Regularized Estimation in High-Dimensional Vector Autoregressive Models

Master Thesis

Department of Economics

University of Mannheim

submitted to:
Prof. Dr. Carsten Trenkler

submitted by:
Ratmir Miftachov

Student ID: 1632051
Degree Programme: Master of Science in Economics (M.Sc.)

Mannheim, 25.09.2020

Contents

Abstract

This thesis focuses on forecasting as well as structural analysis using high-dimensional Vector Autoregressive models. We adapt the novel feature weighted elastic net to the VAR model and propose three different grouping schemes on the penalization term. The double hyperparameter cross validation procedure is adapted to time dependent data. Empirically, we obtain more accurate predictions by using many variables for U.S. macroeconomic data. The Curse of Dimensionality is investigated as a major problem in application. All shrinkage methods of consideration result in (mostly) credible structural impulse response functions in a high-dimensional setting. A controlled simulation study shows that regularization is particularly useful when sample size is small.

Key words: Elastic Net, Feature Weighted Elastic Net, VAR, Forecasting, Structural Analysis

1 Introduction

Introduced by Sims (1980), Vector Autoregressions (VARs) are capable of capturing complex dynamics among the variables of a system, cross sectional as well as temporal. They have shown to be particularly useful for describing the behavior of macroeconomic or financial time series data. The major goals of the applied macroeconomic framework are to conduct accurate forecasts as well as structural analyses. Hereby, this work aims to contribute on both disciplines by using regularized estimation methods. First, we will motivate the crucial problems emerging in the literature and then turn to the main objectives and contributions of this work.

The applied macroeconomic framework typically separates into two categories. On the one hand, researchers try to keep the model rather small but interpretable. The first VAR model (Sims, 1980) falls into this category and was successfully estimated via OLS. On the other hand, the Bayesian literature, pioneered by Litterman (1986), is able to estimate large models, without focusing much on the particular interpretation of the included variables. Similar to the Bayesian literature, the Factor Model approach, pioneered by Stock and Watson (2002), is able to extract relevant information from a large space of variables. However, the motivation for large models is not ambiguous. By including many variables, the researcher seeks to improve the forecast accuracy. As the literature shows, a forecast can indeed improve as the number of variables increases (see, Banbura et al. (2010); Carriero et al. 2011; Koop (2013); and many others), but is not guaranteed. On the contrary, if the researcher fails to include relevant variables, an omitted variable bias can occur, becoming especially perceptible in the structural analysis. An example of latter problem is the Price Puzzle, which often occurs in structural impulse response analyses, encountered by Christiano et al. (1999). In our empirical analysis we will illustrate as well as diminish this phenomenon by estimating a large VAR model. An impulse response analysis is a well known technique, used to analyze the impact of a shock to one variable on the whole system of variables. This type of analysis enables to inspect the dynamics of a system of variables.

A major disadvantage of large VAR models is the *Curse of Dimensionality* or *overfitting*. Often, the data provided to the researcher is either measured on a monthly or (more often) on a quarterly frequency, resulting in a limited number of observations. Additionally, the number of variables in a Vector Autoregressive Model can get large quite quickly. The first paper on VAR models, Sims (1980), already mentions this emerging issue. Consequently, Sims suggested the need of "mean-square-error shrinking devices" to circumvent the problem of overfitting. However, if the researcher ignores this requirement it results in severely distorted predictions.

Many articles on forecasting as well as structural analysis have been published in the meantime. De Mol et al. (2008) analyze Bayesian regression methods using Gaussian as well as double exponential priors. Hereby, the double exponential prior brings the important property of variable selection, which is also of major importance in our work.

Subsequently to this paper, Banbura et al. (2010) empirically show that by setting the degree of shrinkage in proportion to model size, the authors are able to evade the problems of overfitting and multicollinearity. The problem of multicollinearity is especially prominent in high dimensional VARs. Again, both of these related issues interestingly come up in our work. Giannone et al. (2015) propose a hierarchical modeling approach on the optimal choice of prior informativeness, resulting in good forecasting performance as well as accurate impulse response functions. All of the previous authors were interested in the estimation of large systems. In addition, many other different approaches have been made to reduce the parameter space. For example, Stock and Watson (2002) use the approximate Dynamic Factor Model to capture the major variation of the data by a small amount of factors, using a principal component analysis approach. This model is particularly well suited for accurate forecasts and we will use it as a benchmark in our work. However, in its bare version, it is not possible to conduct an impulse response analysis using this model.

Despite these methods tailored to the applied macroeconomic literature, a framework aiming for accurate prediction emerges in the statistical literature. The so called ridge regression (Hoerl and Kennard, 1970) imposes an additional penalty term in the least squares regression, reducing the prediction variance while increasing the bias, by shrinking the coefficient magnitudes towards zero. In addition, the ridge is known to handle correlation among the predictors quite well. Another workhorse method to manage the Curse of Dimensionality in the regression framework is the lasso, popularized by Tibshirani (1996). This method is able to shrink the coefficients towards zero, as well as select important variables simultaneously. Hereby, the variable selection characteristic corresponds to selecting a subset of coefficients into the model and shrinking the remaining to zero. Believing that the true model is sparse, i.e. a subset of coefficients is indeed exactly zero, the lasso brings an attractive characteristic delimiting it from the general Bayesian and Factor model literature. Within this work, the term "sparsity" refers to the number of coefficients equal to zero divided by the total number of coefficients in the model.

The VAR model can be formulated as a system of unrelated equations, where each lag of a certain variable represents an independent variable. Hence, the researcher is able to adapt the regularization techniques emerging from the statistical literature on VARs. Accordingly, Hsu et al. (2008) expand the lasso for the VAR model and investigate the theoretical as well as empirical properties. Further work by Song and Bickel (2011) followed up and imposed additional structure on the lasso VAR, enabling to distinguish between variables' own lags and other lags. Consequently, Nicholsen et al. (2014) develop several further extensions of the lasso-VAR focusing particularly on the lag structure of VAR models.

However, the lasso faces two limitations, which possibly limit the analysis for VARs. First, it has difficulties in variable selection, given a group of highly correlated variables. Since high correlation is not a rarity in macroeconomic variables, there is incentive to avoid

this issue. Second, it is not able to select more variables than the number of observations. In extra large VAR models, this problem might also occur. Consequently, Zou and Hastie (2005) publish the elastic net in the statistical literature. This method can be seen as a mixture between the lasso as well as the ridge method, including the advantages of both methods. Accordingly, it might represent a suitable method for macroeconomic data.

This work seeks to contribute on the bridge between the regularization and applied macroeconomic literature. Hereby, the elastic net regularization method is of key interest, due to its attractive features. We begin by introducing this method for the linear regression model, following Zou and Hastie (2005) and Friedman et al. (2010). Intuition for shrinkage as well as variable selection is given. In order to obtain the tuning parameter of the elastic net, we use Cross Validation (CV), instead of information criteria. In addition, we adapt the CV procedure in a "rolling scheme", tailored for time dependent data. By this adjustment, we intend to improve forecasting performance. To our best knowledge, the previous literature does not incorporate such an adjustment for double parameter CV in a time series setting. Despite the attractive characteristics of the elastic net, it does not assign any additional structure on the penalization of coefficients. As already motivated previously, imposing additional structure on the covariates is a welcomed attribute in the VAR framework. Consequently, we introduce the novel feature weighted elastic net (fenet) regularization on the bare linear regression model, following Tay, Aghaeepour, Hastie and Tibshirani (2020). Then, we extend the elastic net as well as the feature weighted elastic net formally to the VAR model. Hereby, we write the VAR model as a system of equations and validate the tuning parameters equation-wise. By ignoring cross equational dependencies, we are able to obtain equation-wise tailored tuning parameter sets. To the best of our knowledge, we are the first to apply the fenet penalization to the VAR model, formally, as well as empirically. Furthermore, we acknowledge the flexibility of imposing additional structure on the penalization term and propose three different grouping schemes for the fenet model. All three schemes are motivated by the Bayesian literature and the lasso VAR related literature. An attractive characteristic of our model is that we do not need to specify the degree of shrinkage among the variables, while simultaneously being able to impose additional structure on the penalization. The degree of shrinkage on our imposed structure is estimated automatically by an extended pathwise Coordinate Descent algorithm.

Subsequently to the novel formal extensions of this work, we compare the elastic net and its novel extension to the pioneer models lasso VAR and ridge VAR in a forecasting exercise. As often seen in the literature (e.g. Giannone et al. (2015) or Koop and Korobilis (2013)), we predict three major macroeconomic variables of the U.S. using increasing sets of variables. Furthermore, the performance of the whole VAR system is assessed based on a multivariate forecasting evaluation measure. We find that a majority of the regularization methods is able to outperform the popular approximate Dynamic Factor Model (Stock and Watson, 2002), which we use as a benchmark. Moreover, we empirically find a pattern in

the degree of sparsity as well as optimal tuning parameters, in dependence of model size. In addition, the variable selection characteristic of lasso VAR and enet VAR is empirically compared.

Besides prediction performance, a structural analysis is of major interest in this work. Thus, we investigate impulse responses to a monetary policy shock following Christiano et al. (1998) for four different sized sets of variables. We will encounter that OLS results in inflated impulse responses for the large set of variables, due to in-sample overfitting. Contrary, the regularized methods are able to circumvent this problem. By specifying a large model, we are able to reduce the "Price Puzzle". Finally, we illustrate, that the regularization methods considered in this work mostly result in credible impulse response functions in a large model setup. However, an empirical structural analysis does not include the true impulse response functions. Consequently, we conduct a controlled Monte Carlo study and evaluate estimated structural impulse responses to their known counterpart, enabling us to compare in-sample accuracy between regularized VARs. Regarding the simulation setup, we variate in sample size and degree of sparsity in the data generating process.

Particularly, we are mainly interested in the following research questions in this work: Do we detect an improvement in prediction accuracy for larger model sizes? Does forecast precision improve by including a second tuning parameter, as in the enet and the fenet, or should the researcher rather stick to the more "simple" models, lasso and ridge? Are the methods incorporating penalization capable of structural analysis in the large model setup? Does a particular model dominate in terms of impulse response accuracy in a controlled Monte Carlo experiment for different levels of sparsity as well as sample sizes? Are the shrinkage methods capable of handling the major problems in high dimensional VAR systems: overfitting and emerging multicollinearity?

In the first two chapters 2 and 3, we elaborate the methodology of this work. In chapter 4.1 the forecasting exercise is conducted. Afterward, we study a structural analysis in terms of a monetary policy shock in chapter 4.7. Consequently, we apply a controlled Monte Carlo study for impulse response functions in chapter 5. Finally, we summarize our findings and conclude.

2 Regularization

2.1 Elastic Net

As introduced by Zou and Hastie (2005), the elastic net penalty combines the penalty terms of two prominent previously introduced models, the lasso (Tibshirani, 1998) and the ridge (Hoerl and Kennard, 1970). In this chapter, we elaborate the properties of the elastic net model from a bare linear regression point of view in order to illustrate its appropriateness for the VAR context. Note that we waive a detailed discussion of the lasso as well as the ridge in terms of properties. However, we give intuition whenever appropriate, since it motivates the elastic net. The precise ridge as well as lasso minimization problem can be found in the mathematical appendix (definition 1 and definition 2). Let y_i be the dependent variable and x_{ij} the ith observation of variable j for $i = 1, ..., N$ observations and $j = 1, ..., p$ independent variables, such that the linear model is

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i \quad \text{for} \quad i = 1, ..., N$$

The coefficient vector is defined as a column vector $\beta = (\beta_1, ..., \beta_p)' \in \mathbb{R}^{p \times 1}$. For simplicity, a location and scale transformation is conducted. Consequently, the researcher is able to assume that the dependent variable is centred and the predictors are standardized, s.t.

$$\sum_{i=1}^{N} y_i = 0, \quad \sum_{i=1}^{N} x_{ij} = 0 \quad \sum_{i=1}^{N} x_{ij}^2 = 1 \quad \text{for} \quad j = 1, 2, ..., p.$$

Then, the initial elastic net Lagrange function is defined as

$$L(\beta_0, \beta)_{\lambda_1, \lambda_2} = \frac{1}{2} \sum_{i=1}^{N} (y_i - \sum_{j=0}^{p} \beta_j x_{ij})^2 + \lambda_1 ||\beta||_1 + \frac{1}{2} \lambda_2 ||\beta||_2^2 \tag{1}$$

, where a convenient interpretation of the tuning parameter emerges by defining $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$, such that the solution for the Lagrange function 1 is equivalently obtained as

$$\{\hat{\beta}_0, \hat{\beta}\} = \underset{\{\beta_0, \beta\}}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{N} (y_i - \sum_{j=0}^{p} \beta_j x_{ij})^2, \quad \text{s.t.} \quad \alpha ||\beta||_1 + \frac{1}{2}(1 - \alpha) ||\beta||_2^2 \leq t. \tag{2}$$

, which once again is written in Lagrange form as

$$L(\beta_0, \beta)_{\alpha, \lambda} = \frac{1}{2} \sum_{i=1}^{N} (y_i - \sum_{j=0}^{p} \beta_j x_{ij})^2 + \lambda(\alpha ||\beta||_1 + \frac{1}{2}(1 - \alpha) ||\beta||_2^2). \tag{3}$$

Hereby, the Residual Sum of Squares (RSS) is defined as $\sum_{i=1}^{N} (y_i - \sum_{j=0}^{p} \beta_j x_{ij})^2$. For the majority of this work, we use notation 3, since it enables a convenient interpretation

of the tuning parameters λ and α. As we can directly see from the penalty[1] $\mathcal{P}^{enet} := \lambda(\alpha||\beta||_1 + \frac{1}{2}(1-\alpha)||\beta||_2^2)$, it nests the ridge L2 norm penalty $\mathcal{P}^{ridge} := ||\beta||_2^2$ for $\alpha = 0$ and the lasso L1 norm penalty $\mathcal{P}^{lasso} := ||\beta||_1$ for $\alpha = 1$. The L1 norm is defined as $||\beta||_1 = \sum_{j=1}^{p}|\beta_j|$ and the squared L2 norm as $||\beta||_2^2 = \sum_{j=1}^{p}\beta_j^2$. The ridge penalty, as introduced by Hoerl and Kennard (1970), can result in more stable results and handle multicollinearity. However, by the nature of the L2 norm, it is not able to select variables and hence results in a misspecified model if the true model is sparse. Consider an utopic case, in which we have k identical explanatory variables. The ridge penalty $\mathcal{P}^{ridge}$, would result in identical estimated coefficients with $\frac{1}{k}$ times the magnitude that the coefficients would get if estimated alone. This extreme example roots an attractive property of the L2 norm that we will investigate in detail soon.

Contrary to the ridge penalty, the lasso partially ignores highly correlated explanatory variables, since it tends to simply select one and shrink the others to zero. It relies on the belief that the true model consists of a (small) subset of coefficients with relatively strong signal, and many coefficients approximately to zero. Following the definition of Hastie et al. (2015, p.2), a true model is called sparse, if only a (small) subset of regressors helps in explaining the dependent variable, where the remaining regressors have a coefficient of zero. In this case, the lasso is an effective estimator, shrinking the irrelevant coefficients exactly to zero with high probability. However, for a sufficiently large tuning parameter λ, the lasso solution will be $\beta^{lasso} = 0$. The ridge penalty in turn, will shrink the coefficients relatively strong in direction of zero, for a large λ, but never exactly to zero. On the other hand, the lasso as well as the ridge estimator are exactly the OLS solution for $\lambda = 0$, since the respective penalization terms vanish.

The elastic net brings the advantages of both pioneer methods together since it simultaneously shrinks the coefficients as well as conducts variable selection. It follows from Lemma 1 (mathematical appendix) that the elastic net optimization problem can be equivalently expressed as the lasso optimization problem on an augmented data set, implying the ability of variable selection. In addition, the main drawbacks of the lasso are avoided. Theorem 2 of Zou and Hastie (2005) shows that the elastic net can even be interpreted as a stabilized version of the lasso. Furthermore, the imposed sparsity on the estimated coefficients grows monotonically from 0 to the sparsity of the lasso as α increases from 0 to 1. As the attractive properties of the elastic net are the motivation for applying this method on Vector Autoregressive models, we provide a concise elaboration in the subsequent sections.

[1] Note that the authors introduced the penalty as *naive elastic net*, which is a rescaled version of the *elastic net*. We follow Friedman et al. (2010) and do not distinguish between *naive elastic net* and *elastic net* in this work, since Zou and Hastie (2005) showed better empirical performance for the *elastic net*. In addition, the attentive reader might see that the original paper by Zou and Hastie (2005) defines the penalty by weighting the L1 norm with $(1-\alpha)$. However, the respective package *glmnet* uses α for weighting the L1 norm, similarly to the respective paper regarding the Coordinate Descent for the elastic net (Friedman et al., 2010). Again, we follow Friedman et al., 2010.

2.1.1 Grouping Effect

Within this section, we will have a closer look at the previously mentioned grouping effect. As a consequence of the grouping effect, the elastic net assigns highly correlated variables a close coefficient magnitude. Theorem 1 results in an upper bound for the difference of two coefficient paths and gives an intuitive explanation of the imposed grouping. In the mathematical appendix 8.3, we derive the corresponding upper bound for the elastic net.

Theorem 1. *Let X be standardized and $\hat{\beta}_i(\lambda_1, \lambda_2)\hat{\beta}_j(\lambda_1, \lambda_2) > 0$, such that both estimators have the same sign and are unequal to zero.*
Define $D_{\lambda_1,\lambda_2}(i,j) = \frac{1}{||Y||_1}|\hat{\beta}_i(\lambda_1, \lambda_2) - \hat{\beta}_j(\lambda_1, \lambda_2)|$, then the upper bound is $D_{\lambda_1,\lambda_2}(i,j) \leq \frac{1}{\lambda_2}\sqrt{2(1-p)}$.

Theorem 1 implies that the higher the pairwise correlation between feature i and j, the smaller the difference between estimators $\hat{\beta}_i$ and $\hat{\beta}_j$, since the upper bound $\frac{1}{\lambda_2}\sqrt{2(1-p)}$ is getting smaller. This characteristic is a desired feature of the elastic net and is not included in the lasso. Furthermore, we use an adjusted stylized example of Hastie et al. (2015) in order to illustrate the grouping effect property. We generate three groups with three variables each and $N = 100$ observations. The features within each group are highly correlated, such that theorem 1 states that the coefficient paths within a group will be close in magnitude. The data generating process (DGP) is

$$\begin{aligned} Z_1, Z_2, Z_3 &\sim N(0,1), \\ Y &= 3Z_1 - 1.5Z_2 - 6Z_3 + 2\epsilon, \text{with } \epsilon \sim N(0,1), \\ X_j &= Z_1 + \epsilon_j/5, \text{with } \epsilon_j \sim N(0,1) \text{ for } j = 1,2,3 \\ X_j &= Z_2 + \epsilon_j/5, \text{with } \epsilon_j \sim N(0,1) \text{ for } j = 4,5,6 \\ X_j &= Z_3 + \epsilon_j/5, \text{with } \epsilon_j \sim N(0,1) \text{ for } j = 7,8,9 \end{aligned}$$

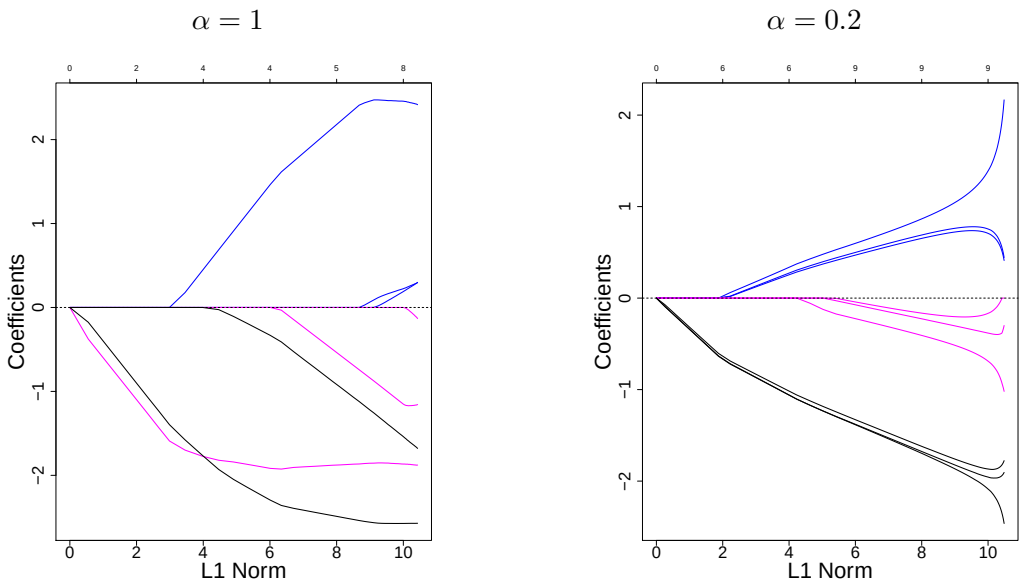
The corresponding coefficient paths for all nine variables are illustrated in figure 1 for the lasso and elastic net with $\alpha = 0.2$. As expected, the lasso does not show a regular pattern, however, the elastic net assigns similar coefficients to variables inside a common group. More precisely, coefficients $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ obtain similar magnitude throughout the path, as well as $\hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6$ and $\hat{\beta}_7, \hat{\beta}_8, \hat{\beta}_9$, respectively.

2.1.2 Pathwise Coordinate Descent

In this section, we explain how the actual solution $\hat{\beta}^{enet}(\lambda)$ is obtained and give reason for using the corresponding procedure, the pathwise Coordinate Descent (CD) algorithm. First, let us write the elastic net minimization problem from equation 3 in Lagrange form as

$$L(\beta_0, \beta)_{\alpha,\lambda} = \frac{1}{2}\text{RSS} + \mathcal{P}^{enet}$$

Figure 1: Lasso and Elastic Net Coefficient paths



Three groups with three variables each. Variables are highly correlated within a group. Lasso estimates on the left ($\alpha = 1$) and elastic net estimates on the right ($\alpha = 0.2$). Each color represents a group of highly correlated variables. The correlation within each group lies around 0.94 to 0.97. The L1 Norm is $||\hat{\beta}(\lambda)||_1$.

, where the loss is specified as $\frac{1}{2}RSS$ and $\mathcal{P}^{enet}$ is as defined previously. It can be shown, that $\hat{\beta}_0 = \bar{y} - \sum_{j=1}^{p} \hat{\beta}_j \bar{x}_j$. Since we assume that y_i is centered and x_j is standardized, $\hat{\beta}_0 = 0$ follows and we are able to waive a discussion around the intercept. The loss function RSS is differentiable and strictly convex. However, the penalty term $\mathcal{P}^{enet} = \lambda(\alpha||\beta||_1 + \frac{1}{2}(1-\alpha)||\beta||_2^2)$ is a combination of two norms, the $L1$ and $L2$ norm. The $L2$ norm, corresponding to the ridge, is differentiable. Hence, for $\alpha = 0$ there exists a closed form solution to the enet minimization problem, since it reduces to the ridge problem. However, the $L1$ norm, as the sum of absolute coefficients, is non-differentiable at zero, which leads to non-differentiability of the convex penalty (convex in β) $\mathcal{P}^{enet}$. Thus there is no closed form solution for the derivative at $\beta = 0$. Consequently, we are not able to use Gradient Type of algorithms, like Gradient Descent, since this approach requires to take the first p derivatives.

Fortunately, the pathwise Coordinate Descent algorithm can be applied to problems like the lasso, elastic net and the fenet, which we will introduce in the next chapter (Friedman et al., 2007). The CD algorithm can be applied to differentiable as well as non-differentiable functions. Hereby, Tseng (1988) establishes the following argument. Consider the functional form of

$$f(\beta_1, ..., \beta_p) = g(\beta_1, ..., \beta_p) + \sum_{j=1}^{p} h_j(\beta_j)$$

, where $g(\cdot)$ is a differentiable and convex function and each $h_j(\cdot)$ is convex. Then, the CD algorithm converges to the minimum of $f(\cdot)$, where the crucial argument is that the

penalty $\sum_{j=1}^{p} h_j(\beta_j)$ can be separated into individual functions of each β_j. Clearly, this argument holds for the penalties introduced earlier, since the RSS represents $g(\beta_1, ..., \beta_p)$ and $\mathcal{P}^{enet}$, $\mathcal{P}^{lasso}$ and $\mathcal{P}^{ridge}$ can be separated in the form of $\sum_{j=1}^{p} h_j(\beta_j)$, as we will do in chapter 3 for the VAR(p) model. However, the ridge regularization has a closed form solution and thus does not require the coordinate descent algorithm.

Going further, the CD algorithm relies on the so called soft thresholding operator, which is responsible for the shrinkage towards zero as well as variable exclusion. We adapt the derivation of the lasso soft thresholding operator to the elastic net procedure, which can be found in appendix 8.4. Note that, similar to the penalty, the soft tresholding operator of the enet problem, reduces to the operator of the lasso for $\alpha = 1$. Algorithm 1 compactly illustrates the steps of the CD algorithm for the elastic net.

Algorithm 1 Pathwise Coordinate Descent

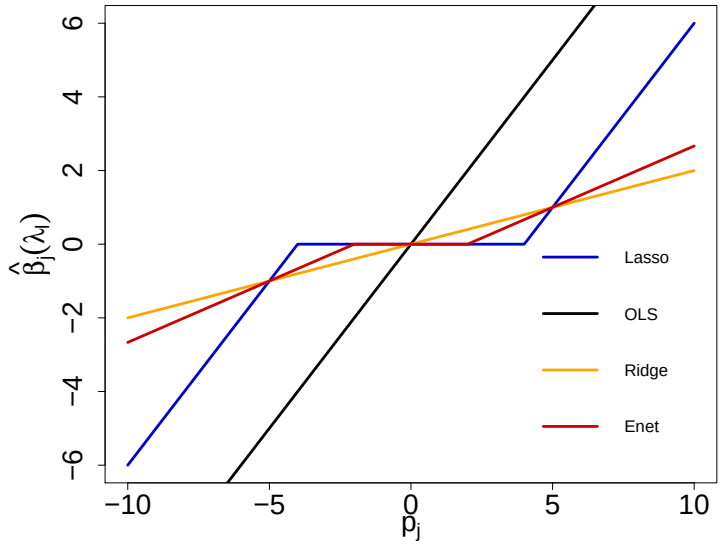
1. Select a value $\alpha \in [0, 1]$ and a sequence $\lambda_1 = \lambda_{max}, ..., \lambda_m = \lambda_{min}$
2. Initialize $\beta^0(\lambda_{max} = \lambda_1)$ for the first run. Otherwise set $\beta(\lambda_l) = \hat{\beta}(\lambda_{l-1})$ as a warm start.
3. For $j \in (1, ..., p)$, use the soft thresholding update form $\hat{\beta}_j(\lambda_l) = \frac{S(\sum_{i=1}^{N} x_{ij}(y_i - \tilde{y}_i^{(j)}),\ \lambda\alpha)}{1+\lambda(1-\alpha)}$
4. Stop if $\sum_{j=1}^{p} |\hat{\beta}_j(\lambda_{l-1}) - \hat{\beta}_j(\lambda_l)| < 10^{-12}$. Otherwise continue with step 1-4 until convergence or until $l = m$.

