Dieppe, Alistair; Legrand, Romain; van Roye, Björn

Working Paper

The BEAR toolbox

ECB Working Paper, No. 1934

Provided in Cooperation with:
European Central Bank (ECB)

Suggested Citation: Dieppe, Alistair; Legrand, Romain; van Roye, Björn (2016) : The BEAR toolbox, ECB Working Paper, No. 1934, ISBN 978-92-899-2182-4, European Central Bank (ECB), Frankfurt a. M., https://doi.org/10.2866/292952

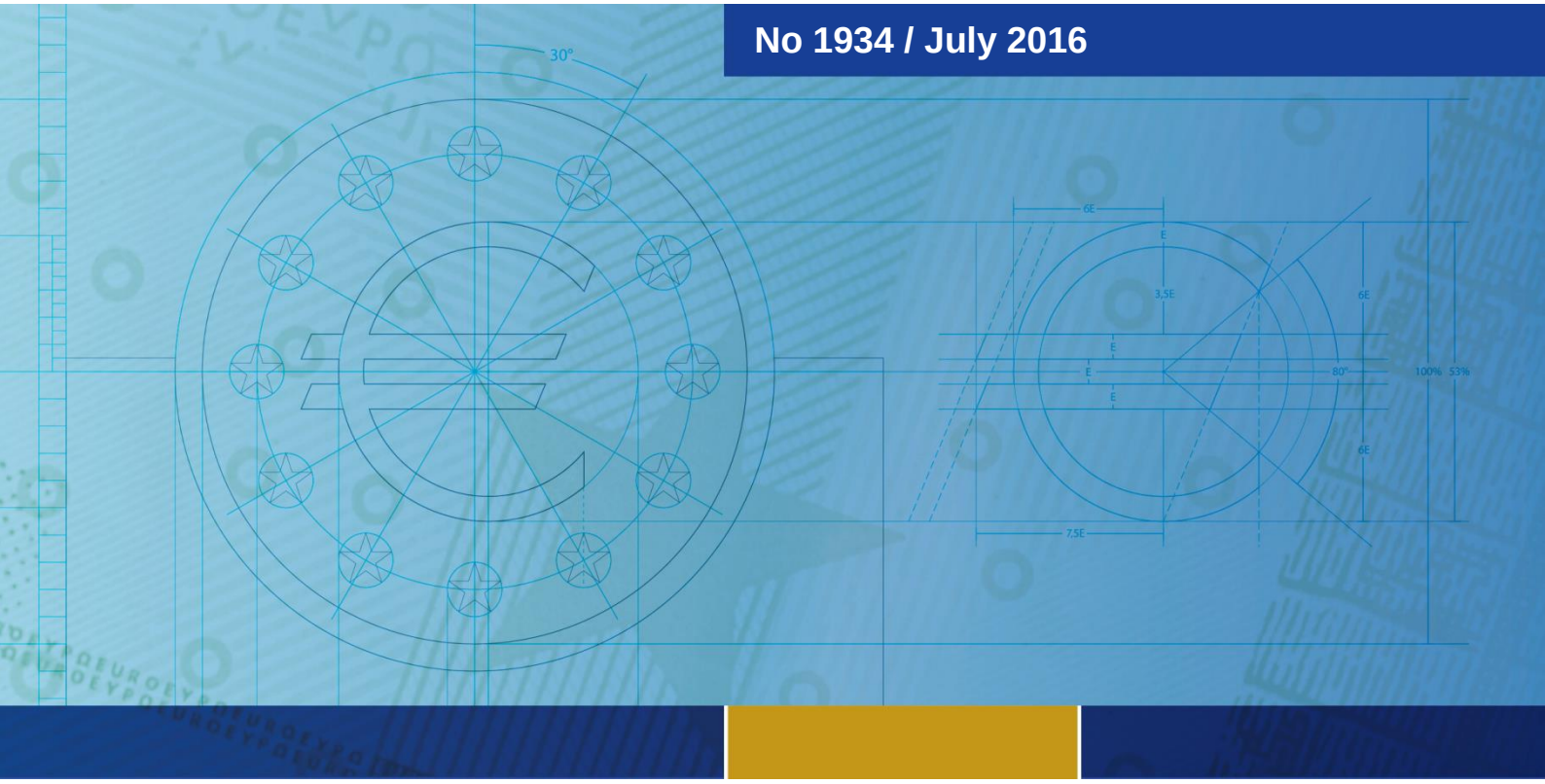
EUROPEAN CENTRAL BANK
EUROSYSTEM

Working Paper Series

Alistair Dieppe, Romain Legrand
and Björn van Roye

The BEAR toolbox

No 1934 / July 2016



Abstract

The Bayesian Estimation, Analysis and Regression toolbox (BEAR) is a comprehensive (Bayesian) (Panel) VAR toolbox for forecasting and policy analysis. BEAR is a MATLAB based toolbox which is easy for non-technical users to understand, augment and adapt. In particular, BEAR includes a user-friendly graphical interface which allows the tool to be used by country desk economists. Furthermore, BEAR is well documented, both within the code as well as including a detailed theoretical and user's guide. BEAR includes state-of-the art applications such as sign and magnitude restrictions, conditional forecasts, Bayesian forecast evaluation measures, Bayesian Panel VAR using different prior distributions (for example hierarchical priors), etc. BEAR is specifically developed for transparently supplying a tool for state-of-the-art research and is planned to be further developed to always be at the frontier of economic research.

Keywords: Bayesian VAR, Panel Bayesian VAR, Econometric Software, Forecasting, Structural VAR.

JEL classification: C11, C30, C87, E00, F00.

1 Non-technical summary

There has been an increasing use of Vector Auto Regressions (VAR) models within academia and central banks to analyse and forecast economic developments. Traditional maximum likelihood VARs, though, are often over-parameterised and imprecisely estimated if data is of questionable quality. For these reasons, Bayesian VAR models have become increasingly popular since their introduction in the seminal work of Doan et al. (1984). Some codes and software applications for Bayesian VAR models already exist, however they offer limited features, are rarely user-friendly, and are difficult to augment with new applications. We have faced these issues, and thus decided to create: the Bayesian Estimation, Analysis and Regression (BEAR) toolbox. BEAR is a comprehensive Matlab package, using Excel as both input and output. The development of BEAR was articulated around three major objectives:

- BEAR should be comprehensive. It should offer both standard features and advanced, state-of-the-art applications.
- BEAR should be easy to use and equally accessible to Bayesian experts and non-specialist desk economists. For this reason, BEAR works with a user-friendly system of graphical interfaces as well as a developers version for advanced users. In addition, BEAR comes with a comprehensive user guide.
- BEAR should be technically flexible and transparent. For this reason, its code is structured in a way that makes it easy to read and adapt. Furthermore, BEAR is accompanied with a technical guide providing complete mathematical derivations for all its applications.

By making the toolbox available, we aim at sharing expertise, and hope BEAR could become a key tool for macroeconomic analysis which exploits synergies and increases efficiency as well as avoids unnecessary duplication of work.

BEAR Version 3.0 offers following applications:

- Estimation techniques of VAR models
  - OLS (maximum likelihood) VAR
  - Standard Bayesian VAR (Doan et al. (1984) and Litterman (1986))
  - Mean-adjusted BVAR with informative prior on the steady-state (Villani (2009))
  - Bayesian Panel VAR (as in Canova and Ciccarelli (2013))
- Alternative priors for Bayesian VAR models

  - Minnesota (Litterman (1986))
  - Normal Wishart (Kadiyala and Karlsson (1997))
  - Independent Normal Wishart with Gibbs sampling
  - Normal diffuse (Kadiyala and Karlsson (1997))
  - Dummy observations (Banbura et al. (2010))
- Prior extensions for Bayesian VARs
  - Hyperparameter optimisation by grid search (similar to Giannone et al. (2015))
  - Block exogeneity
  - Dummy observation extentions: sum-of-coefficient, dummy initial obervation (Banbura et al. (2010))
- Panel models
  - OLS Mean-group estimator (Pesaran and Smith (1995))
  - Bayesian pooled estimator
  - Random effect model, Zellner-Hong (Zellner and Hong (1989))
  - Random effect model, hierarchical (Jarocinski (2010b))
  - Static factor model (Canova and Ciccarelli (2013))
  - Dynamic factor model (Canova and Ciccarelli (2013))
- Structural VARs
  - Choleski factorisation
  - Triangular factorisation
  - Sign, magnitude and zero restrictions (Arias et al. (2014))
- Applications
  - Unconditional forecasts
  - Impulse response functions
  - Forecast error variance decomposition
  - Historical decompositions
  - Conditional forecasts: shock approach (Waggoner and Zha (1999))
  - Conditional forecasts: tilting approach (Robertson et al. (2005))
  - Forecast evaluation: standard and Bayesian-specific criteria

2 Introduction

2.1 Why create a Bayesian Estimation, Analysis and Regression (BEAR) toolbox?

2.1.1 Motivation

There has been an increasing use of Vector Auto Regressions (VAR) models within academia and central banks to analyse and forecast economic developments. In many respects, VAR models have become the workhorse of macroeconometric modelling. Traditional maximum likelihood VARs, though, suffer from two major defects. First, VAR models are often over-parameterised. Too many lags are included in order to improve the in-sample fit, resulting in a significant loss of degrees of freedom and poor out-of-sample forecast performances. Second, central bankers and financial institutions are paying more and more attention to emerging economies for which available datasets are typically short or of questionable quality. Bayesian estimation techniques offer an appealing solution to these issues. Bayesian prior shrinkage allows to reduce the number of lags, hence limiting the over-parameterisation issue. Additionally, the supply of prior information compensates for the possible lack of reliability of the data. For these reasons, Bayesian VAR models have become increasingly popular since their introduction in the seminal work of Doan et al. (1984).

Many codes and software applications for Bayesian VAR already exist, however they suffer from major limitations. Most of them offer very limited features, making it difficult to use them for any advanced research project. Also, such codes are rarely user-friendly, making them hardly accessible to anyone not being an expert in mathematical programming. Finally, Bayesian econometrics is a very dynamic field. As promising applications are published on a regular basis, a good Bayesian tool should be flexible enough to integrate new contributions as they are released. This may not be easily done with existing applications.

We have faced these issues, and thus decided to create: the Bayesian Estimation, Analysis and Regression (BEAR) toolbox. BEAR is a comprehensive Matlab package, using Excel as both input and output. The development of BEAR was articulated around three major objectives:

- BEAR should be comprehensive. It should offer both standard features and advanced, state-of-the-art applications.
- BEAR should be easy to use and equally accessible to Bayesian experts and non-specialist desk economists. For this reason, BEAR works with a user-friendly system of graphical interfaces as well as a developers version for advanced users. In addition, BEAR comes with a comprehensive user guide.

- BEAR should be technically flexible and transparent. For this reason, its code is structured in a way that makes it easy to read and adapt. Furthermore, BEAR is accompanied with a technical guide providing complete mathematical derivations for all its applications.

By making the toolbox available, we aim at sharing expertise, and believe BEAR could become a key tool for macroeconomic analysis which exploits synergies and increases efficiency as well as avoids unnecessary duplication of work.

The remainder of the paper is organized as follows. We continue the introduction by first summarizing main BEAR-applications (section 2.1.2), and then conclude the introduction by presenting an illustrative example on US monetary policy using BEAR (section 2.2). Subsequently, we turn to the core of the paper, which describes the theoretical and econometric underpinnings of BEAR. In section 3 we present the background of BVAR model estimation and evaluation, in section 4 we introduce basic applications under the BVAR methodology. In section 5 we describe advanced applications, and in section 6 we finally introduce Bayesian Panel VAR models.

2.1.2 BEAR-Toolbox applications

We next list an overview of the applications available in BEAR. BEAR Version 3.0 offers the following applications:

- Estimation techniques of VAR models
  - OLS (maximum likelihood) VAR
  - Standard Bayesian VAR (Doan et al. (1984) and Litterman (1986))
  - Mean-adjusted BVAR with informative prior on the steady-state (Villani (2009))
  - Bayesian Panel VAR (as in Canova and Ciccarelli (2013))
- Alternative priors for Bayesian VAR models
  - Minnesota (Litterman (1986))
  - Normal Wishart (Kadiyala and Karlsson (1997))
  - Independent Normal Wishart with Gibbs sampling
  - Normal diffuse (Kadiyala and Karlsson (1997))
  - Dummy observations (Banbura et al. (2010))
- Prior extensions for Bayesian VARs
  - Hyperparameter optimisation by grid search (similar to Giannone et al. (2015))

  - Block exogeneity
  - Dummy observation extentions: sum-of-coefficient, dummy initial obervation (Banbura et al. (2010))

- Panel models
  - OLS Mean-group estimator (Pesaran and Smith (1995))
  - Bayesian pooled estimator
  - Random effect model, Zellner-Hong (Zellner and Hong (1989))
  - Random effect model, hierarchical (Jarocinski (2010b))
  - Static factor model (Canova and Ciccarelli (2013))
  - Dynamic factor model (Canova and Ciccarelli (2013))

- Structural VARs
  - Choleski factorisation
  - Triangular factorisation
  - Sign, magnitude and zero restrictions (Arias et al. (2014))

- Applications
  - Unconditional forecasts
  - Impulse response functions
  - Forecast error variance decomposition
  - Historical decompositions
  - Conditional forecasts: shock approach (Waggoner and Zha (1999))
  - Conditional forecasts: tilting approach (Robertson et al. (2005))
  - Forecast evaluation: standard and Bayesian-specific criteria

2.2 An example using BEAR for US monetary policy analysis

We now introduce an illustrative study on US monetary policy using BEAR. This example application does not aim at producing a major contribution in terms of analysis, but rather at setting an example of the sort of study that can be undertaken. Our setting mostly replicates the seminal work of Christiano et al. (1999) in a simplified fashion. The main dataset comprises 3 series of data: the log of real GDP, the log of the consumer price index, and the target Federal Funds rate, obtained

via Haver Analytics. All the data are quarterly, start in the first quarter of 1960 and end in the last quarter of 2015. As the study also includes some panel applications, the same dataset is replicated for the Euro area, Japan and the United Kingdom.

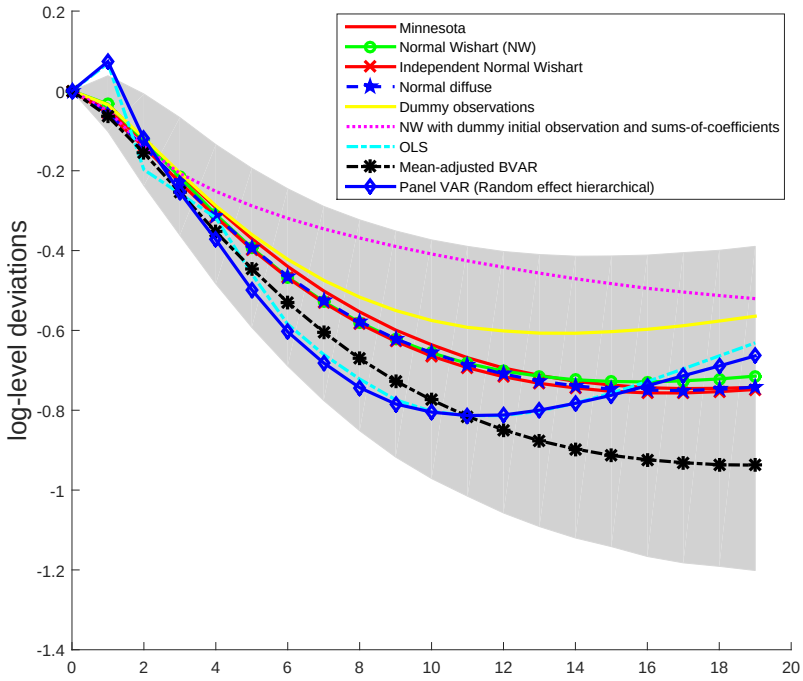
The presentation is divided into two parts. In the first part we illustrate the differences arising between the different models proposed by BEAR. The analysis is carried out by the way of two basic applications: impulse response functions, and unconditional forecasts. In the second part we cover more sophisticated applications including sign restrictions, historical decomposition and a conditional forecast exercise.

2.2.1 BEAR models and basic applications

This section proposes a comparison of different models available in BEAR. The first candidate is the benchmark maximum likelihood (or ordinary least squares) VAR model. The second model is the Bayesian VAR. BEAR proposes no less than 5 different priors for this model: the original Minnesota prior proposed by Litterman (1986) (section 3.3), the natural conjugate normal-Wishart (section 3.4), the independent normal-Wishart prior (section 3.5), the normal-diffuse prior (section 3.6), and the dummy observation prior (section 3.7). Because the data is used in log levels there is a significant possibility of non-stationarity in the results. For this reason, we also estimate a version of the model where the normal-Wishart prior is augmented with dummy initial observation and sums-of-coefficients applications, forming the so-called Sims and Zha (1997) prior. The third model is the mean-adjusted Bayesian VAR model introduced by Villani (2009) (section 5.6). This model makes it possible to explicitly integrate prior information about the long-run or steady-state values of the model. Given that real GDP and the consumer price index are in log levels and that the monotonic increase of the data suggests non-stationarity, we set the priors for the steady state to revolve around the end of sample values. This represents a conservative view based on a limited growth assumption for the variables included in the model. Subsequently, we set a prior mean of 8 for the log of real GDP with a standard deviation of 0.5, and a prior mean of 5.5 for the log of the CPI, with a 0.25 standard deviation. The target Federal Funds rate appears to be stationary, though characterised by ample fluctuations. For this reason, we set a prior mean of 4%, with a unit standard deviation. The final candidate consists of a Bayesian panel VAR (section 6). BEAR proposes 6 different panel models, but for the sake of simplicity we retain only one for this exercise: the random effect model with a hierarchical prior inspired from Jarocinski (2010b). All the models are run with 3 lags. The Bayesian models are all run with a Minnesota-type scheme for the prior distribution of the VAR coefficients. As all the data are in log levels, we follow Litterman (1986) and set to 1 the prior value of the autoregressive coefficients on its own first lag for each variable.

