LUISS Guido Carli

LIBERA UNIVERSITÀ INTERNAZIONALE DEGLI STUDI SOCIALI

Bachelor degree thesis in Applied Statistics and Econometrics

Shrinkage estimators in macroeconomic forecasting. Bayesian VAR

Candidate

Danila Maroz

Coordinator

Giuseppe Ragusa

Academic year 2014/15

Libera Università Internazionale degli Studi Sociali Guido Carli

Faculty of Economics and Finance

Course in Economics and Business

Contents

Abstract

The following work analyzes the reasons and implications of the use of Bayesian VAR, as a model with shrinkage estimators, in short-run macroeconomic forecasting. As well the BVAR model performance relative to benchmark cases is assessed and the proposition on the suitable value of the shrinkage parameter is made. The beginning is devoted to the description of the shrinkage methods in general forecasting setting. Then the features of macroeconomic forecasting are introduced and two competitive models are described. Finally we make a performance comparison and conclusions on the use of BVAR in macro forecasting, suggesting further improvements.

Introduction

The following work first concentrates on the description, contextualization and application assessment of the Bayesian vector autoregression approach to the parametric forecasting modelling problem in the scope of macroeconomic forecasting. The use of the abovementioned method permits to both incorporate the macroeconomic theory in the model and still get all the advantages of the flexibility of non-structural models (models that don't try to mimic the functioning of the economy). The forecasting model in the presented work aims to find differences in the predictions of three main macroeconomic variables: inflation (measured with the CPI), GDP in constant prices and the short term interest rate for the Italian economy.
Chapter 2 is devoted to the description of the shrinkage models used in forecasting. This chapter describes the evolution of the shrinkage methods through the description of three different shrinkage estimation models developed throughout the last half a century. The first model to be described is Ridge regression; then move on to the case of Lasso regression and conclude with the introduction of Bayesian method in the context of the use of shrinkage.
Chapter 3 briefly summarizes the peculiarities of the application of the general forecasting problem to the case of macroeconomic forecasting and then describe two competing methods, namely DFM (dynamic factor model) and Bayesian vector autoregression. The similarities and differences between two models are highlighted and suppositions on the eventual relative performance are made.
Chapter 4 deals with the numerical analysis. We will first describe the data and the precise specifications of the model used in the further analysis. Then we perform the tests of relative performance with regards to different benchmark models and draw conclusions on the applicability and possible improvements of the method.
The essence of the Bayesian approach in the regression parameters estimation is that of weighted average between the prior (the value of the parameter before the estimation was made) and the estimates obtained with the regression. More interestingly, the Bayesian approach permits us to work not only with the point estimates but also directly with the distributions. In such a setting a prior distribution should be specified which then would be "updated" with the new data to provide us the posterior distribution. This method permits us to obtain estimates using an expanding window of the data using the Kalman filter, that is a technique to recalculate the parameters of the model

(in this case update the prior with the newly estimated parameters) without the need to rerun the regression on the data that have already been used in the estimation process. Summing up the abovementioned, the application of the Bayesian method potentially has a lot of improvements over the standard models concerning theoretical base (usage of the distributions instead of the mean-standard error framework) as well as the practical issues.
In our analysis we empirically confirms the ties between the choice of the shrinkage parameter and the degree of predictability (versus theory that macro variables are MDS i.e. expected value of a variable conditional on the past is equal to the lagged value of this variable) of different variables of interest, which persists over different forecasting horizons. We find out that even a simply configured BVAR estimated with a medium dataset outperforms benchmark models for some forecasting horizons, performing relatively better for longer ones. We then present an empirical application of a conditional predictive ability test, describing its implications and results. We conclude with stressing the directions of change and improvement on the path of using Bayesian VAR as a model with shrinkage characteristics that particularly fits the macroeconomic environment.

Chapter 1

Regularization

In the following chapter we will discuss the advantages of Bayesian method in terms of its use of regularization. First we define the notion of regularization used in econometrics, we describe its use and, eventually explain the path undertaken by the development and application of the concept. In particular, after defining regularization, we discuss the particularities of its use in the Ridge regression, followed by the LASSO case, finalized by the Bayesian method.
Traditional regression tries to find optimal estimates of beta by minimizing the loss (MSE in our case). Thus it solves the following problem:

$$\hat{\beta} \;=\; arg\, \min_{\beta} (Y - X\beta)'(Y - X\beta) \tag{1}$$

Which yields the estimate:

$$\hat{\beta}^* \;=\; (X'X)^{-1}X'Y \tag{2}$$

The possible problem that could arise concerns invertibility of the $X'X$ matrix. In fact this matrix is not invertible for the case when X does not have a full rank. While theoretically impossible apart from the case of true perfect multicollinearity, in practice this problem could arise if the imperfect multicollinearity is present (when in form of a correlation matrix $X'X$ is not nearly a unit matrix). In such case the behavior of estimation and prediction outcomes might cause the estimates of the variance and the coefficient values be excessively high in certain cases (Hoerl and Kennard 1970). Thus, in the case of "kitchen sink" regressions with thousands of observations for each one of thousands of regressors, an attempt to run a regression is predisposed to fail[1].
The problem at hand is also related with the risk of overfitting added to the forecasting performance even in absence of the singularity problem. Overfitting derives from the lack of information needed to make a forecast that, being extrapolated out of sample, would take a reasonable value not overly influenced

[1] Even if the modern computational techniques permit to bypass the problem of imperfect multicollinearity, the benefits of using techniques that cope with the problem stretch beyond elimination of those difficulties, as we will see

by the propensity of the model to predict also the error term is the coefficient estimates.
There is a possible remedy (at least partial) to all of these problems, a process called regularization. In the essence regularization is a way of setting restrictions on the parameters of the regression. The initial reasons for this, as has been mentioned before, are the diminution of the possibility for multicollinearity and the lowering of the mean squared forecasting error, which positively depends on the quantity of the parameters to be estimated[2]. The following transformation of Mean Squared Forecasting Error formula helps to understand the influence of regularization on the forecasting performance:

$$MSFE = \sigma_\epsilon^2 + Bias^2(\hat{f}(z)) + Var(\hat{f}(z))$$

where σ_ϵ^2 is the variance of the forecasting error that cannot be influenced by the change of parameters or use of regularization and is usually assumed to take a certain value following econometric theory; $Bias^2(\hat{f}(z))$ is the squared bias of the forecast or, assuming OLS assumptions hold, the squared difference between estimates from the given model and OLS; $Var(\hat{f}(z))$ is the variance of the forecast. Further on we will analyse the influence of regularization on the last two components on the MSFE.
The constraints put on the parameters take on the following form:

$$k(\beta) \leq K$$

while the loss function modifies in order to accommodate for the constraints through the use of the Lagrange multipliers, thus imposing a penalty on the loss:

$$\hat{\beta} \ = \ arg \min_{\beta} (Y - X\beta)'(Y - X\beta) + \lambda(K - k(\beta))$$

where $k(\beta) = \beta' A \beta$. The solution is:

$$\hat{\beta}^* \ = \ (X'X + \lambda A)^{-1} X'Y$$

Thus, we are now facing the choice of A and λ.
The imposition of A and λ as the parameters defining constraints on β estimates has a number of properties regarding the performance of the model in terms of forecasting. As a base we take the MSFE. The performance of the forecast is determined by several components, one being the variance of the error term which could not be influenced neither by the choice of the model nor by the imposition of the constraints, the second component being dependent on the bias in the estimation of the coefficients and the other being determined by the variance in the coefficient estimates. While producing unbiased estimates, OLS method might result in an excessive variance of the parameter estimates, not considering the possibility of misspecification (in which the influence on the bias might be more complex). The problem is of extreme importance in the case of a large number of predictors included in the model since the variance

[2]The implicit assumption of the paper is that we operate with the MSE loss

increases as the number of included regressors grows. Regularization, or putting constraints on the parameters helps reduce their variance thus helping improve the forecasting abilities of the model. Of course the gain on the side of the variance comes at a cost. Introduction of constraints inevitably introduces bias in the estimates, unless λ is equal to zero which would reduce the model to OLS. The new model, which regularizes the coefficient has now two new variables to be determined: λ and A. Let's condition our future elaborations first on the choice of A and then describe the consequences and particularities for any choice of λ, eventually trying to give a meaningful interpretation also to the latter.

