University of Mannheim / Department of Economics

Working Paper Series

Forecasting VARs, Model Selection, and Shrinkage

Christian Kascha Carsten Trenkler

Working Paper 15-07

June 2015

Forecasting VARs, Model Selection, and Shrinkage [*]

Christian Kascha[†] Carsten Trenkler[‡]

University of Zurich *University of Mannheim*

June 8, 2015

Abstract

This paper provides an empirical comparison of various selection and penalized regression approaches for forecasting with vector autoregressive systems. In particular, we investigate the effect of the system size as well as the effect of various prior specification choices on the relative and overall forecasting performance of the methods. The data set is a typical macroeconomic quarterly data set for the US. We find that these specification choices are crucial for most methods. Conditional on certain choices, the variation across different approaches is relatively small. There are only a few methods which are not competitive under any scenario. For single series, we find that increasing the system size can be helpful - depending on the employed shrinkage method.

JEL classification: C32, C53, E47

Keywords: VAR Models, Forecasting, Model Selection, Shrinkage

[*] The research was supported by the Deutsche Forschungsgemeinschaft (DFG) through the SFB 884 'Political Economy of Reforms'.

[†] University of Zurich, Chair for Statistics and Empirical Economic Research, Zürichbergstrasse 14, 8032 Zurich, Switzerland; christian.kascha@econ.uzh.ch

[‡] Address of corresponding author: University of Mannheim, Department of Economics, Chair of Empirical Economics, L7, 3-5, 68131 Mannheim, Germany; trenkler@uni-mannheim.de

1 Introduction

Forecasting future realizations of economic variables is a relevant issue for many policymakers, in particular central banks, but also for economic agents in general. Nowadays, forecasters have literally hundreds of time series at their disposal to obtain predictions. Therefore, we are interested in the problem of forecasting realizations of a $K-$dimensional economic time series (y_t) given an observed sample $y_1, y_2, \ldots, y_T$, where K can be large. A common model which is used for this kind of problem is the stable vector autoregressive (VAR) model:

$$y_t = \mu + A_1 y_{t-1} + \ldots + A_p y_{t-p} + u_t, \quad \text{for } t = 1, \ldots, T, \tag{1}$$

where μ is a $(K \times 1)$ parameter vector, $A_1, \ldots, A_p$ are $(K \times K)$ parameter matrices, p fixed pre-sample values $y_{-p+1}, \ldots, y_0$ are given, and u_t is assumed to be white noise with non-singular covariance matrix Σ_u. A comprehensive treatment of the estimation, specification and various extensions of model (1) is given by Lütkepohl (2005).

Despite its popularity, a problem of this model class is the potentially large number of parameters which have to be estimated before using this model for forecasting purposes. In the general unrestricted model this number is equal to $K^2 p + K$. The estimation of many parameters leads to high estimation uncertainty which typically translates into a large mean squared forecast error (MSFE). A traditional response to this problem was to consider only small- to medium-dimensional systems with carefully chosen variables. However, these VARs do not exploit the potentially useful information contained in the many other variables not included in the VAR but which are nowadays available. It is important to see whether there are gains from incorporating this additional information. Therefore, other solutions have been developed.

The most traditional approach is *subset selection* and an overview can be found in Lütkepohl (2005). Given a suitable maximum lag order, these methods seek to find zero constraints on the parameters μ, A_1, $\ldots$, A_p of the model thereby reducing the number of parameters that have to be estimated in the final model. Computer-automated subset selection strategies can be found in the software package PcGets, see Hendry & Krolzig (2001). In practice most subset selection strategies also involve informal elements which Hoover & Perez (1999) tried to formalize in their paper. Also Brüggemann & Lüktepohl (2001) and Brüggemann (2004) investigated the performance of different subset selection methods.

Another solution is to abandon the VAR framework but stay in a linear framework and consider *factor models* (Geweke 1977) instead. This approach gained popularity due to the increasing amounts of data available to central banks and the general public. In the case of a very large number of series relative to observed time points, the above mentioned problems become very important. Factor models are advocated mainly by Stock & Watson (2002*a*,*b*), Forni, Hallin, Lippi & Reichlin (2000) and Forni, Hallin, Lippi & Reichlin (2004) for forecasting macroeconomic series. In a study similar to ours, Stock & Watson (2012) compare various methods, which are shown to asymptotically possess a shrinkage representation, to the dynamic factor model in a forecasting context with a large number of (orthogonal) predictors. In a forecasting exercise with a quarterly U.S. macro data set they find that the dynamic factor model usually outperforms pretest and information-criterion methods, Bayesian model averaging, empirical Bayes methods as well as bagging.

An alternative to factor models are *large Bayesian VARs* which impose shrinkage induced by a prior. De Mol, Giannone & Reichlin (2008) show that Bayesian estimation with shrinkage-inducing priors applied to a large cross-section of macro time series yields forecasts which are highly correlated with the principal component forecasts advocated by Stock & Watson (2002*a*,*b*). They focus on forecasting a single series with a very large panel of predictors, however. Building on the latter results, Banbura, Giannone & Reichlin (2010) show that Bayesian shrinkage can also deal with high-dimensional VARs by increasing the amount of shrinkage with the system size and find that the largest specifications outperform the small models in forecast accuracy. However, they find that a "medium-sized" system of 20 key macroeconomic indicators usually suffices for the purpose of forecasting the three key macroeconomic variables they consider. Giannone, Lenza & Primiceri (2015) generalize the choice of the priors by introducing hyper parameters and provide further evidence that a medium-sized VAR of about twenty variables improves over smaller VARs.

Apart from Bayesian Shrinkage there also exist classical *shrinkage or regularization methods* which we all summarize under the heading of *penalized regression methods*. The most well known member of this class of techniques is *ridge regression* (Hoerl 1962, Hoerl & Kennard 1970). Recently, however, other methods which combine variable selection and shrinkage have become very popular. The most prominent examples are the *lasso* (Tibshirani 1996) and the *elastic net* (Zou & Hastie 2005). These proposals are rooted in the statistical literature and are traditionally concerned with cross-sectional data rather than time series data. However, in the recent past a number of papers have studied variants of the lasso approach for time series models. Nardi &

Rinaldo (2011) and Kock (2012) study the asymptotic properties of the adaptive lasso applied to univariate autoregressive (AR) models, while Chen (2011, Chapter 4) and Medeiros & Mendes (2012) consider the autoregressive moving-average (ARMA) model and general univariate time series regressions, respectively, in relation to the adaptive lasso.

An early reference on the lasso in a multiple time series set-up is Hsu, Hung & Chang (2008). They explore small sample estimation properties and forecasting performance of lasso applied to VARs in a simulation study with systems of relatively low dimension. Kock & Callot (2014*a*,*b*) and Basu & Michailidis (2013) analyze the asymptotic properties of the lasso and adaptive lasso applied to VARs. Note that Basu & Michailidis (2013) also consider penalization in a likelihood framework in contrast to the least squares set-up on which the other approaches are based. Song & Bickel (2011) and Nicholson, Matteson & Bien (2014) consider various types of (group) lasso methods for VARs that are large in terms of dimension and/or lag order. The simulation evidence and the forecasting performance provided in the aforementioned papers demonstrate that the lasso approach can indeed be beneficial in a VAR framework relative to standard OLS but also relative to factor models; on the latter see Kock & Callot (2014*a*). Finally, Gefang (2014) studies the lasso method as well as the elastic net applied to VARs in a Bayesian context and compares these to standard Bayesian VARs in a forecasting exercise using macroeconomic series. She performed an pseudo out-of-sample forecast study based on US macroeconomic data similar to the ones used in our study. Her results demonstrate that the elastic net and lasso frameworks can lead to improved forecast performance compared to standard Bayesian VARs as considered e.g. in Banbura et al. (2010). To the best of our knowledge, Gefang (2014) is the only paper dealing with elastic net in relation to multiple time series data while Savin & Winker (2012) examine the elastic net for autoregressive distributed-lag models.

This paper adds to the VAR strand of the literature by providing a comparison of the forecast performance of various selection and penalized regression methods which is more comprehensive than the previously mentioned papers in that it includes classical and (empirical) Bayes methods, traditional subset selection methods as well as the more recent lasso and elastic net methods. In particular, we investigate for each estimation method various specification choices, i.e. we analyze the importance of prior lag selection, whether estimation should be based on rolling or expanding estimation windows and we investigate two alternative methods for choosing tuning parameters. Such specification choices have been partially addressed by Carriero, Clark & Marcellino (2015) for Bayesian VARs but have not been systematically considered for the other methods in our study. Furthermore, we are interested in the effect of the system size on the relative performance of the methods as well as on the overall forecast precision across methods. In this sense, this paper also contributes to the growing literature on large VARs.

We find that the initial lag choice is decisive for the forecasting performance of most methods. Also, expanding estimation windows and the use of information criteria for the choice of the tuning parameters are generally preferred. Conditional on these choices, the variation across methods is small - apart from some approaches which are not competitive such as traditional selection methods. With respect to the system size, we find that increasing the dimension of the system can be advantageous for some methods when forecasting single series.

The rest of the paper is organized as follows. Section 2 presents the investigated forecasting methods. Section 3 explain the pseudo out-of-sample exercise undertaken and discusses the results. Section 4 concludes.

2 Forecasting Methods

We call combinations of VAR specification and estimation methods simply (VAR) *forecasting methods*. For all methods, we assume that there is an upper bound on the possible lag order of the system, $p_{\max}$. Thus, a generalized model of (1) is considered

$$y_t = \mu + A_1 y_{t-1} + \ldots + A_{p_{\max}} y_{t-p_{\max}} + u_t, \quad \text{for } t = 1, \ldots, T. \tag{1'}$$

Every method consists of two stages. First, the above model might be reduced by an initial model specification step leading to a VAR with only $p \leq p_{max}$ lags. Second, the model found in the first stage is estimated by one of the selection or penalized regression methods.

