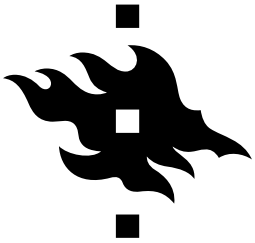

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Conditional BVARX Forecasting Model for Small Open Economies: An Application to Finland

Jetro Anttonen

University of Helsinki
Faculty of Social Sciences
Economics

Master's Thesis

April 2019

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Faculty	Department
Faculty of Social Sciences	Department of Political and Economic Studies

Author
Jetro Anttonen

Title
Conditional BVARX Forecasting Model for Small Open Economies: An Application to Finland

Subject
Economics

Level	Month and year	Number of pages
Master's Thesis	April 2019	71

Abstract

In this thesis, a conditional BVARX forecasting model for short and medium term economic forecasting is developed. The model is especially designed for small-open economies and its performance on forecasting several Finnish economic variables is assessed. Particular attention is directed to the hyperparameter choice of the model. A novel algorithm for hyperparameter choice is proposed and it is shown to outperform the marginal likelihood based approach often encountered in the literature. Other prominent features of the model include conditioning on predictive densities and exogeneity of the global economic variables. The model is shown to outperform univariate benchmark models in terms of forecasting accuracy for forecasting horizons up to eight quarters ahead.

Keywords
forecasting, conditional forecasting, Bayesian VAR, marginal likelihood, hyperparameter choice

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta Valtiotieteellinen tiedekunta	Laitos Politiikan ja talouden tutkimuksen laitos	
Tekijä Jetro Anttonen		
Työn nimi Conditional BVARX Forecasting Model for Small Open Economies: An Application to Finland		
Oppiaine Taloustiede		
Työn laji Maisterin tutkielma	Aika Huhtikuu 2019	Sivumäärä 71

Tiivistelmä

Tässä tutkielmassa kehitetään BVARX-ennustemalli lyhyen ja keskipitkän aikavälin taloudelliseen ennustamiseen. Malli on erityisesti kehitetty pienten avotalouksien erityispiirteet huomioon ottaen ja tutkielmassa testataan mallin kykyä ennustaa useita suomalaisia taloudellisia muuttujia. Erityistä huomiota tutkielmassa kiinnitetään mallin hyperparametrien valintaan. Tutkielmassa esitellään uusi menetelmä hyperparametrien valitsemiseksi ja uuden menetelmän näytetään tuottavan tarkempia ennusteita kuin kirjallisuudessa usein käytetty marginal likelihood funktioon perustuva lähestymistapa. Muita mallin erityispiirteitä ovat mahdollisuus ehdollistaa ennusteita mallin muuttujien tulevilla arvoilla tai tiheyksillä ja globaalien taloudellisten muuttujien eksogeenisuus. Tutkielmassa kehitetyn ennustemallin näytetään tuottavan tarkempia ennusteita kuin vertailukohtana käytettävät yhden muuttujan menetelmät kaikilla tarkastelluilla ennustehorisonteilla yhdestä kahdeksaan neljännestä tulevaisuuteen.

Avainsanat
ennustaminen, ehdollinen ennustaminen, bayesiläinen VAR, marginal likelihood, hyperparametrien valinta

Contents

1 Introduction

It is common for central banks and other economic institutes to regularly publish medium term forecasts concerning the state of the economy within the next few years. These forecasts are usually based on both subjective considerations and econometric models. Nowadays arguably the most prominent tools for econometric forecasting are DSGE- and BVAR-models and they both have their strengths and weaknesses. With the abbreviation DSGE is referred to *dynamic stochastic general equilibrium* models, which consist of multiple behavioral equations, one per endogenous variable. BVAR stands for *Bayesian vector autoregressive* model and BVARX simply refers to a BVAR-model with suitable lag restrictions on the exogenous variables of the model. What motivates the use of a BVARX-model is that when modeling small open economies, like Finland, it is very convenient to treat global economic variables as exogenous, as it is very plausible to assume the small open economy not to have any significant effect on the global economic variables.

However, credible modelling of the whole global economy is not often feasible and thus a model with a few global aggregate variables, measuring factors such as the global economic activity and interest rates, provides little to no informational value on future paths of those variables, expect for their own lags of course. This is undesirable, since knowledge of the future values on exogenous global variables could be used to obtain more accurate forecasts on the endogenous variables as well.

Usually forecasts on global economic variables can be obtained by for example aggregating country specific projections provided by national statistical agencies and central banks. These projections usually contain more accurate information on future paths of the global variables, as it is probable that they are based on more extensive amounts of information than what is available to the forecaster. Therefore, conditioning forecasts of endogenous variables on projections of exogenous global variables can be expected to yield improvements in forecasting accuracy of the model. For example, Bloor & Matheson (2011) successfully condition their forecasts of the GDP, inflation and interest rates in New Zealand on point estimates of global variables. However, by focusing on variables such as import and export flows which are

unarguably highly dependent on the global economy, even greater improvements in terms of forecasting accuracy of the model could be expected. On top of that, instead of using only point estimates for conditioning, using full predictive densities would yield forecasts with full predictive densities *and* reliable predictive intervals.

The conditioning of forecasts with future values or predictive densities of other variables requires no structural assumptions regarding the relationship of the variables. However, it must be borne in mind that conditioning on future values requires very different interpretation from conditioning on structural shocks. Forecast conditioned on future values does not have the interpretation of structural analysis useful for policy evaluation. Conditioning on future values merely provides the most likely forecast on all of the other variables, given the information on the other. No causal conclusions based on these conditional forecasts can be drawn.

The conditioning procedure could also be taken a step further and predictive densities of other endogenous variables could be used in addition to exogenous ones. For example, in some cases forecaster might have a reason to believe that another model would provide more accurate predictions of some other endogenous variable. As an example, medium term projections of inflation figures produced by DSGE-models are often found to yield superior performance to the ones produced by VAR-models (see e.g. Smets & Wouters 2007, Burlon et al. 2015, Wang 2009). In this case, the forecasts of the BVARX-model could very well be conditioned on the inflation projections of the DSGE-model.

The model developed in this thesis allows for conditioning on full predictive densities as discussed above, and the possibilities and technical details of conditioning will be thoroughly discussed. However, in the empirical section of this thesis, the conditioning on future values of exogenous variables was not found to yield significant improvements to the forecasting accuracy of the model and the possible limitations of the approach are also discussed.

In addition to the conditioning discussed above, another topic thoroughly assessed in this thesis is how the hyperparameter choice should be carried out in the context of a Bayesian vector autoregressive model. Hyperparameter choice is arguably the most important single feature affecting the forecast-

ing accuracy of a Bayesian vector autoregressive model and the choice has been approached in the literature in various different ways. Giannone et al. (2015) argue in favor of an approach based on maximization of the marginal likelihood function of the model. In Giannone et al. (2015) they find their proposed approach outperforming the other approaches for hyperparameter choice often encountered in the literature.

However, in this thesis, several deficiencies of the marginal likelihood based methods for hyperparameter choice (in short: ML approach) are presented. Especially, with the very limited nature of macroeconomic time series data the ML approach seems to yield hyperparameter values that cause the model to over-fit the data. With a Monte Carlo simulation study, it can be illustrated how on average the prior suggested by the ML approach tends to become *tighter* as more data becomes available, and hence the *correct* hyperparameter values are only obtained with a large amount data. This feature could be avoided with a hyperprior distribution pulling the hyperparameter estimates themselves towards a cautious zero when the data is scarce. However, Giannone et al. (2015) completely disregard this feature and center their hyperprior distribution around the rule-of-thumb-values originally proposed by Sims & Zha (1998).

A suitable hyperprior distribution discussed could however be difficult to find and therefore an alternative approach is proposed. A novel algorithm presented in this thesis exploits the fact that the mean of the predictive distribution of any typical Bayesian vector autoregressive model has a closed form representation. Therefore, by treating the mean of the predictive distribution as the point-estimate of the forecast and by setting up a pseudo out-of-sample forecasting exercise, it becomes feasible to use numerical optimization methods to search for the vector of hyperaparameters that minimizes any function of *all* the past out-of-sample forecasting errors. The novel algorithm described is found to provide more accurate forecasts with the model developed, than either the ML approach or the rule-of-thumb values of Sims & Zha (1998).

Main contribution of this thesis is to develop a non-structural, conditional, flexible and efficient forecasting model for medium term forecasting of small-open economies. Many of the modeling choices discussed and presented in

this thesis are easily generalizable for different Bayesian vector autoregressive models and work as a solution for several issues of practical importance. These issues include (i) the incorporation of data of different frequencies, (ii) the imposition of exogeneity on global economic variables, (iii) dealing with the so-called *ragged edge* of the data, (iv) conditioning of the forecasts on predictive densities and (v) the proper hyperparameter choice.

The next section provides a comprehensive literature review, where the economic forecasting with Bayesian vector autoregressive models is discussed, with particular attention on conditional forecasting and hyperparameter choice. The model and all its technical details, excluding the hyperparameter related ones, are presented in the third section. The data used for the study is elaborated in the fourth section and in the fifth section the issues related to hyperparameter choice are discussed in depth. The novel algorithm for hyperparameter choice is also presented in the fifth section and an empirical assessment is performed to compare the performance of differently chosen sets of hyperparameters with models of different size. In the sixth section, the forecasting accuracy of the model is assessed with respect to gross domestic product, exports of goods and services, imports of goods and services, inflation and unemployment. The conditioning of the forecasts on future values of exogenous variables is also studied and discussed in the sixth section. In the seventh section the issues addressed and the results obtained are discussed and finally the section eight concludes.

2 Literature review

2.1 Economic forecasting

The medium term economic forecasts produced by central banks and research institutions are often based on structural models. For example in Finland, the Bank of Finland bases it's medium term projections on the large DSGE-model documented in Kilponen et al. (2016) and the Research Institute of the Finnish economy (Etla) bases it's forecasts to a so-called *SEM-model* documented in Lehmus (2018). Also, the Finnish ministry of fi-

nance uses a DSGE-model as a basis of it's forecasts.[1] Usually the published projections are not however based only on these structural models, but are ensembles of structural models, statistical models and subjective considerations. Especially BVAR-models designed for nowcasting purposes can be used as complementary tools in addition to the structural models (see e.g. Itkonen & Juvonen 2017, Bok et al. 2017).

The popularity of DSGE-models is not a surprise given their interpretability and reasonable forecasting accuracy. However, large structural models often incorporate a great amount of strong assumptions and restrictions and can transpire to be an unfeasible option for many forecasters due to their complex nature. Not surprisingly, the developments in more easily implemented BVAR-models (e.g. Banbura et al. 2010, Giannone et al. 2015, Koop 2013) have stemmed a growing interest in statistical methods capable of comparable performance to DSGE-models in terms of forecasting accuracy.

The grown interest in BVAR-modeling accelerated after Banbura et al. (2010) showed that BVAR-models are capable of handling a very large number of variables (131 variables in the paper in question) and that the large BVAR-models could be superior to other statistical methods in terms of forecasting accuracy, at least in forecasting of inflation and unemployment. Koop (2013) compared different prior distributions in the context of BVAR-models with as much as 161 variables. The results further supported the superiority of large BVAR-models compared to other state-of-the-art statistical forecasting methods, such as *dynamic factor models*.

The prevailing view even before the latest developments in BVAR-modeling was that the statistical models often produce more accurate forecasts in short horizons while structural models should be used in longer horizons (see e.g. Wang 2009). The above mentioned studies have however shown that correctly specified BVAR-models are capable of producing forecasts comparable to those of DSGE-models in terms of accuracy for horizons relevant to medium-term forecasting as well. On top of comparable or even superior forecasting accuracy, the BVAR-models have other advantages over popular DSGE-models and subjective considerations as well.

As already mentioned BVAR-models are easier to implement and they do

[1] https://vm.fi/en/economic-forecasts

not require as strong assumptions or restrictions as DSGE-models or other structural models. The problem of forecasts based on subjective considerations is that in general reliable confidence levels of the forecasts can not be produced. Also, the performance of these forecasts can not be assessed via *backtesting* with pseudo out-of-sample forecasting exercises. Also, at least in some cases DSGE-models have even been reported to overestimate the uncertainty around the point estimates (Wolters 2015). The results of structural models are however easier to interpret and their usefulness for policy analysis is unarguable. From a pure forecasting perspective the BVAR-models do however make a strong case for themselves.

No matter which modeling approach is used for forecasting purposes, common issues arise due to global economic fluctuations being an evident factor affecting the development of economic variables in any country involved in the interconnected global markets.

First, rigorous modeling of the global economy is rarely a feasible option to the forecaster. To avoid this issue, exogenous global variables are often added to the model. An example of a more ambiguous approach closely connected to the BVAR-models is the B-GVAR approach (see e.g. Cuaresma et al. 2016, Dovern et al. 2016). Especially with structural models the issue is sometimes addressed by combining several country specific forecasting models to create a massive model for the global economy. For example, the structural forecasting and policy evaluation model (Lehmus 2018) used in the Research Institute of the Finnish Economy is used to update the Finnish module in the *NiGEM*[2] global macroeconomic model. Evidently, a massive structural econometric model consisting of tens or even hundreds of national modules can not always be considered a feasible option. In the context of the BVAR-models, the issue can be addressed by conditioning the forecasts on the most credible projections provided by any institutions concerning the global economy. This is the approach studied in this thesis and earlier applied in Bloor & Matheson (2011).

Second issue concerns the exogeneity of the global economic variables and is especially important when modeling small open-economies, such as Finland. It is easily taken into account, but in the context of VAR- or

[2]For *NiGEM* global econometric macro model, see: https://nimodel.niesr.ac.uk/

BVAR-modeling often ignored for convenience. The issue of exogeneity in VAR-modeling, and the ignorance to it, was first brought up by Zha (1999). The estimation procedure proposed by Zha (1999) concerned strong recursive blocks, which can be seen as a generalization of the so-called BVARX approach used in this thesis and in Bloor & Matheson (2011). The importance of proper incorporation of exogeneity for policy analysis with BVAR-models was illustrated by Zha (1999). It however remains unclear, does the proper incorporation of exogeneity lead to significant improvements in forecasting accuracy of the model. The promising results from the empirical assessment in this thesis suggest so, although this specific issue is not explicitly studied in this thesis.

Despite the advancements in statistical modeling there is not, and arguably will never be, a single model that would outperform other models in every metric imaginable. It has been shown repeatedly that model combinations are likely to outperform any single model (see e.g. Fawcett et al. 2014, Amisano & Geweke 2013, Koop & Koroblis 2012, Timmermann 2006, Yu et al. 2005, Zhang 2003, Bates & Granger 1969). Combination of models is often referred to as *ensemble modeling*. The conditioning of forecasts with (sometimes combined) projections from other models, can as well be viewed as an unorthodox way of ensemble modeling. Essentially, the ensemble modeling in any form allows for incorporating more information more efficiently into the forecasts.

2.2 Conditional forecasting

Here conditional forecast is defined to be a forecast that is conditioned on information from outside the data used for estimation of the forecasting model, excluding prior beliefs. Since statistical models often require the data to be balanced (i.e. to not have any missing observations), conditional forecasts are often forecasts conditioned on the observations from the periods with observations of some other variables missing. Especially in *nowcasting* applications this is essential, since it allows for exploiting all the latest data available. The same idea can be generalized to a case where the observations to be conditioned on are yet to happen, hence there is uncertainty revolv-

ing the information to be conditioned on. Often the term *scenario analysis* is used somewhat substitutively with *conditional forecasting*, but with an emphasis on different scenarios to be conditioned on.

When conditioning on the values of the upcoming observations, a clear distinction between the forecasts conditioned on structural shocks and on reduced form shocks must be made. This is an issue easily forgotten, since in both cases a forecast conditioned on the deceptively alike information can be produced. The interpretation between the two is however completely different.

First, to condition on structural shocks the shocks must be identified, hence a model of at least some structure is required. A bulk of macro econometrics focuses on the identification of economic shocks to produce credible impulse response functions (see e.g. Kilian & Lütkepohl 2017). An impulse response function tells what is the effect of an *exogenous* shock on other variables. Conditional forecasts based on restrictions on structural shocks can thus for example be interpreted as *what would happen to other variables if we were to force one variable to follow a certain path?* This kind of scenario analysis is representative of policy analysis, where causal relations and effects of government policies are of interest. This interpretation must however not be mistaken for the correct one, when dealing with the conditioning of reduced form shocks studied in this thesis.

Conditioning on reduced form shocks requires no identification of structural shocks. The interpretation is however different, and arguably better suited for forecasting as one does not have to impose any assumptions regarding the causes of the shocks to be implicitly conditioned on. As an example, if one were to condition the forecast on very positive believes regarding the future growth rate of the gross domestic product (GDP), the forecast produced would not represent the effects of an economic boom to the economy, but rather illustrate the most likely values of other variables given the economic boom were to happen. This makes conditional forecasting a particularly effective model combination tool, at least in theory, since by conditioning on one variable, some information on all of the other variables from the other model can be incorporated to the forecast.

The one commonly used tool in economics for conditional forecasting

is the *Kalman filter* originally proposed in Kalman (1960) and the various algorithms based on it (see e.g. Banbura et al. 2015, Schorfheide & Song 2015, McCracken et al. 2015). The Kalman filter has a long tradition in economics and it also has many practical properties beyond its applicability to forecasting (see e.g. Athans 1974). In many nowcasting applications (e.g. Schorfheide & Song 2015, McCracken et al. 2015) the Kalman filtering has been used to exploit the information on observations from the latest periods, where the data is still incomplete (i.e. not balanced).

However, as noted in Banbura et al. (2015), the conditional forecasting in more general form, has been left with a very limited amount of attention after the development of the most commonly used method in Waggoner & Zha (1999), apart from a few exceptions (e.g. Bloor & Matheson 2011). Banbura et al. (2015) argue that this is due to unfeasibly computationally demanding nature of the method. The successful implementation of the method in Bloor & Matheson (2011) to a BVAR-model with as much as 41 endogenous variables however suggests otherwise.