1.1 Ridge regularization

When the matrix A is set to be the unit matrix I, the regression takes on the form of Ridge regression. One of the immediate effects of applying the regularization is that sum of the matrices $X'X + \lambda I$ is now invertible even if $X'X$ by itself is not (the case of multicollinearity). Thus the use of the restrictions helps "regularize" the parameter estimation procedure into one with less prior requirements of "good behavior" of the data. Precisely the avoidance of multicollinearity problems was the initial intention of the Ridge regression.
After having chosen A we should proceed to the evaluation of the parameter α. Ridge regression leaves this parameter to be defined by the econometrician. First, let's understand what is the essence of α. For simplicity we assume that values of this parameter are the same for all the coefficients' constraints (the vector of λ consisting of the same constants), so we will speak as if there was a unique value. We start from noting that, if we set λ equal to zero, we obtain a regular OLS regression. As we increase the parameter the more restrictive the conditions on the β will become. To see this we can run the same regressions for different values of λ and see the corresponding effect on the coefficients (see Figure 1.1).

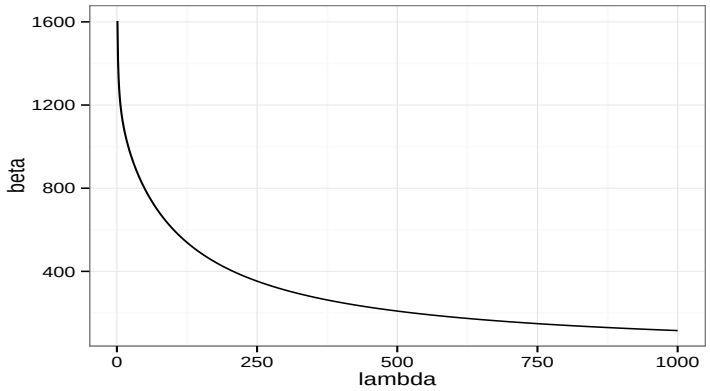


Figure 1.1: Example of the coefficient estimate path for different values of λ

An increase in λ results in the eventual diminution in the absolute values of the all coefficients' estimates apart from the intercept, while the intercept itself approaches the mean value of the variable being predicted. Thus we can conclude that higher values of λ correspond to tighter restrictions on the parameters. This observation corresponds to the theory, which defines λ as a *shrinkage operator*. Such a notation refers to the fact that λ influences the degree to which we want to enforce the regularizing restrictions on the coefficient estimates. The desire to constraint coefficients to be closer to zero would correspond to imposition of higher λ. In the limit case, if λ tends to infinity, the coefficients are constrained to be equal to zero(which is still never reached due to the typology of the constraint).
We can see the overall influence of the change in the parameter on the bias and the coefficient estimate variance on the the following figure.

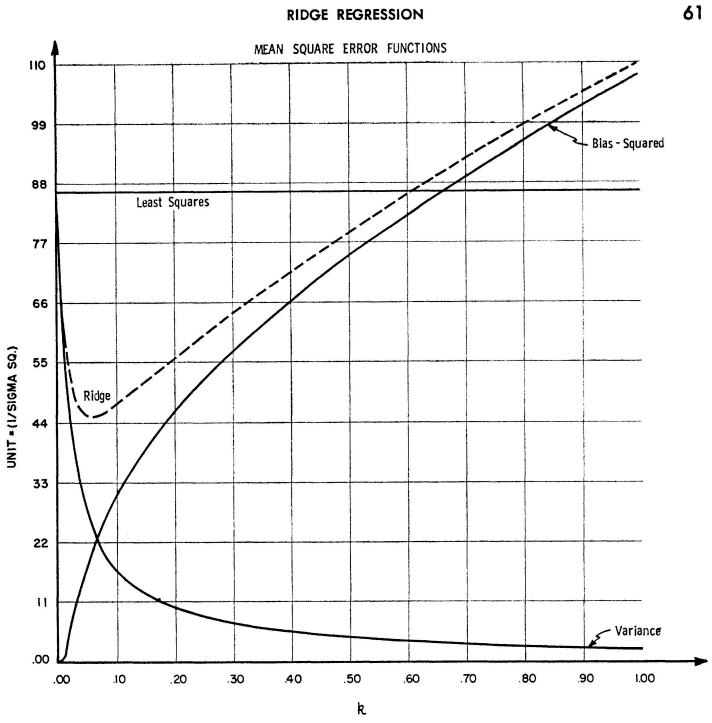


Figure 1.2: Bias versus variance tradeoff over λ

Ways of finding the optimal λ, i.e. such that would optimize the trade-off between the coefficient variance and estimate bias in the framework of assessment of the forecasting performance of the model, are in practice based on computational techniques. The common approach is to bootstrap the optimal value of λ following pseudo out-of-sample forecast assessment and using it for further

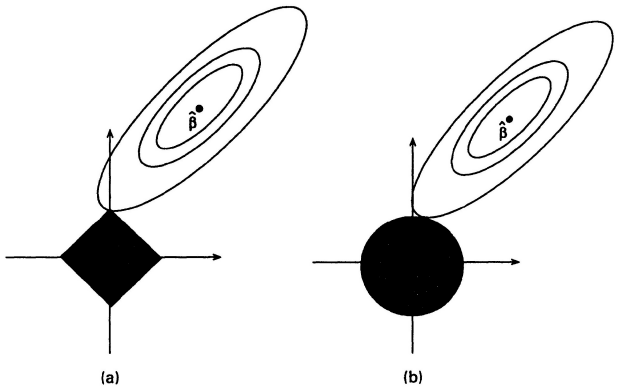
forecasting.
The important conclusion from Ridge regression is a mathematical proof of the fact that a λ exists such that $E[L_f^2(\lambda)] > E[L_f^2(0)]$ where $L_f^2(\lambda)$ is the loss associated with a given level of λ. That is, with a certain value of restriction parameter we are guaranteed to have a lower variance of the loss which, as we have shown before reduces the MSFE and improves forecasting performance. As stated in Hoerl and Kennard (1970), Marquardt (1970) the use of biased/shrinkage estimators in practical problems with nonorthogonal data presents a huge improvement in terms of MSE since the variance of the estimates is significantly reduced. Other advantages include the ability to work with less than full rank data matrices (when rank of $X'X$ is not full) which, as we will see further on, is yet another important aspect of shrinkage estimators.

1.2 LASSO regression

Let us now step away from the case when the matrix A, as one of the parameters of the regularization, is set to an identity matrix, to simplify the analysis. First we start with the general differences between the two models and, eventually, reconstruct the reasoning behind the need for these developments.
The first difference between Lasso and Ridge models is the penalty term included in the general expression to be minimized. While in the Ridge regression the restrictions are put on the squared values of the parameters, Lasso puts restrictions on the absolute values of the parameters:

$$\hat{\beta} \;=\; arg\,\min_{\beta} \sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{ij})^2 \quad subject\,to \sum_j |\beta_j| \leq t$$

As we have seen before, as the shrinkage parameter λ tends to infinity, the estimates of β in the Ridge case are shrunk towards zero. However the procedure does not permit to set them to zero directly even if the eventual values come out to be very small. The shortcoming of the Ridge model is addressed in Lasso. Due to linearity of the constraint on the coefficients, Lasso has a much higher probability for estimated coefficients to be directly equal to zero, which happens when the regressor under consideration has a sufficiently low predictive ability. The intuition to explain such model performance could be easily extracted from the Figure 1.3, representing the constrained minimization problem for both Ridge and Lasso cases.
As we see, the difference in the form of the constraints, which are colored in black, helps visualize why the probability of setting a coefficient precisely equal to zero is much higher in case of Lasso with respect to Ridge, where the constraint imposed precludes this outcome. Such peculiarity permits to feed Lasso regressions with data having a large quantity of the variables directly and let the model perform predictor selection by itself. The result of the regression will be a set of estimates for the coefficient of the variables that were found



