Macroeconomic Forecasting using Large Vector Auto Regressive Model

Master Thesis Submitted to

Prof. Dr. Wolfgang K. Härdle

Dr. Song Song

Ladislaus von Bortkiewicz Chair of Statistics

C.A.S.E.- Centre for Applied Statistics and Economics

Humboldt-Universität zu Berlin

by

Ye Hua

(521952)

in partial fulfillment of the requirements

for the degree of

Master of Economics and Management Science

Berlin, July 15th, 2011

Acknowledgment

I would like to express my deep and sincere gratitude to my supervisor, Prof. Dr. Wolfgang K. Härdle, for his important support and encouragement throughout this work. Many thanks also to my advisor, Dr. Song Song, for his constructive criticism and insightful comments in all the time of writing my thesis which definitely enabled me to develop a deep understanding of the subject.

I would like to extend my thanks to my dear friends, Martin Schelisch and Ceren Önder, who have made available their kind support in a number of ways. I am indeed grateful to them for spending their time in helping me to finish this work with the best possible results. Their loving support has been of great value.

Last but not the least, I would like to thank my parents, Xu Hua and Bing Liu, for supporting me spiritually throughout my life. Without their understanding and encouragement it would have been impossible for me to accomplish this work.

Ye Hua

Abstract

Many macroeconomic problems require the exploitation of large panels of time series. Large vector auto regressions (large VAR), a more integrated method for high dimensional time series, can be applied to conduct impulse response studies, label different time series individualized "endogenous" and "exogeneous" and do the variable selection and lag selection simultaneously. This work investigates the application of large VAR for addressing the challenges coming from a mixture of high dimensional dependence structure, serial correlation and moderate sample size. We introduce the model with three kinds of regularizations first, then discuss the data driven choice of tuning parameters to optimize the forecasting performance, provide a numerical algorithm, and finally summarize comparisons among these regularizations. The new approach is shown to enjoy the oracle properties. Empirical analysis using large panel macroeconomic time series is considered to illustrate this new method outperforms the existing approach.

Keywords: Time Series, Vector Auto Regression, Regularization, Lasso, Group Lasso,Oracle estimator
JEL classification: C13, C14 ,C32, E30, E40, G10

Zusammenfassung

Viele makroökonomische Probleme erfordern die Nutzung von großen Zeitreihenpaneln. Das Large Vector Auto Regression Model (Large VAR), welches ein integriertes Model für hochdimensionale Zeitreihen darstellt, kann u.a. angewandt werden um Impulsantwortsstudien durchzuführen, verschiedene ëndogenünd ëxogenïndividualisierte Zeitreihen zu etikettieren sowie Variablen und Lag gleichzeitig auszuwählen. Diese Arbeit untersucht die Anwendung des Large VAR Prozesses für das Beseitigen der Schwierigkeiten, die aus einer Mischung aus hochdimensionaler Abhängigkeitsstruktur, serieller Korrelation und moderater Stichprobengrß̈e stammen. Zuerst stellen wir drei Regulierungsarten für das Modell vor, dann diskutieren wir die Auswahl der Tuning-Parameter basierend auf der Datenlage, um die Prognose zu optimieren. Weiterhin präsentieren wir einen numerischen Algorithmus dafür und fassen diesen vergleichend mit weiteren Methoden zusammen. Es wird bewiesen, dass diese neue Methode über Orakel Eigenschaften verfügt. Eine empirische Analyse mit einem großen, makroökonomischen Zeitreihenpanel wird durchgeführt, um die überlegenen Ergebnisse zu illustrieren.

Schlagwörter: Zeitreihen, Vector Auto Regression, Regulierung, Lasso, Group Lasso, Orakel Schätzer
JEL Klassifikation: C13, C14 ,C32, E30, E40, G10

Contents

1 Introduction **1**

2 The Large VAR Model and Its Estimation **7**
2.1 The model 7
2.2 Universal Grouping 9
2.3 No Grouping 12
2.4 Segmentized Grouping 14
2.5 Comparison 15

3 Algorithm **17**

4 Empirics **21**
4.1 Data and transformations 21
4.2 Forecasting evaluation 22
4.3 Results on the Large VAR model 23
4.4 Comparison of the FAVAR and BVAR 23

5 Discussion **27**

6 Conclusion **31**

1 Introduction

Over the past few decades, there has been a resurgence of interest in producing accurate forecasts of key macroeconomic variables. Much effort has been put in developing various kinds of forecasting models. By the late 1970s, structural forecasting, which views and interprets economic data through the lens of a particular economic theory and hence rises and falls with theory, receded following the decline of Keynesian theory. In contrast, nonstructural forecasting approach, which attempts to exploit the reduced-form correlations in observed time series, was regarded as an alternative. Since then, besides the previous informal methods, which are still widely used in many forecasting agencies or institutions, macroeconomic forecasts are often based on the results of time-series models, whose family ranges from univariate versions as proposed by Box and Jenkins (1976) to multivariate VARs and cointegrated systems.

The centerpiece of Box-Jenkins program are autoregressive moving average (ARMA) models, which in fact are the combinations of the autoregressive and moving average models of Slutsky (1927) and Yule (1927) . Macroeconomics, however, is crucially concerned with cross-variable relationships, whereas the basic Box-Jenkins models use only the past of a given economic variable to forecast its future. In other words, macroeconomics is concerned particularly with multivariate relationships, whereas ARMA models are univariate. Thus, many extensions of Box and Jenkins framework involve multivariate modeling and VARs have emerged as the central multivariate model.

As one of the standard devices for the exploitation of large dynamic systems, VARs were forcefully advocated by Sims (1980) as a less restrictive alternative to traditional econometric system-of-equations models. Involving a more radical change of direction, VARs enjoy many natural advantages. For example, they do not impose restrictions on the parameters and hence supply a very general representation which allows the capture of the complex data relationships. Moreover, in contrast to the tedious numerical optimization required for estimation of multivariate ARMA models, VARs are numerically stable and simple.

The size of VARs typically used in macroeconomic analysis ranges from three to ten variables, which potentially creates an omitted variable bias with adverse consequences. The reason is the theoretical properties of small-scale VAR methods are rarely met in practice. According to Bernanke et al. (2005), at least two potential sets of problems will be led when using small amount of information.

Firstly, from the economic point of view, the measurement of monetary policy might be contaminated as the small number of variables is unlikely to span the information sets used by financial market observers or central banks. For example, Sims (1986) and Christiano and Eichenbaum (1992) pointed out that "price puzzle", which is in fact a positive reaction of CPI in response to

a tightening monetary policy, is an artefact resulting from the size limitation due to the omission of some forward-looking variables, which could include information about future inflation. Put differently, the conjecture of Sims (1992) is that policy shocks which are associated with substantial so called price puzzles are actually confounded with non-policy disturbances that signal future increases in prices. Sims and Tao (1995) proposed that by including current and past values of commodity prices, one can modify these shock measures and the price puzzle would disappears. Nowadays it is the standard practice not to generate price puzzle. Bańbura et al. (2010) also conjectured that the small sample size is problematic for applications which in fact require the study of a larger set of variables than the key macroeconomic indicators, such as cross-country data or disaggregate information. According to Lucas (1980) and Ljungqvist and Sargent (2004) , the more dimensions on which the model mimics the answers actual economies give to simple questions, the more economists trust its answers to harder questions. Moreover, we analyze the impulse response function only for those variables included in the equation. Consequently, it is difficult for us to observe the responses of multiple indicators. Therefore, researchers are tempted to use as many series as are available for the analysis.

One might expect the large-scale VAR model to include the disaggregated, sectorial and geographical indicators and the corresponding forecasting performance to improve. That is likely to happen. For instance, low-dimensional VARs containing 3 or 4 variables may evolve into richer model determining about 20 variables in equilibrium. The research of Sims et al. (1996) has increased the amount of information to about 20 variables applying Bayesian VARs with Litterman (1986)'s priors. However, the expansion in scale will be likely to stop here for the following reasons although the priors imposition is still not sufficient to deal with larger models.

There are mainly two reasons for that: first, the bigger models are not necessary better. This is the idea enshrined in Zellner (1992)'s "Keep it Complicatedly Simple" principe. According to Diebold (1998), the demise of large models containing so many variables heights professional awareness of the fact that making the model unnecessarily complex can degrade the efficiency of the resulting parameter estimator. Second, historically, VARs or vector autoregressive and moving average (MAVAR) models are not appropriate tools for analyzing large panel data as they involve more than a handful of parameters to estimate. The reason for this is that the nature of dynamic stochastic general equilibrium modeling requires their parameters to be jointly estimated, which limits the complexity of the models that can be entertained. Furthermore, parallel to the theoretical progress, there is a substantial increase in the quality and amount of financial and economic data available. Instead of obtaining predictions easier, the researchers are confronted with the identification problems if they want to exploit the potential benefits of the large panel macroeconomic and financial data set with a limited amount of observations, especially under the situation that the macroeconomic data that persons deal with has relatively low frequencies e.g., monthly, quarterly or yearly. Additionally, due to the structural change points in the economic data (although not explored in this work), the effective number of observations used for estimation could be much smaller than the original one. Take the dataset we will consider for estimation as an example: Given a sample of size $T = 540$, we have 131 dimensionality and number of lags equals P, thus there are a total of 131^2P parameters to estimate in the model. An integrated solution to this sort of questions is not feasible.

To circumvent these problems, the traditional econometrician toolbox offers three classes of methodologies to overcome the curse of over-parameters: the first method is combining or aggregating a large number of forecasts from relatively simple model. The second method uses a small amount of latent factors to summarize the information in the panel. The third class is based on model selection methods, model averaging and shrinkage so as to reduce the sampling error.