BEAR can conveniently estimate impulse response functions (section 4.2). The first base exercise consists of an analysis of the effect of a benchmark contractionary monetary policy shock. We adopt a structural identification by triangular factorisation thanks to which the impulse response functions are directly interpretable as the response to a unit structural shock.

Figure 1: *Impulse response functions to a unit monetary policy shock*



NOTE: Shaded area represents 95 percent credibility intervals for the normal-Wishart prior.

Figure 1 displays the impulse response functions for the selected models. The shaded area represents the 95% credibility interval obtained for the normal-Wishart prior. The results obtained after a benchmark monetary policy shocks are very similar to that of Christiano et al. (1999). A contractionary monetary policy shock leads to a sustained decline in real GDP, the effect becoming significant after roughly two quarters. The response is hump-shaped, with the maximal decline taking place after 10 to 16 quarters.

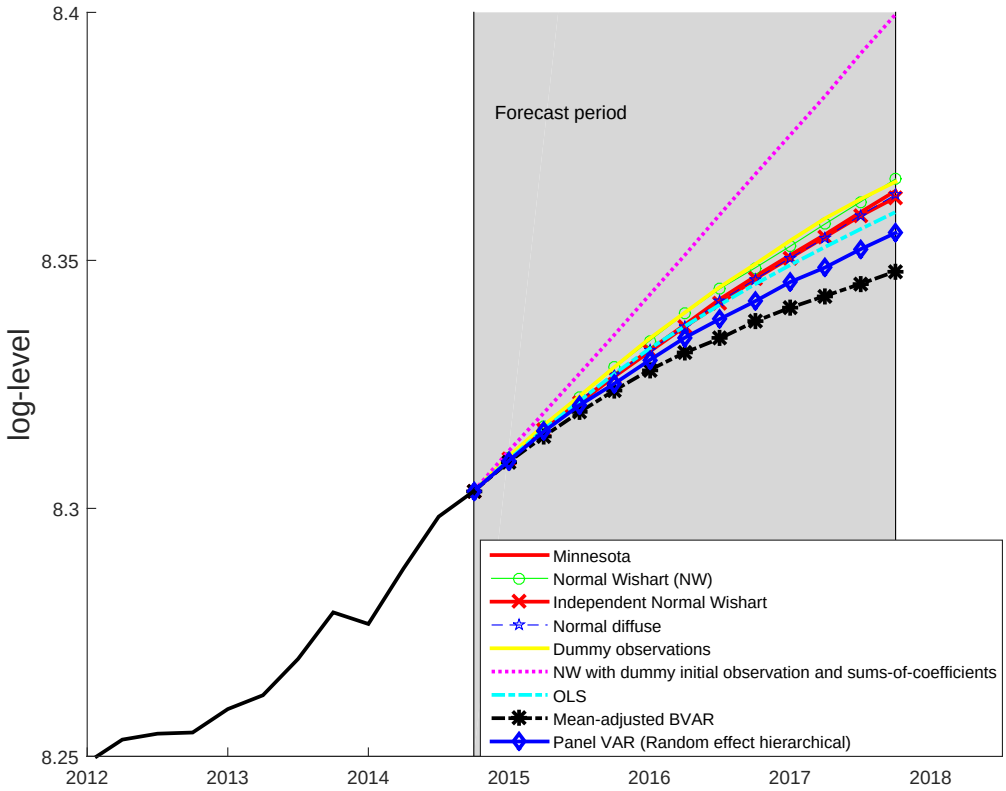
The responses produced by the Bayesian VAR models under the different priors look all very similar, with two noteworthy exceptions. The first is the dummy observation prior which displays a less pronounced decline in GDP and an earlier recovery. This is hardly surprising as the dummy observation prior is the only model for which prior information is transmitted to the model through the likelihood function rather than by the prior distribution. As the two components are attributed

different weight in the posterior, noticeable differences may result in the estimates. The second exception occurs when the dummy initial observation and sums-of-coefficients applications are added to the prior distribution. As the two components together push the model towards a (cointegrated) unit root process, more inertia is generated. The response is slower to reach its minimal value, and suggests a possible permanent effect of the shock.

Examining the responses of the alternative models leads to results which are qualitatively comparable, even though marked differences appear from a quantitative point of view. The response produced by the OLS VAR seems shorter-lived than its Bayesian counterpart, with a more pronounced initial fall in production followed by a faster recovery. This difference is most likely due to the absence of prior information in the model, so that the results represent the information contained in the data alone. The panel model results in responses close to that of the OLS model, but in this case the discrepancy with the Bayesian VAR models most likely results from the spillover effects induced by the multilateral nature of the panel framework (see e.g. Georgiadis (2015) for more details on spillover effects). The mean-adjusted Bayesian VAR finally shows a response which is markedly more negative than any other model in the medium run, perhaps reflecting the fact that the higher steady-state value of the interest rate induces stronger effects of monetary policy in general.

BEAR can also conveniently do unconditional forecasts (section 4.1). The second exercise consists in producing standard unconditional forecasts for the log of real GDP. The forecast period starts in 2014q4 and ends in 2017q4.

Figure 2: *Unconditional forecasts for real GDP*



The results displayed in Figure 2 are qualitatively similar across models: after 2014q4 real GDP grows steadily with a sustained growth until roughly 2016, before a slight slowdown for the rest of the period. The characteristics observed for the different models are overall consistent with that of impulse response functions. The forecasts obtained for the Bayesian VAR under the different priors are all very similar, except once again for the normal-Wishart augmented with the sums-of-coefficients and dummy initial observation extensions for which the growth is significantly more protracted. This is to be expected as the implied unit root favours permanent shifts in the steady-state. The panel VAR produces forecasts which are noticeably lower than the Bayesian VAR models, most likely reflecting the impact of the additional information contained in the external units. Finally, the mean-adjusted produces the lowest forecast values. This is a direct consequence of setting the prior mean for the steady-state as the end of sample value, which biases the forecasts downward.

2.2.2 Advanced applications with BEAR

Beyond standard applications, BEAR makes it possible to run more sophisticated features in a straightforward way. We now build on the previous section by identifying structural shocks and

estimating their impacts. To do so, we adopt the sign and zero restriction methodology proposed by Arias et al. (2014) (section 4.6). We identify 3 shocks: a demand shock, a supply shock, and a monetary policy shock. The following is assumed for the sign of the responses to the different shocks: Following standard theory, demand shocks have a positive effect on output while driving

Table 1: *Sign of the responses to identified shocks*

	demand	supply	monetary
log real GDP	+	+	+
log CPI	+	-	+
Federal Funds rate	+		-

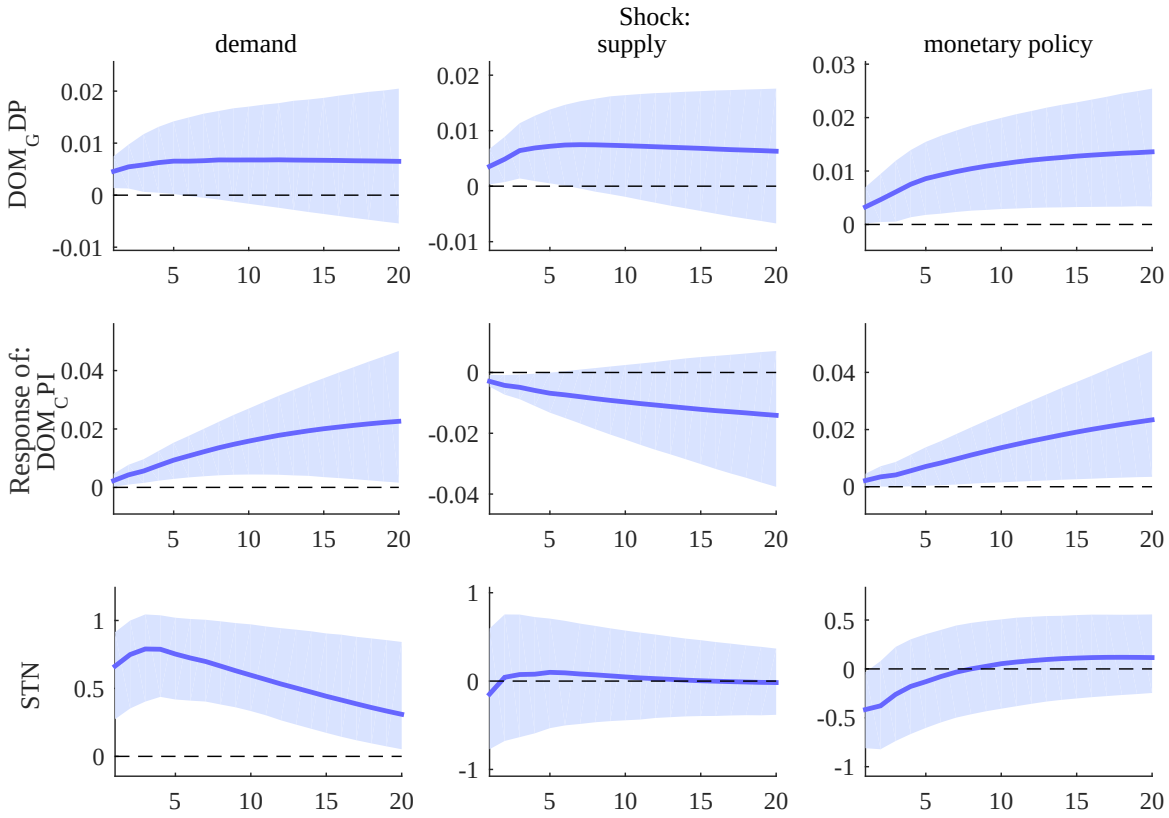
up inflation and the interest rate. Supply shocks impact output positively and contribute to lower prices. The effect on the Federal Funds rate is left undetermined as it is not certain whether the increase in activity or the fall in price will be predominant in the response of the Central Bank to the shock. Finally, an expansionary monetary policy shock translates into a cut in the Federal Fundss rate which boosts output and contributes to increase the price level. With such an identification scheme the shocks are unambiguously defined since they cannot generate similar responses for all the variables. The restrictions are defined over the following periods:

Table 2: *Periods of application of the restrictions*

	demand	supply	monetary
log real GDP	0 3	0 3	0 0
log CPI	0 4	0 3	0 0
Federal Funds rate	1 4		0 0

We obtain the set of impulse responses displayed on Figure 3:

Figure 3: *Impulse responses with sign restrictions*



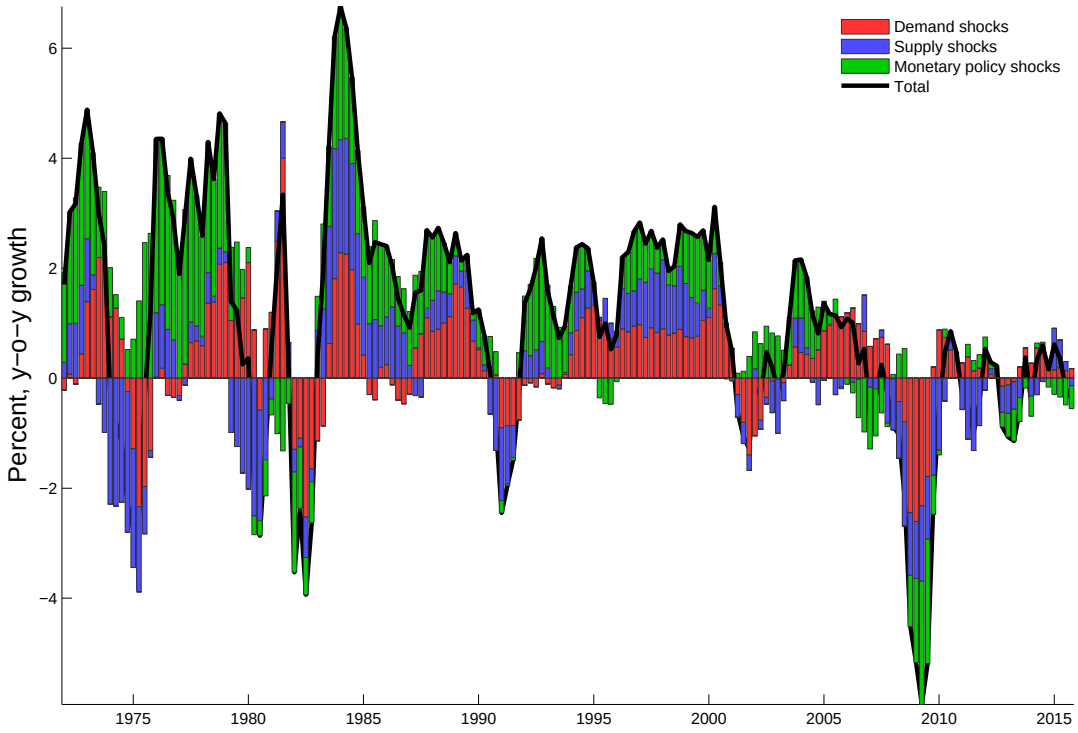
NOTE: Shaded area represents 95 percent credibility intervals for the normal-Wishart prior.

All the responses are initially significant except that of the Federal Funds rate to supply shocks. This is the only response to which no restriction was placed, and such non-significance is quite typical of the Arias et al. (2014) methodology. The main message in terms of transmission of monetary policy shocks is that the effect remains significant on output and the CPI over the whole period of responses, even though the restriction only applies on impact. This confirms the importance and effectiveness of monetary policy to stabilise economic fluctuations in the US. The effect of the shock on the Federal Funds rate is at first negative but becomes positive after roughly two years. This suggests that the initial boost of activity may lead the central bank to reverse their stance in order to counter inflationary pressures. The response however is not significant.

This structural identification scheme also makes it possible to undertake further applications. In particular BEAR offers the possibility to obtain estimates for the sample historical decomposition from the sign restriction framework, which is our second application (section 5.2). The contribution

of each shock is calculated as the median of the posterior distribution, and we also consider the total shock contributions defined as the sum of the individual contributions. These estimates are displayed in Figure 4:

Figure 4: *Historical shock decomposition for US GDP growth*



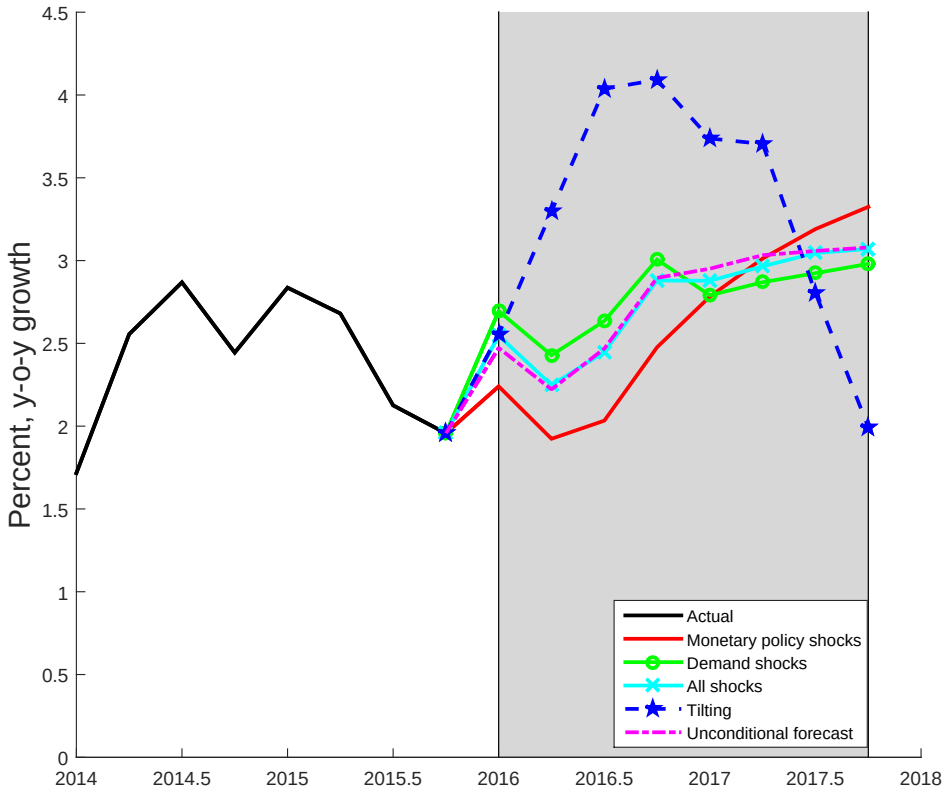
The broad picture provided by the decomposition is that over the whole period, demand shocks and monetary shocks seem to have represented the bulk of real GDP fluctuations, with supply shocks playing a more limited role. There are noteworthy exceptions to this: the 1973 and 1979 oil crises, and the 1985 and 1996 expansions. The 2009 crisis clearly appears as a mostly demand and monetary driven event, with supply contributing more modestly. Supply shocks seem to have gained in importance in the immediate aftermath of the crisis and have continued to play a non negligible role more recently.

The decomposition confirms the importance of monetary policy shocks in US business cycles. For certain periods, their role is actually predominant, as can be seen from the years 1976-1979 and 1992-1994 periods, which corresponds to periods of very accommodating monetary policy. The (negative) contribution was also major over the years 2008 and 2009 of the crisis. The contribution of monetary policy shocks has become less significant since 2010, with the Federal Reserve being limited in its

action by the zero lower bound reached by the Federal Funds rate since 2010.