Since the elastic net Lagrange function $L(\beta)_{\alpha,\lambda}$ is convex in β, the global optimum can be found efficiently even for many observations and predictors. Setting $\beta(\lambda_l) = \hat{\beta}(\lambda_{l-1})$ as a warm start in step 2, enables to calculate the solution for the path $\{\lambda_1, ..., \lambda_m\}$ computationally efficient. As we fix α for each iteration, the CD algorithm of the enet differs to the lasso only in the soft thresholding update form $\hat{\beta}_j(\lambda_l)$, which we illustrate in figure 2 for the lasso, OLS, ridge and enet. The intuitive shrinkage pattern for a single coefficient β_j, c.p., can be observed for fixed $p_j := \sum_{i=1}^{N} x_{ij}(y_i - \tilde{y}_i^{(j)})$. Particularly, the OLS performs no shrinkage at all. The ridge does shrinkage in proportion to $|p_j|$ and does no variable selection. As a consequence of shrinkage towards zero, the ridge results in a smaller coefficient magnitude than OLS, for a fixed p_j. The lasso shrinks the coefficient to zero for sufficiently small $|p_j|$, where a coefficient is shrunk "earlier"[2] to zero as for the enet. The enet shrinks the coefficient towards zero, given $|p_j|$ is relatively large. If $|p_j|$ is relatively small, the respective coefficient is shrunk exactly to zero. Hence, it represents a mixture of the lasso and ridge.

Furthermore, the λ sequence $\lambda_1, ..., \lambda_m$, which is given ex-ante to the coordinate descent algorithm 1 is specified as follows. If $\hat{\beta} = 0$, it results directly from the Soft-Thresholding operator that $\lambda_{max} = \frac{1}{N\alpha} max_l |< x_l, y >|$, using scalar product notation. Friedmann et al. (2010) recommend to set $\lambda_{min} = 0.001\lambda_{max}$. Then a decreasing sequence from λ_{max} to λ_{min} is constructed on the log scale, typically using $K = 100$ steps.

[2] Meaning that the threshold for the lasso to shrink a coefficient to zero is already reached at a larger $|p_j|$ than for the enet.

Figure 2: Soft Thresholding Update Form



Update form used in step 3 of algorithm 1 for lasso ($\alpha = 1, \lambda = 4$), OLS ($\lambda = 0$), ridge ($\alpha = 0, \lambda = 4$) and enet ($\alpha = 0.5, \lambda = 4$).

2.2 Feature Weighted Elastic Net

In this section we move on to an extension of the elastic net, which imposes additional structure on the penalization term and is especially useful for the VAR model. Often, additional information on the variables of a data set is known ex ante. For example, consider a particular group of variables, where we ex ante know that these variables have a correlated effect on the dependent variable. Such information on the predictors of interest is referred to as a common feature that this particular group shares. Then, it might be of interest to incorporate this additional information into the elastic net regularization term introduced in the previous chapters. Fortunately, Tay et al. (2020) recently published a method called "feature-weighted elastic net", which incorporates the idea of "features of variables" by modifying the elastic net penalty P^{enet}. In the following, we will use the terms "assigning several variables into the same group" and "imposing the same common feature for a group of variables", interchangeably. As an exemplary application of the fenet model, Tay et al. (2020) give a genomics example, which is common in the statistical literature. In latter scientific field, the researcher often knows beforehand, which gene corresponds to which pathway. It is assumed that genes included in the same pathway, represent a group with a common feature. However, we believe that an analog argument can be made for time series data, particularly for Vector Autoregressive (VAR) models. Before we extend the fenet model to VARs, a concise introduction of the basic model is required.

Assume we are able to quantify the additional ex-ante information on the feature of variables. For the purpose of grouping the variables together, we introduce the auxiliary matrix $\mathbf{Z} \in \mathbb{R}^{p \times G}$, where p is the number of variables and G is the number of sources of variable information[3]. Next, we define $\mathbf{z}_j \in \mathbb{R}^{G \times 1}$ as the jth row of $\mathbf{Z}$. In order to

[3]Note that the paper by Tay et al. (2020) uses K as the number of different sources. In this work, however, it is notation-wise more consistent to use G.

use the information given to $\mathbf{Z}$, a score $z_j'\theta$ is assigned to each variable $j = 1, ..., p$, where $\theta = [\theta_1, ..., \theta_G]' \in \mathbb{R}^{G\times 1}$ is a hyperparameter (-vector) selected by the algorithm. The fenet algorithm for obtaining the solution $\hat{\beta}^{fenet}$ can be seen as an extension of the pathwise Coordinate Descent algorithm 1. Hereby, in principle, if the researcher finds herself in a situation with only few feature sources (G small), she can run algorithm 1 over a small grid of θ and choose the value with the smallest loss. However, based on the experience of the authors, this approach is computationally not feasible. Thus, the authors propose a modification of algorithm 1, which minimizes the Lagrange function 5 over both, β and θ. The researcher is able to estimate θ, instead of using an additional cross validation step.

For example, assume that our model consists of $p = 4$ variables and $G = 2$ different sources of variable information, where variable x_1, x_2 and x_3, x_4 share a common feature, each. Then the auxiliary matrix Z and the score $z_j'\theta$ for $j = 1$ is

$$Z = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} z_1' \\ z_2' \\ z_3' \\ z_4' \end{bmatrix} \qquad \text{and} \qquad z_1'\theta = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \theta_1$$

As we see, x_1 and x_2 share the same feature and hence $z_1 = z_2$ selects the hyperparameter corresponding to the respective feature, which is θ_1 in this example.

Next, the authors propose to specify the functional form of the score as

$$f(z_j'\theta) = \frac{\sum_{l=1}^{p} exp(z_l'\theta)}{p\exp(z_j'\theta)}. \tag{4}$$

More precisely, $f(z_j'\theta) = w_j(\theta)$ is used as a weighting factor for every coefficient j in the fenet minimization problem, for which the Lagrange function is defined as

$$L(\beta_0, \beta, \theta)_{\alpha,\lambda} = \frac{1}{2}\sum_{i=1}^{N}(y_i - \sum_{j=0}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} w_j(\theta)\left[\alpha|\beta_j| + \frac{1}{2}(1-\alpha)\beta_j^2\right]. \tag{5}$$

The functional specification of $f(z_j'\theta)$ can in principle, be specified differently. In addition, the authors do not provide any optimality condition or well-developed theory on it. However, the present choice is apparently accompanied by several convenient properties. First, given the exponential specification, the fenet objective function 5 nests the enet objective function as introduced in equation 2.1, since $w_j(\theta) = 1$ for all variables j if $\theta = 0$. Next, it gives a lower bound on the weighting, namely $w_j(\theta) \geq \frac{1}{p}$, since the current specification can be written as

$$w_j(\theta) = \frac{\sum_{l=1}^{p} exp(z_l'\theta)}{p\exp(z_j'\theta)} = \frac{\sum_{l=1,l\neq j}^{p} exp(z_l'\theta)}{p\exp(z_j'\theta)} + \frac{\exp(z_j'\theta)}{p\exp(z_j'\theta)}$$

, which reduces to $w_j(\theta) = \frac{1}{p}$ for $z_j'\theta = 0$ and $\exp(z_l'\theta) \to 0$ as $z_l'\theta \to -\infty$ for $l =$

$1, ..., p$, $l \neq j$. This way the authors ensure to avoid a negligible penalty for coefficient j. Furthermore, a relation to the Group Lasso (Yuan and Lin, 2007) is possible, by specifying functional form 4. Theorem 1 of Tay et al. (2020) proves that under certain assumptions the group lasso minimization problem can be written in terms of the fenet problem. The group lasso relies on a similar idea as the fenet. In the group lasso, however, the algorithm either selects or excludes a whole non-overlapping pre specified group of variables from the minimization. On the other hand, the fenet can be seen as a more flexible method. The groups do not need to be disjoint, i.e. we are able to assign a variable into more than one group. In addition, selection is not conducted on the group level. The groups rather obtain the same weight on the group level and selection is done on an individual level.

Finally, the weighting function is accompanied by a convenient intuition. The hyperparameter θ_g contains the importance of a source of information. The more important a common feature is relatively to the remaining sources, the larger is the respective θ_g. A large θ_g, gives a relatively small weight to each coefficient in group g. Hence, the respective coefficients are assigned a smaller regularization penalty. Furthermore, a smaller regularization penalty implies less shrinkage on the respective group of coefficients. Thus it is also less likely that these coefficients will be shrunk exactly to zero.

2.3 Cross Validation

Since the enet and fenet require the tuning of two hyperparameters, namely λ and α, we follow Zou and Hastie (2005) and expand the idea of a single parameter cross validation to that of a double parameter cross validation. First, let us briefly introduce the notion of a single parameter cross validation procedure, which is a popular way of obtaining the tuning parameter λ, e.g. for lasso or ridge. It is common practice to randomly divide the total data set into a training set, validation set and a test set. We are interested in estimating the expected test set error $E[L(Y, \hat{f}_\lambda(X))]$ by applying $\hat{f}_\lambda$ to an independent test set, where $\hat{f}_\lambda$ denotes the estimated fitted model using the training data set τ. The loss function $L(Y, \hat{f}_\lambda(X))$ is often chosen to be either the absolute error $|Y - \hat{f}_\lambda(X)|$ or the squared error $(Y - \hat{f}_\lambda(X))^2$. However, by the law of total expectations the expected test set error consists of

$$E[L(Y, \hat{f}_\lambda(X))] = E[E[L(Y, \hat{f}_\lambda(X))|\tau]]$$

, where $E[L(Y, \hat{f}_\lambda(X))|\tau]$ is the expected test set error conditional on training set τ. The validation set is used to estimate latter conditional expectation. Next, an effective and popular way of re-using the sample in order to obtain K estimates of $E[L(Y, \hat{f}_\lambda(X))|\tau]$ for each ex-ante specified tuning parameter λ is described. First, divide the data set into K equally sized parts, where $K = 5$ or $K = 10$ is often chosen. The kth part is used to calculate the prediction error of the model that is fit on the remaining $K - 1$ parts. This process is repeated for $k = 1, 2, ..., K$. By removing the kth part of the data, we finally obtain K fitted functions $\hat{f}_\lambda^{-k}(x)$, where each prediction is dependent on the tuning

parameter λ. Furthermore, $L(y_i, \hat{f}_\lambda^{-k}(x))$ is an estimate for the conditional test set error. After defining $\hat{f}_\lambda^{-k(i)}(x_i)$ as the prediction of the dependent variable y_i, evaluated at the observation vector x_i on validation set k, we are able to calculate the loss function for each of the k predictions and average them to obtain

$$CV(\hat{f}_\lambda) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}_\lambda^{-k(i)}(x_i)). \tag{6}$$

Expression 6 is an estimate of the expected test set error, since the K-fold procedure results in K sufficiently different training sets τ. However, $CV(\hat{f}_\lambda)$ is still dependent on λ. Subsequently in this work, we select[4] λ as $\lambda_{min} = \text{argmin}_\lambda CV(\hat{f}_\lambda)$.

Moving on from the one tuning parameter methods, the enet and the fenet are accompanied by additional technical as well as computational burdens, since cross validation is conducted on the two dimensional surface of

$$\mathbf{\Lambda} \times \mathbf{A} = \{(\lambda, \alpha) | \lambda \in \mathbf{\Lambda} = \{\lambda_1, ..., \lambda_m\} \text{ and } \alpha \in \mathbf{A} = \{\alpha_1, ..., \alpha_L\}\}.$$

Typically, the λ-grid $\mathbf{\Lambda}$ is specified as introduced in section 2.1.2 and has the cardinality of 100. However, $\mathbf{A}$ represents a rather sparse grid for α, e.g. $\mathbf{A} = \{0, 0.05, ..., 0.95, 1\}$. The validation procedure for both tuning parameters is explained in the following scheme.

1. First, we separate the data into an outer training set and an outer test set. Next, we separate the outer training set into an inner training set and inner test set. The percentage breakdown of the total data set into an outer training set and inner training set is up to the researcher. In this work, we choose 70 % for outer as well as the inner partition. The inner test set is used to validate the two tuning parameters α and λ. The outer test set is meant for model assessment of the procedure incorporating the optimal hyperparameters α^*, λ^*. In a later chapter, we will move on to the time series context and use latter set for a pseudo out of sample prediction exercise.

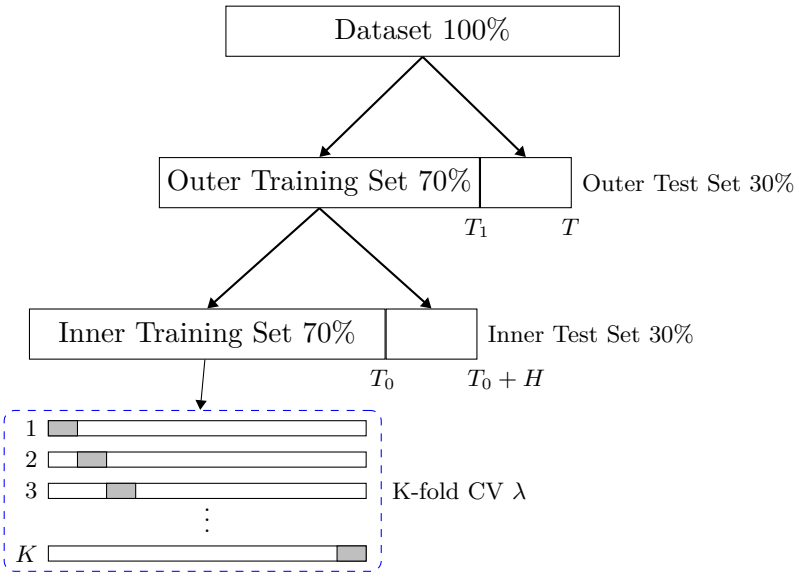
2. Apply K-fold CV of λ on the inner training set for a fixed α_l, as introduced in equation 6. Then predict the set of dependent variables $\{\hat{y}_{T^0+1}, ..., \hat{y}_{T^0+H}\}$ on the inner test set using the resulting λ and fixed α_l. Calculate the resulting $MSFE_{test}(\alpha_l)$. A rough outline of step 1 and 2 is illustrated in figure 3.

3. Repeat step 2. for the whole grid $[\alpha_1, \ldots, \alpha_L]$. Choose the the final tuning parameter combination $\{\alpha^*, \lambda^*\}$ with the smallest inner test set $MSFE_{test}(\alpha_l)$. Use $\{\alpha^*, \lambda^*\}$ in order to assess the model by predicting on the outer test set.

4. Obtain the final model coefficients $\hat{\beta}$ by estimating them using $\{\alpha^*, \lambda^*\}$ on the full data set.

[4] For further details, we refer to Hastie et al. (2009, p. 244).

Figure 3: Data partitioning scheme



2.3.1 Adjustment for Time Dependent Data

The cross validation procedure following chapter 2.3 is based on the assumption that the inner training and inner test set are independent. However, time series variables such as GDP or Inflation have an underlying time dependence. Therefore Nicholsen et al. (2014) propose an extended form of one hyperparameter CV for the lasso VAR estimation. Instead of validating the λ path via traditional K-fold CV, they propose to use a rolling window as an inner training set and sequentially predict the one step ahead forecast on the inner test set for a fixed λ. This procedure is repeated for the whole path of λ and the λ resulting in the smallest inner test set error is chosen to be optimal. By applying the idea of a rolling scheme, the dependence of the inner training and the inner test set is included.

Inspired by Nicholsen et al. (2014), we propose an adjustment of the previously introduced double parameter CV procedure of section 2.3 to the time series context. To the best of our knowledge, previous literature does not include such an adjustment for the enet or fenet in the context of time series. However, we do not validate λ in a rolling window manner[5], but α. By adjusting the procedure in this manner, we seek to improve forecast precision by being conform with the dependencies in the data. For this purpose, we adjust step 2-3 from the previous chapter in the following manner:

2. Apply K-fold CV for λ on the inner training set, given a fixed α_l. The result will be λ_1. Use λ_1 and α_l to obtain the predicted dependent variable $\hat{y}_{T^0+1}$ of the next time period, *instead of the whole inner test set*. Repeat this for the whole test set using a rolling window procedure. The result is $\lambda_1, ..., \lambda_H$ and the predicted dependent observations $\{\hat{y}_{T^0+1}, .., \hat{y}_{T^0+H}\}$ for a given α_l. The predicted observations are used to calculate the $MSFE_{test}(\alpha_l)$. The rolling window procedure is illustrated exemplarily in figure 4.

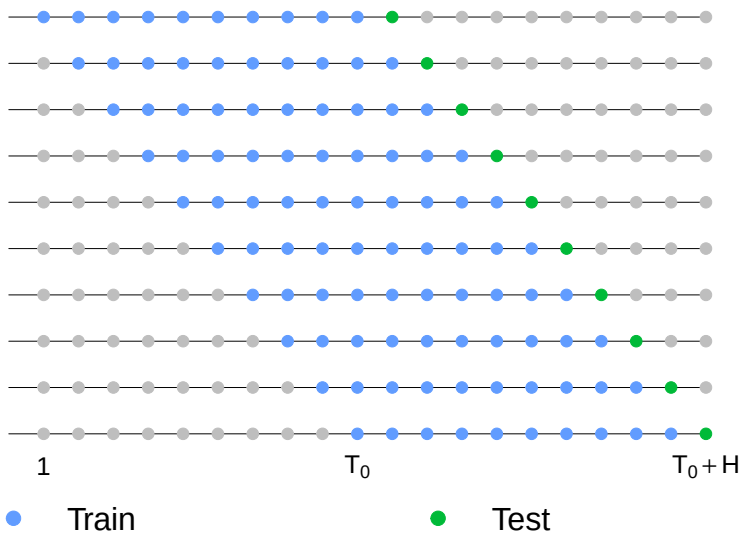
[5] In principle, we could validate both, λ and α using a rolling window scheme. See Bergmeir et al. (2018) on a general discussion of cross validation for AR processes.

3. Repeat step 2. for the whole grid $\{\alpha_1, \ldots, \alpha_L\}$. Choose α^* as the argument that minimizes the inner test set MSFE as

$$\alpha^* = \underset{\alpha_l}{\operatorname{argmin}}\, MSFE_{test}(\alpha_l) \quad \text{for} \quad l = 1, ..., L.$$

Furthermore, instead of having a single λ^* as in chapter 2.3, we obtain a set of optimal $\{\lambda_1^*, ..., \lambda_H^*\}$ for α^*. Based on our experience, $\{\lambda_1^*, ..., \lambda_H^*\}$ does not fluctuate largely around the mean $\bar{\lambda} = \frac{1}{H}\sum_{i=1}^{H} \lambda_i^*$. Consequently, we decide to set $\lambda^* = \bar{\lambda}$ as a heuristic approach. Then, as usual, we use α^*, λ^* to assess the model on the outer test set, as we will see in the forecasting exercise later on.

Figure 4: Rolling Window Scheme



2.4 Bias-Variance Trade Off

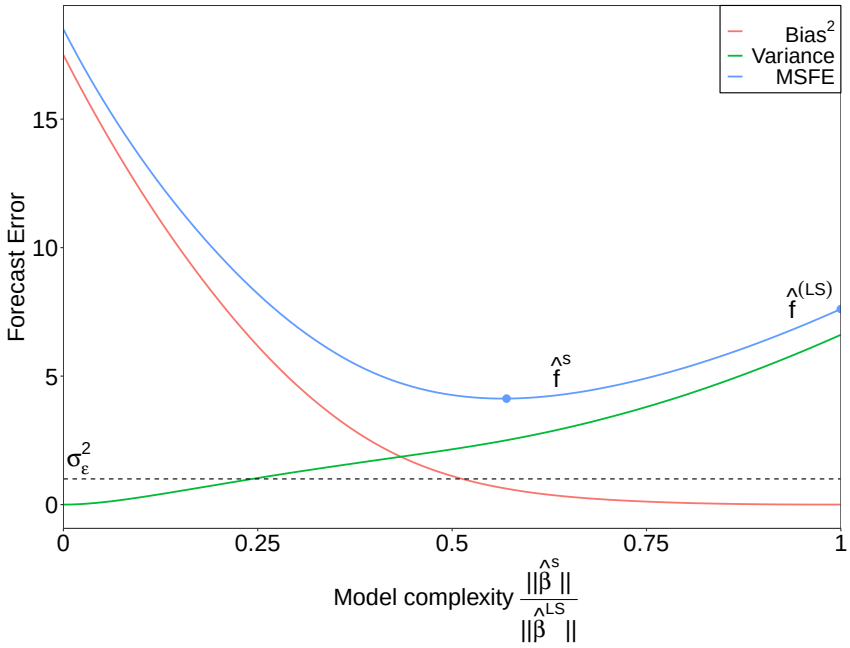
Now that we have a strategy for obtaining the tuning parameter λ^* for the lasso and the ridge, as well as $\{\lambda^*, \alpha^*\}$ for the enet and fenet, we give an intuition for variable shrinkage. Let us generalize $\hat{f}_\lambda(X)$ to $\hat{f}(X)$ in order to represent any statistical model, which can be dependent on λ, $\{\lambda, \alpha\}$, $\{\lambda, \alpha, \theta\}$ or no tuning parameter at all. The previously introduced expected prediction error $E[L(Y, \hat{f}(X))]$ evaluated at the new data point x_0, can be decomposed into three components [6]

$$\begin{aligned} E[L(Y, \hat{f}(X))|X = x_0] &= E[(f(X) - \hat{f}(X))^2|X = x_0] \\ &= \sigma_\epsilon^2 + \text{Bias}[\hat{f}(X)|X = x_0]^2 + \text{Variance}[\hat{f}(X)|X = x_0] \end{aligned} \tag{7}$$

[6] The derivation of the MSFE decomposition can be found in Hastie et al., (2015).

, where the squared bias represents the squared difference between $\hat{f}(x_0)$ and $f(x_0)$ and the variance measures by how much the prediction changes if we apply $\hat{f}(X)$ on a new data point x_0. The irreducible error σ_ϵ^2 can not be avoided. The evaluation point x_0 is assumed to be drawn independently from the same distribution as the data set. Obviously, we are keen to minimize expression 7, by simultaneously achieving a suitable trade-off between bias and variance. This trade-off is graphically illustrated in figure 5. Hereby, model complexity is defined as the relative ratio between the L1 norm of the shrinked solution $||\hat{\beta}^s||_1$ as well as the L1 norm of the Least Squares solution $||\hat{\beta}^{LS}||_1$. The shrinked solution $\hat{\beta}^s$ could, in principle, be obtained by including any of the previously introduced penalties into the minimization problem. This additional penalty term enables to shrink the coefficients towards zero or exactly to zero. Moreover, independent of the induced regularization, a shrinkage parameter of $\lambda = 0$ will lead to $\hat{\beta}^s = \hat{\beta}^{LS}$, which equals a model complexity of one. Thus the ridge, lasso, enet as well as the fenet will reduce to the OLS minimization problem for $\lambda = 0$, meaning that no shrinkage at all is imposed on the coefficients. The LS estimator is known to be the best linear unbiased estimator (BLUE) in the class of linear models. It is called the "best" in the sense that $\hat{\beta}^{LS}$ is an efficient estimator. Since $\hat{\beta}^{LS}$ is unbiased, it follows directly that $\text{Bias}(\hat{f}^{LS}(X)|X = x_0) = 0$. However, the price that we have to pay for the unbiased property is a high variance. On the other hand, shrinkage methods are able to reduce the resulting variance from $\hat{\beta}^s$ by including a penalty term in the minimization problem. This in turn, is accompanied by a higher bias. As long as the variance decreases at a higher rate than the bias increases, shrinkage methods are able to reduce the MSFE. As we see in figure 5, model complexity sinks as we induce shrinkage for $\hat{\beta}^s$, resulting in a smaller MSFE. For example, given that α is fixed and λ increases, model complexity decreases and shifts $\hat{f}^s$ graphically to the left. Thus, by shrinking the coefficients, regularized methods are able to circumvent overfitting and thus lead to more precise predictions. The LS solution $\hat{\beta}^{LS} = (X'X)^{-1}X'Y$ works well in a case where $N >> p$, however, it is problematic if the number of predictors is close to the number of observations. In a high-dimensional case, where $N < p$, the columns of X are not linearly independent and thus $(X'X)$ does not have full rank. Hence the inverse of $(X'X)$ does not exist and $\hat{\beta}^{LS}$ is not well defined. Still, methods including regularization are able to obtain an estimate $\hat{\beta}^s$ in a high-dimensional case.

Figure 5: Bias-Variance Trade Off



3 Vector Autoregressive Model

In this chapter, we move on to a model developed for time dependent data. Afterwards, we will be able to apply the previous regularization methods to the time series context. Let us introduce the VAR(p) model following Luetkepohl (2005). First, a change of notation is necessary, which is more typical for the VAR framework. Let K be the number of variables, p be the number of lagged variables and T be the number of data observations or time periods. The VAR(p) model is defined as

$$Y_t = v + A_1 Y_{t-1} + ... + A_p Y_{t-p} + u_t, \quad u_t \overset{iid}{\sim} N(0, \Sigma) \tag{8}$$

, where $Y_t = (y_{1t}, ..., y_{Kt})'$ is a $(K \times 1)$ vector of endogeneous variables, $u_t = (u_{1t}, ..., u_{Kt})'$ is a $(K \times 1)$ white noise process and A_i for $i = 1, ..., p$ are unknown $(K \times K)$ coefficient matrices. The vector of intercept terms is defined as $v = (v_1, ..., v_K)'$. The unknown covariance matrix of the errors is denoted as $\Sigma = E[u_t u_t']$. A coefficient in the matrix A_i is interpreted as the marginal effect of the jth variable of the ith lag on the current value of the kth dependent variable. Furthermore, we can rewrite equation 8 as

$$Y_t = \mathbf{A} Z_{t-1} + u_t \tag{9}$$

, where $\mathbf{A} = [v, A_1, ..., A_p]$ and $Z_{t-1} = [1, Y_{t-1}', ..., Y_{t-p}']'$.

To ease notation for the regularization methods, we rewrite the VAR(p) process from

equation 8 as a system of K equations

$$
\begin{aligned}
y_{1t} &= v_1 + (\beta_{11}^1 y_{1,t-1} + ... + \beta_{1K}^1 y_{K,t-1}) + ... + (\beta_{11}^p y_{1,t-p} + ... + \beta_{1K}^p y_{K,t-p}) + u_{1t} \\
&\vdots \\
y_{Kt} &= v_K + (\beta_{K1}^1 y_{1,t-1} + ... + \beta_{KK}^1 y_{K,t-1}) + ... + (\beta_{K1}^p y_{1,t-p} + ... + \beta_{KK}^p y_{K,t-p}) + u_{Kt}
\end{aligned}
$$

, which is more compactly written in summation notation as

$$
\begin{aligned}
y_{1t} &= v_1 + \sum_{i=1}^{p} \sum_{j=1}^{K} \beta_{1j}^i y_{j,t-i} + u_{1t} \\
&\vdots \\
y_{Kt} &= v_K + \sum_{i=1}^{p} \sum_{j=1}^{K} \beta_{Kj}^i y_{j,t-i} + u_{Kt}.
\end{aligned}
$$

Then, w.l.o.g. we generalize the kth equation as

$$
y_{kt} = v_k + \sum_{i=1}^{p} \sum_{j=1}^{K} \beta_{kj}^i y_{j,t-i} + u_{kt} \tag{10}
$$

Expressing the kth equation in this way, enables a convenient notational use for adapting shrinkage methods to the VAR(p) model, which we will see soon.

