent werden und eine kritische Interpretation der Testergebnisse ermöglicht werden.

Einige Grundaspekte der Teststatistik und Teststrategien

Für Tests auf Nicht-Linearität (in der Literatur und in dieser Arbeit wird gleichbedeutend auch der Ausdruck Test auf Linearität verwendet) gelten die Grundprinzipien der Teststatistik. Es wird die Null-Hypothese der Linearität H_0 gegen eine Alternativ-Hypothese der Nicht-Linearität H_A getestet. Zentral sind die Begriffe *Größe* und *Macht* von Tests.

Die Größe eines Tests ist die Wahrscheinlichkeit eines Tests die Null-Hypothese abzulehnen, wenn sie richtig ist. Die Größe kann idealerweise durch den Forscher kontrolliert werden. Durch die Vorgabe eines Konfidenzniveaus, typischerweise 5 %, und die Verwendung von zu diesem Konfidenzniveau passenden kritischen Werten wird sichergestellt, dass die Null-Hypothese nur abgelehnt wird, wenn sehr große statistische Evidenz gegen sie spricht. Die korrekte Angabe der Größe eines Tests verlangt die Kenntnis der Verteilung der Teststatistik unter der Null-Hypothese. Oftmals liegen nur Angaben über die asymptotische Verteilung der Teststatistik vor, und der Forscher verlässt sich darauf, dass die Verteilung der Teststatistik für kleine Stichproben in befriedigender Weise durch die asymptotische Verteilung angenähert wird.[16] Darin liegt gerade bei Nicht-Linearitätstests eine große Gefahr, da die Small-Sample-Verteilungen sich mitunter sehr stark von der asymptotischen Verteilung unterscheiden. In diesem Fall weicht auch die *nominelle* Größe beträchtlich von der *tatsächlichen* Größe ab. Dieses Problem ist etwa im Falle des BDS-Tests gravierend. Statt der kritischen Werte aufgrund der asymptotischen Verteilung sollten statt dessen kritische Werte aus Tabellen verwendet werden, die für jede Stichprobengröße (und eventuell zusätzliche Parameter, von denen die Small-Sample-Verteilung abhängt) durch Monte-Carlo-Simulation ermittelt werden.

[16] Zudem basiert die Herleitung der asymptotischen Verteilung auf zusätzlichen Annahmen, z.B. bezüglich des Fehlerprozesses, die nicht unbedingt zutreffen müssen.

Die Macht eines Tests ist die Wahrscheinlichkeit, die Null-Hypothese abzulehnen, wenn sie falsch ist. Die Macht eines Tests kann natürlich nicht allgemein angegeben werden, sondern hängt von der Art und Stärke der Abweichung von der Null-Hypothese ab, also etwa von Art und Stärke der Nicht-Linearität. Man ist jedenfalls interessiert, unter den möglichen Tests (mit gleicher Null- und Alternativ-Hypothese) jenen Test heranzuziehen, der die höchste Macht hat. Dies ist meist nicht möglich, da Tests für bestimmte Unterklassen innerhalb der Alternativ-Hypothese hohe Macht, für andere Unterklassen hingegen geringere Macht haben. Diese Problematik ist vor allem für Nicht-Linearitätstests essentiell, da die Alternativ-Hypothese der Nicht-Linearität in beliebig viele Unterklassen unterteilt werden kann. Je enger die Alternativ-Hypothese auf eine bestimmte Form der Nicht-Linearität eingeschränkt wird, desto eher ist es möglich, Tests mit hoher Macht gegen genau diese Alternativ-Hypothese zu konstruieren. Dies bedeutet jedoch nicht, dass dieser Test nicht auch Macht gegen andere Alternativ-Hypothesen haben kann.

Eine nützliche Einteilung der Nicht-Linearitätstests ist jene in solche, die eine bestimmte nicht-lineare Alternative spezifizieren und solche, die die Alternativ-Hypothese unspezifiziert lassen, also nur ausschließend definieren (Granger und Teräsvirta, 1993, Kap. 6). Diese Einteilung ist nicht strikt, da es in der zweiten Kategorie Tests gibt, die auf bestimmte Weise und bei zusätzlichen Annahmen als Test gegen eine bestimmte Alternative interpretiert werden können. Zur ersten Kategorie gehören jene Tests, die ein bestimmtes nicht-lineares Modell implizieren und einen oder mehrere der Parameter dieses Modells auf Signifikanz testen (etwa der Teräsvirta-Lin-Granger-Test, vgl. unten). Zur zweiten Kategorie zählen jene Tests, die eine bestimmte Eigenschaft von Zeitreihen unter der Alternativ-Hypothese ausnützen, etwa die Korrelation der quadrierten Zeitreihe (McLeod-Li-Test).

Eine weitere im Zusammenhang mit Nicht-Linearitätstest übliche Einteilung ist jene in solche, die auf die Originalzeitreihe angewendet werden (direkte Tests) und solche, die auf die Residuen eines linearen Modells aufbauen (Diagnostik-Tests). Durch Schätzung eines linearen Modells wird die nicht-lineare Struktur in die Residuenzeitreihe transferiert. Natürlich ist es auch möglich mit Hilfe von Diagnostik-Tests die Residuen eines nicht-linearen Modells zu untersuchen. In diesem Fall testet man auf „vernachlässigte“ Nicht-Linearität. Die Vorgangsweise, Nicht-Linearitätstests auf Residuenzeitreihen anzuwenden, ist in der empirischen Forschung allgemein üblich. Es ist jedoch Vorsicht angebracht, da die Qualität des Testverfahrens wesentlich von der Güte des vorgeschalteten linearen Modells abhängt.

Eine wünschenswerte, aber nur eingeschränkt umsetzbare Unterteilung der Nicht-Linearitätstests beruht auf der Unterscheidung zwischen additiver und multiplikativer Nicht-Linearität. Im Zusammenhang dieser Arbeit ist das Ziel, die Vorhersageleistung von ARNNs zu messen. Daraus ergibt sich, dass zur Rechtfertigung des Einsatzes dieses Instruments ein statistischer Nachweis von additiver Nicht-

Linearität gesucht wird. Nur wenige Tests sind ausschließlich gegen additive Nicht-Linearität sensitiv (Hsieh-Test, TLG-Test) und auch für diese Tests kann das Ergebnis durch die Existenz multiplikativer Nicht-Linearität verfälscht sein, da die Verteilung der Teststatistik von Annahmen bezüglich des Fehlerprozesses abhängt (vgl. Lee et al., 1993, S. 288). In dieser Situation ist die Vorgangsweise, zuerst lineare Struktur (durch Schätzung von ARMA) und multiplikative Nicht-Linearität (durch Schätzung von GARCH oder verwandten Modellen) herauszufiltern und dann einen der zur Verfügung stehenden Diagnostik-Tests auf der Residuenzeitreihe anzuwenden, arbeitsintensiv aber prinzipiell empfehlenswert.

Auswahl von Tests auf Nicht-Linearität

Die getroffene Auswahl an Nicht-Linearitätstest versucht den Anforderungen, Klarheit über etwaige additive Nicht-Linearität in makroökonomischen Zeitreihen bzw. in den beiden konkreten untersuchten Zeitreihen zu gewinnen, gerecht zu werden:

- Der McLeod-Li-Test ist ein Test, der sehr große Macht gegen multiplikative und nur geringe Macht gegen additive Nicht-Linearität besitzt. In diesem Sinne kann er als Test gegen multiplikative Nicht-Linearität interpretiert werden. Der McLeod-Li-Test, als einziger Test gegen multiplikative Nicht-Linearität in der getroffenen Auswahl, hilft die anderen Tests zu interpretieren, da er eine Warnung liefert, ob deren Testergebnisse eventuell durch multiplikative Nicht-Linearität beeinflusst werden, bzw. ob im Falle von Diagnostiktests gegebenenfalls auch ein GARCH-Modell vorgeschaltet werden sollte.

- Der Hsieh-Test wurde explizit mit dem Ziel entwickelt, nur auf additive Nicht-Linearität sensibel zu sein. Er ist ein sehr einfach implementierbarer Diagnostiktest und erlaubt eine gut interpretierbare graphische Darstellungsform.

- Der Brock-Dechert-Scheinkmann-Test (BDS-Test), eigentlich ein Test auf unabhängige und identische Verteilung, wird durch Anwendung auf eine Residuenzeitreihe zum Diagnostiktest auf vernachlässigte additive Nicht-Linearität. Aufgrund der großen Beliebtheit des BDS-Tests in der empirischen Forschung ist es möglich, auf zahlreiche Erfahrungen in der richtigen Implementierung und auf vorhandene frei verfügbare Computerprogramme aufzubauen.

- Der Neural-Network-Test (NN-Test) oder seine Weiterentwicklung als Teräsvirta-Lin-Granger-Test (TLG-Test) sind für diese Arbeit besonders aussagekräftig, da sie gegen jenen Typ von Nicht-Linearität sensitiv sind, für den ARNN-Modelle potentielle höhere Prognosekraft haben.

Die verschiedenen ausgewählten Tests sollen einander ergänzen, indem sie von unterschiedlichen Teststrategien Gebrauch machen und verschiedene Spielarten von Nicht-Linearität testen.

McLeod-Li-Test

Wenn in einer Zeitreihe multiplikative Nicht-Linearität auftritt, sind die quadrierten Residuen eines linearen Modells autokorreliert (Granger und Andersen, 1978).[17] Auf dieser Eigenschaft baut der McLeod-Li-Test (McLeod und Li, 1983) auf, um auf multiplikative Nicht-Linearität zu testen. Es seien u_t die N Residuen eines auf der zu untersuchenden Zeitreihe x_t geschätzten linearen Modells. Die empirischen Autokorrelationskoeffizienten zum Lag k von u_t^2 sind gegeben als

$$r_{uu}(k) = \frac{\sum(u_t^2 - \hat{\sigma}^2)(u_{t+k}^2 - \hat{\sigma}^2)}{\sum(u_t^2 - \hat{\sigma}^2)^2} \,, \tag{3.12}$$

wobei $\hat{\sigma}^2$ ein Schätzer für die Varianz von u_t ist. Die Formel für die Teststatistik lautet

$$Q_{uu}(k) = N(N+2)\sum_{i=1}^{k} \frac{1}{N-i} r_{uu}^2(i) \,. \tag{3.13}$$

Unter der Null-Hypothese, dass die u_t^2 iid-verteilt sind, folgt die Teststatistik asymptotisch einer χ^2-Verteilung mit k Freiheitsgraden. Der McLeod-Li-Test entspricht der Durchführung des üblichen Box-Pierce-Tests auf u_t^2 statt u_t. Anstatt die quadrierten Residuen zu verwenden, kann der Test auch durchgeführt werden, indem man die Autokorrelation der Beträge $|u_t|$ untersucht.

Es bestehen enge Zusammenhänge zwischen dem McLeod-Li-Test und dem ARCH-Test von Engle (1982). Während in die Teststatistik des McLeod-Li-Tests die Autokorrelationskoeffizienten von u_t^2 eingehen, verwendet der ARCH-Test die partiellen Autokorrelationskoeffizienten. Ansonsten sind die beiden Tests gleich konstruiert und asymptotisch äquivalent.

Hsieh-Test

Der Hsieh-Test wurde von Hsieh (1989) mit der Zielsetzung entwickelt, ein Verfahren anzubieten, das zwischen additiver und multiplikativer Nicht-Linearität unterscheiden kann. Unter der Null-Hypothese des Tests, dass nicht-lineare Zusammenhänge in einer Zeitreihe x_t nur über die Varianz und nicht über den Mittelwert wirken, gilt, dass die Korrelationkoeffizienten dritter Ordnung der Residuen u_t eines linearen Modells Null sind:

$$\rho_{uuu}(i,j) = \mathrm{E}(u_t, u_{t-i}, u_{t-j})/\mathrm{E}(u_t^3) = 0 \quad \text{für } i,j > 0. \tag{3.14}$$

Es lässt sich leicht ein Schätzer für $\rho_{uuu}(i,j)$ entwickeln:

$$\hat{\rho}_{uuu}(i,j) = \frac{1/N \sum u_t, u_{t-i}, u_{t-j}}{(1/N \sum u_t^2)^{1.5}}. \tag{3.15}$$

[17] Der McLeod-Li-Test kann allerdings auch auf die Originalzeitreihe angewendet werden. Dennoch wird hier der Empfehlung von McLeod und Li (1983) und der dort beschriebenen Vorgangsweise gefolgt und der McLeod-Li-Test als Diagnostiktest verwendet.

Unter der Null-Hypothese nach Gleichung (3.14) und Hilfsannahmen über das Verhalten von u_t ist $\sqrt{N}(1/N)\sum u_t u_{t-i} u_{t-j}$ asymptotisch normalverteilt mit Mittelwert Null und Varianz

$$\omega(i,j) = \operatorname*{plim}_{N\to\infty} \frac{1}{N} \sum u_t^2 u_{t-i}^2 u_{t-j}^2. \tag{3.16}$$

Damit ist $\hat{\rho}_{uuu}(i,j)$ asymptotisch normalverteilt mit Mittelwert Null und Varianz $\omega(i,j)/\sigma_u^6$. Für diesen Ausdruck ist ein konsistenter Schätzer

$$V(i,j) = \big((1/N)\sum u_t^2 u_{t-i}^2 u_{t-j}^2\big) / \big((1/N)\sum u_t^2\big)^3. \tag{3.17}$$

Schließlich ergibt sich die Teststatistik als

$$Q(i,j) = \sqrt{N}\frac{\hat{\rho}_{uuu}(i,j)}{\sqrt{V(i,j)}} \tag{3.18}$$

$Q(i,j)$ ist standardnormalverteilt und erlaubt so ein einfaches Überprüfen der Null-Hypothese. Der Test weist große Ähnlichkeit mit dem Tsay-Test (Tsay, 1986) auf. Ein wichtiger Unterschied besteht allerdings darin, dass der Tsay-Test auf die gemeinsame Verschiedenheit der $\rho_{uuu}(i,j)$ von Null testet, während beim vorliegenden Test für jede Lag-Kombination separat getestet wird. Außerdem gelangt Tsay für seine Teststatistik aufgrund anderer Annahmen bezüglich u_t zu einer Verteilung, die den Test auch sensibel für multiplikative Nicht-Linearität macht.

Brock-Dechert-Scheinkmann-Test

Der Brock-Dechert-Scheinkmann-Test (BDS-Test) (Brock et al., 1987) ist ein Test gegen jede Form von Abhängigkeit in den Daten. Erst in Verbindung mit einem vorgeschalteten linearen Modell kann er als Diagnostiktest auf vernachlässigte Nicht-Linearität angesehen werden. In Verbindung mit einem zusätzlich vorgeschalteten GARCH-Modell kann er als Test auf additive Nicht-Linearität angesehen werden. Die richtige Interpretation der Testergebnisse hängt dann natürlich von der richtigen Spezifizierung der vorgeschalteten Modelle ab.

Der BDS-Test basiert auf dem von Grassberger und Procaccia (1983) entwickelten Konzept des Korrelationsintegrals. Hierbei wird die Zeitreihe x_t in den m-dimensionalen Phasenraum eingebettet, indem m-dimensionale Vektoren $\mathbf{x}_t^m = (x_t, x_{t+1}, \ldots, x_{t+m-1})'$ gebildet werden. Bei einer Länge T der Zeitreihe können $N = T - m + 1$ solche „Historien“ gebildet werden. Nun kann verglichen werden, ob die Distanz zweier ausgewählter Historien mehr als ϵ beträgt. Es können $N(N-1)$ solche Vergleiche angestellt werden. Das Korrelationsintegral $C(\epsilon, m, N)$ ergibt sich als der Anteil der Fälle, in denen dieser Vergleich eine geringere Distanz

als ϵ liefert:

$$C(\epsilon, m, N) = \frac{1}{N(N-1)} \sum_{t \neq \tilde{t}} \mathrm{I}[\mathbf{x}_t^m, \mathbf{x}_{\tilde{t}}^m; \epsilon] \tag{3.19}$$

$$\mathrm{I}[\mathbf{x}_t^m, \mathbf{x}_{\tilde{t}}^m; \epsilon] = \begin{cases} 1 & \text{wenn } \|\mathbf{x}_t^m - \mathbf{x}_{\tilde{t}}^m\| \leq \epsilon \\ 0 & \text{sonst} \end{cases}$$

Als Norm wird die Maximumsnorm verwendet: $\|\mathbf{x}\| = \max |x_i|$. Es lässt sich zeigen, dass unter der Null-Hypothese des BDS-Tests, dass x_t unabhängig und identisch verteilt ist, für $N \to \infty$ gilt:

$$C(\epsilon, m, N) \to C(\epsilon, 1, N)^m \tag{3.20}$$

Davon ausgehend kann die Test-Statistik des BDS-Tests wie folgt definiert werden:

$$Q(\epsilon, m, N) = \sqrt{N} \frac{C(\epsilon, m, N) - C(\epsilon, 1, N)^m}{\hat{\sigma}(\epsilon, m, N)}, \tag{3.21}$$

wobei $\hat{\sigma}(\epsilon, m, N)$ eine Schätzung der asymptotischen Standardabweichung von $C(\epsilon, m, N) - C(\epsilon, 1, N)^m$ ist. Die genaue Formel für $\hat{\sigma}(\epsilon, m, N)$ und ihre Herleitung sowie die Herleitung von $W(\epsilon, m, N)$ insgesamt kann bei Brock et al. (1996) nachgelesen werden. Die Test-Statistik ist unter der Null-Hypothese asymptotisch normalverteilt. Allerdings weicht die Small-Sample-Verteilung der Teststatistik beträchtlich von der asymptotischen Verteilung ab. Daher sollten die kritischen Werte für den Fall $(T - m + 1)/m \leq 200$ aus Tabellen entnommen werden, wie sie bei Brock et al. (1991, S. 232ff) oder Hsieh (1989) zu finden sind.

In der praktischen Anwendung des BDS-Tests muss außerdem die Frage der Wahl von ϵ und m geklärt werden. Für die Wahl von ϵ gibt es aufgrund von Monte-Carlo-Simulationen die dringende Empfehlung sich an der geschätzten Standardabweichung der Zeitreihe, $\hat{\sigma}_x$, zu orientieren. Üblich ist, den Test für mehrere verschiedene ϵ-Werte rund um $\hat{\sigma}_x$ durchzuführen. Meist wird die Test-Statistik auch für verschiedene Werte für m berechnet, so dass der Test für ein Gitter von Werten durchgeführt wird. Es können deshalb Zweideutigkeiten in den Testergebnissen auftreten, wenn für unterschiedliche ϵ- und m-Werte unterschiedlich signifikante Teststatistiken vorliegen. Eine vorsichtige Interpretation ist angebracht.

Neural-Network-Test und Teräsvirta-Lin-Granger-Test

Der Neural-Network-Test (NN-Test) in seiner von White (1989a) entwickelten Form basiert auf der Schätzung eines neuronalen Netzes

$$y_t = \boldsymbol{\alpha}' \tilde{\mathbf{y}}_t + \sum_{k=1}^{K} \beta_k \psi(\boldsymbol{\gamma}_k' \tilde{\mathbf{y}}_t) + u_t \tag{3.22}$$

mit der Aktivierungsfunktion $\psi(\cdot)$ definiert als

$$\psi(x) = (1 + \exp(-x))^{-1} - 1/2. \tag{3.23}$$

Hierbei ist $\tilde{\mathbf{y}}_t = (1, y_{t-1}, \ldots y_{t-p})'$, $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots \alpha_p)'$, $\boldsymbol{\gamma}_j = (\gamma_{j0}, \gamma_{j1}, \ldots \gamma_{jp})'$ und $u_t \sim iid(0, \sigma^2)$. Das neuronale Netz nach (3.22) ist ein Feedforward-Netz mit K verdeckten Neuronen sowie einer direkte Verbindung zwischen den Input-Einheiten und der Output-Einheit. Durch die direkte Verbindung, repräsentiert durch den Term $\boldsymbol{\alpha}'\tilde{\mathbf{y}}_t$ in (3.22), können die linearen Zusammenhänge in der Zeitreihe modelliert werden. Die folgenden Ergebnisse sind mit geringen Änderungen auch für andere Aktivierungsfunktionen als die durch (3.23) gegebene gültig.[18]

Die Null-Hypothese im NN-Test von White (1989a) ist

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_K = 0. \tag{3.24}$$

Man beachte aber, dass eine andere mögliche Null-Hypothese

$$\tilde{H}_0 : \boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2 = \ldots = \boldsymbol{\gamma}_K = \mathbf{0} \tag{3.25}$$

wäre, da auch in diesem Fall der nicht-lineare Teil in (3.22) verschwinden würde. Daher ist unter der Null-Hypothese das Modell (3.22) nicht identifiziert. Das bedeutet, dass die geschätzten Parameter auch asymptotisch keiner Normalverteilung folgen, und dass es unzulässig wäre, Signifikanzniveaus für sie zu berechnen. Im NN-Test wird dieses Problem dadurch gelöst, dass $\boldsymbol{\gamma}_k$ nicht geschätzt, sondern zufällig gewählt wird, wodurch das Modell identifiziert ist. Diese Vorgangsweise hat allerdings gravierende Nachteile, die sich vor allem in der praktischen Anwendung zeigen. Die Macht des Tests hängt stark davon ab, wie effizient die Gewichte der zu den verdeckten Neuronen führenden Schicht zufällig gewählt werden. Eine Verbesserung des Verfahrens kann durch die Durchführung einer Hauptkomponentenanalyse als Zwischenschritt im Schätzverfahren erreicht werden. Der grundsätzliche Mangel, dass Testergebnisse nicht reproduzierbar sind, bleibt jedoch bestehen: die mehrmalige Anwendung des Testverfahrens auf die gleiche Zeitreihe kann zufallsbedingt höchst unterschiedliche Ergebnisse liefern.

Eine Neuformulierung des NN-Tests, die das Problem der Nicht-Identifikation von Grund auf vermeidet, wurde von Teräsvirta et al. (1993) entwickelt. Dieser Test, der auch als NN-Test von Teräsvirta bekannt ist, im folgenden jedoch als Teräsvirta-Lin-Granger-Test (TLG-Test) bezeichnet wird, beruht auf der bloßen Approximation eines Feedforward-Netzes mit einem einzigen verdeckten Neuron

$$y_t = \boldsymbol{\alpha}'\tilde{\mathbf{y}}_t + \beta\psi(\boldsymbol{\gamma}'\tilde{\mathbf{y}}_t) + u_t, \tag{3.26}$$

[18] Zu den Grundlagen von Feedforward-Netzen und zu den speziell für die Modellierung von Zeitreihen entwickelten ARNN-Modellen siehe Abschnitt 4.1.

indem $\psi(\boldsymbol{\gamma}'\tilde{\mathbf{y}}_t)$ im Punkt $\boldsymbol{\gamma} = \mathbf{0}$ in eine Taylorreihe dritter Ordnung entwickelt wird:

$$\begin{aligned}\psi(\boldsymbol{\gamma}'\tilde{\mathbf{y}}_t) \approx \tilde{\psi}(\boldsymbol{\gamma}'\tilde{\mathbf{y}}_t) &= \psi(0) + \sum_{i=0}^{p} \frac{\partial\psi(0)}{\partial\gamma_i}\gamma_i + \frac{1}{2}\sum_{i=0}^{p}\sum_{j=0}^{p} \frac{\partial^2\psi(0)}{\partial\gamma_i\partial\gamma_j}\gamma_i\gamma_j \\ &\quad + \frac{1}{6}\sum_{i=0}^{p}\sum_{j=0}^{p}\sum_{k=0}^{p} \frac{\partial^3\psi(0)}{\partial\gamma_i\partial\gamma_j\partial\gamma_k}\gamma_i\gamma_j\gamma_k\end{aligned} \tag{3.27}$$

Nach algebraischer Auswertung der verschiedenen Ableitungen in (3.27) ergibt sich

$$y_t = \tilde{\boldsymbol{\alpha}}'\tilde{\mathbf{y}}_t + \sum_{i=1}^{p}\sum_{j=i}^{p} \delta_{ij} y_{t-i} y_{t-j} + \sum_{i=1}^{p}\sum_{j=i}^{p}\sum_{k=j}^{p} \delta_{ijk} y_{t-i} y_{t-j} y_{t-k} + \tilde{u}_t \tag{3.28}$$

wobei $\delta_{ij} = g_{ij}\beta\gamma_i\gamma_j\gamma_0$ mit $g_{ij} = 1/36$ wenn $i = j$ und $g_{ij} = 1/18$ sonst, $i, j \geq 1$, sowie $\delta_{ijk} = g_{ijk}\beta\gamma_i\gamma_j\gamma_k$ mit $g_{ijk} = 1/36$ wenn $i = j = k$, $g_{ijk} = 1/18$ wenn $i = j$, $i = k$ oder $j = k$ und $g_{ijk} = 1/6$ sonst, $i, j, k \geq 1$. Die Null-Hypothese im TLG-Test ist nun

$$\tilde{\tilde{H}}_0 : \delta_{ij} = 0, \delta_{ijk} = 0 \qquad i = 1, \ldots, p; j = i, \ldots, p; k = j, \ldots, p. \tag{3.29}$$

Zur Testung der Hypothese wird zunächst y_t auf $1, y_{t-1}, \ldots y_{t-p}$ regressiert und die Residuenzeitreihe $\hat{u}_t$ dieser Regression ermittelt. Dann regressiert man $\hat{u}_t$ auf m Hilfsregressoren, die den nicht-linearen Termen in Gleichung (3.28) entsprechen, und erhält die Residuenzeitreihen $\hat{v}_t$. Die Teststatistik kann als F- oder als χ^2-Statistik definiert werden, die asymptotisch äquivalent sind. Die F-Statistik lautet

$$F = \frac{(SSR_0 - SSR)/m}{SSR/(T - p - 1 - m)}, \tag{3.30}$$

wobei $SSR_0 = \sum \hat{u}_t^2$ und $SSR = \sum \hat{v}_t^2$ Residuensumme der beiden Hilfsregressionen sind.

Neben der hier beschriebenen Version des TLG-Tests existieren noch Modifikationen, die sich z.B. durch Weglassung der Terme 2. Ordnung in (3.28) unterscheiden. Diese Varianten haben gegen spezifische Formen der Nicht-Linearität erhöhte Macht, gegen andere niedrigere. Im angewandten Teil dieses Kapitels wird nur von der hier beschriebenen Version Gebrauch gemacht, da diese am ausführlichsten spezifiziert und am direktesten aus dem Konzept des Feedforward-Netzes abgeleitet ist. Natürlich hat der TLG-Test gegenüber dem NN-Test von White den Nachteil, dass „nur“ eine Approximation eines Feedforward-Netzes Verwendung findet. In Monte-Carlo-Simulationen konnte jedoch nachgewiesen werden, dass dieser Nachteil keine Rolle spielt und dass sogar für Prozesse, die durch ein Feedforward-Netz mit einem oder mehreren verdeckten Neuronen generiert wurden, der TLG-Test an Macht den NN-Test von White übertrifft.

3.3 Ergebnisse für die beiden Beispielzeitreihen

In diesem Abschnitt werden die Ergebnisse der ausgewählten Tests auf Nicht-Linearität für die beiden Zeitreihen Arbeitslosenrate (ALR) und Industrieproduktionsindex (IPI) diskutiert. In diesem Zusammenhang werden auch etwaige Modifikationen der Tests erläutert, die hinsichtlich der Besonderheiten der beiden Zeitreihen angebracht schienen. Die Vorgangsweise und die Ergebnisse sind für beide Zeitreihen in wesentlichen Aspekten ähnlich, so dass meist eine gemeinsame Besprechung möglich ist.

Die untersuchten Zeitreihen werden, aufgrund der Ergebnisse des Kapitels über die lineare Modellierung (Unit-Root-Tests), in erster Linie in ihren saisonalen Differenzen $\nabla_{12}x_t$ untersucht. Um die Robustheit der Testergebnisse hinsichtlich der Alternative des konventionellens Differenzierens und des kombinierten konventionellen und saisonalen Differenzierens zu überprüfen, wird jedoch am Ende des Abschnitts auch kurz auf die Ergebnisse anhand ∇x_t und $\nabla_{12}\nabla x_t$ eingegangen.

Einige der besprochenenen Tests setzen das Vorliegen einer Residuenzeitreihe voraus bzw. können zusätzlich zur Originalzeitreihe auch auf diese angewendet werden. Wenn nicht ausdrücklich anders erwähnt, finden in diesem Falle die Residuen eines ausführlich spezifizierten AR-Modells (ALR: Ordnung $p = 26$, kein Interzept; IPI: Ordnung $p = 49$, mit Interzept) Verwendung. Zusätzlich wird gegebenenfalls auch eine Bereinigung der Residuen um ARCH-Effekte vorgenommen.

McLeod-Li-Test und ARCH-Effekte

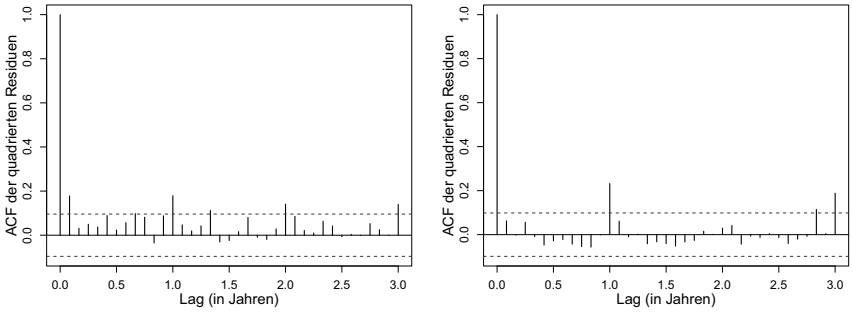
Beide untersuchten Zeitreihen weisen sehr deutlich multiplikative Nicht-Linearitäten auf, die zufriedenstellend durch ein ARCH-Modell modelliert werden können. Im folgenden wird daher, etwas unscharf, der Ausdruck ARCH-Effekte verwendet.[19] Dies sei zunächst anhand eines sehr aussagekräftigen Instruments, der ACF der quadrierten Residuen des geschätzten linearen Modells illustriert (Abb. 6).

In der Abbildung werden die quadrierten Residuen eines ausführlich spezifizierten AR-Modells der Ordnung 26 bzw. 49 verwendet. Die gleiche Untersuchung für die Residuen des entsprechenden ARMA-Modells bzw. des (auf der konventionell differenzierten Zeitreihe geschätzten) ARDS-Modells ergibt kein wesentlich anderes Bild, weshalb auf eine eigene Abbildung verzichtet wurde.

Die ACF der quadrierten Residuen weist in beiden Fällen signifikante Werte zum Lag 1 und zu den saisonalen Lags auf. Wie zu erwarten verwirft der McLeod-Li-Test für beide Zeitreihen klar die Hypothese der Unabhängigkeit der quadrierten Residuen (Tab. 9).

Es kann versucht werden, die Residuenzeitreihen durch die Schätzung eines ARCH-Modells um die ARCH-Effekte zu bereinigen. Die neuen Residuen ergeben sich dann als $v_t = u_t/\sqrt{\hat{z}_t}$, wobei u_t die ursprünglichen Residuen sind und $\hat{z}_t$ die

[19] Von einem unscharfen Ausdruck kann deswegen gesprochen werden, weil auch andere Formen multiplikativer Nicht-Linearität existieren (vgl. Abschnitt 3.1).



(a) ALR (b) IPI

Abb. 6: Autokorrelationsfunktion der quadrierten Residuen eines ausführlich spezifizierten AR-Modells auf der saisonal differenzierten Zeitreihe: (a) Arbeitslosenrate (ALR) und (b) Industrieproduktionsindex (IPI)

Tab. 9: Ergebnisse des McLeod-Li-Tests auf Residuen eines AR-, ARMA- bzw. ARDS-Modells, ALR und IPI

		ALR	
k	AR-Modell	ARMA-Modell	ARDS-Modell
6	18.8079 (0.0045)**	21.5804 (0.0014)**	12.1394 (0.0589)
12	44.0738 (0.0000)***	48.7818 (0.0000)***	42.0497 (0.0000)***
24	63.3971 (0.0000)***	65.9125 (0.0000)***	76.0062 (0.0000)***
36	78.6484 (0.0000)***	81.1913 (0.0000)***	130.5916 (0.0000)***
		IPI	
k	AR-Modell	ARMA-Modell	ARDS-Modell
6	12.4233 (0.0532)	17.7143 (0.0070)**	10.1318 (0.1192)
12	35.4352 (0.0004)***	31.1887 (0.0018)**	20.7235 (0.0546)
24	40.3831 (0.0195)*	35.5074 (0.0612)	28.9228 (0.2231)
36	68.1461 (0.0009)***	61.1856 (0.0055)**	79.2751 (0.0000)***

Die Teststatistik ist χ^2-verteilt mit k Freiheitsgraden. Wahrscheinlichkeitswerte in Klammern. ***, ** und * bezeichnen Werte, die bei einem Konfidenzniveau von 0.1 %, 1 % bzw. 5 % signifikant sind.

Tab. 10: Struktur der auf Residuen geschätzten ARCH-Modelle, ALR und IPI

Residuen des	ALR	IPI
AR-Modells	α_0 (=Interzept), α_1, α_{12}	α_0 (=Interzept), α_1, α_{12}
ARMA-Modells	α_0 (=Interzept), α_1, α_2, α_{12}	α_0 (=Interzept), α_1, α_{12}
ARDS-Modells	α_0 (=Interzept), α_1, α_2, α_{12}	α_0 (=Interzept), α_1, α_{12}

geschätzte bedingte Varianz des Prozesses (vgl. Gleichung (3.8) in Abschnitt 3.1). Diese Vorgangsweise liefert auf den beiden Zeitreihen zufriedenstellende Ergebnisse, d.h. der McLeod-Li-Test zeigt auf den neuen Residuen in keinem der Fälle signifikante Werte an, solange man k geringer als die Ordnung des linearen Modells wählt.

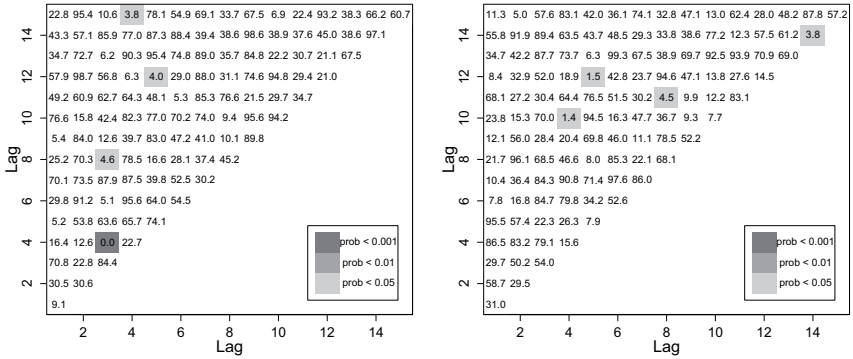
Die Struktur der geschätzten ARCH-Modelle ist in Tab. 10 wiedergegeben. Auf eine genaue Angabe der geschätzten Werte wird verzichtet, da die ARCH-Modellierung in dieser Arbeit dem Ziel dient, bereinigte Residuenzeitreihe zu liefern und nur begrenzt versucht wurde, ein optimales Modell zu finden.

Insgesamt liefert diese Vorgangsweise einen starken statistischen Hinweis, dass sowohl in ALR als auch in IPI eine multiplikative nicht-lineare Struktur vom ARCH-Typus vorhanden ist. Da der McLeod-Li-Test auch gegen bestimmte Formen additiver Nicht-Linearität sensibel ist, handelt es sich aber um keinen statistischen Test auf multiplikative Nicht-Linearität. Jedenfalls erscheint es aufgrund dieses Ergebnisses angebracht, im Rahmen der Diagnostiktests auf additive Nicht-Linearität (Hsieh-Test, BDS-Test) um ARCH-Effekte bereinigte Residuen zu verwenden.

Hsieh-Test

Der Hsieh-Test untersucht die dritten Momente einer Zeitreihe und wird jeweils für eine Kombination von zwei Lags durchgeführt. Für die beiden Zeitreihen ALR und IPI ergeben sich aufgrund des Tests teilweise widersprüchliche und schwer interpretierbare Ergebnisse. Aufgrund des saisonalen Charakter und der hohen Ordnung des vorgeschalteten AR- bzw. ARMA-Modells ist es notwendig, bei der Durchführung des Hsieh-Tests höhere Lags zu berücksichtigen. Nun zeigt sich, dass sich unter den zahlreichen möglichen Lag-Kombinationen vereinzelt signifikante Teststatistiken finden, ohne dass ein Muster erkennbar wäre. Hinzu kommt, dass die Testergebnisse leicht variieren, je nachdem ob die Residuen des AR-, des ARMA- oder des ARDS-Modells herangezogen werden.

Um die Ergebnisse des Hsieh-Test für eine größere Anzahl möglicher Lag-Kombinationen übersichtlich darzustellen, wurde zu einer graphischen Darstellung der Ergebnisse gegriffen. In Abb. 7 wurden die Wahrscheinlichkeitswerte der Teststatistik unter der Null-Hypothese, dass die Residuen des linearen Modells iid-verteilt sind, eingetragen und zusätzlich durch Grauschattierungen kodiert.



(a) ALR (b) IPI

Abb. 7: Ergebnisse des Hsieh-Tests für die Residuen eines ausführlich spezifizierten AR-Modells auf der saisonal differenzierten Zeitreihe (Wahrscheinlichkeitswerte der Teststatistik in Prozent): (a) Arbeitslosenrate ALR und (b) Industrieproduktionsindex (IPI)

In der Abbildung sind die Ergebnisse für die beiden Zeitreihen ALR und IPI jeweils anhand der Residuen des (ausführlich spezifizierten) AR-Modells angeführt. Im Vergleich dazu weisen die hier nicht wiedergegebenen Ergebnisse anhand der Residuen des ARMA-Modells einige zusätzliche signifikante Werte auf, es ergeben sich jedoch immer für jene Lag-Kombinationen signifikante Werte, für die sie auch anhand der AR-Residuen signifikant waren. Dies ist plausibel dadurch zu erklären, dass ein ausführliches AR-Modell lineare Zusammenhänge vollständiger herausfiltert als ein (richtig spezifiziertes) ARMA-Modell. Weiters sind die Testergebnisse auch hinsichtlich einer vorgeschalteten Bereinigung der Residuenzeitreihen um ARCH-Effekte robust, was allerdings aufgrund der Konstruktion des Tests zu erwarten ist.

Aufgrund der Testergebnisse scheint der Schluss auf vorhandene Nicht-Linearitäten nicht zulässig, da bei der 120-maligen Durchführung eines Tests erwartet werden darf, dass in sechs Fällen das 5 %-Signifikanzniveau unterschritten wird. Wenn also in einer der beiden untersuchten Zeitreihen Nicht-Linearitäten eine Rolle spielen, dann sind sie von einer Art, die nicht in den dritten Momenten der Zeitreihe sichtbar wird.

Brock-Dechert-Scheinkman-Test

Die Ergebnisse des BDS-Tests weisen auf Nicht-Linearitäten in beiden untersuchten Zeitreihen hin. Die Tabellen 11 und 12 listen die Ergebnisse für die beiden untersuchten Zeitreihen auf. Da der BDS-Test auch auf multiplikative Nicht-Linearität

anspricht, wurden jeweils nur die um ARCH-Effekte bereinigten Residuen herangezogen. Empfehlungen in der Literatur folgend wurde ϵ auf 0.5σ, σ, 1.5σ bzw. 2σ gesetzt, wobei σ die Standardabweichung der zu untersuchenden Zeitreihe ist. Die Einbettungsdimension wurde angesichts der Ergebnisse der linearen Zeitreihenanalyse zwischen $m = 1$ und dem relativ hohen Wert $m = 12$ gewählt.

Tab. 11: Ergebnisse des BDS-Tests auf ARCH-bereinigten Residuen eines ausführlich spezifizierten AR-Modells, ALR

m	$\epsilon = 0.4769$	$\epsilon = 0.9538$	$\epsilon = 1.4308$	$\epsilon = 1.9077$
2	-1.6843 (0.0921)	-1.1320 (0.2577)	-1.4424 (0.1492)	-1.3954 (0.1629)
3	-2.6739 (0.0075)**	-2.1472 (0.0318)*	-2.5500 (0.0108)*	-2.3765 (0.0175)*
4	-2.7629 (0.0057)**	-1.9367 (0.0528)	-2.1986 (0.0279)*	-2.1601 (0.0308)*
5	-2.9087 (0.0036)**	-2.1769 (0.0295)*	-2.3481 (0.0189)*	-2.2648 (0.0235)*
6	-2.6647 (0.0077)**	-2.2245 (0.0261)*	-2.2528 (0.0243)*	-2.2205 (0.0264)*
7	-3.0668 (0.0022)**	-2.4007 (0.0164)*	-2.2775 (0.0228)*	-2.2996 (0.0215)*
8	-3.1796 (0.0015)**	-2.5877 (0.0097)**	-2.3107 (0.0208)*	-2.3721 (0.0177)*
9	-3.3921 (0.0007)***	-2.5305 (0.0114)*	-2.2801 (0.0226)*	-2.3602 (0.0183)*
10	-2.7988 (0.0051)**	-2.4389 (0.0147)*	-2.2082 (0.0272)*	-2.3136 (0.0207)*
11	-2.3469 (0.0189)*	-2.1779 (0.0294)*	-2.0398 (0.0414)*	-2.1024 (0.0355)*
12	-1.9939 (0.0462)*	-2.0585 (0.0395)*	-1.9637 (0.0496)*	-1.9691 (0.0489)*

***, ** und * bezeichnen Werte, die bei einem Konfidenzniveau von 0.1 %, 1 % bzw. 5 % signifikant sind.

Tab. 12: Ergebnisse des BDS-Tests auf ARCH-bereinigten Residuen eines ausführlich spezifizierten AR-Modells, IPI

m	$\epsilon = 0.4900$	$\epsilon = 0.9799$	$\epsilon = 1.4699$	$\epsilon = 1.9598$
2	-0.2157 (0.8292)	-1.1272 (0.2597)	-1.1477 (0.2511)	-1.3106 (0.1900)
3	-0.4434 (0.6575)	-1.3545 (0.1756)	-1.3733 (0.1696)	-1.3752 (0.1691)
4	-0.4979 (0.6185)	-0.9570 (0.3385)	-0.9907 (0.3218)	-1.0906 (0.2755)
5	0.1092 (0.9130)	-0.8094 (0.4183)	-1.0247 (0.3055)	-1.2297 (0.2188)
6	0.7139 (0.4753)	-1.0057 (0.3146)	-1.1883 (0.2347)	-1.4852 (0.1375)
7	2.8296 (0.0047)**	-1.0757 (0.2821)	-1.3301 (0.1835)	-1.6618 (0.0966)
8	7.6781 (0.0000)***	-1.2738 (0.2027)	-1.4470 (0.1479)	-1.8189 (0.0689)
9	7.7199 (0.0000)***	-1.3742 (0.1694)	-1.5899 (0.1119)	-1.9853 (0.0471)*
10	7.8038 (0.0000)***	-1.4311 (0.1524)	-1.6431 (0.1004)	-2.0877 (0.0368)*
11	-3.1053 (0.0019)**	-1.3220 (0.1862)	-1.7577 (0.0788)	-2.2635 (0.0236)*
12	-2.6563 (0.0079)**	-1.1420 (0.2535)	-1.8244 (0.0681)	-2.3570 (0.0184)*

***, ** und * bezeichnen Werte, die bei einem Konfidenzniveau von 0.1 %, 1 % bzw. 5 % signifikant sind.

Für die Arbeitslosenrate ergeben sich hochsignifikante Werte für alle Einbettungsdimensionen ab $m = 3$ und für alle ϵ-Werte, insbesondere jedoch für den kleinsten der vier ϵ-Werte. Dieses Ergebnis ist einigermaßen robust hinsichtlich Variationen des zugrundeliegenden linearen Modells und der Methode der Bereinigung

um ARCH-Effekte. Für die Residuen des ARMA-Modells ergeben sich ebenso signifikante Werte, wobei allerdings nur ein Signifikanzniveau von etwa 1 % erreicht wird.

Für den Industrieproduktionsindex liefert der BDS-Test signifikante Werte für eine Einbettungsdimension $m = 7, \ldots, 12$. Auch diese Ergebnisse sind robust hinsichtlich der Methode der Gewinnung der Residuenzeitreihe. In Tab. 12 werden die Ergebnisse für die Residuen des AR-Modells dargestellt. Legt man das ARMA-Modell zugrunde, unterscheiden sich die Testergebnisse nur unwesentlich.

In beiden Zeitreihen spielen Nicht-Linearitäten eine Rolle, wobei die statistische Beweiskraft für die Arbeitslosenrate bedeutend stärker ist. Die Ergebnisse hängen jedoch stark davon ab, inwiefern durch das lineare Modell bzw. das ARCH-Modell lineare Strukturen und multiplikative Nicht-Linearitäten herausgefiltert wurden und sind daher vorsichtig zu interpretieren. Auch das Problem des Strukturbruchs, das in Abschnitt 3.4 erläutert wird, kann für die Erklärung signifikanter Werte des BDS-Tests relevant sein.

Teräsvirta-Lin-Granger-Test

Der TLG-Test ist in der bei Teräsvirta et al. (1993) beschriebenen Version auf makroökonomische Zeitreihen, die typischerweise durch hohe Modellordnung charakterisiert sind, kaum anwendbar. Daher wird im folgenden eine Modifikation des Tests verwendet und kurz skizziert. Eine detailliertere Beschreibung der Vorgangsweise findet sich bei Koller und Fischer (2002).

In der Beschreibung des TLG-Tests (Abschnitt 3.2) wurde, der Literatur folgend, davon ausgegangen, dass der Lag-Parameter p des Tests sowohl die Ordnung des im Rahmen des Tests geschätzten linearen Modells bezeichnet, als auch die Anzahl der Lags, die in den Termen zweiter und dritter Ordnung des nicht-linearen Modells berücksichtigt werden sollen.

Für Modelle hoher Ordnung ist diese Vorgangsweise allerdings problematisch, da leicht eingesehen werden kann, dass die Anzahl der zu berücksichtigenden Terme im Rahmen des nicht-linearen Modells sehr rasch mit p wächst. Daher wurde für diese Untersuchung die folgende Modifikation des Tests vorgenommen, die die Eigenschaften der Teststatistik nicht beeinträchtigt. Statt im Rahmen des nicht-linearen Teils des Modells alle Lags $1, \ldots, p$ zu berücksichtigen, kann eine Auswahl von Lags $\mathbf{q} = (q_1, \ldots, q_k)'$, $q_1, \ldots, q_k \leq p$ angegeben werden. Die zu berücksichtigenden Lags können etwa aufgrund von Vorwissen über die Zeitreihe ausgewählt werden. Eine naheliegende Auswahl besteht, darin für $\mathbf{q}$ nur ein bis zwei wichtige Lags zu nehmen.

Der so modifizierte TLG-Test ergibt für beide Zeitreihen ALR und IPI für beinahe alle „sinnvolle“ Einstellungen von p und $\mathbf{q}$ signifikante Werte. Eine Auswahl dieser Ergebnisse wird in Tab. 13 zusammengestellt.

Tab. 13: Ergebnisse des modifizierten TLG-Tests für ausgewählte Testeinstellungen, ALR und IPI

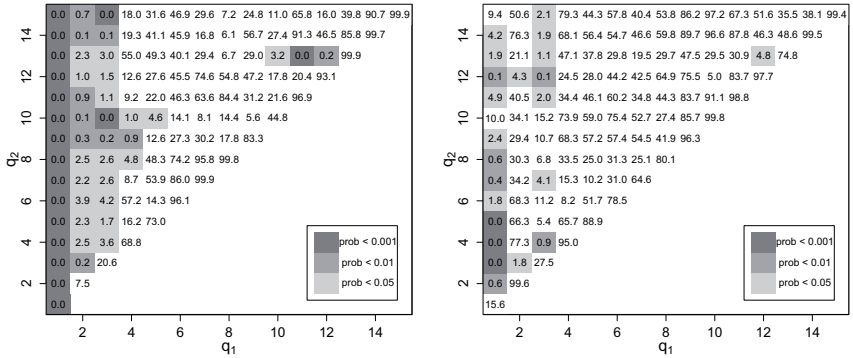
p	$\mathbf{q}'$	ALR	IPI
12	(1)	12.4307 (0.0020)**	13.1322 (0.0014)**
12	(2)	6.3697 (0.0414)*	0.4977 (0.7797)
12	(12)	0.7505 (0.6871)	1.2595 (0.5327)
12	(1, 2, 12)	37.6097 (0.0017)**	36.1994 (0.0027)**
12	(1, 2, 6, 12)	54.2729 (0.0043)**	50.7134 (0.0104)*
24	(1)	19.1210 (0.0001)***	10.7153 (0.0047)**
24	(2)	8.1591 (0.0169)*	1.0715 (0.5852)
24	(12)	4.4977 (0.1055)	0.4149 (0.8127)
24	(1, 2, 12)	40.0049 (0.0008)***	32.3539 (0.0090)**
24	(1, 2, 6, 12)	50.4484 (0.0111)*	46.7177 (0.0265)*
36	(1)	28.7881 (0.0000)***	10.6114 (0.0050)**
36	(2)	12.8805 (0.0016)**	0.8425 (0.6562)
36	(12)	2.4433 (0.2947)	1.6414 (0.4401)
36	(1, 2, 12)	53.3485 (0.0000)***	33.3977 (0.0065)**
36	(1, 2, 6, 12)	61.2999 (0.0006)***	50.7968 (0.0102)*

***, ** und * bezeichnen Werte, die bei einem Konfidenzniveau von 0.1 %, 1 % bzw. 5 % signifikant sind.

Die Testergebnisse belegen, dass in beiden untersuchten Zeitreihen starke additive Nicht-Linearitäten existieren. Für die Arbeitslosenrate dürften sie eine stärkere Rolle spielen als für den Industrieproduktionsindex. Insbesondere wenn im Rahmen des nicht-linearen Teils des im Rahmen des TLG-Tests geschätzten Modells mehrere Lags einbezogen werden, verwirft der Test die Null-Hypothese der Linearität mit großer Sicherheit.

Um gezielt der Frage nachzugehen, welche Lags für die Nicht-Linearität besonders relevant sind, wurde der Test für alle möglichen Kombinationen von jeweils bloß zwei Lags, q_1 und q_2 durchgeführt. Die Ordnung des linearen Modells p wurde hierbei mit 36 gewählt. Durch die entsprechende graphische Darstellung werden die Testergebnisse so unmittelbar mit denen des Hsieh-Tests vergleichbar.

Wie Abb. 8 zeigt, genügt es bereits, eine Kombination von nur zwei Lags in das nicht-lineare Modell, das im Rahmen des TLG-Tests geschätzt wird, aufzunehmen, um das Vorhandensein von Nicht-Linearitäten nachzuweisen. Besonders relevant sind in diesem Sinne bei der ALR die Lags 1, 2, 3, 10, und 13, beim IPI die Lags 1, 3 und 12. Der Vergleich mit dem Hsieh-Test demonstriert die Überlegenheit des TLG-Tests beim Nachweis additiver Nicht-Linearität. Diese Überlegenheit dürfte zum einen in der unterschiedlichen Null-Hypothese der beiden Tests liegen zum anderen in der unterschiedlichen Macht der Tests in Bezug auf ihre jeweilige Null-Hypothese.



(a) ALR (b) IPI

Abb. 8: Ergebnisse des TLG-Tests für die Zeitreihen ALR (a) und IPI (b) bei Beschränkung des nicht-linearen Teils des Tests auf nur zwei Lags, q_1 und q_2: Wahrscheinlichkeitswerte der Teststatistik in Prozent

Testergebnisse auf den anders differenzierten Zeitreihen

Die bisherige Diskussion der Ergebnisse der Tests auf Nicht-Linearität beschränkte sich, wenn nicht anders vermerkt, auf die saisonal differenzierte Zeitreihe, $\nabla_{12}x_t$, bzw. auf Residuenzeitreihen von auf $\nabla_{12}x_t$ geschätzten Modellen. Um die Robustheit der Testergebnisse zu überprüfen wurden die Testverfahren auch auf $\nabla_1 x_t$ und $\nabla_{12}\nabla_1 x_t$ angewandt. Angesichts der auch auf den anders differenzierten Zeitreihen durchwegs insignifikanten Testergebnisse für den Hsieh-Test wird im folgenden nur auf die Ergebnisse des BDS-Tests und des NN-Tests eingegangen.

Für die Arbeitslosenrate zeigt der BDS-Test für die ARCH-bereinigten Residuen von auf $\nabla_1 x_t$ und $\nabla_{12}\nabla_1 x_t$ geschätzten Modellen durchwegs (für beinahe alle m und ϵ) signifikante Werte an. Im zweiten Fall sind im Vergleich die verzeichneten Signifikanzwerte besonders hoch. Auch der TLG-Test liefert sowohl für $\nabla_1 x_t$ und $\nabla_{12}\nabla_1 x_t$ deutliche Hinweise auf additive Nicht-Linearität (relevante Lags für $\nabla_1 x_t$: 1, 3, 4, 7, 13 und für $\nabla_{12}\nabla_1 x_t$: 1, 3, 12).

Für den Industrieproduktionsindex ergibt der BDS-Test sowohl für die ARCH-bereinigten Residuen von auf $\nabla_1 x_t$ und $\nabla_{12}\nabla_1 x_t$ geschätzten Modellen ca. ab $m = 9$ signifikante Testergebnisse, wobei das Muster der Abhängigkeit der Signifikanz von m und ϵ ähnlich jenem in Tab. 12 ist. Auch der TLG-Test liefert für $\nabla_1 x_t$ und $\nabla_{12}\nabla_1 x_t$ im Vergleich zur saisonal differenzierten Zeitreihe, $\nabla_{12}x_t$, etwas schwächere Evidenz für additive nicht-lineare Struktur (für $\nabla_1 x_t$ keine speziell relevante Lags, für $\nabla_{12}\nabla_1 x_t$ relevante Lags 1, 3 und 12). Insgesamt kann damit der Nachweis additiver Nicht-Linearität als relativ robust für die Methode des Differenzierens gelten.

3.4 Strukturbrüche und scheinbare Nicht-Linearitäten

Als eine für die richtige Diagnose von Nicht-Linearität sehr wichtige Problematik ist die Möglichkeit der Instabilität der Parameter des Modells bzw. des Prozesses anzusehen. Koop und Potter (2001) betonen, dass scheinbar gefundene Nicht-Linearität eventuell durch strukturelle Veränderungen erklärt werden kann. Man kann Modelle unterscheiden, bei denen die strukturelle Änderung zu einem bestimmten Zeitpunkt stattfindet, und solche bei denen sie kontinuierlich stattfindet, wobei letzteres eine Verallgemeinerung von ersterem ist. Wenn etwa ein oder mehrere Parameter eines AR-Modells sich im Zeitverlauf zufällig ändern (vgl. Nicholls und Quinn, 1982), so ist das Modell in jedem einzelnen Zeitpunkt linear. Dennoch können statistische Tests Nicht-Linearität anzeigen. Im Falle des BDS-Tests ist das offensichtlich: die Null-Hypothese der iid-Verteilung wird abgelehnt, weil die Residuenzeitreihe nicht identisch (eventuell aber unabhängig) verteilt ist. Auch die Aussagekraft des NN-Tests wird durch Existenz von Strukturbrüchen eingeschränkt, da das im Testverfahren implementierte nicht-lineare Modell, die Gelegenheit bekommt, zufällige Strukturen anzupassen („Overfitting-Problem“). Im übrigen ist bei den vielen denkbaren Modellen, die variable Parameter beinhalten, eine scharfe Abgrenzung zwischen (additiv) nicht-linearen Modellen und linearen Modellen mit zufällig variablen Koeffizienten schwer und unter anderem davon abhängig, ob die Veränderung der Parameter auch endogen von den Daten mitbestimmt wird.

Um die Möglichkeit struktureller Veränderungen in den Zeitreihen ansatzweise zu untersuchen und für die spätere Analyse im Auge zu behalten, sollen die beiden Beispielzeitreihen auf die Möglichkeit eines Strukturbruchs getestet werden. Es können u.a. zwei verschiedene Vorgangsweisen gewählt werden. Erstens kann man auf Strukturbruch hinsichtlich des Interzepts testen, indem man eine Dummyvariable in das lineare Modell einführt, die den Wert Null vor einem bestimmten Zeitpunkt, der als Strukturbruch in Frage kommt, annimmt und den Wert Eins danach. Der t-Wert dieser zusätzlich eingeführten Variable kann als einfacher Test für die Stabilität des Interzepts im Zeitverlauf verwendet werden. Zweitens kann man auf Strukturbruch hinsichtlich aller Koeffizienten testen, indem man für jeden Regressor (inklusive Interzept) im restringierten (ursprünglichen) Modell in einem erweiterten (alternativen) Modell einen zusätzlichen Regressor einführt, der durch Multiplikation des Regressors mit der erwähnten Dummyvariable gebildet wird. Die Definition der Teststatistik für den Test lautet

$$F = \frac{(RSS - ESS)/k}{ESS/(n-2k)}, \tag{3.31}$$

wobei RSS die Summe der quadrierten Residuen des restringierten Modells, ESS die Summe der quadrierten Residuen des erweiterten Modells, k die Anzahl der Koeffizienten im restringierten Modell und n die Anzahl der Residuen ist. Mit Hilfe des Tests wird getestet, ob irgend einer der Koeffizienten sich im Zeitverlauf ändert. Eine asymptotisch äquivalente Vorgangsweise beruht auf Schätzung von

Tab. 14: Ergebnisse der Tests auf Strukturbruch bei einem angenommenen Strukturbruch im Jänner 1975, ARL und IPI

AR-Modell mit Strukturbruch im Interzept		
	ALR	IPI
Interzept vor 1975/1	-0.00654	0.03074
Interzept ab 1975/1	0.00976	0.01426
t-Stat. (prob-Wert)	2.442 (0.0151)*	-3.756 (0.0002)***
AR-Modell mit Strukturbruch in allen Koeffizienten		
	ALR	IPI
$\hat{\sigma}^2$(restringiert)	0.00295	0.000572
$\hat{\sigma}^2$(unrestringiert)	0.00272	0.000489
F-Stat. (prob-Wert)	2.6558 (0.0014)**	4.8571 (0.000)***
Freiheitsgr.	13, 392	13, 369

Wahrscheinlichkeitswerte in Klammern; ***, ** und * bezeichnen t bzw. F-Werte, die bei einem Konfidenzniveau von 0.1 %, 1 % bzw. 5 % signifikant sind.

zwei getrennten Modellen für die durch den Strukturbruch getrennten Zeiträume. Testverfahren nach diesem Prinzip sind unter dem Namen Chow-Test auf Strukturbruch bekannt (Chow, 1983). Es existieren verschiedene ausgefeilte Verfahren, unter anderem solche, die den Zeitpunkt des Strukturbruchs optimal bestimmen (z.B. Andrews, 1993). Für den Zweck dieser Arbeit erscheint jedoch ein einfaches Verfahren ausreichend.

Tabelle 14 enthält die Ergebnisse für diese beiden Vorgangsweisen anhand der saisonal differenzierten Zeitreihen. Als Zeitpunkt für den Strukturbruch wurde hierbei aufgrund der visuellen Auswertung der Zeitreihe der Jänner 1975 gewählt. Ausgangspunkt der Tests stellt das sparsam spezifizierte AR-Modell aus Abschitt 2.4 dar (Tab. 3 und 4). Um einen fairen Test zu erreichen wurde das AR-Modell im Falle der ALR noch um ein Interzept erweitert, obwohl dieses im restringierten Modell nicht signifikant ist.

Die Ergebnisse weisen in beiden Zeitreihen auf einen statistisch signifikanten Strukturbruch hin. Ob dieser tatsächlich im Jahre 1975 stattgefunden hat sei dahingestellt. In beiden Fällen ist das Interzept vor und nach dem Strukturbruch signifikant unterschiedlich. Hierbei ist zu beachten, dass sich das in der Tabelle wiedergegebene Interzept für den Zeitraum ab 1975/1 als die Summe des Interzepts für den Zeitraum vor 1975/1 und des Koeffizienten für die zusätzlich eingeführte Variable errechnet. Auch der Chow-Test auf Strukturbruch liefert signifikante Ergebnisse. Die genaue Inspektion der einzelnen Koeffizienten zeigt jedoch, dass sich nur einige wenige unter den AR-Koeffizienten statistisch signifikant zwischen den beiden Perioden unterscheiden.

Das vorliegende Ergebnis entwertet keineswegs die Diagnose von Nicht-Linearitäten, die im vorhergegangenen Abschnitt präsentiert wurde. Die Diagnose eines

Strukturbruchs nach der vorliegenden Methode, lässt für sich gesehen noch keine Schlussfolgerung zu, ob der Strukturbruch nicht in einer bestimmten Weise, etwa mit Hilfe eines ARNN-Modells, aus den Daten selbst prognostiziert werden kann. Allerdings ist die Frage zu klären, ein wie großer Anteil der anhand der statistischen Tests diagnostizierten Nicht-Linearität durch Strukturbrüche erklärt wird und wieviel davon genuine (additive) Nicht-Linearität ist, die für Prognosezwecke genutzt werden kann. Letztendlich kann diese Frage nur anhand von Out-of-Sample-Prognosen beantwortet werden (vgl. Kap. 5).

3.5 Schlussfolgerungen

Im Rahmen dieses Kapitels wurden einige wichtige Voraussetzungen für die Anwendung von nicht-linearen Prognoseinstrumenten, insbesondere neuronalen Netzen, im Anwendungsgebiet der makroökonomischen Zeitreihen diskutiert. Zunächst war zu klären, welcher Art die nicht-lineare Struktur in Zeitreihen ist, die man durch nicht-lineare Instrumente zu prognostizieren hoffen kann. Angesichts der Tatsache, dass neuronale Netze und andere innovative nicht-lineare Prognoseinstrumente vor allem in naturwissenschaftlichen Applikationen und in jenen ökonomischen und finanzwirtschaftlichen Gebieten Anwendung gefunden haben, in denen ein Überfluss an Daten herrscht und die Entdeckung nicht-linearer Struktur a priori wahrscheinlich ist, ist es auch wichtig für makroökonomische Zeitreihen die Möglichkeit nicht-linearer Struktur im vorhinein abzuschätzen. Das wurde unter Bezugnahme auf theoretische Ansätze in der Ökonomie und bisherige empirische Befunde versucht.

Bevor man das im Vergleich zu linerer Modellierung mühsame Unterfangen der Anpassung nicht-linearer Modelle beginnt, empfiehlt es sich das Vorhandensein von Nicht-Linearität in den Daten anhand eines statistischen Tests zu überprüfen. Ansonsten könnte man Gefahr laufen, zufällige Strukturen fälschlicherweise als nicht-lineare Struktur zu nehmen, was zu schlechteren Prognoseergebnissen führt. Es steht heute eine Reihe von Tests zu diesem Zweck zur Verfügung, unter denen die Auswahl schwerfällt, da die Einschränkungen vor allem bei kurzen Zeitreihen und bei gleichzeitigem Vorhandensein von additiver und multiplikativer Nicht-Linearität, wie sie für makroökonomische Zeitreihen typisch ist, vielfältig sind.

In diesem Kapitel wurden mögliche Teststrategien und eine Reihe von konkreten Tests beschrieben, sich anbietende Modifikationen aufgezeigt und Interpretationsempfehlungen gegeben. Hierbei war die Vorgangsweise auf den Nachweis von additiver Nicht-Linearität ausgerichtet. Die verwendeten Tests unterscheiden sich von einander hinsichtlich Methode und Hypothesen und ergänzen einander dadurch. Es zeigen sich aber auch Unterschiede in der Macht der Tests. Gegen jenen Typ von additiver Nicht-Linearität, der in makroökonomischen Zeitreihen häufig vorkommt, haben aufgrund der Ergebnisse anhand der beiden Beispielzeitreihen vermutlich der BDS-Test und der TLG-Test besonders hohe Macht.

Aufgrund der Testergebnisse kann das Vorhandensein von additiver Nicht-Linearität in den beiden Beispielzeitreihen angenommen werden. Die Testergebnisse sind in einem vernünftigen Rahmen hinsichtlich unterschiedlicher Formen der Differenzenbildung und hinsichtlich unterschiedlicher Gestaltung des vorgeschalteten linearen Modells sowie der Bereinigung um ARCH-Effekte (im Falle der Diagnostik-Tests) robust. Besonders wichtig ist im Hinblick auf die Modellierung mit Hilfe von neuronalen Netzen das statistisch hochsignifikante Testergebnis anhand des TLG-Test.

Dennoch sind einige wichtige Warnungen bei der Interpretation der Testergebnisse zu beachten. Einerseits weichen die Verteilungen der Teststatistiken in der praktischen Anwendung oft von der theoretisch unterstellten ab, wofür in erster Linie multiplikative Nicht-Linearitäten und unvollständige Entfernung linearer Abhängigkeiten verantwortlich sein können. Wichtiger scheint jedoch die Möglichkeit struktureller Veränderungen hinsichtlich der Modellparameter zu sein, die fälschlicherweise als Nicht-Linearität interpretiert werden können. Die Relevanz dieser Möglichkeit wurde durch die Schätzung eines einfachen Strukturbruchmodells auf den beiden Beispielzeitreihen illustriert. Je nachdem welchen Anteil die strukturellen Veränderungen haben und ob sie endogen aus den Daten erklärbar sind, werden nicht-lineare Prognoseinstrumente unterschiedlich erfolgreich sein. Dieser Befund rechtfertigt die Anwendung dieser Instrumente und weist auf die Notwendigkeit von systematischen Vergleichen anhand der Out-of-Sample-Prognosegüte hin, die im letzten Kapitel dieser Arbeit behandelt werden.

4 Neuronale Netze und Zeitreihenanalyse

Mit der Wiederentdeckung der neuronalen Netze (NN) als Forschungsgebiet etwa in den 80er Jahren des vorigen Jahrhunderts und der darauffolgenden großen Anzahl von Forschungsarbeiten war bald auch die Zeitreihenanalyse und -prognose eines der Anwendungsgebiete. Während auf diesem Gebiet anfangs unklare Erwartungen an die Möglichkeiten neuronaler Netze bestanden, ist nunmehr die vorherrschende Sichtweise, dass es sich um nicht-lineare Zeitreihenmodelle handelt, für die die betreffenden statistischen Theorien zur Anwendung kommen sollten.

Dieses Kapitel stellt die Autoregressiven neuronale-Netz-Modelle (ARNN) und eine Generalisierung dieser zur Berücksichtigung auch von deterministischer Saisonalität (ARNNDS) sowie ihre Anwendung vor.[20]

In den ersten beiden Abschnitten werden die Grundlagen, die verwendete Terminologie, das Generalisierungsproblem und die grundsätzlichen Möglichkeiten zur Lösung desselben vorgestellt, wobei auf den Bezug der Konzepte zu jenen, die bereits aus den vorigen Kapiteln bekannt sind, hingewiesen werden soll. Einige der Konzepte und Problembereiche, z.B. Overfitting und die Generalisierungsfähigkeit von Modellen, waren schon lange bekannt, bevor ihnen durch die NN-Forschung neue Aufmerksamkeit geschenkt wurde. Auch die Frage der Modellselektion bzw. der Entwicklung einer Modellierungsstrategie wurde durch die NN-Forschung wesentlich bereichert, obwohl die grundlegenden Konzepte in der nicht-linearen Statistik vorgeformt waren.

In den Abschnitten 4.3–4.5 werden die grundlegenden Techniken der Initialisierung der Gewichte eines ARNN-Modells sowie der lokalen und globalen Optimierung der Fehlerfunktion eines ARNN-Modells vorgestellt. Diese Techniken werden für die Bestimmung der Gewichte eines NN mit gegebener Architektur benötigt, also für jenes Problem, das man in der Ökonometrie als Modellschätzung kennt.

Bei der Verwendung von neuronalen Netzen ist jedoch die Unterscheidung von Modellschätzung einerseits und Modellspezifikation andererseits nicht so scharf wie bei den üblichen ökonometrischen Modellen. So sollte hier die Vorgangsweise der *Modellierung* insgesamt betrachtet werden. Ein Modellierungsansatz umfasst nicht nur Regeln für die Modellspezifikation, sondern auch für die Modellschätzung. Je nach verwendetem Modellierungsansatz kommt den einzelnen Techniken zur Bestimmung der Gewichte im Rahmen der verschiedenen Modellierungsansätze ein

[20] Die Verwendung der Abkürzungen ARNN und ARNNDS betont die spezifische Modelldefinition. Da im weiteren Verlauf der Arbeit immer aus dem Zusammenhang klar ist, wann die generalisierte Modellspezifikation zur Berücksichtigung von deterministischer Saisonalität benötigt wird, wird die Abkürzung ARNNDS nur dann verwendet, wenn auf die generalisierte Modellspezifikation ausdrücklich hingewiesen werden soll. Daher wird die Abkürzung ARNN in dem Sinne verwendet wird, dass sie auch die allgemeineren ARNNDS-Modelle einschließt. Wenn zudem bei der Besprechung von Methoden und Konzepten der Zeitreihenaspekt der Modelle nicht im Vordergrund steht, sondern Aussagen auf neuronale Netze im allgemeinen zutreffen, wird im Rahmen dieses Kapitels von NN-Modellen gesprochen.

ganz unterschiedlicher Stellenwert zu. Im Rahmen der vorliegenden Arbeit werden vier verschiedene Modellierungsansätze unterschieden, denen in diesem Kapitel jeweils ein Abschnitt (Abschnitte 4.6–4.9) gewidmet ist:

- Statistisch-parametrischer Ansatz,
- Klassischer Ansatz mit Early-Stopping,
- Ansatz mit Regularisierung,
- Bayesianischer Ansatz.

In den einzelnen Abschnitten werden die verschiedenen Techniken und Modellierungsansätze in kleinen Anwendungen anhand einer simulierten nicht-linearen Zeitreihe und anhand der beiden Beispielzeitreihen ALR und IPI illustriert und untersucht. Insoferne es hierbei um die Untersuchung der Generalisierungsfähigkeit der Modelle geht, stellt dies bereits einen ersten Schritt in Richtung Evaluation der Modelle anhand eines Evaluierungssets und eines Vergleichs mit linearen Modellen dar. Ein systematischer Vergleich der Prognosegüte ist jedoch dem Kapitel 5 vorbehalten. Im letzten Abschnitt des Kapitels werden einige Schlussfolgerungen zu ARNN-Modellen und den verschiedenen Modellierungsansätzen gezogen.

4.1 Grundlagen und Terminologie von ARNN-Modellen

In diesem Abschitt werden die Grundlagen und die Terminologien von ARNN- und ARNNDS-Modellen präsentiert. In einem kurzen Literaturüberblick werden zunächst die relevanten Bezüge der vorliegenden Arbeit zur angewandten Forschung auf dem Gebiet der Modellierung und Prognose von ökonomischen Zeitreihen hergestellt. Danach werden die ARNN- und ARNNDS-Modelle definiert, wobei eine Notation eingeführt wird, die in den späteren Teilen der Arbeit weiter verwendet wird.

Neuronale Netze zur Zeitreihenmodellierung und -prognose im Überblick

Das Gebiet der neuronale Netze kann heute als reifes Forschungsgebiet angesehen werden, für das es gelungen ist, die Erkenntnisse aus verschiedenen Disziplinen und aus der genuinen NN-Forschung zusammenzuführen. Der Stand der Forschung ist in verschiedenen Lehrbüchern und Monographien zugänglich gemacht worden (Hertz et al., 1991; Haykin, 1994; Bishop, 1995; Rojas, 1995, 1996; Ripley, 1996; Anders, 1997; Fine, 1999).[21] Wichtige Überblicksartikel sind White (1989c), Cheng und Titterington (1994) sowie Kuan und White (1994).

[21] Gemäß der Literaturrecherche des Autors kamen nach der Jahrtausendwende keine neuen umfassenden Lehrbücher und Monographien hinzu, was sicherlich als Hinweis auf den Reifungsprozess des Forschungsgebiets gesehen werden kann. Gleichzeitig ist jedoch eine Verschiebung des Fokusses der NN-Forschung auf Anwendungen wie Finanzmarktprognose und Data-Mining festzustellen, wobei NN nur als eines von mehreren verfügbaren Instrumenten angesehen werden.

NN-Modelle sind flexibel und nicht-linear, haben universelle Approximationseigenschaften (Hornik et al., 1989) und können generalisieren, sofern geeignete Lernverfahren eingesetzt werden. Dies macht sie auch für Zeitreihenmodellierung interessant.[22] Zhang et al. (1998) gibt einen umfassenden Überblick über den Einsatz von NN zur Vorhersage von Zeitreihen. Es herrscht eine große Heterogenität der verwendeten Ansätze und Sichtweisen. NN zur Modellierung und Prognose von Zeitreihen wurden in zahlreichen Anwendungsgebieten eingesetzt, z.B. Sonnenfleckenzeitreihen (Weigend et al., 1990; Medeiros et al., 2006), Flugverkehrzeitreihen (Faraway und Chatfield, 1998), betriebswirtschaftliche Zeitreihen (Balkin und Ord, 2000), Zeitreihen zu Fangzahlen von Luchsen (Zhang, 2003; Kajitani et al., 2005), Elektrizitätsverbrauchszeitreihen (Darbellay und Slama, 2000; Hippert et al., 2001, 2005), Aktien- und Wechselkurse (Weigend et al., 1990; Refenes et al., 1994; Franses und van Griensven, 1997).[23] Auch im Gebiet der makroökonomischen Zeitreihen liegen mittlerweile viele Studien vor, so dass auch hier nur eine Auswahl gemacht werden kann: Swanson und White (1997b), Swanson und White (1997a), Stock und Watson (1999), Moshiri und Cameron (2000), Balkin und Ord (2000), Rech (2002), Heravi et al. (2004), Teräsvirta et al. (2005). Unter diesen Artikeln sind die beiden letztgenannten für die vorliegene Untersuchung wichtig, da sie besonders umfassend sind und ähnliche Modellierungsansätze verfolgen wie die in diesem Kapitel beschriebenen.

Eine Reihe von Aspekten der NN-Modellierung von Zeitreihen ist besonders kritisch zu sehen, wenn man NN zur Prognose makroökonomischer Zeitreihen einsetzt. Es sind dies i) Saisonalität und Trend ii) sparsame Modellspezifikation iii) Entwicklung von Modellierungsstrategien und iv) Evaluierung der Prognosegüte. Diese werden im folgenden kurz behandelt, wobei neben einer Skizzierung des Stands der Forschung jeweils auf die Beiträge der vorliegenden Untersuchung hingewiesen wird.

Saisonalität und Trend. Obwohl Saisonalität und Trend wichtige Aspekte von makroökonomischen Zeitreihen sind, denen in der Zeitreihenökonometrie große Aufmerksamkeit gewidmet wurde, ist in der Literatur zur NN-Modellierung von Zeitreihen dieses Thema lange relativ pragmatisch behandelt worden. Die meisten angewandten Studien lösen das Problem durch vorgeschaltete Trend- und Saisonbereinigungsverfahren. Doch einige Arbeiten stellen diesen Aspekt ins Zentrum und fragen etwa, ob die Zeitreihen tatsächlich zuerst saisonbereinigt werden sollen (Nelson et al., 1999) und ob NN in der Lage sind, sich ändernde saisonale Muster zu erkennen (Franses und Draisma, 1997). Es scheint, dass sich in diesem Bereich unrealistische Erwartungen an die Fähigkeiten von neuronalen Netzen besonders

[22] Auf die vielfältigen Anwendungsgebiete von NN in anderen Bereichen wie z.B. Mustererkennung und räumliche Interaktionsmodellen soll an dieser Stelle nur verwiesen werden (Fischer, 2002, 2006).

[23] Einige der in diesen Arbeiten behandelten Zeitreihen haben aufgrund ihrer Nicht-Linearität die Rolle von Benchmark-Problemen übernommen und werden immer wieder von Forschungsarbeiten und Lehrbüchern zu nicht-linearer Zeitreihenanalyse herangezogen.

lange gehalten haben. So berufen sich z.B. die beiden erwähnten Arbeiten und auch noch spätere Arbeiten auf die universellen Approximationseigenschaften von neuronalen Netzen um die Notwendigkeit von saisonaler und Trendbereinigung in Frage zu stellen. (Zhang und Qi, 2005) liefert einen gründlichen Literaturüberblick zum Thema und führt eine empirische Untersuchung anhand von simulierten und realweltlichen ökonomischen Zeitreihen durch, die klar die Notwendigkeit der separaten Behandlung von Saisonalität und Trend belegt, so dass Gegenstand der eigentlichen NN-Modellierung nur die trend- und saisonbereinigte Zeitreihe ist. Aus theoretischer Sicht sollte durch die Arbeiten von Leisch et al. (1999) und Trapletti et al. (2000) klar geworden sein, dass ARNN-Prozesse (ohne einen linearen Teil) immer stationär sind und daher die Modellierung von nicht-stationären Zeitreihen durch ARNN-Modelle fragwürdig ist, wenn man in einem NN-Modell mehr als eine Black Box sieht.

In der vorliegenden Arbeit wird die Frage von Saisonalität und Trend in neuartiger Weise behandelt, indem ARNN-Modelle spezifiziert werden, die einen deterministischen Teil enthalten, dessen Parameter *simultan* mit den Parametern des NN-Teils (sowie gegebenenfalls eines linearen Teils) geschätzt werden. Als Inputvariable für den NN-Teil (und den linearen Teil) des Modells wird demnach die um deterministische Saisonalität und Trend bereinigte Zeitreihe herangezogen. Um stochastische Trends und Saisonalität sollte hingegen durch geeignete Differenzenfilter bereinigt werden (vgl. Abschnitt 2.3).

Sparsame Modellspezifikation. Die ersten Anwendungen von NN-Modellen auf Zeitreihen verwendeten meist ein Fenster von zeitverzögerten Variablen als Inputs, wobei alle Lags von 1 bis zur Ordnung des Modells aufgenommen wurden. Bei höherer Modellordnung und einer vollen Vernetzung in einem Netz mit mehreren verdeckten Neuronen ergibt sich daraus schnell eine sehr hohe Parameterzahl. Angesichts des Overfittingproblems muss man nach einer verfeinerten Vorgangsweise suchen, die darin besteht, die Lags, zu denen die Zeitreihe als erklärende Variable in das Modell aufgenommen wird, gezielt zu bestimmen. Zhang et al. (1998) betonen, dass das Problem der Bestimmung der Inputknoten in Zeitreihenanwendungen wahrscheinlich noch wichtiger ist als jenes der Bestimmung der Anzahl der verdeckten Knoten. Gerade für makroökonomische und saisonale Zeitreihen ist es offensichtlich, dass bestimmte Lags von größerer Wichtigkeit sind als andere. Die Bestimmung der Lag-Struktur in NN-Modellen wird meist in ähnlicher Weise vorgenommen wie in den lineraren AR-Modellen, ist aber durch die mögliche Nicht-Linearität in der Zeitreihe sowie durch die höhere Modellkomplexität in NN-Modellen schwieriger. Eine exemplarische und heuristische Vorgangsweise findet sich z.B. bei Zhang und Qi (2005), der für seine Modelle von monatlichen Zeitreihen die Lags 1 bis 4, 12 bis 14, 24, 25 und 36 berücksichtigt und in der Validierungsphase die Modellordnung bestimmt. Lags in der Lag-Menge, die die vorbestimmte Modellordnung übersteigen, werden demnach nicht ins Modell aufgenommen. Eine Reihe von ähnlichen Heuristiken existieren, ohne dass sich eine vorherrschende Methodologie für die Auswahl

der Lags etabliert hätte. Möglichkeiten, die Lags, in denen nicht-lineare Struktur vermutet wird und die als Inputs in das NN-Modell aufzunehmen sind, vorab zu ermitteln, liegen in der Automutual-Information (vgl. Granger und Lin, 1994) oder in der in Abschnitt 3.3 vorgeschlagenen Adaptation des Teräsvirta-Lin-Granger-Tests (TLG-Tests) vor (Teräsvirta et al., 1993). Eine weitere Möglichkeit, eine sparsame Spezifikation der Lag-Struktur des Modells herbeizuführen besteht in Pruning (Cottrell et al., 1995).

In der vorliegenden Arbeit sind Möglichkeiten der sparsamen Modellspezifikation in verschiedener Hinsicht vorgesehen. Zum einen werden die ARNN-Modelle so spezifiziert, dass die Lag-Struktur für den linearen AR-Teil und den NN-Teil des Modells getrennt vorgegeben werden kann. Zum anderen besteht im Rahmen des Prunings die Möglichkeit, bestimmte Gewichte des Netzes auf den Wert Null zu restringieren, so dass die Lag-Struktur effektiv für jedes einzelne verdeckte Neuron unterschiedlich spezifiziert werden kann. Für die Vorauswahl der in Frage kommenden Lags wird die Heuristik angewandt, dass jene Lags, die im Rahmen der AR-Modellierung als signifikant ermittelt wurden, auch im NN-Modell ausgewählt werden.

Entwicklung von Modellierungsstrategien. Eine Modellierungsstrategie bezeichnet eine Abfolge von Schritten und Entscheidungsregeln, um zu einem Modell zu gelangen. Sie betreffen bei der NN-Modellierung einerseits die Modellspezifikation (Bestimmung der Inputs bzw. der Lag-Struktur, Anzahl der verdeckten Neuronen, Aufnahme von direkten Verbindungen zwischen Inputs und Outputknoten) und andererseits auch die Methode des Lern- bzw. Schätzverfahrens. Obwohl „Modellierungsstrategie“ (engl. model selection strategy bzw. model selection approach) ein Ausdruck ist, der erst in späteren Arbeiten (Swanson und White, 1997b; Anders und Korn, 1999; Medeiros et al., 2006) stärker betont wurde, die vornehmlich einer statistisch-parametrischen Sichtweise der ARNN-Modellierung folgen, ist das Bemühen um nachvollziehbare Regeln der Modellierung von Anfang an in der NN-Forschung ein zentrales Anliegen gewesen.

Mehr als bei anderen nicht-linearen Zeitreihenmodellen ist es bei NN erforderlich, nicht nur das Modell selbst zu beschreiben, sondern auch das Verfahren, wie man zu diesem Modell gelangt ist. Anderenfalls können die Ergebnisse und Schlussfolgerungen einer präsentierten Forschungsarbeit nicht beurteilt werden. Dies ist auch ein Schwachpunkt zahlreicher angewandter NN-Papiere gerade im Bereich der Zeitreihenanalyse, die nur einzelne Teile der verfolgten Modellierungsstrategie dokumentieren und nachvollziehbar machen. Ein Hauptanliegen der Formulierung von Modellierungsstrategien ist auch, die Prognosefähigkeit von NN-Modellen mit alternativen linearen und nicht-linearen Modellen in systematischer Weise vergleichen zu können. Zu diesem Zweck ist es notwendig, die Modellierung zumindest teilweise zu automatisieren, da nur so größere Anzahl von Zeitreihen bewältigt werden können. Dies ist z.B. für Teräsvirta et al. (2005) relevant, die 47 monatliche makroökonomische Zeitreihen für ihren Vergleich verwendet haben.

In der vorliegenden Arbeit werden vier Ansätze der Modellierung mit ARNN-Modellen unterschieden, die jeweils zu klar definierten Modellierungsstrategien führen. Diese betreffen nicht nur den statistisch-parametrischen Ansatz, sondern decken auch die im Rahmen der traditionellen NN-Forschung entwickelten Verfahren (Early-Stopping) und den Regularisierungsansatz ab, zu denen jeweils neue Möglichkeiten, die Methode nachvollziehbarer und regelbasierter zu gestalten, entwickelt werden. Im Bereich des Bayesianischen Ansatzes, ist die Frage der Modellierungsstrategie bereits relativ weit entwickelt, und hier wurde im Rahmen der vorliegenden Arbeit eine relativ getreue Implementation derselben angestrebt.

Prognosegüte-Evaluierung. Ein weiteres Gebiet, das in der NN-Forschung im Vergleich zur angewandten Statistik von Anfang an einen relativ hohen Stellenwert genießt ist das Bestreben, die Prognosegüte auf beiseite gehaltenenen Daten zu testen, die in keiner Weise zur Modellierung oder Schätzung des Modells verwendet werden. Dieses Datenset wird als Evaluierungsset oder als Out-of-Sample-Set bezeichnet. Die für die Modellbildung zur Verfügung stehenden Daten werden als Lernset oder In-Sample-Set bezeichnet.[24] In der Zeitreihenmodellierung mit NN hat das Bemühen um Out-of-Sample-Evaluierung zu einer besonders engen Verbindung der NN-Forschung mit der Forschung zu linearen und nicht-linearen Zeitreihenmodellen geführt. Für ökonomische Anwendungen ist besonders auf die Arbeiten von Swanson und White (1997b,a), Stock und Watson (1999) und Teräsvirta et al. (2005) hinzuweisen.

In der vorliegenden Arbeit wird die Frage der Evaluierung der Prognosegüte erst im Kapitel 5 systematisch behandelt. Auch bestimmte theoretische Aspekte von ARNN-Modellen, die thematisch eher zur Erstellung eines Evaluierungsdesigns passen, wie etwa die Frage der Erstellung von Mehrschritt-Prognosen, werden im Kapitel 5 behandelt.