The final exercise consists of a conditional forecast experiment. The objective is to analyse the effect of monetary policy on real GDP growth. The experiment consists in assuming a rise of the Federal Funds rate to 0.5% percent over the period 2016q1-2017q4. For the sake of clarity, data for real GDP is turned into year-on-year growth rate. 4 different estimation settings are explored. The first 3 of them rely on the standard methodology developed by Waggoner and Zha (1999) which builds on structural shocks (section 5.3). For the first experiment the conditions are generated only by demand shocks; for the second experiment the conditions are generated by monetary shocks only; for the third one, the conditions are generated by all the shocks jointly, including supply shocks. Finally, a fourth set of conditional forecast is produced using the tilting methodology proposed by Robertson et al. (2005) (section 5.5). This methodology is agnostic about shocks, which represents an interesting alternative for our experiment. The results are shown in Figure 5:

Figure 5: *Conditional forecasts: effect of Federal Funds rate increase on real GDP growth*



The first noticeable characteristics is that the results differ quite significantly according to the selected methodology. This highlights the importance of choosing a suitable setting in order to estab-

lish meaningful results. The lowest conditional forecast values are produced by the pure monetary policy shock scenario. In this case, the perspective in terms of growth is even more pessimistic than for the unconditional forecasts. This is easily explanable: as the monetary authorities implement a set of contractionary monetary policies, economic activity is negatively impacted which results in a noticeable drop of real GDP growth. In this case, monetary policy precedes real activity. By contrast, the pure demand shock scenario leads to an anticipated real GDP growth which is more optimistic than the unconditional forecasts. This is because in this case the economic rationale behind the results is reversed: an initial increase in demand leads to a fueling in real activity, pushing the central authorities to increase the interest rate to prevent inflationary pressures. In this case, real activity precedes monetary policy. The all shocks scenario is somewhere in between: a mixture of shocks hits the economy, some of them enhancing activity (supply and demand shocks) while others hamper it (contractionary monetary shocks). The final forecast is, in this case, fairly close to the unconditional forecast. The final methodology is tilting and it induces an initial GDP growth which is significantly higher than with the standard methodology. The methodology is agnostic about shocks and considers only distributions. Therefore, this results indicates that from a purely statistical point of view the distribution of real GDP growth needs to shift by this much in order to be consistent with the specified path for the Federal Funds rate. Compared with the shock-based methodology this implies a much weaker initial response of monetary authorities to real activity, followed however by a more sustained action resulting in the interest rate to remain high even though the initial rise in GDP growth vanishes.

In what follows, we present the underlying econometric methodologies and principles used in BEAR and provide a concise theoretical background. In particular, we provide thorough derivations and describe in detail the technical details of all applications that can be implemented in BEAR.

3 Model estimation and evaluation

3.1 VAR models: formulation and estimation

A general VAR model with n endogenous variables, p lags, and m exogenous variables can be written as:

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \\ \vdots \\ y_{n,t} \end{pmatrix} = \begin{pmatrix} a_{11}^1 & a_{12}^1 & \cdots & a_{1n}^1 \\ a_{21}^1 & a_{22}^1 & \cdots & a_{2n}^1 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}^1 & a_{n2}^1 & \cdots & a_{nn}^1 \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \\ \vdots \\ y_{n,t-1} \end{pmatrix} + \cdots + \begin{pmatrix} a_{11}^p & a_{12}^p & \cdots & a_{1n}^p \\ a_{21}^p & a_{22}^p & \cdots & a_{2n}^p \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}^p & a_{n2}^p & \cdots & a_{nn}^p \end{pmatrix} \begin{pmatrix} y_{1,t-p} \\ y_{2,t-p} \\ \vdots \\ y_{n,t-p} \end{pmatrix}$$
$$+ \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2p} & \cdots & c_{nm} \end{pmatrix} \begin{pmatrix} x_{1,t} \\ x_{2,t} \\ \vdots \\ x_{m,t} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{1,t} \\ \vdots \\ \varepsilon_{1,t} \end{pmatrix} \tag{3.1.1}$$

In compact form, the model rewrites:

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + C x_t + \varepsilon_t, \text{ where } t = 1, 2, ..., T \tag{3.1.2}$$

$y_t = (y_{1,t}, y_{2,t}, \cdots, y_{n,t})$ is a $n \times 1$ vector of endogenous data, $A_1, A_2, \ldots, A_p$ are p matrices of dimension $n \times n$, C is a $n \times m$ matrix, and x_t is a $m \times 1$ vector of exogenous regressors which can be e.g. constant terms, time trends, or exogenous data series. $\varepsilon_t = (\varepsilon_{1,t}\ \varepsilon_{2,t}\ \cdots\ \varepsilon_{n,t})$ is a vector of residuals following a multivariate normal distribution:

$$\varepsilon_t \sim \mathcal{N}(0, \Sigma) \tag{3.1.3}$$

ε_t is assumed to be non-autocorrelated, so that $E(\varepsilon_t \varepsilon_t^{'}) = \Sigma$ while $E(\varepsilon_t \varepsilon_s^{'}) = 0$ if $t \neq s$. Σ is a $n \times n$ symmetric positive definite variance-covariance matrix, with variance terms on the diagonal and covariance terms off diagonal. T is the size of the sample used for the regression, and the structure of the VAR implies that there are $k = np + m$ coefficients to estimate for each equation, leaving a total of $q = nk = n(np + m)$ coefficients to estimate for the full VAR model.

For further computation, a convenient reformulation of 3.1.2 consists in writing the VAR in transpose form as:

$$y_t^{'} = y_{t-1}^{'} A_1^{'} + y_{t-2}^{'} A_2^{'} + ... + y_{t-p}^{'} A_p^{'} + x_t^{'} C^{'} + \varepsilon_t^{'} \text{ where } t = 1, 2, ..., T \tag{3.1.4}$$

Because 3.1.4 holds for any t, one can stack observations in the usual way to reformulate the model for the whole data set:

$$\underbrace{\begin{pmatrix} y_1' \\ y_2' \\ \vdots \\ y_T' \end{pmatrix}}_{T\times n} = \underbrace{\begin{pmatrix} y_0' \\ y_1' \\ \vdots \\ y_{T-1}' \end{pmatrix}}_{T\times n} \underbrace{A_1'}_{n\times n} + \underbrace{\begin{pmatrix} y_{-1}' \\ y_0' \\ \vdots \\ y_{T-2}' \end{pmatrix}}_{T\times n} \underbrace{A_2'}_{n\times n} + \ldots + \underbrace{\begin{pmatrix} y_{1-p}' \\ y_{2-p}' \\ \vdots \\ y_{T-p}' \end{pmatrix}}_{T\times n} \underbrace{A_p'}_{n\times n} + \underbrace{\begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_T' \end{pmatrix}}_{T\times m} \underbrace{C'}_{m\times n} + \underbrace{\begin{pmatrix} \varepsilon_1' \\ \varepsilon_2' \\ \vdots \\ \varepsilon_T' \end{pmatrix}}_{T\times n} \tag{3.1.5}$$

Gathering the regressors into a single matrix, one obtains:

$$\underbrace{\begin{pmatrix} y_1' \\ y_2' \\ \vdots \\ y_T' \end{pmatrix}}_{T\times n} = \underbrace{\begin{pmatrix} y_0' & y_{-1}' & \cdots & y_{1-p}' & x_1' \\ y_1' & y_0' & \cdots & y_{2-p}' & x_2' \\ \vdots & \vdots & & \vdots & \vdots \\ y_{T-1}' & y_{T-2}' & \cdots & y_{T-p}' & x_T' \end{pmatrix}}_{T\times k} \underbrace{\begin{pmatrix} A_1' \\ A_2' \\ \vdots \\ A_p' \\ C' \end{pmatrix}}_{k\times n} + \underbrace{\begin{pmatrix} \varepsilon_1' \\ \varepsilon_2' \\ \vdots \\ \varepsilon_T' \end{pmatrix}}_{T\times n} \tag{3.1.6}$$

Or, in more compact notation:

$$Y = XB + \mathcal{E} \tag{3.1.7}$$

with:

$$Y = \begin{pmatrix} y_1' \\ y_2' \\ \vdots \\ y_T' \end{pmatrix}, X = \begin{pmatrix} y_0' & y_{-1}' & \cdots & y_{1-p}' & x_1' \\ y_1' & y_0' & \cdots & y_{2-p}' & x_2' \\ \vdots & \vdots & & \vdots & \vdots \\ y_{T-1}' & y_{T-2}' & \cdots & y_{T-p}' & x_T' \end{pmatrix}, B = \begin{pmatrix} A_1' \\ A_2' \\ \vdots \\ A_p' \\ C' \end{pmatrix}, \text{ and } \mathcal{E} = \begin{pmatrix} \varepsilon_1' \\ \varepsilon_2' \\ \vdots \\ \varepsilon_T' \end{pmatrix} \tag{3.1.8}$$

In model 3.1.7, subscripts t have been dropped to emphasize the fact that this formulation encompasses the whole sample. Once the model has been stacked this way, obtaining OLS estimates of the VAR is straightforward. An estimate $\hat{B}$ of the parameter B in 3.1.7 obtains from:

$$\hat{B} = (X'X)^{-1}X'Y \tag{3.1.9}$$

Following, an OLS estimate $\hat{\mathcal{E}}$ of the residual matrix $\mathcal{E}$ can be computed from direct application

of 3.1.7, and a (degree of freedom adjusted) estimate $\hat{\Sigma}$ of the covariance matrix Σ in 3.1.3 may be obtained from:

$$\hat{\Sigma} = \frac{1}{T-k-1}\left(\hat{\mathcal{E}}'\hat{\mathcal{E}}\right) \tag{3.1.10}$$

Alternatively, using A.1.5, one can vectorise 3.1.6 to reformulate the model as:

$$\underbrace{\begin{pmatrix} y_{1,1} \\ \vdots \\ y_{1,T} \\ \vdots \\ y_{n,1} \\ \vdots \\ y_{n,T} \end{pmatrix}}_{nT\times 1} = \underbrace{\begin{pmatrix} y_0' & y_{-1}' & \cdots & y_{1-p}' & x_1' & 0 & \cdots & 0 \\ y_1' & y_0' & \cdots & y_{-p}' & x_2' & & & \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ y_{T-1}' & y_{T-2}' & \cdots & y_{T-p}' & x_T' & 0 & \cdots & 0 \\ & & & & & \ddots & & \\ 0 & \cdots & & 0 & & y_0' & y_{-1}' & \cdots & y_{1-p}' & x_1' \\ & & & & & y_1' & y_0' & \cdots & y_{-p}' & x_2' \\ \vdots & & & \vdots & & \vdots & \vdots & & \vdots & \vdots \\ 0 & \cdots & & 0 & & y_{T-1}' & y_{T-2}' & \cdots & y_{T-p}' & x_T' \end{pmatrix}}_{nT\times q} \underbrace{\begin{pmatrix} A_1^{(1)} \\ \vdots \\ A_p^{(1)} \\ C^{(1)} \\ \vdots \\ A_1^{(n)} \\ \vdots \\ A_p^{(n)} \\ C^{(n)} \end{pmatrix}}_{q\times 1}$$

$$+ \underbrace{\begin{pmatrix} \varepsilon_{1,1} \\ \vdots \\ \varepsilon_{1,T} \\ \vdots \\ \varepsilon_{n,1} \\ \vdots \\ \varepsilon_{n,T} \end{pmatrix}}_{nT\times 1} \tag{3.1.11}$$

where in the above formulation, $A_i^{(j)}$and $C^{(j)}$ respectively denote the transpose of row j of matrix A_i and C. 3.1.11 reformulates compactly as:

$$y = \bar{X}\beta + \varepsilon \tag{3.1.12}$$

with:

$$y = vec(Y),\ \bar{X} = I_n \otimes X,\ \beta = vec(B),\ \varepsilon = vec(\mathcal{E}) \tag{3.1.13}$$

Also, from 3.1.3, one obtains:

$$\varepsilon \sim \mathcal{N}(0, \bar{\Sigma}), \text{ where } \bar{\Sigma} = \Sigma \otimes I_T \tag{3.1.14}$$

An OLS estimate $\hat{\beta}$ of the vectorised form β in 3.1.12 can be obtained as:

$$\hat{\beta} = (\bar{X}'\bar{X})^{-1}\bar{X}'y \tag{3.1.15}$$

Note that one can also simply use 3.1.9 and vectorise $\hat{B}$ to recover $\hat{\beta}$. This solution is often preferred in practice, since the computation of $\hat{B}$ involves smaller matrices and thus produces more accurate estimates. Similarly, OLS estimates $\hat{\varepsilon}$ for the residuals can be obtained either by direct application of 3.1.12, or by vectorising $\hat{\mathcal{E}}$ calculated from 3.1.7.

It should be clear that 3.1.7 and 3.1.12 are just alternative but equivalent representations of the same VAR model 3.1.2. In the incoming developments, one representation or the other will be chosen according to which one is most convenient for computational purposes. 3.1.7 is typically faster to compute, while the main appeal of 3.1.12 resides in the fact that Bayesian analysis typically works with β rather than with B.

3.2 Bayesian VAR estimation: principles

In Bayesian econometrics, every parameter of interest is treated as a random variable, characterized by some underlying probability distribution. The aim of the econometrician is thus to identify these distributions in order to produce estimates and carry inference on the model. This differs from the traditional, frequentist approach which assumes that there exist "true" parameter values, so that the work of the econometrician is limited to the identification of these "true" values.

In a VAR framework, the parameters of interest for the econometrician are the coefficients of the model, gathered in the vector β in 3.1.12, along with the residual covariance matrix Σ defined in 3.1.3 (though in some instances, it may be assumed that it is known). The principle of Bayesian analysis is then to combine the prior information the econometrician may have about the distribution for these parameters (the prior distribution) with the information contained in the data (the likelihood function) to obtain an updated distribution accounting for both these sources of information, known as the posterior distribution. This is done by using what is known as Bayes rule, which represents the cornerstone of Bayesian Analysis. For a general (vector of) parameter(s) θ and a data set y, Bayes rule can be obtained from basic definitions of conditional probabilities, by noting that:

$$\pi(\theta\,|y\,) = \frac{\pi(\theta, y)}{\pi(y)} = \frac{\pi(\theta, y)}{\pi(y)}\frac{\pi(\theta)}{\pi(\theta)} = \frac{\pi(y, \theta)}{\pi(\theta)}\frac{\pi(\theta)}{\pi(y)} = \frac{\pi(y\,|\theta\,)\pi(\theta)}{\pi(y)} \tag{3.2.1}$$

As it is common practice to denote data density by $f(y\,|\theta)$ rather than by $\pi(y\,|\theta)$, Bayes rule is typically written as:

$$\pi(\theta\,|y) = \frac{f(y\,|\theta)\pi(\theta)}{f(y)} \tag{3.2.2}$$

Formula 3.2.2 says that $\pi(\theta|y)$, the posterior distribution of θ conditional on the information contained in y, is equal to the product of the data likelihood function $f(y\,|\theta)$ with the prior distribution $\pi(\theta)$, divided by the density $f(y)$ of the data. Since the denominator $f(y)$ is independent of θ, it only plays the role of a normalizing constant with respect to the posterior $\pi(\theta\,|y)$, so that it is often convenient to ignore it and rewrite 3.2.2 as:

$$\pi(\theta\,|y) \propto f(y\,|\theta)\pi(\theta) \tag{3.2.3}$$

In essence, any Bayesian estimation of econometric models reduces to an application of 3.2.3. This expression allows to obtain the posterior distribution $\pi(\theta\,|y)$, which represents the central object for inference as it combines in one single expression all the information we have about θ. It is this posterior distribution which is then used to carry inference about the parameter values, compute point estimates, draw comparisons between models, and so on.

A preliminary remark may be done about the prior distribution $\pi(\theta)$. Most of the times, θ will not represent a single parameter, but rather several different parameters - or blocks of parameters - considered by the model. This then implies that $\pi(\theta)$ represents the joint prior distribution for all the parameters considered simultaneously, which may be difficult to determine. For example, in a typical Bayesian VAR model, θ will include two blocks: the VAR coefficients β on the one hand, and the residual variance-covariance matrix Σ on the other hand. What should be a joint distribution for a vector of VAR coefficients and a variance-covariance matrix is a question with no obvious answer.

A simple way to overcome this issue is to assume independence between parameters or blocks, so that the joint density simply becomes the product of the individual densities. This then reduces the problem to the determination of one distribution for each individual element, an easier and more meaningful strategy than looking for a joint density. For a general model with d parameters or blocks, $\pi(\theta)$ can then be rewritten as:

$$\pi(\theta) = \pi(\theta_1) \times \pi(\theta_2) \times \cdots \times \pi(\theta_d) \tag{3.2.4}$$

For instance, in the typical VAR example made of the two blocks or parameters $\theta_1 = \beta$ and $\theta_2 = \Sigma$, this allows to rewrite 3.2.3 as:

$$\pi(\theta\,|y) \propto f(y\,|\theta)\pi(\theta_1)\pi(\theta_2) \tag{3.2.5}$$

In most practical Bayesian VAR applications, it is 3.2.5 rather than 3.2.3 that will be applied to obtain the posterior distribution.

A similar issue arises with the posterior distribution $\pi(\theta\,|y\,)$: it is also a joint distribution for the parameters or blocks of the model, while the typical objects of interest for the statistician are the individual posterior distributions. To derive the marginal distributions of a particular element, one simply integrates out the remainder of the parameters from the joint posterior distribution:

$$\pi(\theta_i\,|y\,) = \int \pi(\theta_1, \theta_2, ..., \theta_d\,|y\,) \underbrace{d\theta_1 d\theta_2 ... d\theta_d}_{j \neq i} \tag{3.2.6}$$

For example, in the typical 2-block VAR model, one finds the distribution for θ_1 by integrating the joint distribution over θ_2:

$$\pi(\theta_1\,|y\,) = \int \pi(\theta_1, \theta_2\,|y\,) d\theta_2 \tag{3.2.7}$$

Similarly, one will obtain the distribution for θ_2 by integrating the joint distribution over θ_1.