3.1 Curse of Dimensionality and related problems

In many macroeconomic models the researcher prefers a rather small and interpretable model, enabling to estimate the coefficients of process 9 via the OLS formula

$$
\begin{aligned}
\hat{\mathbf{A}} &= (\sum_{t=1}^{T} Y_t Z_{t-1}^{'})(\sum_{t=1}^{T} Z_{t-1} Z_{t-1}^{'})^{-1} \\
&= Y Z^{'} (Z Z^{'})^{-1}
\end{aligned}
$$

, where $Y = [Y_1, ..., Y_T]$ is a $K \times T$ matrix and $Z = [Z_0, ..., Z_{T-1}]$ is a $(Kp+1) \times T$ matrix. This estimator works fine in the case of $T >> (Kp+1)$, however, otherwise a degrees-of-freedom problem might arise quickly. In the extreme case, where $(Kp+1) > T$ the matrix Z does not have full rank and thus the inverse of $(ZZ^{'})$ does not exist. Given that $(ZZ^{'})$ is singular, we are not able to obtain an unique solution for $\hat{\mathbf{A}}$. Borderline cases in which we still have full rank, but the number of coefficients $(Kp+1)$ is close to the number of time periods T can result in computational inaccuracies in terms of inverting $(ZZ^{'})$, leading to an inaccurately estimated coefficient matrix $\hat{\mathbf{A}}$. Furthermore, as $(Kp+1)$ is large in relation to T, dramatic consequences for prediction accuracy might emerge. This well known problem is also known as overfitting or the curse of dimensionality and is often found in high dimensional data settings. Following on the discussion from chapter 2.4, the variance of an out of sample forecast based on $\hat{\beta}_{kj}^i$ for $i = 1, ..., p$ and $j = 1, ..., K$ increases

as the number of coefficients pK gets relatively large. The reason is that by including too many variables in the model, the estimated model fails to generalize from the training data to new samples, which results in a poor prediction performance, as we will encounter in chapter 4 for the unregularized least squares case. More precisely, the residuals u_{kt} get noticeably small as a consequence, resulting in a tremendously small residual variance for variable k. By imposing shrinkage on the coefficients, the researcher is able to filter the noise in the data and to enable the model to extract the relevant information from the training data. Thus, a VAR(p) model which either includes too many lags or too many variables, rather fits the unsystematic part of the data, instead of the signal. As already explained in 2.4, we include an additional penalty in the OLS problem. Thus, the bias of the prediction increases, however, the variance of the prediction reduces, leading to more accurate out of sample forecasts in terms of a smaller MSFE.

Another closely related problem arising in high dimensional models is multicollinearity among the predictors. This problem especially arises in large VAR systems, containing macroeconomic variables and their respective lags. Strong multicollinearity means that one or more predictors can be explained by a nearly linear combination of the remaining explanatory variables. Although the researcher can not exactly be sure, by which degree the set of explanatory variables is subject to multicollinearity, the consequences can be quite perceptible. Technically speaking, for OLS, the consequences of the presence of multicollinearity arise, as a result of a decreasing determinant of ZZ', which inflates the inverse $(ZZ')^{-1}$. This in turn inflates the coefficient variances as well as covariances, since $(ZZ')^{-1}$ is a factor of the variance matrix of $\hat{\mathbf{A}}$. Consequently, estimated coefficients with unreasonable magnitudes or even the wrong sign might arise. The power of the model is reduced and coefficients are displayed as less significant (smaller t-values). Moreover, the estimated coefficients change noticeably in dependence of inclusion/exclusion of variables and observations into the model. Although this problem does not have tremendous consequences for an out of sample prediction, it can lead to severe distortions for in-sample analyses. Later in this work, we will notice in the simulation study regarding structural impulse responses, that emerging multicollinearity leads to unpleasant consequences for OLS. In order to assess the severity of the present muticollinearity in a set of explanatory variables, we could pay attention to the previously mentioned symptoms. However, this approach is rather informal. Consequently, we use the popular Variance Inflation Factor (VIF) as a measure of collinearity (see Wooldridge (2002), p. 98)

$$VIF_j = \frac{1}{1 - R_j^2} \quad \text{for} \quad j = 1, 2, ..., p$$

, where R_j^2 is the coefficient of determination of regressing the predictor x_j on the remaining $p-1$ predictors[7]. Given that a large part of the variation of x_j can be explained by a linear combination of the remaining predictors, the resulting partial coefficient of

[7]For simplicity, we use linear regression notation here.

determination, R_j^2 is close to 1. Particularly, a rule of thumb is that $0.9 < R_j^2$ represents severe multicollinearity for x_j. The VIF is interpreted as the factor by which the variance of x_j inflates, compared to a situation, where $R_j^2 = 0$ (no multicollinearity at all).

In the next chapters, we will adapt the previously introduced shrinkage methods to the VAR(p) model and, hopefully, perform precise predictions as well as a structural analysis in a high dimensional data setting accompanied by the burdens of this section.

3.2 Regularized VAR

In this section, we adapt $\mathcal{P}^{ridge}$, $\mathcal{P}^{lasso}$, $\mathcal{P}^{enet}$ and $\mathcal{P}^{fenet}$ to the VAR(p) model. Fortunately, the previously introduced characteristics of these penalization terms do translate to the respective regularized VAR(p) model. Thus, we avoid a detailed discussion of these if no particularities for the VAR context are present.

In principle, it is possible to estimate all pK^2 coefficients of the system at once using a pooled estimator. However, assuming that we do not have cross equational restrictions on the parameters, the VAR system reduces to a seemingly unrelated regression (SUR) model, since each equation has the same set of regressors. This enables us to focus our attention on each equation individually, instead of the whole system. Thus, we ignore the cross-equational dynamics of the VAR and validate the tuning parameter(s) separately for each equation k. This approach can be seen as more precise, since we allow the amount of shrinkage as well as the weighting between ridge and lasso to differ for each equation k. Otherwise, we would obtain a pooled average across K equations in form of a single pair of $\{\alpha, \lambda\}$, which might not be accurate. For example, figure 11 indicates that the tuning parameter α does differ substantially across the equations. Although we will transform all variables to stationarity, each equation k contains a different level of sparsity as well as persistence. However, we are aware of the potential efficiency loss by conducting an equation-wise estimation and are willing to increase the computational burden for a more accurate analysis. Representing the VAR(p) as a system of equations (as in equation 10) enables to characterize the solution to the generalized minimization problem for ridge, lasso, enet and fenet as

$$\begin{aligned}\{\hat{v}_k,\ \hat{\beta}_k\} &= \underset{\{v_k,\ \beta_k\}}{\operatorname{argmin}}\ \frac{1}{2}\sum_{t=1}^{T}(y_{kt} - v_k - \sum_{i=1}^{p}\sum_{j=1}^{K}\beta_{kj}^{i}y_{jt-i})^2 + \mathcal{P}_k \\ &= \underset{\{v_k,\ \beta_k\}}{\operatorname{argmin}}\ \frac{1}{2}\text{RSS} + \mathcal{P}_k\end{aligned}$$

, where $\beta_k = [\beta_{k1}^1, ..., \beta_{kK}^1, ..., \beta_{k1}^p, ..., \beta_{kK}^p]' \in \mathbb{R}^{Kp\times 1}$ represents Kp coefficients from equation k. To give the reader a convenient overview of the adjusted penalties used in this work, we summarize the equation-wise penalties in a table as

Table 1: Penalty $\mathcal{P}_k$

$\mathcal{P}_k^{ridge}(\lambda)$	$\lambda\|\|\beta_k\|\|_2^2 = \lambda\left[\sum_{i=1}^{P}\sum_{j=1}^{K}(\beta_{kj}^i)^2\right]$
$\mathcal{P}_k^{lasso}(\lambda)$	$\lambda\|\|\beta_k\|\|_1 = \lambda\left[\sum_{i=1}^{P}\sum_{j=1}^{K}\|\beta_{kj}^i\|\right]$
$\mathcal{P}_k^{enet}(\lambda,\alpha)$	$\lambda\left[\alpha\|\|\beta_k\|\|_1 + \frac{1}{2}(1-\alpha)\|\|\beta_k\|\|_2^2\right] = \lambda\left[\sum_{i=1}^{P}\sum_{j=1}^{K}\left[\alpha\|\beta_{kj}^i\| + \frac{1}{2}(1-\alpha)(\beta_{kj}^i)^2\right]\right]$
$\mathcal{P}_k^{fenet}(\lambda,\alpha,\theta)$	$\lambda\left[\sum_{i=1}^{P}\sum_{j=1}^{K} w_{ij}(\theta^k)\left[\alpha\|\beta_{kj}^i\| + \frac{1}{2}(1-\alpha)(\beta_{kj}^i)^2\right]\right]$

Adjusted penalties for ridge, lasso, enet and fenet for the VAR(p) model, which we will apply to each equation k. These penalties are used for the forecasting exercise as well as to estimate the impulse response functions later on.

3.2.1 Lasso VAR and Ridge VAR

The Lagrange function, of the lasso VAR(p) model is

$$L(v_k,\beta_k)_\lambda = \frac{1}{2}RSS + \lambda\left[\sum_{i=1}^{P}\sum_{j=1}^{K}|\beta_{kj}^i|\right]$$

and for the ridge VAR(p) it is

$$L(v_k,\beta_k)_\lambda = \frac{1}{2}RSS + \lambda\left[\sum_{i=1}^{P}\sum_{j=1}^{K}(\beta_{kj}^i)^2\right].$$

In order to find the global minimum of the lasso VAR(p) objective function, the Coordinate Descent algorithm is used, as explained in algorithm 1. For the ridge, we are able to obtain a closed form solution. However, we will not go into further detail on the ridge VAR(p) in this section, but rather on the lasso VAR(p), due to the variable selection characteristic.

As mentioned in Sims (1980), the VAR(p) model is overparameterized, such that coefficient shrinkage in A_i is even more important the larger K gets. Hereby, we will encounter an increase in the number of estimated zero coefficients for an increase in the model size in our empirical application. Since the VAR(p) system contains Kp coefficients and T observation points per equation, reducing the dimensionality of the system is highly desired. If variable j for a certain lag i contains sufficiently low explanatory power for the dependent variable k, lasso shrinks β_{kj}^i exactly to zero. Even if the explanatory power is not sufficiently low to set β_{kj}^i exactly to zero, the penalization term induces shrinkage towards zero. Intuitively, since the goal is to minimize the Lagrange function, the gain of a decrease in RSS does not outweigh the costs induced by the penalty term. Thus resulting in shrinkage of the respective coefficient, which leads to a lower model complexity and, hopefully, reduces the forecasting error of poorly estimated coefficients. After applying the lasso on K separate equations, we obtain $\hat{\beta}_k$ for $k = 1, ..., K$ and are able to construct the coefficient matrices $\hat{A}_i$ for $i = 1, ..., p$ of the VAR(p).

Furthermore, by construction $\beta_k = [\beta_{k1}^1, ..., \beta_{kK}^1, ..., \beta_{k1}^p, ..., \beta_{kK}^p]'$ includes coefficients of

lagged independent variables, such that the L_1 penalty will either select or shrink some of these variables to zero. For example, if and only if the lasso shrinks all coefficient of $[\beta_{k1}^1, ..., \beta_{k1}^p]$ to zero, then the first variable is said to not Granger-cause the kth dependent variable. Specifically, the L1 penalty naturally includes a Granger-Non Causality interpretation in the estimation procedure. Note that the variable selection discussion is not relevant in the OLS as well as Bayesian framework. In general, Bayesian methods mainly focus on shrinking the coefficients towards zero, instead of selecting them.

3.2.2 Elastic Net VAR

In analogy to the lasso and ridge, we adapt the bare elastic net model from chapter 2.1 to the VAR context and introduce the enet VAR(p) Lagrange function as

$$L(v_k, \beta_k)_{\lambda,\alpha} = \frac{1}{2} RSS + \lambda \left[\sum_{i=1}^{P} \sum_{j=1}^{K} \left[\alpha |\beta_{kj}^i| + \frac{1}{2}(1-\alpha)(\beta_{kj}^i)^2 \right] \right]$$

, where we validate the set of hyperparameters $\{\lambda, \alpha\}$ using our double parameter CV procedure adjusted for time dependent data as elaborated in chapter 2.3.1. This model is particularly useful for the VAR context for the following reasons. Often we find that large VAR models are accompanied by a high degree of pairwise correlation among the predictors. Hereby, we derived the formal argument of the grouping effect in chapter 2.1.1, which enables a convenient interpretation. Particularly, it shows that the elastic net has an upper bound on the absolute difference in magnitude between two estimated coefficients, which is smaller, the larger the pairwise correlation between them is. Hence, in the case of highly correlated variables, the elastic net will select both highly correlated variables and also assign them a similar magnitude. This grouping effect is induced by the ridge part in $\mathcal{P}_k^{enet}(\lambda, \alpha)$ and is desirable for a high-dimensional VAR model. As a part of this discussion, note that a high degree of pairwise correlation can be seen as an indicator for multicollinearity among the predictors of the model. Often, a large VAR system includes several variables engaging in similar behavior, i.e. they move closely together. For example, we might include two measures of Inflation, CPI and GDP Deflator, in our system. Since both measures contain similar signal, in case of selection, the enet will select both variables and assign similar magnitude. The bare lasso would fail in this case and either select CPI or GDP Deflator.

Furthermore, we might increase the number of macroeconomic variables as well as lag length, with the hope to increase the information represented in the VAR(p) model. Unfortunately, not every additional coefficient provides enough predictive power to be considered useful. The selection ability of the enet, however, shrinks the sufficiently irrelevant coefficients to zero. In the empirical forecasting exercise we will encounter, that these properties are accompanied by an accurate prediction.

Another desired property of the enet emerges, when we are interested in a relatively

large model, but do not have a data rich situation. Imagine having T time periods of observations and $pK > T$ coefficients to estimate for equation k. Of course, OLS will not be possible to use anymore, since the solution is not well defined. Furthermore, the lasso has an upper bound on the number of coefficients which it can select, since it is restrained to select at most T coefficients. However, the elastic net is able to circumvent this constraint, since it can select more coefficients than the number of time periods in equation k. We refer to Lemma 1 in the mathematical appendix for more details on this characteristic of the elastic net. Although we do not include a model large enough for this property to play its role, in principle, it enables to estimate much larger model sizes than we consider in this work.

Besides, when looking at the Bayesian literature, such as Litterman (1986), Bayesian methods are able to induce additional structure on the process of variable shrinkage. Specifically, Litterman (1986) assumes that diagonal coefficients of A_i contain more signal than off-diagonal coefficients. The underlying idea is that own lags of variable k explain relatively more variation in the dependent variable than other variable lags. In addition, the fundamental Minnesota prior associated with the author is specified such that the prior standard deviation of each coefficient declines the more distant the respective lag is. The motivation is that recent lags are more informative than more distant lags in explaining the dependent variable. Consequently, GDP_{t-2} will obtain more shrinkage than GDP_{t-1}. Moreover, if the dependent variable in equation k is GDP_t, all lagged variables of GDP_t will obtain less shrinkage than other variables in the system. Contrary, the enet is lacking coefficient specific penalization. Namely, the researcher might find herself in a situation, in which prior knowledge of the data is present. Such knowledge could be formal or informal economic theory. For example, we might have an underlying economic theory, which supports the idea of a particular group of highly correlated variables. Or the researcher might believe that a certain lag structure contains more signal than another. Especially the latter assumption is a common finding in the literature. In such situations, other methods than the elastic net might be more desirable, imposing more structure on the penalized estimation. For example the group lasso (Yuan and Lin, 2006), adaptive lasso (Zou, 2006) or the feature weighted elastic net (Tay et al., 2020). Yet if we do not have such information or do not want to impose further assumptions on our model, the elastic net can be the better choice. Finally, the elastic net penalty $P_k^{enet}(\lambda, \alpha)$ does not discriminate the coefficients in terms of different weighting, since it does not include coefficient specific penalization. This might be too restrictive and let us seek for an alternative model.

3.2.3 Feature Weighted Elastic Net VAR

Finally, we turn to the fenet VAR and write the respective Lagrange function as

$$L(\theta^k, v_k, \beta_k)_{\lambda,\alpha} = \frac{1}{2}RSS + \lambda \left[\sum_{i=1}^{P} \sum_{j=1}^{K} w_{ij}(\theta^k) \left[\alpha|\beta_{kj}^i| + \frac{1}{2}(1-\alpha)(\beta_{kj}^i)^2 \right] \right].$$

, where $w_{ij}(\theta^k) = f(z_{ij}^{'}\theta^k)$ is the weight given to the i'th lag of variable j in equation k. The specified functional form for $w_{ij}(\theta^k)$ is in analogy to chapter 2.2, in which $\theta^k = [\theta_1^k, ..., \theta_G^k]'$ is specific for equation k. Hereby, the fenet VAR incorporates all advantages of the enet VAR. Moreover, it allows us to incorporate the desired additional structure on penalization. In chapter 2.2 we elaborated the fenet formally, however, in this chapter the main interest lies on the application to Vector Autoregressions. The auxiliary matrix Z is accompanied by additional flexibility in terms of grouping the coefficients in our system. In principle the researcher is free to group the Kp coefficients in equation k as she wishes. Note that, in general, the assigned grouping structures are allowed to overlap. The flexibility of the fenet VAR has similarities to the prior specification in Bayesian VAR models, such as Litterman (1980). One possibility for the researcher is to orientate herself on the Bayesian literature, and adjust the ideas of certain hyperparameter specifications for prior distributions in Bayesian models. For example, Banbura et al. (2010) specify the hyperparameter in their prior such that the ad-hoc belief of more relevant recent lags than distant lags is included. The idea is initially proposed by Litterman (1980), in which a hyperparameter controls the degree of shrinkage for more distant lags. Song and Bickel (2011) implemented this idea in the frequentist regularization framework using a lasso type VAR model. However, there is a major distinction between the approach the previous authors take and ours. The researcher using fenet VAR is not able to arbitrarily set the degree of shrinkage imposed on the coefficients. In the Bayesian literature, as well as in Song and Bickel (2011), the model requires both, to arbitrarily decide which coefficients to shrink differently (e.g. the more distant the lag, the higher the shrinkage) and to specify the relation among each other (the relation in shrinkage differs dependent on the size of the hyperparameter). Typically, either a harmonic or geometric decay is assumed for higher lag orders. On the contrary, the only arbitrary part in fenet VAR is to assign the coefficients in groups, which we assume to have common explanatory power for the dependent variable. This belief is quantified in the auxiliary matrix Z. It is not possible to assign an arbitrary amount of shrinkage on the prespecified groups of coefficients in Z. However, this can be seen as an attractive property of the fenet VAR. The amount of shrinkage for each group is determined by the algorithm, not the researcher. Thus, if our belief about the importance of a specific group of coefficients is true, we will observe a higher resulting θ_g^k in the output. However, there are several useful ways in assigning the coefficients into groups. Consequently, we propose three different grouping schemes, which are accompanied by different beliefs on the structure of the VAR system. Remember, that a major goal is to obtain an accurate forecast of important macroeconomic variables, such as GDP, Inflation and Federal Funds Rate. In chapter 4 we will examine if a particular scheme results in a superior forecasting performance.

1. *Scheme 1*

   The first scheme is motivated by the belief that coefficients corresponding to more distant lags should obtain more shrinkage, since they contain less signal. However,

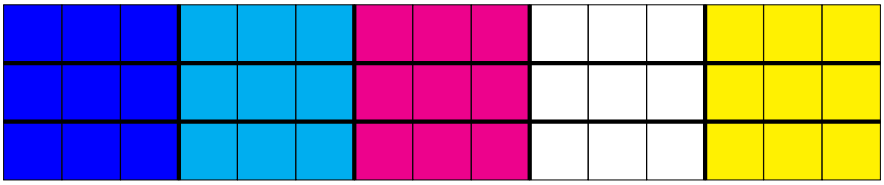
the fenet VAR does not specify ex-ante the relation of shrinkage for more distant lags. By partitioning all variables corresponding to the same lag order into the same group, we allow the algorithm to estimate θ_g^k such that each group in equation k is allowed to have a different amount of shrinkage. More formally, we set the number of different sources of feature information, previously denoted as G, to the number of lags p. Hence the auxiliary matrix Z is a matrix of dimension $(pK \times p)$.

Consider a simple example with $K = 3$ and $p = 2$, then the auxiliary matrix Z as well as the score $z_{ij}^{'}\theta^k$ (e.g. lag $i = 1$ and variable $j = 1$) for equation k are

$$Z = \begin{bmatrix} z_{11}^{'} \\ z_{12}^{'} \\ z_{13}^{'} \\ z_{21}^{'} \\ z_{22}^{'} \\ z_{23}^{'} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \qquad \text{and} \qquad z_{11}^{'}\theta^k = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \theta_1^k \\ \theta_2^k \end{bmatrix} = \theta_1^k$$

An exemplary grouping for three variables of a VAR(5) system is illustrated in figure 6, where each row corresponds to equation k. Hereby, the groups are specific for each equation k and are not linked across equations[8]. Thus, we would need to estimate $\theta^k = [\theta_1^k, \theta_2^k, \theta_3^k, \theta_4^k, \theta_5^k]'$ for each equation $k = 1,\ 2,\ 3$.

Figure 6: Example of *Scheme 1* for VAR(5) with $K = 3$



Every color in each row corresponds to a common imposed feature of the respective coefficients. Each equation is assumed to have five common features.

2. *Scheme 2*

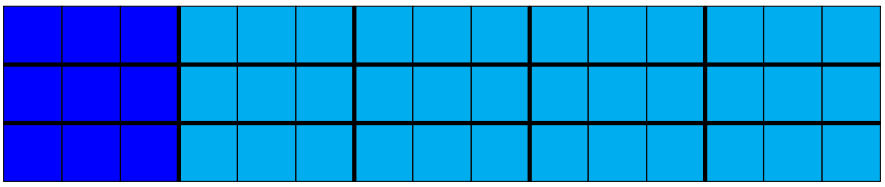
   This scheme includes a similar belief to *scheme 1*, however, the researcher is only interested in assigning a different weight between all variables of the first lag order and all variables of the remaining lag orders. Thus, we separate the first lag of all K variables in the same group. The remaining $pK - K$ coefficients represent the second group. By this grouping scheme, the researcher beliefs that θ_1^k will be relatively large and hence assign less shrinkage to the first group, however, θ_2^k should be relatively small and shrink the respective coefficients in the second group relatively more[9].

[8] Note that this argument holds as well for *scheme 2* and *scheme 3*, since we implement the VAR as a system of seemingly unrelated equations.

[9] We refer to section 2.2 for a nuanced elaboration of the implied regularization in dependence of θ_g^k.

Figure 7 represents an exemplary grouping of the current scheme for a small system of $K = 3$ variables and $p = 5$ lags.

Figure 7: Example of *Scheme 2* for VAR(5) with $K = 3$



Every color in each row corresponds to a common imposed feature of the respective coefficients. Independent of the lag order of the VAR(p) system, we only assume to have two common features per equation. This assumption is rather parsimonious regarding the dimension of θ^k, since it is always two.

3. *Scheme 3*

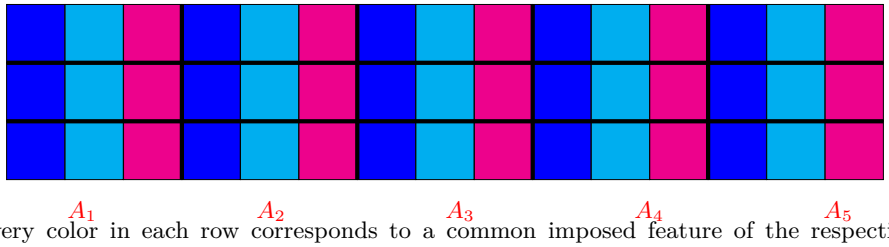
   Motivated by another property of the Litterman prior, namely assigning the diagonal elements less shrinkage as the off-diagonal elements in the VAR, we introduce *scheme 3*. For example, a VAR(5) system with {GDP, GDP Deflator, FFR} as variables, should give less shrinkage on all five lags of GDP in the first equation. The remaining 10 coefficients should obtain relatively more shrinkage since they are assumed to contain less explanatory power for the first dependent variable. However, we modify this approach and separate the remaining 10 coefficients one more time, such that each individual variable has an own group, including all five lags. In other words, we assume that all lags of the same variable have a common underlying feature. Thus, we have $G = K$ groups in total for each equation k. In analogy to the example of the first scheme, where $K = 3,\ p = 2$, the auxiliary matrix and the score are represented as

$$Z = \begin{bmatrix} z_{11}' \\ z_{12}' \\ z_{13}' \\ z_{21}' \\ z_{22}' \\ z_{23}' \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \text{and} \qquad z_{11}'\theta^k = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \theta_1^k \\ \theta_2^k \\ \theta_3^k \end{bmatrix} = \theta_1^k$$

As in the previous two schemes, we illustrate *scheme 3* exemplary in figure 8.

In the following chapters, we will compare fenet *scheme 1* to *scheme 3* to five different estimation methods in terms of a pseudo out of sample forecasting exercise. This exercise will focus on three main macroeconomic variables as well as on a system-wide performance evaluation using a multivariate forecast evaluation measure.

Figure 8: Example of *Scheme 3* for VAR(5) with $K = 3$



Every color in each row corresponds to a common imposed feature of the respective coefficients. Since we assume each equation to have K common features, this grouping might demand a relatively large dimension of θ^k for the high dimensional variable case, as we will encounter later in the empirical exercise.

4 Empirical Study

In this chapter we empirically investigate the forecast precision of the competing regularization methods by applying them on three key variables of interest: GDP, GDP Deflator and the Federal Funds Rate (FFR). For comparison, we include the OLS estimation as well as the approximate Dynamic Factor Model (Stock and Watson, 2002). We expect the OLS prediction to result in an inaccurate forecast as the model size increases, due to the problem of overfitting. On the contrary, the DFM is known to perform very well in prediction. Furthermore, we elaborate on sparsity among different model sizes. Then, we discuss an observed pattern for the tuning parameter α as well as variable selection in terms of binary coefficient matrices. Finally, we evaluate system wide performance using a multivariate forecasting evaluation measure.