2.1 Initial Model Specification

For the initial model specification, we consider either simply selecting the highest, a priori given, order $p_{\max}$ or choosing the lag order via standard information criteria, see Lütkepohl (2005, Ch. 4). The criteria we consider in this stage are

AIC: The order p is selected by minimizing the Akaike information criterion (AIC)

$$AIC(m) = \ln|\widetilde{\Sigma}_u(m)| + \frac{2}{T} m K^2$$

over $m \in \{1, \dots, p_{\max}\}$, where $\widetilde{\Sigma}_u(m)$ is the maximum likelihood (ML) estimator of Σ_u obtained by fitting an *unrestricted* VAR(m) model.

BIC: The order p is selected by minimizing the Bayesian information criterion (BIC)

$$BIC(m) = \ln|\widetilde{\Sigma}_u(m)| + \frac{\ln T}{T} m K^2$$

over $m \in \{1, \dots, p_{\max}\}$, by fitting again an *unrestricted* VAR(m) model.

In order to get get a "reasonably" sized VAR, we let $p_{\max}$ vary according to the dimensions of the systems we consider in our forecast study. Specifically, we set $p_{\max} = 8$ for systems up to dimension $K = 3$, $p_{\max} = 7$ for $K = 4$, $p_{\max} = 6$ for $K = 7$ and $p_{\max} = 3$ for $K = 22$.

2.2 Estimation Methods

Given the specified lag order p, the resulting VAR(p) can be estimated with different selection or penalized regression methods. We consider here standard *subset selection methods* such as *top down selection* and a *single equation testing procedure*. For the penalized regression methods, we investigate *ridge regression*, a *Bayesian VAR*, the *lasso*, a *single equation lasso* method as well as the *elastic net*.

In order to discuss the above methods in a common framework, we use the following notation - and variations thereof. Denote by $A := [\mu\ A^0]$, where $A^0 := [A_1, \dots, A_p]$, and its vectorized version by $a := \text{vec}(A)$. Write the model as $y_t = [\mu\, A_1 \dots A_p] Z_{t-1} + u_t$ for $t = 1, \dots, T$, where $Z_{t-1} = (1,\ Z^{0'}_{t-1})'$ with $Z^0_{t-1} = (y'_{t-1}, \dots,\ y'_{t-p})'$. Then,

$$y \quad = \quad (Z' \otimes I_K)a + u = Xa + u,$$

where $y = \text{vec}([y_1, \dots, y_T])$, $Z = [Z_0, \dots, Z_{T-1}]$, $u = \text{vec}([u_1, \dots, u_T])$, and $X = (Z' \otimes I_K)$. Throughout, I_d denotes the identity matrix of dimension d and, likewise, 0_d denotes a vector of zeros of dimension $d \times 1$. Each of the methods yields a parameter estimator $\hat{a}$ which is used in the standard (iterative) way to yield $h-$step-ahead forecasts.

2.2.1 Top Down Selection (TopDown)

This description follows to a large extent Lütkepohl (2005) which should be consulted for further details. Restrictions are found equation by equation. Given a lag order p, the coefficients of the

k-th equation are *ordered* first by time and then by the order in which the respective variables enter the VAR equations, i.e,

$$\begin{aligned} y_{kt} &= \mu_k + \alpha_{k1,1} y_{1,t-1} + \ldots + \alpha_{kK,1} y_{K,t-1} + \\ &\vdots \\ &+ \alpha_{k1,p} y_{1,t-p} + \ldots + \alpha_{kK,p} y_{K,t-p} + u_{kt}, \end{aligned} \tag{2}$$

where $\alpha_{ij,l}$ is the coefficient in the (i,j)-position of the matrix A_l. A subset model is found by imposing zero restrictions sequentially until no further improvement can be achieved in terms of minimizing a pre-specified information criterion. In each step, one coefficient is deleted according to the order above - starting with the last coefficient, $a_{kK,p}$, up to the intercept - and the value of the criterion is evaluated and compared to the value of the criterion when the equation is estimated with the coefficient. If the criterion is improved, the zero restriction is maintained and the next variable is tried given the restrictions imposed from the previous steps - until all coefficients have been tried.

For concreteness, the exact definitions and steps are as follows. Define $a_k = (\mu_k, \alpha_{k1,1}, \ldots, \alpha_{k1,p}, \ldots, \alpha_{kK,p})'$, $y_{(k)} = (y_{k1}, \ldots, y_{kT})'$ and $u_{(k)} = (u_{k1}, \ldots, u_{kT})'$. Then, the k-th equation can be written as

$$y_{(k)} = Z' a_k + u_{(k)}.$$

Potential zero restrictions are formulated as $a_k = R_k \gamma_k$ as in Lütkepohl (2005, Chapter 5) for a suitable $((K \cdot p+1) \times n_\gamma)$ restriction matrix R_k and a $(n_\gamma \times 1)$ vector of free parameters γ_k. Denote the restricted least squares estimator by $\hat{\gamma}(R_k) = (R_k' Z\, Z' R_k)^{-1} R_k' Z y_{(k)}$ and the ML estimator of the variance by $\tilde{\sigma}^2(R_k) = (y_{(k)} - Z'\, R_k \hat{\gamma}(R_k))'(y_{(k)} - Z'\, R_k \hat{\gamma}(R_k))/T$. The resulting information criterion is $CRIT(R_k) = \ln \tilde{\sigma}^2(R_k) + C_T \mathrm{rk}(R_k)$ with either $C_T = \frac{2}{T}$ (AIC) or $C_T = \frac{2 \ln T}{T}$ (BIC).

The algorithm for the $k-$th equation can then be described as follows:

1. Initialize R_k to be the identity matrix $R_k^{(0)} = I_{K \cdot p+1}$, set $i = 1$
2. For j running from $(K \cdot p + 1)$ to 1
   (a) Form $R_k^{(i)}$ by deleting the jth column of $R_k^{(i-1)}$.
   (b) Compute $\hat{\gamma}(R_k^{(i)})$ and $CRIT(R_k^{(i)})$.

(c) Set

$$R_k^{(i)} = \begin{cases} R_k^{(i)}, & \text{if } CRIT(R_k^{(i)}) < CRIT(R_k^{(i-1)}), \\ R_k^{(i-1)}, & \text{if } CRIT(R_k^{(i)}) \geq CRIT(R_k^{(i-1)}). \end{cases}$$

(d) Set $i = i + 1$

After the determination of the restrictions for all k equations, call the implied restriction matrix for the whole system R such that all zeros restrictions on a can again be formulated as $a = R\gamma$ and we can can write $y = (Z' \otimes I_K)R\gamma + u$ for all equations jointly. The corresponding estimated generalized least squares (EGLS) estimator is $\hat{\gamma}(R) = (R'(Z\,Z' \otimes \widehat{\Sigma}_u^{-1})R)R'(Z \otimes \widehat{\Sigma}_u^{-1})y$, where $\widehat{\Sigma}_u$ is computed from an unrestricted least squares estimator and the top down estimator of a is $\hat{a}_{TD} = R\hat{\gamma}(R)$. Depending on which information criteria is used, the variants of this selection method are labeled *AIC TopDown* or *BIC TopDown*.

2.2.2 Single Equation Testing Procedure (TP)

The description of this selection strategy follows closely the one in Lütkepohl (2005, section 5.2). This procedure is applied again to each equation (2) of the VAR such that regressors are sequentially deleted, one at a time, according to which regressor has the smallest t-ratio. Then, new t-ratios are computed for the reduced model. One stops when all t-values are greater than some threshold value η.

This procedure is computationally much less expensive than an alternative method proposed by Brüggemann & Lüktepohl (2001) that sequentially eliminates those regressors which lead to the largest reduction in a pre-specified information criterion. Both methods are equivalent provided the threshold is chosen such that $\eta = \{[\exp(c_T/T) - 1] \cdot (T - N + j - 1\}^{1/2}$ at the j-th step of the elimination procedure. In this study, we use $c_T = 2$ (AIC) and $c_T = \ln T$ (BIC) and label the corresponding test procedures by *AIC TP* and *BIC TP*, respectively.

After the determination of the restrictions for all K equations, we again collect these and estimate the whole system by EGLS.

2.2.3 Ridge Regression (RR)

Ridge regression goes back to Hoerl (1962) and Hoerl & Kennard (1970). For the implementation, we first standardize the time series variables by subtracting their means and dividing by their standard deviations. Generally, we denote standardized variables by placing a $\sim$ on top of the respective symbols. However, we keep on using the same notation for the corresponding parameters and only use $\sim$ to indicate parameter *estimators* based on the standardized variables. Mean-adjusting the time series provides us with an easy way of excluding the intercept term from penalization. The scaling is motivated by the fact that a common shrinkage parameter is used with respect to all parameters.

Let $a^0 = vec(A^0)$ denote the parameter vector for the standardized model without the intercept. Let C be a selector matrix and denote by c a vector towards which the parameter estimates are shrunk such that deviations from the restriction $Ca^0 = c$ are penalized. Further, let $\tilde{X}^0$ denote the regressors matrix with the standardized variables (without an intercept). Then, the ridge regression can be formulated as

$$\min_{a^0} (\tilde{y} - \tilde{X}^0 a^0)'(\tilde{y} - \tilde{X}^0 a^0) + \lambda \cdot (Ca^0 - c)'(Ca^0 - c).$$

We consider two variants of ridge regression which we label *all* and *allbutdiag*. The first variant restricts all coefficients in $A_1, \dots, A_p$ towards zero and consequently $C = I_{K^2 \cdot p}$, $c = 0_{K^2 \cdot p}$. The second variant restricts all coefficients but the diagonal elements in A_j, $j = 1, \dots, p$. Consequently, the columns of the previously defined matrix C that correspond to the diagonal elements are removed and $c = 0_{K^2 \cdot p - K \cdot p}$.

The above problem leads to the solution

$$\tilde{a}^0_{RR} = \left((\tilde{X}^{0\prime} \tilde{X}^0) + \lambda (C'C) \right)^{-1} \left(\tilde{X}^{0\prime} \tilde{y} + \lambda C'c \right). \tag{3}$$

Then, the parameter estimator in terms of the original scaling is recovered. This re-scaling is also applied to all other estimation approaches which are based on standardized data.