3.2.2 represents the simplest formulation of Bayes rule. It is however possible to build richer and more sophisticated versions of it by using what is known as hierarchical prior distributions. To understand this concept, consider the case of the prior distribution $\pi(\theta)$ set for some parameter of interest θ. This prior distribution itself depends on some other parameter values that we may denote by λ. For instance, if $\pi(\theta)$ is the multivariate normal distribution, it depends on the set of parameters $\lambda = (\mu, \Sigma)$, which respectively represent the mean and covariance of the multivariate normal distribution. To be perfectly rigorous, one should hence denote the prior distribution for θ by $\pi(\theta\,|\lambda\,)$, but in practice the implicit parameters μ and Σ are often omitted to lighten notation so that the prior distribution is simply written as $\pi(\theta)$. The parameters λ, known as hyperparameters (they are the parameters determining the prior distribution of the parameters of interest θ), are usually assumed to be fixed and known, with values provided by the Bayesian practitioner. It is however possible to assume that λ is also a random variable, and as such to also characterize it with some prior distribution. This way, an additional layer of uncertainty is added to the model.

Because λ provides additional random variables to the model, those supplementary random variables must be added to the full posterior distribution, which thus becomes $\pi(\theta, \lambda\,|y\,)$ and not just $\pi(\theta\,|y\,)$. It is straightforward to obtain a formula for $\pi(\theta, \lambda\,|y\,)$ by starting from Bayes rule 3.2.2, and

then use basic algebra:

$$\begin{aligned}\pi(\theta,\lambda\,|y\,) &= \frac{f(y\,|\theta,\lambda)\pi(\theta,\lambda)}{f(y)}\\ &= \frac{f(y\,|\theta,\lambda)}{f(y)}\frac{\pi(\theta,\lambda)}{\pi(\lambda)}\pi(\lambda)\\ &= \frac{f(y\,|\theta,\lambda)}{f(y)}\pi(\theta\,|\lambda)\pi(\lambda)\\ &\propto f(y\,|\theta,\lambda)\pi(\theta\,|\lambda)\pi(\lambda)\end{aligned} \tag{3.2.8}$$

Note that the hyperparameter λ is only used to determine the prior distribution of θ. Therefore, once the value of θ is determined, λ becomes redundant and does not give anymore any useful information for the computation of the likelihood $f(y\,|\theta,\lambda)$. It can thus be omitted so that 3.2.8 rewrites:

$$\pi(\theta,\lambda\,|y\,) \propto f(y\,|\theta)\pi(\theta\,|\lambda)\pi(\lambda) \tag{3.2.9}$$

(3.2.9) says that to obtain the full posterior distribution of the hierarchical model, it suffices to multiply the likelihood function $f(y\,|\theta)$ with the (conditional) prior $\pi(\theta\,|\lambda)$ for θ, and the prior distribution $\pi(\lambda)$ for λ. If one is then interested only in the posterior distribution of θ, a marginalisation process similar to that of 3.2.6 is directly applicable:

$$\pi(\theta\,|y\,) = \int_{\lambda}\pi(\theta,\lambda\,|y)d\lambda \tag{3.2.10}$$

Hierarchical priors can extend to more than one stage. It is possible for instance to add a third layer of uncertainty. Indeed, since λ is a random variable, it also depends on some set of hyperparameters, say γ. In the one-stage hierarchical model, γ was implicitly assumed to be fixed and known. However, it is possible to assume that γ is actually also a random variable, and thus generate a two-stage hierarchical model. Then, Bayes formula becomes:

$$\pi(\theta,\lambda,\gamma\,|y\,) \propto f(y\,|\theta)\pi(\theta\,|\lambda)\pi(\lambda\,|\gamma)\pi(\gamma) \tag{3.2.11}$$

And the posterior distribution for θ obtains from:

$$\pi(\theta\,|y\,) = \int_{\lambda}\int_{\gamma}\pi(\theta,\lambda,\gamma\,|y)d\lambda d\gamma \tag{3.2.12}$$

Any number of additional layers can be added to the model, with the same logic to extend Bayes rule and the marginalisation process.

Once the posterior distribution is obtained, either in a standard or hierarchical way, the question becomes how to handle the posterior distribution. The latter contains all the information that the statistician has about θ, but as such it is hardly of any use, since an entire distribution represents something too complicated to be conveniently used in practical applications. One may thus want to summarize the information contained in the whole distribution in a few criteria only.

For instance, one may typically want to obtain a point estimate for θ. This is done by using a loss function $L(\hat{\theta}, \theta)$, which specifies the loss incurred if the true value of the parameter is θ, but is estimated as $\hat{\theta}$. An example of loss function is the quadratic loss function $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$. The Bayes estimator (or point estimate) of θ is then defined as the value of $\hat{\theta}$ which minimizes the expected loss over the posterior distribution of θ. That is, $\hat{\theta}$ is chosen to minimise:

$$E\left[L(\hat{\theta}, \theta)\right] = \int L(\hat{\theta}, \theta)\pi(\theta\,|y)d\theta \tag{3.2.13}$$

With the quadratic loss function, $\hat{\theta}$ is thus obtained by minimising:

$$E\left[L(\hat{\theta}, \theta)\right] = \int (\hat{\theta} - \theta)^2\pi(\theta\,|y)d\theta \tag{3.2.14}$$

Taking the derivative, setting it to 0 and rearranging, one finds:

$$2\int (\hat{\theta} - \theta)\pi(\theta\,|y)d\theta = 0 \tag{3.2.15}$$

or

$$\hat{\theta} = \int \theta\pi(\theta\,|y)d\theta = E(\theta\,|y) \tag{3.2.16}$$

That is, the point estimate is given by the mean of the posterior distribution. Other values are possible with different loss functions. For example, using an absolute value loss function yields the median as the Bayes estimator, while the so-called step-loss function yields the mode. One may also want to compute interval estimates for θ, that is:

$$P(\theta_L \leq \theta \leq \theta_U) = \alpha \tag{3.2.17}$$

which indicates that $\theta_L \leq \theta \leq \theta_U$ with a probability of α, for instance $\alpha = 0.95$. Such intervals are referred to as credibility intervals, since they reflect plausible values for θ, values outside the interval being considered too uncommon or unlikely to be plausible. The credibility interval can be derived from the posterior distribution, either by trimming probability from both tails of the distribution, or by selecting the pair yielding the shortest interval.

In practice, the median will typically be preferred to the mean as a point estimate, for two reasons. The first is that the median is less sensitive than the mean to extreme values. Therefore, choosing the median avoids selecting a point estimate which can be very remote from the centre of the distribution, as can be the case with the mean if the posterior distribution is strongly skewed. The second is that being the 50% quantile, the median is ensured to be comprised within the bounds of a credibility interval, while the mean can produce an estimate outside these bounds in the case, once again, of a skewed distribution.

A final feature of interest is the comparison of different models. Imagine for example that one wants to compare model 1 and model 2, and determine which one is the true model. Model 1 is characterized by the prior belief or prior probability $P(M_1) = p_1$ that it is indeed the true model, by a set of parameters θ_1, a prior distribution $\pi(\theta_1 \,|M_1\,)$ over these parameters, and a likelihood function $f_1(y\,|\theta_1, M_1)$. Similarly, model 2 is characterized by $P(M_2) = p_2$,θ_2, $\pi(\theta_2\,|M_2\,)$ and $f_2(y\,|\theta_2, M_2)$. The Bayesian methodology then consists in computing for each model the posterior probability $P(M_1|y)$, which is interpreted as the probability that model i is indeed the true one, given the information contained in the data. Using Bayes rule 3.2.2, one obtains this posterior probability as:

$$P(M_i\,|y\,) = \frac{f_1(y\,|M_i)P(M_i)}{f(y)} \tag{3.2.18}$$

After some use of rules of marginal and conditional probabilities, this rewrites as:

$$P(M_i|y) = \frac{p_i \int f_i(y\,|\theta_i, M_i)\pi_i(\theta_i\,|M_i)d\theta_i}{f(y)} \tag{3.2.19}$$

The numerator term in the integral is of particular interest and is known as the marginal likelihood for model i :

$$m_i(y) = \int f_i(y\,|\theta_i, M_i)\pi_i(\theta_i\,|M_i)d\theta_i \tag{3.2.20}$$

Note that this function involves the likelihood $f_i(y\,|\theta_i, M_i)$ and the prior $\pi(\theta_i\,|M_i\,)$. The marginal likelihood is a crucial element for model comparison. Indeed, to compare model 1 with model 2 and determine which one is more likely to be the true one, the simplest method is to compute the ratio of their posterior probabilities. Using 3.2.19 and 3.2.20, one obtains:

$$R_{12} = \frac{P(M_1\,|y\,)}{P(M_2\,|y\,)} = \left(\frac{p_1}{p_2}\right)\left(\frac{m_1}{m_2}\right) \tag{3.2.21}$$

This shows that the ratio is made of two elements: the prior odds ratio p_1/p_2 which reflects the prior belief of the statistician in favour of model 1, and the ratio of the marginal likelihoods m_1/m_2, known as the Bayes factor. If the statistician has no preconceived idea on which model should be

true, he will set so that the whole burden of model comparison will fall on the Bayes factor. Once R_{12} is calculated, the last remaining issue is to determine which rule of thumb should be followed to determine whether model 1 should be deemed as the true one.

Jeffreys (1961) proposes the following guidelines:

Table 3: *Jeffrey's guideline*

$log_{10}(R_{12}) > 2$	Decisive support for M_1
$3/2 < log_{10}(R_{12}) < 2$	Very strong evidence for M_1
$1 < log_{10}(R_{12}) < 3/2$	Strong evidence for M_1
$1/2 < log_{10}(R_{12}) < 1$	Substantial evidence for M_1
$0 < log_{10}(R_{12}) < 1/2$	Weak evidence for M_1

Any negative value of $\log_{10}(R_{12})$ has of course to be interpreted as evidence against model M_1.

This subsection summarized in a nutshell all the principles underlying the practice of Bayesian econometrics. The following subsections mostly build on these principles, developing the details of the Bayesian procedures used to estimate the general VAR model introduced in subsection 3.1.

3.3 The Minnesota prior

This subsection initiates the presentation of the different prior distributions used in Bayesian VAR analysis, along with the derivations of their posterior distributions. The main text provides only the essential steps of the reasoning, but detailed derivations can be found in appendix A.3 and following (for the subsequent priors). Also, appendices A.1 and A.2 provide some calculus and statistical background, if required.

The simplest form of prior distributions for VAR models is known as the Minnesota (or Litterman) prior. In this framework, it is assumed that the VAR residual variance-covariance matrix Σ is known. Hence, the only object left to estimate is the vector of parameters β. To obtain the posterior distribution for β from 3.2.3, one needs two elements: the likelihood function $f(y\,|\beta)$ for the data, and a prior distribution $\pi(\beta)$ for β.

Start with the likelihood function. For the Minnesota prior, 3.1.12 turns out to be the most convenient formulation for the VAR model. As stated in 3.1.3, this formulation implies that the residuals follow a multivariate normal distribution with mean 0 and covariance matrix $\bar{\Sigma}$. This in turn implies from 3.1.12 that y also follows a multivariate normal distribution with mean $\bar{X}\beta$ and covariance $\bar{\Sigma}$. Therefore, one may write the likelihood for y as:

$$f(y\,|\beta,\bar{\Sigma}) = (2\pi)^{-nT/2}\left|\bar{\Sigma}\right|^{-1/2}\exp\left[-\frac{1}{2}\left(y-\bar{X}\beta\right)'\bar{\Sigma}^{-1}\left(y-\bar{X}\beta\right)\right] \tag{3.3.1}$$

Ignoring terms independent from β relegated to proportionality constants, 3.3.1 simplifies to:

$$f(y\,|\beta,\bar{\Sigma}) \propto \exp\left[-\frac{1}{2}\left(y-\bar{X}\beta\right)'\bar{\Sigma}^{-1}\left(y-\bar{X}\beta\right)\right] \tag{3.3.2}$$

Now turn to the prior distribution for β. It is assumed that β follows a multivariate normal distribution, with mean β_0 and covariance matrix Ω_0:

$$\pi(\beta) \sim \mathcal{N}(\beta_0, \Omega_0) \tag{3.3.3}$$

To identify β_0 and Ω_0, Litterman (1986) proposed the following strategy. As most observed macroeconomic variables seem to be characterized by a unit root (in the sense that their changes are impossible to forecast), our prior belief should be that each endogenous variable included in the model presents a unit root in its first own lags, and coefficients equal to zero for further lags and cross-variable lag coefficients. In the absence of prior belief about exogenous variables, the most reasonable strategy is to assume that they are neutral with respect to the endogenous variables, and hence that their coefficients are equal to zero as well. These elements translate into β_0 being a vector of zeros, save for the entries concerning the first own lag of each endogenous variable which are attributed values of 1. Note though that in the case of variables known to be stationary, this unit root hypothesis may not be suitable, so that a value around 0.8 may be preferred to a value of 1.

As an example, consider a VAR model with two endogenous variables and two lags, along with one exogenous variables (for instance a constant, or an exogenous data series). Each equation will involve $k = np + m = 2 \times 2 + 1 = 5$ coefficients to estimate, which implies a total of $q = nk = 2 \times 5 = 10$ coefficients for the whole model, so that β_0 will be a $q \times 1$ vector. For our example, given the structure described by 3.1.11-3.1.12, it is given by:

$$\beta_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{3.3.4}$$

For the variance-covariance matrix Ω_0, it is assumed that no covariance exists between terms in β, so that Ω_0 is diagonal. Also, Litterman (1986) argued that the further the lag, the more confident we should be that coefficients linked to this lag have a value of zero. Therefore, variance should be smaller on further lags. Also, this confidence should be greater for coefficients relating variables to past values of other variables. Finally, it should be assumed that little is known about exogenous variables, so that the variance on these terms should be large. Based on these principles, Litterman (1986) distinguished three different cases:

1. For parameters in β relating endogenous variables to their own lags, the variance is given by:

$$\sigma^2_{a_{ii}} = \left(\frac{\lambda_1}{l^{\lambda_3}}\right)^2 \tag{3.3.5}$$

   where λ_1 is an overall tightness parameter, l is the lag considered by the coefficient, and λ_3 is a scaling coefficient controlling the speed at which coefficients for lags greater than 1 converge to 0 with greater certainty.

2. For parameters related to cross-variable lag coefficients, the variance is given by:

$$\sigma^2_{a_{ij}} = \left(\frac{\sigma_i^2}{\sigma_j^2}\right)\left(\frac{\lambda_1\lambda_2}{l^{\lambda_3}}\right)^2 \tag{3.3.6}$$

   where σ_i^2 and σ_j^2 denote the OLS residual variance of the auto-regressive models estimated for variables i and j, and λ_2 represents a cross-variable specific variance parameter.

3. For exogenous variables (including constant terms), the variance is given by:

$$\sigma^2_{c_i} = \sigma_i^2(\lambda_1\lambda_4)^2 \tag{3.3.7}$$

where σ_i^2 is again the OLS residual variance of an auto-regressive model previously estimated for variable i, and λ_4 is a large (potentially infinite) variance parameter.

Ω_0 is thus a $q \times q$ diagonal matrix with three different types of variance terms on its main diagonal. For instance, for the VAR model with 2 variables, 2 lags and one exogenous variable specified above, Ω_0 is given by:

$$\Omega_0 = \begin{pmatrix} (\lambda_1)^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \left(\frac{\sigma_1^2}{\sigma_2^2}\right)(\lambda_1\lambda_2)^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \left(\frac{\lambda_1}{2^{\lambda_3}}\right)^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \left(\frac{\sigma_1^2}{\sigma_2^2}\right)\left(\frac{\lambda_1\lambda_2}{2^{\lambda_3}}\right)^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_1^2(\lambda_1\lambda_4)^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \left(\frac{\sigma_2^2}{\sigma_1^2}\right)(\lambda_1\lambda_2)^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & (\lambda_1)^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \left(\frac{\sigma_2^2}{\sigma_1^2}\right)\left(\frac{\lambda_1\lambda_2}{2^{\lambda_3}}\right)^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \left(\frac{\lambda_1}{2^{\lambda_3}}\right)^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_2^2(\lambda_1\lambda_4)^2 \end{pmatrix} \tag{3.3.8}$$

Different choices are possible for λ_1,λ_2,λ_3 and λ_4. However, values typically found in the literature revolve around:

$$\lambda_1 = 0.1 \tag{3.3.9}$$

$$\lambda_2 = 0.5 \tag{3.3.10}$$

$$\lambda_3 = 1 \text{ or } 2 \tag{3.3.11}$$

$$\lambda_4 = 10^2 \text{ to } \infty \tag{3.3.12}$$

Finally, since the Minnesota prior assumes that the variance-covariance matrix of residuals Σ is known, one has to decide how to define it. The original Minnesota prior assumes that Σ is diagonal which, as will be seen later, conveniently implies independence between the VAR coefficients of different equations. This property was useful at a time of limited computational power as it allows estimating the model equation by equation (this possibility is not used here). A first possibility is thus to set the diagonal of Σ equal to the residual variance of individual AR models run on each variable in the VAR. A second possibility is to use the variance-covariance matrix of a conventional VAR estimated by OLS, but to retain only the diagonal terms as Σ. Finally, as the model estimates

all the equations simultaneously in this setting, the assumption of a diagonal matrix is not required. Therefore, a third and last possibility consists in using directly the entire variance-covariance matrix of a VAR estimated by OLS.