The conditioning method discussed above was originally proposed in its rawest form by Doan et al. (1984). The approach had however its drawbacks since it did not minimize the mean squared errors of the forecasts conditional on the restrictions and the parameter estimates were not consistent with the path to be conditioned on (see Karlsson 2012). These issues were addressed in Waggoner & Zha (1999) as they proposed a *Gibbs sampler* to draw the reduced form errors that the conditional forecasts would be constructed of. Still, the greatest deficiency of the method remained. Incorporation of uncertainty in the restrictions to be conditioned on was not possible, leading to too narrow and unreliable predictive densities. This can be argued to be the main reason for the lack of attention regarding conditional forecasting, as opposed to what was implied in Banbura et al. (2015).

The issue of unreliably narrow predictive densities was addressed in Robertson et al. (2005) by a computationally very demanding method including exponential tilting and moment conditions. The same method was also applied in Cogley et al. (2005) and as an example of later developments in conditional forecasting, Jarocinski (2010) proposed minor modifications to the algorithm of Waggoner & Zha (1999) for improved computational efficiency.

In this thesis, the revision of the method in Waggoner & Zha (1999), proposed by Andersson et al. (2010), is used. The algorithm proposed in Andersson et al. (2010) is otherwise analogous to the original one in Waggoner & Zha (1999), but it allows for conditioning on predictive densities of future values rather than only point estimates, thus providing full reliable predictive densities itself. The original method can be acquired as a special case of the revised method by setting the variance or standard deviation, of the predictive density to be conditioned on, to zero. This possibility for uncertainty in the restrictions comes with minimal computational costs, as opposed to to the method of exponential tilting in Robertson et al. (2005).

The Kalman filter based method in Banbura et al. (2015) is unarguably computationally more efficient than the method of Andersson et al. (2010). However, the latter mentioned method has many desirable properties. Most importantly, although imposed in this thesis, in general the method does not require normality in the predictive densities to be conditioned on, as Kalman filter based algorithms do. On top of that, the methodology of Andersson et al. (2010) allows for more diverse linear restrictions and is, arguably, more intuitive and easier to implement.

Due to these desirable properties, the algorithm from Andersson et al. (2010) is used to produce the conditional forecasts in this thesis. Also, in practice the computational efficiency of the method becomes an issue only in the most unconventional applications such as the one in the appendix of Banbura et al. (2015) containing 24 variables to be conditioned on and a forecasting horizon of 60 periods. In author's opinion, in conditional economic forecasting, the method of Andersson et al. (2010) should be studied further and applied when applicable, in addition to the algorithm proposed by Banbura et al. (2015).

2.3 Bayesian vector autoregressive models

Until the 1980s, economic time series analysis, and hence forecasting, was performed using a variety of techniques such as structural multiple equation models and statistical univariate time series models. These techniques appeared however to be insufficient for economic analysis as none of them

seemed to be able to produce trustworthy results in the economic turbulence of the 1970s. (see e.g. Stock & Watson 2001). In a seminal paper, Sims (1980) criticized the *"incredible"* identifying assumptions of the structural models of the day and argued in favor of a more statistical approach with minimum assumptions required. Sims (1980) then introduced the vector autoregressive models (VARs) into economics, which have since then proved to be very successful especially in economic forecasting applications (e.g. Stock & Watson 2001).

However, the number of parameters in a VAR-model grows quickly as more variables and lags are added to the model. As in time series applications there is always a very limited amount of data available, this increases the estimation uncertainty, and hence greatly reduces the capabilities of the model to efficiently incorporate more than a few variables. One particular solution to this problem was proposed by Litterman (1979, 1980). Litterman introduced a so-called *Litterman's prior* and brought Bayesian vector autoregressive models (BVARs) into economics. Convinced by the US macroeconomic data, he believed that most of the economic variables closely resemble a univariate random walk process and that the estimation uncertainty could be reduced by shrinking the parameter estimates towards this belief.

The Litterman's prior included other *soft* restrictions as well, such as shrinking the coefficients of the further lags more towards zero, reflecting the belief that more distant observations are less important. After five years of true out-of-sample forecasts Litterman (1986) provided promising evidence of the forecasting performance of his model, comparing the forecasting errors of his models with those from commercial forecasts of the day. The forecasts produced by the BVAR-model were superior for the real variables such as gross domestic product, but could not provide the same accuracy as structural models for some variables, such as inflation.

After Litterman (1979, 1980) had introduced the BVAR-models into economics, several modifications and improvements were made to the original model. Additional priors to be included by adding *pseudo-observations* to the data were proposed by Doan et al. (1984) and Sims (1993). The *sum-of-coefficients* prior introduced by Doan et al. (1984) allows the data to a priori follow a higher order unit root process, than of order one implied by the orig-

inal Litterman's prior. The other additional prior proposed by Sims (1993) is called a *dummy-initial-observations* prior and it introduced the possibility of prior correlation among the parameters of the same equation. Both of these additional priors were proven to enhance the forecasting performance of the BVAR-models and have since become a regular addition to BVAR-models.

Further advancements in BVAR-modeling were provided in the 1990s by Kadiyala & Karlsson (1993, 1997) and Sims & Zha (1998). Kadiyala & Karlsson (1993, 1997) relaxed the assumption of a diagonal covariance matrix and provided empirical evidence in favor of the priors allowing for dependencies between equations, such as the natural conjugate *Normal-Wishart* prior used in this thesis as well. Sims & Zha (1998) further improved the methodology of Kadiyala & Karlsson (1997) by providing new insights to the estimation of the error bands and showed that exploitation of the Kronecker product structure of the covariance matrix allows for computationally efficient estimation of larger systems than what had been possible at the time.

The development of BVAR-models took another big step forward when Banbura et al. (2010) showed that with a correctly specified Normal-Wishart prior it is possible, and efficient, to include even hundreds of variables into a BVAR-model. Along with the improved capability of incorporating more information into the model, the BVAR-models were shown to produce superior forecasts compared with the other state-of-the-art statistical methods, such as the *dynamic factor models* (see e.g. Banbura et al. 2010, Koop 2013).

In addition to the *Litterman-based* priors, such as the one in Banbura et al. (2010), there have been other successful approaches as well for imposing different prior believes through different hyperparameterization of the model. As an example Villani (2009) proposed a so-called steady state prior, which has proven to be especially efficient for example in the context of seasonal BVAR-models (Stelmasiak & Szafranski 2016). However, the efficient estimation of a large number of variables requires for exploitation of the Kronecker product structure of the covariance matrix made possible for example by the natural conjugate Normal-Wishart prior. Thus, not very different priors from Normal-Wishart priors have been successfully applied to large systems.

The Normal-Wishart prior used in this thesis allows for computationally

efficient direct sampling from the posterior distribution of the coefficients, which is extremely convenient in the context of already computationally moderately burdensome conditional forecasts discussed earlier. No other prior distributions are thus considered in this thesis.

2.3.1 Hyperparameter choice

Hyperparameter choice can often be described as specifying the strength of the prior believes. When properly executed, no other choices of importance regarding the specification of the BVAR-model with a given prior distribution need to be done. Thus, well executed hyperparameter choice has the potential to greatly increase the performance of the model, whereas careless assessment of the issue can lead to a significant deterioration in performance. Essentially, with optimal hyperparameter values the maximum amount of information is extracted from the data, while still giving enough importance for the prior to avoid *over-fitting* of the model.

In the literature, various approaches have been proposed for choosing the optimal hyperparameters. Sims & Zha (1998) provide the so called *rule-of-thumb* values often used in the literature. Fixed hyperparameter values cannot however take into account the growing number of the variables in the larger models. Other popular and intuitively appealing technique is to compute the out-of-sample forecasting errors of the model for some pre-specified time interval over some set of hyperparameter alternatives and then choose the set of hyperparameters that minimizes some function of the forecasting errors (e.g. Litterman 1986). The third popular technique for hyperparameter choice is the one from Banbura et al. (2010). They first estimate the model with *ordinary least squares* (OLS) and only three variables, to make sure the model does not suffer from overfitting, and compute the in-sample forecasting errors. Then they define the optimal hyperparameters for the larger models to be the ones that produce in-sample forecasting errors of the same size as the ones with only three variables and no shrinkage.

As noted in Anttonen (2018) all those approaches however lack a solid theoretical foundation and none of them have been proven to consistently outperform the rule-of-thumb values provided by Sims & Zha (1998). To ac-

count for the issue with a theoretically sound and well performing solution, Giannone et al. (2015) emphasize the fact that "the distinction between parameters and hyperparameters is mostly fictious and made only for convenience". The hyperparameter choice could therefore be approached similarly to the estimation of the other parameters in the model. Imposing then a hierarchical structure on the model and defining an informative or uninformative prior to the hyperparamters themselves then allows one to exploit the marginal likelihood function to choose, in a sense, the optimal hyperparameters, that account for the addition of variables into the model. Giannone et al. (2015) proceed to show in their study that the hyperparameters chosen this way, should in theory minimize the one-step-ahead out-of-sample forecasting error of the model. However, the empirical results of this thesis suggest that in practice this might not be the case. The issue is revisited in the fifth section.

The method of Giannone et al. (2015) however requires for *metropolis-algorithm* to account for hyperparameter uncertainty, which increases the computational complexity of the estimation procedure. Therefore, as the method is applied to the model in this thesis, a slight modification of it is used to preserve the computationally convenient features of the model allowing for direct sampling from the posterior distribution of the coefficients, as in Anttonen (2018). Most importantly, only the numerical mode of the marginal posterior distribution of the hyperparameters is used to set the hyperparameter values.

Although the method of Giannone et al. (2015) is theoretically well founded and the authors provide convincing empirical evidence in favor of the method, it still has it's drawbacks and most importantly it is not immune to over-fitting as opposed to what is implied by the authors. As mentioned above, the shortcomings of the method are more elaborately discussed in the fifth section.

As an alternative method for hyperparameter choice, a computationally feasible algorithm minimizing any function of the past out-of-sample forecasting errors with respect to a vector of hyperparameters is proposed in this thesis. The performance of the newly proposed method is compared empirically to that of the marginal likelihood based approach of Giannone et al.

(2015) and to rule-of-thumb values proposed by Sims & Zha (1998) in the fifth section. The results of this assessment provide empirical evidence in favor of the newly proposed algorithm, although it must be acknowledged that the time-interval in which the comparison is performed is fairly short.

3 Conditional BVARX-model

The model developed in this thesis is fairly technical and various econometric methods are involved. Essentially, the statistical model developed is a Bayesian vector autoregressive model (BVAR) with a few additional bells and whistles adding to the model complexity.[3]

In this section, the required econometric methods are described. First, the underlying BVAR-model and the structure of the prior are briefly described. Next, a way to impose exogeneity on certain variables is illustrated and the conditioning procedure is discussed in depth. Finally, these methods are put together and the complete algorithm for estimation of the model is presented. The hyperparamter choice can also be considered a part of the estimation procedure but it is covered in depth later in the fifth section.

3.1 BVAR-model

BVAR-model is a Bayesian version of a celebrated vector autoregressive model (VAR) widely used in the economics. In a VAR model, a variable is modeled as a linear function of the past p values of itself but also of the other variables in the model plus the (usually) normally distributed error term $\boldsymbol{\varepsilon}_t$. The covariance matrix of the error term, $\boldsymbol{\Sigma}$, is in general not diagonal and thus the values of different variables from the same period can also depend on each other. The VAR-model can be represented as follows:

$$\boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{A}_1\boldsymbol{y}_{t-1} + ... + \boldsymbol{A}_p\boldsymbol{y}_{t-p} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim N(0, \boldsymbol{\Sigma}), \tag{1}$$

where $\boldsymbol{y}_t$ is an n-dimensional vector of observed variables at time t, $\boldsymbol{c}$ is a vector of constant factors, $\boldsymbol{A}_1, ..., \boldsymbol{A}_p$ are the coefficient matrices and $\boldsymbol{\varepsilon}_t$ is

[3]All the methods used in the this thesis are implemented in R (R Core Team 2018) and written by the author of the thesis. The source code is available upon request.

the normally distributed vector of error terms with the covariance matrix $\boldsymbol{\Sigma}$.

Evidently, the VAR-models are not very parsimonious in nature and therefore, to avoid *overfitting*, the coefficients to be estimated need to be shrunk towards a more parsimonious prior distribution as the number of variables and lags increases. This is where the Bayesianity comes in. The prior to be used is Normal-Wishart-distributed and very common in the literature. The coefficients themselves are assumed to be normally distributed, whereas the covariance matrix is assumed to follow an Inverse-Wishart distribution. The prior mean of the coefficients is then set to unity for the own first lags of the variables, and to zero otherwise.

Often at this point a distinction between stationary and non-stationary variables is made, and the prior mean of the coefficients on the own first lags of the stationary variables are set to zero, or to some positive number below unity. It can be argued however, that when not differentiated, the mean reversion of stationary economic variables is in most cases reasonably sluggish and that the random walk walk prior (i.e prior with the coefficient on the own first lag set to unity) better fits our perception of the reality as opposed to the white noise prior (i.e prior with all the coefficients set to zero). Natural choice would of course be to set the coefficients on the own first lags of the stationary variables a priori to some value close to but below unity. Specification of this value without leaning on the data would not however be an arbitrary task and for these reasons the random walk prior is used for all the variables in the model. The main charasteristics of the prior mean can be formalized as follows:

$$\mathbb{E}[\,(\boldsymbol{A}_l)_{ij} \mid \gamma\,] = \begin{cases} \gamma, & \text{if } j = i, l = 1 \text{ (i.e. for own first lags)} \\ 0, & \text{otherwise} \end{cases}, \qquad (2)$$

where $\gamma = 1$, as discussed above, coefficient matrices $\boldsymbol{A}_1, ..., \boldsymbol{A}_p$ are assumed to follow a normal distribution, p is the number of lags in the model, $i, j \in \{1, ..., n\}$ and n is the number of variables. The prior variance of the normally distributed coefficients is then set as:

$$\mathrm{Var}[\,(\boldsymbol{A}_l)_{ij} \mid \boldsymbol{\delta}, \boldsymbol{\Psi}\,] = \left(\frac{\lambda_1}{l}\right)^2 \frac{\boldsymbol{\Psi}_{ii}}{\boldsymbol{\Psi}_{jj}}, \quad \text{for all } l \in \{1, ..., p\}, \qquad (3)$$

where λ_1 accounts for the overall tightness of the prior and thus essentially controls the variance of the prior and $\boldsymbol{\delta}$ is the vector of hyperparameters $\lambda_1, \lambda_2, \lambda_3$. Larger the λ_1, closer the estimates will be of coinciding with the OLS-estimates and smaller the λ_1, closer the estimates will be of coinciding with the random walk prior. $\boldsymbol{\Psi}$ is a diagonal square matrix of size n including the prior for the covariance matrix $\boldsymbol{\Sigma}$, and the term $\frac{\boldsymbol{\Psi}_{ii}}{\boldsymbol{\Psi}_{jj}}$ accounts for different variances of the dependent and explanatory variables. The latter term can also be interpreted as the *scaling* factor taking into account the different scales of the variables as the data in the model is not normalized. Finally the prior on the intercept is considered to be non-informative and the variance is set to an arbitrarily large number.

The other hyperparameters in $\boldsymbol{\delta}$ include λ_2 and λ_3. These two hyperparameters control the weight given for two additional very standard priors, the *sum-of-coefficients* prior and the *dummy-initial-observation* prior, that are shown to improve the predictive performance of BVAR-models (see e.g. Karlsson 2012). Notation-wise, the diagonal elements of $\boldsymbol{\Psi}$ could also be included in the vector of hyperparameters $\boldsymbol{\delta}$, but as they (diagonal of $\boldsymbol{\Psi}$) are chosen differently to the hyperparameters in $\boldsymbol{\delta}$ they are treated as a separate set of hyperparameters. The diagonal elements of $\boldsymbol{\Psi}$ are chosen according to residuals of an AR-process of order p. Although this data driven approach is not completely innocuous as the same data is used for the prior and the estimation of the posterior distribution, this is a very common practice in the literature (see e.g. Banbura et al. 2010).