Estimation picture for (a) the lasso and (b) ridge regression

Figure 1.3: Ridge vs Lasso estimation frameworks

relevant in predicting the dependent variable while the coefficients for the irrelevant variables would be equal to zero. Data rich environments, especially those with the number of variables exceeding the quantity of observations, can benefit extensively from the Lasso setting, regarding both the variable selection problem and the potential presence of multicollinearity.
Let us move on to the choice description of the influence of the shrinkage parameter in Lasso regression. The eventual optimal forecasts solve the following equation:

$$\hat{\beta}_{lasso} = \ arg \ \min_{\beta} \Big[(\hat{\beta}_i^{ols} - \beta_i)^2 + \lambda |\beta_i| \Big]$$

Again, λ plays the role of the shrinkage parameter, with higher values corresponding to the imposition of tighter restrictions on the minimized function. We can see the development path of different coefficients with different values of λ. The X-axis corresponds to the inverse of the value of λ.

The importance of Lasso model, as we see it from Figure 1.4, is that some estimated coefficients are set directly to zero for sufficiently high values of the shrinkage parameter. Thus, Lasso permits us to achieve two objectives at the same time: assess the parameters of the model to be used in forecasting and to perform the model selection whose results could be used in some other model or for theoretical purposes. Having discussed the benefits and specialities of Lasso, it is time to move to the case of Bayesian methods.

1.3 Bayesian shrinkage

In the previous two sections we have seen two different though logically similar techniques to impose restrictions on the estimated coefficients. Both methods' final effect is of curbing the estimates towards zero, or shrinking, which gave rise

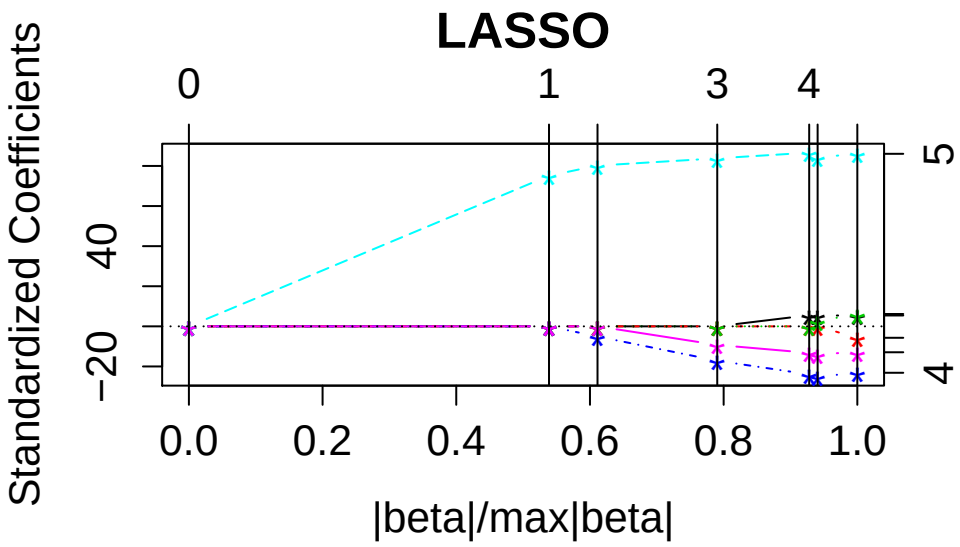


Figure 1.4: Ridge vs Lasso estimation frameworks

to the use of the notion. The passage from the imposition of the restriction on the target function to the derivation of the optimal estimates of the coefficients (under MSE) is easy to trace and understand. The presence of a single tuning parameter(or vector of parameters), λ, which regulates the tightness of the restrictions, and, consequently, the degree of shrinkage of the parameters, has two opposite arguments. The understanding of the model and the comparative evaluation remains simple, however the degree of customization or adjustment does not stretch too much further from the regular OLS estimation. Having argued for the case of usefulness of the shrinkage estimators as the ones producing biased but better performing estimates, we propose the Bayesian procedure for the shrinkage estimation.
Bayesian method requires introduction. The theory of Bayesian estimation is based on the concept underlying the Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

The idea of obtaining the probability of event A (the Y, or the variable to be predicted) conditional on event B (the X, or the data used in the regression) could be stretched from the case of probabilities of singular events to the density functions of the given variable. The resulting equality has the following form:

$$\pi(\theta|z) = \frac{p_Z(z|\theta)\pi(\theta)}{\int_\theta p_Z(z|\theta)\pi(\theta)d\theta}$$

In words, we can obtain the posterior distribution for a set of parameters θ of a given model, from the specification of the model for the data(the likelihood function), $p_Z(z|\theta)$, the marginal likelihood of the data, that lies in the denominator of the formula, and the prior distribution of the parameters, $\pi(\theta)$.
It is important to trace the differences in the imposition of information in the three models. The first two exercise a direct constraint on the estimates, which transform also in the constraint on the parameter variances. In the Bayesian case the imposition of the restricting information is done through the prior, which is a joint distribution of the parameters of the model. Thus we can explicitly introduce our shrinkage targets as in terms of the mean, so in terms of the variance. Having this in mind, we proceed with theoretical elaborations.
The target function of interest takes on the following form:

$$r(\pi, f) = \int_z \left(\int_\theta \left\{ \int_y L(f(z), y) p_Y(y|z, \theta) dy \right\} \pi(\theta|z) d\theta \right) m(z) dz$$

In case we have $y \sim N(X\beta, V)$ and a prior on β is normal: $\beta \sim N(\beta_0, \Sigma_0)$, the marginal distribution of X does not depend on β and V is known (for simplicity), the posterior is $\beta|Y, X \sim N(\tilde{\beta}, \Sigma)$ where

$$\tilde{\beta} = (X'V^{-1}X + \Sigma_0^{-1})^{-1} X'V^{-1}X\hat{\beta} + (X'V^{-1}X + \Sigma_0^{-1})^{-1}\Sigma_0^{-1}\beta_0$$

$$\Sigma = X'V^{-1}X + \Sigma_0^{-1}$$

Where $\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y$ is the Generalised Least Squares estimator. Here we can notice the similarity between the shrinkage concept and the Bayesian implementation. The coefficient estimates are curbed towards the prior values as expressed by mean and variance-covariance matrix. Thus, the degree of shrinkage of β estimates depends both on the prior expected value and the degree of our belief in the prior as expressed by the Σ_0 matrix. Following intuition, there might even be a direct relationship between Bayesian method and Ridge/Lasso implementations. In fact, for example, setting $V = \sigma^2 I$, $\beta_0 = 0$ and $\Sigma_0 = \tau^2 I$ results in the same outcomes as of Ridge regression ($\lambda = \sigma^2/\tau^2$):

$$\tilde{\beta} = (\sigma^{-2}X'X + \tau^{-2}I)^{-1}\sigma^{-2}X'X(\sigma^{-2}X'X)^{-1}\sigma^{-2}X'y = (X'X + \frac{\sigma^2}{\tau^2}I)^{-1}X'Y = \hat{\beta}_{ridge}$$

Concentrating on forecasting, the minimization under MSE loss and normally distributed priors gives us the following result:

$$E(y_{T+1}|z) = \frac{T\tau^2}{T\tau^2 + \sigma^2}\bar{y_T} + \frac{\sigma^2}{T\tau^2 + \sigma^2}\mu_0$$

Where τ^2 is the variance of the prior, μ_0 is the prior mean (assuming the prior is distributed normally with the stated parameters). As we see, the forecast of variable is a weighted average of the prior mean and the sample mean, whether weights depend on the quantity of the data available, T, the variance of the

sample data, σ^2, and the diffusion, or variance of the prior, τ^2. For a given set of data the first two parameters determining the degree of shrinkage (since the sample mean, that is the forecast for the OLS is shrunk to the prior mean) can not be determined by the forecaster. However τ^2, can be adjusted. We can think of it as of the credibility of the prior. The lower we set it, i.e. the more sure we are that the prior reflects the true value of the parameter, the closer the resulting estimated forecast will be to the prior mean.
Thus we conclude the chapter with the important parallel in mind: all of the abovementioned methods, using different means, perform the same task of reducing variance of the estimates while introducing bias. This tradeoff presents a possibility of improvement of the forecasting performance and, as well, carries a number of useful features, such as dealing with overfitting, controlling for stability of the estimates and permitting to deal with underidentified models.