Once β_0 and Ω_0 are determined, and that proper values re-attributed to Σ, one may compute the prior distribution of β. The normality assumption implies that its density is given by:

$$\pi(\beta) = (2\pi)^{-nk/2}|\Omega_0|^{-1/2} \exp\left[-\frac{1}{2}(\beta - \beta_0)'\Omega_0^{-1}(\beta - \beta_0)\right] \tag{3.3.13}$$

Relegating terms independent of β to the proportionality constant, 3.3.13 rewrites:

$$\pi(\beta) \propto \exp\left[-\frac{1}{2}(\beta - \beta_0)'\Omega_0^{-1}(\beta - \beta_0)\right] \tag{3.3.14}$$

Now, directly applying 3.2.3, that is, combining the likelihood 3.3.2 with the prior 3.3.14, the posterior distribution for β obtains as:

$$\begin{aligned}
\pi(\beta\,|y) &\propto f(y\,|\beta)\pi(\beta) \\
&\propto \exp\left[-\frac{1}{2}\left(y - \bar{X}\beta\right)'\bar{\Sigma}^{-1}\left(y - \bar{X}\beta\right)\right] \times \exp\left[-\frac{1}{2}(\beta - \beta_0)'\Omega_0^{-1}(\beta - \beta_0)\right] \\
&= \exp\left[-\frac{1}{2}\left\{\left(y - \bar{X}\beta\right)'\bar{\Sigma}^{-1}\left(y - \bar{X}\beta\right) + (\beta - \beta_0)'\Omega_0^{-1}(\beta - \beta_0)\right\}\right]
\end{aligned} \tag{3.3.15}$$

Equation 3.3.15 represents the kernel of the posterior distribution, but it does not have the form of a known distribution. Yet, it is possible to show that after some manipulations, it reformulates as:

$$\pi(\beta\,|y) \propto \exp\left[-\frac{1}{2}\left\{\left(\beta - \bar{\beta}\right)'\bar{\Omega}^{-1}\left(\beta - \bar{\beta}\right)\right\}\right] \tag{3.3.16}$$

with:

$$\bar{\Omega} = \left[\Omega_0^{-1} + \Sigma^{-1} \otimes X'X\right]^{-1} \tag{3.3.17}$$

and:

$$\bar{\beta} = \bar{\Omega}\left[\Omega_0^{-1}\beta_0 + (\Sigma^{-1} \otimes X')y\right] \tag{3.3.18}$$

This is the kernel of a multivariate normal distribution with mean $\bar{\beta}$ and covariance matrix $\bar{\Omega}$. Therefore, the posterior distribution of β is given by:

$$\pi(\beta\,|y) \sim \mathcal{N}(\bar{\beta}, \bar{\Omega}) \tag{3.3.19}$$

From this posterior distribution, point estimates and credibility intervals for β can obtain by direct application of the methods developed in subsection 3.2.

3.4 The normal-Wishart prior

Although the Minnesota prior offers a simple way to derive the posterior distribution of the VAR coefficients, it suffers from the main drawback of assuming that the residual covariance matrix Σ is known. One possibility to relax this assumption is to use a normal-Wishart prior distribution. In this setting, it is assumed that both β and Σ are unknown.

The analysis starts again with the likelihood function $f(y\,|\beta,\bar{\Sigma})$ for the data sample. Because there is no change in the assumptions relative to the data, its density is still given by 3.3.1, repeated here for convenience:

$$f(y\,|\beta,\Sigma) = (2\pi)^{-nT/2}\left|\bar{\Sigma}\right|^{-1/2}\exp\left[-\frac{1}{2}(y-\bar{X}\beta)'\bar{\Sigma}^{-1}\left(y-\bar{X}\beta\right)\right] \tag{3.4.1}$$

However, since Σ is now assumed to be unknown, $\bar{\Sigma} = \Sigma \otimes I_T$ cannot be relegated anymore to the proportionality constant. Therefore, 3.4.1 now simplifies to:

$$f(y\,|\beta,\Sigma) \propto \left|\bar{\Sigma}\right|^{-1/2}\exp\left[-\frac{1}{2}(y-\bar{X}\beta)'\bar{\Sigma}^{-1}\left(y-\bar{X}\beta\right)\right] \tag{3.4.2}$$

After quite a bit of manipulations, one can show that this density rewrites as:

$$\begin{aligned} f(y\,|\beta,\Sigma) \propto\ & |\Sigma|^{-k/2}\exp\left[-\frac{1}{2}\left(\beta-\hat{\beta}\right)'\left(\Sigma\otimes(X'X)^{-1}\right)^{-1}\left(\beta-\hat{\beta}\right)\right] \\ & \times|\Sigma|^{-[(T-k-n-1)+n+1]/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right\}\right] \end{aligned} \tag{3.4.3}$$

where Y and X are defined in 3.1.8, and $\hat{B}$ and $\hat{\beta}$ are defined as in 3.1.9 and 3.1.15. Equation 3.4.3 can be recognised as the kernel of a multivariate normal distribution (for β), and the kernel of an inverse Wishart distribution (for Σ), both being centred around OLS estimators. It seems then natural to assume similar prior distributions for β and Σ, in the hope that it could yield distributions of the same families for the posterior distribution, which will indeed be the case (such identical families for the prior and the posterior are known as conjugate priors).

For β, one thus assumes a multivariate normal distribution for the prior:

$$\beta \sim \mathcal{N}(\beta_0, \Sigma\otimes\Phi_0) \tag{3.4.4}$$

Similarly to the Minnesota prior, β_0 is an $q \times 1$ vector. Φ_0 is a $k \times k$ diagonal matrix, and Σ is the usual VAR residual variance-covariance matrix, which implies that $\Sigma \otimes \Phi_0$ is a $nk \times nk$ or $q \times q$ covariance matrix.

The choice of β_0 is usually simple, while Φ_0 raises some issues. For β_0, a conventional Minnesota scheme will be typically adopted, setting values around 1 for own first lag coefficients, and 0 for cross variable and exogenous coefficients. For Φ_0, note the difference between 3.4.4 and the Minnesota parameter Ω_0 in 3.3.3: while Ω_0 represents the full variance-covariance matrix of β, now Φ_0 only represents the variance for the parameters of one single equation in the VAR. Each such variance is then scaled by the variable-specific variance contained in Σ. This Kronecker structure implies that the variance-covariance matrix of β cannot be specified anymore as in 3.3.8: the variance-covariance matrix of one equation has now to be proportional to the variance-covariance matrix of the other equations. As shown in Appendix A.4, without this structure, it would not be possible to obtain a well identified posterior distribution. One may however approach a Minnesota type of variance matrix by adopting the following strategy (see e.g. Karlsson (2012)):

For lag terms (both own and cross-lags), define the variance as:

$$\sigma^2_{a_{ij}} = \left(\frac{1}{\sigma_j^2}\right)\left(\frac{\lambda_1}{l^{\lambda_3}}\right)^2 \tag{3.4.5}$$

where σ_j^2 is the unknown residual variance for variable j in the BVAR model, approximated by individual AR regressions. For exogenous variables, define the variance as:

$$\sigma_c^2 = (\lambda_1 \lambda_4)^2 \tag{3.4.6}$$

For instance, with the two-variable VAR with two lags and one exogenous variable used as an example in subsubsection 3.3, Φ_0 would be:

$$\Phi_0 = \begin{pmatrix} \left(\frac{1}{\sigma_1^2}\right)(\lambda_1)^2 & 0 & 0 & 0 & 0 \\ 0 & \left(\frac{1}{\sigma_2^2}\right)(\lambda_1)^2 & 0 & 0 & 0 \\ 0 & 0 & \left(\frac{1}{\sigma_1^2}\right)\left(\frac{\lambda_1}{2^{\lambda_3}}\right)^2 & 0 & 0 \\ 0 & 0 & 0 & \left(\frac{1}{\sigma_2^2}\right)\left(\frac{\lambda_1}{2^{\lambda_3}}\right)^2 & 0 \\ 0 & 0 & 0 & 0 & (\lambda_1\lambda_4)^2 \end{pmatrix} \tag{3.4.7}$$

If one assumes a diagonal Σ as in the original Minnesota prior, 3.4.7 then implies that $\Sigma \otimes \Phi_0$ is given by:

$$\Sigma \otimes \Phi_0 = \begin{pmatrix} (\lambda_1)^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \left(\frac{\sigma_1^2}{\sigma_2^2}\right)(\lambda_1)^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \left(\frac{\lambda_1}{2^{\lambda_3}}\right)^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \left(\frac{\sigma_1^2}{\sigma_2^2}\right)\left(\frac{\lambda_1}{2^{\lambda_3}}\right)^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_1^2(\lambda_1\lambda_4)^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \left(\frac{\sigma_2^2}{\sigma_1^2}\right)(\lambda_1)^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & (\lambda_1)^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \left(\frac{\sigma_2^2}{\sigma_1^2}\right)\left(\frac{\lambda_1}{2^{\lambda_3}}\right)^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \left(\frac{\lambda_1}{2^{\lambda_3}}\right)^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_2^2(\lambda_1\lambda_4)^2 \end{pmatrix} \tag{3.4.8}$$

Comparing with 3.3.8, one can see that the normal-Wishart variance-covariance matrix of β is a special case of the Minnesota variance-covariance matrix where Σ is diagonal and the parameter λ_2 is constrained to take a value of 1. In this sense, the normal-Wishart prior appears as a Minnesota prior that would not be able to provide tighter priors on cross-variable parameters, which may be an undesirable assumption. For this reason, it is advised to set λ_1 at a smaller value than for the Minnesota prior (e.g. between 0.01 and 0.1), in order to compensate for the lack of extra shrinkage from λ_2. For the remaining hyperparameters λ_3 and λ_4, the same values as the Minnesota prior may be attributed.

With β_0 and Φ_0 at hands, the prior density for β writes as:

$$\pi(\beta) \propto |\Sigma|^{-k/2} \exp\left[-\frac{1}{2}(\beta - \beta_0)'(\Sigma \otimes \Phi_0)^{-1}(\beta - \beta_0)\right] \tag{3.4.9}$$

Turn now to the prior for Σ. The retained distribution is an inverse Wishart distribution characterised as:

$$\Sigma \sim \mathcal{IW}(S_0, \alpha_0) \tag{3.4.10}$$

where S_0 is the $n \times n$ scale matrix for the prior, and α_0 is prior degrees of freedom. While any choice can be made for these hyperparameters according to prior information, the literature once again proposes standard schemes. For instance, following Karlsson (2012), S_0 can be defined as:

$$S_0 = (\alpha_0 - n - 1)\begin{pmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sigma_n^2 \end{pmatrix} \tag{3.4.11}$$

On the other hand, the prior degrees of freedom α_0 is defined as:

$$\alpha_0 = n + 2 \tag{3.4.12}$$

This specifies the prior degrees of freedom as the minimum possible to obtain well-defined mean and variance. Indeed, this value implies that:

$$E(\Sigma) = \begin{pmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sigma_n^2 \end{pmatrix} \tag{3.4.13}$$

In other words, the expectation of Σ is the diagonal covariance matrix obtained from individual AR regressions and used as an estimate for Σ in the Minnesota prior. As with the Minnesota prior, it is possible to implement alternative schemes. For instance, the matrix 3.4.11 can be simply replaced with an identity matrix of size n.

With these elements, the kernel of the prior density for Σ is given by:

$$\pi(\Sigma) \propto |\Sigma|^{-(\alpha_0+n+1)/2} \exp\left[-\frac{1}{2} tr\left\{\Sigma^{-1} S_0\right\}\right] \tag{3.4.14}$$

From 3.2.5, the posterior obtains by combining the likelihood 3.4.3 with the priors 3.4.9 and 3.4.14. After lengthy manipulations, one obtains:

$$\begin{aligned} \pi(\beta, \Sigma \,|y\,) \propto |\Sigma|^{-(k)/2} \exp\left[-\frac{1}{2}(\beta - \bar{\beta})' \left(\Sigma \otimes \bar{\Phi}\right)^{-1} (\beta - \bar{\beta})\right] \\ \times |\Sigma|^{-(\bar{\alpha}+n+1)/2} \exp\left[-\frac{1}{2} tr\left\{\Sigma^{-1} \bar{S}\right\}\right] \end{aligned} \tag{3.4.15}$$

with:

$$\bar{\Phi} = \left[\Phi_0^{-1} + X'X\right]^{-1} \tag{3.4.16}$$

$$\bar{\beta} = vec(\bar{B}), \;\; \bar{B} = \bar{\Phi}\left[\Phi_0^{-1} B_0 + X'Y\right] \tag{3.4.17}$$

$$\bar{\alpha} = T + \alpha_0 \tag{3.4.18}$$

and

$$\bar{S} = Y'Y + S_0 + B_0' \Phi_0^{-1} B_0 - \bar{B}' \bar{\Phi}^{-1} \bar{B} \tag{3.4.19}$$

This is recognised as the kernel of a multivariate normal distribution for β (conditional on Σ), multiplied by the kernel of an inverse Wishart distribution for Σ. The fact that the posterior takes the same form as the prior justifies the denomination of conjugate prior. One then wants to use the joint posterior 3.4.16 to derive the marginal distributions for β and Σ. This is done by direct application of 3.2.7.

Obtaining the marginal for Σ is trivial: integrating out β is easy as it appears only in the first term as a multivariate normal. Following integration, only the second term remains, which determines straight away the posterior density for Σ as:

$$\pi(\Sigma\,|y) \propto |\Sigma|^{-(\bar{\alpha}+n+1)/2} \exp\left[-\frac{1}{2} tr\left\{\Sigma^{-1}\bar{S}\right\}\right] \tag{3.4.20}$$

This is once again immediately recognised as the kernel of an inverse Wishart distribution:

$$\pi(\Sigma\,|y) \sim \mathcal{IW}\left(\bar{\alpha}, \bar{S}\right) \tag{3.4.21}$$

Integrating out Σ to derive the marginal for β is a more complicated matter, but it can be shown after some work on 3.4.15 that:

$$\pi(B\,|y) \propto \left|I_n + \bar{S}^{-1}(B-\bar{B})'\bar{\Phi}^{-1}(B-\bar{B})\right|^{-\frac{[T+\alpha_0-n+1]+n+k-1}{2}} \tag{3.4.22}$$

where B is defined in 3.1.8. This is the kernel of a matrix-variate student distribution with mean $\bar{B}$, scale matrices $\bar{S}$ and $\bar{\Phi}$, and degrees of freedom $T+\alpha_0-n+1$:

$$B \sim \mathcal{MT}(\bar{B}, \bar{S}, \bar{\Phi}, \tilde{\alpha}) \tag{3.4.23}$$

with:

$$\tilde{\alpha} = T + \alpha_0 - n + 1 \tag{3.4.24}$$

This then implies that each individual element $B_{i,j}$ of B follows a univariate student distribution with mean $\bar{B}_{i,j}$, scale parameter $\bar{\Phi}_{i,i} \times \bar{S}_{j,j}$ and degrees of freedom $\tilde{\alpha}$.

$$B_{i,j} \sim t(\bar{B}_{i,j}, \bar{\Phi}_{i,i} \times \bar{S}_{j,j}, \tilde{\alpha}) \tag{3.4.25}$$

3.4.21 and 3.4.25 can then eventually be used to compute point estimates and draw inference for β and Σ, using once again the methods developed in subsection 3.2.

3.5 An independent normal-Wishart prior with unknown Σ and arbitrary Ω_0

Even though the normal-Wishart prior is more flexible than the Minnesota prior in the sense that Σ is not assumed to be known, it has its own limitations. As shown by 3.4.4, assuming an unknown Σ comes at the cost of imposing a Kronecker structure on the prior distribution for β, constraining its covariance matrix to be equal to $\Sigma \otimes \Phi_0$. This structure creates, for each equation, a dependence between the variance of the residual term and the variance of the VAR coefficients, which may be an undesirable assumption. An alternative way to see the restrictions generated by this specific formulation is to notice that the covariance matrix for the VAR coefficients $\Sigma \otimes \Phi_0$ (given by 3.1.4) corresponds to the more general covariance matrix used for the Minnesota (see equation 3.4.8) in the special case where $\lambda_2 = 1$, that is, where the variance on cross-variable coefficients is as large as the variance on its own lags, for each equation.

Ideally, one would thus like to estimate a BVAR model where at the same time Σ would be treated as unknown, and an arbitrary structure could be proposed for Ω_0, with no assumed dependence between residual variance and coefficient variance. Such a prior, known as the independent normal-Wishart prior, is feasible but implies the sacrifice of analytical solutions in favour of numerical methods. The analysis starts the usual way: first obtain the likelihood from the data. There is no change here and likelihood is still given by 3.4.3:

$$f(y\,|\beta,\Sigma) \propto |\Sigma|^{-T/2} \exp\left[-\frac{1}{2}\left(\beta-\hat{\beta}\right)'\left(\Sigma\otimes(X'X)^{-1}\right)^{-1}\left(\beta-\hat{\beta}\right)\right]$$
$$\times \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right\}\right] \tag{3.5.1}$$

Concerning the prior for β, it now departs from the normal-Wishart assumption by assuming that β follows a multivariate normal distribution with mean β_0 and covariance matrix Ω_0, but Ω_0 is now an arbitrary $q \times q$ matrix, not necessarily adopting the Kronecker structure described by 3.4.4. β_0 on the other hand is the usual $q \times 1$ mean vector. Hence:

$$\beta \sim \mathcal{N}(\beta_0, \Omega_0) \tag{3.5.2}$$

In typical applications, Ω_0 will take the form of the Minnesota covariance matrix described in 3.3.8, but any choice is possible. Similarly, β_0 will typically be defined as the Minnesota β_0 vector 3.3.4, but any structure of vector β_0 could be adopted.