ARNN-Modell

ARNN-Modelle sind nicht-lineare autoregressive Zeitreihenmodelle, wobei die Nicht-Linearität durch ein neuronales Netz des Feedforward-Typs verwirklicht wird (vgl. auch Trapletti et al., 2000; Medeiros et al., 2006). In dieser Arbeit werden das ARNN-Modell sowie eine Generalisierung des ARNN-Modells, das ARNNDS-Modell, welches auch ein deterministisches saisonales Muster berücksichtigt, in der folgenden Spezifikation verwendet:

ARNN-Modell:

$$y_t = \tilde{y}_t + d,$$

[24] In manchen Modellierungsstrategien wird das Lernset weiter in ein Trainings- und ein Validierungsset unterteilt, wobei das Validierungsset zur Auswahl von alternativen Modellen, zur Bestimmung von Hyperbarametern oder zum Stoppen des Trainings verwendet wird (siehe hierzu die Abschnitte zu den einzelnen Modellierungsansätzen).

$$\tilde{y}_t = \sum_{k=1}^{K} c_k \psi[b_k + \sum_{j \in \mathcal{J}} a_{kj} \tilde{y}_{t-j}] + \sum_{l \in \mathcal{L}} f_l \tilde{y}_{t-l} + \varepsilon_t, \tag{4.1}$$

wobei

y_t	der zum Zeitpunkt t beobachtete Wert der Zeitreihe,
d	das Interzept (spezifiziert als Durchschnittsbereinigung),
c_k, b_k, a_{kj}	die Parameter des neuronalen Netzes,
f_l	die Parameter des linearen Teils (Shortcuts),
K	die Anzahl der verdeckten Neuronen,
$\mathcal{J}$ und $\mathcal{L}$	die Menge der Lags j bzw. l, zu denen vergangene y_t eingehen,
ψ	die Aktivierungsfunktion (z.B. Tangens Hyperbolicus) und
ε_t	ein Fehlerterm.

ARNN-Modell mit deterministischer Saisonalität (ARNNDS-Modell):

$$y_t = \tilde{y}_t + d_t,$$

$$d_t = d + d^t t + \sum_{i=2}^{s} d_i^s s_{it},$$

$$\tilde{y}_t = \sum_{k=1}^{K} c_k \psi[b_k + \sum_{j \in \mathcal{J}} a_{kj} \tilde{y}_{t-j}] + \sum_{l \in \mathcal{L}} f_l + \tilde{y}_{t-l} \varepsilon_t, \tag{4.2}$$

wobei d_t ein deterministischer Mittelwertprozess ist, in dessen Spezifikation abgesehen von dem bereits oben definierten Interzept d die folgenden Symbole definiert sind:

d^t und d_i^s	Parameter des deterministischen Teils (Trend und Saison),
t	Trendvariable (gleichzeitig zu Verwendung als Zeitindex)
s_{it}	saisonale Dummyvariable für Saison i und
s	Anzahl der Saisonen.

Die obige Formulierung des ARNN- bzw. ARNNDS-Modells berücksichtigt als einen wichtigen Aspekt die Möglichkeit eines sparsam spezifizierten Modells, indem nicht alle vergangenen Lags der Zeitreihe in das Modell eingehen. Man beachte, dass das Modell aus einem deterministischen und einem stochastischen Teil besteht, deren Koeffizienten simultan geschätzt werden. Im Rahmen des stochastischen Teils, der selbst wiederum aus einem linearen AR und einem NN-Teil besteht, wird die um die deterministische Komponente bereinigte Zeitreihe $\tilde{y}_t$ verarbeitet. Die deterministische Komponente besteht im einfacheren Falle des ARNN-Modells bloß aus dem Interzept d, welches somit, im Unterschied zu den üblichen Modellformulierungen, als Abzugsterm (vgl. engl. „demeaning“) spezifiziert ist.

Es ist klar, dass es sich beim ARNN-Modell, und soferne man davon absieht, dass für die Modellierung der deterministischen Saisonalität saisonale Dummy-Variable

zum Einsatz kommen, auch beim ARNNDS-Modell um ein nicht-lineares univariates Zeitreihenmodell handelt:

$$y_t = \varphi(y_{t-1}, \ldots, y_{t-p}) + \varepsilon_t, \tag{4.3}$$

wobei φ hier eine nicht-lineare Funktion bezeichnet.

In der vorliegenden Arbeit werden die in der Ökonometrie einerseits und der NN-Forschung andererseits verwendeten Terminologien mehr oder weniger gleichberechtigt nebeneinander verwendet. Bei der Modellspezifikation nach Gleichung (4.1) bzw. (4.2) handelt es sich eher um eine knappe und ökonometrisch orientierte Notation. In der Terminologie der NN-Welt liegt ein um zusätzliche Elemente erweitertes Feedforward-Netz mit drei Schichten von Neuronen bzw. Knoten vor. Die Verbindungen zwischen den einzelnen Knoten bzw. ihre Stärken werden in NN-Terminologie als Gewichte bezeichnet. Die Knoten in der ersten Schicht, der Input-Schicht sind voll mit jenen der zweiten Schicht, den verdeckten Neuronen (engl. Hidden Units), verbunden, wobei a_{kj} das Gewicht der Verbindung ist zwischen dem j-ten Inputknoten (im vorliegenden Fall genaugenommen dem Inputknoten, der die um j Zeiteinheiten verzögerten Beobachtung von $\tilde{y}_t$ als Input hat) und dem k-ten verdeckten Knoten. Der Parameter b_k wird auch als Bias bezeichnet.[25] Die zweite Schicht umfasst die Verbindungen zwischen den verdeckten Neuronen und einem einzelnen Outputknoten, die Gewichte c_k. Die Funktion ψ wird als Aktivierungsfunktion bezeichnet. Neben der Tangenshyperbolicus-Funktion können auch andere geeignete nicht-lineare Funktionen, z.B. die logistische Funktion zum Einsatz kommen. Die Aktivierungsfunktion bildet die biologische Funktionsweise eines Neurons nach, das erst ab Erreichen eines bestimmten Schwellwerts zu „feuern“ beginnt. Am Output-Neuron wird gemäß der Modelldefinition nach Gleichung (4.1) bzw. (4.2) keine Aktivierungsfunktion angewendet, d.h. die Aktivierungsfunktion ist hier implizit die Identitätsfunktion. Die Parameter f_l werden in der NN-Terminologie als direkte Verbindungen oder (engl. Shortcuts) bezeichnet, weil man sie sich als direkte Verbindungen zwischen den Inputknoten und dem Outputknoten denken kann.

Fehlerfunktion (Zielfunktion)

Bei der Formulierung sowohl von linearen als auch nicht-linearen Modellen werden üblicherweise auch Annahmen über die Eigenschaften des Fehlerterms ε_t getroffen, wenn nicht explizit, so implizit durch das angewandte Schätzverfahren. Im vorliegenden Falle wird angenommen, dass ε_t unabhängig, identisch und normalverteilt ist. Aus dieser Annahme folgt, dass das Kleinst-Quadrate-Verfahren zur Bestimmung der Parameter des Modells dem Maximum-Likelihood-Schätzer entspricht.

[25] Dies ist nicht mit dem statistischen Bias-Begriff zu verwechseln, der die Verzerrung eines Schätzers bezeichnet und in dieser Arbeit im Zusammenhang mit der Diskussion des „Bias-Variance“ eine Rolle spielt, vgl. Abschnitt 4.2.

Man spricht in diesem Zusammenhang daher auch von einem Quasi-Maximum-Likelihood-Schätzer.

Dieser Zugang basiert auf der Verwendung der Summe der quadrierten Fehler (Sum of Squares Error, SSE) als Fehlerfunktion E (auch: Zielfunktion) im Rahmen des Verfahrens zur Bestimmung der Gewichte:

$$E(\mathbf{w}) = \sum (y_t - \hat{y}_t(\mathbf{w}, y_{t-1}, y_{t-2}, \ldots))^2, \tag{4.4}$$

wobei $\mathbf{w}' = (a_{11}, a_{12}, \ldots, a_{KJ}, b_1, \ldots, b_K, c_1, \ldots, c_K, d, f_1, \ldots, f_L, d^t, d_2^s, \ldots, d_s^s)$ den Vektor der Parameter des Modells bezeichnet und $\hat{y}_t$ den prognostizierten Wert für y_t, der hier als Funktion der Parameter und der vergangenen Werte der Zeitreihe dargestellt wird. Je nach gewähltem Ansatz handelt es sich hierbei um ein Minimierungsproblem oder um ein modifiziertes Lernverfahren.

Stationarität von ARNN- und ARNNDS-Prozessen

Eine wichtiger Aspekt des ARNN-Modells ist jener der Stationarität. Obwohl hierzu theoretische Ergebnisse vorliegen, werden diese in praktischen Studien kaum beachtet. Dies ist im Gegensatz zur Modellierung mit linearen Modellen, wo es üblich ist, die Stationarität eines geschätzten Modells zu untersuchen, bevor es für Prognosezwecke verwendet wird.

Die Bedingungen der Stationarität eines ARNN-Modells werden von Leisch et al. (1999) angegeben, die sich für dieses Ergebnis auf Chan und Tong (1985) stützen. Ein ARNN-Modell ohne linearen Teil ist immer stationär. Die Stationarität eines ARNN-Modells mit linearem Teil hängt nur von den Parametern des linearen Teils ab. Es ist stationär, wenn die Wurzeln des charakteristischen Polynoms des linearen Teils innerhalb des Einheitskreises liegen. Diese Ergebnisse können ohne weiteres auf die allgemeineren ARNNDS-Modelle übertragen werden, wobei die Stationarität den um den deterministische Komponente bereinigten Prozess $\tilde{y}_t$ betrifft.

Trotz diesem einfachen Befund gibt es in der Praxis von auf Zeitreihen angepassten ARNN-Modellen häufig ein Modellverhalten, das dem theoretischen Kriterium nicht zu entsprechen scheint. Man kann zwei Fälle unterscheiden:

Pseudo-nicht-stationäres Verhalten Die Parameter des ARNN-Modells erfüllen die theoretischen Kriterien für Stationarität. Doch bei Verwendung des Modells für die Prognose kommt es zu explosivem Verhalten. Dennoch ist, wie sich an simulierten Zeitreihen von ausreichender Länge zeigt, das Modell asymptotisch stationär, d.h. nach einer Einschwingphase befindet sich die Zeitreihe in einer stationären Verteilung.

Pseudo-stationäres Verhalten Obwohl die Parameter des ARNN-Modells nicht die theoretischen Kriterien für Stationarität erfüllen, zeigen erstellte Prognosezeitreihen und auch simulierte Zeitreihen ein typisch stationäres Verhalten

bzw. wirken bei visueller Prüfung stationär. Eine genau Inspizierung von solchen Modellen ergibt typischerweise, dass der NN-Teil durch die Ausprägung der Parameterwerte nahezu linear ist und dem linearen Teil des Modells entgegenwirkt. Mithilfe eines solchen Modells simulierte Zeitreihen sind mitunter über ein Vielfaches der typischen Länge von makroökonomischen Zeitreihen unauffällig in ihrem Stationaritätsverhalten.

Die Konsequenzen der beiden Fälle für die Anwendung im Bereich der Modellierung und Vorhersage von makroökonomischen Zeitreihen sind unterschiedlich. Der Fall der Pseudo-Nicht-Stationarität ist sehr störend, da explosive Prognosen die Prognosequalität stark beeinträchtigen können. Man kann sie eigentlich nicht ausschließen, da man sich nicht sicher sein kann, ob sich die Zeitreihe in ihrem bekannten Bereich tatsächlich in ihrer stationären Verteilung befindet. Bei der Verwendung von pseudo-stationären Modelle für kurz- und mittelfristige Prognosen ergeben sich für den Praktiker keine unmittelbaren Probleme. Doch will man eine Interpretation der Parameter des Modells im Sinne eines „wahren“ Modells erreichen, so muss man das Modell verwerfen.

4.2 Generalisierungsproblem und Modellierungsstrategien

In diesem Abschnitt geht es darum, die Frage der Generalisierungsfähigkeit von ARNN-Modellen zu erläutern und die grundsätzlichen Möglichkeiten, diese sicherzustellen. Neuronale Netze bzw. ARNN-Modelle als besonders flexible nicht-lineare Modelle neigen zum sogenannten Overfitting, d.h. dem Erlernen unwesentlicher Aspekte in den Daten, die für die Generalisierungsfähigkeit, d.h. für die Prognosegüte auf neuen, unbekannten Daten, nicht hilfreich sind. Zunächst wird diese Problematik anhand der Bias-Varianz-Zerlegung beleuchtet. Danach werden verschiedene Sichtweisen von ARNN-Modellen und Modellierungsstrategien vorgestellt, die in der vorliegenden Arbeit für die Prognose von makroökonomischen Zeitreihen eingesetzt werden. Schließlich wird am Ende dieses Abschnitts noch die Erstellung der simulierten Zeitreihe SIM erläutert. Diese soll in den späteren Abschnitten des Kapitels zusätzlich zu den beiden Beispielzeitreihen ALR und IPI (vgl. Abschnitt 2.2) zur Illustration der Funktionsweise der Modellierungsstrategien und der ihnen zuzuordnenden Techniken herangezogen werden.

Generalisierungsfähigkeit und die Bias-Varianz-Zerlegung

Prognosen von Zeitreihen, gleich ob diese mit Hilfe von linearen oder nicht-linearen Modellen erstellt werden, sollen einen möglichst geringen Fehler auf dem zu prognostizierenden, zukünftigen Abschnitt der Zeitreihe erreichen, wobei dieser zum Zeitpunkt der Prognoseerstellung entweder noch nicht bekannt ist oder bewusst unbeachtet bleibt, etwa um die Prognosequalität überprüfen zu können. Das Ziel ist also weniger die Minimierung der Fehlerfunktion auf dem In-Sample-Set, d.h. der

zur Erstellung der Prognose verwendeten Zeitreihe, sondern die möglichst hohe Generalisierungsfähigkeit, d.h. Genauigkeit der Vorhersagen auf neuen, unbekannten Daten.

Um dieses Ziel zu erreichen, sollte ein Kompromiss hinsichtlich der Komplexität des Modells gefunden werden. Einerseits sollte die Flexibilität und die Anzahl der Parameter des Modells nicht so groß sein, dass es unwesentliche, d.h. nicht generalisierungsfähige Aspekte in den Daten modelliert. Diese Gefahr, das sogenannte Overfitting, ist für NNs besonders groß, da durch Hinzufügen von verdeckten Neuronen die Anzahl der Parameter schnell sehr hoch werden kann. Der Kern des Overfitting-Problems besteht darin, dass die Prognosen zu stark von den Zufälligkeiten der konkreten Stichprobe abhängig werden. Andererseits sollte die Modellkomplexität auch nicht zu gering sein, da dann wesentliche Strukturen in der Zeitreihe nicht berücksichtigt werden können.

Die Problematik um Overfitting und Generalisierungsfähigkeit ist in der angewandten Statistik seit langem bekannt, wurde aber in der NN-Forschung besonders intensiv erforscht. Theoretische Klarheit bekommt man, wenn man die sog. Bias-Varianz-Zerlegung verwendet, die von Geman et al. (1992) im Zusammenhang mit NNs aufgearbeitet wurde. Sie beruht auf der Zerlegung des Fehlers in einen (quadrierten) Bias und eine Varianz-Komponente. Während zu einfache Modelle einen hohen Bias haben, d.h. im Durchschnitt den zu modellierenden Zusammenhang nicht richtig abbilden, haben zu komplexe Modelle eine hohe Varianz-Komponente, d.h. die Prognosen schwanken stärker um ihren Erwartungswert. Diese Aussagen sollen im folgenden präzisiert werden.

Die Bias-Varianz-Zerlegung geht aus vom erwarteten quadrierten Vorhersagefehler, *ESFE* (engl. expected squared forecasting error),

$$ESFE = \mathrm{E}[(y - \hat{y}(\mathbf{w}, \mathbf{x}))^2]. \tag{4.5}$$

Zu diesem Ausdruck müssen zwei wesentliche Anmerkungen gemacht werden. Erstens wird hier, im Gegensatz zu Formel (4.4), statt den verzögerten Zeitreihenwerten $y_{t-1}, y_{t-2}, \ldots$ einfach $\mathbf{x}$ geschrieben und auch bei y der Zeitindex fallen gelassen. Das Problem wird somit zum Zwecke der folgenden Ausführungen als (nichtlineares) Regressionsproblem mit einer abhängigen Variable y, den unabhängigen Variablen $\mathbf{x}$ und den Koeffizienten $\mathbf{w}$ der nicht-linearen Regression betrachtet. Zweitens wird in der Formulierung offen gehalten, worüber der Erwartungswert gebildet wird. Entweder ist die Bildung des Erwartungswertes über die Verteilung von y oder über die gemeinsame Verteilung von y und $\mathbf{x}$ gemeint. Im ersten Fall könnte man den Ausdruck als

$$ESFE = \mathrm{E}[(y - \hat{y}(\mathbf{w}, \mathbf{x}))^2 | \mathbf{x}] \tag{4.6}$$

präzisieren, im zweiten Falle als

$$ESFE = \int \mathrm{E}[(y - \hat{y}(\mathbf{w}, \mathbf{x}))^2 | \mathbf{x}] p(\mathbf{x})\, d\mathbf{x}, \tag{4.7}$$

wobei $p(\mathbf{x})$ die Wahrscheinlichkeitsdichte von $\mathbf{x}$ ist. Die folgenden Überlegungen treffen im Prinzip auf beide Fälle zu; um die Notation einfach zu halten wird mit der Schreibweise in Gleichung (4.5) gearbeitet und eine Interpretation im Sinne des ersten Falles nahegelegt.

Der erwartete quadrierte Fehler kann in einen unsystematischen und einen systematischen Fehler zerlegt werden:

$$\begin{aligned} \mathrm{E}[(y - \hat{y}(\mathbf{w}, \mathbf{x}))^2] &= \mathrm{E}[(y - \bar{y}(\mathbf{x}) + \bar{y}(\mathbf{x}) - \hat{y}(\mathbf{w}, \mathbf{x}))^2] \\ &= \mathrm{E}[(y - \bar{y}(\mathbf{x}))^2] + \mathrm{E}[(\hat{y}(\mathbf{w}, \mathbf{x}) - \bar{y}(\mathbf{x}))^2] \\ &= \mathit{ESFE}_u + \mathit{ESFE}_s, \end{aligned} \tag{4.8}$$

wobei $\bar{y}(\mathbf{x})$ die kurze Schreibweise für $\mathrm{E}[y|\mathbf{x}]$ ist. ESFE_u ist der unsystematische Fehler und entspricht der Varianz des Fehlerprozesses. Er kann, da rein zufälliger Natur, durch kein Modell vorhergesagt werden. Die Prognosequalität lässt sich daher allein mit dem systematischen Fehler ESFE_s messen. Dieser gibt an, wie hoch die erwartete Abweichung zwischen dem „wahren“ Modell und seiner Approximation ist.

Im Falle, dass der Erwartungswert nur über die Verteilung von y gebildet wird, vereinfacht sich der Ausdruck zu

$$\mathit{ESFE}_s = (\hat{y}(\mathbf{w}, \mathbf{x}) - \bar{y}(\mathbf{x}))^2, \tag{4.9}$$

welcher nun weiter zerlegt wird. Um die Abhängigkeit des Schätzers von der jeweiligen Auswahl der für die Modellierung verwendeten Daten $\mathcal{D}$ deutlich zu machen wird nun $w(\mathcal{D})$ statt w geschrieben und der obige Ausdruck um die Erwartungsbildung bezüglich $\mathcal{D}$ ergänzt. Sodann kann ESFE_s in ähnlicher Weise wie zuvor ESFE zerlegt werden:

$$\begin{aligned} \mathit{ESFE}_s &= \mathrm{E}_{\mathcal{D}}[(\hat{y}(\mathbf{w}(\mathcal{D}), \mathbf{x}) - \bar{y}(\mathbf{x}))^2] \\ &= \mathrm{E}_{\mathcal{D}}[(\hat{y}(\mathbf{w}(\mathcal{D}), \mathbf{x}) - \mathrm{E}_{\mathcal{D}}[\hat{y}(\mathbf{w}(\mathcal{D}), \mathbf{x})] + \mathrm{E}_{\mathcal{D}}[\hat{y}(\mathbf{w}(\mathcal{D}), \mathbf{x})] - \bar{y}(\mathbf{x}))^2] \\ &= (\mathrm{E}_{\mathcal{D}}[\hat{y}(\mathbf{w}(\mathcal{D}), \mathbf{x}) - \bar{y}(\mathbf{x}))^2 + \mathrm{E}_{\mathcal{D}}[(\hat{y}(\mathbf{w}(\mathcal{D}), \mathbf{x}) - \mathrm{E}_{\mathcal{D}}[\hat{y}(\mathbf{w}(\mathcal{D}), \mathbf{x})])^2] \\ &= \mathrm{Bias}^2 + \mathrm{Varianz} \end{aligned} \tag{4.10}$$

In der obigen Schreibweise kommt sehr klar zum Ausdruck, dass die Zerlegung in Bias und Varianz nur im Zusammenhang mit der Abhängigkeit des Vorhersagefehlers von der gewählten Stichprobe verstanden werden kann. Ein einfach gehaltenes Modell hat eine geringe Varianz, da es von den zufälligen Mustern der Stichprobe kaum abhängt. Im Extremfall würde z.B. ein Modell, das immer einen konstanten Wert $\hat{y}(\mathbf{x}) = \mathit{const}$ voraussagt, eine Varianz von Null aufweisen. Aber es hätte offensichtlich einen sehr hohen Bias. Umgekehrt haben komplizierte Modelle mit vielen Parametern zwar einen geringen Bias, neigen aber dazu, in Bereichen des Eingaberaums, die nicht ausreichend dicht mit Daten besetzt sind, Vorhersagen weitab des „wahren“ bedingten Erwartungswertes $\bar{y}(\mathbf{x})$ zu liefern. Bei hochdimensionalen

Eingaberäumen (und daher auch Parameterräumen), ist das Auftreten solcher Bereiche aber unvermeidbar, weshalb Geman et al. (1992) darauf hinweisen, dass es sich dann eher um Extrapolation als um Intrapolation handelt. Die Problematik, einen guten Ausgleich zwischen Bias und Verzerrung zu finden, wurde auch als das Bias-Varianz-Dilemma bezeichnet, da man das eine nur auf Kosten des anderen senken kann. Im folgenden Unterabschnitt werden die grundsätzlichen Möglichkeiten, diesen Ausgleich in der Modellierung und Vorhersage mit ARNN und ARNNDS-Modellen zu bewerkstelligen, skizziert und ein Überblick über die verschiedenen Modellierungsstrategien gegeben werden, die diese Prinzipien verwirklichen und die in der vorliegenden Arbeit erprobt werden.

Sichtweisen und Modellierungsstrategien

Ein optimaler Trade-off zwischen Bias und Varianz kann gefunden werden, indem die Flexibilität bzw. Modellkomplexibilität optimal gewählt wird. Wenn man die Flexibität des Modells als eine Folge der Anzahl der Parameter ansieht, so wird ein optimaler Trade-off durch die Geringhaltung der Anzahl der Parameter des Modells erreicht. Daneben kann man aber auch grundsätzlich andere Wege beschreiten, die die Flexibität auch bei hoher Anzahl der Parameter einschränken. Man spricht von der *effektiven* Modellkomplexität, auf deren Geringhaltung diese Ansätze abzielen.

Dies steht in einem grundsätzlichen Zusammenhang mit der Unterscheidung zwischen parametrischen und nicht-parametrischen Verfahren. Bei einem parametrischen Verfahren unterstellt man für den zu modellierenden Zusammenhang eine bestimmte funktionale Form. Jeder einzelne Parameter im Rahmen dieser funktionalen Form ist prinzipiell einer Interpretation zugänglich, z.B. im Rahmen einer ökonomischen Theorie, die für die funktionale Form bestimmend ist. Im Gegensatz dazu erlaubt ein nicht-parametrisches Verfahren die Modellierung eines Zusammenhangs, ohne über dessen funktionale Form a-priori-Annahmen treffen zu müssen. Die Form ergibt sich aus den zur Verfügung stehenden Beobachtungen. Man lässt sozusagen die Daten für sich selbst sprechen. Den einzelnen Parametern, die zur Festlegung des funktionalen Zusammenhangs benötigt werden, kommt keine theoretisch fundierbare Bedeutung zu. Die Anzahl der Parameter dieser Verfahren und damit ihre Flexibilität passt sich jeweils an die Anzahl der vorhandenen Daten an. Typische Vertreter sind Kerndichteschätzungen und Glättungsverfahren.

Während bei parametrischen Verfahren versucht wird, die Anzahl der Parameter und damit die Modellkomplexität auf die „wahre“ Anzahl einzuschränken, stehen bei nicht-parametrischen Verfahren andere Methoden zur Verfügung, so etwa die weiter unten zu beschreibenden Verfahren des Early-Stopping und der Regularisierung.

Neuronale Netze ordnete man anfangs eher den nicht-parametrischen Verfahren zu (vgl. Geman et al., 1992) oder sah sie in einer zwischen den parametrischen und den nicht-parametrischen Verfahren anzusiedelnden Gruppe der semi-

parametrischen Verfahren. Bei diesem Zugang wird die Modellkomplexität durch die Daten selbst bestimmt. Erst durch Arbeiten wie White (1989c), Kuan und White (1994) und Cheng und Titterington (1994) wurde gezeigt, dass NN auch als parametrische Verfahren aufgefasst werden können. Insbesondere durch die Entwicklung von Hypothesentests bezüglich einzelner Gewichte bzw. verdeckter Neuronen (White, 1989c,a; Lee et al., 1993; Teräsvirta et al., 1993) wurde es möglich, die Frage der Modellselektion nach dem Vorbild ökonometrischer Modelle zu behandeln. Dieser Sichtweise wird in wichtigen neueren Arbeiten deutlich der Vorzug gegeben (Anders und Korn, 1999; Medeiros et al., 2006).

Diese zu beobachtende Favorisierung der parametrischen Sichtweise ist auch motiviert durch die Kritik, dass die traditionellen bzw. nicht-parametrischen Ansätze der NN-Modellierung keine (statistische) theoretische Basis aufweisen, sondern bloße Heuristiken sind (vgl. Anders, 1997). Diese Kritik ist nur teilweise berechtigt, da z.B. für den Regularisierungsansatz in der Form der Bayesianischen Theorie eine theoretische Untermauerung vorliegt. Außerdem zeigt sich bei der genauen Betrachtung und der praktischen Umsetzung der parametrischen Verfahren, dass auch sie nicht ohne Heuristiken auskommen. Insbesondere dann, wenn die „hehren“ Annahmen der Asymptotik nicht zutreffen, können auch die Vorgangsweisen im Rahmen des statistisch-parametrischen Ansatzes nur als Heuristiken gelten.

Der wesentliche Punkt ist, dass neuronale Netze nicht per se parametrische oder nicht-parametrische Verfahren darstellen, sondern erst durch die jeweilige *Sichtweise* und durch die Gesamtheit der zur Bestimmung der Modellarchitektur und der Modellschätzung verwendeten Methoden zu solchen Verfahren werden. Die Sichtweise bestimmt also auch, welche Vorgangsweisen zur Gewährleistung der Generalisierungsfähigkeit eingesetzt werden. In diesem Sinne werden in der vorliegenden Arbeit vier verschiedene Ansätze unterschieden und näher auf ihr Potential zum Einsatz für die Modellierung und Vorhersage von makroökonomischen Zeitreihen untersucht.

Statistisch-parametrischer Ansatz Das ARNNDS-Modell wird als parametrisches Modell interpretiert. Jedem Parameter können Standardfehler und Konfidenzniveaus zugeordnet werden. Die Komplexität des Modells wird durch rigorose auf Hypothesentests beruhende Verfahren bestimmt. Als Voraussetzung für den Einsatz dieser kommen im Rahmen der Modellschätzung konvergente Optimierungsverfahren zum Einsatz, die ein Minimum der Fehlerfunktion finden, welches nicht bloß ein lokales sondern ein globales Minimum sein soll.

Klassischer Ansatz mit Early-Stopping Das ARNNDS-Modell wird als Black-Box betrachtet (nicht- bzw. semi-parametrische Sichtweise). Eine potentiell überdimensionierte Modellarchitektur wird in Kauf genommen. Die effektive Modellkomplexität wird durch das vorzeitige Stoppen des Trainings gesteuert (nicht-konvergentes Optimierungsverfahren).

Regularisierungsansatz Es wird ebenfalls eine nicht- bzw. semi-parametrische Sichtweise des ARNNDS-Modells eingenommen. Eine potentiell überdimensionierte Modellarchitektur wird in Kauf genommen. Die effektive Modellkomplexität wird durch Bestrafung von hohen Parameterausprägungen im Rahmen der Zielfunktion reduziert, die neben dem SSE auch einen Regularisierungterm enthält.

Bayesianischer Ansatz Dieser Ansatz ist eine Erweiterung des Regularisierungsansatzes. Die sogenannten Regularisierungsparameter, die das Ausmaß der Bestrafung von hohen Parameterausprägungen bestimmen, werden mit Hilfe des Bayesianischen „Evidence Framework“ bestimmt.

Diese Ansätze können auch als *Modellierungsstrategien* bezeichnet werden. In diesem und dem nächsten Kapitel wird der Ausdruck Modellierungsstrategie vor allem verwendet um eine Konkretisierung im Zusammenhang mit einer bestimmten Anwendung zum Ausdruck zu bringen. So kann es, je nachdem wie bestimmte Detailfragen zur Methodik gelöst werden, im Rahmen eines Ansatzes auch mehrere verschiedene Modellierungsstrategien geben. Die Ansätze bzw. Modellierungsstrategien werden in den Abschnitten 4.6 bis 4.9 ausführlich beschrieben und in anwendungsorientierter Weise anhand einer simulierten nicht-linearen Zeitreihe und der beiden Beispielzeitreihen untersucht.

Eine simulierte nicht-lineare Zeitreihe

Für die Entwicklung von Methoden der nicht-linearen Zeitreihenanalyse werden als Spieldaten oft simulierte Zeitreihen verwendet, deren Eigenschaften man in einer Experimentanordnung kontrollieren kann. Dies ist deshalb sinnvoll, weil neue Methoden nicht gut auf Zeitreihen, deren Eigenschaften nicht bekannt sind, getestet werden können. Auch in der vorliegenden Arbeit empfiehlt es sich, die verwendeten und für die vorliegenden Zwecke weiterentwickelten Methoden, nicht gleich auf dem realweltlichen Anwendungsgebiet der beiden Beispielzeitreihen ALR und IPI zu testen, sondern zunächst ihre Funktionsweise auf einer simulierten Zeitreihe sicher zu belegen, deren wichtigste Eigenschaften bekannt sind.

Die simulierte Zeitreihe soll einen deutlichen nicht-linearen Charakter aufweisen, und die Nicht-Linearität soll von einer Art sein, die von ARNN-Modellen modelliert werden kann, d.h. Nicht-Linearität im bedingten Mittelwert (vgl. Abschnitt 3.1). Neben anderen Möglichkeiten zur Simulation einer solchen Zeitreihe wird dies in geeigneter Weise durch die folgende Formel erreicht:

$$x_t = \begin{cases} \alpha + x_{t-1}^2 + \varepsilon_t & \text{wenn } x_{t-1} \leq \beta, \\ \alpha + \beta + \varepsilon_t & \text{sonst} \end{cases} \tag{4.11}$$

Hierbei sind α und β die zu wählenden Parameter des Simulationsmodells und ε_t ist eine standardnormalverteilte Zufallsvariable. Die Varianz, der lineare und der

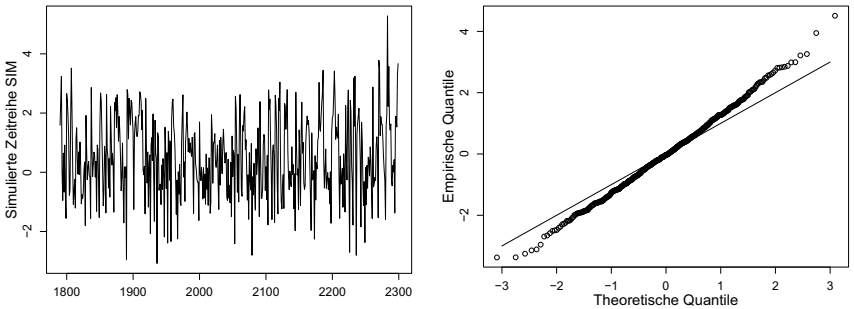


Abb. 9: Simulierte Zeitreihe (SIM): (a) Zeitreihenplot einer Realisation und (b) Quantilsplot der Residuen eines geschätzten AR(10)-Modells

nicht-lineare Varianzanteil der simulierten Zeitreihe können durch Monte-Carlo-Simulation bestimmt werden. Wenn diese Parameter als $\alpha = 1$ und $\beta = -0.55$ gewählt werden, so erhält man eine Zeitreihe ohne lineare Struktur, d.h. sie weist eine Autokorrelation von Null zu allen Lags auf. Ihre Varianz beträgt jedoch 1.167, was einen durch ein ARNN-Modell potentiell erklärbaren Varianzanteil von 0.167 impliziert.[26]

Für die Generierung der in den folgenden Abschnitten verwendeten Zeitreihe (SIM) wurde das Modell (4.11) mit den Parametern $\alpha = 2.3$ und $\beta = -0.6$ verwendet. Mit diesen Parametern hat SIM eine Varianz von 1.858, wovon 0.173 durch ein AR-Modell erklärbar sind. Der größte Anteil der erklärbaren Varianz, 0.685, ist nicht-linearer Natur.

Zu illustrativen Zwecken werden in Abb. 9 ein Zeitreihenplot von SIM und ein Quantilsplot der Residuen eines auf der Zeitreihe SIM geschätzten AR-Modells zusammengefasst. Im Quantilsplot werden die empirischen Quantile der (standardisierten) Residuen gegen die theoretischen Quantile einer Standardnormalverteilung abgetragen. Als Ausdruck der Nicht-Linearität der zugrundeliegenden Zeitreihe zeigt sich, wie zu erwarten, eine deutliche Abweichung der Verteilung von der Normalverteilung. Die empirischen Quantile sind betragsmäßig höher als die theoretischen, die Verteilung ist endlastig bzw. hat „heavy tails“.

4.3 Datentransformation und Parameterinitialisierung

Der erste Schritt der Schätzung eines ARNN-Modells ist die Wahl der Startwerte für die Parameter, ein Problem, das in der NN-Forschung als Initialisierung der Ge-

[26] Im Rahmen der Funktion `simnlts` aus dem R-Paket `NNUTS` werden Funktionalitäten zur Erstellung simulierter Zeitreihen nach Modell (4.11) sowie Informationen zu den entsprechenden linearen und nicht-linearen Varianzanteilen bereitgestellt.

wichte bekannt ist. Dieses Problem ist unabhängig von der Verwendung des nachfolgenden Optimierungsverfahrens von großer Wichtigkeit, da sich zeigt, dass bei komplexen Modellen die Wahl der Startwerte der Parameter das Ergebnis oder zumindest die Geschwindigkeit des Schätzvorgangs beeinflussen kann. Bis zu einem gewissen Grad darf dieses Problem bereits bei komplexen linearen Modellen, z.B. ARMA, nicht unterschätzt werden, mehr noch bei nicht-linearen Modellen. Zudem ist bei NN und ARNN-Modellen ein im Vergleich zu linearen Modellen stärkere Verschränkung mit der Frage der Datentransformation gegeben. Aufgrund der praktischen Bedeutung dieser Fragen insbesondere für die Implementierung und der relativen Besonderheiten, die sich für ARNN- und ARNNDS-Modelle ergeben, wird dieser Schritt nun vergleichsweise ausführlich behandelt.

Verfahren der Initialisierung der Gewichte eines neuronalen Netzes sollten erstens die Skalierung der Daten berücksichtigen, zweitens den Anfangsgewichtsvektor möglichst nahe zu jenem Bereich bringen, in dem das optimale Netz vermutet wird, und drittens dafür Sorge tragen, dass sich von Beginn des Optimierungsverfahrens aussagekräftige, d.h. nicht verschwindende partielle Ableitungen der Fehlerfunktion nach den Gewichten ermitteln lassen. Diese drei Aspekte werden im folgenden erläutert. Bei der Besprechung des ersten Punktes muss zunächst ausführlicher darauf eingegangen werden, inwiefern eine vorgeschaltete Datentransformation (Preprocessing) durch eine geeignete Initialisierung überflüssig gemacht werden kann.

Transformation der Daten

In der Tradition der Datenmodellierung mit neuronalen Netzen ist es üblich, die Eingabe- und Zieldaten eines Neuronalen Netzes einer linearen Transformation

$$\tilde{x}_t = \frac{x_t - \mu_x}{\sigma_x} \tag{4.12}$$

zu unterziehen, wobei die Parameter der Transformation, μ_x und σ_x, in geeigneter Weise gewählt werden. Beispielsweise wird mit

$$\mu_x = 0.5(x_{\max} + x_{\min}) \text{ und } \sigma_x = 0.5(x_{\max} - x_{\min}),$$

wobei $x_{\max}$ und $x_{\min}$ das Maximum bzw. Minimum der Beobachtungen von x_t bezeichnen, erreicht, dass die Daten auf dem Intervall $[-1; 1]$ zu liegen kommen. Und mit

$$\mu_x = T^{-1} \sum_1^T x_t \text{ und } \sigma_x = \sqrt{(T-1)^{-1} \sum_1^T (x_t - \mu_x)^2}$$

wird um den empirischen Mittelwert und die empirische Standardabweichung bereinigt. Solche Transformationen sind in vielen Anwendungen vorteilhaft. Anders (1997, S. 29f), der als Verfechter der vorgeschalteten Transformation angesehen

werden kann, führt eine Reihe wichtiger Vorteile an. Wie im folgenden argumentiert werden soll, ist es jedoch im Kontext von Zeitreihenmodellierung und ARNN-Modellierung besser, auf eine solche vorgeschaltete Transformation *zu verzichten*.

Zunächst ist leicht nachzuprüfen, dass durch die geeignete Wahl der Parameter eines neuronalen Netzes die Skalierung überflüssig gemacht werden kann. Es seien a_{kj}, b_k, c_k und d, die Parameter eines neuronalen Netzes, wobei $j = 1, \ldots, J$, $k = 1, \ldots, K$, J die Anzahl der Eingabeneuronen und K die Anzahl der verdeckten Neuronen ist. Weiters seien x_j, y, $\tilde{x}_j = (x_j - \mu_{x_j})/\sigma_{x_j}$, $\tilde{y} = (y - \mu_y)/\sigma_y$ die erklärenden Variable und die abhängige Variable bzw. ihre jeweiligen Transformationen. Zwischen dem auf den nicht transformierten Daten verwendeten (geschätzten) neuronalen Netz

$$y = d + \sum_{k=1}^{K} c_k \tanh[b_k + \sum_{j=1}^{J} a_{kj} x_j] + \varepsilon \tag{4.13}$$

und seiner Entsprechung auf den transformierten Daten

$$\tilde{y} = \tilde{d} + \sum_{k=1}^{K} \tilde{c}_k \tanh[\tilde{b}_k + \sum_{j=1}^{J} \tilde{a}_{kj} \tilde{x}_j] + \tilde{\varepsilon} \tag{4.14}$$

bestehen folgende Relationen:

$$\tilde{d} = \frac{d - \mu_y}{\sigma_y},\ \tilde{c}_k = \frac{c_k}{\sigma_y},\ \tilde{a}_{kj} = a_{kj}\sigma_{x_j},\ \tilde{b}_k = b_k + \sum_{j=1}^{J} a_{kj}\mu_{x_j},\ \tilde{\varepsilon} = \varepsilon/\sigma_y. \tag{4.15}$$

Die beiden Zugänge sind somit, abgesehen von numerischen Aspekten, äquivalent. Wählt man einen Zugang *ohne vorgeschaltete Transformation* können numerische Probleme vor allem dann auftreten, wenn die einzelnen Inputvariablen unterschiedliche Lage und Streuung aufweisen, was allerdings bei ARNN-Modellen praktisch nicht der Fall ist, da die Inputvariablen hier verzögerte abhängige Variable sind. In einem ARNN-Modell

$$y_t = d + \sum_{k=1}^{K} c_k \psi[b_k + \sum_{j=1}^{J} a_{kj} y_{t-j}] + \varepsilon_t, \tag{4.16}$$

genügt es daher, ein und dieselbe Transformation $\tilde{x}_t = (x_t - \mu_x)/\sigma_x$ auf die abhängige Variable und die erklärenden Variablen anzuwenden. Wenn die ursprüngliche Zeitreihe eine Varianz hat, die nicht sehr stark, etwa um einige Zehnerpotenzen, von Eins abweicht, kann σ_x in der vorgeschalteten Transformation auf Eins gesetzt werden, ohne numerische Probleme zu riskieren. Dann verbleibt in der Spezifikation der vorgeschalteten Transformation nur mehr der Parameter μ_x. Dieser wird nur dann benötigt, wenn das ARNN-Modell ohne Interzept geschätzt werden soll, also $d = 0$ in Gleichung (4.16). Andernfalls kann die vorgeschaltete Transformation

unterbleiben, indem die Mittelwertbereinigung auf die folgende Weise in das Modell integriert wird:

$$y_t - d = \sum_{k=1}^{K} c_k \psi[b_k + \sum_{j=1}^{J} a_{kj}(y_{t-j} - d)] + \varepsilon_t. \tag{4.17}$$

Dies ist die Spezifikation, die bereits im vorigen Abschnitt eingeführt wurde, vgl. Gleichung (4.1) bzw. (4.2).

Betrachtet man den allgemeinen Fall eines ARNNDS-Modells,

$$y_t = \tilde{y}_t + d_t$$

$$d_t = d + d^t t + \sum_{i=2}^{s} d_i^s s_{it}$$

$$\tilde{y}_t = \sum_{k=1}^{K} c_k \psi[b_k + \sum_{j=1}^{J} a_{kj}\tilde{y}_{t-j}] + \sum_{l=1}^{L} f_l \tilde{y}_{t-l} + \varepsilon_t, \tag{4.18}$$

wobei $a_{kj}, b_k, c_k, d, f_l, d^t$ und d_i^s die Parameter des Modells sind, so zeigt sich noch deutlicher, dass eine vorgeschaltete Transformation nicht sinnvoll ist, weil sie mit der Schätzung der deterministischen Komponente im Rahmen des Modells in Konkurrenz treten würde. Ziel einer vorgeschalteten Transformation muss sein, dass die dem neuronalen Netz präsentierten Daten eine gewünschte Skalierung haben. Dieses Ziel kann im ARNNDS-Modell nach Gleichung (4.18) nicht erreicht werden, da vor dem Start des Schätzprozesses die Parameter d, f_l, d^t und d_i^s unbekannt sind. In einem ARNNDS-Modell muss daher die Initialisierung des NN-Teils (a_{kj}, b_k und c_k) bedingt auf die Initialisierung des AR-Teils (Parameter f_l) und des linear-deterministischen Teils (d, d^t und d_i^s) geschehen.

In kurzen Worten zusammengefasst bedeutet dies, dass bei dieser Sichtweise die Modellspezifikation den Daten angepasst wird und nicht umgekehrt.

Initialisierung der Gewichte

Die Initialisierung der Gewichte bzw. Parameter sollte etwaiges Vorwissen nutzen um die Anfangswerte möglichst nahe zu jenem Bereich des Parameterraums zu bringen, in dem das optimale ARNNDS-Modell vermutet wird. Im Falle eines ARNN-Modells bedeutet dies, dass jedenfalls das Wissen, das aus der linearen Modellierung der Zeitreihe gewonnen wurde, in die Initialisierung einbezogen werden soll, z.B. indem die Anfangswerte für d, f_l, d^t und d_i^s mit den entsprechenden Parameterwerten eines geschätzten ARDS-Modells initiert werden. Für die Initialisierung des NN-Teils steht üblicherweise kein Vorwissen zur Verfügung. Deshalb sollten die Gewichte klein gewählt werden.

Die Gewichte des NN-Teils sollten klein gewählt werden, weil auf diese Weise die Aktivierungswerte der verdeckten Neuronen nicht in ihrem saturierten Bereich liegen. Dies ist die Hauptvoraussetzung für nicht verschwindende Ableitungen der Fehlerfunktion nach den Gewichten. Andererseits erscheint es für die optimale Ausnützung der potentiellen Nicht-Linearität wünschenswert, dass die Anfangsgewichte so gewählt werden, dass die Aktivierungsfunktion nicht nur in ihrem linearen Bereich angesprochen wird. Dies wiederum impliziert, dass die Parameter b_k nicht zu klein initialisiert werden sollten.

Als Ergebnis dieser Überlegungen wird nun eine mögliche Vorgangsweise bei der Bestimmung der Anfangsparameter, $\dot{a}_{kj}$, $\dot{b}_k$, $\dot{c}_k$, $\dot{d}$, $\dot{f}_l$, $\dot{d}^t$ und $\dot{d}_i^s$, eines ARNNDS-Modells präsentiert:

(i) Schätze ein linear-deterministisches Modell zur Berechnung der Startwerte für den deterministischen Teil

$$y_t = \dot{d} + \dot{d}^t t + \sum_{i=2}^{s} \dot{d}_i^s s_{it} + u_t. \tag{4.19}$$

(ii) Auf den Residuen u_t dieses Modells schätze ein AR-Modell, dessen Koeffizienten die Startwerte für die Shortcuts liefern,

$$u_t = \chi \sum_{l=1}^{L} \dot{f}_l u_{t-l} + v_t, \tag{4.20}$$

wobei $\chi = 1.2$ eine heuristische Konstante ist, die bewirken soll, dass die Residuen $\tilde{v}_t$ des verzerrten Modells

$$u_t = \sum_{l=1}^{L} \dot{f}_l u_{t-l} + \tilde{v}_t, \tag{4.21}$$

noch einen kleinen Anteil der ursprünglich vorhandenen linearen Struktur beinhalten. Der heuristische Gedanke ist hierbei, dass es zu einer gewissen Arbeitsteilung zwischen dem linearen Teil und dem nicht-linearen Teil, dessen Startwerte im nächsten Schritt bestimmt werden, kommen kann.

(iii) Benütze die Standardabweichung σ_u der Residuen u_t und die Standardabweichung $\sigma_{\tilde{v}}$ der Residuen $\tilde{v}_t$ zur Wahl von $\dot{a}_{kj}$, $\dot{b}_k$ und $\dot{c}_k$ in folgender Weise:

$$\dot{a}_{kj} = \xi_1 \sigma_u^{-1} (1+J)^{-0.5} z_{kj}^a, \quad \dot{b}_k = \xi_2 (1+J)^{-0.5} z_k^b, \quad \dot{c}_k = \sigma_{\tilde{v}} K^{-0.5} z_k^c, \tag{4.22}$$

wobei z_{kj}^a, z_k^b und z_k^c aus der Standardnormalverteilung gezogen und die heuristischen Konstanten ξ_1 und ξ_2 im Intervall $(0;1)$ gewählt werden, z.B. $\xi_1 = 0.3$ und $\xi_2 = 0.1$.

Diese Vorgangsweise verwirklicht einen guten Kompromiss zwischen einer zu vorsichtigen Initialisierung der Modellparameter und einer solchen, in der die potentielle Nicht-Linearität des Modells bereits durch die Anfangsparameter zum Tragen kommt. Die Vorgangsweise wurde auch im Rahmen der für diese Arbeit entwickelten Software `NNUTS` implementiert und für den empirischen Teil eingesetzt.

4.4 Lokale Optimierung

Die Schätzung eines ARNN-Modells stellt ein Optimierungsproblem eines Typs dar, für den die Mathematik in den letzten Jahrzehnten zahlreiche ausgefeilte numerische Verfahren entwickelt hat. Die NN-Forschung hat diese Verfahren aufgegriffen bzw. ausgehend von der zentralen Idee des Backpropagation erweitert und bereichert.

Im folgenden werden jene Techniken und Verfahren präsentiert, die im Rahmen dieser Arbeit implementiert und für die empirischen Analysen verwendet wurden. Für eine vertiefte formelmäßige Darstellung sei auf Bishop (1995) und Press et al. (1992) bzw. auf die konkrete zitierte Literatur verwiesen.

Grundlagen

Die hier besprochenen Verfahren sind Verfahren der lokalen Optimierung. Dies bedeutet, dass ein Minimum in Bezug auf die lokale Umgebung im Suchraum angestrebt wird. Im Gegensatz dazu ist ein globales Minimum ein Punkt im Suchraum, an dem die Fehlerfunktion minimal im Vergleich zum gesamten Suchraum ist. Verfahren der globalen Optimierung werden im Abschnitt 4.4 besprochen. Die Abgrenzung zwischen lokaler und globaler Optimierung ist nicht immer eindeutig, wie anhand einiger Merkmale und Verbesserungen an den im Folgenden besprochenen Verfahren erläutert werden soll.

Gemeinsam ist fast allen Verfahren der lokalen Optimierung sowie den meisten Verfahren der globalen Optimierung, dass sie iterativ vorgehen. Der zu optimierende Parametervektor sei mit $\mathbf{w}$ bezeichnet, umfasse also im Falle der ARNNDS-Modellierung alle Parameter des Modells, $\mathbf{w}' = (a_{11}, a_{12}, \ldots, a_{KJ}, b_1, \ldots, b_K, c_1, \ldots, c_K, d, f_1, \ldots, f_L, d^t, d_2^s, \ldots, d_s^s)$, wobei K die Anzahl der verdeckten Neuronen, J die Anzahl der Input-Neuronen, L die Anzahl der Direktverbindungen (Shortcuts) und s die Anzahl der Saisonen ist. Beginnend bei einem Startwert $\mathbf{w}^{(0)}$ durchläuft das Verfahren in einer Suche eine Abfolge von $\mathbf{w}^{(\tau)}$, wobei τ den Iterationszeitpunkt bezeichnet. Der Algorithmus besteht nun darin, in jeder Iteration die notwendige Veränderung $\Delta\mathbf{w}^{(\tau)}$ zu bestimmen, mit der sich der neue Parametervektor ergibt:

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} + \Delta\mathbf{w}^{(\tau)} \tag{4.23}$$

Die Suche wird so lange fortgesetzt, bis ein Abbruchkriterium erfüllt ist. Dies ist bei konvergenten Verfahren, die das Erreichen eines lokalen Minimums anstreben, die Konvergenz, insofern das Verfahren erfolgreich ist. Die Konvergenz wird

üblicherweise an der Veränderung der Zielfunktion $f(w^{(\tau)})$ festgemacht. So stoppt etwa die in R enthaltene Optimierungsfunktion `optim`, die auch in `NNUTS` verwendet wird, wenn es in einem Iterationsschritt nicht gelingt, den Wert der Zielfunktion um $\epsilon(|E(\mathbf{w}^{(\tau)})| + \epsilon)$ zu verringern, wobei ϵ eine relative Toleranzgrenze angibt. Es ist auch denkbar, andere oder zusätzliche Abbruchkriterien zu verwenden, die sich etwa auf die Veränderung von $\mathbf{w}$ selbst beziehen. Die sorgfältige und wohldokumentierte Definition des Konvergenzkriteriums ist ein oft unterschätztes Detail der Numerik (vgl. Yalta und Yalta, 2010).

Handelt es sich beim primären Abbruchkriterium nicht um die Konvergenz, so spricht man von nicht-konvergenten Verfahren. Im einfachsten Falle wird eine vorgegebene Anzahl von Iterationen trainiert. Beim „Early-Stopping“ beruht das Abbruchkriterium auf dem Wiederanstieg der Fehlerfunktion auf einem Validierungsset (vgl. Abschnitt 4.7). Was immer das primäre Abbruchkriterium ist, iterative Verfahren benötigen immer auch ein weiteres Abbruchkriterium für den Fall des Misserfolgs, um eine zu lange andauernde Suche zu verhindern, meist die Vorgabe einer maximalen Anzahl von Iterationen.

Es sei darauf hingewiesen, dass bei Erfüllung des Konvergenzkriteriums keineswegs sichergestellt ist, dass der gefundene Parametervektor ein (lokales oder globales) Minimum der Zielfunktion ist. Es könnte sich auch um einen Sattelpunkt oder um ein sehr flaches Tal der „Fehlerlandschaft“ handeln.

Gradientenabstiegsverfahren und Backpropagation

Das grundsätzliche Verfahren der lokalen Optimierung kontinuierlicher Funktionen ist das Gradientenabstiegsverfahren (engl. Gradient Descent, im folgenden teilweise mit GD abgekürzt), welches in jeder Iteration im wesentlichen den Funktionswert und den Gradienten, d.h. den Vektor der partiellen Ableitungen der Zielfunktion nach den Gewichten, im Punkt $\mathbf{w}^{(\tau)}$ verwendet und daher zu den Gradientenverfahren 1. Ordnung gehört. Im Gegensatz dazu verwenden kombinatorische Verfahren und Simulated-Annealing-Methoden (siehe hierzu den Abschnitt 4.5) bloß den Funktionswert, und Gradientenverfahren 2. Ordnung (diese werden weiter unten erläutert) verwenden auch die 2. Ableitungen.

Beim Gradientenabstiegsverfahren wird in jeder Interation ein kleiner Schritt in Richtung des negativen Gradienten gegangen:

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \mathbf{g}^{(\tau)}, \tag{4.24}$$

wobei $\mathbf{g}^{(\tau)} = \nabla E|_{\mathbf{w}^{(\tau)}}$ der Gradient an der Stelle $\mathbf{w}^{(\tau)}$ ist und η die Schrittweite angibt und im Zusammenhang mit dem Training eines neuronalen Netzes als Lernrate bezeichnet wird. Für ein anschauliches Verständnis des GD-Verfahrens im speziellen sowie von Optimierungsverfahren allgemeinen wird oft die Metapher einer Fehlerlandschaft herangezogen. Man denkt sie sich als eine Oberfläche über dem Parameterraum, deren Höhe den Funktionswert angibt und welche Berge und

Täler aufweist. Bei einem Minimierungsproblem gilt es, den tiefstgelegenen Punkt eines Tales zu finden. Das GD-Verfahren schlägt auf seinem Weg durch die Fehlerlandschaft in jedem Schritt die Richtung des steilsten Abstiegs ins Tal ein.

Das Prinzip des Gradientenabstiegs ist bei neuronalen Netzen eng mit dem Begriff des Backpropagation (BP) verbunden. BP bezeichnet entweder eine Technik, die partiellen Ableitungen der Fehlerfunktion nach den Gewichten eines NN zu bestimmen, oder einen Lernalgorithmus nach dem Prinzip von Gleichung (4.24), der jedoch nur unter Einschränkungen als Gradientenabstiegsverfahren interpretiert werden kann. Die beiden Deutungsweisen sollen nun getrennt erläutert werden.

BP als *Technik* ist eine Methode, die Ableitungen der Fehlerfunktion nach den einzelnen Gewichten zu berechnen, wobei nur lokal bei den jeweiligen Gewichten vorhandene Informationen verwendet werden. Während im Feedforward-Modus der Ausgabewert eines NN berechnet wird, indem jeder Knoten nur jene Informationen verarbeitet, die er von den vorgelagerten Knoten bekommt, wird bei BP die Flussrichtung der Informationen umgekehrt und jeder Knoten verarbeitet die Informationen, die er von den ihm nachgelagerten Knoten bekommt. Der Fehler wird ausgehend vom bzw. von den Ausgabeknoten durch das Netz zurückpropagiert. Dieses Prinzip funktioniert nicht nur für Feedforward-Netze mit nur einer verdeckten Schicht, wie sie als Teil eines ARNN-Modell eingesetzt werden, sondern generell in Feedforward-Netzen mit beliebig vielen Schichten und Ausgabeknoten, sowie mit beliebigen Direktverbindungen zwischen den Knoten. (Rojas, 1993, 1996) zeigt dies anschaulich anhand eines graphenanalytischen Ansatzes.

Das BP-Prinzip ermöglicht die Implementierbarkeit im Rahmen von parallelen Strukturen und ist damit für die Künstliche-Intelligenz-Forschung wichtig. Es bringt allerdings für die NN-Modellierung im Rahmen der nicht-linearen Statistik und Zeitreihenanalyse kaum Vorteile. Wie im Anhang A demonstriert wird, können die partiellen Ableitungen der Fehlerfunktion nach den Parametern eines ARNN-Modells relativ leicht unter Benutzung der Summen-, Produkt- und Kettenregel der Differentialrechnung berechnet werden. Bei einer effizienten Softwareimplementation muss u.a. beachtet werden, dass Terme, die in den Formeln für die verschiedenen Parameter mehrmals aufscheinen, zwischengespeichert werden und so nur einmal berechnet werden. In Anhang A wird darüber hinaus aufgezeigt, dass auch Erweiterungen der Netzarchitektur im Sinne eines ARNN- bzw. ARNNDS-Modells, d.h. saisonale Terme, Trends usw., in effizienter Weise bei der Berechnung der Ableitungen berücksichtigt werden können.

Backpropagation als *Lernalgorithmus* existiert in verschiedenen Versionen. In der Batch-Version wird in jeder Iteration das gesamte Datenset ausgewertet, und das Verfahren wird daher durch Gleichung (4.24) vollständig beschrieben. In der Online-Version des Verfahrens wird hingegen in jeder Iteration der Gradient bezüglich dem quadrierten Fehlers für nur einen einzelnen Datenpunkt ausgewertet und der Gewichtsvektor sofort angepasst:

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \mathbf{g}^{(i,\tau)}, \tag{4.25}$$

wobei $\mathbf{g}^{(i,\tau)} = \nabla E^{(i)}\big|_{\mathbf{w}^{(\tau)}}$ der Gradient bezüglich des i-ten Datenpunktes an der Stelle $\mathbf{w}^{(\tau)}$ ist. Die Datenpunkte können sequentiell oder in zufälliger Reihenfolge präsentiert werden. Eine weitere Möglichkeit, die als epochenbasiert bezeichnet wird, besteht darin, in jeder Iteration eine Teilmenge des Datensets zu präsentieren. Die Mächtigkeit dieser Menge wird als Epochengröße bezeichnet.

Es ist klar, dass die Online- und die epochenbasierte Version von Backpropagation nicht als GD-Verfahren gelten können. In approximativen Anwendungen können sie nicht konvergieren, auch wenn die Datenpunkte beliebig oft dem Algorithmus präsentiert werden. Ihre Vorteile kommen einerseits dann zum Tragen, wenn die Daten zum Zeitpunkt des Beginns des Trainings noch nicht vollständig bekannt sind oder wenn Redundanz in den Daten vorhanden ist. Diese Vorteile spielen in der Künstliche-Intelligenz-Forschung eine Rolle, wenig hingegen in der nicht-linearen Zeitreihenanalyse. Ein anderer Vorteil ist es, welcher diese Methoden auch für die vorliegende Arbeit interessant macht, nämlich, dass beim Online- oder epochenbasierten Lernen auch Schritte in eine Richtung möglich sind, in der die Fehlerfunktion, also E, nicht $E^{(i)}$, ansteigt. Dadurch entsteht das Potenzial, ein inferiores lokales Minimum zu verlassen. Die Verfahren weisen damit ansatzweise Eigenschaften der globalen Optimierung auf. Um zu erreichen, dass das Verfahren dennoch konvergiert oder zumindestens eine Region mit niedrigen Funktionswerten mit hoher Wahrscheinlichkeit nicht mehr verlässt, kann man z.B. während des Trainings die Lernrate abnehmen oder die Epochengröße anwachsen lassen. Diese Möglichkeit wird weiter unten aufgegriffen.

GD-Verfahren sind generell relativ ineffiziente Optimierungsverfahren. In flachen Regionen der Fehlerlandschaft kann das Training nur sehr geringe Fortschritte machen. In langgestreckten Tälern wiederum kommt es zu oszillierenden Iterationspfaden, d.h. zu große Schrittweiten für einzelne Gewichte. Diesen Schwächen soll durch verschiedene Erweiterungen abgeholfen werden, unter denen das Lernen mit Momentum-Term die einfachste und bekannteste ist. Hierbei handelt es sich um eine zweite Komponente, die in die Iterationsformel eingeht, welche den vorangegangenen Iterationsschritt berücksichtigt:

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} + \Delta\mathbf{w}^{(\tau)} = \mathbf{w}^{(\tau)} - \eta\mathbf{g}^{(\tau)} + \zeta\Delta\mathbf{w}^{(\tau-1)}, \tag{4.26}$$

wobei ζ der Momentum-Parameter ist. Der Momentum-Term bewirkt, dass der Lernfortschritt in relativen flachen Regionen, in denen sich der Gradient über mehrere Iterationen hinweg kaum ändert, beschleunigt wird. So ergibt sich im angenommenen Falle, dass der Gradient gleich bleibt, eine theoretische Beschleunigung um den Faktor $1/(1-\zeta)$. Andererseits werden Oszillationen gedämpft. BP mit Lernrate und Momentum ist damit etwas robuster und schneller als der reine BP-Algorithmus. Allerdings ist nun nicht nur für die Lernrate sondern auch für den Momentum-Parameter ein numerischer Wert vorzugeben. Für diese Wahl stehen allenfalls Heuristiken und die Methode des Ausprobierens zur Verfügung, da sie in unüberschaubarer Weise von den Eigenschaften der Fehlerfunktion abhängt. Wei-

tere Verbesserungen des GD-Verfahrens, die von der NN-Forschung entwickelt wurden, werden unten skizziert.

Implementation eines GD-Verfahrens mit optimierter Lernrate, Momentum und abnehmender Epochengröße

Hier wird nun eine einfache Modifikation von BP mit Lernrate und Momentum vorgeschlagen, die in der vorliegenden Arbeit für die Implementierung des klassischen Ansatzes der ARNN-Modellierung mit Early-Stopping (vgl. Abschnitt 4.7) zum Einsatz kommt. Bei diesem Ansatz wird in jeder Iteration die optimale Schrittweite ermittelt, auf die sich die Lernrate sodann bezieht:

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} + \Delta\mathbf{w}^{(\tau)} = \mathbf{w}^{(\tau)} - \eta\kappa^{(\tau)}\mathbf{g}^{(\tau)} + \zeta\Delta\mathbf{w}^{(\tau-1)}, \tag{4.27}$$

wobei $\kappa^{(\tau)}$ die optimale Schrittweite bezeichnet, d.h. den Wert, für den die Funktion $\tilde{E}(\kappa^{(\tau)}) = E(\mathbf{w}^{(\tau)} - \kappa^{(\tau)}\mathbf{g}^{(\tau)})$ ein Minimum annimmt. Die optimale Schrittweite wird mit dem Linesearch-Algorithmus bestimmt. Für eine Beschreibung dieses Verfahrens siehe Bishop (1995), Press et al. (1992) oder die Dokumentation und den Sourcecode der Funktion `optstepwidth` im R-Paket `NNUTS`, in dem ein einfacher Linesearch implementiert ist.

Gleichung 4.27 ist die Batch-Version des Verfahrens, die epochenbasierte Version lautet

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} + \Delta\mathbf{w}^{(\tau)} = \mathbf{w}^{(\tau)} - \eta\kappa^{(\mathcal{I},\tau)}\mathbf{g}^{(\mathcal{I},\tau)} + \zeta\Delta\mathbf{w}^{(\tau-1)}, \tag{4.28}$$

wobei $g^{(\mathcal{I},\tau)} = \nabla E^{(\mathcal{I})}\big|_{\mathbf{w}^{(\tau)}}$ der Gradient an der Stelle $\mathbf{w}^{(\tau)}$ bezüglich der Fehlerfunktion über der mit $\mathcal{I}$ bezeichneten Epoche ist und $\kappa^{(\mathcal{I},\tau)}$ die optimale Schrittweite im Hinblick auf $g^{(\mathcal{I},\tau)}$ und die Epoche $\mathcal{I}$ ist.

Neben dem Aspekt der Berücksichtigung der optimalen Schrittweite, die auch die Anwenderfreundlichkeit des Verfahrens erhöht, ist es vor allem der Aspekt der variablen Epochengröße, der das verbesserte Backpropagation-Verfahren attraktiv für die ARNN-Modellierung macht. Durch die Möglichkeit, die Epochengröße zu Beginn klein zu wählen und im Verlauf des Trainings anwachsen zu lassen, bekommt das Verfahren eine stochastische Komponente und bleibt nicht so leicht in einem inferioren Minimum hängen. Im Rahmen der vorliegenden Arbeit wurde folgende Formel für das Anwachsen der Epochengröße gewählt:

$$\pi^{(\tau+1)} = \begin{cases} \pi^{(\tau)} & \text{wenn } \tau \bmod \tau^* > 0 \\ \pi^{(\tau)} + \xi(N - \pi^{(\tau)}) & \text{sonst} \end{cases} \tag{4.29}$$

Hierbei ist $\pi^{(\tau)}$ die Epochengröße in Iteration τ gemessen in Datenpunkten, τ^* gibt an, alle wieviel Iterationen die Epochengröße angepasst werden soll, N ist die Anzahl der Datenpunkte, d.h. die maximal verfügbare Epochengröße und ξ bestimmt die Geschwindigkeit des Anwachsens der Epochengröße. Für ξ empfiehlt sich ein kleiner Wert unter Eins, z.B. $\xi = 0.05$. Außerdem muss die initiale Epochengröße

$\pi^{(0)}$ vorgegeben werden. Auf die Berücksichtigung der notwendigen Rundung der Epochengröße auf ganze Zahlen wurde in der obigen Formel verzichtet. Je nach Wahl der Parameter konvergiert die Epochengröße unterschiedlich schnell nach N und nähert sich so dem Batchverfahren an. Auf diese Weise wird bei ausreichend langem Training eine Konvergenz des Lernverfahrens ermöglicht.

Überblick über andere verbesserte GD-Verfahren

Die Gradientenabstiegsverfahren wurden in vielerlei Hinsicht über die bereits besprochenen Aspekte hinaus erweitert und verbessert, insbesondere mit dem Ziel, das Training zu beschleunigen. Ausgehend von der zentralen Idee des Backpropagation wurden von der NN-Forschung spezifisch für neuronale Netze Methoden entwickelt, die teilweise Aspekte der Verfahren 2. Ordnung aufweisen. So gibt es verschiedene Verfahren, die separate Lernraten η_i für jedes Gewicht w_i verwenden und die für die iterative Anpassung von η_i berücksichtigen, ob die partielle Ableitung der Fehlerfunktion für aufeinander folgende Iterationen gleiche oder unterschiedliche Vorzeichen haben. Unter diesen Verfahren sind die delta-delta- bzw. delta-bar-delta-Regel zu nennen (Jacobs, 1988; Silva und Almeida, 1990). Das bekannte Quickprop-Verfahren (Fahlmann, 1989) passt für die Bestimmung von η_i eine nach oben geöffnete Parabel an die Fehlerfunktion an. Zwar mögen diese Verfahren innerhalb der Zielsetzungen der NN-Forschung Bedeutung haben (z.B. in Hinblick auf Implementierbarkeit in parallelen Architekturen, Verknüpfung mit anderen NN-spezifischen Techniken wie dem Online-Lernen). Für die vorliegende Arbeit und die Schätzung von ARNN-Modellen auf makroökonomischen Zeitreihen fallen diese Vorteile hingegen kaum ins Gewicht. Wenn die Konvergenz zu einem (lokalen) Minimum der Fehlerfunktion angestrebt wird, wird daher besser auf bedeutend leistungsfähigere Verfahren der Lokalen Optimierung zurückgegriffen, auf die nun der Fokus gerichtet wird.

Newton-Verfahren und Quasi-Newton-Verfahren

Im Gegensatz zu Gradientenverfahren benutzen Optimierungsverfahren 2. Ordnung auch die zweiten partiellen Ableitungen der Fehlerfunktion und erreichen daher eine im Vergleich deutlich schnellere Konvergenz. Diese Verfahren wurden bereits ab den 50er Jahren für Berechnungen auf den damaligen Hochleistungsrechnern im Zusammenhang mit physikalischen Experimenten entwickelt (im Falle des Quasi-Newton-Verfahrens siehe z.B. Davidon, 1991) und liegen inzwischen in sehr ausgefeilten Versionen und Softwareimplementierungen vor.

Zwischen den beiden Gruppen der Verfahren der konjugierten Gradienten und der Quasi-Newton-Methoden (engl. quasi-Newton bzw. variable metric methods, im Folgenden wird teilweise die Abkürzung QN verwendet), wird meist den letzteren der Vorzug gegeben, ohne dass ein zwingender Grund angegeben werden kann. Im

Rahmen der vorliegenden Arbeit wurde das Quasi-Newton-Verfahren in der Variante BFGS verwendet, das als Softwareimplementation im Rahmen von R in der Funktion `optim` zur Verfügung steht. Dieses Verfahren sei in seinen Grundsätzen und der dahinterstehenden Intuition kurz skizziert. Verschiedene ausführliche Darstellungen finden sich bei Luenberger (1984), Press et al. (1992), Bishop (1995) und Polak (1997).

Das QN-Verfahren ist eine Weiterentwicklung des grundsätzlicheren Newtonverfahrens, welches daher zuerst besprochen werden soll. Es beruht auf der lokalen quadratischen Approximation der Fehlerfunktion durch eine Taylorreihenentwicklung 2. Ordnung (ohne Restglied):

$$E(\mathbf{w}) = E(\hat{\mathbf{w}}) + (\mathbf{w} - \hat{\mathbf{w}})'\mathbf{g}_{\hat{\mathbf{w}}} + \frac{1}{2}(\mathbf{w} - \hat{\mathbf{w}})'\mathbf{H}(\mathbf{w} - \hat{\mathbf{w}}). \tag{4.30}$$

Hierbei ist $\hat{\mathbf{w}}$ der Punkt, nach dem die Taylorreihe entwickelt wird, $\mathbf{g}_{\hat{\mathbf{w}}}$ bezeichnet den Gradienten und $\mathbf{H}$ die Matrix der 2. partiellen Ableitungen (Hesse-Matrix) im Punkt $\hat{\mathbf{w}}$:

$$\mathbf{g}_{\hat{\mathbf{w}}} = \nabla E|_{\hat{\mathbf{w}}}\,, \quad (\mathbf{H})_{ij} = \left.\frac{\partial^2 E}{\partial w_i \partial w_j}\right|_{\hat{\mathbf{w}}} \tag{4.31}$$

Wird für $\hat{\mathbf{w}}$ ein lokales Minimum der Fehlerfunktion, $\mathbf{w}^*$, gewählt, ist $\mathbf{g}_{\hat{\mathbf{w}}} = \mathbf{0}$ und es entfällt der zweite Term auf der rechten Seite. Die Gleichung lautet nun

$$E(\mathbf{w}) = E(\mathbf{w}^*) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)'\mathbf{H}(\mathbf{w} - \mathbf{w}^*). \tag{4.32}$$

Der Gradient im Punkt $\mathbf{w}$ ergibt sich damit als

$$\mathbf{g} = \nabla E = \mathbf{H}(\mathbf{w} - \mathbf{w}^*), \tag{4.33}$$

woraus unmittelbar als Lösung für das Minimum der quadratischen Approximation folgt:

$$\mathbf{w}^* = \mathbf{w} - \mathbf{H}^{-1}\mathbf{g}. \tag{4.34}$$

Der Vektor $-\mathbf{H}^{-1}\mathbf{g}$ wird als Newton-Schritt bezeichnet und ist die Basis der Newton- bzw. Quasi-Newton-Verfahren. Dieses Ergebnis zeigt, dass wenn es sich bei der zu minimierenden Funktion tatsächlich um eine quadratische Funktion der Parameter handelt, das Minimum in einem einzigen Schritt erreicht werden kann. Da die Taylorreihenentwicklung jedoch nur annähernd zutrifft, muss im Newton-Verfahren der Schritt mehrmals gesetzt werden und in jeder Iteration die inverse Hesse-Matrix neu aktualisiert werden. Da zudem weitab vom tatsächlichen Minimum die Approximation wahrscheinlich relativ ungenau ist, kann die inverse Hesse-Matrix negative Eigenwerte haben oder der Newton-Schritt über das Minimum hinausführen, woraus möglicherweise eine Anstieg der Fehlerfunktion durch den Newton-Schritt bewirkt wird. Entsprechende numerische Modifikationen verhindern dies.

Der Hauptnachteil des gewöhnlichen Newton-Verfahrens ist in der praktischen Anwendung jedoch, dass die Inverse der Hesse-Matrix in jeder Iteration oder zumindest periodisch neu berechnet werden muss. Dies ist sehr rechenintensiv. Die Quasi-Newton-Verfahren (QN) vermeiden diesen Berechnungsschritt und bauen statt dessen über die einzelnen Interationsschritte hinweg eine Annährerung der inversen Hesse-Matrix auf, wobei die in jedem Schritt τ aus der Evaluierung der Funktion und des Gradienten in $\mathbf{w}^{(\tau)}$ gewonnnenen Informationen akkumuliert werden. Das Akkumulationsprinzip ergibt sich aus der sogenannten Newton-Bedingung

$$(\mathbf{w}^{(\tau+1)} - \mathbf{w}^{(\tau)}) = -\mathbf{H}^{-1}(\mathbf{g}^{(\tau+1)} - \mathbf{g}^{(\tau)}), \tag{4.35}$$

die man erhält, wenn man Gleichung (4.33) für zwei verschiedene τ substrahiert und dabei annimmt, dass die Hesse-Matrix konstant ist. Definiert man

$$\mathbf{s} = \mathbf{w}^{(\tau+1)} - \mathbf{w}^{(\tau)} \text{ und } \mathbf{v} = \mathbf{g}^{(\tau+1)} - \mathbf{g}^{(\tau)}, \tag{4.36}$$

und nimmt man an, dass die Evaluation nach Gleichung (4.35) p mal durchgeführt wird, wobei p hier die Anzahl der Paramter in $\mathbf{w}$ ist, so erhält man p Gleichungen $\mathbf{s}^{(i)} = -\mathbf{H}^{-1}\mathbf{v}^{(i)}$, $i = 1, \ldots, p$, die zu einem Gleichungssystem

$$\mathbf{S} = -\mathbf{H}^{-1}\mathbf{V} \tag{4.37}$$

zusammengefasst werden. Dies zeigt, dass im Falle der Suche nach dem Minimum einer quadratischen Form die negative Inverse der Hesse-Matrix nach p Evaluationen von Gleichung (4.35) als $-\mathbf{H}^{-1} = \mathbf{S}\mathbf{V}^{-1}$ ermittelt werden kann und liefert die Intuition, wie das QN-Verfahren dasselbe auf iterativem Wege erreicht.

Im iterativen Verfahren muss die Newton-Bedingung auch gelten, wenn man statt $\mathbf{H}^{-1}$ die Annäherung $\mathbf{G}^{(\tau)}$ einsetzt. Die Aktualisierung von $\mathbf{G}^{(\tau)}$ erfolgt gemäß der Broyden-Fletcher-Goldfarb-Shanno-Version (BFGS) des QN-Verfahrens nach folgender Formel:

$$\mathbf{G}^{(\tau+1)} = \mathbf{G}^{(\tau)} + \frac{\mathbf{s}\mathbf{s}'}{\mathbf{s}'\mathbf{v}} - \frac{(\mathbf{G}^{(\tau)}\mathbf{v})\mathbf{v}'\mathbf{G}^{(\tau)}}{\mathbf{v}'\mathbf{G}^{(\tau)}\mathbf{v}} + (\mathbf{v}^{\mathrm{T}}\mathbf{G}^{(\tau)}\mathbf{v})\mathbf{u}\mathbf{u}', \tag{4.38}$$

wobei

$$\mathbf{u} = \frac{\mathbf{s}}{\mathbf{s}'\mathbf{v}} - \frac{\mathbf{G}^{(\tau)}\mathbf{v}}{\mathbf{v}'\mathbf{G}^{(\tau)}\mathbf{v}}. \tag{4.39}$$

Diese Formel enthält als dritten und vierten Term auf der rechten Seite Korrekturterme, die dafür sorgen, dass die Matrix $\mathbf{G}^{(\tau)}$ in jedem Iterationsschritt positiv-definit bleibt, sofern man mit $\mathbf{G}^{(1)} = \mathbf{I}$ beginnt, wobei $\mathbf{I}$ die Einheitsmatrix ist. Die Ausgestaltung dieser Korrekturterme ist der Hauptunterschied zum sehr ähnlichen Davidon-Fletcher-Powell Verfahren, das als numerisch etwas weniger robust gilt.

Der Iterationsschritt des QN-Verfahrens ist

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \kappa^{(\tau)}\mathbf{G}^{(\tau)}\mathbf{g}^{(\tau)}, \tag{4.40}$$

wobei $\kappa^{(\tau)}$ die in jeder Iteration durch das Linesearch-Verfahren bestimmte optimale Schrittweite ist.

In der Praxis der Schätzung von ARNN-Modellen ist QN bzw. BFGS sehr gut einsetzbar. Dennoch können sich bei ARNN-Modellen je nach den Eigenschaften der Zeitreihe und bei komplexen und nicht-identifizierten Modellen beträchtliche Abweichungen der Fehlerfunktion von einer quadratischen Form ergeben, was zu Konvergenzproblemen führt. In solchen Fällen kann auch das QN-Verfahren, durchaus ähnlich wie das GD-Verfahren, in Regionen steckenbleiben, in denen die Fehlerfunktion nur sehr langsam abnimmt und es zu keiner Konvergenz kommt.

4.5 Globale Optimierung

Es wurde bereits betont, dass die Schätzung eines neuronalen Netzes mit mehreren Parametern ein hochdimensionales, nicht-lineares Optimierungsproblem darstellt. Eine besondere Schwierigkeit besteht darin, dass das Auftreten von mehreren lokalen Minima in der Fehlerfunktion hierbei die Regel ist. Neben der Anzahl der zu bestimmenden Parameter hängt das Auftreten von mehreren lokalen Minima insbesondere vom Verhältnis des Signals zum Fehlerprozess in der Zeitreihe (Signal-to-Noise-Ratio) und vom Umfang des Datensets ab. In beider Hinsicht sind makroökonomische Anwendungen ungünstig gelagert (niedriges Signal-to-Noise-Ration, geringer Umfang des Datensets).

Im allgemeinen will man nicht-globale lokale Minima vermeiden, insoferne man überhaupt nach einem Optimum sucht. Insbesondere im Rahmen des statistisch-parametrischen Ansatzes ist es wichtig, nicht nur ein lokales sondern ein globales Minimum zu finden. Denn in diesem Fall hängt die Berechnung der Konfidenzintervalle von der Annahme eines globalen Optimums ab.

Die Literatur über globale Optimierung ist sehr umfangreich und breitgefächert, da globale Optimierungsprobleme in zahlreichen Anwendungsgebieten auftreten. Die entwickelten Algorithmen sind im allgemeinen um ein vielfaches rechenintensiver als die Methoden der lokalen Optimierung. Der Praktiker muss daher einen Kompromiss zwischen Verlässlichkeit und (Zeit-)Effizienz finden. Mit Simulated Annealing und verschiedenen Varianten des Multistart-Verfahrens werden im folgenden einige der gebräuchlichsten Algorithmen kurz vorgestellt und auf ihre praktische Verwendbarkeit zur Schätzung von ARNN-Modellen geprüft.

Simulated Annealing

Simulated Annealing (dt. simulierte Abkühlung, im Folgenden kurz SA) ist ein stochastisches globales Optimierungsverfahren. Das Verfahren beruht auf einer Nachbildung des physikalischen Abkühlungsprozesses, bei dem durch langsame Abkühlung erreicht wird, dass ein Werkstoff einen energiearmen, d.h. optimalen Zustand annimmt. Die physikalischen Energie entspricht der Zielfunktion. Die Temperatur bestimmt die Akzeptanzschwelle mit der beim Übergang zum nächsten Kandidaten

auch ein Kandidat mit einem höheren Energieniveau akzeptiert wird. Simulated Annealing basiert auf dem Metropolis-Algorithmus (Metropolis et al., 1953). Man befinde sich in Iteration τ am Punkt $\mathbf{w}^{(\tau)}$ im Suchraum und habe als nächsten Kandidaten $\dot{\mathbf{w}}^{(\tau+1)}$ zufällig ausgewählt. Der neuer Kandidat wird nach folgender Akzeptanzregel angenommen:

$$\text{Setze } \mathbf{w}^{(\tau+1)} = \begin{cases} \dot{\mathbf{w}}^{(\tau+1)} & \text{wenn } \Delta E < 0 \text{ oder } z < \exp(-\Delta E / \mathit{Temp}), \\ \mathbf{w}^{(\tau)} & \text{sonst} \end{cases} \tag{4.41}$$

wobei $\Delta E = E(\dot{\mathbf{w}}^{(\tau+1)}) - E(\mathbf{w}^{(\tau)})$ ist, $z, 0 \leq z \leq 1$ eine soeben ermittelte Zufallszahl und *Temp* die Temperatur ist. Bei einem Metropolis-Algorithmus mit im Zeitverlauf abnehmender Temperatur handelt es sich um Simulated Annealing.

Bei globalen Optimierungsproblemen in kontinuierlichen Parameterräumen (z.B. Schätzung eines ARNN-Modells) kommen spezielle Varianten des SA zum Einsatz. Der Algorithmus muss eine Methode vorsehen, wie der nächste Kandidat im Parameterraum bestimmt wird, der einer Überprüfung auf Akzeptanz unterzogen werden soll. Der Kandidat wird aus einer Zufallsverteilung gezogen, deren Skalierung (ebenfalls) von der Temparatur abhängig gemacht wird. Bélisle (1992) präsentiert theoretische Ergebnisse zur Konvergenz eines Algorithmus mit einem Schema logarithmischer Abkühlung. Der in R in der Funktion `optim` implementierte SA-Algorithmus beruht auf Bélisle (1992).

Da der SA-Algorithmus nur auf einer Auswertung der Zielfunktion in jeder Iteration beruht, ist er robust aber relativ langsam. Da sich das Verfahren gut eignet, um einen großen Kandidatenraum abzudecken, aber bei der Konvergenz zu einem lokalen Minimum vergleichsweise versagt, empfiehlt es sich, SA nicht bis zur Konvergenz durchzuführen sondern nur so lange, bis man erwartet, in die Nähe eines globalen Minimums gelangt zu sein. Danach führt man eine lokale Optimierung durch, z.B. mit einem Quasi-Newton-Verfahren. Was allerdings die genauen Kriterien sind, die die Kombination der beiden Schritte bestimmen, kann nicht gesagt werden und muss der Erfahrung des Forschers überlassen bleiben.

Multistart-Verfahren

Ein weiteres, sehr naheliegendes Verfahren der globalen Optimierung besteht darin, die lokale Suche (etwa mit dem QN-Verfahren) mehrmals hintereinander von zufällig gewählten Punkten des Parameterraums aus zu starten, und ist unter dem Namen „Multistart“ bekannt. Unter den vereinfachenden Einschränkungen dass

(i) angenommen werden darf, dass die Funktion stetig ist und ein globales Optimum im beschränkten Suchraum $\mathcal{S} = \{\mathbf{w} : \mathbf{w}^u \leq \mathbf{w} \leq \mathbf{w}^o\}$ für vorgegebene untere bzw. obere Schranken $\mathbf{w}^u$ bzw. $\mathbf{w}^o$ hat,

(ii) die Startwerte aus einer Gleichverteilung über $\mathcal{S}$ gezogen werden und

(iii) die Aufgabe nicht darin besteht, das Optimum selbst, sondern bloß einen Punkt in einer ϵ-Umgebung des globalen Optimums mit klein gewählten ϵ zu finden,

hat das Multistart-Verfahren sehr günstige theoretische Eigenschaften. So konvergiert der Funktionswert des gefundenen besten lokalen Minimums mit wachsender Stichprobengröße N von Startwerten gegen denjenigen des globalen Optimums. Weiters kann man Bayesianische Schätzungen für die ingesamt vorhandene Anzahl an lokalen Minima angeben und daraus eine Bayesianische Stoppregel ableiten (siehe für einen Überblick Rinnooy Kan und Timmer, 1987a).

Die Nachteile des Multistart-Verfahrens liegen auf der Hand: es besteht die Gefahr, dass die gleichen lokalen Minima immer wieder gefunden werden. Um dieser Schwäche abzuhelfen und dadurch die Effizienz des Verfahrens zu steigern, gleichzeitig aber auch seine Verlässlichkeit, d.h. die Wahrscheinlichkeit bei ausreichend langer Suche tatsächlich das globale Optimum zu finden, zu erhalten, wurden verschiedene Verfeinerungen des Verfahrens entwickelt. Diese beruhen meist im Kern auf einem Clustering-Prinzip, das die Startwerte einem oder mehreren Clustern zuordnet, so dass jedes Cluster möglichst einem regionalen Attraktionsbecken entspricht. Von jedem Cluster wird nur eine lokale Suche gestartet, und optimaler Weise muss jedes lokale Minimum nur einmal gefunden werden. Unter den verschiedenen Verfahren hat sich in empirischen Vergleichen (z.B. Csendes, 1988) die Multilevel-Single-Linkage-Methode als besonders leistungsstark bewährt. Die Bemühungen, diese Verfahren weiter zu verbessern sind auch in den letzten Jahren fortgeführt worden (vgl. Marti et al., 2010).

Die in der vorliegenden Arbeit verwendete Version des Multistart-Verfahrens ist von den bei Rinnooy Kan und Timmer (1987a,b) vorgestellten Verfahren inspiriert, weicht aber unter anderem in zwei Aspekten von diesen ab. Erstens beruht die Auswahl der Startwerte für die lokalen Suchen nicht auf einer zufälligen Ziehung aus einer Gleichverteilung über dem Suchraum $\mathcal{S}$, sondern verwendet die speziell für die Initialisierung von ARNNDS-Modellen entwickelten Ansätze. Dies erscheint unter anderem opportun, da im Kontext der ARNN-Modellierung a-priori-Erwartungen für die Parameterwerte existieren, gleichzeitig aber auch sehr große Parameterwerte nicht prinzipiell ausgeschlossen werden können. Zweitens erfolgt die Zuordnung der Startwerte nicht wie bei den von Rinnooy Kan und Timmer (1987a,b) beschriebenen Verfahren iterativ und unter Berücksichtigung der bereits gefundenen lokalen Minima sowie, gegebenenfalls, der Funktionswerte, sondern stellt eine separate Phase dar und beruht auf einem nicht-hierarchischen Clusterverfahren.