Chapter 2

Macroeconomic forecasting

There exist several major points upon which we could judge upon the difference of forecasting methods in macroeconomics. First one is the loss function upon which the forecasting performance assessment is made. In the practical applications we do not have too much difference from the regular forecasting setting, even if there might be sufficient ground for introducing less popular alternatives. The most commonly used measure is the Mean Squared Forecasting Error (MSFE). [1]

The second issue at hand is the specificity of the data. While there are plenty of economic variables that are updated on a monthly basis (inflation) and some even have a less-than-daily frequency (interest rates, exchange rates), many major economic variables (GDP and components) have trimestral frequencies. Thus, supposing the trimestral data availability, the effective data span turns out to be not than long. For example in Europe [2], where the accounting processes converged only around the 1990s, in the best cases are able to provide no more than 100 observations. This problem of short data span (which was even more acute in earlier years) drags under the light the problem of underidentification. Richly specified models are deemed to suffer from that problem in macroeconomic environment. Because of this all the models are deemed to search for a solution and somehow extract the most information without creating the predisposition to excessive forecasting outcomes.

The third issue is the macroeconomic theory explanation of the processes underlying macroeconomic evolution and data generation. There exists a huge array of macroeconomic models that try to mimic the structure of the economy, thus pretending to approximate the true framework of the interactions in the system. Their initial reason was the assessment of causal and linking relationships be-

[1] We will not concentrate on this issue in our analysis but it should be mentioned that macroeconomic forecasting, as a major issue affecting economic policy, investment decisions and many other fields has plenty of ground for introducing non-symmetric, non-quadratic and other types of loss functions (for example over-valuation of GDP might be more harmful than undervaluation)

[2] In this paper we use data on Italy in the forecasting assessment

tween major economic variables and components. Then attempts were made to extrapolate this reasoning into the forecasting realm. Most of them had shown relative poor performance but some (DSGE models) successfully function and are currently used in important economic institutions (e.g. Boivin and Giannoni (2006)).
However there is a thick line between the forecasting and structural analysis. Another theory (which also had good practical approval) states that many economic processes (usually the logged differences of the variables) follow the Random Walk model (with or without drift):

$$Y_{t+1} = \beta_0 + Y_t + \epsilon_{t+1}$$

In this setting the forecasting problem reduces to extrapolating the behavior of the last period on the next period and the specification of the distribution of the error term. Seemingly simple, this model is hard to beat in practice (especially for the slowly changing variables, such as inflation), especially in the short run that is defined as a timespan of up to four periods/trimesters ahead. Such theoretical approach, proven by the practice, presents a valuable source of information which, as we will see, will be used in order to implement better forecasting.
In the rest of this chapter we will discuss and compare two approaches in macroeconomic forecasting: Bayesian vector autoregression (BVAR) and dynamic factor models (DFM). The issue at hand concerning the selection of the model type is to manage to incorporate most of the useful information for forecasting purposes without loss of the precision in the estimates and avoiding overfitting. The problem for both regular OLS and VAR that try include many regressors is the apparition of nonorthogonality issues and the physical impossibility to solve for coefficients in when the number of coefficients exceeds the data span. The two techniques address differently the problems of macroeconomic forecasting although both of them possess classical forecasting model features, such as combination of the structural approach (which might also be almost absent) and purely forecasting oriented methods, that aim for no other purpose but predicting the future values. The comparison follows.

2.1 Dynamic factor models

The general expression of dynamic factor models goes as following:

$$X_t = \lambda(L) f_t + \epsilon_t$$

$$f_t = \Psi(L) f_t + \eta_{t-1}$$

Where f_t are the latent factors following a dynamic path (the quantity of factors is much lower than the quantity of the variables in the set of data); L is the lag operator, while $\lambda(L)$ and $\Psi(L)$ stand for the matrices of respective coefficients for the lagged variables. The theoretical essence of the following equations is the idea that much of the variation in the macroeconomic variables could be

explained through the interaction of a small group of latent factors, and thus predictions can be modelled using the factors directly, which provides certain benefits described further on.
The actual reason for the use of this type of models is that in the case the factors are known and the respective assumptions are made on the distribution of the error terms η and ϵ, the forecaster can reduce the dimensions of the model from one assessing coefficients for all the regressors available to the one having to assess only the coefficients for the factors while still using data for all the variables. This manipulation permits to reduce the variance of the forecast thus improving upon the case with all the regressors included directly in the model.
It is important to note that although the dynamic factor model seemingly does not represent by itself a variation on the topic of models with shrinkage, in the reality it does perform the functions similar to those intended for shrinkage estimators. The first point to mention is the ability to obtain consistent forecasts when the number of variables is higher than the length of the time series. Such a problem arises acutely when using VAR, where the number of coefficients rises proportionally to the square of the number of variables included. In case of DMF the eventual regression for the variable of interest will include only the factors, their lags and the lags of the variable of interest.

$$Y_{t+1} = \alpha(L)F_t + \delta(L)T_t$$

Thus with the increase in the lag order the resulting number of coefficients to be estimated rises proportionally to the square of only the quantity of factors (considering that Y_t is a vector of the variables of interest), while rising linearly with lag order of regressants pf interest. In this perspective the use of factors acts as the regularization procedure in the Ridge and Lasso regressions, allowing to bypass the irregularity (insufficiency or multicollinearity) of the data under consideration.
The Dynamic Factor approach has a number of advantages and possible shortcomings. Among the negative aspects we could list the possibility of excessive reduction of information due to the fact of using factors. The insufficient number of factors included might result in underconsideration of some important components managing the underlying processes and eventually lead to poor forecasting performance. The same argument could be seem also from the opposite perspective. The act of information reduction (extraction from the data using factors) could be of beneficial influence due to discarding the influence of shocks to the realizations of the singular variables. Thus, the proper factor usage should grasp the dynamics of latent factors (the most important processes governing the realizations of the observable variables) and contribute positively to the forecasting performance especially in case of aggregate variables, such as GDP.
Important to remember is that usage of factors is a purely forecasting related instrument. While attempted to be rationalized and brought under theoretical explanation, factor usage is practically assessed with the actual performance of the models using it, which show robust effective forecasting (e.g. Ng and Boivin (2005), Bernanke, Boivin and Eliasz (2005))

Moreover, factors could be used also as auxiliary variables in many types of the models, e.g. Factor Augmented Vector Autoregressions (FAVAR) and Dynamic Standard General Equilibrium models (DSGE). Being flexible and non-constrictive in its use, dynamic factor framework presents a substantial space for improvements, especially in macroeconomic forecasting.

