Erasmus University Rotterdam

Erasmus School of Economics
Bachelor Thesis Financial Econometrics
Quantitative Finance

Subset Selection for VAR Processes: a Comparison of Various Penalised Regression Techniques

Author:
Gerben van der Schaaf (416661)

Supervisor:
Dr. A.M. Schnücker
Second assessor:
Mr. J. Klooster

July 5, 2020

Abstract

Vector autoregressive (VAR) models are great for capturing the dynamics of financial time series. This paper researches the effectiveness of various model selection techniques in VAR processes. By comparing past and modern methods that generally increase in complexity, this research offers an extensive empirical analysis of the development of penalised regressions. We evaluate each technique through its prediction performance for U.S. macroeconomic data. We incorporate the Lasso, adaptive Lasso, SCAD, elastic net and adaptive elastic net and compare these to several basic model selection methods. Furthermore, we look at possible combinations of techniques and evaluate whether these hybrid procedures produce promising results. We find that for a large sample size a hybrid technique consisting of an information criteria and the adaptive elastic net, where the model is estimated equation by equation, leads to the best prediction performance.

Keywords: vector autoregressive processes, penalised regression, model selection

Contents

1 Introduction

Time series analysis has always played a key factor in the financial world. Especially, the forecasting component which is the primary goal when applying time series analysis in financial econometrics. In complex problems with numerous variables, it can be quite challenging to develop an uncomplicated model that generates accurate estimates and possesses a relatively low computation time. Autoregressive (AR) models are a great fit in financial analysis. These models argue that the dependent variable has a linear dependency on its own lagged values and on a stochastic error term. Research shows that there often exists some type of interrelations between various time series in a model, which AR models simply cannot specify. This displays a major limitation of the single equation AR model. In order to solve this problem, a generalisation of the univariate autoregressive model is needed, which leads to the necessity of vector autoregressive (VAR) models.

The corresponding framework ensures that the model contains multiple dependent variables. This allows VAR models to capture the linear interdependencies and describe the dynamics of multivariate time series. However, it remains challenging to accurately determine the extent of the dependencies between various time series. Sims (1980) is the first to test and find success in implementing VAR models in macroeconomics related research. A significant benefit of VAR models is that they are not excessively complex and therefore quite straightforward to implement.

Nonetheless, a VAR model that incorporates all original parameters is rarely effective. A VAR model, with a k-dimensional time series and p as order of lags, consists of matrices with a substantial number of parameters, even for relatively small k and p. Adding all parameters can result in unnecessary computation time. This would be a crucial downside of the VAR model. To solve the issue, some parameters need to be eliminated because of redundancy. The statistical process of testing various alterations to find the optimal model is called model selection. This procedure returns the best model, which is often a subset of the full model. Through model selection, we obtain the best performing predictions while implementing the advantageous VAR structure.

In this work, similar to Hsu et al. (2008), we are focused on subset selection for VAR processes. We create a VAR model and implement the lagged variables for all incorporated time series as the explanatory variables in the regression. This changes the procedure of subset selection to a variable selection problem. We examine the prediction performance of the information criterion, searching procedures, Lasso, adaptive Lasso, SCAD, elastic net, adaptive elastic net and various hybrid methods. Our evaluation is based on prediction errors and estimation errors. We apply these techniques to a U.S. macroeconomic data set where we examine the dynamic relationship between the nominal GDP, unemployment rate (UER) and the gross rate of M1.

This paper extends former work by creating an extensive comparative analysis of various model selection techniques in a VAR environment. By starting at basic subset selection methods and constantly increasing the complexity of the models, we obtain an accurate and intriguing overview of the history and evolution of model selection. Where Hsu et al. (2008) and Ren et al. (2010) focus on the introduction of the Lasso and

adaptive Lasso in VAR processes, this paper adds several modern techniques that are built on the foundation of the aforementioned works. Furthermore, this research questions the robustness of the conclusions that have been drawn in similar papers. We analyse whether certain methods always outperform other methods, or only under certain circumstances. And finally, we investigate how the introduction of hybrid procedures hold up against the individual techniques.

Previous research, such as Burnham et al. (2002), shows that there are generally three main means for model selection in statistics. Firstly, there is the information criterion that measures the quality of a statistical model. To accomplish this, the complexity of the model and the extent to which the model fits the data are taken into account. This trade-off between simplicity and the goodness of fit represents the risk of underfitting versus overfitting. The most popular information criteria are the Akaike information criterion (AIC), as done by Akaike (1974), and the Bayesian information criterion (BIC), as done by Schwarz (1978). The latter is also known as the Schwarz information criterion (SIC). Model selection implements these information criteria to minimise the estimated information loss. In general, information criteria only focus on the optimal order of the variables inside of the model, which brings up a disadvantage of this method. It is not feasible to test all the subsets of the available variables due to the extreme number of subsets. Even for a relative small number of parameters, there are too many subsets to calculate the estimation of the information criterion.

To solve the aforementioned computational problem, the number of estimated subsets needs to be considerably reduced. This can be done by introducing searching procedures with parameter constraints, which in our context is the second technique for model selection. This paper implements two strategies from Lütkepohl (2005). The first being the top-down strategy, and the second being the bottom-up strategy. Other research, such as Brüggemann et al. (2000), prove the existence of other subset selection methods for VAR processes. However, these results are often suboptimal to the exhaustive search. The third and final mean of model selection is based on hypothesis testing. However, this method generally focuses on eliminating incorrect models, instead of finding the correct one and is often computationally intensive. Therefore, this paper does not incorporate this method and only implements the first two model selection techniques, which can be viewed as benchmarks.

As mentioned before, an optimal performing model possesses the best trade-off between simplicity and goodness of fit. A technique that revolves around these two characteristics is the least absolute shrinkage and selection operator, also known as Lasso. Shrinkage is in layman's terms a statistical procedure where the estimated coefficients are shrunk to increase prediction performance. This is usually accomplished by increasing the bias of the estimator. The bias is the difference between the expected value of an estimator and its true value. In practice, small biases are often accepted if it results in lower values of estimation error. Shrinkage revolves around the bias–variance trade-off, which represents the conflict of attempting to simultaneously minimise these two measurements of error.

The Lasso is a regression analysis method that functions both in variable selection and regularisation. This method is popularised in Tibshirani (1996). He found that data analysts are often not satisfied with

OLS estimates. They often have low bias but large variance. At the time of Tibshirani (1996), there were two standard techniques for improving OLS estimates: subset selection and ridge regression. The former was during this time nothing more than a discrete process. Parameters were either incorporated or dropped. This made it very data dependent, which is a significant liability. The second technique is a continuous process that shrinks coefficients which can improve prediction accuracy. Ridge regression, also known as Tikhonov regularisation, originates from the concepts that are introduced in Tikhonov et al. (1977). Lasso is to a certain extent a combination of these two methods. This is why the Lasso technique is often viewed as an attractive version of OLS. Lasso constrains the sum of the absolute regression coefficients, which forces certain coefficients to be equal to zero and might shrink other coefficients. An important difference between Lasso and ridge is that ridge regression often shrinks certain coefficients close to zero, yet never equal to zero. This is however the case for the Lasso procedure. Applying Lasso generates a simplified model that ideally, only loses a relatively small amount of information by the shrinkage process.

An enormous step in the research area of this paper is the implementation of the Lasso technique in VAR processes, as introduced by Hsu et al. (2008). VAR models often contain a large number of parameters due to the dimensions and order of the model. In theory, VAR models and the Lasso technique seem like a great fit. To perform the Lasso method, Hsu et al. (2008) applies the least angle regression algorithm, also known as (LARS). Efron et al. (2004) developed LARS as an improved version of traditional forward selection methods. The LARS algorithm has an identical computational order to OLS estimates for a model that contains all possible variables. Hsu et al. (2008) found that the Lasso method in combination with the AIC has remarkable performance and that implementation of the LARS algorithm allows high dimensional data to be fitted in a computationally efficient manner for high order VAR models. Hsu et al. (2008) is quite an innovative research and was the main cause of many promising ventures in the years after.

The Lasso method is however not flawless. Meinshausen et al. (2006) states that in some situations, the Lasso generates inconsistent results. Zou (2006) continued on this idea and developed an adaptation of the Lasso method, called the adaptive Lasso. He argues that the inconsistency of the Lasso was to be found in the variable selection process. By proposing to incorporate adaptive weights in the penalisation of certain coefficients, the adaptive lasso solves this inconsistency problem. When the weights are chosen properly, the adaptive Lasso even possesses the oracle properties. This means that its performance is as well as if the true underlying model was known before estimation. Ren et al. (2010) is the first research to implement the adaptive Lasso in VAR processes. He found the optimal parameters in the penalty term by implementing a two-dimensional grid search.