Das erweiterte Multistart-Verfahren durchläuft mehrere Phasen:

Voroptimierung: Gegeben sei eine Menge von N initialen Parametervektoren $\mathbf{w}_0$, im folgenden einfach als Punkte bezeichnet. Von jedem dieser Punkte aus wird eine wenige Schritte umfassende lokale Suche mit dem BFGS-Verfahren durchgeführt. Die solcherart gefundenen voroptimierten Punkte seien mit $\mathbf{w}_1$

bezeichnet. Durch die Voroptimierung wird erreicht, dass ein Großteil des Suchraums, in dem gar keine oder keine annähernd optimalen Minima zu finden sind, von der weiteren Betrachtung ausgeschlossen werden kann und dass sich die $\mathbf{w}_1$ nun mehr oder weniger nahe um die vorhandenen lokalen Minima herum häufen.

Reduktion Die Menge der Punkte $\mathbf{w}_1$ wird nun anhand der ihnen zugeordneten Werte der Zielfunktion $E(\mathbf{w}_1)$ sortiert und auf die besten ξ_1 % reduziert, z.B. auf 70 %. Die verbleibenden Punkte seien mit $\mathbf{w}_2$ bezeichnet.

Clusteranalyse Auf die Menge der $\mathbf{w}_2$ wird nun eine nicht-hierarchische Clusteranalyse mit dem k-means-Verfahren angewendet, wobei für die Anzahl der Cluster ein Wert vorgegeben wird.

Konzentration Die Cluster werden anhand ihrer jeweiligen Medianwerte bezüglich $E(\mathbf{w}_2)$ sortiert und auch die Anzahl der Cluster reduziert, indem nur die besten ξ_2 % von Clustern weiter betrachtet werden. Von jedem der ausgewählten Cluster wird der gemessen an der Zielfunktion beste Punkt ermittelt. Die solcherart ausgewählten Clusterfavoriten seien mit $\mathbf{w}_3$ bezeichnet.

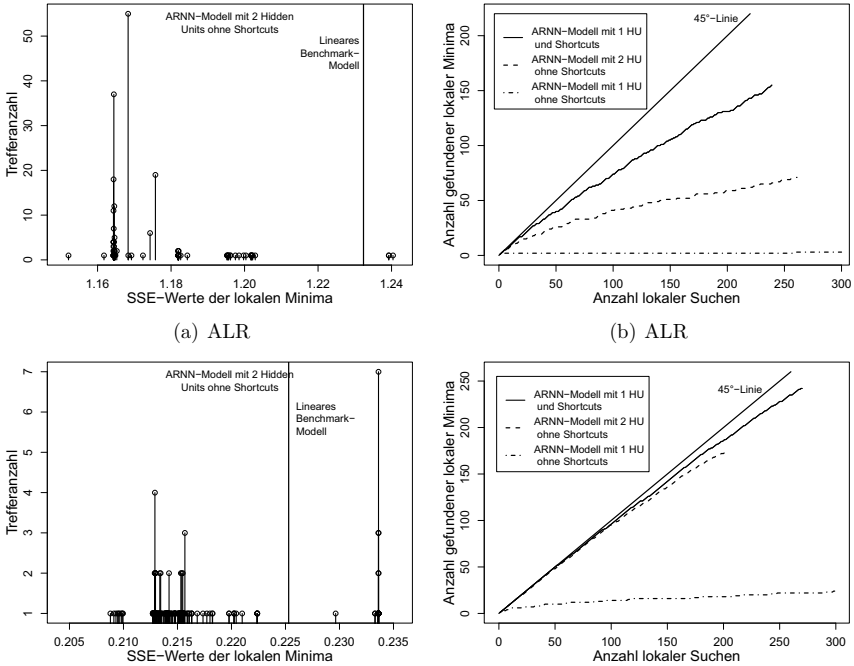
Optimierung Von jedem der Punkte $\mathbf{w}_3$ aus wird eine lokale Suche mit dem BFGS-Verfahren möglichst bis zur Konvergenz durchgeführt. Das beste der gefundenen Minima ist der Kandidat für das globale Minimum.

Das Verfahren ist als Funktion `multistart` im Rahmen des R-Pakets `NNUTS` implementiert. Es ist als Heuristik anzusehen, stellt jedoch gegenüber der üblichen Methode, ein ARNN-Modell wiederholt zu schätzen und bei einem „gut aussehenden“ Schätzergebnis zu verbleiben, ein systematisches Vorgehen dar und erlaubt auch deutliche Effizienzgewinne.

Exploration lokaler Minima in der ARNN-Modellierung auf den Beispielzeitreihen

Um die praktische Relevanz der Problematik des Auftretens mehrerer lokaler Minima zu illustrieren, soll im folgenden eine explorative Analyse der beiden Beispielzeitreihen ALR und IPI durchgeführt werden. Hierbei wird von der saisonal differenzierten Zeitreihe ausgegangen und anhand von drei unterschiedlichen, einfach gehaltenen ARNN-Modellen durch eine Multistart-Simulation die Varietät der lokalen Minima exploriert.

Für die Modellfindung im Rahmen dieser Simulation wurde für beide Zeitreihen jeweils vom AR-Modell (vgl. Tab. 3 und 4 in Abschnitt 2.4) ausgegangen. Die Lags in diesem Modell wurden dann als Lags in einem ARNN-Modell mit einer Hidden Unit und ohne Shortcuts verwendet. Danach wurden schrittweise insignifikante Gewichte eliminiert und derart die Modellstruktur vereinfacht (vgl. Abschnitt 4.2).



(a) ALR (b) ALR

(c) IPI (d) IPI

Abb. 10: Exploration der lokalen Minima: (a) und (b) Arbeitslosenrate (ALR) sowie (c) und (d) Industrieproduktionsindex (IPI)

Auf diese Weise umfasst das auf ALR geschätzte Modell die Lags 1, 10, 12, 13, 16, 24 und 25 sowie ein Interzept, jenes auf IPI die Lags 1, 2, 3, 12, 14, 15, 24, 26, 48, 49 (ohne Interzept). Neben diesen sehr einfach gehaltenen Modellen wurden jeweils auch Modelle mit zwei verdeckten Neuronen bzw. mit Direktverbingungen (Shortcuts) betrachtet. Es wurden jeweils 300 ARNN-Modelle unter Verwendung des BFGS-Verfahrens geschätzt. Die bis zu einer vorgegebenen Anzahl von Iterationen nicht konvergierten Netze wurden von der weiteren Betrachtung ausgeschieden. Für die verbleibenden geschätzten Modelle wurde untersucht, ob das gefundene lokale Minimum bereits bekannt oder neu ist. Hierzu wurde einerseits die Fehlerfunktion und andererseits der Gewichtsvektor mit jenen der zuvor geschätzten Modelle verglichen. Abbildung 10 fasst die Ergebnisse zusammen.

Für das einfachste der betrachteten Modelle (eine Hidden Unit, keine Shortcut) tritt das Problem mehrerer lokaler Minima kaum auf. Im Falle der ALR gibt es nur drei lokale Minima, wobei eines von diesen bei weitem häufiger getroffen wird

als die anderen. Die drei Lösungen sind im Gewichtsraum sehr nah bei einander. Im Falle des IPI sind zwar 24 verschiedene lokale Minima anzutreffen, dennoch ist die Situation ähnlich, da diese sowohl hinsichtlich der Fehlerfunktion als auch hinsichtlich des Gewichtsraums nahe bei einander liegen.

Anders präsentiert sich die Situation im Falle eines neuronalen Netzes mit zwei verdeckten Neuronen. Im Falle der ALR (IPI) fanden die 239 (203) konvergierten neuronalen Netze 155 (174) lokale Minima, die sich sowohl hinsichtlich des Wertes der Fehlerfunktion als auch hinsichtlich des Gewichtsraums von einander unterschieden, wenn auch eine Häufung von mehreren nah bei einander liegenden lokalen Minima festgestellt werden kann. Für das Modell mit einem vereckten Neuron und mit Direktverbindungen ist die Vielfalt lokaler Minima noch größer und unüberschaubarer.

Auf der rechten Seite in Abb. 10 befinden sich Graphiken zur Entwicklung der Anzahl der gefundenen lokalen Minima im Verlauf der Multistart-Simulation. Wenn die Anzahl der lokalen Minima beschränkt ist, so sollten die Kurvenverläufe von einer zunehmenden Abwendung von der 45°-Kurve gekennzeichnet sein und schließlich zum wahren Wert der Anzahl der lokalen Minima, M, konvergieren, der nach einer von Boender (1984) angegebenen Formel (siehe auch Rinnooy Kan und Timmer, 1987a) abschätzbar ist:

$$\mathrm{E}(M) = \frac{m(N-1)}{N-m-2}. \tag{4.42}$$

In dieser Formel ist N die Anzahl der durchgeführten lokalen Suchen, also der in Abb. 10 (rechte Seite) auf der X-Achse abgetragene Wert, und m ist die Anzahl der nach N lokalen Suchen gefundenen verschiedenen lokalen Minima, also der auf der Y-Achse abgetragene Wert. Eine Konvergenz der Kurvenverläufe ist jedoch nicht erkennbar und die Anwendung der Formel ergibt denn auch sehr hohe Werte für die erwartete Anzahl der lokalen Minima (für das Modell mit einem verdecktem Neuron und Direktverbindungen sind es mehr als 1300 bzw. mehr als 380).

Die Analyse zeigt auf, dass die Fehlerlandschaft sogar für einfache ARNN-Modelle vom Vorhandensein zahlreicher lokaler Minima geprägt ist, von denen der bei weitem größte Anteil nur relativ kleine Attraktionsregionen aufweist. Nach diesem Befund müssen die Erwartungen, für neuronale Netze, außer in den einfachsten Modellspezifikationen, das globale Minimum zu finden, relativiert werden. Der Verwendung von ausgefeilten Multistart-Verfahren und anderen globalen Optimierungsverfahren ist bei mehr als 100 lokalen Minima und einem hochdimensionalen Suchraum wenig Erfolg beschieden.

4.6 Statistisch-parametrischer Ansatz

NN wurden nach ihrer Wiederentdeckung in den 80er Jahren bald auch von Statistikern und Ökonometrikern aufgegriffen, die in den entwickelten Innovationen aus

dem eigenen Fachgebiet bekannte Prinzipien und Verfahren erkannten. Während anfangs die nicht-parametrische Sichtweise der NN vorherrschte, stellten wichtige Arbeiten die parametrische Sichtweise in den Vordergrund, bei welcher das NN nicht als „Black Box“ betrachtet wird, sondern als Modell, dessen einzelne Parameter separat einer Untersuchung zugänglich sind. Wie für andere nicht-lineare zeitreihenanalytische Modelle auch können für die einzelnen Parameter des NN-Modells Standardfehler berechnet und Hypothesentests ausgeführt werden. Die zentrale Voraussetzung für die Anwendung dieses Instrumentariums bildet die Asymptotik und ihre Annahmen.

Diese Sichtweise bestimmt auch, wie bei der Modellierung insgesamt vorzugehen ist. Es kann klar zwischen den Schritten der Modellspezifikation, Modellschätzung unterschieden werden, und auch für den dritten Schritt, die Modelldiagnose, sind die üblichen Instrumente verfügbar. Die Schätzung des Modells besteht in der Suche des optimalen Parametervektors. Die Modellspezifikation zielt darauf ab, die in Hinblick auf die Generalisierungsfähigkeit optimale Modellstruktur zu finden. Die Modellkomplexität wird im wesentlichen anhand der Anzahl der Parameter gemessen. Für den Vergleich von unterschiedlichen Modellstrukturen stehen verschiedene Zugänge zur Verfügung (Hypothesentests, Modellselektionskriterien und Kreuzvalidierung), die jeweils voraussetzen, dass die verschiedenen in Frage kommenden Modelle in zulässiger Weise geschätzt werden können.

In der breiter gefassten NN-Forschung wurde die Frage der Bestimmung der optimalen Modellstruktur bzw. der Netzwerkarchitektur immer schon als besonders wichtig angesehen. Sie umfasst u.a. die Wahl der Inputs, die Anzahl der verdeckten Neuronen, die Frage nach der Aufnahme von direkten Verbindungen zwischen Input- und Outputneuronen sowie eventuell den Ausschluss einzelner Verbindungen zwischen Input- und verdeckten Neuronen. Zwei grundsätzlich verschiedene Zugänge sind hierbei unter den Bezeichnungen „Growing“ und „Pruning“ bekannt. Growing bedeutet, dass man die Modellstruktur schrittweise anwachsen lässt, solange hinzugefügte Elemente zusätzlichen Erklärungsgewinn liefern. Pruning bedeutet, dass aus einem bestehenden NN jene Elemente herausgeschnitten werden, die in ungenügender Weise zum Erklärungsziel beitragen. Die Elemente können prinzipiell einzelne verdeckte Neuronen oder einzelne Gewichte.

Es wurden verschiedene Techniken und Heuristiken für das Growing und Pruning entwickelt, von denen im Folgenden nur jene behandelt werden, die der statistischen Sichtweise zuzuordnen sind. Hier wird die Frage, ob der Netzwerkarchitektur ein Element hinzugefügt bzw. weggeschnitten wird, anhand eines statistischen Kriteriums, vorzugsweise eines Hypothesentests, entschieden. Growing und Pruning sind hierbei nicht zwei alternative oder austauschbare Vorgangsweisen, sondern ergänzen einander und können in einer kombinierten Growing-Pruning-Strategie aufeinander folgen, wobei sich das Growing auf das Hinzufügen von verdeckten Neuronen bezieht, während das Pruning die einzelnen Verbindungen zwischen den Neuronen zum Gegenstand hat. Bei diesem Modellierungsansatz wird zuerst im Rahmen des Growing

die Anzahl der verdeckten Neuronen des Modells bestimmt und hiernach das Modell im Rahmen des Pruning von überflüssigen Verbindungen befreit. Die Notwendigkeit bei der Bestimmung der Anzahl der verdeckten Neuronen in anwachsender Weise vorzugehen, ergibt sich aus dem unten noch zu besprechenden Identifikationsproblem bei NN-Modellen. Umgekehrt gelangt man für die Entscheidung über einzelne Verbindungen in natürlicher Weise zu einer Pruning-Vorgangsweise, da nach dem Abschluss des Growing die verdeckten Neuronen üblicherweise in vollzähliger Weise mit den vorhandenen Inputneuronen verbunden sind.

Die hier skizzierte Vorgangsweise der Abfolge von Growing und Pruning wurde unter anderem von Anders und Korn (1999) und Medeiros et al. (2006) beschrieben und mit Anwendungsstudien untersucht. In ähnlicher Weise wie in diesen beiden Arbeiten werden in diesem Abschnitt die theoretischen Grundlagen des statistisch-parametrischen Ansatzes dargelegt und seine Funktionsweise anhand von Anwendungen auf einer simulierten Zeitreihe und auf den beiden Beispielzeitreihen illustriert. Es werden hierbei verschiedene Aspekte von ARNN-Modellen und deren Anwendung auf makroökonomische Zeitreihen besonders herausgearbeitet und verschiedene Erweiterungen des Instrumentariums vorgestellt.

Asymptotische Eigenschaften und statistische Inferenz

In diesem Unterabschnitt geht es um die asymptotischen Eigenschaften der Parameter eines mit der Kleinstquadratemethode (KQ) geschätzten ARNN-Modells. Da es sich um ein nicht-lineares Modell handelt, wird auch von der nicht-linearen Kleinstquadratemethode (NKQ, bzw. engl. Nonlinear least squares, NLS) gesprochen. Bereits in Abschnitt 4.1 wurde bei der Besprechung der Fehlerfunktion darauf hingewiesen, dass mit der zusätzlichen Annahme eines unabhängig, identisch und normalverteilten Fehlerprozesses diese Methode mit Maximum-Likelihood äquivalent ist und daher auch der Ausdruck Quasi-Maximum-Likelihood-Schätzung verwendet wird. Die im folgenden zu besprechenden asymptotischen Eigenschaften der geschätzten Gewichte besagen, dass bei Identifiziertheit des Modells und bei wachsender Stichprobengröße n erstens der mit NKQ geschätzte Parametervektor immer besser mit dem „wahren“ Parametervektor übereinstimmt und zweitens, dass die Verteilung der Schätzungen asymptotisch einer Normalverteilung folgt. Diese beiden asymptotischen Eigenschaften sind die wesentlichen Voraussetzungen für die Ableitung von Hypothesentests bezüglich einzelner oder Gruppen von Parametern und damit auch für den statistisch-parametrischen Ansatz insgesamt.

Für die theoretische Basis der Asymptotik in nicht-linearen Modellen sei auf Lehrbücher wie Gallant und White (1988), Davidson und MacKinnon (1993) und White (1994) verwiesen. Zur Herleitung der asymptotischen Eigenschaften der Parameter in einem NN-Modell stammen wesentliche Arbeiten von White (1989c,d,b) und Kuan und White (1994). Diese Ergebnisse wurden von Trapletti et al. (2000) für die Erfordernisse von ARNN-Modellen erweitert.

Es sei $\mathbf{w}$ der Vektor der p Parameter eines ARNN-Modells. Wenn $\mathbf{w}$ existiert, kann es durch NKQ konsistent geschätzt werden, d.h. dass die Schätzung $\hat{\mathbf{w}}$ bei zunehmender Stichprobengröße n immer näher bei $\mathbf{w}$ zu liegen kommt:

$$\hat{\mathbf{w}} \xrightarrow{\text{f. s.}} \mathbf{w}, \quad n \longrightarrow \infty, \tag{4.43}$$

wobei $\xrightarrow{\text{f. s.}}$ *fast sichere Konvergenz* bezeichnet.

In diesem Zugang spiegelt sich die Annahme des „wahren“ Modells wider: es wird angenommen, dass die Daten durch ein ARNN-Modell mit der entsprechenden Form generiert wurden. Für den Fall, dass diese Annahme nicht zutrifft, das Modell also fehlspezifiziert ist, müssten die im Folgenden präsentierten asymptotischen Ergebnisse generalisiert werden. Hierfür ist es erforderlich, dass statt dessen die Existenz eines optimalen $\mathbf{w}^*$ angenommen werden kann, das für das gegebene ARNN-Modell die beste Approximation an das wahre Modell liefert. Dem Problem der Existenz von $\mathbf{w}$ bzw. $\mathbf{w}^*$, das auch als Identifikationsproblem bezeichnet wird, wird weiter unten ein eigener Unterabschnitt gewidmet, da im Kontext von ARNN-Modellen hier einige Besonderheiten auftreten.

Neben der Konsistenz ist die zweite wichtige asymptotische Eigenschaft die asymptotische Normalität von $\hat{\mathbf{w}}$. Unter ziemlich allgemeinen Bedingungen gilt, dass die geschätzten Parameter asymptotisch normalverteilt sind mit Mittelwert $\mathbf{w}$:

$$\sqrt{n}(\hat{\mathbf{w}} - \mathbf{w}) \Rightarrow \mathrm{N}(\mathbf{0}, \mathbf{C}), \quad n \longrightarrow \infty, \tag{4.44}$$

wobei $\Rightarrow$ *Konvergenz in Verteilung* bezeichnet.

Für die Kovarianzmatrix gibt es mit

$$\hat{\mathbf{C}} = \frac{1}{2\hat{\sigma}_\varepsilon^2}\mathbf{H} \tag{4.45}$$

einen asymptotisch konsistenten Schätzer, wobei man die in dieser Formel auftauchenden Größen aus der Schätzung auf der Stichprobenzeitreihe erhält:

$$\hat{\sigma}_\varepsilon^2 = n^{-1}\sum_{t=1}^{n}\hat{\varepsilon}_t^2 = n^{-1}\sum_{t=1}^{n}(y_t - \hat{y}_t(\mathbf{w}, y_{t-1}, y_{t-2}, \ldots))^2, \tag{4.46}$$

$$\mathbf{H} = \frac{\partial^2 Q_n}{\partial \mathbf{w} \partial \mathbf{w}'}. \tag{4.47}$$

Man beachte, dass sich die Hesse-Matrix $\mathbf{H}$, die Matrix der zweiten Ableitungen der Fehlerfunktion nach den Parametern des Modells, hier nicht auf die übliche Fehlerfunktion (*SSE*) sondern auf $Q_n = n^{-1}SSE$ bezieht.

Diese Ergebnisse bilden die Basis für Hypothesentests für einzelne Parameter und Gruppen von Parametern. Im einfachsten Falle kann etwa ein t-Test verwendet

werden, um zu testen, ob ein bestimmter Parameter w_i von einem vorgegebenen Wert w_i^*, meist $w_i^* = 0$, verschieden ist. Die Teststatistik lautet

$$\frac{\hat{w}_i - w_i^*}{\hat{\sigma}_{\hat{w}_i}}, \tag{4.48}$$

wobei $\hat{\sigma}_{\hat{w}_i}$ die Wurzel des i-ten Elements der Hauptdiagonale von $\frac{1}{n}\hat{\mathbf{C}}$ ist. Es sei erwähnt, dass die Annahme der t-Verteilung dieser Teststatistik nur im Falle der richtigen Spezifizierung des Modells ableitbar ist, Davidson und MacKinnon (1993) sprechen im Falle einer Fehlspezifikation von „Pseudo-t-Werten“.

Um eine ganze Gruppe von Parametern zu testen, wird der Wald-Test verwendet. Hier wird die Null-Hypothese $H_0 : \mathbf{Rw} = \mathbf{r}$ gegen die Alternative $H_1 : \mathbf{Rw} \neq \mathbf{r}$ gestestet, wobei $\mathbf{R}$, eine $r \times p$ Matrix vollen Rangs, und der r-Vektor $\mathbf{r}$ zusammen die r linearen Restriktionen angeben. Die Wald-Statistik lautet

$$n(\mathbf{R}\hat{\mathbf{w}} - \mathbf{r})'(\mathbf{R}\hat{\mathbf{C}}\mathbf{R}')^{-1}(\mathbf{R}\hat{\mathbf{w}} - \mathbf{r}) \Rightarrow \chi_r^2 \tag{4.49}$$

Die beschriebenen Tests erlauben das Entfernen eines Parameters aus einem Modell, wenn dieser nicht signifikant von Null verschieden ist („Pruning“). Diese Vorgangsweise hat jedoch ihre Grenze, wenn die Signifikanz eines verdeckten Neurons insgesamt getestet werden soll, da die Annahme der Identifiziertheit des Modells bereits einschließt, dass im ARNN-Modell alle Parameter $c_k \neq 0$ sind. Führt man in ein gegebenes Modell ein zusätzliches Neuron ein, so ist hingegen nicht sichergestellt, dass mit dem zusätzlichen Neuron das Modell identifiziert ist. Dies führt zur Notwendigkeit einer besonderen Vorgangsweise, für die Modellierung mit ARNN-Modellen, wenn von einem einfachen Modell schrittweise zu einem allgemeineren Modell ausgebaut werden soll und die weiter unten als „Growing“ näher beschrieben werden soll. Zuvor muss allerdings das Identifikationsproblem genauer erläutert werden.

Abschließend sei die Aufmerksamkeit auf einen wichtigen und potentiell unterschätzten Aspekt der asymptotischen Theorie gelenkt, der Annahme, dass es einen Optimierungsalgorithmus gibt, der immer das globale Minimum findet. Für die Entwicklung der asymptotischen Theorie erscheint diese Annahme unproblematisch, da es sich zunächst bloß um eine rein theoretische Annahme handelt. Doch bei der Anwendung der aus der asymptotischen Theorie gewonnenen Verteilungen von $\hat{\mathbf{w}}$ auf den Fall endlicher Stichproben ist diese Annahme höchst problematisch. White (1989c) spricht in diesem Zusammenhang davon, dass in endlichen Stichproben die Schätzung $\hat{\mathbf{w}}$ in zweifacher Weise vom wahren Wert $\mathbf{w}$ bzw. optimalen Wert $\mathbf{w}^*$ entfernt ist, einmal durch die Zufälligkeit der Stichprobe andererseits durch das Optimierungsproblem auf Basis der jeweiligen Stichprobe. Die asymptotische Theorie hat zu diesem Problem kaum konkrete Empfehlungen zu machen, außer das vage Erfordernis, dass die Stichprobengröße ausreichend groß sein sollte. White (1989c) diskutiert verschiedene Ad-hoc-Regeln, was dies in der Praxis bedeuten könnte, und

verweist auf Daumenregeln, die besagen, dass der Stichprobenumfang mindestens $p \times 10$ sein sollte. Bedenkt man, dass die ARNN-Modelle, die im Rahmen der Anwendungsbeispiele dieser Arbeit auf den beiden Zeitreihen ALR und IPI mit dem statistisch-parametrischen Ansatz geschätzt werden, bis zu 25 Parameter aufweisen, so wäre die Stichprobengröße ($n = 456$) nur knapp ausreichend groß.

Bereits in den Abschnitten zur lokalen und globalen Optimierung wurde gezeigt, dass sowohl die lokale als auch die globale Optimierungsaufgabe im Rahmen der Schätzung eines ARNN-Modells auf makroökonomischen Zeitreihen anspruchsvolle Probleme darstellen. Das Auftreten von mehreren lokalen Minima in der Fehlerlandschaft, wie im Anwendungsbeispiel von Abschnitt 4.5 belegt, kann unter anderem auf ein Identifikationsproblem oder auf eine zu geringe Stichprobengröße hindeuten. Beides sind Probleme, die der statistischen Inferenz auf der Basis der asymptotischen Theorie die Berechtigung entziehen.

Identifikationsproblem

Eine wichtige Voraussetzung der Verwendung des statistisch-parametrischen Ansatzes zur ARNN-Modellierung ist die Identifiziertheit des Modells. Ein Modell ist identifiziert, wenn der wahre Parametervektor eindeutig bestimmt ist, d.h. wenn es keinen anderen Parametervektor gibt, mit dem die gleiche Funktion implementiert wird. Identifiziertheit ist eine Voraussetzung dafür, dass ein Modell konsistent geschätzt werden kann, d.h. dass die Schätzung bei zunehmender Stichprobengröße immer besser mit dem wahren Wert übereinstimmt und zum wahren Wert konvergiert. Die Problematik der Identifiziertheit ist zwar prinzipiell für alle linearen und nicht-linearen Modelle als Voraussetzung für statistische Inferenz von Bedeutung, erfordert aber bei NN- bzw. ARNN-Modellen besondere Aufmerksamkeit. Auf der Basis der bereits in den 80er Jahren weit entwickelten Theorie der statistischen Inferenz nicht-linearer Modelle konnte White (1989c,d,b) zeigen, dass aus der Identifiziertheit von NN-Modellen ihre Konsistenz folgt. Für ARNN-Modelle wurde von Trapletti et al. (2000) der entsprechende Beweis gebracht, wobei neben der Identifiziertheit auch die Stationarität des Modells gegeben sein muss.

Bei NN-Modellen sind zwei verschiedene Ursachen von Nicht-Identifiziertheit zu unterscheiden, erstens Gewichtssymmetrien und zweitens redundante Gewichte. Gewichtssymmetrie ist ein bekanntes und relativ leicht in den Griff zu bekommendes Problem von neuronalen Netzen. Sie beruht einerseits darauf, dass (bei Verwendung des Tangens Hyperbolicus als Aktivierungsfunktion) die Vorzeichen der zu und von einem bestimmten verdeckten Neuron führenden Gewichte verändert werden können, ohne dass die implementierte Funktion sich ändert. Im Falle eines ARNN-Modells mit einem verdeckten Neuron implementieren also die Parametervektoren $\mathbf{w}' = (a_{11}, a_{12}, \ldots, a_{1J}, b_1, c_1, d)$ und $\tilde{\mathbf{w}}' = (-a_{11}, -a_{12}, \ldots, -a_{1J}, -b_1, -c_1, d)$ das gleiche Modell. In ähnlicher Weise kann die Reihenfolge von verdeckten Neuronen verändert werden ohne eine Änderung der implementierten Netzwerkfunktion

zu bewirken. Insgesamt gibt es zu jedem Netz mit K verdeckten Neuronen $2^K K!$ äquivalente Netze. Um angesichts von Gewichtssymmetrien die Identifiziertheit des Modells sicherzustellen, betrachtet man nur einen Teil des Parameterraums, indem im Rahmen der Schätzung z.B. auf den Bereich $c_1 \geq c_2 \geq \ldots > c_K \geq 0$ eingeschränkt wird. Dieser eingeschränkte Parameterraum, innerhalb dessen die Identifikation nicht von Gewichtssymmetrie beeinträchtigt ist, heißt Hecht-Nielsen-Kegel (Hecht-Nielsen, 1989).

Redundante Gewichte als zweite Ursache von Nicht-Identifiziertheit bereiten für die Anwendung des Instrumentariums der Statistischen Inferenz hingegen ein beträchtliches Problem. Das Problem tritt auf, wenn ein neuronales Netz, d.h. jenes „wahre“ Modell, von dem die Daten generiert werden, ein verdecktes Neuron aufweist, für das entweder das wegführende Gewicht Null ist oder für das gleichzeitig alle hinführenden Gewichte Null sind, $c_k = 0$ oder $a_{kj} = 0$ für alle j. In beiden Fällen kann das betreffende Neuron entfernt werden, ohne dass die Netzwerkfunktion beeinflusst wird. Bei der Präsenz eines solcherart überflüssigen verdeckten Neurons kann die Funktionalität von zumindest einem der restlichen verdeckten Neuronen in beliebiger Weise zwischen diesem und dem überflüssigen verdeckten Neuron aufgeteilt werden, was den Kern der Nicht-Identifiziertheit ausmacht. Auch die Funktionalität eines allfällig vorhandenen linearen Teils („Shortcuts“) kann zu einem beliebigen Anteil vom überflüssigen verdeckten Neuron übernommen werden. Eine Konsequenz ist, dass das Modell nicht konsistent geschätzt werden kann und die asymptotische Theorie nicht angewendet werden kann.

Im Rahmen der schrittweisen Bestimmung der Modellstruktur muss die Identifikation des Modells getestet werden, bevor es um ein zusätzliches verdecktes Neuron erweitert wird. Dieses Verfahren wird weiter unten als „Growing-Strategie“ genauer erläutert. Andererseits soll die Frage der Identifikation auch in der Phase der Modelldiagnose nicht vergessen werden. Sind die Standardfehler eines geschätzten Modells mit dem Instrumentarium der asymptotischen Theorie nicht zu ermitteln, da die Hesse-Matrix singulär oder nicht positiv-definit ist, so kann dies an numerischen Problemen mit der konkreten Schätzung hängen oder auf ein generelles Problem der Nicht-Identifikation hinweisen. Aber auch wenn die Standardfehler der Koeffizienten eines Modells nach den Regeln der asymptotischen Theorie bestimmbar sind, heißt dies noch nicht, dass das Modell identifiziert ist, sondern lässt zunächst nur den Schluss zu, dass die Hesse-Matrix positiv definit ist bzw. dass ein eindeutiges lokales Minimum gefunden wurde.

Gibt es bei der Schätzung eines ARNN-Modells Hinweise auf zahlreiche lokale und nicht-trivial verschiedene Minima der Fehlerfunktion, so lässt dies nicht umgekehrt den Schluss auf Nicht-Identifiziertheit des Modells zu, aber es ist wahrscheinlich ein Teil des Problems. Die intuitive Überlegung ist hierbei, dass bei einem nicht-identifizierten Modell bei kleinen Stichproben im Vergleich zu großen Stichproben die Fehleroberfläche durch den relativ starken Einfluss des Fehlerprozesses aufgerauht wird, und daher in jenen Bereichen der Fehlerlandschaft, die bei großer

Stichprobengröße flach ausfallen würden, lokale Minima leicht entstehen können. Allgemein ist bekannt, dass die Schätzung von nicht-identifizierten Modellen zu numerischen Problemen führt.

Modellselektionskriterien und Kreuzvalidierung

Bevor auf die beiden Techniken des sogenannten „Growing“ und „Pruning“ vertieft eingegangen wird, sollen noch zwei wichtige alternative Möglichkeiten der Modellselektion besprochen werden. Es sind dies Modellselektionskriterien und Kreuzvalidierung. Beide Zugänge sind im Kernbereich einer statistisch-parametrischen Sichtweise anzusiedeln und beruhen u.a. auf den Überlegungen der Asymptotik, obwohl ihre Verwendung in der Praxis oft auch losgelöst von der statistischen Theorie vorkommt.

Modellselektionskriterien, auch als Informationskriterien bezeichnet, unterstützen die richtige Wahl der geeigneten Modellkomplexität, indem sie den Fehler des Modells gegen die Anzahl der verwendeten Parameter abwägen. Sie bestehen dementsprechend aus einem Fehlerterm und einem Term, der die Modellkomplexität in Abhängigkeit von der Anzahl der Parameter abbildet. Bei der Entscheidung zwischen zwei konkurrierenden Modellen ist dasjenige zu wählen, das das Informationskriterium minimiert. Je nach theoretischer Ableitung und Zielsetzung stehen eine Reihe von Modellselektionskriterien zur Verfügung. Die bekanntesten sind das Akaike (1973) Informationskriterium (AIC) und das Schwarz (1978) Informationskriterium (SIC, auch als Bayes'sches Informationskriterium, BIC bekannt):

$$\mathrm{AIC} = -\frac{2}{n} \ln L(\hat{\mathbf{w}}) + \frac{2p}{n} \tag{4.50}$$

$$\mathrm{SIC} = -\frac{2}{n} \ln L(\hat{\mathbf{w}}) + \frac{p \ln n}{n}, \tag{4.51}$$

wobei $L(\hat{\mathbf{w}})$ die geschätzte Likelihood, p die Anzahl der Parameter und n die Anzahl der verwendeten Beobachtungen ist. Die beiden Kriterien sind ähnlich, doch bestraft das SIC (ab $n \geq 8$ Beobachtungen) die Modellkomplexität gemessen an der Anzahl der Parameter stärker.

Bei Verwendung von Quasi-Maximum-Likelihood bzw. nicht-linearem Kleinst-Quadrate-Ansatz wird das AIC auf der Basis der geschätzten Varianz angeschrieben. Eine mögliche Definition des AIC lautet dann

$$\mathrm{AIC} = n \ln(\hat{\sigma}^2) + 2p \tag{4.52}$$

und dies ist auch die Definition, die im Rahmen des R-Pakets `NNUTS` implementiert wurde. In verschiedenen Softwareimplementationen gibt es unterschiedliche Definitionen des AIC, was weiter kein Problem darstellt, solange diese lineare (oder zu-

mindest stetige) Transformationen einer theoretisch richtigen AIC-Definition sind.[27] Wesentlich ist, dass die Modellselektion auf einer einheitlichen Definition des AIC oder eines anderen Informationskriteriums beruht und dass die Modelle hierarchisch zu einander stehen, d.h. dass eines der beiden Modelle das andere umfasst.

Schließlich sei auch auf das von Murata et al. (1994) speziell für NN-Modelle entwickelte Network Infomation Criterion (NIC) hingewiesen. Dieses kann als Generalisierung des AIC für fehlspezifizierte Modelle betrachtet werden. Unter der Annahme, dass ein ARNN-Modell richtig spezifiziert ist, fällt es asymptotisch mit dem AIC zusammen.

Ein weiteres Prinzip, das zur Modellselektion herangezogen werden kann, ist Kreuzvalidierung („Cross-Validation“, CV) (Stone, 1974). Hierbei wird versucht, mit Hilfe von beiseitegelegten Daten die zu erwartende quadrierte Abweichung bezüglich neuer Daten zu prognostizieren. Im Unterschied zu einfacher Validierung bzw. sogenannten Split-Sample-Methoden, wie sie etwa bei Early-Stopping (vgl. Abschnitt 4.7) zur Anwendung kommen, wird dies nicht bloß einmal gemacht, sondern mehrmals mit jeweils unterschiedlicher Aufteilung der Stichprobe in Schätzset und Validierungsset. Im gründlichsten Falle, der sogenannten „Leave-One-Out“-Methode, ist der Umfang des Validierungsset jeweils nur eins. Bezeichnet man den auf dem beiseitegelegten Validierungsset ermittelten mittleren quadrierten Fehler mit $MSFE_m$ so wird der Kreuzvalidierungsfehler als

$$MSFE_{CV} = \frac{1}{M} \sum_{m=1}^{M} MSFE_m, \tag{4.53}$$

wobei M die Anzahl der verschiedenen Validierungssets ist, in die die gesamte Stichprobe aufgeteilt wird.

Wenn Kreuzvalidierung zur Modellselektion verwendet wird, wird unter konkurrierenden Modellen jenes ausgewählt, welches den Kreuzvalidierungsfehler minimiert. Für Kreuzvalidierung liegen verschiedene theoretische und simulative Studien vor, die Zusammenhänge mit Modellselektionskriterien und Hypothesentests aufzeigen. So zeigt Stone (1977) unter bestimmten Bedingungen die asymptotische Äquivalenz der Modellselektion mit AIC und Kreuzvalidierung. Rivals und Personnaz (1999) vergleichen in einem einfachen simulativen Setting Kreuzvalidierung mit Modellselektion auf Basis von AIC. Sie kommen zum Schluss, dass Hypothesentest vorzuziehen sind, da sie die Stichprobe effizienter ausnützen und das Risiko für den Fehler erster Art (die Größe des Tests) im Unterschied zu Kreuzvalidierung angeben.

Zu diesen theoretischen Nachteilen kommt der enorme mit Kreuzvalidierung verbundene Rechenaufwand. Für die Modellierung von Zeitreihen ist zudem die Auftei-

[27] Dennoch sei der Anwender gewarnt, dass zusätzlich zu harmlosen Unterschieden in der Definition des AIC, für die leicht korrigiert werden kann, auch unterschiedliche Berechnungsweisen der einzelnen Terme den Vergleich von AIC-Werten, die für das gleiche, aber mit unterschiedlicher Software geschätzte Modell ermittelt werden, erschweren können.

lung in Schätzset und Validierungsset mit zusätzlichen Schwierigkeiten verbunden, die diese Methode für die vorliegende Arbeit nicht empfehlen.

Aus den erwähnten theoretischen und praktischen Gründen werden Modellselektionskriterien und Kreuzvalidierung für die vorliegende Arbeit nicht für die Entwicklung von Modellierungsstrategien berücksichtigt. Es sei auf Anders (1997) und Anders und Korn (1999) verwiesen, die die Praktikabilität von Modellierungsstrategie auf Basis dieser beiden Zugänge illustrieren und zu Ergebnissen gelangen, die jenen mit dem hypothesentestbasierten Ansatz vergleichbar sind.

Growing

Die mit „Growing“ bezeichnete Technik basiert im Rahmen der vorliegenden Arbeit auf dem TLG-Testverfahren, mit dem für ein bestehendes NN getestet wird, ob vernachlässigte nicht-lineare Struktur vorliegt, die durch eine zusätzliches verdecktes Neuron modelliert werden kann. Nur wenn die Null-Hypothese, dass keine solche vernachlässigte Nicht-Linearität besteht, abgelehnt werden kann, darf man ein verdecktes Neuron hinzufügen. Der Test kann nun mehrmals hintereinander geschaltet werden, um ein Modell schrittweise jeweils um ein verdecktes Neuron zu erweitern, solange bis ein zusätzliches Neuron nicht mehr signifikant zur Erklärung der Daten beiträgt.

Im Original-Papier von Teräsvirta et al. (1993) wird der Test als reiner Linearitäts-Test beschrieben, d.h. die Null-Hypothese ist, dass die Zeitreihe nur lineare Struktur aufweist. Die Testprozedur wurde bereits in Abschnitt 3.2 dargestellt und muss daher in diesem Kapitel nicht nochmal zur Gänze wiedergegeben werden. Für die Verwendung des TLG-Tests als Test auf die Signifikanz eines zusätzlichen verdeckten Neurons, muss das Verfahren um einige Aspekte und spezielle Notation erweitert werden, die bereits in White (1989a) verwendet wurden. Eine solche Beschreibung des Testverfahrens findet sich unter anderem bei Anders (1997), an dem sich die folgende Darstellung teilweise orientiert. Zudem wird, wie bereits in Abschnitt 3.2, auf einige Modifikationen des Testverfahrens hingewiesen, die der Einsatz auf makroökonomische Zeitreihen wie die beiden Beispielzeitreihen ALR und IPI nahelegt. Die verwendete Notation ist hierbei in Übereinstimmung mit jener der Definition von ARNN- bzw. ARNNDS-Modellen (vgl. Abschnitt 4.1).

Will man in einem ARNN-Modell gemäß (4.1) bzw. einem ARNNDS-Modell gemäß (4.2) ein bestimmtes verdecktes Neuron auf Signifikanz testen, so bietet sich die Null-Hypothese

$$H_0 : c_k = 0 \tag{4.54}$$

an. Wenn diese zutrifft, sind jedoch die zu diesem Neuron führenden Inputverbindungen a_{kj}, $j = 1, \ldots, J$ nicht bestimmt. Daher können in diesem Fall die Annahmen der Asymptotik nicht angewandt werden und die übliche Teststrategie versagt. Der von Teräsvirta et al. (1993) vorgeschlagene Lösungsweg besteht darin, dass das

zusätzliche Neuron durch eine Taylorreihenentwicklung dritter Ordnung

$$\psi(b_k + \sum_{j=1}^{J} a_{kj} x_j) \approx \tilde{\psi}(b_k + \sum_{j=1}^{J} a_{kj} x_j) \tag{4.55}$$

ersetzt wird, vgl. die Formeln (3.27) und (3.28) in Abschnitt 3.2. Ist ein Modell mit $K > 0$ verdeckten Neuronen identifiziert, so ist es auch identifiziert, wenn die sich aus der Taylorreihenentwicklung eines zusätzlichen Neurons ergebenden Variablen als Regressoren hinzugefügt werden. Als neue Null-Hypothese kann somit

$$\tilde{H}_0 : \delta_{ij} = 0, \delta_{ijk} = 0 \qquad i = 1, \ldots, q; j = i, \ldots, q; k = j, \ldots, q, \tag{4.56}$$

gewählt werden, wobei δ_{ij} und δ_{ijk} die Koeffizienten der hinzugefügten Regressoren sind und q die Modellordnung ist.

Diese Formulierung von $\tilde{H}_0$ berücksichtigt noch nicht die volle Flexibilität der Definition von ARNN-Modellen, insbesondere die Möglichkeit zur sparsamen Spezifikation eines Modells. In einem weiteren Generalisierungsschritt kann angenommen werden, dass nicht alle Lags von 1 bis zur Ordnung q als Inputs für das zusätzliche Neuron bzw. in der Taylorreihenentwicklung enthalten sein müssen, sondern eine Menge $\mathcal{Q}$ von ausgewählten Lags. Die Null-Hypothese $\tilde{H}_0$ muss dann so interpretiert werden, dass die i, j und k nicht die Lags sondern die Elemente der Menge $\mathcal{Q}$ indizieren, deren Mächtigkeit q ist. Durch die Möglichkeit, dass $\mathcal{Q}$ bloß eine begrenzte Auswahl von Lags enthält, kann das Problem vermieden werden, dass die Anzahl der aus einer Taylorreihenentwicklung resultierenden Regressoren potentiell sehr groß wird.

Die Testprozedur kann nun in folgender Weise ausgeführt werden:

(i) Schätze auf y_t ein ARNN-Modell gemäß der Spezifikation nach (4.1) bzw. (4.2). Berechne die Residuenzeitreihe $\hat{u}_t$. Die Summe der quadrierten Residuen sei $SSR_0 = \sum \hat{u}_t^2$.

(ii) Regressiere $\hat{u}_t$ auf $\nabla\varphi(\tilde{\mathbf{y}}_t, \tilde{\mathbf{w}})$ und die sich aus der Taylorreihenentwicklung ergebenden Variablen. Hierbei ist φ die im Rahmen des stochastischen Teils des Modells geschätzte Funktion, $\mathbf{y}_t$ ist der Vektor der um den geschätzten deterministischen Teil bereinigten und der entsprechend $\mathcal{J}$ und $\mathcal{L}$ zeitverzögerten y_t, $\tilde{\mathbf{w}}$ ist der Vektor der geschätzten Gewichte im stochastischen Teil des Modells und ∇ bezeichnet die Ableitung nach $\tilde{\mathbf{w}}$. Berechne die Residuenzeitreihe $\hat{v}_t$. Die Summe der quadrierten Residuen sei $SSR = \sum \hat{v}_t^2$.

(iii) Berechne die Test-Statistik

$$F = \frac{(SSR_0 - SSR)/m}{SSR/(n - p - 1 - m)}, \tag{4.57}$$

wobei m die Anzahl der aus der Taylorreihenentwicklung gewonnenen Regressorvariablen ist, p die Anzahl der Parameter in $\tilde{\mathbf{w}}$ und n die Anzahl der benutzten Beobachtungen. Unter der Null-Hypothese, dass das Modell zur Erklärung der Zeitreihe ausreicht, ist F asymptotisch F-verteilt mit m und $n - p - 1 - m$ Freiheitsgraden.

Das Testverfahren ist als Funktion `TLGtest` im Rahmen des R-Pakets `NNUTS` implementiert, die in der folgenden Anwendung verwendet wird. Diese Implementation des TLG-Tests zeichnet sich gegenüber anderen Implementationen durch ihre Flexibilität bei der Spezifikation des unter der Null-Hypothese zu schätzenden Modells sowie der im zusätzliche Neuron erfassten Inputs (Lags) aus. Sowohl die im linearen Teil des ARNN- bzw. ARNNDS-Modells als auch die im NN-Teil des Modells enthaltenen Lags können sparsam spezifiziert werden. Diese Spezifikationen, die in (4.1) und (4.2) als Mengen $\mathcal{J}$ und $\mathcal{L}$ notiert werden, können auch in Vektorform vorgenommen werden. Sind $\mathcal{J}$ und $\mathcal{L}$ identisch, so wird, um möglichst nahe bei der originalen Formulierung des TLG-Test zu bleiben, der Vektor $\mathbf{p}$ verwendet, ansonsten $\mathbf{p}_{\text{lin}}$ bzw. $\mathbf{p}_{\text{neur}}$. Analog kann $\mathcal{Q}$ auch als Vektor $\mathbf{q}$ spezifiziert werden.

Anwendung des Growing

Anhand der Anwendung des TLG-Testverfahrens auf die simulierte Zeitreihe SIM einerseits und auf die realweltliche Zeitreihe IPI andererseits soll nun seine Funktionsweise illustriert werden. In Tab. 15 sind ausgewählte Ergebnisse des TLG-Tests auf Signifikanz von zusätzlichen verdeckten Neuronen für die Zeitreihe SIM zusammengefasst.

Es werden für verschiedene Vorgaben der Vektoren $\mathbf{p}$ und $\mathbf{q}$ jeweils die Tests auf die Signifikanz eines zusätzlichen verdeckten Neurons getestet, wobei unter der Null-Hypothese bereits $K - 1$ Neuronen vorhanden sind, somit das K-te Neuron auf Signifikanz getestet wird. Zunächst sei darauf hingewiesen, dass sich bereits in diesem sehr einfachen Anwendungsbeispiel viele unterschiedliche sinnvolle Wahlmöglichkeiten für die Spezifikation des Testverfahrens ergeben. Wenn man zusätzlich die Möglichkeit, eines sparsam spezifizierten Modells unter der Null-Hypothese berücksichtigt, werden es noch mehr Wahlmöglichkeiten. Im Gegensatz zur weiter unten präsentierten Anwendung anhand der Zeitreihe IPI ist hier der Vektor $\mathbf{p}$ voll spezifiziert.

Wie die Tabelle zeigt, ist das Hauptergebnis der Berechnungen, dass mindestens zwei verdeckte Neuronen signifikant bei üblichen Konfidenzniveaus sind. Die Macht des Tests hängt zudem von der richtigen Wahl von $\mathbf{p}$ und $\mathbf{q}$ ab. Wählt man für die Modellordnung den (wahren) Wert 1, so liefert der Test auch für das dritte verdeckte Neuron ein signifikantes Ergebnis.

Weiters sei erwähnt, dass auch in dieser einfachen Anwendung bei $K > 1$, d.h. wenn das unter der Null-Hypothese zu schätzende ARNN-Modell mindestens ein verdecktes Neuron aufweist, Probleme mit lokalen Minima auftreten. Für die Aus-

Tab. 15: Ergebnisse des TLG-Tests auf Signifikanz von zusätzlichen verdeckten Neuronen, SIM

K	$\mathbf{p}'$	$\mathbf{q}'$	Freiheitsgr.	F-Stat.	p-Wert
1	(1,2,3,4)'	(1,2,3,4)'	30, 471	7.1243	0.0000***
2	(1,2,3,4)'	(1,2,3,4)'	30, 465	1.7657	0.0083**
3	(1,2,3,4)'	(1,2,3,4)'	30, 459	0.9712	0.5123
1	(1,2,3,4)'	(1,2)'	7, 494	28.5940	0.0000***
2	(1,2,3,4)'	(1,2)'	7, 488	4.9171	0.0000***
3	(1,2,3,4)'	(1,2)'	7, 482	1.4415	0.1864
1	(1,2)'	(1,2)'	7, 498	28.2945	0.0000***
2	(1,2)'	(1,2)'	7, 494	4.7060	0.0000***
3	(1,2)'	(1,2)'	7, 490	1.7646	0.0923
1	1	1	2, 505	88.2344	0.0000***
2	1	1	2, 502	12.6795	0.0000***
3	1	1	2, 499	3.7948	0.0231*
4	1	1	2, 496	1.4763	0.2295

***, ** und * bezeichnen Werte, die bei einem Konfidenzniveau von 0.1 %, 1 % bzw. 5 % signifikant sind.

sagekraft des Test ist es essentiell, dass das globale oder zumindest ein nahezu optimales Minimum gefunden wird. Im Rahmen dieser Anwendung ist dies durch Probieren oder durch Multistart-Verfahren leicht zu erreichen.

Die nächste Anwendung hat die Zeitreihe IPI, saisonale Differenzen $\nabla_s x_t$, zum Gegenstand. Eine Anwendung auf die Zeitreihe ALR wurde aus Platzgründen nicht aufgenommen, diese liefert jedoch vergleichbare Ergebnisse. Es sei außerdem auf den Abschnitt 3.3 verwiesen, in dem der Test auf die Signifikanz des ersten verdeckten Neurons durchgeführt wird, allerdings mit dem Unterschied, dass dort der Vektor $\mathbf{p}$ voll spezifiziert ist. Da nun die Zielsetzung des Testverfahrens nicht in erster Linie die Feststellung von nicht-linearer Struktur sondern die Unterstützung der Modellierung im Rahmen des statistischen Modellierungsansatzes ist, ist es sinnvoll, die Erkenntnisse der linearen Modellierung in die Spezifizierung des Tests einfließen zu lassen und $\mathbf{p}$ sparsam zu spezifizieren. Wie oben ausgeführt, wird außerdem zwischen $\mathbf{p}_{\text{lin}}$ und $\mathbf{p}_{\text{neur}}$ unterschieden, da die Lags, die unter der Null-Hypothese des Test im linearen Teil und in einem allenfalls vorhandenen NN-Teil des Modells aufscheinen, nicht notwendigerweise identisch sein müssen. Tab. 16 enthält die Ergebnisse der Testanordnung.

In der Tabelle werden verschiedene Möglichkeiten für die Auswahl von $\mathbf{p}_{\text{neur}}$ und $\mathbf{q}$ verwendet. Bereits in Abschnitt 3.3 (vgl. Abb. 8) wurde die Vermutung nahegelegt, dass die für nicht-lineare Struktur besonders relevanten Lags die Lags 1, 3 und 12 sind, die daher zu diesem Zweck herangezogen werden.

Das Hauptergebnis ist, dass die Null-Hypothese eines rein linearen Modells zugunsten eines Modells mit einem verdeckten Neuron, $K = 1$, im Rahmen von verschiedenen Testspezifikationen sicher verworfen werden kann, wobei die Signifikanz

Tab. 16: Ergebnisse des TLG-Tests auf Signifikanz von zusätzlichen verdeckten Neuronen, IPI

K	$\mathbf{p}'_{\text{neur}}$	$\mathbf{q}'$	Freiheitsgr.	F-Stat.	p-Wert
1	–	(1,3)	7, 375	2.8886	0.0060**
1	–	(1,12)	7, 375	2.8276	0.0070**
1	–	(1,3,12)	16, 366	2.5783	0.0008***
2	(1,3)	(1,3)	7, 371	0.9603	0.4601
2	(1,12)	(1,12)	7, 371	2.6020	0.0124*
2	(1,3,12)	(1,3,12)	16, 361	1.6958	0.0438*
2	$\mathbf{p}'_{\text{lin}}$	(1,3)	7, 361	1.1726	0.3176
2	$\mathbf{p}'_{\text{lin}}$	(1,12)	7, 361	2.0177	0.0521
2	$\mathbf{p}'_{\text{lin}}$	(1,3,12)	16, 352	1.5630	0.0765

$\mathbf{p}'_{\text{lin}} = (1, 2, 3, 12, 14, 15, 24, 26, 27, 36, 48, 49)$
***, ** und * bezeichnen Werte, die bei einem Konfidenzniveau von 0.1 %, 1 % bzw. 5 % signifikant sind.

gemessen am p-Wert am höchsten ist, wenn alle drei Lags 1, 3 und 12 in die Taylorreihenentwicklung eines zusätzlichen Neurons aufgenommen werden. Hingegen kann die Frage der Signifikanz eines zweiten Neurons nicht eindeutig entschieden werden und bedarf zusätzlicher Erläuterungen.

Ist unter der Null-Hypothese des Tests bereits ein verdecktes Neuron vorhanden (in der Tabelle betrifft dies die Zeilen mit $K = 2$), so stellt sich die Frage, ob dieses mit allen Inputknoten, d.h. allen Lags, die bereits im linearen Teil des Modells vorhanden sind, verbunden werden soll oder nur mit jenen, die im Vektor $\mathbf{q}$ des zuvor durchgeführten und signifikant ausgefallenen TLG-Test enthalten waren. Im ersten Fall ist also $\mathbf{p}_{\text{neur}} = \mathbf{p}_{\text{lin}}$, im zweiten Falle ist in der Tabelle $\mathbf{p}_{\text{neur}} = (1, 3)$, $\mathbf{p}_{\text{neur}} = (1, 12)$ bzw. $\mathbf{p}_{\text{neur}} = (1, 3, 12)$.

Im Rahmen dieser Arbeit wird der erste Weg bevorzugt, da er im Rahmen einer kombinierten Growing-Pruning-Strategie in leicht nachvollziehbarer Weise zu handhaben ist. Hier zeigt sich, dass die Null-Hypothese, dass ein einzelnes verdecktes Neuron ausreicht, um den wahren Prozess approximieren zu können, nicht abgelehnt werden kann. Die Durchführung des Tests ist allerdings von Schwierigkeiten mit zahlreichen lokalen Minima in der Fehlerfunktion des unter der Null-Hypothese zu schätzenden Modells beeinträchtigt, die nicht verschwiegen werden sollten. Auch unter Zuhilfenahme eines großzügig ausgestatteten Multistart-Verfahrens wird bei mehrmaliger Durchführung des Testverfahrens meist ein anderes Minimum gefunden, so dass die Testergebnisse bloß ungefähr zu replizieren sind. Außerdem fällt auf, dass das Testverfahren dazu tendiert, bei einem *niedrigeren* SSE des unter der Null-Hypothese geschätzten Modells *niedrigere* p-Werte zu liefern. Dies ist kontraintuitiv, da man erwarten müsste, dass ein geschätztes Modell, das einen größeren Anteil der Varianz erklärt, weniger an Varianz für ein zusätzliches Neuron zum Erklären übrig lässt.

Auch die Ergebnisse für sparsam spezifizierte $\mathbf{p}_{\text{neur}}$ sollen näher erläutert werden. Hier wurde die naheliegende Vereinfachung getroffen, dass $\mathbf{q}$ und $\mathbf{p}_{\text{neur}}$ jeweils übereinstimmen sollen. Es zeigt sich, dass bei diesen Einstellungen teilweise signifikante Testergebnisse errechnet werden. Auch hier kommt es zu einer Beeinträchtigung des Testverfahrens durch das Auftreten von mehreren lokalen Minima für das unter der Null-Hypothese zu schätzenden Modells, und aus diesem Grund sind die Teststatistiken nur als typische Werte zu betrachten. Lehnt man aufgrund dieser Testergebnisse die Null-Hypothese eines Modells mit einem einzelnen verdeckten Neuron als optimale Approximation zugunsten eines Modells mit zwei verdeckten Neuronen ab, so würde eigentlich in der TLG-Testanordnung als nächster Schritt die Testung der Signifikanz eines dritten zusätzlichen Neurons folgen. Auf den Bericht dieser Ergebnisse wird jedoch an dieser Stelle verzichtet, da die Testergebnisse noch stärker von den oben erwähnten Problemen beeinträchtigt werden. Außerdem haben Modellierungsversuche mit mehr als zwei Neuronen klar auf stattfindendes Overfitting sowie auf Probleme mit der Stabilität eines nachfolgenden Pruning-Prozesses hingewiesen.

Die Anwendungen zeigen, dass die praktische Anwendung der Growing-Strategie mit Hilfe des TLG-Tests auf makroökonomische Zeitreihen eine Reihe von Problemen berücksichtigen muss. Obwohl sich das Verfahren bei tatsächlich vorhandener Nicht-Linearität und geringer Modellordnung bzw. -komplexität gut anwenden lässt, ist bei realweltlichen Zeitreihen durch die höhere Modellordnung und durch die relativ große Anzahl von Parametern die Anwendbarkeit auf maximal zwei verdeckte Neuronen beschränkt. Für die Modellierung der beiden Beispielzeitreihen mit dem statistisch-parametrischen Ansatz können die Ergebnisse so interpretiert werden, dass die optimale Anzahl von verdeckten Neuronen $K = 1$ ist. Das Neuron wird zunächst mit allen im linearen Teil vorhandenen Lags verbunden und kann in einem nächsten Schritt dem Pruning unterzogen werden.

Pruning

Der Prozess des Prunings wird in der wissenschaftlichen Literatur als relativ unproblematische Vorgangsweise dargestellt, bei der durch den Forscher bzw. Praktiker statistisch nicht signifikante Gewichte in einem interaktiven Prozess ausgeschieden werden, solange bis das geschätzte Modell nur mehr zu einem vorgegebenen Konfidenzniveau statistisch signifikante Gewichte enthält (z.B. Anders und Korn, 1999). In der praktischen Anwendung sind jedoch eine Vielzahl von Detailentscheidungen zu treffen, die im Folgenden erläutert werden sollen. Eine generelle Empfehlung kann lauten, dass besonders bei der Modellierung von makroökonomischen Zeitreihen das Pruning vorsichtig erfolgen sollte. Die Hauptgefahr, welche es zu vermeiden gilt, ist dass sich nach dem Entfernen eines Parameters aus einem ARNN- bzw. ARNNDS-Modell und nach dessen Neuschätzung eine stark unterschiedliche Lösung ergeben könnte. Die Umstände, die dazu beitragen, sind das Vorhandensein von mehreren

Minima in der Fehleroberfläche und die mögliche Nicht-Identifikation des Modells.

Die folgenden Aspekte müssen bei der Konkretisierung der Vorgangsweise beachtet werden:

- *Welche Parameter sollen grundsätzlich dem Pruningverfahren unterworfen werden?* Die Forschungsliteratur beschränkt sich auf das Pruning der Gewichte der ersten Schicht (die a_{kj} im ARNN-Modell). Ist jedoch auch ein linearer Teil (Shortcuts) im Modell enthalten, sind prinzipiell auch diese Gewichte (die f_l im ARNN-Modell) dem Pruning-Prozess zuzuführen. Auch eine Einschränkung des Pruning auf entweder den linearen Teil oder den NN-Teil des Modells kann gegebenenfalls sinnvoll sein. Andere Parameter, z.B die Bias-Gewichte und die Paramter des deterministische Teils, kommen für das Pruning eher nicht in Frage.

- *Wie und in welcher Reihenfolge sollen die Pruning-Kandidaten ausgewählt werden?* Im Idealfall genügt als Auswahlkriterium für die Pruning-Kandidaten, die Frage, ob der Wahrscheinlichkeitswert der Parameter über dem vorgegebenen Konfidenzniveau liegt, und am Ende des Pruningprozesses steht unabhängig von der Reihenfolge der ausgeschiedenen Parameter immer das gleiche geschätzte Modell. Bei komplizierten Modellen hängt allerdings das Ergebnis und mitunter auch die Stabilität des Pruningverfahrens insgesamt von der Reihenfolge ab. Naheliegend ist, Parameter mit höheren Wahrscheinlichkeitswerten zuerst zu eliminieren. Weitere Kriterien, mit denen man die Wirkungsweise und das Ergebnis des Verfahrens beeinflussen kann, sind der Lag (z.B. Parameter mit einem höheren Lag j bzw. l zuerst wegschneiden) und die Frage der Linearität (z.B. Parameter des linearen Teils zuerst, dann erst jene des nicht-linearen Teils).

- *Wie soll bei der Neu-Schätzung des Modells nach einem Pruning-Schritt vorgegangen werden?* Auch dies ist eine Frage, die bei genügend großer Stichprobe irrelevant wird. Bei kleiner Stichprobe und den oben erwähnten Problemen hingegen kann es für die Stabilität des Pruningverfahrens, aber auch für die Beschleunigung des Verfahrens hilfreich sein, verschiedene Hilfsmittel einzusetzen. So ist es zu empfehlen, bei der Neu-Schätzung des Modells nach der Entfernung eines Parameters die bereits gefundenen Werte zur Initialisierung zu verwenden. In der praktischen Anwendung der Schätzung von ARNN-Modellen auf makroökonomischen Zeitreihen hat sich jedoch erwiesen, dass dies nicht ausreicht, um im neuen Parameterraum jenes lokale Minimum wiederzufinden, das dem zuvor gefundenen lokalen Minimum im alten Parameterraum (d.h. vor dem Pruningschritt) entspricht. Dieses Problem scheint insbesondere in nicht identifizierten Modellen aufzutreten, welche allerdings möglichst vermieden werden sollten. Abhilfe kann geschaffen werden, indem der Übergang vom alten Parameterraum zum neuen, niedriger dimensionierten Parameterraum sanft vorgenommen wird. Dies geschieht durch Festlegen

eines Strafterms für den wegzuschneidenen Parameter im Sinne der „Weight Decay“ Regularisierung und mehrmaliger Neuschätzung mit anwachsendem Strafterm. Erst wenn der Parameter nahezu nach Null gedrängt worden ist, wird er gänzlich entfernt. Auf diese Weise hat das neugeschätzte Modell möglichst große Ähnlichkeit mit jenem vor dem Pruningschritt bzw. mit jenem am Beginn des Pruningverfahrens.

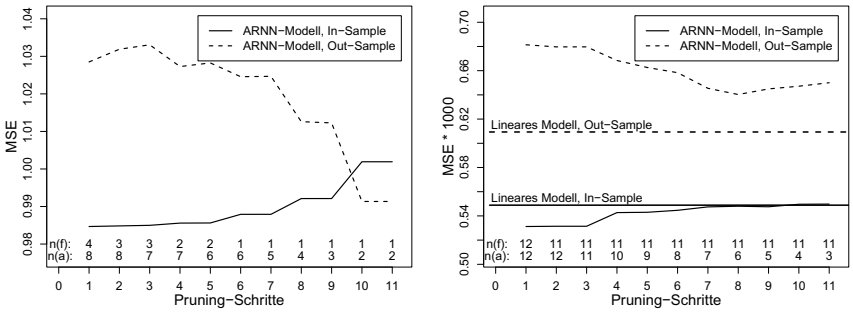
Die hier skizzierten Möglichkeiten sind in der Funktion `prune.arnnds` im Rahmen des R-Pakets `NNUTS` implementiert, welche noch zusätzliche Optionen anbietet, den Pruning-Prozess zu steuern. Dennoch treten bei der Anwendung des Verfahrens mitunter Probleme und Anomalien auf. So ist es nicht auszuschließen, dass ein Modell, das vor dem Pruning-Prozess sowohl in-sample als auch out-sample besser als das lineare Vergleichsmodell ist, im Zuge des Pruningverfahrens diesen Vorsprung verliert. Eine Anwendung soll das Funktionieren des Verfahrens auf der simulierten Zeitreihe und auf der Zeitreihe IPI illustrieren.

Anwendung zum Pruning

Für die Anwendung des Pruning-Ansatzes werden die beiden Zeitreihen SIM ($n = 700$) und IPI ($\nabla_s x_t$, $n = 444$) herangezogen. Um den Zusammenhang des Prunings mit der Generalisierungsfähigkeit darstellen zu können, werden die Zeitreihen in ein Lernset (In-Sample-Set) und ein Evaluierungsset (Out-of-Sample-Set) unterteilt, wobei letzteres zur Überprüfung der Prognosegüte herangezogen wird. Der Umfang des Evaluierungssets beträgt für SIM 200 und für IPI 60.

Zuerst werden die Ergebnisse des Pruningprozesses für die Zeitreihe SIM besprochen, welche in Abb. 11 auf der linken Seite zusammengefasst werden. Es sei zunächst daran erinnert, dass die Zeitreihe eine Varianz von 1.858 aufweist. Der Fehlerprozess des die Daten generierenden Modells hat eine Varianz von 1. Der Großteil der erklärbaren Varianz entfällt auf die nicht-lineare Komponente des Modells. Man kann erwarten, dass bei genügend großer Länge der Zeitreihe ein ausreichend parametrisiertes ARNN-Modell einen MSE von ungefähr 1 aufweist. Dass dies in der Anwendung tatsächlich der Fall ist zeigt, dass ARNN-Modelle grundsätzlich gut mit etwaiger Nicht-Linearität umgehen können.

Das Ausgangsmodell, mit dem der Pruningprozess begonnen wird, hat zwei verdeckte Neuronen. Diese Anzahl wurde mit dem oben geschilderten Testverfahren mit Hilfe des TLG-Tests ermittelt (vgl. Tab.15). Sowohl der NN-Teil als auch der lineare Teil des Modells weisen die Lags 1 bis 4 als Inputs auf. In dieser Hinsicht ist das Modell also überparametrisiert, da das wahre Modell die Ordnung 1 aufweist. Die Abbildung zeigt, wie sich das Modell während des Pruning-Prozesses von insgesamt 11 Schritten (das Ausgangsmodell eingeschlossen) entwickelt. Am unteren Rand sind die Anzahl der Parameter des linearen Teils, $n(f)$, und der ersten Schicht des nicht-linearen Teils, $n(a)$, angegeben. Deren Anzahl verringert sich, bis im endgültigen Modell nur die Parameter zum Lag 1 im Modell verbleiben. Die



(a) SIM (b) IPI

Abb. 11: Entwicklung der Prognosegüte (In-Sample und Out-of-Sample) während des Pruningprozesses, Zeitreihen SIM und IPI

Modellqualität, gemessen am MSE, entwickelt sich wie erwartet. Der MSE auf dem In-Sample-Set steigt in jedem Pruning-Schritt tendenziell an, während er auf dem Evaluierungsset abnimmt. Die Befreiung des Modells von überflüssigen Parametern führt zu einer verbesserten Generalisierungsfähigkeit.

In der Abb. 11, linke Seite wird die Entwicklung des Pruningprozesses für die Zeitreihe IPI dargestellt. Das Ausgangsmodell weist hier einen NN-Teil mit nur einem verdeckten Neuron sowie einen linearen Teil auf. Die Lag-Struktur umfasst zwölf Lags von 1 bis 36. Die genaue Lag-Struktur des Ausgangsmodells wie auch des linearen Benchmark-Modells wurde in ähnlicher Weise bestimmt wie im Abschnitt 2.4 beschrieben. Während des Pruning-Prozesses von insgesamt 11 Schritten wird die Anzahl der Parameter vor allem im NN-Teil des Modells reduziert. Dieses Merkmal des Pruningprozesses lässt sich jedoch nicht verallgemeinern, sondern hängt stark vom Zufall ab. Das endgültige Modell enthält jeden der ursprünglichen Lags entweder im NN-Teil oder im linearen Teil. Die Lags 1 und 3, an denen eventuell nicht-lineare Strukturen hängen, sind sowohl im linearen als auch im NN-Teil enthalten. Anhand der Entwicklung des MSE auf dem Evaluierungsset zeigt sich auch in dieser Illustration die günstige Wirkung des Pruning auf die Generalisierungsfähigkeit, allerdings nur in relativ schwachem Ausmaß. Ein Vergleich mit dem linearen Benchmark-Modell liefert jedoch die ernüchternde Erkenntnis, dass die Out-of-Sample-Prognosegüte sowohl des Ausgangsmodells als auch des endgültigen Modells deutlich schlechter ist als diejenige des linearen Modells.

Die Anwendung belegt, dass der Pruning-Prozess prinzipiell funktioniert, vor allem wenn die Zeitreihe tatsächliche Nicht-Linearität aufweist. Für echte makroökonomische Zeitreihen liefert diese Anwendung nur einen Eindruck. Wenn auch das Ergebnis typisch ist und für die Beispielzeitreihe ALR ähnliche Ergebnisse präsentiert werden könnten, so muss dennoch betont werden, dass das Pruning

in der Anwendung auf realweltliche Zeitreihen auch mit dem Risiko verbunden ist, ein gutes Ausgangsmodell zu verschlechtern. Das schlechte Abschneiden des Modells im Vergleich zum linearen Benchmarkmodell weist eventuell darauf hin, dass das Ausgangsmodell keinen linearen Teil haben sollte. Aber auch dieser alternative Ansatz bringt ähnliche Unwägbarkeiten mit sich, wie die folgenden Ergebnisse und ausführlicher auch die Analysen in Kapitel 5 zeigen.

Ergebnisse der ARNN-Modellierung auf den Beispielzeitreihen

Der statistisch-parametrische Ansatz wird nun auf die beiden Beispielzeitreihen ALR und IPI angewendet. Da der Praktiker in verschiedenen Aspekten der Vorgangsweise Spielraum hat bzw. an verschiedenen Punkten zusätzliche Vereinfachungen und Heuristiken angebracht sind, sei die Modellierungsstrategie hier kurz zusammengefasst:

(i) Ausgangsmodell ist das lineare AR- bzw. ARDS-Modell, dessen Spezifizierung wie in Abschnitt 2.2 beschrieben vorgenommen wird. Die Lag-Struktur des linearen Modells, welche mit Pruning ermittelt wird, definiert die Lags, die grundsätzlich für die ARNN bzw. ARNNDS-Modellierung in Frage kommen. Die endgültige Auswahl der Lags ist jedoch in der weiteren Modellierung durch den Pruning-Prozess vorgesehen. Die Spezifizierung des deterministischen Teils des Modells, insbesondere die Aufnahme eines Interzepts, wird vom linearen Modell in das ARNN bzw. ARNNDS-Modell übernommen und wird danach im Rahmen des Pruning-Prozesses nicht mehr in Frage gestellt.

(ii) Growingprozess: Entscheidung der Frage über die Hinzufügung eines verdeckten Neurons anhand des TLG-Tests über vernachlässigte Nicht-Linearität. Wenn der Test die Null-Hypothese, dass in den Residuen keine nicht-lineare Struktur vorhanden ist, ablehnt, so wird die Modellstruktur um ein mit allen Inputknoten (Lags) verbundenes verdecktes Neuron erweitert. Dies wird so lange wiederholt bis keine weiteres verdecktes Neuron mehr benötigt wird.

(iii) Pruningprozess: schrittweises Entfernen der bei einem vorgegebenen Signifikanzniveau nicht signifikanten Inputknoten. Im Pruning-Prozess werden die Koeffizienten einzeln entfernt, wobei Präferenzen vorgegeben werden, dass Koeffizienten des linearen Teils zuerst eliminiert werden sollen. Nach jedem einzelnen Pruningschritt wird das Modell unter Verwendung der bereits gefundenen Werte für die verbleibenden Koeffizienten und mit Anwendung weiterer Heuristiken neu geschätzt.

In Schritt ii) der obigen Vorgangsweise besteht grundsätzlich die alternative Möglichkeit bei Ansprechen des TLG-Tests unter Vorgabe des linearen Ausgangsmodells, das Modell nicht um ein verdecktes Neuron zu *ergänzen*, sondern den linearen Teil des Modells durch einen NN-Teil zu *ersetzen*. Aus theoretischer Sicht

spricht nichts gegen eine solche Vorgangsweise. Im Rahmen der vorliegenden Arbeit zeigt sich, dass dies große Auswirkungen auf die Struktur und Prognosegüte des Modells hat. Im konkreten Fall wurde diese Option verworfen, da der Pruning-Prozess auf einem Modell ohne linearen Teil nicht einwandfrei funktioniert. Obwohl auf der Basis des TLG-Test das Modell identifiziert sein müsste, zeigt sich in den typischen Anwendungen auf den Beispielzeitreihen der folgende für nicht identifizierte Modelle typische Effekt. Das ohne linearen Teil und mit nur einem verdeckten Neuron geschätzte Modell hat Koeffizienten a_{1j}, die abgesehen von Skalierung weitgehend mit den Koeffizienten f_j des linearen Modells übereinstimmen. Die Berücksichtigung etwaiger nicht-linearer Struktur ist in diesem Modell schwach ausgeprägt, da im Schätzprozess die Anpassung der linearen Struktur dominiert. Im Ergebnis können keine sinnvollen Standardfehler ermittelt werden. Weitergehende Versuche, im Rahmen des statistisch-parametrischen Ansatzes Modelle ohne linearen Teil zu berücksichtigen (z.B. indem für das Modell mit nur einem verdeckten Neuron das Pruning ausgelassen wird), waren im Ergebnis unklar.

In den Tab. 17 und 18 sind die Ergebnisse der Modellierung und Schätzung von ARNN-Modellen auf den Beispielzeitreihen zusammengefasst, wobei es sich hierbei um die saisonalen Differenzen ∇_{12} handelt, die aufgrund der Unit-Root-Testergebnisse (vgl. Kap. 2) die präferierte Zeitreihe für die Modellierung sind. Beide Modelle erfüllen formal die Stationaritätsbedingungen, indem die größte Wurzel des charakteristischen Polynoms knapp unter Eins liegt. Auch in Simulationen auf der Basis dieser Modelle zeigt sich kein explosives oder pseudo-nicht-stationäres Verhalten.

Bei der Betrachtung der Modelle und der durch das Pruning gefundenen Form fällt auf, dass bei der ALR alle Lags, die im linearen Teil aufscheinen auch im NN-Teil des Modells enthalten sind. Von den im Ausgangsmodell enthaltenen Lags wurde hier durch das Pruning nur der Lag 26 weggeschnitten. Im Gegensatz dazu führt bei der Zeitreihe IPI der Pruning-Prozess dazu, dass es zu einer gewissen „Arbeitsteilung“ zwischen dem linearen Teil und dem NN-Teil des Modells kommt, indem aus dem NN-Teil die meisten Lags weggeschnitten werden und der Lag 2 nur im NN-Teil enthalten ist. Es sei aber erwähnt, dass dieses Ergebnis bei mehreren Läufen des Pruning-Prozesses nicht immer reproduzierbar ist. Durch etwaige Zufälligkeiten der Initialisierung usw. kann es auch dazu kommen, dass umgekehrt die meisten Lags im linearen Teil weggeschnitten werden.

Der Q-Statistik nach spricht nichts gegen eine richtige Spezifizierung der Modelle. Gemessen am R^2adj. sind die Modelle jeweils etwas besser als die entsprechenden linearen Modelle (vgl. Tab. 3 und 4 in Abschnitt 2.4): 0.8890 versus 0.8779 im Falle der ALR und 0.7459 versus 0.6976 für die IPI-Zeitreihe. Ob damit Generalisierungsfähigkeit erreicht ist, kann indessen ohne eine Evaluierung auf einer echten Out-of-Sample-Zeitreihe noch nicht gesagt werden.

Tab. 17: Schätzergebnisse für ein ARNN-Modell, ALR

Zeitreihe: $\nabla_{12} \ln ALR - \ln(100 - ALR) + \ln 100$
Zeitraum: 1961/1 – 1997/12 ($N = 444$)
Schätzmethode: Least Squares

Koeff.	Wert	Standardf.	Koeff.	Wert	Standardf.
$a_{1,1}$	1.5149111	0.1506577 ***	c_1	5.2380759	0.0160130 ***
$a_{1,2}$	-0.9098180	0.2152586 ***	f_1	-6.8154956	0.7550873 ***
$a_{1,4}$	-0.8849616	0.1907823 ***	f_2	4.7815547	1.0895032 ***
$a_{1,8}$	1.6164128	0.1890106 ***	f_4	4.4046368	0.9601261 ***
$a_{1,9}$	-0.5001972	0.2042050 *	f_8	-8.3958181	0.9466063 ***
$a_{1,10}$	-0.8317671	0.2632510 **	f_9	2.6787095	1.0309586 **
$a_{1,12}$	-1.0118906	0.3128275 **	f_{10}	4.3900844	1.3294232 ***
$a_{1,13}$	1.0112422	0.2837676 ***	f_{12}	4.6971571	1.5885249 **
$a_{1,16}$	0.5834609	0.1620261 ***	f_{13}	-4.7746759	1.4424492 ***
$a_{1,24}$	-1.2152696	0.2239161 ***	f_{16}	-3.0630629	0.8233321 ***
$a_{1,25}$	1.0546943	0.1748752 ***	f_{24}	5.9660776	1.1237778 ***
b_1	0.0002682	0.0005308	f_{25}	-5.1263596	0.8768968 ***

R^2adj.: 0.8890
F-Stat.: 147.3206*** (mit 24 und 395 Freiheitsgraden)
$\hat{\sigma}^2$: 0.00261
Q-Stat.(Lag=24): 17.3192
AIC: -2592.658

***, ** und * bezeichnen Werte, die bei einem Konfidenzniveau von 0.1 %, 1 % bzw. 5 % signifikant sind.

Schlussfolgerungen

In diesem Abschnitt wurde der statistisch-parametrische Ansatz der ARNN-Modellierung dargestellt. Der Ansatz beruht auf der statistischen Theorie, insbesondere der Asymptotik, die es ermöglicht, Konfidenzintervalle zu berechnen und Hypothesentests abzuleiten. Beim statistisch-parametrischen Ansatz wird die Generalisierungsfähigkeit des Modells durch die Geringhaltung der Anzahl der Parameter erreicht. Hierzu können die Techniken des Growing und des Pruning verwendet eingesetzt werden, welche in der vorliegenden Arbeit für die besonderen Erfordernisse der Modellierung makroökonomischer Zeitreihen erweitert und modifiziert wurden. Anwendungen belegen das Funktionieren des entwickelten Instrumentariums.

Zu den hauptsächlichen Schlussfolgerungen dieses Abschnitts gehört vor allem die Feststellung des enorm hohen theoretischen und rechnerischen Anspruchs des Ansatzes. Die Anforderungen hinsichtlich Stichprobenumfang, Existenz eines globalen Optimierers, Stationarität und Identifiziertheit des Modells sind in der Praxis oft nicht erfüllt bzw. umständlich zu überprüfen. Zudem muss man feststellen, dass sich einige der Motivationen, die mit dem statistisch-parametrischen Ansatz verbunden werden, kaum erreichen lassen, nämlich Reproduzierbarkeit und heuristikfreie Vorgangsweise. Bei komplexen Modellen sind durch die Zufälligkeiten der Initialisierung und des Optimierungsverfahrens die Ergebnisse selten reproduzierbar. Viele

Tab. 18: Schätzergebnisse für ein ARNN-Modell, IPI

Zeitreihe: $\nabla_{12} \ln IPI$					
Zeitraum: 1961/1 – 1997/12 ($N = 444$)					
Schätzmethode: Least Squares					
Koeff.	Wert	Standardf.	Koeff.	Wert	Standardf.
$a_{1,2}$	9.698753	2.732235 ***	f_3	0.312015	0.043691 ***
$a_{1,12}$	45.678866	15.514387 **	f_{12}	-0.558605	0.048271 ***
$a_{1,15}$	-29.647660	10.688209 **	f_{14}	0.249356	0.043216 ***
$a_{1,36}$	-22.083271	8.920982 *	f_{15}	0.223158	0.046841 ***
$a_{1,38}$	39.629174	12.529537 **	f_{24}	-0.455970	0.046893 ***
$a_{1,48}$	-18.113411	7.046157 *	f_{26}	0.205490	0.043205 ***
b_1	30.405072	9.982520 **	f_{27}	0.202577	0.045238 ***
c_1	0.420040	0.001096 ***	f_{36}	-0.186230	0.041910 ***
d	1.502678	0.306751 ***	f_{38}	0.095133	0.041488 *
f_1	0.348722	0.039434 ***	f_{48}	-0.171314	0.038320 ***
f_2	0.298074	0.042034 ***	f_{49}	0.152210	0.038006 ***
R^2adj.: 0.7459					
F-Stat.: 98.0416*** (mit 22 und 373 Freiheitsgraden)					
$\hat{\sigma}^2$: 0.000473					
Q-Stat.(Lag=36): 28.6202					
AIC: -3355.56					

***, ** und * bezeichnen Werte, die bei einem Konfidenzniveau von 0.1 %, 1 % bzw. 5 % signifikant sind.

Detailfragen im Rahmen der Modellierungsstrategie, so z.B. die Frage ob direkte Verbindungen in das Modell aufzunehmen sind oder in welcher Reihenfolge die Gewichte beim Pruning auszuwählen sind, bleiben ohne theoretische Untermauerung und können am besten durch Heuristiken gelöst werden.

Dennoch gelingt es auf den beiden Beispielzeitreihen IPI und ALR plausible Modelle zu schätzen, die im Vergleich mit den linearen Modellen aus Kap. 2 und bei einfachen Modelldiagnosen standhalten. Der letztendliche Wert solcher Modellierungsbemühungen kann jedoch erst anhand von Tests auf Prognosegüte anhand eines echten Evaluierungsset bestimmt werden.

4.7 Klassischer Ansatz mit Early-Stopping

Alternativ zu konvergenten Verfahren haben in der Literatur über neuronale Netze nicht-konvergente Methoden eine reichhaltige Tradition (vgl. Finnoff et al., 1993). Eine einfach zu implementierende Methode ist das vorzeitig gestoppte Training („Early-Stopping“, im Folgenden wird teilweise die Abkürzung ES verwendet). Um eine hohe Generalisierung zu erzielen, wird die effektive Komplexität des neuronalen Netzes durch vorzeitiges Stoppen des Trainings reduziert. Hierfür ist es notwendig, das Datenset in ein Trainingsset und ein Validierungsset zu unterteilen. Der Grundgedanke von ES ist, dass während des Trainings das neuronale Netz ab einem

bestimmten Zeitpunkt beginnt, unwesentliche Aspekte in den Daten zu lernen und an Generalisierungsfähigkeit zu verlieren. Daher erwartet man, dass für das Validierungsset der durchschnittliche Fehler ab diesem Zeitpunkt wieder zunimmt. Das Trainingsset wird zur Bestimmung der Gewichte des neuronalen Netzes verwendet, das Validierungsset nur zur Bestimmung des Stoppzeitpunkts und zur Selektion des optimalen Netzes (also jenes Netzes, für das der Fehler auf dem Validierungsset das Minimum annimmt). Allenfalls kann ein weiteres Datenset beiseite gelegt werden, das nicht für die Schätzung des Modells sondern bloß für die Evaluierung der Prognosegüte dient. Dieses wird als Evaluierungsset bezeichnet.

Zu den Vorteilen von ES zählt seine leichte Verständlichkeit und Implementierbarkeit. Auch hinsichtlich Rechenintensität schneidet dieses Verfahren relativ gut ab, da in vielen Anwendungsfällen nur wenige Iterationen durchgeführt werden müssen, bis das Training gestoppt wird. Umgekehrt könnte es auch zu sehr langen Lernläufen kommen, wenn der Fehler auf dem Validierungsset zufällig sehr lange nicht wieder zunimmt. Zu den Nachteilen von ES wird vor allem der ineffiziente Umgang mit dem Datenset gezählt, da das Validierungsset für die Bestimmung der Modellparameter nicht bzw. nur indirekt benützt wird. Als rein heuristisches Vorgehen beruht ES bloß auf der *Hoffnung*, dass ein niedriger durchschnittlicher Fehler auf dem Validierungsset mit einem niedrigen durchschnittlichen Fehler auf neuen Daten (bzw. auf dem Evaluierungsset) korrespondiert. Ein günstiges Abschneiden auf dem Validierungsset könnte aber auf dem Zufall beruhen.

Early-Stopping wird üblicherweise mit einem verbessertem Gradientenabstiegsverfahren kombiniert. Dieses reicht aus, da keine Konvergenz angestrebt wird. Verschiedene Autoren betonen, dass gerade die Kombination von ES und dem relativ langsamen (verbesserten) Backpropagation-Verfahren der Generalisierungsfähigkeit förderlich sei (z.B. Caruana et al., 2001).

