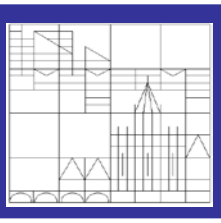
University of Konstanz
Department of Economics



Forecasting Euro Area Macroeconomic Variables with Bayesian Adaptive Elastic Net

Sandra Stankiewicz

Working Paper Series
2015-12

http://www.wiwi.uni-konstanz.de/econdoc/working-paper-series/

Forecasting Euro area macroeconomic variables with Bayesian adaptive elastic net*

Sandra Stankiewicz**

May 13, 2015

Abstract

I use the adaptive elastic net in a Bayesian framework and test its forecasting performance against lasso, adaptive lasso and elastic net (all used in a Bayesian framework) in a series of simulations, as well as in an empirical exercise for macroeconomic Euro area data. The results suggest that elastic net is the best model among the four Bayesian methods considered. Adaptive lasso, on the other hand, shows the worst forecasting performance. Lasso is generally better then adaptive lasso, but worse than adaptive elastic net. The differences in the performance of these models become especially large when the number of regressors grows considerably relative to the number of available observations. The results point to the fact that the ridge regression component in the elastic net is responsible for its improvement in forecasting performance over lasso. The adaptive shrinkage in some of the models does not seem to play a major role, and may even lead to a deterioration of the performance.

Keywords: Elastic net, Lasso, Bayesian, Forecasting

JEL Classification Code: C11, C22, C53

*I would like to thank the participants of the Doctoral Seminar on Econometrics and the Brown Bag Seminar at the University of Konstanz, as well as the participants of the Horn Workshop in Econometrics for useful comments that helped me to improve my paper. I also gratefully acknowledge that this work was performed on the computational resource bwUniCluster funded by the Ministry of Science, Research and Arts and the Universities of the State of Baden-Württemberg, Germany, within the framework program bwHPC.

**University of Konstanz, Department of Economics, Chair of Statistics and Econometrics, Box 129, 78457 Konstanz, Germany, e-mail: sandra.stankiewicz@uni-konstanz.de

1 Introduction

With the growing amount of data, researchers face the challenge of making the best use of the available information for prediction. Including all potential predictors in a model is usually either infeasible or comes at a cost of great parameter estimation uncertainty. Consequently, methods which help to select the relevant variables (see e.g. least absolute shrinkage and selection operator - lasso - of Tibshirani (1996) or stochastic search variable selection - SSVS - of George and McCulloch (1993, 1997)) or estimate large models without loss of prediction accuracy (see e.g. factor models of Stock and Watson (2002) or Mixed Data Sampling Regression of Ghysels, Santa-Clara, and Valkanov (2002)) gain a lot of attention recently.

In this paper I introduce the Bayesian adaptive elastic net, which can be very helpful in forecasting with large datasets. The method uses different shrinkage for different predictors, ideally shrinking the irrelevant ones to zero. The Bayesian framework makes the estimation and statistical inference straightforward and helps to efficiently use the information from a large set of potential predictors in forecasting. The method is tested in terms of forecasting performance in a series of simulations, as well as in an empirical exercise for Euro area macroeconomic data. Its performance is compared to the performance of similar Bayesian methods, such as Bayesian lasso, Bayesian adaptive lasso and Bayesian elastic net, as well as to the performance of two models estimated in a frequentist way: a simple autoregressive model and a factor model. Bayesian adaptive elastic net shows good and robust performance against Bayesian lasso and adaptive lasso. It also performs quite well in comparison to Bayesian elastic net, although it can never beat its counterpart without adaptive shrinkage. Its performance against the two frequentist methods depends on the sample and specification of the models.

Bayesian adaptive elastic net belongs to the lasso-type methods. Lasso was first introduced in the frequentist framework by Tibshirani (1996). It imposes an L_1-penalty on the regression coefficients. As a result some of the coefficients are shrunk, while others are set to zero, which ensures the parsimony of the model and reduces the parameter uncertainty.

Although lasso enables simultaneous shrinkage and variable selection, it has some serious drawbacks. Zou and Hastie (2005) point to the fact that it cannot deal with the problem of multicollinearity, which is often an issue in high-dimensional models. Lasso can also provide unstable results, especially for large datasets. Fan and Li (2001) and Zou (2006) show that lasso provides asymptotically biased results for large coefficients. It can also be inconsistent for model selection (see Zou (2006)).

As an improvement to lasso, two new methods were introduced. The first one is the adaptive lasso of Zou (2006), which lets the lasso shrinkage parameter vary with different coefficients. It has the oracle property, which means that when the shrinkage parameter is properly chosen, it performs as well as if the true underlying model were known. However, it cannot deal with the multicollinearity problem, and, similar to lasso, it can also provide unstable results.

The second method that improves over the lasso is the elastic net of Zou and Hastie (2005),

which combines lasso with ridge regression. It simultaneously performs variable selection and shrinkage, but it handles the multicollinearity problem better than the standard lasso. It also can be seen as a stabilized version of lasso.

Zou and Hastie (2005) test the forecasting performance of elastic net through a series of simulation exercises and for prostate cancer data against such models as linear regressions, estimated by ordinary least squares (OLS), ridge regressions, lasso and naive elastic net. In the empirical comparison elastic net outperforms all the other methods in terms of prediction accuracy and sparsity, although one should keep in mind that OLS, ridge regression and naive elastic net per construction select all variables. The simulation results show that elastic net dominates lasso in terms of prediction accuracy if multicollinearity is present in the data. However, elastic net tends to select more variables than lasso due to the grouping effect. Zou and Hastie (2005) also test the elastic net in gene selection for the case when the number of explanatory variables p is much larger than the number of observations T, and find out that it is the best method among different classification methods used in the analysis.

The elastic net, although being an improvement over lasso, lacks the oracle property. Fan and Peng (2004) and Zou and Zhang (2009) argue that when the dimension of the model is high, the estimation method should not only deal with the problem of multicollinearity, but it also should have the oracle property to ensure the optimal large sample performance of the model. Thus, Zou and Zhang (2009) introduce adaptive elastic net, which is a combination of adaptive lasso and elastic net. It has the oracle property and at the same time it can handle the problem of multicollinearity. They show in a simulation experiment that their method performs better in finite samples than other similar methods (lasso, adaptive lasso, elastic net, smoothly clipped absolute deviation penalty - SCAD (see Fan and Li (2001)). Although the oracle property is irrelevant for the Bayesian approach, one may presume that the method which performs well in the frequentist framework due to its flexibility in the degree of shrinkage for different coefficients, will also perform well in the Bayesian framework.

Bayesian estimation also provides new methods for obtaining shrinkage parameters, which are crucial in lasso-type methods. In the frequentist approach shrinkage parameters are usually chosen through cross-validation. Li and Lin (2010) point to the fact that in the case of methods with two shrinkage parameters (like elastic net) this might result in a double shrinkage problem. This is due to the fact that the cross-validation procedure of the Least Angle Regression algorithm for elastic net (LARS-EN algorithm) of Zou and Hastie (2005), used for the elastic net estimation, chooses the shrinkage parameters sequentially (looks for λ_1 for a given λ_2). In the Bayesian estimation both shrinkage parameters are chosen simultaneously, which helps to avoid the double shrinkage problem.

Finally, Bayesian estimation makes statistical inference straightforward, which can be an advantage for some applications. In the Bayesian estimation, one obtains the posterior distributions of the parameters of interest. Thus, one can construct the credible intervals easily, whereas the calculation of the standard errors for lasso-type methods in the frequentist framework can be very problematic (see e.g. Kyung, Gill, Ghosh, and Casella (2010) for a discussion on the problem).

Park and Casella (2008) were the first to introduce lasso into the Bayesian framework. They proposed Gibbs sampling for lasso with Laplace prior in a hierarchical model. The drawback of the Bayesian lasso is that it cannot automatically do the variable selection. However, it does shrink the parameters and it also provides Bayesian credible intervals which can help with the variable selection. Park and Casella (2008) compare Bayesian lasso to frequentist lasso and OLS regression for diabetes data, finding that Bayesian lasso yields similar estimation results to the frequentist lasso, and at the same time provides credible intervals, which are not available for frequentist lasso applied to a finite sample.

Leng, Tran, and Nott (2014) propose Bayesian adaptive lasso, which allows for different amount of shrinkage for each coefficient. They use a simulation exercise and an empirical analysis for body fat data and prostate cancer data in their paper. They find that the Bayesian adaptive lasso performs better in terms of model selection and prediction then the lasso and adaptive lasso applied in a frequentist framework.

Li and Lin (2010) introduce Bayesian elastic net and describe different ways to do variable selection in a Bayesian approach (e.g. based on credible intervals). Their simulation exercise and empirical analysis for diabetes and prostate cancer data show that Bayesian elastic net performs comparably in prediction to its frequentist counterpart, but it is better in variable selection. It performs slightly worse for simple models, but much better for complicated models.

Kyung et al. (2010) investigate the performance of various Bayesian lasso-type estimators against their frequentist counterparts. They consider fused lasso of Tibshirani, Saunders, Rosset, Zhu, and Knight (2005), group lasso of Yuan and Lin (2006), elastic net of Zou and Hastie (2005) and adaptive lasso by Zou (2006). They conduct their analysis for prostate cancer data, birth weight data and data on state failures[1] in Asia, as well as for a series of simulation exercises. They find that the performance of Bayesian lasso-type methods is comparable or in some cases better than their frequentist counterparts in terms of prediction mean squared errors.

Most of the papers which introduce lasso-type methods come from the fields of biology and chemistry. There are, however, also some implementations in the field of economics. For example, Korobilis (2013) investigates within the Bayesian framework the forecasting performance of lasso, fused lasso, elastic net and adaptive shrinkage with t-prior and Jeffreys prior. In his analysis he uses 129 US macroeconomic variables with quarterly frequency on a time span from 1959 to 2010, and compares different Bayesian shrinkage techniques with factor models in forecasting. He shows that for some cases the Bayesian models outperform factor models.

Mol, Giannone, and Reichlin (2008) also compare Bayesian methods with factor models in terms of forecasting for the US macroeconomic data and find that Bayesian lasso provides comparable results to principal component analysis.

Gefang (2014) introduces Bayesian doubly adaptive elastic net for a VAR model and compares its forecasting performance for US macroeconomic data with the performance

[1]Kyung et al. (2010) define a state failure as a case when the government stops to function effectively, which often results in their collapse in violence and disarray.

of lasso, adaptive lasso, elastic net and adaptive elastic net for Bayesian VARs. She finds that the performance of the Bayesian doubly adaptive elastic net is comparable to the other methods, as well as to the performance of VAR with Minnesota prior.

The approach in this paper is in some ways similar to the approach of Korobilis (2013), who also uses lasso and elastic net to forecast macroeconomic variables in a Bayesian framework. However, the methods I use are slightly different and include also Bayesian adaptive lasso and Bayesian adaptive elastic net. Moreover, I use a different dataset in the empirical exercise, as I apply the Bayesian methods to the Euro-area data. Finally, in my paper I include Monte Carlo simulations to compare the Bayesian methods for different true data generating processes (DGP), whereas Korobilis (2013) focuses on the empirical application only.

The remainder of the paper is structured as follows: Section 2 presents the lasso-type models, and introduces the Bayesian adaptive elastic net and its estimation. Section 3 describes the design and results of the simulation exercise. Section 4 presents the results of the empirical forecasting analysis for Euro area data. Section 5 concludes.