Additionally, we include a penalised likelihood technique to provide a more varied set of penalties. These models can be used for generalised linear models, robust regression, nonparametric models and several more. This makes penalised likelihood quite divers in its applications. We incorporate the smoothly clipped absolute deviation (SCAD) penalty. This penalty behaves like the Lasso for small coefficients. For relatively large coefficients, the SCAD does not penalise at all. And between these values, the SCAD provides a smooth

transition, hence its name. The framework of the SCAD is first introduced in Fan et al. (2001).

Before the introduction of the adaptive Lasso, there were discussions of other limitations the Lasso would have. Zou et al. (2005) argues that there are three scenarios where the drawbacks of the Lasso are exposed. The first is when the number of predictor variables p exceeds the amount of observations n. In this case, the Lasso selects at most n variables. This is due to the framework of the optimisation problem. The second scenario is when there is one or multiple groups of variables present in the data set. The Lasso tends to select only one variable from a group and does not seem to have a consistent selection procedure for this case. And lastly, when predictor variables are highly correlated, the Lasso is vastly outperformed by ridge regression. Zou et al. (2005) tried to find a new method that would keep all the beneficial characteristics of the Lasso and would solve the aforementioned limitations. They proposed a new regularisation technique called the elastic net. This penalty is a convex combination of the Lasso and ridge penalty. Zou et al. (2005) shows that the elastic net often outperforms the Lasso in prediction accuracy, while keeping its strength to simplify models and thereby making them easier to comprehend.

Zou et al. (2009) further develops the elastic net procedure by replacing the Lasso penalty with the adaptive Lasso penalty. This creates the adaptive elastic net. This combines the strength of the quadratic regularisation of the ridge component with the oracle properties of the adaptive Lasso. Therefore, under weak regularity conditions, the adaptive elastic net possesses the oracle property as well. Zou et al. (2009) shows that the adaptive elastic net outperforms other oracle-like methods in prediction accuracy, making it quite a promising technique.

This paper studies a variable selection problem in a VAR context by implementing various model selection techniques. Ultimately, we find that more advanced methods do not necessarily outperform less advanced techniques. However, when these advanced methods are incorporated in a hybrid procedure, then there is a substantial difference in prediction performance. Our results indicate that the AIC combined with the adaptive elastic net leads to the best performance. Furthermore, we show that our methods generally produce more accurate predictions when they are estimated equation by equation, and not simultaneously. The rest of the paper is organised as follows. In Section 2, we create the context around a VAR model and review the theoretical framework behind the different subset selection techniques. In Section 3, we briefly explain the chosen data set and report its interesting features. Next in Section 4, we explain the generated results. And finally, in Section 5 and 6, we draw conclusions and discuss our findings compared to similar research.

2 Methodology

In this section we define the context of our research, discuss the methods that this paper implements and lay the foundation on which our results are built. We start with a description of VAR models, their formulation in a regression context and the least square estimation. Then, we move to the techniques and their descriptions.

2.1 VAR models

A VAR model is a vector of autoregressive (AR) models. We denote a p-order AR model as

$$y_t = v + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + u_t,$$

where y_t is the output variable, v is an intercept, $\phi_1, ..., \phi_p$ are the parameters, u_t is a white noise process. To obtain a VAR model, we now introduce a k-dimensional time series $\{\boldsymbol{y}_t : t = 1, 2, ..., T\}$, where $\boldsymbol{y}_t = (y_{1t}, y_{2t}, ..., y_{kt})'$ and T is the number of observations in the data set. We define a p-order vector autoregressive model as

$$\boldsymbol{y}_t = \boldsymbol{v} + \Phi_1 \boldsymbol{y}_{t-1} + ... + \Phi_p \boldsymbol{y}_{t-p} + \boldsymbol{u}_t, \tag{1}$$

where $\boldsymbol{v}$ is a $k \times 1$ vector that contains intercepts, $\Phi_1, ..., \Phi_p$ represents $k \times k$ coefficient matrices, $\boldsymbol{u}_t$ is a white noise process with covariance matrix $\mathrm{E}(\boldsymbol{u}_t \boldsymbol{u}_t') = \Sigma_u$. Before we can perform a regression, we introduce the following variables:

$$\begin{aligned}
\boldsymbol{Y}^* &= (\boldsymbol{y}_{p+1}, \boldsymbol{y}_{p+2}, ..., \boldsymbol{y}_n), & \boldsymbol{Y} &= \mathrm{vec}(\boldsymbol{Y}^*), \\
X_t &= (\boldsymbol{1}, \boldsymbol{y}_t', ..., \boldsymbol{y}_{t-p+1}')', & \boldsymbol{X}^* &= (X_p, ..., X_{n-1}), \\
\boldsymbol{B} &= (\boldsymbol{v}, \Phi_1, ..., \Phi_p), & \boldsymbol{\beta} &= \mathrm{vec}(\boldsymbol{B}), \\
\boldsymbol{U}^* &= (\boldsymbol{u}_{p+1}, ..., \boldsymbol{u}_n), & \boldsymbol{U} &= \mathrm{vec}(\boldsymbol{U}^*),
\end{aligned}$$

where vec represents the vectorisation of a matrix. This is a linear transformation where a matrix converts into a column vector. As an example, to obtain $\boldsymbol{\beta} = \mathrm{vec}(\boldsymbol{B})$ we stack all columns of the matrix $\boldsymbol{B}$ on top of each other. Hence, $\boldsymbol{\beta}$ is an $(k(kp+1) \times 1)$ column vector. We can now create the matrix notation, derived from (1), as

$$\boldsymbol{Y}^* = \boldsymbol{B}\boldsymbol{X}^* + \boldsymbol{U}^*.$$

This is equivalent to

$$\boldsymbol{Y} = \Big((\boldsymbol{X}^*)' \otimes \boldsymbol{I}_k\Big)\boldsymbol{\beta} + \boldsymbol{U} \equiv \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}, \tag{2}$$

where $\boldsymbol{I}_k$ is the k-dimensional identity matrix, $\otimes$ denotes the Kronecker product and we express the covariance matrix of $\boldsymbol{U}$ as $\Sigma_U = \boldsymbol{I}_{n-p} \otimes \Sigma_u$. As we recognise a regression setup in equation (2), we find that the least squares (LS) estimators of the parameters satisfy

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \Big[\Big(\boldsymbol{X}^*(\boldsymbol{X}^*)'\Big)^{-1}\boldsymbol{X}^* \otimes \boldsymbol{I}_k\Big]\boldsymbol{Y}, \\
\hat{\Sigma}_u &= \frac{1}{n-p}(\boldsymbol{Y}^* - \hat{\boldsymbol{\beta}}\boldsymbol{X}^*)(\boldsymbol{Y}^* - \hat{\boldsymbol{\beta}}\boldsymbol{X}^*)'.
\end{aligned}$$

This minimises the weighted least squares

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'\ \Sigma_U^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Fuller (1996) found that under some regularity conditions, the LS estimator $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normal.

2.2 Subset selection

As we stated earlier, a VAR model that has not discarded or shrunk any parameters is often ineffective. That is why we apply model selection, which gives us a subset of the full model. This subset is deemed as the best performing subset by some criteria. In this section we explain the ideas behind several types of model selection. We begin by discussing the Lasso technique and then move to increasingly more advanced techniques. Then we finish the methodology by introducing the evaluation criteria and the conventional procedures, which can be seen as benchmarks.

2.2.1 Lasso

The Lasso technique is a regression analysis method that performs both variable selection and regularisation. As mentioned before, the Lasso is quite similar to the OLS approach. Both techniques target to minimise the residual sum of squares (RSS), which is the difference between the data and the estimated model. However, there is a major difference between the Lasso and OLS. The framework of the OLS revolves only around the minimisation problem, while the framework of the Lasso consists of an additional component. The Lasso aims to minimise the residual sum of squares subject to the sum of the absolute value of the coefficients. This constraint leads to the possibility that certain coefficients are set equal to zero. When this occurs, we obtain a subset of the full model which increases the interpretability of the model. This is what makes the Lasso method a popular and often favoured subset selection technique. Tibshirani (1996) introduces the framework of the Lasso. The Lasso is a least squares method with an L_1 constraint on the regression parameters. Under the formulation in equation (2), we obtain the Lasso estimator $\tilde{\boldsymbol{\beta}}$ by minimising the RSS:

$$\text{RSS} = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \ ,$$

with respect to $\boldsymbol{\beta}$ subject to the constraint $\sum |\beta_j| \leq s$, where β_j represents the j-th element of $\boldsymbol{\beta}$ and s is a tuning parameter. We can rewrite the aforementioned RSS by expressing it as the square of the norm of residuals. Then we can formulate the Lasso estimator

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\arg\min} \left\{ ||\boldsymbol{Y} - \boldsymbol{X}'\boldsymbol{\beta}||^2 + \lambda \sum_{j=1}^{pk^2} |\beta_j| \right\} \ , \tag{3}$$

where p is the amount of predictor variables which in our context is the number of lags, k is the amount of dimensions in the VAR model and λ is a tuning parameter that depends on s. A tuning parameter, also known as penalty parameter, determines the strength of the penalty term. This represents the level of shrinkage. Tuning parameters are an essential component of regularisation. This process is responsible to prevent overfitting in statistics. The most popular regularisation techniques implement either an L_1 or an L_2 penalty, where the former is associated with the Lasso and the latter is associated with ridge regression. In other words, the L_1 is the sum of the absolute distance, whereas L_2 represents the Euclidean norm. The tuning parameter operates as follows, λ represents a vector of slowly decreasing or increasing values. This creates a series of different models. We evaluate these models and select the one with the best performance.