The choice of the penalty parameter λ is important. It has been mentioned in the literature that traditional cross-validation is less suited for tuning parameter selection in a time-dependent framework, see e.g. Medeiros & Mendes (2012), Nicholson et al. (2014).[1] We employ two alternative methods that have been advocated in the recent literature on penalization in (vector) time series models: λ is either determined by evaluating the fit via information criteria (AIC or BIC) or by evaluating the predictive MSE (PMSE) over the last 20% of observations for a grid of possible values for λ.

For the choice of the grid for λ, we adopt the approach of Friedman, Hastie & Tibshirani (2010)

[1] Kock & Callot (2014*b*), however, have noted that cross-validation performs similarly to the BIC in their simulation study on adaptive and standard lasso estimation of VAR models. However, they found cross-validation to be considerably slower than the BIC. We arrived at similar conclusions in a small pilot forecasting study on our data.

for determining a grid for the lasso penalty parameter.[2] Denote the grid by $\lambda^{(i)},\ i = 1,\dots,G$. First, we determine the highest value, $\lambda^{(G)}$, by $||C\tilde{a}^0_{RR}(\lambda^{(G)})||_2 = 0.1 \cdot ||C\tilde{a}^0_{OLS}||_2$, where $\tilde{a}^0_{OLS}$ is the OLS estimator of a^0 based on the standardized variables and $||\cdot||_2$ is the Euclidean norm. That is, we choose $\lambda^{(G)}$ such that the resulting estimated parameter vector is "small". Then, we set the minimum value to $\lambda^{(1)} = 0.001 \cdot \lambda^{(G)}$ and construct a sequence of 100 values of λ linearly decreasing from $\lambda^{(G)}$ to $\lambda^{(1)}$ - on the log scale.[3] Hence, we have $G = 100$.

The BIC has been used e.g. by Kock & Callot (2014*a*,*b*) and by Medeiros & Mendes (2012). Wang, Li & Tsai (2007) have shown the asymptotic validity of the BIC for tuning parameter specification for lasso estimation of a univariate time series regression with an error term following an AR process. The use of the information criteria works as follows. For each value $\lambda^{(i)}$ of the grid, the information criteria are computed as

$$IC(\lambda^{(i)}) = \ln|\widetilde{\Sigma}_u(\lambda^{(i)})| + C_T \times dof(\lambda^{(i)}), \tag{4}$$

where either $C_T = 2/T$ (AIC) or $C_T = \ln T/T$ (BIC) and $\widetilde{\Sigma}_u(\lambda^{(i)})$ is the usual ML estimator of the covariance matrix computed on the standardized data. The degrees of freedoms are obtained as $dof(\lambda) = \text{tr}\left(\tilde{X}^0\left(\tilde{X}^{0\prime}\tilde{X}^0 + \lambda C'C\right)^{-1}\tilde{X}^{0\prime}\right)$, see Bühlmann & van de Geer (2011, Sect. 2.11). The $\lambda^{(i)}$ that minimizes $IC(\lambda)$ is chosen. For all methods, we employ the same information criterion for the penalty selection that was used for the initial model selection. If $p = p_{\max}$ is chosen, we use the BIC. This is done for simplicity. In principle, we could have tried more combinations.

For the alternative approach using the predictive MSE, we divide the available data points into two parts: an estimation subsample consisting of $\tilde{z}^{(1)} = [\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_{\bar{t}}]$ and an evaluation subsample consisting of $\tilde{z}^{(2)} = [\tilde{y}_{\bar{t}+1}, \tilde{y}_{\bar{t}+2}, \dots, \tilde{y}_T]$ with $\bar{t} = [0.8T]$, where $[x]$ denotes the smallest integer larger than or equal to x. The corresponding ridge estimator of A^0 computed from the estimations sample for fixed λ is denoted by $\tilde{A}^{(1)}_{RR}(\lambda)$. For each candidate $\lambda^{(i)},\ i = 1,\dots,G$, the predictive MSE criterion is computed as

$$PMSE(\lambda^{(i)}) = \text{vec}\left(\tilde{z}^{(2)} - \tilde{A}^{(1)}_{RR}(\lambda^{(i)})\tilde{Z}^{0(2)}\right)' \text{vec}\left(\tilde{z}^{(2)} - \tilde{A}^{(1)}_{RR}(\lambda^{(i)})\tilde{Z}^{0(2)}\right),$$

where $\tilde{Z}^{0(2)} = [\tilde{Z}^0_{\bar{t}}, \tilde{Z}^0_{\bar{t}+1}, \dots, \tilde{Z}^0_{T-1}]$. The $\lambda^{(i)}$ that minimizes $PMSE(\lambda^{(i)})$ is chosen.

The foregoing approach has also been used by Song & Bickel (2011) and Nicholson et al.

[2] Their original approach is briefly outlined in subsection 2.2.5.

[3] That is, we compute 100 equidistant points between $\ln(\lambda^{(1)})$ and $\ln(\lambda^{(G)})$ and then take the exponent of these values.

(2014) for the determination of the tuning parameters of the lasso method applied to a VAR model setup. It can be interpreted as a simplified version of cross-validation that respects the time series nature of the data. Since the observations are kept together with their lags, the predictive MSE approach boils down to a sequence of one-step ahead forecasts for the evaluation subsample using fixed parameter estimates, i.e. the parameters are not re-estimated after proceeding one period in the evaluation subsample.

2.2.4 Bayesian Shrinkage (Bayes)

This approach is similar to a "truly" Bayesian approach of Doan, Litterman & Sims (1983) and Litterman (1986) in so far as the resulting point estimator is the posterior mean of the VAR parameters given the so called Minnesota or Litterman prior. Conceptually, this strategy is very similar to ridge regression and we frame our "Bayesian" strategy in terms of an optimization problem in order to discuss all approaches within a common framework. The estimation problem becomes

$$\min_a (y - Xa)'(I_T \otimes \Sigma_u^{-1})(y - Xa) + \lambda \cdot (Ca - c)'V_a^{-1}(Ca - c)$$

where λ and the matrix $V_a = \text{diag}(v_{11,1}, v_{21,1}, \dots, v_{11,2}, \dots, v_{KK,\hat{p}})$ with

$$v_{ij,l} = \begin{cases} (1/l)^2, & \text{if } i = j \\ (\theta\sigma_i/l\sigma_j)^2, & \text{if } i \neq j, \end{cases} \tag{5}$$

determines the tightness of the prior information, see Lütkepohl (2005, 7.4.2). The matrix $C = [\mathbf{0}_{(K^2 \cdot p) \times K}\ I_{K^2 \cdot p}]$ selects all parameters but the intercepts and c is a $(K^2 \cdot p \times 1)$ vector towards which a is shrunk. In the case of differenced data c is just a vector of zeros. In the case of levels data c is usually such that the VAR is shrunk towards a $K-$dimensional random walk. The solution is given by

$$\hat{a}_{BA} = \left((X'(I_T \otimes \Sigma_u^{-1})X) + \lambda(C'V_a^{-1}C)\right)^{-1} \left(X'(I_T \otimes \Sigma_u^{-1})y + \lambda C'V_a^{-1}c\right).$$

Note that in most descriptions of the prior, λ, as defined here, is actually λ^{-2} but we decided to keep the interpretation of λ as a parameter which is positively related to the strength of the restriction. The above formula as well as parts of the prior contain the unknown covariance Σ_u. In order to apply the above formula, we use the OLS estimator, $\hat{\Sigma}_u$, in its place.

A specific pair (λ, θ) is determined via the information criteria or via a comparison of predictive MSEs. The procedure is analogous to the one described in the section on ridge regression with a few modifications. First, non-standardized data are used and, second, the degrees of freedom used to compute the information criteria are computed analogous to (4) as

$$dof(\lambda, \theta) = \mathrm{tr}\left(X'(I_T \otimes \hat{\Sigma}_u^{-1})X \left[X'(I_T \otimes \hat{\Sigma}_u^{-1})X + \lambda C' V_a^{-1} C \right]^{-1} \right).$$

For the determination of a suitable grid, we let θ take on values in $\{0.2, 0.4, 0.6, 0.8, 1\}$ since this parameter just determines the relative shrinkage of the diagonals versus the off-diagonal elements, see Lütkepohl (2005). The grid for λ is obtained analogous to case of ridge estimation, holding $\theta = 1$ fixed.

2.2.5 Lasso Regression (Lasso)

The lasso procedure was originally proposed by Tibshirani (1996). The approach minimizes the sum of squared residuals subject to an L_1 penalty on the size of the estimated coefficients. In contrast to ridge regression that only continuously shrinks coefficients towards zero, parameter estimates can become zero if the penalty parameter λ is large enough. Hence, lasso allows for sparse solutions and, thereby, also performs model selection. Indeed, lasso is particularly useful for models with many coefficients which are close to zero and a small number of coefficients that are relatively large.

Using the introduced notation, the optimization problem underlying lasso is formulated in terms of standardized variables and is given by

$$\min_{a^0} \, (\tilde{y} - \tilde{X}^0 a^0)'(\tilde{y} - \tilde{X}^0 a^0) + \lambda \cdot ||a^0||_1, \tag{6}$$

where $||a^0||_1 = \sum_{i,j,l} |\alpha_{ij,l}|$ is the L_1-norm of the parameter vector. There exists no closed-form solution for the Lasso estimator but (6) can be solved numerically in order to obtain the estimator $\tilde{a}^0_{LA}$. To this end, we use the `glmnet` package for Matlab, see Friedman et al. (2010). This package is designed to solve elastic net minimization problems of which the lasso problem (6) is a special case. The general elastic net minimization problem is described in subsection 2.2.7.

Again, the penalization parameter λ is chosen by optimizing different criteria for a grid of possible values - starting at the smallest value $\lambda^{(G)}$ for which the entire vector $a^0 = 0$. (Friedman et al. 2010). The minimum value of the grid is chosen such that $\lambda^{(1)} = 0.001 \cdot \lambda^{(G)}$. Then, we proceed as described for the ridge regression set-up, i.e. setting $G = 100$, a sequence from $\lambda^{(G)}$ to $\lambda^{(1)}$ is constructed that linearly decreases on the log scale. Again, we have applied information criteria (AIC, BIC) and the predictive MSE criterion analogous to the procedure for the ridge regression set-up in order to determine the tuning parameter λ. Note, however, that we use the number of non-zero coefficients of the lasso solution for a given penalty λ to measure the degrees of freedom needed for computing the information criteria, see e.g. Zou, Hastie & Tibshirani (2007) and Bühlmann & van de Geer (2011).

