Savin, Ivan; Winker, Peter

Working Paper

Lasso-type and heuristic strategies in model selection and forecasting

Jena Economic Research Papers, No. 2012,055

Provided in Cooperation with:
Max Planck Institute of Economics

JENA ECONOMIC RESEARCH PAPERS

2012 – 055

Lasso-type and Heuristic Strategies in Model Selection and Forecasting

by

Ivan Savin
Peter Winker

www.jenecon.de

ISSN 1864-7057

The JENA ECONOMIC RESEARCH PAPERS is a joint publication of the Friedrich Schiller University and the Max Planck Institute of Economics, Jena, Germany. For editorial correspondence please contact markus.pasche@uni-jena.de.

Impressum:

Friedrich Schiller University Jena
Carl-Zeiss-Str. 3
D-07743 Jena
www.uni-jena.de

Max Planck Institute of Economics
Kahlaische Str. 10
D-07745 Jena
www.econ.mpg.de

Lasso-type and Heuristic Strategies in Model Selection and Forecasting*

Ivan Savin† and Peter Winker‡

Abstract

Several approaches for subset recovery and improved forecasting accuracy have been proposed and studied. One way is to apply a regularization strategy and solve the model selection task as a continuous optimization problem. One of the most popular approaches in this research field is given by Lasso–type methods. An alternative approach is based on information criteria. In contrast to the Lasso, these methods also work well in the case of highly correlated predictors. However, this performance can be impaired by the only asymptotic consistency of the information criteria. The resulting discrete optimization problems exhibit a high computational complexity. Therefore, a heuristic optimization approach (Genetic Algorithm) is applied. The two strategies are compared by means of a Monte–Carlo simulation study together with an empirical application to leading business cycle indicators in Russia and Germany.

Keywords: *Adaptive Lasso, Elastic net, Forecasting, Genetic algorithms, Heuristic methods, Lasso, Model selection*

JEL Classification: *C51, C52, C53, C61, C63*

*Financial support from the German Science Foundation (DFG RTG 1411) is gratefully acknowledged. The present paper will be published as a chapter in: Borgelt C., M.A. Gil, J. Sousa and M. Verleysen (Eds.) Towards Advanced Data Analysis by Combining Soft Computing and Statistics, Springer, Berlin (2013).

†DFG Research Training Program 'The Economics of Innovative Change', Friedrich Schiller University Jena and Max Planck Institute of Economics, Bachstrasse 18k Room 216, D-07743 Jena, Germany, Ivan.Savin@uni-jena.de

‡Justus Liebig University Giessen, Licher Strasse 64, D-35394 Giessen, and Centre for European Economic Research, Mannheim, Germany, Peter.Winker@wirtschaft.uni-giessen.de

1 Introduction

The model selection process is crucial for the further analysis of any multiple regression model and its forecasting performance. Picking up too many regressors increases the variance of the constructed model, and taking fewer regressors than needed might result in biased and even inconsistent estimates. Both of these problems can also have negative effects on the quality of forecasts based on the models obtained through the application of these methods.

During the last years, the least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996) has become a very popular approach for simultaneous model selection and parameter estimation. Its main advantage is seen in obtaining both a high prediction accuracy and a parsimonious model, which is due to the regularization parameter which results in shrinking the coefficients of insignificant regressors towards zero. Hence, the resulting models concentrate on the strongest effects which tends to increase the total accuracy of the model forecast. Furthermore, the Lasso is very computationally efficient (hardly exceeding the complexity of one linear regression (Efron et al, 2004)).

However, the Lasso has some limitations. In particular, inconsistent estimates are obtained for highly correlated regressors. Numerous modifications have been suggested revising and improving the initial Lasso concept (e.g., the elastic net, the adaptive lasso), which can improve its performance under certain conditions, but do not represent a universal remedy from the limitation stated.

An alternative to the shrinkage operator is offered by model selection approaches based on information criteria (IC) which tend to provide a consistent model choice also for correlated predictors. However, even for a moderate number of predictors, these methods might result in substantial computational cost when considering a full enumeration of all alternatives. Fortunately, thanks to advances in heuristic optimization methods mimicking some evolution processes (Gilli and Winker, 2009), there are efficient algorithms able to identify at least a good approximation to the IC's global optimum even for larger problem instances. Furthermore, IC's performance is naturally impaired by small sample sizes due to their only asymptotic consistency.

To the best of our knowledge, this study is the first[1] comparing the Lasso–type and heuristic methods both for model selection and forecasting, and contributing to the literature by demonstrating that in certain situations (e.g., if regressors in a given data set are pairwise highly correlated and for large data sets) heuristic algorithms can outperform the Lasso–type solutions.