Given β_0 and Ω_0, the prior density for β is given by:

$$\pi(\beta) \propto \exp\left[-\frac{1}{2}(\beta-\beta_0)'\Omega_0^{-1}(\beta-\beta_0)\right] \tag{3.5.3}$$

Similarly to the inverse-Wishart prior, the prior distribution for Σ is an inverse Wishart distribution, with scale matrix S_0 and degrees of freedom α_0:

$$\Sigma \sim \mathcal{IW}(S_0, \alpha_0) \tag{3.5.4}$$

In typical applications, S_0 and α_0 will be determined as 3.4.11 and 3.4.12. Following, the prior density of Σ is given by:

$$\pi(\Sigma) \propto |\Sigma|^{-(\alpha_0+n+1)/2} \exp\left[-\frac{1}{2} tr\left\{\Sigma^{-1}S_0\right\}\right] \tag{3.5.5}$$

As usual, the posterior obtains from direct application of 3.2.5, using 3.5.1, 3.5.3 and 3.5.5 :

$$\begin{aligned}
\pi(\beta, \Sigma\,|y) &\propto f(y\,|\beta, \Sigma)\pi(\beta)\pi(\Sigma) \\
&\propto |\Sigma|^{-T/2} \exp\left[-\frac{1}{2}\left(\beta-\hat{\beta}\right)'\left(\Sigma\otimes(X'X)^{-1}\right)^{-1}\left(\beta-\hat{\beta}\right)\right] \\
&\times \exp\left[-\frac{1}{2} tr\left\{\Sigma^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right\}\right] \\
&\times \exp\left[-\frac{1}{2}(\beta-\beta_0)'\Omega_0^{-1}(\beta-\beta_0)\right] \\
&\times |\Sigma|^{-(\alpha_0+n+1)/2} \exp\left[-\frac{1}{2} tr\left\{\Sigma^{-1}S_0\right\}\right]
\end{aligned} \tag{3.5.6}$$

Regrouping terms, 3.5.6 rewrites as:

$$\begin{aligned}
\pi(\beta, \Sigma\,|y) \propto &|\Sigma|^{-(T+\alpha_0+n+1)/2} \exp\left[-\frac{1}{2}\left\{\left(\beta-\hat{\beta}\right)'\left(\Sigma^{-1}\otimes X'X\right)\left(\beta-\hat{\beta}\right)+(\beta-\beta_0)'\Omega_0^{-1}(\beta-\beta_0)\right\}\right] \\
&\times \exp\left[-\frac{1}{2} tr\left\{\Sigma^{-1}\left[S_0+(Y-X\hat{B})'(Y-X\hat{B})\right]\right\}\right]
\end{aligned} \tag{3.5.7}$$

And after quite a bit of manipulations, 3.5.7 becomes:

$$\pi(\beta,\Sigma\,|y\,) \propto |\Sigma|^{-(T+\alpha_0+n+1)/2} \exp\left[-\frac{1}{2}(\beta-\bar{\beta})'\bar{\Omega}^{-1}(\beta-\bar{\beta})\right]$$
$$\times \exp\left[-\frac{1}{2}\hat{\beta}'\left(\Sigma^{-1}\otimes X'X\right)\hat{\beta}+\beta_0'\Omega_0^{-1}\beta_0-\bar{\beta}'\bar{\Omega}^{-1}\bar{\beta}\right]$$
$$\times \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left[\left(Y-X\hat{B}\right)'\left(Y-X\hat{B}\right)+S_0\right]\right\}\right] \tag{3.5.8}$$

with $\hat{B}$ and $\hat{\beta}$ defined the usual way as in 3.1.9 and 3.1.14, and:

$$\bar{\Omega} = \left[\Omega_0^{-1}+\Sigma^{-1}\otimes X'X\right]^{-1} \tag{3.5.9}$$

and

$$\bar{\beta} = \bar{\Omega}\left[\Omega_0^{-1}\beta_0+(\Sigma^{-1}\otimes X')y\right] \tag{3.5.10}$$

Note that the general structure of Ω_0 prevents from reformulating the second row of 3.5.8 as a trace expression, which would have allowed to obtain a term corresponding to $\bar{S}$ in 3.4.19. Consequently, as it is, 3.5.8 provides no way to derive an analytical marginal distribution for β and Σ.

However, even if it is not possible to derive the unconditional marginal distribution for β and Σ, it is possible to derive their conditional distributions. To do so, one considers the joint posteriors distribution for all parameters and retain only terms involving parameters whose conditional distribution must be determined. All terms that do not involve these parameters do not contain information about their distribution and are thus relegated to the proportionality constant.

Apply first this method to β: considering 3.5.8, and ignoring terms not involving β, one is left with:

$$\pi(\beta\,|\Sigma,y\,) \propto \exp\left[-\frac{1}{2}(\beta-\bar{\beta})'\bar{\Omega}^{-1}(\beta-\bar{\beta})\right] \tag{3.5.11}$$

This is recognised as the kernel of a multivariate distribution with mean $\bar{\beta}$ and variance-covariance matrix $\bar{\Omega}$. Hence:

$$\pi(\beta\,|\Sigma,y\,) \sim \mathcal{N}(\bar{\beta},\bar{\Omega}) \tag{3.5.12}$$

Now determine the conditional distribution for Σ. To do so, it is easier to work directly with 3.5.7. Thus, ignore terms not involving Σ in 3.5.7, and reshape the remaining elements to obtain:

$$\pi(\Sigma\,|\beta,y)\propto|\Sigma|^{-[(T+\alpha_0)+n+1]/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left[(Y-XB)'(Y-XB)+S_0\right]\right\}\right] \tag{3.5.13}$$

This is recognised as the kernel of an inverse Wishart distribution:

$$\pi(\Sigma\,|\beta,y)\sim\mathcal{IW}((\hat{S},\hat{\alpha}) \tag{3.5.14}$$

with scale matrix:

$$\hat{S}=(Y-XB)'(Y-XB)+S_0$$

And degrees of freedom:

$$\hat{\alpha}=T+\alpha_0 \tag{3.5.15}$$

With these conditional distributions at hand, it is possible to use a numerical method known as Gibbs sampling to obtain random draws from the unconditional posterior distributions of the parameters of interest. While a rigorous presentation of the theory of Gibbs sampling implies an exposition of Markov Chain Monte Carlo methods and lies beyond the scope of this guide , it is still possible to present the ideas underlying the process.[1]

The presentation is made for the simplest case: a model with only two blocks of parameters θ_1 and θ_2 (corresponding for a VAR model to β and Σ). However, it is straightforward to generalise it to any number of blocks. Assume hence that one considers a model with two blocks of parameters θ_1 and θ_2, and wants to determine the unconditional posterior distribution for each of these blocks. The unconditional distribution has an unknown form, but the conditional distribution has a known form, and numerical softwares are able to sample from it. Then, consider the following procedure:

Algorithm 1.5.1 (Gibbs sampling with two blocks):

1. Fix any arbitrary initial value for $\theta_2^{(0)}$.
2. At iteration 1, determine the conditional distribution $f(\theta_1\left|\theta_2^{(0)}\right.)$, using the value $\theta_2^{(0)}$. Then obtain a draw $\theta_1^{(1)}$ from this distribution.
3. At iteration 1, determine the conditional distribution $f(\theta_2\left|\theta_1^{(1)}\right.)$, using the value $\theta_1^{(1)}$. Then obtain a draw $\theta_2^{(1)}$ from this distribution. This marks the end of iteration 1.

[1] See e.g. Kadiyala and Karlsson (1997) for a brief presentation of the Gibbs sampling methodology, and Greenberg (2008), chapters 6 and 7, for a much more comprehensive treatment of the topic.

4. At iteration 2, determine the conditional distribution $f(\theta_1 \left| \theta_2^{(1)} \right.)$, using the value $\theta_2^{(1)}$. Then obtain a draw $\theta_1^{(2)}$ from this distribution.

5. At iteration 2, determine the conditional distribution $f(\theta_2 \left| \theta_1^{(2)} \right.)$, using the value $\theta_1^{(2)}$. Then obtain a draw $\theta_2^{(2)}$ from this distribution. This marks the end of iteration 2. Then pursue the process, repeating any number of times:

6. At iteration n, determine the conditional distribution $f(\theta_1 \left| \theta_2^{(n-1)} \right.)$, using the value $\theta_2^{(n-1)}$. Then obtain a draw $\theta_1^{(n)}$ from this distribution.

7. At iteration n, determine the conditional distribution $f(\theta_2 \left| \theta_1^{(n)} \right.)$, using the value $\theta_1^{(n)}$. Then obtain a draw $\theta_2^{(n)}$ from this distribution.

The essential property of this process is that after a certain number of iterations, known as the transient, or burn-in sample, the draws will not be realised any more from the conditional distribution, but from the unconditional distribution of each block. This is due to the fact the successive conditional draws result in a gradual convergence of the process towards the unconditional distributions of each block, and then remain at this distribution, hence the denomination of invariant distribution. Once the convergence stage is over, it suffices to pursue the process any number of times to obtain any number of draws from the unconditional posterior distribution, discard the burn-in sample to keep only the draws from the unconditional distribution, and hence build an empirical posterior distribution.

This remarkable property constitutes the core of modern numerical methods applied to Bayesian analysis. All it requires is to know the conditional posterior distributions for the model, and sufficient computer speed to accomplish the steps. Note that the process is flexible. While the example has exposed the case of drawing first θ_1, then θ_2, the converse could have been done: the order is arbitrary and convergence will be achieved anyway. Actually, the choice of the order should always be based on convenience. Also, the process can be adapted to a model comprising more than two blocks. For a model with n blocks, suffice is to condition each draw on each of the other $(n-1)$ block values. For instance, for a model with three blocks, each draw should be realised as : $\theta_1^{(n)}$ from $f(\theta_1 \left| \theta_2^{(n-1)}, \theta_3^{(n-1)} \right.)$, then $\theta_2^{(n)}$ from $f(\theta_2 \left| \theta_1^{(n)}, \theta_3^{(n-1)} \right.)$, then $\theta_3^{(n)}$ from $f(\theta_1 \left| \theta_1^{(n)}, \theta_2^{(n)} \right.)$ (once again, the order could be different).

After this succinct presentation, it is possible to introduce the Gibbs sampling algorithm used for the independent normal-Wishart prior. For this model, the two blocks of parameters of interest are β and Σ, and their conditional distributions are given by 3.5.12 and 3.5.14. Also, while there is no theorem to define what should be the size of the burn-in sample and the total number of iterations,

in practice a total number of iterations (It) of 2000 and a burn-in sample (Bu) of 1000 provide sufficient precision. The following algorithm is thus proposed:

Algorithm 1.5.2 (Gibbs sampling for VAR parameters with an independent normal-Wishart prior):

1. Define the total number of iterations It of the algorithm, and the size Bu of the burn-in sample.
2. Define an initial value $\beta_{(0)}$ for the algorithm. This will typically be the VAR model OLS estimates. Reshape to obtain $B_{(0)}$. Then start running the algorithm.
3. At iteration n, draw the value $\Sigma_{(n)}$, conditional on $B_{(n-1)}$. $\Sigma_{(n)}$ is drawn from an inverse Wishart distribution with scale matrix $\hat{S}$ and degrees of freedom $\hat{\alpha}$, as defined in 3.5.15 and 3.5.15: $\pi(\Sigma_{(n)} \left| \beta_{(n-1)}, y \right.) \sim \mathcal{IW}(\hat{S}, \hat{\alpha})$ with: $\hat{\alpha} = T + \alpha_0$ and $\hat{S} = S_0 + (Y - XB_{(n-1)})'(Y - XB_{(n-1)})$
4. At iteration n, draw $\beta_{(n)}$ conditional on $\Sigma_{(n)}$, and reshape it into $B_{(n)}$ for the next draw of Σ. Draw $\beta_{(n)}$ from a multivariate normal with mean $\bar{\beta}$ and covariance matrix $\bar{\Omega}$, as defined in 3.5.9 and 3.5.10: $\pi(\beta_{(n)} \left| \Sigma_{(n)}, y \right.) \sim \mathcal{N}(\bar{\beta}, \bar{\Omega})$ with: $\bar{\Omega} = \left[\Omega_0^{-1} + \Sigma_{(n)}^{-1} \otimes X'X\right]^{-1}$ and $\bar{\beta} = \bar{\Omega}\left[\Omega_0^{-1}\beta_0 + (\Sigma_{(n)}^{-1} \otimes X')y\right]$
5. Repeat until It iterations are realized, then discard the first Bu iterations.

3.6 The normal-diffuse prior

A possible alternative to the Minnesota and the normal-Wishart priors is the so-called normal-diffuse prior distribution. The specificity of this prior distribution is that it relies on a diffuse (uninformative) prior for Σ. It hence represents a good alternative to the independent normal-Wishart developed in the previous subsection when one wants to remain agnostic about the value that Σ should be given. The likelihood function and the prior distribution for β are similar to those developed in the previous subsection and are thus respectively given by:

$$f(y \left| \beta, \Sigma \right.) \propto |\Sigma|^{-T/2} \exp\left[-\frac{1}{2}\left(\beta - \hat{\beta}\right)'\left(\Sigma \otimes (X'X)^{-1}\right)^{-1}\left(\beta - \hat{\beta}\right)\right] \times \exp\left[-\frac{1}{2} tr\left\{\Sigma^{-1}(Y - X\hat{B})'(Y - X\hat{B})\right\}\right] \tag{3.6.1}$$

and

$$\pi(\beta) \propto \exp\left[-\frac{1}{2}(\beta - \beta_0)'\Omega_0^{-1}(\beta - \beta_0)\right] \tag{3.6.2}$$

The main change intervenes in the prior distribution for Σ, which is now defined as the so-called Jeffrey's or diffuse prior:

$$\pi(\Sigma) \propto |\Sigma|^{-(n+1)/2} \tag{3.6.3}$$

This prior is called an improper prior as it integrates to infinity rather than to one. Yet, this does not necessarily preclude the posterior distribution to be proper, which is indeed the case here: combining the likelihood 3.6.1 with the priors 3.6.2 and 3.6.3, and applying Bayes rule 3.2.5, the joint posterior is given by:

$$\begin{aligned}\pi(\beta, \Sigma\,|y) \propto & |\Sigma|^{-(T+n+1)/2} \exp\left[-\frac{1}{2}\left\{\left(\beta-\hat{\beta}\right)'\left(\Sigma^{-1}\otimes X'X\right)\left(\beta-\hat{\beta}\right)+(\beta-\beta_0)'\Omega_0^{-1}(\beta-\beta_0)\right\}\right] \\ & \times \exp\left[-\frac{1}{2} tr\left\{\Sigma^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right\}\right]\end{aligned} \tag{3.6.4}$$

Also, it can be shown that 3.6.4 may alternatively rewrite as:

$$\begin{aligned}\pi(\beta, \Sigma\,|y) \propto & |\Sigma|^{-(T+n+1)/2} \exp\left[-\frac{1}{2}(\beta-\bar{\beta})'\bar{\Omega}^{-1}(\beta-\bar{\beta})\right] \\ & \times \exp\left[-\frac{1}{2}\left\{\hat{\beta}'\left(\Sigma^{-1}\otimes X'X\right)\hat{\beta}+\beta_0'\Omega_0^{-1}\beta_0-\bar{\beta}'\bar{\Omega}^{-1}\bar{\beta}\right\}\right] \\ & \times \exp\left[-\frac{1}{2} tr\left\{\Sigma^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right\}\right]\end{aligned} \tag{3.6.5}$$

with $\bar{\beta}$ and $\bar{\Omega}$ defined as in 3.5.9 and 3.5.10. The conditional posterior distribution for β obtains by considering 3.6.5 and ignoring any term not involving β. This yields:

$$\pi(\beta\,|\Sigma, y) \propto \exp\left[-\frac{1}{2}(\beta-\bar{\beta})'\bar{\Omega}^{-1}(\beta-\bar{\beta})\right] \tag{3.6.6}$$

which is recognised as the kernel of a multivariate normal distribution:

$$\pi(\beta|\Sigma, y) \sim \mathcal{N}(\bar{\beta}, \bar{\Omega}) \tag{3.6.7}$$

On the other hand, the conditional posterior distribution for Σ obtains from 3.6.4, relegating to the proportionality constant any term not involving Σ, and then rearranging. This yields:

$$\pi(\beta, \Sigma\,|y) \propto |\Sigma|^{-(T+n+1)/2} \exp\left[-\frac{1}{2} tr\left\{\Sigma^{-1}\left[(Y-XB)'(Y-XB)\right]\right\}\right] \tag{3.6.8}$$

This is the kernel of an inverse Wishart distribution:

$$\pi(\Sigma\,|\beta, y) \sim \mathcal{IW}(\tilde{S}, T) \tag{3.6.9}$$

with $\tilde{S}$ the scale matrix defined as:

$$\tilde{S} = (Y - XB)'(Y - XB) \tag{3.6.10}$$

As one can see, the main difference between the results developed in this subsection and those of the previous subsection lie in the expressions related to the posterior distribution of Σ (compare 3.6.9-3.6.10 with 3.5.15-3.5.15). The posterior distribution for Σ is inverse Wishart in both cases, but the results obtained in the previous subsection involve the prior beliefs S_0 and α_0, while those obtained in the present subsection don't, due to agnosticism about Σ.

The Gibbs algorithm used to derive the unconditional posterior is then similar to algorithm 3.5, save for the scale matrix and degrees of freedom of the inverse Wishart distribution which have to be modified, in compliance with 3.6.9 and 3.6.10.

3.7 A dummy observation prior

Most of the Bayesian VAR applications covered so far have been relying on the prior structure specified by Litterman (1986) for the so-called Minnesota prior. That is, for a VAR model with n endogenous variables, m exogenous variables and p lags, the prior mean for the VAR coefficients is a $q \times 1 = n(np + m) \times 1$ vector β_0, while the prior covariance matrix is a $q \times q$ matrix Ω_0 with variance terms on the diagonal, and zero entries off diagonal, implying no prior covariance between the coefficients.

While this representation is convenient, it results in three main shortcomings. The first is technical and linked to the estimation of large models. Indeed, for all the priors adopting this Minnesota structure, estimation of the posterior mean $\bar{\beta}$ and the posterior variance $\bar{\Omega}$ involves the inversion of a $q \times q$ matrix. For instance, in the case of a large model with 40 endogenous variables, 5 exogenous variables and 15 lags $(n = 40, m = 5, p = 15)$, q is equal to 24200, implying that each iteration of the Gibbs sampler requires the inversion of a 24200×24200 matrix, rendering the process so slow that it becomes practically intractable. In the worst case, such very large matrices may even cause numerical softwares to fail the inversion altogether. The second shortcoming is theoretical: with this structure, no prior covariance is assumed among the VAR coefficients, which may be sub-optimal. Of course, one could simply add off-diagonal terms in Ω_0 in order to create prior covariance terms. However, there is no all-ready theory to indicate what those values should be. The third issue is that

with this kind of structure, it is very difficult to impose priors on combinations of VAR coefficients, which can yet be useful when working with unit root or cointegrated processes.

