Mizrach, Bruce

Working Paper
Nonlinear time series analysis

Working Paper, No. 2006-04

Provided in Cooperation with:
Department of Economics, Rutgers University

Suggested Citation: Mizrach, Bruce (2006) : Nonlinear time series analysis, Working Paper, No. 2006-04, Rutgers University, Department of Economics, New Brunswick, NJ

This Version is available at:
http://hdl.handle.net/10419/31303

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.
nonlinear time series analysis

Since the early 1980s, there has been a growing interest in stochastic nonlinear dynamical systems of the form

\[ x_{t+1} = f(x_t, x_{t-1}, \ldots, x_{t-p}) + \sigma(x_t)e_t, \]

(1)

where \( \{x_t\}_{t=0}^{\infty} \) is a zero mean, covariance stationary process, \( f : \mathbb{R}^{p+1} \rightarrow \mathbb{R} \), \( \sigma \) is the conditional volatility, and \( \{e_t\}_{t=0}^{\infty} \) is an independent and identically distributed noise process. The major recent developments in nonlinear time series are described here using this canonical model. The first section develops representation theory for a third order approximation. Nonparametric approaches follow; these rely on series expansions of the general model. Ergodic properties including path dependence and dimension are considered next. I then consider two widely utilized parametric models, piecewise-linear models of \( f \) and autoregressive models for \( \sigma \) and \( \{e_t\} \).

I conclude with a discussion of hypothesis testing and forecasting.

Volterra expansion

There is no general causal representation for nonlinear time series as in the linear case. Series approximations rely on the Volterra expansion,

\[
x_{t+1} \simeq f(0) + p_{i=1} f_i x_{t-i} + \sum_{i_1=1}^{p} f_{i_1 i_2} x_{t-i_1} x_{t-i_2} + \cdots
\]

(2)

Brockett (1976) shows any continuous map over \([0, T]\) can be approximated by a finite Volterra series. Mittnik and Mizrach (1992) examine forecasts using generalized polynomial expansions like (2). Potter (2000) shows that in the cubic case, a one-sided Wold-type representation can be obtained,

\[
x_{t+1} \simeq \sum_{i=1}^{\infty} g_{i_1} v_{t-i} + \sum_{i_1=1}^{\infty} g_{i_1} x_{t-i_1} v_{t-i_2} + \cdots
\]

(3)

Koop, Pesaran and Potter (1996) note that the impulse response functions, \( E[x_{t+i}|x_t, v_t] - E[x_{t+i}|x_t] \) will depend upon the size and sign of \( v_t \) as well as the current state \( x_t \). I now turn to nonparametric approaches which build on approximations like 2.
Nonparametric estimation

Consider the local polynomial approximation to \( f(.) \) around \( x_0 \),
\[
\tilde{f}(x) = \sum_{j=0}^{m} \beta_j (x - x_0)^j.
\]
(4)

In the case \( j = 0 \), this corresponds to the kernel regression estimator of Nadaraya and Watson,
\[
\tilde{f}(x) = \frac{\sum_{i=1}^{T} x_{i+1} K_h(x_i - x_0)}{\sum_{i=1}^{T} K_h(x_i - x_0)}.
\]
(5)

The \( K_h \) are kernels, usually functions with a support on a compact set, assigning greater weight to observations closer to \( x_0 \), \( h \) is the bandwidth parameter, determining the size of the histogram bin. Nearest neighbours estimation is the case where \( h \) is adjusted to find a fixed number of nearby observations \( k \).

More generally, the local linear approximation solves,
\[
\begin{align*}
\alpha_0, \beta_0 \min (x_{i+1} - x_0 - \beta_0 (x_i - x_0))^2 \\
K_h(x_i - x_0).
\end{align*}
\]
(6)

The estimator (5) corresponds to the case where the only regressor in (6) is the constant term.

The application of these methods in the time series case is a fairly recent development. Conditions for consistency and asymptotic normality rely on mixing conditions where the dependence between \( x_{t+j} \) and \( x_t \) becomes negligible as \( j \) grows large.

A closely related approach involves the use of a recurrent neural networks,
\[
\Psi_i(x_t, h_{t-1}) = \Psi(g_{i,0} + g_{i,1} x_t + \cdots + g_{i,k} h_{t-k}),
\]
(13)

Kuan, Hornik, and White (1994) provide convergence results for bounded \( \Psi \) (most commonly the logistic) as \( p \) grows large.