Besondere Aufmerksamkeit kommt der Wahl eines guten Stoppkriteriums zu. Prechelt (1998) analysiert in einer umfangreichen Simulation verschiedene Stoppkriterien hinsichtlich Zeiteffizienz und Effektivität (d.h. Prognosegüte des neuronalen Netzes auf dem Evaluierungsset). Es seien $E_{\text{tr}}^{(\tau)}$ und $E_{\text{va}}^{(\tau)}$ der Fehler auf dem Trainingset bzw. auf dem Validierungsset nach Iteration τ (auch „Lernkurve“ genannt). Der Wert $E_{\text{opt}}^{(\tau)}$ wird definiert als der bis zu einer Iteration τ angetroffene niedrigste Wert auf dem Vaidierungsset:

$$E_{\text{opt}}^{(\tau)} = \min_{\iota \leq \tau} E_{\text{va}}^{(\iota)} \tag{4.58}$$

Der Generalisierungsverlust nach Iterationsschritt τ ergibt sich aus dem Vergleich des Fehlers mit dem bisher angetroffenen Minimum auf dem Validierungsset als

$$Q^{(\tau)} = E_{\text{va}}^{(\tau)} / E_{\text{opt}}^{(\tau)} - 1. \tag{4.59}$$

Prechelt (1998) untersucht drei Klassen von Stoppkriterien:

(i) basierend auf $Q^{(\tau)} > \xi_1$ mit vorgegebenem ξ_1, z.B. $\xi_1 = 0.05$,

(ii) basierend auf $Q^{(\tau)}/P_\kappa^{(\tau)} > \xi_2$, wobei $P_\kappa^{(\tau)}$ eine Kennzahl ist, die die Dynamik von $E_{\text{tr}}^{(\iota)}$ im Abschnitt $\tau - \kappa < \iota \leq \tau$ misst (damit wird bezweckt, dass das Training nicht gestoppt wird, solange noch ein starker Trainingserfolg auf dem Trainingsset erzielt wird, der eine schließliche Verbesserung auch auf dem Validierungsset erwarten lässt)

(iii) basierend auf dem ξ_3-maligen, aufeinanderfolgenden Anstieg von $E_{\text{va}}^{(\iota)}$ in den Abschnitten $\tau-\kappa < \iota \leq \tau$, $\tau-2\kappa < \iota \leq \tau-\kappa, \ldots, \tau-\xi_3\kappa < \iota \leq \tau-(\xi_3-1)\kappa$.

Es zeigt sich u.a., dass auf einer breitgefächerten Auswahl von Beispielanwendungen durch die Wahl der Parameter ξ_1, ξ_2 bzw. ξ_3 ein guter Tradeoff zwischen Zeiteffizienz und Effektivität des ES-Algorithmus erzielt werden kann.

In der Simulation von Prechelt (1998) erwies sich außerdem, dass für einen solchen guten Tradeoff die Dauer des überflüssigen Trainings typischerweise zwischen 50 und 100 % der Dauer des mindestens benötigten Trainings beträgt. „Überflüssig“ bezeichnet das Training über das schließlich selektierte Minimum hinaus. Dies legt die Frage nahe, ob an diesem Befund nicht ein neu konstruiertes Stoppkriterum anknüpfen kann, das sich an dem Ausmaß des zu einem bestimmten Iterationszeitpunkt τ des Trainings als überflüssig erachteten Trainings orientiert.

Ein solches neuartiges Stoppkriterium soll im Folgenden definiert werden und im Rahmen der vorliegenden Arbeit dem ES-Ansatz zugrunde gelegt werden. Es sei $R^{(\tau)}$ die Verhältniszahl von überflüssiger Trainingsdauer zu mindestens benötigter Trainingsdauer in Iteration τ:

$$R^{(\tau)} = \tau/\tau_{\text{opt}}^{(\tau)} - 1, \quad \tau_{\text{opt}}^{(\tau)} = \min(\iota | E_{\text{opt}}^{(\iota)} = E_{\text{opt}}^{(\tau)}) \tag{4.60}$$

Die Stoppregel besagt nun, dass das Training solange fortgesetzt wird, bis $R^{(\tau)}$ einen vorgegebenen Wert, z.B. 0.6, überschreitet; d.h. nach dem Erreichen des Minimums der Lernkurve auf dem Trainingsset wird das Training maximal um weitere 60 % der bis zum Minimum durchlaufenen Iterationsschritte fortgesetzt. Um ein zu frühes Abstoppen oder ein zu langes Andauern des Trainings zu verhindern, muss außerdem eine minimale und maximale Anzahl von Iterationen vorbestimmt werden. Nach dem Abstoppen des Trainings werden jene Gewichte, die das Netz nach der Iteration τ_{opt} hatte, als endgültige Gewichte genommen.

Anwendung

Zur Illustration der Funktionsweise des ES-Ansatzes auf Zeitreihen und der möglicherweise auftretenden Probleme werden nun die Ergebnisse einer Anwendung auf den Zeitreihen SIM ($n = 700$), ALR ($\nabla_s x$, $n = 276$), IPI($\nabla_s x$, $n = 444$) präsentiert. Die Zeitreihen werden in ein Lernset und ein Evaluierungsset unterteilt, welches für SIM den Umfang $n_{\text{ev}} = 200$ und für ALR und IPI $n_{\text{ev}} = 60$ hat. Die Lag-Struktur der auf diesen Zeitreihen geschätzten ARNN-Modelle und linearen Benchmark-Modelle

wurde in ähnlicher Weise bestimmt wie im Abschnitt 2.4 beschrieben. Die ARNN-Modelle weisen eine angemessene Komplexität (Anzahl der verdeckten Neuronen, Aufnahme von Shortcuts) auf, so dass die Vorteile des ES-Ansatzes prinzipiell zum Tragen kommen können.

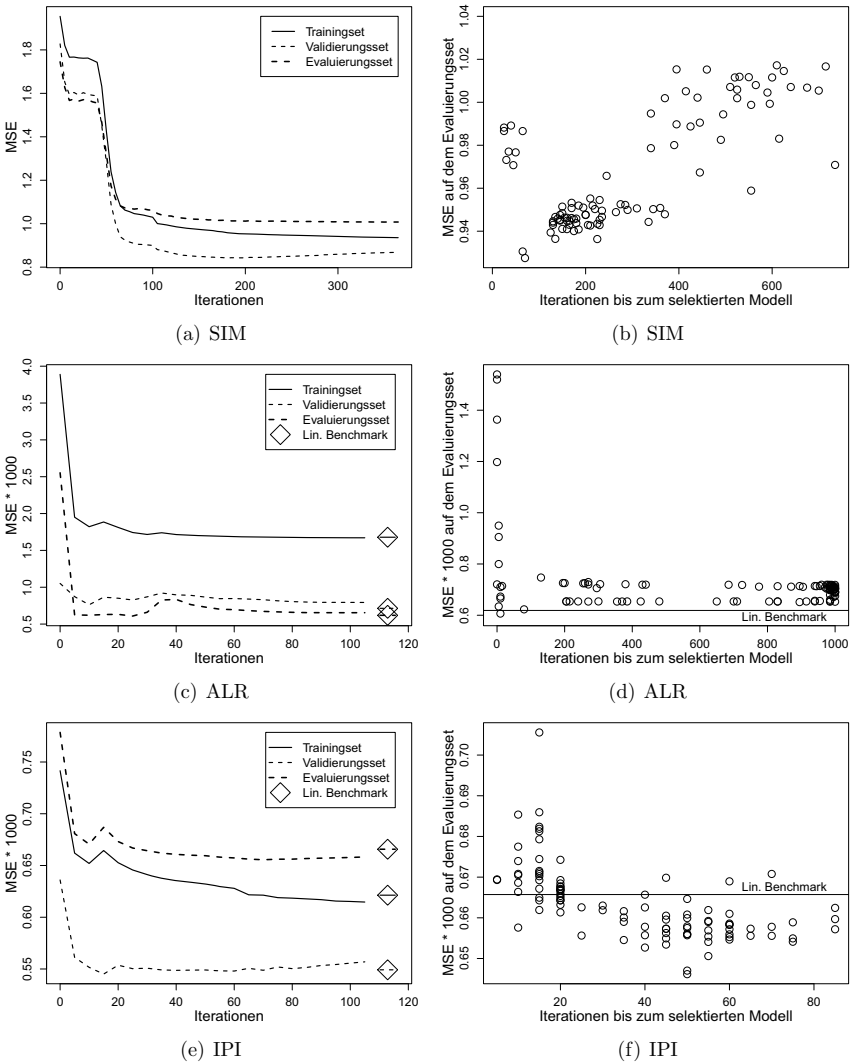
Die in dieser Anwendung gewählten Parameter des Schätzverfahrens sind die folgenden (vgl. Abschnitt 4.4):

- Gradientenabstieg mit Lernrate $\eta = 0.5$ und Momentum $\zeta = 0.5$, wobei sich die Lernrate auf eine mittels Linesearch ermittelte Schrittlänge bezieht,
- Epochbasiertes Lernen mit wachsender Epochengröße, beginnend bei einer Epochengröße von 30 % des Trainingssets, sowie
- ES-Ansatz mit $R^{(\tau)} = 0.5$ sowie minimaler und maximaler Iterationsanzahl 100 bzw. 1000.

In der graphischen Darstellung der Ergebnisse (vgl. Abb. 12) werden jeweils die Lernkurven eines exemplarischen mit dem ES-Ansatz trainierten Modells (linke Seite) sowie die Ergebnisse einer Simulation von 100 wiederholten Läufen gezeigt (rechte Seite).

Anhand der Zeitreihe SIM wird untersucht, wie der ES-Ansatz mit Zeitreihen, die einen tatsächlichen nicht-linearen Anteil haben, zurecht kommt. Die Anzahl der verdeckten Neuronen ist in diesem Beispiel fünf, wobei bereits zwei bis drei Neuronen ausgereicht hätten um die Nicht-Linearität vollständig zu modellieren. An den in Abb. 12 (a) gezeigten Lernkurven ist der gestufte Verlauf typisch, der darauf schließen lässt, dass die Nicht-Linearität erst durch Exploration der Fehlerlandschaft gelernt werden kann (ca. ab Iteration 80). Ca. ab Iteration 100 liegen auf allen der drei Teilsets (Trainingsset, Validierungsset, Evaluierungsset) die MSE-Werte sehr nahe beim wahren Wert der Fehlervarianz $\sigma^2 = 1$. In diesem Beispiel hat der ES-Algorithmus in Iteration 365 gestoppt und das Modell der Iteration 180 ausgewählt.

Weiters wurde eine Simulation mit 100 Läufen auf dieser Realisation der Zeitreihe SIM durchgeführt, die erstens zeigt, wie stark Zufälligkeiten die Ergebnisse des ES-Ansatzes beeinflussen und zweitens den angenommenen Zusammenhang zwischen Trainingsdauer und Overfitting, der ja die grundlegende Motivation des Verfahrens bildet, sichtbar machen. In Abb. 12 (b) wurden auf der X-Achse die Anzahl der Iterationen, die jeweils bis zum selektierten Modell absolviert wurden, aufgetragen und auf der Y-Achse die MSE-Werte auf dem Evaluierungsset der 100 Läufe. Es zeigt sich eine sehr große Variabilität der Anzahl der nötigen Iterationen, in Abhängigkeit von Anfangswerten und der jeweiligen zufälligen Präsentation von Teilsamples im Rahmen des epochenbasierten Lernens. Die positive Korrelation zwischen der Anzahl der Iterationen und dem MSE auf dem Evaluierungsset zeigt, dass es etwa ab Iteration 400 zu Overfitting kommt, das durch die Validierung auf dem Validierungsset offenbar nicht ganz vermieden werden kann. Insgesamt kann jedoch festgestellt werden, dass der ES-Ansatz in allen 100 Läufen der Simulation



(a) SIM (b) SIM

(c) ALR (d) ALR

(e) IPI (f) IPI

Abb. 12: Lernkurven für die Zeitreihen SIM, ALR und IPI, sowie Verteilung der Ergebnisse mit dem Early-Stopping-Ansatz bei wiederholtem Lernen

die tatsächliche Nicht-Linearität der Zeitreihe SIM erfolgreich modelliert und dass die Unterschiede gemessen am MSE gering ausfallen.

Für die beiden Zeitreihen ALR und IPI werden die entsprechenden Ergebnisse mit dem ES-Ansatz in Abb. 12 (c) bis (f) dargestellt. Bei diesen Anwendungen kann zusätzlich mit einem linearen Benchmark-Modell verglichen werden, welches die gleiche Lag-Struktur, aber nur einen linearen Teil aufweist. Bei beiden Zeitreihen sind die typischen Lernkurven von einer schnellen Abnahme des MSE-Werts in den ersten 10 bis 20 Iterationen gekennzeichnet, der ein lang gezogener flacher Bereich folgt. Betrachtet man die Diagramme zu den Simulationsergebnissen, so werden im Falle der ALR gravierende Probleme des Ansatzes offenbar, während sich für IPI ein gutes Funktionieren des ES-Ansatz feststellen lässt.

Die Simulation mit der Zeitreihe ALR zeigt:

- Die Prognosegüte auf dem Evaluierungsset bleibt in allen 100 Läufen hinter derjenigen des linearen Modells zurück. Die erzielten MSE-Werte häufen sich um die Werte 0.00072 und 0.00065, die möglicherweise zwei lokalen Minima oder flachen Regionen der Fehlerlandschaft entsprechen, während der lineare Benchmark 0.00063 beträgt.

- Die Trainingsdauer variiert sehr stark und wird in der Mehrzahl der Fälle erst durch die maximale Anzahl der Iterationen von 1000 begrenzt, d.h. es tritt der Fall auf, dass während des Trainings sehr lange keine Verschlechterung auf dem Validierungsset stattfindet (bzw. dass das Optimum erst nach Iteration 500 auftritt).

- In etwa 10 % der Fälle versagt der ES-Ansatz, indem ein inferiores Netz, das bereits wenige Iterationen nach dem Beginn des Trainings gefunden wird, selektiert wird (dies trotz einer Mindestiterationenanzahl von 100).

Die Simulation mit der Zeitreihe IPI zeigt:

- Die Prognosegüte der mit dem ES-Ansatz geschätzten ARNN-Modelle gemessen am MSE auf dem Evaluierungsset ist im Vergleich zum linearen Benchmark meist besser, hängt aber von der Gestalt der jeweiligen Lernkurve ab. Wenn die Lernkurve so ausfällt, dass ein noch vor Iteration 20 auftretender Gewichtsvektor selektiert wird, so sind die Prognoseergebnisse schlechter, als wenn das selektierte Netz von einer späteren Iteration stammt.

- Im Vergleich zu ALR sind die Ergebnisse deutlich weniger vom Zufall beeinflusst, sowohl was den MSE als auch die notwendige Länge der Lernkurven betrifft. Insbesondere treten keine ungünstigen Ausreißer auf.

Diese Anwendungen lassen zwar keine umfassende Beurteilung des ES-Ansatzes zu, weisen aber auf die grundsätzliche Schwäche des ES-Ansatzes hin, dass er in mehrerer Hinsicht vom Zufälligkeit geprägt ist. Einerseits ist das Ergebnis jeder

einzelnen Schätzung mehr als bei Modellierungsstrategien auf der Basis von konvergenten Verfahren durch den Zufall beeinflusst. Andererseits gibt es auch zwischen den verschiedenen Anwendungen Unterschiede in der Funktionsweise. Es sei jedoch betont, dass die vorhandene Nicht-Linearität in der Zeitreihe SIM durch den ES-Ansatz in befriedigender Weise berücksichtigt wird, während das Vorhandensein von Nicht-Linearität in ALR und IPI unsicher ist. Der ES-Ansatz eignet sich deshalb kaum für die Erstellung von ungeprüft einsetzbaren Prognosen sondern eher als ergänzende Komponente im Rahmen einer explorativen Phase.

4.8 Ansatz mit Regularisierung

Regularisierung ist eine wichtige und in ihren Prinzipien unmittelbar einleuchtende Methode, das Overfitting-Problem zu vermeiden. Die Komplexität des Modells wird gering gehalten, indem zur Fehlerfunktion des Netzes ein Strafterm hinzugefügt wird, der bewirkt, dass Parameter, die zur Modellierung der Daten wenig beitragen, in Richtung Null gedrängt werden. Es bestehen enge Zusammenhänge zum statistisch-parametrischen Ansatz (vgl. Abschnitt 4.6). Während beim diesem u.a. nicht-signifikante Parameter durch ein Pruning-Verfahren entfernt werden und so die Modellkomplexität reduziert wird, werden solche Parameter im Rahmen des Regularisierungsansatzes bloß eingeschränkt, wobei keine Entweder-Oder-Entscheidung notwendig ist, sondern graduell vorgegangen wird. Ein weiterer wichtiger Zusammenhang besteht zwischen dem Regularisierungsansatz und dem Bayesianischen Ansatz, der eine theoretische Rechtfertigung des Regularisierungsansatzes liefert und diesen in verschiedene Richtungen weiterentwickelt (vgl. Abschnitt 4.9).

Grundlagen

Bei Regularisierung wird der Fehlerfunktion des Netzes ein Strafterm $\nu\Omega$ hinzugefügt:

$$\tilde{E} = E + \nu\Omega, \tag{4.61}$$

wobei ν ein Regularisierungsparameter ist, der angibt, wie stark die Modellkomplexität bestraft wird, und Ω die effektive Modellkomplexität abbildet. Ein Schätzverfahren auf der Basis einer derart erweiterten Fehlerfunktion liefert einen Kompromiss zwischen einer guten Anpassung an die Daten, ausgedrückt durch ein kleines E, und einer geringen Modellkomplexität, die sich als kleiner Strafterm zeigt.

Für die Spezifizierung des Strafterms wurden in der NN-Forschung verschiedene Verfahren entwickelt. Die üblichste Technik definiert den Strafterm mit Hilfe der Summe der quadrierten Parameter des Modells und ist unter dem Begriff „Weight Decay“ bekannt (dt. „Gewichteverfall“, im folgenden als WD abgekürzt). Bishop (1995) gibt einen guten Überblick über diese Technik und ihre Motivation. In der einfachsten Version ist

$$\Omega = \sum_i w_i^2. \tag{4.62}$$

Der Name „Weight Decay“ ist dadurch motiviert, dass in einem Gradientenabstiegsverfahren die Gewichte exponentiell gegen Null gehen, wenn man das Training ohne Daten fortführt. Dies wird aus einer Betrachtung der Ableitungen von $\tilde{E}$ nach den Gewichten $\mathbf{w}$, $\nabla\tilde{E} = \nabla E + \nu\mathbf{w}$, deutlich. Unter Annahme einer kontinuierlichen Verfahrenszeit τ ergibt sich als Lösung für die Entwicklung der Gewichte über die Verfahrenszeit

$$\mathbf{w}^{(\tau)} = \mathbf{w}^{(0)} \exp(-\eta\nu\tau), \tag{4.63}$$

wobei η die Lernrate bezeichnet. Somit werden im Rahmen des regularisierten Trainings die Gewichte permanent in Richtung Null gedrängt und auf diese Weise der effektiv verfügbare Parameterraum eingeengt.

Anders (1997) kritisiert die grundlegende Idee der Regularisierung mit WD. Er vermisst erstens eine Erklärung, warum die Gewichte gerade nach Null gezogen werden sollten, wo sie doch optimalerweise auch weit von Null gelegen sein könnten. Dem ist entgegenzuhalten, dass man eben nicht weiß, wo die Gewichte optimalerweise liegen. Zweifellos jedoch handelt es sich bei dem Verfahren um eine Heuristik. Es wären z.B. Verallgemeinerungen des Verfahrens vorstellbar, bei denen die Gewichte statt nach Null zu einem anderen Wert hin gedrängt werden. Zweitens, so Anders, würde sich die Komplexität des Modells durch die Verwendung des Strafterms nicht ändern. Diese Behauptung ist leicht zu entkräften. Es sei auf den Fall eines prohibitiv großen Strafterms verwiesen, der immer zu einem Nullvektor als Parametervektor führt (die Dimensionalität des effektiven Parameterraums ist null).

Ein wesentlicher Aspekt an der Regularisierung mit WD ist, dass signifikante Gewichte im Vergleich zu weniger signifikanten weniger stark gegen Null gedrängt werden. Dies wird von (Bishop, 1995, S. 340) für den Fall einer quadratischen Fehlerfunktion anhand der im Folgenden kurz zusammengefassten Eigenwertüberlegungen gezeigt. Es sei $\mathbf{H}$ die Hesse-Matrix für die ursprüngliche Fehlerfunktion E. Mit Hilfe der Eigenvektorgleichung

$$\mathbf{H}\mathbf{u}_j = \lambda_j \mathbf{u}_j \tag{4.64}$$

wird nun der Vektorraum gebildet, der von den Eigenvektoren $\mathbf{u}_j$ als Orthonormalvektoren aufgespannt wird. Ein Gewichtsvektor $\mathbf{w}$ kann nun als gewichtete Summe der Eigenvektoren angeschrieben werden,

$$\mathbf{w} = \sum_j u_j \mathbf{u}_j, \tag{4.65}$$

wobei u_j das Gewicht des Parametervektors in Bezug auf den Eigenvektor $\mathbf{u}_j$ bezeichnet. Betrachtet man den Gewichtsvektor im Minimum der ursprünglichen Fehlerfunktion, $\mathbf{w}^*$, und jenen im Minimum der Fehlerfunktion mit Regularisierung, $\tilde{\mathbf{w}}$, so lässt sich der folgende Zusammenhang zeigen:

$$\tilde{u}_j = \frac{\lambda_j}{\lambda_j + \nu} u_j^*. \tag{4.66}$$

Dies bedeutet, dass in jene Richtungen im Vektorraum, für die der entsprechende Eigenwert relativ groß ist und daher $\lambda_j \gg \nu$ gilt, $\tilde{u}_j \simeq u_j^*$ resultiert und das Minimum der Fehlerfunktion relativ wenig verschoben wird. Hingegen ergibt sich für jene Richtungen, für die der entsprechende Eigenwert relativ klein ist, also $\lambda_j \ll \nu$, dass $|\tilde{u}_j| \ll |u_j^*|$, so dass die in Richtung dieser Eigenvektoren liegenden Gewichte stark gegen Null gedrängt werden.

Wie unschwer zu erkennen ist, besteht ein enger Zusammenhang zwischen diesen Überlegungen und jenen, die im Rahmen des statistisch-parametrischen Ansatzes (vgl. Abschnitt 4.6) für die Herleitung der statistischen Signifikanz von Koeffizienten gebraucht werden. Auch dort spielt die Hesse-Matrix eine zentrale Rolle.

Regularisierung mit mehreren Regularisierungsparametern

Eine problematische Einschränkung der Regularisierung mit WD in seiner einfachen Form ist seine mangelnde Skaleninvarianz. Dies bedeutet, dass die Wirkungsweise der Regularisierung nicht unabhängig von einer linearen Transformation der Input- und Zielvariablen bzw. in unserem Falle der Zeitreihe ist. Bei einer linearen Transformation der Daten können die Koeffizienten leicht angepasst werden, damit ein äquivalentes Modell verwirklicht wird, vgl. Gleichung (4.15) in Abschnitt 4.3. Der Regulierungsparameter ν kann hingegen nicht so angepasst werden, dass sich das so angepasste Modell erneut als Schätzergebnis ergibt, zumal es unwahrscheinlich ist, dass es ein $\tilde{\nu} = \lambda\nu$ gibt, bei dem die regularisierte Fehlerfunktion des angepassten Modells ein Minimum annimmt, vorausgesetzt das originale Modell liegt in einem lokalen Minimum. Es kommt also zu einer Verlagerung der regulierenden Wirkung zwischen den verschiedenen Parametergruppen des Modells und einer willkürlichen Bevorzugung bestimmter Bereiche des Parameterraums.

Um diesen Mangel zu beheben, werden für verschiedene Parametergruppen des Modells eigene Regularisierungsparameter eingeführt. In der vorliegenden Arbeit werden drei Regulierungsparameter verwendet, je einer für die erste Schicht und die zweite Schicht des NN-Teils sowie den allenfalls vorhandenen linearen Teil. Somit kann für ein ARNN-Modell die regularisierte Fehlerfunktion folgendermaßen angeschrieben werden:

$$\tilde{E} = SSE + \nu_1 \sum_{k,j} a_{k:j}^2 + \nu_2 \sum_k c_k^2 + \nu_3 \sum_l f_l^2 \tag{4.67}$$

Wie aus dieser Spezifikation ersichtlich ist, werden hingegen die Parameter des deterministischen Teils des Modells sowie die Bias-Parameter keiner Regulierung unterzogen. Als Begründung für eine solche Vorgangsweise findet man in der Literatur das Argument, dass diese Parameter zum nicht-linearen Charakter des Modells wenig beitragen. Diese Begründung ist kaum haltbar, da die Nicht-Linearität durch das Zusammenwirken aller Parameter des Modells zustandekommt. Letztlich ist der gewählte Zugang das Ergebnis pragmatischer Überlegungen, die durchaus einer weiteren Prüfung unterzogen werden sollten.

Verfahren zur Bestimmung der Regularisierungsparameter

In der praktischen Umsetzung des Regulierungsansatzes steht man vor dem Problem der Wahl der Regulierungsparameter $\boldsymbol{\nu} = (\nu_1, \nu_2, \nu_3)'$. Die ad-hoc-Vorgangsweise ist, die Regulierungsparameter solange anzupassen und das Modell erneut zu schätzen, bis der Regularisierungsterm einen entsprechend starken Einfluss auf die gesamte Fehlerfunktion ausübt, so dass der Anwender vermutet, dass das „richtige“ Maß an effektiver Komplexität des Parameterraums verwirklicht ist. Ein wissenschaftlicher Zugang und vor allem der Bedarf nach Automatisierung im Rahmen von benutzerfreundlicher Software verlangen jedoch eine Präzisierung der Vorgangsweise. Im folgenden werden einige Heuristiken für die Bestimmung von $\boldsymbol{\nu}$ entwickelt und ihre Funktionsweise anhand von Beispielen analysiert.

Die gesamte Vorgangsweise ist in drei Phasen gegliedert:

Phase 1 Bestimmung von brauchbaren Anfangswerten $\boldsymbol{\nu}^{(1)}$

Phase 2 Bestimmung von verbesserten Werten anhand proportionaler Skalierung der Anfangswerte: $\boldsymbol{\nu}^{(2)} = \chi\boldsymbol{\nu}^{(1)}$

Phase 3 Feinbestimmung der Regulierungsparameter durch Austestung von mehreren Kandidaten in der Umgebung von $\boldsymbol{\nu}^{(2)}$ mit Hilfe von Kreuzvalidierung: $\boldsymbol{\nu}^{(3)} = (\chi_1\nu_1^{(2)}, \chi_2\nu_2^{(2)}, \chi_3\nu_3^{(2)})'$.

In *Phase 1* geht es darum, mit möglichst wenig Rechenaufwand gute Anfangswerte für die Regulierungsparamter zu „erraten“. Hierbei soll sowohl eine „spürbare“ regulierende Wirkung insgesamt erzielt werden, als auch das Verhältnis der Regulierungsparameter untereinander so gewählt werden, so dass von jedem der drei Regulierungsparameter eine ähnlich starke regulierende Wirkung ausgeübt wird.

Für die folgende Darstellung des Algorithmus sei angenommen, dass das zu regularisierende Modell abgesehen von einem Interzept keinen deterministischen Teil aufweist:[28]

(i) Als erster Schritt wird ein lineares Modell M_{Lin} geschätzt, das die gleiche Lag-Struktur aufweist, wie das Modell, das Gegenstand des regularisierten Trainings ist. Der Vektor der autoregressiven Koeffizienten von M_{Lin} sei mit $\mathbf{f}_{Lin}$, der Fehler mit SSE_{Lin} und die Standardabweichung der geschätzten Werte mit $\hat{\sigma}_{Lin}$ bezeichnet. Dieses Modell dient als Orientierung für die Bestimmung der Regularisierungsparameter.

(ii) Auf der Basis von SSE_{Lin} wird nun der Wert, den der Regularisierungsterm annehmen soll, als $R = \xi SSE_{Lin}$ vorgegeben, wobei ξ vom Anwender zu wählen

[28] Eine allgemeine Darstellung, die auch Modelle mit saisonalen Mittelwerten und einen Trend berücksichtigt, würde zusätzliche Notation erfordern, die an dieser Stelle zuviel Platz einnehmen würde. Der interessierte Leser sei auf den Quellcode der Funktion `init.regul` im R-Paket `NNUTS` verwiesen.

ist. Aufgrund der Experimente mit den Beispielzeitreihen liefern Werte zwischen 0.02 und 0.1 brauchbare Ergebnisse.

(iii) Dann wird ein ARNN-Modell M_{NN} als Annäherung an M_{Lin} bestimmt, das die gewünschte Modellstruktur aufweist, so dass die Funktionalität von M_{Lin} gleichmäßig auf die verdeckten Neuronen und den allenfalls vorhandenen linearen Teil verteilt ist. Wenn als Aktivierungsfunktion in M_{NN} statt dem Tangens hyperbolicus die Identitätsfunktion gewählt würde, so wäre M_{NN} eine exakte Replikation von M_{Lin}. Da es sich jedoch um ein ARNN-Modell handelt und die Gewichte so gewählt werden, dass die Aktivierungsfunktion auch ober- und unterhalb ihres linearen Bereichs abgedeckt wird, weicht M_{NN} leicht von M_{Lin} ab. Dies wird erreicht, indem die Parameter folgendermaßen gewählt werden:

$$\mathbf{f} = \mathbf{f}_{Lin}/(K+1), \quad c_k = 2s_{Lin}/(K+m_{\mathbf{f}}), \quad \mathbf{a}_k = \mathbf{f}_{Lin}/2s_{Lin}. \tag{4.68}$$

Hierbei sind $\mathbf{a}_k$ und c_k die Gewichte in der ersten und zweiten Schicht des k-ten verdeckten Neurons. Da alle Neuronen gleich initialisiert werden, genügt es im folgenden $\mathbf{a}$ und c zu schreiben. $\mathbf{f}$ bezeichnet die Gewichte des linearen Teils von M_{NN} („Shortcuts“), falls vorhanden. Weiters ist K die Anzahl der verdeckten Neuronen und $m_{\mathbf{f}}$ eine Indikatorvariable, die angibt, ob M_{NN} einen linearen Teil enthält oder nicht.

(iv) Nun wird das Verhältnis der Regulierungsparameter für die erste und zweite Schicht, $r = \nu_1/\nu_2$, im Rahmen einer Optimierung bestimmt. Es werden die folgenden Überlegungen angestellt, die zu einem einfachen Minimierungsproblem führen. Das Ziel, dass von ν_1 und ν_2 ungefähr die gleiche regularisierende Wirkung ausgeht, bedeutet, dass es im Zuge des Trainings nur schwer möglich sein soll, dass dem Regulierungsdruck durch umgekehrt proportionale Veränderung der Gewichte der ersten und der zweiten Schicht ausgewichen wird. In anderen Worten, es soll bei gegebenen Gewichten für die schließlich gefundenen ν_1 und ν_2 keinen Skalierungsparameter ω geben, so dass der Regularisierungsterm mit alternativen Regularisierungsparametern $\tilde{\nu}_1 = \omega\nu_1$ und $\tilde{\nu}_2 = (1/\omega)\nu_2$ geringer ist. Man definiere die Funktion

$$g(\omega) = \sum_{j\in\mathcal{J}} r(\omega a_j)^2 + (\frac{1}{\omega}c)^2. \tag{4.69}$$

Man beachte, dass ω eine gegengleiche Veränderung der Regularisierungsparameter bewirkt. Es sei daran erinnert, dass bei der Gültigkeit der linearen Annäherung ein verdecktes Neuron mit Gewichten $\tilde{\mathbf{a}} = \omega\mathbf{a}$ und $\tilde{c} = (1/\omega)c$ die gleiche Funktion implementiert wie jenes mit den Gewichten $\mathbf{a}$ und c. Auf diese Weise gibt g die Abhängigkeit des einem bestimmten verdeckten Neuron zugeordneten Regularisierungsterms von der gegengleichen Veränderung der

Gewichte der ersten und zweiten Schicht an. Das oben erklärte Ziel ist dann erreicht, wenn r so gewählt wird, dass die Funktion g bei $\omega = 1$ ein Minimum hat. Also setze man $g'(\omega) = 0$, wobei g' die Ableitung der Funktion bezeichnet und erhält

$$r = c^2/\mathbf{a}'\mathbf{a}. \tag{4.70}$$

(v) Die Regularisierungsparameter erhält man durch Einsetzen von $\nu_1 = r\nu_2$ und der Vorgabe, dass der Regularisierungsterm für jedes verdeckte Neuron bzw. für den linearen Teil $\tilde{R} = R/(K+m_{\mathbf{f}})$ betragen soll. Die Lösung für den Vektor der Regularisierungsparameter der Phase 1, $\boldsymbol{\nu}^{(1)} = (\nu_1^{(1)}, \nu_2^{(1)}, \nu_3^{(1)})'$, lautet:

$$\nu_1^{(1)} = \tilde{R}/2\sum_{j\in\mathcal{J}} a_j^2, \quad \nu_2^{(1)} = \tilde{R}/2c^2, \quad \nu_3^{(1)} = \tilde{R}/\sum_{j\in\mathcal{J}} f_j^2 \tag{4.71}$$

Die *Phase 2* dient dazu, die Wahl der Regularisierungsparameter durch Skalierung zu verbessern, wobei im Vergleich zur Phase 1 stärker auf die Besonderheiten und Struktur der Daten eingegangen wird und der Rechenaufwand höher ist. Die im Folgenden vorgeschlagene Vorgangsweise soll den Vektor der Regulierungsparameter $\boldsymbol{\nu}^{(2)} = \chi\boldsymbol{\nu}^{(1)}$ solcherart bestimmen, dass das mit Hilfe von Regularisierung geschätzte Modell einen SSE innerhalb bestimmter vorgegebener Grenzen aufweist. Das Verfahren ist iterativ und kann in mehrere Schritte gegliedert werden:

(i) Der Anwender gibt gewünschte obere und untere Schranken SSE_1 und SSE_2 vor, zwischen denen der SSE des regularisierten Modells zu liegen kommen soll. In diese Vorgaben geht die Erwartung des Anwenders ein, welche Steigerung der Prognosegüte durch Berücksichtigung der Nicht-Linearität zu holen ist und wie groß die Gefahr des Overfittings bei der gegebenen Modellarchitektur ist. Ist beispielsweise der nicht-lineare Anteil gering und die Gefahr des Overfittings hoch, so ist die Orientierung am linearen Modell naheliegend. Eine geeignete Vorgabe ist möglicherweise $SSE_1 = 0.99SSE_{Lin}$ und $SSE_2 = 0.95SSE_{Lin}$.

(ii) Nun wird ein Modell in der gewünschten Modellspezifikation geschätzt, wobei entweder keine oder eine Regularisierung mit Hilfe $\boldsymbol{\nu}^{(1)}$ verwendet werden kann. Die Parameter dieses Modells werden mit $\dot{\mathbf{w}}$ bezeichnet.

(iii) Der Skalierungsparameter χ wird mit Eins initialisiert.

(iv) Mit Regularisierung unter Verwendung von $\boldsymbol{\nu}^{(2)} = \chi\boldsymbol{\nu}^{(1)}$ und mit Initialisierung der Parameter auf $\dot{\mathbf{w}}$ wird eine Schätzung des Modells durchgeführt und der SSE, SSE_{Reg}, ermittelt.

(v) SSE_{Reg} wird mit den vorgegebenen Schranken verglichen und χ bzw. $\boldsymbol{\nu}^{(2)}$ entsprechend angepasst. Wenn $SSE_{Reg} \leq SSE_2$, so ist die Regularisierung zu

schwach und muss verstärkt werden, indem χ mit einer Zahl zwischen 1 und 2 multipliziert wird. Wenn hingegen $SSE_{Reg} \geq SSE_1$, so muss χ durch diese Zahl dividiert werden. Mit dem so angepassten χ geht man zurück zu Schritt (iv). Wenn $SSE_2 < SSE_{Reg} < SSE_1$, so ist keine Anpassung notwendig und das aktuelle $\boldsymbol{\nu}^{(2)} = \chi\boldsymbol{\nu}^{(1)}$ ist der Finalwert. Bei der iterativen Suche nach einem passenden Finalwert sind verschiedene Vorkehrungen zu beachten, damit die Konvergenz des Verfahrens unterstützt und gleichzeitig seine Effizienz sichergestellt werden. Weiters sollte die Anzahl der Iterationen begrenzt werden. Im Falle, dass bei der vorgegebenen maximalen Anzahl von Iterationen der Fall $SSE_2 < SSE_{Reg} < SSE_1$ nicht eintritt, wird dennoch das aktuelle $\boldsymbol{\nu}^{(2)}$ als Finalwert genommen, da es wahrscheinlich eine Regularisierung repräsentiert, die nahe am Zielbereich liegt.

In *Phase 3* kann nun eine Feinabstimmung der Regularisierungsparameter vorgenommen werden, wobei sowohl die Skalierung als auch das Verhältnis der Regularisierungsparameter untereinander anhand von Kreuzvalidierung ausgetestet wird. Im Gegensatz zu Phase 1 und 2 findet damit eine Überprüfung der tatsächlichen Generalisierungsfähigkeit des Modells statt. Die Vorgangsweise kann wie folgt skizziert werden:

(i) Die gesamte für die Schätzung zur Verfügung stehende Zeitreihe wird in zwei Bereiche unterteilt. Der erste Teil („Schätzset“) ist für die Schätzung vorgesehen, der zweite Teil („Testset“), der typischerweise 20 % der Gesamtlänge umfasst, ist für die Testung vorgesehen.

(ii) Es wird eine Menge von auszutestenden Vektoren $\boldsymbol{\nu}^{(3)} = (\nu_1^{(3)}, \nu_2^{(3)}, \nu_3^{(3)}) = (\chi_1\nu_1^{(2)}, \chi_2\nu_2^{(2)}, \chi_3\nu_3^{(2)})$ erstellt. Die Werte von $\nu_i, i = 1, 2, 3$ liegen hierbei sowohl ober als auch unterhalb der in Phase 2 gefundenen Werte und liegen um einen vom Anwender zu wählenden Faktor, z.B. 1.5, auseinander. Die Menge soll für jedes ν_i eine ausreichende Variation enthalten, darf aber nicht zu umfangreich werden. Werden beispielsweise von jedem Regulierungsparameter fünf alternative Werte für die Austestung vorgesehen, so wäre die Anzahl aller Kombinationen $5^3 = 125$. In solchen Fällen wäre eine reduzierte Auswahl von Kombinationen zu treffen, um den Rechenaufwand gering zu halten.

(iii) Man wähle eine Initialisierung für den Parametervektor $\dot{\mathbf{w}}$. Alternativ kann dieser Schritt auch eine Voroptimierung enthalten, indem $\dot{\mathbf{w}}$ aus der Schätzung eines Modells mit Regularisierung unter Verwendung von $\boldsymbol{\nu}^{(2)}$ als Regularisierungsparameter ermittelt wird.

(iv) Für jedes Element der Menge der auszutestenden $\boldsymbol{\nu}^{(3)}$ wird das entsprechend regularisierte Modell geschätzt und der SSE auf dem Testset ermittelt.

(v) Als Finalwert wird jenes $\boldsymbol{\nu}^{(3)}$ genommen, für das der SSE auf dem Testset das Minimum erreicht. Um zufällige Effekte zu vermeiden kann man alternativ

den Finalwert für $\boldsymbol{\nu}^{(3)}$ als gewichteten Durchschnitt einer kleinen Anzahl der besten Vektoren (gemessen am SSE, der damit auf dem Testset erreicht wurde) bestimmen.

Es sollte erwähnt werden, dass die Methode der Kreuzvalidierung in Phase 3 *nicht* impliziert, dass die Datenpunkte des Testsets nicht für die Schätzung des endgültigen Modells verwendet werden können. Die Aufteilung der Daten in ein Schätzset und ein Testset wird nur für die Ermittlung der Regularisierungsparameter benötigt und kann aufgehoben werden, sobald diese vorliegen.

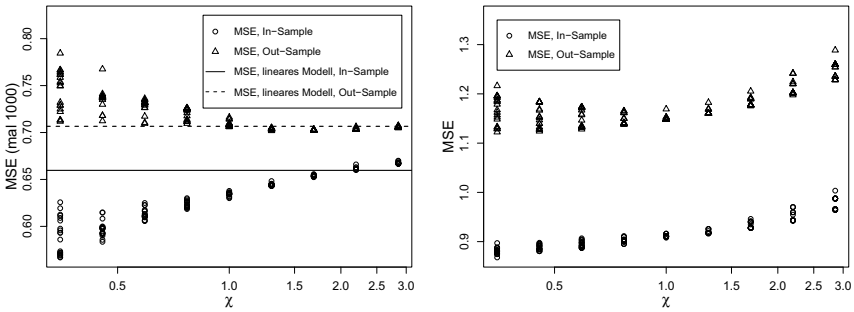
Bei allen Verfahren der Phase 1, 2 und 3 sind zusätzliche Modifikationen und Verfeinerungen denkbar bzw. wurden teilweise im R-Paket `NNUTS` implementiert. Bei den möglichen Modifikationen und bei der Wahl der Einstellungen der oben beschriebenen Verfahren sind insbesondere der Rechenaufwand und die Abhängigkeit von numerischen Aspekten zu beachten, wobei ein sinnvoller Ausgleich von Fall zu Fall gefunden werden muss.

Anwendung

Die Funktionsweise des Regularisierungsansatzes soll im Folgenden anhand zweier Anwendungsbeispiele demonstriert werden. Es wird auf der Zeitreihe des Industrieproduktionsindex (saisonale Differenzen) und einer simulierten Zeitreihe SIM nach Formel (4.11) eine Simulation durchgeführt. Für beide Zeitreihen wird die gesamte Zeitreihe in ein Lernset („In-Sample-Set“) und ein Evaluierungsset („Out-of-Sample-Set“) unterteilt. Die Regularisierungsparameter wurden zunächst mit dem oben beschriebenen Ansatz aus Phase 1 bestimmt und dann skaliert, so dass der Vektor der Regularisierungsparameters als $\tilde{\boldsymbol{\nu}} = \chi\boldsymbol{\nu}$ geschrieben werden kann und χ der Skalierungsparameter ist. Insgesamt wurden so neun verschiedene Regularisierungsparametervektoren erstellt und ausgetestet. Für jede Wahl der Regularisierungsparameter wurden 20 verschiedene Schätzungen eines ARNN-Modells durchgeführt. Hierbei wurde die Anzahl der verdeckten Neuronen mit 5 angesetzt. Dies läuft auf eine Überparametrisierung des Modells hinaus, doch der Regularisierungsansatz sollte in der Lage sein, die daraus resultierende Gefahr des Overfittings in den Griff zu bekommen.

In Abb. 13 werden jeweils die MSE-Werte für das Lernset und das Evaluierungsset abgetragen. Aus dem Verlauf der Punktwolken lässt sich die Abhängigkeit der Prognosegüte und der Streuung bzw. Verlässlichkeit der Ergebnisse von der Wahl der Regularisierungsparameter ablesen.

In der linken Grafik der Abbildung wird das Funktionieren des Regulierungsansatzes anhand des Industrieproduktionsindex illustriert. Bei dieser Zeitreihe ist die Erwartung, dass ein nicht-lineares Modell eine, wenn überhaupt, nur wenig bessere Prognoseleistung als das rein lineare Modell bringt. Daher kann sowohl in Phase 1 als auch Phase 2 der oben beschriebenen Vorgangsweise der Bestimmung der Regularisierungsparameter eine Orientierung am linearen Modell erfolgen. Die



(a) IPI (b) SIM

Abb. 13: Ergebnisse mit dem Regularisierungsansatz in Abhängigkeit von der Wahl der Regularisierungsparameter: (a) Industrieproduktionsindex (IPI) und (b) Simulierte Zeitreihe (SIM)

MSE-Werte für das lineare Modell wurden als horizontale Linien sowohl für das Lernset als auch das Evaluierungsset eingezeichnet.

Die Anfangs-Regularisierungsparameter (Phase 1) wurden mit Hilfe von $\xi = 0.08$ gewählt, d.h. so dass der Regularisierungsterm gemessen am SSE des (ohne Regularisierung geschätzten) linearen Modells 8 % ausmacht. Für diese Wahl der Regularisierungsparameter, welche in der Graphik in der Mitte der X-Achse liegt, $\chi = 1$, ergibt sich ein MSE auf dem In-Sample-Set, der etwa um 5 % niedriger ist als der Benchmarkwert für das lineare Modell. Ganz klar ist zu erkennen, dass es unterhalb dieses χ-Werts zu Overfitting kommt, indem in der Graphik nach links hin die In-Sample-Ergebnisse besser werden, während sich die Out-of-Sample-Ergebnisse verschlechtern. Bei den initialen Regularisierungsparameterwerten ($\chi = 1$) hingegen dürfte die Regularisierung nur wenig zu schwach sein, um ein Overfitting zu verhindern. Der minimale MSE-Wert auf dem Out-of-Sample-Set wird etwa bei $\chi = 1.5$ erreicht. Der Umstand, dass im ungefähren Minimum für das Evaluierungsset der MSE-Wert für das Lernset kaum unterhalb des linearen Modells liegt, kann als Hinweis gesehen werden, dass kaum nicht-lineare Struktur in den Daten vorgefunden wird. Doch auch andere Interpretationen sind nicht auszuschließen. So könnte bei diesem Wert die nicht-lineare Struktur in den Daten sehr wohl im Modell berücksichtigt worden sein und sowohl die lineare als auch der nicht-lineare Teil des Modells durch die Regularisierung in seiner effektiven Komplexität eingeschränkt worden sein. Ein weiteres wichtiges Detail in Abb. 13 a) ist die gleichzeitig mit dem Overfitting auftretende höhere Streuung der ARNN-Schätzergebnisse. Hingegen, zeigt sich bei χ-Werten ab ca. 1.5, dass die 20 Durchläufe jeweils sehr ähnliche Ergebnisse liefern, d.h. dass immer nahezu das gleiche lokale Minimum der Fehlerfunktion gefunden wird. Regularisierung kann daher auch als ein wichtiges

Instrument gesehen werden, um die Zufälligkeit des Schätzverfahrens bei ARNN-Modellierung geringzuhalten.

Die rechte Graphik in Abb. 13 veranschaulicht die Ergebnisse der Simulation auf der simulierten Zeitreihe SIM. Der stochastische Prozess, der diese Zeitreihe generiert, hat eine Varianz von 1.858. Der Fehlerterm des stochastischen Prozesses hat eine Varianz von Eins. Das wahre Modell könnte somit (in großen Stichproben) eine mittlere Varianz von 0.858 erklären. Von dieser mittleren Varianz können (bei genügend großen Stichproben) 0.173 durch ein lineares Modell erklärt werden und 0.685 durch ein nicht-lineares Modell. Bedenkt man diese Struktur in der Zeitreihe, so zeigt die Graphik auf, dass alle in der Simulation in Betracht gezogenen Modelle einen Großteil der erklärbaren Varianz tatsächlich erklären und für die Prognosequalität nützen können. Im In-Sample-Set ist der MSE der Modelle durchwegs deutlich unter Eins, d.h. die ARNN-Modelle sind auch in der Lage, einen Teil der Zufallskomponente der Zeitreihe zu „lernen“. Die optimale Generalisierungsfähigkeit ist etwa bei einem Wert von $\chi = 1$ angesiedelt, der die Wahl der Regularisierungsparameter gemäß dem oben geschilderten Verfahren (Phase 1) mit einem Wert von $\xi = 0.01$ repräsentiert. Auch in dieser Simulation zeigt sich, dass sich eine zu geringe Regularisierung durch eine hohe Streuung der Ergebnisse über die 20 Durchläufe hinweg auswirkt.

Insgesamt zeigt diese Simulation die Wichtigkeit der Wahl der Regularisierungsparameter für das Funktionieren des Regularisierungsansatzes auf. In Phase 1 ist die Wahl des ξ-Wertes wichtig. Jedoch kann in den meisten Anwendungsfällen eine zu kleine oder zu große Wahl durch wenige Iterationen im Rahmen der Phase 2 korrigiert werden. So liefert die Phase 2 im Falle des Industrieproduktionsindex in etwa zwei bis vier Iterationen jene Regularisierungsparameter, die die gewünschte Regularisierungswirkung im Sinne einer vorgegebenen Verbesserung des *SSE* gegenüber dem linearen Modell ausüben und auch im Evaluierungsset die optimale Generalisierungsfähigkeit erzielen. Im Falle der simulierten Zeitreihe SIM mit einem beträchtlichen nicht-linearen Anteil ist allerdings die Orientierung am linearen Modell, um die Zielvorgaben für die Stärke der Regulierungswirkung in Phase 2 zu bestimmen, nicht sinnvoll. Die sinnvolle Wahl der verschiedenen Metaparameter der Phase 1 und 2 der Bestimmung der Regularisierungsparameter erfordert bereits eine ungefähre Vorstellung, wieviel Gewinn an Prognosegüte durch ein ARNN-Modell im Vergleich zu einem linearen Modell zu holen ist.

Die Funktionsweise der Phase 3 wurde ebenso in einer Simulation anhand der beiden Beispielzeitreihen getestet, deren Ergebnisse jedoch nicht in handlicher Form aufbereitet werden können. Es kann weder für die Zeitreihe IPI noch für SIM gesagt werden, dass das in Phase 3 zur Anwendung kommende Verfahren der Bestimmung der Regularisierungsparameter mit Hilfe von Kreuzvalidierung bessere Regularisierungsparamtersets findet als die Verfahren der Phase 1 und 2 alleine. Da das Verfahren der Phase 3 besonders rechenintensiv ist, kann daher von seinem Einsatz in ähnlichen Anwendungsgebieten abgeraten werden.

Diese Anwendung belegt das gute Funktionieren des Regularisierungsansatzes für das Anwendungsgebiete der Prognose makroökonomischer Zeitreihen, vorausgesetzt, es existiert genügend nicht-lineare Struktur in den Zeitreihen. Neben dem Overfitting-Problem gelingt es mit diesem Ansatz auch die Zufälligkeit des Schätzverfahrens eines ARNN-Modells in den Griff zu bekommen. Die vorgeschlagenen Verfahren zur Bestimmung der Regularisierungsparameter bewähren sich in der Praxis, wobei bereits einfach gehaltene Verfahren ausreichen und von umfangreichen Bestimmungsverfahren auf der Basis von Kreuzvalidierung Abstand genommen werden sollte.

4.9 Bayesianischer Ansatz

Bayesianische Sichtweisen finden in der Statistik und Ökonometrie immer mehr Beachtung und dienen insbesondere oft als Rahmenwerk, das verschiedene theoretische Ansätze zu vereinen vermag. So lassen sich für viele der im Zusammenhang mit der Modellierung mit neuronalen Netzen auftauchenden Probleme mit Hilfe bayesianischer Ansätze Lösungen finden. Bishop (1995), dem der Bayesianismus ein besonderes Anliegen ist, führt eine Reihe von Anwendungen für Bayesianische Techniken an, z.B. den Vergleich von alternativen Modellen ohne die Zuhilfenahme von Out-of-Sample-Daten.

Im folgenden soll bei der Besprechung des Bayesianischen Ansatzes vor allem auf zwei Aspekte fokussiert werden, die im Zusammenhang mit der Bestimmung der Parameter eines ARNN-Modells relevant sind: erstens die Interpretation der Regularisierung mit „Weight Decay“ (WD) im Lichte des Bayesianismus und zweitens die Wahl des Regularisierungsparameters ν im Rahmen eines integrierten Bayesianischen Verfahrens, dem sogenannten „Evidence Framework“ (MacKay, 1992a,b), der im folgenden auf deutsch als Bayesianischer Evidenzansatz bezeichnet und mit BEA abgekürzt wird. Während die Darstellung des erstgenannten Aspekts die Gelegenheit gibt, die spezifischen Begriffe und die Notation des Bayesianischen Ansatzes einzuführen, ist die Darstellung des BEA auf die Implementation im Rahmen der vorliegenden empirischen Untersuchung hin orientiert. Die Darstellung folgt eng Bishop (1995), wobei für alternative Überblicksdarstellungen auch das neuere Lehrbuch von Bishop (2006), sowie die Fachartikel von MacKay (1995), Penny und Roberts (1999) und Titterington (2004) hilfreich sind.

Regularisierung aus Bayesianischer Sicht

Bei der Bestimmung der Gewichte ist die Bayesianische Sichtweise prinzipiell an der gesamten bedingten Verteilung der Gewichte $p(\mathbf{w}|\mathcal{D})$ interessiert, wobei $\mathcal{D}$ für die Daten steht. Der Startpunkt der bayesianischen Überlegungen ist immer das Bayestheorem, das bezogen auf das Problem der Bestimmung der Verteilung der

Gewichte folgende Form annimmt:

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}. \tag{4.72}$$

Gemäß dem Bayestheorem wird zunächst eine prior-Verteilung $p(\mathbf{w})$ definiert, in die möglicherweise das Vorwissen und die Erwartungen des Analytikers über die Gewichte einfließen. Aus dieser wird sodann nach Beobachtung der Daten $\mathcal{D}$ die posterior-Verteilung $p(\mathbf{w}|\mathcal{D})$ hergeleitet, wobei eine Verknüpfung mit der bedingten Wahrscheinlichkeit der Daten (Likelihood-Funktion) $p(\mathcal{D}|\mathbf{w})$ erfolgt.

Aus der Wahl einer geeigneten Spezifikation für $p(\mathcal{D}|\mathbf{w})$ und $p(\mathbf{w})$ ergibt sich nun eine Spezifikation von $p(\mathbf{w}|\mathcal{D})$. Für $p(\mathbf{w})$ ist die Annahme einer Gaußverteilung der Gewichte um den Mittelwert Null naheliegend, wobei die Gewichte gleiche Varianz aufweisen:

$$p(\mathbf{w}) = \frac{1}{Z_W(\alpha)} \exp(-\alpha E_W), \tag{4.73}$$

wobei

$$E_W = \frac{1}{2}\sum_{i=1}^{W} w_i^2 \quad \text{und} \quad Z_W(\alpha) = \left(\frac{2\pi}{\alpha}\right)^{W/2} \tag{4.74}$$

und W die Anzahl der Gewichte ist.

In ähnlicher Weise nimmt man auch für die Verteilung der Zielvariable[29] um ihren Erwartungswert eine Gaußverteilung an, sodass sich schreiben lässt:

$$p(\mathcal{D}|\mathbf{w}) = \frac{1}{Z_{\mathcal{D}}(\beta)} \exp(-\beta E_{\mathcal{D}}), \tag{4.75}$$

wobei

$$E_{\mathcal{D}} = \frac{1}{2}\sum_{n=1}^{N} \big(\hat{y}_n(\mathbf{w}, \mathbf{x}_n) - y_n\big)^2 \quad \text{und} \quad Z_{\mathcal{D}}(\beta) = \left(\frac{2\pi}{\beta}\right)^{N/2} \tag{4.76}$$

und N die Anzahl der Datenpunkte ist.

Aus der Zusammenführung der Gleichungen (4.72–4.76) erhält man

$$p(\mathbf{w}|\mathcal{D}) = \frac{1}{Z_S} \exp(-\beta E_{\mathcal{D}} - \alpha E_W) = \frac{1}{Z_S} \exp(-S_W), \tag{4.77}$$

wobei

$$S_W = \frac{\beta}{2}\sum_{n=1}^{N} (\hat{y}_n(w, \mathbf{x}_n) - y_n)^2 + \frac{\alpha}{2}\sum_{i=1}^{W} w_i^2 \tag{4.78}$$

und

$$Z_S = \int \exp(-\beta E_{\mathcal{D}} - \alpha E_W) d\mathbf{w}. \tag{4.79}$$

[29] Es sei erwähnt, dass in diesem Ansatz die Verteilung der unabhängigen Daten $\mathbf{x}$ nicht modelliert wird, sondern diese als fixiert angenommen werden. Somit umfasst $\mathcal{D}$ nur die zu erklärende Variable y.

In dieser Spezifikation kommt den Parametern α und β die Rolle zu, die Varianz der Gewichte bzw. der Zielvariable zu repräsentieren. Da sie die Verteilung von anderen Parametern bestimmen, werden sie Hyperparameter genannt. Zunächst wird angenommen, dass sie bekannt sind.

Es zeigt sich, dass sich aus dieser Spezifikation unmittelbar die Methode der Regularisierung mit WD ableiten lässt, wenn man für $p(\mathbf{w}|\mathcal{D})$ den wahrscheinlichsten Wert $\mathbf{w}_{\mathrm{MP}}$ ermittelt. Dies läuft auf die Minimierung der Fehlerfunktion S_W hinaus, da Z_S als bloßer Normierungsfaktor vom Gewichtsvektor unabhängig ist. Abgesehen von einem multiplikativen Faktor entspricht dies genau der Regularisierung mit WD gemäß Gleichung (4.62) und (4.63), wobei der Regularisierungsparameter als $\nu = \alpha/\beta$ bestimmt wird.

Bereits im Abschnitt 4.8 wurde auf die mangelnde Skaleninvarianz von Regularisierung mit WD bei Verwendung von bloß einem Regularisierungsparameter hingewiesen. Als Problemlösung bot sich eine Verallgemeinerung des Regularisierungsterms mit unterschiedlichen Regularisierungsparametern für verschiedene Gruppen von Gewichten an, vgl. Gleichung (4.67). Auch für den bayesianischen Ansatz der Regularisierung ist die entsprechende Verallgemeinerung naheliegend und spiegelt wider, dass für verschiedene Gruppen von Gewichten $\mathbf{w}_m$, $m = 1, \ldots, M$ unterschiedliche prior-Verteilungen angenommen werden. Statt αE_W wird in den obigen Formeln dann die Verallgemeinerung

$$\sum_m \alpha_m E_{W_m} = \sum_m \frac{\alpha_m}{2} \sum_{i \in W_m} w_i^2 \tag{4.80}$$

eingesetzt. Es sei jedoch erwähnt, dass wenn nicht alle Gewichte bzw. Koeffizienten eines Modells durch prior-Verteilungen berücksichtigt werden, es sich um eine unechte (engl. improper) prior-Verteilung handelt, d.h. sie kann nicht normiert werden. Für die im folgenden dargestellte Anwendung des bayesianischen Ansatzes auf die Bestimmung der Regulierungsparameter ist dies nicht problematisch. Hingegen ergeben sich für andere Anwendungen (z.B. Modellvergleich) Einschränkungen daraus.

Der Bayesianische Evidenzansatz

Bisher wurde angenommen, dass die Hyperparameter α und β bekannt sind. Es ist jedoch ein zentrales Problem bei regularisiertem Lernen, dass die Hyperparameter nicht bekannt sind. Der Bayesianische Zugang bietet mehrere Möglichkeiten an, mit dieser Situation zurecht zu kommen. Zunächst ist der geradlinige Weg über die Hyperparameter auszuintegrieren:

$$p(\mathbf{w}|\mathcal{D}) = \iint p(\mathbf{w}, \alpha, \beta|\mathcal{D})\, d\alpha\, d\beta = \iint p(\mathbf{w}|\alpha, \beta, \mathcal{D}) p(\alpha, \beta|\mathcal{D})\, d\alpha\, d\beta. \tag{4.81}$$

Dieser Ansatz mag zwar als der analytisch exakte Weg betrachtet werden, ist aber nur zum Preis von Vereinfachungen möglich, die seine Brauchbarkeit in prakti-

schen Situationen einschränken. MacKay (1996) favorisiert daher den im folgenden beschriebenen Ansatz des BEA , als dessen zentrale Bausteine die Berechnung der Anzahl der wohl-determinierten Parameter γ und eine iterative Optimierung der Hyperparameter α und β gelten können.

Der BEA wurde von MacKay (1992a,b) in die NN-Forschung eingeführt und beruht auf Arbeiten von Gull (1988, 1989) und Skilling (1991) und ist rechnerisch äquivalent zur sogenannten Typ II Maximum Likelihood Methode (Berger, 1985).

Im BEA werden die Hyperparameter α und β so gewählt, dass die Evidenz (engl. evidence) für α und β, $p(\mathcal{D}|\alpha,\beta)$, maximiert wird. Zu dieser Vorschrift gelangt man durch die beiden folgenden Annahmen bzw. Schritte. Erstens wird angenommen, dass die posterior-Verteilung $p(\alpha,\beta|\mathcal{D})$ eine Spitze bei ihrem wahrscheinlichsten Wert $(\alpha_{\mathrm{MP}},\beta_{\mathrm{MP}})$ hat, die den Großteil des Volumens der Verteilung enthält. Mit dieser Annahme lässt sich die folgende Annäherung rechtfertigen:

$$p(\mathbf{w}|\mathcal{D}) \simeq p(\mathbf{w}|\alpha_{\mathrm{MP}},\beta_{\mathrm{MP}},\mathcal{D}) \iint p(\alpha,\beta|\mathcal{D})\,d\alpha\,d\beta = p(\mathbf{w}|\alpha_{\mathrm{MP}},\beta_{\mathrm{MP}},\mathcal{D}). \tag{4.82}$$

Dies besagt, dass man die wahrscheinlichsten Werte für α und β finden sollte und den Rest der Analyse mit diesen Hyperparameterwerten durchführt.

Zweitens nimmt man für α und β eine sogenannte nicht-informative prior-Verteilung an, d.h. eine solche, die allen Werten von α und β weitgehend die gleiche Wahrscheinlichkeit einräumt. Mit dieser Annahme ergibt sich, dass das Maximum der posterior-Verteilung

$$p(\alpha,\beta|\mathcal{D}) = \frac{p(\mathcal{D}|\alpha,\beta)p(\alpha,\beta)}{p(\mathcal{D})} \tag{4.83}$$

hauptsächlich vom Term $p(\mathcal{D}|\alpha,\beta)$, also der Evidenz für α und β bestimmt wird.

Die Bayesianische Analyse geht nun hierarchisch vor. Auf der ersten Ebene wird die Verteilung der Gewichte unter Zugrundelegung von (4.82) bestimmt, auf der zweiten Ebene die Verteilung der Hyperparameter, wobei die Evidenz für diese maximiert wird. Unter Explizitmachung der Abhängigkeit der Evidenz für α und β von der auf der ersten Ebene des Verfahrens gefundenen Verteilung der Gewichte kann die Evidenz als

$$p(\mathcal{D}|\alpha,\beta) = \int p(\mathcal{D}|\mathbf{w},\beta)p(\mathbf{w}|\alpha)\,d\mathbf{w} \tag{4.84}$$

angeschrieben und durch Einsetzen der Ergebnisse aus den Gleichungen (4.73, 4.75, 4.79) zu

$$p(\mathcal{D}|\alpha,\beta) = \frac{Z_S(\alpha,\beta)}{Z_{\mathcal{D}}(\beta)Z_W(\alpha)} \tag{4.85}$$

umgeformt werden.

Schließlich kommt noch eine weitere wichtige Hilfsannahme des BEA zum Einsatz, nämlich die Annahme, dass die posterior-Verteilung der Gewichte $\mathbf{w}$ durch

eine Gaußverteilung mit Mittelwert $\mathbf{w}_{\mathrm{MP}}$ angenähert werden kann. Über diese Annahme und ihre Relevanz für das Funktionieren des Ansatzes in der Praxis muss später noch einiges gesagt werden. Aufgrund dieser Annahme kann $p(\mathbf{w}|\mathcal{D})$ mit Hilfe der Hesse-Matrix $\mathbf{A}$, der Matrix der zweiten Ableitungen von S_W, neu formuliert werden als

$$p(\mathbf{w}|\mathcal{D}) = \frac{1}{Z_S^*} \exp\bigl(-S(\mathbf{w}_{\mathrm{MP}}) - \frac{1}{2}\Delta\mathbf{w}'\mathbf{A}\Delta\mathbf{w}\bigr), \tag{4.86}$$

wobei Z_S^* der für die Gauß-Approximierung neu formulierte Normalisierungsfaktor ist:

$$Z_S^*(\alpha,\beta) = \exp\bigl(-S(\mathbf{w}_{\mathrm{MP}})\bigr)\,(2\pi)^{\frac{W}{2}}\,|\mathbf{A}|^{-\frac{1}{2}}. \tag{4.87}$$

Unter Verwendung dieser Gleichung, der Gleichungen (4.74) und (4.76) sowie durch Logarithmierung ergibt sich

$$\ln p(\mathcal{D}|\alpha,\beta) = -S(\mathbf{w}_{\mathrm{MP}}) - \frac{1}{2}\ln|\mathbf{A}| + \frac{W}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi). \tag{4.88}$$

Um das Maximum zu ermitteln wird die Ableitung von $\ln p(\mathcal{D}|\alpha,\beta)$ nach α und β gebildet und Null gesetzt. Hierbei taucht der Term

$$\gamma = \sum_{i=1}^{W} \frac{\lambda_i}{\lambda_i + \alpha} \tag{4.89}$$

auf, wobei $\lambda_i + \alpha$ die Eigenwerte der Matrix $\mathbf{A} = \mathbf{H} + \alpha\mathbf{I}$ sind, $\mathbf{H} = \beta\nabla\nabla E_{\mathcal{D}}$ die Hesse-Matrix der unregularisierten Fehlerfunktion ist und λ_i die Eigenwerte von $\mathbf{H}$ sind. γ ist die Anzahl der wohl-determinierten Gewichte, d.h. der Gewichte, deren Wert von den Daten und nicht vom Regularisierer bestimmt wird. Diese Interpretation wird klar, wenn man die einzelnen Summanden in Gleichung (4.89) betrachtet. Im Falle dass der Eigenwert λ_i im Vergleich zu α groß ist, ergibt sich ein Summand von Eins und es handelt sich um ein wohl-determiniertes Gewicht. Anderenfalls ist der Wert des Summanden nahe Null und es ist anzunehmen, dass auch der Wert des entsprechenden Gewichts vom Regulierer in die Nähe von Null gedrängt wird.

Aus diesen Überlegungen erhält man, dass im Maximum

$$2\alpha E_W = \gamma \;\text{ und }\; 2\beta E_{\mathcal{D}} = N - \gamma \tag{4.90}$$

gilt. In der praktischen Applikation des BEA muss sowohl α und β als auch der optimale Gewichtsvektor $\mathbf{w}_{\mathrm{MP}}$ gefunden werden. Zu diesem Zweck geht man so vor, dass man ein lokales Optimierungsverfahren für die Suche nach $\mathbf{w}_{\mathrm{MP}}$ einsetzt und periodisch die Werte von α und β gemäß den Iterationsformeln

$$\alpha^{\mathrm{neu}} = \gamma/2E_W \;\text{ und }\; \beta^{\mathrm{neu}} = (N-\gamma)/2E_{\mathcal{D}} \tag{4.91}$$

aktualisiert und sodann das Suchverfahren an der gleichen Stelle der Fehlerlandschaft fortsetzt.

Abschließend muss für die richtige Interpretation der so gefundenen α und β wie auch des Verfahrens des BEA insgesamt betont werden, dass α und β nur in Bezug auf das lokale Minimum optimal sind, da auch die Gaußapproximierung gemäß Gleichung (4.86) nur in Bezug auf die Umgebung des lokalen Minimums gültig ist. In einfachen linearen Modellen mit nur einem einzigen Minimum der Fehlerfunktion ist dies keine Einschränkung. Für ARNN-Modelle, die aus Gründen der Modellsymmetrie wie auch der allgemeinen Modellkomplexität zahlreiche Minima haben können, resultieren jedoch Einschränkungen sowohl für die theoretische Interpretation als auch die praktische Anwendung.

Anwendung

Im Rahmen der vorliegenden Arbeit wurde der bayesianische Ansatz für ARNN-Modelle als Funktionalität des R-Softwarepakets `NNUTS` implementiert, wobei, wie bereits oben erwähnt, auf den BEA zur Bestimmung der Regularisierungsparameter eingeschränkt wurde. Somit wird ein Spezialfall von Regularisierung verwirklicht. Die Umsetzung berücksichtigt getrennte Regularisierung der ersten und der zweiten Schicht der Gewichte des NN-Teils, $\mathbf{a}$ und $\mathbf{c}$ sowie der direkten Verbindungen, d.h. der Koeffizienten des linearen Teils, $\mathbf{f}$.

Gegenüber der von MacKay beschriebenen Vorgangsweise waren einige kleinere Modifikationen des Algorithmus notwendig um das stabile Funktionieren des Ansatzes auch im komplexen Anwendungsgebieten von ARNN-Modellen auf makroökonomischen Zeitreihen zu gewährleisten. Für die möglichen auftretenden Probleme und die entsprechenden Modifikationen liefert teilweise die Forschungsliteratur Hinweise, teilweise scheinen diese bisher allerdings noch nicht thematisiert worden zu sein.

Zum einen kann bei komplexen Modellen aus numerischen Gründen der Fall auftreten, dass die Hesse-Matrix nicht in einem Minimum der Fehlerlandschaft ausgewertet wird. Die Eigenwerte λ_i können daher negativ sein, sodass der Nenner des Terms $\lambda_i/(\lambda_i+\alpha)$ in Gleichung (4.89) im ungünstigen Falle sehr nahe bei Null liegen kann. Daraus resultiert eine numerische Instabilität der Berechnung von γ. Bishop (1995) erwähnt das Problem nebenbei. Um dem Problem auszuweichen, wurden in der vorliegenden Softwareimplementation einem Vorschlag von Penny und Roberts (1999) folgend negative Eigenwerte der Hesse-Matrix $\mathbf{H}$ auf Null gesetzt.

Weiters kann es in nicht-linearen und stark überparametrisierten Modellen zu einem Ausbleiben der Konvergenz des Algorithmus kommen. Hier spielt der Umstand, dass die Annahme der Gaußverteilung der Modellparameter bei diesen Modellen nicht zutrifft, eine entscheidende Rolle. Bereits MacKay (1992b) stellt fest, dass diese Annahme zusammenbrechen kann, wenn die Anzahl der Gewichte im Vergleich zur Anzahl der Datenpunkte zu groß wird. Da es in der Praxis zudem schwer ist, langsame Konvergenz des Algorithmus von Divergenz zu unterscheiden, muss dem Algorithmus eine maximale Anzahl von Iterationen vorgegeben werden.

Über das Konvergenzverhalten des BEA finden sich in der Forschungsliteratur wenig Hinweise, was angesichts der relativ einfachen Anwendungsprobleme, die zur Illustration des Ansatzes herangezogen werden, nicht verwundert.

Schließlich kann in einem besonders ungünstigen Falle von Überparametrisierung die Anzahl der wohl-determinierten Parameter γ gegen Null gehen, dies obwohl gleichzeitig $E_W > 0$ und Overfitting stattfindet. Diese Gefahr ist naturgemäß dann besonders groß, wenn in den Daten keine oder nur wenig lineare bzw. nicht-lineare Struktur vorhanden ist. Es resultiert eine numerische Instabilität des Algorithmus, da aufgrund des verschwindenden γ auch ν verschwindet und ohne Regularisierung dem Overfitting keine Grenzen gesetzt werden. Im Rahmen der implementierten Softwarelösung wird daher für γ eine untere Schwelle von 1/20 der Anzahl der Parameter vorgegeben.

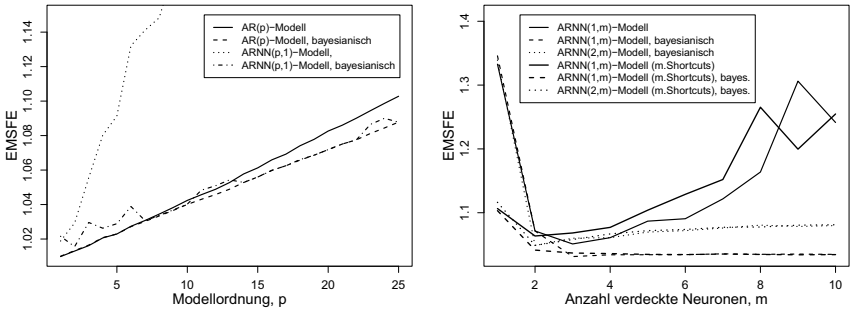
Im Zusammenhang mit dem BEA wurde von MacKay auch eine vereinfachte Version des Algorithmus vorgeschlagen, in der auf die Berechnung der Anzahl der wohl-determinierten Parameter mit Hilfe der Auswertung der Hesse-Matrix verzichtet wird und statt dessen angenommen wird, dass sämtliche Parameter des Modells wohl-determiniert sind. Die in diesem Fall verwendeten Iterationsformeln lauten

$$\alpha^{\text{neu}} = W/(2E_W), \quad \beta^{\text{neu}} = N/(2E_{\mathcal{D}}) \tag{4.92}$$

Auch für diesen vereinfachten Zugang liegt eine Softwareimplementation im Rahmen des R-Pakets `NNUTS` vor. Jedoch ist, wie sich in verschiedenen Versuchsanordnungen gezeigt hat, von diesem Ansatz in Hinblick auf die hier untersuchten Anwendungsgebiete wenig zu erwarten. Einerseits ist bei einfach gehaltenen Modellen, bei denen die Annahme, dass sämtliche Parameter wohl-determiniert sind, nicht weit fehl geht, auch die Berechnung der Hesse-Matrix weder in numerischer Hinsicht noch aufgrund der Rechenzeit ein großes Problem, sodass kein Vorteil zu erwarten ist. Bei komplexeren und potentiell überparametrisierten Modellen andererseits bewirkt ein großer Unterschied zwischen W und γ, dass die Modelle bei Anwendung der Iterationsformeln (4.92) deutlich zu stark regularisiert werden. In dieser Arbeit werden daher nur Ergebnisse mit dem nicht-vereinfachten BEA präsentiert.

Um die Funktionsweise des BEA zu untersuchen, ist insbesondere interessant, wie der Algorithmus mit überflüssigen Parametern eines Modells zurechtkommt, wenn die Spezifikation des datengenerierenden Prozesses bekannt ist. Dies ist nicht mit realweltlichen sondern nur mit simulierten Zeitreihen möglich.

Im folgenden werden daher zwei Simulationsstudien durchgeführt, welche die Effektivität des BEA einerseits bei überflüssig parametrisierter Lag-Struktur in einem AR-Modell und andererseits bei wachsender Anzahl von verdeckten Neuronen in einem ARNN-Modell erkunden. Es wurde $N = 100$ AR(1)-Zeitreihen mit $\phi = 0.6$ bzw. SIM-Zeitreihen nach Gleichung (4.11) jeweils der Länge $T = 2000$ generiert. Die ersten $T_{in} = 300$ Datenpunkte dieser Zeitreihen wurden jeweils zur



(a) AR(1) (b) SIM

Abb. 14: EMSFE-Ergebnisse mit dem BEA bei überparametrisierten Modellen im Vergleich mit nicht-bayesianischen Methoden: (a) auf einem AR(1)-Prozess und (b) auf einer simulierten nicht-linearen Zeitreihe (SIM)

Modellschätzung verwendet, die restlichen $T_{out} = 1700$ zur Berechnung des MSFE

$$MSFE = \frac{1}{T_{out}} \sum_{t=T_{in}+1}^{T} (\hat{y}_t - y_t)^2. \tag{4.93}$$

Die Mittelung über die 100 Durchläufe stellt einen Schätzer für den erwarteten MSFE dar, $EMSFE = \frac{1}{N} \sum_n MSFE_n$. In Abb. 14 werden die Ergebnisse der beiden Simulationstudien zusammengefasst.

Im Diagramm a) von Abb. 14 wurde auf der X-Achse die Modellordnung p abgetragen. Der EMSFE des Basismodells, des mit der Kleinstquadrat-Methode (KQ) geschätzten AR(p)-Modells, ist als kontinuierliche Linie eingetragen. Im Falle von $p = 1$ entspricht dieses Modell abgesehen von einem zusätzlich aufgenommenen Interzept-Term dem datengenerierenden Prozess, und der EMSFE liegt mit 1.0098 sehr nahe bei der Varianz des datengenerierenden Prozesses von 1. Wie zu erwarten, steigt der EMSFE mit wachsendem p aufgrund von Overfitting an (bis auf 1.1028). Im Vergleich dazu liefert das mit dem BEA geschätzte AR(p)-Modell bis zu einer Modellordnung von $p = 4$ marginal schlechtere EMSFE-Werte und ab $p = 5$ zunehmend bessere EMSFE-Werte. Bei der höchsten gerechneten Modellordnung liegt der EMSFE bei der bayesianischen Schätzung bei 1.0878. Außerdem wurden im Diagramm die EMSFE-Werte für ein geschätztes ARNN(p,1)-Modell eingetragen, d.h. für ein Modell mit Input-Knoten für die Lags 1 bis p und mit einem verdeckten Neuron. Hier erweist sich das Overfitting-Problem bei der Schätzung mit KQ als so gravierend, dass bereits bei $p = 9$ der EMSFE über 1.15 liegt und für höhere p-Werte nicht im Diagramm eingetragen werden konnte. Der bayesianische Zugang liefert hingegen meist die gleichen EMSFE-Werte wie bei einem linearen

Modell.[30] Die Simulation zeigt, dass die Gefahr von Overfitting bei rein linearen Modellen von vorneherein relativ gering ist, jedoch durch den bayesianischen Ansatz bei Überparametrisierung teilweise verhindert werden kann.

Das Diagramm b) von Abb. 14 präsentiert eine Analyse, wobei der datengenerierende Prozess nicht-linear ist und entlang der X-Achse die Anzahl der verdeckten Neuronen m variiert wird. Die Nicht-Linearität ist offenbar von einer Art, die mit einem verdeckten Neuron allein nicht befriedigend approximiert werden kann, da die EMSFE-Kurvenverläufe ihr Minimum bei 2 bzw. 3 haben. Es wurden sowohl Modelle ohne als auch mit direkten Verbindungen geschätzt, deren EMSFE-Kurven ähnlich verlaufen. Wie zu erwarten ist die optimale Anzahl der verdeckten Neuronen bei letzteren niedriger. Im Vergleich der mit KQ geschätzten Modelle mit jenen des bayesianischen Ansatzes erweisen sich letztere stark überlegen, dies nicht nur bei stark überparametrisierten Modellen, sondern auch bereits beim optimalen $m = 3$. Hier beträgt der EMSFE 1.0313 (im Gegensatz zu 1.0683 beim mit KQ geschätzten Modell). Auch wenn zusätzlich zu überflüssigen verdeckten Neuronen auch die Lag-Struktur der geschätzten Modelle überparametrisiert wird (2 statt ansonsten bloß ein Lag), ist der bayesianisch Zugang in der Lage, das Overfitting-Problem in den Griff zu bekommen, wie sich an einer EMSFE-Kurve zeigt, die bei wachsendem m nicht über ca. 1.08 steigt. Der bayesianische Ansatz scheint also bei nicht-linearen Modellschätzungen besonders hilfreich zu sein.

4.10 Schlussfolgerungen

In diesem Kapitel wurden die Grundlagen des Einsatzes von NN zur Modellierung und Prognose von makroökonomischen Zeitreihen präsentiert, die ARNN- und ARNNDS-Modelle eingeführt und die Techniken und Ansätze zu ihrer Anwendung erarbeitet. ARNN- und ARNNDS-Modelle sind durch die Möglichkeit der Berücksichtigung einer deterministischen Saison- und Trendkomponente, Flexibilität und Nicht-linearität des NN-Modellteils, Aufnahme eines linearen Modellteils sowie der sparsamen Spezifikation der Lag-Struktur besonders geeignet für die Anwendung auf makroökonomische Zeitreihen.

Es werden vier verschiedene Modellierungsansätze unterschieden, die unterschiedlichen Sichtweisen und Interpretationsformen von ARNN-Modellen entsprechen und die Entwicklung von Modellierungsstrategien zur Gewährleistung der Generalisierungsfähigkeit von ARNN-Modellen ermöglichen. Sowohl bei den verschiedenen Techniken der Initialisierung, lokalen und globalen Otpimierung als auch im Bereich der Modellierungsansätze wurden nicht nur die in der Literatur bekannten Methoden analysiert, sondern auch neue Algorithmen und Heuristiken eingeführt. Z.B. sind die Anpassung des TLG-Tests im Rahmen des Growing, die Entwicklung

[30] Wie eine Nachprüfung zeigte, sind Abweichungen durch einzelne Fälle von suboptimalen Minima verursacht worden, die in einem realweltlichen Anwendungsfall ohne weiteres als solche erkannt werden, jedoch in einer Simulation schwer eliminiert werden können.

einer Pruning-Strategie, Stoppregeln für das gestoppte Training und Heuristiken zur Bestimmung der Regularisierungsparameter zu nennen. Die Funktionsweisen dieser Methoden wurden auf einer simulierten nicht-linearen Zeitereihe SIM sowie auf den beiden makroökonomischen Zeitreihen ALR und IPI illustriert und belegt.

Dieses Kapitel dient auch dazu, die Voraussetzungen für das Kap. 5 zu schaffen, in welchem ein systematischer Prognosevergleich von linearen Modellen und NN-Modellen anhand konkreter Modellierungsstrategien auf der Basis der entwickelten Methoden durchgeführt werden soll. Erst auf der Basis eines umfassenden Vergleichs anhand von Out-of-Sample-Daten lässt sich sagen, ob das lineare oder nicht-lineare Instrumentarium auf den beiden ausgewählten Beispielzeitreihen überlegen ist.

Eine Reihe von Schlussfolgerungen zum Einsatz von ARNN-Modellen in diesem Anwendungsgebiet sind jedoch weitgehend unabhängig von einem strengen Out-of-Sample-Vergleich und haben sich bereits in den Anwendungen dieses Kapitels klar genug gezeigt. Es betrifft dies die folgenden Aspekte: i) tatsächliches Vorhandensein nicht-linearer Strukturen in den Zeitreihen, ii) Nachvollziehbarkeit der Methoden und iii) Rechen- und Modellierungsaufwand für den Anwender.

Das Vorhandensein von Nicht-Linearität in einem Ausmaß, das für bessere Vorhersagen genutzt werden kann, ist fraglich. Allenfalls ist für die Zeitreihe IPI in geringem Ausmaß Nicht-Linearität vorhanden. Zu diesem Befund führt vor allem der Umstand, dass auf der simulierten Zeitreihe SIM mit allen vier Modellierungsansätzen die Nicht-Linearität im bekannten, nahezu vollständigen Ausmaß mit ARNN modelliert werden kann und für die Prognosegüte genützt werden kann, wobei keine numerischen Probleme auftreten. Im Gegensatz dazu tauchen auf den Beispielzeitreihen bei allen vier Ansätzen gewisse numerische Probleme auf, und es ergeben sich kaum Hinweise auf eine bessere Prognosegüte der ARNN-Modelle. In den Anwendungsfällen zu ES und Regularisierung, bei denen durch eine zusätzliche Unterteilung der Daten eine begrenzte Out-of-Sample-Testung durchgeführt wurde, lag für die Zeitreihe IPI die Prognosegüte auch bei günstigem Verlauf der Lernkurven (ES) bzw. bei optimaler Einstellung der Regularisierungsparameter nur äußerst knapp unter dem linearen Benchmark, der durch das entsprechende AR-Modell vorgegeben wurde. Bei der ALR wurde mit den ARNN-Modellen nicht einmal der lineare Benchmark erreicht. Dies mag auf den Zufälligkeiten des als Evaluierungsset verwendeten Abschnitts der Zeitreihe gründen. Die Erwartungen über die erreichbaren Verbesserungen der Prognosegüte von nicht-linearen gegenüber linearen Modellen sollten jedoch bereits aufgrund dieser Ergebnisse mäßig sein.

Die Nachvollziehbarkeit der Modellierung mit ARNN ist ein großes Anliegen der vorliegenden Arbeit wie auch der NN-Forschung insgesamt und betrifft alle vier behandelten Modellierungsansätze. Wenn Nachvollziehbarkeit jedoch nicht nur den Prozess der Modellbildung betrifft, also das Befolgen von klar vorgegebenen Regeln, sondern auch das Ergebnis, so muss festgestellt werden, dass mit keinem der vier Ansätze in den Anwendungen auf den beiden Zeitreihen ALR und IPI das Ziel der Nachvollziehbarkeit erreicht wird. Die gefundenen Modelle sind in vielerlei Hinsicht

von Zufälligkeiten abhängig, insbesondere jenen der Initialisierung und der Problematik von suboptimalen lokalen Minima. Auch im statistisch-parametrische Ansatz, der die Frage der Modellspezifikation und -schätzung am rigorosesten behandelt, kann das Ergebnis zweier Modellierungsversuche unter Verfolgung der gleichen Modellierungsstrategie beträchtliche Unterschiede aufweisen. Dies muss natürlich in Hinblick auf die zuvor erwähnte Möglichkeit relativiert werden, dass die Beispielzeitreihen ALR und IPI keine ausreichende Nicht-Linearität enthalten. Denn bei klar vorhandener Nicht-Linearität, so zeigt sich am Beispiel der Zeitreihe SIM, liefern die verschiedenen Modellierungsansätze verlässlich sehr ähnliche Ergebnisse und Prognosegüten.

Hinsichtlich des Rechen- und Modellierungsaufwands für den Anwender sind ARNN-Modelle im Vergleich zu linearen Modellen deutlich anspruchsvoller. Doch angesichts der Rechenstärke moderner Computer und der zunehmenden Qualität statistischer Softwarepakete im allgemeinen und von NN-Software im speziellen tritt dieser Aspekt mehr und mehr in den Hintergrund und ist auch bei der Modellierung von makroökonomischen Zeitreihen nahezu zu vernachlässigen. Dennoch konnten im Rahmen dieses Kapitels bestimmte Schritte von Modellierungsstrategien als unverhältnismäßig zeit- und rechenaufwendig ausgemacht werden, z.B. der Versuch das globale und nicht nur ein nahezu optimales lokales Minimum der Fehlerfunktion eines ARNN-Modell hoher Modellkomplexität zu finden, oder die Bestimmung der Regularisierungsparameter mithilfe einer Validierungsstrategie. Von solchen Teilalgorithmen sollte daher nicht Gebrauch gemacht werden sondern statt dessen die alternativen Verfahren und Heuristiken verwendet werden. Die im Rahmen der vorliegenden Arbeit entwickelten Ansätze und Heuristiken werden im R-Paket `NNUTS` frei verfügbar gemacht und sollen damit einen Beitrag leisten, den Rechen- und Modellierungsaufwand mit ARNN überschaubar zu halten. Sollten in einzelnen Fällen mit Routinen erhöhter Rechenaufwand verbunden sein, so liegt das oft an numerischen und Konvergenzproblemen. Wie bereits bei den beiden vorigen Punkten erwähnt, ist dies aber nicht ein grundsätzliches Problem von ARNN-Modellen, sondern weist auf Problemen der Modellspezifikation angesichts ungenügend vorhandener Nicht-Linearität hin.