2.2 Bayesian VAR

The alternative to the use of dynamic factor models is the use of Bayesian approach in the estimation of vector autoregressions. Vector Autoregression (VAR) approach has gained extensive application in the macroeconomic setting. First, VAR models allow to incorporate interactions of all the variables at the same time thus being logical from the structural point of view (even if chaotic from the theoretical perspective). Second, vector autoregressions have a developed set to instruments which permits to analyse the structural shock consequences and perform scenario analysis (forecasting conditional on fixing the value of certain variables).
The advantage of the Bayesian approach is the absence of the requirement of the limitations on the quantity of the coefficients to be estimated. This problem is resolved through use of prior information that acts as a regularization parameter for the estimation procedure and thus permits to use all the available data to the full extent (unlike DFM that tries to extract the common root of the data generating processes which results very costly in case of misspecification of the factors).
One of the other advantages of the BVAR is that it permits to incorporate the economic theory in the model, which is undertaken by the setting of a specific prior on the data. One of the most popular and theoretically elaborated is the Minessota prior first introduced in Litterman(1979). The essence of the Minnesota prior is using the random walk as the prior for all the VAR equations. Following the idea of Litterman, there is a fair reason to suppose that macroeconomic processes do not follow any kind of predictable path, that is are random. Using this supposition as the point of departure, the BVAR adjusts the coefficients in such a way to concord with the data observed, with the tightness of the prior regulated by the shrinkage parameter. The tightness of the prior is autoregulated with regards to the obsolescence of the data, setting the more distant lag prior variances (or the tightness) to be lower and lower.

$$Var(A_l)_{ij} = \begin{cases} \frac{\pi_1}{l^2} & for\ i = j \\ \frac{\pi_1 \pi_2}{l^2} \left(\frac{\sigma_i}{\sigma_j} \right)^2 & for\ i \neq j \end{cases}$$

Where the variance of elements of the matrix of the autoregressive coefficients A_l, is dependent on the lag length l, parameter characterizing the tightness of the prior with respect to the random walk (shrinkage towards random walk) π_1, and parameter π_2 which regulates the tightness of priors on cross-variable effects. Such a setting permits to impose stricter priors on the data laying too far

in time from the forecasting date, which allows to model the fact that influence of the data on the variables fades away with passage of time.
Other advantage of the Bayesian approach is that the result of the optimization is not the point estimates of the coefficients which thus results in the production of point forecasts, but the distributions of the coefficients which eventually results in the production of the distribution of the forecasted variable. Density forecasting has a number of advantages over point forecasting since many more forecasting estimation procedures can be undertaken in the former case, not to mention that point forecasts themselves could be easily obtained from the predictive distribution.
Yet another radical difference between the Bayesian approach to the estimation problem and the frequentist approach (approach of maximizing the likelihood of observing the given data realization considering that model specification is known) is that the former explicitly addresses the uncertainty resulting from researchers' inexact knowledge of the "true" specification of the model, as put in Doan, Litterman Sims (1983). Thus Bayesian framework and the usage of shrinkage ought to be regarded not as the attempts to find a seemingly better answer with the wrong means but rather to address the inherent and ineradicable imperfection of pure frequentist inference.
Having discussed and compared different models that are related to the concept of shrinkage (and the respective effects of its usage) in the view of the forecasting problem in macroeconomic context, we now explore further our model of interest - Bayesian VAR.

Chapter 3

BVAR in macroeconomic forecasting

This chapter will entail the analysis of the application of the Bayesian vector autoregression to the problem of the forecasting of three major macroeconomic variables: GDP, inflation and short term interest rate, using the data for Italy. We begin by describing the model and its specifications. Then we briefly describe the data used in the regression coefficient evaluation and eventually proceed with the analytical part. Analysis will consist of the comparison of the model forecasting performance relative to several benchmark models using the unconditional and conditional specifications of the forecasting ability test described in Giacomini and White (2006). Next step will consider the analysis of the sensitivity of the forecasting performance of the BVAR model to the choice of the shrinkage parameter λ and propositions will be made on account of the λ choice implications.

The choice of the target variables we mention before is common in the macroeconomic setting since they present the main anchors of the reasoning about the performance of the economy. Moreover there is another interesting point of view regarding such choice. Each variable has quite different dynamics and thus by represents a different degree of inherent predictability and different decisions applied in the forecasting reasoning. As we see in the Appendix A, the degree of sensitivity and the dispersion of the variables is quite variant. In the absence of excessive economic shocks log difference in Gross National Product might seem the least erratic but in the time of strong instant shock it's reaction is most excessive. Instead inflation seems to have stronger persistence in the face of the shock, which reflects its slow-to-adjust nature. Short run interest rate, being the closest to be affected by the monetary policy and tied to the interest rate on main refinancing operations, exhibits quite heterogeneous behavior over time, however, without large consistent swings from the trend (several trends identifiable over the period).

We will use these perspective as one of the plains of our description of the model

performance, using, the unique setting of the model for modelling all the three variables. Such approach should permit us to note the structural differences and perfection the adjustment from the holistic perspective.
We first describe the specificities of the model used in our computations and then describe the data used in the model. Second, we analyse the choice of the shrinkage parameter, coordinating ourselves with the eventual measures of the forecasting performance and while also trying to grasp the underlying relationship between the inherent features of the processes governing the data generation and the eventual optimal levels of the shrinkage parameter. We make use of the visualization of the relationship and try to elaborate on the idea and framework behind the use of shrinkage in forecasting. Third, we will compare the performance of the model with different benchmarks: its predecessor, ordinary vector autoregression, and the typical standard models with low quantity of parameters (ARIMA, Random Walk).

3.1 Description of the model and the data

The model used in the analysis is Bayesian vector autoregression, which uses the Minnesota prior as the prior for the regression coefficients. Following macroeconomic theory it assumes the distribution of all the coefficients to be normal and centered in zero, apart from coefficients that stand for the lagged value of the regressant, which are centered on 1 (thus imposing random walk behavior prior). Even though our analysis presents the case of a conjugate prior[1], we will estimate the posterior using a sampling procedure, the Gibbs sampler. Even though it presents a simplification, usage of this method leaves space for improvement, usage of different priors and different assumptions on the likelihood function (which is normal in our case). The essence of the procedure is that by drawing samples from the underlying distributions we obtain a sample distribution of the posterior upon which we directly perform the analysis.
The list of the data used in the regression is given in Appendix B. In short, the major macroeconomic statistics, describing the structure of the economy (i.e. short run interest rate, CPI, GDP, consumption, investment, net exports and the decomposition of the latter GDP components) are used in combination with several other indices describing the overall economy. All the data have quarterly frequency and span from the first quarter of 1995 till the first quarter of 2014 included.
The following approach is used: first $41 - h$ observations are used in the initial regression and the further observations are used to evaluate the performance of the model for the forecast horizon h from 1 to 8 periods. Then, using the rolling window of the data, we recompute the regression, adding each time a new layer of the observations (while subtracting the last one) and again compare the resulting forecasts with the pseudo out-of-sample data for h periods ahead. Such

[1] A conjugate prior is one with which there is a closed-form formula for the computation of the posterior given the choice of the likelihood function

procedure results in obtainment of an array of matrices of forecasting errors for different forecasting horizons over different models. These results are further used in the evaluation of forecasting performance.

3.2 Sensitivity analysis. Choice of λ

Having decided on the choice of the prior distribution for the parameters of the regression, we now need to turn to the determining of the optimal level of the shrinkage coefficient, λ. First, let us observe the development of mean squared forecasting error (which we will use to assess the predictive performance of the model) for forecasts of different variables for different forecasting periods over different values of λ (Appendix C). As we can see, the optimal values of λ vary highly for different forecasting periods and different variables. Even if there there are only two variables of interest and a unique forecasting horizon is specified, an empirical choice of λ presents a difficulty. In our case we need to select such parameter value that would be suitable for different variables and forecasting horizons. Econometric theory does not possess a framework which solves the general problem of finding optimal shrinkage coefficients. Instead, numerical techniques, such as cross-validation are used in practice to find the best suiting parameter. Cross-validation implies finding the optimal values of the parameter for different subsamples of data and then averaging to obtain a unique value.

We, however, try to select the best value of the coefficient based on our empirical analysis of the forecasting error behavior, while coordinating our reasoning with the specificities of the processes governing the data viewed from the economic perspective.

First observation that comes into play is that for the forecasting of GNP, the model performance changes significantly with the change in λ only for the cases of longer forecasting horizons (marks on top of the plots show the specific forecasting horizon). Moreover we cannot observe stable optimal value thus presuming better long term forecasts could be made with even higher values of the parameter. Since lower values of the parameter correspond to tighter restriction towards the prior, we conclude that the BVAR benefits a lot in the long term from a greater use of the structural part (the information extracted from the data).