To summarize, the Normal-Wishart prior can be characterized by equations 4 and 5.

$$vec(\boldsymbol{A}) \mid \boldsymbol{\Sigma} \sim N\left(vec(\boldsymbol{A}_0), \boldsymbol{\Sigma} \otimes \boldsymbol{\Omega}_0\right) \tag{4}$$

$$\boldsymbol{\Sigma} \sim iW\left(\boldsymbol{\Psi}, v_0\right), \tag{5}$$

where $\boldsymbol{A}_0$ and $\boldsymbol{\Omega}_0$ are chosen to satisfy equations 2 and 3, v_0 is set to $n+2$ to ensure the existence of a finite prior variance, $\otimes$ denotes a Kronecker product and *vec* denotes a vectorization operator that stacks the columns of a matrix into a single vector. Assuming a normal model according to equation 1, it is straightforward to show that the unnormalized posterior distribution of the coefficients $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$ can be expressed as a function of the prior and the data,

as illustrated in equations 6 and 7 (see e.g. Karlsson (2012) or appendix of Giannone et al. (2015)).

$$vec(\boldsymbol{A}) \mid \boldsymbol{\Sigma}, \boldsymbol{Y} \sim N\left(vec(\widetilde{\boldsymbol{A}}), \boldsymbol{\Sigma} \otimes \left(\boldsymbol{X}^{\intercal}\boldsymbol{X} + \boldsymbol{\Omega}_0^{-1}\right)^{-1}\right), \tag{6}$$

where $\boldsymbol{X}$ is the data matrix of predictors constructed as a function of $\boldsymbol{Y}$ according to equation 1, $\widetilde{\boldsymbol{A}} = \left(\boldsymbol{X}^{\intercal}\boldsymbol{X} + \boldsymbol{\Omega}_0^{-1}\right)^{-1}\left(\boldsymbol{X}^{\intercal}\boldsymbol{Y} + \boldsymbol{\Omega}_0^{-1}\widehat{\boldsymbol{A}}\right)$ and $\widehat{\boldsymbol{A}} = \left(\boldsymbol{X}^{\intercal}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\intercal}\boldsymbol{Y}$. Also,

$$\boldsymbol{\Sigma} \mid \boldsymbol{Y} \sim iW\left(\widetilde{\boldsymbol{S}}, T + v_0\right), \tag{7}$$

where T is the number of observations, i.e the number of rows in $\boldsymbol{Y}$ or $\boldsymbol{X}$, and $\widetilde{\boldsymbol{S}} = \boldsymbol{\Psi} + \widehat{\boldsymbol{S}} + \left(\boldsymbol{A}_0 - \widehat{\boldsymbol{A}}\right)^{\intercal}\left(\boldsymbol{\Omega}_0 + \left(\boldsymbol{X}^{\intercal}\boldsymbol{X}\right)^{-1}\right)^{-1}\left(\boldsymbol{A}_0 - \widehat{\boldsymbol{A}}\right)$

The model can be estimated by explicitly constructing the matrices $\boldsymbol{A}_0$ and $\boldsymbol{\Omega}_0$, but especially with larger systems it is often more convenient, and in some cases necessary for numerical stability, to follow Banbura et al. (2010) and impose the prior structure on the model through artificial observations. Augmenting the data matrices $\boldsymbol{Y}$ and $\boldsymbol{X}$ with $\boldsymbol{Y}_d$ and $\boldsymbol{X}_d$ (see equation 8) and estimating the system with ordinary least squares (OLS) is equivalent to imposing the prior through $\boldsymbol{A}_0$ and $\boldsymbol{\Omega}_0$ and using equations 6 and 7.

$$\boldsymbol{Y}_d = \begin{pmatrix} \boldsymbol{\Psi}/\lambda_1 \\ \boldsymbol{0}_{n(p-1)\times n} \\ \boldsymbol{\Psi} \\ \boldsymbol{0}_{1\times n} \\ \boldsymbol{Y}_a \end{pmatrix} \qquad \boldsymbol{X}_d = \begin{pmatrix} \boldsymbol{J}_p \otimes \boldsymbol{\Psi}/\lambda_1 & \boldsymbol{0}_{np\times 1} \\ \boldsymbol{0}_{n\times np} & \boldsymbol{0}_{n\times 1} \\ \boldsymbol{0}_{n\times 1} & \epsilon \\ \multicolumn{2}{c}{\boldsymbol{X}_a} \end{pmatrix}, \tag{8}$$

where $\boldsymbol{J}_p = \text{diag}(1, ..., p)$, ϵ is some very small number close to zero (uninformative prior on the intercept) and $\boldsymbol{0}_{i\times j}$ is a matrix of zeros with i rows and j columns. The additional priors are imposed through $\boldsymbol{Y}_a$ and $\boldsymbol{X}_a$ to be specified below. Thus, the unnormalized posterior distribution of the coefficients $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$ can be expressed as:

$$vec(\boldsymbol{A}) \mid \boldsymbol{\Sigma}, \boldsymbol{Y} \sim N\left(vec(\widehat{\boldsymbol{A}_*}), \boldsymbol{\Sigma} \otimes \left(\boldsymbol{X}_*^{\intercal}\boldsymbol{X}_*\right)^{-1}\right) \tag{9}$$

$$\boldsymbol{\Sigma} \mid \boldsymbol{Y} \sim iW\left(\widehat{\boldsymbol{S}}_*, T+2+T_d-k\right), \tag{10}$$

where $\boldsymbol{Y}_* = (\boldsymbol{Y}^\intercal, \boldsymbol{Y}_d^\intercal)^\intercal$, $\boldsymbol{X}_* = (\boldsymbol{X}^\intercal, \boldsymbol{X}_d^\intercal)^\intercal$, T_d is the number of artificial observations, k is the number of coefficients per equation, $\widehat{\boldsymbol{A}_*} = \left(\boldsymbol{X}_*^\intercal \boldsymbol{X}_*\right)^{-1} \boldsymbol{X}_*^\intercal \boldsymbol{Y}_*$, and $\widehat{\boldsymbol{S}}_* = \left(\boldsymbol{Y}_* - \boldsymbol{X}_* \widehat{\boldsymbol{A}}\right)^\intercal \left(\boldsymbol{Y}_* - \boldsymbol{X}_* \widehat{\boldsymbol{A}}\right)$.

The additional priors are imposed through $\boldsymbol{Y}_a$, $\boldsymbol{X}_a$, as discussed above, and following closely the notation in Giannone et al. (2015) they are defined as follows:

$$\boldsymbol{Y}_a = \begin{pmatrix} \boldsymbol{Y}^+ \\ \boldsymbol{y}^{++} \end{pmatrix} \qquad \boldsymbol{X}_a = \begin{pmatrix} \boldsymbol{X}^+ \\ \boldsymbol{x}^{++} \end{pmatrix} \tag{11}$$

$$\boldsymbol{Y}^+ \quad = \quad diag\left(\frac{\bar{\boldsymbol{y}}_0}{\lambda_2}\right) \tag{12}$$

$$\boldsymbol{X}^+ \quad = \quad \left[0, \boldsymbol{Y}^+, ..., \boldsymbol{Y}^+\right], \tag{13}$$

where $\bar{\boldsymbol{y}}_0$ is an $n \times 1$ vector that contains the average of the p first observations for each variable, $\boldsymbol{Y}^+$ is an $n \times n$ dimensional matrix, $\boldsymbol{X}^+$ is an $n \times (1 + np)$ dimensional matrix and λ_2 is a positive hyperparameter controlling the strength of the *sum-of-coefficients* (SOC) prior. Smaller the λ_2, more weight is given to the SOC prior. Also,

$$\boldsymbol{y}^{++} \quad = \quad \left(\frac{\bar{\boldsymbol{y}}_0}{\lambda_3}\right)^\intercal \tag{14}$$

$$\boldsymbol{x}^{++} \quad = \quad \left[\frac{1}{\lambda_3}, \boldsymbol{y}^{++}, ..., \boldsymbol{y}^{++}\right], \tag{15}$$

where $\boldsymbol{x}^{++}$ is a $1 \times (1 + np)$ vector, and λ_3 controls the strength of the *dummy-initial-observation* (DIO) prior. Thus, setting hyperparameters λ_2 and λ_3 to infinity would equal to ignoring the additional priors altogether, whereas setting the hyperparameters to zero would put all the weight on these priors, therefore ignoring the data entirely.

The SOC prior enables the coefficients of a variable to have correlation among their own lags by allowing for unit processes of higher order than one implied by the Litterman-styled prior. The SOC prior was originally proposed by Doan et al. (1984) and it was shown to be able to significantly

improve the forecasting accuracy of a BVAR model. The other additional prior, the DIO prior, was originally proposed by Sims (1993) and it is motivated by the fact that the SOC prior alone is not consistent with the possible cointegration of macroeconomic variables.

Although the strength of these additional priors is separately specified through hyperparameters λ_2 and λ_3, in this thesis the relative strength of these additional priors have been fixed according to equation 16, as in Banbura et al. (2010). The hyperparameters λ_2 and λ_3 are treated as fixed for simplicity and most importantly to make the discussion regarding the hyperparameter choice more clear.

$$10 \times \lambda_1 = \lambda_2 = \lambda_3 \tag{16}$$

3.2 Exogeneity

An additional feature of the model explaining the abbreviation BVARX instead of the usual BVAR stems from the exogeneity imposed on global economic variables. This exogeneity is obtained as in Bloor & Matheson (2011), by first estimating a BVAR-model on the exogenous variables alone, and then using the forecasts acquired from this model to append the estimation equations of the endogenous variables with lags and current values of the exogenous variables. Thus in this case, there are two separate blocks of equations, the endogenous and the exogenous block. The exogenous block of equations is first estimated as a typical BVAR-model consisting of only exogenous variables (see equation 1). The endogenous block of equations is then constructed by adding exogenous variables as exogenous predictive variables to a model with only endogenous variables. The equation 17 illustrates.

$$\boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{A}_1\boldsymbol{y}_{t-1} + ... + \boldsymbol{A}_p\boldsymbol{y}_{t-p} + \boldsymbol{B}_0\boldsymbol{z}_t + ... + \boldsymbol{B}_p\boldsymbol{z}_{t-p} + \boldsymbol{\varepsilon}_t, \tag{17}$$

where $\boldsymbol{z}_t$ is a vector of exogenous variables at time t, $\boldsymbol{B}_0, ..., \boldsymbol{B}_p$ are the coefficient matrices on exogenous variables and everything else is as in equation 1.

Evidently, the exogenous block remains symmetric, as all the variables entering the equations are to be forecast. However the block of endogenous

variables becomes non-symmetric as the exogenous variables are added to the estimation equations. This calls for slight adjustments in the prior. The coefficients for the lags of exogenous variables are set to have a prior mean of zero and variance analogous to the endogenous variables. Finally, the coefficients on the exogenous variables from the current period are chosen to have an analogous prior to those from the previous period.

The generalization of this method for more than two blocks of equations is formalized in Zha (1999).

3.3 Conditionality

The final and the major additional feature of the model is the conditional nature of it. To illustrate this feature, following closely the exposition in Bloor & Matheson (2011), the h-step out-of-sample forecast at time T can be decomposed as follows:

$$\boldsymbol{y}_{T+h} = \boldsymbol{d}_{T+h} + \sum_{j=1}^{h} \boldsymbol{M}_{h-j} \boldsymbol{\epsilon}_{T+j}, \tag{18}$$

where the vector $\boldsymbol{d}_{T+h}$ includes the unconditional dynamic forecasts produced by the model. The future shocks effect the forecast through the second term, where $\boldsymbol{M}_{h-j}$ illustrates the impulse response functions and the future structural shocks are presented by $\boldsymbol{\epsilon}_{T+j}$. For forecasting purposes, the structural errors can always be identified with a recursive identification scheme, since as shown by Waggoner & Zha (1999) the predictive distribution is invariant to orthonormal transformation of the system. Hence, the ordering of the equations within a given block has no effect on the predictive distributions produced. Therefore, one can implicitly operate with reduced form errors instead of structural ones.

As $\boldsymbol{M}$ is a function of already estimated $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$, the conditional forecast can be acquired by restricting the reduced form shocks to be drawn from a distribution satisfying the restriction to be conditioned on. This can be obtained by following the conditioning algorithm of Andersson et al. (2010). As the forecasts are to be conditioned on normally distributed future projec-

tions, the conditions are to be formed according to the equation 19.

$$\boldsymbol{C}\boldsymbol{y}_{T+1,T+h} \sim N\left(\boldsymbol{f}_{T+1,T+h},\, \boldsymbol{\Omega}_{\boldsymbol{f}}\right), \tag{19}$$

where $\boldsymbol{y}_{T+1,T+h}$ includes all the forecasts from horizon $T+1$ to $T+h$, $\boldsymbol{f}_{T+1,T+h}$ represents the mean or central tendency of the restrictions, $\boldsymbol{\Omega}_{\boldsymbol{f}}$ is the covariance matrix of the restrictions and the matrix $\boldsymbol{C}$ completes the restrictions on the first moments by *picking* the right future values to be restricted. Thus, usually $\boldsymbol{C}$ consists of one element per row set to unity, while all the other elements are set to zero. However, more versatile linear restrictions can be imposed through different structure of $\boldsymbol{C}$, which underlines the flexibility of this conditioning method. To elaborate further, the central tendency of the restrictions, $\boldsymbol{f}_{T+1,T+h}$, is a vector of all future values to be conditioned on. If for example, one were to condition the forecast on only future values of one particular variable, the vector $\boldsymbol{f}_{T+1,T+h}$ would consist of only those values and the matrix $\boldsymbol{C}$ would be constructed to pick only the corresponding elements of $\boldsymbol{y}_{T+1,T+h}$.

Also, $\boldsymbol{\Omega}_{\boldsymbol{f}} = \boldsymbol{\Xi} \circ \left(\boldsymbol{\sigma}_f \boldsymbol{\sigma}_f^{\intercal}\right)$, where $\circ$ stands for Hadamard product (i.e. element-wise matrix multiplication) and the restrictions regarding the second moments of the predictive distribution (i.e. the uncertainty around the values to be conditioned on) enter into the algorithm through $\boldsymbol{\sigma}_f$. Alternatively, one could set $\boldsymbol{\Omega}_{\boldsymbol{f}} = \boldsymbol{D}\boldsymbol{D}^{\intercal}$, where $\boldsymbol{D} = \boldsymbol{C}\boldsymbol{M}^{\intercal}$ and $\boldsymbol{M}$ is as in equation 20. This way the uncertainty around the values to be conditioned on is itself estimated from the data, which makes the conditioning of the forecast on uncertain future values sensible even in the absence of prior information on the uncertainty regarding those future values.

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{M}_0 & \boldsymbol{M}_1 & \boldsymbol{M}_2 & \dots & \boldsymbol{M}_{h-1} \\ \boldsymbol{0} & \boldsymbol{M}_0 & \boldsymbol{M}_1 & \dots & \boldsymbol{M}_{h-2} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{M}_0 & \dots & \boldsymbol{M}_{h-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{M}_0 \end{pmatrix}, \tag{20}$$

where $\boldsymbol{M}_{h-j}$ is $n \times n$ matrix of impulse responses as in equation 18, $j \in \{1, ..., h\}$ and subscript $h-j$ denotes the number of periods from the shock.

For example, $\boldsymbol{M}_0$ includes the response of every variable to a shock from every variable from the same period.

One is able to formulate so-called *hard restrictions* of Waggoner & Zha (1999) as a special case of density restrictions elaborated in this section. Hard restrictions can be obtained by setting every element of $\boldsymbol{\sigma}_f$ to zero, which is equivalent to imposing no uncertainty around the values to be conditioned on. The hard restrictions are particularly useful when dealing with the so called *ragged edge* of the data. A data set of time series is said to have a ragged edge if for the most recent periods there are observations for only some of the variables (i.e. the data is unbalanced). Ordinary VAR models (or almost any other multivariate time series models) are not capable of utilizing the data from those incomplete periods, but are forced to discard the latest observations. With hard restrictions however, the forecast can be conditioned on those latest observations and all the available information can be preserved in order to produce more accurate forecasts.

Given one has specified the central density $\boldsymbol{f}_{T+1,T+h}$ and covariance matrix $\boldsymbol{\Omega_f}$ of the restrictions, the unconditional dynamic forecast *without* the error term can be transformed into a conditional forecast by adding the restricted errors according to equation 18. The distribution of the restricted errors $\widetilde{\boldsymbol{\epsilon}}_{T+1,T+h}$ can be derived from 19 and is presented in equation 21. For full derivation of this result, see Andersson et al. (2010).

$$\widetilde{\boldsymbol{\epsilon}}_{T+1,T+h} \sim N\left(\boldsymbol{\mu}_\epsilon,\, \boldsymbol{\Sigma}_\epsilon\right), \tag{21}$$

where

$$\boldsymbol{\mu}_\epsilon = \boldsymbol{D}^* \boldsymbol{f}_{T+1,T+h} - \boldsymbol{D}^* \boldsymbol{C} \boldsymbol{d}_{T+1,T+h},$$
$$\boldsymbol{\Sigma}_\epsilon = \boldsymbol{D}^* \boldsymbol{\Omega_f} \left(\boldsymbol{D}^*\right)^\intercal + \widehat{\boldsymbol{D}}^\intercal \widehat{\boldsymbol{D}}$$

$\boldsymbol{D}^*$ denotes the generalized inverse of $\boldsymbol{D}$, the rows of the matrix $\widehat{\boldsymbol{D}}$ are chosen to form an orthonormal basis for the null space of $\boldsymbol{D}$ and $\boldsymbol{d}_{T+1,T+h}$ denotes the unconditional dynamic forecast *without* the error term (see equation 18). The other elements are as previously defined.

Although the posterior distribution of the coefficients is known and can be sampled from directly, sampling from the conditional predictive distribution

requires for a Gibbs sampler. Gibbs sampler is required since the conditions laid out in equation 19 are nonlinear functions of the coefficients themselves and the correct distribution of the coefficients conditional on equation 19 must therefore be derived from the joint distribution of the coefficients and $\boldsymbol{Y}_{T+1,T+h}$. This joint distribution is generally unknown, but as shown in Waggoner & Zha (1999), it is feasible to simulate the distribution with a Gibbs sampler by augmenting the data used for the estimation of the coefficients with the conditioned forecast drawn in the previous period.

Conditionality is a remarkably attractive feature for a forecasting model since as already discussed it allows for efficient processing of the latest observations and flexible conditioning of the forecasts on uncertain information coming from outside the model. Moreover, the developed conditional model could be used for example to produce full predictive densities around pre-specified point-estimates of other forecasters, by forcing the conditional forecasts around those point-estimates but allowing the model to estimate the higher moments of the predictive distribution. This could prove to be practical since the official forecasts of many forecasting agencies are usually at least partly based on subjective considerations and are therefore usually not accompanied by any metrics regarding the uncertainty of the forecast presented. However, for the predictive densities not to be too wide or narrow, the underlying conditional model should on average be able to produce as accurate forecasts as the ones based on subjective considerations.

3.4 Forecasting algorithm

After this overview of the econometric methods required for estimating the model and producing the forecast, summarizing the complete algorithm is in order. The data used in the model consists of both monthly and quarterly variables, which are further divided into exogenous and endogenous variables. Therefore, the estimation of the model comprises four separate conditional BVAR-models. Each conditional BVAR-model is estimated analogously, only with different data and with different hyperparameters.