5 Evaluierung der Prognosen

In diesem Kapitel wird eine systematische Evaluierung der Prognosequalität der ARNN-Modelle unter dem Einsatz von verschiedenen Modellierungszugängen sowie der alternativen linearen Modelle anhand der beiden Beispielzeitreihen ALR und IPI durchgeführt. Die Evaluierung der Prognosequalität erfolgt auf einem beiseitegelegten Datenset der beiden Zeitreihen, das bis zu diesem Zeitpunkt noch nicht für die Zwecke der Entwicklung der Modellierungszugänge verwendet wurde. Auf diese Weise soll eine echte Prognosesituation nachgebildet werden, wie sie in der angewandten Wirtschaftsforschung typischerweise vorkommt. Auch in anderen Aspekten der Prognoseerstellung, z.B. in der periodischen Aktualisierung der für die Prognose zur Verfügung stehenden Informationen und Modelle und in der Berücksichtigung von Mehr-Schritt-Prognosen sollen diese Anforderungen berücksichtigt werden. Die Zielsetzung der Evaluierung liegt in der Abschätzung der Brauchbarkeit des Instrumentariums der neuronalen Netze für typische Anwendungsfälle der Prognose makroökonomischer Zeitreihen.

Im ersten Abschnitt werden zunächst die Möglichkeiten zur Evaluierung der Prognosegüte überblicksmäßig dargestellt und die Grundlagen zu Gütemaßen und die statistischen Tests zum Vergleich der Prognosegüte gelegt. Danach (Abschnitt 5.2) werden die Besonderheiten erläutert, die sich aus der Erstellung von Mehr-Schritt-Prognosen ergeben. Um Mehr-Schritt-Prognosen mit ARNN-Modellen durchführen zu können, wird eine teilweise neuartige Notation eingeführt und eine Erweiterung der ARNN-Modelle vorgenommen, die durch die Berücksichtigung von direkten Mehr-Schritt-Prognosen notwendig wird. Abschnitt 5.3 dient der detaillierten Beschreibung der Vorgangsweise der Erstellung der Prognosen auf Basis der verschiedenen Modelle und Modellierungsstrategien. Die Ergebnisse der Evaluierung werden in Abschnitt 5.4 analysiert. Da die Anzahl der verschiedenen Prognosen aufgrund der verschiedenen Modelle, Modellierungsansätze und Modellierungsstrategien, Zeitreihen, Differenzenfilter usw. sehr groß ist, können einige Teilergebnisse aus Darstellungsgründen nicht in diesem Abschnitt besprochen werden, sondern werden in einen eigenen Anhang der Arbeit aufgenommen. Abschnitt 5.5 zieht die wichtigsten Schlussfolgerungen aus der Evaluierung.

5.1 Grundlagen

Der Vergleich von verschiedenen Prognosemethoden hat auf einem beiseite gelegten Datenset, dem Evaluierungsset (engl. „Out-of-Sample“ bzw. „Hold-Out-Sample“) zu geschehen, von dem im Zuge der Modellierung noch nicht Gebrauch gemacht wurde. Obwohl dies eine Selbstverständlichkeit ist, ist die Gefahr dennoch groß, gegen diesen Grundsatz auf subtile Art zu verstoßen. Mit „Data Peeking“ bezeichnet man es, wenn der Prognostiker im Zuge der Modellierungsphase auf das Evaluierungsset hinüberspäht, und – vielleicht nur unwillkürlich – das so erlangte Wissen

in die Modellierungsphase eingehen lässt (z.B. Tashman, 2000). Auch der Ausdruck „Data Snooping“ wird verwendet, um die Gefahr einer ungenügenden Trennung zwischen Modellierung und Evaluierung aufzuzeigen (White, 2000). In der vorliegenden Arbeit wird dieser Gefahr begegnet, indem rigoros für Entwicklung der Algorithmen und Software sowie für Analyse und Modellierung nur simulierte Zeitreihen oder ökonomische Zeitreihen, die 1997/12 enden, verwendet werden. Auf dieser Basis wird ein weitgehend automatisiertes Evaluierungsschema entwickelt, das zur Evaluierung der verschiedenen Prognosetechniken und Modellierungsstrategien auf der beiseite gelegten Zeitreihe 1998/1–2002/12 dient.

Prognosegütemaße

Für Vergleiche der Prognosegüte können verschiedene Gütemaße eingesetzt werden, für die in der Literatur an vielen Stellen Definitionen gegeben werden, u.a. bei Zhang et al. (1998). Diskussionen der Eigenschaften der gebräuchlichsten Prognosegütemaße finden sich bei Granger und Newbold (1986), Granger und Teräsvirta (1993), Clements und Hendry (1998). Diese Arbeit hält sich an die grundsätzliche Empfehlung von Granger (1993), dass das gleiche Kriterium sowohl für die Schätzung als auch für die Evaluierung der Prognosegüte herangezogen werden sollte, und verwendet ausschließlich das aus dem SSE abgeleitete Gütemaß MSE (Mean Squared Error) bzw. MSFE (Mean Squared Forecasting Error). In diesem Zusammenhang spielt auch das Konzept der Kostenfunktion eine Rolle, für die im Zusammenhang mit makroökonomischen Prognosen allgemein eine quadratische Form angenommen wird, was die Wahl von aus dem SSE abgeleiteten Maßen rechtfertigt (vgl. Clements und Hendry, 1998, S. 53).

Alternative Prognosegütemaße beruhen z.B. auf dem absoluten Fehler (z.B. MAD, Mean Absolute Deviation) oder auf dem Vorzeichenfehler. Für den (aggregierten) Vergleich von Prognosemethoden auf unterschiedlichen Zeitreihen, wie sie bei Prognosewettbewerben vorgenommen werden, haben auch Maße, die eine Art der Normierung vornehmen, eine Berechtigung (z.B. MAPE, Mean Absolute Percentage Error). Hyndman und Koehler (2006) bieten einen umfassenden Überblick insbesondere über solche alternative Prognosegütemaße.

Der MSFE ist der mittlere quadrierte Prognosefehler von mit der gleichen Prognosemethode und mit dem gleichen Prognosehorizont erstellten Prognosen. Die Definition des MSFE wurde in leicht unterschiedlichem Kontext bereit in Kap. 4 in Gleichung (4.93) angegeben und wird für den Zweck der Out-of-Sample-Evaluierung nun neu formuliert:

$$MSFE = \frac{1}{N_{ev}} \sum_{t=T^*+1}^{T^*+N_{ev}} (\hat{y}_t - y_t)^2, \tag{5.1}$$

wobei N_{ev} der Umfang des Evaluierungssets und T^* die letzte Beobachtung des Trainingsets ist. $\hat{y}_t$ und y_t bezeichnen den prognostizierten Wert und den tatsächlichen Wert der Zeitreihe.

Tests auf Prognosegüte

Schließlich drängt sich die Frage auf, ob ein ermittelter Unterschied zwischen der Prognosegüte zweier alternativen Prognosen auf tatsächliche Überlegenheit einer der beiden Methoden oder bloß auf den Zufall zurückzuführen ist. Für diese Fragestellung wurde eine Reihe von statistischen Tests entwickelt, unter denen der Diebold-Mariano-Test (DM-Test) von Diebold und Mariano (1995) der vielseitigste und gebräuchlichste ist. Andere Tests, die ebenfalls angewendet werden können, wurden von Harvey et al. (1997) und Clark (1999) entwickelt. Hier wird im folgenden die Berechnung des DM-Tests beschrieben.

Unter der Null-Hypothese des DM-Tests stammen zwei vorliegende Prognosezeitreihen von Methoden gleicher Prognosegüte. Bezeichnet man mit $e_{1,t}$ den Prognosefehler der ersten Prognosemethode und mit $e_{2,t}$ den Prognosefehler der zweiten Prognosemethode, so beruht der Test auf der Zeitreihe der Prognosegütedifferenz

$$z_t = g(e_{1,t}) - g(e_{2,t}), \tag{5.2}$$

wobei g eine beliebige Verlustfunktion ist. Analog zur Wahl des MSFE als Prognosegütemaß wird in dieser Arbeit, wie allgemein üblich, eine quadratische Verlustfunktion bevorzugt, womit sich

$$z_t = e_{1,t}^2 - e_{2,t}^2 \tag{5.3}$$

ergibt. Die DM-Teststatistik lautet

$$Z = \hat{\sigma}_z^{-1} \bar{z}, \tag{5.4}$$

wobei $\bar{z} = N_{ev}^{-1} \sum_t z_t$ und $\hat{\sigma}_z^2$ ein Schätzer für die asymptotische Varianz von $\bar{z}$ ist. Unter der Null-Hypothese ist die Teststatistik asymptotisch standardnormalverteilt. Der Test kann gegen eine einseitige oder zweiseitige Alternative durchgeführt werden.

5.2 Mehr-Schritt-Prognosen

Eine wichtige Unterscheidung, die für dieses Kapitel zentral ist, betrifft Ein-Schritt-Prognosen und Mehr-Schritt-Prognosen. Wenn für die Prognosenerstellung Information bis zur Periode T genutzt wird, betrifft die Ein-Schritt-Prognose („One-Step-Ahead“), $x_{T+1|T}$, die Periode $T+1$, die Mehr-Schritt-Prognose („Multi-Step-Ahead“), $x_{T+h|T}$, die Periode $T+h$, wobei h der Prognosehorizont ist. Der Zeitpunkt der Prognoseerstellung T wird auch als Prognose-Ursprung (engl. forecasting-origin) bezeichnet.

Die Mehr-Schritt-Prognosen haben (bei monatlichen Zeitreihen) in der Prognosepraxis, insbesondere in der institutionalisierten Wirtschaftsprognose, als Entscheidungsgrundlage für die Politik eine im Vergleich zu Ein-Schritt-Prognosen ungleich

höhere Bedeutung, da die Wirtschaftspolitik deutlich weiter als nur einen Monat in die Zukunft planen muss. Dieses Erfordernis steht im Gegensatz zu den Modellformulierungen des vorangegangenen Kapitels, die sich an einem stochastischen Prozess orientieren, von dem angenommen wird, dass er die Zeitreihe generiert („Datengenerierender Prozess“) und der in den ausgewählten Fällen (AR, ARMA, ARMADS, ARNN und ARNNDS) immer nur bestimmt, wie der jeweils nächstfolgende Wert einer Zeitreihe zustandekommt. Die Ableitung von optimalen Mehr-Schritt-Prognosen ist bei linearen Prozessen bzw. Modellen keine Schwierigkeit, stellt hingegen bei nicht-linearen Prozessen bzw. Modellen ein Problem dar, das im folgenden erläutert werden soll. Insbesondere sind für ARNN-Modelle spezielle Erweiterungen und Ergänzungen notwendig, um ihr Potenzial nützen zu können.

Mehr-Schritt-Prognosen mit nicht-linearen Modellen

Prognosen mit linearen Methoden, z.B. ARMA-Modellen, haben die Eigenschaft, dass Mehr-Schritt-Prognosen durch Iterierung von Ein-Schritt-Prognosen erstellt werden können. Am Beispiel eines AR(1)-Prozesses und einer Zwei-Schritt-Prognose $x_{T+2|T}$:

$$\begin{aligned} x_{T+2|T} &= \mathrm{E}_T(x_{T+2}) = \mathrm{E}_T(\phi x_{T+1} + \varepsilon_{T+2}) = \phi \mathrm{E}_T(x_{T+1}) \\ &= \phi x_{T+1|T} = \phi \mathrm{E}_T(\phi x_T + \varepsilon_{T+1}) = \phi^2 x_T, \end{aligned} \tag{5.5}$$

wobei E_T den bedingten Erwartungswert unter Nutzung der zum Zeitpunkt T erhältlichen Information bezeichnet. Wenn man sich vor Augen führt, dass $x_{T+2|T+1} = \phi x_{T+1}$, so wird klar, dass die Zwei-Schritt-Prognose als Ein-Schritt-Prognose mit neuem Prognose-Ursprung $T+1$ aufgefasst werden kann, wobei statt dem wahren x_{T+1} die Prognose $x_{T+1|T}$ eingesetzt wird. Daher wird die iterierte Prognose auch als „Plug-In-Prognose“ bezeichnet. Der Prozess des iterierten Einsetzens muss für Modelle höherer Ordnung und für höhere Prognosehorizonte entsprechend verallgemeinert werden.

Für nicht-lineare Modelle hingegen entsteht die Schwierigkeit, dass eine solche Iteration nicht korrekt ist. Dies lässt sich anhand eines NLAR(1)-Prozesses $x_t = g(x_{t-1}) + \varepsilon_t$ illustrieren, wobei $g()$ eine nicht-lineare Funktion ist:

$$x_{T+1|T} = \mathrm{E}_T(x_{T+1}) = \mathrm{E}_T(g(x_T) + \varepsilon_{T+1}) = g(x_T) \tag{5.6}$$

$$\begin{aligned} x_{T+2|T} &= \mathrm{E}_T(x_{T+2}) = \mathrm{E}_T(g(x_{T+1}) + \varepsilon_{T+2}) = \mathrm{E}_T(g(g(x_T) + \varepsilon_{T+1})) \\ &\neq \mathrm{E}_T(g(g(x_T))) = g(g(x_T)). \end{aligned} \tag{5.7}$$

Der Grund für die Ungleichheit ist der Umstand, dass der Erwartungswert einer nicht-linearen Funktion (einer Verteilung) nicht die nicht-lineare Funktion des Erwartungswertes (einer Verteilung) ist.

Diese Schwierigkeit wird auch heute noch in vielen angewandten Forschungsarbeiten zur nicht-linearen Zeitreihenprognose, insbesondere mit NN-Methoden, nicht

entsprechend gewürdigt. Die Tragweite des Problems hängt von Art und Ausmaß der Nicht-Linearität und vom Prognosehorizont ab und kann nur für den einzelnen Fall abgeschätzt werden.

Bei der Erstellung von Mehr-Schritt-Prognosen mit nicht-linearen Modellen verfügt man prinzipiell über fünf verschiedene Alternativen (Granger und Teräsvirta 1993, S. 131):

(i) naive Methode: iterierte Prognose unter Missachtung des oben dargestellten Problems,

(ii) exakte Methode: Anwendung von iterierten Integralen über die Verteilungen der Fehler $\varepsilon_{T+1}, \varepsilon_{T+2}, \ldots, \varepsilon_{T+h-1}$,

(iii) Monte-Carlo-Methode: Ziehung von N zufälligen Fehlerzeitreihen $\varepsilon_t, t = T + 1, T + 2, \ldots, T + h - 1$ und Ermittlung von $x_{T+h|T} = N^{-1} \sum_{i=1}^{N} \dot{x}_{T+h}$, wobei $\dot{x}_{T+h}$ durch Simulation mit Hilfe des geschätzten Modells und unter Einsetzen von ε_t generiert wird,

(iv) Boot-Strap-Methode: analog zu (iii), wobei ε_t aus den Residuen des geschätzten Modells gezogen wird,

(v) direkte Methode: Erstellung der Prognose $x_{T+h|T}$ anhand eines für den Prognosehorizont h maßgeschneiderten nicht-linearen Modells, z.B. einer nicht-linearen Regression von x_{t+h} auf $x_t, x_{t-1}, \ldots$.

Eine Beurteilung der Vor- und Nachteile dieser Methoden wird zunächst auf ihre technische Umsetzbarkeit und den damit verbundenen Zeitaufwand achten. Während die naive Methode hierbei klar am besten abschneidet, wird die exakte Methode wohl für die meisten nicht-linearen Modelle nicht umsetzbar sein, weil das entstehende mathematische Problem analytisch zu komplex ist. Diese Einschätzung scheint jedenfalls auf die ARNN-Modelle zuzutreffen. In der vorliegenden Arbeit wurde nicht versucht, die exakte Methode umzusetzen.

Die beiden simulativen Methoden sind mit modernen Rechnern und Softwarelösungen gut umsetzbar und wurden im Rahmen des R-Paket `NNUTS` in der Funktion `iterated.multistep` implementiert. Weiter unten in diesem Abschnitt folgt eine Anwendung anhand der Zeitreihe IPI und der simulierten nicht-linearen Zeitreihe SIM, die einen Vergleich der naiven und der simulierten iterierten Methode erlaubt.

In der Methode der direkten Mehr-Schritt-Prognose liegt ein grundsätzlicher Weg vor, das erläuterte Problem zu lösen. In diesem Fall wird unter Aufgabe des Ziels, den wahren Prozess abzubilden, das zu schätzende Modell so spezifiziert, dass die gewünschte Mehr-Schritt-Prognose unmittelbar möglich ist. Am Beispiel eines NLAR(p)-Prozesses impliziert das ein Modell $x_{t+h} = \tilde{g}(x_t, x_{t-1}, \ldots, x_{t-\tilde{p}}) + \tilde{\varepsilon}_{t+h}$ und $x_{T+h|T} = \tilde{g}(x_t, x_{t-1}, \ldots, x_{t-\tilde{p}})$. Klarerweise gilt $\tilde{g} \neq g$ und wahrscheinlich $\tilde{p} \neq p$. Der mit dieser Methode verbundene Arbeitsaufwand ist enorm, da für jeden gewünschten Prognosehorizont h ein eigenes Modell spezifiziert und geschätzt

werden muss. Aus den weiter unten folgenden Erläuterungen wird zudem klar, dass eine Reihe von Verallgemeinerungen der Modellformulierung speziell im Fall der ARNN-Modelle notwendig werden.

Die Frage, ob und unter welchen Bedingungen die direkte Methode den verschiedenen Varianten der iterierten Methode überlegen ist, wird in der aktuellen Forschung nicht eindeutig beantwortet. Clements und Hendry (1998) favorisieren die direkte Methode aufgrund allgemeiner Überlegungen, vor allem wegen der Gefahr, dass sich Probleme mit einer Missspezifikation des geschätzen Modells im Zuge der Iteration von Ein-Schritt-Prognosen verstärken können. Clive W. Granger fasst in einem Interview seine Einschätzung zu dieser Frage zusammen und meint, dass zumindest für nicht-lineare Modelle die direkte Methode besser sein sollte (Phillips, 1997). Im Zusammenhang mit neuronalen Netzen berichtet Zhang (1994), dass die direkte Methode in einer Anwendung überlegen war, anders hingegen Weigend et al. (1992). Eine Simulationsstudie von Lin und Granger (1994), die in Granger und Teräsvirta (1993) zusammengefasst wird, zeigt immerhin klar, dass die naiv-iterative Methode die schlechteste Prognosegüte aufweist. Die anderen Methoden sind jedoch mit bedeutend mehr Rechenaufwand verbunden.

Marcellino et al. (2006) untersuchen die Frage, ob iterierte oder direkte Prognosen besser sind, unter der Einschränkung auf lineare Modelle. Dies sind AR-Modelle für die Erstellung der iterierten Prognosen, so dass die Notwendigkeit für simulativ-iterierte Methoden wegfällt, sowie lineare Regressionen für die Erstellung der direkten Prognosen. In diesem Setting hängt die relative Vorteilhaftigkeit der beiden Ansätze ausschließlich von der Frage der Missspezifikation der AR-Modelle ab. Eine auf einem richtig spezifizierten AR-Modell beruhende iterierte Prognose sollte immer überlegen sein. Die empirische Fragstellung wird durch die Autoren anhand eines Sets von insgesamt 170 monatlichen US-amerikanischen makroökonomischen Zeitreihen untersucht. Sie kommen zum Ergebnis, dass die iterierten Prognosen auf Basis von AR-Modellen gemessen am MSFE tendenziell überlegen sind. Zudem nimmt die relative Güte der direkten Prognosen mit zunehmendem Prognosehorizont ab. Es ergeben sich somit keine starken Hinweise auf eine Fehlspezifikation von AR-Modellen (und damit auch kaum Hinweise auf Nicht-Linearität) auf dem Großteil der untersuchten Zeitreihen.

Mehr-Schritt-Prognosen auf differenzierten Zeitreihen

Die Darstellung der Problematik von Mehr-Schritt-Prognosen basierte bisher auf der vereinfachenden Annahme, dass die modellierte Zeitreihe mit der zu prognostizierenden Zeitreihe übereinstimmt. Dies ist dann nicht der Fall, wenn die Zeitreihe einmal oder mehrmals differenziert wird um Stationarität herzustellen und das Modell auf der differenzierten Zeitreihe geschätzt wird, wie es in der Praxis häufig vorkommt. In diesem Fall muss man, um zur Prognose der Originalzeitreihe zu gelangen, die Prognose der differenzierten Zeitreihe nehmen und mit Hilfe der bekannten Werte

der originalen Zeitreihe auf die Prognose der Originalzeitreihe zurückrechnen. Wie sich zeigt, verkompliziert sich die Situation in Abhängigkeit vom angewendeten Differenzenfilter und vom gewünschten Prognosehorizont. So kann die Rückrechnung sowohl mehrere prognostizierte Werte der differenzierten Zeitreihe als auch mehrere vergangene Werte der Originalzeitreihe erfordern.

Marcellino et al. (2006) behandeln das Problem sowohl für iterierte AR-Prognosen als auch für direkte Prognosen jeweils für die Fälle von $I(0)$, $I(1)$ und $I(2)$ Zeitreihen, auf die ihre Anwendungsstudie einschränkt. Es sei Y_t die Originalzeitreihe (in Niveaus oder Logarithmen) und y_t die stationäre Transformation, also $y_t = \nabla^d Y_t$, wobei $d = 0, 1, 2$. Die Vorhersage, die zum Zeitpunkt T mit einem Prognosehorizont h erstellt wird, sei mit $Y_{T+h|T}$ bzw. $y_{T+h|T}$ bezeichnet. Damit lässt sich die Formel für die Rückrechnung von den Prognosen der differenzierten Zeitreihe auf die Prognose der Originalzeitreihe relativ übersichtlich in der folgenden Formel ausdrücken:[31]

$$Y_{T+h|T} = \begin{cases} y_{T+h|T} & \text{wenn } Y_t\ I(0) \text{ ist,} \\ Y_T + \sum_{i=1}^{h} y_{T+i|T} & \text{wenn } Y_t\ I(1) \text{ ist,} \\ Y_T + h\nabla Y_T + \sum_{i=1}^{h} \sum_{j=1}^{i} y_{T+j|T} & \text{wenn } Y_t\ I(2) \text{ ist.} \end{cases} \tag{5.8}$$

In ähnlicher Weise legen Marcellino et al. (2006) auch die Vorgangsweise bei der Erstellung von direkten Prognosen formelmäßig nieder. Hier ist die abhängige Variable des Modells, y_{t+h}^h, zu bestimmen, wobei die unabhängigen Variablen $y_t, y_{t-1}, \ldots$ sind:

$$y_{t+h}^h = \begin{cases} y_{t+h} = Y_{t+h} & \text{wenn } Y_t\ I(0) \text{ ist,} \\ \sum_{i=1}^{h} y_{t+i} = Y_{t+h} - Y_t & \text{wenn } Y_t\ I(1) \text{ ist,} \\ \sum_{i=1}^{h} \sum_{j=1}^{i} y_{t+j} = Y_{t+h} - Y_t - h\nabla Y_t & \text{wenn } Y_t\ I(2) \text{ ist.} \end{cases} \tag{5.9}$$

Bezeichnet man die Prognose von y_{t+h}^h, welche auf der Basis des geschätzten Modells und mit dem Prognoseursprung T erstellt wird, mit $y_{T+h|T}^h$, so ergeben sich die Formeln für die Rückrechnung auf die Prognose der Originalzeitreihe als geeignete Zusammenfassungen von $y_{T+h|T}^h$ und $Y_T, Y_{T-1}, \ldots$:

$$Y_{T+h|T} = \begin{cases} y_{T+h|T}^h & \text{wenn } Y_t\ I(0) \text{ ist,} \\ Y_T + y_{T+h|T}^h & \text{wenn } Y_t\ I(1) \text{ ist,} \\ Y_T + h\nabla Y_T + y_{T+h|T}^h & \text{wenn } Y_t\ I(2) \text{ ist.} \end{cases} \tag{5.10}$$

Aus mehreren Gründen ist die Behandlung des Problems der Mehr-Schritt-Prognosen nach Gleichung (5.8) bis (5.10) ungenügend für die Zwecke der vorliegenden Arbeit. Zum einen ist es wünschenswert, eine Generalisierung der Darstellung auch für saisonale Differenzenfilter und für die mehrmalige Anwendung

[31] Die Notation in dieser und in den beiden folgenden Gleichungen enthält im Vergleich zum Original bei Marcellino et al. (2006) einige kleinere Anpassungen.

der Differenzenfilter zu erreichen. Zum anderen erfordert die Verallgemeinerung des ARNN-Modells für direkte Mehr-Schritt-Prognosen auch eine Berücksichtigung des deterministischen Teils bei der Rückrechnung von der Prognose der differenzierten Zeitreihe auf die Prognose der originalen Zeitreihe (siehe hierzu weiter unten). Daher wird nun eine neuartige Notation eingeführt, die auf Lag-Polynomen beruht und es erlaubt, das Problem in knapper und übersichtlicher Weise darzustellen.

Es seien die Originalzeitreihe Y_t und die differenzierte Zeitreihe y_t gegeben. Mit $\tilde{Y}_t$ sei die um h Perioden verzögerte Zeitreihe bezeichnet: $\tilde{Y}_t = B^h Y_t = Y_{t-h}$, wobei B der Lag-Operator ist (vgl. Abschnitt 2.1). Dann lässt sich die Originalzeitreihe aus $\tilde{Y}_t$ und y_t durch die Anwendung von geeigneten Lag-Polynomen, $\theta(B)$ und $\phi(B)$, zusammensetzen:

$$Y_t = \theta(B)\tilde{Y}_t + \phi(B)y_t. \tag{5.11}$$

Damit wird gezeigt, dass die Wiedergewinnung der Originalzeitreihe ohne die direkte Kenntnis der jeweils letzten h Werte möglich ist. Im Unterschied zu den Lag-Polynomen, die in Abschnitt 2.1 eingeführt wurden, werden die Koeffizienten der Lag-Polynome $\theta(B)$ und $\phi(B)$ nicht geschätzt sondern bestimmt (siehe genaueres weiter unten).

Nach dem gleichen Muster kann auch $Y_{T+h|T}$, die zum Prognoseursprung T erstellte h-Schritt-Prognose, als die Summe eines Lag-Polynoms von Y_T und eines Lag-Polynoms von $y_{T+h|T}$ angeschrieben werden, wobei allerdings beim Lag-Operator danach unterschieden werden muss, welcher der beiden im Subskript vorkommenden Zeitreihenindizes verzögert werden soll, was durch die Verwendung eines entsprechenden Subskripts beim Lag-Operator zum Ausdruck gebracht wird:

$$Y_{T+h|T} = \theta(B_T)Y_T + \phi(B_h)y_{T+h|T}. \tag{5.12}$$

Es ist somit zu beachten, dass sich die durch das Lag-Polynom spezifizierten Verzögerungen im Falle von $\theta(B_T)$ auf den Zeitindex T beziehen, während durch die Schreibweise $\phi(B_h)$ angezeigt wird, dass der Zeitindex durch die Veränderung von h bei gleichbleibendem T angepasst wird. Mit dieser Schreibweise ist Gleichung (5.12) eine Verallgemeinerung der Gleichung (5.8).

Zur Bestimmung der Ordnung der beiden Lag-Polynome und ihrer einzelnen (ganzzahligen) Koeffizienten können in der hier gebotenen Kürze die folgenden Aussagen gemacht werden. $\phi(B_h)$ ist maximal vom Grade h, so dass

$$\phi(B_h)y_{T+h|T} = \phi_1 y_{T+h|T} + \phi_2 y_{T+h-1|T} + \ldots + \phi_h y_{T+1|T}. \tag{5.13}$$

Der maximale Grad von $\theta(B_T)$ wird durch die angewendete Differenzierung bestimmt. Schreibt man die Art der Differenzierung allgemein als $\nabla_s^n \nabla^m$ an, wobei ∇_s^n die n-malige Anwendung des saisonalen Differenzenfilters und ∇^m die m-malige Anwendung des gewöhnlichen Differenzenfilters bezeichnet, so ist der Grad von $\theta(B_T)$ maximal $ns + m$, kann aber auch in Abhängigkeit von h geringer sein. Für einfache Differenzenfilter und geringe Werte für den Prognosehorizont h ist die Gestalt

Tab. 19: Koeffizienten der für Mehr-Schritt-Prognosen benötigten Lag-Polnome in Abhängigkeit von Differenzenfilter und Prognosehorizont

	∇:		∇_4:		$\nabla_4\nabla$:	
	θ	ϕ	θ	ϕ	θ	ϕ
$h=1$	(1)	(1)	$(0,0,0,1)$	(1)	$(1,0,0,1,-1)$	(1)
$h=2$	(1)	$(1,1)$	$(0,0,1)$	(1)	$(1,0,1,0,-1)$	$(1,1)$
$h=3$	(1)	$(1,1,1)$	$(0,1)$	(1)	$(1,1,0,0,-1)$	$(1,1,1)$
$h=4$	(1)	$(1,1,1,1)$	(1)	(1)	$(2,0,0,0,-1)$	$(1,1,1,1)$
$h=5$	(1)	$(1,1,1,1,1)$	$(0,0,0,1)$	$(1,0,0,0,1)$	$(2,0,0,1,-2)$	$(1,1,1,1,2)$
$h=6$	(1)	$(1,1,1,1,1,1)$	$(0,0,1)$	$(1,0,0,0,1)$	$(2,0,1,0,-2)$	$(1,1,1,1,2,2)$
$h=7$	(1)	$(1,1,1,1,1,1,1)$	$(0,1)$	$(1,0,0,0,1)$	$(2,1,0,0,-2)$	$(1,1,1,1,2,2,2)$
$h=8$	(1)	$(1,1,1,1,1,1,1,1)$	(1)	$(1,0,0,0,1)$	$(3,0,0,0,-2)$	$(1,1,1,1,2,2,2,2)$
$h=9$	(1)	$(1,1,1,1,1,1,1,1,1)$	$(0,0,0,1)$	$(1,0,0,0,1,0,0,0,1)$	$(3,0,0,1,-3)$	$(1,1,1,1,2,2,2,2,3)$
$h=10$	(1)	$(1,1,1,1,1,1,1,1,1,1)$	$(0,0,1)$	$(1,0,0,0,1,0,0,0,1)$	$(3,0,1,0,-3)$	$(1,1,1,1,2,2,2,2,3,3)$

der Lag-Polynome offensichtlich, kann aber bei komplizierten Differenzenfiltern und wachsendem h schnell unübersichtlich werden.

Eine allgemeine Formel für die Wahl der Koeffizienten der Polynome ist nicht leicht zu gewinnen. Die Bestimmung der Koeffizienten der Lag-Polynome erfolgt, indem zunächst $Y_{T+h|T}$ gemäß dem zugrundeliegenden Differenzenfilter in $y_{T+h|T}$ und ein Lag-Polynom auf $Y_{T+h-1|T}$ zerlegt wird und dann rekursiv alle in dieser Zerlegung auftauchenden $Y_{t|T}$ solange weiter zerlegt werden, bis kein $Y_{t|T}$ mit $t > T$ mehr auftaucht. In diesem Prozess ist $Y_{t|T} = Y_t$ für $t \leq T$ definiert. Die Koeffizienten der Lag-Polynome ergeben sich dann durch Abzählen der auftauchenden Terme.

In Tab. 19 werden für die drei in der vorliegenden Arbeit verwendeten Differenzenfilter die Koeffizientenvektoren θ und ϕ, die sich je nach Prognosehorizont ergeben, in Übersichtsform dargestellt, wobei allerdings um der Kürze der Darstellung willen vierteljährliche statt monatliche Zeitreihen angenommen werden.

Direkte Mehr-Schritt-Prognosen mit ARNN-Modellen

Die Erstellung von direkten Mehr-Schritt-Prognosen erfordert eine grundlegende Anpassung der ARNN-Modelle (bzw. ARNNDS-Modelle), die wie im Fall der iterierten Mehr-Schritt-Prognosen auf der Notation mit Hilfe von Lag-Polynomen beruht. Sowohl die Art der Differenzierung der Zeitreihe als auch Wahl des Prognosehorizonts müssen in der Spezifikation des Modells berücksichtigt werden.

Zuvor soll noch aufgezeigt werden, worin die Einschränkungen beim Versuch liegen, direkte Mehr-Schritt-Prognosen mit der herkömmlichen ARNN-Modellformulierung zu erstellen. Nur in einzelnen einfachen Fällen würde man mit einer adhoc Modifikation der ARNN-Modellformulierung auf der Basis von Gleichung (4.1) bzw. (4.2) auf Seite 75 das Auslangen finden. So wäre es bei einem Modell auf der gewöhnlich und saisonal differenzierten Zeitreihe, $y_t = \nabla_s Y_t$, eine Option, eine Mehr-Schritt-Prognose für Prognosehorizonte bis zu $h = 12$ Monate voraus zu erstellen, indem man dem Modell die Restriktion vorgibt, dass es keine verzögerten Werte bis zum Lag $h - 1$ enthalten darf. Denn in diesem Fall kann mit Hilfe der zum Zeitpunkt der Erstellung der Prognose T bekannten Werte eine Prognose $y_{T+h|T}$ erstellt werden (die eigentlich eine Ein-Schritt-Prognose $y_{T+h|T+h-1}$ darstellt, welche von den Werten $y_{T+1}, y_{T+2}, \ldots, y_{T+h-1}$ keinen Gebrauch macht). Die Prognose auf der Originalzeitreihe ergibt sich dann als $Y_{T+h|T} = Y_T + y_{T+h|T}$. Aber die Unübersichtlichkeit dieses Ansatzes lässt ihn für den Praktiker nicht empfehlenswert erscheinen. Wie bereits oben gezeigt wurde, steht bei den Differenzierungen ∇ und $\nabla_s \nabla$ diese Option nicht zur Verfügung, da die Rückrechnung auf die Originalzeitreihe Y_t der Prognose eines Aggregats von y_t bedarf.

Ein umfassender Ansatz der Erstellung von Mehr-Schritt-Prognosen erweitert das ARNN-Modell um die integrierte Berücksichtigung des benötigten Aggregats von y_t in Form eines Lag-Polynoms $\phi(B)$ der Ordnung h, gemäß den in Tab. 19 dargestellten Erfordernissen. Bezeichnet man dieses Aggregat mit $y^h_{t+h-1} = \phi(B) y_{t+h-1}$, so ist die Zielsetzung des Modells eine Anpassung bzw. Prognose von y^h_{t+h-1}. Das generalisierte ARNN-Modell wird wie folgt angeschrieben:[32]

$$\tilde{y}_t = y_t - d_t,$$

$$\phi(B)\tilde{y}_{t+h-1} = \sum_{k=1}^{K} c_k \psi[b_k + \sum_{j \in \mathcal{J}} a_{kj} \tilde{y}_{t-j}] + \sum_{l \in \mathcal{L}} f_l \tilde{y}_{t-l} + \varepsilon_t, \tag{5.14}$$

wobei d_t den deterministischen Teil der Zeitreihe bezeichnet. Im Falle des ARNN-Modells ist dieser ein konstanter Mittelwert, $d_t = d$. Beim ARNNDS-Modell gilt $d_t = d + d^t t + \sum_{i=2}^{s} d_i^s s_{it}$. Ansonsten sei für die Erklärung der Symbole auf die Erläuterungen zu den Gleichungen (4.1) und (4.2) in Kapitel 4 verwiesen.

Es ist hilfreich, bei dieser Modellformulierung zu unterscheiden zwischen dem deterministischen bzw. stochastischen Teil der Zeitreihe y_t einerseits und dem deterministischen bzw. stochastischen Teil des Aggregats y^h_{t+h-1} andererseits. Wie y_t lässt sich auch y^h_{t+h-1} in einen stochastischen und deterministischen Teil, $\tilde{y}^h_{t+h-1}$ bzw. d^h_{t+h-1}, zerlegen:

$$y^h_{t+h-1} = \tilde{y}^h_{t+h-1} + d^h_{t+h-1} = \phi(B)\tilde{y}_{t+h-1} + \phi(B) d_{t+h-1}. \tag{5.15}$$

[32] Die Notation und die Setzung der Zeitindizes wurde so gewählt, dass für den einfachen Fall der Ein-Schritt-Prognose, $h = 1$, das gewünschte Aggregat zu $y^h_{t+h-1} = y_t$ kollabiert, was der üblichen Notation entspricht.

Mit dieser Modellformulierung können unter Verwendung der geschätzten Koeffizienten die angepassten Werte für y^h_{t+h-1} wie folgt angeschrieben werden:

$$\hat{y}^h_{t+h-1} = \hat{\tilde{y}}^h_{t+h-1} + \hat{d}^h_{t+h-1}, \tag{5.16}$$

wobei

$$\hat{\tilde{y}}^h_{t+h-1} = \sum_{k=1}^{K} \hat{c}_k \psi[\hat{b}_k + \sum_{j \in \mathcal{J}} \hat{a}_{kj} \hat{\tilde{y}}_{t-j}] + \sum_{l \in \mathcal{L}} \hat{f}_l \hat{\tilde{y}}_{t-l}$$

und

$$\hat{d}^h_{t+h-1} = \phi(B)\hat{d}_{t+h-1}.$$

Die zum Zeitpunkt T erstellte Prognose von y^h_{T+h} erhält man als

$$y^h_{T+h|T} = \hat{\tilde{y}}^h_{T+h} + \hat{d}^h_{T+h}, \tag{5.17}$$

wobei man sich durch Einsetzen davon überzeugen kann, dass tatsächlich nur Information bis zum Zeitpunkt T verwendet wurde. Die Rückrechnung auf die Prognose der Originalzeitreihe erfolgt mit

$$Y^h_{T+h|T} = \theta(B)Y_T + y^h_{T+h|T}, \tag{5.18}$$

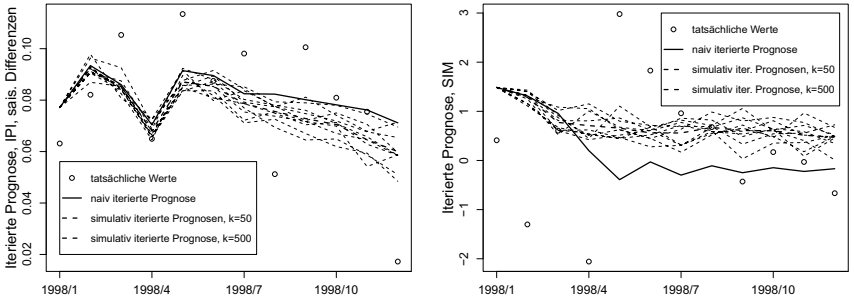
wobei $\theta(B)$ den Erfordernissen gemäß Tab. 19 entspricht.

Im R-Paket `NNUTS` unterstützen die Funktion `arnnds` und einige weitere Funktionen die Möglichkeit für direkte Prognosen von Aggregaten $y^h_{t+h-1} = \phi(B)y_{t+h-1}$, wobei die Koeffizienten des Lag-Polynoms $\phi(B)$ vom Benutzer spezifiziert werden müssen. Die Rückrechnung auf die Originalzeitreihe muss separat bewerkstelligt werden.

Iterierte Mehr-Schritt-Prognosen mit ARNN-Modellen

Zum Abschluss dieses Abschnitts wird die Methode der iterierten Mehr-Schritt-Prognose in einer Anwendung auf Besispielzeitreihen untersucht. Es soll gezeigt werden, wie die naiv-iterierte Methode einerseits und die simulativ iterierte Methode andererseits funktionieren, wenn in der Zeitreihe Nicht-Linearität in einem Ausmaß vorhanden ist, wie man sie in makroökonomischen Zeitreihen ungefähr erwarten kann. Die Methode der direkt iterierten Prognosen kann hier nicht berücksichtigt werden, da sie die Problemstellung der Modellspezifikation spezifisch für jeden einzelnen Prognosehorizont benötigt und daher erst im Abschnitt 5.3 aus einer angewandten Perspektive behandelt wird.

Es wird auf den Zeitreihen SIM und IPI jeweils ein ARNN-Modell geschätzt und damit zum Prognoseursprung $T = 1997/12$ eine Prognose bis zum Prognosehorizont $h = 12$ erstellt. Abb. 15 illustriert die Ergebnisse. Für die Erstellung der Vorhersage von SIM hat das Modell ein einzelnes verdecktes Neuron sowie Direktverbindungen,



(a) IPI (b) SIM

Abb. 15: Iterierte Prognosen auf Basis eines ARNN-Modells: (a) Industrieproduktionsindex (IPI), saisonale Differenzen und (b) simulierte nichlineare Zeitreihe (SIM)

die jeweils mit einem Input zum Lag 1 verbunden sind. Das Modell zur Vorhersage von IPI wurde auf den saisonalen Differenzen mit den bereits im vorigen Kapitel beschriebenen Methoden geschätzt. Die naiv-iterierten Vorhersagen wurden in fetter Linie eingezeichnet, die simulativ-iterierten mit strichlierter Linie. Es wurde die Bootstrap-Methode verwendet, die in dieser Anwendung im Vergleich zur Monte-Carlo-Methode keine erkennbaren Unterschiede liefert. Die Anzahl der im Rahmen der Simulation generierten Vorhersagen wird mit k bezeichnet. Im Falle der grau strichlierten Linien wird die berechnete Vorhersage aus Mittelung über $k = 50$ Prognosezeitreihen ermittelt, welche erwartungsgemäß um die schwarz strichlierte Linie herum schwanken, die ihrerseits aus Mittelung über $k = 500$ Prognosezeitreihen gewonnen wurde. Zusätzlich sind die tatsächlichen Realisationen der Zeitreihe als Punkte eingezeichnet.

Die Anwendung zeigt, dass relevante Unterschiede zwischen der naiv-iterierten und der simulativ-iterierten Prognosemethode bestehen. Diese werden allerdings erst ab einem mittleren Prognosehorizont von vier bzw. sechs Monaten (bei monatlichen Daten) deutlich. Die Unterschiede fallen indes im Verhältnis zum gesamten Vorhersagefehler kaum ins Gewicht. Für die Zeitreihe SIM ist die Abweichung der beiden Prognosemethoden voneinander besonders stark. Dies ist ein Hinweis, dass die Wahl der Methode insbesondere dann relevant ist, wenn tatsächliche Nicht-Linearität, wie sie von ARNN-Modellen modelliert werden kann, in den Daten vorhanden ist.

Für den Praktiker stellt sich auch die Frage, wie groß bei der simulativ-iterierten Methode der Simulationsumfang der Simulation k gewählt werden soll. Die Antwort hängt in erster Linie von der verfügbaren Rechenkraft ab. Da jedoch der Rechenaufwand (gemessen an der CPU-Zeit) für die simuliert-iterativen Methoden in typischen Szenarien nur einen Bruchteil der Rechenzeit der ARNN-Modellierung

ausmacht, sollte der Simulationsumfang eher groß gewählt werden. Die Beispiele in Abb. 15 legen einen Simulationsumfang von mindestens 100 nahe, da erst ab diesem Bereich die Unterschiede zur naiv-iterierten Methode deutlich sichtbar werden.

5.3 Anordnung der Prognoseerstellung zum Zweck der Evaluierung

In diesem Abschnitt wird die Anordnung zur Erstellung der Prognosen beschrieben, die für die Durchführung eines fairen und umfassenden Vergleichs der Prognosegüte herangezogen werden. Zuerst wird die grundlegende Struktur der Prognoseanordnung beschrieben. Dies umfasst die Vorgangsweise bei der Unterteilung der Zeitreihe in das Lernset („In-Sample“) und das Evaluierungsset(„Out-of-Sample“), sowie andere Aspekte, die unabhängig von der gewählten Modellierungsstrategie sind. Sodann wird auf die Frage der Bestimmung der Lags bei Ein- und Mehr-Schritt-Prognosen mit ARNN-Modellen eingegangen. Auch diese Frage wird für alle Modellierungsstrategien gleich gehandhabt, wenn auch in diesem Falle aus Gründen der Vereinfachung der Prognoseanordnung. Schließlich werden die getesteten Modellierungsstrategien überblicksmäßig aufgestellt.

Grundstruktur der Prognoseanordnung

Die Grundstruktur der Prognoseanordnung betrifft insbesondere die Unterteilung der gesamten zur Verfügung stehenden Zeitreihe in das Lernset und das Evaluierungsset. Das Lernset beginnt im Jänner 1960 und endet im Dezember 1997. Als alternative Modellierungsoption wird die Evaluierung auch mit einem verkürzten Lernset durchgeführt, welches erst im Jänner 1975 beginnt. Das Evaluierungsset umfasst die Jahre 1998 bis 2002, das sind 60 Datenpunkte. Doch für die Prognoseanordnung ist dies nicht die einzige verwendete Unterteilung, sondern es wird im Rahmen eines rollierenden Prognoseursprungs die Unterteilung dynamisch gestaltet.

Es interessieren Prognosen für die Prognosehorizonte $h = 1, \ldots, 12$. Prognosen darüber hinaus, d.h. bei monatlichen Zeitreihen mehr als ein Jahr in die Zukunft, werden in der angewandten Wirtschaftsprognoseforschung selten benötigt und werden daher auch hier nicht berücksichtigt. Um möglichst viele Prognosen zu einem bestimmten Prognosehorizont h zu erstellen, durchläuft der Prognoseursprung T alle Zeitpunkte von 1997/12 bis 2002/11. Bei jedem gewähltem T werden jeweils für die zwölf verschiedenen Prognosehorizonte Prognosen erstellt, wobei das Prognosziel $T + h$ noch innerhalb des Evaluierungssets liegen muss. Fasst man die Prognosen, die mit einer bestimmten Methode und mit gleichem Prognosehorizont h erstellt werden, zu einer Zeitreihe zusammen, so liegt jede solche Zeitreihe mit einer Länge von $60 - h + 1$ Datenpunkten vor. Die Prognosezeitreihe beginnt im h-ten Monat des ersten Jahres des (ersten) Evaluierungssets und endet im Dezember 2002.

Die Modellierung könnte prinzipiell alle bis zum Zeitpunkt T vorhandenen Beobachtungen der Zeitreihe für Modellspezifikation, Modellschätzung und Prognose nutzen. Doch wird dies aus Gründen der Vereinfachung und nicht vollständig getan, indem der Prozess der Modellspezifikation und -schätzung für jedes Quartal nur einmal durchgeführt wird.[33] Für die beiden nachfolgenden Monate des Quartals wird nur die Zeitreihe aktualisiert und nicht das Modell selbst. Dies kann folgendermaßen verdeutlicht werden. Man bezeichne mit $Y_{T+h|M_t,T}$ die Prognose, die zum Prognoseursprung T mit einem Prognosehorizont h auf der Basis eines Modells M_t und der bis T aktualisierten Zeitreihe erstellt wird, wobei das Modell auf einem im Zeitpunkt t endenden Zeitreihenabschnitt geschätzt wurde. Dann kann die Zeitreihe der im Rahmen der Evaluierungsanordnung erstellten Prognosen mit Prognosehorizont h so angeschrieben werden:

$$\begin{aligned}&(Y_{T^*+h|M_{T^*},T^*},\ Y_{T^*+h+1|M_{T^*},T^*+1},\ Y_{T^*+h+2|M_{T^*},T^*+2},\\ &Y_{T^*+h+3|M_{T^*+3},T^*+3},\ Y_{T^*+h+4|M_{T^*+3},T^*+4},\ Y_{T^*+h+5|M_{T^*+3},T^*+5},\ \ldots),\end{aligned}$$

wobei T^* für den Dezember 1997 steht.

Diese Vorgangsweise beruht auf der plausiblen Annahme, dass die Unterschiede zwischen den zu verschiedenen Prognoseursprüngen erstellten Prognosen hinsichtlich der Ausnützung der Informationsmenge für die Prognosegüte zu vernachlässigen sind und dass daher die Verteilung der Prognosefehler nur vom Prognosehorizont und der gewählten Modellstrategie abhängt.

Neben der oben beschriebenen Vereinfachung verwendet die Prognoseanordnung noch an einigen anderen Stellen Vereinfachungen. Zum Beispiel werden einzelne Hyperparameter des Regularisierungsansatzes nicht für jedes Quartal neu ermittelt, sondern bloß einmal für jedes Jahr. Eine weitere Vereinfachung betrifft die Menge der Lags, die für die Spezifizierung der Lags im den linearen und ARNN-Modellen in Frage kommen. Auch diese werden nicht für jedes Quartal neu bestimmt, sondern bloß für jedes Jahr insgesamt. Da dieser Aspekt wichtiger erscheint, wird er im folgenden Unterabschnitt detaillierter behandelt.

Bestimmung der Lags bei Ein- und Mehr-Schritt-Prognosen

Die prinzipielle Vorgangsweise bei der Bestimmung der in einem AR- oder ARNN-Modell (bzw. deren Erweiterungen zu ARDS- und ARNNDS-Modellen) enthaltenen Lags wurde bereits in Kap. 2 und 4 behandelt. Beim AR-Modell wird zunächst die Modellordnung bestimmt und danach in einem Pruningprozess ein sparsam spezifiziertes Modell gesucht, das nur signifikante Lags enthält. Für die ARNN- bzw. ARNNDS-Modelle wird die vereinfachende Annahme getroffen, dass die so gefundene Lag-Liste auch jene Lags enthält, die möglicherweise für nicht-lineare Strukturen verantwortlich sind. Die Lag-Liste des linearen Modells stellt deshalb

[33] Diese Vorgangsweise entspricht ungefähr der Praxis der Konjunkturprognose durch Wirtschaftsforschungsinstitute, die jeweils einmal im Quartal Prognosen erstellen.

auch den Ausgangspunkt für die Spezifikation der Lags im entsprechenden ARNN-Modell dar.

Im Falle der Erstellung von direkten Mehr-Schritt-Prognosen und der dafür speziell entwickelten Modelle ist für jeden Prognosehorizont h die Lag-Liste neu zu spezifizieren. Das Problem ist von einer höheren Komplexität und verdient an dieser Stelle eine gesonderte Behandlung, zumal sich an den Beispielzeitreihen interessante Muster zeigen. Der Fall der direkten Mehr-Schritt-Prognose schließt dabei als Spezialfall die Ein-Schritt-Prognose, $h = 1$, ein.

In Tab. 20 werden die Ergebnisse für die Modellordnung und die Anzahl der enthaltenen Lags der Modelle der Prognoseanordnung überblicksmäßig dargestellt. Die Tabelle berichtet jeweils zusammenfassende Statistiken zur Modellordnung und der Anzahl der enthaltenen Lags in den Modellspezifikationen, die für die fünf verschiedenen berücksichtigten Endzeitpunkte des Lernsets, T = Dez1997, Dez1998, ..., Dez2001, ermittelt werden. Neben dem Median sind in Klammer der kleinste und der größte Wert angegeben. Die Tabelle zeigt, wie Modellordnung und Anzahl der enthaltenen Lags bei der gegebenen Spezifizierungsmethode vom zugrundeliegenden Differenzenfilter und dem Prognosehorizont h abhängen.

Die Tabelle offenbart unterschiedliche Muster bei den verschiedenen Differenzenfiltern. Im Falle der Zeitreihe IPI wachsen für Modelle auf der Basis der konventionellen Differenzierung, ∇, die Modellordnung und die Anzahl der Koeffizienten, die nach dem Pruningverfahren in der Modellspezifikation verbleiben, zunächst an, um ab einem Prognosehorizont von $h = 6$ wieder zu sinken. Für Prognosehorizonte $h = 11, 12$ ist nur mehr ein einziger Koeffizient (bei Lag 1) bei einem Wahrscheinlichkeitsniveau von 5 % signifikant. Bei der ALR nimmt sowohl die Modellordnung als auch die Anzahl der signifikant enthaltenen Koeffizienten mit wachsendem h kontinuierlich ab. Im Falle der Zeitreihe IPI und dem Differenzenfilter ∇_{12} fällt die Modellordnung mit wachsendem h von anfangs 49 auf schließlich 13. Ähnlich präsentiert sich die Situation für die ALR, wo die Modellordnung von anfangs 50 auf 20 sinkt. Auch die Anzahl der nach dem Pruningverfahren verbleibenden Lags sinkt bei Modellen unter Verwendung des Differenzenfilters ∇_{12} auf niedrige Werte ab, dies bereits ab mittleren Prognosehorizonten. In jedem dieser vier Fälle erhält man für mittlere und hohe Prognosehorizonte relativ sparsame Lag-Spezifikationen.

Gänzlich anders präsentiert sich das Bild für die Differenzierungsform $\nabla_{12}\nabla_1$. Hier bleibt die Modellordnung mit wachsendem h über 30 während die Anzahl der signifikanten Koeffizienten durchwegs zunimmt und bei $h = 12$ im Median 34 beträgt. Im Extremfall findet die automatisierte AR-Spezifizierung ein AR-Modell der Ordnung 50, dessen sämtliche Koeffizienten bei einem Niveau von 5 % signifikant von Null verschieden sind. Das gleiche Muster der Abhängigkeit der Modellordnung und Lag-Spezifikation vom Prognosehorizont h erhält man auch, wenn man die obenstehende Analyse auch für Zeitreihen durchführt, die erst ab dem Jahr 1975 beginnen.

Tab. 20: Modellordnung und Anzahl der enthaltenen Lags in den Modellen der Evaluierungsanordnung, ALR und IPI

	∇:		∇_{12}:		$\nabla_{12}\nabla$:	
	Ordnung	Anzahl Lags	Ordnung	Anzahl Lags	Ordnung	Anzahl Lags
IPI:						
$h = 1$	14(14,59)	5(5,11)	49(49,49)	16(15,17)	48(48,48)	13(12,15)
$h = 2$	19(13,19)	9(5,11)	37(37,41)	16(13,17)	36(36,36)	12(10,16)
$h = 3$	19(18,19)	10(9,10)	34(34,34)	13(12,14)	35(35,35)	11(11,18)
$h = 4$	30(30,30)	11(10,14)	38(35,39)	14(14,15)	39(34,39)	16(16,20)
$h = 5$	31(31,31)	13(13,13)	29(25,32)	11(8,13)	33(33,36)	18(15,19)
$h = 6$	30(30,30)	13(13,13)	31(27,31)	10(9,10)	32(32,35)	19(19,22)
$h = 7$	29(17,30)	13(12,13)	18(18,18)	7(7,7)	34(31,42)	20(18,22)
$h = 8$	16(16,28)	10(9,11)	17(17,17)	7(7,7)	30(30,30)	21(21,21)
$h = 9$	15(15,16)	8(7,11)	16(16,16)	6(4,6)	26(26,40)	24(21,24)
$h = 10$	3(3,6)	3(3,6)	15(15,15)	6(6,6)	25(25,48)	23(23,43)
$h = 11$	1(1,1)	1(1,1)	14(14,14)	4(4,5)	34(26,48)	27(26,35)
$h = 12$	1(1,1)	1(1,1)	13(13,13)	2(1,2)	34(26,50)	30(24,50)
ALR:						
$h = 1$	60(60,60)	10(10,16)	50(49,50)	18(18,19)	52(49,55)	15(14,18)
$h = 2$	42(42,48)	13(13,18)	32(24,32)	13(10,14)	62(62,62)	13(13,14)
$h = 3$	35(35,35)	15(15,15)	28(28,31)	9(7,10)	24(24,55)	11(11,19)
$h = 4$	35(34,35)	18(16,18)	25(25,25)	6(6,6)	24(24,24)	13(13,13)
$h = 5$	35(35,35)	16(16,16)	25(25,25)	5(5,5)	32(32,32)	14(14,14)
$h = 6$	35(35,35)	17(16,19)	25(25,25)	5(5,6)	32(32,32)	18(17,18)
$h = 7$	21(17,33)	15(14,18)	22(22,25)	4(4,4)	24(24,24)	17(17,17)
$h = 8$	20(14,29)	15(13,16)	21(21,21)	4(4,4)	32(32,32)	18(18,19)
$h = 9$	16(16,16)	12(11,13)	20(20,20)	4(4,4)	23(22,24)	19(17,21)
$h = 10$	18(15,29)	9(6,20)	22(22,22)	4(4,4)	24(24,24)	22(22,22)
$h = 11$	17(14,17)	9(3,9)	21(21,21)	4(4,4)	24(24,24)	23(23,23)
$h = 12$	16(15,29)	6(4,8)	20(20,20)	3(3,4)	28(28,28)	27(25,27)

Mediane sowie in Klammern Minimum und Maximum der Modellordnung bzw. der Anzahl der im Modell enthaltenen Lags von AR- bzw. ARDS-Modellen, gerechnet über fünf unterschiedliche Samples

Die Untersuchung der Ursachen für diese auffälligen Unterschiede zwischen den verschiedenen Differenzierungsformen würde über die Fragestellung der Arbeit hinausgehen. Eine zentrale Rolle dürfte jedoch die Überdifferenzierung durch die Anwendung von $\nabla_{12}\nabla_1$ und die dadurch bedingte Einführung von MA-Anteilen in die Zeitreihe spielen.

Für die Anwendbarkeit der Modellierungsstrategien auf den Fall von direkten Mehr-Schritt-Prognosen hat die Unterschiedlichkeit der ermittelten Lag-Listen große Konsequenzen. Während für ein rein lineares Modell eine große Anzahl von Parametern zumindest aus numerischer Sicht nicht prohibitiv ist, ist die Modellierung und Schätzung eines ARNN-Modells mit mehr als 30 Lags auch bei der Anwendung der im Rahmen der vorliegenden Arbeit vorgeschlagenen Techniken schwer vorstellbar. Im konkreten wurde nach verschiedenen fehlgeschlagenen Versuchen, numerisch stabile Methoden für ARNN-Modelle mit $h > 1$ anhand der Differenzierungsform $\nabla_{12}\nabla_1$ zu entwickeln, beschlossen, in die Evaluierungsanordnung direkte Mehr-Schritt-Prognosen nur für ∇_1 und ∇_{12} aufzunehmen.

Modellstrategien der Prognoseanordnung

In diesem Unterabschnitt werden die Modellstrategien beschrieben, die für die linearen und ARNN-Modellierungsansätze im Rahmen der Prognoseanordnung verwendet werden.

Lineare Ansätze. Hierbei handelt es sich einerseits um AR-Modelle, andererseits um ARMA-Modelle, wobei erstere auch für die Erstellung von direkten Mehr-Schritt-Prognosen angepasst werden. Im Falle der konventionell differenzierten Zeitreihe, die deterministische Saisonalität enthält, wird zusätzlich ein entsprechender Term in das Modell aufgenommen (Erweiterung von AR zu ARDS und von ARMA zu ARMADS). Die Vorgangsweise der Modellierung umfasst im wesentlichen die Bestimmung der Lag-Struktur, die für die AR-Modelle bereits im vorangegangenen Unterabschnitt besprochen wurde und auch für die ARMA-Modelle anhand eines Pruning-Verfahrens auf der Basis einer ausreichend groß bestimmten Modellordnung erfolgt.

Statistisch-parametrischer Ansatz der ARNN-Modellierung. Es werden vier verschiedene Modellierungsstrategien definiert, die sich darin unterscheiden, ob Direktverbindungen (Shortcuts) in der Modellspezifikation enthalten sind oder nicht und welche Pruning-Methode zum Einsatz kommt. Das Pruning basiert auf den in Abschnitt 4.6 beschriebenen Vorgangsweisen und verlangt von den im Modell verbleibenden Parametern eine Signifikanz beim Konfidenzniveau von 5 %. Beim Pruning können die gefundenen Modellparameter nach jedem einzelnen Pruningschritt weiterverwendet werden, wodurch ein einmal gefundenes lokales Minimum über die Pruning-Schritte hinweg beibehalten wird, vorausgesetzt es existiert eine Entsprechung im reduzierten Parameterraum. Wird hingegen nach jedem Pruningschritt das ARNN-Modell neu initialisiert, so wird potentiell in jedem Schritt ein neues

lokales Minimum gefunden, für welches sich möglicherweise andere Lags als statistisch signifikant erweisen. Alle Modelle mit dem statistisch-parametrischen Ansatz werden mit nur einem einzigen verdeckten Neuron gerechnet, da der Teräsvirta-Lin-Granger-Test die Null-Hypothese der Linearität in jedem Fall sicher ablehnen kann, während er als Test auf Signifikanz eines zusätzlichen verdeckten Neurons bei bereits einem (unter der Null-Hypothese) vorhandenen verdeckten Neuron keine eindeutigen Ergebnisse bringt. Wie bei den linearen Ansätzen, so wird auch beim ARNN-Modell, d.h. nicht nur unter Anwendung des statistisch-parametrischen Ansatzes sondern auch der anderen Modellierungsansätze, im Falle der konventionell differenzierten Zeitreihe ein Modellteil zur Berücksichtigung von deterministischer Saisonalität aufgenommen (Erweiterung von ARNN zu ARNNDS). Als lokales Suchverfahren kommt das Quasi-Newton-Verfahren zum Einsatz. Die folgende Aufstellung gibt eine Übersicht über die vier Modellierungsstrategien:

Bezeichnung	Shortcuts	Pruningmethode
Param1	keine	kein Pruning
Param2	keine	mit Weiterverwendung der Parameter
Param3	vorhanden	mit Neuinitialisierung der Parameter
Param4	vorhanden	mit Weiterverwendung der Parameter

Klassischer Ansatz mit Early-Stopping. Es werden acht verschiedene Modellierungsstrategien verwendet, die sich hinsichtlich der Anzahl der eingesetzten verdeckten Neuronen, K, und der Shortcuts unterscheiden und ansonsten die in Abschnitt 4.7 dargelegte Vorgangsweise implementieren: epochenbasierter Backpropagation mit Lernrate, $\eta = 0.5$, und Momentum, $\zeta = 0.5$, sowie anwachsender Epochengröße. Es sei daran erinnert, dass sich die Lernrate auf die in jedem Iterationsschritt mittels Linesearch ermittelte optimale Schrittweite bezieht. Die Epochengröße beträgt zu Beginn des Trainings 50 % des Traininsset und konvergiert langsam nach 100 %. Für die Stoppregel wird eine Verhältniszahl des überflüssigen zum mindestens notwendigen Training von 50 % vorgegeben, wobei allerdings die Anzahl der mindestens zu durchlaufenden Iterationen mit 1000 recht hoch gewählt wird. Das Validierungsset umfasst die letzten 20 % des In-Sample-Bereichs der Zeitreihe. Von den acht Modellierungsstrategien weisen die ersten vier einen linearen Modellteil (Direktverbindungen bzw. Shortcuts) auf. Es kommen Modelle mit einer, zwei, fünf und zehn verdeckten Neuronen zum Einsatz. Die folgende Aufstellung zeigt, worin sich die acht Modellierungsstrategien unterscheiden:

Bezeichnung	Shortcuts	K	Bezeichnung	Shortcuts	K
Klass1	vorhanden	1	Klass5	keine	1
Klass2	vorhanden	2	Klass6	keine	2
Klass3	vorhanden	5	Klass7	keine	5
Klass4	vorhanden	10	Klass8	keine	10

Ansatz mit Regularisierung. Auch bei diesem Ansatz werden acht verschiedene Modellierungsstrategien gerechnet, die sich einerseits durch die Anzahl der eingesetzten verdeckten Neuronen, K, unterscheiden, andererseits danach, ob ein linearer Modellteil (Direktverbindungen bzw. Shortcuts) vorhanden ist. Es kommen die in Abschnitt 4.8 dargelegten Vorgangsweisen zum Einsatz. Die Bestimmung des Vektors der Regularisierungsparameter, $\boldsymbol{\nu}$, erfolgt zur Verringerung des Rechenaufwands der Evaluation bloß einmal für jedes Jahr. Hierbei wird in Phase 1 des Verfahrens der heuristische Parameter $\xi = 3 + K/2$ gewählt, d.h. für ein Modell mit nur einem verdeckten Neuron würde bei Annäherung des geschätzten linearen Modells durch das ARNN-Modell der Regularisierungsterm 3 % des SSE des linearen Modells betragen, und bei zehn verdeckten Neuronen wären es 8 %. In Phase 2 wird in einer iterativen Suche der Vektor $\boldsymbol{\nu}$ so skaliert, dass der SSE des damit geschätzten ARNN-Modells zwischen 92 % und 99 % des SSE des linearen Vergleichsmodells ausmacht. Die Phase 3 der Bestimmung des Vektors der Regularisierungsparameter (Verfahren mit Kreuzvalidierung) wird wegen des hohen Rechenaufwands und der unbefriedigenden Ergebnisse in der In-Sample-Testung (vgl. Abschnitt 4.8) ausgelassen. Als lokales Suchverfahren wird Quasi-Newton herangezogen. Die folgende Aufstellung zeigt, worin sich die acht Modellierungsstrategien unterscheiden:

Bezeichnung	Shortcuts	K	Bezeichnung	Shortcuts	K
Regul1	vorhanden	1	Regul5	keine	1
Regul2	vorhanden	2	Regul6	keine	2
Regul3	vorhanden	5	Regul7	keine	5
Regul4	vorhanden	10	Regul8	keine	10

Bayesianischer Ansatz. Wie bei den beiden vorangegangenen Ansätzen werden acht verschiedene Modellierungsstrategien berücksichtigt, die sich anhand der Anzahl der verdeckten Neuronen, K, und der Shortcuts unterscheiden und ansonsten die in Abschnitt 4.9 beschriebene Vorgangsweise ausführen. Im Rahmen der vorliegenden Implementation des Bayesianischen Evidenzansatz werden in einem iterativen Verfahren die sogenannten Hyperparameter α und β ermittelt, die wiederum die Regularisierungsparameter bestimmen. Die maximale Anzahl an Iterationen in diesem Verfahren wird auf 30 begrenzt. Innerhalb jeder einzelnen dieser Iterationen erfolgt eine lokale Suche nach den optimalen Modellparametern, für die hier Quasi-Newton verwendet wird. Die Anzahl der wohl-determinierten Modellparameter, γ, die für die Bestimmung von α und β benötigt wird, benutzt die Hesse-Matrix, da sich in der In-Sample-Austestung des Verfahrens Methoden ohne die Verwendung der Hesse-Matrix als inferior erwiesen haben. Die folgende Aufstellung zeigt, worin sich die acht Modellierungsstrategien unterscheiden:

Bezeichnung	Shortcuts	K	Bezeichnung	Shortcuts	K
Bayes1	vorhanden	1	Bayes5	keine	1
Bayes2	vorhanden	2	Bayes6	keine	2

Bayes3	vorhanden	5	Bayes7	keine	5
Bayes4	vorhanden	10	Bayes8	keine	10

Aus diesen Modellierungsstrategien wird für jeden der ARNN-Ansätze jeweils eine *primäre* Modellierungsstrategie herausgehoben. Es sind dies Param3, Klass2, Regul2 und Bayes2. Diese Wahl spiegelt die Erfahrungen bei der Entwicklung und In-Sample-Austestung der Methoden wider (vgl. Kap. 4). So hat sich bei allen Modellstrategien gezeigt, dass ARNN-Modelle mit einem linearen Teil tendenziell bessere oder zumindest stabilere Prognosen liefern als solche ohne lineren Teil. Für den parametrisch-statistischen Ansatz konnte das Pruning-Verfahren als wichtige Methode zur Sicherstellung der Generalisierungsfähigkeit nachgewiesen werden, doch die verschiedenen Verfeinerungen des Pruning-Verfahren haben sich auf den realweltlichen Zeitreihen als wenig lohnend herausgestellt. Weiters hat sich bereits in der Entwicklungsphase der anderen, nicht-parametrisch ausgerichteten Ansätze gezeigt, dass die theoretische Erwartung nicht ganz erfüllt werden kann, dass diese Methoden auch für Modelle mit großer Anzahl von verdeckten Neuronen die Generalisierungsfähigkeit gewährleisten können. Deshalb werden auch für diese Ansätze Modelle mit nur wenigen verdeckten Neuronen in der praktischen Anwendung favorisiert. Unter diesen Gesichtspunkten ist für die Prognoseanordnung zu erwarten, dass die ausgewählten primären Modellierungsstrategien das Potential der jeweiligen Ansätze am besten ausschöpfen. Auf sie wird der Fokus des systematischen Vergleichs der Prognosegüte gelegt.

5.4 Ergebnisse

In diesem Abschnitt werden die Ergebnisse der Evaluierung der Modelle bzw. Modellierungsstrategien auf dem Evaluierungsset der beiden Beispielzeitreihen Arbeitslosenrate und Industrieproduktionsindex besprochen. Das Evaluierungsset umfasst die Jahre 1998-2002. Unter der Berücksichtigung aller Modellfamilien und Modellierungsstrategien sind für die beiden Beispielzeitreihen je 466 Prognosezeitreihen auszuwerten und zu vergleichen. Außerdem gibt es jede der 466 Prognosen für zwölf verschiedene Prognosehorizonte h.

Der Schwerpunkt der folgenden Auswertung liegt auf der Frage, ob neuronale Netze in der Prognose von makroökonomischen Zeitreihen einen Vorteil gegenüber den linearen Methoden aufweisen. Ein wesentlicher Teil der folgenden Darstellungen ist daher der Vergleich der Prognosen auf der Basis von ARNN-Modellen mit der verfügbaren linearen Prognose. Da eine große Anzahl von Prognosen vorliegt, die sich nach verschiedenen Modellierungsentscheidungen unterscheiden, kann man naturgemäß nicht alle Varianten in gleicher Weise mit den linearen Alternativen bzw. untereinander vergleichen sondern muss strukturiert vorgehen.

Im Vordergrund des systematischen Prognosevergleichs stehen jene Prognosen, die auf den saisonalen differenzierten Zeitreihen, unter Verwendung der bis ins

Jahr 1960 zurückreichenden Zeitreihendaten und unter Verwendung der simulativ-iterierten oder der direkten Mehr-Schritt-Prognosemethode ermittelt werden. Diese Ergebnisse werden in den folgenden fünf Unterabschnitten getrennt nach den einzelnen Modellierungsstrategien untersucht. Im Zentrum der Auswertung steht der MSFE bzw. der relative MSFE, d.h. bezogen auf den MSFE des linearen Vergleichsmodells.

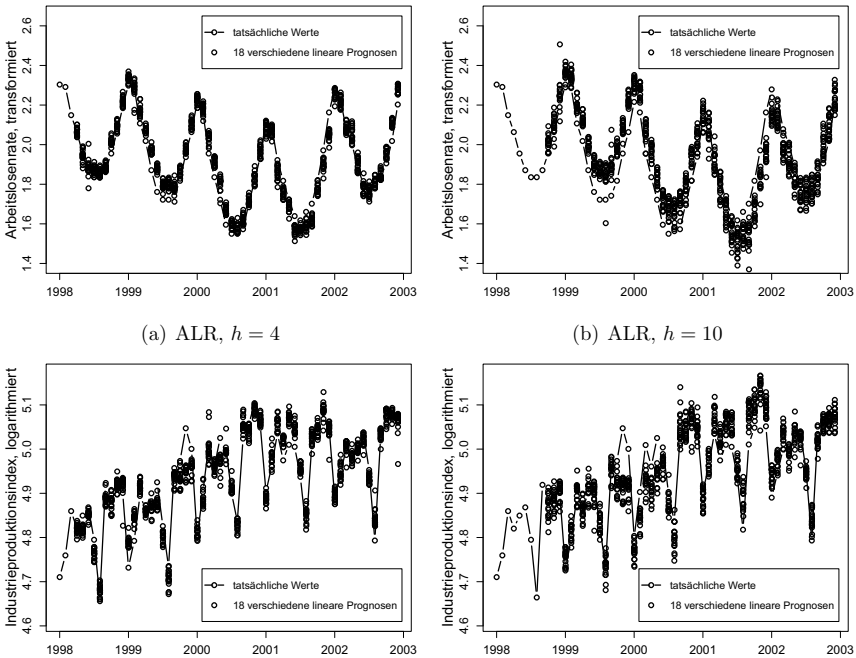
Prognosen unter abweichenden Modellierungsoptionen finden in weiteren Unterabschnitten Berücksichtigung. Dies geschieht im Rahmen von Sensitivitätsanalysen, die gezielt danach fragen, wie es sich auf die Prognosegüte auswirkt, wenn die einfachen Differenzen oder die saisonalen und einfachen Differenzen zur Modellbildung verwendet werden, wenn die naiv-iterierte Methode der Mehr-Schritt-Prognose eingesetzt wird, wenn Direktverbindungen weggelassen werden oder wenn das Lernset erst im Jahr 1975 beginnt.

Ein eigener Unterabschnitt führt die Hauptergebnisse zusammen, indem von jedem der verschiedenen Modellierungsansätze nur die primäre Modellierungsstrategie herangezogen wird. Bei einer solchen Einschränkung ist es auch möglich, in sinnvoller Weise einen statistischen Test auf Unterschiedlichkeit der Prognosegüte durchzuführen.

Lineare Prognosen

Die AR-, ARDS-, ARMA- und ARMADS-Modelle liefern die Prognosen, mit denen jene der ARNN- und ARNNDS-Modelle verglichen werden sollen. Bevor die Ergebnisse zur Prognosegenauigkeit der linearen Modelle präsentiert werden, kann noch kurz der Evaluierungsbereich der beiden Beispielzeitreihen und die Aufgabe der Prognosestellung vorgestellt werden. Zu diesem Zweck soll exemplarisch für zwei ausgewählte Prognosehorizonte gezeigt werden, in welcher Weise die Prognosezeitreihen typischerweise von der realisierten Zeitreihe im zu prognostizierenden Bereich abweichen. In Abb. 16 wird dies jeweils für ALR und IPI sowie ausgewählte Prognosehorizonte visualisiert (je ein mittlerer und ein großer Prognosehorizont, $h = 4$ bzw. $h = 10$).

Aus der Abbildung sind zunächst die Verläufe der beiden Beispielzeitreihen ALR und IPI im Evaluierungszeitraum ersichtlich (schwarze Linien). In beiden Fällen ist etwa in der Mitte des betrachteten Zeitraums die Verschlechterung des wirtschaftlichen Umfelds festzustellen, die sich ab dem Jahr 2001 in einem Wiederanstieg der Arbeitslosigkeit bzw. einer Stagnation der Industrieproduktion äußert. Vom tatsächlichen Verlauf der Zeitreihe weichen die 18 verschiedenen linearen Prognosen unterschiedlich stark ab. Naturgemäß sind bei $h = 10$ die Abweichungen im Vergleich zu $h = 4$ viel stärker. Der Wirtschaftsabschwung ab 2001 lässt sich offenbar besonders schwer vorhersagen, da die Arbeislosigkeit für die Zeit nach 2001 zu niedrig und die Industrieproduktion zu hoch prognostiziert wird.



(a) ALR, $h = 4$

(b) ALR, $h = 10$

(c) IPI, $h = 4$

(d) IPI, $h = 10$

Abb. 16: Vergleich der realisierten Zeitreihe mit Mehr-Schritt-Prognosen auf Basis der getesteten linearen Modellierungsansätze, mit $h = 4$ bzw. $h = 10$: (a) und (b) Arbeitslosenrate (ALR) sowie (c) und (d) Industrieproduktionsindex (IPI)

Es können einige weitere bemerkenswerte empirische Regelmäßigkeiten anhand der Abbildung ausgemacht werden:

- aufeinanderfolgende Prognosen weichen in die gleiche Richtung von den tatsächlichen Werten ab (serielle Korreliertheit der Prognosefehler),
- die Prognosen weichen regelmäßig in die gleiche Richtung ab (Korreliertheit der Prognosefehler über unterschiedliche Prognosemethoden),
- die Varianz der Prognosefehler steigt mit wachsendem h an, wobei dieser Anstieg im Falle der ALR viel höher ist, als bei der IPI; die ALR ist offenbar eine vergleichsweise schwierige Herausforderung für Mehr-Schritt-Prognosen.

Generell ist die soeben gegebene Charakterisierung weitgehend repräsentativ für sämtliche der getesteten Modelle und Modellierungsstrategien, nicht nur die

linearen Modelle. Anhand der Prognosen mit AR-Modellen lassen sich jedoch diese Muster besonders eindeutig beobachten, da hier viele der zusätzlichen numerischen Schwierigkeiten der nicht-linearen Modelle wegfallen.

In Tab. 21 werden die Ergebnisse für die Prognosegüte von neun verschiedenen linearen Modellen zusammengefasst. Das primäre lineare Modell ist das AR-Modell auf der saisonal differenzierten Zeitreihe, da der Differenzenfilter ∇_{12} in Kapitel 2 als optimal identifiziert wurde. Dieses Modell dient als Referenzmodell. Es wird im Folgenden als lineares Basismodell (LBM) und die Prognose, die es liefert, als lineare Basisprognose (LBP) bezeichnet. In der Tabelle wird die Prognosegüte für die Prognosehorizonte $h = 1, \ldots, 12$ angegeben, wobei für die LBP der MSFE eingetragen ist und durch Kursivstellung hervorgehoben wird, während bei den anderen Modellen die Prognosegüte als Prozentsatz des MSFE der LBP ausgedrückt wird.

Es seien zunächst die Ergebnisse für die Arbeitslosenrate betrachtet. Für das AR-Modell auf den saisonalen Differenzen (LBM) steigt der MSFE mit wachsendem Prognosehorizont von 65.0 ($h = 1$) auf 1450 ($h = 12$) an. Das LBM liefert jedoch nicht die besten verfügbaren linearen Prognosen. Je nach Prognosehorizont haben das auf $\nabla_{12}\nabla Y_t$ geschätzte AR-Modell (bei $h = 1, 3, \ldots, 7$), das auf ∇Y_t geschätzte direkte ARDS-Modell (bei $h = 2$) und das auf $\nabla_{12} Y_t$ geschätzte ARMA-Modell (bei $h = 8, \ldots, 12$) die jeweils besten Prognosen. Die meisten Modelle liegen recht nahe an der Prognosegüte des LBM. Das ARMA-Modell auf $\nabla_{12}\nabla Y_t$ weist hingegen eine auffallend schlechte Prognosegüte auf.[34]

Beim Industrieproduktionsindex steigt für das LBM, d.h. das auf $\nabla_{12} Y_t$ geschätzte AR-Modell, der MSFE von 70.5 bei Ein-Schritt-Prognosen auf 201.0 (bei $h = 12$) an. Hinsichtlich der Prognosegüte ist die LBP nur für hohe Prognosehorizonte, $h = 10, 11, 12$, am besten. Für $h = 1, \ldots, 9$ schneidet die ARMA-Prognose, die ebenfalls auf saisonalen Differenzen ermittelt wurde, am besten ab. Auffällig ist weiters die relative schlechte Prognosegüte der meisten anderen Modelle im Bereich der mittleren und hohen Prognosehorizonte. Das ARMADS-Modell liefert durchgängig die schlechtesten Prognosen.

Zusammengefasst zeigen diese Ergebnisse, dass es nicht *die* lineare Prognose schlechthin gibt, sondern dass auch bei linearen Modellen die Prognosegüte entscheidend von der Wahl des Differenzenfilters abhängt und je nach verwendetem Modell und je nach Prognosehorizont variiert. Die Verwendung der saisonalen Differenzen für die primären Modelle, welche, wie bereits erläutert, vorab getroffen wurde, stellt sich im Lichte dieser Ergebnisse im Falle der Arbeitslosenrate als nicht ganz optimal heraus, da mit einem anderen Differenzenfilter eine bessere Prognosegüte erzielbar wäre. Für den Industrieproduktionsindex hingegen liefert die Wahl die besten Prognosen. In den folgenden Unterabschnitten zur Analyse der Prognosegenauigkeit der verschiedenen ARNN-Modelle wird die Auswertung auf Modelle

[34] Dies liegt zumindest teilweise daran, dass die Zeitreihe in dieser Form überdifferenziert sein dürfte, woraus eine Einheitswurzel im Lag-Polynom des MA-Terms und in weiterer Folge numerische Probleme beim Pruning-Verfahren folgen.

Tab. 21: Out-of-Sample-Prognosegüte mit linearen Modellen, ALR und IPI

Arbeitslosenrate									
	AR (bzw. ARDS), iteriert			AR (bzw. ARDS), direkt			ARMA (bzw. ARMADS)		
	$\nabla_{12}Y_t$	∇Y_t	$\nabla_{12}\nabla Y_t$	$\nabla_{12}Y_t$	∇Y_t	$\nabla_{12}\nabla Y_t$	$\nabla_{12}Y_t$	∇Y_t	$\nabla_{12}\nabla Y_t$
$h = 1$	*65.0*	106.5	98.2	100.0	106.5	98.2	100.9	122.9	106.9
$h = 2$	*135.9*	108.2	96.2	102.5	86.2	115.5	102.2	116.6	116.3
$h = 3$	*203.2*	111.1	91.7	102.0	99.1	107.9	106.5	108.9	119.8
$h = 4$	*302.6*	112.6	91.2	101.1	93.6	109.8	107.6	104.9	139.0
$h = 5$	*422.2*	111.7	93.9	100.8	98.7	109.8	106.0	97.4	121.2
$h = 6$	*565.3*	109.4	95.0	102.7	102.5	112.6	101.2	101.6	117.3
$h = 7$	*714.4*	106.3	96.4	101.8	105.5	103.4	99.9	100.2	150.8
$h = 8$	*864.5*	103.2	98.5	101.8	101.4	103.8	96.7	102.1	136.3
$h = 9$	*995.0*	100.7	101.5	100.1	104.4	115.0	91.8	105.9	126.2
$h = 10$	*1141.6*	99.9	104.5	101.1	111.0	118.3	86.9	112.8	130.1
$h = 11$	*1284.0*	99.4	106.5	99.8	109.4	123.5	82.8	113.3	123.5
$h = 12$	*1450.0*	100.0	108.7	97.6	107.2	130.6	82.6	112.6	134.8
Industrieproduktionsindex									
	AR (bzw. ARDS), iteriert			AR (bzw. ARDS), direkt			ARMA (bzw. ARMADS)		
	$\nabla_{12}Y_t$	∇Y_t	$\nabla_{12}\nabla Y_t$	$\nabla_{12}Y_t$	∇Y_t	$\nabla_{12}\nabla Y_t$	$\nabla_{12}Y_t$	∇Y_t	$\nabla_{12}\nabla Y_t$
$h = 1$	*70.5*	105.7	100.9	100.0	105.7	100.9	86.0	221.6	86.8
$h = 2$	*82.0*	101.1	101.3	96.1	110.5	102.8	86.0	225.2	92.9
$h = 3$	*92.0*	102.0	104.6	93.9	109.5	107.7	90.1	206.8	97.8
$h = 4$	*99.3*	97.1	108.9	98.7	106.9	98.5	94.3	206.7	98.6
$h = 5$	*108.5*	108.2	111.5	96.2	101.5	113.4	96.2	198.1	104.0
$h = 6$	*122.7*	108.8	114.0	99.3	100.6	113.8	96.7	171.1	104.6
$h = 7$	*150.3*	108.3	117.1	103.5	100.6	122.1	96.9	183.0	111.7
$h = 8$	*166.9*	110.2	119.4	100.8	101.3	120.9	95.9	195.6	114.6
$h = 9$	*173.6*	115.6	122.3	104.8	109.5	129.6	98.8	174.5	117.7
$h = 10$	*183.9*	119.1	127.8	107.2	109.8	134.7	100.6	200.3	121.3
$h = 11$	*180.5*	120.7	129.7	114.0	111.7	139.5	121.6	191.0	145.6
$h = 12$	*201.0*	116.2	131.8	110.7	104.2	147.8	129.4	219.4	156.3

In der ersten Datenspalte (in Kursivschrift) stehen MSFE-Werte, in den restlichen Spalten relative MSFE-Werte, d.h. in Prozent der ersten Spalte.

beschränkt, die auf der Basis von saisonal differenzierten Zeitreihen erstellt werden. Erst später wird im Rahmen von Sensibilitätsanalysen auch wieder der Einfluss des Differenzenfilters in die Analyse miteinbezogen werden.

Prognosen mit dem statistisch-parametrischen Ansatz

Es werden vier verschiedene Modellierungsstrategien des statistisch-parametrischen Ansatzes berücksichtigt. Tab. 22 fasst die Auswertung der Out-of-Sample-Prognosegüte für ALR und IPI zusammen.