4.1 Forecasting Setup

The data set is a subset from Stock and Watson (2008) and includes 22 quarterly macroeconomic US variables. In order to compare findings, it is common in the macroeconomic forecasting literature to use standard data sets containing similar variables and analyze similar model sizes. The focus rather lies on the performance and properties of the forecasting methods, than nuances of the data set. Thus we construct a data set, motivated by Giannone et al. (2015), Kascha and Trenkler (2015) and Koob and Korobilis (2013). Similar to these authors, we classify the variables into different model sizes of $K = 3$, 7 and 22, referring to these as *small*, *medium* and *large*, respectively. Furthermore, we include an additional step in model size by considering $K = 15$ variables, calling it the *intermediate* size model. The data ranges from period 1959:Q1 to period 2012:Q2. Thus it includes the last recession and consists of $T = 214$ observations per variable. We follow the variable transformation similar to Koob and Korobilis (2013), where the variables are transformed to stationarity. Following on Kascha and Trenkler (2015), we annualize most of the variables. Some variables need to be differentiated twice, leading to a sample size of $T = 212$, since periods 1959:Q1 and 1959:Q2 are excluded. However, a part of the signal gets lost by differencing variables twice and is translated into noise. Thus we follow Christiano et al. (1998) and take the first difference of GDP Deflator, instead of the sec-

ond. In other words, we include Inflation rather in level than in first differentiated form. The individual variables and transformation codes are summarized in table 3. The three transformed variables of interest are illustrated in figure 12. Despite the transformation of the variables, we will still refer to the variables in their original name, such as "GDP Deflator" instead of "annualized Inflation". Following Giannone et al. (2015) we set the number of lags in each model size to five. The forecast of each estimation method and model size combination is evaluated using an expanding window forecasting exercise. For tuning parameter validation, we initialize the sample $\{y_{k,t}\}_{t=1}^{T_0}$, where $T_0 = 149$, representing 70 % of the data set. Hereby, the estimation window expands one step at a time with the endpoint being $T^* = \{T_0, ..., 212\}$, where period 149 corresponds to 1996:Q3 and period 212 to 2012:Q2. For the regularization methods, we use the first sample of the estimation window, i.e. $\{y_{k,t}\}_{t=1}^{T_0}$ for validating the respective tuning parameter(s). However, the coefficient vector β_k is iteratively estimated using $\{y_{k,t}\}_{t=1}^{T^*}$ for $T^* = T_0, ..., 212$ for all estimation methods. After obtaining $\hat{\beta}_k$ we predict the pseudo out of sample observation y_{k,T^*+h}. We consider three different forecast horizons, $h = 1, 4, 8$, where the prediction for $h = 4$ and $h = 8$ is computed recursively. The one period forecast contains $H = 63$ predicted observations ranging from 1996:Q4 up to period 2012:Q2. The four period forecast contains $H = 60$ predicted observations ranging from 1997:Q3 up to period 2012:Q2. The eight period forecast contains $H = 56$ predicted observations ranging from 1998:Q3 up to period 2012:Q2. In order to evaluate the forecast, we define the MSFE for equation k as

$$MSFE_k = \frac{1}{T - T_0 - h + 1}\{\sum_{t=T_0}^{T-h}(y_{k,t+h} - \hat{y}_{k,t+h}|\Omega_t)^2\}$$

, where $\hat{y}_{k,t+h|\Omega_t}$ is the predicted value of variable k, conditioned on the information set $\Omega_t = \{Y_1, ..., Y_t\}$ for a particular estimation method and model size combination. Often, it is appropriate to use a trivial benchmark model for comparison. Two popular choices are to either use a random walk or a conditional mean forecast. We decide to use the conditional mean forecast (CMF) for this role, since it leads to an interesting argument, as we will see soon. The CMF is defined as

$$\hat{y}_{k,T^*+h}|\Omega_{k,T^*} = \frac{\sum_{t=1}^{T^*} y_{k,t}}{T^*} \quad \text{for} \quad T^* = \{T^0, ..., T - h\}$$

Finally, we write the absolute criterion $MSFE_k$ in relative terms as

$$MSFE_k^R = \frac{MSFE_k}{MSFE_k^{cmf}}$$

, where $MSFE_k^{cmf}$ is calculated using the predicted set of conditional means $\{\hat{y}_{k,T^*+h}|\Omega_{k,T^*}\}_{T_0}^{T-h}$.

4.1.1 Model Confidence Sets

Another procedure to compare the prediction accuracy of several models is the Model Confidence Sets (MCS) approach by Hansen et al. (2011). Particularly, this approach can be seen in analogy to confidence intervals for individual coefficients. Instead of coefficients, we are interested in the significance of different models. The result of the MCS approach is a set of "superior" models $\mathcal{M}^*$, which is a subset of the total set of models $\mathcal{M}$. The final set $\mathcal{M}^*$ contains the best model with confidence level $1-\alpha$. The MCS procedure is an iterative procedure. At each iteration, pairwise comparisons are made under the null hypothesis of an equal relative performance measure in expectation, where the performance measure is the difference in loss functions of the respective models. If any of these null hypotheses is rejected, the model with the highest pairwise difference is eliminated from $\mathcal{M}$. This iteration is repeated until the null hypothesis can not be rejected anymore. By construction, further distinctions in forecast performance can not be made from the final set of models in $\mathcal{M}^*$. This superior set contains the best models with coverage probability $1-\alpha$. We determine the superior set of models[10] to quantify significance between the competing models within the individual variable forecasts as well as the multivariate forecasts.

4.1.2 Dynamic Factor Model

As a competing benchmark model for the forecasting exercise, we use the approximate Dynamic Factor Model proposed by Stock and Watson (2002)

$$\begin{aligned} Y_t &= \Lambda F_t + e_t \\ y_{k,t+h} &= \beta_F^{'} F_t + \beta_w^{'} w_t + \epsilon_{t+h} \end{aligned} \tag{11}$$

, where w_t are observable variables (such as lags of $y_{k,t}$), F_t is a r dimensional vector of latent factors and ϵ_{t+h} is the forecast error. It can be shown that the RSS of equation 11 is minimized by the principal component estimator $\hat{F}_t(\hat{\Lambda}) = \hat{\Lambda}' Y_t$, due to $\hat{\Lambda}\hat{\Lambda}' = I$, where $\hat{\Lambda}$ is of dimension $(K \times r)$ and contains the first r eigenvectors of the sample covariance matrix. The h step ahead forecast is obtained in two steps. In the first step, we obtain the first r principal components of $\{Y_t\}_{t=1}^{T}$. These are used as estimates of the factors F_t. In the second step, we fit $\hat{y}_{k,T+h} = \hat{\beta}_F^{'} \hat{F}_T + \hat{\beta}_w^{'} w_T$, where $\hat{\beta}_F^{'}$ and $\hat{\beta}_w^{'}$ are obtained by regressing $\{y_{k,t+h}\}_{t=1}^{T-h}$ on $\{\hat{F}_t\}_{t=1}^{T-h}$ and $\{w_t\}_{t=1}^{T-h}$. Since the DFM is of minor importance in this work, we follow Banbura et al. (2015) and set the number of factors to three, without experiment with additional explanatory variables[11] w_t.

[10] In order to obtain MCS, we use the R package *modelconf* and a significance level of $\alpha = 0.15$. Thus the superior set $\mathcal{M}^*$ contains the true set of models with coverage probability 0.85. The block length used for the moving-block bootstrap is obtained via the function *bstar* from the package *np*, which uses the methods introduced in Patton, Politis and White (2009).

[11] In principle, we could include further lags of the dependent variable or exogeneous variables into the regression. Note that we use a basic approximate DFM here, although the literature gives plenty of extensions and nuances on.

4.2 Results

Now, we turn our attention to an empirical forecast comparison of the regularization methods introduced in the previous chapters. Tables 4 to 6 illustrate the $MSFE_k^R$ for GDP, GDP Deflator and the FFR. This set of tables expresses the forecast in relation to the conditional mean forecast. Hereby, values smaller than 1 indicate better performance than the benchmark. First, we are interested in the best performance between *scheme 1* to *scheme 3* of the fenet VAR model. Second, we want to see if the ex-ante knowledge induced by the fenet model outperforms the enet estimation. Third, it is of interest if the advantages accompanied by both of these methods result in better out of sample prediction than the classical shrinkage techniques such as lasso and ridge. In general, it is of major interest for all penalized techniques if the forecast precision improves as we move to the large model. Finally, we compare the regularization methods with the OLS method and the popular approximate Dynamic Factor Model with three factors.

As expected, we denote that OLS performs worse for almost all horizon and variable combinations the larger the model gets. This aligns with the curse of dimensionality as described in chapter 3.1 that arises with many variables and/or many lags in the VAR model. The more variables we include in our model (keeping the number of observations fixed), the more the estimated coefficients explain the noise in the particular data set rather than the signal. Thus, the estimated model fails to generalize on new data points. As a result, predicted values differ substantially from the true dependent variable, resulting in an absurdly large MSFE. Consequently, OLS is outperformed by the regularized techniques for model sizes of $K > 3$ for all horizon combinations. Additionally, most of the results are not included in the respective MCS $\mathcal{M}^*$.

In chapter 3.2.3, we motivated three different beliefs which allow us to impose structure on the penalization of the coefficients in the VAR system. All three beliefs were inspired by the Bayesian literature. Consequently, we are keen to compare the forecasting precision of *scheme 1* - *scheme 3* among each other. For GDP, we see in table 4 that *scheme 1* and *scheme 2* slightly outperform *scheme 3*, without a visible difference between the first and second scheme. For the DGP Deflator, fenet 3 outperforms fenet 1 and fenet 2 for most of the horizon/model size combinations, which is even more pronounced the larger the model and the horizon get. *scheme 1* performs slightly better than *scheme 2*. For the third variable, the FFR, we do not see a difference for $h = 1$, however, a 2-10 % better relative performance arises for both multistep horizons using, again, *scheme 3*. Once more, the improvement is more pronounced the more variables we include in our model. In summary, although not uniformly, it seems that *scheme 3* performs especially well among the three different imposed structures. Remember, that it groups the coefficients by variables, instead of lags.

Next, we are interested in the added value by imposing additional structure on the penalization, ergo using the fenet VAR instead of the enet VAR[12]. For GDP we denote a

[12]Note that we used the same folds for hyperparameter cross validation in all mixture penalizations, which

relative improvement of around 2-6 % for $h = 1$ and $h = 4$ by using *scheme 1*, compared to the enet VAR. A perceptible improvement is especially noticeable for the second variable, GDP Deflator, where *scheme 3* outperforms the enet VAR by up to 12 %. The reduction in $MSFE_k^R$ is especially visible for the multistep forecasts and for the intermediate to large model. For the FFR, *scheme 3* performs better than the enet for up to 4 %, where the improvement behaves in the same pattern as for the GDP Deflator. Thus, we conclude that imposing additional structure on penalization results in a better forecast, compared to the same model, with uniform shrinkage of the coefficients (enet VAR).

For the first variable, we see that lasso and ridge are outperformed by using mixture penalizations for almost all model sizes and horizon combinations. On the contrary, lasso and ridge result in a smaller forecast error compared to the remaining shrinkage methods for the GDP Deflator for horizon $h = 4$ and $h = 8$, as we see in table 5. Especially ridge performs noticeably well in these situations. For the third variable, the lasso and ridge are outperformed by the enet and fenet schemes for most of the combinations.

In addition, most of the forecasts across all methods, which result in relatively good performance in terms of $MSFE_k^R$, are simultaneously included in the respective model confidence sets $\mathcal{M}^*$. Moreover, we do not find noteworthy anomalies regarding the inclusion in $\mathcal{M}^*$.

An emphasis of our analysis lies on the research question if we can improve forecast performance by including further variables in our model. As table 4 - 6 illustrate, this is not uniformly the case. For the enet VAR and fenet VAR, we reach the smallest forecast error in the large model case across all horizons for the GDP variable, however, the FFR reaches the smallest error for the large model only for the $h = 1$ horizon. Furthermore, the GDP Deflator performs best in the small model case, uniformly for all horizons. This is not surprising, since a visual inspection of figure 12 reveals that the GDP Deflator is highly persistent. A persistent variable largely depends on the state of the previous periods of the same variable. Thus, the majority of variation in GDP Deflator is already explained by the small model (mainly by the own five lags). The persistent behavior is also observed in the structural impulse response function in figure 15, since GDP Deflator seems to converge to zero at a slow rate. However, by including further variables, the model does not gain any forecast accuracy, but rather results in overfitting relatively early. This finding aligns with the resulting tuning parameter α for enet VAR and fenet VAR. Precisely, figure 10 illustrates an increase in α as we increase the model size for the GDP Deflator. For the large model size, it even results in $\alpha = 1$, which simplifies the enet VAR to the lasso VAR. This pattern confirms our expectation, since the persistent variable GDP Deflator, is rather explained by a sparse model.

In addition, we notice an almost consistent increase in prediction error when moving from the medium model to the intermediate model for all regularization methods and forecast horizons. A potential explanation might be that the additional eight variables

enables a detailed comparison among them.

included in the intermediate model (see table 3) do not include enough novel information to explain the respective dependent variable. In other words, the trade-off in decreasing the degrees of freedom on the one hand and increasing the information provided by the total set of variables on the other hand, seems to weaken prediction accuracy. In this case, the intermediate model rather brings the model closer to overfitting than improve prediction.

Given that the one step ahead forecast is of special interest among the three different horizons, we notice that the approximate Dynamic Factor Model is outperformed by up to 6 % for the GDP variable and up to 13 % for the GDP Deflator. However, it does result in a lower MSFE for the FFR, by around 7 %, compared to the respective strong performing methods, the enet VAR and fenet VAR *scheme 1 - scheme 3*. Furthermore, the DFM performs especially well for the multistep forecast horizons, even though there is almost always a regularization method which gives a better prediction. Although this comparison is not particularly fair, since we compare six different variations of penalization to one variation of the DFM. In general, the researcher could try to improve forecast accuracy of the DFM by choosing the number of factors r via information criteria, instead of an ad-hoc specification and/or include further lagged dependent variables in the regression.

Finally, we notice that for almost all combinations, the forecast gets worse when moving from an one step ahead horizon to a multistep horizon. This aligns with our intuition, since the h-step ahead forecast is based on Ω_t and thus closer horizons should, in general, be predicted with greater accuracy[13] than farther horizons. However, by increasing the horizon from $h = 4$ to $h = 8$, an occasional contradiction occurs, since only the GDP Deflator forecast accuracy gets considerably worse. The first and third variable prediction even get slightly better. Although it seems ambiguous at first sight, there is a reasonable explanation. A visible inspection of figure 12 immediately reveals, that the first and third variable move around their respective unconditional mean μ_k for most of the time periods. We know that a predicted variable, $\hat{y}_{k,t+h}$, converges to its conditional mean $E[y_k|\Omega_t]$, as $h \to \infty$, implying the forecast is less distant to the conditional mean for a larger horizon. It seems that this argument leads to a slight decrease in mean squared forecast error for GDP and FFR, in our case. On the contrary, the second variable has large spikes from 1971-1984 (see figure 12). Broadly speaking, the conditional mean including information up to time period 1984, will be "inflated", and not represent the following time periods very well. Consequently, our expanding window forecasts, initializing Ω_{1992} as the first information set, would converge to an "inflated" conditional mean as $h \to \infty$. In general, this "inflated" conditional mean is not close (in MSE sense) to the true out of sample observations, resulting in a larger MSFE for $h = 8$ than for $h = 4$. This pattern in the data[14] is so present, that we are able to say, the distorted predictions for $h = 8$, underlie

[13] Note that although this explanation is tailored to the absolute $MSFE_k$ results, we do not include these in this work, since it would be redundant with the relative results. Still, the present elaboration translates to the relative results.

[14] The particular transformation of variables plays a significant role in this discussion.

an "upward bias", compared to $h = 4$.

4.3 On Sparsity

Next, we are interested in the sparsity pattern between the lasso, elastic net and feature weighted elastic net *scheme 1 - scheme 3*. Figure 9 represents the percentage of sparsity for the three variables of interest and the whole VAR(5) system containing $K = 22$ dependent variables. We obtain the percentage of sparsity as follows. First, we use cross validation on $T_0 = 149$ observations as described in chapter 2.3.1 to validate α and λ for the enet and variations, or only λ for the lasso. Using the resulting optimal tuning parameter(s), we conduct a pseudo out of sample forecast in an expanding window scheme, as described in section 4.1. At each forecast step, we extract the coefficients of equation k and calculate

$$\mathbb{S}_{k,T^*} = \frac{\#\{\hat{\beta}_{kj}^i = 0\}_{i=1,j=1}^{p,K}}{pK}.$$

Finally, we calculate the average using the whole forecast interval as

$$\bar{\mathbb{S}}_k = \frac{1}{T - T_0 + 1} \sum_{T^*=T_0}^{T} \mathbb{S}_{k,T^*}$$

For the full model, we average $\bar{\mathbb{S}}_k$ over all equations in an additional step, resulting in $\bar{\mathbb{S}}$. Subsequently, we refer to both, $\bar{\mathbb{S}}_k$ as well as the full model measure $\bar{\mathbb{S}}$ as the "percentage of sparsity".

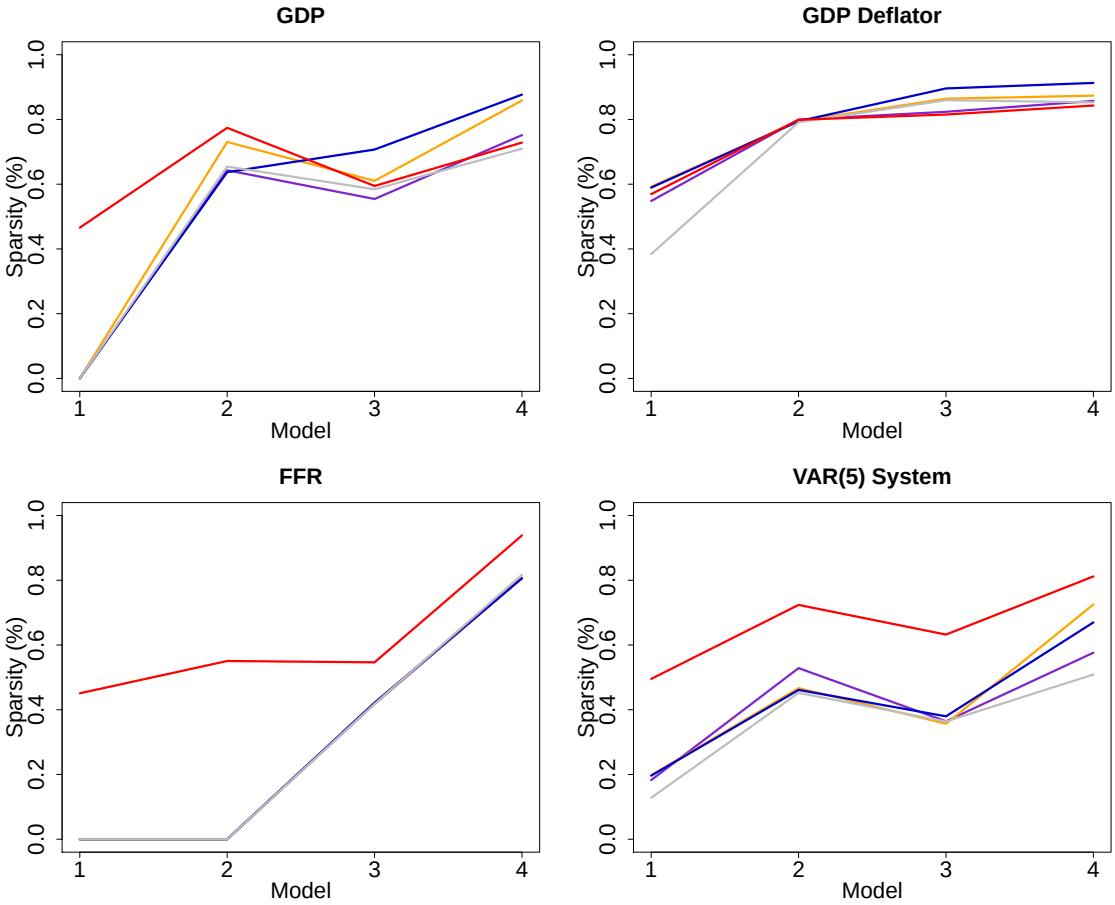
In general, the percentage of sparsity increases as we include further variables in the model. The intuitive reason is that as we increase the model size, in tendency, we include relatively more coefficients without sufficient signal to be selected. In other words, the relative amount of "non-significant" coefficients increases as we include further variables[15]. Note that this pattern does not always hold. Although we observe it for the whole VAR system, GDP Deflator and the FFR, the percentage of sparsity does not increase consistently with model size for GDP. However, it seems the induced sparsity of the whole system reveals a drop when moving from the medium sized model to the intermediate size. A possible reason is that the majority of the additional eight variables bring sufficient signal to be selected, given the respective tuning parameter combination. Hereby, figure 11 illustrates the optimal α for each equation k, used for the coefficient estimation in this analysis. A welcomed pattern gets visible. Namely, the intermediate model brings additional equations $k = 8, ..., 15$ into the system, which result in $\alpha = 0$ in four of these equations. This in turn, leads to no variable selection at all for these equations. In total, the percentage of sparsity reduces for the whole system.

Finally, we denote that the lasso contains the largest percentage of sparsity for the whole system. This is no surprise, since variable selection is guaranteed for each equation k. The

[15] In this context, "non-significant" describes a coefficient with not enough signal to be selected by the lasso, enet or fenet soft-thresholding operator.

sparsity pattern for the elastic net and feature elastic net schemes are similar, although the fenet *scheme 2* puts on slightly more sparsity for the GDP and GDP Deflator. In the next two sections, we distance ourselves from the forecasting exercise and focus on the tuning parameter α as well as the variable selection pattern, using the full set of observations ($T = 212$) for tuning parameter(s) validation.

Figure 9: Sparsity



Sparsity pattern for GDP, GDP Deflator, FFR and the whole VAR(5) system. Model 1-4 correspond to small, medium, intermediate and large model. Included methods are lasso (red), enet (purple), fenet *scheme 1* (orange), fenet *scheme 2* (blue) and fenet *scheme 3* (grey).

4.4 On Tuning Parameter α

From the previous chapters, we already know that $\mathcal{P}_k^{enet}$ nests $\mathcal{P}_k^{lasso}$ as well as the $\mathcal{P}_k^{ridge}$ penalty. This nesting is captured by the tuning parameter α. Now, the researcher might be interested if the weighting between these two penalties in the four model sizes reveals a pattern in the empirical application. Thus, we investigate the resulting tuning parameter α on the full data set ($T = 212$ observations). In theory, as we include a new set of variables into the already large model, ceteris paribus, two extreme cases might occur. First, each coefficient in the additional set contains small signal. A model with relatively many coefficients containing small signal would be the result, compared to the initial model without additional coefficients. In other words, there are not sufficiently many

coefficients with relatively strong signal for the enet VAR to assign $\alpha = 1$. The enet VAR will in tendency provide a higher weight on the ridge penalty, i.e. we would expect α to be close to zero. On the contrary, the additional set of coefficients might contain a few coefficients with strong signal and many with approximately zero signal. This results in a relatively sparse structure of the final model, compared to a model without additional coefficients. Then, the enet VAR should give more weight to the lasso penalty. These two cases represent the boundaries of our expectations, represented by $\alpha = 0$ and $\alpha = 1$.

Figure 10 illustrates the validated α for every equation k of all four model sizes for enet and fenet *scheme 1*. By increasing the number of variables, especially when moving from the intermediate to the large model, we denote a pattern in the validation of α. It seems, that both regularization methods tend to select α more often on the boundary of $[0, 1]$, which becomes more visible as we include further variables into the model. Although this observation does not hold uniformly for all equations k, it is especially noticeable for equation 1, 2, 5, 6, 7 and 15. As we increase the model size, these equations result in an α, which is closer to the boundary. Particularly, the resulting tuning parameter α tends to move in direction of zero. This can be seen as an empirical result, indicating that as we move to the large model, many dependent variables are better explained by a rather non-sparse model structure. The enet VAR as well as the fenet VAR seem to capture this effect. Going further, a resulting ridge penalty ($\alpha = 0$) is observed for a large part of dependent variables. However, we also notice $\alpha = 1$ for several variables k, indicating that the respective dependent variables are better explained by a sparse structure.

In addition, the elastic net penalty as well as the fenet penalty tend to have the same tuning parameter α in many equations. However, it is not rare to see a slight difference between these two models. Finally, note that this pattern becomes visible by using the complete number of observations and is not clearly pronounced for $T_0 = 149$, as we see in figure 11. A possible reason is that as we increase the number of observations (c.p.) the elastic net has more observations to validate as well as assess the prediction error over the two dimensional surface $\mathbf{\Lambda} \times \mathbf{A}$, resulting in more precise tuning parameters $\{\lambda, \alpha\}$.

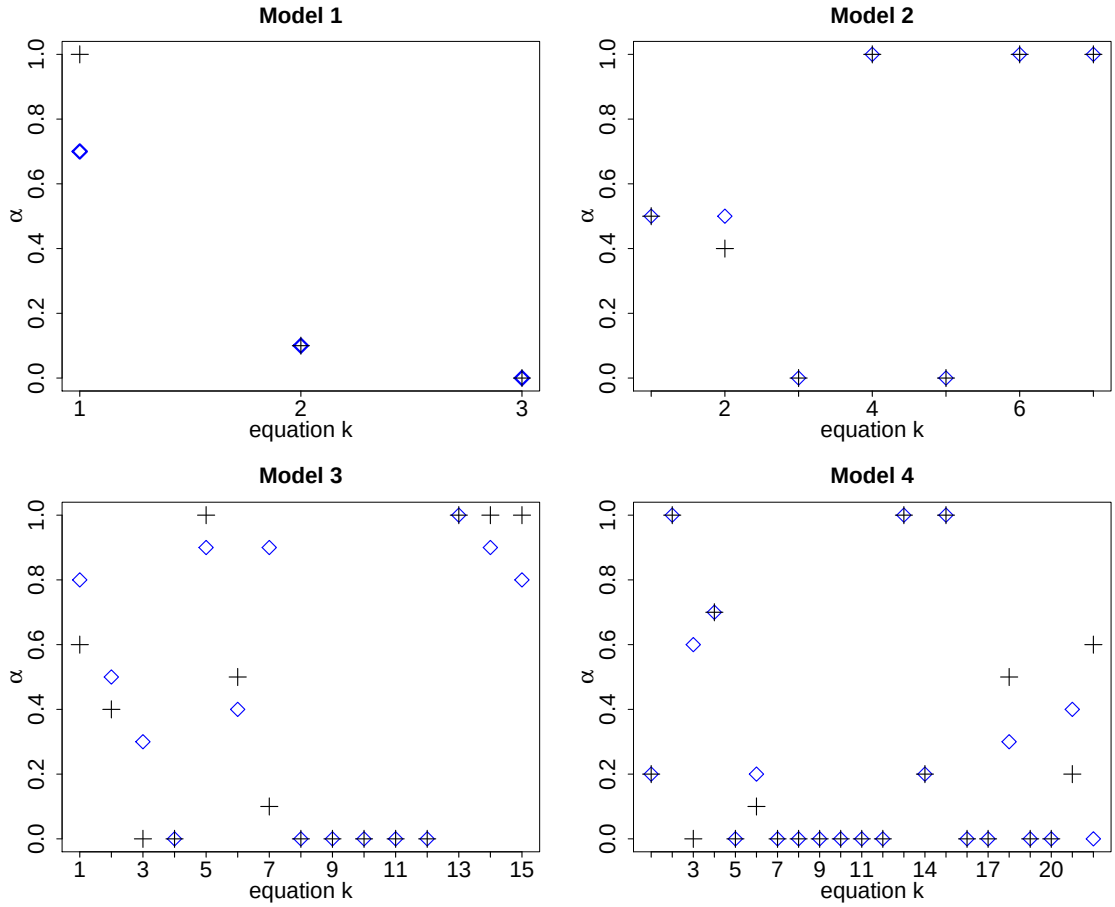
4.5 On Variable Selection

Figures 13 and 14 include the binary[16] coefficient matrices $\hat{A}_1, ..., \hat{A}_5$ for enet VAR and lasso VAR, respectively. Both figures are able to demonstrate the respective selected coefficients for the large model size with $K = 22$ variables. As already mentioned in section 3.2.1, the variable selection ability of both methods enables a Granger-Non Causality interpretation, which is directly observable from the respective binary plots.