A popular approach in the frequency domain is wavelets. The discrete wavelet transform
\[
x_{t+1} = \frac{1}{\sqrt{s_0}} \sum_{j=0}^{\infty} \gamma(j, k) \Psi_{j,k}(t),
\]
(8)

where the mother wavelet \( \Psi(t) \),
\[
\Psi_{j,k}(t) = \frac{1}{\sqrt{s_0}} \Psi \left( \frac{t - k\tau s_0}{s_0} \right),
\]
(9)

is parameterized by scale \( s_0 \) and translation \( \tau \), and the wavelet coefficients are given by
\[
\gamma(j, k) = \langle \Psi_{j,k}(t), x(t) \rangle.
\]
(10)

Daubechies (1992) orthonormal basis functions,
\[
E[\Psi_{j,k}(t)\Psi_{m,n}(t)] = 0, \quad \forall j \neq m, k \neq n,
\]
(11)

have received the widest application.

Even when very little is known about \( f \) or \( \sigma \), nonlinear time series analysis can shed light on the long run average or ergodic properties of the dynamical system.

Ergodic properties

Mathematicians have known since Poincaré that even simple maps like (1) can produce very complex dynamics. The nonlinear time series literature has developed tools for estimation of ergodic properties of these systems. Denote by \( Df(x) \) the Jacobian matrix of partial derivatives of (1),
\[
\begin{bmatrix}
\frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_p} \\
\vdots & \ddots & \vdots \\
\frac{\partial f_p}{\partial x_1} & \cdots & \frac{\partial f_p}{\partial x_p}
\end{bmatrix}
\]
(12)
evaluated at \( x \). Replacing 12 with a sample analog,
\[
J_t = \begin{bmatrix}
\Delta f_1/\Delta x_{1,t} & \cdots & \Delta f_1/\Delta x_{p,t} \\
\vdots & \ddots & \vdots \\
\Delta f_p/\Delta x_{1,t} & \cdots & \Delta f_p/\Delta x_{p,t}
\end{bmatrix}
\]
(13)

we compute eigenvalues \( V_i \),
\[
V_i(Q_t'Q_t)
\]
(14)
rank ordered from 1, \ldots, \( p \), where
\[
Q_t = J_{t-p} \cdots J_1
\]
(15)
The Lyapunov exponents are defined for the positive eigenvalues \( V_i^+ \) as
\[
T \to \infty \lim \lambda_i = \frac{1}{2(T-p)} \ln V_i^+,
\]
(16)
and a single exponent greater than 1 characterizes a system with sensitive dependence. Popularly known as ‘chaos’, this property implies that dynamic trajectories become unpredictable even when the state of the system is known with certainty. Gentçay and Dechert (1992) and Shintani and Linton (2004) provide methods for estimating these. Shintani and Linton (2003; 2004) reject the presence of positive Lyapunov exponents in both real output and stock returns.

The sum of the Lyapunov exponents also provides a measure of the Kolmogorov–Sinai entropy of the system. This tells the researcher how quickly trajectories separate. Mayfield and Mizrach (1991) estimate this time at about 15 minutes for the S&P 500 index.
A final quantity of interest is the dimension $p$ of the dynamical system. Nonlinear econometricians try to estimate the dimension from a scalar $m$-history. A powerful result due to Takens (1981) says this can be done as long as $m \geq 2p + 1$. Diks (2004) has shown that the scaling of correlation exponents seems to be consistent with the stochastic volatility model.

A great deal of progress has been made with parametric models of (1) as well. I begin with the widely utilized piecewise linear models.

**Piecewise linear models**

The most widely applied parametric nonlinear time series specification has been the Markov switching model introduced by James Hamilton (1989). The function $f$ is a piecewise linear function,

$$
\begin{align*}
  f(x_t) &= \begin{cases}
    \mu^{(1)} + \sum_{j=0}^{p} \phi_j^{(1)}(x_{t-j} - \mu^{(1)}), & S_t = s_t^{(1)} \\
    \vdots & \\
    \mu^{(m)} + \sum_{j=0}^{p} \phi_j^{(m)}(x_{t-j} - \mu^{(m)}), & S_t = s_t^{(m)}
  \end{cases}
\end{align*}
$$

(17)

where the changes among states are governed by an unobservable regime switching process, $S_t = s_t^{(i)}$, $i = 1, \ldots, m$, an $m \times m$ transition matrix $\Pi$, and $E[x_t|S_t = s_t^{(i)}] = \mu^{(i)}$. When $S_t$ is unobserved, Pr($S_t|x_{t-1}$) is nonlinear in $x_{t-1}$. Hamilton has shown that a two-dimensional switching model describes well the business cycle dynamics in the United States. This model has been extended to include regime dependence in volatility (Kim, 1994) and time varying transition probabilities (Filardo, 1994).