2 Methodology

2.1 Lasso-type methods in the frequentist framework

In this paper I consider the following model:

$$\boldsymbol{y} = \mu \mathbf{1}_T + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\boldsymbol{y}$ is the $T \times 1$ vector of the dependent variable values, μ is the mean, $\boldsymbol{X}$ is the $T \times p$ matrix of standardized predictors, and $\boldsymbol{\varepsilon} \overset{iid}{\sim} N_T(\mathbf{0}, \sigma^2 \boldsymbol{I}_T)$. Thus one can write:

$$\boldsymbol{y}|\mu, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 \sim N_T(\mu \mathbf{1}_T + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_T).$$

Lasso

Lasso was introduced in the frequentist framework by Tibshirani (1996) as a method for variable selection and shrinkage. The main idea behind this method is to shrink the coefficients of the irrelevant predictors to zero. This is achieved by imposing an L_1-penalty on the regression coefficients. The residual sum of squares is minimized under the condition that the sum of the absolute values of the regression coefficients is smaller than a specific value. As a result some of the coefficients are shrunk, while others are set to zero. This ensures the parsimony of the model and reduces the parameter estimation uncertainty. The lasso estimator for model in equation (2.1) can be expressed as follows:

$$\hat{\boldsymbol{\beta}}_L = \arg\min_{\boldsymbol{\beta}} \left\{ (\boldsymbol{y} - \mu \mathbf{1}_T - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \mu \mathbf{1}_T - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

Estimation is done with the help of the LARS algorithm of Efron, Hastie, Johnstone, and Tibshirani (2004) (see also Osborne, Presnell, and Turlach (2000) for earlier algorithms).

The parameter λ regulates the degree of shrinkage. In the lasso approach, the same λ is used for different parameters β_j, which is one of the limitations of this method, as it could be beneficial if the irrelevant predictors were shrunk more than the relevant ones.

Zou and Hastie (2005) also point to other drawbacks of this method. When the number of predictors p is larger than the number of observations T, lasso selects at most T variables due to the nature of the convex optimization problem. It also does not deal well with multicollinearity, which is often an issue when handling large datasets. If there is a group of highly correlated variables in the dataset, lasso tends to select only one of them (without caring which one, e.g. it can select a dummy), and ignores the rest. Tibshirani (1996) also finds that when the number of observations is higher than the number of predictors, and some variables are highly correlated, the ridge regression tends to outperform lasso in prediction. Lasso may also provide unstable results, especially for high-dimensional models. Fan and Li (2001) and Zou (2006) show that lasso provides asymptotically biased results for large coefficients, so it does not have oracle properties. Zou (2006) also shows that it can be inconsistent for model selection.

These drawbacks may deteriorate the performance of lasso in variable selection and prediction, especially for large datasets. Therefore, some modifications of this method have been proposed in the literature. Those relevant for my work are: adaptive lasso, elastic net and adaptive elastic net, which is in the special focus of this paper.

Adaptive lasso

The adaptive lasso of Zou (2006) is a more flexible approach than lasso, as it allows the shrinkage parameter to differ for different β_j:

$$\hat{\boldsymbol{\beta}}_{AL} = \arg\min_{\boldsymbol{\beta}} \left\{ (\boldsymbol{y} - \mu \mathbf{1}_T - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \mu \mathbf{1}_T - \boldsymbol{X}\boldsymbol{\beta}) + \sum_{j=1}^{p} \lambda_j |\beta_j| \right\}.$$

Zou (2006) shows that adaptive lasso has the oracle property. However, it can be as unstable as lasso, and it does not solve the problem of multicollinearity in large datasets.

Elastic net

Zou and Hastie (2005) introduce the elastic net, which combines the lasso approach with a ridge regression of Hoerl and Kennard (1970). The elastic net estimator is given by:

$$\hat{\boldsymbol{\beta}}_{EN} = \arg\min_{\boldsymbol{\beta}} \left\{ (\boldsymbol{y} - \mu \mathbf{1}_T - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \mu \mathbf{1}_T - \boldsymbol{X}\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\}.$$

Ridge regression imposes an L_2-penalty on the regression coefficients. It can provide stable results and deal with multicollinearity, but as it keeps all variables in the model, it cannot provide a sparse solution. The elastic net combines the advantages of lasso and ridge regression. It simultaneously does variable selection and shrinkage, but it does not have some of the drawbacks of lasso. When there is a group of highly correlated variables, elastic net allows for the whole group to be selected, thus catching 'all the big fish' (the grouping effect). It also provides more stable results and can be seen as a stabilized version of lasso. Zou and Hastie (2005) propose a LARS-EN algorithm for elastic net estimation, which is a modified version of the LARS algorithm for lasso.

Adaptive elastic net

The elastic net, although being an improvement over lasso, lacks the oracle property. Zou and Zhang (2009) argue that when the dimension of the model is high, the estimation method should not only deal with the problem of high correlation among some of the predictors, but it also should have the oracle property. Thus, they introduce adaptive elastic net, which allows the shrinkage parameter of the L_1-penalty to differ for different coefficients:

$$\hat{\boldsymbol{\beta}}_{AEN} = \arg\min_{\boldsymbol{\beta}} \left\{ (\boldsymbol{y} - \mu \mathbf{1}_T - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \mu \mathbf{1}_T - \boldsymbol{X}\boldsymbol{\beta}) + \sum_{j=1}^{p} \lambda_{1,j}|\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\}.$$

Adaptive elastic net improves over the earlier methods, as it has the oracle property of the adaptive lasso, and at the same time deals with multicollinearity. Both characteristics are especially important when dealing with large datasets. Although for the Bayesian approach, implemented in this paper, the oracle property is not especially important, allowing for more flexibility in the degree of shrinkage may lead to better model performance in forecasting.

2.2 Bayesian estimation of adaptive elastic net

The frequentist lasso-type methods face problems in providing reliable standard errors. Bayesian estimation, on the other hand, makes statistical inference straightforward, as it provides the posterior distribution of the parameters of interest. Apart from that, Bayesian estimation makes it easier to obtain the optimal degree of shrinkage, as one can put a prior on the shrinkage parameters and include their conditional posterior distribution in the Gibbs sampler. This makes the Bayesian lasso-type methods an attractive alternative to their frequentist counterparts.

Tibshirani (1996) suggests that lasso estimates can be seen as posterior mode estimates when the regression parameters have independent and identical Laplace priors. Motivated by this suggestion Park and Casella (2008) use the Laplace prior on parameters $\boldsymbol{\beta}$ in their Bayesian analysis. To ensure unimodality of the full posterior distribution they condition this Laplace prior on σ^2:

$$\pi(\boldsymbol{\beta}|\sigma^2) = \prod_{j=1}^{p} \left[\frac{\lambda}{2\sqrt{\sigma^2}} \exp\left(-\frac{\lambda}{\sigma} |\beta_j| \right) \right].$$

Park and Casella (2008) point to the fact that conditioning on σ^2 is important, because if the unimodality of the full posterior distribution is not present, Gibbs sampler converges more slowly. Moreover, in the case of multiple posterior modes the point estimates are less meaningful, as summarizing a multi-mode posterior distribution with a single posterior mean or median cannot properly reflect the properties of this distribution and consequently of the parameter of interest.

The prior that Park and Casella (2008) use for σ^2 is an inverse gamma conjugate prior with the shape parameter a and scale parameter b:

$$\pi(\sigma^2) \propto (\sigma^2)^{-a-1} \exp\left(-\frac{b}{\sigma^2} \right).$$

It is easy to see that when $a \to 0$ and $b \to 0$, the inverse gamma prior approaches the following limiting prior:

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2}. \tag{2.2}$$

This is a non-informative, improper, scale-invariant prior, which maintains the conjugacy property of the inverse gamma prior and leads to a proper posterior (see Park and Casella (2008) and Kyung et al. (2010)). The prior (2.2) is used in my paper to implement all Bayesian methods.

For the Bayesian elastic net Zou and Hastie (2005) suggest a prior on $\boldsymbol{\beta}$ that is a mixture between Gaussian and Laplacian priors. They also condition this prior on σ^2 to ensure unimodality of the posterior distribution:

$$\pi(\boldsymbol{\beta}|\sigma^2) \propto \prod_{j=1}^{p} \left[\exp\left(-\frac{\lambda_1}{\sigma} |\beta_j| \right) \exp\left(-\frac{\lambda_2}{2\sigma^2} \beta_j^2 \right) \right].$$

Thus, for the adaptive elastic net which allows for different degree of shrinkage for different coefficients, one can write the conditional prior on $\boldsymbol{\beta}$ as follows:

$$\pi(\boldsymbol{\beta}|\sigma^2) \propto \prod_{j=1}^{p} \left[\exp\left(-\frac{\lambda_{1,j}}{\sigma} |\beta_j| \right) \exp\left(-\frac{\lambda_2}{2\sigma^2} \beta_j^2 \right) \right]. \tag{2.3}$$

A standard practice in the literature (see e.g. Park and Casella (2008), Kyung et al. (2010)) is to implement Bayesian lasso methods with the help of the Gibbs sampler. It is, however, computationally difficult to use Gibbs sampler directly with the prior (2.3), as the absolute values $|\beta_j|$ in this prior would result in unknown full conditional posterior distributions. Thus, Gibbs sampler usually exploits the property of the Laplace distribution, which can be written as a scale mixture of a normal distribution with an exponential mixing density (see Andrews and Mallows (1974)):

$$\frac{k}{2} \exp(-k|l|) = \int_0^\infty \frac{1}{\sqrt{2\pi s}} \exp\left(-\frac{l^2}{2s} \right) \frac{k^2}{2} \exp\left(-\frac{k^2}{2} s \right) ds. \tag{2.4}$$

Setting $s = \tau_j^2$, $k = \lambda_{1,j}$, and $l = \beta_j$, we may transform (2.3) into the following:

$$\pi(\boldsymbol{\beta}|\sigma^2, \tau_1^2, ..., \tau_p^2) \propto \prod_{j=1}^{p} \left[\exp\left(-\frac{\lambda_2}{2\sigma^2} \beta_j^2 \right) \times A \right],$$

where

$$A = \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2\tau_j^2}} \exp\left(-\frac{\beta_j^2}{2\sigma^2\tau_j^2} \right) \frac{\lambda_{1,j}^2}{2} \exp\left(-\frac{\lambda_{1,j}^2 \tau_j^2}{2} \right) d\tau_j^2.$$

This gives the conditional prior on $\boldsymbol{\beta}$ for the adaptive elastic net, which is:

$$\beta_j|\sigma^2, \tau_1^2, ..., \tau_p^2 \sim N\left(0, \frac{\sigma^2\tau_j^2}{\lambda_2\tau_j^2 + 1} \right)$$

or written as a vector:

$$\boldsymbol{\beta}|\sigma^2, D \sim N_p(\mathbf{0}_p, \sigma^2 D), \quad \text{where} \quad D = \text{diag}\left\{\frac{\tau_1^2}{\lambda_2\tau_1^2+1}, ..., \frac{\tau_p^2}{\lambda_2\tau_p^2+1}\right\}.$$

The introduction of the hyperparameters $\tau_1^2, ..., \tau_p^2$ enables an easy implementation of the Gibbs sampler, as through the transformation (2.4) a normal conditional prior on $\boldsymbol{\beta}$ is obtained. However, the introduction of the hyperparameters $\tau_1^2, ..., \tau_p^2$ into the model hierarchy has also consequences for the shrinkage of the $\boldsymbol{\beta}$ coefficients.[2] Looking at the prior covariance matrix D one can observe that the degree of shrinkage of the $\boldsymbol{\beta}$ coefficients depends also on the hyperparameters $\tau_1^2, ..., \tau_p^2$.