Die Tabelle enthält die MSFE-Werte, ausgedrückt in Prozent des jeweiligen MSFE-Werts der LBP. Aus den in der Tabelle gezeigten Ergebnissen kann man die folgenden Punkte hervorheben:

Tab. 22: Out-of-Sample-Prognosegüte mit dem statistisch-parametrischen Ansatz, ALR und IPI

Arbeitslosenrate								
	iterierte Prognosen				direkte Prognosen			
	Param1	Param2	Param3	Param4	Param1	Param2	Param3	Param4
$h = 1$	105.5	104.6	91.4	112.0	105.5	104.6	104.9	106.6
$h = 2$	115.1	115.3	82.0	122.6	116.0	117.6	108.3	111.3
$h = 3$	109.7	119.1	85.3	125.7	114.5	115.5	110.7	102.3
$h = 4$	109.2	121.7	85.5	120.3	109.3	114.4	106.1	121.2
$h = 5$	110.7	127.4	85.1	120.5	109.3	114.3	101.0	131.1
$h = 6$	107.1	122.2	84.2	121.8	107.0	108.9	120.7	104.8
$h = 7$	104.1	119.4	85.0	125.5	107.5	107.7	104.8	100.8
$h = 8$	103.2	116.4	88.3	117.1	103.2	101.7	111.5	112.1
$h = 9$	103.5	112.3	90.5	108.3	98.6	101.1	103.2	111.7
$h = 10$	101.4	108.1	94.2	103.9	99.6	99.6	105.5	111.0
$h = 11$	102.9	104.6	94.8	104.3	93.9	95.1	99.5	100.5
$h = 12$	100.4	102.2	94.8	108.3	94.5	94.6	100.5	109.5
Industrieproduktionsindex								
	iterierte Prognosen				direkte Prognosen			
	Param1	Param2	Param3	Param4	Param1	Param2	Param3	Param4
$h = 1$	95.5	94.8	99.3	99.4	95.5	94.6	101.4	102.6
$h = 2$	96.3	97.2	100.0	104.5	94.5	95.4	97.9	87.3
$h = 3$	101.5	103.5	104.9	107.3	94.1	94.6	95.8	105.3
$h = 4$	108.6	107.7	109.8	100.1	101.8	101.1	98.3	97.6
$h = 5$	112.3	107.9	105.5	107.3	100.3	99.7	104.0	108.0
$h = 6$	115.5	110.4	103.3	113.5	100.7	99.9	107.7	102.4
$h = 7$	105.5	108.6	109.8	127.1	105.3	104.1	109.0	102.7
$h = 8$	107.5	112.0	111.1	130.7	102.0	101.2	103.7	108.1
$h = 9$	110.0	107.3	112.9	130.7	105.8	106.4	101.2	110.8
$h = 10$	108.9	111.5	110.6	125.3	110.0	110.2	112.9	117.4
$h = 11$	109.8	108.9	112.7	124.8	117.1	119.4	114.7	123.5
$h = 12$	111.1	107.3	113.6	113.5	109.7	108.0	114.1	113.5

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

- In der überwiegenden Anzahl der Fälle liegt der MSFE knapp über demjenigen der LBP, in einzelnen Bereichen knapp darunter sowie in wenigen Fällen um bis zu 30 % darüber.

- Die Prognose, die für ALR von der Modellierungsstrategie Param3 geliefert wird, sticht durch relativ gute MSFE-Werte hervor. Zur Erinnerung, Param3 bezeichnet die primäre Modellierungsstrategie, die eine Modellspezifikation mit Direktverbindungen umfasst und ein Pruningverfahren beinhaltet, welches nach jedem Pruningschritt eine Neuinitialisierung der Gewichte verwendet. Diese Prognose ist die einzige im Rahmen dieser Auswertung, die für alle Prognosehorizonte deutlich bessere MSFE-Werte als die LBP aufweist. Im Falle $h = 2$ beträgt dieser Vorteil sogar -18 % und ist statistisch signifikant.

- Direkte Prognosen, d.h. solche, die mit einem für den gewünschten Prognosehorizont maßgeschneiderten Modell generiert werden, sind in einzelnen Bereichen leicht besser als die LBP. Interessanterweise sind dies jedoch für ALR und IPI unterschiedliche Bereiche (ALR: ca. ab $h = 9$, IPI: ca. für $h = 1, \ldots, 4$).

Insgesamt erlaubt diese Auswertung kaum eine klare Aussagen über die Vorteilhaftigkeit einer der vier Modellierungsstrategien gegenüber den anderen oder gegenüber der LBP. Allerdings scheint Param3, nicht nur für den Fall von iterierten Prognosen der Arbeitslosenrate, eine relativ gute Wahl zu sein. Es ergeben sich keine starken Hinweise auf Nicht-Linearitäten in den beiden Zeitreihen, die in verlässlicher Weise durch ARNN-Modelle behandelt werden könnten.

Prognosen mit dem klassischen Ansatz

Die Ergebnisse mit dem klassischen Ansatz mit Early-Stopping werden in Tab. 23 angeführt. Es wird die Prognosegüte der Modellierungsstrategien Klass1, Klass2, Klass3 und Klass4 untersucht. Sie betreffen ARNN-Modelle mit Direktverbindungen zwischen Input- und Outputknoten und 1, 2, 5 und 10 verdeckten Neuronen. Die Ergebnisse, die bei Weglassung der Direktverbindungen (Modellierungsstrategien Klass5, Klass6, Klass7 und Klass8) berechnet werden, werden weiter unten in einer Sensitivitätsanalyse besprochen.

Insgesamt liegen die MSFE-Werte mit dem klassischen Ansatz meist über denjenigen der LBP, wobei Unterschiede zwischen den verschiedenen Modellierungsstrategien und den beiden Beispielzeitreihen existieren, aber kaum Regelmäßigkeiten zu erkennen sind:

- Die Prognosegüte relativ zur LBP ist auf IPI besser als auf ALR, vor allem auch durch die höhere Variabilität der MSFE-Werte auf ALR. In einzelnen liefert der klassische Ansatz auf ALR um bis zu 40 % schlechtere MSFE-Werte, während die Prognosen auf IPI maximal um 20 % schlechtere MSFE-Werte aufweisen. Allerdings sind bei für die ALR auch einzelne Fälle von deutlich unter der LBP liegenden MSFE-Werten zu beobachten (direkte Prognosen mit Klass2 und Klass3 für $h = 10$).

- Im Vergleich lässt sich kein klarer Vorteil für eine der vier Modellierungsstrategien ausmachen. Allerdings kann für die ALR ein besseres Abschneiden von Modellern mit nur ein oder zwei verdeckten Neuronen vermutet werden.

- Direkte Prognosen weisen beim klassischen Ansatz im allgemeinen bessere Ergebnisse auf als iterierte Prognosen, insbesondere auf dem IPI. Wie bereits zuvor beim statistisch-parametrischen Ansatz kann festgestellt werden, dass diese relativen Vorteile vom Prognosehorizont abhängen. Auf der ALR sind direkte Prognosen für die Prognosehorizonte $h = 8, \ldots, 12$ teilweise besser als die LBP, in einzelnen Fällen um mehr als 10 %. Für den IPI hingegen

Tab. 23: Out-of-Sample-Prognosegüte mit dem klassischen Ansatz, ALR und IPI

Arbeitslosenrate								
	iterierte Prognosen				direkte Prognosen			
	Klass1	Klass2	Klass3	Klass4	Klass1	Klass2	Klass3	Klass4
$h = 1$	98.9	113.1	104.8	98.9	104.8	113.4	105.1	111.4
$h = 2$	102.1	115.1	106.0	111.3	112.1	112.2	100.4	109.3
$h = 3$	105.7	114.8	115.0	117.4	112.1	108.5	108.6	113.3
$h = 4$	102.6	115.4	117.6	123.1	110.6	109.0	126.7	145.0
$h = 5$	97.9	118.7	116.6	118.6	111.6	108.7	108.7	110.4
$h = 6$	96.2	115.6	114.6	113.3	101.3	109.9	137.1	133.4
$h = 7$	100.6	112.9	111.5	116.7	118.1	106.2	114.1	119.1
$h = 8$	104.0	113.6	110.6	115.8	103.8	104.7	106.4	99.2
$h = 9$	101.4	113.2	111.9	111.9	108.9	93.5	98.6	96.6
$h = 10$	97.5	113.3	115.5	109.0	95.0	93.3	85.9	100.6
$h = 11$	100.0	110.1	116.2	109.3	101.0	103.2	99.6	98.0
$h = 12$	100.3	107.5	115.9	110.2	97.5	91.4	110.0	103.4
Industrieproduktionsindex								
	iterierte Prognosen				direkte Prognosen			
	Klass1	Klass2	Klass3	Klass4	Klass1	Klass2	Klass3	Klass4
$h = 1$	99.0	100.6	101.4	95.9	100.4	98.6	102.3	101.8
$h = 2$	95.6	94.8	99.6	91.7	100.4	99.6	96.3	94.8
$h = 3$	99.9	99.1	104.8	101.8	92.3	93.0	99.1	98.6
$h = 4$	105.1	102.6	109.2	95.3	98.4	100.2	100.6	96.6
$h = 5$	108.6	99.3	107.1	101.7	96.1	92.9	104.9	94.2
$h = 6$	111.1	102.9	110.7	94.5	96.1	99.3	101.9	104.2
$h = 7$	108.8	99.3	109.1	96.8	102.5	100.0	100.1	102.3
$h = 8$	104.9	98.1	110.5	99.0	96.6	97.7	100.4	99.6
$h = 9$	110.5	99.4	111.2	102.2	100.8	98.7	101.0	101.4
$h = 10$	113.9	100.0	116.6	102.3	110.4	102.3	109.8	105.7
$h = 11$	118.3	104.5	116.7	103.6	103.3	111.1	111.4	115.6
$h = 12$	115.6	100.0	119.3	108.8	114.9	109.6	105.5	111.5

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

wird eine Prognosegüte, die die LBP aussticht, eher im Bereich der mittleren Prognosehorizonte beobachtet. Im Vergleich zur Auswertung der Ergebnisse des statistisch-parametrischen Ansatzes fällt auf, dass dieses Muster genau umgekehrt auftritt.

Insgesamt lässt sich sagen, dass die Ergebnisse mit dem klassischen Ansatz zwar vergleichsweise gute Prognoseergebnisse haben können, diese jedoch durch die Zufälligkeiten, die der Methode stärker als anderen Ansätzen inherent sind, und auch durch Ausreißer beeinträchtigt sein können. Wie bereits im Abschnitt 4.7 nahegelegt, stellt der Early-Stopping-Ansatz in diesem Sinne keine besonders gute Strategie zur Erreichung der Generalisierungsfähigkeit dar. Teilweise zeigt sich in den Ergebnissen, dass die Zufallsabhängigkeit der Ergebnisse mit der Anzahl der verdeckten Neuronen zunimmt, sodass der Praktiker bei der Wahl von K, entgegen

manchen anderslautenden Empfehlungen der NN-Literatur, vorsichtig sein sollte und K nicht bedeutend größer als bei den anderen Ansätzen wählen sollte.

Aus dem Umstand, dass sich mit dem klassischen Ansatz der ARNN-Modellierung keine klare Verbesserung gegenüber der LBP erzielen lassen, kann man außerdem keinen Hinweis auf Nicht-Linearitäten in den Zeitreihen ableiten, die mit ARNN-Modellen für Prognosen genutzt werden können.

Prognosen mit dem Ansatz mit Regularisierung

In Tab. 24 werden die Ergebnisse zur Out-of-Sample-Prognosegüte für ALR und IPI zusammengefasst, die mit dem Regularisierungsansatz erreicht werden. Wieder handelt es sich um Modelle mit Direktverbindungen und 1, 2, 5 und 10 verdeckten Neuronen (für die Auswertung der Modelle ohne Direktverbindungen siehe die entsprechende Sensitivitätsanalyse).

Auf den ersten Blick ist leicht die im Vergleich zu den bisher diskutierten Ergebnissen noch schlechtere Prognosegüte zu erkennen. Unabhängig ob man die Ergebnisse für iterierte oder direkte Prognosen, ALR oder IPI betrachtet, finden sich kaum Werte, die nicht mindestens um 5 % schlechter sind als die LBP. Auf der positiven Seite sind allerdings die beiden folgenden Punkte zu vermerken:

- Für die Zeitreihe IPI ergibt sich bei fast allen Prognosen im Bereich $h = 1, 2, 3$ ein leichter Vorteil der ARNN-Modelle gegenüber der LBP.
- Es sind nur vereinzelte Fälle von besonders schlechten Prognosen oder Ausreißern zu beobachten. Dies trifft insbesondere auch auf Modellierungsstrategien mit einer höheren Anzahl von verdeckten Neuronen zu.

Zusammenfassend lässt sich sagen, dass mit dem Regularisierungsansatz zwar keine besseren Prognosen erzielt werden als mit dem linearen Basismodell. Aber die Methode erscheint robust und im Vergleich zu anderen Modellierungsansätzen auch bei unterschiedlichen Modellspezifikationen relativ verlässlich. Der Vollständigkeit soll auch hier erwähnt werden, dass das Ergebnis sicherlich keine Hinweise auf eine etwaige vorhandene Nicht-Linearität in den Daten geben kann.

Prognosen mit dem Bayesianischen Ansatz

Die Ergebnisse der Evaluierung der Out-of-Sample-Prognosegüte mit dem Bayesianischen Ansatz werden in Tab. 25 zusammengefasst. Auch hier handelt es sich um Modelle mit Direktverbindungen und 1, 2, 5 bzw. 10 verdeckten Neuronen (für die Auswertung der Modelle ohne Direktverbindungen siehe die entsprechende Sensitivitätsanalyse).

Unter den untersuchten Modellierungsansätzen hat der Bayesianische Ansatz die schlechteste Prognosegenauigkeit aufzuweisen, mit teilweise doppelt und dreifach so hohen MSFE-Werten wie die LBP. Dies bedeutet allerdings nicht, dass es keine Fälle

Tab. 24: Out-of-Sample-Prognosegüte mit dem Ansatz mit Regularisierung, ALR und IPI

Arbeitslosenrate								
	iterierte Prognosen				direkte Prognosen			
	Regul1	Regul2	Regul3	Regul4	Regul1	Regul2	Regul3	Regul4
$h = 1$	104.9	108.8	103.5	108.2	105.7	108.6	104.2	108.5
$h = 2$	108.8	113.6	105.0	109.1	116.8	113.5	111.7	111.3
$h = 3$	108.4	111.7	108.0	112.7	126.3	123.7	108.5	113.8
$h = 4$	110.0	109.9	110.3	105.4	109.3	113.4	106.3	108.3
$h = 5$	108.7	111.9	111.3	108.7	115.5	111.1	109.1	106.5
$h = 6$	105.6	113.5	108.2	104.5	113.3	114.8	108.0	113.1
$h = 7$	105.5	108.3	109.1	100.1	108.4	106.8	110.1	107.9
$h = 8$	106.6	106.7	106.0	99.8	100.4	105.8	107.9	107.3
$h = 9$	108.7	104.5	108.5	101.1	117.0	111.8	102.3	104.2
$h = 10$	109.9	103.5	112.8	104.3	106.4	115.3	103.7	105.6
$h = 11$	110.4	104.0	116.7	107.0	106.8	98.6	101.9	98.3
$h = 12$	109.6	98.2	115.2	103.9	85.9	85.9	84.5	85.5
Industrieproduktionsindex								
	iterierte Prognosen				direkte Prognosen			
	Regul1	Regul2	Regul3	Regul4	Regul1	Regul2	Regul3	Regul4
$h = 1$	97.3	96.3	95.7	94.0	102.7	103.1	97.4	94.0
$h = 2$	103.5	95.6	96.2	95.1	94.1	95.2	92.7	93.4
$h = 3$	96.8	97.4	100.4	102.8	90.0	94.8	96.3	97.1
$h = 4$	99.0	106.5	106.8	107.9	102.1	105.8	108.3	105.1
$h = 5$	102.7	111.4	109.2	110.7	106.4	107.4	112.7	107.4
$h = 6$	108.7	107.2	107.3	106.2	117.9	128.4	108.6	109.0
$h = 7$	113.4	107.9	108.3	102.9	108.1	105.5	107.5	105.5
$h = 8$	115.3	109.6	107.0	102.6	107.6	105.8	103.1	104.3
$h = 9$	113.7	107.4	112.0	108.8	109.2	107.1	116.9	109.3
$h = 10$	114.4	106.8	111.5	105.4	113.8	114.8	116.9	115.3
$h = 11$	117.3	107.2	111.6	105.3	115.8	129.9	116.6	124.3
$h = 12$	113.0	103.1	109.2	104.4	114.9	110.8	111.1	111.3

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

zu finden gibt, in denen der Bayesianische Ansatz die LBP schlägt. Vor allem bei der Prognose der Arbeitslosenrate versagt der Bayesianische Ansatz fast durchgehend. Beim Industrieproduktionsindex trifft dies nur auf Bayes3 und Bayes4 zu, hängt also von der Anzahl der verdeckten Neuronen K ab. Mit Bayes1 und Bayes2 ($K = 1, 2$) hingegen lassen sich ebensogute Resultate erzielen wie mit dem LBM.

Dies wirft zwei Fragen auf. Zum einen muss man fragen, warum der Bayesianische Ansatz so stark versagt, obwohl er in einfachen Modellen klar sein Funktionieren bewiesen hat. Eine Inspektion der Schätzprozesse dieses Ansatzes zeigt, dass sowohl eine zu große als auch eine zu geringe Regularisierung im Rahmen des Bayesianischen Ansatzes zu den schlechten Ergebnissen beitragen.

Tab. 25: Out-of-Sample-Prognosegüte mit dem Bayesianischen Ansatz, ALR und IPI

Arbeitslosenrate								
	iterierte Prognosen				direkte Prognosen			
	Bayes1	Bayes2	Bayes3	Bayes4	Bayes1	Bayes2	Bayes3	Bayes4
$h = 1$	104.3	116.8	134.0	139.8	101.2	120.5	127.2	127.7
$h = 2$	117.8	139.6	171.8	193.1	110.4	119.2	137.2	155.3
$h = 3$	118.0	155.7	203.2	234.5	123.4	99.8	152.7	135.8
$h = 4$	123.9	166.4	215.4	279.2	109.4	110.7	155.8	182.5
$h = 5$	129.1	171.7	219.4	312.1	127.7	107.7	152.0	196.9
$h = 6$	125.0	164.2	228.9	334.9	110.5	110.8	149.9	125.3
$h = 7$	124.7	159.9	236.4	345.4	103.2	112.5	200.5	152.4
$h = 8$	122.2	149.9	237.8	341.3	78.4	111.4	172.6	147.1
$h = 9$	122.3	140.7	226.7	308.8	101.0	104.3	173.2	157.5
$h = 10$	121.9	136.9	212.1	266.3	104.8	92.7	129.7	148.1
$h = 11$	124.2	131.4	213.7	214.7	95.3	96.1	122.1	176.3
$h = 12$	123.6	124.1	198.9	178.1	91.3	83.9	100.4	150.5
Industrieproduktionsindex								
	iterierte Prognosen				direkte Prognosen			
	Bayes1	Bayes2	Bayes3	Bayes4	Bayes1	Bayes2	Bayes3	Bayes4
$h = 1$	98.7	100.7	119.7	128.7	98.7	98.9	113.6	141.7
$h = 2$	99.5	103.7	117.6	124.1	95.2	99.4	107.1	118.3
$h = 3$	100.7	102.5	132.8	117.8	93.7	96.8	104.1	129.3
$h = 4$	101.4	100.4	146.3	120.5	98.6	100.4	128.5	143.8
$h = 5$	103.7	103.8	164.7	141.8	101.5	107.7	140.0	151.6
$h = 6$	98.9	102.0	183.5	183.1	106.5	108.0	131.9	186.2
$h = 7$	97.8	102.2	192.7	209.0	103.5	103.4	113.8	138.5
$h = 8$	100.6	100.3	209.5	236.7	101.5	108.6	113.8	126.1
$h = 9$	99.5	99.4	242.4	228.5	104.8	109.0	115.9	120.2
$h = 10$	97.5	108.2	241.9	194.9	109.1	108.8	111.7	136.4
$h = 11$	99.9	106.3	271.4	220.9	118.4	116.7	138.7	134.5
$h = 12$	99.2	107.7	243.1	244.4	110.7	108.9	109.8	110.2

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Weiters drängt sich die Frage auf, ob es sich um einzelne Ausreißer von schlechten Prognosen handelt, die den MSFE der Prognose trotz ansonsten gutem Abschneiden beeinträchtigen. Dies ist zwar teilweise der Fall. Im Unterschied zu einer automatisierten Prognoseanordnung kann der Praktiker in einer realweltlichen Anwendung solche schlechten Prognosen vorab ausschließen, wenn sie zu offensichtlich sind. In einer weitgehend automatisierten Prognoseanordnung, wie sie für die vorliegende Arbeit entwickelt wurde, kann dies jedoch nur sehr eingeschränkt nachvollzogen werden. Letztlich zeigt sich, dass das Bemühen Ausreißer vorab (d.h. rein anhand des Lernsets) zu erkennen, nur ein heuristischer Versuch ist, die Schwächen einer Methode auszugleichen, zumal es eine klare Definition von „Ausreißer“ nicht gibt. Wo immer in einer praktischen Anwendung die Grenze eingezogen wird, treten

innerhalb dieser Grenzen dennoch schlechte Prognosen auf, die die Out-of-Sample-Prognosegüte beeinträchtigen.[35]

Sensitivitätsanalyse für unterschiedliche Differenzenfilter

Diese Sensitivitätsanalyse zeigt, dass Prognosen, die auf $\nabla_{12}\nabla Y_t$ erstellt werden, oft die gleiche oder eine bessere Güte aufweisen als jene auf der Basis von $\nabla_{12}Y_t$. Allerdings scheint die relative Prognosegüte auch von der Modellkomplexität und dem Prognosehorizont abzuhängen. Die auf $\nabla_{12}\nabla Y_t$ geschätzten ARNN-Modelle dürften in allen Modellierungssansätzen stärker von Overfitting und numerischen Problemen betroffen sein, sobald die Modelle mehr als zwei verdeckte Neuronen aufweisen. Die auf Prognosen auf $\nabla_{12}\nabla Y_t$ schneiden relativ gut auf niedrigen Prognosehorizonten ab, weniger auf mittleren und hohen Prognosehorizonten. Die Methode der direkten Mehr-Schritt-Prognose funktioniert auf $\nabla_{12}\nabla Y_t$ so schlecht, dass sie ausgelassen werden musste (vgl. dazu die Erläuterungen in Abschnitt 5.2).

Die Modelle auf der Basis der einfachen Differenzen berücksichtigen in ihrer Spezifikation deterministische Saisonalität (ARDS und ARNNDS). Diese Modelle gehen in fast allen Fällen als unterlegen aus dem Vergleich der Prognosegüte hervor. Für die Möglichkeit, dass auch neuronale Netze mit verschiedenen Arten von Saisonalität in den Daten gut zurechtkommen, wenn man sie nur mit geigneten zusätzlichen Termen spezifiziert, kann daher aus dieser Evaluierung kein positiver Hinweis generiert werden.

Insgesamt lässt sich auf Basis dieser Sensitivitätsanalyse wenig sicheres sagen. Jedenfalls kann keine Überlegenheit der Prognosegenauigkeit auf Basis der alternativen Differenzenfilter gezeigt werden.

Sensitivitätsanalyse für unterschiedliche Methoden der iterierten Prognose

Für die Erstellung von iterierten Mehr-Schritt-Prognosen mit nicht-linearen Modellen stehen die naiv-iterierte und die simulativ-iterierte Methode zur Verfügung. Aus theoretischer Sicht ist die simulativ-iterierte Methode vorzuziehen, da sie unter der Annahme, dass das verwendete Modell den DGP abbildet, korrekte Prognosen liefert, während die naiv-iterierte Methode verzerrte Prognosen ergibt (vgl. Abschnitt 5.2). Aus diesem Grund und weil die Methode sich mit relativ begrenztem Rechenaufwand umsetzen lässt, basieren die Hauptergebnisse des Prognose-Vergleichs auf der simulativ-iterierten Methode.

Um die Abhängigkeit der Prognosegüte von der Methode der iterierten Mehr-Schritt-Prognose zu untersuchen, wird eine Sensitivitätsanalyse durchgeführt, deren Ergebnisse in Tab. 38–45 im Anhang B zusammengestellt werden. Damit kann die

[35] Diese grundsätzliche Diskussion gilt in ähnlicher Weise auch für die anderen ARNN-Modellierungsansätze, wo ebenso einzelne Heuristiken zum Einsatz kommen, dazu dienen instabile Modellierungsroutinen und explodierende Prognoseverläufe auszuschließen.

Erwartung einer Prüfung unterzogen werden, dass die simulativ-iterierte Methode bessere Prognosen erlaubt. Ist dies der Fall, insbesondere in Fällen, in denen die Prognosegenauigkeit besser als diejenige des linearen Vergleichsmodells ist, so kann dies auch als Hinweis auf tatsächliche Nicht-Linearität in der Zeitreihe interpretiert werden.

Bei Durchsicht der Ergebnisse findet sich jedoch kein klar erkennbares Muster, das diesen Erwartungen entspricht. Die Werte für die Prognosegüte, die mit der naiv-iterierten Methode erzielt werden, streuen in scheinbar zufälliger Weise um die Werte auf Basis der simulativ-iterierten Methode. Hierbei betragen die Abweichungen nur in seltenen Fällen mehr als 10 %. Wie zu erwarten nehmen die Abweichungen im Betrag tendenziell mit dem Prognosehorizont zu und mit der Qualität des zugrundeliegenden Modells ab, jedoch ohne klare Richtung. Auch eine Fokussierung auf Prognosen, die im Vergleich zur LBP relativ gut sind und die auf der getreuen Modellierung von tatsächlich vorhandenen Nicht-Linearitäten beruhen könnten (z.B. die Modellierungsstrategien Param3 und Param4 auf den saisonalen Differenzen der Arbeitslosenrate), gibt keinen Hinweis auf die Überlegenheit der simulativ-iterierten Methode.

Der Wert der simulativ-iterierten Methode kann anhand der Evaluierungsanordnung nicht angegeben werden, was unter anderem an nicht vorhandenen Nicht-Linearitäten in den Zeitreihen liegen könnte. Auch wenn diese Sensitivitätsanalyse nur sehr eingeschränkt als Test für Nicht-Linearität gelten darf, so ist das Ergebnis dennoch am besten mit der Abwesenheit von Nicht-Linearität in den beiden Zeitreihen vereinbar.

Sensitivitätsanalyse für Weglassung der Direktverbindungen

Die Frage, ob ein ARNN-Modell Direktverbindungen, d.h. einen linearen Teil, der durch die Koeffizienten f_l spezifiziert wird, enthalten soll, ist in der NN-Literatur nicht eindeutig geklärt. Für die Aufnahme von Direktverbindungen spricht, dass es dadurch dem Modell erleichtert wird, vorhandene nicht-lineare Struktur zu erlernen. Aus diesem Grund enthält die primäre Modellstrategie in allen vier Modellansätzen Direktverbindungen. Mit dem Einsatz von Direktverbindungen geht jedoch eine erhöhte Overfitting-Gefahr einher. Insbesondere sind Direktverbindungen dann überflüssig, wenn in den Daten nicht ausreichend nicht-lineare Struktur vorhanden ist. In diesem Falle wäre der NN-Teil des Modells frei, zufällige Merkmale der Daten zu lernen, während er ansonsten die linearen Strukturen abbilden müsste. Aus diesem Grund wird in der hier zusammengefassten Sensitivitätsanalyse die alternative Möglichkeit der Weglassung der Direktverbindungen untersucht. Dies bietet auch die Möglichkeit, die Robustheit der verschiedenen Modellierungsansätze in Hinblick auf unterschiedliche Modellspezifikation zu untersuchen.

In Tab. 46–48 im Anhang B findet man die Ergebnisse dieser Analyse für den klassischen Ansatz mit Early-Stopping, den Ansatz mit Regularisierung und den

Bayesianischen Ansatz. Für den statistisch-parametrischen Ansatz ist keine eigene Sensitivitätsanalyse notwendig, da die Modellierungsstrategien Param1 und Param2 ohnehin keine Direktverbindungen enthalten und bereits besprochen wurden.

Die Weglassung der Direktverbindungen wirkt sich auf die Prognosegüte, von Ausnahmefällen abgesehen, insgesamt nicht sehr stark aus, wobei dieser Befund von der betrachteten Zeitreihe (ALR oder IPI), vom Modellierungsansatz und von der Prognosemethode abhängt. Beim klassischen Ansatz bringen die Modelle ohne Direktverbindungen für die ALR eine etwas schlechtere Prognosegenauigkeit. Auch für den IPI sind sie meist schlechter, wobei das Ergebnis gemischt ist. Beim Regulierungsansatz liefern Modelle mit und ohne Direktverbindungen meist sehr nahe bei einander liegende MSFE-Werte. Allerdings sind einige der iterierten Prognosen der ALR ohne Direktverbindungen deutlich schlechter. Für die Prognose des IPI sind die regularisierten ARNN-Modelle ohne Direktverbindungen in den meisten Fällen jedoch ein wenig besser als solche mit Direktverbindungen. Sieht man sich schließlich die Ergebnisse mit dem Bayesianischen Ansatz an, so zeigt sich zunächst, dass das schlechte Abschneiden mit Modellen mit mehreren verdeckten Neuronen auch durch das Weglassen der Direktverbindungen nicht abgemildert wird. Brauchbare Prognosen, deren MSFE um weniger als 20 Prozentpunkte über denen der LBP liegt, erhält man nur mit einem einzelnen verdeckten Neuron, egal ob mit oder ohne Direktverbindungen. Abgesehen davon sind die Ergebnisse des Bayesianischen Modellierungsstrategien bei Weglassung der Direktverbindungen mal besser mal schlechter als solche mit Direktverbindungen. Zusammenfassend ergibt die Sensitivitätsanalyse keinen deutlichen Befund, ob in bestimmten Modellierungsvarianten Direktverbindungen in der Modellspezifikation enthalten sein sollten.

Sensitivitätsanalyse für unterschiedliche Startzeitpunkte der Zeitreihe

Eine weitere Sensitivitätsanalyse hat schließlich die Frage zum Gegenstand, wie die Ergebnisse der Evaluierung von der Wahl des Startzeitpunkts des Lernsets abhängen. Neben dem Jan1960 wird hier als alternativer Start der Zeitreihe der Jan1975 getestet. Unterschiede, die sich daraus ergeben, könnten in Strukturbrüchen, zeitabhängiger Heteroskedastizität und sich ändernden saisonalen Mustern liegen, auf die bereits in den Kapiteln 2 und 3 hingewiesen wurde.

Ist eine Zeitreihe in relevantem Ausmaß von einem Strukturbruch betroffen, sodass für unterschiedliche beobachtete Bereiche ein unterschiedlicher DGP angenommen werden kann, so ist sowohl das lineare Modell als auch das nicht-lineare Modell fehlspezifiziert, und beide liefern keine optimalen Out-of-Sample-Prognosen. Man kann aber erwarten, dass nicht-lineare Modelle, die auf dem gesamten beobachteten Ausschnitt der Zeitreihe geschätzt werden, in ihrer Prognosegüte vergleichsweise stärker durch einen Bruch im DGP beeinträchtigt werden. Denn nicht-lineare Modelle, insbesondere NN-Modelle sind zwar flexibler, aber damit auch weniger robust und neigen dazu, sich durch Strukturbrüche, Heteroskedastizität, usw. in die Irre

führen zu lassen. Wenn jedoch keine relevanten Strukturbrüche in einer Zeitreihe vorhanden sind und dennoch das verkürzte Lernset für die lineare bzw. nicht-lineare Modellierung und Schätzung verwendet wird, so kann man annehmen, dass die Prognosegüte der nicht-linearen Modelle besonders durch die Verkürzung der verwendeten Zeitreihe leidet. Denn nicht-lineare Modelle, insbesondere NN-Modelle, sind als datenhungrig bekannt.

Um diese Hypothesen zu testen, wurde die gesamte Prognoseanordnung alternativ mit der verkürzten Zeitreihe durchgeführt. Aus Platzgründen und um dem begrenzten Erkenntnisgewinn Rechnung zu tragen, wurde bei der Auswertung der Schwerpunkt darauf gelegt, wie sich die Prognosegüte *relativ* zu derjenigen des linearen Basismodells verändert. Die Tab. 49–64 des Anhang B berichten den relativen MSFE, ausgedrückt in Prozent des MSFE der linearen Basisprognose LBP, welche in Abschnitt 5.3 definiert wurde. Daneben wird auch berücksichtigt, wie sich der MSFE jeweils für eine gegebene Modellierungsstrategie und gegebenen Differenzenfilter bei Veränderung des Startzeitpunkts verändert (keine Tabellen für diese Auswertung aufgenommen). Die Ergebnisse werden nun getrennt nach Arbeitslosenrate und Industrieproduktionsindex besprochen.

Es sei zunächst die relative Prognosegüte der linearen Modelle auf ALR betrachtet. Für den bevorzugten Differenzenfilter, $\nabla_{12}\nabla$, bringt die Verkürzung des Lernsets durchweges eine Verschlechterung der Prognosegenauigkeit, unabhängig davon, ob man ein konventionelles, ein für direkte Prognosen maßgeschneidertes AR-Modell oder ein ARMA-Modell verwendet, und auch unabhängig vom Prognosehorizont. Bezieht man hingegen in den Vergleich auch die anderen Differenzenfilter mit ein, so zeigt sich, dass ein auf $\nabla 12 \nabla Y_t$ geschätztes direktes AR-Modell für die Prognosehorizonte $h = 3, \ldots, 9$ die besten Prognosen liefert.

Das gleiche Muster wiederholt sich auch bei Betrachtung der ARNN-Modelle. Auch hier ist bei Anwendung des Differenzenfilters $\nabla_{12}\nabla$ die Verwendung des erst im Jahr 1975 beginnenden Lernset durchwegs ungünstig für die Prognosegenauigkeit. Bei Einbezug auch der anderen Differenzenfilter hingegen ist das Ergebnis des Vergleichs weniger klar. Oft haben die ARNN-Modelle, die auf ∇Y_t oder $\nabla_{12}\nabla Y_t$ geschätzt wurden, in ausgewählten Prognosehorizonten eine bessere Prognoseleistung sowohl im Vergleich mit der LBP als auch im Vergleich mit den jeweils entsprechenden Prognosen, die mit dem Startjahr 1960 ermittelt wurden. In diesem Sinne stechen besonders die niedrigen relativen MSFE-Werte für die folgenden Prognosen hervor (siehe hierzu Tab. 54 und 56):

- Prognosen für mittlere Prognosehorizonte, ca. $h = 5, \ldots, 9$, die mit einem direkten ARNNDS-Modell auf ∇Y_t mit Hilfe von Modellierungsstrategien des Ansatzes mit Regularisierung oder des Bayesianischen Ansatzes erstellt wurden,

- Prognosen für niedrige und teilweise auch mittlere Prognosehorizonte, die mit dem statistisch-parametrischen Ansatz erstellt wurden.

Zusammenfassend lassen sich diese Ergebnisse als Hinweis darauf werten, dass der DGP für die Arbeitslosenrate ab 1975 im Vergleich zu davor genügend unterschiedlich ist, um eine Verbesserung der Prognosen zu ermöglichen. Sowohl lineare als auch ARNN-Modelle können von der Verkürzung des Lernsets in ihrer Prognosegüte profitieren, wobei dies allerdings auf die Wahl des Differenzenfilters bedingt ist.

Auf dem Industrieproduktionsindex ist die Beurteilung der Abhängigkeit der Prognosegüte von der Wahl des Startpunktes des Lernsets vergleichsweise einfach vorzunehmen. Die Prognosegenauigkeit auf Basis des normalen, in 1960 beginnenden Lernsets ist meist deutlich besser als jene auf Basis der verkürzten Zeitreihe. Insbesondere für den bevorzugten Differenzenfilter ∇_{12} gilt dies. Basieren die Prognosen hingegen auf $\nabla_{12}\nabla Y_t$, gibt es bei niedrigen Prognosehorizonten, $h = 1, 2, 3$, sowohl bei linearen als auch ARNN-Modellen einige Fälle, wo die Prognosegüte besser als die der LBP und auch besser als diejenige der Prognose mit dem gleichem Modell bei unverkürztem Lernset ist. Diese Prognosegütegewinne sind jedoch nur schwach ausgeprägt.

Als Schlussfolgerung ergibt sich, dass im Falle des Industrieproduktionsindex keine Strukturbrüche die Verkürzung der Zeitreihe für Prognosezwecke vorteilhaft zu machen scheinen. Die unangebrachte Verkürzung der Zeitreihe wirkt sich auf die Prognosegenauigkeit der ARNN-Modelle in stärkerem Ausmaß ungünstig aus als auf die linearen Modelle.

Hauptergebnisse und statistische Signifikanz

In diesem Unterabschnitt wird aus der Fülle an Ergebnissen, die aus der Prognose- und Evaluierungsanordnung resultieren, der Kern herausgeschält, der jene Modelle bzw. Modellstrategien betrifft, die vorab als die primären definiert wurden. Es sind dies unter den linearen Modellen das AR-Modell und das ARMA-Modell sowie von den vier ARNN-Modellen jeweils eines (Param3, Klass2, Regul2 und Bayes2). Die Einschränkung auf diese Hauptergebnisse impliziert auch, dass nur Modelle betrachtet werden, die auf den saisonalen Differenzen gebildet werden. Zudem ist der Start der Zeitreihe immer Jan1960. Allerdings werden von dieser Auswahl jeweils auch die entsprechenden direkten Modelle in den Vergleich miteinbezogen. Diese Einschränkung und vor allem der Umstand, dass sie getroffen wurde *bevor* die Ergebnisse der Prognose- und Evaluierungsanordnung vorlagen, erlaubt es, in sinnvoller Weise statistische Tests auf Vergleich der Prognosegüte durchzuführen.

Die Tab. 26 und 27 enthalten die Auswertung der Out-of-Sample-Prognosegüte für die ausgewählten linearen und ARNN-Modelle auf der Arbeitslosenrate. Die erste der beiden Tabellen fasst die Ergebnisse hinsichtlich der Prognosegüte zusammen, wobei hier, im Unterschied zu den früheren Tabellen in diesem Abschnitt, nicht der relative MSFE im Vergleich zur LBP berichtet wird, sondern unmittelbar die jeweiligen MSFE-Werte. In der ersten Datenspalte ist der MSFE der

Tab. 26: Hauptergebnisse zur Out-of-Sample-Prognosegüte von linearen und ARNN-Modellen, MSFE-Werte, Arbeitslosenrate

	AR	ARMA	Param3	Klass2	Regul2	Bayes2
iterierte Prognosen						
$h = 1$	65.0	65.6	59.4	73.5	70.7	75.9
$h = 2$	135.9	138.9	111.4	156.4	154.4	189.7
$h = 3$	203.2	216.4	173.3	233.2	227.0	316.3
$h = 4$	302.6	325.7	258.8	349.2	332.6	503.5
$h = 5$	422.2	447.4	359.5	501.2	472.3	724.9
$h = 6$	565.3	572.0	475.9	653.5	641.4	928.4
$h = 7$	714.4	713.5	606.9	806.3	774.0	1142.0
$h = 8$	864.5	836.2	763.3	982.0	922.5	1295.5
$h = 9$	995.0	913.6	900.3	1125.9	1039.7	1400.2
$h = 10$	1141.6	992.3	1075.7	1293.6	1181.3	1562.5
$h = 11$	1284.0	1063.7	1216.6	1413.7	1335.2	1686.9
$h = 12$	1450.0	1197.9	1375.3	1559.1	1424.2	1799.7
direkte Prognosen						
$h = 1$	65.0		68.2	73.7	70.6	78.3
$h = 2$	139.3		147.2	152.5	154.2	162.0
$h = 3$	207.3		224.9	220.5	251.4	202.7
$h = 4$	305.8		321.2	329.7	343.2	335.1
$h = 5$	425.4		426.3	459.0	469.2	454.6
$h = 6$	580.6		682.2	621.5	648.7	626.4
$h = 7$	727.5		748.8	758.8	763.2	803.6
$h = 8$	880.1		964.2	904.8	914.4	962.7
$h = 9$	995.7		1027.0	930.0	1112.2	1038.0
$h = 10$	1154.0		1204.5	1065.4	1315.7	1058.4
$h = 11$	1281.2		1277.7	1325.0	1265.7	1234.0
$h = 12$	1415.9		1456.6	1324.6	1246.2	1217.1

iterierten Mehr-Schritt-Prognose mit dem AR-Modell (=LBP) angegeben, sodass der Vergleich mit der ersten Spalte erleichtert wird. In Tab. 27 sind die Ergebnisse des Diebold-Mariano-Tests wiedergegeben, wobei jede der Prognosen gegen die entsprechende Prognose des linearen Basismodells getestet wird. Es wird eine quadratische Verlustfunktion verwendet und für die Berechnung der Wahrscheinlichkeitswerte eine zweiseitige Hypothesebildung unterstellt. Teststatistiken, die bei üblichen Konfidenzniveaus signifikant sind, sind mit einem entsprechenden Symbol versehen.

Die Prognosegüte der linearen Basisprognose wird nur von einem einzelnen der konkurrierenden Vergleichsmodelle durchgängig durch alle Prognosehorizonte hindurch übertroffen, nämlich von Param3. Es bietet sich ansonsten ein differenziertes Bild, bei dem insgesamt der Eindruck entsteht, dass das AR-Modell einen robusten und angemessenen Modellierungszugang bietet. In nur vereinzelten Fällen zeigt der DM-Test eine signifikante Unterschiedlichkeit der Prognosegüte an. Im einzelnen

Tab. 27: Ergebnisse des Diebold-Mariano-Tests für den Vergleich der Prognosegüte von linearen und ARNN-Modellen, Arbeitslosenrate

iterierte Prognosen					
	ARMA	Param3	Klass2	Regul2	Bayes2
$h = 1$	-0.06	0.56	-0.85	-0.57	-1.09
$h = 2$	-0.30	1.76 *	-1.27	-1.03	-2.00 °
$h = 3$	-1.05	0.91	-1.65	-0.50	-1.84 °
$h = 4$	-0.98	0.76	-1.34	-0.37	-1.58
$h = 5$	-0.74	0.78	-1.37	-0.37	-1.45
$h = 6$	-0.16	0.88	-1.10	-0.39	-1.17
$h = 7$	0.02	0.79	-0.88	-0.22	-1.06
$h = 8$	0.46	0.62	-1.11	-0.17	-0.89
$h = 9$	1.14	0.51	-1.07	-0.11	-0.73
$h = 10$	2.07 *	0.28	-1.21	-0.09	-0.66
$h = 11$	3.12 **	0.24	-0.78	-0.10	-0.58
$h = 12$	3.96 **	0.23	-0.60	0.05	-0.47
direkte Prognosen					
	AR	Param3	Klass2	Regul2	Bayes2
$h = 1$	0.00	-0.32	-0.87	-0.56	-1.33
$h = 2$	-0.34	-0.86	-1.28	-1.28	-1.79 °
$h = 3$	-0.41	-0.65	-1.20	-1.28	0.01
$h = 4$	-0.16	-0.36	-0.97	-0.90	-0.42
$h = 5$	-0.10	-0.04	-0.76	-0.55	-0.19
$h = 6$	-0.34	-0.63	-0.71	-0.41	-0.25
$h = 7$	-0.24	-0.24	-0.37	-0.27	-0.26
$h = 8$	-0.23	-0.51	-0.55	-0.29	-0.29
$h = 9$	-0.01	-0.22	0.94	-0.69	-0.32
$h = 10$	-0.14	-0.44	1.68 *	-0.79	0.54
$h = 11$	0.03	0.02	-0.49	0.24	0.22
$h = 12$	0.42	-0.04	1.36	1.25	1.56

Teststatistiken des Diebold-Mariano-Tests mit der LBP (AR-Modell auf saisonalen Differenzen) als Vergleichsprognose. ** (bzw. °°) und * (bzw. °) bezeichnen Werte, die für einen zweiseitigen Test bei einem Konfidenzniveau von 2 % bzw. 10 % signifikant sind, wobei die Verwendung von ° statt * eine Verschlechterung gegenüber der LBP anzeigt.

können die folgenden Aussagen gemacht werden, die teilweise schon aus der Ergebnisbesprechung zu den einzelnen Modellierungsstrategien bekannt sind:

- Das ARMA-Modell liefert ab dem Prognosehorizont $h = 7$ bessere Vorhersagen als das AR-Modell. Diese Überlegenheit ist für $h = 11, 12$ statistisch signifikant.

- Bei den iterierten Mehr-Schritt-Prognosen hat unter den vier verschiedenen ARNN-Modellierungsstrategien die statistisch-parametrische Modellierungsstrategie, Param3, eine gute Prognosegenauigkeit. Dieses gute Abschneiden, durchgehend für alle Prognosehorizonte, ist jedoch nur für $h = 2$ signifikant besser als die LBP.

- Mit direkten Prognosemethoden kann für die Prognosehorizonte $h = 1$ bis ca. $h = 8$ keine zusätzliche Prognosegüte bewirkt werden, allerdings zeigen diese Modellvarianten eine robuste und dem linearen Basismodell vergleichbare Prognosegüte. Hingegen sind für $h = 11, 12$ fast alle Prognosen mit direkten Methoden dem LBM überlegen. Da dies jedoch nicht nur die direkten ARNN-Modelle sondern auch das direkte AR-Modell zutrifft und der Unterschied statistisch bis auf eine Ausnahme (Klass2, $h = 10$) nicht signifikant ist, kann dies nicht als Hinweis auf Nicht-Linearität sondern eher auf eine leichte Fehlspezifikation des konventionellen AR-Modells gedeutet werden.

Die Tab. 28 und 29 geben eine Übersicht über die Out-of-Sample-Prognosegüte für den Industrieproduktionsindex. Die Auswertung ist in gleicher Weise strukturiert wie oben.

Die Auswertung der Prognosegüte der primären Modellierungsstrategien präsentiert sich insgesamt recht gleichmäßig, indem die meisten Modelle ähnlich gut abschneiden wie das LBM, für das die MSFE-Werte in der ersten Datenspalte der Tab. 28 angeführt sind. Die Unterschiede zu dieser Vergleichsprognose werden durch den DM-Test nur in seltenen Fällen als statistisch signifikant ausgewiesen, wobei es sich jedoch um Fälle von signifikant schlechterer Prognosegüte handelt.

Im einzelnen können die folgenden Beobachtungen hervorgehoben werden:

- Das ARMA-Modell zeigt durchgehend für alle Prognosehorizonte $h = 1, \ldots, 9$ eine bessere Prognosegenauigkeit als die LBP und bis auf eine einzige Ausnahme auch als jedes der ARNN-Modelle. Dennoch ist diese relativ gute Prognosegüte des ARMA-Modells nicht statistisch signifikant. Ab $h = 10$ dreht die Prognosegüte des ARMA-Modells schnell in einen ungünstigen Bereich und ist dann statistisch signifikant schlechter als die LBP.

- Unter den ARNN-Modellierungsstrategien weist der klassische Ansatz mit Early-Stopping eine sehr gute Prognosegenauigkeit auf. Die Modellierungsstrategie Klass2 liegt in seinem MSFE-Wert meist knapp unter oder über

Tab. 28: Hauptergebnisse zur Out-of-Sample-Prognosegüte von linearen und ARNN-Modellen, MSFE-Werte, Industrieproduktionsindex

	AR	ARMA	Param3	Klass2	Regul2	Bayes2
iterierte Prognosen						
$h = 1$	70.5	60.6	70.0	70.9	67.9	71.0
$h = 2$	82.0	70.5	82.0	77.7	78.4	85.0
$h = 3$	92.0	82.9	96.5	91.2	89.6	94.3
$h = 4$	99.3	93.6	109.0	101.9	105.8	99.7
$h = 5$	108.5	104.4	114.5	107.7	120.9	112.6
$h = 6$	122.7	118.6	126.7	126.2	131.5	125.2
$h = 7$	150.3	145.7	165.0	149.3	162.2	153.6
$h = 8$	166.9	160.0	185.4	163.8	183.0	167.4
$h = 9$	173.6	171.6	196.0	172.6	186.5	172.6
$h = 10$	183.9	185.0	203.4	183.9	196.4	198.9
$h = 11$	180.5	219.4	203.4	188.6	193.5	191.8
$h = 12$	201.0	260.1	228.3	201.1	207.2	216.5
direkte Prognosen						
$h = 1$	70.5		71.5	69.5	72.7	69.7
$h = 2$	78.8		80.3	81.7	78.1	81.5
$h = 3$	86.4		88.1	85.6	87.2	89.1
$h = 4$	98.0		97.6	99.5	105.1	99.7
$h = 5$	104.4		112.8	100.8	116.5	116.9
$h = 6$	121.8		132.1	121.9	157.5	132.5
$h = 7$	155.5		163.8	150.3	158.6	155.4
$h = 8$	168.3		173.1	163.1	176.6	181.3
$h = 9$	182.0		175.6	171.4	185.9	189.3
$h = 10$	197.1		207.6	188.2	211.1	200.0
$h = 11$	205.7		207.1	200.6	234.5	210.6
$h = 12$	222.5		229.3	220.2	222.7	218.8

demjenigen der LBP, sowohl unter Verwendung der iterierten Methode der Mehr-Schritt-Prognose als auch unter Verwendung eines für direkte Prognosen maßgeschneiderten Modells.

- Auch die anderen Modellierungsansätze Param3, Regul2 und Bayes liegen bei niedrigen h manchmal unter dem Vergleichswert der LBG und fallen zumindest bei mittleren h nicht stark gegenüber der LBP zurück.

- Aus den direkten Prognosen ergeben sich bei allen Modellierungsstrategien in niedrigen Prognosehorizonten leichte Vorteile gegenüber den iterierten Prognosen des AR-Modells.

Im Vergleich der IPI-Auswertung mit derjenigen für die Zeitreihe ALR offenbart sich teilweise eine Umkehrung der beobachteten Muster. Während für die ALR das ARMA-Modell sowie direkte Prognosemethoden für hohe Prognosehorizonte überlegen waren, ist es dies für die IPI genau umgekehrt für niedrige und mittlere

Tab. 29: Ergebnisse des Diebold-Mariano-Tests für den Vergleich der Prognosegüte von linearen und ARNN-Modellen, Industrieproduktionsindex

iterierte Prognosen					
	ARMA	Param3	Klass2	Regul2	Bayes2
$h = 1$	0.98	0.04	-0.04	0.25	-0.05
$h = 2$	1.15	0.00	0.42	0.36	-0.30
$h = 3$	0.90	-0.45	0.08	0.24	-0.24
$h = 4$	0.57	-0.97	-0.26	-0.65	-0.04
$h = 5$	0.41	-0.60	0.08	-1.23	-0.40
$h = 6$	0.42	-0.40	-0.35	-0.88	-0.25
$h = 7$	0.46	-1.47	0.09	-1.19	-0.33
$h = 8$	0.69	-1.85 °	0.23	-1.60	-0.05
$h = 9$	0.20	-2.25 °°	0.06	-1.29	0.09
$h = 10$	-0.10	-1.68 °	0.00	-1.25	-1.46
$h = 11$	-2.86 °°	-2.17 °°	-0.59	-1.07	-0.85
$h = 12$	-1.85 °	-2.73 °°	-0.01	-0.60	-0.99
direkte Prognosen					
	AR	Param3	Klass2	Regul2	Bayes2
$h = 1$	0.00	-0.11	0.09	-0.23	0.07
$h = 2$	0.32	0.17	0.03	0.39	0.05
$h = 3$	0.56	0.38	0.64	0.48	0.29
$h = 4$	0.14	0.17	-0.01	-0.58	-0.04
$h = 5$	0.42	-0.42	0.77	-0.80	-0.55
$h = 6$	0.09	-0.94	0.08	-1.57	-0.98
$h = 7$	-0.46	-1.16	0.00	-0.51	-0.43
$h = 8$	-0.07	-0.30	0.21	-0.41	-0.54
$h = 9$	-0.48	-0.17	0.11	-0.62	-1.05
$h = 10$	-0.89	-1.39	-0.19	-1.96	-1.38
$h = 11$	-1.63	-2.66 °°	-0.74	-2.20 °	-2.21 °
$h = 12$	-1.37	-1.44	-1.35	-1.26	-0.85 °

Teststatistiken des Diebold-Mariano-Tests mit der LBP (AR-Modell auf saisonalen Differenzen) als Vergleichsprognose. ** (bzw. °°) und * (bzw. °) bezeichnen Werte, die für einen zweiseitigen Test bei einem Konfidenzniveau von 2 % bzw. 10 % signifikant sind, wobei die Verwendung von ° statt * eine Verschlechterung gegenüber der LBP anzeigt.

Prognosehorizonte. Dies mag als Warnung dienen, die gefundenen Charakterisierungen der Auswertung zu generalisieren.

Insgesamt lässt sich auf der Basis der primären Modelle keine Überlegenheit der ARNN-Methoden belegen, aber auch in den meisten Fällen keine bedeutende oder statistisch signifikante Unterlegenheit. ARNN umfassen in ihrer Spezifikation AR-Modelle, so dass vermutet werden kann, dass die ähnliche Prognosegüte von linearen und ARNN-Modellen daran liegt, dass sie, gezwungen durch die verschiedenen Techniken der Modellierung, nahezu lineare Zusammenhänge abbilden.

5.5 Schlussfolgerungen

In diesem Kapitel wurde eine systematische und umfangreich angelegte Prognoseanordnung entwickelt, um die Out-of-Sample-Prognosegüte von linearen und ARNN-Modellen in Mehr-Schritt-Prognosen zu testen und zu vergleichen. Für die beiden Zeitreihen der österreichischen Arbeitslosenrate (ALR) und des österreichischen Industrieproduktionsindex (IPI) wurde der Zeitraum Jan1998–Dez2002 als Evaluierungsset bestimmt, der in Mehr-Schritt-Prognosen mit Prognosehorizonten von $h = 1$ bis $h = 12$ vorherzusagen war.

Um diese Aufgabenstellung zu bewältigen, mussten zunächst neben der Darlegung der Grundlagen der Analyse von Prognosegüte in einem eigenen Abschnitt die Theorie der Mehr-Schritt-Prognosen erweitert und zusätzliche Methoden entwickelt werden, damit Mehr-Schritt-Prognosen mit ARNN-Modellen auf verschiedene Weisen durchgeführt werden können. Die beiden für die Praxis relevanten Alternativen sind die iterierte und die direkte Methode der Mehr-Schritt-Prognose. Im Bereich der iterierten Mehr-Schritt-Prognose wurde die Notwendigkeit einer simulativ-iterierten Methode begründet und die entsprechenden Methoden vorgestellt. Im Bereich der direkten Methode, wurde das im vorigen Kapitel definierte ARNN- bzw. ARNNDS-Modell hinsichtlich Notation und Spezifikation so erweitert, dass es die für den gewünschten Prognosehorizont maßgeschneiderte Spezifikation des Modells erlaubt.

Die Anordnung des Prognosevergleichs wurde in einem eigenen Abschnitt strukturiert und festgelegt. Die Zielsetzung ist, die Situation eines angewandten Wirtschaftsforschers nachzubilden, der in regelmäßigen Abständen Prognosen zu verschiedenen Prognosehorizonten erstellt. Daher basiert die Prognoseanordnung auf einer Neumodellierung der Zeitreihe jedes Quartal. Um den Rechenaufwand der Evaluierung nicht unnötig zu steigern, wurden jedoch bestimmte Modellierungsentscheidungen nur einmal für jedes Jahr getroffen, etwa die Bestimmung der Regulierungsparameter oder der Lag-Struktur der Modelle. Für die vier verschiedenen Modellierungsansätze wurde jeweils die Vorgangsweise klar definiert und eine Anzahl von Modellierungsstrategien definiert. Um “Data-Snooping“ zu vermeiden und eine unverzerrte Auswertung der Prognosegüte zu ermöglichen, wurde von diesen jeweils eine Modellierungsstrategie als die primäre Modellstrategie festgelegt.

Ausgerüstet mit diesem Instrumentarium werden im Rahmen der Evaluierung für jede der beiden Zeitreihen und jeden Prognosehorizont insgesamt 466 Prognosen erstellt. Der Kern der Auswertung beruht auf dem Vergleich des AR-Modells, des ARMA-Modells und der vier primären ARNN-Modellstrategien, die auf den saisonalen Differenzen der Zeitreihe geschätzt werden. Für diesen Kern zeigt sich kaum eine Überlegenheit der ARNN-Modelle im Vergleich zum AR-Modell, welches als Referenz und als Bezugspunkt im Rahmen der Durchführung eines statistischen Tests auf Prognosegüte verwendet wird. Nur vereinzelt und in einzelnen Bereichen der abgedeckten Prognosehorizonte ist die Prognosegenauigkeit besser. Statistische Signifikanz der Unterschiedlichkeit wird nur sehr selten festgestellt. Die untersuchten primären Modellstrategien schneiden allerdings auch nicht deutlich schlechter als das AR-Modell ab.

Eine vielversprechende Prognosegenauigkeit weisen die direkten Prognosemethoden auf. Dies betrifft allerdings sowohl die linearen als auch die ARNN-Modelle. Bei Verwendung von direkten Prognosemethoden ist die Prognosegenauigkeit in den meisten Prognosehorizonten ähnlich gut wie die konventionelle AR-Modellierung und übertrifft diese für einzelne Prognosehorizonte sogar.

Nimmt man in den Vergleich auch das ARMA-Modell auf, so wird der Wert der ARNN-Modelle als Alternative zur AR-Modellierung zusätzlich in Frage gestellt. Denn die ARMA-Prognosen liefern sowohl auf der Arbeitslosenrate als auch den Industrieproduktionsindex unter allen untersuchten Modellen für einen Großteil der Prognosehorizonte die beste Out-of-Sample-Prognosegüte. Diese Überlegenheit der ARMA-Modelle hängt wesentlich vom Prognosehorizont ab, wobei man im Vorhinein nicht wissen kann, in welcher Weise.

Neben der Abhängigkeit der Prognosegüte vom Prognosehorizont wurde auch die Abhängigkeit von anderen Modellierungsoptionen im Rahmen von mehreren Sensitivitätsanalysen getestet, z.B. die Verwendung von anderen Differenzenfiltern und der unterschiedliche Beginn des Lernsets. Die Ergebnisse unterstreichen zusätzlich, dass neben der sorgfältigen Gestaltung der (linearen und nicht-linearen) Modellierungsstrategie auch andere Modellierungsoptionen mitentscheidend für den Prognoseerfolg sind.

Durch die mangelnde Überlegenheit der ARNN-Modelle in Hinblick auf die Out-of-Sample-Prognosegüte liegt der Schluss nahe, dass in keiner der beiden Zeitreihen ausreichend Nicht-Linearität vorhanden ist, um den Einsatz von ARNN-Modellen vorteilhaft zu machen. Dennoch kann festgestellt werden, dass beim geeigneten Einsatz der in Kapitel 4 entwickelten Strategien zur Sicherstellung der Generalisierungsfähigkeit die Einbußen an Prognosegüte gegenüber einer AR-Modellierung gering sind. Für einen Praktiker, der mit dem Instrumentarium der ARNN-Modelle vertraut ist, kann daher auch bei unsicherem Vorhandensein von nicht-linearen Strukturen der Einsatz von ARNN-Modellen durchaus empfehlenswert sein.

6 Schlussfolgerungen und Ausblick

Die vorliegende Arbeit untersucht die Möglichkeiten, ökonomische Zeitreihen mit neuronalen Netzen (NN), im konkreten mit ARNN-Modellen (Autoregressive Neuronal Network), vorherzusagen, und vergleicht ihre Eignung bei dieser Aufgabenstellung mit jener der AR- und ARMA-Modelle (Autoregressive bzw. Autoregressive Moving Average) als den wichtigsten Vertretern der linearen univariaten Zeitreihenmodelle. Als Anwendungsgebiet werden die beiden monatlichen Zeitreihen der österreichischen Arbeitslosigkeit und des österreichischen Industrieproduktionsindex herangezogen. Die Zielsetzung der Arbeit umfasst sowohl die Gewinnung von Erkenntnissen über die untersuchten Zeitreihen (Was sind die linearen und nichtlinearen Strukturen und sonstigen Eigenschaften der Zeitreihen, die ihre Vorhersagbarkeit bestimmen?) als auch über das eingesetzte Instrumentarium der NN (Welche Techniken und Vorgangsweisen bewähren sich? Mit welchen Modellierungsstrategien können ARNN-Modelle die AR- bzw. ARMA-Modelle in der Prognosegüte übertreffen?).

Diese doppelte Perspektive durchzieht mit wechselnder Schwerpunktsetzung alle vier Hauptkapitel der Arbeit. Bereits am Ende jedes einzelnen Hauptkapitels findet sich eine Zusammenfassung einschließlich der wichtigsten Schlussfolgerungen, die aus den jeweiligen Teilanalysen gezogen werden können. Im Folgenden werden diese Schlussfolgerungen zusammengeführt und eine Gesamtbewertung unternommen. Hierbei soll zusätzlich auch die Ableitung von Empfehlungen für den Praktiker und der Ausblick auf weitere mögliche Schwerpunkte zukünftiger Forschung im Bereich der Prognose von ökonomischen Zeitreihen berücksichtigt werden.

6.1 Erkenntnisse zum Instrumentarium

In der Arbeit werden sowohl lineare als auch NN-Methoden in Hinblick auf die Modellierung von ökonomischen Zeitreihen untersucht. Die Besonderheiten von ökonomischen Zeitreihen, denen die Methoden gerecht werden müssen, sind u.a. mögliche Nicht-Stationarität, Saisonalität, relativ geringe Zeitreihenlänge bei gleichzeitig hoher Stochastizität und möglichen Strukturbrüchen. Obwohl das lineare Instrumentarium bereits seit Jahrzehnten gut erforscht ist, zeigt sich im Rahmen der vorliegenden Untersuchung, dass einige Modellierungsentscheidungen selbst bei linearer Modellierung nicht-trivial sind und vom Anwender große Aufmerksamkeit erfordern. Insbesondere sind dies die richtige Auswahl der Transformation und des geeigneten Differenzenfilters zur Herstellung der Stationarität der Zeitreihe, die richtige Bestimmung der Modellordnung und gegebenenfalls einer sparsamen Modellspezifikation. Um der Saisonalität in den Zeitreihen gerecht zu werden, werden in die Modellformulierung Terme für deterministische Saisonalität aufgenommen (Erweiterung von AR und ARMA zu ARDS bzw. ARMADS).

Im Bereich der ARNN-Modellierung werden sowohl an der Spezifikation des Modells selbst als auch an den verschiedenen Ansätzen zur Schätzung und Modellierung im Vergleich zum Stand der Forschung Erweiterungen vorgenommen, die in Hinblick auf das Anwendungsgebiet geboten erscheinen. So wie wichtige theoretische und angewandte Referenzarbeiten auf dem Gebiet der ARNN-Modellierung (Swanson und White, 1997b; Leisch et al., 1999; Trapletti et al., 2000; Teräsvirta et al., 2005; Medeiros et al., 2006), sieht auch die in der vorliegenden Arbeit verwendete Modellspezifikation die Möglichkeit eines linearen Modellteils vor. Auch die sparsame Modellspezifikation (d.h. nicht alle Koeffizienten bzw. Gewichte sind enthalten) ist nicht neu. Als Neuerung ist jedoch die Möglichkeit eines deterministischen Modellteils zur Berücksichtigung von deterministischer Saisonalität und Trends anzusehen, dessen Koeffizienten gleichzeitig mit den anderen geschätzt werden (Erweiterung von ARNN zu ARNNDS). Die weitreichendste Neuerung ist schließlich die Verallgemeinerung der ARNN-Modelle zur Ermöglichung von direkten Mehr-Schritt-Prognosen für gewünschte Prognosehorizonte $h > 1$.

Für die Schätzung und Modellierung der ARNN bzw. ARNNDS-Modelle wurden vier verschiedene Modellierungsansätze,

- Statistisch-parametrischer Ansatz,
- Klassischer Ansatz mit Early-Stopping,
- Ansatz mit Regularisierung und
- Bayesianischer Ansatz,

verwendet. Diese Modellierungsansätze haben unterschiedliche Ansprüche. Während der statistisch-parametrische Ansatz ein ARNN- bzw. ARNNDS-Modell als ein ökonometrisches Modell betrachtet, dessen Modellspezifikation den in der Ökonometrie üblichen Anforderungen genügen soll, begnügen sich die anderen Ansätze im wesentlichen mit einer Sichtweise des geschätzten Modells als Black-Box. Die Generalisierungsfähigkeit wird auf jeweils unterschiedliche Weise hergestellt. Die Methoden wurden in Kapitel 4 entwickelt und in theoretischer wie praktischer Hinsicht untersucht. Bereits hierbei, mehr noch aber in der systematischen Out-of-Sample-Evaluation (Kapitel 5) zeigen sich die Stärken und Schwächen der Ansätze. Auf einer simulierten Zeitreihe mit bekannter nicht-linearer Struktur kann mit allen Ansätzen eine gute Modellierung und Generalisierungsfähigkeit erreicht werden. Auf den beiden realweltlichen Zeitreihen ist ein Urteil hingegen schwieriger zu treffen, und es zeigen sich auch Unterschiede zwischen den Ansätzen.

Der statistisch-parametrische Ansatz stellt ein Instrumentarium dar, dessen anspruchsvolle Annahmen in der Praxis kaum gegeben sind. Growing- und Pruning-Strategien, mit denen die Generalisierungsfähigkeit hergestellt werden soll, erfordern zusätzlich die Festlegung von Vorgangsweisen, die selbst nicht mehr theoretisch sondern bloß heuristisch hergeleitet werden können. Bereits bei relativ einfachen Modellen treten mehrfache lokale Minima in der Fehlerlandschaft auf, ein

Problem, das in der asymptotischen Theorie keinen großen Stellenwert einnimmt, die praktische Verwendbarkeit des statistisch-parametrischen Ansatzes jedoch einschränkt. Die Ergebnisse mit dem statistisch-parametrischen Ansatz in Hinblick auf die Generalsierungsfähigkeit sind dennoch als gut einzuschätzen. Die Ergebnisse der Out-of-Sample-Prognosegüte liegen entweder knapp ober oder unter dem linearen Vergleichsfall.

Mit den anderen Ansätzen zeigt sich zwar im Prinzip eine ebenfalls brauchbare Generalisierungsfähigkeit, die jedoch teilweise stärker von den Details der Modellspezifikation abhängt. Insbesondere kann hier die Entscheidung, direkte Verbindungen in das Modell aufzunehmen, und die Anzahl der verdeckten Neuronen eine beträchtliche Auswirkung auf die Prognosegüte haben. Ein grundsätzliches Problem ist hierbei, dass bei diesen Ansätzen der Zufall, der mit der Initialisierung der Parameter und der Ausgestaltung des Lernverfahrens einhergeht, eine teilweise beträchtliche Auswirkung haben kann. Der in der Forschungsliteratur oft behauptete Anspruch, dass bei vorzeitig gestopptem Training, bei regularisiertem Lernen und beim Bayesianischen Ansatz die Prognosegüte weitgehend unabhängig von der Anzahl der verdeckten Neuronen bzw. der Parameter sein sollte, kann in der Anwendung auf realweltlichen Zeitreihen in nur sehr eingeschränktem Ausmaß eingelöst werden. In der Out-Sample-Evaluierung dieser Modellierungsansätze zeigt sich nämlich, dass die erzielten Ergebnisse tendenziell etwas schlechter im Vergleich zu denen mit der linearen Basismodellierung und auch zu denen des statistisch-parametrischen Ansatzes sind und bei jenen Modellstrategien mit mehreren verdeckten Neuronen vereinzelt sehr schlechte Prognosen vorkommen. Insbesondere der Bayesianische Ansatz weist bei einer höheren Anzahl von verdeckten Neuronen eine schlechte Prognosequalität auf. Dies lässt auf grundsätzliche Schwächen des Instrumentariums in der Modellierung von überparametrisierten Modellen und bei Zeitreihen ohne ausreichend vorhandener nicht-linearer Struktur schließen.

6.2 Erkenntnisse zu den Zeitreihen

Unter den Erkenntnissen zu den Zeitreihen interessiert vor allem die Antwort auf die Frage, ob in den Daten ausreichende nicht-lineare Strukturen vorhanden sind, sodass sie mit ARNN-Modellen im Vergleich zu AR- und ARMA-Modellen besser vorhergesagt werden können. Zwar ergeben sich auch aus den Analysen der Kapitel 2–4 Hinweise, dass dies nur in beschränktem Ausmaß der Fall ist. Doch das Hauptergebnis der Arbeit in Hinblick auf diese Frage wird von der systematischen Evaluation der Out-of-Sample-Prognosegüte der verschiedenen Modelle bzw. Modellstrategien im Rahmen des Kapitels 5 geliefert. Und hier zeigt sich, dass die ARNN- bzw. ARNNDS-Modelle im Vergleich zu den linearen AR- und ARMA-Modellen kaum relevante zusätzliche Leistungssteigerung bei der Prognose der beiden Beispielzeitreihen bringen. Allenfalls vorhandene Nicht-Linearitäten sind demnach nur schwach ausgeprägt oder ihre Nutzbarmachung für Prognosezwecke wird von an-

deren Eigenschaften der Zeitreihen wie Heteroskedastizität und Strukturbrüchen erschwert.

Angesichts der Vielzahl von Modellierungsstrategien und -optionen, die in der Evaluation erprobt werden, fokussiert die Auswertung der systematischen Prognoseevaluation auf jeweils eine primäre Modellierungsstrategie für jeden der vier Ansätze der ARNN-Modellierung, wobei die saisonal differenzierten Zeitreihen die Basis der Modellierung bilden. Durch diese Einschränkung soll es vermieden werden, dass die zuvor erwähnten Schwächen der verschiedenen Modellierungsansätze, die sich teilweise bereits im Zuge der Entwicklung des Instrumentariums im Rahmen des Kapitels 4 gezeigt haben, den Blick auf die Beurteilung der tatsächlichen Prognostizierbarkeit der beiden Zeitreihen der Arbeitslosenrate und des Industrieproduktionsindex verstellen. Auch bei dieser Fokussierung ist eine große Anzahl von Prognosen zu berücksichtigen, da sowohl die iterierten als auch die direkten Mehr-Schritt-Prognosen jeweils für die Prognosehorizonte $h = 1, \ldots, 12$ zu erstellen sind. Dies muss bei der Interpretation des Vergleichs der Prognoseperformance und insbesondere der Ergebnisse des Diebold-Mariano-Tests berücksichtigt werden. Es finden sich unter den zahlreichen Prognosen nur wenige Fälle, in denen die Zeitreihen mit ARNN-Modellen besser prognostizierbar sind als mit dem hauptsächlich betrachteten linearen Modell eines AR-Modells. In noch weniger Fällen weist das Testergebnis des Diebold-Mariano-Tests auf eine statistisch-signifikante Unterschiedlichkeit der Prognoseperformance hin.

Neben der Frage nach den nicht-linearen Strukturen liefert die Arbeit auch Ergebnisse zu verschiedenen anderen Eigenschaften der Zeitreihen, die auch für die lineare Modellierung relevant sind. So kann hier insbesondere hervorgehoben werden, dass sich die Annahme eines saisonalen stochastischen Trends in den beiden Zeitreihen, der durch die Anwendung des saisonalen Differenzenfilters bereinigt werden kann, nicht nur aufgrund der In-Sample-Analysen empfiehlt, sondern sich auch in der Out-of-Sample-Evaluierung bewährt. Eine Besonderheit sowohl der linearen als auch der ARNN-Modelle, die für die beiden Zeitreihen angepasst werden, ist die relativ hohe Modellordnung bzw. hohe Anzahl von Modellparameter. In der systematischen Out-of-Sample-Prognoseevaluation erweist sich jedoch, dass diese bei direkten Mehr-Schritt-Prognosen mit wachsendem Prognosehorizont abnimmt. Insgesamt ergibt sich, dass für Prognosezwecke auch die scheinbar einfachen Modellierungsoptionen im Vergleich zur Frage des Einsatzes eines nicht-linearen Modells nicht unterschätzt werden sollten.

6.3 Empfehlungen für den Praktiker

Letztlich ist das Ergebnis der vorliegenden Arbeit nicht so sehr die Entscheidung der Frage ob nun die neuronalen Netze oder die traditionellen linearen Modelle „besser“ sind oder ob in den beiden untersuchten Zeitreihen nicht-lineare Struktur enthalten ist, sondern die Vermittlung eines ganzheitlichen Überblicks über die theoretischen

und praktischen Probleme, die bei der Anwendung linearer (AR, ARMA) sowie ARNN-Modelle im Anwendungsbereich der ökonomischen Zeitreihenmodellierung und -prognose auftreten. Denn für anderere Zeitreihen mögen die Ergebnisse zur Vorteilhaftigkeit der einen oder anderen Modellvariante anders ausfallen, die Probleme, die sich dem Praktiker stellen, bleiben dennoch die gleichen.

Angesichts dessen stellt sich die Frage, ob das entwickelte Instrumentarium der ARNN-Modelle für praktische Anwendungsfälle im Bereich ökonomischer Zeitreihen geeignet ist und ob für den Praktiker Empfehlungen abgegeben werden können. Unter einer Reihe von Vorbehalten, auf die in der vorliegenden Arbeit immer wieder, insbesondere in den Schlussfolgerungen zu den einzelnen Hauptkapiteln, hingewiesen wurde, kann diese Frage positiv beantwortet werden: Die ARNN-Modelle können in der hier vorgestellten Spezifizierung und in Verbindung mit den beschriebenen Modellierungsstrategien in zielführender Weise zur Prognose von ökonomischen Zeitreihen eingesetzt werden. Die Gründe, die zu diesem Urteil führen, können wie folgt zusammengefasst werden:

- Das Instrumentarium der ARNN-Modellierung ist theoretisch gut abgesichert. Dies gilt insbesondere für den statistisch-parametrischen Ansatz, für den in Abschnitt 4.6 die entsprechende Literatur aufgearbeitet wurde. Aber auch für die anderen Ansätze lässt sich feststellen, dass sie das Generalisierungsproblem prinzipiell in befriedigender Weise zu lösen im Stande sind. Ein erfahrener Zeitreihenanalytiker sollte daher keine großen Schwierigkeiten haben, das Potenzial von ARNN-Modellen in konkreten Anwendungen richtig einzuschätzen, dies umso mehr, als die meisten der für die ARNN-Modellierung zu lösenden Probleme auch in der linearen Modellierung von prinzipieller und praktisch relevanter Bedeutung sind. Wie anhand der beiden Beispielzeitreihen gezeigt werden konnte, sind viele der Probleme (Identifiziertheit, Spezifikation, ausreichende Samplegröße, Erstellung von Mehr-Schritt-Prognosen) grundsätzlicher Natur und beeinflussen auch im linearen Fall die Prognosegüte entscheidend mit.

- Das Instrumentarium der ARNN-Modellierung kann heute als gut zugänglich für die praktische Anwendung gelten. Neben den im Rahmen der vorliegenden Arbeit entwickelten Software-Implementationen des R-Pakets `NNUTS` existieren eine Vielzahl von alternativen Möglichkeiten, u.a. weitere Funktionen und Pakete in der statistisch-mathematischen Programmiersprache R. Die zusätzlich erforderliche Rechenzeit, die in früheren Zeiten für NN-Methoden ein Hindernis darstellen konnte, ist heute durch die verfügbare Rechenkraft, mehr aber noch durch die professionelle und benutzerfreundliche Implementierung der Algorithmen, meist vernachlässigbar. Praktische Probleme mit Rechenzeit und numerische Schwächen der Schätzverfahren sollten nicht dem Instrumentarium zugeschrieben werden, sondern als Hinweis auf Eigenschaften der Zeitreihen genommen werden.

Die Vorbehalte, die für eine fruchtbare praktische Anwendung beachtet werden sollten, werden nun ebenso zusammengefasst:

- Die Literatur zu NN im allgemeinen und ARNN-Methoden im speziellen ist noch immer uneinheitlich und vielfältig in ihrer Ausrichtung, so dass teilweise veraltete Vorstellungen und Empfehlungen vorzufinden sind, denen der Praktiker nicht Folge leisten sollte. Dies betrifft zum Beispiel die Vorstellung, dass NN auch ohne entsprechende Transformation der Zeitreihen oder sorgfältige Spezifizierung entsprechender Modellterme mit nicht-stationären Komponenten zurechtkommen können. Die vorliegende Arbeit bemüht sich, die relevante Literatur aufzuarbeiten und in der Spezifizierung der ARNNDS-Modelle einen gangbaren Weg vorzugehen. Ein weiteres Beispiel ist die Vorstellung, dass die Modellierungsstrategien des klassischen Modellierungssansatzes mit Early-Stopping und mit regularisiertem Training (inkl. mit dem Bayesianischen Ansatz) auch bei höherer Anzahl von verdeckten Neuronen die Generalisierungsfähigkeit sicherstellen können. Dies ist nur in sehr beschränktem Ausmaß der Fall, wie sich insbesondere aus dem systematischen Prognosevergleich (Kapitel 5) zeigt, und die Anzahl der verdeckten Neuronen sowie der Netzparameter insgesamt sollte daher immer möglichst gering gehalten werden.

- Bei der Modellierung mit NN spielen numerische Aspekte eine im Vergleich zu linearen Modellen ungleich größere Rolle. In dieser Arbeit wird gezeigt, dass bereits bei relativ einfach gehaltenen Modellen mehrere lokale Minima in der Fehlerfunktion der zu schätzenden Modelle auftreten. Dies erfordert den Einsatz besonderer Methoden und allgemeine Vorsicht durch den Zeitreihenanalytiker. Bei Verwendung des statistisch-parametrischen Ansatzes der Modellierung sollte das Auftreten von mehreren lokalen Minima sowie von anderen numerischen Schwierigkeiten im Zusammenhang mit der Hesse-Matrix als dringender Hinweis auf mögliche Nicht-Identifiziertheit und Fehlspezifikation genommen werden.

- Das Instrumentarium der NN kann zwar prinzipiell auch bei nicht vorhandener Nicht-Linearität eingesetzt werden, jedoch ist in diesem Falle mit besonderer Vorsicht vorzugehen. Die Nicht-Linearität lässt sich mit den Nicht-Linearitätstests des Kapitels 3 sowie mit dem Teräsvirta-Lin-Granger-Test im Rahmen des statistisch-parametrischen Ansatzes jedoch nicht leicht nachweisen, da unklar ist, wie sich die Macht und Größe der Tests bei Verletzung der Annahmen der Tests verhalten. Insbesondere Heteroskedastizität, Strukturbrüche spielen hier eine Rolle.

- Als ein sehr grundsätzlicher Vorbehalt bzw. Nachteil ist die teilweise beträchtliche Zufallsabhängigkeit der Modellierungsergebnisse mit NN anzusehen, d.h. der Umstand, dass bei mehrmaliger Wiederholung des Modellierungs

bzw. des Schätzvorgangs mitunter stark unterschiedliche Ergebnisse erhalten werden. Diese Zufallsabhängigkeit ist einigen der Modellierungsansätze inhärent (insbesondere dem klassischen Ansatz mit Early-Stopping). Doch auch im statistisch-parametrischen Ansatz, obwohl dieser durch den Aufbau auf den Annahmen der Asymptotik die Annahme der Nicht-Zufälligkeit des Optimierungsverfahrens impliziert, ist die Zufälligkeit eine meist unvermeidbare Folge des Auftretens von lokalen Minima. Der Praktiker sollte daher die Schätzung mit ARNN und bei gleicher oder variierter Modellierungsstrategie mehrmals wiederholen und auch mit dem linearen Fall vergleichen, um die Implikationen der Zufälligkeit richtig einschätzen zu können.

- Wie die Ergebnisse der Prognose- und Evaluierungsanordnung zeigen, spielen neben der Berücksichtigung möglicher Nicht-Linearitäten durch ARNN und neben einer angemessenen Stategie zur Sicherstellung der Generalisierungsfähigkeit auch andere Modellierungsoptionen eine wichtige Rolle für die Prognosequalität. Es sind dies die Berücksichtigung von Saisonalität, die Wahl des Differenzenfilters, die Abgrenzung der In-Sample-Zeitreihe und die Methode der Mehr-Schritt-Prognose. Über den zusätzlichen Aufwand, der mit der ARNN-Modellierung verbunden ist, sollte der Zeitreihenanalytiker nicht versäumen, auch diesen Aspekten die gebührende Aufmerksamkeit zu widmen.

6.4 Ausblick auf mögliche weitere Forschung

In dieser Arbeit werden viele Möglichkeiten, lineare (AR, ARMA) oder ARNN-Modelle zur Modellierung und Prognose von ökonomischen Zeitreihen einzusetzen und weiterzuentwickeln vorgestellt. Aufgrund der konkreten Ergebnisse und der allgemeinen Erfahrungen, die im Verlauf dieser Arbeit gesammelt wurden, kann auf zahlreiche zusätzliche Möglichkeiten und Wege hingewiesen werden, die zukünftiger Forschung vorbehalten sein müssen. Dies betrifft einerseits Fragen der Modellierung und Prognose von ökonomischen Zeitreihen insgesamt, die auch beim Einsatz linearer Modelle zum Tragen kommen, als auch die gezielte Verbesserung des Instrumentariums der neuronalen Netze.

Die gegenwärtige Arbeit zeigt, dass viele der Methoden und Modellierungsstrategien, die speziell im Zusammenhang mit neuronalen Netzen entwickelt wurden, auch für die Modellierung mit linearen Modellen eingesetzt werden können. Auch bei linearen Modellen ist es angesichts möglicher Überparametrisierung und anderer grundlegender Problem wünschenswert, dem Zeitreihenprognostiker mehrere Alternativen zur Sicherung der Generalisierungsfähigkeit von Modellen bereitstellen zu können. Aus der Sicht der vorliegenden Arbeit besteht eine noch nicht ausreichend erforschte Lücke z.B. im Bereich der Anwendung des Ansatzes mit Regularisierung und des Bayesianischen Ansatzes auf lineare Modelle. Zusätzliche Aspekte wie die Berücksichtigung von Mehr-Schritt-Prognosen durch iterative oder direkte Methoden tragen zur Attraktivität dieser Forschungsrichtung bei.

Für die ARNN-Modellierung im Anwendungsgebiet der ökonomischen Zeitreihen, in denen allenfalls vorhandene nicht-lineare Strukturen nur schwach ausgeprägt sind und eventuell störende Merkmale wie Strukturbrüche und Heteroskedastizität hinzutreten, entsteht der Eindruck, dass diese tendenziell zu mächtig und flexibel sind und die verschiedenen in dieser Arbeit untersuchten Strategien zur Sicherstellung der Generalisierungsfähigkeit relativ umständlich und teilweise fehleranfällig sind. In der Suche nach grundsätzlicheren und einfacheren Methoden sind verschiedene Ansätze denkbar, die einerseits bei der Spezifikation der Modelle ansetzen können oder andererseits die Vorgangsweisen der Modellierung betreffen.

Man könnte in die Modellspezifikation zusätzliche Restriktionen einführen, die das Ausmaß der Nicht-Linearität deutlich einschränken. Dies wäre etwa erreichbar, indem die Inputschicht des Neuronalen-Netz-Teils eines ARNN-Modells nicht vollständig, sondern in einer vorgegebenen sparsamen Weise (z.B. eins-zu-eins) mit den Neuronen der verdeckten Schicht verbunden sind. Andere Parameterrestriktionen könnten gezielt entwickelt werden, um die potentielle Nicht-Linearität oder andere Aspekte des Modells wie das Stationaritätsverhalten einzuschränken. Durch solche Methoden würden die neuronalen Netze noch näher in das Kerngebiet der nicht-linearen Zeitreihenanalyse geholt werden und unmittelbarer an die Seite von anderen nicht-linearen Modellen wie z.B. TAR-Modellen treten.

Mögliche Innovationen bei den Vorgangsweise der Modellierung sind insbesondere beim regularisierten Lernen denkbar. Mit neuartigen Definitionen des Regularisierungsterms könnte versucht werden, ein ARNN-Modell in eine bestimmte Richtung zu beeinflussen, etwa um Nicht-Stationarität oder das Phänomen der Pseudostationarität zu vermeiden. Weitere Bereiche, in denen zusätzliche Forschung vielversprechend sein könnte, sind die Bedingungen für das Auftreten vom multiplen lokalen Minima und die Möglichkeiten deren Anzahl zu schätzen, sowie spezifische Verbesserungen für Growing- und Pruning-Verfahren.

Bereits in der Einleitung wurde darauf hingewiesen, dass das in den letzten Jahren etwas nachlassende Interesse an neuronalen Netzen möglicherweise zu einem guten Teil als Modetrend zu interpretieren ist und nicht auf die Unattraktivität des Forschungsgebietes per se zurückzuführen ist. Tatsächlich wurden in dieser Arbeit viele interessante Forschungsfragen aufgezeigt, die neuronale Netze im allgemeinen und ARNN-Modelle im speziellen weiterhin zu einem attraktiven Forschungsgebiet machen. Was jedoch besonders notwendig ist, ist eine noch stärkere Integration dieser Methoden und Modelle in die Zeitreihenökonometrie.

A Ableitungen der Fehlerfunktion im ARNN-Modell

Die Schätzung eines ARNN-Modells impliziert die Optimierung der Fehlerfunktion, im Rahmen dieser Arbeit die Minimierung der Summe der quadrierten Abweichungen (SSE). Die in Kap. 4 beschriebenen Optimierungsalgorithmen beinhalten die Berechnung der partiellen Ableitungen der Fehlerfunktion nach den einzelnen Parametern des Modells. Im folgenden werden für zwei Fälle die Berechnungsformeln für die Ableitungen präsentiert: erstens für den einfachen Spezialfall eines ARNN-Modell ohne linearen Teil , das einem neuronalen Netz mit einer Bias-Einheit in beiden Schichten entspricht, und zweitens für den generellen Fall eines ARNN-Modells, in dem der nicht-lineare Teil um einen linearen autoregressiven Teil und um einen deterministischen Teil erweitert wurde (ARNNDS-Modell). Der deterministische Teil umfasst hierbei einen Trend und einen saisonalen Teil. Für alle Unterarten dieses generell formulierten Modells, also etwa für das reine AR-Modell oder das reine ARNN-Modell, kann man die Ableitungen durch Weglassen der entsprechenden Terme gewinnen. Die Ableitungen wurden für den einzelnen Datenpunkt formuliert. Durch Summierung über alle Datenpunkte gewinnt man die Ableitung der gesamten Fehlerfunktion. Dies ist insbesondere für die Umsetzung im Rahmen einer matrizenorientierten Programmiersprache wichtig. Die Berechnungsformeln der Ableitungen machen intensiven Gebrauch von der Summenregel, der Produktregel und der Kettenregel.