The latent state vector requires forming prior and posterior estimates of which regime you are in. The EM algorithm (Hamilton, 1990) and Bayesian Gibbs sampling methods (Albert and Chib, 1993) have proven fruitful in handling this problem. Hypothesis testing is also non-standard because under the alternative of $m - 1$ regimes, the conditional mean parameters are nuisance parameters. Hansen (1996) has explored carefully these issues.

A closely related framework is the threshold autoregressive (TAR) model,

$$
\begin{align*}
  f(x_t) &= \begin{cases}
    \left[ \mu^{(1)} + \sum_{j=0}^{p} \phi_j^{(1)}(x_{t-j} - \mu^{(1)}) \right] I(q(x_{t-d},Z_t) \leq \gamma_1) \\
    \left[ \mu^{(2)} + \sum_{j=0}^{p} \phi_j^{(2)}(x_{t-j} - \mu^{(2)}) \right] I(\gamma_1 < q(x_{t-d},Z_t) \leq \gamma_2) \\
    \vdots & \\
    \left[ \mu^{(m)} + \sum_{j=0}^{p} \phi_j^{(m)}(x_{t-j} - \mu^{(m)}) \right] I(q(x_{t-d},Z_t) > \gamma_{m-1})
  \end{cases}
\end{align*}
$$

(18)

$I(\cdot)$ is the indicator function, and $q(x_{t-d}, Z_t)$, the regime switching variable, is assumed to be an observable function of exogenous variables $Z_t$ and lagged $x$’s. The integer $d$ is known as the delay parameter. When $q$ depends only upon $x$, the model is called self-exciting.

Teräsvirta (1994) has developed a two-regime version of the TAR model in which regime changes are governed by a smooth transition function $g(x_{t-d}, Z_t) : \mathbb{R}^k \rightarrow [0, 1]$,

$$
\begin{align*}
  f(x_t) &= g(x_{t-d}, Z_t)p_{i=0}^{p} \phi^{(1)}(x_{t-j} - \mu^{(1)}) \\
    &+ (1 - g(x_{t-d}, Z_t))p_{i=0}^{p} \phi^{(2)}(x_{t-j} - \mu^{(2)}).
\end{align*}
$$

(19)

Luukkonen, Saikkonen and Teräsvirta (1988) have shown that inference and hypothesis testing in this model is often much simpler than in the piecewise linear models. Van Dijk and Franses (1999) have extended this model to multiple regimes. Applications of this framework have been widespread from macroeconomics (Teräsvirta and Anderson, 1992) to empirical finance (Franses and van Dijk, 2000).

Krolzg (1997) considers the multivariate case where $x_t = (x_{1,t}, x_{2,t}, \ldots, x_{k,t})^t$ is $k \times 1$. Balke and Fomby (1997) introduced threshold cointegration by incorporating error correction terms into the thresholds. Koop, Pesaran and Potter (1996) develop a bivariate model of US GDP and unemployment where the threshold depends upon the depth of the recession.

I now turn to models that introduce nonlinearity through the error term.

**Models of volatility**

Engle and Bollerslev have introduced the generalized autoregressive conditional heteroskedasticity (GARCH) model,

$$
\begin{align*}
  h_t &= \alpha_0 + \sum_{i=1}^{q} \alpha_i \sigma^2(x_{i-1}) + \sum_{i=1}^{p} \beta_i h_{t-i},
\end{align*}
$$

(20)

where $h_t = E[(x_t - E[x_t|\Omega_{t-1}])^2|\Omega_{t-1}]$ is the conditional variance. This is just a Box–Jenkins model in the squared residuals of 1 of order $(\max[p, q], p)$. The model is nonlinear because the disturbances are uncorrelated, but their squares are not.

The GARCH model describes the volatility clustering and heavy-tailed returns in financial market data, and has found wide application in asset pricing and risk management applications.

Volatility modelling has been motivated by the literature on options pricing. Popular alternatives to the GARCH model include the stochastic volatility (SV) model (Ghysels, Harvey and Renault, 1996), and the realized volatility approach of Andersen et al. (2003) and Barndorff-Nielsen and Shephard (2002). The discrete-time
SV model takes the form,

\[ x_t = \sigma_x \exp \left( h_t/2 \right) e_t, \quad (21) \]

\[ h_t = \beta h_{t-1} + \sigma_b \eta_t, \]

where \( x_t \) is the demeaned log asset return, and \( e_t \) and \( \eta_t \) are noise terms. Realized volatility sums high-frequency squared returns as an approximation of lower frequency volatility. Both GARCH and SV have been successful in explaining the departures from the Black–Scholes volatility. Both GARCH and SV have been successful in explaining the departures from the Black–Scholes volatility. Both GARCH and SV have been successful in explaining the departures from the Black–Scholes volatility. Both GARCH and SV have been successful in explaining the departures from the Black–Scholes volatility.