In the Bayesian approach there are two popular methods of obtaining the shrinkage parameters $\lambda_{1,j}$ for $j = 1, ..., p$ and λ_2. One may use the Empirical Bayes approach with marginal maximum likelihood estimation (MLE) (see Casella (2001) or Atchade (2011)). Alternatively one may treat the shrinkage parameters as random variables and include them in the Gibbs sampler by imposing priors on them (see Park and Casella (2008)). The advantage is the easy implementation of the Gibbs sampler. Also Kyung et al. (2010) argue that using a Gibbs sampler is faster than the marginal MLE and the estimates they obtain with both methods are very similar in all considered examples. In addition, the choice of the hyperparameters of the priors for the degree of shrinkage should have in theory a relatively small influence on inference, as these are the parameters which are deeper in the hierarchy (see Lehmann and Casella (1998), p. 260). I choose to follow this approach, because on the one hand it is a convenient way of obtaining the shrinkage parameters, and on the other hand, it also helps to account for the uncertainty connected to its selection, which affects the parameters of interest.

Following the literature (see e.g. Kyung et al. (2010), Park and Casella (2008), Gefang (2014) or Korobilis (2013)) I set the exponential prior on τ_j^2 and the gamma priors on $\lambda_{1,j}^2$ and λ_2 as:

$$\tau_1^2, ..., \tau_p^2 \sim \prod_{j=1}^{p}\left[\frac{\lambda_{1,j}^2}{2}\exp\left(-\frac{\lambda_{1,j}^2\tau_j^2}{2}\right)\right], \text{ where } \tau_1^2, ..., \tau_p^2 > 0$$

$$\lambda_{1,1}^2, ..., \lambda_{1,p}^2 \sim \prod_{j=1}^{p}\left[\frac{\delta_{1,j}^{r_{1,j}}}{\Gamma(r_{1,j})}(\lambda_{1,j}^2)^{r_{1,j}-1}\exp(-\delta_{1,j}\lambda_{1,j}^2)\right]$$

$$\lambda_2 \sim \frac{\delta_2^{r_2}}{\Gamma(r_2)}\lambda_2^{r_2-1}\exp(-\delta_2\lambda_2).$$

For the gamma priors put on the shrinkage parameters $\lambda_{1,j}^2$ and λ_2 the hyperparameters $r_{1,j}$ and r_2 are the shape parameters, while $\delta_{1,j}$ and δ_2 are the inverse scale parameters. For the parametrization of the gamma distribution used in this paper,[3] smaller values

[2] I thank Christoph Frey for drawing my attention to this point.

[3] The parametrization of the gamma distribution used in this paper is the following:

$$f(z; c, d) = \frac{1}{\Gamma(c)}z^{c-1}d^c e^{-dz}, \quad \text{with} \quad z \geq 0, \quad c > 0 \quad \text{and} \quad d > 0,$$

where $\Gamma(c) = (c-1)!$ is the gamma function.

of the hyperparameters $r_{1,j}$, r_2, and $\delta_{1,j}$, δ_2 lead to larger shrinkage of the coefficients (parameters $\lambda_{1,j}$ and λ_2 are larger).

To obtain the posterior distribution, one needs the likelihood function, which is given by:

$$L(\mu, \boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{T/2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \mu\mathbf{1}_T - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \mu\mathbf{1}_T - \boldsymbol{X}\boldsymbol{\beta})\right).$$

To simplify the calculations one may integrate μ out. For this purpose, I set $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \bar{y}\mathbf{1}_T$, where $\bar{y}$ is the average of the elements in $\boldsymbol{y}$. As the columns of $\boldsymbol{X}$ are standardized, one can write following Park and Casella (2008):

$$\begin{aligned}
&(\boldsymbol{y} - \mu\mathbf{1}_T - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \mu\mathbf{1}_T - \boldsymbol{X}\boldsymbol{\beta}) \\
&= (\tilde{\boldsymbol{y}} + \bar{y}\mathbf{1}_T - \mu\mathbf{1}_T - \boldsymbol{X}\boldsymbol{\beta})'(\tilde{\boldsymbol{y}} + \bar{y}\mathbf{1}_T - \mu\mathbf{1}_T - \boldsymbol{X}\boldsymbol{\beta}) \\
&= (\bar{y}\mathbf{1}_T - \mu\mathbf{1}_T)'(\bar{y}\mathbf{1}_T - \mu\mathbf{1}_T) + (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})'(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}) \\
&= T(\bar{y} - \mu)^2 + (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})'(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}).
\end{aligned}$$

Thus the the full conditional distribution of μ is $N(\bar{y}, \sigma^2/T)$, and $\tilde{\boldsymbol{y}} \sim N_T(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_T)$, which gives the modified likelihood function:

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{(T-1)/2}} \exp\left(-\frac{1}{2\sigma^2}(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})'(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})\right).$$

Combining the likelihood function with the prior information on the parameters, gives the following full posterior distribution:

$$\begin{aligned}
&\pi(\boldsymbol{\theta}|\tilde{\boldsymbol{y}}, \boldsymbol{X}) \propto \frac{1}{(2\pi\sigma^2)^{(T-1)/2}} \exp\left(-\frac{1}{2\sigma^2}(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})'(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})\right) \\
&\times \frac{1}{\sqrt{(2\pi)^p \prod_{j=1}^p \left(\frac{\sigma^2\tau_j^2}{\lambda_2\tau_j^2+1}\right)}} \exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\beta}' D^{-1}\boldsymbol{\beta}\right) \times \frac{1}{\sigma^2} \\
&\times \prod_{j=1}^p \left[\frac{\lambda_{1,j}^2}{2} \exp\left(-\frac{\lambda_{1,j}^2\tau_j^2}{2}\right)\right] \times \prod_{j=1}^p \left[\frac{\delta_{1,j}^{r_{1,j}}}{\Gamma(r_{1,j})}(\lambda_{1,j}^2)^{r_{1,j}-1} \exp(-\delta_{1,j}\lambda_{1,j}^2)\right] \\
&\times \frac{\delta_2^{r_2}}{\Gamma(r_2)}\lambda_2^{r_2-1} \exp(-\delta_2\lambda_2),
\end{aligned}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \sigma^2, \tau_1^2, ..., \tau_p^2, \lambda_{1,1}^2, ..., \lambda_{1,p}^2, \lambda_2\}$.

As the full posterior distribution does not have a form of any known distribution, and is quite complicated, it is not possible to draw directly from it. Instead a Gibbs sampler is implemented, which draws random values from the distributions of $\boldsymbol{\beta}$, σ^2, $\left\{\tau_j^2\right\}_{j=1}^p$, $\left\{\lambda_{1,j}^2\right\}_{j=1}^p$ and λ_2, conditional on the current values of the rest of the parameters. The full conditional posterior distributions for all the parameters of the model are given by:

$$\boldsymbol{\beta}|\,\sigma^2, \left\{\tau_j^2\right\}_{j=1}^p, \left\{\lambda_{1,j}^2\right\}_{j=1}^p, \lambda_2, \boldsymbol{X}, \tilde{\boldsymbol{y}} \sim N_p\left((\boldsymbol{X}'\boldsymbol{X} + D^{-1})^{-1}\boldsymbol{X}'\tilde{\boldsymbol{y}}, \sigma^2(\boldsymbol{X}'\boldsymbol{X} + D^{-1})^{-1}\right)$$

$$\sigma^2 \big| \boldsymbol{\beta}, \left\{\tau_j^2\right\}_{j=1}^p, \left\{\lambda_{1,j}^2\right\}_{j=1}^p, \lambda_2, \boldsymbol{X}, \tilde{\boldsymbol{y}} \sim \text{inv. gamma}\Bigg(\frac{T-1+p}{2}, \frac{1}{2}\left((\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})'(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{\beta}' D^{-1} \boldsymbol{\beta}\right)\Bigg)$$

$$\frac{1}{\tau_j^2} \Bigg| \boldsymbol{\beta}, \sigma^2, \left\{\lambda_{1,j}^2\right\}_{j=1}^p, \lambda_2, \boldsymbol{X}, \tilde{\boldsymbol{y}} \sim \text{inverse Gaussian}\left(\sqrt{\frac{\lambda_{1,j}^2 \sigma^2}{\beta_j^2}}, \lambda_{1,j}^2\right) \text{ for } \; j = 1, ..., p$$

$$\lambda_{1,j}^2 \big| \boldsymbol{\beta}, \sigma^2, \left\{\tau_j^2\right\}_{j=1}^p, \lambda_2, \boldsymbol{X}, \tilde{\boldsymbol{y}} \sim \text{gamma}\left(r_{1,j} + 1, \frac{\tau_j^2}{2} + \delta_{1,j}\right) \quad \text{for } \; j = 1, ..., p$$

$$\lambda_2 | \, \boldsymbol{\beta}, \sigma^2, \left\{\tau_j^2\right\}_{j=1}^p, \left\{\lambda_{1,j}^2\right\}_{j=1}^p, \boldsymbol{X}, \tilde{\boldsymbol{y}} \sim \text{gamma}\left(\frac{p}{2} + r_2, \frac{1}{2\sigma^2}\sum_{j=1}^p \beta_j^2 + \delta_2\right).$$

Drawing from the normal, inverse gamma and gamma distributions is straightforward and can be easily implemented with most statistical programs. To sample from the inverse Gaussian distribution I use the method introduced by Michael, Schucany, and Haas (1976).

The full conditional posterior distributions for lasso, adaptive lasso and elastic net can be found in the Appendix.

3 Simulation exercise

3.1 Simulation design

The data used in the simulations come from the following true data generating process (DGP):

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon},$$

where $\boldsymbol{y}$ is the $T \times 1$ vector of the dependent variable values, $\boldsymbol{X}$ is the $T \times p$ matrix of predictors, and $\varepsilon_i \overset{iid}{\sim} N(0,1)$ for $i = 1, ..., T$.

In each simulation set I use 500 Monte Carlo replications. The first four simulation sets are based on the simulations carried out in the papers of Tibshirani (1996), Zou and Hastie (2005) and Kyung et al. (2010). The fifth simulation set is based on the values estimated from the data, used in the empirical analysis in this paper. In all simulation sets, datasets of $T = 100$ are generated, where the first 70 observations are used for the in-sample estimation, and the last 30 for forecasting. I describe the simulation sets in the following:

- Set 1: The number of predictors is $p = 20$. I set $\sigma = 3$ and $\beta_1 = 3$, $\beta_2 = 1.5$, $\beta_5 = 2$. The other parameters β_j are set to zero. The pairwise correlation between x_i and x_j is set to be $corr(i, j) = 0.5^{|i-j|}$.

- Set 2: Here $\beta_j = 0.85$ for all $j = 1, ..., 20$. Everything else is just as in Set 1.

- Set 3: The number of predictors is $p = 40$. I set $\sigma = 15$ and $\boldsymbol{\beta} = (\mathbf{0}', \mathbf{2}', \mathbf{0}', \mathbf{2}')'$, where $\mathbf{0}$ and $\mathbf{2}$ are vectors of length 10 and elements 0 and 2 respectively. The pairwise correlation between x_i and x_j is set to be $corr(i, j) = 0.5$ for $i \neq j$.

- Set 4: The number of predictors is $p = 40$. I set $\sigma = 15$ and $\beta_j = 3$ for $j = 1, ..., 15$, whereas $\beta_j = 0$ for $j = 16, ...40$. The variables $\boldsymbol{X}$ are generated as follows:

$$
\begin{aligned}
x_j &= Z_1 + \varepsilon_j^x, \quad Z_1 \sim N(0,1), \quad j = 1, ..., 5 \\
x_j &= Z_2 + \varepsilon_j^x, \quad Z_2 \sim N(0,1), \quad j = 6, ..., 10 \\
x_j &= Z_3 + \varepsilon_j^x, \quad Z_3 \sim N(0,1), \quad j = 11, ..., 15 \\
x_j &\overset{iid}{\sim} N(0,1), \quad j = 16, ..., 40,
\end{aligned}
$$

  where $\varepsilon_j^x \overset{iid}{\sim} N(0, 0.01)$ for $j = 1, ..., 15$.
  In order to be able to test the capability of the models to deal with multicollinearity, three groups of 5 highly correlated variables and 25 pure noise features are generated in this set.