We describe three situations for our tuning parameter:

$\lambda = 0$, no parameters are eliminated and the estimates are equal to the ones found by OLS.

λ increases which means more coefficients are shrunk and set to zero.

$\lambda \to \infty$, all coefficients are set to zero which eliminates all parameters.

We observe a trade-off in bias and variance. When λ increases, the bias increases and the variance decreases, when λ decreases, the bias decreases and the variance increases. In the entirety of this work, we implement an approach proposed by Tibshirani (1996) to find the optimal tuning parameter λ using a tenfold cross-validation method.

2.2.2 Adaptive Lasso

As we observe in the framework of the Lasso, the penalty term of the Lasso is linear in the size of the regression coefficient. Therefore, Meinshausen et al. (2006) argues that the Lasso tends to give substantially biased estimates for large regression coefficients. Zou (2006) introduced an alteration of the Lasso, called the adaptive Lasso. Ren et al. (2010) is the first to extend this framework to VAR processes. In this technique, the addition of the weights changes the operation of the Lasso procedure. The adaptive Lasso estimator is given by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ ||\boldsymbol{Y} - \boldsymbol{X}'\boldsymbol{\beta}||^2 + \lambda \sum_{j=1}^{pk^2} \hat{\omega}_j \ |\beta_j| \right\}, \tag{4}$$

where $\hat{\omega}_j$ represents adaptive weights that ensure an oracle property. We set $\hat{\omega}_j = 1/|\hat{\beta}_j|^\gamma$, where $\hat{\beta}_j$ can be any root-n consistent estimator of the true β, as long as $\gamma > 0$. Therefore, we choose to use the consistent OLS estimator $\hat{\beta}_{OLS}$, which we can obtain quite easily. Instead of implementing a grid search, as done by Ren et al. (2010) to find an optimal γ, we choose to apply the adaptive Lasso by setting γ equal to the widely used values 0.5, 1, 1.5 and 2 as observed in similar contexts. We find the preferred solution for γ per technique by a simple trial-and-error method. Furthermore, when we set $\gamma = 0$ in equation (4), we observe that the adaptive Lasso estimator is equal to the regular Lasso estimator.

2.2.3 SCAD

The framework of the SCAD as discussed in Fan et al. (2001) is as follows

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ ||\boldsymbol{Y} - \boldsymbol{X}'\boldsymbol{\beta}||^2 + p_\lambda^{SCAD}(\beta_j) \right\}, \tag{5}$$

where

$$p_\lambda^{SCAD}(\beta_j) = \begin{cases} \lambda|\beta_j|, & \text{if } \beta_j \leq \lambda, \\ -\left(\frac{|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2)}{2(a-1)} \right), & \text{if } \lambda < \beta_j \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2}, & \text{if } \beta_j > a\lambda. \end{cases}$$

The parameter a must be greater than 2 at all times. We observe that for relatively small values of β_j, SCAD penalises similarly to Lasso. With correct regularisation parameters SCAD possesses, similar to adaptive

Lasso, the oracle property. Instead of finding optimal a by using a grid search, as in Fan et al. (2001). We choose to apply the SCAD by setting a equal to 2.5, 4, 6, 8 and 10. We find the preferred value for a by a simple trial-and-error method.

2.2.4 Elastic net

The elastic net is introduced in Zou et al. (2005) and can be viewed as a convex combination of the L_1 Lasso and L_2 ridge penalty. This gives the elastic net estimator a great amount of flexibility. When implemented on a VAR model, this procedure can adapt to the necessity of the model. This gives the elastic net on paper an advantage over the Lasso. Furthermore, two of the three limitations of the Lasso in Zou et al. (2005) are revolved around strongly correlated predictor variables. An environment where VAR models theoretically excel. Therefore, the elastic net seems quite an intriguing technique for VAR processes. However, the elastic net does have disadvantages. The first is that by adding α in the framework, the computational cost increases tremendously. Additionally, the flexibility of the elastic net increases the probability of overfitting, which is always undesirable. We define the aforementioned penalties as follows

$$||\beta||_1 = \sum_{j=1}^{pk^2} |\beta_j| \ , \qquad ||\beta||^2 = \sum_{j=1}^{pk^2} \beta_j^2 \ ,$$

This leads to the following framework for the elastic net estimator

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ ||\boldsymbol{Y} - \boldsymbol{X}'\boldsymbol{\beta}||^2 + \alpha \ \lambda \sum_{j=1}^{pk^2} |\beta_j| + (1-\alpha) \ \lambda \sum_{j=1}^{pk^2} \beta_j^2 \right\} , \tag{6}$$

where α is a parameter that determines the proportion of which the elastic net is similar to the Lasso. For example, when we set $\alpha = 1$ in equation (6) we observe that the elastic net is identical to the Lasso. When $\alpha = 0$, the elastic net is equivalent to ridge regression. An excellent method to find the optimal value of α would be by minimising cross-validated mean squared error by implementing a grid search. This research chooses to set α in equation (6) equal to the values 0.25, 0.5 and 0.75. We find the preferred solution for α per technique by a simple trial-and-error method.

2.2.5 Adaptive elastic net

The adaptive elastic net is first introduced in Zou et al. (2009). This method can be viewed as a combination of the elastic net and the adaptive Lasso. Therefore, it essentially enjoys all beneficial attributes of both individual techniques, yet only possesses the disadvantages of the elastic net. Hence, the adaptive elastic net can be seen as an improvement of the elastic net. The adaptive elastic net estimator is given by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ ||\boldsymbol{Y} - \boldsymbol{X}'\boldsymbol{\beta}||^2 + \alpha \ \lambda \sum_{j=1}^{pk^2} \hat{\omega}_j \ |\beta_j| + (1-\alpha) \ \lambda \sum_{j=1}^{pk^2} \beta_j^2 \right\} , \tag{7}$$

where $\hat{\omega}_j$ represents adaptive weights. We set $\hat{\omega}_j = 1/|\hat{\beta}_j|^\gamma$. When $\alpha = 1$ in equation (7), we observe that the adaptive elastic net is identical to the adaptive Lasso. When $\alpha = 0$, the adaptive elastic net is equivalent

to ridge regression. We select γ in the same manner as we do for the adaptive Lasso in Section 2.2.2 and select α similar to the elastic net in Section 2.2.3.

2.2.6 Conventional procedures

Information criteria weigh the simplicity of a model against the extent to which the model fits the data. This trade-off is expressed in numerical values in which a lower value is favoured. In this research, we implement the AIC and BIC. For a VAR model as specified in equation (1), AIC and BIC select the best model with order

$$\hat{p}_{aic} = \arg\min_{p} \Big\{ n \ \ln \Big| \ \tilde{\Sigma}_u(p) \ \Big| + 2pk^2 \Big\}, \tag{8}$$

$$\hat{p}_{bic} = \arg\min_{p} \Big\{ n \ \ln \Big| \ \tilde{\Sigma}_u(p) \ \Big| + (\ln \ n) \ pk^2 \Big\}, \tag{9}$$

where $\hat{\Sigma}_u(p)$ represents the maximum likelihood (ML) estimate of Σ_u, which is the covariance matrix of u, under the fitted VAR(p) model. We observe in equations (8) and (9) that the difference between the AIC and BIC is in the end of the equations, which is known as the penalised factor for model complexity. Or in other words, the penalty for the number of parameters. Considering that the penalised factor $\ln(n)$ is larger than two for n larger than seven, the BIC tends to gravitate towards lower order models than the AIC. Generally, information-based criteria only function as order selection and not as subset selection in a VAR context.