To remedy these shortcomings, this subsection proposes a specific prior known as the dummy coefficient prior, closely following the methodology introduced by Banbura et al. (2010) for large VAR models. The text first introduces the simplified prior used as the basis of the exercise. It then describes the process of creation of the dummy variable that enables to match the Minnesota moments. Finally, it develops the two extensions (sum of coefficients, and initial dummy observation) enriching the prior, and making them consistent with unit root or cointegration processes.

Consider first the prior distribution. As shown by A.4.8, it is possible to express the likelihood function for the data as:

$$f(y\,|\beta,\Sigma) \propto |\Sigma|^{-T/2} \exp\left[-\frac{1}{2} tr\left\{\Sigma^{-1}(B-\hat{B})'(X'X)(B-\hat{B})\right\}\right] \times \exp\left[-\frac{1}{2} tr\left\{\Sigma^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right\}\right] \tag{3.7.1}$$

with $\hat{B}$ the OLS estimate for B , defined by 3.1.9 as:

$$\hat{B} = (X'X)^{-1}X'Y \tag{3.7.2}$$

This likelihood function is then combined with a joint improper prior for β and Σ :

$$\pi\,(\beta,\Sigma) \propto |\Sigma|^{-(n+1)/2} \tag{3.7.3}$$

This prior is the simplest and least informative prior that one can propose for a VAR model. Combining the likelihood function 3.7.1 with the improper prior 3.7.3, one obtains the posterior distribution as:

$$f(\beta,\Sigma\,|y) \propto |\Sigma|^{-(T+n+1)/2} \exp\left[-\frac{1}{2} tr\left\{\Sigma^{-1}(B-\hat{B})'(X'X)(B-\hat{B})\right\}\right] \times \exp\left[-\frac{1}{2} tr\left\{\Sigma^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right\}\right] \tag{3.7.4}$$

3.7.4 appears as being the product of a matrix normal distribution with an inverse-Wishart distribution.

Four remarks can be done about this posterior. First, as the product of a matrix normal distribution with an inverse-Wishart distribution, this posterior is immediately comparable to that obtained for the normal-Wishart prior (see 3.4.15 and A.4.16). It can then be shown (see Appendix A.7 for details) that similarly to the normal-Wishart prior, the marginal posterior distributions for Σ and B are respectively inverse-Wishart and matrix student. They are parameterized as:

$$\Sigma \sim \mathcal{IW}\left(\hat{S}, \hat{\alpha}\right) \tag{3.7.5}$$

with:

$$\hat{S} = (Y - X\hat{B})'(Y - X\hat{B}) \tag{3.7.6}$$

and

$$\hat{\alpha} = T - k \tag{3.7.7}$$

On the other hand:

$$B \sim \mathcal{MT}(\hat{B}, \hat{S}, \hat{\Phi}, \widehat{\alpha}) \tag{3.7.8}$$

with $\hat{B}$ and $\hat{S}$ defined by 3.7.2 and 3.7.6, and:

$$\hat{\Phi} = (X'X)^{-1} \tag{3.7.9}$$

$$\widehat{\alpha} = T - n - k + 1 \tag{3.7.10}$$

The second remark is that this prior solves the dimensionality issue. While the Minnesota requires the inversion of a $q \times q$ matrix, it is apparent from 3.7.9 that this prior only requires the inversion of a $k \times k$ matrix, with $k = np + m$. The intuition behind the result is similar to that of the normal-Wishart: the posterior is computed at the scale of individual equations, rather than for the full model simultaneously. In the case of the example VAR model with 40 endogenous variables, 5 exogenous variables and 15 lags ($n = 40, m = 5, p = 15$), while a prior in the Minnesota fashion requires the inversion of a 24200×24200 matrix, which is practically infeasible, the present prior only requires inversion of a 605×605 matrix, a size that remains tractable for numerical softwares.

The third point of interest is that, not surprisingly, an uninformative prior for β and Σ yields posterior estimates centered at OLS (maximum likelihood) values. By not providing any prior information on the mean of the estimates, and setting a flat distribution with infinite variance, one does hardly more than performing OLS estimation, using only the information provided by the data.

The final comment on this prior is that it is getting estimates that are basically OLS estimates which creates an issue. The strength (and main interest) of Bayesian estimation is precisely to be able to supplement the information contained in the data with personal information, in order to inflect the estimates provided by the data and improve the accuracy of the model. If one does not provide any information at all, there is, in fact, very little point into using Bayesian methods. Ideally, one would thus like to provide prior information for the model, despite the diffuse prior. This is possible thanks to what is known as dummy observations, or artificial observations.

Consider thus the possibility of generating artificial data for the model. The aim of this generated data is to provide information to the model that would be equivalent to that supplied traditionally by the prior distribution. Precisely, the following data matrices Y_d and X_d, corresponding to Y and X in 3.1.7, are created:

$$Y_d = \begin{pmatrix} diag(\rho\sigma_1/\lambda_1, ..., \rho\sigma_n/\lambda_1) \\ 0_{n(p-1)\times n} \\ 0_{m\times n} \\ diag(\sigma_1, ..., \sigma_n) \end{pmatrix} \tag{3.7.11}$$

and

$$X_d = \begin{pmatrix} J_p \otimes diag\left(\sigma_1/\lambda_1, ..., \sigma_n/\lambda_1\right) & 0_{np\times m} \\ 0_{m\times np} & (1/\lambda_1\lambda_4) \otimes I_m \\ 0_{n\times np} & 0_{n\times m} \end{pmatrix} \tag{3.7.12}$$

ρ denotes the value of the autoregressive coefficient on first lags in the Minnesota prior, and $\sigma_1, ..., \sigma_n$ denotes as usual the standard deviation of the OLS residual obtained from individual autoregressive models. J_p is defined as: $J_p = diag\left(1^{\lambda_3}, 2^{\lambda_3}, ..., p^{\lambda_3}\right)$. Y_d is of dimension $(n(p+1)+m)\times n$, and X_d is of dimension $(n(p+1)+m)\times(np+m)$. Considering that each row of Y_d (or X_d) corresponds to an artificial period, one obtains a total of $T_d = n(p+1)+m$ simulated time periods. Note that unlike the canonical VAR model, X_d here does not correspond to lagged values of Y_d.

Both matrices Y_d and X_d are made of three blocks. The first block, made of the first np rows, is related to the moment of the VAR coefficients corresponding to the endogenous variables of the model. The second block, made of the next m rows, represents the moments of the coefficients on the exogenous variables. Finally, the last block, made of the last n rows, deals with the residual variance-covariance matrix.

To make this more concrete, consider a simple example: a VAR model with two endogenous variables and two lags, along with one exogenous variable ($n = 2, m = 1, p = 2$). This formulates as:

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \end{pmatrix} = \begin{pmatrix} a_{11}^1 & a_{12}^1 \\ a_{21}^1 & a_{22}^1 \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \end{pmatrix} + \begin{pmatrix} a_{11}^2 & a_{12}^2 \\ a_{21}^2 & a_{22}^2 \end{pmatrix} \begin{pmatrix} y_{1,t-2} \\ y_{2,t-2} \end{pmatrix} + \begin{pmatrix} c_{11} \\ c_{21} \end{pmatrix} (x_{1,t}) + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix} \tag{3.7.13}$$

For T_d periods, reformulated in the usual stacked form 3.1.6, one obtains:

$$\begin{pmatrix} \rho\sigma_1/\lambda_1 & 0 \\ 0 & \rho\sigma_2/\lambda_1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} = \begin{pmatrix} 1^{\lambda_3}\sigma_1/\lambda_1 & 0 & 0 & 0 & 0 \\ 0 & 1^{\lambda_3}\sigma_2/\lambda_1 & 0 & 0 & 0 \\ 0 & 0 & 2^{\lambda_3}\sigma_1/\lambda_1 & 0 & 0 \\ 0 & 0 & 0 & 2^{\lambda_3}\sigma_2/\lambda_1 & 0 \\ 0 & 0 & 0 & 0 & 1/\lambda_1\lambda_4 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} a_{11}^1 & a_{21}^1 \\ a_{12}^1 & a_{22}^1 \\ a_{11}^2 & a_{21}^2 \\ a_{12}^2 & a_{22}^2 \\ c_{11} & c_{21} \end{pmatrix}$$
$$+ \begin{pmatrix} \varepsilon_{1,1} & \varepsilon_{2,1} \\ \varepsilon_{1,2} & \varepsilon_{2,2} \\ \varepsilon_{1,3} & \varepsilon_{2,3} \\ \varepsilon_{1,4} & \varepsilon_{2,4} \\ \varepsilon_{1,5} & \varepsilon_{2,5} \\ \varepsilon_{1,6} & \varepsilon_{2,6} \\ \varepsilon_{1,7} & \varepsilon_{2,7} \end{pmatrix} \tag{3.7.14}$$

To see why this formulation implies a structure comparable to the normal-Wishart prior, develop the system 3.7.14, row after row. Start with the third block (the last two rows), and consider the entries related to the first variable (the first column). Developing, one obtains:

$$\varepsilon_{1,6} = \sigma_1 \tag{3.7.15}$$

and

$$0 = \varepsilon_{1,7} \tag{3.7.16}$$

Taking expectation over 3.7.16, one obtains:

$$E(\varepsilon_1) = 0 \tag{3.7.17}$$

Then, using 3.7.17 to compute the variance over 3.7.15, one concludes:

$$Var(\varepsilon_1) = \sigma_1^2 \tag{3.7.18}$$

This simply replicates the prior variance for ε_1 in the normal-Wishart prior (see equation 3.7.13).

Now, consider blocks 1 and 2, starting with block 1. The first row of (3.7.14) develops as:

$$\frac{\rho\sigma_1}{\lambda_1} = \frac{1^{\lambda_3}\sigma_1}{\lambda_1}a_{11}^1 + \varepsilon_{1,1} \Rightarrow a_{11}^1 = \frac{\rho}{1^{\lambda_3}} - \frac{\lambda_1}{1^{\lambda_3}\sigma_1}\varepsilon_{1,1} \tag{3.7.19}$$

And from 3.7.19, it is straightforward to conclude:

$$E(a_{11}^1) = \rho \quad \text{and} \quad Var(a_{11}^1) = (\lambda_1)^2 \tag{3.7.20}$$

The second entry of the first row of 3.7.14 yields:

$$0 = \frac{1^{\lambda_3}\sigma_1}{\lambda_1}a_{21}^1 + \varepsilon_{2,1} \Rightarrow a_{21}^1 = -\frac{\lambda_1}{1^{\lambda_3}\sigma_1}\varepsilon_{2,1} \tag{3.7.21}$$

From which one obtains:

$$E(a_{21}^1) = 0 \quad \text{and} \quad Var(a_{11}^1) = \left(\frac{\sigma_2^2}{\sigma_1^2}\right)(\lambda_1)^2 \tag{3.7.22}$$

Go for block 2. Develop the first entry of row 5:

$$0 = \frac{c_{11}}{\lambda_1\lambda_4} + \varepsilon_{1,5} \Rightarrow c_{11} = -\lambda_1\lambda_4\varepsilon_{1,5} \tag{3.7.23}$$

And from this, one obtains:

$$E(c_{11}) = 0 \quad \text{and} \quad Var(c_{11}) = (\lambda_1\lambda_4)^2\sigma_1^2 \tag{3.7.24}$$

Going on the same way with the other entries of blocks 1 and 2, it is straightforward to see that one will recover the full diagonal of 3.4.8, the prior covariance matrix for β implemented in the normal-Wishart prior.

Note however that 3.7.14 implies more than 3.4.8. Using for instance 3.7.19 and 3.7.23 one concludes that:

$$Cov(a_{11}^1, c_{11}) = (\lambda_1)^2\lambda_4\sigma_1 \tag{3.7.25}$$

(3.7.25) shows that unlike the strict normal-Wishart prior, the dummy observation setting allows to implement some prior covariance between the VAR coefficients of the same equation. In this

respect, the dummy observation scheme is even richer than the normal-Wishart prior.

To conclude this presentation of the basic dummy observation strategy, it is now shown how this setting combines with the simplified prior introduced at the beginning of the subsection. This is done in a simple way. Define:

$$Y^* = \begin{pmatrix} Y \\ Y_d \end{pmatrix}, \; X^* = \begin{pmatrix} X \\ X_d \end{pmatrix}, \; T^* = T + T_d \tag{3.7.26}$$

That is, Y^* and X^* are obtained by concatenating the dummy observation matrices at the top of the actual data matrices Y and X, and T^* is the total number of time periods, obtained from adding the actual and simulated time periods. Using this modified data set for the prior, one then obtains the same posterior distributions as 3.7.5 and 3.7.8, except that the posterior parameters 3.7.2, 3.7.6, 3.7.7, 3.7.9, and 3.7.10 are computed using Y^*, X^* and T^* rather than Y, X and T.

Possible extensions to this basic methodology are now developed. Indeed, as already discussed already, one of the main asset of the dummy coefficient prior is the convenience it offers to estimate large models. However, one specific issue may typically arise in large models when variables are introduced in level. Because such variables typically include unit roots, the model itself should be characterized by one (or more) unit roots, that is, roots with a value of one. However, with large models, each draw from the posterior distribution produces VAR coefficients for a large number of equations. This significantly increases the risk that for any draw, at least one equation will obtain coefficients that are actually explosive (have a root greater than one in absolute value) rather than comprising a strict unit root. This may result, for instance, on explosive confidence bands for the impulse response functions, which becomes larger at longer horizons.

It would thus be desirable to set a prior that would force the dynamic to favor unit root draws, rather than explosive draws. This can be realized, once again, with the use of dummy observations. The literature mainly focuses on two strategies: the "sum-of-coefficients" approach, initially introduced by Doan et al. (1984), and the "initial dummy observation" approach due to Sims (1992). To understand the mechanism at work, the traditional VAR model 3.1.2 is first re-expressed into what is known as an error correction form. Considering again the example VAR model used so far in this subsection, it is possible to obtain:

$$
\begin{aligned}
& y_t = A_1 y_{t-1} + A_2 y_{t-2} + C x_t + \varepsilon_t \\
& \Rightarrow y_t - y_{t-1} = -y_{t-1} + A_1 y_{t-1} + A_2 y_{t-2} + C x_t + \varepsilon_t \\
& \Rightarrow y_t - y_{t-1} = -y_{t-1} + A_1 y_{t-1} + (A_2 - A_2) y_{t-1} + A_2 y_{t-2} + C x_t + \varepsilon_t \\
& \Rightarrow y_t - y_{t-1} = (-y_{t-1} + A_1 y_{t-1} + A_2 y_{t-1}) - (A_2 y_{t-1} - A_2 y_{t-2}) + C x_t + \varepsilon_t \\
& \Rightarrow \Delta y_t = -(I - A_1 - A_2) y_{t-1} - A_2 \Delta y_{t-1} + C x_t + \varepsilon_t
\end{aligned} \tag{3.7.27}
$$

3.7.27 comprises two main parts on its right-hand side. The first term, $-(I - A_1 - A_2)y_{t-1}$, is the error correction term. It guaranties stationarity of the model by correcting the current-period change in y_t with part of the preceding period deviation from the long-run value of the model. The second part, $-A_2 \Delta y_{t-1} + C x_t + \varepsilon_t$, is made of stationary terms and hence does not impact the stationarity of the model.

In general, a VAR model with p lags in the form of 3.1.2 can be rewritten in error correction form as (see Appendix A.7 for details):

$$
\Delta y_t = -(I - A_1 - A_2 ... - A_p) y_{t-1} + B_1 \Delta y_{t-1} + B_2 \Delta y_{t-2} + ... + B_{p-1} \Delta y_{t-(p-1)} + C x_t + \varepsilon_t \tag{3.7.28}
$$

with $B_i = -(A_{i+1} + A_{i+2} + ... + A_p)$.

It is now possible to introduce the "sums-of-coefficients" approach. Assume that for a VAR model with p lags, the following holds:

$$
I - A_1 - A_2 ... - A_p = 0 \tag{3.7.29}
$$

Then from 3.7.28, it follows that:

$$
\Delta y_t = B_1 \Delta y_{t-1} + B_2 \Delta y_{t-2} + ... + B_{p-1} \Delta y_{t-(p-1)} + C x_t + \varepsilon_t
$$

Or, equivalently:

$$
y_t = y_{t-1} + B_1 \Delta y_{t-1} + B_2 \Delta y_{t-2} + ... + B_{p-1} \Delta y_{t-(p-1)} + C x_t + \varepsilon_t \tag{3.7.30}
$$

In this case, y_t is equal to its previous value, plus a sum of stationary terms not affecting stationarity. It follows that each variable in y_t contains a unit root. Also, note that the absence of error correction term rules out the possibility of cointegration relations.