2.2.6 Single Equation Lasso Regression (Lasso SE)

The foregoing version of lasso ignores the structure of the VAR model. Song & Bickel (2011) have suggested a version of lasso that considers each of the K equations of the VAR separately and, in terms of penalization, distinguishes between the different lags as well as between the variables' own lags and the corresponding other variables' lags. This "no grouping" version is their preferred lasso variant and shares some of the ideas of the Bayesian shrinkage approach discussed above.

To illustrate the approach, consider the k-th equation of the VAR in terms of the standardized variables but still using the same notation as in 2.2.1. Hence, a_k is the parameter vector of the k-th equation. Denote by $\alpha_{kk,i}$ the element of a_k that belongs to the k-th variable at lag i and denote by $\alpha_{kj,i}$ the parameters that belong to the other regressors' - at lag i. The optimization problem for this equation is

$$\min_{a_k} (\tilde{y}_{(k)} - \tilde{Z}'a_k)'(\tilde{y}_{(k)} - \tilde{Z}'a_k) + \lambda_k \left(\theta_k \sum_{i=1}^{p} i^{\nu_k} |\alpha_{kk,i}| + \sum_{i=1}^{p} i^{\nu_k} \sum_{j \neq k} |\alpha_{kj,i}| \right).$$

The penalty of the parameters associated with lag i is scaled with i^{ν_k} such that the coefficients of higher-order lags are penalized more strongly if $\nu_k > 0$. The value $\nu_k = 2$ would correspond to the Bayesian shrinkage approach, compare equation (5). Similarly, the penalty of the parameters associated with the k-th variable are scaled with θ_k whereby $\theta_k < 1$ assures that these are less strongly penalized than the parameters associated with the other variables. The scaling factor θ_k has the same function as in the Bayesian shrinkage approach.

However, in the current set-up θ_k as well as ν_k and λ_k are individually determined for each equation k, for details see Song & Bickel (2011). The grids for ν_k and θ_k are chosen as follows: ν_k can take on values in $\{0, 1, 2\}$ and θ_k in $\{0.1, 0.2, \ldots, 0.9, 1\}$. Conditional on a particular pair

(ν_k, θ_k), we determine the grid for λ_k as for the general lasso.

We use the univariate versions of the information criteria and the PMSE criterion to jointly select the three tuning parameters for each equation separately, searching over the three-dimensional product space obtained from the three individual grids.

2.2.7 Elastic Net Regression (Elastic Net)

This approach has been suggested by Zou & Hastie (2005). The elastic net combines the L_1 and L_2 penalties used in the lasso approach and ridge regression, respectively. This combination is motivated by some problems from which the lasso approach suffers in the case of correlated regressors. Ridge regression shrinks coefficients of correlated regressors towards each other introducing the so-called grouping effect. By contrast, the lasso tends to pick one of the regressors and ignores the rest of them. Moreover, lasso may show weird behavior in case of extreme correlations. In fact, it breaks down in case of perfect regressor correlation. Accordingly, lasso cannot pick more variables than observations are available. See Zou & Hastie (2005) for more details on the latter issues.

The joint consideration of the penalties introduces both automatic model selection and the described grouping effect. Thereby, it is hoped that the elastic net performs as well as lasso whenever lasso works well but fixes the highlighted problems of lasso, compare Zou & Hastie (2005).

Following Friedman et al. (2010), the relevant minimization problem for the standardized variables can be written as

$$\min_{a^0} \, (\tilde{y} - \tilde{X}^0 a^0)'(\tilde{y} - \tilde{X}^0 a^0) + \lambda \left(\alpha ||a^0||_1 + 0.5 \cdot (1-\alpha) a^{0\prime} a^0 \right) \tag{7}$$

The last term is the elastic net penalty leading to the classical ridge penalty if $\alpha = 0$ and the lasso problem if $\alpha = 1$. For a fixed value of α, (7) can be interpreted as a re-scaled lasso minimization problem.

Using again the `glmnet` package for Matlab to numerically solve the minimization problem (7) we obtain the elastic net solutions for a fixed value of α. We let α take on values in $\{0, 0.05, 0.10, \ldots, 0.9, 0.95, 1\}$ to create a grid for α. As regards λ, we set up a grid in the same way as for lasso conditional on a particular value for α. We jointly determine the values for λ and α by searching over the Cartesian product of the two grids using the information criteria as well as the predictive MSE criterion analogous to the case of the ridge regression. However, the degrees of freedom for the computation of the information criteria have to be adopted. Following Tibshirani & Taylor (2012), we set $dof(\lambda, \alpha) = \operatorname{tr}\left(\tilde{X}^0_{\mathcal{A}} \left[\tilde{X}^{0\prime}_{\mathcal{A}} \tilde{X}^0_{\mathcal{A}} + \lambda \cdot 0.5 \cdot (1-\alpha) I\right]^{-1} (\tilde{X}^0_{\mathcal{A}})'\right)$, where $\tilde{X}^0_{\mathcal{A}}$ is formed from the columns of $\tilde{X}^0$ associated with non-zero coefficients - given λ, α. Accordingly, I is an identity matrix of dimension equal to the cardinality of the active set.

3 Empirical Forecast Comparison

3.1 Data and Setup

We empirically compare the selection and penalized regression methods using different systems of quarterly US macroeconomic data. The data are taken from the Federal Reserve Bank of St. Louis and the respective series IDs are given in table 1. Our data sample spans from 1959 Q1 to 2012 Q2. Thus, we have $\bar{T} = 214$ quarterly observations. The first 102 observations from 1959 Q1 to 1984 Q2 are used for initial estimation.[4] That is, the first 1-step-ahead forecast is for 1984 Q3 and so on.

We follow Carriero et al. (2015), Stock & Watson (2008) and Koop & Korobilis (2013) by only considering variables transformed to stationarity. The composition of the VARs is inspired by Giannone et al. (2015) and Koop & Korobilis (2013) and similar papers in the area. We consider here a variety of systems ranging from very small systems to a system that comprises 22 variables. This setup allows us to evaluate the effect of increasing the system size on the forecast performance of the methods. The used transformations as well as the composition of the VARs are given in table 1 in the appendix.

Most results in the tables are given for expanding estimation windows. However, to account for structural breaks we consider in addition rolling estimation windows of size 100 which corresponds to 25 years. While one can think of more elaborate ways of dealing with structural breaks, the focus of this paper is different and the results in Bauwens, Koop, Korobilis & Rombouts (2015) indicate that for MSFE comparisons rolling estimation windows have reasonable forecasting performance in the presence of structural breaks.

We will measure the forecasts' precision in two ways. First, we are interested in a performance measure for the whole system. Second, we take a closer look at three variables which are important from an economic point of view.

[4] Since some variables have to be differenced twice, this ensure that there is always a minimum of 100 sample observations for estimation.

To measure system-wide performance we use the generalized forecast second moment criterion proposed by Clements & Hendry (1993). It is defined as follows for a maximum horizon h

$$\mathrm{GFESM}_h = \left(\det(E[\mathrm{vec}(e_1, \dots, e_h)\mathrm{vec}(e_1, \dots, e_h)'])\right)^{1/(K \cdot h)}, \tag{8}$$

where e_j is $(K \times 1)$-dimensional j-th-step-ahead forecast error, $j = 1, \dots, h$. We take $h = 4$ in the following. The main advantage of this measure is that it is invariant to non-singular, scale-preserving linear transformations of the variables, see Clements & Hendry (1993) for details.

Second, as it is common in the literature, we consider the root mean squared forecast error (RMSFE) for the following three variables: annualized real GDP growth, $400\Delta \ln \mathrm{rgdp}_t$, annualized inflation as measured by the GDP deflator, $400\Delta \ln \mathrm{pgdp}_t$, and the (raw) federal funds rate in levels i_t.

3.2 Results

The tables 2 and 3 show the results on the overall forecasting performance as measured by the GFESM measure (8) for different forecasting methods. For all tables, each row displays the results for a combination of an initial model specification step and an estimation method. For example, the fourth row "AIC Lasso SE" shows the result for an initial model choice with AIC followed by the application of the lasso single equation method. Table 2 contains the GFESM measures obtained when using the predictive MSE for choosing the tuning parameters while table 3 gives the corresponding results when using an information criterion.[5] More precisely, the tables contain percentage differences of the GFESM measure of the listed methods relative to the forecasts from a benchmark VAR(0), that is, a model which only contains an intercept. Negative values indicate an improvement over this benchmark. For example, a value of -0.02 means that the GFESM measure of the particular methods is 2 % smaller than the GFESM measure of the benchmark. In addition, the six lowest numbers are marked bold. We chose to first present in tables 2 and 3 the results for the case of expanding estimation windows. Later on, we comment on tables 4 and 5 that show the performance of the methods for expanding estimation windows relative to rolling estimation windows.

The unconstrained autoregression performs well for the smallest system but their precision

[5] Note that the GFESM measures are the same for the unconstrained VARs and the VARs specified via subset selection methods because they do not depend on tuning parameters.

deteriorates when the system size increases. The AIC in particular does not do better than the benchmark "unconditional" VAR(0) forecast and it does particularly bad when the system size is maximal.

When the VAR is estimated by a *single equation lasso* method, the forecast performance is mixed and depends on the system it is applied to as well as on the initial model selection method and on the method for selecting the tuning parameter. Generally, however, the single equation lasso method works better when the BIC is used for the initial model selection. In this case, the method works generally better than the benchmark.

Applying the *lasso* to the estimation of the *entire* system is not in all cases better than the single equation approach. However, it clearly outperforms the *single equation lasso* method when the system dimension is large (System V). The method performs best when combined with the BIC at the initial model selection stage.

When the VAR is estimated via the *elastic net* method, the resulting forecasts are in general as precise as the forecasts resulting from the lasso. Also in this case, the initial model selection stage is important and using BIC is preferable.