Next we introduce the top-down and bottom-up strategies. In this paragraph, we explain how these procedures reduce a VAR(p) model for each individual series. We start by denoting $\boldsymbol{\beta}_j$ as the j-th parameter vector in $\boldsymbol{\beta}$. This is a $(kp + 1)$-dimensional column vector that corresponds to the j-th row in matrix $\boldsymbol{B}$, for j = 1,2,...,k, where k is of course the number of dimensions or individual time series. Hence, $\boldsymbol{\beta}_j$ is linked with the j-th series. We define the i-th element in $\boldsymbol{\beta}_j$ as $\boldsymbol{\beta}_{j,i}$. We denote the full AR(p) model that contains all parameters in $\boldsymbol{\beta}_j$ for the j-th series as $\mathcal{M}(\boldsymbol{\beta}_j)$. Moreover, we define $\mathcal{M}(\boldsymbol{\beta}_j \backslash \{\beta_{j,i_1}, ..., \beta_{j,i_m}\})$ as a subset model of $\mathcal{M}(\boldsymbol{\beta}_j)$, where $\beta_{j,i_1} = ... = \beta_{j,i_m} = 0$. We apply the following algorithm to find the optimal subset model for the j-th series.

Top-down strategy

1. initialisation: $\mathcal{A} = \boldsymbol{\beta}_j, i = kp + 1$,
2. if $AIC(\mathcal{M}(\mathcal{A} \backslash \beta_{j,i})) \leq AIC(\mathcal{M}(\mathcal{A}))$, update $\mathcal{A}$ to $\mathcal{A} \backslash \beta_{j,i}$,
3. update i to i - 1,
4. repeat steps 2-3 until $i = 0$.

We end up with a subset for the j-th series that contains all exploratory variables with non-zero coefficients in $\mathcal{A}$. This algorithm begins by taking all parameters into account. Then it evaluates the quality of a parameter with a criterion value. Hence, when the exclusion of a parameter improves the information criterion value of the model, the aforementioned parameter is permanently eliminated.

We now introduce the algorithm of the other searching procedure to find the optimal subset model for the j-th series.

Bottom-up strategy

1. initialisation: $\mathcal{A} = \emptyset,\ i = 1$,
2. maintain the explanatory variables in $\mathcal{A}$ and add lags of the i-th series as additional explanatory variables, then select the optimal order for the new series based on AIC, call this order $p_{j,i}$,
3. update $\mathcal{A}$ by including the lag variables of the i-th series up to lag $p_{j,i}$,
4. update i to $i + 1$,
5. repeat steps 2–4 until $i = k$.

We obtain a model for the j-th series that contains variables in $\mathcal{A}$ that includes the lag variables of i-th series up to lag $p_{j,i}$ for $i = 1, 2, ..., k$. That is,

$$y_{jt} = v_j + \sum_{l=1}^{p_{j,1}} \phi_{1,j,l} y_{1,t-l} + \sum_{l=1}^{p_{j,2}} \phi_{2,j,l} y_{2,t-l} + ... + \sum_{l=1}^{p_{j,k}} \phi_{k,j,l} y_{k,t-l} + u_{jt},$$

for some coefficients v_j and $\{\phi_{i,j,l}\}$. In essence, this algorithm begins from the empty set and then sequentially adds a univariate AR structure from the multiple series.

The results of the top-down and bottom-up strategies are often suboptimal in practice and altered by the dependence of the created search paths. Or in other words, the order of the series in the system. An excellent model selection method should not depend on this order and always generate an optimal subset. That is why searching procedures are only a great starting point.

2.2.7 Hybrid procedures

The aforementioned model selection techniques are often quite distinct in their approach. This offers the possibility to combine certain strategies in the hope to find a new method with better performance. This includes the information-based criteria, top-down, bottom-up and methods involving the Lasso technique. Following the work of Hsu et al. (2008), we implement the following hybrid procedures:

Bottom-up + top-down strategy (BU + TD)

1. implement the bottom-up strategy to built a model for each series.
2. implement the top-down strategy to reduce this model for each series.

AIC + top-down strategy (AIC + TD)

1. use AIC to determine the order for VAR model fitting when multiple series are considered simultaneously.
2. implement the top-down strategy to reduce the VAR($\hat{p}_{aic}$) model for each series.

AIC + Lasso for multiple series (AIC + Lasso-f)

1. use AIC to determine the order for VAR model fitting when multiple series are considered simultaneously.
2. find Lasso estimates for multiple series (full system) under the VAR($\hat{p}_{aic}$) model.

AIC + Lasso for individual series (AIC + Lasso-s)

1. use AIC to determine the order for VAR model fitting when multiple series are considered simultaneously.
2. find Lasso estimates for each individual series under the VAR($\hat{p}_{aic}$) model.

One may notice the different notation in the form of Lasso-f and Lasso-s. These are distinct techniques. The former, Lasso-f, denotes when we apply the Lasso technique on the full system where all series are taken into account simultaneously. Lasso-s on the other hand, is when we apply the Lasso technique for each series separately, also known as single equation Lasso. By introducing these variants of the Lasso, we add an extra dimension to our research. We implement a similar single equation v. full system framework for all other techniques that are discussed in Section 2.2.2 to 2.2.5.

2.3 Evaluation criteria

We evaluate the performance of the predictions by implementing the empirical prediction mean squared error (PMSE). The PMSE is the expected value of the squared difference between the fitted values and the observed values. That is why it is an excellent measurement of the explanatory power of an estimated model. As we discuss more in-depth in Section 3, our data set consists of three individual series. This means that our VAR model is three dimensional. Furthermore, we have 176 in-sample observations and evaluate nine out-of-sample observations. We define the PMSE for individual series and all series as follows:

$$\text{PMSE}_i = \frac{1}{9}\sum_{h=1}^{9}(y_{i,176+h} - \hat{y}_{i,176+h})^2, \qquad i = 1, 2, 3, \tag{10}$$

$$\text{PMSE}_{all} = \frac{1}{9}\sum_{h=1}^{9}(y_{176+h} - \hat{y}_{176+h})' \; \hat{\Sigma}_{aic}^{-1} \; (y_{176+h} - \hat{y}_{176+h}), \tag{11}$$

where $\hat{\Sigma}_{aic}$ is the estimation of the covariance matrix Σ_u based on AIC. We use this estimate in equation (11) to adjust the scales for different series. We observe in equation (10) and (11) that we compute our prediction errors as the sum of all differences between the real values and the n-step predictions. To realise this, we implement a moving window. This means that the sample period which we use to make predictions, moves one period per prediction. The sample size however, does not change. Essentially, after we generated a prediction, we add this most recent prediction to our sample period and remove the observation of the oldest time period. We do not re-estimate the coefficients given to certain lagged variables.

Additionally, we report the proportion of zero. For every method, we start with a p order VAR model. Then following some subset selection technique, we remove redundant variables. This criterion represents the fraction of coefficients that have been set to zero after a subset selection procedure.

3 Data

The data set that we use to evaluate the various techniques consists of several indicators of the U.S. economy. More precisely, we will examine a model that captures the relationship of the nominal GDP, the unemployment rate and the gross rate of M1. The nominal GDP is a seasonally adjusted annual rate that is expressed as a percent change from a year ago. The unemployment rate is seasonally adjusted and expressed in percentages. And lastly, the gross rate of M1 is not seasonally adjusted and is, similar to the GDP, expressed as a percent change from a year ago. We select these three indicators because we assume that they are strongly correlated. During economic expansions, we expect the gross rates of GDP and M1 to increase and the unemployment rate to decrease. During economic recessions, we expect the exact opposite. This strong correlation between the time series would be a great environment for a VAR process to excel in prediction performance. This data set consists of three-dimensional time series and contains observations for the period January 1960 to March 2006. We choose a large sample period as we want to incorporate the effect of all rare economic circumstances, such as wars, economic crises and other events of this nature. Therefore, we do not use structural breaks or time variation in our data set. This creates an utmost realistic perspective and gives us a higher probability to draw relevant and reliable conclusions about the methods that we implement.

All three indicators are retrieved from FRED.[1] We favour quarterly observations. For the unemployment and M1 rate, we use the end of the period and not the average of a quarter as the aggregation method. This does not concern the nominal GDP. We obtain a sample size n = 185. Our in-sample period is n = 176, which automatically makes our out-of-sample period n = 9. We use the first 176 observations to develop and improve the model that estimates the nine out-of-sample observations.