- Set 5: This set is designed to imitate the correlation structure of the dataset used in the empirical analysis in Section 4, so that the results obtained from simulations and in the empirical exercise are comparable. I simulate datasets with 37 predictors. I include 2 lags of each predictor in the model, so that there are altogether 74 regressors. The matrix $\boldsymbol{X}$ is constructed in such a way that the first 37 columns include the first lag of all predictors, and the next 37 columns include the second lag of all predictors. The coefficient matrix $\boldsymbol{\beta}$ is chosen arbitrary and does not reflect the actual dependencies in the dataset from Section 4. I set $\boldsymbol{\beta} = (\mathbf{0}'_{10\times1}, \mathbf{2}'_{10\times1}, \mathbf{0}'_{10\times1}, \mathbf{2}'_{10\times1}, \mathbf{0}'_{10\times1}, \mathbf{1}'_{10\times1}, \mathbf{0}'_{10\times1}, \mathbf{1}'_{4\times1},)'$, where $\mathbf{0}$, $\mathbf{1}$ and $\mathbf{2}$ are vectors with elements 0, 1 and 2 respectively. The variables $\boldsymbol{X}$ are generated as an AR(1) process:

$$
X_{t,j} = \alpha_j + \rho_j \cdot X_{t-1,j} + \varepsilon_{t,j}, \quad j = 1, ..., 37,
$$

  where α_j and ρ_j are AR(1) estimates for the variables in the empirical dataset, while $\boldsymbol{\varepsilon}_j$ are drawn from a multivariate normal distribution with zero mean and covariance matrix estimated from the empirical dataset.

Following Park and Casella (2008) I set the hyperparameters of the gamma priors for $\lambda_{1,j}$ and λ_2 in all simulation sets to $r_{1,j} = r_2 = 1.0$ and $\delta_{1,j} = \delta_2 = 1.78$ for all $j = 1, ..., p$.

3.2 Simulation results

The out-of-sample performance of the models is evaluated in terms of the Mean Squared Error (MSE) and the log-score. MSE is defined as the average of the squared differences between the predicted values of the dependent variable $\hat{y}_{t+h|t}$ and the actually observed

Table 1: Out-of-sample Mean Squared Error of adaptive elastic net (AEN), elastic net (EN) and adaptive lasso (AL) relative to MSE of lasso (L)

Simulation set	**p**	**AEN vs. L**			**EN vs. L**			**AL vs. L**		
		$\mathbf{25^{th}}$	$\mathbf{50^{th}}$	$\mathbf{75^{th}}$	$\mathbf{25^{th}}$	$\mathbf{50^{th}}$	$\mathbf{75^{th}}$	$\mathbf{25^{th}}$	$\mathbf{50^{th}}$	$\mathbf{75^{th}}$
h=1										
1	20	0.993	1.003	1.014	0.963	0.983	0.999	1.018	1.030	1.044
2	20	0.972	0.979	0.987	0.958	0.974	0.987	1.002	1.013	1.021
3	40	0.929	0.947	0.962	0.835	0.869	0.900	1.130	1.168	1.210
4	40	0.978	0.990	1.006	0.904	0.931	0.959	1.097	1.143	1.198
5	74	0.856	0.886	0.912	0.715	0.767	0.818	1.382	1.509	1.653
h=2										
1	20	0.989	1.002	1.012	0.956	0.975	0.994	1.023	1.034	1.046
2	20	0.967	0.976	0.983	0.952	0.968	0.982	1.006	1.014	1.023
3	40	0.926	0.945	0.960	0.826	0.861	0.897	1.134	1.175	1.227
4	40	0.975	0.991	1.006	0.904	0.933	0.962	1.092	1.141	1.208
5	74	0.853	0.883	0.910	0.715	0.765	0.815	1.388	1.510	1.654
h=4										
1	20	0.990	1.002	1.013	0.957	0.979	0.998	1.017	1.033	1.048
2	20	0.967	0.977	0.986	0.951	0.968	0.984	1.005	1.014	1.024
3	40	0.926	0.944	0.964	0.830	0.868	0.906	1.120	1.168	1.222
4	40	0.979	0.992	1.011	0.903	0.936	0.963	1.082	1.134	1.191
5	74	0.847	0.880	0.916	0.703	0.760	0.825	1.373	1.513	1.677

Note: The table presents the Bayesian estimation results of simulations for 500 Monte Carlo replications. For each replication data are generated according to one of the five simulation sets described above. The Bayesian adaptive elastic net, elastic net, adaptive lasso and lasso are estimated and their out-of-sample fit for forecasting horizons $h = 1, 2$ and 4 is measured by the Mean Squared Error. The table presents the 25th, 50th and 75th percentiles of the ratio of the MSE of the adaptive elastic net, elastic net and adaptive lasso to the MSE of lasso over all 500 simulated datasets. Values below/above 1 indicate a better/worse out-of-sample performance of the model in comparison to lasso.

values of y_{t+h}:

$$MSE = \frac{\sum_{t=\tau}^{T-h}(\hat{y}_{t+h|t} - y_{t+h})^2}{T - \tau - h},$$

where τ denotes the beginning of the evaluation period, and h is the forecasting horizon.

The second performance measure is based on the predictive likelihood. The predictive likelihood is the predictive density for y_{t+h} (conditional on the data available until t), evaluated at the actually observed value of the dependent variable in time $t + h$. For every Gibbs sampler iteration one can calculate the value of the predictive likelihood for the given value of the model parameters. Then the mean predictive likelihood over all Gibbs sampler iterations can be calculated. The log-score is the sum of the logarithms of the mean predictive likelihood over all forecasting periods:

$$\sum_{t=\tau}^{T-h} \log p(y_{t+h}|y_t, ..., y_1),$$

where $p(y_{t+h}|y_t, ..., y_1)$ is the predictive density.

The results of the simulation exercise can be found in Tables 1 and 2. In terms of the MSE the results of the simulations clearly show that adaptive lasso is the worst performing model, while elastic net is the best one among the tested models. Looking at the median of

Table 2: Out-of-sample average log-scores of adaptive elastic net (AEN), elastic net (EN), adaptive lasso (AL) and lasso (L)

Simulation set	**p**	**Log-scores**			
		AEN	**EN**	**AL**	**L**
h=1					
1	20	-79.408	-78.845	-80.478	-79.639
2	20	-79.371	-79.031	-80.524	-79.948
3	40	-130.571	-128.415	-136.978	-132.148
4	40	-129.632	-128.033	-134.111	-130.272
5	74	-84.459	-80.873	-106.245	-88.476
h=2					
1	20	-76.852	-76.219	-78.071	-77.141
2	20	-77.006	-76.607	-78.301	-77.655
3	40	-126.081	-123.900	-132.507	-127.684
4	40	-125.643	-124.057	-130.199	-126.317
5	74	-82.134	-78.510	-103.625	-86.135
h=4					
1	20	-71.603	-71.017	-72.716	-71.865
2	20	-71.683	-71.306	-72.880	-72.278
3	40	-117.492	-115.520	-123.361	-118.947
4	40	-117.028	-115.493	-121.019	-117.628
5	74	-76.332	-72.968	-96.729	-80.626

Note: The table presents the Bayesian estimation results of simulations for 500 Monte Carlo replications. For each replication data are generated according to one of the five simulation sets described above. The Bayesian adaptive elastic net, elastic net, adaptive lasso and lasso are estimated and their out-of-sample fit for forecasting horizons $h = 1, 2$ and 4 is evaluated. The table presents the average log-scores of adaptive elastic net, elastic net, adaptive lasso and standard lasso over all 500 simulated data sets. The higher the value, the better performance of the model.

the MSEs for all 500 replications, one can see that for all forecasting horizons $h = 1, 2$ and 4, adaptive lasso is dominated by lasso, elastic net and adaptive elastic net. Elastic net, on the other hand, dominates the other three models for all considered cases. Adaptive elastic net also shows good performance, beating lasso in all but the first simulation set.

One also observes that the differences in the performance of all four models are relatively small for sets 1 and 2. In other words, for the case when there are few predictors relative to the number of observations available for estimation the methods tend to perform equally well. When the number of predictors grows (with constant number of available observations), the differences in the performance of single models become considerably larger. This suggests that while all four models perform quite similarly when the considered case is relatively easy in terms of estimation, the adaptive elastic net and especially elastic net show better forecasting performance than lasso and adaptive lasso in the more challenging case (i.e. when the number of predictors grows considerably).

The results of the forecasting exercise in terms of MSE are confirmed for the log-scores, although the differences in the performance of the models are less pronounced than in the case of MSEs. Adaptive lasso is again outperformed by all three other models for all considered simulation sets, whereas elastic net is the best performing model among the considered ones. However, the differences in the performance of the models are small in terms of log-scores for sets 1 and 2. When the number of predictors grows, these differences become more pronounced.

To summarize the simulation results, one may conclude that adding the ridge regression part to lasso improves the forecasting performance of the model, especially when the number of regressors is high relative to the number of available observations. However, allowing for different shrinkage parameters λ_1 for different coefficients does not improve the results in comparison to the corresponding model without the adaptive shrinkage. Adaptive lasso performs worse than lasso, and adaptive elastic net performs worse than elastic net for all considered simulations sets. This might be due to the fact that through the introduction of the hyperparameters $\tau_1, ..., \tau_p$ into the Bayesian framework for all the lasso-type models, some kind of adaptive shrinkage is already introduced automatically for the parameters $\boldsymbol{\beta}$ through the prior covariance matrix D, even if the parameter λ_1 does not vary for different coefficients. Thus, introducing additional adaptive shrinkage through varying λ_1 might not be necessary, and might even lead to a deterioration of the model performance. All in all, it seems that the ridge regression part in the elastic net and adaptive elastic net is more important for the improvement of the results than the adaptive shrinkage.

4 Empirical analysis

4.1 Data and the description of the forecasting exercise

For the empirical analysis, I use data from the Area-wide Model (AWM) database,[4] which is in use e.g. at the European Central Bank (ECB). The data cover the period from 1970Q1 to 2013Q4 (176 quarters) and include aggregated macroeconomic information on the Euro area, e.g. on GDP, consumption, investment, interest rates, unemployment and prices. Although there are 46 variables in the dataset, only 37 are reported for the whole time period and only these are used in the following analysis.

I divide the available sample into the in-sample part from 1970Q1 to 2005Q4 (144 quarters) and out-of-sample part from 2006Q1 to 2013Q4 (32 quarters), for which the forecasting performance evaluation is done. For the forecasting exercise I use 8 series as dependent variables: the real gross domestic product - GDP (YER), private consumption (PCR), exports (XTR), imports (MTR), commodity prices (COMPR), long-term interest rate (LTN), labor productivity (LPROD) and investment (ITR).[5] When one of these variables is used as the dependent variable, all the other 36 variables are considered as potential predictors. As in reality the values of the predictors are rarely available for the period for which the forecasts are made, I do not include contemporaneous values of the explanatory variables in the regression. Instead, I include 2 lags of every explanatory variable in the model, so for each regression I get 72 potential regressors. This number is still considerably smaller than the available number of observations for estimation (144 quarters). To make the exercise more challenging and to investigate whether the Bayesian elastic net and Bayesian adaptive elastic net perform better than Bayesian lasso and Bayesian adaptive lasso, especially in the case when $p > T$ (which is often claimed in the frequentist

[4]The AWM data, as well as the details of the aggregation of different time series and the method of updating historical data can be found at the following website: http://www.eabcn.org/area-wide-model.