The rest of this paper is structured as follows. Section 2 introduces both the Lasso–type and the heuristic methods. Section 3 provides results of a Monte–Carlo analysis, and

[1] An exception, however, only with regard to the comparison of the two strategies for model selection can be found in Savin (forthcoming).

Section 4 illustrates an application to leading business cycle indicators. Finally, Section 5 concludes.

2 Model Selection Methods

The least absolute shrinkage and selection operator (Lasso), introduced by Tibshirani (1996), is a constrained version of the ordinary least squares estimator, but has also been applied to GMM–estimators. Numerous applications of this technique can be found in medicine, economics and other scientific fields (Hastie et al, 2009) including also time series forecasting (see, among others, Bai and Ng (2008)).

Consider the model selection problem for the following regression function:

$$y = \alpha + X^{opt}\beta + \varepsilon, \tag{1}$$

where α is an n-vector with all elements equal, X is an $n \times k$ matrix of k regressors and their values for n observations, β is a $k \times 1$ vector of their coefficients and ε is an $n \times 1$ vector of residuals. In (1) X^{opt} refers either to the 'true' model in a Monte–Carlo simulation set–up or to an optimal approximation to the unknown real data generating process. Standardizing the predictors so that they have mean 0 and standard deviation equaling 1, and the response having mean 0, one can omit α without loss of generality.

2.1 Lasso–type Strategies

For (1) the Lasso objective function can be presented as follows:

$$\widehat{\beta}_{Lasso} = \arg\min_{\beta} \left[\| y - X\widehat{\beta} \|_2^2 + \lambda \| \widehat{\beta} \|_1 \right]. \tag{2}$$

While the first term in the right part of equation (2) measures the fit of the model by the residual sum of squares (RSS), the second one with $\lambda > 0$ is the shrinkage applied to the sum of the absolute values of the coefficients. Hence, the Lasso can be referred to as a special case of the Bridge regression approach (Frank and Friedman, 1993) imposing an upper bound on the L^q-norm of the parameters ($0 < q < \infty$) with $q = 1$:

$$\| \widehat{\beta} \|_q = \left[\sum_{j=1}^{k} |\beta_j|^q \right]^{1/q}. \tag{3}$$

There are different approaches to solve (2) including quadratic programming, coordinate–wise optimization and gradient projection (see, e.g., Gasso et al (2009)). For the sake of brevity we do not discuss any of those methods, so that the interested reader is advised to consult the literature. In this study we use a modification of the *LARS* algorithm

suggested by Efron et al (2004) and popularized among practitioners.[2] The algorithm provides a piecewise–linear solution path in the tuning parameter $\lambda \in [0, \infty)$ with all $\widehat{\beta}$'s set to zero at $\lambda = \infty$ and equal to the OLS estimate at $\lambda = 0$. To select a single solution, λ is chosen by tenfold cross–validation minimizing the prediction error (PE) of the model.

Setting $\lambda > 0$ by cross–validation one insures the Lasso solution to have a parsimony property, i.e. only a subset of resulting predictors in (2) has non–zero coefficients. This feature of the Lasso might increase the total accuracy of the model forecast and improves the interpretability of the selected model.

However, the Lasso has substantial limitations. First, it cannot identify all 'true' predictors in a data set with pairwise highly correlated regressors (Zou and Hastie, 2005). The latter can be referred to as the 'irrepresentable condition' (Zhao and Yu, 2006, p. 2544). Thus, Lasso is consistent in low correlation settings only, when

$$\max_{j>r} \| cov(X_j, X^{true}) cov(X^{true})^{-1} \|_1 < 1, \tag{4}$$

while in presence of high correlations between 'true' and irrelevant variables, the Lasso cannot recover the correct sparsity pattern ($\widehat{\beta}_{Lasso} \nrightarrow \beta^{true}$).

However, as Meinshausen and Yu (2008) show, even failing to discover the correct sparsity pattern (when (4) does not hold), the Lasso can provide good approximations of the 'true' model for large sample sizes ($\| \beta - \widehat{\beta}_{Lasso} \|_2 \to 0$ as $n \to \infty$). In other words, Lasso selects 'true' variables with high probability and irrelevant ones have only marginal coefficients (L^2-norm consistency).

Second, Lasso is inconsistent for $k \gg n$ (underdetermined linear system), where (2) can identify not more than $n - 1$ (standardized) predictors (Efron et al, 2004).