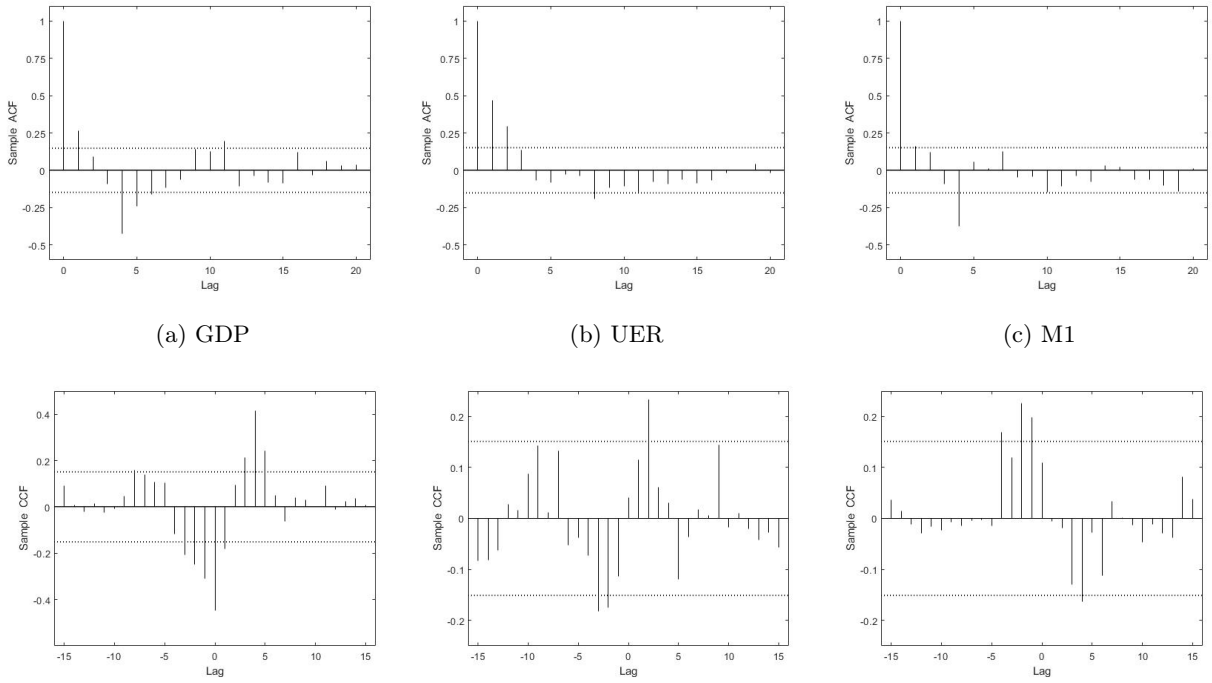
Generally, almost all statistical procedures in time series analysis are built on the assumption of stationarity. When we look at the first-order differenced time series, we observe a 'saw-tooth' pattern. This could indicate non-stationarity. To test whether our times series carry this feature, we apply the Dickey–Fuller test. The result suggests that there is substantial evidence that there is non-stationarity present in all individual series. Therefore, we alter each time series by implementing first-order differencing. This will dispose of any non-stationarity in our data set. The graphs of the original and differenced series during the in-sample period can be found in Appendix A.

As mentioned earlier, the data set is selected based on the expected strong correlation between the time series. To confirm this expectation, we calculate the sample autocorrelation function (ACF) and the cross-correlations function (CCF) of the three time series. The results are shown below in Figure 1. We observe in Figure 1(a)-(c) that the autocorrelation values are often significantly different from zero. This suggests that there is dependence between the lags within each series. The values of the CCF in Figure 1(d)-(f) are often significantly different from zero as well. This represents dependence between the lags between the three

[1] https://fred.stlouisfed.org/ The Federal Reserve Economic Data is a database revolving around the U.S. economy. FRED is owned and maintained by the Research department of the Federal Reserve Bank of St. Louis.

series. Overall, we observe strong correlation within and between the three series in our data set, which suits the VAR structure very well.

Figure 1: Sample ACF (a)-(c) and sample CCF (d)-(f) for the difference series of the GDP, UER, and M1.



(a) GDP (b) UER (c) M1

(d) UER v. GDP (e) M1 v. GDP (f) M1 v. UER

Note. The horizontal axis represents the amount of lags. The data ranges from the in-sample period 1960Q2-2003Q4. We observe that the ACF and CCF are often significantly different from zero.

4 Results

As we explained in Section 2.1, VAR models incorporate the lag values of the outcome variables as the predictor variables. When we apply model selection techniques as the bottom-up or top-down strategies, we emphasize the importance of the construction of the data set, which impacts the structure of the VAR model. This is due to the fact that the aforementioned methods implement a given model in a consistent manner. For example, the top-down strategy starts the initialisation process by beginning at the coefficient of the highest order lag of the last series in the model. When the order of the series in the model is altered, it could have a major impact on the evaluation of the model and therefore the optimal model that this strategy generates. This research chooses to select the order GDP, then UER and lastly M1.

In the following tables we denote several characteristics of the created models using various model selection techniques. We first report the order of the VAR model. This represents the optimal order of lags p that we include in the generated model. For computational purposes, we do not allow lag orders larger than the set maximum, which we denote with $p_{\max}$. In this research, we set $p_{\max} = 10$. Then, we denote the proportion of zero. And finally, we report the PMSE and how they rank compared to the other techniques. For implementations of all aforementioned penalised regression techniques, this paper uses McIlhagga (2016).[2]

4.1 Basic models

Table 1: Characteristics of various model selection methods.

	AIC	BIC	AIC+TD	BU+TD	Lasso-s	Lasso-f	AIC+Lasso-s	AIC+Lasso-f
Selected VAR order	8	4	8	9	10	10	8	8
proportion of zero	-	-	0.587	0.688	0.344	0.473	0.240	0.333
$PMSE_1$	0.474	0.219	0.306	0.262	0.272	0.348	0.230	0.320
Rank	8	1	5	3	4	7	2	6
$PMSE_2$	0.048	0.038	0.027	0.031	0.032	0.029	0.029	0.029
Rank	8	7	1	5	6	2	4	3
$PMSE_3$	3.089	2.222	2.596	2.667	2.699	2.862	1.897	2.190
Rank	8	3	4	5	6	7	1	2
$PMSE_{all}$	3.467	2.392	2.547	2.616	2.714	2.855	2.075	2.403
Rank	8	2	4	5	6	7	1	3

Note. The results shown in this table are the selected VAR order, proportion of variables which coefficient has been set to zero, the PMSE and the corresponding rankings of each model selection technique. These are computed using quarterly differenced U.S. macroeconomic data where all methods are evaluated out-of-sample.

Looking at Table 1, we notice several interesting findings. We begin with the Lasso procedure. It is clearly visible that combining the Lasso procedure with an information criteria improves the performance tremendously. Especially, regarding the M1 series and all series together. We observe that the intuitive methods, such as the BU and TD generate good results with stable rankings. They roughly outperform the Lasso-s and Lasso-f in each PMSE evaluation, which is rather impressive. However, the two hybrid procedures involving the Lasso often outperform all of our benchmark models. This does not hold up for every series. Therefore, we cannot state that the Lasso method is superior. Based on Table 1, we do strongly prefer hybrid procedures that incorporate the Lasso. We state that the AIC combined with the Lasso for single equations has the best performance. In some series this method does not have the number one rank. On closer inspection,

[2]https://www.jstatsoft.org/article/view/v072i06 provides a flexible, extensible, and efficient MATLAB toolbox for penalised regressions.

we do observe that for these series the differences between the PMSE of the number one method and the AIC+Lasso-s are relatively modest.

Starting at a lower p order VAR model seems to benefit the effectiveness of the Lasso procedure. We recognise that the Lasso-f eliminates more variables relative to the Lasso-s. This could be due to the fact that the Lasso-f selects variables from a substantially larger set than the Lasso-s. It could be that certain variables are necessary in a single equation model, yet replaceable when other variables are taken into account. When the Lasso starts with a lower order VAR model, we notice that the proportion to zero decreases. This is expected as higher lagged variables commonly have a minor impact compared to lower lagged variables. Moreover, in this situation the total set of variables decreases as well. This means that proportionally more variables add value to the model. And lastly looking at the PMSE, we argue that generally in the context of Table 1 single equation Lasso outperforms full system Lasso.

To a certain extent, the findings in Table 1 are quite in line with Hsu et al. (2008). Similar to this research, we recommend a hybrid procedure involving the Lasso. Hsu et al. (2008) found as well that the standard Lasso is often outperformed by the benchmark models. Overall, the selected VAR orders and the proportions of zeroes are considerably similar. However, one major difference is the numerical values of all PMSE, which are rarely identical. This is rather unexpected since this research implements the exact same data set as Hsu et al. (2008). These numerical differences are proportional for most methods. This means that even though the values of PMSE do not match, the rankings are relatively constant. Another distinction that needs to be made between this research and Hsu et al. (2008), revolves around the best performing technique. Where we advocate the AIC+Lasso-s, Hsu et al. (2008) undoubtedly finds the best results when implementing the AIC+Lasso-f method.