Furthermore, when we look at plot 14, we see that the lasso shrinks all coefficients of equation 10, 11 and 20 exactly to 0. In the literature, this is known to be a rather uncommon behavior of the lasso. Obviously, the binary matrices for enet VAR align with the validated tuning parameter α in figure 10. If $\alpha = 0$, then all $pK = 110$ coefficients of

[16] Binary in the sense that, an estimated coefficient is either equal or unequal to zero.

Figure 10: Tuning parameter α for $T = 212$



+ represents fenet *scheme 1*; ◇ represents enet. Model 1 - Model 4 correspond to small, medium, intermediate and large model, respectively.

the respective equation are selected, which is observed for a good amount of equations[17], including equation 10, 11 and 20. From chapter 2.1, we know that it reduces $P_k^{enet}(\lambda, \alpha)$ to $P_k^{ridge}(\lambda)$. This tells us that the ridge penalty is a better choice than the lasso penalty, since it gives a smaller inner test set MSFE (chapter 2.3.1). We are interested in finding an explanation for this, at first sight, ambiguous selection pattern for equation 10, 11 and 20, which correspond to 5 year bond rate, 10 year bond rate and oil price. Although this result is surprising at first sight, it rather illustrates the intuition behind the lasso, ridge and enet penalty. Apparently, all pK coefficients in equation 10, 11 and 20 do not have enough signal to be selected for the lasso and thus to sufficiently explain the respective dependent variables. Hence we rather have many coefficients with small signal, instead of a sparse model. More precisely, the signal has to be so small, such that the lasso Soft-Thresholding Operator shrinks these values towards zero. In addition, the high degree of multicollinearity in the large system plays a role. It is well known that the lasso shrinkage can be quite unintuitive, in case of collinear predictors. However, the ridge is known to handle this problem well, leading the enet to select $\alpha = 0$. A further experiment[18] reveals,

[17] These equations are 5, 7, 8, 9, 10, 11, 12, 16, 17, 19, 20 and 22.

[18] We estimated the lasso solution $M = 500$ times and calculated the percentage of non-zero coefficients in each estimation for equation 10 for various lag lengths and numbers of variable combinations. The

as we reduce the lag length and/or reduce the number of variables in the VAR system, the degree of multicollinearity (measured by the VIF from chapter 3.1) shrinks. This, partially, enables us to circumvent the radical shrinkage behaviour of the lasso, since more coefficients are selected.

4.6 Multivariate Forecast Evaluation

In order to compare the predictive performance of the introduced models for the whole VAR(5) system, instead of only three main variables of interest, we define the Multivariate Weighted MSFE (introduced by Christoffersen and Diebold (1998) and used by Koop et al. (2017)) as

$$WMSFE = \frac{1}{T - T_0 - h + 1} \sum_{t=T_0}^{T-h} w\epsilon_{t+h}$$

Hereby, $w\epsilon_{t+h} = \epsilon'_{t+h} W \epsilon_{t+h}$ denotes the weighted forecast error of a particular model at period $t+h$ and $\epsilon_{t+h} = (\epsilon_{1,t+h}, ..., \epsilon_{K,t+h})'$ is a K dimensional forecast error vector, where $\epsilon_{k,t+h} = y_{k,t+h} - \hat{y}_{k,t+h}$. We follow Koop et al. (2017) and specify the $(K \times K)$ weighting matrix W with the inverse of the series variances on the diagonal. Note that by using WMSFE as a measure of forecast evaluation, the comparison across different model sizes is limited, since $w\epsilon_{t+h}$, in general, gets larger as we increase the number of variables K. In analogy to the univariate forecast evaluation, we define the relative WMSFE as

$$WMSFE^R = \frac{WMSFE}{WMSFE^{cmf}}$$

, where $WMSFE^{cmf}$ is based on the conditional mean forecast.

Table 2 includes the $WMSFE^R$ for the different forecasting models as well as horizon and model size combinations. In order to construct a fair comparison to the approximate DFM, we use as many variables to determine the three common factors, as the size of the respective system is. The table reveals several things. First, as expected, OLS yields a more inaccurate prediction, relative to the other methods, the larger the model size gets. This is not surprising and is, as in the univariate prediction exercise, due to overfitting[19]. Consequently, the OLS forecasts for the intermediate and large model are not included in the respective model confidence set $\mathcal{M}^*$. Second, we notice that fenet VAR *scheme 3* slightly outperforms the other schemes, which is more pronounced for the intermediate and large model case as well as the multistep forecasts. It seems, that grouping the coefficients by variable, instead of by lags, improves performance in the large model case. In addition, imposing additional structure on penalization has a positive effect on forecast precision, as the fenet VAR *scheme 3* outperforms the enet VAR slightly. Finally, we notice that lasso

mean of percentage of non zero coefficients across M estimations, increased as we, hopefully, reduced the degree of collinearity among the predictors (using the percentage of covariates with a $VIF > 10$ and the average VIF as rough indicators).

[19] See in chapter 3.1 and 4.2 for a detailed discussion. The same arguments are applicable for the multivariate evaluation.

and ridge outperform the methods including two tuning parameters for the intermediate as well as large model. This is partially surprising, since the enet VAR and fenet VAR should, in principle, enjoy better performance due to incorporated advantages of both, the L1 and the L2 penalty. A reason for this happening might be the following. As we validate not only λ, but also α, we give the CV algorithm less observations for λ to validate on. In turn, it might result in less reliable λ for the enet/fenet schemes. Apparently, this gets visible as we conduct a multivariate forecast evaluation instead of an univariate, as in section 4.2. Finally, the approximate DFM performs best in the small model case (for $h = 1$), and is outperformed for almost all other horizon and model size combinations.

Table 2: Multivariate Forecast Evaluation

Hor.				$K = 3$				
	OLS	Ridge	Lasso	Enet	Fenet	Fenet 2	Fenet 3	DFM
h=1	0.7303*	0.6303*	0.7003*	0.6157*	0.6158*	0.6171*	0.6187*	**0.5638***
h=4	0.8191*	0.7950	0.8353	0.7645*	0.7653*	0.7682*	**0.7516***	0.7715*
h=8	0.9647*	0.8873*	0.8855*	0.8516*	0.8543*	0.8556*	**0.8351***	0.9045*
				$K = 7$				
	OLS	Ridge	Lasso	Enet	Fenet	Fenet 2	Fenet 3	DFM
h=1	0.8972	0.7425	0.7494	0.7215*	**0.7117***	0.7130*	0.7221*	0.7535*
h=4	1.1089*	0.9752*	0.9632*	0.9618*	0.9603*	0.9657*	**0.9595***	0.9926*
h=8	1.1153*	0.9987*	**0.9786***	0.9820*	0.9818*	0.9836*	0.9797*	1.0490
				$K = 15$				
	OLS	Ridge	Lasso	Enet	Fenet	Fenet 2	Fenet 3	DFM
h=1	1.4147	0.6846*	**0.6815***	0.8205	0.8286	0.8320	0.7571	0.7926
h=4	1.7582	1.0178*	1.0517*	1.0133*	1.0275*	1.0300*	1.0031*	**1.0021***
h=8	1.4522	1.0124*	1.0466*	1.0135*	1.0280*	1.0226*	**1.0049***	1.0330*
				$K = 22$				
	OLS	Ridge	Lasso	Enet	Fenet	Fenet 2	Fenet 3	DFM
h=1	3.0257	0.7199*	**0.7115***	0.7784*	0.7762	0.7782	0.7400*	0.8008*
h=4	3.3387	0.9785*	0.9647*	0.9578*	0.9650*	0.9658*	**0.9402***	0.9509*
h=8	4.0546	0.9859*	0.9728*	0.9915*	0.9931*	0.9902*	**0.9596***	1.0165*

Entries containing "*" are part of the MCS with $\alpha = 0.15$. **Bold** entries have the lowest $WMSFE^R$ for the particular model size and horizon combination. Values are in relation to conditional mean forecast.

4.7 Structural Impulse Response Analysis

In the previous sections, we empirically proved that regularized VARs are capable of an accurate prediction in a high-dimensional variable setting. However, in this section we want to investigate if regularized techniques are capable of a structural analysis. Initially, shrinkage methods are by design constructed for predicting accurately. However, in the context of applied macroeconomics, it is an attractive feature of the model, when it is capable of both, an accurate out of sample prediction as well as trustworthy structural analysis. Therefore, a structural impulse response analysis is often chosen. The purpose of this analysis is to capture the response of the variables as a reaction to an exogenous shock in one or several other variables of the system. The researcher tries to identify the shock in the VAR system, by using economically motivated identification assumptions. Thus, we are able to model the dynamic relationships of the variables included in our system, which,

at least in theory, represent the economy. Practitioners often handle this type of structural analysis as a well-suited tool for policy evaluations. However, following Luetkepohl (2008), researchers typically work with low dimensional VAR models in terms of a structural impulse response analysis. Consequently, an omitted variable bias problem, is not a rarity. Every variable that is not included in the system is assumed to be in the residuals, which might lead to severe inaccuracies in the impulse response functions. Corresponding to this issue, we encounter the so called Price Puzzle in this chapter (Christiano et al., 1998), which is dampened by including further variables into the system.

Before we apply our methods to a real data example, it is convenient to introduce the required methodology. Let us transform the VAR(p) process from equation 8 in structural form as

$$B_0 Y_t = B_1 Y_{t-1} + ... + B_p Y_{t-p} + w_t$$

, where $B_i = B_0 A_i$ for $i = 1, ..., p$ and $w_t = B_0 u_t$ is a K dimensional vector containing structural shocks, given Σ_w is a diagonal matrix and has economic meaning. Furthermore, w.l.o.g we assume that $\Sigma_w = I_K$, implying that the shocks are uncorrelated among each other. By construction the residual covariance matrix is $\Sigma_u = E[u_t u_t'] = B_0^{-1} E[w_t w_t'] B_0^{-1\prime} = B_0^{-1} \Sigma_w B_0^{-1\prime} = B_0^{-1} B_0^{-1\prime}$. Furthermore, the VAR(p) can be written in Wold MA representation as

$$Y_t = \sum_{i=0}^{\infty} \Phi_i u_{t-i} = \sum_{i=0}^{\infty} \Phi_i B_0^{-1} B_0 u_{t-i} = \sum_{i=0}^{\infty} \Theta w_{t-i}$$

, which results in the structural impulse response function $\Theta_i = \Phi_i B_0^{-1}$ with $\Phi_1 = I_K$. In order to construct Θ_i, we require identification of B_0^{-1}. Therefore $\frac{K(K-1)}{2}$ equality restrictions are a necessary, but not sufficient condition. For identification of the monetary policy shock, we use the standard recursive identification scheme, which we adapt to models of different sizes. Following Banbura et al. (2010), Christiano et al. (1999) and Bernanke et al. (2005), we separate the set of variables into slow moving variables (s_t), the monetary policy instrument (r_t) and fast moving variables (f_t). The variables are ordered as $Y_t = (s_t', r_t, f_t')'$. B_0^{-1} is estimated via a lower triangular Cholesky decomposition of the estimator of Σ_u. The underlying identification assumption is that slow moving variables such as real variables and prices do not respond contemporaneously to an exogenous change in the monetary policy instrument. On the contrary, fast moving variables such as financial variables do respond contemporaneously. A precise classification of variables is given in table 3. Latter identification assumption is consistently adjusted for every model size of consideration, such that the ordering of s_t, r_t and f_t is kept. Following Banbura et al. (2010), we rescale the shock such that the FFR is increased contemporaneously by 100 basis points. We use the full data set of $T = 212$ observations for validation of the respective tuning parameter(s) as well as coefficient estimation. Figure 15 - 16 illustrate the impulse responses of a monetary policy shock on the FFR for the small, medium,

intermediate and large model, using OLS, lasso VAR, ridge VAR, enet VAR and fenet VAR[20] *scheme 1*. Figure 17 - 19 include the s-IRFs of the remaining 19 variables of the large model for the respective estimation methods.

In general, we notice that all s-IRFs contemporaneously react, as the theory dictates. Namely, the slow moving variables such as real GDP or Industry Production do not react contemporaneously, but the fast moving variables (e.g. five year bond price, oil price) react immediately. Since the FFR represents the monetary policy instrument, by construction, a contemporaneous increase of 1 % is evident. Broadly speaking, most of the impulse response functions look intuitive and as dictated by the theory. Not only at the shock's impact, but also in the form of shapes over the course of the next $h = 30$ periods. This holds for the three main variables, GDP, GDP Deflator and FFR, as well as the remaining 19 variables in the large model case. Similar shapes are observed in the literature, such as Banbura et al. (2010), Giannone et al. (2015) and Furman (2014). Particularly, we obtain similar impulse responses as Furman (2014), who similar to our analysis, investigated the adaptive elastic net for a large system of variables[21].

In the following, a qualitative interpretation of the structural impulse responses is given. First, we observe that real activity (real GDP) declines as expected for the large model for almost all estimation methods as response to a monetary tightening. The reason is that an increase in interest rates makes it more expensive for consumers and firms to borrow money whilst encouraging them to save more money as a result of a greater return. Thus, real economic activity dampens and less money is spent on goods and services and less investments are conducted by firms. As we see in figure 17, this explanation aligns with the decline in variables such as real Consumption, real Private Investment, Hours, real Wages, Industrial Production and real Personal Income. On the other hand, unemployment increases, since less production reduces the demand for workers. Second, the inflationary variable, GDP Deflator, displays the Price Puzzle (Christiano et al., 1999) for the VAR systems with small information set (figure 15). The so called Prize Puzzle is an empirical phenomenon stating that a monetary policy shock in form of an increase in interest rates such as the Federal Funds Rate is often followed by a sustained increase in prices. The conjecture is that monetary policy shocks which are associated with substantial prize puzzles are confounded with non-policy disturbances that signal a future increase in prices. In other words, we do not include all relevant variables in our VAR system. The monetary policy shock has an effect on those variables and these in turn lead to an increase of the inflationary variable. This is against our general expectation, that a rise in interest rate is followed by a decrease in Inflation. However, this phenomenon exhibits especially in the small model, for all estimation methods. The larger the model gets, the less severe is the Price Puzzle. Especially for the large variable model, the regularization

[20] Note that, in principle, we could as well include *scheme 2* and *scheme 3*. However, to give the reader a better overview, we focus on *scheme 1* in the structural analysis and simulation.

[21] However, the author did not annualize the data, such that the resulting s-IRF differ by a factor of 4 from ours. Also, the used data set and transformations are slightly different. Still, the impulse responses are alike to ours.

methods counteract the Price Puzzle for GDP Deflator, as well as for CPI. Third, the FFR increases contemporaneously by 1 %, but then flattens after a couple of periods.

Interestingly, we notice that the OLS method performs rather erratic and reacts in an extensive magnitude to the monetary policy shock in the large model, as illustrated in figure 15 - 16. Finding an unambiguous reason is not trivial, since, in general, OLS results in the best in-sample performance. More precisely, structural impulse responses rely on the Moving Average coefficients Φ_i for $i = 1, ..., h$ as well as a cholesky decomposition of the estimated residual covariance matrix $\hat{\Sigma}$, resulting in a non-linear transformation. By construction, OLS results in the smallest residual vector $\hat{u}_t$, meaning that the estimated residual variance for each variable is smaller using OLS compared to penalized methods. As we move to the large model size, the in-sample fit of OLS becomes too close to the true observations, resulting in the previously mentioned problem of overfitting. Since we scale by the standard deviation[22] of the monetary policy instrument, the resulting structural impulse responses are inflated, due to a tremendously small standard deviation. However, by having a look at the forecast error impulse responses (FEIR), we denote that these are, indeed erratic and have large magnitudes compared to the shrinkage methods. It seems, that multiplying the relatively large FEIR by a "small" B_0^{-1}, works against the large magnitude. Finally, the scaling inflates the s-IFRs, as a result of an in-sample overfit. Consequently, since we do not know the true FEIR, it remains unclear, if the relatively[23] large magnitude is indeed a distortion, or rather represents an accurate estimation. Potentially, multicollinearity could partially be a cause that distorts the estimated coefficients[24]. However, the problem remains ambiguous and would require a controlled simulation setup for further investigations.

Furthermore, we want to emphasize that a cross model comparison between the four different model sizes in figure 15 to figure 16 is limited in a structural analysis, since the composition of the system is crucial. It should rather be seen as a detailed illustration of shrinkage methods in four different economy scenarios, compared to OLS. A comparison regarding changing estimation uncertainty in dependence of number of variables is not possible by conducting a single estimation and would, again, require a controlled simulation setup comparing coverage ratios.

In the following, we exclude the OLS method from the structural analysis of the remaining 19 macroeconomic variables and rather focus on the regularization methods[25]. So far, all four regularization methods are able to estimate reasonable impulse responses, even in a high-dimensional model setting. As figure 19 shows, the interest rates such as 3-month Tbill, five year bond rate and ten year bond rate behave as the FFR. Regarding

[22] The used standard deviation is, of course, method specific to enable a comparison, given a certain model size.

[23] In relation to the FEIR functions obtained by regularized methods in the large model size.

[24] Particularly, we denote that the average VIF is 45.63 for the large model, being a sign of severe multicollinearity in the regressors. For comparison, the small model has an average VIF of 3.74, the medium model of 5.52 and the intermediate model of 13.73.

[25] Note that the same described pattern for OLS is also present for the remaining 19 functions in the large model, although not illustrated in this work.

the stock market, the S&P 500 index first falls due to less investments, but then converges to zero after a couple of periods. The money base M1 and M2 decrease at the beginning, which indicates liquidity effects. Interestingly, the contemporaneous response of Oil Price is ambiguous in magnitude as well as sign across the estimation techniques. Still, all corresponding responses converge to zero after a couple of periods. As the figure 15 and 16 show, the impulse responses to a monetary policy shock change in shape as we add more information in terms of additional variables, which underlines the importance of the variable composition for structural analysis. Finally, all shrinkage methods perform, in general, quite similar in a structural analysis. However, the elastic net occasionally reacts in a relatively larger magnitude compared to the remaining methods. Otherwise, we are not able to identify a clear pattern in impulse response functions among the different regularization methods.

5 Simulation

In the previous empirical exercise we demonstrated that the regularized VAR methods can be used to obtain reliable structural impulse response functions. However, the accuracy of the estimated impulse responses can not be evaluated since we do not know the data generating process and thus the true impulse response function. Hence, we are only able to judge the resulting impulse responses by economic reasoning. Accordingly, a quantitative evaluation measurement is desired. Motivated by Giannone et al. (2015), we consequently conduct a controlled Monte Carlo (MC) experiment. We simulate empirical processes for the variables of our medium sized model, which includes the variables GDP, GDP Deflator, Consumption, Investment, Hours, Wage and FFR. This convenient choice of variables enables us reference[26] to the previous empirical structural exercise. Moreover, the enet as well as the fenet face computational constraints for the large model case in a simulation setup. Ideally, the large model case is of main interest. Furthermore, the desired grouping effect of the mixture models is more useful the larger the model size is. However, a Monte Carlo study for $K = 22$ variables is not computationally feasible in our work. The major reason is the detailed, although computationally burdensome double parameter cross validation procedure, adapted to the time series context (chapter 2.3.1). A possible solution is to use information criteria such as AIC or BIC for this purpose and/or waive an adjustment for time dependent data. However, restricting the number of simulated time periods, is another way of limiting the degrees of freedom in the model. Although the advantages of the elastic net and fenet *scheme 1* might not be that pronounced in the medium sized model, we are expecting the shrinkage methods to perform competitively to the OLS, especially in a small data setup. Thus, we simulate data using two different sizes, a small dataset ($T = 80$) as well as a large dataset ($T = 200$). The large size is designed to be in a size range common in the empirical macroeconomics literature for

[26] The same transformation of variables is used. The monetary policy shock is rescaled such that we denote an increase of 100 basis points in the FFR contemporaneously, as in the empirical exercise.

quarterly data. On the contrary, we intend to design the small set as small as possible to sufficiently restraint the degrees of freedom, but not too low such that the estimation gets too inaccurate. By experience, the current choice of $T = 80$ meets our demands[27]. We estimate the response to a monetary policy shock as introduced in the previous chapter. Following Furman (2014), we restrict the analysis to the first $h = 8$ periods after the shock occurs. By doing so, we absent our analysis from the rate at which an estimated response converges to zero. Mainly, we are interested in two questions. First, we are keen to quantify the estimation accuracy of s-IRFs for the regularization methods considered in this work. Second, we want to compare estimation accuracy between an observation rich and a small sample. For each Monte Carlo run, we estimate $\hat{\beta}$ for the respective shrinkage method, as well as for the OLS case. Next, we calculate the squared error, between the estimated as well as the original impulse response function. Then, the average over M MC replications is calculated for period $h = 1, ..., 8$. This enables us to evaluate the bias as well as the deviation of the errors across all M runs. Although the average MSE over all horizons is of limited informativeness, we include it as a broad indicator for estimation accuracy. Each MSE as well as the average MSE are represented relative to the OLS estimation. More precisely, the OLS can be seen as a benchmark model in this setup. The data generating process is described as follows.

1. Estimate $\hat{\beta}^{LS}$ based on the data set from the medium model, as in chapter 4.7.
2. Set all coefficients to zero, for which p-value> 0.85. The result is $\hat{\beta}^{LS}_{sparse}$, containing 20% sparsity[28].
3. Using $\hat{\beta}^{LS}_{sparse}$, generate $M = 40$ VAR(5) processes by drawing from $u_t \sim N(0, \hat{\Sigma})$, where $\hat{\Sigma}$ is the estimated residual covariance matrix[29].
4. Estimate $\Theta_i = \Phi_i B_0^{-1}$ using $\hat{\beta}^{LS}_{sparse}$ and $\hat{\Sigma}$.

Table 7 and 8 contain the simulation results of the present exercise. The tables uncover the following findings. Generally speaking, the shrinkage methods seem to perform well, since the MSE is competitive with OLS in both setups. Furthermore, all four shrinkage methods perform relatively better in relation to OLS when we decrease the number of observations. This finding aligns with the results obtained in Furman (2014). A potential reason is the problem of multicollinearity. Although it is not unpleasantly noticeable in the data rich setup, the consequences of collinear predictors come up in the small data setup, resulting in a less precise impulse response estimation (larger MSE). More precisely, as multicollinearity is more pronounced in the small sample, the variability in $\hat{\beta}$ increases, leading to more fluctuation of estimates across Monte Carlo replications. This in turn translates into higher variability of s-IRFs and finally, in a higher MSE, for all

[27] E.g. for $T = 45$ the OLS impulse responses can not be computed since $T = 40$ observations is not enough to compute the cholesky decomposition of $\hat{\Sigma}$, because the leading minor is not positive definite. However, the remaining methods still are able to estimate impulse responses, although inaccurate.

[28] Percentage of sparsity corresponds to the percentage of zero coefficients in the model.

[29] The estimation of $\hat{\Sigma}$ is based on $\hat{\beta}^{LS}_{sparse}$.

methods of consideration. However, this data problem is more pronounced for OLS, than for the shrinkage methods. This is not surprising, since by shrinking the coefficients, the multicollinearity problem is transmitted into the IRFs to a lesser degree.

In order to quantify the different degrees of multicollinearity among small and large data samples, we calculate the VIF for the simulated data sets of both sample sizes. First, the percentage of variables with VIF > 10 is calculated for every simulated data set. Then we calculate the average percentage across M data sets. As expected, the VIF for the small sample is, in tendency, larger than for the large sample size. For the small sample, we obtain a measurement of 40 % and for the large sample of 17 %, validating our suspicion. The average VIF across the predictors over M data sets is 9.3 for the small sample and 6.4 for the large sample. Although these measurements support our conjecture, they should rather be seen as broad indicators of multicollinearity.

On the other hand, we know that OLS results in smaller residual variances than the shrinkage methods due to a better in-sample fit. In the previous chapter, we elaborated that in a situation of in-sample overfit, we inflate the structural impulse responses by scaling them with a relatively small standard deviation of the monetary policy instrument. In this simulation setup, however, the resulting residual covariance matrix of OLS does not differ critically to the shrinkage methods, as in the previous chapter. Thus, we might suppose that an in-sample overfit is not present[30] here, or at least not noticeably in this type of analysis. The potential reason is that we keep the model size moderate.

Next, a comparison among the methods of interest reveals that the ridge tends to outperform the remaining models for almost all variables in both setups. Again, the multicollinearity argument might play a significant role, since the ridge is known to handle it quite well.

Consequently, the ability of variable selection does not seem to be the crucial criterium for an accurate s-IRF estimation in this setting. A potential reason is that 20% imposed sparsity in the DGP is not sufficient to make the variable selection ability of the lasso, enet and fenet stand out in terms of an accurate structural IRF estimation. This gives rise to our third question of interest: what is the relative performance of the shrinkage methods among each other in terms of relative MSE, as we increase the induced sparsity in the DGP? To answer this question, we generate $M = 40$ data sets using $\hat{\beta}^{LS}_{sparse}$, where all coefficients are set to zero, for which p-value> 0.1 holds, resulting in approximately 80% sparsity. The results are contained in table 9 and 10. As expected, the variable selection methods tend to detect the sparsity in the set of coefficients, shrink the respective coefficients to zero and result in a more accurate estimation of the true impulse response functions (in relative MSE). In tendency, ridge performs relatively worse compared to the previous setting with only 20 % sparsity. Moreover, the lasso and fenet result in noticeably accurate estimations for the high sparsity setting. As expected, the OLS estimation precision deteriorates as we increase the sparsity in the DGP.

[30] Note that, in general, it is not simple to completely identify overfitting, especially to single out a "threshold" after which the model overfits.