VARs estimated with *ridge regression* yield forecasts whose precision depends on a number of factors. First, restricting *all* coefficients or only the off-diagonal coefficients (*allbutdiag*) is not decisive. Often the simpler variant that restricts all coefficients seems, however, slightly preferable but not by a large amount. Second, the initial model selection step is still important with similar results as in the previous cases. Excluding the generally inferior results for the VARs that use the maximal lag length (*Pmax*), the VARs with ridge regression appear to perform worse or not much better than the VARs estimated with the lasso or elastic net techniques.

When the VARs are estimated with the empirical *Bayesian* estimation method, the resulting forecasts are generally more precise than the benchmark - with very few exceptions. Furthermore, the forecast precision varies much less over different initial model selection methods. This could be a consequence of the particular shrinkage that penalizes long lags more than shorter lags. While the VARs estimated with the lasso or the elastic net methods yield similarly precise estimates they only do so when the BIC is used initially.

Using some of the more *traditional selection methods* seems less advantageous relative to the benchmark. The forecasts' precision can be very bad when the initial model selection and the selection method is too "liberal", in the sense of allowing for too many non-zero parameters. However, if one uses the BIC at the initial stage together with a *top down* or *testing* procedure using the BIC as well, the resulting forecasts are usually more precise than the benchmark.

From table 2, one can see that it is often more important how the VAR lag length is chosen rather than which specific shrinkage or regularization method is used. It turns out that using the BIC for initial model specification is typically the best choice. Table 3 shows that applying the BIC also for selecting the tuning parameters additionally increases forecast precision compared to the AIC. This applies in particular in case of large systems. We have obtained corresponding findings in relation to the individual series which we discuss below.

The tables 4 and 5 contain relative GFESM numbers for comparing methods which rely on an expanding estimation window versus methods which rely on rolling estimation windows. The results represent percentage differences in GFESM such that negative entries indicate that the expanding estimation window is preferable. The six largest values (in modulus) are marked bold. In general the expanding estimation window is advantageous, sometimes quite clearly, no matter whether the PMSE (table 4) or an information criterion (table 5) is used to select the tuning parameters.

The tables 6 and 7 contain relative GFESM numbers for comparing methods which specify the tuning parameter via the PMSE or via the use of information criteria. Negative entries indicate that the PMSE is advantageous. Again, the six largest values (in modulus) are marked bold. No obvious pattern is seen. However some larger positive values indicate that the PMSE approach might not be overly stable and occasionally leads to unfavorable choices.

The results for the *single series* are given in the tables 8 - 10. We consider real GDP growth, the GDP deflator and the federal funds rate. This corresponds to the three key macroeconomic variables considered by Banbura et al. (2010).[6] The displayed numbers are the estimated RMSFEs for different horizons. The six lowest numbers are marked bold. Not all results for the whole system carry over to the single series. However, it is generally preferable to use BIC at the initial model selection stage, i.e. typically to choose a small lag length. Also the empirical Bayesian method performs reliably well over different settings. Relative to forecasts from a benchmark VAR(0) the improvements are typically confined to short forecast horizons.

Whether an increase in the system size is beneficial depends on the series as well as the employed penalized estimation and selection methods. It can help for example for real GDP growth when a shrinkage method like lasso or the elastic net is applied while for the other series the shrinkage rather ensures that the results do not worsen much and sometimes improve. Note

[6] As Banbura et al. (2010) rely on monthly data, they use employment as an indicator of real economic activity rather than real GDP growth. Moreover, they consider the consumer price index.

that increasing the system size is, in relative terms, more beneficial at short forecast horizons than at longer horizons.

Generally, the improvement in forecast performance due to an increase of the system size are less pronounced in comparison to the Bayesian VAR set-up of Banbura et al. (2010). Yet, some results are rather similar. First, forecasts on a macroeconomic price index seem to profit the least when considering a medium-sized VAR with about 20 variables. Second, the reduction in the RMSFE for our *Bayes* methods regarding the federal funds rate observed when extending system IV ($K = 7$) to the larger VAR system V ($K = 22$) corresponds very well to the respective findings of Banbura et al. (2010) on their similar sized VARs. Note, however, that Banbura et al. (2010) use monthly data for a period that just runs until 2003.

Overall, we can derive the following three main results. First, it is most beneficial to use BIC for initial model specification and for deciding on the values for the tuning parameters. Indeed, model and tuning parameter specification is often more important than the issue of which selection or penalized regression method should be applied. The fact that BIC is preferred is an indication that using AIC may lead to an in-sample over-fit which is negatively correlated with out-of-sample forecast performance. Second, as an exception, the empirical Bayesian approach is relatively robust to different initial model specification methods. Third, the relative performance of the lasso approach (in combination with BIC) clearly improves with the size of the system. As a consequence, it is often the best approach for the large system V in terms of absolute forecast performance.

4 Conclusion

In this paper, we compared the forecasting performance of some traditional and some newly proposed selection and penalized regression methods for estimating small to medium-sized VARs. The comparison was conducted with quarterly US macroeconomic data. For this data set, we found that some specification choices such as the overall lag order can be more important than the choice of the estimation or selection method. That said, subset selection methods did not perform very well for our data sets, while the other methods yielded comparable forecasts. We also confirm the results in the previous literature that increasing the dimension of the VAR can be beneficial provided that some shrinkage is applied to account for the quickly increasing number of parameters.

References

Banbura, M., Giannone, D. & Reichlin, L. (2010), 'Large Bayesian vector auto regressions', *Journal of Applied Econometrics* **25**, 71–92.

Basu, S. & Michailidis, G. (2013), Estimation in high-dimensional vector autoregressive models, arXiv: 1311.4175v1.
URL: *http://arxiv.org/abs/1311.4175*

Bauwens, L., Koop, G., Korobilis, D. & Rombouts, J. V. (2015), 'The contribution of structural break models to forecasting macroeconomic series', *Journal of Econometrics,* **forthcoming**.

Brüggemann, R. (2004), *Model Reduction Methods for Vector Autoregressive Processes*, Springer Verlag.

Brüggemann, R. & Lüktepohl, H. (2001), *Econometric Studies: A Festschrift in Honour of Joachim Frohn*, LIT Verlag, chapter Lag Selection in subset VAR models with an application to a U.S. monetary system, pp. 107–128.

Bühlmann, P. & van de Geer, S. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer Verlag: New York.

Carriero, A., Clark, T. & Marcellino, M. (2015), 'Bayesian VARs: specification choices and forecast accuracy', *Journal of Applied Econometrics* **30**, 46–73.

Chen, K. (2011), Regularized multivariate stochastic regression, PhD thesis, University of Iowa.

Clements, M. P. & Hendry, D. F. (1993), 'On the limitations of comparing mean squared forecast errors', *Journal of Forecasting* **12**, 617–637.

De Mol, C., Giannone, D. & Reichlin, L. (2008), 'Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components?', *Journal of Econometrics* **146**(2), 318–328.

Doan, T., Litterman, R. B. & Sims, C. A. (1983), Forecasting and conditional projection using realistic prior distributions, Technical Report 1202.

Forni, M., Hallin, M., Lippi, M. & Reichlin, L. (2000), 'The generalized dynamic-factor model: Identification and estimation', *The Review of Economics and Statistics* **82**(4), 540–554.

Forni, M., Hallin, M., Lippi, M. & Reichlin, L. (2004), 'The generalized dynamic factor model consistency and rates', *Journal of Econometrics* **119**(2), 231–255.

Friedman, J., Hastie, T. & Tibshirani, R. (2010), 'Regularization paths for generalized linear models via coordinate descent', *Journal of Statistical Software* **33**(1), 1–22.

Gefang, D. (2014), 'Bayesian doubly adaptive elastic-net lasso for VAR shrinkage', *International Journal of Forecasting* **30**(1), 1 – 11.

Geweke, J. (1977), The dynamic factor analysis of economic time series, *in* D. J. Aigner & A. S. Goldberger, eds, 'Latent Variables in Socio-Economic Models', North-Holland, Amsterdam.

Giannone, D., Lenza, M. & Primiceri, G. E. (2015), 'Prior selection for vector autoregressions', *The Review of Economics and Statistics* **2**(97), 436–451.

Hendry, D. F. & Krolzig, H.-M. (2001), *Automatic Econometric Model Selection*, Timberlake Consultants Press.

Hoerl, A. E. (1962), 'Application of ridge analysis to regression problems', *Chemical Engineering Progress* **58**, 54–59.

Hoerl, A. E. & Kennard, R. W. (1970), 'Ridge regression: Biased estimation for nonorthogonal problems', *Technometrics* **12**(1), 55–67.

Hoover, K. D. & Perez, S. J. (1999), 'Data mining reconsidered: encompassing and the general-to-specific approach to specification search', *Econometrics Journal* **2**(2), 167–191.

Hsu, N.-J., Hung, H.-L. & Chang, Y.-M. (2008), 'Subset selection for vector autoregressive processes using Lasso', *Computational Statistics & Data Analysis* **52**, 3645–3657.

Kock, A. B. (2012), On the oracle property of the adaptive lasso in stationary and nonstationary autoregressions, CREATES Research Papers 2012-05, School of Economics and Management, University of Aarhus.

Kock, A. B. & Callot, L. A. (2014*a*), Oracle efficient estimation and forecasting with the adaptive lasso and the adaptive group lasso in vector autoregressions, *in* N. Haldrup, M. Meitz & P. Saikkonen, eds, 'Essays in Nonlinear Time Series Econometrics', Oxford University Press, chapter 10.

Kock, A. B. & Callot, L. A. (2014*b*), Oracle inequalities for high dimensional vector autoregressions, arXiv: 1311.0811v2.
URL: *http://arxiv.org/abs/1311.0811*

Koop, G. & Korobilis, D. (2013), 'Large time-varying parameter VARs', *Journal of Econometrics* **177**(2), 185–198.

Litterman, R. B. (1986), 'Forecasting with Bayesian vector autoregressions-five years of experience', *Journal of Business & Economic Statistics* **4**(1), 25–38.

Lütkepohl, H. (2005), *New Introduction to Multiple Time Series Analysis*, Berlin: Springer-Verlag.