First, the *symmetric* system of exogenous monthly variables is estimated as an ordinary BVAR-model, leaving the latest periods for which the data is

incomplete out of the data. The dynamic central tendency of the unconditional predictive density is then produced according to the equation 1 *without* the error term. Next, the conditional forecast is acquired by drawing the reduced form errors according to the conditioning algorithm of Andersson et al. (2010) discussed above and summing them with the unconditional forecast (see equation 19). The drawn errors are conditioned on the observations from the incomplete periods *and* on the possible additional restrictions consisting of the predictive densities of certain variables. These additional restrictions might include projections from outside the model on the future path of the global interest rates or on the economic activity of the export markets of an individual country, for example.

Then, as proposed by Waggoner & Zha (1999), to ensure that the parameter estimates are consistent with the conditional forecast produced, the forecast is appended to the data used for estimation of the coefficients before the next draw. These steps are repeated N times, after which the first half of the sample is burned in order to make sure the chain has converged. This comprises the Gibbs sampler required for producing the conditional forecasts and the same steps are followed with the other three sub-models.

Algorithm 1 Sampling from the conditional predictive distribution

1: **for** i ∈ 1,...,N **do**
2: Draw the coefficients $\boldsymbol{A}^{(i)}, \boldsymbol{\Sigma}^{(i)} \mid \boldsymbol{Y}, \boldsymbol{Y}_{T+1,T+h}^{(i-1)}$ according to equations 9 and 10.
3: Draw the unconditional forecast $\boldsymbol{d}_{T+1,T+h}^{(i)} \mid \boldsymbol{A}^{(i)}, \boldsymbol{\Sigma}^{(i)}, \boldsymbol{Y}$ by simulating from equation 1 or 17 *without* the error term.
4: Draw the restricted errors $\widetilde{\boldsymbol{\epsilon}}_{T+1,T+h}^{(i)} \mid \boldsymbol{d}_{T+1,T+h}^{(i)}, \boldsymbol{f}_{T+1,T+h}, \boldsymbol{\Omega}_{\boldsymbol{f}}, \boldsymbol{C}$ according to equation 21.
5: Compute the conditional forecast $\boldsymbol{Y}_{T+1,T+h}^{(i)} \mid \boldsymbol{d}_{T+1,T+h}^{(i)}, \widetilde{\boldsymbol{\epsilon}}_{T+1,T+h}^{(i)}, \boldsymbol{A}^{(i)}, \boldsymbol{\Sigma}^{(i)}$ according to equation 18.
6: **end for**
7: Discard the first half of the sample $\{\boldsymbol{Y}_{T+1,T+h}^{(1)}, ..., \boldsymbol{Y}_{T+1,T+h}^{(N)}\}$

The *non-symmetric* system of endogenous monthly variables is then estimated analogously, only this time producing the dynamic central tendency of the unconditional predictive density according to the equation 17, using the newly produced forecasts of exogenous variables $\boldsymbol{z}_{t+h}^{(i)}$ for each draw i,

where $i \in 1, ..., N$.

At this point, the series of monthly frequency are aggregated to a quarterly frequency, and the newly produced forecasts of monthly variables are used to approximate a predictive distribution of the quarters for which there are only some observations of monthly variables. In other words, the forecasts from the sub-models for monthly variables constitute a sample of the latest quarters in the data of aggregated variables. The uncertainty revolving these quarters constructed of the incomplete monthly observations is preserved by randomly sampling from the predictive distribution of these quarters prior to every draw from the posterior predictive distribution of the set of variables consisting of both quarterly frequency and the monthly variables aggregated to a quarterly frequency.

To illustrate, after the forecasts for the variables of monthly frequency are produced and both the series and the forecasts are aggregated to a quarterly frequency, the forecast quarters for which there were no observations of any of the variables are dropped from the sample, as the monthly forecasting model is only used to produce a predictive distribution of the incomplete quarters. Then the data of quarterly frequency is augmented with the newly aggregated series originally of monthly frequency producing a system of exogenous and endogenous variables. This new data is then used to produce the conditional quarterly forecast analogously to the monthly forecasts, with only one difference. Prior to every draw from the posterior predictive distribution of quarterly variables, new aggregated quarters are drawn randomly from the sample produced by the model of monthly frequency.

This kind of random sampling with replacement from a sample drawn itself from some distribution (in this case the predictive posterior distribution produced by the monthly BVARX-model) is often referred to as *bootstrapping*. One could set the number of draws from the predictive distribution of monthly variables to exactly match the number of draws needed by the quarterly model and use this sample directly, but by bootstrapping one allows the number of draws to vary, which might be desirable for computational reasons. Another approach for conditioning the quarterly forecasts on latest monthly observations would be to assume that the predictive distribution of aggregated monthly variables is approximately normal and then use the con-

ditioning algorithm laid out above. This assumption however does not hold very well. Even in the case of unconditional model, the predictive distribution would not be normal, but *t-distributed*, and there is no reason to believe that the predictive distribution of the quarters acquired by aggregating the monthly forecasts would be normal either. The bootstrapping procedure thus allows one to preserve the non-normal distribution of the quarters containing uncertainty due to missing monthly observations.

Algorithm 2 Sampling from the predictive distribution of the conditional BVARX-model with monthly and quarterly variables

1: Draw a sample $\{\boldsymbol{Y}^{M,Exo}_{T+1,T+h}\}$ of conditional forecasts of *exogenous monthly* variables according to algorithm 1.

2: Draw a sample $\{\boldsymbol{Y}^{M,Endo}_{T+1,T+h} \mid \{\boldsymbol{Y}^{M,Exo}_{T+1,T+h}\}$ of conditional forecasts of *endogenous monthly* variables conditioning $\boldsymbol{Y}^{M,Endo,(i)}_{T+1,T+h}$ on $\boldsymbol{Y}^{M,Exo,(i)}_{T+1,T+h}$, for all $i \in \{1, ..., N_m\}$ according to the algorithm 1 and equation 17.

3: Augment the obtained samples with the data $\boldsymbol{Y}^M$ and aggregate the monthly series to a quarterly frequency to obtain $\{\boldsymbol{Y}^{M,exo}\}$ and $\{\boldsymbol{Y}^{M,endo}\}$.

4: Discard the quarters for which there were *no* observations on any of the monthly variables as the monthly model is only used to fill in the latest quarters.

5: Augment the exogenous and endogenous quarterly data with $\{\boldsymbol{Y}^{M,exo,(r)}\}$ and $\{\boldsymbol{Y}^{M,endo,(r)}\}$, where r is randomly drawn from $\{1, ..., N_m\}$.

6: To produce the final forecast, draw the samples $\{\boldsymbol{Y}^{Q,Exo}_{T+1,T+h}\}$ and $\{\boldsymbol{Y}^{Q,Endo}_{T+1,T+h}\}$ of quarterly forecasts analogously to the steps 1 and 2. Prior to every draw, repeat the step 5.

The model presented in this section is highly flexible and it seeks to preserve the uncertainty of the estimates as precisely as possible in every step of the forecasting algorithm. The objective of the model is to produce as accurate full forecasting densities as possible, not limiting the forecast to mere point-estimates. The exogeneity of the global variables is imposed for more efficient parameter estimation and the conditional nature of the model allows for efficient utilization of both monthly and quarterly variables. The

conditionality of the model also allows the forecaster to impose uncertain believes regarding the predictive variables into the model in a highly flexible manner and the model could even be used for producing full predictive densities around pre-specified point-estimates.

4 Data

Although the model developed in this thesis is applicable for medium term forecasting of economic variables in any small open economy, in this thesis the focus has been directed to Finland. Also, as small open economies are often highly dependent of international trade, the export and import flows as variables are of special interest, in addition to gross domestic product (GDP) of course.

As already discussed, the model consists of both monthly and quarterly variables which have been further divided into exogenous and endogenous variables. The endogenous variables consist mostly of official statistics acquired from Statistics Finland. These include the quarterly national account and monthly economic indicators such as price indices, confidence indicators and labour market statistics. Also the monthly statistics on the volume of imports and exports of goods provided by the custom authorities of Finland are used in the largest version model.

A priori the most important exogenous variable is the quarterly indicator of economic activity of trade partners of Finland (Trade market GDP in table 1). This indicator is acquired as a weighted sum of the gross domestic product figures of trade partners of Finland acquired from OECD. The weights for the countries are acquired from the custom authorities of Finland as *shares of bilateral trade of goods between Finland and the respective country* of the total trade of goods by Finland. Unfortunately there is no corresponding data available to produce a set of weights for the trade in services, but as the trade in goods comprises over two thirds of the total trade in Finland[4] this should suffice.

Other exogenous variables are of monthly frequency and consist of interest

[4]International trade in goods and services 2018Q3, Statistics Finland

rates, exchange rates, price indices and confidence indicators. The export and import prices of a small open economy are treated as exogenous, as in the Bloor & Matheson (2011). A full list of variables can be found in table 1 and in the Appendix A.

Model	Endogenous variables	Exogenous variables	Number of variables
Small	GDP Exports of goods Exports of services Imports of goods Imports of services Unemployment rate Inflation rate		7
Medium	*Small*	Export price index Import price index ESI (EU) USD/EUR Trade market GDP	12
Large	*Medium* + Employment rate Consumer confidence Industrial confidence Exports (Customs) Imports (Customs) Building permits Private consumption Public consumption Investments (residental) Investments (residental exc.)	*Medium* + Oil price Euribor 3kk CPI (EU)	25

Table 1: *Variables of the Small, Medium and Large model.*

As is evident after inspecting table 1, there are three models of different size. The smallest model called *Small* includes only 7 variables with no exogenous variables. The second model is called *Medium* and it is constructed of the variables in the small model plus five additional exogenous variables, totaling 12 variables. The largest model is called *Large* and it consists of 25

variables in total. Although models up to only 25 variables are studied, the hyperparameterization of the model should allow for an efficient exploitation of an arbitrarily large number of variables. At some point the conditional forecasting algorithm presented earlier would however become in practice computationally too burdensome.

The data starts from the beginning of the European monetary union, which is considered to be the date from which there is official exchange change rate data available for the Euro. The data thus starts from the beginning of the year 1999 and spans for approximately 20 years to the third quarter of 2018. The starting date is used to avoid structural breaks in the estimation period as well as possible, and to ensure the availability of as many useful time series as possible, while still keeping the data sufficiently long for the estimation and out-of-sample study of the model.

All the data is in levels and logarithmic transformation is carried out for all the variables that are not already expressed in rates or percentage points.

5 Hyperparameter choice

The hyperparameter choice is arguably the most important single thing affecting the performance of a BVAR-model. As discussed earlier in the previous section, the hyperparameter values determine the extent to which the prior believes are weighted as compared to the information gathered from the data. As more variables are added to a parameter rich VAR-model, the model becomes less parsimonious and the importance of the hyperparameter choice increases. Without *shrinking* the parameter estimates towards a suitable prior distribution, the limited amount of data causes the parameter estimates to *over-fit* the data and the model loses it's capability to predict anything outside the sample used for estimation. As one has only one timeline at disposal to sample time series data from, the problem of limited amount of data is not easily avoidable.

For these reasons, traditional VAR-models rarely perform on a par with BVAR-models beyond a very limited amount of variables and lags. The optimal amount of shrinkage towards the prior is however non straightforward to determine as too loose prior (too little shrinkage) causes the parameter

estimates to over-fit, whereas too tight prior (too much shrinkage) prevents the model from exploiting all the information available. Several factors affect the optimal amount of shrinkage in a BVAR-model: the number of observations, the number of variables, the number of lags, the noisiness of the data and of course the suitability of the prior. To make things even more difficult, the optimal amount of shrinkage may sometimes vary when comparing the out-of-sample forecasting errors of different variables within the same model.

In the literature review, several different approaches for hyperparameter choice were discussed. The approaches can be roughly divided to four classes: (i) fixed values from the literature (usually the rule-of-thumb values proposed by Sims & Zha (1998)), (ii) minimization of the out-of-sample forecasting errors over pre-specified time interval and grid of parameter values, (iii) the in-sample-fit based method by Banbura et al. (2010) and (iv) the marginal likelihood based methods, e.g. the one formalized in Giannone et al. (2015).

In this section the marginal likelihood function based method of Giannone et al. (2015) later simplified and applied for a small BVAR-model nowcasting the unemployment rate in Finland by Anttonen (2018) is studied in depth. As an alternative to that method a novel algorithm based on the minimization of the past out-of-sample forecasting errors is proposed. Finally, both the performance of the newly proposed algorithm and the marginal likelihood based approach are empirically assessed against the fixed rule-of-thumb values proposed by Sims & Zha (1998).

5.1 Marginal likelihood based approach

Marginal likelihood based approach (ML approach) builds upon the idea of treating the hyperparameters as just another set of parameters to be estimated in a hierarchical setting. Giannone et al. (2015) formalize this approach by imposing a full hierarchical structure on a BVAR-model by specifying a prior distribution on the hyperparameters themselves and centering it around the rule-of-thumb values of Sims & Zha (1998). The full hierarchical structure however complicates the sampling process as Metropolis algorithm is required for sampling from the posterior distribution. The fairly informative but fixed prior on the hyperparameters is not a very appealing feature

either, since bigger models are *a priori* believed to need much tighter priors than the smaller ones.

Anttonen (2018) simplifies the approach of Giannone et al. (2015) by numerically estimating the mode of the *hyperposterior* distribution prior to estimating the model, thus parting ways with the full hierarchical structure of Giannone et al. (2015) and allowing for direct sampling from the posterior distribution of the BVAR-model. The equation 22 illustrates the structure of the hyperposterior distribution and follows directly from the Bayes' theorem.

$$\underbrace{p\left(\boldsymbol{\delta} \mid \boldsymbol{y}\right)}_{\text{Hyperposterior}} \propto \underbrace{p\left(\boldsymbol{y} \mid \boldsymbol{\delta}\right)}_{\text{ML}} \underbrace{p\left(\boldsymbol{\delta}\right)}_{\text{Hyperprior}} \tag{22}$$

Evidently, the unnormalized hyperposterior distribution can be expressed in a closed form if one possesses a closed form solution for the marginal likelihood function. The closed form solution also makes the numerical search for the mode of the hyperposterior distribution very straightforward to execute. Under a flat (i.e non-informative) hyperprior the problem reduces to a maximization problem of the marginal likelihood over the vector of hyperparameters $\boldsymbol{\delta}$.

Giannone et al. (2015) derive a moderately complex closed form representation of the marginal likelihood function by integrating the coefficients $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$ out of the joint posterior distribution of the coefficients after imposing the prior beliefs through $\boldsymbol{A}_0$ and $\boldsymbol{\Omega}_0$, as discussed in the third section. However, by imposing the prior distribution through artificial observations only, one is able to derive considerably simpler closed form representation of the proportional marginal likelihood function (see equation 23)[5], that can be shown to produce exactly the same estimates of the optimal vector of hyperparameters $\boldsymbol{\delta}$ as the one in Giannone et al. (2015).

$$p\left(\boldsymbol{Y} \mid \boldsymbol{\delta}\right) \propto \frac{\mid \widehat{\boldsymbol{S}}_d \mid^{\frac{v_d}{2}} \mid \boldsymbol{X}_d^{\intercal}\boldsymbol{X}_d \mid^{\frac{m}{2}}}{\mid \widehat{\boldsymbol{S}}_* \mid^{\frac{v_*}{2}} \mid \boldsymbol{X}_*^{\intercal}\boldsymbol{X}_* \mid^{\frac{m}{2}}}, \tag{23}$$

where $v = T - p - k - m - 1$, all the other elements are as described in the third section and subscripts d and $*$ implicate that the respective element is

[5]For derivation, see Appendix B.

constructed of only dummy observations or of all observations including the dummy observations, respectively.

The ML approach however has its deficiencies, not considered in Giannone et al. (2015). The idea of Bayesian shrinkage is to reduce the parameter estimation uncertainty by shrinking the parameter estimates towards a cautious prior, controlling the shrinkage of a potentially high dimensional system through a lower dimensional vector of hyperparameters. By a suitable choice of those hyperparameters the *curse of dimensionality* can be dealed with and no over-fitting takes place. However, if non-infinite amount of data is used to estimate the hyperparameters themselves, they can not be in any way immune to over-fitting the data themselves, no matter how low dimensional the vector of hyperparameters. This holds of course in practice for almost any method used for choosing the hyperparameters, but it needs to be emphasized that the ML approach even with a hierarchical structure of Giannone et al. (2015) is not an exception, especially when the hyperprior is centered around fixed and *non-zero* values.

Monte Carlo simulation experiment shows, that the ML approach tends to give less weight to the prior distribution (by means of hyperparameter values) when there is less data, which is exactly at odds with what would be desirable with a limited amount of data available. In figure 1 are the results of the simulation experiment. To keep things simple, the generated data is bivariate and the model is thus reasonably parsimonious. On the left hand side of the figure 1, the data was generated from a harmonic VAR-process of order 13, as this is the order often chosen with monthly BVAR-models. The order of the model to be estimated is also chosen to impose some dimensionality to the parameter space. Still, a model with 27 parameters per equation is fairly parsimonious as opposed to the BVAR-models discussed in this thesis, which might involve hundreds of parameters per equation. On the right hand side of the figure 1, both variables follow an independent random walk process.