[5]The acronyms for the dependent variables come from the AWM database.

literature, see e.g. Zou and Zhang (2009)), I repeat the whole exercise for a shorter sample, starting from 1990Q1. This leaves 64 in-sample observations for the shorter sample case, which is less than the number of predictors (72). For both samples I calculate recursive direct h-step ahead forecasts for different forecasting horizons: $h = 1, 2$ and 4.

All the time series chosen for further analysis are transformed either by taking first differences of the series, first differences of the logarithms of the series or second differences of the logarithms of the series. The detailed description of the variables used in the analysis and the information on their transformation can be found in Table A.1 in the Appendix.

The explanatory variables are standardized (with mean zero and variance one), which is a standard practice in the lasso literature. The reason is that different predictors are usually reported in different units, and thus there can be huge differences in their variances. These, however, do not necessarily correspond to the amount of information the predictors contain for predicting the dependent variable. The dependent variable is demeaned before the estimation, as μ is not of interest in this paper.

The parameter estimates in every Bayesian regression are means from 10,000 iterations of the Gibbs sampler after 1000 burn-in iterations. For each parameter draw I compute 10 forecasts, so that I get 100,000 draws from the predictive density of each of the dependent variables. According to Korobilis (2013) this can help to reduce the sampling error of the predictive density.

The four Bayesian methods (lasso, adaptive lasso, elastic net and adaptive elastic net) are compared to a simple AR(2) model and to a factor model with four factors, extracted through principal component analysis (see e.g. Stock and Watson (2002), Forni, Hallin, Lippi, and Reichlin (2005) and Bai and Ng (2002)). Just as in the simulation exercise I set the hyperparameters of the gamma prior for $\lambda_{1,j}$ and λ_2 to $r_{1,j} = r_2 = 1.0$ and $\delta_{1,j} = \delta_2 = 1.78$ for all $j = 1, ..., p$.

The benchmark specification of the Bayesian methods for the long (from 1970Q1) and the short (from 1990Q1) sample is modified by allowing for different shrinkage for higher lags of the predictors (see Section 4.2.2), as well as by including autoregressive terms in the model (Section 4.2.3).

4.2 Results of the forecasting exercise

4.2.1 The benchmark specification

The results of the forecasting exercise can be found in Table 3 for the long sample (from 1970Q1) and Table 4 for the short sample (from 1990Q1). They fully confirm the results of the simulation exercise. Among the Bayesian methods, both in terms of MSE and log-score, the adaptive lasso turns out to be the worst, whereas elastic net is the best performing model for all considered dependent variables. The differences between the performance of the models grow considerably when the sample gets smaller (short sample) and the exercise becomes more challenging in terms of estimation. While for the

Table 3: Out-of-sample Mean squared errors and log-scores of adaptive elastic net (AEN), elastic net (EN), adaptive lasso (AL) and lasso (L). Sample: 1970Q1 - 2013Q4.

	MSE					**Log-score**			
	AEN	**EN**	**AL**	**AR(2)**	**4 fac.**	**AEN**	**EN**	**AL**	**L**
YER									
h = 1	0.991	0.950	1.104	0.955	0.945	-36.746	-35.320	-41.322	-37.792
h = 2	1.004	0.976	1.071	1.019	0.995	-36.006	-35.198	-40.171	-37.275
h = 4	1.016	0.988	1.073	1.008	1.013	-36.281	-35.007	-39.412	-36.652
PCR									
h = 1	0.983	0.937	1.170	0.817	1.587	-18.018	-17.934	-19.415	-18.066
h = 2	0.970	0.913	1.198	0.772	1.496	-18.134	-17.876	-19.997	-18.296
h = 4	0.978	0.947	1.185	0.792	1.584	-17.239	-17.184	-18.598	-17.262
XTR									
h = 1	0.993	0.985	1.064	1.265	1.022	-72.932	-72.637	-74.160	-73.018
h = 2	0.994	1.001	1.043	1.313	1.029	-73.205	-72.853	-74.010	-71.895
h = 4	1.000	0.993	1.082	1.245	1.006	-71.167	-71.162	-73.474	-71.204
MTR									
h = 1	0.973	0.919	1.115	1.221	1.102	-62.790	-61.683	-65.256	-63.341
h = 2	0.967	0.926	1.081	1.212	1.100	-62.087	-61.184	-64.339	-62.801
h = 4	0.967	0.914	1.136	1.158	1.088	-60.452	-59.186	-63.415	-61.172
COMPR									
h = 1	0.979	0.977	1.128	0.924	1.071	-131.610	-130.965	-136.408	-132.227
h = 2	0.979	0.978	1.140	0.914	1.070	-127.989	-127.416	-132.563	-128.626
h = 4	0.991	0.994	1.109	0.978	1.080	-121.277	-121.288	-124.680	-122.028
LTN									
h = 1	0.986	0.932	1.101	0.679	0.828	-162.241	-161.184	-164.485	-162.608
h = 2	0.985	0.920	1.145	0.664	0.850	-158.195	-156.893	-160.893	-158.487
h = 4	0.983	0.909	1.170	0.609	0.814	-148.806	-147.377	-151.766	-149.181
LPROD									
h = 1	0.996	0.949	1.112	0.918	0.980	-28.955	-27.733	-32.164	-29.297
h = 2	1.005	0.971	1.075	0.981	1.044	-28.542	-27.574	-30.772	-28.907
h = 4	1.018	0.983	1.074	0.985	1.065	-28.367	-27.764	-30.794	-28.899
ITR									
h = 1	0.981	0.929	1.092	1.022	0.925	-64.743	-62.565	-68.477	-65.427
h = 2	0.984	0.946	1.073	1.054	0.958	-63.846	-62.163	-67.802	-64.646
h = 4	0.989	0.949	1.066	1.062	0.984	-60.697	-58.860	-63.543	-60.778

Note: The table presents the values of the log-scores, as well as the ratio of the MSE of the adaptive elastic net, elastic net and adaptive lasso to the MSE of lasso for 32 forecast periods. Values of the MSE below/above 1 indicate a better/worse out-of-sample performance of model in comparison to lasso. The dependent variables are: YER, PCR, XTR, MTR, COMPR, LTN, LPROD and ITR. The original sample is divided into the in-sample period between 1970Q1 and 2005Q4 (144 quarters) and the out-of-sample period between 2006Q1 and 2013Q4 (32 quarters). For the out-of-sample period the forecasts are calculated recursively. The forecasting horizons are $h = 1, 2$ and 4.

long sample all four Bayesian methods perform quite similarly, for the short sample the domination of the elastic net becomes very pronounced. Also the adaptive elastic net performs much better than lasso and adaptive lasso. The performance of the adaptive lasso in comparison to the elastic net becomes extremely poor for the short sample.

Compared to a simple AR(2) model and a factor model with 4 factors the Bayesian methods do not show especially good performance. For the long sample the AR(2) model outperforms all considered Bayesian methods for private consumption and long-term interest rate. For commodity prices and labor productivity it performs comparably to the elastic net. For the rest of variables it is outperformed by lasso, elastic net and adaptive

Table 4: Out-of-sample Mean squared errors and log-scores of adaptive elastic net (AEN), elastic net (EN), adaptive lasso (AL) and lasso (L). Sample: 1990Q1 - 2013Q4.

	MSE					**Log-score**			
	AEN	**EN**	**AL**	**AR(2)**	**4 fac.**	**AEN**	**EN**	**AL**	**L**
YER									
h = 1	0.935	0.873	1.240	0.842	0.751	-49.830	-43.804	-84.965	-59.189
h = 2	0.913	0.828	1.319	0.821	0.716	-48.367	-42.860	-89.235	-56.403
h = 4	0.880	0.795	1.457	0.759	0.653	-49.957	-43.674	-105.750	-60.872
PCR									
h = 1	0.777	0.569	1.749	0.363	0.576	-24.239	-20.292	-45.574	-29.234
h = 2	0.788	0.585	1.699	0.391	0.585	-23.880	-19.928	-42.169	-28.522
h = 4	0.769	0.543	1.674	0.324	0.485	-24.595	-20.139	-43.297	-29.126
XTR									
h = 1	0.884	0.779	1.295	0.753	0.675	-88.522	-80.170	-126.675	-98.676
h = 2	0.879	0.775	1.331	0.721	0.651	-89.460	-80.223	-127.761	-100.934
h = 4	0.879	0.770	1.269	0.714	0.623	-91.790	-79.566	-134.202	-109.082
MTR									
h = 1	0.910	0.789	1.394	0.981	0.660	-72.350	-66.931	-91.866	-76.702
h = 2	0.911	0.789	1.370	0.922	0.583	-74.299	-68.642	-95.186	-79.224
h = 4	0.931	0.818	1.289	0.945	0.605	-68.782	-63.759	-83.692	-72.884
COMPR									
h = 1	0.866	0.775	1.589	0.601	0.850	-139.148	-134.630	-166.660	-143.014
h = 2	0.867	0.782	1.605	0.585	0.865	-133.416	-130.948	-159.527	-137.157
h = 4	0.875	0.768	1.614	0.668	0.817	-126.935	-123.506	-152.091	-131.896
LTN									
h = 1	0.883	0.763	1.734	0.560	0.723	-161.679	-159.257	-180.982	-164.208
h = 2	0.908	0.790	1.591	0.516	0.751	-158.522	-156.171	-174.731	-160.877
h = 4	0.901	0.765	1.703	0.504	0.743	-146.687	-144.355	-164.843	-148.829
LPROD									
h = 1	0.920	0.835	1.317	0.784	0.722	-40.861	-34.588	-82.511	-49.240
h = 2	0.885	0.784	1.370	0.769	0.703	-39.968	-34.155	-83.251	-48.666
h = 4	0.852	0.752	1.512	0.705	0.614	-42.918	-35.467	-94.071	-53.132
ITR									
h = 1	0.884	0.787	1.511	1.008	0.733	-68.345	-63.174	-97.348	-73.666
h = 2	0.870	0.764	1.584	1.018	0.694	-68.407	-62.626	-102.050	-74.422
h = 4	0.838	0.729	1.813	0.974	0.649	-65.129	-59.201	-103.702	-72.051

Note: The table presents the values of the log-scores, as well as the ratio of the MSE of the adaptive elastic net, elastic net and adaptive lasso to the MSE of lasso for 32 forecast periods. Values of the MSE below/above 1 indicate a better/worse out-of-sample performance of model in comparison to lasso. The dependent variables are: YER, PCR, XTR, MTR, COMPR, LTN, LPROD and ITR. The original sample is divided into the in-sample period between 1990Q1 and 2005Q4 (64 quarters) and the out-of-sample period between 2006Q1 and 2013Q4 (32 quarters). For the out-of-sample period the forecasts are calculated recursively. The forecasting horizons are $h = 1, 2$ and 4.

elastic net. For exports also for adaptive lasso. The factor model outperforms the considered Bayesian methods only for long-term interest rate, and it performs comparably for GDP and investment. For the rest of the variables it is outperformed by lasso, elastic net and adaptive elastic net. In the case of private consumption also by adaptive lasso.

The situation changes when the short sample is considered. The performance of the AR(2) and the factor model improves a lot in comparison to lasso or adaptive lasso. These Bayesian methods are practically dominated by the two frequentist models for all analyzed variables and all forecasting horizons. A similar result is seen in the case of adaptive elastic net, although not for all variables, and the differences in the performance

between this model and AR(2) or factor model are much smaller than in the case of lasso or adaptive lasso. Only elastic net is comparable to the AR(2) and the factor model.