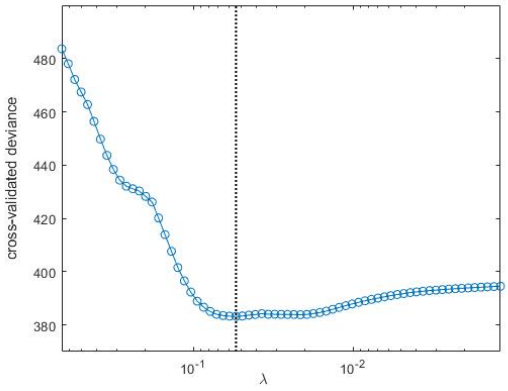
The differences between the results generated by this paper and Hsu et al. (2008) can occur for many reasons. One of them is a possible difference in the data set. This could mean a small alteration in the cut-off date, or even slightly adjusted data due to a different computation of the units in which the data is expressed. For example, when we excessively analyse the graph of the first differences of the GDP in Figure 3 in Appendix A. We observe small, yet visible differences, e.g. kinks, with Fig. 2.d in Hsu et al. (2008) which should present the same graph. Another difference regarding the data set is found in the sample cross-correlations function of M1 v. GDP. Fig 3.e in Hsu et al. (2008) shows that several lags are significantly different from zero. However, some of these lags are not significantly different from zero in our research, see 1 panel e. Since we do not find these irregularities for all series, we assume that there might be one series that is slightly different from Hsu et al. (2008). This small difference can eventually have major impact on model selection through some type of ripple effect.

Another possibility could be that Hsu et al. (2008) computes the methods in a different matter. For example, small alterations in algorithms can have a major impact on generated models. However, due to lack

of information in Hsu et al. (2008), we cannot further argue this claim.

Additionally, one may notice that the p order of lags is considerably higher for the AIC than the BIC. This is due to the fact that the BIC penalises the complexity of a model more heavily than the AIC. In other words, BIC is stricter in preventing overfitting than the AIC. Since both methods only decide the optimal lag order p, there is no evaluation of individual variables. The PMSE performance of the BIC is not only noteworthy when compared to the AIC. When we look at the entire table, we observe that the PMSE performance of the BIC is outstanding. Therefore after analysing the first two columns, one can wonder why none of the hybrid procedures that we introduced in Section 2.2.7 incorporate the BIC. The reason for this is quite simple. We believe that by reducing the VAR order from ten to four, the latter part of the hybrid procedure is too restricted. Otherwise, it could be challenging to efficiently determine the effect of the second part of the hybrid procedure. That is why we choose AIC. This works as a conservative first step, which offers us the opportunity to test our methods on higher lagged variables. Nonetheless, the BIC acts as an excellent benchmark, since it is an uncomplicated method and not expensive to execute. We add Table 3 in Appendix B, which is similar to Table 1. In this table we replace the AIC in hybrid procedures with the BIC. These techniques produce remarkable results which are quite interesting, yet not essential to the objective of this research.

Figure 2: Cross-validation error for the AIC+Lasso-f method.



Note. The horizontal axis represents the lambda sequence. The data ranges from the in-sample period 1960Q2-2003Q4. The vertical line indicates the lambda value that generates the minimum cross-validation error.

We observe in Figure 2 an example of how the cross-validation error alters for different values of λ. The λ that we incorporate in the corresponding model, is selected based on which λ generates the lowest cross-validation error.

4.2 Advanced models

In Table 2, we observe the results generated by more advanced model selection techniques. We added the results of a penalised regression where all λ are set equal to zero. Additionally, we added the best performing method from Table 1, which is the AIC combined with the single equation Lasso. These two methods acts as reliable benchmarks, which we compare to the advanced techniques.

Table 2: Characteristics of various advanced model selection methods.

	VAR order	prop. zero	$PMSE_1$	R.	$PMSE_2$	R.	$PMSE_3$	R.	$PMSE_{all}$	R.
No penalty	10	-	0.458	14	0.066	16	4.444	16	4.741	16
AIC+Lasso-s	8	0.240	0.230	4	0.029	7	1.897	5	2.075	4
SCAD-s	10	0.495	0.220	1	0.030	10	2.126	10	2.191	7
SCAD-f	10	0.495	0.536	16	0.030	9	4.190	15	4.050	15
Adap-Lasso-s	10	0.366	0.312	10	0.035	15	3.204	14	3.055	13
Adap-Lasso-f	10	0.473	0.414	13	0.028	3	3.166	13	3.140	14
AIC+Adap-Lasso-s	8	0.467	0.246	5	0.029	5	2.039	8	2.082	6
AIC+Adap-Lasso-f	8	0.533	0.303	9	0.031	11	1.949	7	2.242	9
E-net-s	10	0.301	0.270	7	0.032	12	2.681	11	2.704	11
E-net-f	10	0.376	0.326	11	0.029	4	2.833	12	2.827	12
AIC+E-net-s	8	0.213	0.230	3	0.029	6	1.856	4	2.051	3
AIC+E-net-f	8	0.240	0.292	8	0.028	2	2.089	9	2.304	10
Adap-E-net-s	10	0.534	0.500	15	0.035	14	1.353	1	2.220	8
Adap-E-net-f	10	0.720	0.359	12	0.033	13	1.417	2	2.000	1
AIC+Adap-E-net-s	8	0.547	0.222	2	0.028	1	1.908	6	2.003	2
AIC+Adap-E-net-f	8	0.613	0.248	6	0.029	8	1.837	3	2.080	5

Note. The results shown in this table are the selected VAR order, proportion of variables which coefficient has been set to zero, the PMSE and the corresponding rankings of each model selection technique. These are computed using quarterly differenced U.S. macroeconomic data where all methods are evaluated out-of-sample.

By adding weights in the penalty term, the adaptive Lasso should eliminate the problems that could occur when applying the Lasso. For the adaptive Lasso in Table 2, the proportion of variables that has been set to zero increases as the order of the VAR decreases, which is rather unexpected. We assume that the impact of the weights in the penalty term influence the coefficients in the lower order model by such an amount that relatively more variables are deemed redundant. We found that for the Adap-Lasso-s and Adap-Lasso-f γ

= 0.5 produces the best PMSE. For AIC+Adap-Lasso-s and AIC+Adap-Lasso-f we found that $\gamma = 1$ is the best choice. We observe that the AIC+Adaptive-Lasso-s generates the overall best adaptive Lasso results. However, for the adaptive Lasso the difference between the single equation and full system framework is less apparent than it was in Table 1. Moreover, the addition of the AIC to the adaptive Lasso substantially improves the generated PMSE.

Additionally, we observe that the single equation SCAD produces relatively stable predictions. Especially, the performance for the GDP stands out. The full system SCAD however produces predictions that are close to a model with no penalty parameter, which is rather undesirable.

By adding the ridge regression component to the Lasso procedure, the elastic net is formed. We find that the proportion of zeroes in Table 2 decreases as the order of the VAR model decreases, as we would expect. However, in this instance we notice that the hybrid procedures generate a modest proportion to zero compared to former techniques. We observed that for the two single equation elastic net techniques, an α of 0.75 generates the best results. For the full system elastic net, we set α equal to 0.5. Similar to before, we detect that the hybrid procedures have superior performance over the standard methods. Regarding the elastic net, we state that the AIC+E-net-s has the best performance in terms of PMSE.

The adaptive elastic net is a convex combination of the adaptive Lasso and ridge regression. Looking at Table 2, we immediately notice the large proportions of zero. Especially compared to other methods that involve the Lasso. This results in a smaller and therefore simpler model which is incredibly desirable. Similar to other tables, we see the benefits that hybrid theories bring. However, regarding the third series (M1), the standard methods have an outstanding performance. This could be due to some rare computational occurrence involving α and γ. Table 2 shows that the standard methods perform relatively poor for other series than M1 and all series together. We observe in the third column that Adap-E-net-s has a substantially higher PMSE than a model with no penalty parameter at all, which is quite unusual. This statement is supported by the fact that we find the best PMSE for Adap-E-net-s and Adap-E-net-f when $\alpha = 0.25$ and $\gamma = 2$. While on the other hand, the best results are obtained for AIC+Adap-E-net-s and AIC+Adap-E-net-f when $\alpha = 0.5$ and $\gamma = 0.5$. Therefore, we assume for all methods in Table 2 that for a relatively great performance for one series, it is not uncommon to perform relatively poor for different series. Overall, we observe in Table 2 that AIC+Adap-E-net-s generates the best results.

As discussed in Section 3, we find our results with an in-sample size of n = 176. It is quite possible that certain methods thrive on larger sample sizes, or the exact opposite. Therefore, as a robustness check, we added two tables where we divide the sample size in two. Both tables are found in Appendix C. Table 4 contains the in-sample time period January 1960 to December 1984. Whereas Table 5 contains the in-sample time period January 1985 to December 2003. The reason for these different in-sample sizes is that if we would divide the full in-sample size in half, our first out-of-sample period starts during the rise of a large-scale economic crisis. We find that the results in Table 2 are not as robust as we would prefer. We observe that in both sample

periods, the adaptive elastic net still generates the best predictions. However, in the context of Table 4 and 5, we favour the full system framework over the single equation framework. Additionally, we see that hybrid procedures do not increase the performance in the same manner as in Table 2.