Furthermore, denote that no horizon specific MSE of any shrinkage method is consistently lower than using OLS, which indicates a sporadic pattern. It seems, that this observation is DGP dependent and can not be explained by multicollinearity. By shrinking the coefficients, it might occur that "decisive coefficients" are shrunk exactly to zero, which is dependent on the simulated data in each replication. Consequently, the estimated structural impulse response might be distorted, leading to low performance in terms of MSE for certain horizons. In addition, even shrinkage towards zero might lead to such a distortion. Finally, the sporadic behavior does not clearly follow a pattern, but rather is dependent on variable, sample size and imposed percentage of sparsity on the DGP.

Additional[31] analyses of the MSE for large horizons show, that as the forecast horizon gets large, the period specific MSE based on the shrinkage methods converges faster to zero than for OLS. This is especially noticeable in both small sample setups. The reason is that as the horizon increases the s-IRFs based on shrinkage methods converge faster to 0, being closer to the true s-IRFs. This relatively fast converging behavior, seems to be a natural consequence of shrinking the coefficients in this setting. Whereas OLS still moves in relatively larger magnitudes, being more distant to the true s-IRF.

In summary, the simulation shows that the regularization methods considered in this work are able to perform an accurate structural impulse response analysis in a medium sized variable setting, using the relative MSE as a measure. In addition, ridge performs noteworthy well in the 20 % sparsity case. However, variable selection methods rather excel in a setting containing 80% sparsity, particularly lasso and fenet *scheme 1*. Finally, the relative strength of the shrinkage methods is especially visible in a small data case, due to multicollinearity.

6 Conclusion

On the methodological side, we extend the novel feature weighted elastic net to the VAR model. With the intention to improve forecast precision, we adapt the double parameter cross validation procedure to time series data. Motivated by the Bayesian literature, we propose three different grouping schemes on the penalization term of the fenet. In principle, we either group the coefficients by lags or by variables. The different schemes are empirically compared to other regularization methods in a forecast exercise. In the univariate forecast exercise, we are partially able to improve forecast precision by using a high-dimensional model size. The fenet VAR *scheme 3* performs especially well across different horizon and model size combinations. However, the dominance of the mixture models does not seem to be as pronounced for the multivariate forecast. The fenet *scheme 3* still gives accurate predictions, however, the remaining mixture models are less precise than the lasso and ridge for the intermediate and large model setup. Consequently, it remains ambiguous which regularization method performs best universally. Nevertheless, we are able to obtain more precise predictions than the approximate DFM for the ma-

[31] Due to the limited scope of this work, we do not include corresponding graphical illustrations.

jority of forecasts. As expected, OLS results in tremendously imprecise predictions, due to overfitting. However, all shrinkage methods handle this problem well. Throughout the forecasting exercise, we investigate shrinkage behavior and the tuning parameter α. Particularly, we denote that the variable selection methods shrink relatively more coefficients towards zero as we increase mode size. In the large model setup, the mixture models tend to select α on the boundary, resulting in the ridge regression relatively often. We empirically show that under potential multicollinearity and low signal predictors, the lasso shows uncommon selection behavior.

In a structural analysis we demonstrate that the shrinkage methods of consideration result in (mostly) credible structural impulse responses to a monetary policy shock. We are able to diminish the Prize Puzzle by estimating a high-dimensional model. Next, we find that OLS results in inflated structural impulse responses as we move to higher dimensions, due to in-sample overfitting. As a natural consequence of shrinkage, the impulse responses of regularized methods do not suffer severely from it. Our controlled Monte Carlo experiment simulates structural impulse responses in a moderate model size for a small and a large data setup. We are able to illustrate, that estimation precision (measured by MSE) is more pronounced for shrinkage methods relative to OLS as we move to the small data setup, due to a higher degree of collinearity. Finally, the ridge performs most accurately in a low degree of sparsity setting. Contrary, the lasso and fenet *scheme 1* perform exceptionally well in a model setup with a high degree of sparsity.

However, there is still room for future research in the context of applied macroeconomics: One major area of future research is inference for structural analysis. Due to the limited scope of this work, we did not extensively analyze estimation uncertainty for structural impulse responses. It is of potential interest, if a certain regularization method is accompanied by less variation in structural impulse response estimation given a specific set of variables. Moreover, it is of interest, if estimation uncertainty changes in dependence of model size, raising the previously mentioned problem of multicollinearity. Both research questions could be addressed by comparing actual coverage ratios across methods as well as model sizes. Correspondingly, confidence intervals need to be bootstrapped. However, bootstrap techniques regarding structural impulse responses are a vibrant research area with many nuances, which can get computationally intensive for regularized methods and have theoretical limitations.

Furthermore, our high-dimensional model is represented by 110 unknown coefficients with $K = 22$ variables. This setup is quite common in the literature. However, it would be of interest to push further into higher model sizes. Particularly, we did not make use of the elastic net property being able to select $pK > T$ coefficients. Given the lag order and number of observations of this work, we would require more than 42 variables, to investigate this characteristic. Moreover, we would be interested if the effect of imposing structure on penalization gets even more pronounced as we move to higher dimensions. Several researchers such as Banbura et al. (2010) or Bloor and Matheson (2010) have

already been investigating higher dimensional VAR models with more than 100 macroeconomic variables, for the U.S. and New Zealand, respectively. In principle, this model size is feasible for the enet VAR as well as the fenet VAR. Subsequently, moving on to higher dimensions is of methodological as well as empirical interest.

7 References

[1] Banbura, M., Giannone, D., Reichlin, L. (2010). Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1), 71-92.

[2] Bergmeir, C., Hyndman, R.J., Koo, B., 2018. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics Data Analysis* 120, 7083.

[3] Bloor, C., Matheson, T. (2010). Analysing shock transmission in a data-rich environment: A large BVAR for New Zealand. *Empirical Economics*, 39(2), 537-558.

[4] Brüggemann, R., Lütkepohl, H. (2001). Lag selection in subset VAR models with an application to a US monetary system. *Econometric Studies*, 107-128.

[5] Christiano, L., Eichenbaum, M., Evans, C. (1998). Monetary Policy Shocks: What Have We Learned and to What End? *National Bureau of Economic Research.*

[6] Christoffersen, P. F., Diebold, F. X. (1998). Cointegration and long-horizon forecasting. *Journal of Business Economic Statistics*, 16(4), 450-456.

[7] 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.

[8] Friedman, J., Hastie, T., Höfling, H., Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2), 302-332.

[9] Friedman, J., Hastie, T., Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1).

[10] Furman, Y. (2014). VAR Estimation with the Adaptive Elastic Net. *SSRN Electronic Journal.*

[11] Giannone, D., Lenza, M., Primiceri, G. E. (2015). Prior Selection for Vector Autoregressions. *Review of Economics and Statistics*, 97(2), 436-451.

[12] Hansen, P. R., Lunde, A., Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2), 453-497.

[13] Hoerl, A. E., Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.

[14] Hsu, N.-J., Hung, H.-L., Chang, Y.-M. (2008). Subset selection for vector autoregressive processes using Lasso. *Computational Statistics Data Analysis*, 52(7), 3645-3657.

[15] James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning (Vol. 103). *Springer New York.*

[16] Kascha, C., Trenkler, C. (2015). Forecasting VARs, Model Selection, and Shrinkage. *Working Paper*.

[17] Koop, G. (2017). Bayesian Methods for Empirical Macroeconomics with Big Data. *Review of Economic Analysis*, 9, 33-56.

[18] Koop, G., Korobilis, D. (2013). Large time-varying parameter VARs. *Journal of Econometrics*, 177(2), 185-198.

[19] Litterman, R. (1986). Forecasting with Bayesian Vector Autoregressions: Five Years of Experience. *Journal of Business Economic Statistics*, 4(1), 2538.

[20] Lütkepohl, H. (2005). New introduction to multiple time series analysis. *Springer New York*.

[21] Nicholson, W. B., Matteson, D. S., Bien, J. (n.d.). Structured Regularization for Large Vector Autoregression. *Cornell University*.

[22] Patton, A., Politis, D. N., and White, H. (2009). Correction to Automatic block-length selection for the dependent bootstrap. *Econometric Reviews*, 28(4), 372375.

[23] Sims, C. (1980). Macroeconomics and Reality. *Econometrica*, 48, 148.

[24] Song, S., Bickel, P. J. (n.d.). Large Vector Auto Regressions. *ArXiv:1106.3915v1.*

[25] Stock, J. H., Watson, M. W. (2002). Forecasting Using Principal Components From a Large Number of Predictors. *Journal of the American Statistical Association*, 97(460), 1167-1179.

[26] Stock, J. H., Watson, M. W. (2012). Generalized Shrinkage Methods for Forecasting Using Many Predictors. *Journal of Business Economic Statistics*, 30(4), 481-493.

[27] Tay, J. K., Aghaeepour, N., Hastie, T., Tibshirani, R. (n.d.). Feature-weighted elastic net: Using features of features for better prediction. *Working Paper*.

[28] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288.

[29] Tibshirani, R., Suo, X. (2016). An Ordered Lasso and Sparse Time-Lagged Regression. *Technometrics*, 58(4), 415-423.

[30] Wooldridge, Jeffrey M., (2013). Introductory econometrics: a modern approach. Mason, Ohio, *South-Western Cengage Learning*.

[31] Hastie, T., Tibshirani, R., Wainwright, M. (2015). Statistical learning with sparsity: the lasso and generalizations. *CRC press*.

[32] Hsu, N., Hung, H., and Chang, Y. (2008). Subset Selection for Vector Autoregressive Processes using Lasso. *Computational Statistics and Data Analysis*, 52, 36453657.

[33] Yuan, M. and Lin, Y. (2006). Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society*, Series B, 68(1), 4967.

[34] Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101, 14181429.

[35] 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.

8 Appendix

8.1 On Ridge and Lasso

Definition 1. *The ridge estimator is defined as*

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N}(y_i - \sum_{j=0}^{p}\beta_j x_{ij})^2, \ s.t. \ ||\beta||_2^2 \leq t$$

, where $||\beta||_2^2 = \sum_{j=1}^{p}\beta_j^2$ *is the squared Euclidean distance. Using the Lagrange multiplier we rewrite definition 1 as*

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \ L(\lambda, \beta) = \underset{\beta}{\operatorname{argmin}} \ RSS + \lambda \sum_{j=1}^{p}\beta_j^2$$

, where we define $\sum_{i=1}^{N}(y_i - \sum_{j=0}^{p}\beta_j x_{ij})^2$ *as the Residual Sum of Squares (RSS).*

Definition 2. *The lasso is defined as*

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N}(y_i - \sum_{j=0}^{p}\beta_j x_{ij})^2, \ s.t. \ ||\beta||_1 \leq t$$

, where $||\beta||_1 = \sum_{j=1}^{p}|\beta_j|$ *is the* L_1 *norm. We rewrite definition 2 as*

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \ L(\lambda, \beta) = \underset{\beta}{\operatorname{argmin}} \ RSS + \lambda \sum_{j=1}^{p}|\beta_j|$$

8.2 Equivalence between Lasso and Elastic Net

Here, we want to give a lemma, which states an important property of the elastic net optimization problem. Namely, the equivalent to the lasso problem on an augmented data set.

Lemma 1. *Define the augmented data set* $X^* = \frac{1}{\sqrt{1+\lambda_2}}\begin{pmatrix} X \\ \sqrt{\lambda_2} I \end{pmatrix}$, $y^* = \begin{pmatrix} y \\ 0 \end{pmatrix}$ *and let* $\gamma = \frac{\lambda_1}{\sqrt{1+\lambda_2}}$, $\beta^* = \sqrt{1+\lambda_2}\beta$, *where* X^* *has dimension* $(N+p) \times p$ *and* y *is a column vector of length* $N+p$. *Then, the elastic net optimization problem is equivalent to* $L(\gamma, \beta^*) = ||y^* - X^*\beta^*||_2^2 + \gamma||\beta^*||_1$. *The solution is* $\hat{\beta}^* = \operatorname{argmin}_{\beta^*} L(\gamma, \beta^*)$ *and we obtain the naive elastic net estimator as* $\hat{\beta} = \frac{1}{\sqrt{1+\lambda_2}}\hat{\beta}^*$.

The corresponding proof can be found in Zou and Hastie (2005). Since the artificial design matrix X^* contains $N+p$ rows and p columns, X^* has rank p, which illustrates an important property of the elastic net estimator. It avoids a limitation of the lasso since $p >> N$ features of the data can potentially be selected. Furthermore, Lemma 1 implies that the elastic net is also capable of automatically selecting variables as the lasso does.

8.3 On Grouping Effect

For this derivation we are using the Lagrange specification $L(\beta)_{\lambda_1,\lambda_2}$ of the elastic net, introduced in equation 1. For simplicity, matrix notation is used such that Y is $(N \times 1)$, $X = [X_1, ..., X_p]$ is a $(N \times p)$ design matrix and X_j is $(N \times 1)$ for $j = 1, ..., p$. Since $\hat{\beta}$ is the minimizer of $L(\beta)_{\lambda_1,\lambda_2}$, it satisfies

$$\frac{\delta}{\delta\beta_k} L(\beta)_{\lambda_1,\lambda_2}|_{\beta=\hat{\beta}(\lambda_1,\lambda_2)} = 0, \text{ if } \hat{\beta}_k(\lambda_1, \lambda_2) \neq 0. \tag{12}$$

Taking the derivative w.r.t. β_i gives

$$-2X_i^{'}(Y - X\hat{\beta}(\lambda_1, \lambda_2)) + \lambda_1 \text{ sign}(\hat{\beta}_i(\lambda_1, \lambda_2)) + 2\lambda_2\hat{\beta}_i(\lambda_1, \lambda_2) = 0 \tag{13}$$

and w.r.t. to β_j gives

$$-2X_j^{'}(Y - X\hat{\beta}(\lambda_1, \lambda_2)) + \lambda_1 \text{ sign}(\hat{\beta}_j(\lambda_1, \lambda_2)) + 2\lambda_2\hat{\beta}_j(\lambda_1, \lambda_2) = 0. \tag{14}$$

Next we subtract equation 13 from equation 14, which after simple algebra results in

$$\hat{\beta}_i(\lambda_1, \lambda_2) - \hat{\beta}_j(\lambda_1, \lambda_2) = \frac{1}{\lambda_2}\hat{r}(\lambda_1, \lambda_2)(X_i^{'} - X_j^{'}) \tag{15}$$

, where we define $\hat{r}(\lambda_1, \lambda_2) = Y - X\hat{\beta}(\lambda_1, \lambda_2)$.
Then, 12 implies that $L(\hat{\beta}(\lambda_1, \lambda_2))_{\lambda_1,\lambda_2}$ can not be larger than any other $L(\beta^*(\lambda_1, \lambda_2))_{\lambda_1,\lambda_2}$, particularly

$$\begin{aligned} & L(\hat{\beta}(\lambda_1, \lambda_2))_{\lambda_1,\lambda_2} \leq L(\beta = 0)_{\lambda_1,\lambda_2} \\ \Leftrightarrow\ & ||\hat{r}(\lambda_1, \lambda_2)||_2^2 + \lambda_2||\hat{\beta}(\lambda_1, \lambda_2)||_2^2 + \lambda_1||\hat{\beta}(\lambda_1, \lambda_2)||_1 \leq ||Y||_2^2 \end{aligned}$$

, which implies that $||\hat{r}(\lambda_1, \lambda_2)||_2^2 \leq ||Y||_2^2$ and hence $||\hat{r}(\lambda_1, \lambda_2)||_2 \leq ||Y||_2$. Next, we take the absolute value of 15 and by using the Cauchy-Schwarz Inequality we obtain

$$|\hat{\beta}_i(\lambda_1, \lambda_2) - \hat{\beta}_j(\lambda_1, \lambda_2)| \leq \frac{1}{\lambda_2}|\hat{r}(\lambda_1, \lambda_2)||X_i^{'} - X_j^{'}|$$

Finally, we divide by $||Y||_2$ and write the result as

$$\begin{aligned} D_{\lambda_1,\lambda_2}(i, j) = \frac{1}{||Y||_2}|\hat{\beta}_i(\lambda_1, \lambda_2) - \hat{\beta}_j(\lambda_1, \lambda_2)| &\leq \frac{1}{\lambda_2}\frac{|\hat{r}(\lambda_1, \lambda_2)|}{||Y||_2}|X_i^{'} - X_j^{'}| \\ &\leq \frac{1}{\lambda_2}|X_i^{'} - X_j^{'}| \\ &= \frac{1}{\lambda_2}\sqrt{2(1 - \rho)} \end{aligned}$$

, where we use the result that $\frac{|\hat{r}(\lambda_1,\lambda_2)|}{||Y||_2} \leq 1$ and $(X_i^{'} - X_j^{'})^2 = 2(1 - \rho)$ since X is standardized. Finally, the desired upper bound on $D_{\lambda_1,\lambda_2}(i, j)$ is $\frac{1}{\lambda_2}\sqrt{2(1 - \rho)}$.

8.4 Elastic Net Soft-Thresholding Update Form

We are interested in deriving the elastic net soft-thresholding update form used in algorithm 1. We use Friedman et al. (2010) for orientation. First, the derivative of RSS is calculated:

$$\begin{aligned}
\frac{\delta}{\delta\beta_j} RSS(\beta)|_{\beta=\hat{\beta}} &= -\sum_{i=1}^{N} x_{ij}(y_i - \sum_{j=0}^{p} \hat{\beta}_j x_{ij}) \\
&= -\sum_{i=1}^{N} x_{ij}(y_i - \sum_{k\neq j}^{p} \hat{\beta}_k x_{ik} - \hat{\beta}_j x_{ij}) \\
&= -\sum_{i=1}^{N} x_{ij}(y_i - \sum_{k\neq j}^{p} \hat{\beta}_k x_{ik}) + \hat{\beta}_j \sum_{i=1}^{N} x_{ij}^2 \\
&= -p_j + \hat{\beta}_j z_j = -p_j + \hat{\beta}_j
\end{aligned}$$

, where $z_j = \sum_{i=1}^{N} x_{ij}^2 = 1$ since we assume that the data is normalized. Second, we write the L_1 term as

$$\lambda\alpha \sum_{j=1}^{p} |\beta_j| = \lambda\alpha|\beta_j| + \lambda\alpha \sum_{k\neq j}^{p} |\beta_k|$$

, where the derivative is

$$\frac{\delta}{\delta\beta_j} \lambda\alpha \sum_{j=1}^{p} |\beta_j| = \begin{cases} -\lambda\alpha & \text{if } \hat{\beta}_j < 0 \\ [-\lambda\alpha, \lambda\alpha] & \text{if } \hat{\beta}_j = 0 \\ \lambda\alpha & \text{if } \hat{\beta}_j > 0 \end{cases}$$

Now, we calculate the derivative of the L_2 term as

$$\frac{\delta}{\delta\beta_j} \frac{1}{2} \lambda(1-\alpha) \sum_{j=1}^{p} \beta_j^2 = \hat{\beta}_j \lambda(1-\alpha)$$

Using properties of subdifferential theory, particularly Moreau-Rockafellar theorem results in

$$\frac{\delta}{\delta\beta_j} L(\lambda, \alpha, \beta)|_{\beta=\hat{\beta}} = \frac{\delta}{\delta\beta_j} \frac{1}{2} RSS + \frac{\delta}{\delta\beta_j} \lambda\alpha ||\beta||_1 + \frac{\delta}{\delta\beta_j} \frac{1}{2} \lambda(1-\alpha) ||\beta||_2^2 = 0$$

$$0 = \begin{cases} -p_j + \hat{\beta}_j - \lambda\alpha + \hat{\beta}_j \lambda(1-\alpha) & \text{if } \hat{\beta}_j < 0 \\ [-p_j - \lambda\alpha, -p_j + \lambda\alpha] & \text{if } \hat{\beta}_j = 0 \\ -p_j + \hat{\beta}_j + \lambda\alpha + \hat{\beta}_j \lambda(1-\alpha) & \text{if } \hat{\beta}_j > 0 \end{cases}$$

For the second case distinction it follows that since $0 \in [-p_j - \lambda\alpha, -p_j + \lambda\alpha]$ that $-\lambda\alpha \leq p_j \leq \lambda\alpha$. Finally, solving for the estimator results in

$$\hat{\beta}_j^{enet} = \begin{cases} \frac{p_j + \lambda\alpha}{1+(1-\alpha)\lambda} & \text{if } p_j < -\lambda\alpha \\ 0 & \text{if } -\lambda\alpha \leq p_j \leq \lambda\alpha \\ \frac{p_j - \lambda\alpha}{1+(1-\alpha)\lambda} & \text{if } p_j > \lambda\alpha \end{cases}$$

Furthermore, we make use of the threshold operator $S(...)$, which results in the desired soft thresholding update form used in step 3 of algorithm 1:

$$\begin{aligned} \hat{\beta}_j^{enet} &= \frac{1}{1+\lambda(1-\alpha)} S(p_j, \lambda\alpha) \\ &= \frac{1}{1+\lambda(1-\alpha)} S(\sum_{i=1}^{N} x_{ij}(y_i - \sum_{k \neq j}^{p} \hat{\beta}_k x_{ik}), \lambda\alpha) \\ &= \frac{1}{1+\lambda(1-\alpha)} S(\sum_{i=1}^{N} x_{ij}(y_i - \tilde{y}_i^j), \lambda\alpha) \\ &= \frac{1}{1+\lambda(1-\alpha)} S(\sum_{i=1}^{N} x_{ij} r_{ij}, \lambda\alpha). \end{aligned}$$

8.5 Plots and Tables

Table 3: Data series description

Variables	Mnemonic	Small VAR	Medium VAR	Intermediate VAR	Large VAR	Trans.	s/f
Real GDP	GDPC96	×	×	×	×	5	s
GDP Deflator	GDPDEF	×	×	×	×	5	s
FFR	FEDFUNDS	×	×	×	×	2	r
R. Consumption	PCECC96		×	×	×	5	s
R. private Inv.	GPDIC96		×	×	×	5	s
Hours	HOANBS		×	×	×	5	s
Real wage	COMPRNFB		×	×	×	5	s
CPI	CPIAUCSL			×	×	6	s
3-Month Tbill	TB3MS			×	×	2	f
Five year bond rate	GS5			×	×	2	f
Ten year bond rate	GS10			×	×	2	f
M2	M2SL			×	×	6	f
R. Personal Income	RPI			×	×	5	s
Industrial Production	INDPRO			×	×	5	s
Unemployment Rate	UNRATE			×	×	2	s
Producer Price Index	PPIFCG				×	5	s
PCE Price Index	PCECTPI				×	6	s
Average hourly earnings	CES3000000008				×	6	s
M1	M1SL				×	6	f
Oil price	OILPRICE				×	5	f
Real Gov. Cons. and Inv.	GCEC96				×	5	s
S&P 500 Index	SP500				×	5	f

Data is downloaded from the FRED database. Variables observed on a monthly frequency were averaged over the respective three months to obtain the quarterly observation. Let $z_{i,t}$ be the untransformed data point and $x_{i,t}$ the transformed observation. Then the transformation codes are defined as (Koob and Korobilis, 2013): 2 - first difference, $x_{i,t} = z_{i,t} - z_{i,t-1}$; 3 - second difference $x_{i,t} = z_{i,t} - z_{i,t-2}$; 4 - logarithm $x_{i,t} = 400(logz_{i,t})$; 5 - first difference of logarithm, $x_{i,t} = 400(logz_{i,t} - logz_{i,t-1})$; 6 - second difference of logarithm, $x_{i,t} = 400(logz_{i,t} - logz_{i,t-2})$. s/f corresponds to slow/fast moving variable classification.

Table 4: Relative MSFE for GDP of an Expanding Window Forecast

Setup	GDP							
	$h = 1$							
	OLS	Ridge	Lasso	Enet	Fenet	Fenet 2	Fenet 3	DFM
K=3	**0.8692***	0.9380*	0.9120*	0.9472*	0.9449*	0.9481*	0.9548*	-
K=7	0.7802*	0.7630*	0.6913*	0.7285*	**0.6732***	0.6816*	0.7200*	-
K=15	1.7201	0.8295*	0.8043*	0.6770*	0.6518*	**0.6498***	0.6772*	-
K=22	4.0236	0.7605	0.7226*	0.6732*	**0.6529***	0.6549*	0.6792*	0.7144*
	$h = 4$							
	OLS	Ridge	Lasso	Enet	Fenet	Fenet 2	Fenet 3	DFM
K=3	**1.0702***	1.1582	1.1665	1.1269	1.1221	1.1271	1.1239	-
K=7	1.1676*	1.1254*	1.1226*	1.1102*	**1.0781***	1.1050*	1.0959*	-
K=15	2.3967	1.2776	1.1947*	1.1462*	**1.1175***	1.1197*	1.1524*	-
K=22	4.7577	1.2239	1.1849	1.1182	1.0539*	**1.0375***	1.1578	1.0507*
	$h = 8$							
	OLS	Ridge	Lasso	Enet	Fenet	Fenet 2	Fenet 3	DFM
K=3	1.0787*	1.1187*	1.1348*	1.0678*	**1.0659***	1.0660*	1.0674*	-
K=7	**1.0377***	1.0932	1.0889	1.0745	1.0619	1.0694	1.0700	-
K=15	1.5560	1.1469	1.1649	1.0427	1.0504	**1.0383***	1.0656	-
K=22	7.1557	1.1838	1.1076	1.0475*	1.0371*	**1.0365***	1.0404*	1.2367

Relative MSFE for OLS, Ridge, Lasso, Enet, Fenet *scheme 1*, Fenet *scheme 2*, Fenet *scheme 3* and the Dynamic Factor Model with three factors. All setups include p=5 lags. The MSFE is illustrated in relation to the conditional mean benchmark. Entries containing "*" are part of the MCS with $\alpha = 0.15$. **Bold** entries have the lowest $MSFE_1^R$ for the particular model size and horizon combination.