The need for novel approach to deal with macroeconomic variables forecasting using many predictors has firstly concentrated research efforts on the field of factor models. Bernanke et al. (2005) investigated FAVAR which combines the advantages of both factor model and standard structure VAR analysis. Here factor model is the fundamental contribution of Burns and Mitchell (1946). The basic idea that stands behind this approach is that the movement of one time series can be characterized as the sum of two mutually orthogonal components: The first component (common component) explains the main part of the variance of the time series and is a linear combination of the common factors. The second component (idiosyncratic component), contains the remaining specific information and is only weakly correlated across the data set. Sims and Sargent (1977) and Geweke (1977) introduced a desirable improvement of the factor model by exploiting the dynamic interrelationship among variables and by reducing common factors' number even further. Forni et al. (2000) then developed a generalized dynamic factor model which allows for a limited amount of cross correlation among the idiosyncratic principal components and proposed this methodology for exploiting the potential information in the analysis of large panels of time series data. They imposed restrictions on the covariance structure so as to limit the number of parameters to estimate. Stock and Watson (2002a) have introduced an approximate dynamic factor model using frequency domain analysis to summarize the information and showed that the forecasts based on the factors outperform univariate/ small vector autoregressions.

Over the past decade there are an increasing number of macroeconomic forecasting which rely on dynamic factor models such as Gosselin and Tkacz (2001), Masten (2010). Commonly used estimation procedures are principal components methods, state space models(Stock and Watson (1998)) and cointegration frameworks (Gonzalo and Granger (1995)). Other recently developed methods include Forni et al. (2005), Giannone et al. (2004), Park et al. (2009) and Song et al. (2010) for nonstationary case.

Another device for the over-parametrization problem in the literature that we review is the shrinkage. Firstly, in factor models a small number of weighted linear combinations are used to summarize the information of all variables in the data set. In other words, credible cross-variable impulse responses can not be produced and that is not suitable for structural analysis. Macroeconomics, however, is crucially concerned with variable - to - variable relationships. From the economic point of view, we should find other methodologies which could facilitate corresponding interpretation. Secondly, compared to factor models, the VAR framework with shrinkage is able to do observation and estimation in one step directly, which leads to greater efficiency. While for dynamic factor models, we should do a two step procedure: dimension reduction first and low dimensional time series modeling.

In nonstructural modeling and forecasting, VAR with shrinkage has a long history of productive use. It is a distinct, but intimately related and equally important strand of VAR literature compared to FAVAR and has a long been known that VAR estimated using Bayesian shrinkage techniques produces forecasts superior drastically to those from unrestricted ones. Nowadays, it has already emerged as a key component of estimation.

Sims (1980) spearheaded to impose Bayesian restrictions and then calibrate or estimate tuning parameters. The basic idea is to estimate the VARs by empirical Baysian methods. An ongoing flood of work followed Sims (1980). Doan et al. (1984) and Litterman (1986) advocated that VARs with Bayesian shrinkage containing six variables could lead to better forecast performance. What they used is the "Minnesota prior", which is a simple vector random walk. F.Canova and Ciccarelli (2004).considered exclusion, exogeneity or homogeneity restrictions for data sets with a panel structure in global VARs and panel VARs separately. Bańbura et al. (2010) constructed a model containing 40 variables for policy analysis and imposed exclusion and exogeneity restrictions in addition to shrinkage. However, these restrictions are unnecessary.

The work of Bańbura et al. (2010) has already reached this point. Their paper shows that bringing the additional information content will yield more accurate predictions. Building on the asymptotic analysis in DE Mol et al. (2008), Bańbura et al. (2010) made an empirical study and showed that shrinkage is sufficient to deal with large models and restrictions are not necessary. They increased the cross-sectional dimension to 131 variables, which amounts to increasing the tightness of priors, and showed that the Bayesian regression tends to select factors that could explain most of the variation of the predictors when data are characterized by stronger collinearity. Therefore, they controlled for over-fitting while preserving the relevant sample information by setting the degree of shrinkage in relation to the model size. Considering the Bayesian approach requires the choice of priors, from the computational point of view, we should find a more efficient method.

Other solutions include so called marginal approach, see for example, Christiano et al. (1999) and S.Kim (2001). Their key insight is to define a core set of indicators and to add one variable (or one group of variables). However, as suggested by the analysis in Bańbura et al. (2010), impulse responses comparison across models is also problematic with this approach. The non-Bayesian examples include Chudik and Pesaran (2007), which studied "neighboring" procedure for the case $p = 1$ when J and T are large, and Wang et al. (2007), which considered the regression coefficient and autoregressive order shrinkage and selection via the lasso when p is large and T is small for the univariate case.

This work is an empirical version of Song and Bickel (2011), who pursued one approach called large VARs to answering the challenges which come from a mixture of the high dimension, temporal dynamics and moderate sample size for the analysis of relative large data set. Contrary to the factor literature they model the variables in levels to retain all the information in the trends. It is a more integrated method which has the following advantages while in Bayesian case, it is still not clear to find out the prior function for the previous three points: Firstly, from the variable selection and regularization point of view, the new method can address the time

dependence issue together with high dimensionality and moderate sample size, and could still obtain the consistency of variable selection even under the dependent scenario, i.e., to reveal the equilibrium among them.

Secondly, the useful lag of each piece of data is small due to structure change point and, importantly, unknown to the forecaster in real time. Because indicators quite useful over history may break down when used in forecasting, while indicators with not so good predictive characteristics in the past may proved to have better predictive ability in future. In the literature, the variable selection is always performed first and the corresponding estimate's performance w.r.t. different lags is compared each other. Thus, the researchers could select the optimal number of lags. As such our research is thus fraught with difficulties when disregarding the serial correlation. We should believe that there exist some interaction between variable selection and lag selection. Our approach could take the dependent scenario into consideration and do the variable selection and lag selection simultaneously.

Thirdly, large VARs allow different variables in the high dimensional time series to have individualized endogenous and exogenous properties, especially when the differences between various time series are significant. While in the general case, in Baysian approach, it is unclear to find proper prior distribution functions to consider the previous three points. These three key features differentiate Song and Bickel (2011)'s product from others.

Fourthly, the new approach of selection, broadly defined, is based on the Least Absolute Shrinkage and Selection Operator (LASSO) proposed by Tibshirani (1996), which already complied some packages to make the computation more efficient. Lasso is developed in the last century but is not widely used due to the limitation of computation speed. In recent years, however, powerful new packages have been developed. Several algorithms which work without limitations of dimensionality have been proposed for Lasso regression recently, for example, the Least Angle Regression (LARS) packages developed in 2004 by Efron and Glmnet packages written by Schmidt (2005) for Matlab software.

The rest of work is organized as follows. Following the introduction to the large VAR model, three types of estimators are carefully studied in Section 2. In this section, we treat each variable's own lags differently from other variables' lags, distinguish various lags over time, and select the variables and lags simultaneously. After that, we evaluate the forecasting performance for a VAR with 131 variables, containing sectoral data, several financial variables and conjunctural information besides macroeconomic information. The data set is the one used by Stock and Watson (2005) for forecasting based on factor analysis. The method is applied to solve the empirical macroeconomic problem, which shows that the new method outperforms the existing methods. The last section contains conclusion with a brief discussion.

2 The Large VAR Model and Its Estimation

2.1 The model

The mechanics of VARs are not difficult. Recall that we approximate time dynamics with a univariate autoregression by regressing a variable Y_T on itself given a sample of size T. Here, its own past values $Y_{T-1}, Y_{T-2}, \ldots$ of this variable are used as regressors. By logical extension method, we then regress each of the set of variables on lagged values of itself as well as the lagged values of every other variable in this system. That is the case of a vector regression of $Y_t^\top$. As we include lags of all variables in every equation and we allow for correlations among the disturbances of the various equations, structural relationships among these variables are automatically incorporated.

Let $\{Y_{t,j}\}_{t=1,j=1}^{T,J}$ be a potentially large vector of random variables. We consider the following VAR model:

$$Y_t^\top = Y_{t-1}^\top B_1 + \ldots + Y_{t-P}^\top B_P + U_t^\top \tag{2.1}$$

which could be written as:

$$\underbrace{\begin{pmatrix} Y_T^\top \\ Y_{T-1}^\top \\ \ldots \end{pmatrix}}_{T\times J} = \underbrace{\begin{pmatrix} Y_{T-1}^\top & Y_{T-2}^\top & \ldots & Y_{T-P}^\top \\ Y_{T-2}^\top & Y_{T-3}^\top & \ldots & Y_{T-1-P}^\top \\ \ldots & \ldots & \ldots & \ldots \end{pmatrix}}_{T\times JP} \underbrace{\begin{pmatrix} B_1 \\ B_2 \\ \ldots \end{pmatrix}}_{JP\times J} + \underbrace{\begin{pmatrix} U_T^\top \\ U_{T-1}^\top \\ \ldots \end{pmatrix}}_{T\times J} \tag{2.2}$$

where $Y = (Y_T^\top, Y_{T-1}^\top, \ldots, Y_1^\top)^\top$ with $Y_t^\top = (Y_{t1}, Y_{t2}, \ldots, Y_{tJ})$. $Y_t^\top$ is an $J \times 1$ vector of observable J economic and financial variables which are assumed to have pervasive effects on the economic situation of a country at time t. $B_1, B_2, \ldots, B_P$ are $J \times J$ autoregressive coefficient matrices where P is the number of lags which are initially prespecified. Without loss of generality, we set a large enough P here. Let us suppose the T-dimensional vector of errors $U_t^\top = (U_T^\top, U_{T-1}^\top, \ldots, U_1^\top)$ is mean zero with covariance matrix Q which is independent of t. It could be observed that our framework here imposes no restrictions on the parameters and would like to seek some general representations.

As the variables in the panel we would like to consider for empirical application will be standardized and demeaned, the $J \times J$ covariance matrix Σ is assumed to be an identity matrix I_{J*J} although homogenous variance and zero mean are very naive. The same transformation

for regressors has also been applied by Mol et al. (2008). The relaxation will be discussed in the penultimate section. Therefore, we have statistically efficient one-equation-at-a-time least squares estimation of VARs in spite of the potential correlation of disturbances.