Lasso Modifications

Many proposals have been made on how to improve the Lasso concept. Due to space restrictions, we concentrate only on two such modifications. For a more complete overview, the interested reader is referred, e.g., to Hastie et al (2009) and Gasso et al (2009).

We consider two extensions of the Lasso: the elastic net (EN) using a combination of the Lasso (λ_1) and the ridge regression (λ_2) penalty (Zou and Hastie, 2005):

$$\widehat{\beta}_{EN} = \arg\min_{\beta} \left[\| y - X\widehat{\beta} \|_2^2 + \lambda_1 \| \widehat{\beta} \|_1 + \lambda_2 \| \widehat{\beta} \|_2^2 \right], \tag{5}$$

and the adaptive Lasso (aLasso) applying different amounts of shrinkage for each regression coefficient (Zou, 2006):[3]

[2] Related codes are available at `http://www.stanford.edu/~hastie/Papers/LARS` for R and `http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=3897` for Matlab.

[3] For more details the reader is referred to the literature. See also Savin (forthcoming).

$$\widehat{\beta}_{aLasso} = \arg\min_{\beta} \left[\| y - X\widehat{\beta} \|_2^2 + \lambda \sum_{j=1}^{k} \widehat{\omega}_j |\beta_j| \right]. \tag{6}$$

Thus, the selected extensions are particularly designed to deal with the limitations stated and operate in a continuous space remaining computationally efficient. Furthermore, in line with Candès and Tao (2007) we also perform unregularized restricted estimation (i.e. OLS estimation on the selected set of regressors) for all Lasso–type methods tested to alleviate a potential bias that results from the regularized estimation ($\lambda > 0$).

2.2 Heuristic Optimization Methods

Alternatively, information criteria (IC) can be used to identify X^{opt} in (1). IC rank different models according to their fitness, while penalizing model complexity. Hence, they can be interpreted as a L^0-constraint, penalizing not the coefficients' values, but only their number:

$$\widehat{\beta}_{IC} = \arg\min_{\beta} \left[\| y - X\widehat{\beta} \|_2^2 + \lambda \, \| \widehat{\beta} \|_0 \right]. \tag{7}$$

IC have become a standard instrument in model selection ranging from lag order selection in multivariate linear and nonlinear autoregression models to selection between rival nonnested models (Winker, 1995). In this study, the Bayesian IC (BIC) and the Hannan-Quinn IC (HQIC) are employed. For infinitely large sample sizes these IC are consistent model selection instruments and, as noted by (Zhao and Yu, 2006, p. 2553), the solution of (7) remains consistent even for data sets with correlated regressors.

Given that the search space of candidate models in (7) is discrete, standard gradient methods cannot be applied. Also the full enumeration of all possible solutions is only feasible for a moderate k. Consequently, in the last decade many studies have been devoted to the problem in (1): sequential bottom–up (top–down) inclusion (deletion) of individual regressors (Perez-Amaral et al, 2003; Hendry and Krolzig, 2005); usage of certain prior probabilities shrinking the parameter search space and resulting in model averaging (Kapetanios et al, 2008). However, these methods investigate only a specific fraction of all submodels, whereas there is no guarantee to find the 'true' model in this way.

In order to tackle the highly complex integer optimization problem, one can take advantage of optimization heuristics that mimic natural evolution processes. These methods are called 'heuristic' or 'meta-heuristics' because of their stochastic nature that helps them to converge to a model which at least represents a good approximation to the IC optimum. For an overview of these optimization techniques see Gilli and Winker (2009). In Savin and Winker (forthcoming) a similar subset selection problem was handled by

two algorithms: Threshold Accepting and Genetic Algorithms (GA). Since GA provided slightly better results in terms of both CPU time and solution quality, only GA are considered in the following.

GA are population–based heuristics that investigate the search space in many directions simultaneously, performing jumps in the search space by means of crossover and mutation mechanisms. Thereby, the probability of getting stuck in a local optimum is reduced. The members in the population are represented as bit strings of ones and zeros corresponding to the predictor variables included and not included in the candidate model. In each generation GA replace parts of a population with new solutions aimed to be better for a given problem. The GA algorithm implemented is very similar to the one in Savin and Winker (forthcoming).[4] The only difference is that 1000 generations are found to be sufficient in this study for GA to converge.

3 Monte–Carlo Study

The goal of this section is to determine in what set–ups which of the two strategies, Lasso–type methods (Lasso, EN, aLasso) or GA tined by IC, provide superior results (in terms of correctly recovered subsets, forecasting and estimation accuracy) and what is the corresponding CPU-time required.