Table 5: Relative MSFE for GDP Deflator of an Expanding Window Forecast

Setup	GDP Deflator							
	$h = 1$							
	OLS	Ridge	Lasso	Enet	Fenet	Fenet 2	Fenet 3	DFM
K=3	0.1719	0.1508*	0.1467*	0.1495*	0.1498*	0.1497*	**0.1465***	-
K=7	0.2321	0.1764	0.1514*	0.1514*	0.1510*	0.1512*	**0.1497***	-
K=15	0.3886	0.1867*	0.1527*	**0.1414***	0.1588	0.1700*	0.1457*	-
K=22	0.7542	0.1721*	**0.1561***	0.1610*	0.1711*	0.1768	0.1612*	0.2972
	$h = 4$							
	OLS	Ridge	Lasso	Enet	Fenet	Fenet 2	Fenet 3	DFM
K=3	0.4031*	0.2699*	0.2824*	0.2913*	0.2984*	0.3001*	**0.2630***	-
K=7	0.5059	0.3166*	0.2884*	0.2871*	0.2869*	0.2891*	**0.2760***	-
K=15	0.6553	**0.3876***	0.3876*	0.5016	0.6538	0.7033	0.4197*	-
K=22	0.9334	**0.2933***	0.3460*	0.4382	0.5137	0.5502	0.3759*	0.3535*
	$h = 8$							
	OLS	Ridge	Lasso	Enet	Fenet	Fenet 2	Fenet 3	DFM
K=3	0.7983*	0.5200*	**0.4961***	0.5282*	0.5379*	0.5413*	0.4858*	-
K=7	1.4048	0.6196	0.5292*	0.5295*	0.5331*	0.5398*	**0.5064***	-
K=15	1.8802	**0.5572***	0.6428*	0.8292*	0.9359	0.9536*	0.7053*	-
K=22	2.3479	**0.3782***	0.5064*	0.6953	0.7949	0.8047	0.6113	0.3992*

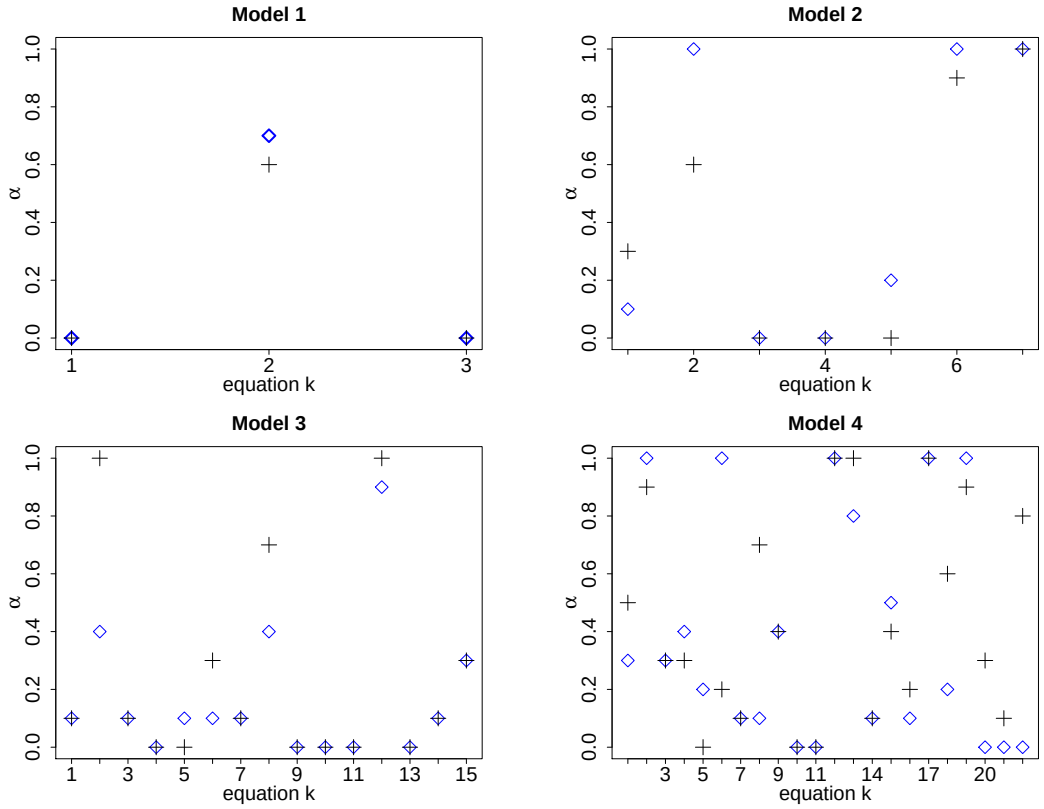
Relative MSFE for OLS, Ridge, Lasso, Enet, Fenet *scheme 1*, Fenet *scheme 2*, Fenet *scheme 3* and the Dynamic Factor Model with three factors. All setups include p=5 lags. The MSFE is illustrated in relation to the conditional mean benchmark. Entries containing "*" are part of the MCS with $\alpha = 0.15$. **Bold** entries have the lowest $MSFE_2^R$ for the particular model size and horizon combination.

Table 6: Relative MSFE for FFR of an Expanding Window Forecast

Setup	FFR							
				$h = 1$				
	OLS	Ridge	Lasso	Enet	Fenet	Fenet 2	Fenet 3	DFM
K=3	1.8614*	**0.9510***	1.5429*	0.9690*	0.9752*	0.9752*	0.9752*	-
K=7	2.9887	1.1354	1.3859	1.0571	1.0556*	1.0563*	**1.0542***	-
K=15	7.3961	1.1658	**1.1420**	1.5586	1.5551	1.5521*	1.5611	-
K=22	7.5802	1.4301	0.9685	0.8393*	0.8412*	0.8451*	0.8220*	**0.7505***
				$h = 4$				
	OLS	Ridge	Lasso	Enet	Fenet	Fenet 2	Fenet 3	DFM
K=3	1.2085	1.0346*	1.0430	0.9998*	1.0000*	1.0000*	**0.9996***	-
K=7	2.3692	**1.2168**	1.2673	1.1477*	1.1461*	1.1484*	1.1439*	-
K=15	4.8395	**1.1827***	1.2865	1.3177*	1.3752*	1.3646*	1.2711*	-
K=22	11.3453	1.3482	1.1675	1.0331	1.0306	1.0334	1.0057*	**1.0032***
				$h = 8$				
	OLS	Ridge	Lasso	Enet	Fenet	Fenet 2	Fenet 3	DFM
K=3	1.0358*	1.0169	**1.0012***	1.0045*	1.0039*	1.0038*	1.0038*	-
K=7	1.4652	1.0221*	**1.0029***	1.0125*	1.0130*	1.0168*	1.0100*	-
K=15	3.6081	1.0591*	**1.0343***	1.0804*	1.1128*	1.0969*	1.0553*	-
K=22	12.6603	1.0508*	1.1014*	1.0526*	1.0370*	1.0353*	**1.0159***	1.0644*

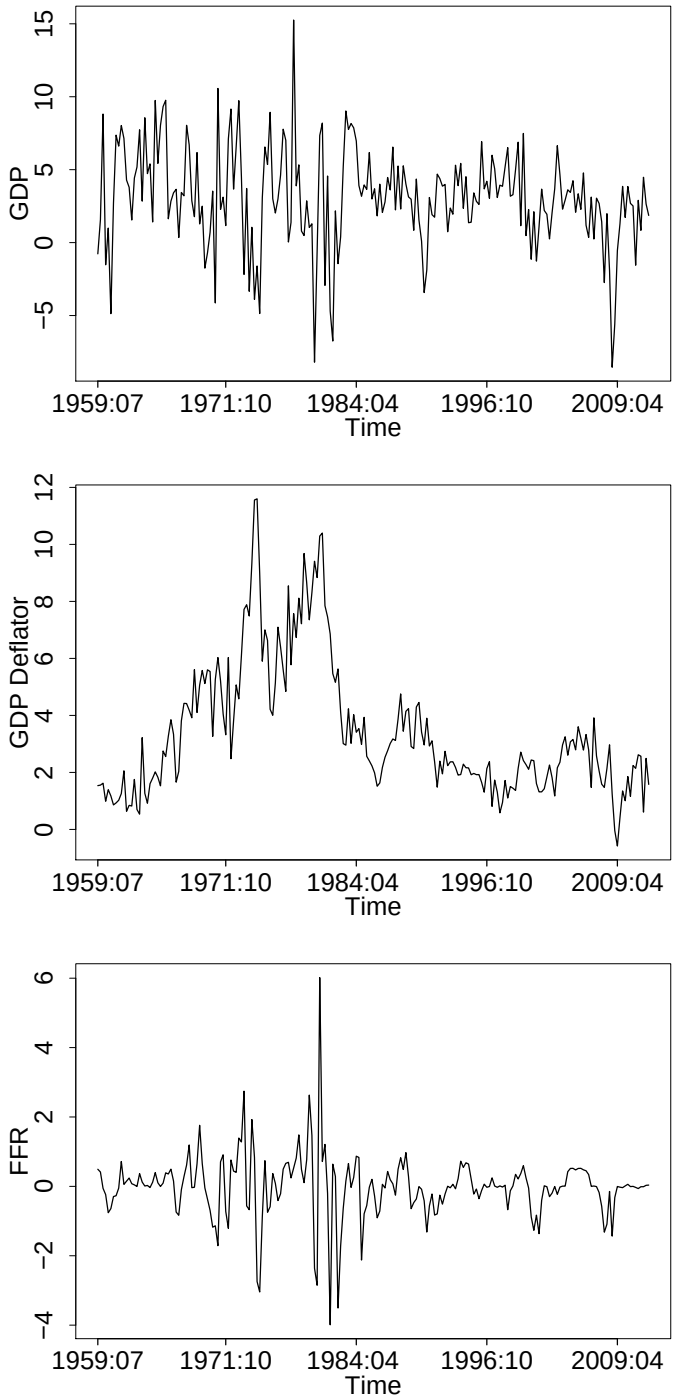
Relative MSFE for OLS, Ridge, Lasso, Enet, Fenet *scheme 1*, Fenet *scheme 2*, Fenet *scheme 3* and the Dynamic Factor Model with three factors. All setups include p=5 lags. The MSFE is illustrated in relation to the conditional mean benchmark. Entries containing "*" are part of the MCS with $\alpha = 0.15$. **Bold** entries have the lowest $MSFE_3^R$ for the particular model size and horizon combination.

Figure 11: Tuning parameter α for $T = 149$



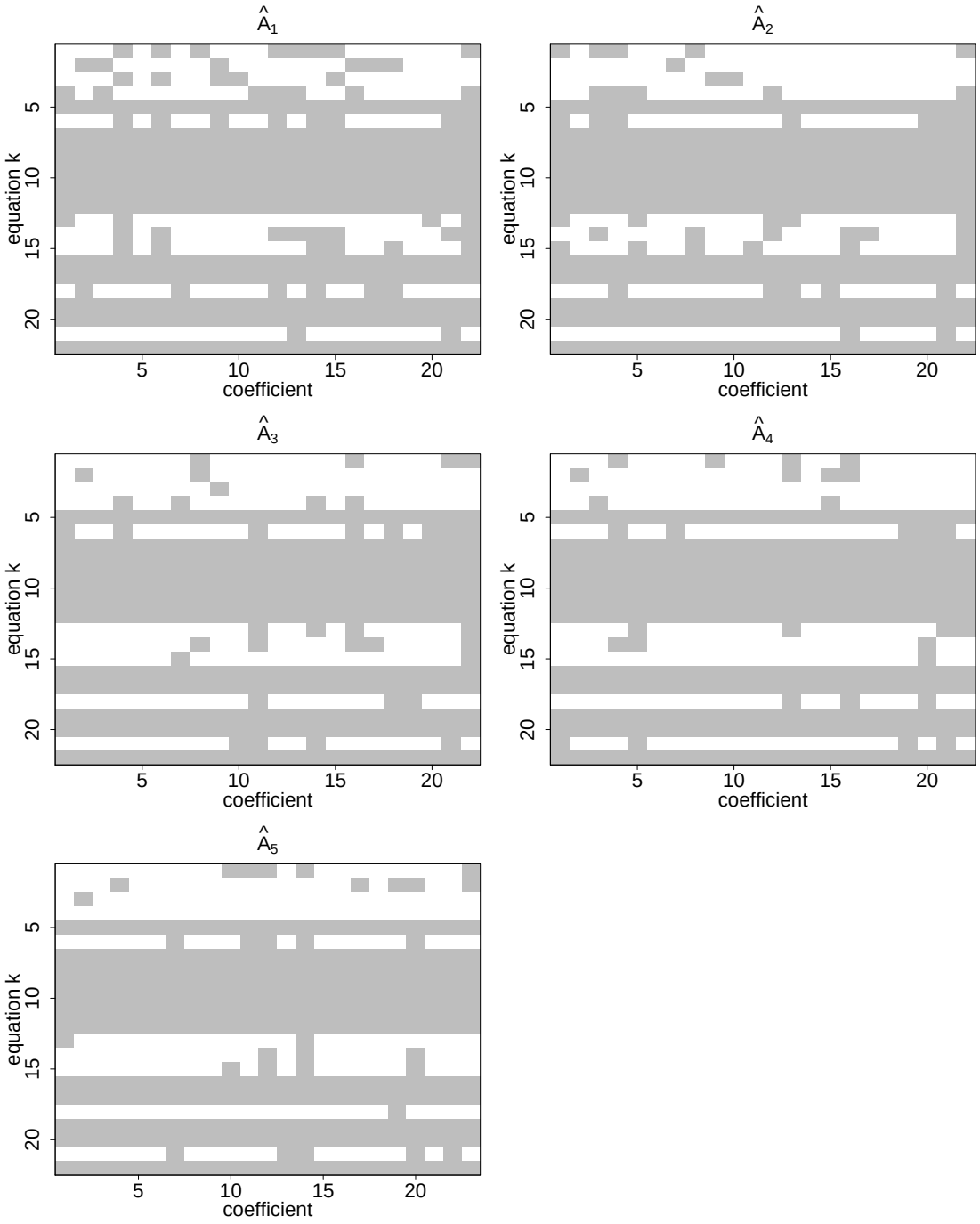
+ represents Fenet *scheme 1*; ◇ represents Enet. Model 1 - Model 4 correspond to small, medium, intermediate and large model, respectively.

Figure 12: Variables of interest: GDP, GDP Deflator and FFR



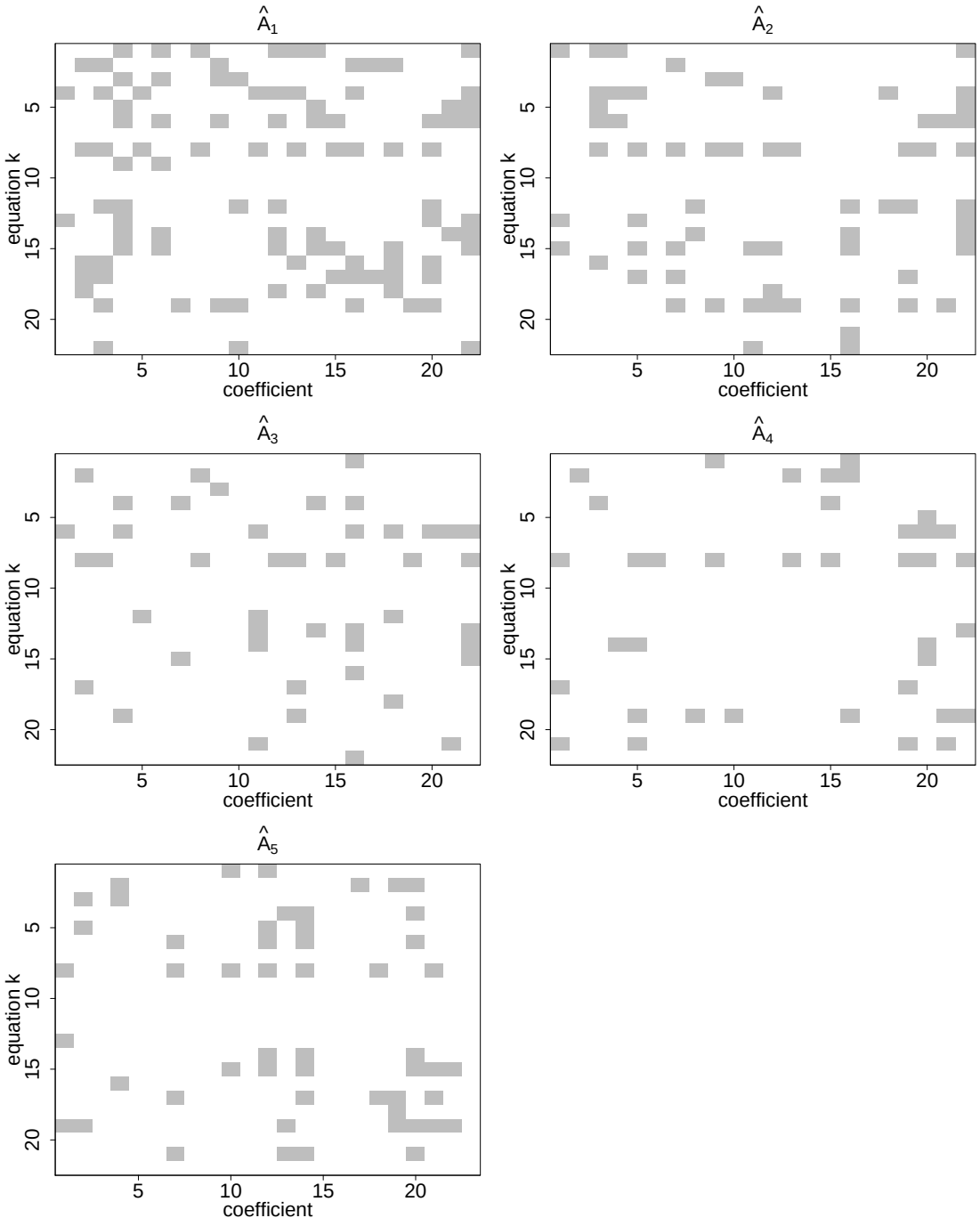
Three variables of main interest. Period of observation is 1959:07 to 2012:04. All three variables are transformed. See table 3 for the transformation codes. GDP is interpreted as annualized growth rate in %. GDP Deflator is transformed s.t. plot interpretation is of an annualized logarithmic Inflation in %. FFR is transformed s.t. the change to the previous period is displayed.

Figure 13: Binary Coefficient Matrices for Enet VAR(5)



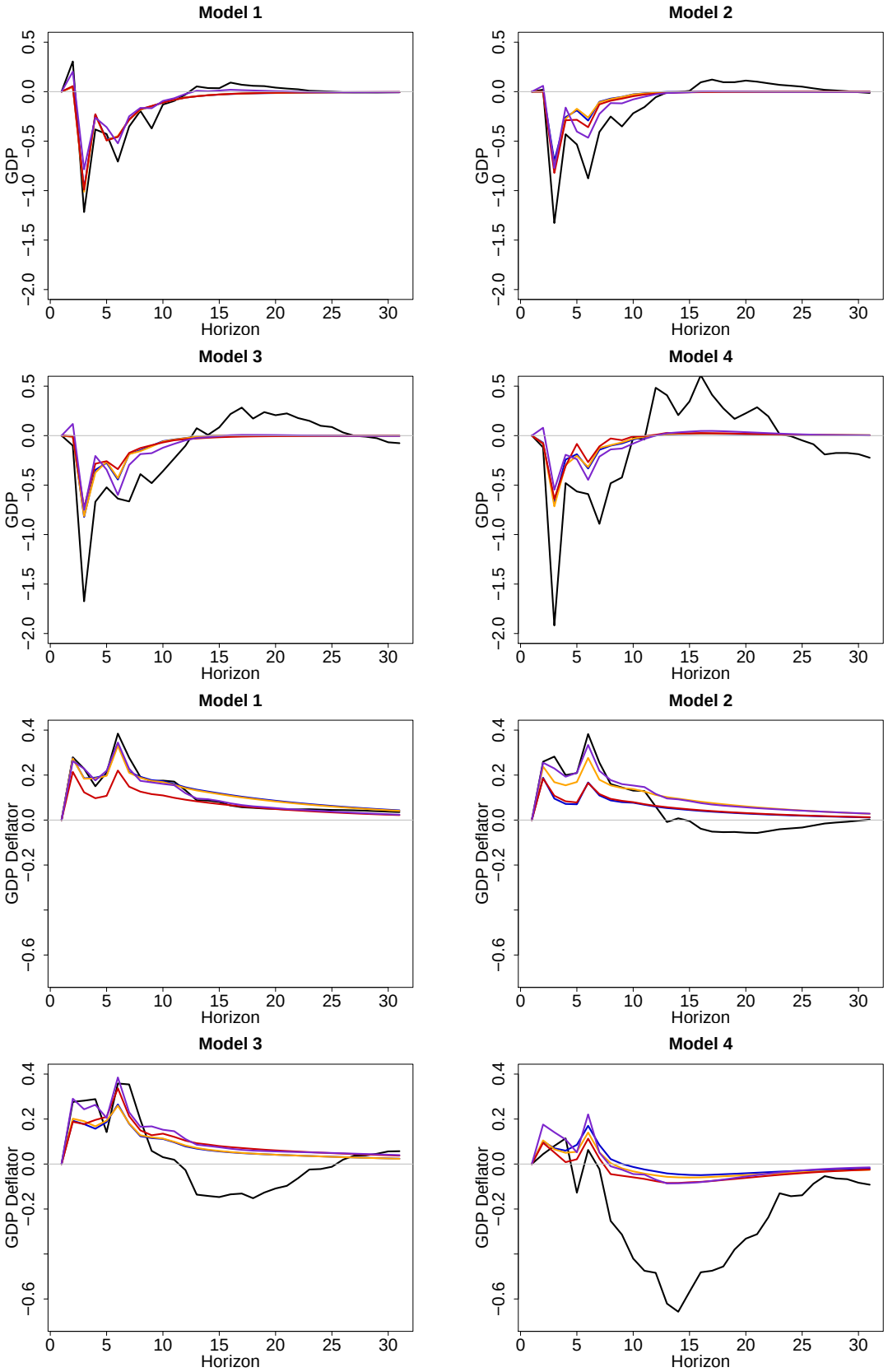
Grey square means the respective coefficient is selected; white square means no selection. Intercept is not included in the figure. $T = 212$ observations are used for CV.

Figure 14: Binary Coefficient Matrices for Lasso VAR(5)



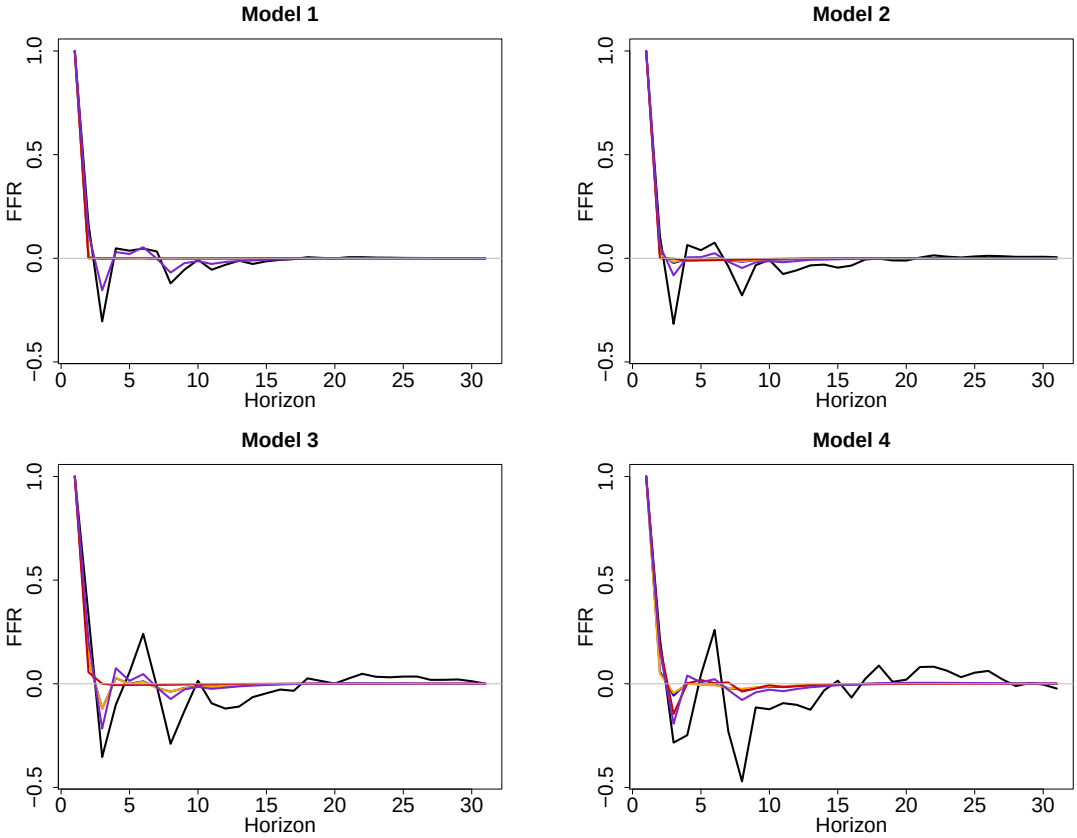
Grey square means the respective coefficient is selected; white square means no selection. Intercept is not included in the figure. $T = 212$ observations are used for CV.

Figure 15: Impulse Response Functions for GDP and GDP Deflator



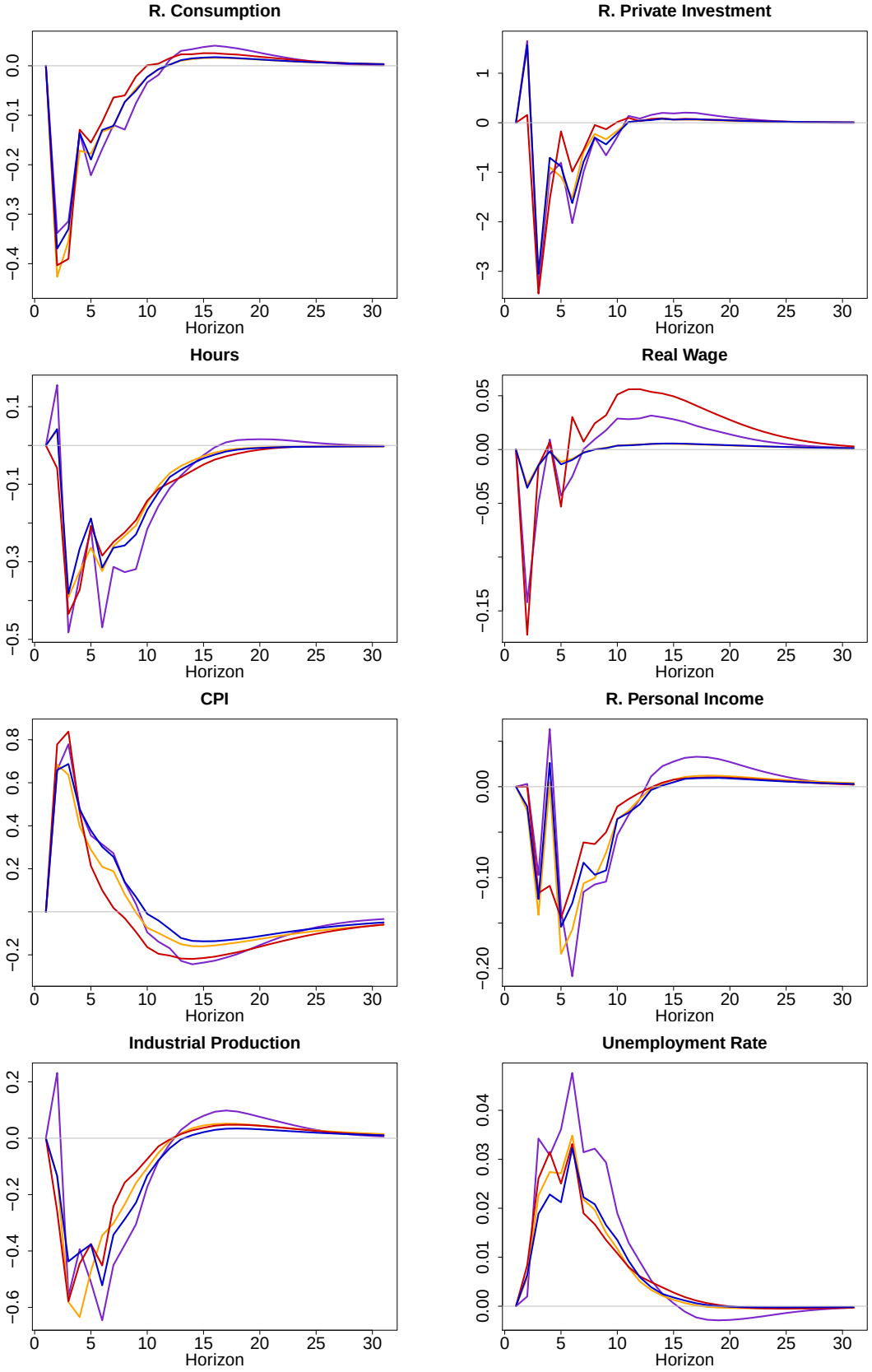
Impulse response functions of GDP and GDP Deflator to a monetary policy shock on the federal funds rate for Model 1-4, where each model includes $p = 5$ lags and $K = 3, 7, 15, 22$ variables, respectively. Estimation is done by OLS (black), Ridge (purple), Lasso (red), Enet (blue) and Fenet *scheme 1* (orange).