A.1 Fall A: ARNN-Modell

Modell:

$$y_t = \tilde{y}_t + d$$

$$\tilde{y}_t = \sum_{k=1}^{K} c_k \psi[b_k + \sum_{j \in \mathcal{J}} a_{kj} \tilde{y}_{t-j}] + \varepsilon_t, \tag{A.1}$$

wobei y_t der zum Zeitpunkt t beobachtete Wert der Zeitreihe ist, d, c_k, b_k und a_{kj} die Parameter des Modells sind, K die Anzahl der verdeckten Neuronen (Hidden-Units) ist, $\mathcal{J}$ die Menge der Lags j ist, zu denen vergangene $\tilde{y}_t$ als Regressoren in das Modell eingehen, ψ die Tangenshyperbolicus-Funktion ist und ε_t ein Fehlerterm ist.

Fehlerfunktion:

$$E = (\tilde{y}_t - \sum_{k=1}^{K} c_k \psi[b_k + \sum_{j \in J} a_{kj} \tilde{y}_{t-j}])^2 \tag{A.2}$$

Um eine kürzere Schreibweise zu erreichen, werden im folgenden die Terme H_k und

G definiert,

$$H_k = b_k + \sum_{j \in \mathcal{J}} a_{kj} \tilde{y}_{t-j} \tag{A.3}$$

$$G = \sum_{k=1}^{K} c_k \psi[H_k], \tag{A.4}$$

sodass $E = (\tilde{y}_t - G)^2$ geschrieben werden kann.

Partielle Ableitungen:

$$\begin{aligned} \frac{\partial E}{\partial d} &= 2(\tilde{y}_t - G)(\frac{\partial \tilde{y}_t}{\partial d} - \frac{\partial G}{\partial d}) = 2(\tilde{y}_t - G)(-1 - \sum_{k=1}^{K} c_k \psi'[H_k] \frac{\partial H_k}{\partial d}) = \\ & -2(\tilde{y}_t - G)(1 - \sum_{k=1}^{K} (c_k \psi'[H_k] \sum_{j \in J} a_{kj})), \end{aligned} \tag{A.5}$$

wobei ψ' die Ableitung der Tangenshyperbolicus-Funktion ist.

$$\frac{\partial E}{\partial c_k} = 2(\tilde{y}_t - G)(\frac{\partial \tilde{y}_t}{\partial c_k} - \frac{\partial G}{\partial c_k}) = -2(\tilde{y}_t - G)\psi[H_k] \tag{A.6}$$

$$\frac{\partial E}{\partial b_k} = 2(\tilde{y}_t - G)(\frac{\partial \tilde{y}_t}{\partial b_k} - \frac{\partial G}{\partial b_k}) = -2(\tilde{y}_t - G)c_k \psi'[H_k] \tag{A.7}$$

$$\frac{\partial E}{\partial a_{kj}} = 2(\tilde{y}_t - G)(\frac{\partial \tilde{y}_t}{\partial a_{kj}} - \frac{\partial G}{\partial a_{kj}}) = -2(\tilde{y}_t - G)c_k \psi'[H_k]\tilde{y}_{t-j} \tag{A.8}$$

A.2 Fall B: ARNNDS-Modell

Modell:

$$y_t = \tilde{y}_t + d_t,$$

$$d_t = d + d^t t + \sum_{i=2}^{s} d_i^s s_{it},$$

$$\tilde{y}_t = \sum_{k=1}^{K} c_k \psi[b_k + \sum_{j \in \mathcal{J}} a_{kj} \tilde{y}_{t-j}] + \sum_{l \in \mathcal{L}} f_l \tilde{y}_{t-l} + \varepsilon_t, \tag{A.9}$$

wobei zusätzlich zu den oben in Fall A definierten Symbolen folgende hinzukommen: f_l, d^t und d_i^s sind Parameter des Modells, $\mathcal{L}$ bezeichnet die Menge der Lags l, zu denen vergangene $\tilde{y}_t$ als Regressoren in das Modell eingehen, und s_{it} ist eine saisonale Dummyvariable.

Fehlerfunktion:

$$E = (\tilde{y}_t - \sum_{k=1}^{K} c_k \psi[b_k + \sum_{j \in \mathcal{J}} a_{kj} \tilde{y}_{t-j}] - \sum_{l \in \mathcal{L}} f_l \tilde{y}_{t-l})^2 \tag{A.10}$$

Im folgenden werden zur Abkürzung der Schreibweise neben H_k (siehe oben Fall A) G_1, G_2 und G wie folgt definiert:

$$G_1 = \sum_{k=1}^{K} c_k \psi[H_k],\ G_2 = \sum_{l \in \mathcal{L}} f_l \tilde{y}_{t-l}, \text{und } G = G_1 + G_2, \tag{A.11}$$

sodass $E = (\tilde{y}_t - G_1 - G_2)^2 = (\tilde{y}_t - G)^2$ geschrieben werden kann.

Partielle Ableitungen:

$$\begin{aligned} \frac{\partial E}{\partial d} &= 2(\tilde{y}_t - G_1 - G_2)(\frac{\partial \tilde{y}_t}{\partial d} - \frac{\partial G_1}{\partial d} - \frac{\partial G_2}{\partial d}) = \\ & -2(\tilde{y}_t - G)(1 - \sum_{k=1}^{K}(c_k \psi'[H_k] \sum_{j \in \mathcal{J}} a_{kj}) - \sum_{l \in \mathcal{L}} f_l), \end{aligned} \tag{A.12}$$

$$\frac{\partial E}{\partial c_k} = 2(\tilde{y}_t - G_1 - G_2)(\frac{\partial \tilde{y}_t}{\partial c_k} - \frac{\partial G_1}{\partial c_k} - \frac{\partial G_2}{\partial c_k}) = -2(\tilde{y}_t - G)\psi[H_k] \tag{A.13}$$

$$\frac{\partial E}{\partial b_k} = 2(\tilde{y}_t - G_1 - G_2)(\frac{\partial \tilde{y}_t}{\partial b_k} - \frac{\partial G_1}{\partial b_k} - \frac{\partial G_2}{\partial b_k}) = -2(\tilde{y}_t - G)c_k \psi'[H_k] \tag{A.14}$$

$$\frac{\partial E}{\partial a_{kj}} = 2(\tilde{y}_t - G_1 - G_2)(\frac{\partial \tilde{y}_t}{\partial a_{kj}} - \frac{\partial G_1}{\partial a_{kj}} - \frac{\partial G_2}{\partial a_{kj}}) = -2(\tilde{y}_t - G)c_k \psi'[H_k]\tilde{y}_{t-j} \tag{A.15}$$

$$\frac{\partial E}{\partial f_l} = 2(\tilde{y}_t - G_1 - G_2)(\frac{\partial \tilde{y}_t}{\partial f_l} - \frac{\partial G_1}{\partial f_l} - \frac{\partial G_2}{\partial f_l}) = -2(\tilde{y}_t - G)\tilde{y}_{t-l} \tag{A.16}$$

$$\begin{aligned} \frac{\partial E}{\partial d^t} &= 2(\tilde{y}_t - G_1 - G_2)(\frac{\partial \tilde{y}_t}{\partial d^t} - \frac{\partial G_1}{\partial d^t} - \frac{\partial G_2}{\partial d^t}) = \\ & -2(\tilde{y}_t - G)(t - \sum_{k=1}^{K}(c_k \psi'[H_k] \sum_{j \in \mathcal{J}} a_{kj}(t-j)) - \sum_{l \in \mathcal{L}} f_l(t-l)) \end{aligned} \tag{A.17}$$

$$\begin{aligned} \frac{\partial E}{\partial d_i^s} &= 2(\tilde{y}_t - G_1 - G_2)(\frac{\partial \tilde{y}_t}{\partial d_i^s} - \frac{\partial G_1}{\partial d_i^s} - \frac{\partial G_2}{\partial d_i^s}) = \\ & -2(\tilde{y}_t - G)(s_{it} - \sum_{k=1}^{K}(c_k \psi'[H_k] \sum_{j \in \mathcal{J}} a_{kj} s_{i(t-j)}) - \sum_{l \in \mathcal{L}} f_l s_{i(t-l)}) \end{aligned} \tag{A.18}$$

B Ergänzende Ergebnisse zu den Sensitivitätsanalysen

In diesem Anhang werden die ergänzenden Tabellen für verschiedene Sensitivitätsanalysen zusammengestellt, die in Abschnitt 5.4 besprochen werden. Wie schon die Tabellen zu den hauptsächlichen Ergebnissen der Prognose- und Evaluierungsanordnung, so berichten auch diese Tabellen die Prognosegüte ausgedrückt in relativen MSFE-Werten, d.h. in Prozent der Linearen Basisprognose, die vom AR-Modell auf den saisonalen Differenzen geliefert wird, bzw. in Prozent des MSFE eines anderen angegebenen Referenzmodells. Auch für diesen Teil der Evaluierung werden die Vorgangsweisen und Modellierungsstrategien verwendet, die in Abschnitt 5.3 beschrieben wurden.

Es sind vier verschiedene Sensitivitätsanalysen durch Tabellen dokumentiert:

Unterschiedliche Differenzenfilter Die hauptsächliche Auswertung (vgl. 5.4) beruht auf dem bevorzugten Differenzenfilter ∇_{12}. Ergänzend dazu sind Ergebnisse, die alternativ auf ∇Y_t und $\nabla_{12}\nabla Y_t$ erzielt werden, Gegenstand der Tabellen 30–37.

Unterschiedliche Methoden der iterierten Prognose Für iterierte Mehr-Schritt-Prognosen mit ARNN-Modellen stehen als prinzipielle Alternativen die naiv-iterierte und die simulativ-iterierte Prognosemethode zur Verfügung (vgl. Abschnitt 5.2). In der Prognoseanordnung wird aus prinzipiellen Gründen der simulativ-iterierten Prognosemethode der Vorzug gegeben, da sie bei überschaubarem Rechenaufwand Prognosen liefert, die für ein gegebenes geschätztes Modell korrekt sind. Die Tabellen 38–45 enthalten die Auswertung, wieviel sich hinsichtlich der Prognosegüte ändert, wenn statt dessen die naiv-iterierte Prognose verwendet wird.

Weglassung von Direktverbindungen Zur Frage der Aufnahme von Direktverbindungen (Koeffizienten f_l des ARNN-Modells) gibt es in der NN-Literatur keine eindeutigen Empfehlungen. Im Rahmen der Prognoseanordnung dieses Kapitels enthalten die primären Modellstrategien Direktverbindungen. Für den klassischen Ansatz mit Early-Stopping, den Ansatz mit Regularisierung und den Bayesianischen Ansatz wird in den Tabellen 46–48 untersucht, wieviel sich an der Prognosegüte ändert, wenn man die Direktverbindungen weglässt.

Unterschiedlicher Startzeitpunkt der Zeitreihe Die Beispielzeitreihen ALR und IPI starten Jan1960, was im Rahmen der Prognoseanordnung In-Samples von mehr als 400 Datenpunkten erlaubt. Setzt man den Start der Zeitreihe, die für die Schätzung der Modelle benutzt wird, erst mit Jan1975 an, verkürzt sich die Zeitreihe beträchtlich. Die Tabellen 49–64 dokumentieren die Änderungen für die Prognosegüte, die sich bei Jan1975 als Startzeitpunkt ergeben.

Tab. 30: Out-of-Sample-Prognosegüte mit dem statistisch-parametrischen Ansatz auf Basis von alternativen Differenzenfiltern, ALR

	Param1 ∇Y_t	Param2 ∇Y_t	Param3 ∇Y_t	Param4 ∇Y_t	Param1 $\nabla_{12}\nabla Y_t$	Param2 $\nabla_{12}\nabla Y_t$	Param3 $\nabla_{12}\nabla Y_t$	Param4 $\nabla_{12}\nabla Y_t$
iterierte Prognosen								
$h = 1$	97.2	100.6	115.2	115.8	98.2	98.2	102.2	104.5
$h = 2$	99.6	109.3	118.4	98.8	99.7	99.3	90.6	97.5
$h = 3$	89.3	111.8	119.6	115.3	91.7	94.7	91.0	101.6
$h = 4$	92.6	113.9	123.8	121.1	90.2	100.9	95.0	94.4
$h = 5$	97.8	109.5	139.2	127.9	94.5	105.4	102.7	90.9
$h = 6$	101.4	106.9	144.7	125.4	93.2	107.3	113.8	87.6
$h = 7$	102.6	103.8	141.3	124.2	93.3	107.0	111.4	87.6
$h = 8$	101.0	99.1	135.0	121.4	96.3	112.9	116.6	91.1
$h = 9$	98.9	95.3	131.6	118.6	99.0	119.7	119.2	96.6
$h = 10$	96.9	97.8	131.6	113.9	104.0	120.9	125.9	101.9
$h = 11$	96.2	95.0	128.4	113.8	106.2	124.3	130.1	106.2
$h = 12$	91.1	94.7	123.5	112.3	109.4	125.0	129.9	110.3
direkte Prognosen								
$h = 1$	95.8	110.6	111.7	136.2				
$h = 2$	102.1	97.2	101.7	94.0				
$h = 3$	103.1	113.3	109.0	148.1				
$h = 4$	109.8	113.7	124.8	118.3				
$h = 5$	102.3	104.1	122.0	131.7				
$h = 6$	124.9	123.2	123.1	111.3				
$h = 7$	105.3	119.1	142.6	117.4				
$h = 8$	103.6	110.2	115.8	124.5				
$h = 9$	110.4	126.0	118.8	124.9				
$h = 10$	111.6	107.3	120.8	117.9				
$h = 11$	114.9	118.8	115.2	127.5				
$h = 12$	127.4	137.0	117.5	122.7				

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Tab. 31: Out-of-Sample-Prognosegüte mit dem klassischen Ansatz auf Basis von alternativen Differenzenfiltern, ALR

	Klass1	Klass2	Klass3	Klass4	Klass5	Klass6	Klass7	Klass8
gewöhnliche Differenzen, ∇Y_t, iterierte Prognosen								
$h = 1$	114.2	116.6	125.5	118.5	141.4	142.6	129.2	140.6
$h = 2$	112.7	120.1	128.7	114.7	152.1	161.0	117.5	159.5
$h = 3$	118.8	129.7	128.3	123.8	150.1	161.8	111.6	169.0
$h = 4$	115.9	136.1	137.3	129.1	128.8	128.4	102.1	169.9
$h = 5$	120.8	131.5	139.6	130.0	111.2	109.3	103.7	151.5
$h = 6$	112.9	124.7	131.0	131.1	106.7	101.8	105.5	126.7
$h = 7$	109.3	119.1	121.4	128.7	108.1	97.1	109.4	106.0
$h = 8$	102.4	114.2	116.3	120.5	109.0	99.6	104.4	91.7
$h = 9$	97.7	109.5	110.4	112.7	108.3	102.3	99.0	89.4
$h = 10$	97.1	108.3	111.8	112.6	104.5	102.5	105.1	89.1
$h = 11$	97.9	108.5	109.3	112.5	103.3	99.8	107.1	98.6
$h = 12$	97.0	107.8	105.1	112.4	97.0	96.3	106.5	112.4
saisonale und gewöhnliche Differenzen, $\nabla_{12}\nabla Y_t$, iterierte Prognosen								
$h = 1$	101.5	97.8	101.4	94.6	97.7	98.0	101.5	102.2
$h = 2$	101.0	92.9	106.2	99.3	101.5	100.6	108.1	103.4
$h = 3$	101.8	92.3	102.8	93.4	99.3	101.1	108.3	109.2
$h = 4$	96.2	93.2	101.1	96.5	107.7	105.1	111.3	105.6
$h = 5$	98.5	92.2	104.9	99.0	117.5	107.7	115.5	111.9
$h = 6$	103.6	96.4	112.4	100.7	121.0	103.8	118.7	113.6
$h = 7$	107.5	99.7	117.6	100.8	126.8	108.9	122.0	117.0
$h = 8$	113.1	100.9	118.4	102.4	128.6	114.7	126.2	115.9
$h = 9$	117.3	104.2	121.2	106.9	133.1	120.4	132.7	116.1
$h = 10$	121.8	110.1	122.6	108.3	133.5	124.3	131.8	116.0
$h = 11$	123.1	110.6	125.3	109.7	139.1	129.4	132.3	113.9
$h = 12$	127.1	112.3	125.7	116.2	142.6	131.3	135.9	112.8
gewöhnliche Differenzen, ∇Y_t, direkte Prognosen								
$h = 1$	113.5	115.7	120.3	122.9	151.2	127.5	129.7	128.3
$h = 2$	92.7	97.6	124.9	119.4	101.3	129.1	104.6	124.7
$h = 3$	106.8	106.4	109.6	118.0	125.0	138.0	128.6	139.4
$h = 4$	88.4	102.4	83.6	105.7	165.3	121.2	125.2	95.6
$h = 5$	91.9	105.8	99.9	105.5	136.0	136.0	144.3	105.9
$h = 6$	98.2	104.8	124.0	101.7	182.6	119.6	129.6	125.3
$h = 7$	104.6	121.1	112.2	115.1	135.2	109.8	104.1	143.0
$h = 8$	93.6	98.7	94.4	108.3	116.7	123.8	95.7	103.4
$h = 9$	107.8	109.8	109.9	113.0	114.1	106.3	102.2	109.4
$h = 10$	103.6	112.2	110.1	122.4	132.2	125.1	122.9	111.7
$h = 11$	105.8	108.2	110.4	111.7	129.3	119.3	107.3	112.0
$h = 12$	122.9	120.9	123.2	120.0	128.7	132.6	128.7	126.2

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Tab. 32: Out-of-Sample-Prognosegüte mit dem Ansatz mit Regularisierung auf Basis von alternativen Differenzenfiltern, ALR

	Regul1	Regul2	Regul3	Regul4	Regul5	Regul6	Regul7	Regul8
gewöhnliche Differenzen, ∇Y_t, iterierte Prognosen								
$h = 1$	95.5	105.1	101.2	98.6	104.3	109.1	97.1	101.7
$h = 2$	97.4	116.3	106.3	104.2	110.4	124.4	101.8	112.1
$h = 3$	93.3	117.2	106.2	106.8	108.8	130.1	105.5	119.1
$h = 4$	94.1	113.6	108.3	105.0	111.8	135.8	106.0	120.3
$h = 5$	94.4	112.8	105.6	103.0	108.7	129.9	110.6	120.0
$h = 6$	102.4	115.2	99.2	99.0	108.7	125.3	109.6	113.6
$h = 7$	101.9	114.9	96.8	96.9	104.3	115.3	107.4	106.8
$h = 8$	99.6	111.4	95.3	92.0	103.1	103.7	106.4	100.5
$h = 9$	96.1	110.2	93.8	87.2	101.3	90.3	103.5	96.2
$h = 10$	97.0	108.6	93.1	86.0	97.3	80.4	101.9	95.3
$h = 11$	92.8	104.9	93.4	86.1	95.0	76.1	100.5	93.0
$h = 12$	88.2	106.3	92.7	87.4	94.3	77.0	101.0	89.1
saisonale und gewöhnliche Differenzen, $\nabla_{12}\nabla Y_t$, iterierte Prognosen								
$h = 1$	109.1	88.6	99.7	99.5	95.5	115.7	102.0	98.5
$h = 2$	104.3	89.0	108.2	99.3	80.9	115.2	100.0	95.7
$h = 3$	106.2	83.6	97.8	96.9	83.5	110.4	96.9	88.6
$h = 4$	103.0	97.3	95.6	98.9	79.7	114.9	97.2	92.6
$h = 5$	114.7	103.0	103.0	99.1	85.3	118.1	103.0	101.8
$h = 6$	124.4	99.8	101.7	102.3	90.5	112.3	101.9	103.0
$h = 7$	135.1	99.8	104.0	109.8	91.4	110.7	108.6	105.2
$h = 8$	137.9	106.8	108.3	113.4	93.8	108.6	114.4	107.8
$h = 9$	141.8	109.4	110.0	116.3	99.2	109.3	120.4	113.5
$h = 10$	147.4	111.3	116.1	121.0	104.1	112.1	125.2	118.4
$h = 11$	148.8	120.4	121.3	124.1	105.9	115.7	133.1	122.9
$h = 12$	151.0	126.6	127.5	128.4	107.7	120.1	137.8	122.1
gewöhnliche Differenzen, ∇Y_t, direkte Prognosen								
$h = 1$	99.8	105.2	98.8	108.0	103.7	119.5	100.8	102.0
$h = 2$	96.0	84.8	94.7	89.6	94.4	95.0	107.2	102.0
$h = 3$	104.2	105.3	104.7	110.3	115.7	110.0	100.2	104.3
$h = 4$	96.5	108.8	109.8	101.1	112.1	113.0	98.9	107.8
$h = 5$	109.4	116.7	108.1	106.8	104.4	118.9	104.2	118.2
$h = 6$	111.8	102.0	110.0	108.3	97.8	108.7	104.9	107.3
$h = 7$	98.7	112.3	110.1	101.9	123.2	109.0	98.5	111.5
$h = 8$	101.1	104.3	106.8	107.9	106.6	113.3	104.2	106.8
$h = 9$	99.8	105.1	105.3	118.6	103.7	101.5	104.4	105.1
$h = 10$	109.6	112.2	115.7	110.5	106.7	114.4	115.1	113.4
$h = 11$	115.4	106.9	118.7	117.7	111.9	114.2	114.2	117.0
$h = 12$	106.7	116.5	113.1	117.0	117.1	112.8	122.0	117.2

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Tab. 33: Out-of-Sample-Prognosegüte mit dem Bayesianischen Ansatz auf Basis von alternativen Differenzenfiltern, ALR

	Bayes1	Bayes2	Bayes3	Bayes4	Bayes5	Bayes6	Bayes7	Bayes8
gewöhnliche Differenzen, ∇Y_t, iterierte Prognosen								
$h = 1$	102.3	103.2	110.5	130.2	118.0	109.2	107.5	112.2
$h = 2$	115.7	119.9	128.1	138.9	135.3	128.7	116.0	116.3
$h = 3$	128.2	130.2	117.1	126.0	153.2	127.7	110.6	107.2
$h = 4$	133.9	126.0	98.8	107.6	146.3	120.6	104.4	97.7
$h = 5$	127.9	124.2	91.0	102.2	132.3	121.6	97.3	96.2
$h = 6$	127.5	115.0	88.0	102.9	123.3	119.1	96.6	92.1
$h = 7$	123.8	111.8	87.9	103.8	120.5	117.6	93.9	88.9
$h = 8$	120.6	102.2	85.1	97.0	118.6	110.6	88.4	86.2
$h = 9$	117.7	99.7	87.2	90.1	115.5	103.1	83.3	83.6
$h = 10$	115.8	100.5	86.3	82.0	107.1	95.3	81.0	80.6
$h = 11$	116.5	102.5	88.8	76.7	103.9	94.2	80.8	81.6
$h = 12$	118.3	108.4	86.8	75.0	101.2	93.5	79.3	84.1
saisonale und gewöhnliche Differenzen, $\nabla_{12}\nabla Y_t$, iterierte Prognosen								
$h = 1$	98.3	99.7	98.0	97.8	100.5	97.4	105.2	101.1
$h = 2$	98.6	107.4	95.2	94.6	99.0	110.5	111.7	91.9
$h = 3$	100.3	103.1	100.3	83.9	95.2	112.3	112.9	95.3
$h = 4$	105.8	103.9	100.2	84.3	95.0	110.2	115.9	102.2
$h = 5$	101.9	109.3	107.3	86.1	99.1	115.2	116.6	103.6
$h = 6$	106.6	112.8	111.7	87.0	102.5	120.9	118.0	107.1
$h = 7$	109.7	115.4	117.5	88.0	102.5	127.1	118.7	110.4
$h = 8$	112.0	125.9	117.4	95.2	104.3	132.6	120.2	113.8
$h = 9$	120.1	132.9	119.4	103.7	108.4	140.7	123.3	120.5
$h = 10$	124.5	135.6	124.6	113.8	109.6	144.4	127.5	123.4
$h = 11$	127.0	138.8	128.8	122.6	109.6	153.7	128.8	127.5
$h = 12$	130.6	144.2	130.6	127.5	112.0	159.9	130.3	130.1
gewöhnliche Differenzen, ∇Y_t, direkte Prognosen								
$h = 1$	117.2	102.5	93.5	114.3	110.3	111.8	109.8	106.9
$h = 2$	93.2	101.5	98.9	90.8	83.7	83.1	97.3	96.5
$h = 3$	114.7	103.4	99.8	123.1	109.1	107.7	120.2	114.0
$h = 4$	111.5	106.0	114.5	99.8	106.2	100.9	109.6	95.9
$h = 5$	101.6	114.4	103.2	100.9	105.3	95.6	105.3	97.4
$h = 6$	124.4	92.5	93.8	102.0	114.4	110.1	106.7	109.6
$h = 7$	107.8	108.9	108.6	91.1	112.7	89.8	98.5	86.1
$h = 8$	113.9	103.0	98.1	96.0	117.4	101.9	94.8	95.8
$h = 9$	110.1	119.0	94.9	99.8	117.2	108.2	94.9	95.8
$h = 10$	107.1	105.8	115.0	105.9	111.0	123.4	113.0	112.2
$h = 11$	116.1	116.7	121.7	122.6	107.9	132.1	110.8	123.8
$h = 12$	107.6	124.3	120.1	129.3	120.1	118.3	117.3	125.3

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Tab. 34: Out-of-Sample-Prognosegüte mit dem statistisch-parametrischen Ansatz auf Basis von alternativen Differenzenfiltern, IPI

	Param1 ∇Y_t	Param2 ∇Y_t	Param3 ∇Y_t	Param4 ∇Y_t	Param1 $\nabla_{12}\nabla Y_t$	Param2 $\nabla_{12}\nabla Y_t$	Param3 $\nabla_{12}\nabla Y_t$	Param4 $\nabla_{12}\nabla Y_t$
iterierte Prognosen								
$h = 1$	112.6	107.1	108.5	115.7	100.9	101.7	101.3	107.5
$h = 2$	111.1	105.4	101.5	113.0	103.7	103.8	104.1	110.6
$h = 3$	109.3	101.8	102.2	121.5	107.1	103.0	102.6	108.5
$h = 4$	113.2	101.3	92.8	135.6	110.5	108.6	108.2	113.2
$h = 5$	126.1	111.2	108.8	149.4	113.5	113.0	111.1	113.1
$h = 6$	128.4	126.2	108.3	127.9	116.3	114.0	106.6	115.2
$h = 7$	128.0	115.8	110.5	122.2	122.0	116.5	116.6	113.8
$h = 8$	121.4	121.7	112.9	121.3	117.1	120.1	114.2	117.1
$h = 9$	121.8	122.9	122.8	121.8	121.5	124.3	125.9	117.7
$h = 10$	123.3	127.0	121.2	115.6	128.0	126.2	128.5	123.2
$h = 11$	126.6	130.2	122.4	120.6	124.4	132.5	127.4	126.1
$h = 12$	122.2	128.4	116.9	121.4	127.3	134.0	135.0	136.3
direkte Prognosen								
$h = 1$	108.2	105.7	110.4	102.4				
$h = 2$	127.4	125.1	123.8	113.8				
$h = 3$	119.5	116.7	132.0	138.5				
$h = 4$	112.0	110.8	131.1	129.0				
$h = 5$	118.7	117.7	138.2	99.2				
$h = 6$	110.9	115.3	125.3	126.2				
$h = 7$	110.3	125.0	147.2	155.5				
$h = 8$	125.2	128.1	160.7	140.9				
$h = 9$	146.3	155.7	146.5	127.4				
$h = 10$	120.3	125.0	148.8	148.0				
$h = 11$	125.4	118.8	127.2	137.5				
$h = 12$	115.2	108.4	116.4	117.1				

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Tab. 35: Out-of-Sample-Prognosegüte mit dem klassischen Ansatz auf Basis von alternativen Differenzenfiltern, IPI

	Klass1	Klass2	Klass3	Klass4	Klass5	Klass6	Klass7	Klass8
gewöhnliche Differenzen, ∇Y_t, iterierte Prognosen								
$h = 1$	114.9	115.5	115.5	124.0	165.7	141.8	129.2	132.6
$h = 2$	110.5	111.1	110.7	120.2	184.1	132.1	119.4	121.5
$h = 3$	110.4	115.1	109.6	114.2	178.8	99.8	117.7	121.3
$h = 4$	105.1	103.4	102.4	102.5	130.5	100.3	108.5	113.9
$h = 5$	117.0	121.0	114.7	112.5	152.4	128.0	117.5	119.3
$h = 6$	117.1	110.3	113.2	105.3	122.2	127.8	127.6	113.4
$h = 7$	112.7	116.4	115.9	103.8	123.1	115.0	128.3	117.3
$h = 8$	114.2	113.8	116.7	102.3	126.7	115.8	130.7	118.3
$h = 9$	125.9	121.0	123.3	112.2	128.3	111.2	133.6	124.2
$h = 10$	126.6	122.7	126.8	111.9	125.3	118.6	140.1	128.2
$h = 11$	125.0	125.1	135.0	110.5	130.7	109.6	137.0	124.8
$h = 12$	114.1	118.6	123.1	102.4	118.5	102.2	131.6	115.4
saisonale und gewöhnliche Differenzen, $\nabla_{12}\nabla Y_t$, iterierte Prognosen								
$h = 1$	102.3	101.0	102.3	103.0	124.7	103.4	101.7	110.6
$h = 2$	99.1	101.3	101.7	106.7	114.1	103.4	106.2	105.1
$h = 3$	100.4	107.9	107.1	108.2	101.3	110.8	121.3	110.4
$h = 4$	109.4	112.1	111.8	116.5	131.0	119.2	129.4	118.0
$h = 5$	114.6	115.3	114.5	113.5	132.2	130.2	135.1	139.4
$h = 6$	111.9	116.8	115.1	119.6	126.0	130.1	147.0	131.7
$h = 7$	115.8	118.5	117.4	116.4	131.7	124.3	145.1	130.9
$h = 8$	121.4	116.9	124.8	120.1	136.8	123.7	146.9	136.5
$h = 9$	124.3	119.2	130.2	120.4	134.7	141.6	142.7	140.1
$h = 10$	127.4	127.3	134.7	128.5	158.8	150.8	143.3	141.4
$h = 11$	139.5	130.6	142.5	130.1	151.5	152.2	146.1	144.2
$h = 12$	144.2	132.3	143.8	134.5	163.4	158.8	156.6	143.3
gewöhnliche Differenzen, ∇Y_t, direkte Prognosen								
$h = 1$	121.6	122.8	116.6	117.4	161.0	150.8	121.6	124.7
$h = 2$	124.4	129.3	140.9	126.6	193.2	192.7	170.6	153.8
$h = 3$	119.9	124.0	132.2	130.7	231.6	216.0	192.6	141.1
$h = 4$	114.3	113.7	114.8	111.3	191.5	179.4	149.8	151.3
$h = 5$	110.2	110.7	114.1	116.0	220.8	204.4	184.4	125.3
$h = 6$	105.7	105.6	116.7	109.1	162.3	156.4	146.4	129.2
$h = 7$	103.9	100.5	102.3	126.1	155.0	175.0	120.9	104.7
$h = 8$	101.6	104.1	98.7	94.6	150.8	159.1	151.2	101.7
$h = 9$	114.1	114.9	113.2	113.9	190.4	157.0	143.2	126.3
$h = 10$	115.2	109.5	119.6	120.1	174.5	143.2	142.0	119.5
$h = 11$	110.6	112.8	110.4	109.8	110.6	114.4	110.2	108.3
$h = 12$	99.9	102.0	102.9	103.0	110.3	104.8	102.9	105.2

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Tab. 36: Out-of-Sample-Prognosegüte mit dem Ansatz mit Regularisierung auf Basis von alternativen Differenzenfiltern, IPI

	Regul1	Regul2	Regul3	Regul4	Regul5	Regul6	Regul7	Regul8
gewöhnliche Differenzen, ∇Y_t, iterierte Prognosen								
$h = 1$	113.8	109.8	113.0	118.4	117.6	116.3	112.2	129.2
$h = 2$	109.1	108.5	115.4	112.4	109.3	112.4	117.8	132.1
$h = 3$	104.6	107.4	109.8	128.5	107.3	117.5	116.1	134.2
$h = 4$	92.2	105.8	111.0	111.2	114.4	120.1	110.5	120.2
$h = 5$	104.3	110.6	118.4	117.6	118.7	129.5	127.5	128.4
$h = 6$	110.0	112.6	117.8	133.2	123.0	124.3	130.8	126.5
$h = 7$	107.5	113.2	118.4	131.6	131.1	117.9	125.1	114.6
$h = 8$	108.7	111.7	120.9	132.5	122.3	117.0	130.3	109.1
$h = 9$	111.4	112.8	127.4	135.0	129.1	120.3	131.9	121.1
$h = 10$	111.4	118.1	129.8	143.7	129.6	123.2	129.1	119.6
$h = 11$	109.6	120.0	131.2	142.8	132.7	123.3	136.8	117.7
$h = 12$	105.8	112.9	116.8	134.4	122.8	111.8	117.7	112.5
saisonale und gewöhnliche Differenzen, $\nabla_{12}\nabla Y_t$, iterierte Prognosen								
$h = 1$	100.9	93.9	94.3	94.0	102.1	100.3	98.0	97.9
$h = 2$	97.9	93.7	98.5	97.0	104.9	106.2	105.6	99.1
$h = 3$	103.5	103.7	106.6	104.5	112.5	107.2	115.3	109.6
$h = 4$	114.3	106.4	103.7	109.9	111.5	110.7	109.4	112.1
$h = 5$	115.6	108.7	107.4	111.2	114.4	121.7	102.7	111.8
$h = 6$	120.5	109.9	103.6	117.8	118.7	132.8	107.2	117.6
$h = 7$	119.2	112.8	102.3	120.2	121.5	138.6	118.8	117.4
$h = 8$	120.9	118.2	116.1	129.5	121.9	141.6	123.7	123.7
$h = 9$	131.0	120.4	118.8	133.6	125.9	143.7	123.8	132.2
$h = 10$	138.9	125.1	121.8	143.1	129.6	145.7	131.3	135.4
$h = 11$	144.2	125.4	125.9	146.4	133.4	149.1	130.1	138.4
$h = 12$	151.2	128.5	126.2	145.8	132.8	157.5	132.9	136.5
gewöhnliche Differenzen, ∇Y_t, direkte Prognosen								
$h = 1$	109.6	108.5	117.6	111.6	103.8	109.8	110.8	107.2
$h = 2$	140.9	116.5	121.3	123.0	126.3	125.2	123.8	127.3
$h = 3$	128.5	131.7	121.4	125.3	127.0	125.4	122.0	122.6
$h = 4$	121.5	111.5	127.5	132.9	113.6	130.1	132.0	138.3
$h = 5$	108.0	116.5	110.8	105.2	109.2	103.5	100.7	109.4
$h = 6$	115.6	123.1	119.4	124.2	120.3	120.3	120.9	123.2
$h = 7$	108.1	108.7	112.8	117.8	101.3	119.2	108.6	106.1
$h = 8$	116.4	126.8	133.1	132.3	122.6	118.8	132.0	129.4
$h = 9$	140.3	126.8	129.7	119.4	135.8	160.9	130.0	123.8
$h = 10$	121.3	138.5	130.3	122.3	118.2	136.2	133.9	118.8
$h = 11$	123.7	123.2	124.2	122.2	130.2	126.9	128.8	125.0
$h = 12$	103.4	108.2	109.9	107.5	109.8	104.6	105.3	110.5

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Tab. 37: Out-of-Sample-Prognosegüte mit dem Bayesianischen Ansatz auf Basis von alternativen Differenzenfiltern, IPI

	Bayes1	Bayes2	Bayes3	Bayes4	Bayes5	Bayes6	Bayes7	Bayes8
gewöhnliche Differenzen, ∇Y_t, iterierte Prognosen								
$h = 1$	109.5	109.8	108.7	120.1	110.1	100.7	114.2	109.5
$h = 2$	110.1	102.0	112.7	118.9	106.5	104.4	130.9	112.6
$h = 3$	107.5	92.8	115.9	117.8	105.1	106.3	128.3	113.6
$h = 4$	108.1	93.1	108.9	115.9	99.6	103.7	130.9	120.4
$h = 5$	115.8	105.5	126.2	126.5	100.0	114.3	140.6	127.4
$h = 6$	117.4	112.7	125.2	124.9	108.1	114.4	136.8	116.7
$h = 7$	116.0	111.0	122.2	117.7	108.2	114.7	131.4	107.9
$h = 8$	114.0	108.4	126.2	121.7	111.4	110.9	130.0	112.9
$h = 9$	117.6	115.4	128.3	132.0	116.1	120.1	141.2	114.4
$h = 10$	120.4	117.4	126.7	136.8	117.2	124.2	136.4	111.1
$h = 11$	116.9	115.0	127.4	134.8	119.5	118.9	131.2	118.0
$h = 12$	107.7	108.4	118.9	127.2	113.0	113.8	122.4	107.2
saisonale und gewöhnliche Differenzen, $\nabla_{12}\nabla Y_t$, iterierte Prognosen								
$h = 1$	96.7	97.2	110.4	156.7	100.1	100.9	112.3	138.7
$h = 2$	98.5	99.0	112.1	164.1	101.6	100.0	131.5	153.2
$h = 3$	100.3	99.2	120.2	146.8	105.3	107.2	135.9	159.6
$h = 4$	104.4	100.8	127.3	161.5	109.2	110.0	137.2	178.3
$h = 5$	101.7	99.8	135.9	169.4	114.0	112.2	138.1	207.6
$h = 6$	105.5	102.4	137.5	176.7	118.5	116.3	141.5	201.1
$h = 7$	107.7	102.3	135.1	170.3	120.1	117.2	143.4	194.7
$h = 8$	110.8	107.7	143.2	177.1	123.4	120.6	153.0	187.4
$h = 9$	115.6	109.7	147.2	179.8	124.1	125.6	156.7	172.9
$h = 10$	119.1	115.1	156.1	187.9	134.0	125.6	157.2	189.2
$h = 11$	120.2	118.0	158.3	204.5	136.1	130.7	151.1	177.8
$h = 12$	125.2	124.3	167.0	204.1	135.6	138.0	157.3	186.6
gewöhnliche Differenzen, ∇Y_t, direkte Prognosen								
$h = 1$	107.8	106.8	110.6	120.3	106.7	110.4	110.4	108.7
$h = 2$	122.4	129.3	127.2	139.4	120.7	134.3	136.6	134.8
$h = 3$	127.6	145.3	144.3	139.5	111.4	125.0	144.0	126.2
$h = 4$	114.8	125.7	154.4	148.9	114.2	142.1	156.1	152.5
$h = 5$	106.2	135.6	140.9	147.1	110.8	124.2	136.1	145.9
$h = 6$	130.3	120.0	135.1	151.3	100.1	117.5	145.4	144.8
$h = 7$	106.1	113.2	144.0	153.4	107.1	128.7	155.8	149.4
$h = 8$	137.1	145.9	144.3	149.4	130.0	149.9	138.8	152.5
$h = 9$	113.1	154.8	154.9	149.9	145.1	149.8	179.8	140.6
$h = 10$	143.2	147.2	153.7	151.9	118.4	120.8	138.8	135.2
$h = 11$	129.0	126.0	127.5	129.9	126.8	129.8	124.3	128.1
$h = 12$	112.5	112.8	111.9	105.6	109.0	116.7	110.7	105.4

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Tab. 38: Out-of-Sample-Prognosegüte mit dem statistisch-parametrischen Ansatz bei Verwendung der naiv-iterierten Prognosemethode, im Vergleich zur simulativ iterierten Prognosemethode, ALR

	$\nabla_{12}Y_t$				∇Y_t			
	Param1	Param2	Param3	Param4	Param1	Param2	Param3	Param4
$h = 1$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$h = 2$	96.2	97.2	103.9	98.1	101.0	97.9	96.5	100.4
$h = 3$	100.9	96.3	96.5	99.6	106.0	94.1	96.4	94.5
$h = 4$	100.9	94.9	96.1	101.1	105.7	93.4	93.6	93.9
$h = 5$	101.7	93.1	94.5	102.3	101.6	97.4	93.8	91.8
$h = 6$	104.5	96.0	97.1	107.1	99.6	96.9	92.9	92.3
$h = 7$	105.0	95.2	95.7	108.7	98.0	97.2	96.2	93.3
$h = 8$	103.8	94.7	94.5	115.4	98.3	97.6	97.3	93.8
$h = 9$	102.7	96.8	93.2	115.7	96.5	98.0	97.6	94.6
$h = 10$	103.4	99.0	90.5	113.6	95.2	95.2	96.6	96.6
$h = 11$	100.4	100.8	89.2	115.8	94.7	97.3	96.7	96.4
$h = 12$	100.1	99.6	90.2	119.5	97.6	96.6	98.2	97.3

	$\nabla_{12}\nabla Y_t$			
	Param1	Param2	Param3	Param4
$h = 1$	100.0	100.0	100.0	100.0
$h = 2$	96.6	97.0	100.6	94.5
$h = 3$	100.2	97.1	100.1	97.1
$h = 4$	101.6	90.8	103.5	102.6
$h = 5$	99.8	89.5	103.1	103.0
$h = 6$	102.4	89.0	97.3	103.7
$h = 7$	104.0	90.7	101.6	103.6
$h = 8$	103.1	88.0	100.5	103.1
$h = 9$	103.6	85.7	101.2	102.9
$h = 10$	101.6	87.4	99.0	104.2
$h = 11$	101.7	86.9	98.2	105.6
$h = 12$	100.9	88.3	99.8	104.8

MSFE auf Basis der naiv-iterierten Methode in Prozent des MSFE der simulativ-iterierten Methode

Tab. 39: Out-of-Sample-Prognosegüte mit dem klassischen Ansatz bei Verwendung der naiv-iterierten Prognosemethode, im Vergleich zur simulativ iterierten Prognosemethode, ALR

	Klass1	Klass2	Klass3	Klass4	Klass5	Klass6	Klass7	Klass8
saisonale Differenzen, $\nabla_{12} Y_t$								
$h = 1$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$h = 2$	97.9	96.9	101.5	94.1	105.1	95.7	97.7	96.1
$h = 3$	95.5	96.5	96.6	98.7	102.2	97.3	96.6	94.6
$h = 4$	97.6	96.4	96.9	99.6	100.8	96.2	100.2	97.3
$h = 5$	100.2	95.5	96.6	100.8	102.0	98.7	100.5	97.4
$h = 6$	103.4	97.0	96.9	101.5	101.4	102.9	105.1	95.2
$h = 7$	101.8	99.3	98.2	99.6	101.8	105.4	109.1	95.3
$h = 8$	97.8	100.7	101.9	101.6	100.3	106.0	108.8	95.5
$h = 9$	97.9	103.6	103.5	105.8	96.4	108.5	107.8	98.9
$h = 10$	98.6	104.6	102.0	109.2	95.3	108.8	105.6	99.2
$h = 11$	98.3	105.9	101.4	111.2	96.5	109.4	106.1	100.4
$h = 12$	100.3	105.3	99.8	112.7	94.8	107.5	103.7	99.3
gewöhnliche Differenzen, ∇Y_t								
$h = 1$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$h = 2$	101.6	97.4	96.5	103.1	97.6	95.1	99.8	92.4
$h = 3$	102.8	94.2	103.1	102.0	100.9	94.6	99.4	97.4
$h = 4$	107.4	92.5	102.7	101.3	102.7	91.1	99.9	98.2
$h = 5$	103.4	92.6	100.5	100.9	105.2	92.8	99.6	97.0
$h = 6$	108.4	92.6	102.9	98.9	101.5	90.3	99.6	95.4
$h = 7$	107.5	92.3	103.2	98.0	101.6	92.6	97.3	94.0
$h = 8$	108.7	93.3	100.0	101.0	99.6	90.5	96.9	97.8
$h = 9$	109.0	95.0	99.4	103.5	96.9	87.9	97.5	96.8
$h = 10$	108.4	96.5	98.5	101.0	97.1	90.9	96.6	99.4
$h = 11$	105.7	97.2	99.7	101.4	96.3	91.4	96.6	100.4
$h = 12$	105.6	100.7	102.5	103.2	95.4	92.5	97.6	101.7
saisonale und gewöhnliche Differenzen, $\nabla_{12} \nabla Y_t$								
$h = 1$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$h = 2$	101.9	105.6	96.5	95.7	96.3	101.4	91.7	100.0
$h = 3$	100.2	107.4	95.7	99.4	97.3	105.7	87.0	92.4
$h = 4$	104.8	106.9	97.6	94.8	93.4	104.5	85.7	94.7
$h = 5$	105.6	111.5	98.4	95.6	90.6	101.8	86.5	93.6
$h = 6$	103.5	109.1	96.9	94.0	93.7	105.0	87.6	94.5
$h = 7$	103.4	107.7	97.4	95.8	94.3	102.4	90.1	94.2
$h = 8$	101.6	109.8	99.7	97.6	95.0	102.0	92.0	96.7
$h = 9$	101.8	110.6	100.8	97.5	93.2	101.2	92.1	99.2
$h = 10$	101.8	110.0	103.6	99.3	95.4	100.2	96.9	100.8
$h = 11$	102.7	112.9	103.9	101.4	95.8	99.0	98.2	103.0
$h = 12$	100.1	113.4	105.7	99.8	97.5	100.9	98.0	105.0

MSFE auf Basis der naiv-iterierten Methode in Prozent des MSFE der simulativ-iterierten Methode

Tab. 40: Out-of-Sample-Prognosegüte mit dem Ansatz mit Regularisierung bei Verwendung der naiv-iterierten Prognosemethode, im Vergleich zur simulativ iterierten Prognosemethode, ALR

	Regul1	Regul2	Regul3	Regul4	Regul5	Regul6	Regul7	Regul8
saisonale Differenzen, $\nabla_{12} Y_t$								
$h = 1$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$h = 2$	99.1	96.1	97.7	97.2	101.0	99.7	101.9	103.2
$h = 3$	99.2	98.3	96.4	96.0	101.6	101.7	100.2	98.6
$h = 4$	97.6	98.7	97.1	104.0	100.5	104.2	101.4	105.0
$h = 5$	100.4	95.7	97.6	101.3	106.1	114.4	106.4	108.3
$h = 6$	103.6	93.2	101.1	105.2	105.7	119.4	111.3	107.5
$h = 7$	102.8	96.3	102.0	110.9	108.5	126.3	120.7	113.4
$h = 8$	101.2	96.8	106.9	112.5	107.8	131.7	126.3	119.7
$h = 9$	100.8	99.9	107.6	114.2	110.3	137.2	128.6	127.1
$h = 10$	101.0	104.5	109.5	116.8	112.0	140.3	131.0	127.0
$h = 11$	100.9	106.6	110.8	118.6	112.0	142.5	131.8	132.5
$h = 12$	100.3	111.5	112.3	121.3	112.7	141.4	135.7	132.6
gewöhnliche Differenzen, ∇Y_t								
$h = 1$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$h = 2$	98.5	97.2	96.9	98.4	96.2	99.8	97.1	99.3
$h = 3$	94.3	98.9	97.8	93.3	93.0	100.9	94.4	98.2
$h = 4$	93.7	101.3	100.0	92.4	91.9	100.1	94.9	95.6
$h = 5$	96.8	102.0	101.4	90.4	93.6	101.0	92.7	94.2
$h = 6$	94.7	100.2	103.6	90.1	93.8	98.5	93.1	94.2
$h = 7$	98.0	100.8	101.2	89.8	97.9	94.7	96.1	96.7
$h = 8$	99.0	101.8	99.6	90.8	97.1	93.7	96.8	96.9
$h = 9$	100.1	99.9	98.3	92.2	96.2	94.5	97.7	96.8
$h = 10$	98.2	100.4	98.7	93.5	97.0	95.6	99.6	95.0
$h = 11$	99.7	101.5	98.6	96.0	98.9	95.2	101.5	95.9
$h = 12$	100.5	98.4	99.1	96.9	99.4	96.1	101.8	98.6
saisonale und gewöhnliche Differenzen, $\nabla_{12} \nabla Y_t$								
$h = 1$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$h = 2$	102.3	97.9	90.8	100.1	107.9	99.7	99.6	101.8
$h = 3$	98.9	98.8	98.3	102.2	99.1	103.3	100.7	108.7
$h = 4$	101.6	97.5	101.9	101.7	101.7	99.6	101.7	105.8
$h = 5$	102.9	95.5	98.0	105.0	97.8	97.1	99.3	100.0
$h = 6$	103.7	97.0	101.5	104.1	94.0	100.3	101.7	100.4
$h = 7$	103.2	97.1	101.1	100.2	94.7	101.3	98.7	101.0
$h = 8$	104.0	95.3	99.7	100.3	95.7	104.2	98.6	102.0
$h = 9$	103.0	95.1	101.7	101.7	92.8	105.3	99.1	101.0
$h = 10$	102.7	94.5	100.4	101.9	92.1	104.2	100.9	101.3
$h = 11$	104.0	94.8	99.5	103.0	93.9	103.0	99.9	101.0
$h = 12$	104.4	95.9	97.3	102.5	96.3	101.8	100.0	104.8

MSFE auf Basis der naiv-iterierten Methode in Prozent des MSFE der simulativ-iterierten Methode

Tab. 41: Out-of-Sample-Prognosegüte mit dem Bayesianischen Ansatz bei Verwendung der naiv-iterierten Prognosemethode, im Vergleich zur simulativ iterierten Prognosemethode, ALR

	Bayes1	Bayes2	Bayes3	Bayes4	Bayes5	Bayes6	Bayes7	Bayes8
saisonale Differenzen, $\nabla_{12} Y_t$								
$h = 1$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$h = 2$	99.4	97.7	98.8	99.2	103.0	101.3	99.8	103.2
$h = 3$	107.8	97.5	100.1	101.8	98.9	108.0	98.5	101.6
$h = 4$	104.9	96.4	102.3	101.4	96.1	107.6	98.2	100.7
$h = 5$	102.4	97.7	104.4	101.9	95.0	111.8	101.9	102.4
$h = 6$	107.3	104.2	106.0	100.5	91.4	114.8	104.2	102.9
$h = 7$	110.6	108.2	107.5	101.0	92.2	118.9	106.0	102.7
$h = 8$	112.5	113.9	109.5	102.5	95.9	121.9	107.0	101.8
$h = 9$	112.0	120.8	114.7	106.4	96.9	123.3	110.0	101.5
$h = 10$	109.8	126.1	122.3	111.3	98.2	125.3	109.6	100.9
$h = 11$	105.8	133.6	125.8	113.7	98.1	124.4	113.3	100.1
$h = 12$	104.3	133.6	126.0	107.0	100.7	128.1	112.1	101.3
gewöhnliche Differenzen, ∇Y_t								
$h = 1$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$h = 2$	97.5	100.5	94.6	94.3	100.5	93.0	96.3	99.0
$h = 3$	95.2	97.9	99.6	95.2	94.2	95.8	96.2	101.4
$h = 4$	94.9	96.2	97.8	95.1	97.4	95.6	92.8	100.1
$h = 5$	98.4	95.8	98.6	98.2	100.0	95.0	96.0	98.4
$h = 6$	98.2	97.0	97.9	94.7	99.2	96.8	95.3	99.1
$h = 7$	99.9	95.3	98.9	92.0	98.5	95.5	96.7	98.2
$h = 8$	100.9	97.7	102.6	92.2	98.2	95.4	98.5	96.5
$h = 9$	100.8	97.0	99.3	95.8	97.3	96.4	99.7	97.0
$h = 10$	101.9	96.4	100.3	97.3	99.0	98.9	99.4	97.1
$h = 11$	101.1	95.3	96.1	99.2	98.4	99.3	99.1	96.7
$h = 12$	101.1	93.4	97.4	100.0	99.6	99.9	100.2	96.3
saisonale und gewöhnliche Differenzen, $\nabla_{12} \nabla Y_t$								
$h = 1$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$h = 2$	99.6	94.2	102.3	98.0	101.0	93.5	99.1	101.8
$h = 3$	94.5	99.4	102.1	101.1	102.0	92.4	101.2	99.0
$h = 4$	89.9	102.1	102.9	99.5	101.3	95.3	100.3	99.5
$h = 5$	96.2	102.6	102.4	98.0	99.3	97.2	102.7	99.4
$h = 6$	94.1	100.5	102.0	98.2	97.3	96.1	101.5	96.8
$h = 7$	94.5	101.1	100.8	99.1	97.7	96.3	101.4	97.4
$h = 8$	95.9	98.9	101.4	98.3	97.6	96.4	101.7	97.7
$h = 9$	94.0	98.7	101.8	99.0	96.7	95.9	101.8	97.4
$h = 10$	94.5	98.6	101.6	99.0	98.8	97.9	100.5	99.5
$h = 11$	96.5	97.7	100.6	99.3	101.0	95.5	101.5	99.2
$h = 12$	97.1	96.4	101.7	99.7	100.3	95.2	101.9	100.1

MSFE auf Basis der naiv-iterierten Methode in Prozent des MSFE der simulativ-iterierten Methode

Tab. 42: Out-of-Sample-Prognosegüte mit dem statistisch-parametrischen Ansatz bei Verwendung der naiv-iterierten Prognosemethode, im Vergleich zur simulativ iterierten Prognosemethode, IPI

	$\nabla_{12}Y_t$				∇Y_t			
	Param1	Param2	Param3	Param4	Param1	Param2	Param3	Param4
$h = 1$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$h = 2$	100.8	98.5	97.6	100.2	100.0	96.8	98.1	101.6
$h = 3$	100.1	97.0	99.1	98.4	98.5	99.7	99.0	101.4
$h = 4$	97.7	97.4	98.5	100.3	100.2	101.8	99.9	97.0
$h = 5$	97.0	99.5	102.5	96.1	97.4	97.7	98.5	96.2
$h = 6$	94.5	97.6	104.9	95.4	93.9	96.3	99.6	97.8
$h = 7$	100.6	96.7	101.3	98.2	96.0	99.6	99.8	98.9
$h = 8$	100.6	95.6	102.6	96.1	100.0	95.5	98.9	97.5
$h = 9$	98.2	100.5	102.6	98.4	100.9	98.3	95.6	95.4
$h = 10$	99.9	97.0	104.2	97.3	98.6	98.5	96.7	98.8
$h = 11$	97.4	97.9	103.9	95.8	98.4	99.2	96.7	100.7
$h = 12$	95.3	98.7	104.1	99.3	97.6	96.2	96.5	98.1
	$\nabla_{12}\nabla Y_t$							
	Param1	Param2	Param3	Param4				
$h = 1$	100.0	100.0	100.0	100.0				
$h = 2$	97.8	98.0	96.1	96.1				
$h = 3$	97.7	101.5	101.0	97.2				
$h = 4$	98.8	101.4	97.8	101.2				
$h = 5$	98.5	101.2	97.3	102.1				
$h = 6$	98.6	101.7	103.0	100.5				
$h = 7$	96.6	102.5	98.1	100.5				
$h = 8$	102.6	101.4	102.7	99.6				
$h = 9$	101.4	101.0	96.2	101.6				
$h = 10$	100.8	104.0	95.9	101.5				
$h = 11$	105.6	101.5	99.4	103.9				
$h = 12$	105.1	102.2	95.5	101.6				

MSFE auf Basis der naiv-iterierten Methode in Prozent des MSFE der simulativ-iterierten Methode

Tab. 43: Out-of-Sample-Prognosegüte mit dem klassischen Ansatz bei Verwendung der naiv-iterierten Prognosemethode, im Vergleich zur simulativ iterierten Prognosemethode, IPI

	Klass1	Klass2	Klass3	Klass4	Klass5	Klass6	Klass7	Klass8
saisonale Differenzen, $\nabla_{12} Y_t$								
$h = 1$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$h = 2$	99.4	101.7	102.2	101.6	98.4	96.0	98.7	100.0
$h = 3$	97.3	100.7	102.0	98.3	99.8	100.3	101.2	93.5
$h = 4$	94.9	99.3	98.4	101.9	102.1	98.0	101.3	95.9
$h = 5$	97.1	105.0	103.2	97.5	101.2	96.1	100.6	93.9
$h = 6$	96.7	100.3	101.9	100.1	101.5	97.0	100.6	98.0
$h = 7$	97.3	103.5	103.7	98.5	98.4	97.7	100.6	94.8
$h = 8$	100.3	103.1	106.5	98.1	100.9	98.0	101.3	98.7
$h = 9$	99.2	104.6	105.3	101.1	99.3	99.5	99.5	100.2
$h = 10$	99.3	102.4	102.2	103.3	97.3	100.9	103.9	95.5
$h = 11$	101.1	101.0	103.9	103.6	94.4	98.2	101.4	100.7
$h = 12$	102.9	103.1	99.1	100.5	98.6	98.8	100.0	99.0
gewöhnliche Differenzen, ∇Y_t								
$h = 1$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$h = 2$	97.6	99.8	98.9	97.9	99.3	96.7	101.6	100.4
$h = 3$	99.2	95.8	98.9	99.7	98.5	102.7	99.8	100.0
$h = 4$	98.0	98.4	96.9	100.4	101.4	97.0	102.1	97.3
$h = 5$	96.8	93.5	96.9	99.4	104.4	97.0	99.4	98.6
$h = 6$	93.7	99.7	97.2	100.8	101.0	102.3	100.1	97.5
$h = 7$	95.9	93.7	95.9	101.0	96.3	103.2	96.8	93.4
$h = 8$	95.5	94.5	93.6	104.2	100.4	99.0	97.2	93.3
$h = 9$	93.5	96.2	95.6	97.5	100.0	100.6	95.0	96.6
$h = 10$	93.9	96.6	94.9	99.6	101.6	98.5	94.6	92.8
$h = 11$	94.8	95.6	90.4	97.7	94.7	98.3	95.3	94.3
$h = 12$	97.1	94.3	94.2	98.8	97.7	92.5	94.2	93.6
saisonale und gewöhnliche Differenzen, $\nabla_{12} \nabla Y_t$								
$h = 1$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$h = 2$	101.2	98.6	98.6	97.8	99.9	97.2	96.1	96.8
$h = 3$	104.2	99.4	98.1	102.0	102.8	98.9	97.5	96.9
$h = 4$	101.7	99.3	99.3	101.6	98.5	101.2	98.6	92.2
$h = 5$	99.2	100.3	98.3	105.8	103.6	103.5	98.0	91.1
$h = 6$	100.6	96.9	100.6	102.6	102.8	100.1	93.9	90.3
$h = 7$	100.1	99.3	101.7	104.0	102.7	102.5	94.9	93.7
$h = 8$	99.5	103.1	101.0	102.8	103.5	102.0	95.5	92.6
$h = 9$	99.4	101.9	98.2	105.0	106.7	105.2	98.3	92.5
$h = 10$	103.9	103.5	101.5	104.4	104.6	104.9	96.3	95.0
$h = 11$	96.5	104.6	99.9	104.9	112.2	107.6	95.8	90.5
$h = 12$	93.5	106.0	98.1	103.8	107.7	106.2	94.7	90.9

MSFE auf Basis der naiv-iterierten Methode in Prozent des MSFE der simulativ-iterierten Methode

Tab. 44: Out-of-Sample-Prognosegüte mit dem Ansatz mit Regularisierung bei Verwendung der naiv-iterierten Prognosemethode, im Vergleich zur simulativ iterierten Prognosemethode, IPI

	Regul1	Regul2	Regul3	Regul4	Regul5	Regul6	Regul7	Regul8
saisonale Differenzen, $\nabla_{12} Y_t$								
$h = 1$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$h = 2$	94.7	99.5	98.5	99.5	98.5	97.6	99.1	99.0
$h = 3$	99.0	104.8	100.0	98.3	97.9	100.2	102.1	101.8
$h = 4$	96.5	102.3	100.7	101.1	100.4	97.5	96.1	100.3
$h = 5$	101.1	99.4	101.8	101.7	97.0	100.6	97.9	102.7
$h = 6$	101.1	104.1	104.7	106.2	99.0	102.4	99.3	109.5
$h = 7$	102.9	100.6	99.9	103.6	99.7	102.7	100.1	108.3
$h = 8$	103.4	103.4	104.0	107.4	98.7	105.2	99.6	107.6
$h = 9$	102.4	108.0	102.6	103.0	100.5	104.9	106.0	113.1
$h = 10$	103.4	108.4	104.2	107.8	102.3	108.8	102.9	111.6
$h = 11$	100.5	107.1	104.1	109.6	100.5	107.4	103.3	111.3
$h = 12$	101.3	106.3	103.9	108.1	104.0	112.0	103.9	115.2
gewöhnliche Differenzen, ∇Y_t								
$h = 1$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$h = 2$	100.0	100.0	97.6	97.5	99.0	101.7	99.7	100.2
$h = 3$	100.0	99.9	98.4	96.0	99.0	97.2	103.8	99.1
$h = 4$	99.8	98.8	100.5	102.7	97.7	98.8	105.3	101.1
$h = 5$	99.1	101.8	100.0	100.5	101.4	100.6	101.7	102.4
$h = 6$	97.5	100.4	101.9	97.5	99.0	99.3	100.8	104.1
$h = 7$	102.6	100.7	101.1	100.6	96.2	102.7	105.1	102.1
$h = 8$	102.4	101.6	100.6	100.4	102.1	101.4	101.1	103.1
$h = 9$	102.3	101.1	99.9	101.4	98.3	101.2	102.1	96.8
$h = 10$	102.8	98.5	99.5	100.1	100.2	102.6	106.3	101.5
$h = 11$	102.2	96.9	100.1	101.3	97.8	100.0	100.3	99.3
$h = 12$	102.2	101.1	101.6	103.0	98.5	99.0	101.0	99.7
saisonale und gewöhnliche Differenzen, $\nabla_{12} \nabla Y_t$								
$h = 1$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$h = 2$	97.8	100.0	97.0	98.4	98.0	95.5	97.1	100.4
$h = 3$	100.9	96.2	100.7	99.6	94.7	98.7	99.7	98.2
$h = 4$	99.6	97.6	101.9	98.3	99.9	100.0	100.1	99.6
$h = 5$	97.9	96.7	97.9	98.5	99.4	98.6	104.5	102.0
$h = 6$	99.7	97.9	103.9	95.5	99.5	102.1	106.2	101.0
$h = 7$	101.6	97.4	104.7	98.0	99.5	103.0	104.9	104.1
$h = 8$	104.5	96.2	100.3	97.0	101.5	104.9	104.9	102.4
$h = 9$	101.1	99.4	100.6	97.2	100.9	106.1	107.4	100.9
$h = 10$	101.3	101.2	103.4	94.8	103.4	109.3	106.2	101.4
$h = 11$	100.8	103.8	102.4	96.4	103.7	112.8	108.6	102.2
$h = 12$	99.7	103.1	105.6	98.4	106.1	111.2	108.5	104.8

MSFE auf Basis der naiv-iterierten Methode in Prozent des MSFE der simulativ-iterierten Methode

Tab. 45: Out-of-Sample-Prognosegüte mit dem Bayesianischen Ansatz bei Verwendung der naiv-iterierten Prognosemethode, im Vergleich zur simulativ iterierten Prognosemethode, IPI

	Bayes1	Bayes2	Bayes3	Bayes4	Bayes5	Bayes6	Bayes7	Bayes8
saisonale Differenzen, $\nabla_{12} Y_t$								
$h = 1$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$h = 2$	98.9	95.8	100.5	99.0	99.0	98.8	99.2	98.8
$h = 3$	98.4	97.9	101.1	99.4	98.7	97.2	100.2	100.3
$h = 4$	98.4	100.9	102.2	97.5	99.4	97.7	100.2	97.6
$h = 5$	97.0	95.0	100.7	101.0	98.4	95.1	99.2	100.3
$h = 6$	101.6	98.8	103.6	101.2	101.1	94.9	98.0	99.9
$h = 7$	102.0	96.3	105.2	102.7	99.3	99.1	101.8	102.5
$h = 8$	98.7	97.8	101.4	100.4	100.9	96.8	100.9	101.7
$h = 9$	100.1	99.8	103.8	98.7	98.0	97.7	101.1	103.1
$h = 10$	102.7	95.4	103.6	101.2	98.9	97.4	102.9	104.8
$h = 11$	101.2	99.3	104.1	101.2	99.0	97.8	105.3	103.1
$h = 12$	101.5	96.9	104.0	98.9	98.4	94.3	106.1	105.0
gewöhnliche Differenzen, ∇Y_t								
$h = 1$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$h = 2$	97.9	99.9	96.6	99.1	99.9	96.1	96.0	98.7
$h = 3$	99.5	99.4	98.9	97.0	100.6	95.9	101.0	102.6
$h = 4$	92.5	97.2	102.3	98.7	98.4	95.2	99.7	97.7
$h = 5$	95.2	99.1	100.6	98.5	105.8	94.9	101.8	97.5
$h = 6$	93.6	97.3	101.8	101.0	103.4	97.4	99.5	97.2
$h = 7$	94.0	97.8	102.3	101.1	103.7	95.3	100.1	97.7
$h = 8$	97.3	99.8	99.9	100.8	101.5	98.8	98.6	96.9
$h = 9$	95.7	99.8	102.8	99.3	101.2	94.6	96.5	94.3
$h = 10$	96.0	101.3	101.3	96.6	102.1	95.1	96.1	98.1
$h = 11$	97.6	100.7	100.7	98.0	100.0	97.3	94.3	93.4
$h = 12$	97.7	100.0	100.1	99.9	103.3	98.1	94.8	96.1
saisonale und gewöhnliche Differenzen, $\nabla_{12} \nabla Y_t$								
$h = 1$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$h = 2$	100.0	99.1	98.9	98.5	98.7	102.6	98.1	97.8
$h = 3$	97.9	102.2	95.9	99.7	98.6	98.3	98.9	99.3
$h = 4$	97.2	100.7	96.4	99.2	99.5	99.3	99.6	101.2
$h = 5$	100.1	101.3	96.5	101.1	97.7	99.9	100.9	100.9
$h = 6$	98.9	101.3	101.5	102.3	96.4	98.4	104.1	100.2
$h = 7$	99.4	103.6	104.1	101.7	97.6	101.2	102.7	100.5
$h = 8$	100.3	101.8	101.1	102.4	96.7	100.6	102.6	100.3
$h = 9$	98.9	105.1	100.8	103.4	98.5	100.1	103.8	101.0
$h = 10$	100.8	103.2	101.8	104.4	95.9	105.2	109.2	101.6
$h = 11$	102.4	103.8	104.7	103.2	96.3	101.8	110.4	105.6
$h = 12$	102.3	104.7	103.0	104.7	98.2	97.3	110.8	103.9

MSFE auf Basis der naiv-iterierten Methode in Prozent des MSFE der simulativ-iterierten Methode

Tab. 46: Out-of-Sample-Prognosegüte mit dem klassischen Ansatz unter Weglassung von Direktverbindungen, ALR und IPI

Arbeitslosenrate								
	iterierte Prognosen				direkte Prognosen			
	Klass5	Klass6	Klass7	Klass8	Klass5	Klass6	Klass7	Klass8
$h = 1$	105.8	114.2	103.4	106.0	106.2	101.5	122.2	111.4
$h = 2$	106.3	128.1	103.5	116.5	112.0	118.2	115.8	133.8
$h = 3$	108.1	129.2	100.8	122.2	109.4	111.6	129.1	124.6
$h = 4$	108.2	134.2	101.7	118.0	97.6	116.5	134.8	151.9
$h = 5$	109.2	134.5	120.2	122.9	97.0	108.4	113.5	119.6
$h = 6$	108.9	128.2	132.3	129.1	113.1	105.6	114.5	115.1
$h = 7$	104.8	123.5	133.8	130.4	103.2	92.3	110.4	118.4
$h = 8$	103.4	119.2	128.0	128.0	130.3	93.6	109.3	97.3
$h = 9$	105.9	112.9	120.3	123.2	125.6	89.7	93.1	95.6
$h = 10$	104.9	108.2	121.3	125.6	126.1	100.1	102.3	96.4
$h = 11$	101.5	106.8	122.1	128.0	128.7	109.0	110.1	106.6
$h = 12$	100.3	107.3	125.7	130.0	120.5	109.8	99.2	91.3
Industrieproduktionsindex								
	iterierte Prognosen				direkte Prognosen			
	Klass5	Klass6	Klass7	Klass8	Klass5	Klass6	Klass7	Klass8
$h = 1$	96.5	105.5	98.6	115.7	90.2	93.5	111.9	111.2
$h = 2$	99.4	107.8	99.6	106.8	91.7	98.0	97.7	100.7
$h = 3$	97.6	104.0	108.0	115.1	92.4	98.0	94.7	104.0
$h = 4$	97.3	108.6	124.7	118.4	98.6	98.3	101.4	104.1
$h = 5$	99.4	109.9	130.3	123.2	95.6	98.2	102.2	97.2
$h = 6$	97.3	111.2	137.4	113.0	94.6	99.0	102.6	106.6
$h = 7$	97.0	103.9	134.2	126.0	97.0	104.7	106.5	102.8
$h = 8$	96.5	106.4	133.0	122.6	96.7	107.4	103.3	99.9
$h = 9$	98.2	105.2	140.4	116.2	118.6	113.8	106.9	105.7
$h = 10$	101.3	104.5	133.2	119.0	112.4	119.1	106.8	108.3
$h = 11$	103.8	107.2	138.9	115.7	115.3	115.6	118.0	125.1
$h = 12$	100.8	100.6	131.6	116.4	106.1	110.3	114.5	111.2

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Tab. 47: Out-of-Sample-Prognosegüte mit dem Ansatz mit Regularisierung unter Weglassung von Direktverbindungen, ALR und IPI

Arbeitslosenrate								
	iterierte Prognosen				direkte Prognosen			
	Regul5	Regul6	Regul7	Regul8	Regul5	Regul6	Regul7	Regul8
$h = 1$	103.7	119.8	116.6	120.6	103.5	125.4	117.1	122.2
$h = 2$	107.5	138.6	121.0	129.0	114.1	109.5	112.7	113.4
$h = 3$	104.6	154.3	135.3	152.7	111.8	103.4	105.6	110.4
$h = 4$	104.5	163.0	142.3	156.1	104.5	106.3	104.4	108.6
$h = 5$	102.2	155.1	138.4	160.2	105.1	118.0	109.0	108.6
$h = 6$	103.6	153.0	133.1	166.4	101.8	115.6	114.1	115.8
$h = 7$	99.5	146.1	124.2	163.2	104.3	114.7	109.4	108.2
$h = 8$	99.1	138.4	118.3	156.1	101.6	105.9	100.7	104.1
$h = 9$	98.6	130.3	116.0	147.8	97.6	106.1	97.5	101.4
$h = 10$	99.1	128.0	115.6	150.3	99.6	96.9	101.7	100.9
$h = 11$	100.1	126.5	115.5	143.5	94.4	91.7	99.0	99.4
$h = 12$	98.1	123.4	108.2	136.5	96.8	82.0	87.1	86.7
Industrieproduktionsindex								
	iterierte Prognosen				direkte Prognosen			
	Regul5	Regul6	Regul7	Regul8	Regul5	Regul6	Regul7	Regul8
$h = 1$	95.3	95.5	92.8	94.6	95.3	93.8	98.3	91.9
$h = 2$	98.4	99.6	95.6	97.1	93.9	97.1	93.0	92.7
$h = 3$	103.8	101.8	98.7	101.6	93.9	94.1	97.4	95.4
$h = 4$	105.8	115.9	113.9	111.3	114.8	110.6	110.1	105.5
$h = 5$	112.4	114.7	114.6	112.2	100.8	111.4	106.2	109.5
$h = 6$	110.4	112.9	112.8	105.8	107.2	107.2	109.7	108.2
$h = 7$	106.7	106.1	105.9	101.3	105.3	110.3	105.7	107.1
$h = 8$	109.6	109.3	110.7	106.5	102.2	102.8	108.3	104.7
$h = 9$	107.7	113.7	107.2	105.8	105.9	116.2	111.1	114.3
$h = 10$	106.5	111.5	112.0	109.4	110.0	112.2	114.7	113.1
$h = 11$	106.4	113.1	114.0	111.2	119.8	124.0	118.3	115.6
$h = 12$	101.7	103.1	110.5	104.7	110.8	111.0	109.3	109.0

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Tab. 48: Out-of-Sample-Prognosegüte mit dem Bayesianischen Ansatz unter Weglassung von Direktverbindungen, ALR und IPI

Arbeitslosenrate								
	iterierte Prognosen				direkte Prognosen			
	Bayes5	Bayes6	Bayes7	Bayes8	Bayes5	Bayes6	Bayes7	Bayes8
$h = 1$	100.6	100.3	144.6	126.2	99.7	108.0	150.5	172.9
$h = 2$	97.6	111.6	177.6	138.1	110.5	116.6	152.6	155.6
$h = 3$	103.0	122.2	241.5	171.2	110.0	112.2	167.9	150.4
$h = 4$	106.2	134.8	307.0	203.9	106.4	110.6	156.7	167.7
$h = 5$	107.2	132.6	326.3	218.0	105.8	107.9	159.9	239.7
$h = 6$	110.8	131.8	343.2	223.6	106.2	133.2	116.5	163.1
$h = 7$	109.1	126.3	333.1	231.2	103.8	143.6	207.6	171.4
$h = 8$	104.3	117.3	345.0	235.5	101.2	125.4	122.5	225.1
$h = 9$	102.7	107.7	308.9	229.7	103.0	119.0	163.6	129.3
$h = 10$	100.9	104.9	316.9	226.8	102.7	92.0	120.5	121.8
$h = 11$	100.8	109.4	281.3	243.5	101.8	102.3	121.4	126.7
$h = 12$	97.8	110.1	277.6	253.6	96.9	81.0	118.8	123.4
Industrieproduktionsindex								
	iterierte Prognosen				direkte Prognosen			
	Bayes5	Bayes6	Bayes7	Bayes8	Bayes5	Bayes6	Bayes7	Bayes8
$h = 1$	98.9	102.6	125.0	129.8	98.7	101.7	119.3	124.7
$h = 2$	99.4	104.0	121.3	130.0	95.2	97.8	107.8	122.1
$h = 3$	100.3	106.1	126.5	165.8	93.6	95.9	129.1	122.6
$h = 4$	100.4	108.9	136.8	177.7	98.6	107.6	154.0	162.0
$h = 5$	102.2	118.8	145.3	195.9	97.3	113.3	150.0	185.3
$h = 6$	99.3	118.6	174.2	216.2	99.9	104.6	164.4	181.0
$h = 7$	100.5	114.5	171.1	227.7	103.4	105.4	124.7	138.9
$h = 8$	98.4	116.4	183.6	253.7	100.6	103.7	111.9	123.1
$h = 9$	101.6	111.3	191.8	294.2	104.8	104.3	123.0	130.4
$h = 10$	101.1	111.6	188.8	295.9	107.2	109.5	105.9	137.7
$h = 11$	102.1	109.5	210.5	366.3	114.5	117.2	131.4	126.6
$h = 12$	102.2	110.6	209.5	376.8	110.4	111.1	115.8	112.4

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Tab. 49: Out-of-Sample-Prognosegüte mit linearen Modellen mit Jan1975 als Start des In-Samples, ALR

	iterierte Prognosen:			direkte Prognosen:		
	AR $\nabla_{12}Y_t$	AR ∇Y_t	AR $\nabla_{12}\nabla Y_t$	AR $\nabla_{12}Y_t$	AR ∇Y_t	AR $\nabla_{12}\nabla Y_t$
$h=1$	98.8	106.6	100.6	98.8	106.6	100.6
$h=2$	107.9	109.5	92.8	90.0	108.0	91.2
$h=3$	123.3	125.0	98.8	99.3	111.4	91.7
$h=4$	130.7	130.5	101.6	118.7	113.8	90.8
$h=5$	131.4	132.5	102.4	119.5	98.7	85.7
$h=6$	131.5	132.5	102.7	112.8	91.4	87.7
$h=7$	132.7	133.4	104.5	103.8	85.1	82.0
$h=8$	132.4	134.2	107.4	133.6	118.5	85.3
$h=9$	134.6	138.5	112.6	127.5	106.2	86.0
$h=10$	138.8	144.3	122.5	126.5	100.9	93.3
$h=11$	142.5	150.3	129.3	141.4	123.7	96.1
$h=12$	144.7	153.9	133.6	136.4	120.8	106.7
	ARMA $\nabla_{12}Y_t$	ARMA ∇Y_t	ARMA $\nabla_{12}\nabla Y_t$			
$h=1$	124.9	118.0	133.4			
$h=2$	129.7	107.5	149.3			
$h=3$	147.6	104.3	140.2			
$h=4$	151.2	108.5	140.5			
$h=5$	147.8	107.2	134.6			
$h=6$	141.2	106.9	138.2			
$h=7$	140.9	109.1	134.8			
$h=8$	137.7	109.8	137.0			
$h=9$	134.0	111.9	136.0			
$h=10$	134.9	117.3	149.9			
$h=11$	135.7	121.2	162.4			
$h=12$	137.2	122.7	167.7			

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Tab. 50: Out-of-Sample-Prognosegüte des statistisch-parametrischen Ansatzes mit Jan1975 als Start des In-Samples, ALR

	Param1 iteriert	Param2 iteriert	Param3 iteriert	Param4 iteriert	Param1 direkt	Param2 direkt	Param3 direkt	Param4 direkt
saisonale Differenzen, $\nabla_{12} Y_t$								
$h = 1$	105.1	102.8	100.6	110.3	105.1	102.9	98.3	115.5
$h = 2$	119.6	114.4	108.6	124.9	100.5	100.7	91.2	94.9
$h = 3$	136.7	133.8	122.2	165.0	112.1	115.1	111.7	121.6
$h = 4$	141.2	140.3	127.2	190.7	135.2	132.5	89.4	150.8
$h = 5$	137.5	134.7	122.2	203.5	134.7	132.0	144.8	144.4
$h = 6$	131.9	129.6	113.9	203.6	122.0	123.4	110.6	107.5
$h = 7$	128.4	125.8	115.8	199.6	114.1	115.2	162.6	162.0
$h = 8$	120.5	122.1	117.6	190.3	142.6	123.4	140.8	144.4
$h = 9$	118.1	119.6	119.7	187.4	150.4	129.5	137.6	152.1
$h = 10$	120.3	123.0	126.3	182.2	177.3	176.2	150.2	175.8
$h = 11$	122.5	125.2	130.6	183.8	162.4	170.2	167.2	151.1
$h = 12$	122.2	123.7	134.8	178.6	148.5	152.6	171.5	147.8
gewöhnliche Differenzen, ∇Y_t								
$h = 1$	109.5	127.1	94.3	125.2	112.3	104.8	110.0	114.0
$h = 2$	118.7	134.4	97.0	127.2	126.3	132.5	110.8	130.5
$h = 3$	140.0	156.4	121.9	141.0	118.8	124.0	131.2	148.3
$h = 4$	148.2	159.5	130.5	142.6	101.1	100.3	96.6	145.3
$h = 5$	152.9	159.7	133.2	139.2	104.8	112.5	97.7	109.1
$h = 6$	151.2	151.7	129.4	133.8	94.3	107.2	99.9	87.1
$h = 7$	149.0	148.5	129.8	129.8	75.2	87.8	79.7	114.0
$h = 8$	148.1	146.7	127.4	124.7	103.7	109.0	99.6	104.0
$h = 9$	149.5	146.1	130.5	123.2	88.7	98.9	102.7	107.3
$h = 10$	152.4	148.7	138.1	124.8	93.8	103.7	94.7	103.1
$h = 11$	157.1	156.6	146.2	130.1	133.8	124.2	97.9	96.1
$h = 12$	162.2	163.1	155.4	137.1	169.0	134.3	98.7	115.1
saisonale und gewöhnliche Differenzen, $\nabla_{12} \nabla Y_t$								
$h = 1$	100.9	103.1	87.1	112.5				
$h = 2$	97.4	96.6	84.5	103.9				
$h = 3$	102.3	105.1	94.3	97.5				
$h = 4$	105.6	105.1	97.5	91.6				
$h = 5$	106.3	104.4	99.9	91.3				
$h = 6$	105.3	104.6	92.4	88.2				
$h = 7$	111.2	109.6	90.7	87.1				
$h = 8$	114.1	112.7	93.0	88.6				
$h = 9$	118.5	119.9	96.7	93.3				
$h = 10$	129.5	133.2	105.0	100.4				
$h = 11$	137.0	143.3	110.0	106.4				
$h = 12$	141.7	150.1	114.2	107.3				

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Tab. 51: Out-of-Sample-Prognosegüte des klassischen Ansatzes, iterierte Prognosen, mit Jan1975 als Start des In-Samples, ALR

	Klass1	Klass2	Klass3	Klass4	Klass5	Klass6	Klass7	Klass8
saisonale Differenzen, $\nabla_{12} Y_t$								
$h = 1$	120.9	109.8	115.8	111.7	118.5	136.8	135.7	105.8
$h = 2$	146.5	119.5	140.6	123.4	137.7	164.8	151.1	106.5
$h = 3$	178.3	142.8	168.9	145.8	162.0	195.3	183.7	106.8
$h = 4$	194.2	158.5	188.2	152.5	170.7	201.8	204.8	99.1
$h = 5$	189.6	161.3	196.2	147.3	161.6	200.4	214.2	97.7
$h = 6$	176.5	161.3	195.6	137.4	151.3	194.3	217.1	99.4
$h = 7$	167.7	155.7	195.1	127.9	140.9	191.8	218.9	105.0
$h = 8$	162.9	147.3	191.7	128.1	131.8	188.8	236.1	111.8
$h = 9$	160.1	145.8	192.9	133.5	126.4	184.7	245.9	122.7
$h = 10$	156.7	146.6	190.3	138.7	123.7	177.9	275.4	135.6
$h = 11$	155.9	149.1	187.9	138.9	122.7	171.6	291.6	146.0
$h = 12$	154.9	150.3	184.6	138.2	121.2	161.6	304.3	159.6
gewöhnliche Differenzen, ∇Y_t								
$h = 1$	103.1	108.9	113.8	122.6	142.5	136.9	123.5	126.0
$h = 2$	108.3	108.8	116.6	130.8	116.9	126.4	117.7	140.3
$h = 3$	116.2	120.2	130.1	145.1	126.0	129.2	136.9	158.8
$h = 4$	120.4	119.8	137.3	142.3	135.1	132.2	146.8	163.2
$h = 5$	116.2	123.6	137.3	142.6	132.1	121.2	143.3	163.7
$h = 6$	115.8	126.8	131.8	133.3	131.5	120.3	140.1	155.2
$h = 7$	118.7	133.4	125.3	131.0	135.8	125.8	144.0	148.1
$h = 8$	120.3	134.7	121.7	121.1	148.4	136.0	141.4	140.2
$h = 9$	128.9	140.2	122.2	121.6	160.8	146.5	147.6	140.0
$h = 10$	136.0	147.7	125.3	126.1	174.6	161.7	155.1	143.2
$h = 11$	139.7	153.0	127.0	134.7	179.3	167.6	158.4	151.6
$h = 12$	140.4	154.3	127.6	141.4	175.5	161.3	159.6	162.0
saisonale und gewöhnliche Differenzen, $\nabla_{12} \nabla Y_t$								
$h = 1$	104.2	106.5	109.5	110.0	112.5	102.5	99.1	97.2
$h = 2$	99.6	102.9	96.8	102.9	104.4	94.8	103.5	88.4
$h = 3$	101.7	107.0	107.7	112.2	105.3	110.8	110.8	93.6
$h = 4$	107.1	108.0	116.0	117.1	108.4	118.0	119.3	98.3
$h = 5$	107.9	110.6	114.2	115.5	111.9	117.1	124.8	94.7
$h = 6$	108.8	110.0	117.2	119.2	114.1	117.9	124.3	94.5
$h = 7$	112.8	110.8	119.1	118.7	114.1	115.1	132.6	96.6
$h = 8$	115.4	113.2	123.3	120.2	118.3	115.0	135.2	102.4
$h = 9$	121.4	120.6	129.0	124.3	126.6	120.1	142.7	109.5
$h = 10$	129.2	133.1	140.9	135.5	139.9	128.5	153.9	123.3
$h = 11$	137.1	141.8	146.4	140.5	145.8	138.0	160.9	129.3
$h = 12$	141.0	145.5	148.0	142.1	148.1	140.1	166.9	137.6

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Tab. 52: Out-of-Sample-Prognosegüte des klassischen Ansatzes, direkte Prognosen, mit Jan1975 als Start des In-Samples, ALR

	Klass1	Klass2	Klass3	Klass4	Klass5	Klass6	Klass7	Klass8
saisonale Differenzen, $\nabla_{12} Y_t$								
$h = 1$	109.1	109.8	119.2	118.6	117.1	132.0	151.5	128.6
$h = 2$	117.5	115.9	119.9	123.0	117.4	138.9	130.3	136.7
$h = 3$	126.9	132.1	112.4	128.6	125.6	114.8	133.5	156.8
$h = 4$	151.5	156.9	181.3	148.6	159.7	155.8	133.3	150.7
$h = 5$	134.7	137.5	95.0	104.9	158.3	177.5	108.8	103.7
$h = 6$	130.6	165.2	146.5	97.9	126.4	132.1	148.5	110.3
$h = 7$	110.0	133.1	113.5	147.7	140.8	129.9	103.6	111.0
$h = 8$	135.9	183.9	169.1	162.0	175.6	165.4	161.6	144.9
$h = 9$	149.3	130.0	143.1	144.3	169.4	150.3	140.8	168.1
$h = 10$	168.5	146.7	210.5	172.1	164.4	140.9	209.3	188.1
$h = 11$	155.2	188.3	217.6	204.6	144.1	151.8	162.2	161.5
$h = 12$	192.7	220.2	221.4	250.1	172.7	175.8	189.8	219.1
gewöhnliche Differenzen, ∇Y_t								
$h = 1$	106.6	114.8	118.3	112.8	155.4	117.1	131.4	122.0
$h = 2$	103.7	111.3	115.5	98.3	138.7	156.7	140.7	126.6
$h = 3$	116.7	118.5	125.0	116.1	170.6	158.4	144.1	123.5
$h = 4$	126.8	109.6	118.5	111.4	156.4	124.3	114.8	128.2
$h = 5$	104.8	105.8	112.7	118.0	130.6	128.3	133.9	118.6
$h = 6$	101.7	109.4	101.3	105.1	115.0	96.0	114.7	107.2
$h = 7$	91.7	88.0	91.1	75.7	121.8	94.2	98.8	98.6
$h = 8$	105.7	110.0	103.1	99.7	125.1	132.9	125.3	104.4
$h = 9$	103.3	108.6	92.3	85.3	106.4	141.0	94.0	103.8
$h = 10$	111.4	103.3	87.7	101.7	137.1	115.7	118.1	121.5
$h = 11$	118.3	109.2	116.7	107.1	167.0	133.8	140.3	105.7
$h = 12$	148.0	134.8	133.9	132.4	136.7	136.6	105.2	129.5