Looking at the inflation we note the same phenomenon of the increase of the forecasting performance for the higher values of λ i.e. lesser shrinkage towards the prior. The effect is even more accentuated and is easily noted also for shorter forecasting horizons, starting to be visually relevant for horizons including and longer than 1 year.

The interest rate forecasting response presents an interesting case as well. We can observe that there is a stable optimal position around the λ value of 0.005 which is to a certain degree independent from the forecasting horizon length (even though it is more pronounced for longer horizons). The fact that optimal shrinkage level is tighter for the short term interest rate coincides with the

reasoning about its fast-to-adjust manner and thus lesser inherent predictability. Therefore random walk forecast should perform relatively better for this variable and respectively more accounted for when shrinking towards it in the BVAR with the given prior.
For the sake of consistency of the analysis we choose one value of λ to be used in further comparison tests against other models. For this sake we computed an average of the optimal values over all the forecasting horizons and variables giving equal weight to each of the entries (which of course could be easily adjusted in light of the forecasting horizon and variable of interest choice). The value for λ we calculate is 0.18375. Next we will look at the comparative performance of the BVAR model.

3.3 Comparison of performance

In the comparison of forecasting performance one should first choose the parameter to be used as a proxy for the ability of the model to fulfil the forecasting objective. In our case we will use the mean squared forecasting error. After having computed the squared forecasting errors for the BVAR and the benchmark models (small specification VAR, ARIMA, random walk), we use the Giacomini and White approach for predictive ability evaluation and perform the test for the difference in MSFE of different forecasting models for given forecasting horizons.
First, a short description of the test is needed to ground the reasons for further inference. The peculiarity of the forecasting test described in Giacomini and White (2006) is in its ability to assess the performance of the forecast in the case of possible misspecification of any kind due to the fact that it does not try to compare the unconditional performance of two models [2] but rather concentrates on the comparison of the estimates provided by given models with the given information set from which parameter estimates are made. Thus both forecasting models and the estimation procedures are taken into consideration in the test. Important to note is that the pure model comparison (testing the hypothetical performance of two models) would fail in the case of the use of shrinkage estimators since this procedure produces biased estimates.
A particular stress is made on the use of rolling window forecasts which simplifies the analysis to bounded data frame forecasts. Using such approach we obtain the $h\,period$ ahead forecasts using the window of size $41 - h$ where h is ranging from $1\ to\ 8$. The reasons for using different window sizes allows us to obtain equal ranges of pseudo out-of-sample forecasts for different forecasting horizons so that we obtain commensurate data for the forecasting ability test. The loss function choice did not present good reasons to be considered in deep and thus the standard Mean Squared Error is used as the test statistic.

[2] Here by the models we mean the dependence structure. Even presuming the knowledge of the true structure, two models are difficult to compare since the "true" coefficients of the models are unknown and are only assessed using a finite sample of data

The null hypothesis, given the MSE loss and information set G_t is

$$H_0 : E[(Y_{t+h} - \hat{f}_{1,t,h})^2 - (Y_{t+h} - \hat{f}_{2,t,h})^2 | G_t] = 0$$

where $\hat{f}_{i,t,h}$ is the $h\,period$ ahead forecast produced at t for model i.
The first test we make is the unconditional test for the difference in means of the squared forecasting error (as assuming an empty information set G_t). Results of this test are presented in Appendix D.
The relative performance against random walk model and VAR model show quite significant improvements in forecasting GNP for forecasting horizons exceeding one year. This confirms the idea that GNP has to exhibit more structural predictability for longer forecasting horizons. However, the performance is not that good with comparison to ARMA model. In our comparison we did not use constant values of the parameters of ARMA model (the lag length and the moving average term) while those were instead automatically determined considering the information criteria for each individual case. Instead the parameters of BVAR (lambda, lag length = 5) were held constant, which could explain the relative performance obtained.
What concerns inflation, we again observe an improvement over the VAR and random walk with the use of Bayesian VAR. The best performance lies in the mid of our horizon span, which should be due to the choice of particular shrinkage parameter and thus could be adjusted as well to perform even better for other forecasting horizons.
The picture of interest rate forecasting is quite mixed. Bayesian VAR showed relatively good performance for the horizons of up to one year in comparison with all the three other models and exhibited much worse performance for the larger horizons. To our view the reason is again the choice of λ since, as we've seen, the influence on interest rate forecasting performance is quite pronounced and consistent (with an optimal value ranging around 0.05 rather then 0.18 as used in our case).
Next issue of our analysis is the test of conditional performance of the models. We perform the test using the wald statistic testing the joint hypothesis of the influence of the test function, where the test function, following Stinchcombe and White (1998) is chosen to be the matrix of the respective model errors spanning over a chosen quantity of prediction errors. In our case we span the vector over 8 periods (prediction errors from t to $t-7$). We thus perform the regression of $(Y_{t+h} - \hat{f}_{1,t,h})^2 - (Y_{t+h} - \hat{f}_{2,t,h})^2$ on the difference of the prediction errors $(Y_{t-i} - \hat{f}_{1,t-i})^2 - (Y_{t-i} - \hat{f}_{2,t-i})^2$ for $i = 0{:}7$ and the intercept.
Next we test the hypothesis of the joint influence of the in-sample prediction performance on the forecasting performance. The results of the test are presented in Appendix E. As we see, it is hard to notice a pattern in the predictability of model performance over difference forecasting horizons. However sometimes relative model performance is found to be correlated with the values of the difference in the predicted squared errors (e.g. inflation for the first 4 quarters ahead in case of comparison of BVAR with RW and ARMA or GNP 7 and 8 period ahead in case of comparison with RW). Thus, according to the results of

this test, sometimes it is possible to expect different relative model forecasting performance given its in-sample performance history. We will now look in one of those cases more closely (one for which null of unpredictability is rejected). We choose the case of comparison of BVAR and RW for 5 period ahead inflation forecasts.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	0.0000	0.0000	0.14	0.887
f.fit1	0.3017	0.2201	1.37	0.182
f.fit2	0.0290	0.2396	0.12	0.905
f.fit3	0.2391	0.1635	1.46	0.155
f.fit4	0.0507	0.0907	0.56	0.581
f.fit5	0.4895	0.8593	0.57	0.574
f.fit6	0.0898	0.2703	0.33	0.742
f.fit7	-0.1781	0.6703	-0.27	0.792
f.fit8	-0.4219	0.6378	-0.66	0.514

Table 3.1: test for conditional predictive ability, BVAR vs RW, 5 period ahead, inflation

As we see, according to the regression results, the worse relative performance of BVAR for last 6 predicted periods will, on average, signify the worse 5 period ahead forecast. However, the opposite could be said about the relation of prediction error 7 and 8 periods earlier than the last period observation. By themselves the coefficients are not seen to be significantly different from zero, which is largely due to the fact of short data span available to be used for assessment. However, this method could represent a useful additional tool for the decision on the use of particular forecasting model.