Figure 16: Impulse Response Functions for FFR



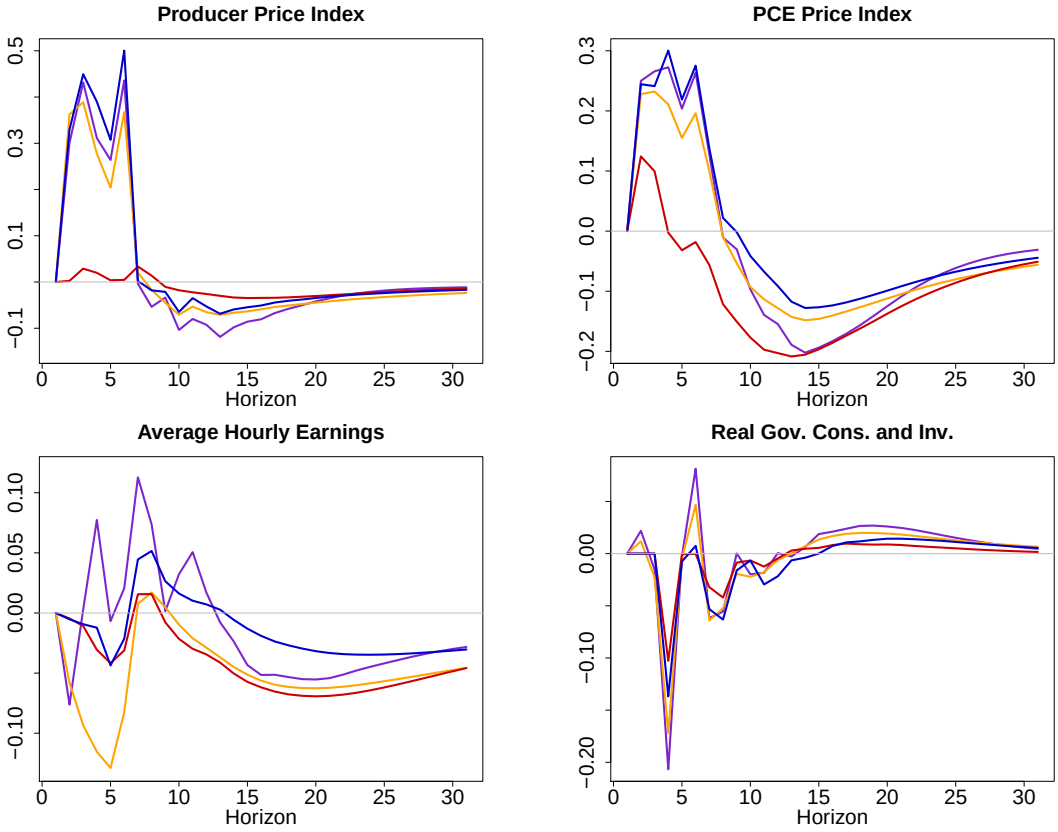
Impulse response functions of FFR to a monetary policy shock on the FFR for Model 1-4, where each model includes $p = 5$ lags and $K = 3, 7, 15, 22$ variables, respectively. Estimation is done by OLS (black), Ridge (purple), Lasso (red), Enet (blue) and Fenet *scheme 1* (orange).

Figure 17: Impulse Response Functions for slow moving variables



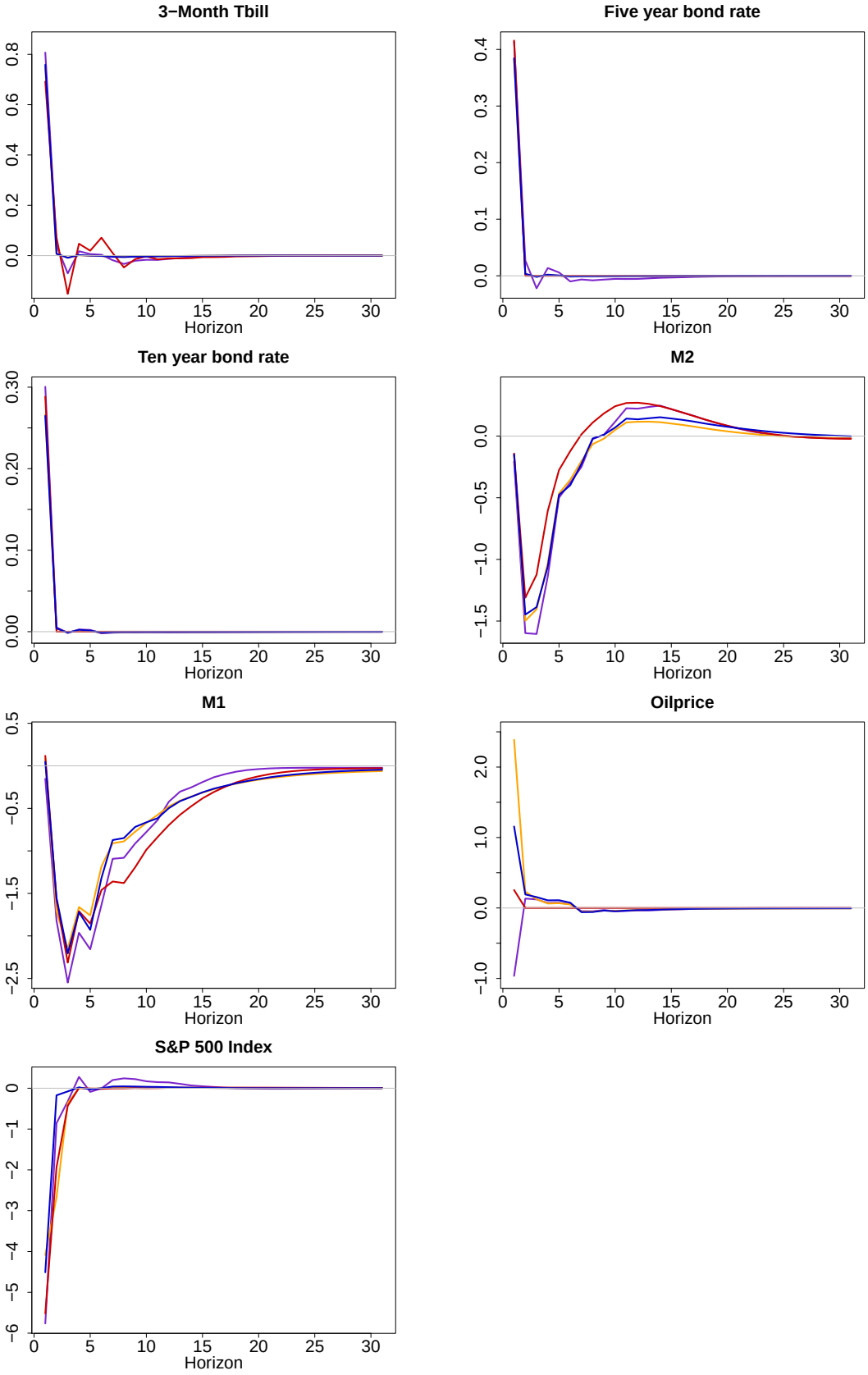
Impulse response functions of slow moving variables to a monetary policy shock on the FFR for the large model. Estimation is done by Ridge (blue), Lasso (red), Enet (purple) and Fenet *scheme 1* (orange).

Figure 18: Impulse Response Functions for slow moving variables



Impulse response functions of slow moving variables to a monetary policy shock on the FFR for the large model. Estimation is done by Ridge (blue), Lasso (red), Enet (purple) and Fenet *scheme 1* (orange).

Figure 19: Impulse Response Functions for fast moving variables



Impulse response functions of fast moving variables to a monetary policy shock on the FFR for the large model. Estimation is done by Ridge (blue), Lasso (red), Enet (purple) and Fenet *scheme 1*(orange).

Table 7: Simulation Results for Variable 1-4 (20 % sparsity)

	GDP								GDP Deflator							
Setup	$T = 80$				$T = 200$				$T = 80$				$T = 200$			
	Ridge	Lasso	Enet	Fenet	Ridge	Lasso	Enet	Fenet	Ridge	Lasso	Enet	Fenet	Ridge	Lasso	Enet	Fenet
h=1	0.1545	0.0906	**0.0453**	0.0634	0.4436	0.2829	**0.1443**	0.1723	**0.2576**	0.5326	0.3925	0.3673	**0.6579**	1.1289	0.9413	0.7413
h=2	2.1089	**1.8488**	2.6379	2.6231	2.2836	**1.9453**	2.971	2.9282	**0.4718**	0.6726	0.6411	0.5957	**0.593**	1.0026	0.8855	0.7467
h=3	0.5133	**0.3575**	0.4654	0.4384	0.6633	0.6057	**0.4992**	0.5107	0.3296	0.3892	0.2976	**0.2703**	**0.7513**	1.4521	1.4392	1.2644
h=4	**0.4018**	0.5227	0.5057	0.5281	**0.7338**	0.9980	1.1183	1.1033	0.2235	0.2936	0.2272	**0.1947**	0.6858	**0.6717**	0.7711	0.7745
h=5	**1.2213**	1.4635	1.7664	1.863	**1.301**	1.7343	2.2056	2.316	**0.3883**	0.5820	0.4685	0.4783	**0.6190**	1.0199	0.9921	0.8706
h=6	**0.3166**	0.3211	0.3246	0.3181	**0.5236**	0.6098	0.6017	0.6028	0.2423	0.2788	0.2356	**0.2286**	0.5519	0.6508	0.5813	**0.5347**
h=7	**0.1415**	0.2087	0.1673	0.1529	0.5736	0.6193	0.5311	**0.5145**	0.1623	0.1587	0.1108	**0.0964**	0.5775	0.5058	**0.4778**	0.4785
h=8	**0.4097**	0.4607	0.4224	0.418	**0.6769**	0.8709	0.8474	0.8581	0.1863	0.1858	0.1203	**0.1058**	0.5755	0.5518	**0.5455**	0.5500
Av.	**0.6584**	0.6592	0.7919	0.8006	**0.8999**	0.9583	1.1148	1.1257	**0.2827**	0.3867	0.3117	0.2921	**0.6265**	0.873	0.8292	0.7451

	Consumption								Investment							
Setup	$T = 80$				$T = 200$				$T = 80$				$T = 200$			
	Ridge	Lasso	Enet	Fenet	Ridge	Lasso	Enet	Fenet	Ridge	Lasso	Enet	Fenet	Ridge	Lasso	Enet	Fenet
h=1	1.2355	**1.1183**	1.5264	1.4437	1.4875	**1.2177**	1.7072	1.6402	**0.2933**	0.4141	0.3217	0.3186	**0.4659**	0.7769	0.621	0.5561
h=2	1.7091	**1.6926**	2.2594	2.2146	1.6524	**1.3967**	2.0117	1.964	0.7812	**0.7558**	0.9732	1.2265	1.1707	**1.0095**	1.1224	1.3999
h=3	**0.4058**	0.4471	0.4318	0.4076	**0.7533**	1.0093	0.9544	0.9275	0.5308	**0.4512**	0.5036	0.5631	0.7152	0.6734	**0.6378**	0.6399
h=4	**0.5901**	0.6925	0.7527	0.755	**1.0082**	1.3693	1.4578	1.4684	0.3035	0.2648	**0.1145**	0.125	0.6667	0.5429	0.4583	**0.4451**
h=5	**0.6042**	0.7936	0.7884	0.7654	0.7494	**0.6707**	0.8804	0.873	0.4657	0.5035	0.4526	**0.4484**	**0.7770**	0.9147	0.9138	0.9256
h=6	**0.3648**	0.3875	0.4628	0.4401	**0.7455**	0.899	0.9169	0.9176	**0.4486**	0.5489	0.4639	0.4988	**0.6328**	0.7138	0.8446	0.9054
h=7	**0.4900**	0.5523	0.5934	0.5849	**0.9904**	1.2823	1.4393	1.4629	0.0760	0.1070	0.0379	**0.0311**	0.4256	0.4352	**0.242**	0.2641
h=8	0.1002	0.1270	0.0772	**0.0717**	0.4160	0.3678	0.2793	**0.2735**	0.2788	0.3398	0.2822	**0.2711**	**0.6328**	0.8141	0.767	0.7579
Av.	**0.6875**	0.7264	0.8615	0.8354	**0.9753**	1.0266	1.2059	1.1909	0.3972	0.4231	**0.3937**	0.4353	**0.6858**	0.7351	0.7009	0.7368

MSE of structural IRFs for a system of $K = 7$ variables for $h = 1, ..., 8$ using $M = 40$ Monte Carlo replications. Methods of interest are the Ridge, Lasso, Elastic Net and Feature Elastic Net *scheme 1*, which are displayed relative to the OLS estimation. The DGP includes 20 % of sparsity. A value smaller than 1, corresponds to better performance than OLS.

Table 8: Simulation Results for Variable 5-7 (20 % sparsity)

	Hours								Wage							
Setup	$T = 80$				$T = 200$				$T = 80$				$T = 200$			
	Ridge	Lasso	Enet	Fenet	Ridge	Lasso	Enet	Fenet	Ridge	Lasso	Enet	Fenet	Ridge	Lasso	Enet	Fenet
h=1	0.1908	0.1860	**0.0762**	0.1086	0.5961	0.4913	**0.3832**	0.4002	**0.2030**	0.2521	0.2425	0.2398	**0.6049**	0.8965	1.0707	0.9507
h=2	0.8063	**0.7632**	1.0538	1.0563	**1.6214**	1.8956	2.3142	2.4108	**0.4306**	0.4748	0.4450	0.4373	**1.1020**	1.2878	1.6648	1.5933
h=3	**0.5986**	0.6630	0.8429	0.8276	**0.7823**	0.8072	0.8961	0.9192	0.1186	0.1060	0.0145	**0.0130**	0.3150	0.3291	**0.1013**	0.1046
h=4	0.3240	**0.2714**	0.2823	0.2746	0.728	0.6762	**0.568**	0.5828	0.2985	0.3169	**0.1978**	0.2042	0.6949	**0.6848**	0.7950	0.7734
h=5	**1.0554**	1.2222	1.3039	1.3443	**1.1402**	1.586	1.8729	1.9988	0.0940	0.1296	**0.0286**	0.0301	0.4403	0.3921	**0.2768**	0.2849
h=6	**0.6221**	0.7510	0.7637	0.7538	**0.7326**	0.7498	0.8856	0.9242	0.2549	0.2361	0.1974	**0.1924**	**0.5424**	0.6525	0.6057	0.5949
h=7	**0.5027**	0.6615	0.6427	0.6391	**0.9085**	1.1814	1.2800	1.3267	**0.1126**	0.1903	0.1327	0.1327	**0.4384**	0.4997	0.5707	0.5708
h=8	**0.7878**	0.9275	1.0077	1.0062	**1.0592**	1.5741	1.8149	1.8819	0.0352	0.0500	**0.0039**	0.0051	0.2863	0.2602	0.1043	**0.1006**
Av.	**0.6110**	0.6807	0.7466	0.7513	**0.946**	1.1202	1.2519	1.3056	0.1934	0.2195	0.1578	**0.1568**	**0.5530**	0.6253	0.6487	0.6217

	FFR							
Setup	$T = 80$				$T = 200$			
	Ridge	Lasso	Enet	Fenet	Ridge	Lasso	Enet	Fenet
h=1	**0.2758**	0.5248	0.3568	0.3548	0.5695	0.6553	0.4884	**0.4819**
h=2	1.4724	**1.2747**	2.3019	2.3341	1.5824	**1.2589**	1.8980	1.8595
h=3	0.2295	0.3111	0.1910	**0.1852**	**0.6192**	0.8455	0.7079	0.6968
h=4	0.4193	0.3864	0.1805	**0.1734**	0.5950	0.6672	0.5713	**0.5415**
h=5	0.3317	0.4158	0.2919	**0.2872**	0.6793	0.7956	**0.5868**	0.6221
h=6	0.1162	0.1952	**0.1003**	0.1016	0.4842	0.5053	0.4785	**0.4527**
h=7	**1.2421**	1.3901	1.7649	1.7676	**1.5962**	1.6515	2.3162	2.2817
h=8	0.1205	0.1735	**0.0674**	0.0691	0.4039	0.3752	**0.2385**	0.2404
Av.	**0.5259**	0.5839	0.6568	0.6591	**0.8162**	0.8443	0.9107	0.8971

MSE of structural IRFs for a system of $K = 7$ variables for $h = 1, ..., 8$ using $M = 40$ Monte Carlo replications. Methods of interest are the Ridge, Lasso, Elastic Net and Feature Elastic Net *scheme 1*, which are displayed relative to the OLS estimation. The DGP includes 20 % of sparsity. A value smaller than 1, corresponds to better performance than OLS.

Table 9: Simulation Results for Variable 1-4 (80 % sparsity)

	GDP								GDP Deflator							
Setup	$T = 80$				$T = 200$				$T = 80$				$T = 200$			
	Ridge	Lasso	Enet	Fenet	Ridge	Lasso	Enet	Fenet	Ridge	Lasso	Enet	Fenet	Ridge	Lasso	Enet	Fenet
h=1	0.1284	0.0749	**0.0705**	0.0936	0.4309	**0.1613**	0.2253	0.3147	**0.2437**	0.4087	0.3049	0.3029	**0.5682**	0.8541	0.6674	0.6563
h=2	1.8926	**1.3087**	1.9337	1.9108	1.7583	**1.0548**	1.8218	1.8137	0.2489	0.1971	**0.1387**	0.1403	0.4502	**0.2597**	0.3113	0.3161
h=3	0.5073	**0.4834**	0.6249	0.6306	1.1311	**0.6924**	1.0605	1.1160	0.3431	0.1379	0.1192	**0.1146**	0.5027	**0.3114**	0.3205	0.3277
h=4	0.1706	0.1607	0.114	**0.1024**	0.4386	**0.2319**	0.2685	0.2429	0.3069	**0.0953**	0.1171	0.1189	0.4416	**0.2394**	0.3349	0.3399
h=5	0.3789	0.2694	0.2053	**0.1517**	0.4319	0.2585	0.2777	**0.2293**	0.3037	**0.1784**	0.2059	0.2008	**0.4476**	0.6893	0.4978	0.4704
h=6	**0.1332**	0.1703	0.1619	0.1768	0.4155	**0.3031**	0.3429	0.3394	0.1874	0.1875	**0.1732**	0.1739	**0.4816**	0.7029	0.5343	0.5104
h=7	**0.2247**	0.3025	0.3288	0.3406	**0.4699**	0.6275	0.6678	0.6907	0.1543	0.0914	**0.0758**	0.0780	0.4850	**0.3241**	0.3677	0.3500
h=8	**0.1088**	0.1251	0.1228	0.1211	0.2925	**0.2505**	0.2719	0.2776	0.1354	0.0863	**0.0562**	0.0572	0.3898	**0.2985**	0.3265	0.3264
Av.	0.4430	**0.3619**	0.4452	0.4410	0.6711	**0.4475**	0.6171	0.6280	0.2404	0.1728	0.1489	**0.1483**	0.4709	0.4599	0.4201	**0.4122**
	Consumption								Investment							
Setup	$T = 80$				$T = 200$				$T = 80$				$T = 200$			
	Ridge	Lasso	Enet	Fenet	Ridge	Lasso	Enet	Fenet	Ridge	Lasso	Enet	Fenet	Ridge	Lasso	Enet	Fenet
h=1	0.6438	**0.4835**	0.7365	0.6965	1.8344	**1.019**	1.6343	1.5855	**0.2635**	0.5892	0.4949	0.4256	**0.5318**	0.8960	0.6516	0.6229
h=2	1.0227	**0.9262**	1.2425	1.2781	2.2510	**1.6638**	2.3918	2.431	1.5118	**1.396**	1.7309	1.8731	1.2080	**0.7838**	1.2026	1.5281
h=3	**0.3255**	0.3404	0.3483	0.3432	0.6508	**0.4533**	0.6005	0.5925	0.2899	0.2776	0.2624	**0.2559**	0.6871	0.5311	0.5147	**0.4721**
h=4	0.1380	0.1923	**0.1136**	0.1265	0.3765	0.3242	0.2599	**0.2504**	0.4803	0.4900	0.4471	**0.4381**	0.9794	**0.7149**	0.9523	0.8577
h=5	0.1072	0.1215	**0.0813**	0.082	0.4939	**0.2418**	0.2903	0.2702	0.3862	0.4156	0.2188	**0.1776**	0.5105	0.3720	0.3558	**0.2813**
h=6	**0.2327**	0.3007	0.2898	0.2899	**0.5056**	0.5106	0.5922	0.597	0.1193	0.2028	**0.0757**	0.1036	0.2973	**0.1347**	0.1630	0.1542
h=7	**0.1774**	0.2121	0.1920	0.1950	**0.4506**	0.5155	0.5904	0.6248	0.0908	0.1177	**0.0543**	0.0560	0.2587	0.2466	0.1819	**0.1639**
h=8	0.2290	0.1930	0.1813	**0.1754**	0.4745	**0.2623**	0.3674	0.3631	**0.1976**	0.3140	0.2766	0.2650	**0.5022**	0.5928	0.7041	0.6928
Av.	0.3595	**0.3462**	0.3982	0.3983	0.8797	**0.6238**	0.8409	0.8393	**0.4174**	0.4754	0.4451	0.4494	0.6219	**0.534**	0.5907	0.5966

MSE of structural IRFs for a system of $K = 7$ variables for $h = 1, ..., 8$ using $M = 40$ Monte Carlo replications. Methods of interest are the Ridge, Lasso, Elastic Net and Feature Elastic Net *scheme 1*, which are displayed relative to the OLS estimation. The DGP includes 80 % of sparsity. A value smaller than 1, corresponds to better performance than OLS.

Table 10: Simulation Results for Variable 5-7 (80 % sparsity)

	Hours								Wage							
Setup	$T = 80$				$T = 200$				$T = 80$				$T = 200$			
	Ridge	Lasso	Enet	Fenet	Ridge	Lasso	Enet	Fenet	Ridge	Lasso	Enet	Fenet	Ridge	Lasso	Enet	Fenet
h=1	0.1521	0.1561	**0.0929**	0.1143	0.5002	**0.1510**	0.3436	0.4132	0.0572	0.0340	**0.0074**	0.0196	0.1832	0.1756	**0.0222**	0.0245
h=2	0.7546	**0.6005**	0.8372	0.9166	1.3979	**0.8927**	1.513	1.5831	0.0545	0.0278	**0.0038**	0.0059	0.1815	0.1931	**0.0836**	0.0870
h=3	0.3817	**0.3248**	0.4032	0.3842	0.5145	**0.3781**	0.4484	0.4695	0.2094	**0.1841**	0.2371	0.2136	0.5698	**0.5430**	0.6625	0.6101
h=4	0.2129	0.2267	**0.1938**	0.1995	0.7034	**0.5277**	0.6346	0.6144	0.0905	0.1015	0.0650	**0.0586**	0.1958	0.1967	**0.1254**	0.1305
h=5	0.3328	0.2740	0.2648	**0.2463**	0.3819	**0.193**	0.2412	0.2309	0.0971	0.0688	**0.0191**	0.0256	0.1908	0.1995	**0.0516**	0.0575
h=6	0.0656	0.0642	**0.0356**	0.0405	0.2529	**0.1767**	0.1947	0.1844	0.0524	0.0537	**0.0151**	0.0178	0.1839	0.1929	0.0509	**0.0468**
h=7	**0.1822**	0.1949	0.2136	0.2091	**0.4761**	0.4903	0.5348	0.5282	0.0402	0.0562	0.0317	**0.0269**	0.2789	0.2956	0.2574	**0.2405**
h=8	**0.1843**	0.2092	0.2203	0.2206	**0.4194**	0.4882	0.5362	0.5601	0.0394	0.0351	**0.0161**	0.0205	0.1013	0.1522	0.0507	**0.0501**
Av.	0.2833	**0.2563**	0.2827	0.2914	0.5808	**0.4122**	0.5558	0.5730	0.0801	0.0701	0.0494	**0.0486**	0.2357	0.2436	0.1630	**0.1559**

	FFR							
Setup	$T = 80$				$T = 200$			
	Ridge	Lasso	Enet	Fenet	Ridge	Lasso	Enet	Fenet
h=1	0.1562	0.1785	**0.0784**	0.0805	0.5521	**0.1781**	0.3988	0.3985
h=2	**1.3627**	1.3906	1.6965	1.6881	1.4159	**1.0587**	1.9003	1.8953
h=3	**0.324**	0.4229	0.3273	0.3437	**0.5854**	0.7103	0.6776	0.6677
h=4	0.1697	0.1667	**0.0855**	0.0886	0.4316	0.3714	**0.2888**	0.3012
h=5	0.2231	0.3477	**0.1712**	0.1726	0.4660	0.6109	**0.4104**	0.4146
h=6	0.1923	0.2651	**0.0819**	0.0936	0.3631	**0.3185**	0.3203	0.3371
h=7	0.6794	**0.5402**	0.7540	0.7569	1.1276	**0.7295**	1.3607	1.3769
h=8	0.1021	0.1531	**0.0725**	0.0745	0.3676	0.3152	**0.2791**	0.2842
Av.	**0.4012**	0.4331	0.4084	0.4123	0.6637	**0.5366**	0.7045	0.7094

MSE of structural IRFs for a system of $K = 7$ variables for $h = 1, ..., 8$ using $M = 40$ Monte Carlo replications. Methods of interest are the Ridge, Lasso, Elastic Net and Feature Elastic Net *scheme 1*, which are displayed relative to the OLS estimation. The DGP includes 80 % of sparsity. A value smaller than 1, corresponds to better performance than OLS.

Affidavit

I affirm that this Master thesis was written by myself without any unauthorised third-party support. All used references and resources are clearly indicated. All quotes and citations are properly referenced. This thesis was never presented in the past in the same or similar form to any examination board. I agree that my thesis may be subject to electronic plagiarism check. For this purpose an anonymous copy may be distributed and uploaded to servers within and outside the University of Mannheim.

Mannheim, 25.09.2020 Signature