5 Conclusion

This paper researches the effectiveness of various model selection techniques in VAR processes. By comparing past and modern methods that generally increase in complexity, this research offers an extensive analysis of the development of penalised regressions. We investigate if certain methods have superior performance over others. Furthermore, we look at possible combinations of techniques and evaluate whether these hybrid procedures produce promising results. Hsu et al. (2008) and Ren et al. (2010) laid the foundation on which this research is built. We implement the Lasso, adaptive Lasso, SCAD, elastic net, adaptive elastic net, hybrid procedures and various old-fashioned model selection techniques. Our data set consists of several quarterly U.S. macroeconomic time series from the period January 1960 to March 2006.

The results generated by the models in this paper present a substantial amount of information. First and foremost, we observe that hybrid theories, in the form of information criteria, considerably improve the performance of model selection techniques. We find that for a large sample size the combination of AIC and the adaptive elastic net for single equations has the best overall prediction performance. This technique possesses a rather large proportion of zeroes as well, which represents a smaller and simpler model. VAR models are by definition relatively large in size and consist of a substantial number of parameters. Therefore, a superior model selection technique for VAR processes is evaluated on prediction performance as well as the simplicity of the model. Since the AIC and adaptive elastic net for single equations method show excellent results for both these criteria, we strongly favour this technique. However, it remains a difficult task to conclude the superiority of one technique when it does not outperform all methods for all series. This is due to the fact that it is challenging to comprehend to which extent a great performance of a method at one series relates to a relatively poor performance at a different series.

Nonetheless, this research prefers the single equation framework over the full system framework for all methods. This is quite surprising, considering the strongly correlated data set in combination with the use of VAR models. We would expect that in the full system framework, the VAR model would excel due to its ability to capture the linear interdependencies among multiple time series. However, the evidence indicates that this is, in terms of prediction performance, more of a liability than an asset.

6 Discussion

Similar to Ren et al. (2010), the AIC combined with the single equation Lasso has excellent performance compared to the more advanced techniques. This brings us to a significant limitation in this research. When implementing the adaptive Lasso, SCAD, the elastic net and the adaptive elastic net, the outcome of these

methods are incredibly dependent on the values given to parameters α, a and γ. Since we do not apply a grid search where we change these parameters to hyperparameters, the flexibility for α, a and γ are limited. This tells us that the results that we obtained for these techniques are only an extremely limited set of all possible results. Therefore, these models most probably generate substantially lower PMSE values when implementing optimal α, a and γ. This supports our claim that modern, complex hybrid techniques produce superior predictions.

This research has several more limitations. For example, we set $p_{\max}$ equal to ten for computational purposes. However, none of our advanced methods eliminate all parameters that belong to the maximum lag order in that particular context. This could indicate that there exist relatively better fitted VAR models when we select a larger value for $p_{\max}$. Furthermore, we question the robustness of our results for advanced model selection methods. Dividing the in-sample period in two produces similar results, yet we draw slightly different conclusions. Therefore, it can be relevant to study the sample size effects more in-depth.

One could argue as well that model evaluation is a complex matter, and that the proportion of redundant variables and prediction accuracy are not telling the full story about the quality of a model selection method. Especially, regarding the objective of this paper to offer an extensive comparative analysis. By adding several statistical hypothesis tests, for instance the Diebold-Mariano test, the assertions made in this paper are more substantiated. Additionally, it could be interesting to further explore the performance of variations of the Lasso or elastic net in combination with different information criteria, e.g. BIC or the HQ criterion. By implementing a similar research with various data sets, one can argue whether certain information criteria are more suited for hybrid procedures in model selection or if it is completely data dependent.

Finally, the results presented by this paper could ignite new interest in subset selection for VAR processes. Since VAR models are popular in the financial world, any innovative ideas in this research area could create a great deal of real world applications. Excluding analysing the possibilities to solve the aforementioned limitations, there are still quite some ventures to explore. For example, Song et al. (2011) distinguishes various types of lags in large VAR models. They handle the own lags of a variable differently than other lags and implement different regularisation for different lags over time. Furthermore, one could explore alternative penalised regression techniques. For instance, the group Lasso as discussed in Yuan et al. (2006), which allows predefined groups to be selected into or out of a model together. One could additionally look into the prior Lasso, introduced in Jiang et al. (2016), which incorporates prior information in the penalised regression. This method outperforms Lasso by a significant amount when the prior information is relatively accurate. Lastly, since the introduction of the successful adaptive elastic net, this technique has been further developed in several researches. Xiao et al. (2015) create a multi-step adaptive elastic net estimation algorithm. Gefang (2012) constructs a Bayesian doubly adaptive elastic net Lasso approach that is designed for VAR processes. This leads us to an intriguing venture called Bayesian VAR (BVAR). The BVAR is an increasingly popular framework for preventing overfitting and initiates numerous opportunities to research.

References

Akaike. H. (1974). "A New Look at the Statistical Model Identification". *IEEE Transactions on Automatic Control*, 19(6): pp. 716–723, 1974.

Brüggemann, R., and H. Lütkepohl. (2000). "Lag Selection in Subset VAR Models with an Application to a U.S. Monetary System". *Econometric Society*, 8(21): pp. 107–128, 2000.

Burnham, K., and D. Anderson. *"Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach" - Second Edition.* §6.3. Springer-Verlag, 2002.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani. (2010). "Least Angle Regression". *The Annals of Statistics*, 32(2): pp. 407–499, 2004.

Fan, J., and R. Li. (2001). "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties". *Journal of the American Statistical Association*, 96(456): pp. 1348–1360, 2001.

Fuller. *"Introduction to Statistical Time Series" - Second Edition.* §8. John Wiley and Sons, 1996.

Gefang. D. (2012). "Bayesian Doubly Adaptive Elastic-Net Lasso For VAR Shrinkage". *International Journal of Forecasting*, 30(1): pp. 1–11, 2012.

Hsu, N., H. Hung, and Y. Chang. (2008). "Subset selection for vector autoregressive processes using Lasso". *Computational Statistics and Data Analysis*, 52: pp. 3645–3657, 2008.

Jiang, Y., Y. He, and H. Zhang. (2016). "Variable Selection With Prior Information for Generalized Linear Models via the Prior Lasso Method". *Journal of the American Statistical Association*, 111(513): pp. 355–376, 2016.

Lütkepohl. *"New Introduction to Multiple Time Series Analysis".* Springer-Verlag, 2005.

McIlhagga. W. (2016). "penalized: A MATLAB Toolbox for Fitting Generalized Linear Models with Penalties". *Journal of Statistical Software*, 72(6), 2016.

Meinshausen, N., and P. Bühlmann. (2006). "High-dimensional Graphs and Variable Selection with the Lasso". *The Annals of Statistics*, 34(3): pp. 1436–1462, 2006.

Ren, Y., and X. Zhang. (2010). "Subset selection for vector autoregressive processes via adaptive Lasso". *Statistics and Probability Letters*, 80: pp. 1705–1712, 2010.

Schwarz. G. (1978). "Estimating the Dimension of a Model". *The Annals of Statistics*, 6(2): pp. 461–464, 1978.

Sims. C. (1980). "Macroeconomics and Reality". *Econometrica*, 48(1): pp. 1–48, 1980.

Song, S., and P.J. Bickel. (2011). "Large Vector Auto Regressions". *ArXiv e-prints arXiv:1106.3915*, 2011.

Tibshirani. R. (1996). "Regression Shrinkage and Selection via the Lasso". *Journal of the Royal Statistical Society*, 58(1): pp. 267–288, 1996.

Tikhonov, A.N., and V.Y. Arsenin. *"Solutions of Ill-Posed Problems"*. Winston, 1977.

Xiao, N., and Q. Xu. (2015). "Multi-step Adaptive Elastic-net: reducing False Positives in High-dimensional Variable Selection". *Journal of Statistical Computation and Simulation*, 85(18): pp. 3755–3765, 2015.

Yuan, M., and Y. Lin. (2006). "Model Selection and Estimation in Regression with Grouped Variables". *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 68(1): pp. 49–67, 2006.

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

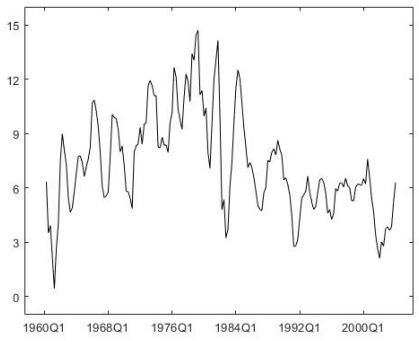
Zou, H., and T. Hastie. (2005). "Regularization and Variable Selection via the Elastic Net". *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2): pp. 301–320, 2005.