The final two sections address the marginal contribution of nonlinear modelling to goodness of fit and forecasting.

Testing for linearity and Gaussianity
There is a large literature on testing the importance of the nonlinear components of a model. The most widely used test is due to Brock, Dechert, Scheinkman and LeBaron (BDSL, 1996). Their nonparametric procedure is based upon U-statistics. Serfling (1980) is a good introduction.

The first step is to form \( m \)-histories of the data,

\[ x_t^m = (x_t, x_{t+1}, \ldots, x_{t+m-1}), \quad (22) \]

with joint distribution \( F(x_t^m) \). Introduce the kernel \( h : R^m \times R^m \rightarrow R \),

\[ h(x_t^m, x_s^m) = I(x_t^m, x_s^m, \epsilon) \equiv I(\|x_t^m - x_s^m\| < \epsilon), \quad (23) \]

where \( I(.) \) is the indicator function. The correlation integral of Grassberger and Procaccia (1983),

\[ C(m, \epsilon) \equiv \int_X \int_X I(x_t^m, x_s^m, \epsilon) dF(x_t^m) dF(x_s^m), \quad (24) \]

is the expected number of \( m \)-vectors in an \( \epsilon \) neighbourhood. A U-statistic,

\[ C(m, N, \epsilon) \equiv \frac{2}{N(N-1)} \sum_{t=1}^{N-1} I(x_t^m, x_s^m, \epsilon), \quad (25) \]

is a consistent estimator of 24. BDSL demonstrate the asymptotic normality of the statistic

\[ \sqrt{N} \frac{S(m, N, \epsilon)}{\sqrt{\text{Var}[S(m, N, \epsilon)]}} \overset{d}{\rightarrow} N(0, 1), \quad (26) \]

where

\[ S(m, N, \epsilon) = C(m, N, \epsilon) - C(1, N, \epsilon)^m. \quad (27) \]

There is a multi-dimensional extension due to Baek and Brock (1992). De Lima (1997) explores the use of the BDSL under moment condition failure.

There is a direct relationship between nonlinear and non-Gaussian time series. In the model (1), even if the disturbance term \( e_t \) is normal, nonlinear transformations of Gaussian noise will make \( x_t \) non-Gaussian. Testing for Gaussianity is then an instrumental part of the nonlinear time series toolkit.

Hinich (1982) has developed testing in the time domain using the biconrelation,

\[ \gamma(r,s) = \frac{1}{s} \sum_{t=1}^{s} x_t x_{t+r} x_{t+s}/(N-s), \quad 0 \leq r \leq s, \quad (28) \]

and in the frequency domain using the bispectrum,

\[ B(\omega_1, \omega_2) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \gamma(r,s) \exp [-i(\omega_1 r + \omega_2 s)]. \quad (29) \]

For a Gaussian time series, the biconrelation should be close to zero, and the bispectrum should be flat across all frequencies. Both tests have good power against skewed alternatives.

Ramsey and Rothman (1996) have proposed a related time domain procedure that looks for time reversibility,

\[ F(X_t, X_{t+1}, \ldots, X_{t+r}) = F(X_{t-r}, X_{t-s-1}, \ldots, X_{t-1}) \quad (30) \]

for any \( r, s \) and \( t \), where \( F(.) \) is the joint distribution. This condition is stronger than stationarity because of the triple index. The authors find evidence of business cycle asymmetry using this diagnostic.

Forecasting
For many, the bottom line on nonlinear modelling is the ability to generate superior forecasts. In this respect, the results from the nonlinear literature are decidedly mixed. Harding and Pagan (2002) are prominent sceptics. Teräsvirta, van Dijk and Medeiros (2005) provide a very wide set of evidence in favour of nonlinear models.

Aside from the comparison of point forecasts from model \( h \), \( u_{t+1} = f_i(x_t) \), with a particular loss function \( g(.) \),

\[ H_0 : E[g(u_{t+1}) - g(u_{t+1})] = 0, \quad (31) \]

there has been growing interest in comparing forecast densities \( p_i(x_{t+1}|f_i(x_t)) \),

\[ H_0 : \int |p_i(x_{t+1}|f_i(x_t)) - p_j(x_{t+1}|f_j(x_t))| dx = 0. \quad (32) \]

Corradi and Swanson (2005) provide a comprehensive overview of available tools.
See also forecasting; linear models; stochastic volatility models.