Data Generating Process

To this end, different artificial data sets are generated varying the sample size (n) from 100 (frequent in macroeconomics) to 1000 (which is mostly available only in finance and natural sciences) and fixing the number of potential regressors to 50. First, we generate 4 predictors with a joint Gaussian distribution and covariance matrix Σ. We choose either $\Sigma_{i,j} = 0.5^{|i-j|}$ or $0.75^{|i-j|}$ with $1 \leq i, j \leq k$, corresponding to a 'low' and 'high' correlation setting, respectively. Second, the data matrix consisting of lags 1 to 10 of these predictors is formed (X^{mc}). Third, we select a small number of elements $k^{true} = 5$ of the coefficient vector β^{mc}, which are set to non–zero values.[5] These non–zero coefficient values are randomly distributed between -1 and 1, and divided by the respectively chosen lag order so that lags of higher order are (on average) assigned with smaller coefficients.[6] Fourth, the initial value of the response variable (y_0^{mc}) is set to zero, and based on β_j^{mc}, one recursively generates y_t^{mc} and adds an i.i.d. normal random error term:[7]

[4]Thus, a population of 500 solutions, the uniform crossover mechanism and a mutation operator applied to 5 randomly chosen genes with 50% probability are employed.

[5]One ensures that one lag of each variable (including the dependent one) is included.

[6]This appears reasonable since in empirical studies lags of lower order are found to be more important.

[7]Finally, the first 11 observations in y^{mc} and X^{mc} are discarded.

$$y_t^{mc} = \sum_{i=1}^{10} \beta_{0,i}^{mc} y_{t-i}^{mc} + \sum_{j=1}^{4} \sum_{i=1}^{10} \beta_{j,i}^{mc} x_{j,t-i}^{mc} + \varepsilon_t, \quad \varepsilon \sim N(0, \sigma_\varepsilon^2). \tag{8}$$

In (8) one chooses σ_ε such that the corresponding noise–to–signal ratio (NSR, for details see (Frank and Friedman, 1993, p. 125)) equals either 1/5 ('low noise') or 2 ('high noise'). Obviously, (8) represents an Autoregressive Distributed Lag model with 10 lags for both, one dependent and four explanatory variables, where no current values of the explanatory variables are involved. Thus, for a general ADL(p_1, p_2, p_3) we consider ADL(10,4,10).

Simulation Results

The quality of the results in terms of model identification is assessed by the True Positive Rate (TPR) and the False Negative Rate (FNR)[8], whereas mean–squared error (MSE = $E[(\widehat{\beta} - \beta^{mc})'\Sigma(\widehat{\beta} - \beta^{mc})]$) is used as a measure of the estimation accuracy.[9] For this purpose, 90% of the observations are used as a training set. The CPU time corresponding to a single restart using Matlab 7.11 on a Pentium IV 3.3 GHz is reported.[10]

Furthermore, the remaining 10% of observations are left for an out–of–sample forecast, where root mean–squared forecast error, and its standard deviation computed over 50 replications (in parentheses),

$$RMSFE = \sqrt{\frac{1}{T_2 - T_1 + 1} \sum_{t=T_1}^{T_2} (y_t^{mc} - \widehat{y}_t)^2}, \tag{9}$$

is used to assess the forecast quality. Thereby, T_1 and T_2 indicating the first and the last period of the forecasting period.

Simulation results obtained for different set–ups are reported in Table 1.[11] For medium–sized samples ($n = 500$) heuristics clearly outperform Lasso–type methods in subset recovery and estimation accuracy,[12] which eventually results in a better forecasting performance. However, the difference in RMSFEs is not that large. Among the Lasso methods considered, aLasso provides in general superior results, and this dominance holds for different correlation and noise settings. For 'high correlation' some marginal improvements as compared to classical Lasso are obtained via EN, which is due to the more robust ridge penalties.

[8] TPR is the percentage of 'true' regressors from all variables selected and FNR is the portion of rejected 'true' regressors among correctly selected and correctly rejected ones.

[9] Standard deviations computed over 50 replications are given in parenthes in Table 1. Unregularized restricted estimations for Lasso–type methods are reported as MSE_2.

[10] For each method, averages over 50 replications of the procedure are reported.

[11] Due to space constraints, here we report only results for the BIC, but qualitatively similar findings based on HQIC are available on request.

[12] Even accounting for MSE_2 an improvement for all scenarios is depicted in Table 1.