The compact form is

$$Y_t^\top = X^\top B + U \tag{2.3}$$

where $X = (X_T^\top, X_{T-1}^\top, \ldots, X_1^\top)^\top$ with $X_t = (Y_{t-1}^\top, Y_{t-2}^\top, \ldots, Y_{t-P}^\top)$. X are the lags of Y. We assume that Y_t, X_t both have mean zero and proceed by the VAR estimation using all variables as regressors. For now, it is unnecessary to specify whether the ultimate interest of us is in uncovering cross-variable linkages or the forecasting of $Y_T^\top$. What we are interested in is estimating the coefficient matrix $B_1, B_2, \ldots, B_P$ with coefficients $\{B_{p,i,j}\}_{p=1,i=1,j=1}^{P,I,J}$. $B_{\cdot\cdot j}$, $B_{p\cdot j}$, $B_{\cdot i\cdot}$, $B_{pi\cdot}$ is the jth column of B and B_p, ith row of B and B_p respectively.

Here we do not impose restrictions on the parameters and hence supply a very general representation which allows the capture of the complex data relationships. However, the high level of generality implies a large number of parameters for estimation. Given J dimensionality and P lags, there will be a total of J^2P parameters to estimate, while we have only JT observations which is much smaller than J^2P. Considering the structure change points,which is a typical characteristic of macroeconomic sample, the effective number will be reduced severely compared to the original T. Thus, the number of unrestricted parameters that can be estimated reliably is very limited. In order to handle the proliferation of parameters, we will discuss the estimation procedure considering three different kinds of regularization techniques. Before moving on, we incorporate two mild beliefs according to Bańbura et al. (2010).

1. Compared to the more distant lags, the more recent ones should provide more reliable information.
2. The own lags explain more (less) of the variation of a given variable than the lags of other variables in the equation.

These beliefs are imposed when choosing the tuning parameters. According to Bańbura et al. (2010), the prediction accuracy can be improved by adding additional spatial information. They applied shrinkage as a regularization solution for the problem of inverting an otherwise unstable large covariance matrix for their empirical application. In order to overcome the curse of dimensionality, they imposed prior beliefs on parameters. Our approach is also to include a large enough number of variables firstly and then shrink insignificant regression coefficients towards zero exactly using proper penalty so as to get a parsimonious model. Via this regularization approach, we realized the purpose for observing and controlling significant economic variables in the macroeconomic process. The basic principle behind this approach is that we distinguish various lags over time and treat each variable's own lags different from other variables' lags. According to Song and Bickel (2011). the basic methods to select variables and lags used in this work can be outlined as follows:

2.2 Universal Grouping

Firstly, "universal grouping" regularization is considered. We estimate the whole coefficient matrix, say B_p when lag equals p with entries $B_{pij}, 1 \leq i, j \leq J$:

$$\begin{pmatrix} \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ 0 & \bullet & 0 & 0 & 0 & 0 \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & 0 & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & 0 \end{pmatrix}$$

The diagonal terms $B_{p11}, B_{p22}, \ldots, B_{pjj}, \ldots, B_{pJJ}$ are the variables'own lags with B_{pjj} is used to describe the effect of the jth variable to itself. The off-diagonal terms of B_p reflect the others' lags. According to the second belief, we impose different regularizations for the variables' own lags and the others' lags.

Towards the diagonal term of B_p, labeled "endogenous", we base our selection on the Lasso technique proposed by Tibshirani (1996) . One important feature of this technique is that it can be used for variable selection. It can be seen as a penalized regression with a penalty on the coefficients involving the $L1$ norm and is a popular technique for simultaneous variable selection and estimation. Because of this natural of the constraint, it tents to produce some coefficients that are exactly zero and hence give a more interpretable model. Lasso has at least two advantages compared to the classical variable selection methods, for example, subset selection. Firstly, the lasso selection process is continuous and more stable than the classical methods. Second, lasso is regraded computationally feasible for high-dimensional data. If we use subset selection here for example, computation is very combinatorial and will not be feasible when J is very large. Lasso idea is quite general and hence has been broadly applied in a variety of statistical models. Here we apply it for the study of macroeconomics.

For the off-diagonal coefficients of B_p, suppose that they are not only sparse, but also have the same sparsity pattern across different columns, which is called "group sparsity" by Song and Bickel (2011). B_{pj-j} is used to denote the vector composed of $B_{pji_{i \neq j}}$ and W_{-j} is used to denote the $(J-1) \times (J-1)$ diagonal matrix diag$[w_1, \ldots, w_{j-1}, w_{j+1}, \ldots, w_J]$ where w_i is a positive real-valued weight associated with the ith variable for $1 < i < J$. Tibshirani (1996) also prefer to use this method for comparisons to prior works.

We include it here mainly for the practical implementation since if w_i is chosen as the $Std(Y_i)$, it is equivalent to standardize the predictors so that they have zero mean and unit variance. Thus for the off-diagonal terms, labeled "exogenous", we apply the group Lasso type regularization proposed by Yuan and Lin (2006), where variables are included or excluded in groups. In fact, there has been a enormous amount of research activity devoted to related regularization methods such as the group Lasso in the past decade. Other related approaches include regu-

larization paths for the support-vector machine(Hastie et al. (2004)), the elastic net (Zou and Hastie (2005)), L_1 regularization paths for generalized linear models(Park and Hastie (2006)), the Dantzig selector (candes and Tao (2005)) and the graphical Lasso (Friedman et al. (2007)).

Specifically, given the notations above, we use Lasso type penalty $\mu \sum_{j=1}^{J} w_j \left|B_{pjj}\right|$ to impose the regularization on predicted variables' own lags and the group Lasso type penalty $\sum_{j=1}^{J} \left\|B_{pj-j}W_{-j}\right\|_2$ for the implementation of the regularization on other regressors'lags respectively.

The tuning parameter μ governs the extent to which the lags of other variables are less (more) important than the lags of the predicted variable itself. When this hyperparameter μ is large, the penalty assigned to own lags is more than to others' lags. As a result, the off-diagonal terms are more likely to be shrunk to zero than the diagonal ones. In this case, we believe the variable is more edogeneous while the variables's dynamic is driven by itself rather than other variables. When μ is small, it is more likely that the diagonal entries are shrunk to zero instead of the off-diagonal ones, which means these variables are mainly driven by others. The "edogeneity" property is easy to understand. For the exogeneity case, if the dynamic of one variable is driven by itself, while for a different variable, it might be driven by the dynamics of others. One "exogeneity" example is from Ben McCallum's notes: suppose a policy variable labeled as "exogeneous", we believe that this variable could have been managed exogeneously by policymakers if they had been unorthodox enough to do so.

For the whole coefficient matrix B_p we have the following penalty:

$$\mu \sum_{j=1}^{J} w_j \left|B_{pjj}\right| + \sum_{j=1}^{J} \left\|B_{pj-j}W_{-j}\right\|_2 \leq p^{-\alpha} \tag{2.4}$$

The third block $p^{-\alpha}$ reflects the different regularizations for different lags. This item will increase when p gets smaller. This is consistent with the first belief: compared to the more distant lags, the more recent ones should provide more reliable information. We use larger amounts of shrinkage for the more distant lags compared to the more recent lags. The idea is simple: for example, the autocorrelation between today's and yesterday's GDP is assumed to be larger than that between today's GDP and the GDP one year ago. The tuning parameter α controls the relative importance of more distant lags with respect to the more recent lags.

Here $p^{-\alpha}$ is a decreasing function which could reflect the fact that with the increase of α, the value of this function will decrease. We will impose a stronger restriction to the more distent lags if $p^{-\alpha}$ is bigger. But that does not mean this is the unique function that could be applied here. For example, $\log(p)^{-\alpha}$, $\exp(p)^{-\alpha}$ can also be used here to express the relationship. In order to avoid too many hyperparameters, we do not consider a general representation for this block (use a data driven way to estimate the appropriate functions $f(1),\ldots,f(p),\ldots,f(P)$ for different lags correspondingly), especially when $P \to \infty$.

After that, we multiply p^{α} on both sides so that the right side will be 1.

$$\sum_{j=1}^{J} p^{\alpha} \left\| B_{pj-j} W_{-j} \right\|_2 + \mu \sum_{j=1}^{J} w_j p^{\alpha} \left| B_{pjj} \right| \leq 1$$

We then sum the two blocks $\sum_{p=1}^{P} \sum_{j=1}^{J} p^{\alpha} \left\| B_{pj-j} W_{-j} \right\|_2$ and $\sum_{p=1}^{P} \mu \sum_{j=1}^{J} w_j p^{\alpha} \left| B_{pjj} \right|$ up over p since there are P matrices $B_1, \ldots, B_p$ for B_p and have:

$$\sum_{p=1}^{P} \sum_{j=1}^{J} p^{\alpha} \left\| B_{pj-j} W_{-j} \right\|_2 + \mu \sum_{p=1}^{P} \sum_{j=1}^{J} w_j p^{\alpha} \left| B_{pjj} \right| \leq P \tag{2.5}$$

After adding them to the original mean squares, the equation can be expressed by a primal Lagrangian:

$$\min_B L(B) = \min_B \{2J(T-P)\}^{-1} \sum_{t=P+1}^{T} \left\| Y_t^{\top} - X_t^{\top} B \right\|_2^2$$
$$+\lambda \left(\sum_{p=1}^{P} \sum_{j=1}^{J} p^{\alpha} \left\| B_{pj-j} W_{-j} \right\|_2 + \mu \sum_{p=1}^{P} \sum_{j=1}^{J} w_j p^{\alpha} \left| B_{pjj} \right| \right)$$

Specifically, we couple (2.5) to the quadratic loss and have the equation (2.6). Our algorithm is proposed for finding a minimizer of the following function:

$$\min_B \{J(T-P)\}^{-1} \sum_{t=P+1}^{T} \left\| Y_t^{\top} - X_t^{\top} B \right\|_2^2 + \lambda \sum_{p=1}^{P} \sum_{j=1}^{J} p^{\alpha} \left\| B_{pj-j} W_{-j} \right\|_2 + \gamma \sum_{p=1}^{P} \sum_{j=1}^{J} w_j p^{\alpha} \left| B_{pjj} \right| \tag{2.6}$$

where $\widehat{B}$ is the universal grouping estimator with three tuning parameters λ, γ and α. Here $\gamma = \lambda \mu$. As increase of the dimensionality J increases, the parameters will be shrunk more in order to avoid over-fitting. It is also applied by Mol et al. (2008).