In the simulation exercise 100 artificial time series, 200 observations per series, were generated. With every series, the optimal value for λ_1 implied by the ML approach was computed at intervals of ten observations, starting from 30 observations. The model fitted was an ordinary BVAR-model discussed in the third section, without any additional priors (sum-of-coefficients- and

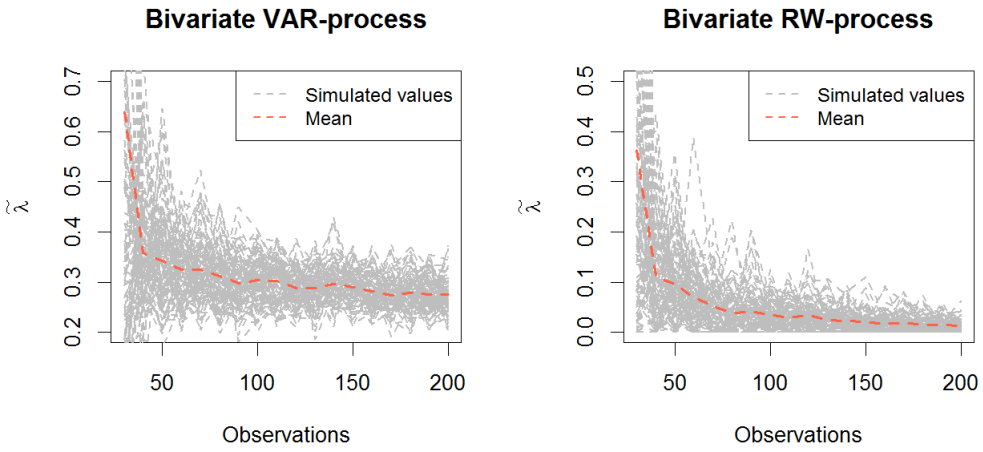

Figure 1: *Monte Carlo simulation experiment illustrating the tendency of the hyperparameters chosen by the marginal likelihood based approach to over-fit the data on average when the amount of data is limited. The dashed grey lines are the hyperparameter values implied during 100 simulations and the red dashed line represents the mean of all those values given the number of observations in the data.*

dummy-initial-observation prior).

Even with a model as parsimonious as this, on average, the implied optimal hyperparameter value decreases for at least eighty observations, i.e. the strength of the prior increases as more observations become available. With data of monthly frequency there is usually more than eighty observations available to fit a model of economic variables with, but with larger models and with data of quarterly frequency, the danger of over-fitting with a limited amount of data implied by the Monte Carlo exercise might become an issue. Although the exercise was performed as if the data was of monthly frequency, eighty observations of quarterly data means 20 years of data, which is coincidentally the length of the data available for the model developed in this thesis.

One way to deal with the above illustrated deficiency of the ML approach would be to use a more suitable hyperprior than the one proposed by Giannone et al. (2015). More suitable hyperprior would shrink the parameter estimates towards the cautious prior, just like the *ordinary prior*, thus pre-

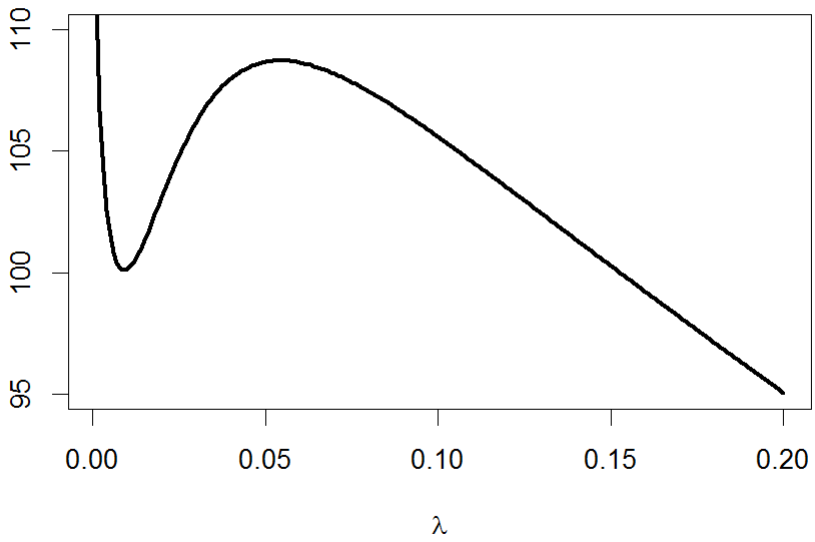

Figure 2: *Logarithmic marginal likelihood function or model evidence as a function of one hyperparameter controlling the overall shrinkage of the medium sized model.*

venting the hyperparameters from over-fitting the model when there is not enough data available for reliable estimation of the low-dimensional vector of hyperparameters $\boldsymbol{\delta}$. The proper choice of the variance of the hyperprior would however remain a cumbersome task to solve and this approach is not pursued further in this thesis.

Another practical issue regarding the ML approach that can be solved with a proper choice of hyperprior, is the behavior of the marginal likelihood function with hyperparameter values close to zero. As illustrated in figure 2, the marginal likelihood function may yield higher values with hyperparameter values very close to zero than the desired local maximum of the function further away from the zero. The hyperparameters set according to the local solution away from the zero would also seem to produce more accurate forecasts than the random walk model implied by the solution of the marginal likelihood function with hyperparameters close tho zero. Therefore, the behavior of the marginal likelihood function can be deemed undesirable and

the hyperparameter values should be restricted from getting values too close to zero when using marginal likelihood based methods for hyperparameter choice. The source of this behavior of the marginal likelihood function is not however clear and this matter is not studied to a further extent in this thesis.[6]

5.2 New algorithm

To address the issues regarding the ML approach to hyperparameter choice, alternative methods must be considered. Essentially, the optimal hyperparameter values can be defined as the ones producing the most accurate out-of-sample forecast, given the model at hand. Therefore, the idea of simply minimizing the past out-of-sample forecasting error of the model over some pre-specified grid of hyperparameters strikes as a compelling approach to choose the hyperparameters. The approach however has some impractical properties and lacks the theoretical elegance of the ML approach.

First, it is often computationally unfeasible to compute the out-of-sample forecasting errors over the grid of pre-specified hyperparameter values, especially if the vector of hyperparameters is multi-dimensional, as it would require estimating the potentially computationally burdensome model as many times as there are elements in the grid of hyperparameters. Second, it is not an arbitrary task to choose the proper support or density of this grid of hyperparameters, especially given the computational limitations at hand.

While the elegance of the ML approach may not be matched with the novel algorithm to be proposed, the above mentioned practical issues concerning the minimization of the past out-of-sample forecasting errors over the hyperparameter values can be dealt with, and the algorithm can be shown to produce in many cases forecasts of superior out-of-sample performance when compared to the ML approach.

The first feature of the algorithm builds upon the fact that the mean of the predictive distribution of a BVAR-model with a normal error term is available in a closed form for any forecasting horizon, even if the model is conditional.

[6]It is possible that the observed behavior of the marginal likelihood function is merely a result of numerical instability of the function with hyperparameter values close to zero.

Thus defining the mean of the predictive distribution as the point-estimate of the forecast, one is able to produce forecasts with no uncertainty in a computationally highly efficient way. Then, instead of pre-specifying a grid of hyperparameters to optimize over, one can specify a closed set (i.e infinitely dense grid) of almost arbitrary size of hyperparameter values to optimize over, and use numerical optimization methods to minimize the past out-of-sample forecasting error over that set of hyperparameter values.

First, let's define:

$$f(\boldsymbol{\delta}) = \sum_{t=t_1}^{T} g\left(\left(\widehat{\boldsymbol{y}}_{t+1,t+h} \mid \boldsymbol{y_t}\right) - \boldsymbol{y}_{t+1,t+h}\right), \tag{24}$$

where the vector $\widehat{\boldsymbol{y}}_{t+1,t+h} \mid \boldsymbol{y_t}$ includes all the point-estimates of the future values up to period $t+h$ at time t, conditional on the information available up to period t, and the vector $\boldsymbol{y}_{t+1,t+h}$ includes the realized values of those periods. The difference of those vectors is therefore a vector of the out-of-sample forecasting errors of the model. This vector can be denoted as $\boldsymbol{z}_t = \left(\widehat{\boldsymbol{y}}_{t+1,t+h} \mid \boldsymbol{y_t}\right) - \boldsymbol{y}_{t+1,t+h}$. In the sum, t_1 is the first period at which the out-of-sample forecasting error is evaluated in order to choose the optimal set of hyperparameters. The period t_1 should be chosen to allow for both, (i) enough observations at time t_1 for reliable estimation of the model and (ii) enough periods $t_1, ..., T$ for the evaluation of the forecasting errors.

Given the data and the structure of the model, $\boldsymbol{z}_t$ only depends on the vector of hyperparameters, $\boldsymbol{\delta}$. Hence, given some functional form of $g(.)$, $f(.)$ is only a function of $\boldsymbol{\delta}$, and $g(.)$ can be defined as any function of the out-of-sample forecasting errors, $\boldsymbol{z}_t$, of the model. For example, in the next subsection, the optimal set of hyperparameters is defined to minimize the squared one-step-ahead out-of-sample forecasting error of the gross domestic product and therefore $g(.)$ is defined as:

$$g(\boldsymbol{z}_t) = \left(z_t^{GDP,\, h=1}\right)^2, \tag{25}$$

where $z_t^{GDP,\, h=1}$ denotes the element of the vector $\boldsymbol{z}_t$ corresponding to the one-step-ahead forecasting error of the gross domestic product.

Finally, the optimal set of hyperparameters $\widetilde{\boldsymbol{\delta}}$ can be obtained as the

solution for the following minimization problem:

$$\min_{\boldsymbol{\delta}} f(\boldsymbol{\delta}), \tag{26}$$

subject to, $\delta_j > 0$ for every element j of $\boldsymbol{\delta}$.

As the algorithm enables one to specify the objective function to be minimized over, the hyperparameters can be chosen to minimize the out-of-sample forecasting error of a certain variable and forecasting horizon or of a combination of variables and horizons. Although it is theoretically compelling with the ML approach to maximize the *model evidence* (i.e marginal likelihood function) of the whole model, it might not be as desirable feature in practice. Especially in a large model there might be several predictive variables that are practically impossible to forecast themselves and some of these variables might even have very little predictive power on the variable of interest. The addition of a variable like this is detrimental to the forecasting performance of the model and the detrimental effect can not be avoided even with Bayesian shrinkage, since any tightening of the prior reduces also the amount of useful information that can be extracted from the data. For this very same reason, the addition of variables to a large BVAR model with a standard prior structure often does not seem to enhance the performance of the model after a very limited number of variables, as noted before in Banbura et al. (2010). ML approach amplifies this effect as there is no distinction in importance between the variables and the additional shrinkage required by the addition of an unnecessary predictive variable to a model is even greater.

The algorithm proposed is very straightforward to implement, but there is no guarantee that the objective function behaves well and multimodularity might become an issue if the closed set of hyperparameter values to be optimized over is not carefully selected. Figure 3 illustrates a situation in which a typical gradient based numerical optimization algorithm might get stuck in the local minimum near zero, never finding the global minimum, leading to a suboptimal set of hyperparameters implied. Solution to this problem is to simply restrict the set of hyperparameter values not to include values close to zero. This is equivalent to setting a uniform prior distribution over positive values of hyperparameters with probability zero near zero.

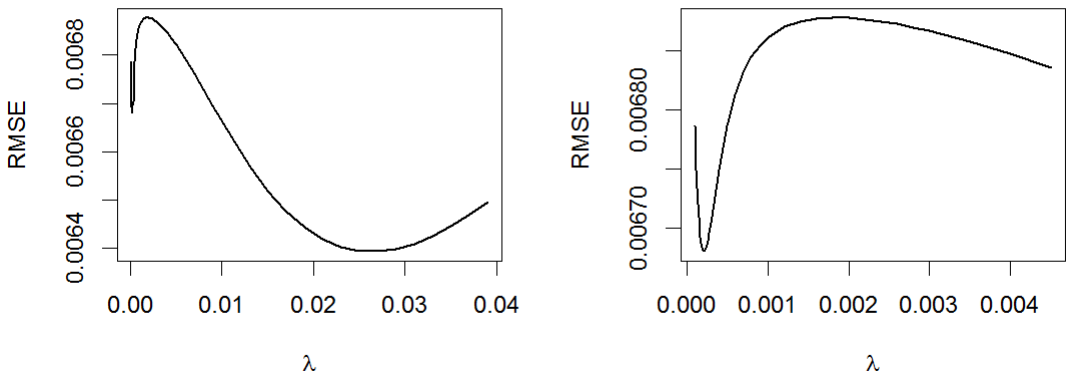

Figure 3: *The root mean squared one-step-ahead out-of-sample forecasting error of the medium sized model from 2010Q1 to 2018Q3 as a function of the hyperparameter λ_1 controlling the shrinkage of the fourth sub-model keeping all the other hyperparameter values fixed. On the right hand side is a zoomed version of the picture on the left hand side.*

L-BFGS-B (Byrd et al. 1995), an extension of the celebrated *BFGS* algorithm, is used for solving the numerical optimization problems in this thesis, as the method conveniently allows for specifying the lower and upper bounds for the set of values to be optmized over.

5.3 Empirical assessment

To empirically assess the performance of the discussed methods for hyperparameter choice, a pseudo out-of-sample forecasting exercise is performed. In this exercise, the out-of-sample forecasts are computed as if the forecaster would have only the information that was available at the time. This requires some assumptions regarding the publication lags of the variables in the model and all these assumptions are listed in the Appendix A. For example, the first official publication of quarterly GDP figure in Finland is published approximately 90 days after the end of the respective quarter. The GDP is thus listed to have a publication lag of one quarter, as in the beginning of 2019Q1

the latest observation at one's disposal would not be from 2018Q4, but from 2018Q3. The forecasting horizons are also defined with respect to the latest observation.[7]

Because there are always several revisions to the data for a few years after the first official figures, the forecasting performance from this exercise does not exactly match the *real* forecasting performance of the model. The latest observations of economic data usually involve some amount of uncertainty, which is not usually accounted for explicitly in the forecasting models. This uncertainty is not accounted for in the model developed in this thesis either, since accounting for it would require data on the revisions not easily available. This can be regarded as a serious shortcoming, as it causes the predictive densities produced by the model in real time to be most probably a bit too narrow.

The pseudo out-of-sample forecasting exercise is ran from 2012Q1 to 2018Q4 and forecasts for horizons of length one and four are produced for every quarter. The time interval is chosen to always ensure at least ten years of data for the estimation of the model. The algorithm for hyperparameter choice discussed above uses periods starting from 2009Q4 for computation of the out-of-sample forecasting errors.

As with this exercise the interest lies in the hyperparameter choice rather than in the forecasting accuracy of different variables per se, only forecasts regarding the GDP are compared and discussed. Focusing on only one variable at the time allows for clear and simple representation and analysis of the results. GDP is the natural choice of variable for this analysis, as it is the most commonly forecast economic variable and usually the economic variable of most interest. Later in the next section, the forecasting performance of the model with respect to the other variables as well is assessed in depth.

Other feature streamlining the forecasting exercise discussed has to do with the hyperparameters λ_2 and λ_3 when studying the ML approach and the novel algorithm proposed. As discussed in the third section, for computational efficiency and ease of interpretation the hyperparameters λ_2 and λ_3 are fixed to $\lambda_1 \times 10$, as in Banbura et al. (2010). This reduces the dimen-

[7] i.e in the beginning of 2019Q1, $h = 1$ for *backcasting* 2018Q4 and $h = 4$ for forecasting 2019Q3.

sionality of the optimization problem at hand and allows for specification of the prior only in terms of its tightness. This is not an innocuous choice and this might come with deterioration of the forecasting performance.

Two metrics are reported to assess the forecasting accuracy of the model with different schemes for hyperparameter choice and with different number of variables. The names *Small, Medium* and *Large* refer to the models of different sizes presented in table 1.[8] The two metrics reported are the *Root Mean Squared Forecasting Error* (RMSE) and the so-called *Log-Score*. RMSE is perhaps the most commonly used metric in the literature for assessing the forecasting performance of a model. It measures the accuracy of the point-estimates produced by the model, giving smaller values for models producing smaller squared forecasting errors. Log-Score on the other hand is a metric for evaluating the accuracy of full predictive densities produced by a model. It is produced by averaging all the values of log-predictive marginal densities produced by the model at the realized values. Thus, larger the value, more accurate the predictive densities produced by the model.

To give some idea of the relative accuracy of the forecasts, results from two naive benchmark models are reported as well. First model is the random walk with drift. With random walk the forecast for the next period is always the same as the observation from the last period. Random walk also happens to be the prior of our model, if the additional dummy-priors are excluded. Random walk with drift is merely an extension of this, with a constant time-dependent factor that usually grows at a constant pace as the time passes. This is sensible as GDP is not stationary, but is expected to grow in the long run. The predictive densities for the random walk are produced by estimating a normally distributed error term of mean zero from the data. The other naive benchmark model is a simple univariate AR-model, with lag length chosen according to Akaike information criterion. For the AR-model the data is transformed to differences to obtain a stationary series and enhance the forecasting performance.

The results of the pseudo out-of-sample forecasting exercise are presented in table 2. In the table, there are four different BVARX-models. Sims-Zha

[8] *Small* = no exogenous variables at all and 7 variables in total, *Medium* = 12 variables in total, *Large* = 25 variables in total.

refers to the model in which the fixed hyperparmeter values from the literature are used, 0.2 for λ_1 and 1 for λ_2 and λ_3. Evidently, those fixed hyperparameters lead to a significant over-fitting and poor forecasting performance for every model expect the smallest one without the exogenous block, in which only slightly poorer results are obtained compared to the other methods. The conditional BVARX-model developed would seem to require much tighter priors and that the tightness of the prior should be specified separately for every sub-model included.

The abbreviation ML stands for the ML approach to the hyperparameter choice, and it would seem to yield comparable results to the novel algorithm proposed, atleast for h = 1. The ML approach would not seem to suffer too badly from the over-fitting related issues discussed above and it proves itself as a fairly successful option for hyperparameter choice in this occasion. Although, as expected, a closer inspection of the hyperparameters implied reveals the prior implied by the ML approach to be slightly looser than the prior implied by the novel algorithm.