Zou, H., and H. Zhang. (2009). "On the Adaptive Elastic-Net with a Diverging Number of Parameters". *The Annals of Statistics*, 37(4): pp. 1733–1751, 2009.

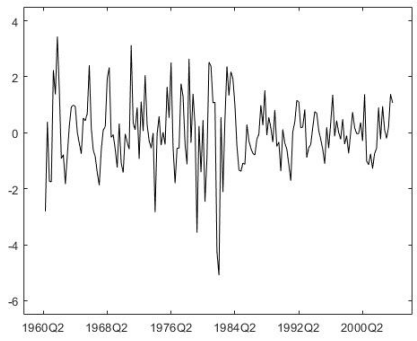
Appendices

A Graphs of individual time series

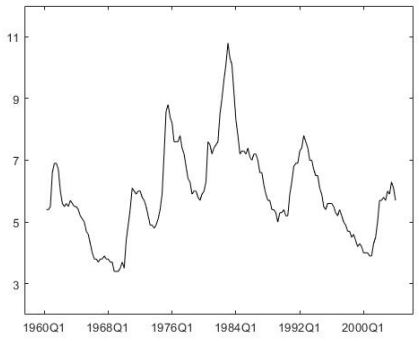
Figure 3: Quarterly nominal GDP, unemployment rates and the gross rates of M1 from the in-sample period 1960Q1 - 2003Q4. The original series are on the left side and the differenced series are on the right side.



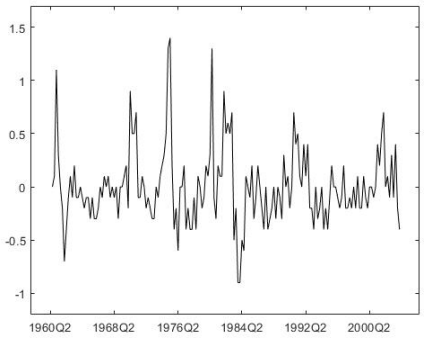
(a) GDP



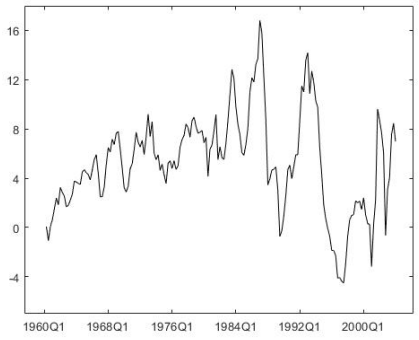
(b) differenced GDP



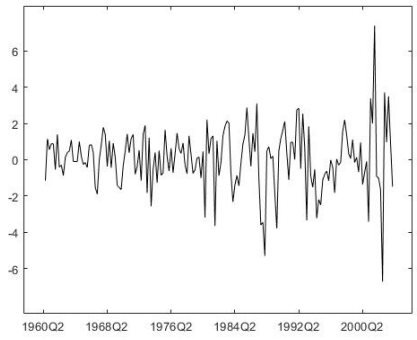
(c) UER



(d) differenced UER



(e) M1



(f) differenced M1

B Basic model selection methods implementing BIC

Table 3: Characteristics of various model selection methods.

	AIC	BIC	BIC+TD	BU+TD	Lasso-s	Lasso-f	BIC+Lasso-s	BIC+Lasso-f
Selected VAR order	8	4	4	9	10	10	4	4
proportion of zero	-	-	0.667	0.688	0.344	0.473	0.385	0.538
$PMSE_1$	0.474	0.219	0.233	0.262	0.272	0.348	0.199	0.211
Rank	8	3	4	5	6	7	1	2
$PMSE_2$	0.048	0.038	0.036	0.031	0.032	0.029	0.028	0.031
Rank	8	7	6	3	5	2	1	4
$PMSE_3$	3.089	2.222	1.664	2.667	2.699	2.862	1.700	1.781
Rank	8	4	1	5	6	7	2	3
$PMSE_{all}$	3.467	2.392	1.989	2.616	2.714	2.855	1.864	1.989
Rank	8	4	2	5	6	7	1	3

Note. The results shown in this table are the selected VAR order, proportion of variables which coefficient has been set to zero, the PMSE and the corresponding rankings of each model selection technique. These are computed using quarterly differenced U.S. macroeconomic data where all methods are evaluated out-of-sample. The hybrid procedures in this table have incorporated the BIC instead of the AIC as in Table 1.

C Robustness of advanced models

Table 4: Characteristics of various advanced model selection methods for the in-sample period January 1960 to December 1984.

	VAR order	prop. zero	$PMSE_1$	R.	$PMSE_2$	R.	$PMSE_3$	R.	$PMSE_{all}$	R.
No penalty	10	-	1.175	16	0.302	16	3.681	16	13.043	16
AIC+Lasso-s	8	0.413	0.870	15	0.126	13	2.627	2	7.969	13
SCAD-s	10	0.753	0.512	8	0.050	1	2.743	8	6.762	2
SCAD-f	10	0.742	0.832	14	0.057	4	3.245	14	8.045	14
Adap-Lasso-s	10	0.567	0.459	6	0.102	11	3.167	13	8.153	15
Adap-Lasso-f	10	0.688	0.319	1	0.055	3	3.333	15	7.710	10
AIC+Adap-Lasso-s	8	0.600	0.517	9	0.102	12	2.697	6	7.413	8
AIC+Adap-Lasso-f	8	0.653	0.756	13	0.082	9	2.708	7	7.439	9
E-net-s	10	0.441	0.434	4	0.067	6	2.873	11	7.061	4
E-net-f	10	0.419	0.439	5	0.070	7	2.897	12	7.184	6
AIC+E-net-s	8	0.280	0.671	11	0.126	14	2.591	1	7.892	11
AIC+E-net-f	8	0.480	0.641	10	0.078	8	2.636	3	7.135	5
Adap-E-net-s	10	0.688	0.394	3	0.129	15	2.800	10	7.948	12
Adap-E-net-f	10	0.785	0.346	2	0.054	2	2.748	9	6.492	1
AIC+Adap-E-net-s	8	0.640	0.465	7	0.102	10	2.659	5	7.363	7
AIC+Adap-E-net-f	8	0.693	0.680	12	0.058	5	2.636	4	6.777	3

Note. The results shown in this table are the selected VAR order, proportion of variables which coefficient has been set to zero, the PMSE and the corresponding rankings of each model selection technique. These are computed using quarterly differenced U.S. macroeconomic data where all methods are evaluated out-of-sample.

Table 5: Characteristics of various advanced model selection methods for the in-sample period January 1985 to December 2003.

	VAR order	prop. zero	$PMSE_1$	R.	$PMSE_2$	R.	$PMSE_3$	R.	$PMSE_{all}$	R.
No penalty	10	-	0.340	16	0.078	16	7.540	16	10.255	16
AIC+Lasso-s	8	0.480	0.172	9	0.037	11	4.242	11	5.655	12
SCAD-s	10	0.677	0.172	10	0.020	2	4.105	10	4.780	8
SCAD-f	10	0.817	0.164	8	0.025	7	5.716	15	5.399	11
Adap-Lasso-s	10	0.366	0.207	14	0.047	13	2.967	2	5.945	13
Adap-Lasso-f	10	0.753	0.131	4	0.025	8	3.421	6	3.715	3
AIC+Adap-Lasso-s	8	0.293	0.205	13	0.029	10	3.611	8	4.734	7
AIC+Adap-Lasso-f	8	0.680	0.143	5	0.022	5	3.474	7	3.761	4
E-net-s	10	0.419	0.172	11	0.057	14	4.600	13	7.420	15
E-net-f	10	0.452	0.114	2	0.019	1	4.783	14	4.402	6
AIC+E-net-s	8	0.133	0.192	12	0.026	9	4.559	12	4.909	10
AIC+E-net-f	8	0.467	0.150	6	0.019	3	3.945	9	3.945	5
Adap-E-net-s	10	0.398	0.267	15	0.060	15	2.834	1	6.765	14
Adap-E-net-f	10	0.774	0.122	3	0.024	6	3.360	5	3.587	2
AIC+Adap-E-net-s	8	0.413	0.159	7	0.039	12	3.275	4	4.806	9
AIC+Adap-E-net-f	8	0.653	0.110	1	0.020	4	3.117	3	3.319	1

Note. The results shown in this table are the selected VAR order, proportion of variables which coefficient has been set to zero, the PMSE and the corresponding rankings of each model selection technique. These are computed using quarterly differenced U.S. macroeconomic data where all methods are evaluated out-of-sample.