However, in the universal grouping regularization case, we actually pose some strong assumptions on the underlying structure, which are rarely met in practice.

Firstly, we have one pair of tuning parameters γ and λ which govern the extent to which the lags of the other variables are "less (more) important" than the lags of the predicted variables. This implicitly means that between the others's lags and the lags of the predicted variable itself, we

apply the same weight, which is not realistic from the economic point of view. The reason for this is that we apply group Lasso regularization for the off-diagonal terms. We have only one hyperparameter μ to govern the relative weights between the own lags and the lags of others. In fact, if we select the optimal pair of γ and λ (one optimal μ for the off-diagonal terms) to optimize the forecasting performance, what we optimize is not the variable of particular interest, but the average forecasting performance for all J variables in the data set. Bańbura et al. (2010) studied the case that the own lags are always more important than others' lags, which might be less general than that of ours.

In other words, the "endogeneity" and "exogeneity" labeling should be systematic rather than arbitrary. If we have univariate time series, it is easy to determine the variables are endogeneous or exogeneous. But for the high dimensional time series analysis, assuming all of the time series in the panel have the same endogeneity or exogeneity throughout the economy will be rather restrictive. When we include a vast number of macroeconomic time series, various weights should be applied instead.

Secondly, group Lasso techniques will shrink all terms in one row to zero simultaneously except the lags of the predicted variable itself. This equivalently means that we assume the corresponding variable is either significant for all the others' lags or not significant for them at all, which is also hardly met in practice.

2.3 No Grouping

In order to diagnose and amend the inadequacies of "universal group" regularization, we consider the "no grouping" case. Here we no longer estimate the whole matrix B all at once, but do the estimation column by column. There are J column vectors to estimate, say $\{B_{\cdot\cdot j}\}_{j=1}^{J}$. For both diagonal and off-diagonal terms, we apply Lasso-type estimate as the jth column $B_{\cdot\cdot j}$ is a vector. An illustration of the no grouping type of estimate is showed as follows:

$$\left(\begin{array}{c|c|c|c|c|c}
\bullet & 0 & \bullet & \bullet & \bullet & \bullet \\
\bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
\bullet & \bullet & 0 & \bullet & \bullet & \bullet \\
0 & \bullet & \bullet & \bullet & \bullet & \bullet \\
\bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
0 & \bullet & \bullet & \bullet & \bullet & \bullet
\end{array}\right)$$

Given $Y_{tj} = y_t$ with $t = 1, \ldots T$, following the similar abbreviations in subsection universal grouping, we have the primal Lagrangian (2.7):

$$\min_{B_{..j}} L(B_{..j}) = \min_{B_{..j}} (T-P)^{-1} \sum_{t=P+1}^{\top} (Y_{tj} - X_t^\top B_{..j})^2$$
$$+\lambda_j (\sum_{p=1}^{P} \sum_{i \neq j} p^\alpha w_i \left|B_{pij}\right| + \upsilon_j \sum_{p=1}^{P} p^\alpha w_j \left|B_{pjj}\right|)$$

Substitute γ_j with $\mu_j \times \lambda_j$, we rewrite the Lagrangian function as:

$$\min_{B_{..j}} L(B_{..j}) = \min_{B_{..j}} (T-P)^{-1} \sum_{t=P+1}^{\top} (Y_{tj} - X_t^\top B_{..j})^2 + \lambda_j \sum_{p=1}^{P} \sum_{i \neq j} p^\alpha w_i \left|B_{pij}\right| + \gamma_j \sum_{p=1}^{P} p^\alpha w_j \left|B_{pjj}\right| \tag{2.7}$$

with tuning parameters λ_j, γ_j and α. $\widehat{B}$ is the no grouping estimate. Here we add the subindex j to λ_j and γ_j and hence they could vary when estimating different $B_{..j}$s when $1 \leq j \leq J$.

Then we drop the common subindex j by writing $B_{..j} = \beta$ and $Y_{tj} = y_t$ for simplicity. Besides we write $B_{pij} = c_i, i = 1, \ldots, P(J-1); i \neq j$ for the off-diagonal terms and $B_{pjj} = d_p, p = 1, \ldots, P$ for the diagonal terms. $\lambda_j = \lambda$, $\gamma_j = \gamma$. We now have the corresponding abbreviated estimation equation (2.8) for (2.7):

$$\min_{\beta} Q_T(\beta) = \min_{\beta} (T-P)^{-1} \sum_{t=P+1}^{T} (y_t - X_t^\top \beta)^2$$
$$+\lambda \sum_{p=1}^{P} \sum_{i \neq j} p^\alpha w_i \left|c_i\right| + \gamma \sum_{p=1}^{P} p^\alpha \left|d_p\right|$$

we substitute $\lambda p^\alpha w_i$ with λ_i and $\gamma p^\alpha w_p$ with γ_p, therefore:

$$\min_{\beta} Q_T(\beta) = \min_{\beta} (T-P)^{-1} \sum_{t=P+1}^{T} (y_t - X_t^\top \beta)^2 + \sum_{i=1}^{P(J-1)} \lambda_i \left|c_i\right| + \gamma_p \sum_{p=1}^{P} \left|d_p\right| \tag{2.8}$$

Here individualized weights between others' lags and owns' are considered because of separating λs and γs for the estimations for different columns. This is a reflection of individualized endogeneity and exogeneity. By the same reason, we amend the deficiencies that all off-diagonal terms in one row might be shrunk to 0 simultaneously.

2.4 Segmentized Grouping

Here we consider the problem of tracing out the explanatory groups of variables (factors) for accurate prediction in regression. As there exist some natural "segment" information the large panel data set, we will take the lags of the predicted variable itself, others'(but in the same segment i) lags and the others' (outside the segment i) lags into consideration when estimating the coefficients corresponding to the ith segment. Now we estimate the coefficient matrix B neither column by column nor all at once. Instead, we do the estimation segment by segment.

Now let us now turn to the next question: how we group them? Considering the interest rates at different maturities for example, it is natural to believe that there exist some relationships among the federal fund rate, the 3-month interest rate, T-bill rate and 1-and 10-year bond rate. Another example is about (un)employment rate: the number of employees with respect to different industrial sectors could be regarded as a natural segment in the macroeconomic data.

$$\left(\begin{array}{ccc|cc|c} \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ 0 & \bullet & 0 & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ 0 & 0 & 0 & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \end{array}\right)$$

For the ith segment $B_{\cdot\cdot\mathcal{N}}$, $\mathcal{N}_i$ denotes the index set for the ith segment and $N_i = |\mathcal{N}_i|$ denotes the cardinality of the set $\mathcal{N}_i$. The corresponding part of $Y_i^\top$ is $Y_{t\mathcal{N}_i}^\top$. We use $W_{\mathcal{N}_i}$ to denote the $N_i \times N_i$ diagonal matrix with diagonal terms $\{w_i\}_{i\in\mathcal{N}_i}$ and $W_{\mathcal{N}_i - j}$ to denote the $(N_i-1)\times(N_i-1)$ diagonal matrix with diagonal terms $\{w_i\}_{i\in\mathcal{N}_i, i\neq j}$. Chudik and Pesaran (2007) considered the case that $P=1$ and large J and T through some neighboring procedures, which can be viewed as a special case in this section. Following the similar ideas and abbreviations in subsection 2.2, we have the following regularization :

$$\min_{B_{\cdot\cdot\mathcal{N}_i}} \{N_i(T-P)\}^{-1} \sum_{t=P+1}^{T} \left\| Y_{t\mathcal{N}_i}^\top - X_t{}^\top B_{\cdot\cdot\mathcal{N}_i} \right\|_2^2$$
$$+\lambda_{\mathcal{N}_i} \left(\sum_{p=1}^{P} \sum_{j\notin\mathcal{N}_i} p^\alpha \left\| B_{pj\cdot} W_{\mathcal{N}_i} \right\|_2 + \mu_{1\mathcal{N}_i} \sum_{p=1}^{P} p^\alpha w_j \left| B_{pjj} \right| + \mu_{2\mathcal{N}_i} \sum_{p=1}^{P} \sum_{j\in\mathcal{N}_i} p^\alpha \left\| B_{pj-j} W_{\mathcal{N}_i} \right\|_2 \right)$$

There are four hyperparameters $\lambda_{\mathcal{N}_i}$, $\gamma_{\mathcal{N}_i}$, $\eta_{\mathcal{N}_i}$, α for the segmentized grouping estimate. We wrote $\gamma_{\mathcal{N}_i} = \lambda_{\mathcal{N}_i}\mu_{1\mathcal{N}_i}$ and $\eta_{\mathcal{N}_i} = \lambda_{\mathcal{N}_i}\mu_{2\mathcal{N}_i}$ for the simplicity of the notation, $i = 1,\ldots,I$ where I is the overall number of segments. The corresponding estimation equation is:

$$\min_{B_{\cdot\cdot\mathcal{N}_i}} \{N_i(T-P)\}^{-1} \sum_{t=P+1}^{T} \left\| Y_{t\mathcal{N}_i}^{\top} - X_t{}^{\top} B_{\cdot\cdot\mathcal{N}_i} \right\|_2^2$$
$$+\lambda_{\mathcal{N}_i} \sum_{p=1}^{P} \sum_{j \notin \mathcal{N}_i} p^{\alpha} \left\| B_{pj\cdot} W_{\mathcal{N}_i} \right\|_2 + \gamma_{\mathcal{N}_i} \sum_{p=1}^{P} p^{\alpha} w_j \left| B_{pjj} \right| + \eta_{\mathcal{N}_i} \sum_{p=1}^{P} \sum_{j \in \mathcal{N}_i} p^{\alpha} \left\| B_{pj-j} W_{\mathcal{N}_I} \right\|_2 \tag{2.9}$$

2.5 Comparison

In the following, we will decide which kind of regularizations will be suggested for the practical implementation. The selection of the proper regularization should make a compromise between flexibility and realization of assumptions. Considering the four following reasons, we suggest the no grouping estimate for practical implementation. Table 2.1 provides an overview of the comparisons.