It can be observed that heuristics improve in performance relative to the Lasso methods in low noise settings and for larger sample sizes. The former is due to a more restrictive selection performed by the shrinkage strategies, which for 'low noise' translates in substantially more type II errors, i.e., ignoring too many relevant predictors, whereas the latter results from the asymptotic consistency of IC allowing to identify the correct sparsity pattern. In contrast, for $n = 100$ the difference between shrinkage and heuristic methods becomes less evident. Furthermore, for 'high noise' and small n Lasso–type strategies indisputably beat heuristics both in estimation and forecasting.

4 Application to Leading Business Cycle Indicators in Germany and Russia

Being particularly interested in the usefulness of the strategies from the forecasting point of view, we also show their application to real economic data.

Leading indicators (LI) are nowadays a standard tool for the analysis and forecasting of business cycles due to the publication delay of data on real production. While for Germany (as for other industrial countries) there is a large body of empirical evidence that models forecasting industrial production (IP), which include LI, outperform forecasts of univariate time series models (Vogt, 2007; Ozyldrim et al, 2010), there is less such evidence for developing countries.

For the empirical application we use two LI (business expectations and business climate) and IP for Germany and Russia for the period 02/1999–09/2009. More information on the properties of the data can be found in Savin and Winker (2012). Important is that the German LI are seasonally adjusted, while for Russia they are not. Furthermore, a potential structural break in Russian data must be accounted. Hence, we consider the IP indices for both countries also as unadjusted and introduce seasonal and shift dummies, and their interaction terms (for details see Savin and Winker (2012)) to account for these data features.

Similar to Section 3, ADL models (augmented with seasonal and shift dummies) are our modelling framework to identify predictors and construct forecasts, while an AR(2) process serves as benchmark. The latter is found to be a hard competitor in business cycle forecasting for small data sets (Savin and Winker, 2012).

We only employ 1–step–ahead forecasts of IP growth rates (log differences) for periods of one and two years length between 11/2006 and 09/2009 (this also allows one to consider the forecasting performance both prior and during the crisis) and increasing estimation windows (IEW). Furthermore, in contrast to Savin and Winker (2012), we allow for lags from both LI and both countries to be included in (9) for each IP, so that a selection out of 65 predictors (5 variables and 13 lags) has to be made. As a result, two data sets with