Table 5: Out-of-sample Mean squared errors and log-scores of adaptive elastic net (AEN), elastic net (EN), adaptive lasso (AL) and lasso (L). Higher lags are shrunk more. Sample: 1970Q1 - 2013Q4.

	MSE		**Log-score**	
	AEN	**AL**	**AEN**	**AL**
YER				
h = 1	0.988	1.075	-36.705	-40.615
h = 2	1.007	1.063	-36.453	-39.864
h = 4	1.021	1.045	-36.144	-38.437
PCR				
h = 1	0.938	1.063	-17.704	-18.479
h = 2	0.923	1.066	-17.767	-18.828
h = 4	0.944	1.067	-16.969	-17.712
XTR				
h = 1	1.023	1.077	-73.308	-74.771
h = 2	1.031	1.068	-73.590	-74.659
h = 4	1.033	1.074	-71.812	-72.562
MTR				
h = 1	0.983	1.137	-62.937	-65.613
h = 2	0.985	1.122	-62.387	-64.718
h = 4	0.973	1.126	-60.410	-63.846
COMPR				
h = 1	0.989	1.088	-131.504	-134.839
h = 2	0.988	1.093	-127.975	-131.445
h = 4	1.006	1.069	-121.772	-123.685
LTN				
h = 1	0.950	1.044	-161.708	-163.765
h = 2	0.942	1.058	-157.475	-159.857
h = 4	0.929	1.059	-147.902	-150.524
LPROD				
h = 1	0.986	1.084	-28.680	-31.601
h = 2	1.004	1.062	-28.453	-30.605
h = 4	1.015	1.049	-28.324	-30.099
ITR				
h = 1	0.967	1.069	-64.051	-67.707
h = 2	0.978	1.055	-63.355	-66.444
h = 4	0.987	1.050	-60.217	-63.113

Note: The table presents the values of the log-scores, as well as the ratio of the MSE of the adaptive elastic net, elastic net and adaptive lasso to the MSE of lasso for 32 forecast periods. Values of the MSE below/above 1 indicate a better/worse out-of-sample performance of model in comparison to lasso. The dependent variables are: YER, PCR, XTR, MTR, COMPR, LTN, LPROD and ITR. The original sample is divided into the in-sample period between 1970Q1 and 2005Q4 (144 quarters) and the out-of-sample period between 2006Q1 and 2013Q4 (32 quarters). Different shrinkage is applied to higher lags of explanatory variables. The hyperparameters of the gamma prior for $\lambda_{1,j}$ for the first lag of all explanatory variables are set to $r_{1,j} = 1$ and $\delta_{1,j} = 1.78$ for all $j = 1, ..., \frac{p}{2}$, whereas for the second lag, the hyperparameters are set to $r_{1,j} = 0.01$ and $\delta_{1,j} = 0.01$ for all $j = \frac{p}{2} + 1, ..., p$. For the out-of-sample period the forecasts are calculated recursively. The forecasting horizons are $h = 1, 2$ and 4.

Table 6: Out-of-sample Mean squared errors and log-scores of adaptive elastic net (AEN), elastic net (EN), adaptive lasso (AL) and lasso (L). Higher lags are shrunk more. Sample: 1990Q1 - 2013Q4.

	MSE		**Log-score**	
	AEN	**AL**	**AEN**	**AL**
YER				
h = 1	0.912	1.207	-46.716	-79.933
h = 2	0.865	1.272	-45.134	-84.507
h = 4	0.831	1.352	-46.183	-93.394
PCR				
h = 1	0.674	1.510	-22.237	-39.812
h = 2	0.682	1.444	-21.731	-37.331
h = 4	0.649	1.470	-22.048	-36.438
XTR				
h = 1	0.846	1.253	-84.978	-126.338
h = 2	0.832	1.305	-85.416	-137.756
h = 4	0.832	1.285	-85.545	-166.218
MTR				
h = 1	0.875	1.268	-70.281	-85.716
h = 2	0.879	1.260	-72.600	-88.752
h = 4	0.902	1.207	-67.305	-79.283
COMPR				
h = 1	0.814	1.392	-134.501	-148.789
h = 2	0.814	1.417	-130.759	-147.262
h = 4	0.826	1.396	-125.314	-139.269
LTN				
h = 1	0.816	1.559	-160.463	-174.465
h = 2	0.841	1.454	-157.366	-169.882
h = 4	0.816	1.583	-145.353	-160.197
LPROD				
h = 1	0.868	1.208	-37.217	-71.240
h = 2	0.815	1.252	-37.114	-71.480
h = 4	0.779	1.353	-38.456	-80.757
ITR				
h = 1	0.857	1.448	-66.078	-94.443
h = 2	0.834	1.579	-65.880	-104.390
h = 4	0.798	1.695	-62.290	-107.645

Note: The table presents the values of the log-scores, as well as the ratio of the MSE of the adaptive elastic net, elastic net and adaptive lasso to the MSE of lasso for 32 forecast periods. Values of the MSE below/above 1 indicate a better/worse out-of-sample performance of model in comparison to lasso. The dependent variables are: YER, PCR, XTR, MTR, COMPR, LTN, LPROD and ITR. The original sample is divided into the in-sample period between 1990Q1 and 2005Q4 (144 quarters) and the out-of-sample period between 2006Q1 and 2013Q4 (32 quarters). Different shrinkage is applied to higher lags of explanatory variables. The hyperparameters of the gamma prior for $\lambda_{1,j}$ for the first lag of all explanatory variables are set to $r_{1,j} = 1$ and $\delta_{1,j} = 1.78$ for all $j = 1, ..., \frac{p}{2}$, whereas for the second lag, the hyperparameters are set to $r_{1,j} = 0.01$ and $\delta_{1,j} = 0.01$ for all $j = \frac{p}{2} + 1, ..., p$. For the out-of-sample period the forecasts are calculated recursively. The forecasting horizons are $h = 1, 2$ and 4.

4.2.2 More shrinkage on higher lags

One can presume that higher lags are less relevant for prediction and should be shrunk more than the low lags. Therefore, in this section the benchmark case hyperparameters of the gamma prior for $\lambda_{1,j}$ are used for the first lag of all explanatory variables: $r_{1,j} = 1$ and $\delta_{1,j} = 1.78$ for all $j = 1, ..., \frac{p}{2}$, whereas for the second lag, the hyperparameters are

set to $r_{1,j} = 0.01$ and $\delta_{1,j} = 0.01$ for all $j = \frac{p}{2} + 1, ..., p$. As different degree of shrinkage is only allowed in adaptive lasso and adaptive elastic net, only these two methods are used in this section. Both methods are compared to the same benchmark as in Section 4.2.1. The results of the exercise with larger shrinkage for higher lags can be found in Table 5 for the long sample and Table 6 for the short sample.

The comparison to the results of Section 4.2.1 shows that there are no considerable differences between the two specifications for the long sample, neither in terms of MSE nor in terms of log-score. For the short sample, some improvement in forecasting performance can be seen both for adaptive lasso and adaptive elastic net, when higher lags are allowed to be shrunk more. This can be observed for all considered variables, although for some of them the improvement is very small. All in all, it seems that shrinking higher lags more can improve the performance of the models with adaptive shrinkage in the short sample case, but the improvement for most variables is rather small.

4.2.3 Autoregressive terms included

In the forecasting literature adding autoregressive terms to the model is often reported to improve its out-of-sample performance considerably. Therefore, I also include autoregressive terms in the models considered in this paper. For the four analyzed Bayesian methods $q = 2$ autoregressive lags are included in the model, but autoregressive terms (two lags of the dependent variable) are also added to the factor model, such that the comparison between all considered models is fair. The degree of shrinkage for the coefficients of the autoregressive terms in the Bayesian models is the same as the degree of shrinkage for the rest of the explanatory variables, that is $r_{1,j} = 1$ and $\delta_{1,j} = 1.78$ for all $j = 1, ..., p + q$. However, in contrast to the rest of the explanatory variables, the autoregressive terms are not standardized but only demeaned to match the dependent variable. The results of the forecasting exercise for all models with autoregressive terms can be found in Table 7 for the long sample and Table 8 for the short sample.

For the long sample the elastic net turns out to have the best out-of-sample performance in terms of MSE among the four Bayesian models, the adaptive lasso, on the other hand, shows the worst performance among all considered Bayesian methods. However, the differences between these four models are quite small for all considered dependent variables. Both frequentist models included in the analysis, the AR(2) and the factor model, show quite good performance. For some variables (private consumption and long-term interest rate) the AR(2) model shows considerably better performance than the best Bayesian method - elastic net. For exports, imports and investment, however, it shows considerably worse performance than elastic net, adaptive elastic net and even lasso. The factor model, on the other hand, shows good performance for the long-term interest rate. For other variables, its performance is comparable or worse than the performance of elastic net, and for exports, imports and commodity prices even worse than adaptive elastic net and lasso. The results of the comparison between the four Bayesian methods in terms of MSE are confirmed also for the log-scores.

For the short sample, just as in the benchmark case from Section 4.2.1, the differences between all considered models grow considerably. In terms of MSE the adaptive lasso

Table 7: Out-of-sample Mean squared errors and log-scores of adaptive elastic net (AEN), elastic net (EN), adaptive lasso (AL) and lasso (L). Models with autoregressive terms. Sample: 1970Q1 - 2013Q4.

	MSE					**Log-score**			
	AEN	**EN**	**AL**	**AR(2)**	**4 fac.**	**AEN**	**EN**	**AL**	**L**
YER									
h = 1	0.993	0.958	1.113	0.959	0.943	-36.673	-35.168	-41.014	-37.762
h = 2	1.001	0.973	1.070	1.017	0.986	-36.264	-35.069	-40.020	-37.122
h = 4	1.013	0.985	1.071	1.006	0.997	-35.938	-34.884	-39.490	-36.488
PCR									
h = 1	0.876	0.803	1.232	0.838	0.992	-15.322	-15.165	-18.550	-15.889
h = 2	0.898	0.825	1.205	0.851	1.010	-15.316	-15.098	-18.067	-15.795
h = 4	0.897	0.850	1.205	0.905	1.099	-14.252	-14.272	-16.759	-14.604
XTR									
h = 1	0.998	0.990	1.062	1.242	1.036	-73.485	-73.260	-75.152	-73.835
h = 2	0.995	0.999	1.032	1.278	1.048	-73.637	-73.664	-73.993	-73.923
h = 4	0.998	0.995	1.067	1.221	1.033	-70.614	-71.518	-73.770	-71.776
MTR									
h = 1	0.970	0.920	1.105	1.199	1.088	-63.190	-62.090	-65.849	-63.980
h = 2	0.969	0.926	1.077	1.196	1.090	-62.741	-61.399	-64.979	-63.267
h = 4	0.974	0.918	1.120	1.150	1.084	-60.768	-59.382	-63.570	-61.553
COMPR									
h = 1	0.981	0.960	1.145	0.944	1.085	-131.394	-130.245	-136.389	-131.700
h = 2	0.982	0.959	1.156	0.933	1.087	-127.543	-126.634	-132.623	-127.791
h = 4	0.988	0.977	1.121	0.993	1.094	-121.016	-120.301	-124.681	-121.739
LTN									
h = 1	0.983	0.941	1.059	0.738	0.818	-162.409	-161.283	-164.544	-162.859
h = 2	0.980	0.930	1.084	0.720	0.814	-158.632	-157.240	-161.254	-159.112
h = 4	0.984	0.926	1.084	0.710	0.807	-148.072	-146.627	-150.517	-148.566
LPROD									
h = 1	0.997	0.954	1.118	0.922	0.924	-28.788	-27.759	-32.236	-29.256
h = 2	1.005	0.968	1.077	0.980	0.966	-28.728	-27.558	-30.948	-28.723
h = 4	1.017	0.983	1.079	0.987	0.980	-28.250	-27.395	-30.519	-28.465
ITR									
h = 1	0.984	0.956	1.069	1.070	0.962	-63.814	-62.299	-67.055	-64.448
h = 2	0.994	0.970	1.045	1.102	1.001	-63.255	-61.740	-65.897	-63.437
h = 4	0.996	0.976	1.034	1.106	1.029	-59.566	-58.600	-62.127	-60.334