Medeiros, M. C. & Mendes, E. F. (2012), Estimating high-dimensional time series models, CREATES Research Papers 2012-37, School of Economics and Management, University of Aarhus.

Nardi, Y. & Rinaldo, A. (2011), 'Autoregressive process modeling via the Lasso procedure', *Journal of Multivariate Analysis* **102**, 528–549.

Nicholson, W., Matteson, D. & Bien, J. (2014), Structured regularization for large vector autoregressions. Cornell University.

Savin, I. & Winker, P. (2012), Lasso-type and heuristic strategies in model selection and forecasting, Technical report, Jena Economic Research Papers 2012-055.
URL: *http://ssrn.com/abstract=2161793*

Song, S. & Bickel, P. J. (2011), Large vector auto regressions, Discussion Paper arXiv:1106.3915v1, University of California, Berkeley.

Stock, J. H. & Watson, M. W. (2002*a*), 'Forecasting using principal components from a large number of predictors', *Journal of the American Statistical Association* **97**, 1167–1179.

Stock, J. H. & Watson, M. W. (2002*b*), 'Macroeconomic forecasting using diffusion indexes', *Journal of Business & Economic Statistics* **20**(2), 147–62.

Stock, J. H. & Watson, M. W. (2008), Forecasting in dynamic factor models subject to structural instability, *in* J. Castle & N. Shephard, eds, 'The Methodology and Practice of Econometrics, A Festschrift in Honour of Professor David F. Hendry', Oxford University Press, Oxford.

Stock, J. H. & Watson, M. W. (2012), 'Generalized shrinkage methods for forecasting using many predictors', *Journal of Business & Economic Statistics* **30**(4), 481–493.

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

Tibshirani, R. J. & Taylor, J. (2012), 'Degrees of freedom in Lasso problems', *The Annals of Statistics* **40**(2), 1198–1232.

Wang, H., Li, G. & Tsai, C.-L. (2007), 'Regression coefficient and autoregressive order shrinkage and selection via the lasso', *Journal of the Royal Statistical Society Series B* **69**(1), 63–78.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2), 301–320.

Zou, H., Hastie, T. & Tibshirani, R. (2007), 'On the "degrees of freedom" of lasso', *Annals of Statistics* **35**(5), 21732192.

A Data

Denote by x_{kt} the raw variable and by y_{kt} the transformed variable taken in logarithms. The transformation codes (Tcode) are: 1 - no transformation (levels), $y_{k,t} = x_{k,t}$; 2 - first difference, $y_{k,t} = \Delta x_{k,t}$; 3 - second difference, $y_{k,t} = \Delta^2 x_{k,t}$; 4 - logarithm, $y_{k,t} = 400 \cdot \ln(x_{k,t})$; 5 - first difference of logarithm, $y_{k,t} = 400 \cdot \Delta \ln(x_{k,t})$; 6 - second difference of logarithm, $y_{k,t} = 400 \cdot \Delta^2 \ln(x_{k,t})$. The transformations are taken from Koop & Korobilis (2013).

Table 1: Data and Systems

			System				
Series	Series ID	Tcode	I	II	III	IV	V
Real GDP	GDPC96	5	x	x	x	x	x
GDP Deflator	GDPDEF	6		x		x	x
Federal Funds Rate	FEDFUNDS	2		x		x	x
Real Consumption	PCECC96	5	x		x	x	x
pReal Investment	GPDIC96	5				x	x
Hours	HOANBS	5				x	x
Real Wages	COMPRNFB	5				x	x
CPI	CPIAUCSL	6			x		
3-Month Tbill	TB3MS	2			x		
One year bond rate	GS5	2					x
Five years bond rate	GS10	2					x
M2 Money Stock	M2SL	6					x
S&P 500 Index	SP500	5					x
ISM Manufacturing: Prices Index	NAPMPRI	1					x
Real Personal Income	RPI	5					x
Industrial Production Index	INDPRO	5					x
Civilian Unemployment Rate	UNRATE	2					x
Housing Starts	HOUST	4					x
Producer Prixe Index	PPIFCG	5					x
PCE Price Index	PCECTPI	6					x
Average Hourly Earnings	CES3000000008	6					x
M1 Money Stock	M1SL	6					x
Oilprice	OILPRICE	5					x
Real Gov. Consumption & Investment	GCEC96	5					x
K			2	3	4	7	22

Note: Description of the series used in the forecasting exercise. Series ID refers to the identification in the St. Louis' FRED database.

B Forecast Error Measures

B.1 System Forecasts

Table 2: GFESM percentage differences relative to benchmark for expanding estimation window using predictive MSE

Method \ System			I	II	III	IV	V
AIC	no		**-0.05**	0.22	0.08	0.03	0.33
BIC	no		-0.04	-0.01	0.01	-0.02	0.01
Pmax	Lasso SE		-0.04	0.02	0.04	0.01	-0.02
AIC	Lasso SE		-0.04	**-0.01**	**-0.01**	-0.02	-0.01
BIC	Lasso SE		-0.03	**-0.04**	**-0.01**	-0.01	-0.03
Pmax	Lasso		-0.02	0.05	0.06	0.00	**-0.04**
AIC	Lasso		**-0.05**	0.05	0.04	-0.02	**-0.04**
BIC	Lasso		-0.03	**-0.01**	0.00	**-0.03**	**-0.05**
Pmax	Elastic Net		-0.02	0.08	0.06	0.01	-0.03
AIC	Elastic Net		**-0.05**	0.04	0.03	**-0.03**	-0.03
BIC	Elastic Net		-0.03	**-0.03**	-0.00	**-0.03**	**-0.04**
Pmax	RR	all	0.03	0.36	0.40	0.36	0.42
AIC	RR	all	**-0.05**	0.21	0.08	0.02	0.29
BIC	RR	all	-0.04	-0.01	0.01	**-0.02**	0.01
AIC	RR	allbutdiag	**-0.05**	0.21	0.08	0.02	0.30
BIC	RR	allbutdiag	-0.04	-0.01	0.01	**-0.02**	0.01
Pmax	Bayes		-0.04	0.01	0.01	0.02	-0.02
AIC	Bayes		**-0.05**	**-0.01**	**-0.01**	-0.02	-0.01
BIC	Bayes		-0.03	**-0.05**	**-0.01**	**-0.03**	**-0.04**
Pmax	AIC TopDown		0.00	0.31	0.30	0.21	0.33
AIC	AIC TopDown		-0.03	0.22	0.08	0.04	0.24
Pmax	BIC TopDown		-0.02	0.16	0.14	0.14	0.14
BIC	BIC TopDown		-0.03	0.01	**-0.00**	-0.02	**-0.05**
AIC	AIC TP		-0.04	0.21	0.08	0.05	0.27
BIC	BIC TP		-0.03	0.00	**-0.00**	-0.02	-0.04

Note: Relative performance is computed as $(\text{GFESM}_i/\text{GFESM}_{BM}) - 1$.

Table 3: GFESM percentage differences relative to benchmark for expanding estimation window using information criteria

Method \ System			I	II	III	IV	V
AIC	no		**-0.05**	0.22	0.08	0.03	0.33
BIC	no		-0.04	-0.01	0.01	**-0.02**	0.01
Pmax	Lasso SE		-0.04	0.00	**-0.02**	-0.00	-0.01
AIC	Lasso SE		**-0.05**	0.13	0.03	0.03	0.13
BIC	Lasso SE		-0.03	-0.02	-0.00	**-0.03**	**-0.05**
Pmax	Lasso		-0.03	-0.01	-0.01	-0.02	-0.04
AIC	Lasso		-0.05	0.13	0.05	0.01	0.17
BIC	Lasso		-0.04	**-0.03**	-0.01	**-0.03**	**-0.05**
Pmax	Elastic Net		-0.04	-0.01	0.00	-0.02	-0.04
AIC	Elastic Net		**-0.05**	0.12	0.04	-0.01	0.19
BIC	Elastic Net		-0.04	**-0.04**	**-0.02**	**-0.03**	-0.04
Pmax	RR	all	-0.02	-0.00	0.01	-0.01	0.01
AIC	RR	all	**-0.06**	0.11	0.05	-0.02	0.22
BIC	RR	all	-0.04	**-0.04**	**-0.02**	**-0.03**	-0.01
AIC	RR	allbutdiag	**-0.06**	0.11	0.04	-0.00	0.17
BIC	RR	allbutdiag	-0.04	**-0.06**	**-0.01**	**-0.04**	**-0.07**
Pmax	Bayes		-0.03	**-0.05**	**-0.03**	-0.02	**-0.05**
AIC	Bayes		**-0.06**	0.05	0.01	-0.01	0.11
BIC	Bayes		-0.04	**-0.05**	**-0.02**	-0.02	**-0.05**
Pmax	AIC TopDown		0.00	0.31	0.30	0.21	0.33
AIC	AIC TopDown		-0.03	0.22	0.08	0.04	0.24
Pmax	BIC TopDown		-0.02	0.16	0.14	0.14	0.14
BIC	BIC TopDown		-0.03	0.01	-0.00	-0.02	**-0.05**
AIC	AIC TP		-0.04	0.21	0.08	0.05	0.27
BIC	BIC TP		-0.03	0.00	-0.00	-0.02	-0.04

Note: Relative performance is computed as $(\text{GFESM}_i/\text{GFESM}_{BM}) - 1$.