Table 1: Monte–Carlo simulation results

			Lasso	EN	aLasso	BIC	Lasso	EN	aLasso	BIC
			Low correlation				High correlation			
Results for $n = 100$	Low noise	TPR	72.7%	71.2%	70.9%	56.0%	73.9%	73.5%	60.7%	53.8%
		FNR	2.6%	2.5%	2.3%	1.0%	3.0%	2.9%	3.0%	1.6%
		MSE	.0212	.0209	.0102	.0035	.0235	.0233	.0104	.0076
			(.0413)	(.0408)	(.0160)	(.0033)	(.0431)	(.0422)	(.0140)	(.0166)
		CPU	.4s	1.0s	1.0s	32s	.3s	.7s	1.0s	31s
		MSE_2	.0166	.0165	.0076		.0189	.0188	.0083	
			(.0374)	(.0373)	(.0135)		(.0374)	(.0372)	(.0143)	
		RMSFE	.0114	.0113	.0112	.0113	.0119	.0117	.0126	.0122
			(.0053)	(.0051)	(.0048)	(.0039)	(.0049)	(.0049)	(.0052)	(.0053)
	High noise	TPR	79.3%	79.3%	62.7%	36.6%	71.2%	70.2%	50.0%	31.9%
		FNR	7.1%	7.1%	7.1%	6.1%	7.2%	7.3%	7.0%	6.6%
		MSE	.0339	.0336	.0247	.0529	.0363	.0359	.0286	.0521
			(.0718)	(.0719)	(.0447)	(.0542)	(.0856)	(.0857)	(.0527)	(.0518)
		CPU	.3s	.7s	1.1s	28s	.3s	.7s	.7s	29s
		MSE_2	.0234	.0234	.0224		.0338	.0338	.0234	
			(.0451)	(.0451)	(.0413)		(.0844)	(.0844)	(.0421)	
		RMSFE	.1078	.1078	.1107	.1191	.1088	.1089	.1123	.1234
			(.0402)	(.0402)	(.0424)	(.0496)	(.0415)	(.0413)	(.0449)	(.0547)
Results for $n = 500$	Low noise	TPR	68.5%	71.5%	68.3%	84.5%	66.1%	67.0%	74.6%	87.1%
		FNR	2.1%	2.1%	1.6%	.7%	2.3%	2.1%	1.6%	.7%
		MSE	.0187	.0187	.0069	3.0×10^{-4}	.0209	.0208	.0063	2.3×10^{-4}
			(.0299)	(.0299)	(.0158)	(4.7×10^{-4})	(.0376)	(.0376)	(.0196)	(3.1×10^{-4})
		CPU	.3s	.8s	.8s	88s	.3s	1.0s	.9s	93s
		MSE_2	.0120	.0120	.0023		.0154	.0151	.0042	
			(.0235)	(.0235)	(.0052)		(.0328)	(.0329)	(.0182)	
		RMSFE	.0102	.0102	.0102	.0098	.0107	.0107	.0107	.0103
			(.0041)	(.0041)	(.0042)	(.0041)	(.0043)	(.0043)	(.0046)	(.0041)
	High noise	TPR	95.6%	95.6%	75.4%	81.8%	91.7%	91.7%	78.4%	79.7%
		FNR	6.3%	6.3%	5.0%	3.5%	6.4%	6.4%	5.7%	4.1%
		MSE	.0269	.0266	.0214	.0054	.0270	.0266	.0225	.0060
			(.0420)	(.0413)	(.0488)	(.0056)	(.0426)	(.0415)	(.0432)	(.0063)
		CPU	.3s	.8s	.8s	78s	.3s	.8s	.8s	81s
		MSE_2	.0166	.0166	.0087		.0167	.0167	.0108	
			(.0276)	(.0276)	(.0105)		(.0277)	(.0277)	(.0191)	
		RMSFE	.0985	.0985	.0962	.0957	.1007	.1006	.0992	.0981
			(.0407)	(.0407)	(.0392)	(.0397)	(.0421)	(.0421)	(.0407)	(.0413)
Results for $n = 1000$	Low noise	TPR	54.2%	54.3%	80.8%	90.8%	60.1%	60.1%	72.2%	88.7%
		FNR	1.4%	1.4%	.9%	.4%	1.7%	1.7%	.9%	.3%
		MSE	.0107	.0107	.0041	6.9×10^{-5}	.0122	.0122	.0030	1.3×10^{-4}
			(.0181)	(.0181)	(.0077)	(1.1×10^{-4})	(.0171)	(.0172)	(.0048)	(2.6×10^{-4})
		CPU	.4s	1.0s	1.7s	157s	.5s	1.5s	1.8s	156s
		MSE_2	.0044	.0044	6.7×10^{-4}		.0080	.0080	.0010	
			(.0097)	(.0097)	(.0019)		(.0132)	(.0132)	(.0033)	
		RMSFE	.0103	.0103	.0102	.0101	.0095	.0095	.0098	.0092
			(.0038)	(.0038)	(.0037)	(.0037)	(.0040)	(.0040)	(.0067)	(.0039)
	High noise	TPR	93.7%	93.9%	75.7%	85.9%	92.9%	93.0%	76.3%	83.2%
		FNR	5.2%	5.2%	3.6%	2.3%	5.3%	5.4%	4.4%	2.7%
		MSE	.0199	.0199	.0089	.0029	.0199	.0198	.0153	.0030
			(.0247)	(.0247)	(.0115)	(.0042)	(.0245)	(.0245)	(.0248)	(.0043)
		CPU	.5s	1.5s	1.4s	147s	.5s	1.4s	1.3s	153s
		MSE_2	.0132	.0133	.0063		.0133	.0132	.0086	
			(.0178)	(.0178)	(.0109)		(.0178)	(.0178)	(.0145)	
		RMSFE	.0925	.0927	.0912	.0902	.0951	.0949	.0944	.0926
			(.0378)	(.0386)	(.0375)	(.0372)	(.0403)	(.0397)	(.0403)	(.0395)

highly correlated potential predictors are generated (see Figure 1).[13]