Large	h = 1		h = 4	
	RMSE	Log-Score	RMSE	Log-Score
Random Walk	0.0069	3.2588	0.0215	2.3917
AR	0.0064	3.2948	0.0205	2.3362
BVARX (Sims-Zha)	0.0083	3.2719	0.0299	1.8523
BVARX (ML)	**0.0061**	**3.7036**	0.0166	2.6418
BVARX (A1)	0.0067	3.5890	0.0171	2.5515
BVARX (A4)	0.0063	3.6477	**0.0144**	**2.8299**

Medium	h = 1		h = 4	
	RMSE	Log-Score	RMSE	Log-Score
Random Walk	0.0069	3.2588	0.0215	2.3917
AR	0.0064	3.2948	0.0205	2.3362
BVARX (Sims-Zha)	0.0072	3.4838	0.0202	2.4401
BVARX (ML)	0.0060	3.6836	0.0158	2.7165
BVARX (A1)	**0.0054**	**3.7289**	0.0132	2.8895
BVARX (A4)	0.0055	3.7242	**0.0123**	**2.9298**

Small	h = 1		h = 4	
	RMSE	Log-Score	RMSE	Log-Score
Random Walk	0.0069	3.2588	0.0215	2.3917
AR	0.0064	3.2948	0.0205	2.3362
BVARX (Sims-Zha)	0.0065	3.6112	0.0166	2.6696
BVARX (ML)	0.0062	**3.6359**	0.0154	2.7433
BVARX (A1)	**0.0061**	3.6204	0.0153	2.7191
BVARX (A4)	0.0062	3.6118	**0.0129**	**2.8427**

Table 2: *The out-of-sample forecasting errors of natural logarithm of GDP from 2011Q4 to 2018Q3 for h = 1 and 2012Q3 to 2018Q3 for h = 4, h denoting the length of the forecasting horizon, with* ***out-of-sample*** *hyperparameter estimates. Smaller the root mean squared forecasting error (RMSE), more accurate the point-estimates of the model are, and larger the Log-Score, more accurate are the full predictive densities produced by the model. Descriptions of the models of different size can be found from table 1. The univariate benchmark-models (Random Walk and AR) are the same in each of the three tables as they are not dependent of the number of variables included in the model. Four BVARX-models of the table differ only in how their hyperparameters are chosen and those differences are elaborated in the text.*

For example, in the large BVARX-model and in its fourth BVAR-model requiring the tightest prior as all the variables are included, the optimal hyperparameter value λ_1 implied by the ML approach hovered from around 0.045 to 0.035 (see figure 4). The slow tightening of the prior as more data becomes available is very similar to what was observed in the Monte Carlo exercise above. The optimal hyperparameter value implied by the algorithm (A4) had some jumps in the beginning of the sample caused by the very limited amount of observations[9], but varied mostly between 0.02 and 0.03. These would be very small hyperparameter values for a traditional BVAR-model of this size[10], which implies that the modified prior structure imposed by the BVARX-model requires a fairly tight prior and further developments in the prior formulation could yield potential improvements in forecasting accuracy due to allowing for a looser prior. The need for a fairly tight prior could also be explained by the relative lack of quality of the predictive variables, which could partly explain why the largest model is not able to match the smaller ones in terms of forecasting accuracy.

In table 2, A1 stands for the novel algorithm approach in which the one step ahead out-of-sample forecasting errors are minimized in order to choose the hyperparameters, whereas in A4 the four steps (i.e. one year) ahead out-of-sample forecasting errors are minimized. The performance of both of these methods and the ML approach is very comparable for h = 1 and no significant differences in performance become evident. Especially for A1 and ML approach this is not surprising since both of those two methods are essentially based on the minimization of the one step ahead out-of-sample forecasting errors. A4 on the other hand is based on the minimization of the out-of-sample forecasting errors one year ahead and it still seems to perform atleast as well as the other methods when h = 1. When h = 4, A4 yields significantly better forecasting performance as the other methods for a model of any size.

Based on this assesment the A4 seems therefore to be the best suited

[9]only 8 observations (2009Q4-2011Q3) used for the minimization of the out-of-sample forecasting error for the first forecast of the sample.

[10]In Giannone et al. (2015) for a model of 22 variables, corresponding hyperparameter value implied by the ML approach is approximately 0.09. However that estimate is drawn towards a hyperprior centered at 0.2.

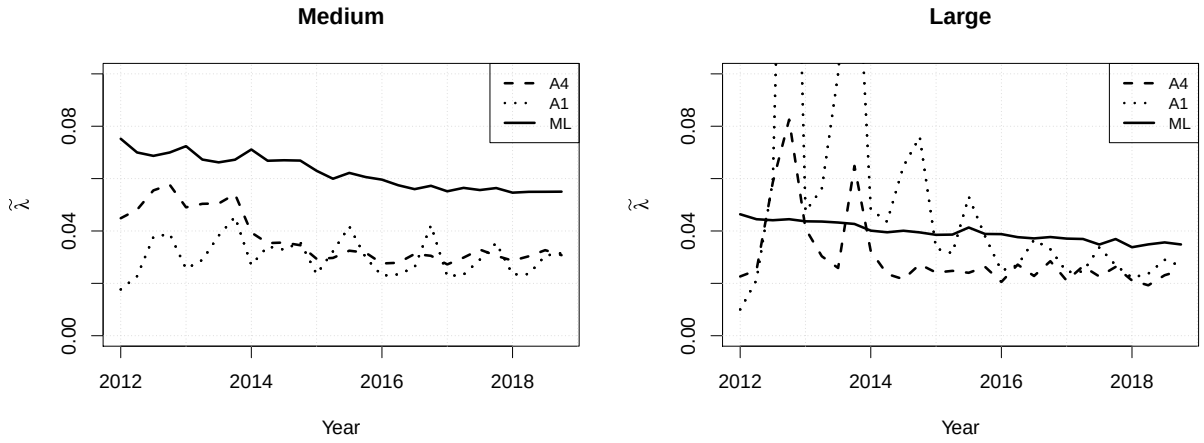

Figure 4: *Implied optimal hyperparameter values λ_1 for the fourth sub-model with three different methods for medium and large model from 2012Q1 to 2018Q4. Descriptions of the methods can be found in the text.*

method for medium-term forecasting with the conditional BVARX-model developed. Inspection of the implied hyperparameters reveals that the prior implied by A4 varies less from period to period than the one implied by A1. The superior performance of A4 may thus atleast partly be attributed to this feature. Minimization of the four steps ahead out-of-sample forecasting error would seem to be more robust against the misspecification of hyperparameters than the one step ahead approach of A1 and ML.

Figure 4 illustrates the optimal hyperparameter values implied by the three methods discussed, over different points in time as more data becomes available. The hyperparameter values implied by the ML approach stay much more stable over time since the relative amount of data available for the method alters much less over the time interval. With the ML approach the whole data beginning from the year 1999 can be used *in a sense* to minimize the out-of-sample forecasting error. Thus, the length of the data available for the method grows only approximately by 70 percent. With A1 and A4 the model must be estimated to produce the out-of-sample errors to minimize, which restricts the amount of data to be used in the minimization process to start from 2009Q4 and the length of the data available for the method grows approximately by 500 percent over the time interval of the

exercise.

Given this, especially the hyperparameter values implied by A4 seem to converge surprisingly fast towards a stable perimeter to be hovered around from approximately 2014Q2 onward. Although the values implied by the ML approach vary less, the values decrease in a very persistent manner as was also observed with the Monte Carlo experiment. The values from the last period implied by A1 and A4 also naturally coincide with the values minimizing the out-of-sample forecasting errors over the entire sample from 2011Q4 to 2018Q3. The hyperparameter values implied by the ML approach are significantly larger at the end of the sample and can thus be deemed to cause the model to over-fit the data, atleast based on this sample of seven years. Therefore, the almost comparable accuracy of the ML method to the approach with the novel algorithm seems to stem, atleast to a some extent, from the more stable behavior of the implied hyperparameter values especially in the beginning of the sample.

Large	λ_1			
Method	Exo (M)	Endo (M)	Exo (Q)	Endo (Q)
ML	0.121	0.158	0.185	0.035
A1	0.010	0.184	0.090	0.027
A4	0.098	0.163	0.010	0.025

Medium	λ_1			
Method	Exo (M)	Endo (M)	Exo (Q)	Endo (Q)
ML	0.161	0.173	0.220	0.055
A1	0.505	1.000	0.083	0.032
A4	0.130	0.447	0.139	0.031

Small	λ_1			
Method	Exo (M)	Endo (M)	Exo (Q)	Endo (Q)
ML	-	0.324	-	0.231
A1	-	1.000	-	0.174
A4	-	1.000	-	0.055

Table 3: *Optimal hyperparameter values at 2018Q4 implied by three different methods.*

Table 3 presents the optimal hyperparameters at the end of the sample, implied by three different methods. The lower and upper bounds for the hyperparameters are set to 0.01 and 1, respectively, and the smallest model has no exogenous variables at all and therefore possesses only two hyperparameters controlling the shrinkage. Not all the hyperparameters are of equal importance and it is the last column of table 3 that deserves attention the most, since it controls the shrinkage of the final sub-model including all the variables and producing the final forecast. Focusing on the last column, it seems that the ML method changes the most when moving from the 12 variable model (medium) to the 25 variable model (large). This reflects the fact that the ML method pursues to minimize the out-of-sample forecasting error of *every* variable in the model, whereas the other two methods are specified to be concerned only of the variable of interest. Also, since the hyperparameter values implied by the algorithm (A1 and A4) are less affected by the addition of the variables, it would seem that not all the variables added had much impact on the predictions regarding the variable of interest.

Table 4 presents the same forecasting errors as table 2, only this time the hyperparameter values from table 3 are used for the entire length of the forecastin exercise. In other words, table 4 presents the out-of-sample forecasting errors using the *in-sample* estimates of the hyperparameters. As expected, the forecasts are slightly more accurate with in-sample estimates of the hyperparameters, however the difference to the errors reported in 2 is not huge. Interestingly, now the A1 seems to produce the most accurate forecasts for both horizons and for a model of any size. This implies that the minimization of the one step ahead out-of-sample forecasting error instead of the four step approach is more sensitive to the amount of data used for estimation, but that after the hyperparameter values have converged in some sense, the one step ahead approach becomes as valid as, or even superior to, the four step approach.

Also, the in-sample estimates of the hyperparameters improve the relative performance of the large model and to the same extent worsen the relative performance of the small model. This implies that a great extent of the some times observed relatively inferior performance of the larger BVAR-models could in general be accounted to their greater sensitivity to the hyperparam-

eter choice controlling the shrinkage and to the insufficient way of choosing the hyperparameters.

Overall, the medium sized model with 12 variables seems to be outperforming the other model specifications. According to the out-of-sample forecasting errors with the in-sample hyperparameter estimates, both the medium and the large model also seem to outperform the small model, which suggests that the exogenous variables of the model are indeed useful, atleast in forecasting the gross domestic product.

The medium sized model with the hyperparameters chosen by the novel algorithm minimizing the one step ahead out-of-sample forecasting error of the gross domestic product seems to be yielding more accurate forecasts than the other models. Therefore, the medium sized model and the corresponding hyperparameter values (A1 and ML) from table 3 are used in the next section of the thesis, where the forecasting accuracy of the model is empirically assessed in more depth.

Large	h = 1		h = 4	
	RMSE	Log-Score	RMSE	Log-Score
Random Walk	0.0069	3.2588	0.0215	2.3917
AR	0.0064	3.2948	0.0205	2.3362
BVARX (Sims-Zha)	0.0083	3.2719	0.0299	1.8523
BVARX (ML)	0.0062	3.6625	0.0154	2.7309
BVARX (A1)	**0.0058**	**3.7407**	0.0131	**2.9163**
BVARX (A4)	0.0062	3.6912	**0.0130**	2.9097

Medium	h = 1		h = 4	
	RMSE	Log-Score	RMSE	Log-Score
Random Walk	0.0069	3.2588	0.0215	2.3917
AR	0.0064	3.2948	0.0205	2.3362
BVARX (Sims-Zha)	0.0072	3.4838	0.0202	2.4401
BVARX (ML)	0.0057	3.7339	0.0146	2.8026
BVARX (A1)	**0.0053**	**3.7573**	**0.0115**	**2.9838**
BVARX (A4)	0.0055	3.7374	0.0122	2.9384

Small	h = 1		h = 4	
	RMSE	Log-Score	RMSE	Log-Score
Random Walk	0.0069	3.2588	0.0215	2.3917
AR	0.0064	3.2948	0.0205	2.3362
BVARX (Sims-Zha)	0.0065	3.6112	0.0166	2.6696
BVARX (ML)	0.0061	**3.6482**	0.0153	2.7439
BVARX (A1)	**0.0060**	3.6326	**0.0142**	**2.7860**
BVARX (A4)	0.0062	3.6323	0.0154	2.7261

Table 4: *The out-of-sample forecasting errors of natural logarithm of GDP from 2011Q4 to 2018Q3 for h = 1 and 2012Q3 to 2018Q3 for h = 4, h denoting the length of the forecasting horizon, with* ***in-sample*** *hyperparameter estimates. Smaller the root mean squared forecasting error (RMSE), more accurate the point-estimates of the model are, and larger the Log-Score, more accurate are the full predictive densities produced by the model. Descriptions of the models of different size can be found from table 1. The univariate benchmark-models (Random Walk and AR) are the same in each of the three tables as they are not dependent of the number of variables included in the model. Four BVARX-models of the table differ only in how their hyperparameters are chosen and those differences are elaborated in the text.*

6 Forecasting accuracy

In this section, the forecasting accuracy of the BVARX-model developed is assessed empirically. The capability of the model to forecast five different economic variables in Finland is studied. Those five variables include, gross domestic product (GDP), exports of goods and services (Exports), imports of goods and services (Imports), inflation and unemployment. Forecasting horizons assessed are one-step-ahead ($h = 1$), one-year-ahead ($h = 4$) and two-years-ahead ($h = 8$).

The medium sized model and two sets of hyperparameter values are included in the assessment of this section. A1 refers to the hyperparameter values deemed most successful in the previous section. These hyperparameter values were however chosen to specifically minimize the out-of-sample forecasting errors of the GDP forecasts while the maximum likelihood based method pursues to minimize the out-of-sample forecasting errors of the whole model. Thus, the hyperparameters chosen that way (ML) could outperform the A1 values with respect to other variables than GDP. Therefore, forecasts produced with both sets of hyperparameters, namely A1 and ML, are assessed (for hyperparameter values of sets A1 and ML see table 3).

For benchmarking the same naive univariate models from the previous section are used. Namely, the univariate autoregressive model (AR) with lag length chosen according to the Akaike information criterion and the random walk with drift[11] (Random Walk).

Table 5 presents the forecasting errors and log-score measures for the forecasts of GDP, exports and imports. As already discussed in the previous section, the BVARX-model seems to outperform the univariate models by a wide margin when forecasting the GDP. With exports and imports the performance of the nowcasts ($h = 1$) produced by the univariate models is comparable to those produced by the BVARX-model.

[11]Drift coefficient is set to zero for stationary variables, i.e. for inflation and unemployment.

GDP	h = 1		h = 4		h = 8	
	RMSE	Log-Score	RMSE	Log-Score	RMSE	Log-Score
Random Walk	0.0069	3.2588	0.0215	2.3917	0.0370	1.8984
AR	0.0064	3.2948	0.0205	2.3362	0.0358	0.2197
BVARX (ML)	0.0057	3.7339	0.0146	2.8026	0.0163	2.5284
BVARX (A1)	**0.0053**	**3.7573**	**0.0115**	**2.9838**	**0.0127**	**2.6464**

Exports	h = 1		h = 4		h = 8	
	RMSE	Log-Score	RMSE	Log-Score	RMSE	Log-Score
Random Walk	0.0303	1.8153	0.0408	1.1910	0.0701	0.8068
AR	**0.0283**	1.8448	0.0408	**1.6925**	0.0702	**1.2272**
BVARX (ML)	0.0338	1.9621	0.0483	1.4497	0.0542	1.0703
BVARX (A1)	0.0333	**1.9897**	**0.0379**	1.5446	**0.0424**	1.1629

Imports	h = 1		h = 4		h = 8	
	RMSE	Log-Score	RMSE	Log-Score	RMSE	Log-Score
Random Walk	**0.0285**	**2.1252**	0.0435	1.5877	0.0672	1.1931
AR	0.0326	1.9942	0.0486	1.4053	0.0699	0.2176
BVARX (ML)	0.0316	1.4931	0.0459	1.6371	0.0594	1.3906
BVARX (A1)	0.0307	1.6208	**0.0389**	**1.8264**	**0.0464**	**1.5741**

Table 5: *The accuracy of the out-of-sample forecasts for the natural logarithm of gross domestic product (GDP), exports of goods and services (Exports) and imports of goods and services (Imports) from 2011Q4 to 2018Q3 for h = 1, 2012Q3 to 2018Q3 for h = 4 and 2013Q3 to 2018Q3 for h = 8. Letter h refers to the forecasting horizon and descriptions of the models are in the text. Root mean squared forecasting error (RMSE) measures the accuracy of the point estimates while Log-Score measures the overall performance of the marginal predictive distribution.*

However, for longer forecasting horizons, atleast the other BVARX-model (A1) seems to outperform the the univariate benchmark-models. The inferior log-score measure of the exports forecasts produced by the BVARX-model compared with the one produced by the AR-model is somewhat surprising, since the RMSE of the BVARX-model implies much more accurate forecasts especially for $h = 8$ (0.0424 against 0.0702). This is probably caused by the sensitivity of the log-score measure for misspecification of the tails of the predictive distribution. Normality assumption of the error terms in the model can cause the approximately t-distributed[12] predictive distribution to have thinner tails than the true data generating process. Consequently, the near zero probability mass further in the tails causes the occasional observations with large deviations from the mean (or mode) to affect the mean log-score in a disproportional manner.

Although the set A1 of hyperparameters were chosen only to minimize the out-of-sample forecasting error of the GDP, the set still seems to outperform the ML set of hyperparameters for other variables as well. This implies that the minimization of the forecasting errors of only one variable could produce sufficient hyperparameter estimates for the analysis of also other variables in the same model.