Firstly, the no grouping case allows different λs and γs for different variables's forecasting while in the universal grouping case we have only one pair of λ and γ, which will produce some suboptimal forests. The reason for this is that we allow individualized weights between the lags of the predicted variable and the lags of others. Therefore we could tune λ_j's and γ_j's to produce optimal forecasting performance for each variable of interest, say jth. From allowing individualized endogeneity or exogeneity point of view, no grouping is the best approach and "segmentized grouping" seconds.

Secondly, if we apply the group Lasso technique, all off-diagonal terms in one row will be shrunk to zero, whereas for the no grouping case, we could get rid of the disadvantage. From the "universal shrinkage" point of view, the "no grouping" still tops.

Thirdly, from the optimization point of view, the parameter for universal grouping, no grouping and segmentized grouping case are selected to optimize the averaged forecasting performance of all variables, the forecasting performance of specific variable, and the averaged forecasting performance of the variables in the same segment respectively. Individualized optimization for the no grouping case still tops here because different variables' time series have very distinct patterns.

Fourthly, we could either estimate the coefficient matrix all at once or do the estimation column by column, which will not change the final result. From this point of view, all these three methods are the same.

In the end, from the "efficiency" point of view, on the one hand, considering the efficiency of the theory, the universal grouping estimate might have smaller estimation error because the group Lasso type estimator actually has a sharper upper risk bound due to the strongly group-sparse assumption. According to Huang and Zhang (2009), group Lasso is more robust to noise due to the stability associated with group structure and thus requires a relatively smaller sample size to

	Universal Grouping	Segmentized Grouping	No Grouping
Individual endogeny/exogeny	–	+	++
Universal shrinkage	–	+	++
Forecasting necessity	+	+	+
Optimization	–	+	++
Efficiency(theory, computation)	++	+	–
$\hookrightarrow$	+	+	+

Table 2.1: Comparison table

satisfy the sparse eigenvalue condition required in the modern sparsity analysis. However, the statistical error is a combination of the modeling error and the estimation error. The underlying structure of group Lasso type estimate has a stronger assumption which might increase the modeling error. Thus the overall risk of the universal grouping case might not be smaller. On the other hand, we also consider the computational efficiency. As we have already noted, the macroeconomic data typically has a low frequency, for example, monthly data. Thus we only need to update the model once a month at most. In this case, the computational cost is not a severe problem and could be feasible. Due to all these, we suggest the "no grouping" approach to be used in practice.

3 Algorithm

Let's define:
$P = diag[1^{\alpha}, 2^{\alpha}, \ldots, P^{\alpha}] \otimes I_{J\times J}$, where $diag[1^{\alpha}, 2^{\alpha}, \ldots, P^{\alpha}]$ is the diagonal matrix with diagonal entries $1^{-\alpha}, 2^{-\alpha}, \ldots, P^{-\alpha}$, $\otimes$ refers to the Kronecker product and $I_{J\times J}$ is the $J \times J$ identity matrix; $W = diag[w_1, w_2, \ldots w_J] \otimes I_{P\times P}$; $\widetilde{X}^{\top} = X^{\top} W^{-1} P^{-1}$ and $\widetilde{B}^{\top} = P^{\top} W^{-1} B^{-1}$. Note $X^{\top} B$ in (2.3) could be written as $X^{\top} W^{-1} P^{-1} PWB = \widetilde{X}^{\top} \widetilde{B}$.

We applied the "rescale in" and "rescale out" technique introduced by Zou (2006) for adaptive Lasso research procedure, where adaptive weights are applied for penalizing different coefficients in the $L1$ norm. W is a known weights vector. Zou (2006) pointed that the weighted Lasso can have the oracle properties if the weights depend on data and are cleverly chosen. Oracle means this adaptive Lasso performs as well as if the true underlying model was given in advance. Therefore we could simultaneously discover relevant predictive variables and ensure high prediction accuracy(optimal estimation).

Zou (2006) discussed that adaptive Lasso is essentially a convex optimization problem with an L_1 constraint and hence can be solved by the same efficient algorithm for solving Lasso problem. Their result showed that the L_1 penalty is at least as competitive as other concave oracle penalties and is also computationally feasible.

The efficient path algorithm makes the adaptive Lasso technique an attractive method for macroeconomic applications. For our case, we transform the own lags of the predicted variables by the factor $\frac{\gamma}{\lambda}$ and then transform back the estimated coefficient matrix in the rescale step by $\frac{\gamma}{\lambda}$ respectively. This may solve the iteration problems and tackle the question faster than the case of the standard Lasso case. After that, motivated by Wang et al. (2007), we do the iteration between multiple penalty terms.

Noting that $\gamma_j = \mu_j \lambda_j$, $\gamma_j \sum_{p=1}^{P} \left|\widetilde{B}_{pjj}\right| = \lambda_j \sum_{p=1}^{P} \left|\mu_j \widetilde{B}_{pjj}\right|$. The estimation procedure for no grouping case are as follows:

1. Generate $\widetilde{X}^{\top} = X^{\top} W^{-1} P^{-1}$.
2. Solve the Lasso problem through iteration among multiple penalty terms. Corresponding

to the estimator (2.7), solve

$$\min_{\widetilde{B}_{\cdot\cdot j}} L(\widetilde{B}_{\cdot\cdot j}) = \min_{\widetilde{B}_{\cdot\cdot j}} (T-P)^{-1} \sum_{t=P+1}^{\top} (Y_{tj} - \widetilde{X}_t^{\top} \widetilde{B}_{\cdot\cdot j})^2 + \lambda_j \sum_{p=1}^{P} \sum_{i \neq j} \left| \widetilde{B}_{pij} \right| + \gamma_j \sum_{p=1}^{P} \left| \widetilde{B}_{pjj} \right|$$

Correspondingly,

$$\min_{\widetilde{B}_{\cdot\cdot j}} (T-P)^{-1} \sum_{t=P+1}^{\top} (Y_{tj} - \widetilde{X}_t^{\top} \widetilde{B}_{\cdot\cdot j})^2 + \lambda_j \sum_{p=1}^{P} \sum_{i=1}^{J} \left| \widetilde{B}_{pij} \right|$$

3. Output $\widehat{B} = P^{-1} W^{-1} \widehat{\widetilde{B}}$ with $\widehat{\widetilde{B}} = (\widehat{\widetilde{B}}_{\cdot\cdot 1}, \ldots, \widehat{\widetilde{B}}_{\cdot\cdot J})$. $(\widehat{\widetilde{B}}_{\cdot\cdot j})$ here is used to minimize the first equation at step (2).
4. Implement the OLS for selected regressors

Given the progress in computing power, estimation does not present any numerical problems now. The computation of Lasso solutions is a quadratic programming problem and could be tackled by standard numerical analysis algorithms. We use Glmnet package Schmidt (2005), which includes a set of Matlab routines. This algorithm exploits the special structure of the Lasso problem, and hence provides an efficient way to compute the solutions simultaneously for all values of P here. Alternatives include least angle regression (LARS) procedure, which uses forward stepwise regression. After adding the Lasso directories to the Matlab path we could load the data set.

According to Wainwright (2009) and Huang et al. (2008), shrinkage with penalization in model selection often cause non-negligible estimation bias which is related to the penalty parameter. When we apply the penalty parameter selection approach, the significant coefficients are shrunk (but not to exact zero) as well as those insignificant ones. Although those significant coefficients are retained in the model the estimation bias will increase. For this reason, although the Lasso could be used for variable selection and simultaneously estimation, we use ordinary least square estimation (OLS) after implementing our method for variable and lag selections. There are also some other advanced bias reduction methods, e.g., the resampling approach for bias corrected selection and estimation method, which deserves further research.

Without loss of generality, we consider the universal grouping case and segmented grouping case as well. The basic method to select variables and do estimation used in this work can be outlined as follows:

1. Generate $\widetilde{X}^{\top} = X^{\top} W^{-1} P^{-1}$.
2. Solve the (group) Lasso problem through iteration among multiple penalty terms. Corresponding to the estimators (2.6) and (2.9) , solve

$$\min_{\widetilde{B}}\{J(T-P)\}^{-1}\sum_{t=P+1}^{T}\left\|Y_t^\top-\widetilde{X}_t^\top\widetilde{B}\right\|_2^2+\lambda\sum_{p=1}^{P}\sum_{j=1}^{J}\left\|\widetilde{B}_{pj-j}\right\|_2+\gamma\sum_{p=1}^{P}\sum_{j=1}^{J}\left|\widetilde{B}_{pjj}\right| \tag{3.1}$$

and

$$\min_{B_{\cdot\cdot\mathcal{N}_i}}\{N_i(T-P)\}^{-1}\sum_{t=P+1}^{T}\left\|Y_{t\mathcal{N}_i}^\top-X_t^\top\widetilde{B}_{\cdot\cdot\mathcal{N}_i}\right\|_2^2$$
$$+\lambda_{\mathcal{N}_i}\sum_{p=1}^{P}\sum_{j\notin\mathcal{N}_i}\left\|\widetilde{B}_{pj\cdot}\right\|_2+\gamma_{\mathcal{N}_i}\sum_{p=1}^{P}\left|\widetilde{B}_{pjj}\right|+\eta_{\mathcal{N}_i}\sum_{p=1}^{P}\sum_{j\in\mathcal{N}_i}\left\|\widetilde{B}_{pj-j}\right\|_2 \tag{3.2}$$

respectively.