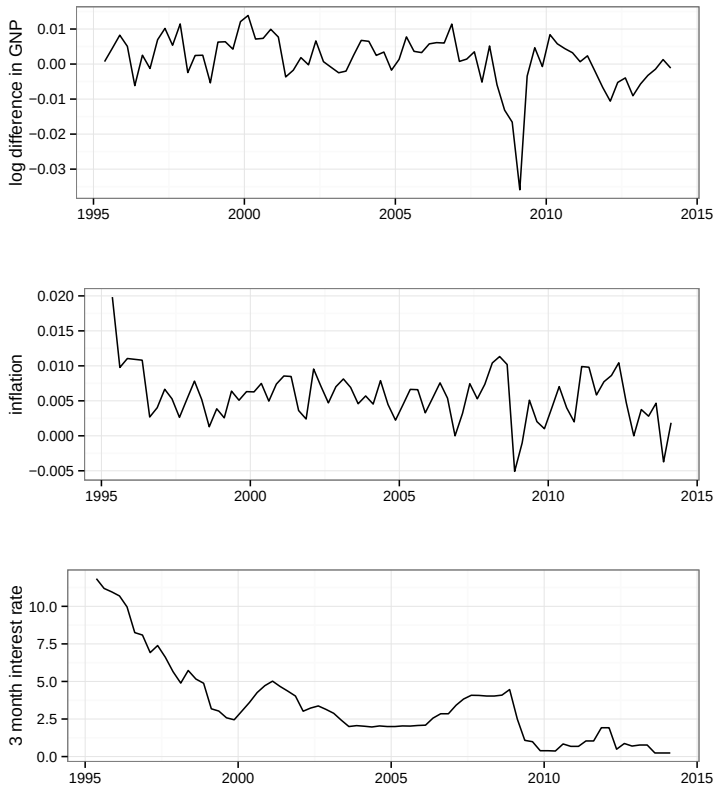
3.4 Conclusions

As we have described before, the use of the shrinkage methods in macroeconomic forecasting represents a very useful tool to handle specificity of the data (short span, overparametrisation of the models). The use of Bayesian VAR, beyond coping with the data scale difficulties, permits the incorporation of economic theory in the modelling of the forecasting problem, which inevitably improves upon usage of purely frequentist methods since the adjustment of BVAR (shrinkage parameter) nests also the case of absolutely diffuse prior (which is a reduction to the frequentist case).
As we can observe from the relative performance analysis, BVAR outperforms small scale VAR by a most of the times for most forecasting horizon settings and as well manages to compete with the random walk model, which proves the potential of this model to be a useful macroeconomic forecasting tool. The comparison with ARMA model does not show relatively better performance, but the fact of usage of single shrinkage parameter for the forecasting of all the variables over all the forecasting horizons should be taken in the account.

As well, the possibility of adding different and more comprehensive data in the BVAR model should be considered in view of its automatic coping with over-parametrization using the prior.
A more elaborate approach, including thorough analysis and specification of the prior imposed on the coefficients, coupled with inclusion of more data in the regression could constitute a significant improvement to the performance of Bayesian VAR described in our simple case. Moreover the relationship between the prior imposition/adjustment and the inherent data-generating processes should be examined and incorporated in modelling framework.
The importance of BVAR as the shrinkage estimation model in general cannot be neglected, giving due to its flexibility and ability to cope with difficulties arousing in the specific environment of the macroeconomic forecasting problem.