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Tab. 53: Out-of-Sample-Prognosegüte des Ansatzes mit Regularisierung, iterierte Prognosen, mit Jan1975 als Start des In-Samples, ALR

	Regul1	Regul2	Regul3	Regul4	Regul5	Regul6	Regul7	Regul8
saisonale Differenzen, $\nabla_{12} Y_t$								
$h = 1$	119.8	140.5	162.5	173.8	112.0	186.5	225.1	154.8
$h = 2$	144.7	169.3	190.8	192.6	133.4	263.4	244.7	186.2
$h = 3$	173.1	198.2	217.5	210.9	155.8	348.7	265.2	209.5
$h = 4$	186.1	204.5	209.4	206.1	162.8	399.8	250.4	205.8
$h = 5$	180.9	208.0	196.2	191.6	158.8	400.6	229.9	195.7
$h = 6$	178.4	210.5	183.0	178.3	152.4	379.5	210.0	184.9
$h = 7$	174.3	217.3	178.7	173.5	146.7	362.1	196.8	184.1
$h = 8$	169.2	211.6	176.4	166.9	139.8	333.9	185.8	183.0
$h = 9$	166.7	212.8	178.1	166.0	136.0	315.5	182.8	186.0
$h = 10$	167.6	210.4	179.1	168.4	137.9	299.9	182.2	189.8
$h = 11$	167.3	196.8	179.9	166.7	138.0	280.7	173.0	189.7
$h = 12$	163.5	181.5	176.1	158.7	133.8	258.4	160.8	182.2
gewöhnliche Differenzen, ∇Y_t								
$h = 1$	101.1	103.5	104.8	111.1	112.2	111.2	110.3	107.4
$h = 2$	99.3	101.7	102.4	115.4	121.0	117.6	122.2	105.9
$h = 3$	109.0	109.4	118.0	128.3	139.6	134.2	137.6	124.7
$h = 4$	110.5	109.8	122.2	124.3	151.0	143.2	144.0	134.0
$h = 5$	107.8	109.8	122.7	121.0	144.1	137.8	143.3	137.0
$h = 6$	104.0	108.5	121.7	117.3	144.0	137.4	143.1	137.5
$h = 7$	101.0	109.8	119.4	112.3	141.7	135.1	143.0	138.2
$h = 8$	101.5	113.4	121.9	110.4	142.6	133.2	140.8	136.0
$h = 9$	104.3	119.5	125.7	114.3	142.9	136.7	141.4	138.5
$h = 10$	109.5	125.1	133.8	119.9	146.4	139.7	142.0	140.2
$h = 11$	113.4	128.9	135.5	128.9	153.0	143.5	143.7	139.0
$h = 12$	116.9	131.3	137.0	135.6	160.9	146.4	146.2	139.5
saisonale und gewöhnliche Differenzen, $\nabla_{12} \nabla Y_t$								
$h = 1$	101.5	103.7	100.8	101.5	101.8	96.0	102.0	99.1
$h = 2$	94.2	95.0	94.1	88.0	101.5	93.2	92.2	94.6
$h = 3$	100.0	106.4	100.8	97.7	110.4	102.5	90.7	100.2
$h = 4$	106.1	111.6	106.8	103.5	115.6	106.2	93.9	108.8
$h = 5$	109.4	110.0	109.2	101.3	120.3	108.0	95.6	107.6
$h = 6$	108.8	109.3	110.7	101.5	121.7	103.9	94.4	103.0
$h = 7$	110.1	111.9	113.2	104.7	126.6	104.9	97.8	107.2
$h = 8$	113.4	115.4	118.5	109.3	127.8	109.2	100.5	111.2
$h = 9$	119.2	120.2	124.2	115.6	134.0	118.1	106.3	117.4
$h = 10$	129.4	133.2	136.2	126.5	143.0	132.5	115.4	128.5
$h = 11$	135.3	142.8	144.3	134.6	150.5	139.7	122.0	136.6
$h = 12$	144.3	145.8	149.1	139.2	155.8	147.2	127.4	141.3

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Tab. 54: Out-of-Sample-Prognosegüte des Ansatzes mit Regularisierung, direkte Prognosen, mit Jan1975 als Start des In-Samples, ALR

	Regul1	Regul2	Regul3	Regul4	Regul5	Regul6	Regul7	Regul8
saisonale Differenzen, $\nabla_{12} Y_t$								
$h = 1$	122.0	142.3	145.4	151.8	112.0	194.5	189.2	209.7
$h = 2$	77.6	91.7	133.5	129.6	111.6	155.0	155.6	153.6
$h = 3$	132.2	151.0	148.9	149.7	118.1	175.5	152.9	157.1
$h = 4$	189.3	150.6	158.0	187.0	139.1	159.7	167.9	184.7
$h = 5$	173.1	169.8	155.2	156.8	137.2	166.0	158.5	156.1
$h = 6$	157.3	151.2	142.6	144.8	124.3	157.1	141.0	141.3
$h = 7$	139.3	133.3	128.6	134.1	117.3	123.1	125.7	124.1
$h = 8$	148.4	157.0	157.0	147.6	146.9	146.2	150.1	150.4
$h = 9$	155.8	156.7	154.6	151.6	149.0	156.9	154.9	139.9
$h = 10$	148.6	148.5	145.2	148.0	143.6	159.8	143.6	144.6
$h = 11$	147.5	148.6	148.2	144.8	146.0	151.6	136.0	149.0
$h = 12$	145.5	143.9	147.8	143.1	132.9	158.7	145.0	139.6
gewöhnliche Differenzen, ∇Y_t								
$h = 1$	103.2	102.3	108.8	104.8	109.5	102.5	117.5	108.5
$h = 2$	123.7	116.7	113.0	124.8	140.1	115.2	112.8	122.8
$h = 3$	111.7	111.5	114.6	108.8	127.6	143.2	109.4	117.7
$h = 4$	90.4	87.2	88.6	92.1	108.0	101.2	106.8	110.3
$h = 5$	83.3	95.5	90.7	85.5	118.5	105.5	85.4	91.4
$h = 6$	80.2	70.8	81.4	68.8	89.5	93.5	85.6	85.2
$h = 7$	79.8	75.2	72.2	79.8	72.4	77.1	79.1	77.8
$h = 8$	109.2	107.0	110.0	107.9	105.5	110.7	108.8	106.9
$h = 9$	88.7	99.2	101.8	101.8	93.4	103.0	99.0	92.4
$h = 10$	107.3	107.1	105.3	104.1	104.4	110.7	103.2	105.7
$h = 11$	114.8	118.1	120.7	125.7	108.5	100.7	106.4	112.9
$h = 12$	115.5	121.6	116.9	116.8	122.6	115.4	125.0	133.5

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Tab. 55: Out-of-Sample-Prognosegüte des Bayesianischen Ansatzes, iterierte Prognosen, mit Jan1975 als Start des In-Samples, ALR

	Bayes1	Bayes2	Bayes3	Bayes4	Bayes5	Bayes6	Bayes7	Bayes8
saisonale Differenzen, $\nabla_{12} Y_t$								
$h = 1$	111.5	100.2	107.4	122.5	100.2	104.2	124.6	147.4
$h = 2$	131.3	110.8	131.3	145.4	110.7	114.9	160.3	173.6
$h = 3$	155.4	124.7	162.1	163.2	123.7	132.2	227.6	200.6
$h = 4$	167.9	129.6	194.6	174.1	132.9	145.3	296.9	198.6
$h = 5$	173.3	137.4	203.5	190.0	131.1	148.5	375.2	172.2
$h = 6$	175.3	138.0	212.4	216.8	131.6	146.2	495.9	141.3
$h = 7$	176.8	142.2	235.2	252.3	132.8	144.1	687.7	125.9
$h = 8$	176.5	141.7	252.4	293.4	132.4	140.6	956.3	121.0
$h = 9$	178.8	140.7	271.1	337.2	134.7	140.9	1306.5	126.3
$h = 10$	182.7	143.6	296.7	407.1	138.0	144.1	1509.4	136.2
$h = 11$	176.7	146.0	327.0	541.9	142.0	148.1	1183.8	143.8
$h = 12$	164.0	150.9	378.1	857.6	144.0	149.3	530.0	148.6
gewöhnliche Differenzen, ∇Y_t								
$h = 1$	105.8	108.0	98.5	106.3	113.5	100.9	103.5	112.6
$h = 2$	108.4	107.9	104.0	102.1	120.9	102.7	106.0	102.6
$h = 3$	124.4	110.1	104.6	101.8	143.4	112.3	111.8	103.7
$h = 4$	129.2	108.6	102.0	98.1	152.7	113.7	102.0	98.3
$h = 5$	132.9	106.2	103.8	101.3	151.4	107.2	102.0	96.4
$h = 6$	129.0	100.0	105.8	100.2	148.2	103.0	99.0	93.3
$h = 7$	130.2	91.8	107.1	97.1	146.5	98.0	98.3	90.4
$h = 8$	127.6	87.0	111.5	97.4	143.0	98.0	102.5	92.9
$h = 9$	124.9	88.0	117.6	98.9	141.3	102.2	106.5	99.0
$h = 10$	124.0	89.9	123.8	102.2	144.4	110.8	111.7	106.8
$h = 11$	125.4	93.3	128.1	105.6	149.9	118.6	114.0	113.2
$h = 12$	128.5	97.0	133.4	110.2	152.0	125.8	115.4	119.9
saisonale und gewöhnliche Differenzen, $\nabla_{12} \nabla Y_t$								
$h = 1$	100.6	100.8	100.0	102.0	100.8	103.2	100.6	105.1
$h = 2$	95.4	93.9	95.8	94.2	90.9	98.4	98.0	99.6
$h = 3$	104.7	103.0	104.8	103.5	100.3	105.3	108.1	106.1
$h = 4$	106.9	110.4	107.8	107.9	103.2	111.7	109.2	107.5
$h = 5$	112.4	109.5	106.9	106.8	105.3	115.2	108.8	106.3
$h = 6$	112.5	110.3	106.8	106.3	107.1	113.9	105.9	107.4
$h = 7$	112.2	113.3	110.2	108.5	109.3	116.2	107.4	110.6
$h = 8$	116.8	117.9	115.3	109.9	111.5	122.2	108.6	117.1
$h = 9$	123.8	126.7	122.5	115.0	116.0	126.4	117.0	126.3
$h = 10$	131.3	135.8	132.0	123.4	126.6	137.1	127.7	135.2
$h = 11$	140.3	141.9	137.3	127.4	131.0	142.3	133.6	142.8
$h = 12$	145.8	147.8	142.6	130.8	134.1	144.0	137.8	146.2

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Tab. 56: Out-of-Sample-Prognosegüte des Bayesianischen Ansatzes, direkte Prognosen, mit Jan1975 als Start des In-Samples, ALR

	Bayes1	Bayes2	Bayes3	Bayes4	Bayes5	Bayes6	Bayes7	Bayes8
saisonale Differenzen, $\nabla_{12} Y_t$								
$h = 1$	108.5	105.2	134.5	106.5	100.0	103.5	101.7	117.8
$h = 2$	97.6	81.1	141.8	233.1	90.5	115.3	153.2	159.5
$h = 3$	113.6	143.3	159.5	187.7	101.4	126.5	131.0	177.4
$h = 4$	220.3	155.1	150.5	113.5	121.8	152.5	136.2	147.6
$h = 5$	194.2	177.8	152.1	162.4	124.5	170.1	148.7	163.2
$h = 6$	181.1	227.6	88.4	141.1	109.0	152.0	100.1	114.3
$h = 7$	182.6	155.8	101.9	137.4	110.9	126.4	85.5	194.7
$h = 8$	110.1	140.3	117.3	157.5	139.9	129.0	121.1	230.0
$h = 9$	154.0	146.5	147.8	215.1	134.9	139.6	153.4	184.8
$h = 10$	161.2	233.9	216.8	245.3	161.5	193.4	261.0	224.4
$h = 11$	127.8	194.1	154.3	164.0	154.2	120.5	140.0	168.4
$h = 12$	155.4	182.6	162.4	290.2	143.3	164.3	145.4	313.0
gewöhnliche Differenzen, ∇Y_t								
$h = 1$	101.4	99.4	103.5	97.8	108.5	108.5	102.9	107.5
$h = 2$	128.3	107.7	106.3	102.3	120.7	106.4	94.6	101.8
$h = 3$	123.4	112.1	104.2	109.3	112.6	109.9	98.1	102.2
$h = 4$	119.1	110.7	92.6	92.7	110.3	98.8	97.6	87.9
$h = 5$	90.8	91.8	91.2	119.6	103.5	98.6	115.9	101.9
$h = 6$	81.9	79.1	86.6	87.1	89.7	88.7	90.0	91.6
$h = 7$	85.7	75.4	97.5	89.2	88.7	76.6	84.3	92.3
$h = 8$	102.0	100.0	97.6	101.5	101.4	88.8	101.1	102.7
$h = 9$	91.4	80.3	95.7	101.9	91.0	73.7	90.0	94.1
$h = 10$	95.3	87.9	115.3	111.4	103.5	89.5	113.3	113.8
$h = 11$	109.1	95.6	94.0	92.3	113.9	104.7	99.5	89.8
$h = 12$	113.2	104.5	91.4	92.8	141.6	123.1	106.1	100.2

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Tab. 57: Out-of-Sample-Prognosegüte mit linearen Modellen mit Jan1975 als Start des In-Samples, IPI

	iterierte Prognosen:			direkte Prognosen:		
	AR $\nabla_{12}Y_t$	AR ∇Y_t	AR $\nabla_{12}\nabla Y_t$	AR $\nabla_{12}Y_t$	AR ∇Y_t	AR $\nabla_{12}\nabla Y_t$
$h=1$	112.9	115.2	97.0	112.9	115.2	97.0
$h=2$	120.5	125.6	98.3	121.3	115.9	104.8
$h=3$	131.5	122.2	100.0	102.0	122.1	106.6
$h=4$	141.6	113.4	104.4	129.6	125.0	121.2
$h=5$	155.4	114.7	110.7	134.4	133.1	111.4
$h=6$	162.9	114.7	117.0	130.2	117.2	111.7
$h=7$	159.8	122.0	119.0	127.7	122.9	129.1
$h=8$	162.0	123.7	124.3	136.3	129.1	109.8
$h=9$	170.6	126.4	129.1	136.0	135.9	147.1
$h=10$	174.4	121.2	141.3	140.5	129.9	155.8
$h=11$	182.6	122.2	146.9	159.0	123.8	141.8
$h=12$	173.1	123.4	151.7	150.1	119.6	149.9
	ARMA $\nabla_{12}Y_t$	ARMA ∇Y_t	ARMA $\nabla_{12}\nabla Y_t$			
$h=1$	98.4	215.5	94.0			
$h=2$	104.1	220.4	92.6			
$h=3$	110.7	155.3	97.7			
$h=4$	122.4	151.9	106.7			
$h=5$	132.4	159.9	109.3			
$h=6$	136.5	134.6	108.3			
$h=7$	141.7	159.9	114.2			
$h=8$	146.4	173.2	112.3			
$h=9$	158.2	141.3	118.9			
$h=10$	165.3	155.8	121.9			
$h=11$	187.3	145.2	150.2			
$h=12$	182.7	123.5	159.9			

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Tab. 58: Out-of-Sample-Prognosegüte des statistisch-parametrischen Ansatzes mit Jan1975 als Start des In-Samples, IPI

	Param1 iteriert	Param2 iteriert	Param3 iteriert	Param4 iteriert	Param1 direkt	Param2 direkt	Param3 direkt	Param4 direkt
saisonale Differenzen, $\nabla_{12} Y_t$								
$h = 1$	110.4	111.9	119.6	131.5	110.2	112.3	112.8	101.3
$h = 2$	119.5	120.4	122.6	143.2	116.0	114.1	128.4	136.7
$h = 3$	129.9	131.1	138.2	155.1	96.2	100.5	118.2	116.1
$h = 4$	140.5	143.1	139.9	154.2	123.2	127.1	120.7	187.4
$h = 5$	154.6	154.4	155.2	172.6	133.1	131.6	127.6	139.5
$h = 6$	161.2	162.9	164.1	188.9	128.4	128.9	131.0	131.9
$h = 7$	155.0	153.6	162.8	176.2	127.2	131.1	138.1	139.7
$h = 8$	159.2	160.1	160.5	184.5	139.4	128.8	131.6	141.3
$h = 9$	164.1	164.7	171.1	200.9	133.1	136.4	149.9	141.0
$h = 10$	165.5	171.2	180.4	200.9	137.4	139.8	165.3	144.0
$h = 11$	174.1	180.3	200.9	197.5	161.9	115.7	141.8	143.9
$h = 12$	162.1	166.3	202.7	192.4	163.8	147.3	148.2	150.2
gewöhnliche Differenzen, ∇Y_t								
$h = 1$	113.0	114.2	118.6	138.6	112.8	113.9	125.7	127.2
$h = 2$	115.9	119.3	138.9	130.9	127.3	139.3	117.6	116.7
$h = 3$	113.8	118.0	133.0	129.5	133.5	137.7	109.6	126.5
$h = 4$	108.4	110.7	129.6	118.6	136.0	139.0	146.8	142.6
$h = 5$	113.2	112.3	134.0	130.0	135.5	141.5	135.6	142.7
$h = 6$	111.4	109.6	140.1	143.6	118.4	125.7	131.2	157.4
$h = 7$	117.4	116.0	144.2	140.1	126.9	129.1	142.6	150.4
$h = 8$	122.0	114.3	142.3	138.9	138.6	145.2	127.1	142.1
$h = 9$	120.2	120.2	145.6	138.8	165.0	172.8	203.9	187.0
$h = 10$	115.8	115.3	140.7	136.4	147.5	159.1	145.9	168.4
$h = 11$	118.0	116.2	152.3	145.1	123.5	119.4	120.9	142.9
$h = 12$	123.3	115.9	160.2	151.3	121.9	122.8	152.6	138.5
saisonale und gewöhnliche Differenzen, $\nabla_{12} \nabla Y_t$								
$h = 1$	98.3	98.0	86.2	117.9				
$h = 2$	102.8	99.8	94.6	117.1				
$h = 3$	103.3	103.2	108.2	108.4				
$h = 4$	109.2	111.7	110.9	126.3				
$h = 5$	116.1	118.1	112.9	137.2				
$h = 6$	123.6	128.5	122.8	134.0				
$h = 7$	127.0	127.0	123.5	137.3				
$h = 8$	131.3	131.9	127.1	143.3				
$h = 9$	135.9	142.6	131.1	146.2				
$h = 10$	153.5	156.0	139.2	155.1				
$h = 11$	160.3	162.4	156.0	164.8				
$h = 12$	167.3	167.5	164.7	174.8				

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Tab. 59: Out-of-Sample-Prognosegüte des klassischen Ansatzes, iterierte Prognosen, mit Jan1975 als Start des In-Samples, IPI

	Klass1	Klass2	Klass3	Klass4	Klass5	Klass6	Klass7	Klass8
saisonale Differenzen, $\nabla_{12} Y_t$								
$h = 1$	103.1	107.4	114.3	103.1	111.6	104.3	113.0	112.2
$h = 2$	109.8	112.0	122.4	115.6	109.5	112.7	120.6	111.3
$h = 3$	118.5	124.3	127.5	132.7	112.9	117.8	134.1	111.5
$h = 4$	135.0	141.8	139.3	149.8	119.2	141.3	141.1	115.4
$h = 5$	145.9	150.0	157.3	149.8	130.4	153.4	148.2	132.4
$h = 6$	152.8	149.1	164.3	149.5	127.4	171.4	140.7	138.6
$h = 7$	145.6	140.6	152.9	148.0	129.3	169.5	142.4	145.6
$h = 8$	143.5	140.4	151.8	152.4	129.7	172.9	144.8	149.8
$h = 9$	152.6	145.6	161.8	163.0	145.0	171.1	157.4	161.9
$h = 10$	148.1	148.3	160.0	161.4	141.2	176.0	156.2	167.8
$h = 11$	152.5	146.4	169.0	167.3	149.5	192.5	166.8	186.8
$h = 12$	137.7	138.9	162.5	154.8	142.2	184.4	160.2	179.9
gewöhnliche Differenzen, ∇Y_t								
$h = 1$	127.5	134.0	133.0	139.4	189.8	174.5	159.4	131.9
$h = 2$	143.8	142.6	137.4	137.0	178.8	175.0	162.1	148.2
$h = 3$	143.3	137.4	138.8	134.1	163.8	198.4	148.2	146.4
$h = 4$	124.4	128.0	123.5	129.4	140.3	168.4	129.4	124.0
$h = 5$	124.3	129.4	118.0	127.7	153.1	167.6	141.7	124.9
$h = 6$	116.9	128.6	118.6	122.9	125.7	137.3	133.3	119.3
$h = 7$	126.5	135.8	122.8	128.5	137.6	150.0	140.8	128.2
$h = 8$	136.4	133.9	129.5	129.1	157.2	167.3	141.3	129.5
$h = 9$	144.0	139.5	122.4	125.2	178.5	201.3	149.4	128.4
$h = 10$	131.8	129.7	111.9	117.9	170.0	195.8	138.0	111.6
$h = 11$	127.1	133.5	108.0	114.1	147.9	177.5	132.6	109.8
$h = 12$	122.6	128.0	108.8	109.9	116.3	155.4	127.9	109.2
saisonale und gewöhnliche Differenzen, $\nabla_{12} \nabla Y_t$								
$h = 1$	96.2	94.8	96.7	95.2	110.6	105.8	112.2	91.3
$h = 2$	97.8	97.1	93.8	97.4	112.6	106.7	116.0	101.3
$h = 3$	102.6	103.3	100.7	106.8	116.4	109.9	112.7	108.5
$h = 4$	109.4	108.1	103.9	113.9	114.4	114.3	120.3	119.9
$h = 5$	117.2	113.3	109.3	118.6	116.8	136.1	136.8	123.1
$h = 6$	122.5	119.6	120.0	120.4	125.3	129.0	153.7	124.0
$h = 7$	129.9	120.7	117.2	125.2	130.2	154.8	159.8	127.6
$h = 8$	131.5	130.9	128.5	128.9	139.0	155.1	158.5	133.8
$h = 9$	136.9	137.2	132.1	136.8	140.1	171.1	153.1	138.0
$h = 10$	157.5	150.4	148.3	148.7	156.6	186.8	173.8	155.2
$h = 11$	162.9	159.6	152.9	157.0	171.1	190.1	193.5	162.7
$h = 12$	165.6	160.3	153.2	162.4	174.3	190.1	202.5	163.5

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Tab. 60: Out-of-Sample-Prognosegüte des klassischen Ansatzes, direkte Prognosen, mit Jan1975 als Start des In-Samples, IPI

	Klass1	Klass2	Klass3	Klass4	Klass5	Klass6	Klass7	Klass8
saisonale Differenzen, $\nabla_{12} Y_t$								
$h = 1$	110.2	118.0	110.4	113.0	114.5	114.2	101.8	118.7
$h = 2$	124.4	126.1	129.1	136.8	119.6	125.4	134.3	119.0
$h = 3$	104.5	107.4	114.1	107.1	103.2	111.5	106.8	145.8
$h = 4$	114.2	145.1	119.8	139.8	126.6	135.6	134.5	122.1
$h = 5$	142.2	133.3	142.0	138.9	168.6	147.9	135.9	143.9
$h = 6$	139.4	128.9	138.5	133.1	130.6	106.6	127.2	134.4
$h = 7$	141.8	127.1	151.6	148.2	150.8	153.3	116.2	156.4
$h = 8$	124.3	123.5	135.5	128.5	133.4	123.0	121.0	135.5
$h = 9$	151.4	148.2	152.7	148.0	144.6	117.4	136.6	134.2
$h = 10$	156.9	154.3	140.2	157.3	123.4	117.0	138.3	153.7
$h = 11$	167.6	154.3	183.0	238.7	119.5	157.5	154.2	183.3
$h = 12$	151.1	162.5	177.2	179.0	128.2	142.1	143.7	136.1
gewöhnliche Differenzen, ∇Y_t								
$h = 1$	135.0	130.8	140.9	138.3	221.0	174.5	150.8	142.8
$h = 2$	138.8	139.9	138.9	138.3	190.4	204.4	156.1	124.0
$h = 3$	140.9	142.2	128.3	141.5	225.7	224.0	149.0	136.7
$h = 4$	148.5	148.8	133.3	138.5	191.0	157.9	153.5	135.6
$h = 5$	145.2	140.1	141.2	150.2	212.3	194.0	155.2	140.7
$h = 6$	119.8	131.5	120.5	141.3	157.0	150.5	148.0	132.7
$h = 7$	146.4	141.0	142.8	140.6	151.4	177.8	138.8	142.9
$h = 8$	151.3	149.6	144.0	144.9	146.9	146.3	136.7	139.0
$h = 9$	147.4	147.2	139.0	146.5	186.1	180.2	170.1	155.1
$h = 10$	121.1	135.8	127.7	131.4	171.6	179.8	167.6	129.9
$h = 11$	126.3	127.2	122.8	126.9	140.0	134.3	133.4	130.3
$h = 12$	115.9	114.3	117.3	121.7	108.5	104.7	102.4	115.8

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Tab. 61: Out-of-Sample-Prognosegüte des Ansatzes mit Regularisierung, iterierte Prognosen, mit Jan1975 als Start des In-Samples, IPI

	Regul1	Regul2	Regul3	Regul4	Regul5	Regul6	Regul7	Regul8
saisonale Differenzen, $\nabla_{12} Y_t$								
$h = 1$	108.2	106.4	107.1	106.7	108.9	104.0	111.9	106.5
$h = 2$	113.8	118.2	112.6	112.3	121.1	108.9	118.5	113.0
$h = 3$	121.8	129.2	120.8	126.0	130.0	121.5	134.2	123.8
$h = 4$	133.2	140.8	133.8	140.1	142.4	133.4	156.8	139.2
$h = 5$	154.0	159.8	147.1	152.3	154.7	147.5	168.1	151.8
$h = 6$	163.5	165.3	150.7	153.5	160.0	156.1	178.3	155.5
$h = 7$	155.1	160.3	144.3	144.2	151.4	148.8	166.4	145.4
$h = 8$	155.7	157.5	139.7	145.5	151.7	142.8	171.4	142.1
$h = 9$	163.9	165.1	148.5	155.9	158.1	151.0	188.9	153.6
$h = 10$	162.8	164.3	152.7	154.1	156.6	154.7	194.5	158.0
$h = 11$	171.4	176.7	159.2	165.8	168.1	161.4	208.3	165.0
$h = 12$	163.9	167.1	148.2	153.1	154.5	149.5	200.7	155.7
gewöhnliche Differenzen, ∇Y_t								
$h = 1$	125.0	124.7	125.5	126.8	114.0	122.3	123.5	127.1
$h = 2$	138.3	124.6	137.0	132.9	117.8	123.8	126.8	139.1
$h = 3$	130.8	130.8	130.0	121.2	114.8	122.1	124.2	135.5
$h = 4$	123.1	136.5	116.7	109.3	106.6	112.1	120.9	126.7
$h = 5$	119.1	135.5	121.3	114.0	110.6	113.5	123.4	120.7
$h = 6$	111.5	143.1	117.0	113.0	109.5	112.5	123.6	119.7
$h = 7$	127.6	145.2	123.7	113.8	113.6	121.4	134.9	128.5
$h = 8$	132.7	151.3	126.1	116.7	119.4	123.1	138.5	137.9
$h = 9$	130.4	166.9	138.0	121.5	123.3	126.0	144.9	142.1
$h = 10$	127.8	161.0	128.2	116.1	118.4	119.9	137.7	138.1
$h = 11$	117.0	161.3	124.8	119.4	121.3	114.7	141.2	139.3
$h = 12$	115.8	166.4	120.5	116.4	125.9	114.4	142.1	140.5
saisonale und gewöhnliche Differenzen, $\nabla_{12} \nabla Y_t$								
$h = 1$	101.8	101.0	98.0	99.6	97.4	103.3	104.7	104.3
$h = 2$	108.0	103.3	93.7	101.6	98.5	104.4	92.4	115.5
$h = 3$	112.4	112.8	103.3	108.8	102.3	106.1	105.2	117.8
$h = 4$	118.5	116.0	113.3	111.0	104.9	108.8	115.8	123.1
$h = 5$	123.5	129.2	120.2	123.2	114.1	115.0	121.9	132.0
$h = 6$	126.1	134.5	132.4	125.3	122.8	119.6	121.1	146.3
$h = 7$	126.5	132.9	129.7	131.2	125.2	125.3	128.9	148.0
$h = 8$	137.3	141.1	140.8	142.4	129.8	131.4	145.1	153.5
$h = 9$	144.9	148.4	148.7	152.8	140.3	137.4	157.8	158.9
$h = 10$	160.3	166.8	163.2	171.5	154.2	151.3	175.6	179.0
$h = 11$	160.0	173.2	176.0	176.7	158.8	161.0	185.0	190.1
$h = 12$	161.3	172.9	178.9	182.3	165.4	164.8	187.7	192.4

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Tab. 62: Out-of-Sample-Prognosegüte des Ansatzes mit Regularisierung, direkte Prognosen, mit Jan1975 als Start des In-Samples, IPI

	Regul1	Regul2	Regul3	Regul4	Regul5	Regul6	Regul7	Regul8
saisonale Differenzen, $\nabla_{12} Y_t$								
$h = 1$	105.8	106.5	107.5	108.4	108.8	105.0	119.1	107.2
$h = 2$	123.9	115.0	114.6	118.3	116.6	116.5	115.9	116.0
$h = 3$	99.2	104.0	103.2	104.2	97.2	105.5	96.4	101.0
$h = 4$	122.6	127.6	127.8	121.6	122.0	118.6	118.1	123.3
$h = 5$	128.8	135.9	131.4	128.1	146.2	131.8	130.3	129.1
$h = 6$	127.5	127.9	129.7	129.3	128.1	126.4	125.5	124.3
$h = 7$	129.9	126.3	121.8	126.5	124.7	127.9	125.4	126.7
$h = 8$	134.6	126.0	126.7	126.4	131.2	124.7	135.4	128.4
$h = 9$	132.0	133.4	132.1	132.6	132.9	133.2	133.9	131.9
$h = 10$	140.2	133.2	133.0	141.9	135.4	131.2	134.4	132.1
$h = 11$	141.8	139.3	138.6	140.8	152.8	165.8	141.1	146.8
$h = 12$	144.2	144.7	144.3	136.6	149.3	141.7	140.7	137.3
gewöhnliche Differenzen, ∇Y_t								
$h = 1$	123.4	133.2	122.1	125.2	116.3	124.0	128.7	125.2
$h = 2$	107.6	114.1	117.4	135.4	139.0	127.2	124.6	130.7
$h = 3$	130.8	129.1	137.3	134.9	136.0	125.5	137.1	126.4
$h = 4$	129.2	140.2	135.0	135.5	134.2	129.3	142.4	145.7
$h = 5$	152.7	141.0	128.6	135.9	127.3	139.2	139.2	136.6
$h = 6$	117.3	128.9	126.7	139.7	119.8	135.5	136.0	133.2
$h = 7$	120.0	125.1	138.5	127.9	131.3	134.1	132.1	128.3
$h = 8$	136.8	131.9	128.5	140.0	131.7	135.2	133.6	136.7
$h = 9$	155.7	144.5	164.8	154.4	180.6	144.5	153.8	147.4
$h = 10$	125.2	135.1	133.8	144.0	131.0	120.2	157.2	142.9
$h = 11$	117.6	118.3	117.7	117.0	128.3	120.8	116.0	118.1
$h = 12$	115.6	117.8	127.8	118.1	116.2	142.6	121.5	136.8

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Tab. 63: Out-of-Sample-Prognosegüte des Bayesianischen Ansatzes, iterierte Prognosen, mit Jan1975 als Start des In-Samples, IPI

	Bayes1	Bayes2	Bayes3	Bayes4	Bayes5	Bayes6	Bayes7	Bayes8
saisonale Differenzen, $\nabla_{12} Y_t$								
$h = 1$	110.2	110.5	112.2	114.9	110.4	109.8	120.4	122.6
$h = 2$	116.0	115.2	116.5	122.6	119.0	116.1	123.2	132.3
$h = 3$	128.6	126.7	129.1	152.8	127.5	123.8	132.5	147.4
$h = 4$	141.1	138.6	135.3	155.1	139.8	132.3	150.1	160.3
$h = 5$	152.5	156.7	145.7	162.8	156.6	152.6	172.6	188.1
$h = 6$	157.5	159.4	154.1	163.4	157.7	159.7	188.6	198.0
$h = 7$	154.1	156.3	154.9	167.7	151.6	155.8	180.4	187.0
$h = 8$	157.9	157.0	162.3	183.6	155.4	153.2	199.8	212.4
$h = 9$	161.1	164.9	175.5	206.6	166.0	166.7	220.0	229.1
$h = 10$	168.0	167.8	186.8	217.4	166.2	171.7	233.8	235.8
$h = 11$	175.7	176.6	213.2	245.3	173.0	176.3	258.7	255.5
$h = 12$	163.8	165.2	230.2	252.0	162.7	166.4	272.9	283.8
gewöhnliche Differenzen, ∇Y_t								
$h = 1$	119.7	121.0	116.9	118.9	126.5	117.9	125.7	131.6
$h = 2$	122.6	124.4	121.3	123.8	143.7	135.6	121.2	156.7
$h = 3$	121.6	117.0	126.0	113.8	138.2	138.4	127.8	145.4
$h = 4$	114.3	116.2	131.7	121.8	127.0	130.3	128.4	154.4
$h = 5$	119.1	128.9	141.1	108.3	128.4	133.5	147.2	164.1
$h = 6$	125.7	126.9	137.5	102.9	123.9	126.5	139.5	155.8
$h = 7$	128.5	132.8	136.7	116.6	130.3	140.1	144.6	151.2
$h = 8$	130.7	134.8	147.9	116.1	138.6	150.0	145.7	148.4
$h = 9$	137.8	133.1	147.7	101.9	151.0	155.0	152.8	127.9
$h = 10$	135.0	133.6	152.7	104.0	144.0	150.2	155.8	118.7
$h = 11$	144.1	148.9	156.8	106.1	135.7	154.2	170.8	134.7
$h = 12$	144.5	156.1	163.6	118.0	129.7	151.4	178.1	141.1
saisonale und gewöhnliche Differenzen, $\nabla_{12} \nabla Y_t$								
$h = 1$	96.6	98.4	98.9	107.1	96.7	97.2	105.7	128.1
$h = 2$	98.5	104.6	96.5	103.5	98.9	95.2	109.5	112.1
$h = 3$	101.7	108.8	98.0	110.9	100.4	99.8	116.4	119.5
$h = 4$	109.9	114.2	105.9	131.9	106.8	108.9	122.3	133.1
$h = 5$	113.0	122.7	108.1	132.4	112.4	117.6	131.8	136.6
$h = 6$	127.4	131.5	124.7	139.0	127.5	126.2	132.1	141.6
$h = 7$	127.4	131.6	126.1	142.2	126.2	133.7	128.3	155.4
$h = 8$	134.3	136.5	136.7	151.5	132.4	137.7	140.0	167.1
$h = 9$	140.0	142.5	139.5	160.0	142.6	147.5	160.2	184.7
$h = 10$	148.3	160.0	152.4	183.4	151.3	160.1	180.3	204.1
$h = 11$	162.7	168.3	168.4	193.5	160.1	170.5	188.5	211.1
$h = 12$	166.3	174.9	175.2	201.6	162.5	174.9	181.5	222.8

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Tab. 64: Out-of-Sample-Prognosegüte des Bayesianischen Ansatzes, direkte Prognosen, mit Jan1975 als Start des In-Samples, IPI

	Bayes1	Bayes2	Bayes3	Bayes4	Bayes5	Bayes6	Bayes7	Bayes8
saisonale Differenzen, $\nabla_{12} Y_t$								
$h = 1$	110.4	110.5	113.9	110.4	110.4	110.4	111.5	116.5
$h = 2$	118.8	118.2	124.5	139.0	130.6	119.0	130.2	132.4
$h = 3$	102.6	104.5	106.1	106.2	102.2	104.7	110.3	120.3
$h = 4$	129.9	127.1	128.4	183.9	127.6	124.4	127.5	201.0
$h = 5$	137.2	139.7	145.3	167.6	144.1	149.4	146.8	154.7
$h = 6$	133.9	137.4	125.8	144.3	130.2	127.1	137.3	147.1
$h = 7$	142.7	140.7	129.5	164.7	128.6	127.7	137.3	171.7
$h = 8$	132.7	147.7	151.9	176.4	135.4	140.4	154.2	159.7
$h = 9$	136.3	135.5	144.1	149.4	133.5	128.5	139.4	153.9
$h = 10$	135.7	141.2	175.4	170.7	135.1	145.8	171.3	172.9
$h = 11$	144.2	149.9	159.7	219.4	157.7	147.3	171.9	202.1
$h = 12$	146.2	149.8	150.1	146.0	149.8	153.2	152.5	158.1
gewöhnliche Differenzen, ∇Y_t								
$h = 1$	113.2	119.3	120.9	126.4	115.5	111.8	144.4	120.3
$h = 2$	117.7	112.0	111.6	130.0	123.9	109.3	121.6	144.4
$h = 3$	132.6	131.0	131.3	127.7	120.1	133.7	134.6	111.7
$h = 4$	128.2	128.8	148.6	166.6	131.5	134.5	155.9	198.7
$h = 5$	132.8	150.5	158.6	142.0	137.1	151.6	156.8	149.8
$h = 6$	124.3	125.7	139.3	146.4	129.9	128.1	149.6	132.8
$h = 7$	130.1	136.5	146.4	147.8	122.7	136.5	130.6	142.0
$h = 8$	139.3	136.8	136.1	135.8	134.8	135.2	140.3	138.3
$h = 9$	148.9	153.9	156.1	152.9	163.5	162.4	160.5	166.5
$h = 10$	128.9	124.2	123.1	175.4	122.9	152.8	154.6	155.6
$h = 11$	116.2	116.2	124.3	142.4	123.3	132.1	146.6	135.2
$h = 12$	120.6	145.8	117.6	115.2	124.7	120.0	118.7	115.2

Die Zahlen bezeichnen den MSFE in Prozent des MSFE der Linearen Basisprognose, LBP

Verzeichnis der Abkürzungen

ACF	Autokorrelationsfunktion
ADF	Augmented Dickey-Fuller
AIC	Akaike-Informationskriterium
ALR	Arbeitslosenrate
AR	Autoregressive
ARCH	Autoregressive Conditional Heteroskedasticity
ARDS	Autoregressiv mit deterministischer Saisonalität
ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
ARMADS	Autoregressive Moving Average mit deterministischer Saisonalität
ARNN	Autoregressives neuronales Netz
ARNNDS	Autoregressives neuronales Netz mit deterministischer Saisonalität
BCA	Business Cycle Assymetry
BDS	Brock-Dechert-Scheinkman
BEA	Bayesianischer Evidenzansatz
BFGS	Broyden-Fletcher-Goldfarb-Shanno
BL	bilinear
BP	Backpropagation
CV	Cross-Validation
DHF	Dickey-Hasza-Fuller
DGP	datengenerierender Prozess
EMSFE	Expected Mean Square Forecasting Error
ES	Early-Stopping
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
GD	Gradient Descent
HEGY	Hylleberg-Engle-Granger-Yoo
IPI	Industrieproduktionsindex
KPSS	Kwiatkowski-Phillips-Schmidt-Shin
KQ	Kleinstquadrat
LBM	lineares Basismodell
LBP	lineare Basisprognose
MA	Moving Average
ML	Maximum-Likelihood
MSE	Mean Square Error
MSFE	Mean Square Forecasting Error
NIC	Netzwerk-Informationskriterium
NKQ	nicht-lineare Kleinstquadratmethode
NLAR	nicht-linear autoregressiv
NN	neuronales Netz
NNUTS	Neural Networks for Univariate Time Series

PACF	Partielle Autokorrelationsfunktion
PP	Phillips-Perron
QN	Quasi-Newton
SA	Simulated Annealing
SARMA	Seasonal Autoregressive Moving Average
SIC	Schwarz-Informationskriterium
SIM	simulierte Zeitreihe
SSE	Sum of Squared Errors
TAR	Threshold Autoregressive
TLG	Teräsvirta-Lin-Granger
WD	Weight Decay

Verzeichnis der mathematischen Notation

Die in der folgenden Auflistung gegebenen Erläuterungen betreffen jeweils auch die mit Superskripten versehenen Symbole, $\tilde{\cdot}$, $\bar{\cdot}$, $\hat{\cdot}$, $\dot{\cdot}$, $\cdot^*$, $\cdot^{(\tau)}$, die (in dieser Reihenfolge) eine Modifikation des ursprünglichen Terms, Durchschnittsbildung oder Erwartungswertbildung, empirische Schätzung, Vorläufigkeit oder Ziehung aus einer Simulation, Hervorhebung oder Optimierung sowie Abhängigkeit vom Iterationszeitpunkt zum Ausdruck bringen. Im allgemeinen gelten die Erläuterungen auch für die entsprechenden mit Subskripten versehenen Symbole. Symbole für Zeitreihen, zu erkennen am Subskript t, sind eigens angeführt. Fett- und Aufrechtschreibung bezeichnet Vektoren und Matrizen. Für Notation in lateinischer Schrift können fett- und aufrechtgeschriebene Symbole eine separate Bedeutung aufweisen, weshalb für sie in der Auflistung eine eigene Abteilung zusammengestellt wurde, nicht so jedoch für griechische Symbole.

a, b, c	Koeffizienten im NN-Teil des ARNN-Modells
d, d^t, d^s	Koeffizienten im deterministischen Teil des ARNN-Modells
d	auch: Ordnung des Differenzierens (z.B. im ARIMA-Modell)
e_t	Prognosefehler
f	Koeffizienten im linearen Teil des ARNN-Modells
$f(\cdot)$, $g(\cdot)$	Funktionen
g	im Kontext wechselnde Verwendungen
h	Prognosehorizont
i	Verwendung als Laufindex; auch: die imaginäre Zahl i
j, k, l	Verwendung als Laufindex
m, n	im Kontext wechselnde Verwendungen
p, q	(Teil-)Modellordnung, Anzahl der Parameter
$p(\cdot)$	Wahrscheinlichkeitsdichtefunktion
r	im Kontext wechselnde Verwendungen
s	Anzahl Saisonen
s_{it}	saisonale Dummyvariable für Saison i
t	Zeitindex, Trendvariable
u	Gewicht des Parametervektors in Bezug auf einen Eigenvektor
u_t, v_t	Residuenzeitreihe, abgeleitete Zeitreihe
w	einzelner Modellparameter
x, y	zu modellierende Variable, Zufallsvariable
x_t, y_t	zu modellierende Zeitreihe
z	Zufallsvariable (aus Verteilung gezogen)
z_t	Zeitreihe, in wechselndem Kontext
B, B_t	Backshift-Operator bzw. Backshift-Operator in Bezug auf Zeitindex t
D	Ordnung des saisonalen Differenzierens (z.B. im SARIMA-Modell)
C	im Kontext wechselnde Verwendungen
$E(\cdot)$, E	Fehlerfunktion bzw. Wert der Fehlerfunktion

G, H	Hilfsterme für Ableitungen der ARNN-Fehlerfunktion
H_0, H_A	Null-Hypothese bzw. Alternativ-Hypothese
$I(i)$	integrierte Zeitreihenvariable der Ordnung i
J	Anzahl verzögerter Variablen im NN-Teil des ARNN-Modells
K	Anzahl verdeckter Neuronen
L	Anzahl verzögerter Variablen im linearen Teil des ARNN-Modells
$L(\cdot)$	Likelihood-Funktion
N	Stichprobenumfang, Länge einer Zeitreihe
M, P, Q	im Kontext wechselnde Verwendungen
R, S	im Kontext wechselnde Verwendungen
T	Zeitreihenende, Prognoseursprung; auch: Länge der Zeitreihe
V	im Kontext wechselnde Verwendungen
Y_t	Originalzeitreihe (vor Anwendung des Differenzenfilters)
W	Anzahl der Parameter (im Bayesianischen Evidenzansatz)
Z	im Kontext wechselnde Verwendung
$\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$	Koeffizienten im NN-Teil des ARNN-Modells (als Vektoren)
$\mathbf{f}$,	Koeffizienten im linearen Teil des ARNN-Modells (als Vektor)
$\mathbf{g}$	Gradient
$\mathbf{p}$, $\mathbf{q}$	Vektoren aus ausgewählten Lags
$\mathbf{r}$	Vektor von Parameterrestriktionen
$\mathbf{s}$, $\mathbf{u}$, $\mathbf{v}$	Zwischengrößen im Rahmen des Quasi-Newton-Verfahrens (QN)
$\mathbf{u}$	auch: Eigenvektor
$\mathbf{w}$	Parametervektor
$\mathbf{x}$	Vektor der unabhängigen Variablen
$\mathbf{x}_t^m$	Vektor von m eingebetteten Zeitreihen
$\mathbf{y}_t$	Vektor von verzögerten Zeitreihen
$\mathbf{A}$	Hesse-Matrix in Bezug auf die regularisierte Fehlerfunktion
$\mathbf{C}$	Korrelationsmatrix
$\mathbf{G}$	Annäherung der inversen Hesse-Matrix im Rahmen des QN-Verfahrens
$\mathbf{I}$	Einheitsmatrix
$\mathbf{H}$	Hesse-Matrix
$\mathbf{R}$	Matrix von Parameterrestriktionen
$\mathbf{S}$, $\mathbf{V}$	Zwischengrößen im Rahmen des QN-Verfahrens
$\mathrm{Cov}(\cdot)$	Kovarianz
$\mathrm{E}(\cdot)$	Erwartungswert
$\mathrm{I}(\cdot)$	Indikatorfunktion
$\mathrm{N}(\cdot)$	Normalverteilung
$\mathrm{Var}(\cdot)$	Varianz
$\mathcal{D}$	Daten, d.h. die in den Daten enthaltenen Informationen
$\mathbb{N}$	Menge der natürlichen Zahlen
$\mathcal{I}$	Epoche, d.h. die in den Daten der Epoche enthaltenen Informationen
$\mathcal{J}$, $\mathcal{L}$	Menge der Lags im NN-Teil bzw. im linearen Teil des ARNN-Modells

$\mathcal{Q}$	Menge der Lags zum zusätzlichen verdeckten Neuron im TLG-Test
$\mathcal{S}$	Suchraum
α, β	Koeffizienten verschiedener Modelle und Hilfsregressionen
α, β	auch: Hyperparameter im Bayesianischen Evidenzansatz (BEA)
γ, δ	Koeffizienten verschiedener Modelle und Hilfsregressionen
γ	auch: Anzahl wohldeterminierter Parameter im BEA
ϵ	Toleranzgröße
ε_t	Fehler- bzw. Innovationsprozess
ζ	Momentumrate
η	Schrittweite im interativen Verfahren, Lernrate
θ	Koeffizienten im MA-Modell
$\theta(B)$	Lag-Polynom
ι, κ	Laufindizes (nur für Iterationsschritte verwendet)
κ	auch: (optimierte) Schrittweite in iterativen Verfahren
λ	Eigenwert
μ	Mittelwert; auch: Parameter der linearen Transformation
μ_t	Mittelwertprozess oder deterministischer Prozess
ν	Regularisierungsparameter
ξ	heuristische Konstanten (mit wechselndem Kontext)
π	die Zahl Pi
π	auch: Epochengröße im Backpropagation-Lernalgorithmus
ρ	Autokorrelationskoeffizient und verwandte Statistiken
σ, σ^2	Standardabweichung bzw. Varianz
σ	auch: Parameter der linearen Transformation
τ	Iterationsschritt; auch: kontinuierliche Verfahrenszeit
ϕ	Koeffizient im AR-Modell und in linearen Modellen
$\phi(B)$	Lag-Polynom
ϕ_{kk}	partieller Autokorrelationskoeffizient
$\varphi(\cdot)$	nicht-lineares Modell
χ	wechselnde Verwendungen
$\psi(\cdot)$	Aktivierungsfunktion eines neuronalen Netzes
ω	wechselnde Verwendungen
Δ	Veränderung (zwischen zwei Iterationsschritten)
$\Theta(B^s)$	saisonales Lag-Polynom
$\Phi(B^s)$	saisonales Lag-Polynom
Ω	Modellkomplexität
∇,	Differenzenoperator; auch: Gradientenoperator
∇_k^m	Differenzenoperator der Ordnung m zum Lag k
∇_s	saisonaler Differenzenoperator

Literatur

Acemoglu, D. und Scott, A. (1994). Asymmetries in the cyclical behaviour of UK labour markets. *The Economic Journal*, **104**(427), 1303–1323.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. und Csaki, F. (eds.), *Second International Symposium on Information Theory*, pp. 267–281. Budapest: Akademiai Kiado.

Anders, U. (1997). *Statistische Neuronale Netze*. München: Vahlen.

Anders, U. und Korn, O. (1999). Model selection in neural networks. *Neural Networks*, **12**(2), 309–323.

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica*, **61**(4), 821–856.

Balkin, S. D. und Ord, J. K. (2000). Automatic neural network modeling for univariate time series. *International Journal of Forecasting*, **16**(4), 509–515.

Banerjee, A., Dolado, J. J., Galbraith, J. W. und Hendry, D. F. (1993). *Cointegration, Error Correction, and the Econometric Analysis of Non-Stationary Data*. Oxford: Oxford University Press.

Barnett, W. A., Medio, A. und Serletis, A. (1997). Nonlinear and complex dynamics in economics. Washington University St. Luis, Working Paper ewp-em/9709001.

Beaulieu, J. J. und Miron, J. A. (1993). Seasonal unit roots in aggregate U.S. data. *Journal of Econometrics*, **55**(1-2), 305–328.

Bélisle, C. J. P. (1992). Convergence theorems for a class of simulated annealing algorithms on R^d. *Journal of Applied Probability*, **29**(4), 885–895.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Boender, C. (1984). The generalized multinomial distribution: A Bayesian analysis and applications. Ph.D. Dissertation, Erasmus Universiteit Rotterdam.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, **31**(3), 307–327.

Box, G. E. und Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

Brock, W. A. (2000). Whither nonlinear? *Journal of Economic Dynamics and Control*, **24**(5–7), 663–678.

Brock, W. A., Hsieh, D. A. und LeBaron, B. (1991). *Nonlinear dynamics, chaos, and instability: Statistical theory and economic evidence*. Cambridge, Mass. and London: MIT Press.

Brock, W. A. und Sayers, C. L. (1988). Is the business cycle characterized by deterministic chaos? *Journal of Monetary Economics*, **22**(1), 71–90.

Brock, W. A., Scheinkman, J. A. und Dechert, W. D. (1987). A test for independence based on the correlation dimension. Department of Economics, University of Wisconsin, Economics Working Paper SSR I-8702.

Brock, W. A., Scheinkman, J. A., Dechert, W. D. und LeBaron, B. (1996). A test for independence based on the correlation dimension. *Econometric Reviews*, **15**(3), 197–235.

Brockwell, P. J. und Davis, R. A. (1991). *Time Series: Theory and Methods*. New York: Springer.

Burgess, S. M. (1992). Asymmetric employment cycles in Britain: Evidence and an explanation. *Economic Journal*, **102**(411), 279–290.

Burns, A. F. und Mitchell, W. C. (1946). *Measuring Business Cycles*. New York: Columbia University Press.

Caballero, R. J. und Hammour, M. L. (1994). The cleansing effect of recessions. *American Economic Review*, **84**(5), 181–186.

Caruana, R., Lawrence, S. und Giles, L. (2001). Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In Leen, T. K., Dietterich, T. G. und Tresp, V. (eds.), *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO*, pp. 402–408. MIT Press.

Chan, K. S. und Tong, H. (1985). On the use of the deterministic Lyapunov function for the ergodicity of stochastic difference equations. *Advances in Applied Probability*, **17**(3), 666–678.

Cheng, B. und Titterington, D. M. (1994). Neural networks: A review from a statistical perspective. *Statistical Science*, **9**(1), 2–30.

Chow, G. C. (1983). *Econometrics*. New York: McGraw-Hill.

Clark, T. E. (1999). Finite-sample properties of tests for equal forecast accuracy. *Journal of Forecasting*, **18**(7), 489–504.

Clements, M. P. und Hendry, D. F. (1997). An empirical study of seasonal unit roots in forecasting. *International Journal of Forecasting*, **13**(3), 341–355.

Clements, M. P. und Hendry, D. F. (1998). *Forecasting Economic Time Series*. Cambridge: Cambridge University Press.

Cottrell, M., Girard, B., Girard, Y., Mageas, M. und Muller, C. (1995). Neural modeling for time series: A statistical stepwise method for weight elimination. *IEEE Transactions on Neural Networks*, **6**(6), 1355–1364.

Cromwell, J. B., Labys, W. C. und Terraza, M. (1994). *Univariate Tests for Time Series Models*. Thousand Oaks, CA: SAGE Publications.

Csendes, T. (1988). Nonlinear parameter estimation by global optimization – efficiency and reliability. *Acta Cybernetica*, **8**(4), 361–370.

Darbellay, G. A. und Slama, M. (2000). Forecasting the short-term demand for electricity. Do neural networks stand a better chance? *International Journal of the Forecasting*, **16**(1), 71–83.

Davidon, W. C. (1991). Variable metric method for minimization. *Siam Journal on Optimization*, **1**(1), 1–17.

Davidson, R. und MacKinnon, R. G. (1993). *Estimation and Inference in Econometrics*. Oxford: Oxford University Press.

De Gooijer, J. G. und Kumar, K. (1992). Some recent developments in non-linear time series modelling, testing, and forecasting. *International Journal of Forecasting*, **8**(2), 263–276.

Dickey, D. A. und Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, **74**(366), 427–431.

Dickey, D. A., Hasza, D. P. und Fuller, W. A. (1984). Testing for unit roots in seasonal time series. *Journal of the American Statistical Association*, **79**(386), 355–367.

Diebold, F. X. und Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economics Statistics*, **13**(3), 253–263.

Enders, W. (1995). *Applied Econometric Time Series*. New York: Wiley.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. *Econometrica*, **50**(4), 987–1007.

Engle, R. F., Lilien, D. M. und Robins, R. P. (1987). Estimating time varying risk premia in the term structure: The ARCH-M model. *Econometrica*, **55**(2), 391–407.

Fahlmann, S. E. (1989). Faster learning variations on back-propagation: An empirical study. In Touretzky, D., Hinton, G. und Sejnowski, T. (eds.), *Proceedings of the 1988 Connectionist Models Summer School*, pp. 38–51, San Mateo, CA. Morgan Kaufmann.

Faraway, J. und Chatfield, C. (1998). Time series forecasting with neural networks: A comparative study using the airline data. *Applied Statistics*, **47**(2), 231–250.

Fine, T. L. (1999). *Feedforward Neural Network Methodology*. Berlin: Springer Verlag.

Finnoff, W., Hergert, F. und Zimmermann, H. G. (1993). Improving model selection by nonconvergent methods. *Neural Networks*, **6**(6), 771–783.

Fischer, M. M. (2002). Learning in neural spatial interaction models: A statistical perspective. *Journal of Geographical Systems*, **4**(3), 287–299.

Fischer, M. M. (2006). Neural networks. a general framework for non-linear function approximation. *Transactions in GIS*, **10**(4), 521–533.

Fornari, F. und Mele, A. (1997). Asymmetries and non-linearities in economic activity. *Applied Financial Economics*, **7**(2), 203–206.

Franses, P. H. (1991). Seasonality, non-stationarity and the forecasting of monthly time series. *International Journal of Forecasting*, **7**(2), 199–208.

Franses, P. H. (1996a). *Periodicity and Stochastic Trends in Economic Time Series*. Oxford: Oxford University Press.

Franses, P. H. (1996b). Recent advances in modelling seasonality. *Journal of Economic Surveys*, **10**(3), 299–345.

Franses, P. H. und De Bruin, P. (2000). Seasonal adjustment and the business cycle in unemployment. *Studies in Nonlinear Dynamics and Econometrics*, **4**(2), 73–84.

Franses, P. H. und van Dijk, D. (2000). *Non-linear time series models in empirical finance*. Cambridge: Cambridge University Press.

Franses, P. H. und Draisma, G. (1997). Recognizing changing seasonal patterns using artificial neural networks. *Journal of Econometrics*, **81**(1), 273–280.

Franses, P. H. und van Griensven, K. (1997). Forecasting exchange rates using neural networks for technical trading rules. *Studies in Nonlinear Dynamics and Econometrics*, **2**(4), 109–116.

Gallant, A. R. und White, H. (1988). *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Oxford: Basil Blackwell.

Gallegati, M. und Mignacca, D. (1995). Nonlinearities in business cycle: SETAR models and G7 industrial production data. *Applied Economics Letters*, **2**(11), 287–294.

Geman, S., Bienenstock, E. und Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, **4**(1), 1–58.

Granger, C. und Lin, J.-L. (1994). Using the mutual information coefficients to identify lags in nonlinear models. *Journal of Time Series Analysis*, **15**(4), 371–384.

Granger, C. W. (1991). Developments in the nonlinear analysis of economic series. *Scandinavian Journal of Economics*, **93**(2), 263–276.

Granger, C. W. (1999). Aspects of research strategies for time series analysis. Presentation to the conference on New Developments in Time Series Economics, October 1999, Cowles Foundation for Research in Economics at Yale University, New Haven.

Granger, C. W. und Andersen, A. (1978). *An Introduction to Bilinear Time Series Models*. Göttingen: Vandenhoeck & Ruprecht.

Granger, C. W. und Newbold, P. (1986). *Forecasting Economic Time Series*. London: Academic Press.

Granger, C. W. und Teräsvirta, T. (1993). *Modelling Nonlinear Economic Relationships*. Oxford: Oxford University Press.

Granger, C. W. J. (1993). Comments on 'On the limitations of comparing mean squared forecast errors', by M. P. Clements and D. F. Hendry. *Journal of Forecasting*, **12**(8), 651–652.

Grassberger, P. und Procaccia, I. (1983). Measuring the strangeness of strange attractors. *Physica D*, **9**(1–2), 189–208.

Gull, S. F. (1988). Bayesian inductive inference and maximum entropy. In Smith, C. R. (ed.), *Maximum-Entropy and Bayesian Methods in Science and Engineering, Vol. 1: Foundations*, pp. 53–74. Dordrecht: Kluwer.

Gull, S. F. (1989). Developments in maximum entropy data analysis. In Skilling, J. (ed.), *Maximum-Entropy and Bayesian Methods, Cambridge 1988*, pp. 53–71. Dordrecht: Kluwer.

Hamermesh, D. S. und Pfann, G. A. (1996). Adjustment costs in factor demand. *Journal of Economic Literature*, **34**(3), 1264–1292.

Harvey, A. C. (1993). *Time Series Models.* New York: Harvester Whearsheaf.

Harvey, D., Leybourne, S. und Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, **13**(2), 281–291.

Haykin, S. (1994). *Neural Networks. A Comprehensive Foundation.* New York: Macmillan.

Hecht-Nielsen, R. (1989). Theory of the back-propagation network. In *Proceedings of the International Joint Conference on Neural Networks, Washington, DC.*, vol. 1, pp. 893–606, New York. IEEE Press.

Heravi, S., Osborn, D. R. und Birchenhall, C. R. (2004). Linear versus neural network forecasts for European industrian production series. *International Journal of the Forecasting*, **20**(3), 435–446.

Hertz, J. A., Krogh, A. und Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation.* Redwood City, CA: Addison-Wesley.

Hippert, H. S., Bunn, D. W. und Souza, R. C. (2005). Large neural networks for electricity load forecasting: Are they overfitted? *International Journal of Forecasting*, **21**(3), 425–434.

Hippert, H. S., Pedreira, C. E. und Souza, R. C. (2001). Neural networks for short term load forecasting: a review and evaluation. *IEEE Transactions on Power Systems*, **16**(1), 44–55.

Hornik, K., Stinchcombe, M. und White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, **2**(5), 359–366.

Hornik, K., Stinchcombe, M. und White, H. (1990). Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*, **3**(5), 551–560.

Hsieh, D. A. (1989). Testing for nonlinear dependence in daily foreign exchange rates. *Journal of Business*, **62**(3), 339–368.

Hylleberg, S. (1992). General introduction. In Hylleberg, S. (ed.), *Modelling Seasonality*, pp. 3–14. Oxford: Oxford University Press.

Hylleberg, S., Engle, R., Granger, C. W. J. und Yoo, B. S. (1990). Seasonal integration and cointegration. *Journal of Econometrics*, **44**(1-2), 215–238.

Hylleberg, S., Jorgensen, C. und Sorensen, N. K. (1993). Seasonality in macroeconomic time series. *Empirical Economics*, **18**(2), 321–335.

Hyndman, R. J. und Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, **22**(4), 679–688.

Jacobs, R. A. (1988). Increased rates of convergence through learning rate adaptation. *Neural Networks*, **1**(4), 295–307.

Kajitani, Y., Hipel, K. W. und Mcleod, A. I. (2005). Forecasting nonlinear time series with feed-forward neural networks: A case study of Canadian lynx data. *Journal of Forecasting*, **24**(2), 105–117.

Keuzenkamp, H. A. (1995). The econometrics of the holy grail — A review of "Econometrics: Alchemy or Science? Essays in Econometric Methodology" by D. Hendry. *Journal of Economic Surveys*, **9**(2), 233–248.

Koller, W. und Fischer, M. M. (2001). Testing for non-linear dependence in univariate time series. An empirical investigation to the Austrian unemployment rate. Vortrag an der 41. Konferenz der European Regional Science Association, Zagreb, 29.8.–1.9.2001.

Koller, W. und Fischer, M. M. (2002). Testing for non-linear dependence in univariate time series. An empirical investigation of the Austrian unemployment rate. *Networks and Spatial Economics*, **2**(2), 191–209.

Koop, G. und Potter, S. (2001). Are apparent findings of nonlinearity due to structural instability in economic time series? *Econometrics Journal*, **4**(1), 37–55.

Kuan, C.-M. und White, H. (1994). Artificial neural networks: An econometric perspective. *Econometric Reviews*, **13**(4), 1–91.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P. und Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. How sure are we that economic time series have a unit root? *Journal of Econometrics*, **54**(1-3), 159–178.

Layard, R., Nickell, S. und Jackmann, R. (1991). *Unemployment. Macroeconomic Performance and the Labour Market*. Oxford: Oxford University Press.

Lee, T.-H., White, H. und Granger, C. W. (1993). Testing for neglected nonlinearity in time series models. *Journal of Econometrics*, **56**(3), 269–290.

Leisch, F., Trapletti, A. und Hornik, K. (1999). Stationarity and stability of autoregressive neural network processes. In Kearns, M. S., Solla, S. A. und Cohn, D. A. (eds.), *Advances in Neural Information Processing Systems 11, Papers from Neural Information Processing Systems (NIPS) 1988, Denver*, vol. 11, pp. 267–273. MIT Press.

Lin, J.-L. und Granger, C. W. (1994). Forecasting from non-linear models in practice. *Journal of Forecasting*, **13**(1), 1–9.

Luenberger, D. G. (1984). *Linear and Nonlinear Programming*. New York: Springer.

MacKay, D. J. C. (1992a). Bayesian interpolation. *Neural Computation*, **4**(3), 415–447.

MacKay, D. J. C. (1992b). A practical Bayesian framework for backpropagation networks. *Neural Computation*, **4**(3), 448–472.

MacKay, D. J. C. (1995). Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, **6**(3), 469–505.

MacKay, D. J. C. (1996). Hyperparameters: optimise or integrate out? In Heidbreder, G. (ed.), *Maximum Entropy and Bayesian Methods, Santa Barbara 1993*, pp. 43–59. Dordrecht: Kluwer.

Marcellino, M., Stock, J. H. und Watson, M. W. (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics*, **135**(1–2), 499–526.

Marti, R., Moreno-Vega, J. und Duarte, A. (2010). Advanced multi-start methods. In Gendreau, M. und Potvin, J.-Y. (eds.), *Handbook of Metaheuristics. 2nd Edition*, pp. 265–282. New York: Springer.

McLeod, A. und Li, W. (1983). Diagnostic checking of ARMA time series models using squared residuals autocorrelations. *Journal of Time Series Analysis*, **4**(4), 269–273.

Medeiros, M. C., Teräsvirta, T. und Rech, G. (2006). Building neural network models for time series: A statistical approach. *Journal of Forecasting*, **25**(1), 49–75.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. und Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**(4), 1087–1092.

Mills, T. C. (1990). *Time Series Techniques for Economists*. Cambridge: Cambridge University Press.

Mitchell, W. C. (1927). *Business Cycles: The Problem and Its Setting*. New York: National Bureau of Economic Research.

Montgomery, A. L., Zarnowitz, V., Tsay, R. S. und Tiao, G. C. (1998). Forecasting the U.S. unemployment rate. *Journal of the American Statistical Association*, **93**(442), 478–493.

Moshiri, S. und Cameron, N. (2000). Neural network versus econometric models in forecasting inflation. *Journal of Forecasting*, **19**(1), 201–217.

Murata, N., Yoshizawa, S. und Amari, S.-I. (1994). Network information criterion—determining the number of hidden units for an artificial neural network model. *Neural Networks*, **5**(6), 865–872.

Neftci, S. (1984). Are economic time series asymmetric over the business cycle? *Journal of Political Economy*, **92**(2), 307–328.

Nelson, C. R. und Plosser, C. I. (1982). Trends and random walks in macroeconomic time series: Some evidence and implications. *Journal of Monetary Economics*, **10**(2), 139–162.

Nelson, M., Hill, T., Remus, W. und O'Connor, M. (1999). Time series forecasting using neural networks: Should the data be deseasonalized first? *Journal of Forecasting*, **18**(5), 359–367.

Newbold, P., Agiakloglou, C. und Miller, J. (1994). Adventures with ARIMA software. *International Journal of Forecasting*, **10**(4), 573–581.

Nicholls, D. F. und Quinn, B. G. (1982). *Random Coefficient Autoregressive Models: An Introduction*. New York: Springer.

Osborn, D. R. (1990). A survey of seasonality in UK macroeconomic variables. *International Journal of Forecasting*, **6**(3), 327–336.

Peat, M. und Stevenson, M. (1996). Asymmetry in the business cycle: Evidence from the Australian labour market. *Journal of Economic Behavior and Organization*, **30**(3), 353–368.

Peel, D. A. und Speight, A. E. H. (1998a). Modelling business cycle nonlinearity in conditional mean and conditional variance: Some international and sectoral evidence. *Economica*, **65**(258), 211–229.

Peel, D. A. und Speight, A. E. H. (1998b). The non-linear time series properties of unemployment rates: some further evidence. *Applied Economics*, **30**(2), 287–294.

Peel, D. A. und Speight, A. E. H. (2000). Threshold nonlinearities in unemployment rates: further evidence for the UK and G3 economies. *Applied Economics*, **32**(6), 705–715.

Penny, W. D. und Roberts, S. J. (1999). Bayesian neural networks for classification: how useful is the evidence framework? *Neural Networks*, **12**(6), 877–892.

Perron, P. (1988). Trends and random walks in macroeconomic time series. *Journal of Economic Dynamics and Control*, **12**(2-3), 297–332.

Phillips, P. C. und Xiao, Z. (1998). A primer on unit root testing. Cowles Foundation, Discussion Paper 1189.

Phillips, P. C. B. (1987). Time series regression with a unit root. *Econometrica*, **55**(2), 277–301.

Phillips, P. C. B. (1997). The ET interview: Professor Clive Granger. *Econometric Theory*, **13**(2), 253–303.

Phillips, P. C. B. und Perron, P. (1988). Testing for a unit root in time series regression. *Biometrica*, **75**(2), 335–346.

Polak, E. (1997). *Optimization. Algorithms and Consistent Approximations.* New York: Springer.

Potter, S. M. (1994). Asymmetric economic propagation mechanism. In Semmler, W. (ed.), *Business Cycles: Theory and Empirical Methods*, pp. 313–330. Boston: Kluwer.

Prechelt, L. (1998). Automatic early stopping using cross validation: Quantifying the criteria. *Neural Networks*, **11**(4), 761–767.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. und Flannery, B. P. (1992). *Numerical Recipes in C. The Art of Scientific Computing.* Cambridge: Cambridge University Press.

Pötscher, B. M. (1991). Effects of model selection on inference. *Econometric Theory*, **7**(2), 163–185.

Rech, G. (2002). Forecasting with artificial neural network models. Stockholm School of Economics, SSE/EFI Working Paper 491.

Refenes, A. N., Zapranis, A. und Franciss, G. (1994). Stock performance modeling using neural networks: A comparative study with regression models. *Neural Networks*, **7**(2), 375–388.

Rinnooy Kan, A. H. G. und Timmer, G. T. (1987a). Stochastic global optimization methods. Part i: Clustering methods. *Mathematical Programming*, **39**(1), 27–56.

Rinnooy Kan, A. H. G. und Timmer, G. T. (1987b). Stochastic global optimization methods. Part ii: Multi level methods. *Mathematical Programming*, **39**(1), 57–78.

Ripley, B. (1996). *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Rivals, I. und Personnaz, L. (1999). On cross validation for model selection. *Neural Computation*, **11**(4), 863–870.

Rojas, R. (1993). Backpropagation in general networks. In *Proceedings of the Joint Meeting of the AMS and MAA, San Antonio, 13–16 January*.

Rojas, R. (1995). *Theorie der neuronalen Netze. Eine systematische Einführung*. Berlin: Springer Verlag.

Rojas, R. (1996). *Neural Networks. A Systematic Introduction*. Berlin: Springer Verlag.

Rothman, P. (1998). Forecasting asymmetric unemployment rates. *Review of Economics and Statistics*, **80**(1), 164–168.

Rudebusch, G. D. (1993). The uncertain unit root in real GNP. *American Economic Review*, **83**(1), 264–272.

Said, S. E. und Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models. *Biometrica*, **71**(3), 599–607.

Schlittgen, R. und Streitberg, B. H. J. (1997). *Zeitreihenanalyse*. München: Oldenbourg.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**(2), 461–464.

Semmler, W. (ed.) (1994). *Business Cycles: Theory and Empirical Methods*. Boston: Kluwer.

Silva, F. M. und Almeida, L. B. (1990). Speeding-up backpropagation. In Eckmiller, R. (ed.), *Advanced Neural Computers*, pp. 151–156. Amsterdam: North-Holland.

Skalin, J. und Teräsvirta, T. (1999). Modelling asymmetries and moving equilibria in unemployment rates. Stockholm School of Economics, Working Paper 262.

Skilling, J. (1991). On parameter estimation and quantified maxent. In Skilling, J. (ed.), *Maximum Entropy and Bayesian Methods, Laramie 1990*, pp. 267–273. Dordrecht: Kluwer.

Sorensen, N. K. (1999). Modelling the seasonality of hotel nights in Denmark by county and nationality. *Tourism Economics*, **5**(1), 9–23.

Stanca, L. (1999). Asymmetries and non-linearities in economic activity. *Applied Economics*, **31**(4), 483–491.

Stock, J. H. und Watson, M. W. (1999). A comparison of linear and nonlinear univariate time series. In Engle, R. F. und White, H. (eds.), *Cointegration, Causality and Forecasting: A Festschrift in Honour of Clive W. J. Granger*, pp. 1–44. Oxford: Oxford University Press.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, **36**(2), 111–147.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society. Series B*, **39**(1), 44–47.

Swanson, N. R. und White, H. (1997a). Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models. *International Journal of Forecasting*, **13**(4), 439–461.

Swanson, N. R. und White, H. (1997b). A model selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks. *Review of Economics and Statistics*, **79**(4), 540–550.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, **16**(4), 437–450.

Teräsvirta, T., van Dijk, D. und Medeiros, M. C. (2005). Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic time series: A re-examination. *International Journal of Forecasting*, **21**(4), 755–774.

Teräsvirta, T., Lin, C.-F. und Granger, C. (1993). Power of the neural network linearity test. *Journal of Time Series Analysis*, **14**(2), 209–220.

Titterington, D. M. (2004). Bayesian methods for neural networks and related models. *Statistical Science*, **19**(1), 128–139.

Tong, H. (1990). *Non-linear Time Series*. Oxford: Clarendon Press.

Tong, H. und Lim, K. S. (1980). Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society, Series B*, **42**(3), 245–292.

Trapletti, A., Leisch, F. und Hornik, K. (2000). Stationary and integrated autoregressive neural network processes. *Neural Computation*, **12**(10), 2427–2450.

Tsay, R. (1986). Nonlinearity tests for time series. *Biometrika*, **73**(2), 461–466.

Wallis, K. F. (1987). Time series analysis of bounded economic variables. *Journal of Time Series Analysis*, **8**(1), 115–123.

Weigend, A. S., Huberman, B. A. und Rumelhart, D. E. (1990). Predicting the future: A connectionist approach. *International Journal of Neural Systems*, **1**(3), 193–209.

Weigend, A. S., Huberman, B. A. und Rumelhart, D. E. (1992). Predicting sunspots and exchange rates with connectionist networks. In Casdagli, M. und Eubank, S. (eds.), *Nonlinear Modeling and Forecasting*, pp. 395–432. Reading, MA: Addison-Wesley.

White, H. (1989a). An additional hidden unit test for neglected non-linearity in multilayer feedforward networks. In *Proceedings of the International Joint Conference on Neural Networks, Washington, DC*, vol. I, pp. 451–455, San Diego, CA. SOS Printing.

White, H. (1989b). Correction: Some asymptotic results for learning in single hidden-layer feedforward network models. *Journal of the American Statistical Association*, **87**(408), 1252.

White, H. (1989c). Learning in neural networks: A statistical perspective. *Neural Computation*, **1**(4), 425–464.

White, H. (1989d). Some asymptotic results for learning in single hidden-layer feedforward network models. *Journal of the American Statistical Association*, **84**(408), 1003–1013.

White, H. (1994). *Estimation, Inference and Specification Analysis*. Cambridge: Cambridge University Press.

White, H. (2000). A reality check for data snooping. *Econometrica*, **68**(5), 1097–1126.

Yalta, A. T. und Yalta, A. Y. (2010). Should economists use open source software for doing research? *Computational Economics*, **35**(4), 371–394.

Zhang, G., Patuwo, B. E. und Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of the Forecasting*, **14**(1), 35–62.

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, **50**(1–4), 159–175.

Zhang, G. P. und Qi, M. (2005). Neural network forecasting for seasonal and trend time series. *European Journal of Operational Research*, **160**(2), 501–514.

Zhang, X. (1994). Time series analysis and prediction by neural networks. *Optimization Methods and Software*, **4**(2), 151–170.

Forschungsergebnisse der WU Wirtschaftsuniversität Wien

Herausgeber: WU Wirtschaftsuniversität Wien –
vertreten durch Univ. Prof. Dr. Barbara Sporn

INFORMATION UND KONTAKT:

WU Wirtschaftsuniversität Wien
Department of Finance, Accounting and Statistics
Institute for Finance, Banking and Insurance
Welthandelsplatz 1, D 4, 1. OG, 1020 Wien
Tel.: 0043-1-313 36/4556
Fax: 0043-1-313 36/904556
valentine.wendling@wu.ac.at
www.wu.ac.at/finance

Band 1 Stefan Felder: Frequenzallokation in der Telekommunikation. Ökonomische Analyse der Vergabe von Frequenzen unter besonderer Berücksichtigung der UMTS-Auktionen. 2004.

Band 2 Thomas Haller: Marketing im liberalisierten Strommarkt. Kommunikation und Produktplanung im Privatkundenmarkt. 2005.

Band 3 Alexander Stremitzer: Agency Theory: Methodology, Analysis. A Structured Approach to Writing Contracts. 2005.

Band 4 Günther Sedlacek: Analyse der Studiendauer und des Studienabbruch-Risikos. Unter Verwendung der statistischen Methoden der Ereignisanalyse. 2004.

Band 5 Monika Knassmüller: Unternehmensleitbilder im Vergleich. Sinn- und Bedeutungsrahmen deutschsprachiger Unternehmensleitbilder – Versuch einer empirischen (Re-)Konstruktion. 2005.

Band 6 Matthias Fink: Erfolgsfaktor Selbstverpflichtung bei vertrauensbasierten Kooperationen. Mit einem empirischen Befund. 2005.

Band 7 Michael Gerhard Kraft: Ökonomie zwischen Wissenschaft und Ethik. Eine dogmenhistorische Untersuchung von Léon M.E. Walras bis Milton Friedman. 2005.

Band 8 Ingrid Zechmeister: Mental Health Care Financing in the Process of Change. Challenges and Approaches for Austria. 2005.

Band 9 Sarah Meisenberger: Strukturierte Organisationen und Wissen. 2005.

Band 10 Anne-Katrin Neyer: Multinational teams in the European Commission and the European Parliament. 2005.

Band 11 Birgit Trukeschitz: Im Dienst Sozialer Dienste. Ökonomische Analyse der Beschäftigung in sozialen Dienstleistungseinrichtungen des Nonprofit Sektors. 2006

Band 12 Marcus Kölling: Interkulturelles Wissensmanagement. Deutschland Ost und West. 2006.

Band 13 Ulrich Berger: The Economics of Two-way Interconnection. 2006.

Band 14 Susanne Guth: Interoperability of DRM Systems. Exchanging and Processing XML-based Rights Expressions. 2006.

Band 15 Bernhard Klement: Ökonomische Kriterien und Anreizmechanismen für eine effiziente Förderung von industrieller Forschung und Innovation. Mit einer empirischen Quantifizierung der Hebeleffekte von F&E-Förderinstrumenten in Österreich. 2006.

Band 16 Markus Imgrund: Wege aus der Insolvenz. Eine Analyse der Fortführung und Sanierung insolventer Klein- und Mittelbetriebe unter besonderer Berücksichtigung des Konfigurationsansatzes. 2007.

Band 17 Nicolas Knotzer: Product Recommendations in E-Commerce Retailing Applications. 2008.

Band 18 Astrid Dickinger: Perceived Quality of Mobile Services. A Segment-Specific Analysis. 2007.

Band 19 Nadine Wiedermann-Ondrej: Hybride Finanzierungsinstrumente in der nationalen und internationalen Besteuerung der USA. 2008.

Band 20 Helmut Sorger: Entscheidungsorientiertes Risikomanagement in der Industrieunternehmung. 2008.

Band 21 Martin Rietsch: Messung und Analyse des ökonomischen Wechselkursrisikos aus Unternehmenssicht: Ein stochastischer Simulationsansatz. 2008.

Band 22 Hans Christian Mantler: Makroökonomische Effizienz des Finanzsektors. Herleitung eines theoretischen Modells und Schätzung der Wachstumsimplikationen für die Marktwirtschaften und Transformationsökonomien Europas. 2008.

Band 23 Youri Tacoun: La théorie de la valeur de Christian von Ehrenfels. 2008.

Band 24 Monika Koller: Longitudinale Betrachtung der Kognitiven Dissonanz. Eine Tagebuchstudie zur Reiseentscheidung. 2008.

Band 25 Marcus Scheiblecker: The Austrian Business Cycle in the European Context. 2008.

Band 26 Aida Numic: Multinational Teams in European and American Companies. 2008.

Band 27 Ulrike Bauernfeind: User Satisfaction with Personalised Internet Applications. 2008.

Band 28 Reinhold Schodl: Systematische Analyse und Bewertung komplexer Supply Chain Prozesse bei dynamischer Festlegung des Auftragsentkopplungspunkts. 2008.

Band 29 Bianca Gusenbauer: Öffentlich-private Finanzierung von Infrastruktur in Entwicklungsländern und deren Beitrag zur Armutsreduktion. Fallstudien in Vietnam und auf den Philippinen. 2009.

Band 30 Elisabeth Salomon: Hybrides Management in sino-österreichischen Joint Ventures in China aus österreichischer Perspektive. 2009.

Band 31 Katharina Mader: Gender Budgeting: Ein emanzipatorisches, finanzpolitisches und demokratiepolitisches Instrument. 2009.

Band 32 Michael Weber: Die Generierung von Empfehlungen für zwischenbetriebliche Transaktionen als gesamtwirtschaftliche Infrastrukturleistung. 2010.

Band 33 Lisa Gimpl-Heersink: Joint Pricing and Inventory Control under Reference Price Effects. 2009.

Band 34 Erscheint nicht.

Band 35 Dagmar Kiefer: Multicultural Work in Five United Nations Organisations. An Austrian Perspective. 2009.

Band 36 Gottfried Gruber: Multichannel Management. A Normative Model Towards Optimality. 2009.

Band 37 Rainer Quante: Management of Stochastic Demand in Make-to-Stock Manufacturing. 2009.

Band 38 Franz F. Eiffe: Auf den Spuren von Amartya Sen. Zur theoriegeschichtlichen Genese des Capability-Ansatzes und seinem Beitrag zur Armutsanalyse in der EU. 2010.

Band 39 Astrid Haider: Die Lohnhöhe und Lohnstreuung im Nonprofit-Sektor. Eine quantitative Analyse anhand österreichischer Arbeitnehmer-Arbeitgeber-Daten. 2010.

Band 40 Maureen Lenhart: Pflegekräftemigration nach Österreich. Eine empirische Analyse. 2010.

Band 41 Oliver Schwank: Linkages in South African Economic Development. Industrialisation without Diversification? 2010.

Band 42 Judith Kast-Aigner: A Corpus-Based Analysis of the Terminology of the European Union's Development Cooperation Policy, with the African, Caribbean and Pacific Group of States. 2010.

Band 43 Emel Arikan: Single Period Inventory Control and Pricing. An Empirical and Analytical Study of a Generalized Model. 2011.

Band 44 Gerhard Wohlgenannt: Learning Ontology Relations by Combining Corpus-Based Techniques and Reasoning on Data from Semantic Web Sources. 2011.

Band 45 Thomas Peschta: Der Einfluss von Kundenzufriedenheit auf die Kundenloyalität und die Wirkung der Wettbewerbsintensität am Beispiel der Gemeinschaftsverpflegungsgastronomie. 2011.

Band 46 Friederike Hehle: Die Anwendung des Convenience-Konzepts auf den Betriebstyp Vending. 2011.

Band 47 Thomas Herzog: Strategisches Management von Koopetition. Eine empirisch begründete Theorie im industriellen Kontext der zivilen Luftfahrt. 2011.

Band 48 Christian Weismayer: Statische und longitudinale Zufriedenheitsmessung. 2011.

Band 49 Johannes Fichtinger: The Single-Period Inventory Model with Spectral Risk Measures. 2011.

Band 50 Isabella R. Hatak: Kompetenz, Vertrauen und Kooperation. Eine experimentelle Studie. 2011.

Band 51 Birgit Gusenbauer: Der Beitrag der Prospect Theory zur Beschreibung und Erklärung von Servicequalitätsurteilen und Kundenzufriedenheit im Kontext von Versicherungsentscheidungen. 2012.

Band 52 Markus A. Höllerer: Between Creed, Rhetoric Façade, and Disregard. Dissemination and Theorization of Corporate Social Responsibility in Austria. 2012.

Band 53 Jakob Müllner: Die Wirkung von Private Equity auf das Wachstum und die Internationalisierung. Eine empirische Impact-Studie des österreichischen Private Equity Marktes. 2012.

Band 54 Heidrun Rosič: The Economic and Environmental Sustainability of Dual Sourcing. 2012.

Band 55 Christian Geier: Wechselkurssicherungsstrategien exportorientierter Unternehmen. Effizienzmessung von regelgebundenen Selektionsentscheidungen. 2012.

Band 56 Ernst Gittenberger: Betriebsformenwahl älterer KonsumentInnen. 2012.

Band 57 Michael Pichlmair: Miete, Lage, Preisdiktat. Strukturelle Effekte der Lageregulierung im mietrechtlich geschützten Wiener Wohnmarkt. 2012.

Band 58 Anna Katherina Guserl: Internationalisierungsprozesse und Finanzstrategien. Ansätze und empirische Analysen. 2013.

Band 59 Christian Idinger: Konsumentenpreiswissen. Eine empirische Studie im österreichischen Lebensmitteleinzelhandel. 2013.

Band 60 Dennis Jancsary: Die rhetorische Konstruktion von Führung und Steuerung. Eine argumentationsanalytische Untersuchung deutschsprachiger Führungsgrundsätze. 2013.

Band 61 Nicolas Hoffmann: Loyalty Schemes in Retailing. A Comparison of Stand-alone and Multi-partner Programs. 2013.

Band 62 Jose Gabriel Delgado Jimenez: Grenzüberschreitende Patientenmigration im zahnmedizinischen Bereich. Eine ökonomische Analyse am Beispiel Österreich und Ungarn. 2013.

Band 63 Wolfgang Koller: Prognose makroökonomischer Zeitreihen: Ein Vergleich linearer Modelle mit neuronalen Netzen. 2014.

www.peterlang.com