Note: The table presents the values of the log scores, as well as the ratio of the MSE of the adaptive elastic net, elastic net and adaptive lasso to the MSE of lasso for 32 forecast periods. Values of the MSE below/above 1 indicate a better/worse out-of-sample performance of model in comparison to lasso. The dependent variables are: YER, PCR, XTR, MTR, COMPR, LTN, LPROD and ITR. The original sample is divided into the in-sample period between 1970Q1 and 2005Q4 (144 quarters) and the out-of-sample period between 2006Q1 and 2013Q4 (32 quarters). 2 lags of the dependent variable are included in all the models. For the out-of-sample period the forecasts are calculated recursively. The forecasting horizons are $h = 1, 2$ and 4.

performs much worse than all other analyzed models. Adaptive elastic net dominates the performance of lasso, and elastic net dominates the rest of the Bayesian methods for all considered variables. However, also AR(2) and the factor model perform very well. The AR(2) considerably beats elastic net for some of the variables (private consumption, commodity prices, long-term interest rates and labor productivity). The factor model outperforms elastic net for almost all variables (GDP, private consumption, exports, imports, long-term interest rate, labor productivity and investment). Thus, it seems that in the more challenging case (short sample) the considered Bayesian methods cannot out-

Table 8: Out-of-sample Mean squared errors and log-scores of adaptive elastic net (AEN), elastic net (EN), adaptive lasso (AL) and lasso (L). Models with autoregressive terms. Sample: 1990Q1 - 2013Q4.

	MSE					**Log-score**			
	AEN	**EN**	**AL**	**AR(2)**	**4 fac.**	**AEN**	**EN**	**AL**	**L**
YER									
h = 1	0.934	0.875	1.252	0.842	0.763	-49.794	-43.561	-83.188	-56.072
h = 2	0.907	0.819	1.317	0.814	0.724	-47.377	-42.407	-90.160	-55.286
h = 4	0.879	0.791	1.454	0.756	0.657	-49.352	-43.751	-101.943	-61.262
PCR									
h = 1	0.968	0.768	1.447	0.585	0.589	-19.980	-17.392	-38.941	-21.680
h = 2	0.971	0.774	1.512	0.618	0.630	-19.380	-16.962	-37.935	-20.364
h = 4	0.994	0.748	1.567	0.538	0.518	-20.017	-17.207	-43.012	-21.029
XTR									
h = 1	0.893	0.790	1.258	0.775	0.717	-88.040	-80.330	-127.869	-100.483
h = 2	0.876	0.773	1.311	0.734	0.669	-90.510	-81.465	-127.703	-98.350
h = 4	0.875	0.769	1.275	0.729	0.646	-90.570	-80.032	-137.322	-104.648
MTR									
h = 1	0.915	0.809	1.408	1.021	0.694	-71.560	-67.020	-90.545	-76.026
h = 2	0.912	0.804	1.389	0.958	0.608	-73.576	-68.226	-94.618	-78.313
h = 4	0.927	0.832	1.294	0.969	0.625	-68.234	-64.079	-83.356	-72.267
COMPR									
h = 1	0.852	0.733	1.649	0.621	0.872	-136.925	-133.252	-168.029	-143.595
h = 2	0.843	0.722	1.669	0.602	0.879	-133.080	-128.687	-160.381	-137.028
h = 4	0.860	0.727	1.656	0.676	0.816	-126.828	-122.618	-153.504	-132.573
LTN									
h = 1	0.865	0.761	1.622	0.595	0.671	-164.985	-161.597	-192.671	-169.904
h = 2	0.873	0.763	1.460	0.556	0.649	-162.581	-158.729	-186.503	-167.594
h = 4	0.867	0.764	1.619	0.668	0.727	-145.782	-143.290	-169.039	-149.656
LPROD									
h = 1	0.914	0.834	1.313	0.781	0.729	-40.366	-34.384	-82.759	-49.375
h = 2	0.874	0.777	1.366	0.760	0.682	-39.497	-34.133	-78.605	-48.598
h = 4	0.846	0.747	1.514	0.702	0.606	-41.932	-36.082	-103.572	-53.659
ITR									
h = 1	0.877	0.782	1.563	0.987	0.697	-68.844	-63.555	-97.218	-74.165
h = 2	0.867	0.757	1.619	0.995	0.666	-68.492	-62.852	-102.695	-74.502
h = 4	0.835	0.724	1.856	0.946	0.607	-65.477	-59.712	-103.579	-72.145

Note: The table presents the values of the log scores, as well as the ratio of the MSE of the adaptive elastic net, elastic net and adaptive lasso to the MSE of lasso for 32 forecast periods. Values of the MSE below/above 1 indicate a better/worse out-of-sample performance of model in comparison to lasso. The dependent variables are: YER, PCR, XTR, MTR, COMPR, LTN, LPROD and ITR. The original sample is divided into the in-sample period between 1990Q1 and 2005Q4 (64 quarters) and the out-of-sample period between 2006Q1 and 2013Q4 (32 quarters). 2 lags of the dependent variable are included in all the models. For the out-of-sample period the forecasts are calculated recursively. The forecasting horizons are $h = 1, 2$ and 4.

perform the standard frequentist techniques, such as AR(2) and the factor model. For the less challenging case (long sample) the comparison is more in favor of the best among the Bayesian models (elastic net), at least for some of the considered variables.

4.3 Robustness checks

The empirical analysis from Section 4.2.1 was repeated for all Bayesian models for a few other specifications of the hyperparameters of the gamma priors for $\lambda_{1,j}$ and λ_2.[6] The hyperparameters used in the robustness checks are the following ones:

(a) $r_{1,j} = r_2 = 0.01$ and $\delta_{1,j} = \delta_2 = 0.01$

(b) $r_{1,j} = r_2 = 0.1$ and $\delta_{1,j} = \delta_2 = 0.1$

(c) $r_{1,j} = r_2 = 1$ and $\delta_{1,j} = \delta_2 = 0.1$

(d) $r_{1,j} = r_2 = 1$ and $\delta_{1,j} = \delta_2 = 3$

(e) $r_{1,j} = r_2 = 3$ and $\delta_{1,j} = \delta_2 = 1$

for all $j = 1, ..., p$. The results for hyperparameters in (d) and (e) are very similar to the results obtained for the benchmark specification. When the degree of shrinkage becomes large (hyperparameters in (a), (b), and (c)) the results for the short sample are still similar to the results for the benchmark case, but adaptive elastic net tends to be slightly better than elastic net. Both methods are still better than lasso and much better than adaptive lasso. For the long sample the differences between all four Bayesian methods become quite small, but adaptive elastic net and lasso show the best forecasting performance. Thus, for moderate degree of shrinkage the results of the empirical exercise are robust to the specifications of the hyperparameters of the gamma priors for $\lambda_{1,j}$ and λ_2. For large degree of shrinkage the adaptive elastic net in many cases outperforms the elastic net in forecasting. However, imposing very high degree of shrinkage on the $\boldsymbol{\beta}$ coefficients is reasonable only when one can assume that the predictors have very little predictive power in relation to the dependent variable, and thus can be shrunk very strongly. However, when one assumes that at least some of the explanatory variables can help to forecast the dependent variable, a moderate degree of shrinkage seems more reasonable. Consequently, the whole analysis in this paper is carried out for moderate degree of shrinkage.

5 Conclusion

In this paper I introduce adaptive elastic net in a Bayesian framework. Then I test its forecasting performance against Bayesian lasso, Bayesian adaptive lasso and Bayesian elastic net in a series of simulations, as well as in an empirical exercise for Euro area data. In the empirical comparison I also include two frequentist models, a simple AR model and a factor model.

The results of the simulations, as well as of the empirical exercise suggest that elastic net is the best model among the four Bayesian methods considered in the paper. Adaptive elastic net is the second best Bayesian method, while adaptive lasso shows the worst forecasting

[6]The results of the robustness checks are available on request.

performance. Lasso is generally better then adaptive lasso, but worse than adaptive elastic net. The differences in the forecasting performance of these models become especially pronounced when the number of regressors grows considerably relative to the number of available observations. This suggests that the ridge regression component in the elastic net model is mainly responsible for its improvement in forecasting performance over lasso. Allowing for different shrinkage parameters λ_1 for different coefficients does not seem to play a major role, and may even lead to a considerable deterioration of the forecasting performance, as in the case of the adaptive lasso.

Surprisingly, the two selected standard methods, an AR and a factor model, show very good performance in comparison to the Bayesian lasso-type of methods. For some specifications (long sample) elastic net, adaptive elastic net and even lasso show better forecasting performance than the frequentist models for most considered variables. However, for the short sample case (especially when autoregressive terms are included in all models) the AR and the factor model beat the Bayesian methods for most analyzed variables.

All in all, the Bayesian adaptive elastic net shows very good forecasting performance in comparison to the Bayesian lasso and adaptive lasso. It also performs well in comparison to the Bayesian elastic net, although it is never better than its counterpart without the adaptive shrinkage. Thus, in some applications the Bayesian adaptive elastic net might be an interesting alternative to the Bayesian elastic net and other lasso-type methods.

References

Andrews, D. F. and C. L. Mallows (1974). Scale Mixtures of Normal Distributions. *Journal of the Royal Statistical Society 36*, 99–102.

Atchade, Y. F. (2011). A computational framework for empirical Bayes inference. *Statistics and Computing 21*, 463–473.

Bai, J. and S. Ng (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica 70*(1), 191–221.

Casella, G. (2001). Empirical Bayes Gibbs sampling. *Biostatistics 2*, 485–500.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least Angle Regression. *The Annals of Statistics 32*, 407–499.

Fan, J. and R. Li (2001). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association 96*, 1348–1360.

Fan, J. and H. Peng (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics 32 (3)*, 928–961.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2005). The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting. *Journal of the American Statistical Association 100*(471), 830–840.

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

George, E. I. and R. E. McCulloch (1993). Variable selection via Gibbs sampling. *journal of the American Statistical Association 88*, 881–889.

George, E. I. and R. E. McCulloch (1997). Approaches for Bayesian Variable Selection. *Statistica Sinica 7*, 339–374.

Ghysels, E., P. Santa-Clara, and R. Valkanov (2002). The MIDAS touch: Mixed Data Sampling Regression Models. Working Paper, UNC and UCLA.

Hoerl, A. and R. Kennard (1970). Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics 12*, 55–68.

Korobilis, D. (2013). Hierarchical shrinkage priors for dynamic rregression with many predictors. *International Journal of Forecasting 29*, 43–59.

Kyung, M., J. Gill, M. Ghosh, and G. Casella (2010). Penalized Regression, Standard Errors, and Bayesian Lassos. *Bayesian Analysis 5 (2)*, 369–412.

Lehmann, E. L. and G. Casella (1998). *Theory of Point Estimation (2nd edition)*. New York: Springer.

Leng, C., M.-N. Tran, and D. Nott (2014). Bayesian Adaptive Lasso. *Annals of Institute of Statistical Mathematics 66*, 221–244.

Li, Q. and N. Lin (2010). The Bayesian Elastic Net. *Bayesian Analysis 5*, 151–170.