Appendix A

Variables of interest



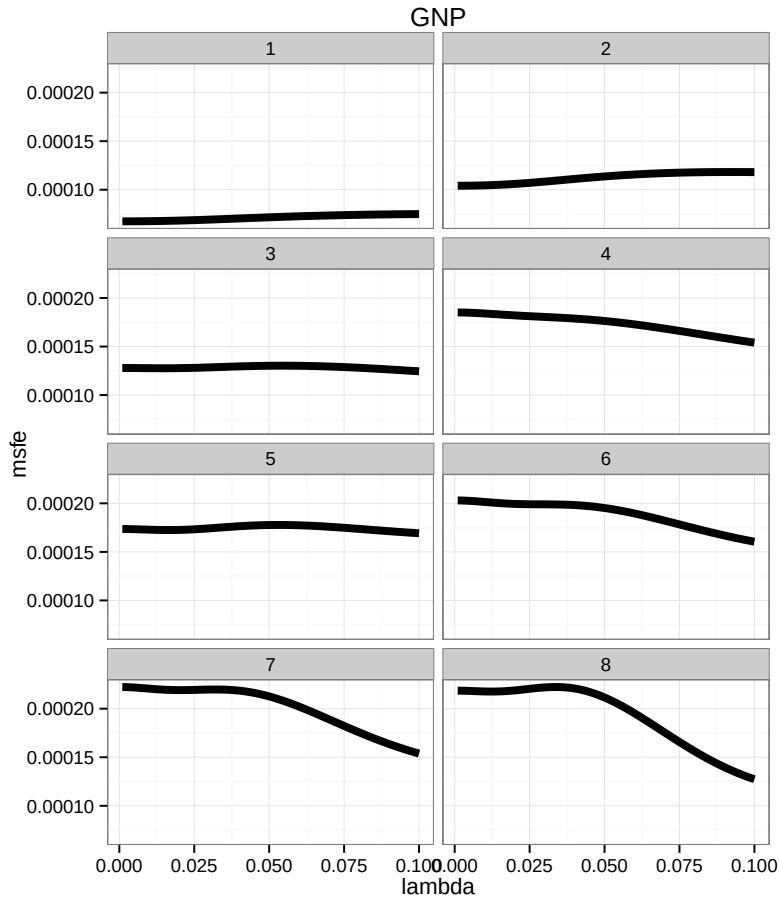
Appendix B

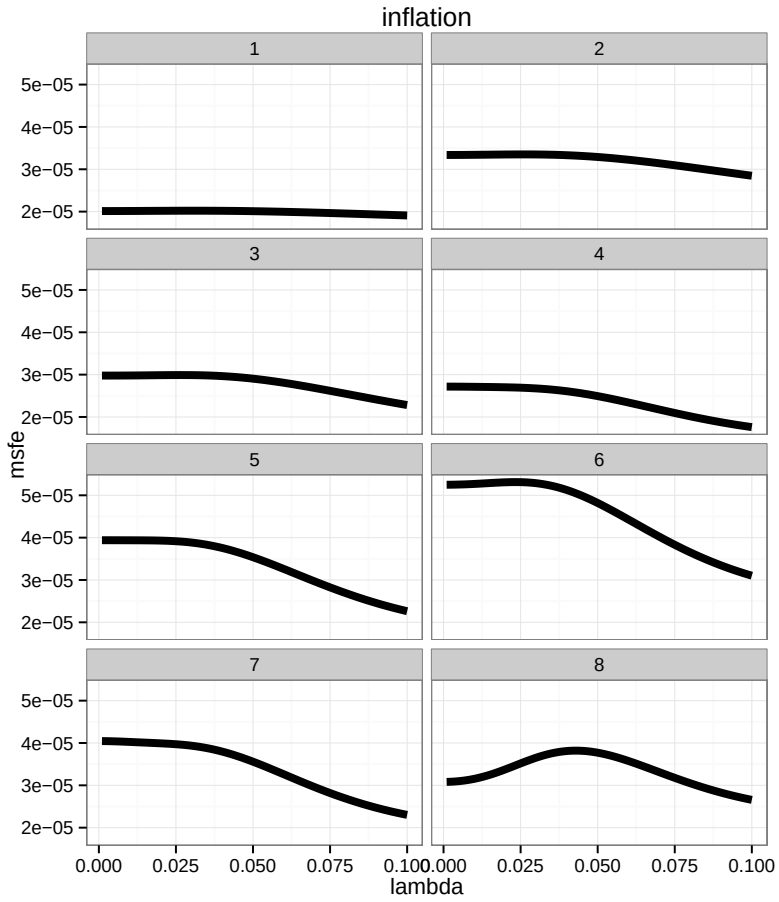
list of variables

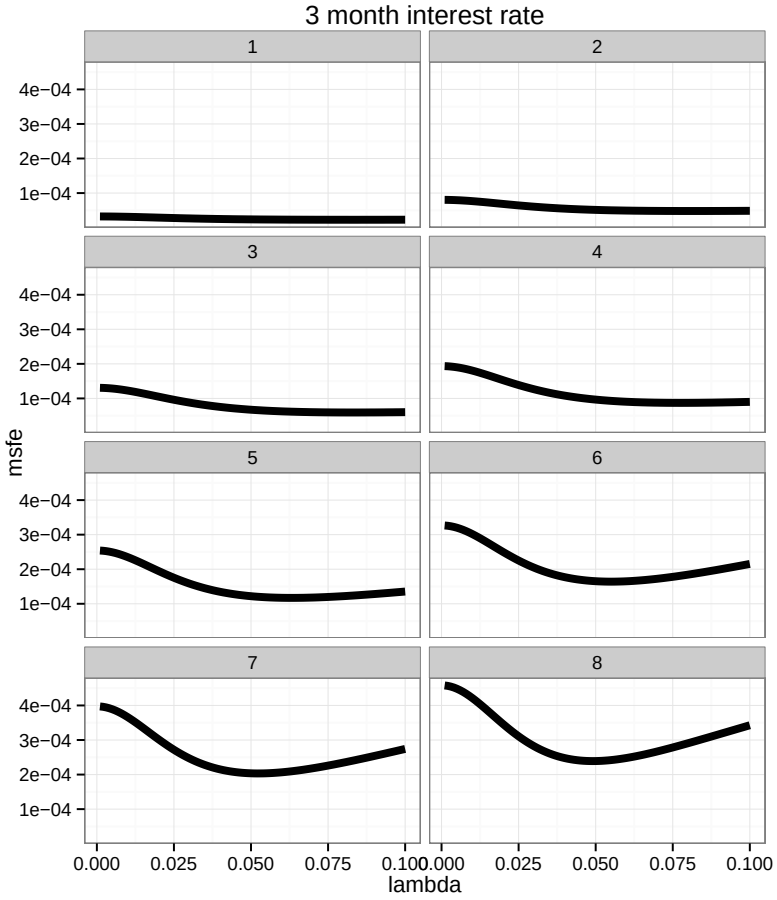
	variable	description
1.	GNPK	Gross National Product
2.	IMPK	Imports of Goods and Services
3.	IMPBEN	Imports of Goods
4.	CFIN	Final National Consumption
5.	CFAM	Consumption of Resident Families
6.	CFTER	Consumption of the Families on the Economic Territory
7.	CODUZ	Private Consumption of Durable Goods
8.	CNDZ	Private Consumption of Non-Durable Goods
9.	COSERZ	Private Consumption of Services
10.	INFLV	Total Fixed Investment
11.	INVMAC	Investment in Machinery and Equipment
12.	INVTRASP	Investment in Means of Transport
13.	INVCOSTR	Investment in Construction
14.	EXPK	Export of Goods and Services
15.	EXPBEN	Export of Goods
16.	IPCOSTR	Industrial Production index in Construction Sector
17.	OCCERV	Employment in Services
18.	CPIIT	Consumer Price Index
19.	X3MBOT	average of 3-month government bond (BOT)

Appendix C

Choice of λ




inflation
1
2
3
4
5
6
7
8
msfe
5e−05
4e−05
3e−05
2e−05
0.000
0.025
0.050
0.075
0.100
lambda


3 month interest rate
1
2
3
4
5
6
7
8
4e−04
3e−04
2e−04
1e−04
msfe
0.000
0.025
0.050
0.075
0.100
lambda

Appendix D

Test of predictive ability results

period ahead	1	2	3	4	5	6	7	8
GNP	0.87	0.90	0.42	0.17	0.81	0.30	0.25	0.19
inflation	0.26	0.04	0.01	0.02	0.00	0.04	0.05	0.51
interest rate	0.25	0.18	0.06	0.09	0.45	0.67	0.71	0.75

Table D.1: p-value, BVAR vs RW

	1	2	3	4	5	6	7	8
GNP	0.81	0.36	0.18	0.12	0.13	0.07	0.04	0.03
inflation	0.29	0.14	0.04	0.07	0.13	0.11	0.06	0.13
interest rate	0.05	0.00	0.02	0.19	0.30	0.82	0.80	0.78

Table D.2: p-value, BVAR vs VAR

	1	2	3	4	5	6	7	8
GNP	0.60	0.29	0.21	0.35	0.95	0.71	0.63	0.97
inflation	0.84	1.00	0.63	0.43	0.70	0.87	0.74	0.83
interest rate	0.28	0.20	0.05	0.10	0.53	0.71	0.70	0.72

Table D.3: p-value, BVAR vs ARMA

Appendix E

Conditional test results

period ahead	1	2	3	4	5	6	7	8
GNP	1.00	1.00	1.00	0.97	1.00	0.91	0.00	0.00
inflation	0.01	0.01	0.00	0.10	0.00	0.85	0.83	1.00
interest rate	0.98	0.99	0.94	0.04	0.00	0.95	0.07	0.13

Table E.1: p-value, joint test of the coefficients for prediction error difference, BVAR vs RW

period ahead	1	2	3	4	5	6	7	8
GNP	0.10	0.03	0.94	0.93	0.94	0.98	0.97	0.97
inflation	0.90	0.96	0.91	0.79	0.92	0.97	0.97	1.00
interest rate	0.98	0.01	0.75	0.98	0.99	0.99	0.54	0.97

Table E.2: p-value, joint test of the coefficients for prediction error difference, BVAR vs VAR

period ahead	1	2	3	4	5	6	7	8
GNP	1.00	1.00	0.28	0.99	0.95	1.00	0.89	0.00
inflation	0.00	0.35	0.05	0.01	1.00	1.00	0.00	0.98
interest rate	0.83	0.97	0.92	0.84	0.15	0.98	0.50	0.59

Table E.3: p-value, joint test of the coefficients for prediction error difference, BVAR vs ARMA

Bibliography

[1] Arthur E. Hoerl and Robert W. Kennard *Ridge Regression: Biased Estimation for Nonorthogonal Problems* 1970: Technometrics, Vol 12, 55-67.

[2] Arthur E. Hoerl and Robert W. Kennard *Ridge Regression: Applications to Nonorthogonal Problems* 1970: Technometrics, Vol 12, 69-82.

[3] Donald W. Marquardt *Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation* 1970:Tehnometrics, Vol 12, 591-612.

[4] Robert Tibshirani *Regression Shrinkage and Selection via the Lasso* 1996: Journal of Royal Statistical Society, 267-288.

[5] Rafaella Giacomini and Halbert White *Tests of Conditional Predictive Ability* 2006: Econometrica, Vol 74, No 6, 1545 - 1578.

[6] James H. Stock and Mark W. Watson *Dynamic Factor Models* 2010: Oxford Handbook of Economic Forecasting.

[7] Robert B. Litterman *Techniques of Forecasting Using Vector Autoregressions* 1979: Working paper no. 115, Federal Reserve Bank of Minneapolis.

[8] Maxwell B. Stinchombe and Halbert White *Consistent Specification Testing with Nuisance Parameters Present Only Under the Alternative* 1998: Econometric Theory, 295 - 325.

[9] Boivin, J. and M. Giannoni DSGE Models in a Data-Rich Environment 2006: NBER Working Paper no. 12772.

[10] Boivin, J. and S. Ng Understanding and Comparing Factor-Based Forecasts 2005: International Journal of Central Banking 1, 117-151.

[11] Bernanke, B.S., J. Boivin, and P. Eliasz Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach 2005: Quarterly Journal of Economics, 120, 387-422.

[12] Thomas Doan, Robert Litterman and Christopher R. Sims Forecasting and Conditional Projection Using Realistic Prior Distributions 1983: NBER Working Paper no. 1202.

[13] Francis X. Diebold The Past, Present and Future of Macroeconomic Forecasting 1997: NBER Working Paper no. 6290.

[14] M.J. Bayarri and T.O. Berger The Interplay of Bayesian and Frequentist Analysis 2004: Statistical Science, Vol. 19, No. 1 (Feb., 2004), pp. 58-80.

[15] K. Rao Kadiyala and Sune Karlsson Numerical Methods For Estimation and Inference in Bayesian VAR-models 1997: Journal of Applied Econometrics, vol. 12, 99-132.

[16] Robert B. Litterman Forecasting with Bayesian Vector Autoregressions: Five Years of Experience 1986: Journal of Business and Economic Statistics, Vol. 4, No. 1 (Jan., 1986), pp. 25-38.

[17] Massimiliano Marcellino, James H. Stock and Mark W. Watson Macroeconomic forecasting in the Euro area: Country specic versus area-wide information 2003: European Economic Review 47 (2003) 1 18.