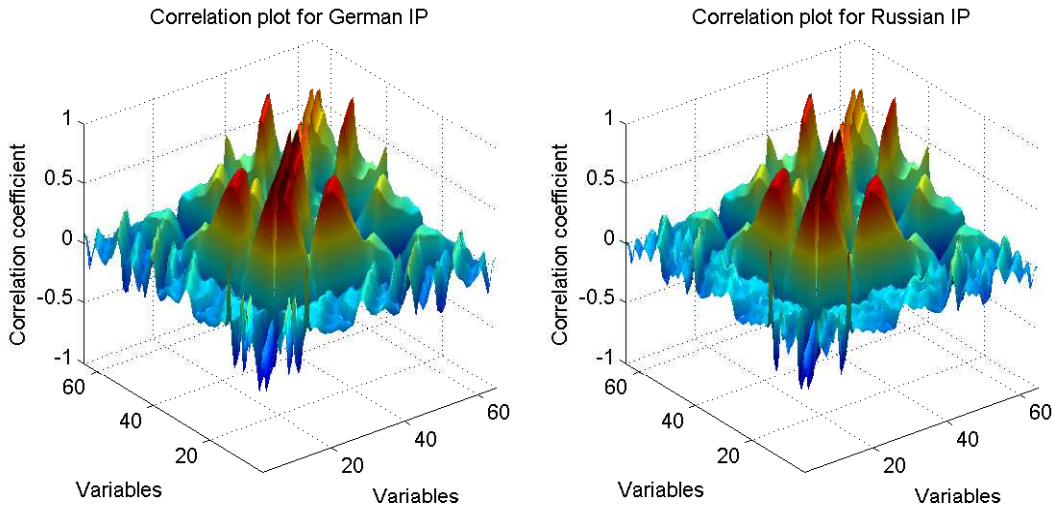

Figure 1: Pairwise correlations in the empirical data sets

Results for the two model selection strategies are exhibited in Table 2. As one can see, for the final forecasting period the shrinkage strategies mostly dominate the benchmark for both 12– and 24–month period forecasts, while heuristics fail do so. The main reason for this performance is seen to be the small estimation sample available: there are merely 128 observations for estimation and forecasting in total, which is most comparable with the upper panel in Table 1. Furthermore, since the IP growth rates can be to a large extent ($R^2 \approx 70 - 80\%$) explained by the set of lags selected (together with the seasonal and shift dummies), which corresponds to the low noise setting, EN and Lasso outperform aLasso. Finally, since particularly for small noise and high correlation among predictors in small samples Lasso–type methods provide some better forecasts than heuristics, the advantage of the shrinkage methods could be expected from the Monte–Carlo results.

Table 2: Forecasting performance of the ADL models

Model specification			Germany	Russia	Germany	Russia
			10/2007–09/2009		10/2008–09/2009	
RMSFE in relation to AR(2)	Lasso		0.9064	0.8296	0.7479	0.7800
	EN		0.9469	0.8296	0.7479	0.7748
	aLasso		0.9849	0.9744	1.1355	0.8422
	Genetic Algorithms	BIC	1.1763	0.9629	1.1832	0.9322
		HQIC	1.1558	1.1191	1.2473	1.1169

We also consider the performance of the two strategies over a set of forecasting periods

[13]The main diagonal in the correlation matrix is removed.

shifting by one month (rolling windows). The results are provided in Figure 2 (upper panel for 12– and lower for 24–month forecasts).

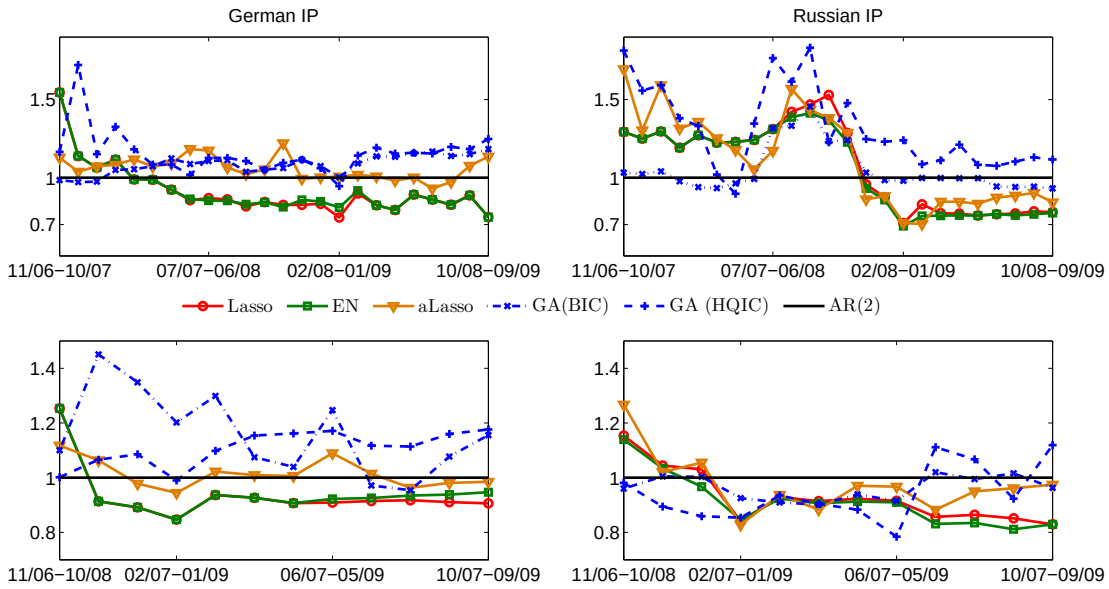

Figure 2: Forecast accuracy in relation to AR(2) with IEW (RMSFE in relation to AR(2))