B.2 Specification Choices

Table 4: GFESM percentage differences for expanding estimation windows (EEW) relative to rolling estimation windows (REW) using predictive MSE

Method \ System			I	II	III	IV	V
AIC	no		**-0.02**	-0.02	-0.07	**-0.13**	**-0.29**
BIC	no		-0.01	-0.03	-0.01	-0.02	-0.05
Pmax	Lasso SE		-0.02	**-0.05**	**-0.11**	-0.06	-0.06
AIC	Lasso SE		-0.01	-0.00	-0.04	-0.03	-0.05
BIC	Lasso SE		-0.01	-0.02	-0.02	-0.01	-0.02
Pmax	Lasso		-0.00	-0.02	-0.02	-0.01	-0.01
AIC	Lasso		-0.01	**0.04**	-0.01	-0.01	-0.01
BIC	Lasso		-0.00	-0.02	0.01	-0.01	-0.01
Pmax	Elastic Net		-0.01	-0.00	-0.00	-0.02	-0.02
AIC	Elastic Net		-0.01	0.04	-0.01	-0.02	-0.02
BIC	Elastic Net		-0.00	-0.01	0.01	-0.01	-0.02
Pmax	RR	all	**-0.04**	**-0.05**	**-0.12**	**-0.16**	-0.19
AIC	RR	all	**-0.02**	-0.02	-0.06	**-0.12**	-0.26
BIC	RR	all	-0.01	-0.03	-0.01	-0.02	-0.05
AIC	RR	allbutdiag	**-0.02**	-0.02	-0.07	**-0.12**	**-0.27**
BIC	RR	allbutdiag	-0.01	-0.03	-0.01	-0.02	-0.05
Pmax	Bayes		-0.02	**-0.05**	-0.03	-0.04	-0.09
AIC	Bayes		-0.01	-0.03	-0.04	-0.04	-0.08
BIC	Bayes		-0.01	-0.03	0.00	-0.01	-0.02
Pmax	AIC TopDown		-0.02	**-0.04**	**-0.12**	**-0.14**	**-0.29**
AIC	AIC TopDown		-0.00	-0.01	**-0.07**	-0.07	**-0.34**
Pmax	BIC TopDown		0.01	**-0.07**	**-0.14**	**-0.07**	**-0.31**
BIC	BIC TopDown		**-0.04**	-0.04	-0.04	-0.02	-0.03
AIC	AIC TP		0.00	-0.03	**-0.08**	-0.07	**-0.35**
BIC	BIC TP		**-0.03**	-0.04	-0.03	-0.01	-0.03

Note: Relative performance is computed as $(\text{GFESM}_{EEW}/\text{GFESM}_{REW}) - 1$.

Table 5: GFESM percentage differences for expanding estimation windows (EEW) relative to rolling estimation windows (REW) using information criteria

Method \ System			I	II	III	IV	V
AIC	no		**-0.02**	-0.02	**-0.07**	**-0.13**	-0.29
BIC	no		-0.01	**-0.03**	-0.01	-0.02	-0.05
Pmax	Lasso SE		**-0.02**	-0.03	-0.06	-0.01	-0.01
AIC	Lasso SE		-0.01	**0.03**	-0.04	-0.04	-0.30
BIC	Lasso SE		-0.01	-0.01	-0.00	-0.02	-0.01
Pmax	Lasso		0.01	-0.01	0.00	0.00	-0.01
AIC	Lasso		-0.00	-0.00	-0.04	-0.06	**-0.34**
BIC	Lasso		**-0.02**	-0.03	-0.01	-0.01	-0.01
Pmax	Elastic Net		0.00	-0.02	0.01	0.01	-0.01
AIC	Elastic Net		-0.00	0.00	-0.05	-0.06	**-0.32**
BIC	Elastic Net		-0.01	-0.02	-0.01	-0.00	-0.00
Pmax	RR	all	0.01	0.00	0.01	-0.00	-0.03
AIC	RR	all	-0.01	-0.00	-0.04	-0.06	-0.30
BIC	RR	all	-0.01	-0.02	-0.01	-0.00	-0.00
AIC	RR	allbutdiag	-0.02	-0.01	**-0.06**	**-0.06**	**-0.33**
BIC	RR	allbutdiag	-0.01	-0.03	-0.01	-0.01	-0.00
Pmax	Bayes		0.01	-0.01	-0.01	0.00	-0.04
AIC	Bayes		-0.01	-0.01	-0.05	-0.05	**-0.32**
BIC	Bayes		-0.01	-0.03	-0.01	-0.00	-0.01
Pmax	AIC TopDown		**-0.02**	**-0.04**	**-0.12**	**-0.14**	-0.29
AIC	AIC TopDown		-0.00	-0.01	**-0.07**	**-0.07**	**-0.34**
Pmax	BIC TopDown		0.01	**-0.07**	**-0.14**	**-0.07**	-0.31
BIC	BIC TopDown		**-0.04**	**-0.04**	-0.04	-0.02	-0.03
AIC	AIC TP		0.00	-0.03	**-0.08**	**-0.07**	**-0.35**
BIC	BIC TP		**-0.03**	**-0.04**	-0.03	-0.01	-0.03

Note: Relative performance is computed as $(\text{GFESM}_{EEW}/\text{GFESM}_{REW}) - 1$.

Table 6: GFESM percentage differences for predictive MSE (PMSE) relative to information criteria (IC) using expanding estimation windows

Method \ System			I	II	III	IV	V
Pmax	Lasso SE		0.00	0.02	**0.07**	0.01	-0.01
AIC	Lasso SE		**0.01**	**-0.13**	**-0.04**	**-0.04**	**-0.12**
BIC	Lasso SE		0.00	-0.02	-0.01	0.02	0.02
Pmax	Lasso		**0.02**	0.06	**0.07**	0.02	-0.00
AIC	Lasso		0.00	-0.07	-0.02	**-0.03**	**-0.18**
BIC	Lasso		**0.01**	0.02	0.02	0.00	0.00
Pmax	Elastic Net		**0.02**	**0.09**	**0.05**	**0.03**	0.00
AIC	Elastic Net		0.00	**-0.07**	-0.01	-0.02	**-0.19**
BIC	Elastic Net		0.01	0.01	0.02	-0.00	0.00
Pmax	RR	all	**0.05**	**0.36**	**0.38**	**0.37**	**0.40**
AIC	RR	all	0.00	**0.09**	0.03	**0.04**	0.06
BIC	RR	all	-0.00	0.03	0.03	0.01	0.02
AIC	RR	allbutdiag	0.00	**0.09**	0.03	0.02	**0.11**
BIC	RR	allbutdiag	0.00	0.05	0.02	0.01	0.08
Pmax	Bayes		**-0.01**	0.07	**0.05**	**0.04**	0.03
AIC	Bayes		0.01	-0.06	-0.03	-0.01	**-0.10**
BIC	Bayes		0.00	0.00	0.02	-0.01	0.01

Note: Relative performance is computed as $(\text{GFESM}_{PMSE}/\text{GFESM}_{IC}) - 1$.

Table 7: GFESM percentage differences for predictive MSE (PMSE) relative to information criteria (IC) using rolling estimation windows

Method \ System			I	II	III	IV	V
Pmax	Lasso SE		-0.00	0.03	**0.14**	**0.07**	0.05
AIC	Lasso SE		0.01	-0.09	-0.05	-0.06	**-0.35**
BIC	Lasso SE		0.00	-0.01	0.00	0.01	0.03
Pmax	Lasso		**0.03**	0.07	**0.09**	0.04	0.00
AIC	Lasso		0.01	**-0.11**	-0.04	**-0.08**	**-0.45**
BIC	Lasso		-0.01	0.01	-0.00	-0.00	0.00
Pmax	Elastic Net		**0.03**	0.08	**0.07**	0.06	0.02
AIC	Elastic Net		**0.01**	**-0.10**	**-0.06**	-0.07	**-0.44**
BIC	Elastic Net		-0.00	-0.00	-0.00	0.01	0.02
Pmax	RR	all	**0.10**	**0.43**	**0.58**	**0.63**	**0.67**
AIC	RR	all	**0.01**	**0.10**	0.05	**0.11**	0.00
BIC	RR	all	0.00	0.05	0.02	0.03	0.06
AIC	RR	allbutdiag	0.01	**0.10**	0.04	**0.09**	0.00
BIC	RR	allbutdiag	0.00	0.06	0.01	0.03	**0.13**
Pmax	Bayes		**0.02**	**0.12**	**0.06**	**0.08**	0.09
AIC	Bayes		0.01	-0.05	-0.04	-0.03	**-0.34**
BIC	Bayes		-0.00	0.01	0.00	0.00	0.02

Note: Relative performance is computed as $(\text{GFESM}_{PMSE}/\text{GFESM}_{IC}) - 1$.

Table 8: RMSFE for real GDP growth for expanding estimation windows using information criteria

			II				IV				V			
Method \Horizon			1	2	4	8	1	2	4	8	1	2	4	8
AIC	no		2.59	2.81	2.77	2.66	2.35	2.62	2.58	2.63	3.03	3.59	3.15	2.95
BIC	no		2.39	2.56	2.59	2.63	**2.08**	**2.38**	**2.55**	**2.62**	2.47	2.76	2.98	2.78
Pmax	Lasso SE		**2.29**	**2.46**	2.60	2.64	**2.12**	**2.43**	2.58	2.66	**2.01**	**2.39**	**2.55**	**2.61**
AIC	Lasso SE		2.52	2.71	2.71	2.66	2.26	2.51	**2.56**	2.63	2.40	2.71	2.86	2.69
BIC	Lasso SE		**2.29**	**2.47**	**2.58**	2.63	**2.04**	**2.37**	**2.56**	**2.62**	**1.99**	**2.38**	2.58	2.65
Pmax	Lasso		2.42	2.56	2.64	2.65	2.31	2.56	2.60	2.65	2.15	2.46	2.56	**2.61**
AIC	Lasso		2.52	2.74	2.72	2.66	2.25	2.56	**2.57**	2.63	2.49	3.08	2.97	2.79
BIC	Lasso		2.33	2.50	2.59	**2.63**	2.15	2.49	2.57	**2.62**	**2.05**	**2.44**	2.58	2.63
Pmax	Elastic Net		2.41	2.57	2.62	2.65	2.29	2.55	2.60	2.65	2.15	2.46	**2.56**	**2.61**
AIC	Elastic Net		2.51	2.73	2.72	2.66	2.23	2.52	2.57	2.63	2.52	3.08	2.97	2.79
BIC	Elastic Net		**2.32**	2.50	**2.58**	**2.63**	2.16	**2.46**	2.57	**2.62**	**2.02**	**2.44**	2.58	2.62
Pmax	RR	all	2.50	2.58	2.63	2.65	2.34	2.52	2.63	2.65	2.21	2.55	2.61	**2.62**
AIC	RR	all	2.49	2.69	2.72	2.65	2.20	2.51	2.57	2.63	2.58	3.16	2.96	2.81
BIC	RR	all	2.32	**2.50**	**2.58**	**2.63**	2.13	**2.45**	2.57	2.62	**2.05**	**2.37**	**2.56**	2.62
AIC	RR	allbutdiag	2.43	2.64	2.69	2.64	2.27	2.55	2.58	2.63	2.44	3.06	2.95	2.79
BIC	RR	allbutdiag	**2.26**	**2.43**	**2.56**	**2.63**	**2.09**	**2.39**	**2.56**	**2.62**	2.11	**2.39**	**2.55**	2.62
Pmax	Bayes		**2.29**	**2.48**	2.60	2.64	2.51	2.54	2.59	2.65	2.46	2.49	**2.55**	**2.61**
AIC	Bayes		2.42	2.62	2.67	2.65	2.17	2.48	**2.56**	2.63	2.40	2.93	2.85	2.72
BIC	Bayes		**2.27**	**2.45**	**2.58**	**2.63**	2.45	2.53	2.58	**2.62**	2.47	2.51	**2.56**	2.62
Pmax	AIC TopDown		2.60	2.78	2.77	2.65	2.43	2.80	2.67	2.71	2.74	3.02	3.05	2.84
AIC	AIC TopDown		2.59	2.82	2.75	2.65	2.23	2.59	2.58	2.66	2.61	2.89	3.05	2.78
Pmax	BIC TopDown		2.44	2.67	2.73	2.67	2.28	2.61	2.67	2.69	2.27	2.73	2.67	**2.61**
BIC	BIC TopDown		2.35	2.54	**2.57**	**2.62**	**2.11**	2.47	2.62	2.69	**2.06**	2.48	2.59	2.65
AIC	AIC TP		2.59	2.80	2.75	2.66	2.29	2.59	2.57	2.63	2.60	2.97	3.16	2.90
BIC	BIC TP		2.34	2.53	2.60	2.63	**2.13**	2.52	2.60	2.65	2.06	2.53	2.70	2.70