I would like to thank Cees Diks, James Hamilton, Sebastiano Manzan, Simon Potter, Phil Rothman, Dick van Dijk and Steven Durlauf for helpful comments.

Bibliography


Hansen, B.E. 1996. Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica* 64, 413–30.


non-nested hypotheses

In economics, as in many other disciplines, there are competing explanations of the same phenomena, often characterized by alternative statistical models. Different models may represent, for example, different theoretical paradigms, or could be the result of alternative formulations from the same paradigm. Within the classical framework, the problem of model adequacy is approached through ‘general specification tests’, the ‘diagnostic tests’, and the ‘non-nested tests’. All three approaches can be used to test the same explanation or hypothesis of interest (the null or the maintained hypothesis), but they differ in their consideration of the alternative(s). General specification tests intentionally consider a broad class of alternatives, while the alternatives considered under diagnostic and non-nested testing procedures are much more specific. In the case of non-nested tests the null hypothesis is contrasted to a specific alternative. Non-nested tests are appropriate when rival hypotheses are advanced for the explanation of the same economic phenomenon, and the aim is to devise a powerful test against a specific alternative.

When the null hypothesis is nested within the alternative, standard classical procedures such as those based on the likelihood ratio, Wald and Lagrange multiplier (or score) principles can be utilized. But if the null and the alternative hypotheses belong to ‘separate’ families of distributions, classical testing procedures cannot be applied directly and need to be suitably modified.

This article provides an overview of the concepts and some of the most widely used non-nested hypotheses tests and applies these procedures to the classical regression models. Our discussion of non-nested hypothesis testing will necessarily omit many topics. Survey articles on this subject include McAler and Pesaran (1986), Gourieroux and Monfort (1994), and Pesaran and Weeks (2001).

Non-nested models

Suppose the object of interest is the process generating the random variable $Y$, observed over a sample of size $n$, $y = (y_1, y_2, \ldots, y_n)^T$. Assume that the true process generating $y$ is characterized by a joint probability density function, $f_0(y)$, which is unknown, and two models (hypotheses) are advanced as possible explanations of $Y$, represented by the joint probability density functions:

$$H_g = \{g(y; \theta), \theta \in \Theta\},$$

$$H_h = \{h(y; \gamma), \gamma \in \Gamma\}. \quad (1)$$

These functions are known but depend on a finite number of unknown parameters denoted by $\theta \in \Theta$ and $\gamma \in \Gamma$, respectively. The sets $\Theta$ and $\Gamma$ represent the ‘admissible’ parameter space for which the respective densities $g(y; \theta)$ and $h(y; \gamma)$ are well defined. The aim is to ascertain which of the two alternatives, $H_g$ and $H_h$, if any, can be viewed as belonging to $f_0(y)$. In this set-up there is no natural null hypothesis; either of the two hypotheses under consideration can be taken as the null.

In practice, the analysis of non-nested hypotheses is carried out with both alternatives taken in turn as the null hypothesis. Four outcomes are possible: (i) $H_g$ rejected against $H_h$ and not vice versa, (ii) $H_h$ rejected against $H_g$ and not vice versa, (iii) neither hypothesis is rejected against the other, and finally (iv) both hypotheses are rejected against one another. The first two outcomes are familiar from the classical test results and are straightforward to interpret. The third outcome can arise when the two models are very close to $f_0(y)$, and hence equivalent observationally. The fourth outcome suggests the existence of a third possible model which shares important features from both models under consideration.

Pseudo-true values and closeness measures

Given the observations $y$, the maximum likelihood (ML) estimators of $\theta$ and $\gamma$ are given by

$$\hat{\theta}_n = \arg \max_{\theta \in \Theta} L_g(\theta), \quad \hat{\gamma}_n = \arg \max_{\gamma \in \Gamma} L_h(\gamma),$$

where the corresponding log-likelihood functions are defined by $L_g(\theta) = \log (g(y; \theta))$ and $L_h(\gamma) = \log (h(y; \gamma))$. Throughout we shall assume that probability densities satisfy the usual regularity conditions as established, for example in White (1982), such that $\hat{\theta}_n$ and $\hat{\gamma}_n$ have asymptotically normal limiting distributions under the ‘true’ model, $f_0(y)$. In the general case where neither of the models under consideration coincide with $f_0(y)$, $\hat{\theta}_n$ and $\hat{\gamma}_n$ are known as quasi-ML estimators and their