5 Conclusions

Since the correct dynamic specification of time series models is often unknown, the use of model selection strategies is required. We consider two classes of model selection approaches, one based on shrinkage estimators such as Lasso and the other one – a subset selection method making use of optimization heuristics, to solve the corresponding highly complex discrete optimization problem.

A Monte–Carlo simulation is used to assess the merits of the different methods in the context of univariate autoregressive distributed lag models. The simulation setting is chosen to mimic realistic situations found in the framework of forecasting business cycles. In particular, the number of available observations is often small compared to the number of potentially relevant predictors. Due to the high persistence in many economic variables, different lags of these predictors might be highly correlated rendering the model selection problem more challenging. The results from the Monte–Carlo simulation suggest that the use of information criteria in the subset selection approach is impaired by the small number of observations, while the shrinkage estimators still perform remarkably well despite of the high correlation of potential predictors.

Furthermore, we consider to what extent a proper model selection might help to improve forecasts of business cycle indicators for Russia and Germany. While the im-

provements compared to a simple autoregressive process are small in all settings, we find again slight advantages of the shrinkage estimators.

Based on these findings, several questions emerge naturally which we will consider in future research. In particular, we will test whether larger sample sizes improve the relative performance of information criteria based selection as these criteria are asymptotically consistent. Furthermore, we will study a situation with a larger number of relevant regressors in the model. Finally, further real applications will be studied to learn about performance gains to be expected when moving away from simplistic univariate time series models.

References

Bai J, Ng S (2008) Forecasting economic time series using targeted predictors. Journal of Econometrics 146(2):304–317

Candès EJ, Tao T (2007) The Dantzig selector: statistical estimation when p is much larger than n. Annals of Statistics 35(6):2313–2351

Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. Annals of Statistics 32:407–489

Frank IE, Friedman JH (1993) A statistical view of some chemometrics regression tools. Technometrics 35(2):109–135

Gasso G, Rakotomamonjy A, Canu S (2009) Recovering sparse signals with a certain family of non-convex penalties and DC programming. IEEE Transactions on Signal Processing 57(12):4686–4698

Gilli M, Winker P (2009) Heuristic optimization methods in econometrics. In: Belsley D, Kontoghiorghes E (eds) Handbook of Computational Econometrics, Wiley, Chichester, pp 81–119

Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Verlag, New York

Hendry DF, Krolzig HM (2005) The properties of automatic "GETS" modelling. The Economic Journal 115(502):C32–C61

Kapetanios G, Labhard V, Price S (2008) Forecasting using Bayesian and information-theoretic model averaging: An application to U.K. inflation. Journal of Business & Economic Statistics 26(1):33–41

Meinshausen N, Yu B (2008) Lasso-type recovery of sparse representations for high-dimensional data. Annals of Statistics 37(1):246–270

Ozyldrim A, Schaitkin B, Zarnowitz V (2010) Business cycles in the Euro area defined with coincident economic indicators and predicted with leading economic indicators. Journal of Forecasting 29(1–2):6–28

Perez-Amaral T, Gallo GM, White H (2003) A flexible tool for model building: The relevant transformation of the inputs network approach (RETINA). Oxford Bulletin of Economics and Statistics 65(1):821–838

Savin I (forthcoming) A comparative study of the lasso–type and heuristic model selection methods. Journal of Economics and Statistics

Savin I, Winker P (2012) Heuristic optimization methods for dynamic panel data model selection. Application on the Russian innovative performance. Computational Economics 39(4):337–363

Savin I, Winker P (forthcoming) Heuristic model selection for leading indicators in Russia and Germany. Journal of Business Cycle Measurement and Analysis

Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society B 58(1):267–288

Vogt G (2007) The forecasting performance of ifo-indicators under realtime conditions. Journal of Economics and Statistics 227(1):87–101

Winker P (1995) Identification of multivariate AR-models by threshold accepting. Computational Statistics & Data Analysis 20:295–307

Zhao P, Yu B (2006) On model selection consistency of lasso. Journal of Machine Learning Research 7:2541–2563

Zou H (2006) The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101:1418–1429

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society B 67(2):301–320