The inflation and unemployment need to be assessed separately from the three variables above to make the comparison to univariate models meaningful. In the above pseudo out-of-sample forecasting studies in this thesis the forecasts are assumed to be performed at the end of the respecting quarter. As an example, the one-step-ahead predictions for 2018Q4 would be established using the data available at the end of the year 2018 (i.e. nowcast). The last data points available for GDP, exports and imports would be from 2018Q3 and thus the comparison of the multi- and univariate forecasts is meaningful. However, the monthly variables (e.g. inflation and unemployment) have shorter publication lags and at the end of the year the BVARX-model would use the information on these variables from October and November as well. This same monthly information regarding the variables is not however available for the univariate models which would invali-

[12]The predictive distribution is conditioned and is therefore not exactly t-distributed as in the case of an unconditional BVAR-model.

date the comparison of the models. Therefore, when assessing the forecasting performance of the inflation and unemployment the pseudo out-of-sample forecasting study is performed assuming that the forecasts are established at the end of the first month of the respective quarter. This way both the multi- and univariate models have the same information set at disposal regarding the variable of interest. Alternatively, the univariate models could have been specified for the monthly series and aggregated only after the forecasts had been established. This could however affect the accuracy of the univariate forecasts potentially unfavorably, especially with longer forecasting horizons.

Table 6 presents the results of this pseudo out-of-sample forecasting exercise. For inflation the BVARX-model seems to be outperforming the univariate models by a wide margin and the A1 set of hyperparameter values seems to yield more accurate forecasts for every forecasting horizon than the ML set of hyperparameter values.

Inflation	h = 1		h = 4		h = 8	
	RMSE	Log-Score	RMSE	Log-Score	RMSE	Log-Score
Random Walk	0.3276	-0.5656	0.8578	-1.3635	1.4344	-1.8019
AR	0.3194	-0.4110	0.9747	-1.7810	1.6440	-3.7035
BVARX (ML)	0.1498	0.4516	0.4021	-0.5313	1.1044	-1.6505
BVARX (A1)	**0.1491**	**0.4578**	**0.3661**	**-0.4059**	**1.0510**	**-1.4966**

Unemployment	h = 1		h = 4		h = 8	
	RMSE	Log-Score	RMSE	Log-Score	RMSE	Log-Score
Random Walk	0.2592	-0.0818	0.6587	-1.0795	0.9346	**-1.4118**
AR	0.2654	-0.1085	0.7566	-3.6672	1.1408	-8.3756
BVARX (ML)	**0.1704**	**0.3523**	0.3712	-1.0526	0.8183	-2.1614
BVARX (A1)	0.1779	0.0446	**0.3455**	**-0.9431**	**0.7007**	-1.7195

Table 6: *The accuracy of the out-of-sample forecasts for the inflation and unemployment from 2011Q4 to 2018Q3 for h = 1, 2012Q3 to 2018Q3 for h = 4 and 2013Q3 to 2018Q3 for h = 8. Letter h refers to the forecasting horizon and descriptions of the models are in the text. Root mean squared forecasting error (RMSE) measures the accuracy of the point estimates while Log-Score measures the overall performance of the marginal predictive distribution.*

For unemployment as well, the BVARX-model seems to be producing more accurate forecasts than the univariate benchmark-models, and the ML set of hyperparameter values slightly outperforms the A1 set for one-step-ahead forecasts, while A1 yet again produces the most accurate forecasts for longer forecasting horizons.

The log-score measures regarding the unemployment forecasts seem to suffer from a fairly unstable behavior and the measure suggests a different model for every forecasting horizon assessed. This is at odds with the RMSE measures, which imply that the BVARX-models are clearly outperforming the univariate models. The reason for the unstable behavior of the log-score measures probably stems from the same issue of the sensitivity of the measure for misspecification of the tails of the predictive distribution, briefly discussed earlier. This could imply that particularly unemployment forecasting requires accounting for more probability mass in the tails of the predictive distribution, or simply that there just happened to be more extreme values than usual in the short sample of unemployment figures used for this study.

Overall, the medium sized BVARX-model with 12 variables and hyperparameters chosen according to the novel algorithm presented in the previous section seems to provide the most accurate forecasts for every variable concerned. Not only does the multivariate BVARX-model with conditional features and exogenous variables seem to perform remarkably better than the univariate benchmarking models, but the proper choice of hyperparameter values seems to provide further considerable yields over the model with less suitable set of hyperparameter values.

The empirical assessment of this section provides evidence in favor of the modeling choices presented in this thesis. However, due to data limitations, the time interval of the pseudo out-of-sample forecasting exercises performed in this thesis is fairly short[13], which should be borne in mind when making conclusions based on these results. In terms of future research, the applicability of the novel algorithm for hyperparameter choice presented in this thesis could also be tested against alternative approaches with different datasets.

[13] 21 ($h = 8$) to 28 ($h = 1$) quarterly observations

6.1 Conditional forecasts

Conditional framework of the model developed in this thesis allows for utilizing all the information from the latest statistical publications in an efficient way, but as discussed it also allows for conditioning of the forecasts on predictive densities. By conditioning the forecast of one variable on predictions regarding the evolution of other variables, the forecast may be enhanced, but only if the additional information acquired is relevant. The relevance of this information depends on whether (i) the changes in the variables to be conditioned are associated with changes in the variable to be forecast, directly or indirectly, according to the model, and whether (ii) the predictions to be conditioned on differ from those produced by the unconditional model.

In this subsection, the accuracy of the forecasts conditioned on the information on future values of the exogenous variables is compared to the forecasts produced in the previous section. Specifically, the set A1 of hyperparameter values and the medium sized model are used and the variables of interest are those from table 5 (i.e GDP, exports and imports). The forecasting exercise is performed as the one reported in table 5, but this time the forecaster is assumed to have a perfect foresight on the future values of the exogenous variables, up to third quarter of 2018 of course, which is the last observation in the dataset of quarterly frequency. First, the perfect foresight was assumed on only the future values of trade market GDP (Conditional) and then on all the exogenous variables (Conditional - all). Table 7 presents the results of this exercise.

The conditioning of the forecasts on future values of exogenous variables did not seem to yield any significant improvements to the forecasts produced without this information. Therefore, it can be concluded that conditioning of the forecasts on predictions of future values of exogenous variables would not have improved the forecasts during the time interval of our exercise either. However, the number of observations used in the exercise is very small and the subtle differences in forecasting performance, if any, might become evident with a different dataset or with more data.

As the medium sized model studied was previously observed to produce more accurate GDP forecasts than the otherwise identical smaller model *with-*

GDP	h = 1		h = 4		h = 8	
	RMSE	Log-Score	RMSE	Log-Score	RMSE	Log-Score
Unconditional	0.0053	3.7573	**0.0115**	**2.9838**	0.0127	2.6464
Conditional	0.0054	3.7516	0.0116	2.9718	0.0127	2.6510
Conditional - all	**0.0052**	**3.7583**	0.0117	2.9703	**0.0125**	**2.6558**

Exports	h = 1		h = 4		h = 8	
	RMSE	Log-Score	RMSE	Log-Score	RMSE	Log-Score
Unconditional	0.0333	1.9897	**0.0379**	1.5446	0.0424	1.1629
Conditional	0.0336	**1.9923**	0.0401	1.5321	0.0440	1.1486
Conditional - all	**0.0332**	1.9826	0.0380	**1.5467**	**0.0394**	**1.1782**

Imports	h = 1		h = 4		h = 8	
	RMSE	Log-Score	RMSE	Log-Score	RMSE	Log-Score
Unconditional	**0.0307**	**1.6208**	**0.0389**	**1.8264**	**0.0464**	1.5741
Conditional	0.0308	1.6128	0.0392	1.8163	0.0471	1.5750
Conditional - all	0.0310	1.5451	0.0395	1.8084	0.0467	**1.5817**

Table 7: *The accuracy of the conditional and unconditional out-of-sample forecasts for the natural logarithm of gross domestic product (GDP), exports of goods and services (Exports) and imports of goods and services (Imports) from 2011Q4 to 2018Q3 for h = 1, 2012Q3 to 2018Q3 for h = 4 and 2013Q3 to 2018Q3 for h = 8. 'Unconditional' refers to the forecasts for which no information regarding the future values of exogenous variables was used, whereas 'Conditional' refers to the forecasts where information on the future values of trade market GDP was used. In 'Conditional - all' information on future values of all the exogenous variables was used. Letter h refers to the forecasting horizon.*

out the exogenous variables, the exogenous variables should be directly or indirectly associated with changes in the GDP. Therefore, the first relevance-condition of the information conditioned on (laid out above) should be satisfied. Hence, the results imply that the second relevance-condition must have been violated. The realized future values of the exogenous variables did not seem to differ enough from those predicted by the model. It is possible, that conditioning on future values significantly improves the forecast only in a case of highly surprising information. This would be in line with the results in Bloor & Matheson (2011) where they found conditioning on the sharp

and surprising rise in commodity prices over the year 2007 to significantly improve the forecasts on those periods.

However, conditioning on some variables could turn out to be more useful than the others. Many economic variables can be forecast to a sufficient degree with reasonably simple models and when they cannot, the changes are often fundamentally unpredictable. Conditioning on future values of these kind of variables may not yield significant improvements to the forecasts as no relevant information, that would not already be included in the model, on future values of these variables is usually available.

On the other hand, there are some variables for which there may be plenty information available on the expectations of their future values, that cannot directly be derived from the past data. For example, the policy rates controlled by the central banks have significant effects on the economy and the central banks seek to communicate the future policy rate changes as accurately as possible. This kind of information could be conveniently included in the forecast with the conditioning framework presented in this thesis. Future values of policy rates, or any other interest rates for that matter, should also usually be restricted to account for any possible lower bounds[14] on these rates, which makes the addition of the conditioning framework presented to any model including policy rates even more appealing. However, the assessment of the applicability of this conditioning framework to the models with policy rates falls beyond the scope of this thesis and is left for future research.

No strong conclusions regarding the usefulness of the conditioning on future predictions to acquire more accurate forecasts can be made on the basis of the results presented in this section.

7 Discussion

Although in the previous section, the conditioning on the future values of exogenous variables was not found to yield significant improvements to the forecasts of endogenous variables, the conditioning framework studied in this thesis offers several possibilities for future research. The results obtained

[14] e.g. zero (or near zero) lower bound of policy rates.

in this thesis were not highly surprising, since the possible improvements to the forecasting accuracy were expected to be subtle and the number of observations in the assessment of this thesis is fairly small. Similar forecasting exercise could be repeated for a number of different datasets to attain better understanding of the usefulness of the conditioning approach.

However, the conditioning framework studied allows for a great number of versatile applications beyond the studied conditioning on future values of exogenous variables. As an example, the conditioning framework could be used to impose restrictions to ensure the consistence of forecasts with some known or observed properties of economic variables, such as lower bounds of policy rates and accounting identities.

Another, more unorthodox application of the conditioning framework has to do with the predictive densities of forecasts based on subjective considerations. As official forecast published by different economic institutes are usually based on subjective considerations, it is difficult to produce sensible estimates of uncertainty around the published point-estimates. With the model developed in this thesis, the mean of the posterior predictive distribution could be restricted to follow a pre-specified path (i.e. the official point-estimates based on subjective considerations), while not restricting the estimation of higher moments of the predictive distribution. In other words, with the above portrayed approach it is possible to produce full predictive densities for the forecasts based on subjective considerations.

Predictive densities produced this way should however be dealt with caution, as the interpretation of these densities as *true* predictive densities of the forecasts based on subjective considerations would require reasonably strong assumptions. For one, given the forecasts obtained with the help of subjective considerations were to be more accurate on average than the ones produced by the data-driven model, the predictive densities produced by the model would be incorrigibly too wide. However, if the data-driven model can be expected on average to produce approximately as accurate predictions, then the predictive densities obtained this way could very well be used to communicate the uncertainty revolving the forecast, or even to answer more subtle questions than what the point-estimates are capable of, such as *what is the probability of the economic growth in Finland to exceed a certain*

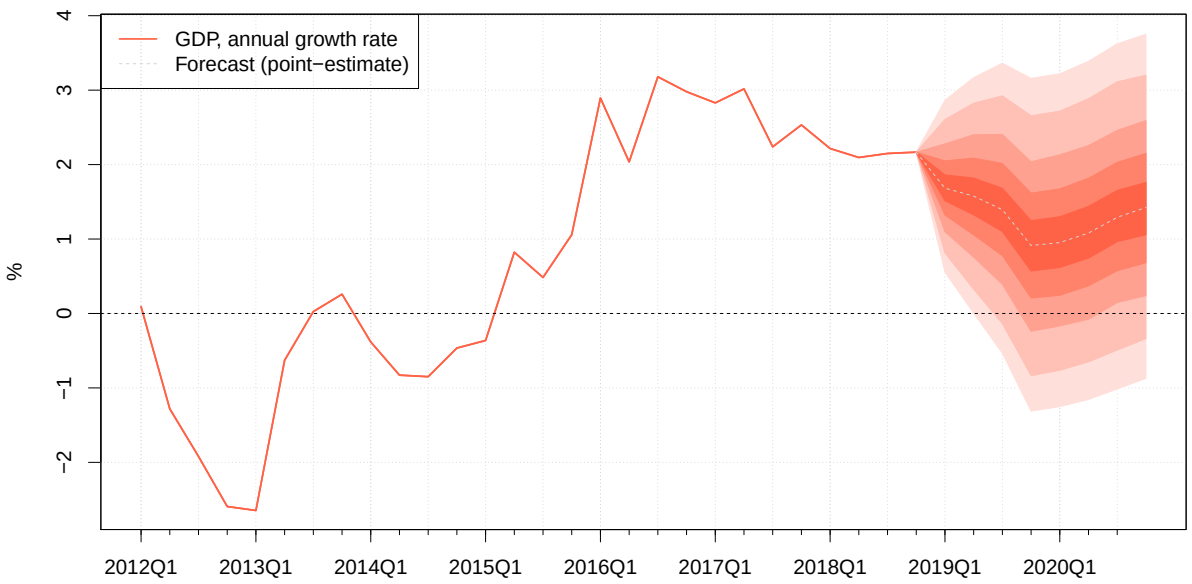

Figure 5: *Fan plot of a forecast for annual growth rate of the Finnish economy produced by the BVARX model, that is restricted to produce the same point-estimates as the official subjectively adjusted forecast published by the Research Institute of the Finnish Economy (Etla). The outermost lines of the fan portray the 90 percent credible interval of the forecast, while the innermost area covers 20 percent of the probability mass of the posterior predictive distribution.*

growth rate on a given period?.

In addition to the conditional forecasting framework, another feature thoroughly assessed in this thesis is the hyperparameter choice. The hyperparameter choice is of the essence in Bayesian modelling and the main driver of the forecasting accuracy of BVAR forecasting models. However, the issue has been left with not enough attention in many of the previous studies concerning BVAR modelling, which may have caused the forecasting accuracy of BVAR models to have been underestimated. For instance, the seminal paper of Banbura et al. (2010) on large BVAR models is often used as a reference when comparing the forecasting performance of BVAR and dynamic factor

models. However, in Banbura et al. (2010) the hyperparameters are chosen in an extremely non-rigorous way, by fixing the in-sample fit of the model to that of a model with three variables. Later Giannone et al. (2015) have shown the hyperparameters chosen that way to perform in a very suboptimal manner. The approach for hyperparameter choice presented in Giannone et al. (2015) can arguably be reckoned as the state-of-the-art approach, yet as shown in this thesis, the approach has its deficiencies as well. The novel, reasonably simple, approach for hyperparameter choice presented in the fifth section was shown to generate sizeable improvements to the forecasting accuracy of the model compared to any other approach for hyperparameter choice. This result suggests that there may be plenty of space to improve the performance of BVAR models in general, via more careful selection of the hyperparameter values.

Although the proposed novel approach resulted in the most accurate forecasts, the marginal likelihood based approach for hyperparameter choice was also observed to lead to a reasonable forecasting accuracy and the compact expression for the unnormalized marginal likelihood function derived in the Appendix B can be used conveniently for hyperparameter choice. However, these hyperparameter estimates can be expected to lead to a over-fitting of some degree and the addition of a hyperprior distribution to drag the parameter estimates towards the prior is recommended.

Due to considerable uncertainty in medium term economic forecasting, simple univariate autoregressive and random walk models have often been shown to produce forecasts on a par with, or even beyond, the forecasting accuracy of the more complex models. However, with a suitable prior and optimally chosen hyperparameter values, a BVAR model should in principle always be able to produce forecasts *at least* as accurate as the simpler model. Model combination point-of-view may be used to illuminate the argument, since a BVAR model can be constructed as a combination of a simpler model (e.g. random walk prior) and the underlying VAR model. Therefore, with optimally chosen hyperparameter values, the simpler model should match the performance of the BVAR model only when the optimal hyperparameter values would shrink the variance of the prior distribution to zero. The empirical assessment of the fifth and sixth section also suggest that the predictive

densities produced by the BVARX model are far more accurate than the ones produced by the univariate benchmark models.

All things considered, the BVARX forecasting model developed in this thesis appears to be a highly suitable choice for medium term forecasting of a small open economy. Further improvements to the model could include non-fixed hyperparameter values λ_2 and λ_3 controlling the strength of the additional priors (i.e. $10 \times \lambda_1 \neq \lambda_2 \neq \lambda_3$) and the addition and conditioning of policy rates as discussed above.

In terms of future research, the possibilities of conditioning on future values of predictive variables could be studied further, since no strong conclusions regarding the matter can be drawn from the very limited empirical assessment of this thesis. Another interesting topic left for future research is the *proper* formulation of a hyperprior distribution, when the hyperparameters of the model are chosen according to the marginal likelihood function. With a proper formulation of the hyperprior distribution, the marginal likelihood function could possibly be used to choose the hyperparameters more efficiently than with any other method, and in a computationally highly feasible manner.

8 Conclusions

In this thesis, a conditional BVARX forecasting model for short and medium term forecasting of small open economies is developed. The proposed model offers a framework to deal with several practical issues of data-driven economic forecasting. The conditioning framework of the model (i) allows for efficient incorporation of time series of different frequencies, (ii) provides a way to deal with the so-called ragged edge of the multivariate time series data by conditioning the forecasts on the latest observations and (iii) allows for imposing versatile restrictions on the forecasts by conditioning on marginal predictive densities, or linear combinations of them, of any variable in the model. The model is also especially well suited for modelling small open economies as it allows for (iv) imposing exogeneity on the global economic variables.