Michael, J. R., W. R. Schucany, and R. W. Haas (1976). Generating Random Variates Using Transformations with Multiple Roots. *The American Statistician 30 (2)*, 88–90.

Mol, C. D., D. Giannone, and L. Reichlin (2008). Forecasting using a large number of predictors: is Bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics 146*, 318–328.

Osborne, M. R., B. Presnell, and B. A. Turlach (2000). A New Approach to Variable Selection in Least Squares Problems. *IMA Journal of Numerical Analysis 20*, 389–404.

Park, T. and G. Casella (2008). The Bayesian Lasso. *Journal of the American Statistical Association 103*, 681–686.

Stock, J. and M. Watson (2002). Forecasting Using Principal Components From a Large Number of Predictors. *Journal of the American Statistical Association 97*(460), 1167–1179.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society 58*, 267–288.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and Smoothness via the Fused Lasso. *Journal of the Royal Statistical Society Series B (67)*, 91–108.

Yuan, M. and Y. Lin (2006). Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society Series B (68)*, 49–67.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association 101*, 1418–1429.

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

Zou, H. and H. H. Zhang (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics 37 (4)*, 1733–1751.

A Appendix

Below one can find the prior distributions and the full conditional posterior distributions for all the parameters in the model, estimated by Bayesian lasso, Bayesian adaptive lasso and Bayesian elastic net. For all three methods a non-informative, scale-invariant prior on σ^2 is used, which takes the following form: $\pi(\sigma^2) \propto \frac{1}{\sigma^2}$.

Lasso

Prior distributions:

$$\boldsymbol{\beta}|\,\sigma^2, \tau_1^2, ..., \tau_p^2 \sim N_p(\mathbf{0}, \sigma^2 V), \text{ where } V = \text{ diag}\left\{\tau_1^2, ...\tau_p^2\right\}$$

$$\tau_j^2 \sim \text{exponential}\left(\frac{\lambda^2}{2}\right) \text{ for } j = 1, ..., p$$

$$\lambda^2 \sim \text{gamma}(r, \delta)$$

Full conditional posterior distributions:

$$\boldsymbol{\beta}|\,\sigma^2, \left\{\tau_j^2\right\}_{j=1}^p, \lambda, \boldsymbol{X}, \tilde{\boldsymbol{y}} \sim N_p\left((\boldsymbol{X'X} + \boldsymbol{V}^{-1})^{-1}\boldsymbol{X'}\tilde{\boldsymbol{y}}, \sigma^2(\boldsymbol{X'X} + \boldsymbol{V}^{-1})^{-1}\right)$$

$$\sigma^2\Big|\,\boldsymbol{\beta}, \left\{\tau_j^2\right\}_{j=1}^p, \lambda, \boldsymbol{X}, \tilde{\boldsymbol{y}} \sim \text{inv. gamma}\Bigg(\frac{T-1+p}{2}, \frac{1}{2}\left((\tilde{\boldsymbol{y}} - \boldsymbol{X\beta})'(\tilde{\boldsymbol{y}} - \boldsymbol{X\beta}) + \boldsymbol{\beta'V}^{-1}\boldsymbol{\beta}\right)\Bigg)$$

$$\frac{1}{\tau_j^2}\Bigg|\,\boldsymbol{\beta}, \sigma^2, \lambda, \boldsymbol{X}, \tilde{\boldsymbol{y}} \sim \text{inverse Gaussian}\left(\sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}}, \lambda^2\right) \text{ for } \; j = 1, ..., p$$

$$\lambda^2\Big|\,\boldsymbol{\beta}, \sigma^2, \left\{\tau_j^2\right\}_{j=1}^p, \boldsymbol{X}, \tilde{\boldsymbol{y}} \sim \text{gamma}\left(p + r, \frac{1}{2}\sum_{j=1}^p \tau_j^2 + \delta\right)$$

Adaptive lasso

Prior distributions:

$$\boldsymbol{\beta}|\,\sigma^2, \tau_1^2, ..., \tau_p^2 \sim N_p(\mathbf{0}, \sigma^2 V), \text{ where } V = \text{ diag}\left\{\tau_1^2, ...\tau_p^2\right\}$$

$$\tau_j^2 \sim \text{exponential}\left(\frac{\lambda_j^2}{2}\right) \text{ for } j = 1, ..., p$$

$$\lambda_j^2 \sim \text{gamma}(r_j, \delta_j) \text{ for } j = 1, ..., p$$

Full conditional posterior distributions:

$$\boldsymbol{\beta}|\,\sigma^2, \left\{\tau_j^2\right\}_{j=1}^p, \left\{\lambda_j^2\right\}_{j=1}^p, \boldsymbol{X}, \tilde{\boldsymbol{y}} \sim N_p\left((\boldsymbol{X'X} + \boldsymbol{V}^{-1})^{-1}\boldsymbol{X'}\tilde{\boldsymbol{y}}, \sigma^2(\boldsymbol{X'X} + \boldsymbol{V}^{-1})^{-1}\right)$$

$$\sigma^2\Big|\,\boldsymbol{\beta}, \left\{\tau_j^2\right\}_{j=1}^p, \left\{\lambda_j^2\right\}_{j=1}^p, \boldsymbol{X}, \tilde{\boldsymbol{y}} \sim \text{inv. gamma}\Bigg(\frac{T-1+p}{2}, \frac{1}{2}\left((\tilde{\boldsymbol{y}} - \boldsymbol{X\beta})'(\tilde{\boldsymbol{y}} - \boldsymbol{X\beta}) + \boldsymbol{\beta'V}^{-1}\boldsymbol{\beta}\right)\Bigg)$$

$$\frac{1}{\tau_j^2}\Big|\,\boldsymbol{\beta},\sigma^2,\left\{\lambda_j^2\right\}_{j=1}^p,\boldsymbol{X},\tilde{\boldsymbol{y}}\sim\text{inverse Gaussian}\left(\sqrt{\frac{\lambda_j^2\sigma^2}{\beta_j^2}},\lambda_j^2\right)\text{ for }j=1,...,p$$

$$\lambda_j^2\big|\,\boldsymbol{\beta},\sigma^2,\left\{\tau_j^2\right\}_{j=1}^p,\boldsymbol{X},\tilde{\boldsymbol{y}}\sim\text{gamma}\left(r_j+1,\frac{\tau_j^2}{2}+\delta_j\right)\text{ for }j=1,...,p$$

Elastic net

Prior distributions:

$$\boldsymbol{\beta}\big|\,\sigma^2,\tau_1^2,...,\tau_p^2\sim N_p(\mathbf{0},\sigma^2 D)$$

$$\tau_j^2\sim\text{exponential}\left(\frac{\lambda_1^2}{2}\right)\text{ for }j=1,...,p$$

$$\lambda_1^2\sim\text{gamma}(r_1,\delta_1)$$

$$\lambda_2^2\sim\text{gamma}(r_2,\delta_2)$$

Full conditional posterior distributions:

$$\boldsymbol{\beta}\big|\,\sigma^2,\left\{\tau_j^2\right\}_{j=1}^p,\lambda_1,\lambda_2,\boldsymbol{X},\tilde{\boldsymbol{y}}\sim N_p\left((\boldsymbol{X}'\boldsymbol{X}+D^{-1})^{-1}\boldsymbol{X}'\tilde{\boldsymbol{y}},\sigma^2(\boldsymbol{X}'\boldsymbol{X}+D^{-1})^{-1}\right)$$

$$\sigma^2\big|\,\boldsymbol{\beta},\left\{\tau_j^2\right\}_{j=1}^p,\lambda_1,\lambda_2,\boldsymbol{X},\tilde{\boldsymbol{y}}\sim\text{inv. gamma}\left(\frac{T-1+p}{2},\right.$$

$$\left.\frac{1}{2}\left((\tilde{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta})'(\tilde{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta})+\boldsymbol{\beta}'D^{-1}\boldsymbol{\beta}\right)\right)$$

$$\frac{1}{\tau_j^2}\Big|\,\boldsymbol{\beta},\sigma^2,\lambda_1,\lambda_2,\boldsymbol{X},\tilde{\boldsymbol{y}}\sim\text{inverse Gaussian}\left(\sqrt{\frac{\lambda_1^2\sigma^2}{\beta_j^2}},\lambda_1^2\right)\text{ for }j=1,...,p$$

$$\lambda_1^2\big|\,\boldsymbol{\beta},\sigma^2,\left\{\tau_j^2\right\}_{j=1}^p,\lambda_2,\boldsymbol{X},\tilde{\boldsymbol{y}}\sim\text{gamma}\left(p+r_1,\frac{1}{2}\sum_{j=1}^p\tau_j^2+\delta_1\right)$$

$$\lambda_2\big|\,\boldsymbol{\beta},\sigma^2,\left\{\tau_j^2\right\}_{j=1}^p,\lambda_1,\boldsymbol{X},\tilde{\boldsymbol{y}}\sim\text{gamma}\left(\frac{p}{2}+r_2,\frac{1}{2\sigma^2}\sum_{j=1}^p\beta_j^2+\delta_2\right)$$

Table A.1: Data used in the empirical exercise for the Euro area

No.	Data code	Data description	Transformation code
1	COMPR	Commodity Prices (in USD)	5
2	EEN	Effective exchange rate	5
3	GCD	Government Consumption Deflator	6
4	GCR	Government Consumption (Real)	6
5	GON	Gross Operating Surplus (Nominal)	6
6	HICP	Overall HICP (Non-seasonally adjusted)	6
7	ITD	Gross Investment Deflator	6
8	ITR	Gross Investment (Real)	5
9	LEN	Employees (persons)	5
10	LFN	Labor Force (persons)	5
11	LNN	Total Employment (persons)	5
12	LPROD	Labor Productivity (YER/LNN)	5
13	LTN	Long-Term Interest Rate (Nominal)	2
14	MTD	Imports of Goods and Services Deflator	6
15	MTR	Imports of Goods and Services (Real)	5
16	PCD	Consumption Deflator	6
17	PCOMU	Non-oil commodity prices (in USD)	5
18	PCR	Private Consumption (Real)	6
19	POILU	Oil prices (in USD)	5
20	SAX	Household's savings ratio	2
21	STN	Short-Term Interest Rate (Nominal)	2
22	TIN	Indirect Taxes (net of subsidies)	6
23	ULC	Unit Labor Costs (WIN/YER)	6
24	UNN	Number of Unemployed	5
25	URX	Unemployment rate (as a percentage of labor force)	2
26	WIN	Compensation to Employees	6
27	WRN	Wage per head	6
28	XTD	Exports of Goods and Services Deflator	6
29	XTR	Exports of Goods and Services (Real)	5
30	YED	Gross Domestic Product Deflator	6
31	YER	Gross Domestic Product (Real)	5
32	YFD	Gross Domestic Product at Factor Costs Deflator	6
33	YFN	Gross Domestic Product at Factor Costs (WIN + GON)	6
34	YIN	Gross Domestic Product, Income Side	6
35	YWD	World Gross Domestic Product Deflator	6
36	YWR	World Gross Domestic Product (Real)	5
37	YWRX	World Demand, Composite Indicator	5

Note: The above data come from the Area Wide Model (AWM) dataset, available at the following website: http://www.eabcn.org/area-wide-model.
The variables have been transformed to achieve stationarity. The transformations together with their codes are:
2 - data transformed by taking first differences
5 - data transformed by taking first differences of logarithms
6 - data transformed by taking second differences of logarithms

UNIVERSITY OF KONSTANZ
Department of Economics

Universitätsstraße 10
78464 Konstanz
Germany

Phone: +49 (0) 7531-88-0
Fax: +49 (0) 7531-88-3688

www.wiwi.uni-konstanz.de/econdoc/working-paper-series/