3. Output $\widehat{B}=P^{-1}W^{-1}\widehat{\widetilde{B}}$ with $\widehat{\widetilde{B}}$ minimizing (3.1) for universal grouping and $\widehat{\widetilde{B}}=(\widehat{\widetilde{B}}_{\cdot\cdot\mathcal{N}_i},\ldots,\widehat{\widetilde{B}}_{\cdot\cdot\mathcal{N}_i})$, $\widehat{\widetilde{B}}_{\cdot\cdot\mathcal{N}_i}$ minimizing (3.2) for segmentized grouping.
4. Implement the OLS for selected regressors

After applying OLS method to construct the adaptive weights, we now are trying to find the optimal pair of λ_i and γ_i. The two-dimensional cross validation is used here to tune the Lasso, or say, adaptive Lasso in fact. For a given tuning parameter, we can just use cross validation along with our algorithm provided in the second step to exclusively search for the optimal another tuning parameter. We will explain this process in detail in the next section. Note that in principle, other consistent estimators could also be used here instead of OLS estimator, say $\widehat{\beta}$. Thus according to Zou (2006), we could treat this estimator as the third tuning parameter and perform a three-dimensional cross-validation to find an optimal triple $(\gamma_i,\lambda_i,\widehat{\beta})$. Zou (2006) also suggested still using the OLS estimator unless collinearity is a concern, in which we could try $\widehat{\beta}(ridge)$ using the best ridge regression fit, which is much more stable than $\widehat{\beta}(ols)$.

4 Empirics

In this section we discuss the application issues. The empirical performance of the estimation procedure will be studied in an out-of-sample forecast exercise based on a large panel of macroeconomic time series. The estimation procedure is discussed and a data driven recursive scheme is used here so as to choose the hyperparameters which could minimize the corresponding MSFE.

4.1 Data and transformations

The data set employed for our study is the same as the one used in Stock and Watson (2002b). The panel includes a total of 131 variables covering real quantity variables (employment and hours worked), nominal interest variables (wages), asset prices (stock prices), and the yield curve and surveys. In detail, the predictors include series in 14 categories: employment and hours; consumption; real retail, manufacturing and trade sales; real output and income; money and credit quantity aggregates; housing starts and sales; stock prices; exchange rates; real inventories; orders; interest rates and spreads; price indexes; average hourly earnings; and miscellaneous. The sample has a monthly frequency and ranges from January 1959 to December 2003. All 131 variables' lags are used as regressors.

Series are transformed by taking logarithms and/or differencing to obtain approximately stationarity. Generally speaking, first differences of logarithms are used for real variables, first differences are used for series already expressed in rates (nominal rates) and second differences of logarithms for price series following Mol et al. (2008), Giannone et al. (2004) and Giannone et al. (2008).

As in Tudebusch (1995), we use the federal fund rate as an indicator of the exogenous changes in monetary policy (monetary policy instrument). He recognized that a component of the unanticipated move in the FFR would reflect the Federal Reserve's endogenous response to the economy. According to Christiano et al. (1999), the level of real economic activity is well measured by changes in the number of employees on non-farm payrolls (EMPL). The level of prices is measured by the consumer price index. Let us define CPI as the consumer price index, EMPL as the employees on non-farm payrolls and FFR as the federal fund rate. These three series are the variables of special interest.

4.2 Forecasting evaluation

The penalization methods depend critically on penalty parameter selection for their performance in model selection, parameter estimation and prediction accuracy. To simulate real-time forecasting, we conduct an out-of sample experiment. Let us denote the longest forecast horizon to be evaluated by H, and the beginning and the end of the evaluation sample period by T_0 and T_1, respectively. We choose an optimal band by cross validation procedure. The point estimate of the h-step-ahead forecast is computed as $\widehat{y}^{(\lambda,\gamma,\alpha)}_{t+h|\sigma(t)} = (y^{(\lambda,\gamma,\alpha)}_{1,t+h|\sigma(t)}, y^{(\lambda,\gamma,\alpha)}_{2,t+h|\sigma(t)}, \ldots, y^{(\lambda,\gamma,\alpha)}_{J,t+h|\sigma(t)})^\top$, where $J = 131$ is the number of variables included in the equation. The point estimate of the jth variable's forecast is denoted by $\widehat{y}^{(\lambda,\gamma,\alpha)}_{j,t|\sigma(t)}$. Forecasts h steps ahead are computed recursively. For our case, the point estimate of the one-step-ahead forecast is computed as in equation (2.7). For a given forecast horizon h, in each period $T = T_0, \ldots, T_{1-h}$, we compute h-step-forecasts $Y^{(\lambda,\gamma,\alpha)}_{T+h|\sigma(T)}$ using the information only up to time T.

The accuracy of predictions is measured using the mean-square forecast error (MSFE) metric, given by:

$$MSFE^{(\lambda,\gamma,\alpha)}_{j,h} = \frac{1}{T_1 - T_0 - h - 1} \sum_{t=T_0}^{T_1-h} (\widehat{y}^{(\lambda,\gamma,\alpha)}_{j,t+h|\sigma(t)} - y_{j,t+h|\sigma(t)})^2$$

We report results for MSFE relative to the benchmark,i,e.,

$$RMSFE^{(\lambda,\gamma,\alpha)}_{j,h} = \frac{MSFE^{(\lambda,\gamma,\alpha)}_{j,h}}{MSFE^{(0)}_{j,h}}$$

Note that when the number of RMSFE is smaller than one, the large VAR model with tuning parameters λ, γ, α performs better than the naive model, as also considered by Bańbura et al. (2010)

The parameters are set to yield a desired fit for the variables of interest between T_0 and T_1. In other words, to obtain the desired magnitude of fit, the search is performed over a grid of three tuning parameters γ, λ and α to minimize the $\sum_{j=1}^{J} RMSFE^{(\lambda,\gamma,\alpha)}_{j,h}$, tuning parameters γ_j, λ_j and α to minimize $RMSFE^{(\lambda,\gamma,\alpha)}_{j,h}$ for the no grouping case, and tuning parameters λ_j, γ_j, η_j and α for the segmentized grouping respectively.

Now let's explain the strategy for how to choose the three shrinkage hyperparameters. We prefix α to be 1 and 2 firstly, and then do the search of other two tuning parameters (λ's and γ's) over loose grids. After the large stepwise search. we do a denser search around them for the nice performing λ's and γ's. The $parfor$ command in Matlab is used to facilitate parallel computations to fasten this process. Besides, recent improvements in computing and algorithms broaden the approach to allow for discovering relevant predictive variables when ensuring high prediction performance. For example, we make the computation time together with the parameter selection very moderate in this experience using new developed packages.

4.3 Results on the Large VAR model

We evaluate the forecast performance of large VARs for three key series (FFR, CPI, EMPL) over the period between $T_0 = January1970$ and $T_1 = December2003$. Forecast horizons are set to be one month, three months, half a year and one year ($h = 1,3,6,12$). The rolling estimates with a window of 10 years are considered in this thesis. For example, parameters are estimated at each time T using the data of the most recent 10 years.

We report results for FFR, EMPL and CPI in this work. All procedures are applied to standardized data. Mean and variance are re-attributed to the forecasts accordingly. Now we get the value of MSFE with respect to different choices of the forecast horizon and the order of VAR (lags), which is set to be $P = 1,4,7,13,25$ for comparison purpose.

Table 4.1 represents the RMSFE for forecast horizon $h = 1,3,6,12$ and an comparison to the result of Bayesian VAR (listed in the last column) is also made. Our empirical result shows that large VAR can provide significant gains in forecasting precision, which has been shown to lead to significant improvement for forecasting using BVAR. Three main results emerge from Table 4.1.

As we can see, firstly, large VAR method is very satisfactory in terms of prediction accuracy and stability. This result shows the "value" of large information set for forecasting and "value" of the new method for handling the large-dimensionality problem. Compared to the information criteria based on lag selection techniques, the RMSFE for large VAR is very robust to the initial choice of P, which primarily benefits from the "re-weighting over lags" technique we used before. We pay less attention to the more distant lags and more attention to the more recent lags. Since some variables have relatively long-term effect, the lags for them might be large. In order to retaining long-term information, a large enough P is initially considered to allow flexibility. Thus, we avoid the problems of over-fitting.

Secondly, our empirical analysis shows, for FFR especially, the outcome of large VAR is quite satisfactory acoss all horizons, which mainly results from the fact different time series might have quite different behaviors.

However, and this is the third important result, for the one-step-ahead forecast, this method gives superior forecasting hallmark to that of BVAR for the forecasting of all variables included in the table, while when for the p-step-ahead forecast ($p \geq 3$), it outperforms for the forecasting of EMPL and FFR. The forecast results of CPI do not improve over the BVAR, and even get worse.