Finally, as the most important single feature of the model affecting the

forecasting accuracy, (v) a novel approach for hyperparameter choice is proposed and shown to lead to a more accurate forecasts than any of the alternative approaches tested, when forecasting five economic variables in Finland from the last quarter of 2011 to the third of 2018. All the features of the model should be easily generalizable for a wide range of BVAR models as a solution for many issues of practical importance, providing more accurate forecasts.

The marginal likelihood based approach for the hyperparameter choice of a BVAR model is also shown to be prone to cause the model to over-fit the data, if the hyperprior distribution is not carefully selected. The proper formulation of the hyperprior distribution is however left for future research.

The accuracy of the model is assessed via pseudo out-of-sample forecasting exercises. The model is shown to outperform the univariate benchmark models by a wide margin for all the five economic variables tested. All things considered, the model developed in this thesis provides practical and effective data-driven tools for economic forecasting and could be used by economic forecasters from nowcasting to medium term forecasting either independently or collectively with other models and subjective considerations.

References

Amisano, G. & Geweke, J. (2013), Prediction Using Several Macroeconomic Models, Working Paper 1537, European Central Bank (ECB).

Andersson, M. K., Palmqvist, S. & Waggoner, D. F. (2010), Density-Conditional Forecasts in Dynamic Multivariate Models, Working paper series 247, Sveriges Riksbank (Central Bank of Sweden).

Anttonen, J. (2018), Nowcasting the Unemployment Rate in the EU with Seasonal BVAR and Google Search Data, Working paper series 62, Research Institute of the Finnish Economy (Etla).

Athans, M. (1974), 'The importance of Kalman filtering methods for economic systems', *Annals of Economic and Social Measurement* **3**, 49–64.

Banbura, M., Giannone, D. & Lenza, M. (2015), 'Conditional forecasts and scenario analysis with vector autoregressions for large cross-sections', *International Journal of Forecasting* **31**, 739–756.

Banbura, M., Giannone, D. & Reichlin, L. (2010), 'Large bayesian vector auto regressions', *Journal of Applied Econometrics* **25**, 71–92.

Bates, J. M. & Granger, C. W. J. (1969), 'The Combination of Forecasts', *Operational Research Quarterly* **20**, 451–468.

Bloor, C. & Matheson, T. (2011), 'Real-time conditional forecasts with Bayesian VARs: An application to New Zealand', *The North American Journal of Economics and Finance* **22**(1), 26–42.

Bok, B. . D. C., Giannone, D., Sbordone, A. M. & Tambalotti, A. (2017), Macroeconomic nowcasting and forecasting with big data, Staff Reports 830, Federal Reserve Bank of New York.

Burlon, L., Emiliozzi, S., Notarpietro, A. & Pisani, M. (2015), Medium-Term Forecasting of Euro-Area Macroeconomic Variables with DSGE and BVARX Models, Occasional Paper 257, Bank of Italy.

Byrd, R. H., Lu, P., Nocedal, J. & Zhu, C. (1995), 'A Limited Memory Algorithm for Bound Constrained Optimization', *SIAM Journal on Scientific Computing* **16**(5), 1190–1208.

Cogley, T., Morozov, S. & Sargent, T. J. (2005), 'Bayesian fan charts for U.K. inflation: Forecasting and sources of uncertainty in an evolving monetary system', *Journal of Economic Dynamics and Control* **29**(11), 1893–1925.

Cuaresma, J. C., Feldkircher, M. & Huber, F. (2016), 'Forecasting with Global Vector Autoregressive Models: a Bayesian Approach', *Journal of Applied Econometrics* **31**, 1371–1391.

Doan, T., Litterman, R. B. & Sims, C. A. (1984), 'Forecasting and Conditional Projection Using Realistic Prior Distributions', *Econometric Reviews* **3**(1), 1–100.

Dovern, J., Feldkircher, M. & Huber, F. (2016), 'Does joint modelling of the world economy pay off? Evaluating global forecasts from a Bayesian GVAR', *Journal of Economic Dynamics and Control* **70**, 86–100.

Fawcett, N., Kapetanios, G., Mitchell, J. & Price, S. (2014), Generalised Density Forecast Combinations, Working Paper 492, Bank of England.

Giannone, D., Lenza, M. & Primiceri, G. E. (2015), 'Prior Selection for Vector Autoregressions', *The Review of Economics and Statistics* **97(2)**, 436–451.

Itkonen, J. & Juvonen, P. (2017), 'Nowcasting the Finnish economy with a large Bayesian vector autoregressive model', *BoF Economics Review* .

Jarocinski, M. (2010), 'Conditional forecasts and uncertainty about forecast revisions in vector autoregressions', *Economics Letters* **108**, 257–259.

Kadiyala, K. R. & Karlsson, S. (1993), 'Forecasting with Generalized Bayesian Vector Autoregressions', *Journal of Forecasting* **12**(3-4), 365–378.

Kadiyala, K. R. & Karlsson, S. (1997), 'Numerical Methods for Estimation and Inference in Bayesian VAR-Models', *Journal of Applied Econometrics* **12**(2), 99–132.

Kalman, R. E. (1960), 'A New Approach to Linear Filtering and Prediction Problems', *Journal of Basic Engineering* **82**, 35–45.

Karlsson, S. (2012), Forecasting with Bayesian Vector Autoregressions, Working Papers 12, Örebro University, School of Business.

Kilian, L. & Lütkepohl, H. (2017), *Structural Vector Autoregressive Analysis*, Themes in Modern Econometrics, Cambridge University Press.

Kilponen, J., Orjasniemi, S., Ripatti, A. & Verona, F. (2016), The Aino 2.0 model, Research Discussion Papers 16, Bank of Finland.

Koop, G. & Koroblis, D. (2012), 'Forecasting Inflation Using Dynamic Model Averaging ', *International Economic Review* **53**, 867–886.

Koop, G. M. (2013), 'Forecasting with Medium and Large Bayesian VARS', *Journal of Applied Econometrics* **28**(2), 177–203.

Lehmus, M. (2018), 'ETLA macro model for forecasting and policy simulations', *Economic Modelling* **74**, 142–166.

Litterman, R. (1979), Techniques of forecasting using vector autoregressions, Working Papers 115, Federal Reserve Bank of Minneapolis.

Litterman, R. (1980), A Bayesian Procedure for Forecasting with Vector Autoregression, Working papers, Massachusetts Institute of Technology.

Litterman, R. (1986), 'Forecasting with Bayesian Vector Autoregressions-Five Years of Experience', *Journal of Business & Economic Statistics* **4**(1), 25–38.

McCracken, M. W., Owyang, M. T. & Sekhposyan, T. (2015), Real-Time Forecasting with a Large, Mixed Frequency, Bayesian VAR, Reserve bank of st. louis working paper series.

R Core Team (2018), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Robertson, J. C., Tallman, E. W. & Whiteman, C. H. (2005), 'Forecasting Using Relative Entropy', *Journal of Money, Credit and Banking* **37**(3), 383–401.

Schorfheide, F. & Song, D. (2015), 'Real-Time Forecasting With a Mixed-Frequency VAR', *Journal of Business I& Economic Statistics* **33**, 366–380.

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

Sims, C. (1993), A Nine-Variable Probabilistic Macroeconomic Forecasting Model, *in* 'Business Cycles, Indicators and Forecasting', National Bureau of Economic Research, Inc, pp. 179–212.

Sims, C. & Zha, T. (1998), 'Bayesian Methods for Dynamic Multivariate Models', *International Economic Review* **39**(4), 949–68.

Smets, F. & Wouters, R. (2007), 'Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach', *American Economic Review* **97**(3), 586–606.

Stelmasiak, D. & Szafranski, G. (2016), 'Forecasting the Polish Inflation Using Bayesian VAR Models with Seasonality', *Central European Journal of Economic Modelling and Econometrics* **8**(1), 21–42.

Stock, J. H. & Watson, M. W. (2001), 'Vector Autoregressions', *Journal of Economic Perspectives* **15**(4), 101–115.

Timmermann, A. (2006), Forecast Combinations, 1 edn, Vol. 1, Elsevier, chapter 04, pp. 135–196.

Villani, M. (2009), 'Steady-state priors for vector autoregressions', *Journal of Applied Econometrics* **24**(4), 630–650.

Waggoner, D. F. & Zha, T. (1999), 'Conditional Forecasts in Dynamic Multivariate Models', *The Review of Economics and Statistics* **81**(4), 639–651.

Wang, M.-C. (2009), 'Comparing the DSGE model with the factor model: an out-of-sample forecasting experiment', *Journal of Forecasting* **28**, 167–182.

Wolters, M. H. (2015), 'Evaluating Point and Density Forecasts of DSGE Models', *Journal of Applied Econometrics* **30**, 74–96.

Yu, L., Wang, S. & Lai, K. K. (2005), 'A novel nonlinear ensemble forecasting model incorporating GLAR and ANN for foreign exchange rates', *Computers & Operations Research* **32**, 2523–2541.

Zha, T. (1999), 'Block recursion and structural vector autoregressions', *Journal of Econometrics* **90**(2), 291–316.

Zhang, G. P. (2003), 'Time series forecasting using a hybrid ARIMA and neural network model', *Neurocomputing* **50**, 159–175.

Appendices

A Data

Quarterly variables (1999Q1 - 2018Q3)

Name	Exogenous	Pub. lag (quarters)
Gross domestic product (B1GMH)		1
Private consumption expenditure (P31S14+S15)		1
Public consumption expenditure (P3S13)		1
Gross fixed capital formation, res. buildings (P51N111)		1
Exports of goods (P61)		1
Exports of services (P62)		1
Imports of goods (P71)		1
Imports of services (P72)		1
Gross fixed capital formation, exc. res. buildings (P51S1-P51N111)		1
Trade market GDP	x	1

Monthly variables (1999M01 - 2018M09)

Variable	Exogenous	Pub. lag (months)
Employment rate		1
Unemployment rate		1
Inflation		1
Consumer confidence indicator		1
Industrial confidence indicator		1
Exports of goods according to custom authorities		3
Imports of goods according to custom authorities		3
Building permits		3
USD to EUR exchange rate	x	1
Crude Oil Spot Price (BFOE)	x	1
Euribor (3 months)	x	1
Economic sentiment indicator (EU)	x	1
Consumer price index (EU)	x	1
Import price index	x	2
Export price index	x	2

Table 8: *Publication lags (Pub. lag) refer to assumptions used in the pseudo out-of-sample forecasting exercises. For example, if a monthly figures for a variable are published less than a month after the end of the respective month, the publication lag is set to one. For some variables (e.g. exchange rates) there could be real time information available and the publication lag of those variables could therefore be set to zero. However in this thesis those variables are treated as if they were available not before the beginning of the subsequent month.*

B Marginal likelihood function

This appendix derives a compact analytical expression for the unnormalized marginal likelihood function (ML) of a BVAR model with the prior imposed solely through dummy observations (e.g. large BVAR models). The derived expression can be used conveniently for hyperparameter choice.

If the entire prior of a BVAR model is imposed via dummy observations, one can obtain the ML directly by integrating an expression equivalent to the likelihood function of a multivariate normal model (i.e. the underlying VAR model). Conditioning on the initial p observations of the sample (standard assumption), the likelihood function can be written as

$$p(\boldsymbol{Y} \mid \boldsymbol{A}, \boldsymbol{\Sigma}) \propto \prod_{i=1}^{T-p} \left[\mid \Sigma \mid^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{y}_i - \boldsymbol{x}_i \boldsymbol{a})^{\intercal} \Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{x}_i \boldsymbol{a}) \right\} \right] \quad \text{(B.1)}$$

$$= \mid \Sigma \mid^{-\frac{T-p}{2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{T-p} (\boldsymbol{y}_i - \boldsymbol{x}_i \boldsymbol{a})^{\intercal} \Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{x}_i \boldsymbol{a}) \right\}$$

$$= \mid \Sigma \mid^{-\frac{T-p}{2}} \exp\left\{ -\frac{1}{2} tr\left[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{A})^{\intercal} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{A}) \Sigma^{-1} \right] \right\},$$

where $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$ are the coefficient matrices, lower-case symbols without subscripts represent the vectorized versions of the upper-case symbols, i.e. $vec(\boldsymbol{A}) = \boldsymbol{a}$ and tr denotes the trace-operator. By *completing the square* one can write the term within the trace operator as

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{A})^{\intercal} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{A}) \Sigma^{-1} = \left[(\boldsymbol{A} - \widehat{\boldsymbol{A}})^{\intercal} \boldsymbol{X}^{\intercal} \boldsymbol{X} (\boldsymbol{A} - \widehat{\boldsymbol{A}}) + \widehat{\boldsymbol{S}} \right] \Sigma^{-1}. \quad \text{(B.2)}$$

The equation B.2 and the properties of the trace-operator and Kronecker product yield the following expression for the unnormalized likelihood function:

$$p(\boldsymbol{Y} \mid \boldsymbol{A}, \boldsymbol{\Sigma}) \propto \quad \mid \Sigma \mid^{-\frac{T-p}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{a}-\widehat{\boldsymbol{a}})^{\intercal}(\boldsymbol{\Sigma}^{-1}\otimes \boldsymbol{X}^{\intercal}\boldsymbol{X})(\boldsymbol{a}-\widehat{\boldsymbol{a}})\right\} \quad \text{(B.3)}$$

$$\times \exp\left\{-\frac{1}{2} tr(\widehat{\boldsymbol{S}}\boldsymbol{\Sigma}^{-1})\right\},$$

where

$$\widehat{\boldsymbol{a}} = vec((\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1}\boldsymbol{X}^{\intercal}\boldsymbol{Y})$$

$$\widehat{\boldsymbol{S}} = (\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{A}})^{\intercal}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{A}}).$$

The unnormalized posterior distribution of a BVAR model with a perfectly flat prior is obtained as a product of an arbitrary constant prior and the likelihood function derived above. This unnormalized posterior distribution is denoted as

$$p(\boldsymbol{A}, \boldsymbol{\Sigma} \mid \boldsymbol{Y})^{flat} \propto \quad \mid \Sigma \mid^{-\frac{T-p}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{a}-\widehat{\boldsymbol{a}})^{\intercal}(\boldsymbol{\Sigma}^{-1}\otimes \boldsymbol{X}^{\intercal}\boldsymbol{X})(\boldsymbol{a}-\widehat{\boldsymbol{a}})\right\}$$

$$\times \exp\left\{-\frac{1}{2} tr(\widehat{\boldsymbol{S}}\boldsymbol{\Sigma}^{-1})\right\}, \quad \text{(B.4)}$$

From here on, it becomes straight forward to integrate out $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$. Starting from the normally distributed $\boldsymbol{A}$ one obtains

$$\int p(\boldsymbol{A}, \boldsymbol{\Sigma} \mid \boldsymbol{Y})^{flat} \, \mathrm{d}\boldsymbol{A} \propto \quad \mid \boldsymbol{\Sigma} \mid^{-\frac{T-p}{2}} \mid \boldsymbol{\Sigma}^{-1}\otimes \boldsymbol{X}^{\intercal}\boldsymbol{X} \mid^{-\frac{1}{2}} \quad \text{(B.5)}$$

$$\times \exp\left\{-\frac{1}{2} tr(\widehat{\boldsymbol{S}}\boldsymbol{\Sigma}^{-1})\right\}$$

$$= \mid \boldsymbol{\Sigma} \mid^{-\frac{T-p-k}{2}} \mid \boldsymbol{X}^{\intercal}\boldsymbol{X} \mid^{-\frac{m}{2}} \exp\left\{-\frac{1}{2} tr(\widehat{\boldsymbol{S}}\boldsymbol{\Sigma}^{-1})\right\}.$$

Next, $\boldsymbol{\Sigma}$ can be integrated out (as it follows an Inverse-Wishart distribution) to obtain the unnormalized marginal likelihood function as

$$\int p(\boldsymbol{\Sigma} \mid \boldsymbol{Y})^{flat} \, \mathrm{d}\boldsymbol{\Sigma} = p(\boldsymbol{Y})^{flat} \propto \quad \mid \widehat{\boldsymbol{S}} \mid^{-\frac{v}{2}} \mid \boldsymbol{X}^{\intercal}\boldsymbol{X} \mid^{-\frac{m}{2}}, \quad \text{(B.6)}$$

where $v = T-p-k-m-1$.[15]. Finally, as noted in the appendix of Banbura

[15] T = number of total observations, p = number of lags in the model and the number

et al. (2010), based on properties of conditional probabilities, the marginal likelihood function of a BVAR model, for the part of the data *not* containing dummy observations, can be written as a fraction of the marginal likelihood function of every observation and only dummy observations. Hence, one can write

$$p(\boldsymbol{Y}) = \frac{p(\boldsymbol{Y}_*)^{flat}}{p(\boldsymbol{Y}_d)^{flat}}, \tag{B.7}$$

where $\boldsymbol{Y}_*$ is the data set containing both the observations of interest and the dummy observations and $\boldsymbol{Y}_d$ includes only the dummy observations (i.e. the prior). Therefore, denoting the vector of hyperparameters with $\boldsymbol{\delta}$, the unnormalized marginal likelihood function of a BVAR model with the prior imposed solely through dummy observations can be written as

$$p\left(\boldsymbol{Y} \mid \boldsymbol{\delta}\right) \propto \quad \frac{\mid \widehat{\boldsymbol{S}}_d \mid^{\frac{v_d}{2}} \mid \boldsymbol{X}_d^{\intercal}\boldsymbol{X}_d \mid^{\frac{m}{2}}}{\mid \widehat{\boldsymbol{S}}_* \mid^{\frac{v_*}{2}} \mid \boldsymbol{X}_*^{\intercal}\boldsymbol{X}_* \mid^{\frac{m}{2}}}. \tag{B.8}$$

of initial observations to be conditioned on, k = number of parameters in an equation (i.e $k = mp + 1$), m = number of variables in the model.