Table 9: RMSFE for the GDP deflator for expanding estimation windows using information criteria

			II				IV				V			
Method	\Horizon		1	2	4	8	1	2	4	8	1	2	4	8
AIC	no		0.87	0.97	1.17	1.43	0.91	0.99	1.11	1.31	1.19	1.32	1.76	2.56
BIC	no		**0.85**	**0.89**	**1.00**	1.23	0.87	0.95	1.07	1.30	0.97	1.06	1.39	2.05
Pmax	Lasso SE		0.86	0.93	1.01	1.23	**0.87**	**0.94**	**1.04**	**1.24**	**0.86**	**0.94**	**1.06**	**1.30**
AIC	Lasso SE		0.86	0.93	1.11	1.36	0.90	0.98	1.12	1.35	0.98	1.09	1.39	1.88
BIC	Lasso SE		0.86	**0.90**	**0.99**	**1.21**	0.87	0.95	**1.04**	**1.24**	**0.87**	**0.96**	**1.04**	**1.24**
Pmax	Lasso		0.92	1.01	1.09	1.30	0.93	1.01	1.08	1.26	0.92	1.01	1.07	**1.25**
AIC	Lasso		0.87	0.96	1.16	1.41	0.88	0.95	1.06	1.26	1.00	1.11	1.47	2.01
BIC	Lasso		0.87	0.92	1.01	**1.22**	0.91	1.00	1.07	1.26	0.89	0.99	1.10	1.35
Pmax	Elastic Net		0.91	0.99	1.07	1.29	0.93	1.01	1.08	1.27	0.92	1.01	**1.07**	**1.25**
AIC	Elastic Net		0.87	0.95	1.15	1.40	**0.87**	**0.93**	1.05	**1.25**	1.01	1.13	1.46	1.98
BIC	Elastic Net		0.87	0.92	1.02	1.24	0.90	0.98	1.06	1.26	0.89	0.99	1.10	1.36
Pmax	RR	all	0.93	1.00	1.09	1.28	0.93	1.02	1.13	1.31	0.93	1.01	1.14	1.35
AIC	RR	all	0.87	0.96	1.15	1.41	**0.86**	**0.92**	**1.03**	**1.24**	1.03	1.17	1.57	2.16
BIC	RR	all	0.87	0.91	1.01	1.23	0.90	0.97	1.05	1.26	0.91	0.99	1.09	1.32
AIC	RR	allbutdiag	0.86	0.94	1.13	1.39	**0.86**	**0.94**	1.05	1.27	0.99	1.12	1.49	2.06
BIC	RR	allbutdiag	**0.84**	**0.88**	**0.99**	**1.22**	**0.86**	**0.94**	**1.04**	1.27	**0.86**	**0.95**	**1.05**	1.31
Pmax	Bayes		0.86	0.93	1.02	1.24	0.90	0.98	1.05	**1.25**	**0.87**	**0.95**	**1.06**	**1.30**
AIC	Bayes		**0.86**	0.93	1.10	1.36	**0.86**	**0.93**	1.05	1.27	0.95	1.04	1.38	1.92
BIC	Bayes		**0.85**	**0.91**	**1.00**	**1.22**	0.89	0.98	1.07	1.26	**0.88**	0.97	1.07	1.30
Pmax	AIC TopDown		0.91	1.05	1.29	1.57	1.01	1.13	1.41	1.76	1.13	1.23	1.69	2.27
AIC	AIC TopDown		0.89	0.98	1.18	1.44	0.92	0.99	1.07	1.30	1.14	1.22	1.54	2.09
Pmax	BIC TopDown		0.90	1.00	1.21	1.46	1.00	1.15	1.31	1.46	0.99	1.09	1.22	1.65
BIC	BIC TopDown		**0.86**	**0.89**	**0.98**	**1.19**	0.88	0.95	**1.04**	**1.23**	0.91	**0.94**	**1.03**	**1.24**
AIC	AIC TP		0.88	0.99	1.19	1.43	0.93	0.99	1.10	1.30	1.10	1.24	1.51	2.13
BIC	BIC TP		**0.85**	**0.89**	**0.98**	**1.18**	0.88	0.95	**1.04**	1.26	**0.87**	**0.95**	1.10	1.45

Table 10: RMSFE for the federal funds rate for expanding estimation windows using information criteria

			II				IV				V			
Method \Horizon			1	2	4	8	1	2	4	8	1	2	4	8
AIC	no		0.59	1.09	1.69	2.76	0.56	1.07	1.65	2.70	0.91	1.56	2.79	4.28
BIC	no		0.53	1.00	1.65	2.64	**0.49**	**0.88**	**1.53**	**2.59**	0.70	1.28	2.27	3.98
Pmax	Lasso SE		**0.50**	**0.92**	**1.52**	**2.60**	0.51	0.97	1.62	2.78	0.62	1.03	1.70	2.85
AIC	Lasso SE		0.57	1.06	1.64	2.72	0.54	1.02	1.62	2.67	0.78	1.31	2.13	3.54
BIC	Lasso SE		0.52	0.98	1.63	2.63	**0.47**	**0.85**	**1.50**	**2.56**	0.56	1.00	1.78	3.11
Pmax	Lasso		0.51	**0.92**	1.61	2.66	0.51	0.91	1.61	2.67	0.51	0.92	**1.59**	**2.65**
AIC	Lasso		0.56	1.04	1.66	2.72	0.52	1.00	1.60	2.63	0.75	1.32	2.28	3.57
BIC	Lasso		0.50	0.93	1.61	**2.63**	0.50	**0.90**	1.58	2.63	**0.50**	**0.91**	1.59	2.72
Pmax	Elastic Net		0.51	**0.92**	**1.59**	2.65	0.52	0.92	1.62	2.69	**0.51**	**0.92**	**1.59**	**2.65**
AIC	Elastic Net		0.55	1.03	1.65	2.72	0.50	0.98	1.59	2.62	0.77	1.34	2.32	3.57
BIC	Elastic Net		**0.50**	0.93	**1.60**	**2.62**	**0.49**	**0.88**	**1.56**	**2.61**	0.51	0.92	1.61	2.75
Pmax	RR	all	0.51	0.92	1.60	2.67	0.50	0.92	1.61	2.66	0.52	0.98	1.68	2.73
AIC	RR	all	0.54	1.01	1.64	2.71	**0.49**	0.95	**1.57**	**2.61**	0.82	1.41	2.47	3.74
BIC	RR	all	**0.49**	**0.92**	**1.59**	**2.61**	**0.47**	**0.86**	**1.54**	**2.60**	**0.49**	**0.89**	**1.54**	**2.61**
AIC	RR	allbutdiag	0.54	1.02	1.64	2.74	0.52	1.00	1.60	2.64	0.74	1.29	2.30	3.58
BIC	RR	allbutdiag	**0.49**	0.94	1.61	**2.62**	**0.45**	**0.84**	**1.50**	**2.57**	**0.44**	**0.82**	**1.48**	**2.58**
Pmax	Bayes		**0.49**	**0.90**	**1.58**	2.65	0.51	0.93	1.62	2.69	**0.48**	**0.88**	**1.52**	**2.56**
AIC	Bayes		0.54	1.01	1.62	2.70	0.51	0.98	1.58	2.62	0.64	1.16	2.10	3.49
BIC	Bayes		**0.49**	**0.92**	**1.59**	**2.62**	0.51	0.92	1.61	2.66	**0.48**	**0.87**	**1.53**	**2.58**
Pmax	AIC TopDown		0.59	1.08	1.69	2.83	0.65	1.22	1.99	3.29	0.94	1.58	2.54	3.77
AIC	AIC TopDown		0.60	1.09	1.70	2.74	0.59	1.08	1.63	2.72	0.86	1.40	2.38	3.75
Pmax	BIC TopDown		0.59	1.09	1.73	2.85	0.63	1.13	1.72	2.82	0.70	1.13	1.77	2.93
BIC	BIC TopDown		0.56	1.05	1.73	2.80	0.51	0.93	1.69	3.04	0.57	1.02	1.82	3.20
AIC	AIC TP		0.60	1.10	1.67	2.74	0.58	1.09	1.68	2.77	0.89	1.39	2.32	3.79
BIC	BIC TP		0.54	1.00	1.65	2.64	0.51	0.92	1.66	2.94	0.59	1.07	1.90	3.37