4.4 Comparison of the FAVAR and BVAR

Factor Models have been shown to be very successful at forecasting macroeconomic variables with many predictors. Mol et al. (2008) and White (2006) performed a comparison of fore-

		P=1	P=4	P=7	P=13	P=25	BVAR
	EMPL	0.3333	0.3336	0.3338	0.3341	0.3335	0.46
h=1	CPI	0.3623	0.3618	0.3613	0.3621	0.3623	0.50
	FFR	0.4279	0.4281	0.4281	0.4284	0.4287	0.75
	EMPL	0.5191	0.5188	0.5192	0.5191	0.5189	0.38
h=3	CPI	0.4990	0.4992	0.4986	0.4995	0.4996	0.40
	FFR	0.4615	0.4614	0.4619	0.4617	0.4628	0.94
	EMPL	0.4730	0.4730	0.4735	0.4729	0.4736	0.50
h=6	CPI	0.4880	0.4874	0.4884	0.4885	0.4891	0.40
	FFR	0.5237	0.5242	0.5243	0.5243	0.5250	1.29
	EMPL	0.4997	0.4991	0.4992	0.4997	0.5002	0.78
h=12	CPI	0.4689	0.4687	0.4689	0.4694	0.4686	0.44
	FFR	0.4201	0.4199	0.4201	0.4200	0.4216	1.93

Table 4.1: The table reports MSFE relative to that from the benchmark model for employment(EMPL), CPI and federal funds rate(FFR) for different forecast horizons h and different models. LargeVAR

casts on Bayesian regression and principal components regression. In these experiments, they transformed variables to stationarity, as the standard practice in the literature about principal component analysis. The Bayesian regression is estimated as a single equation. Bańbura et al. (2010) made an exercise in which factor models are compared with standard VAR specification in the macroeconomic literature. Instead of using a set of single equations, they treated variables in levels and estimated the model as a system. That is the FAVAR method suggested by Bernanke et al. (2005) and discussed by Stock and Watson (2002b). The key insight behind FAVAR is that they use a small VAR augmented by principal components extracted from the panel data. Bańbura et al. (2010) showed the results of Bayesian VARs with those from the factor augmented VAR (FAVAR) of Bernanke et al. (2005) and showed that the BVAR is an appropriate tool for forecasting and structural analysis when it is desirable to condition on a large information set.

They extracted principal components from the large panel including 131 variables. By taking first differences, variables were also made stationary. As principal components were not scale invariant, factors were computed on standardized variables, recursively at each time point in the sample of evaluation.

The specifications with one and three factors are considered and different lag specification are discussed for the VAR. As Bernanke et al. (2005), they chose $p = 13$ and p was selected by the BIC criterion. In these two methods, variable selection is carried out first, and then the corresponding estimate's performance with respect to different numbers of lags are compared through the criterion above so as to select the optimal number of lags. This amounts to the case that lag numbers are prefixed. Table 4.2 reports the results. BVAR column recalls results with the number of lags chosen by applying Bayesian shrinkage for comparison. As can be seen, the two tables above show that both methods are in general outperformed by the large VAR.

		FAVAR 1 factor			FAVAR 3 factor			Large
		P=13	P=BIC	BVAR	P=13	P=BIC	BVAR	BVAR
	EMPL	1.36	0.54	0.70	3.02	0.52	0.65	0.46
h=1	CPI	1.10	0.57	0.65	2.39	0.52	0.58	0.50
	FFR	1.86	0.98	0.89	2.40	0.97	0.85	0.75
	EMPL	1.13	0.55	0.68	2.11	0.50	0.61	0.38
h=3	CPI	0.80	0.49	0.55	1.44	0.44	0.49	0.40
	FFR	1.62	1.12	1.03	3.08	1.06	0.99	0.94
	EMPL	1.33	0.73	0.87	2.52	0.63	0.77	0.50
h=6	CPI	0.74	0.52	0.55	1.18	0.46	0.50	0.40
	FFR	2.07	1.31	1.40	3.28	1.45	1.27	1.29
	EMPL	1.15	0.98	0.92	3.16	0.84	0.83	0.78
h=12	CPI	0.95	0.58	0.70	1.98	0.54	0.64	0.44
	FFR	2.69	1.43	1.93	7.09	1.46	1.69	1.93

Table 4.2: The table reports MSFE for the FAVAR model relative to that from the benchmark model (random walk with drift) for employment (EMPL), CPI and federal funds rate (FFR) for different forecast horizons h. FAVAR includes 1 or 3 factors and the three variables of interest. The system is estimated by OLS with number of lags fixed to 13 or chosen by the BIC and by applying Bayesian shrinkage. For comparison the results from large Bayesian VAR are also provided. Sourse: Bańbura et al. (2010) and Bernanke et al. (2005)

5 Discussion

Several authors have already studied the properties of the Lasso approach. For example, Knight and Fu (2004) showed that, under some appropriate conditions, the Lasso is consistent for the regression parameters estimation. Zou (2006) further studied the the properties of variable selection and estimation of Lasso. N.Meinshausen and P.Bühlmann (2006) showed that Lasso is variable selection consistent but under some conditions, not in general. Furthermore, even if it is variable selection consistent, the evidence that it is efficient for estimating the nonzero parameters is not enough. Therefore, these studies confirm that the Lasso does not enjoy the oracle property as Fan and Li (2001) and Fan and Peng (2004).

Built on the techniques of the proofs in Lounici (2008), Bickel et al. (2009), Lounici et al. (2009) and Wang et al. (2007), Song and Bickel (2011) have shown that through reweighing over time, our estimator in (2.8) could produce an estimator as efficient as an oracle and this method could produce a sparse solution for significant coefficients consistently. They have also discussed the risk bond of the estimator. See Song and Bickel (2011) for more details.

One of the possible model inadequacies diagnosed here is the non-stationarity, which has not been explored in this work. For a typical macroeconomic data set, the non-stationarity comes from the seasonality, business cycle and economic developments. In order to avoid the "spurious regression" mentioned by Granger and Newbold (1974), we took the difference for variables in the panel following Bernanke et al. (2005). More precisely, in our empirical study, series are transformed by taking logarithms and/or differencing to obtain approximately stationarity. However, the series are misspecified in the presence of cointegration and this will induce the nonstationary variables to lose the long-run information.

Motivated by Banerjee and Marcellino (2008), we may combine the Vector Error Correction Model (ECM) with large VAR. We believe that the FAVECM represents an potentially useful modeling approach, and a natural generalization of the FAVAR and ECM. The first one could include long run information while the second could include information from a large set of cointegrated series.

Banerjee and Marcellino (2008) proposed a Factor-Augmented Vector Error Correction Model, which could be regarded as a general version of the existing FAVAR model. The FAVECM methodology is a refinement of dynamic factor models, since it allows to include the error correction terms into the model, preventing the errors from being non-invertible MA processes. Hence it extends the FAVAR model to analyze long-run and short-run dynamics of non-stationary variables and combines the advantages of factor model and the VECM model. According to the standard practice, VAR(p) could be directly applied if the null hypothesis of unit root test, Augmented Dickey Fuller test for instance, is rejected. Otherwise, it is necessary to investigate

whether there exist cointegration among variables. The basic definition of cointegration raised by Engle and Granger (1987) is that: two or more time series are cointegrated if they each share a common type of stochastic drift: that is, to a limited degree they share a certain type of behavior in terms of their long-term fluctuations, but they do not necessarily move together and may be otherwise unrelated. If there exists no cointegration, we could simply take the difference for these variables. Otherwise, VECM should be used for further estimation. Therefore, the VAR part should be substituted by VECM when the variables are nonstationary as advocated by Banerjee and Marcellino (2008). However, studying this extended model is not so simple and deserves further research.

According to Song et al. (2010), we could add the non-stationary component $U\Gamma$ to equations as below:

$$Y_t = Z_t\Gamma + \widetilde{Y},$$

$$\widetilde{Y}_t = X_t B + U_t,$$

where $Z = (Z_1(t), \ldots, Z_R(t))^\top$ contains R basis functions of time containing fourier series with different frequencies and segment by segment ortho-normal polynomials with the corresponding $R \times J$ coefficient matrix Γ. X_t and B here are the same as in (2.3).

If we want to consider the cointegration, rank test and causal test, what we need for the high dimensional time series are not the case for the univariate time series, but the high dimensional simultaneous tests, which is much more difficult than the traditional ones.

In this work, the heteroscedastic structure of the error term U_t has not been taken into consideration. We carried out a similar approach of Mol et al. (2008) that all variables in the panel are standardized and demeaned in the empirical section. Now let's consider $Cov(U_t) = \Sigma$ with nonzero off-diagonal terms in Σ. Suppose we have a consistent estimate $\widehat{\Sigma}$ for the $J \times J$ matrix Σ with Cholesky decomposition $\widehat{\Sigma} = C^\top C$, where C is an upper triangular matrix with another upper triangular matrix inverse D. D is the off-diagonal terms of C.

Suppose all diagonal terms of the matrices $\widehat{\Sigma}$, C, D are 1. Then we generate $\widetilde{X}_t$ which equals to $X_t D$ by transforming the original X_t by D s.t. $Cov(U_t D)$ is an identity matrix. Therefore, we are selecting the linear transformations of the original variables rather than selecting themselves. Song and Bickel (2011) have showed that this does not affect the inference:

$$
\begin{aligned}
\widetilde{\beta}_1\widetilde{x}_{t1}+\widetilde{\beta}_2\widetilde{x}_{t2}+\ldots+\widetilde{\beta}_J\widetilde{x}_{tJ} &= \widetilde{\beta}_1 x_{t1}+\widetilde{\beta}_2(d_{12}x_{t1}+x_{t2})+\ldots+\widetilde{\beta}_J(\sum_{j=1}^{J-1} d_{jJ}x_{tj}+x_{tJ}) \\
&= (\widetilde{\beta}_1+\sum_{j=2}^{J}\widetilde{\beta}_j d_{1j})x_{t1}+(\widetilde{\beta}_2+\sum_{j=3}^{J}\widetilde{\beta}_j d_{2j})x_{t2}+\ldots+\widetilde{\beta}_J\widetilde{x}_{tJ}
\end{aligned}
$$

If the off-diagonal entries of D are much smaller than 1, the selected sets of local minimizer $\widetilde{\beta}$'s for regression and autoregression coefficients are likely to be the same as the selected nonzero sets $\widehat{S_1}=\{1\le i\le P(J-1),\widehat{c_i}\neq 0\}$ and $\widehat{S_2}=\{1\le p\le P,\widehat{d_p}\neq 0\}$ of β's, which have been shown to be the same as the oracle ones proved in Song and Bickel (2011).

6 Conclusion

In this work, we study the application of the large VAR the case for which a large set of variables, which is a very important question not only from a theoretical point of view, but also for empirical application. The large VAR represents an interesting modeling approach, and the empirical analysis shows that this method produces forecasts drastically superior to BVAR. In this thesis, we assess the performance of the large VAR methods for large panels of macroeconomic data. Based on our favorable results regarding large VAR we can conclude that for large dynamic models the large VAR can be seen as a robust alternative in building macroeconomic prediction models.

The novelty of this article lies in the following perspectives. Firstly, this new methodology could achieve the consistency of variable selection even under the dependent scenario, e.g., to reveal the equilibrium among the serial correlation, high dimensional dependent structure and moderate sample size. Secondly, large VAR method is able to do variable selection and lag selection simultaneously and is rather robust to the initial choice of lags. Therefore, we neglect the "interaction" between variable selection and lag selection. Besides, it allows individualized endogeneity and exogeneity. We differentiate the variable of interest's own lags from the lags of other variables. The relative weights are allowed to be varied when predicting different variables, while in other literature, they are assumed to be the same. All of the advantages have been confirmed by the real forecasting result in the previous section and come with a relatively low computational cost.

To conclude, large VAR approach is of potential usefulness in solving the challenge from a mixture of serial correlation, high dimensional dependence structure and moderate sample size for a wide range of macroeconomic forecasting. Moreover, related developments would occur in many fields well beyond macroeconomics.

Bibliography

Bańbura, M., Giannone, D., and Reichlin, L. (2010). Large bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1):71–92.

Banerjee, A. and Marcellino, M. (2008). Factor-augmented error correction models. *CEPR Discussion Papers 6707*.

Bernanke, B., Boivin, J., and Eliasz, P. (2005). Measuring monetary policy: a factor augmented autoregressive (favar) approach. *Quarterly Journal of Economics*, 120:387–422.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *Annals of Statists*, 37(4):1705–1732.

Box, G. E. P. and Jenkins, G. M., editors (1976). *Time Series Analysis, Forecasting and Control*. Holden-Day, San Francisco.

Burns, A. F. and Mitchell, W. C. (1946). Measuring business cycles. *National Bureau of Economic Research*.

Candes, E. and Tao, T. (2005). The dantzig selector: statistical estimation when p is much larger than n. *Preprint, Department of Computational and Applied Mathematics, Caltech*.

Christiano, L. J. and Eichenbaum, M. (1992). Identification and the liquidity effect of a monetary policy shock. *Political Economy, Growth and Business Cycles*, pages 335–370.

Christiano, L. J., Eichenbaum, M., and Evans, C. (1999). Monetary policy shocks: What have we learned and to what end? *Handbook of Macroeconomics*, 1(1):165–148.

Chudik, A. and Pesaran, M. (2007). Infinite dimensional vars and factor models. *Cambridge Working Papers in Economics 0757, Faculty of Economics, University of Cambridge*.

Diebold, F. X., editor (1998). *Elements of Forecasting in Business, Economics, Government and Finance*. South-Western College Publishing, Cincinnati, Ohio.

Doan, T., Litterman, R., and Sims, C. A. (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews*, 3:1–144.

Engle, R. F. and Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation and testing. *Econometrica*, 55:251–276.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist*, 32:928–961.

F.Canova and Ciccarelli, M. (2004). Forecasting and turning point predictions in a bayesian panel var model. *Journal of Econometrics*, 120(2):327–359.

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

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2005). The generalized dynamic factor model: One-sided estimation and forecasting. *Journal of the American Statistical Association*, 100:830–840.

Friedman, J., Hastie, T., and Tibshirani, R. (2007). Sparse inverse covariance estima- tion with the graphical lasso. *Biostatistics*, 9:432–441.

Geweke, J. (1977). The dynamic factor analysis of economic time-series models. *Latent Variables in Socioeconomic Models, eds. D. J. Aigner and A. S. Goldberg, Amsterdam: North-Holland*, pages 365–383.

Giannone, D., Reichlin, L., and Sala, L. (2004). Monetary policy in real time. *NBER Macroeconomics Annual, National Bureau of Economic Research, Inc, NBER Chapters*, 19:161–224.

Giannone, D., Reichlin, L., and Small, D. (2008). Nowcasting: the real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4):665–676.

Gonzalo, J. and Granger, C. (1995). Estimation of common long-memory components in cointegrated systems. *Journal of Business and Economic Statistics*, 13:27–35.

Gosselin, M. A. and Tkacz, G. (2001). Evaluating factor models:an application to forecasting inflation in canada. *Bank of Canada, Working Paper*.

Granger, C. W. J. and Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics*, 2:111–120.

Hastie, T., Rosset, S., Tibshirani, R., and Zhu, J. (2004). The entire regularization path for the support vector machine. *NIPS*.

Huang, J., Ma, S., and Zhang, C. (2008). Adaptive lasso for sparse highdimensional regression. *Statistica Sinica*, 18:1603–1618.

Huang, J. and Zhang, T. (2009). The benefit of group sparsity. *ArXiv e-prints*.

Knight, K. and Fu, W. J. (2004). Asymptotics for lasso-type estimators. *Ann. Statist*, 28:1356–1378.

Litterman, R. (1986). Forecasting with bayesian vector autoregressions: five years of experience. *Journal of Business and Economic Statistics*, 4:25–38.

Ljungqvist, L. and Sargent, T. J., editors (2004). *Recursive Macroeconomic theory*. The MIT Press, Massachusetts.

Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electronic Journal of Statistics*, 2:90–102.

Lounici, K., Pontil, M., Tsybakov, A. B., and van de Geer, S. (2009). Taking advantage of sparsity in multi-task learning. *Proceedings of Conference on Learning Theory (COLT)*.

Lucas, R. E. (1980). Methods and problems in business cycle theory. *Journal of Money, Credit and Banking*, 12(2):696–715.

Masten, A. B. . M. M. . I. (2010). Forecasting with factor-augmented error correction. *iscussion Papers 09-06r, Department of Economics, University of Birmingham*.

Mol, C. D., Giannone, D., and 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.

N.Meinshausen and P.Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34:1436–1462.

Park, B. U., Mammen, E., Härdle, W., and Borak, S. (2009). Time series modelling with semiparametric factor dynamics. *Journal of the American Statistical Association*, 104(485):284–298.

Park, M. Y. and Hastie, T. (2006). An l1 regularization-path algorithm for generalized linear models. *JRSSB*, 69(4):659–677.

Schmidt, M. (2005). Least squares optimization with l1-norm regularization. *CS542B Project Report*.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48:1–48.

Sims, C. A. (1986). Are forecasting models usable for policy analysis? *Federal Reserve Bank of Minneapolis Quarterly Review*, 10:2–16.

Sims, C. A. (1992). Interpreting the macroeconomic time series facts: The effects of monetary policy. *European Economic Review*, 36:975–1000.

Sims, C. A., Leeper, E. M., and Tao, Z. (1996). What does monetary policy do? *Brookings Papers on Economic Activity*, 2:1–63.

Sims, C. A. and Sargent, T. J. (1977). Business cycle modeling without pretending to have too much a priori theory. *New Methods of Business Cycle Research. Minneapolis: Federal Reserve Bank of Minneapolis.*

Sims, C. A. and Tao, Z. (1995). Does monetary policy generate recessions? *Manuscript*.

S.Kim (2001). International transmission of u.s. monetary policy shocks: evidence from var. *Journal of Monetary Economics*, 48(2):339–372.

Slutsky, E. (1927). The summation of random causes as the source of cyclic processes. *Econometrica*, 5:105–146.

Song, S. and Bickel, P. J. (2011). Large vector auto regressions. *Working Paper*.

Song, S., Härdle, W., and Ritov, Y. (2010). Dynamic factor models for high dimensional nonstationary time series. *Under revision*.

Stock, J. H. and Watson, M. W. (1998). Adiffusion indexes. *NBER Working Papers 6702, National Bureau of Economic Research, Inc*.

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

Stock, J. H. and Watson, M. W. (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics*, 20(2):147–162.

Stock, J. H. and Watson, M. W. (2005). Implications of dynamic factor models for var analysis. *NBER Working Papers 11467, National Bureau of Economic Research*.

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

Tudebusch, G. D. (1995). Federal reserve interest rate targeting, rational expectations, and the term structure. *Journal of Monetary Economics*, 35(2):245–274.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE Trans. Inf. Theor*, 55:2183–2202.

Wang, H., Li, G., and Tsai, C. L. (2007). Regression coeffcient and autoregressive order shrinkage and selection via the lasso. *Journal Of The Royal Statistical Society Series B*, 69(1):63–78.

White, R. G. . H. (2006). Tests of conditional predictive ability. *Econometrica, Econometric Society*, 74(6):1545–1578.

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):49–67.

Yule, G. U. (1927). On a method of investigating periodicities in disturbed series, with special reference to wolfer's sunspot numbers. *Philosophical Transactions*.

Zellner, A. (1992). Statistics, science and public policy. *Journal of the American Statistical Association*, 87:1–6.

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

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B*, 67:301–320.

List of Tables

2.1 Comparison table 16

4.1 MSFE relative to that from the benchmark model for employment(EMPL), CPI and federal funds rate(FFR) for different forecast horizons h and different models 24

4.2 MSFE for the FAVAR model relative to that from the benchmark model (random walk with drift) for employment (EMPL), CPI and federal funds rate (FFR) for different forecast horizons h 25

Declaration of Authorship

I hereby confirm that I have authored this master thesis independently and without use of others than the indicated sources. Where I have consulted the published work of others, in any form (e.g. ideas, equations, figures, text, tables),this is always explicitly attributed.

Berlin, July 15th, 2011 Ye Hua

Selbständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Berlin, den 15.07.2011 Ye Hua