In order to guarantee that the draws obtained from the posterior are characterized by a unit root rather than by explosive behavior, one may thus want to shrink prior information around 3.7.29. This can be done by creating the following dummy observations:

$$Y_s = diag(\bar{y}_1/\lambda_6, ..., \bar{y}_n/\lambda_6) \tag{3.7.31}$$

and

$$X_s = \begin{pmatrix} 1_{1\times p} \otimes Y_s & 0_{n\times m} \end{pmatrix} \tag{3.7.32}$$

where $\bar{y}_i$ denotes the arithmetic mean of variable y_i over the p pre-sample initial conditions, and λ_6 is a sums-of-coefficients specific shrinkage parameter. Y_s is of dimension $n \times n$, and X_s is of dimension $n \times (np+m)$. That is, an additional $T_s = n$ periods of dummy observations are generated. For the example VAR model used so far, this yields:

$$\begin{pmatrix} \bar{y}_1/\lambda_6 & 0 \\ 0 & \bar{y}_2/\lambda_6 \end{pmatrix} = \begin{pmatrix} \bar{y}_1/\lambda_6 & 0 & \bar{y}_1/\lambda_6 & 0 & 0 \\ 0 & \bar{y}_2/\lambda_6 & 0 & \bar{y}_2/\lambda_6 & 0 \end{pmatrix} \begin{pmatrix} a_{11}^1 & a_{21}^1 \\ a_{12}^1 & a_{22}^1 \\ a_{11}^2 & a_{21}^2 \\ a_{12}^2 & a_{22}^2 \\ c_{11} & c_{21} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,1} & \varepsilon_{1,2} \\ \varepsilon_{2,1} & \varepsilon_{2,2} \end{pmatrix} \tag{3.7.33}$$

Develop entry (1,1) to obtain:

$$\frac{\bar{y}_1}{\lambda_6} = \frac{\bar{y}_1}{\lambda_6} a_{11}^1 + \frac{\bar{y}_1}{\lambda_6} a_{11}^2 + \varepsilon_{1,1} \Rightarrow 1 - a_{11}^1 - a_{11}^2 = \frac{\lambda_6}{\bar{y}_1} \varepsilon_{1,1} \tag{3.7.34}$$

And from 3.7.34 one concludes (taking expectation and variances):

$$E(1 - a_{11}^1 - a_{11}^2) = 0 \quad \text{and} \quad Var(1 - a_{11}^1 - a_{11}^2) = \left(\frac{\lambda_6}{\bar{y}_1}\right)^2 \sigma_1^2 \tag{3.7.35}$$

Going on, one recovers 3.7.29. Also, it becomes apparent from 3.7.35 that λ_6 controls the variance over the prior belief: as λ_6 shrinks, so does the prior variance over 3.7.29. The limit case $\lambda_6 \to 0$ implies that there is a unit root in each equation, and cointegration is ruled out, while the limit case $\lambda_6 \to \infty$ implies an uninformative (diffuse) prior. The methodology is then similar to the one used for the basic dummy observation prior. Define:

$$Y^* = \begin{pmatrix} Y \\ Y_d \\ Y_s \end{pmatrix}, \quad X^* = \begin{pmatrix} X \\ X_d \\ X_s \end{pmatrix}, \quad T^* = T + T_d + T_s \tag{3.7.36}$$

Apply then the same posterior distributions 3.7.5 and 3.7.8, but use the values 3.7.36 rather than the original data values.

A shortcoming of the sum-of-coefficients strategy is that it rules out cointegration in the limit, which may be undesirable. Therefore, an additional identification scheme, known as the "dummy initial observation", is now proposed. In this scheme, one single dummy observation is created for each variable. This leads to the creation of the following matrices Y_o and X_o:

$$Y_o = \begin{pmatrix} \bar{y}_1/\lambda_7 & \dots & \bar{y}_n/\lambda_7 \end{pmatrix} \tag{3.7.37}$$

and

$$X_o = \begin{pmatrix} 1_{1\times p} \otimes Y_o & \bar{x}/\lambda_7 \end{pmatrix} \tag{3.7.38}$$

where $\bar{x} = \begin{pmatrix} \bar{x}_1 & \dots & \bar{x}_m \end{pmatrix}$ is the $1 \times m$ vector in which each entry $\bar{x}_i$ denotes the arithmetic mean of exogenous variable x_i over the p pre-sample initial conditions, and λ_7 is a dummy initial observation specific hyperparameter. Y_o is of dimension $1 \times n$, X_o is of dimension $1 \times (np+m)$, and $T_o = 1$ period of dummy observation is generated. For the example VAR model used so far, this yields:

$$\begin{pmatrix} \bar{y}_1/\lambda_7 & \bar{y}_2/\lambda_7 \end{pmatrix} = \begin{pmatrix} \bar{y}_1/\lambda_7 & \bar{y}_2/\lambda_7 & \bar{y}_1/\lambda_7 & \bar{y}_2/\lambda_7 & \bar{x}_1/\lambda_7 \end{pmatrix} \begin{pmatrix} a_{11}^1 & a_{21}^1 \\ a_{12}^1 & a_{22}^1 \\ a_{11}^2 & a_{21}^2 \\ a_{12}^2 & a_{22}^2 \\ c_{11} & c_{21} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,1} & \varepsilon_{1,2} \end{pmatrix} \tag{3.7.39}$$

Developing the first entry of 3.7.39, one obtains:

$$\begin{aligned} \frac{\bar{y}_1}{\lambda_7} &= \frac{\bar{y}_1}{\lambda_7} a_{11}^1 + \frac{\bar{y}_2}{\lambda_7} a_{12}^1 + \frac{\bar{y}_1}{\lambda_7} a_{11}^2 + \frac{\bar{y}_2}{\lambda_7} a_{12}^2 + \frac{\bar{x}_1}{\lambda_7} c_{11} + \varepsilon_{1,1} \\ &\Rightarrow \bar{y}_1 - \bar{y}_1 a_{11}^1 - \bar{y}_2 a_{12}^1 - \bar{y}_1 a_{11}^2 - \bar{y}_2 a_{12}^2 - \bar{x}_1 c_{11} = \lambda_7 \varepsilon_{1,1} \end{aligned} \tag{3.7.40}$$

Taking expectations, obtain:

$$\bar{y}_1 = \bar{y}_1 a_{11}^1 + \bar{y}_2 a_{12}^1 + \bar{y}_1 a_{11}^2 + \bar{y}_2 a_{12}^2 + \bar{x}_1 c_{11} \tag{3.7.41}$$

And computing the variance yields:

$$Var(\bar{y}_1 - \bar{y}_1 a_{11}^1 - \bar{y}_2 a_{12}^1 - \bar{y}_1 a_{11}^2 - \bar{y}_2 a_{12}^2 - \bar{x}_1 c_{11}) = (\lambda_7)^2 \sigma_1^2 \tag{3.7.42}$$

3.7.41 states that a no-change forecast constitutes a good representation of the dynamic of the model. From 3.7.42, it appears that λ_7 represents, again, the shrinkage parameter over the prior. When $\lambda_7 \to \infty$, the prior is diffuse. When $\lambda_7 \to 0$, 3.7.41 holds. Then, either all the variables are at their unconditional mean, which implies that the model is stationary despite the unit roots in the variables (implying cointegration), or the dynamic of the system is characterized by an unspecified number of unit roots, and the variables share a common stochastic trend. The methodology is then similar to the one previously applied. Define:

$$Y^* = \begin{pmatrix} Y \\ Y_d \\ Y_o \end{pmatrix}, \quad X^* = \begin{pmatrix} X \\ X_d \\ X_o \end{pmatrix}, \quad T^* = T + T_d + T_o \tag{3.7.43}$$

Apply then the same posterior distributions 3.7.5 and 3.7.8, but use the values 3.7.43 rather than the original data values.

Note finally that if the two dummy observation extensions (sums-of-coefficients and dummy initial observations) have been introduced in the context of the dummy prior, they can actually apply to any of the other priors developed so far. Adaptation is straightforward. Define:

$$Y^* = \begin{pmatrix} Y \\ Y_s \\ Y_o \end{pmatrix}, \quad X^* = \begin{pmatrix} X \\ X_s \\ X_o \end{pmatrix}, \quad \text{and} \quad T^* = T + T_s + T_o \tag{3.7.44}$$

with Y_s, Y_o, X_s and X_o defined as in 3.7.31, 3.7.32, 3.7.37 and 3.7.38. Then run estimation of the posterior distribution as usual, but replace Y, X and T by Y^*, X^* and T^*. Also, it is obviously possible to run the estimation process with only one of the two dummy extensions, with 3.7.44 modified accordingly.

3.8 Block exogeneity

Before closing the part related to the estimation of BVAR models, it is worth describing a simple and useful feature known as block exogeneity. This concept is closely related to that of Granger causality in VAR models. To make things more concrete, consider again the simple example developed in subsection 3 for the mean and variance of the Minnesota prior, which consisted of a VAR model with two endogenous variables and two lags, along with one exogenous variables. Using 3.1.1, this model formulates as:

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \end{pmatrix} = \begin{pmatrix} a_{11}^1 & a_{12}^1 \\ a_{21}^1 & a_{22}^1 \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \end{pmatrix} + \begin{pmatrix} a_{11}^2 & a_{12}^2 \\ a_{21}^2 & a_{22}^2 \end{pmatrix} \begin{pmatrix} y_{1,t-2} \\ y_{2,t-2} \end{pmatrix} + \begin{pmatrix} c_{11} \\ c_{21} \end{pmatrix} (x_{1,t}) + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix} \tag{3.8.1}$$

Imagine that for some reason, one thinks that the second variable does not affect the first variable, that is, it has no impact on it. In terms of the example model 3.8.1, the fact that $y_{1,t}$ is exogenous to $y_{2,t}$ translates into:

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \end{pmatrix} = \begin{pmatrix} a_{11}^1 & 0 \\ a_{21}^1 & a_{22}^1 \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \end{pmatrix} + \begin{pmatrix} a_{11}^2 & 0 \\ a_{21}^2 & a_{22}^2 \end{pmatrix} \begin{pmatrix} y_{1,t-2} \\ y_{2,t-2} \end{pmatrix} + \begin{pmatrix} c_{11} \\ c_{21} \end{pmatrix} (x_{1,t}) + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix} \tag{3.8.2}$$

If 3.8.2 is believed to be the correct representation of the relation between y_1 and y_2, one would like to obtain this representation from the posterior of the VAR model. In fact, it turns out to be easy to force the posterior distribution of a BVAR model to take the form of 3.8.2: by setting a 0 prior mean on the relevant coefficients, and by implementing an arbitrary small prior variance on them, one can make sure that the posterior values will be close to 0 as well. In practice, this implies the following: first, set the prior mean by using a conventional Minnesota scheme. Here, one would just use 3.3.4:

$$\beta_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{3.8.3}$$

This guarantees that the prior mean of any block exogenous coefficient is 0 (the Minnesota scheme only implements non-zero entries on own lags, which cannot be part of the block exogeneity scheme since a variable cannot be exogenous to itself). Then, use a variance scheme similar to 3.3.8, but multiply the block exogenous variance by an additional parameter $(\lambda_5)^2$, which will be set to an arbitrary small value. This will result in a very tight prior variance on these coefficients. In practice, one may for example use the value: $\lambda_5 = 0.001$. Using this strategy on the above example, one obtains a modified version of 3.3.8:

$$\Omega_0 = \begin{pmatrix} (\lambda_1)^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \left(\frac{\sigma_1^2}{\sigma_2^2}\right)(\lambda_1\lambda_2\lambda_5)^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \left(\frac{\lambda_1}{2^{\lambda_3}}\right)^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \left(\frac{\sigma_1^2}{\sigma_2^2}\right)\left(\frac{\lambda_1\lambda_2\lambda_5}{2^{\lambda_3}}\right)^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_1^2(\lambda_1\lambda_4)^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \left(\frac{\sigma_2^2}{\sigma_1^2}\right)(\lambda_1\lambda_2)^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & (\lambda_1)^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \left(\frac{\sigma_2^2}{\sigma_1^2}\right)\left(\frac{\lambda_1\lambda_2}{2^{\lambda_3}}\right)^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \left(\frac{\lambda_1}{2^{\lambda_3}}\right)^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_2^2(\lambda_1\lambda_4)^2 \end{pmatrix} \tag{3.8.4}$$

Because the prior variance will be very close to 0 (it can actually be made arbitrarily close to 0 by reducing the value of λ_5), the posterior distribution will be extremely tight around 0, as wished. Of course, block exogeneity needs not being limited to one variable only. One may create as many exogenous blocks as required. Suffice is to multiply the prior variance of all the relevant coefficients by $(\lambda_5)^2$ to obtain the desired exogenity on the posterior mean.

A final remark on block exogeneity: it is available with the Minnesota, independent normal-Wishart, and normal diffuse priors, but not with the normal-Wishart prior nor the dummy observation prior. For the dummy observation prior the reason is obvious - the prior is diffuse, so $\Sigma\otimes$ is simply not defined. For the normal-Wishart prior it is the particular Kronecker structure $\Sigma \otimes \Phi_0$ in place of the covariance matrix Ω_0 that causes instability. This structure implies that the variance of one equation has to be proportional with the variance of the other equations. Hence, imposing block exogeneity on one variable for one equation would lead to impose it on all the other equations. Not only would it lead to assume block exogeneity on some equations where it would not be desired, but it would also lead to some of the model variables to be exogenous to themselves, which is impossible (for instance, in the above example, y_2 would become exogenous to itself).

3.9 Evaluating the model: calculation of the marginal likelihood

The previous subsections have underlined different methods used to obtain the posterior distribution of the parameters of interest, given the selected prior. While these methods enable the estimation of Bayesian VAR models, they are of no help to answer a central question: how should the model be specified? For instance, what is the optimal lag number? What are the best shrinkage values for the prior distribution? This issue is central, as one wants to select an adequate model before starting the analysis of forecasts, impulse response functions, and so on.

Answering this question is fundamentally a matter of model comparison: for instance the simplest way to determine if the optimal number of lags for the model is two or three, one will simply compare the two models and determine which one has the highest posterior probability of being the true model. The discussion was initiated in subsection 2, and it was seen (equations 3.2.20 and 3.2.21) that comparing two models requires the computation of the Bayes factor, itself derived from the marginal likelihoods of the models. This subsection is thus dedicated to the methods used to compute the marginal likelihood in practice.

The methodology employed to derive the marginal likelihood varies from one prior to the other. Therefore, the derivations are developed in turn for each prior [2].

Deriving the marginal likelihood for the Minnesota prior:

The derivation of the marginal likelihood for the Minnesota prior essentially follows the strategy proposed by Giannone et al. (2015) for the normal-Wishart prior. To derive the marginal likelihood for model i, first remember that it is defined by 3.2.20 as:

$$m_i(y) = \int f_i(y\,|\theta_i, M_i)\pi_i(\theta_i\,|M_i)d\theta_i \tag{3.9.1}$$

Suppress for the time being the model indexes i and M_i to consider the general situation wherein $m(y)$ is the marginal density for a given model. Then, 3.9.1 reformulates as:

$$m(y) = \int f(y\,|\theta)\pi(\theta)d\theta \tag{3.9.2}$$

The marginal likelihood can thus be seen as the product of the data likelihood function $f(y\,|\theta)$ with the prior distribution $\pi(\theta)$, integrated over parameter values. Note that unlike the application of Bayes rule to derive posterior distribution, it is not sufficient here to work only with the kernels of the distributions. The normalizing constants have to be integrated in the calculations. In the case of the Minnesota prior (and the other natural conjugate prior, the normal-Wishart), it is possible to apply 3.9.2 directly to compute the value of the marginal likelihood.

For the Minnesota, the set of parameters θ reduces to the VAR coefficients β, so that $\theta = \beta$, and 3.9.2 writes as:

$$m(y) = \int f(y\,|\beta)\pi(\beta)d\beta \tag{3.9.3}$$

The likelihood function $f\,(y\,|\beta\,)$ is given by 3.3.1:

[2]Some details of the derivations are omitted in the main text. The missing steps can be found in Appendix A.8

$$f(y\,|\beta,\bar{\Sigma}) = (2\pi)^{-nT/2}\left|\bar{\Sigma}\right|^{-1/2}\exp\left[-\frac{1}{2}\left(y-\bar{X}\beta\right)'\bar{\Sigma}^{-1}\left(y-\bar{X}\beta\right)\right] \tag{3.9.4}$$

The prior density $\pi(\beta)$ is given by 3.3.13:

$$\pi(\beta) = (2\pi)^{-nk/2}|\Omega_0|^{-1/2}\exp\left[-\frac{1}{2}(\beta-\beta_0)'\Omega_0^{-1}(\beta-\beta_0)\right] \tag{3.9.5}$$

Combining 3.9.4 and 3.9.5, one obtains:

$$\begin{aligned} f(y\,|\beta)\,\pi(\beta) &= (2\pi)^{-n(T+k)/2}\left|\bar{\Sigma}\right|^{-1/2}|\Omega_0|^{-1/2} \\ &\times\exp\left[-\frac{1}{2}\left\{\left(y-\bar{X}\beta\right)'\bar{\Sigma}^{-1}\left(y-\bar{X}\beta\right)+(\beta-\beta_0)'\Omega_0^{-1}(\beta-\beta_0)\right\}\right] \end{aligned} \tag{3.9.6}$$

Note that the second row of 3.9.6 is just A.3.1. Then, from A.3.8, it is possible to rewrite 3.9.6 as:

$$\begin{aligned} f(y\,|\beta)\,\pi(\beta) &= (2\pi)^{-n(T+k)/2}\left|\bar{\Sigma}\right|^{-1/2}|\Omega_0|^{-1/2} \\ &\times\exp\left[-\frac{1}{2}\left\{\left(\beta-\bar{\beta}\right)'\bar{\Omega}^{-1}\left(\beta-\bar{\beta}\right)+\left(\beta_0'\Omega_0^{-1}\beta_0-\bar{\beta}'\bar{\Omega}^{-1}\bar{\beta}+y'\bar{\Sigma}^{-1}y\right)\right\}\right] \end{aligned} \tag{3.9.7}$$

Where $\bar{\beta}$ and $\bar{\Omega}$ are defined as in 3.3.17 and 3.3.18. It is then possible to reformulate 3.9.7 as:

$$\begin{aligned} f(y\,|\beta)\,\pi(\beta) &= (2\pi)^{-nT/2}\left|\bar{\Sigma}\right|^{-1/2}|\Omega_0|^{-1/2}\left|\bar{\Omega}\right|^{1/2}\exp\left[-\frac{1}{2}\left(\beta_0'\Omega_0^{-1}\beta_0-\bar{\beta}'\bar{\Omega}^{-1}\bar{\beta}+y'\bar{\Sigma}^{-1}y\right)\right] \\ &\times(2\pi)^{-nk/2}\left|\bar{\Omega}\right|^{-1/2}\exp\left[-\frac{1}{2}\left(\beta-\bar{\beta}\right)'\bar{\Omega}^{-1}\left(\beta-\bar{\beta}\right)\right] \end{aligned} \tag{3.9.8}$$

The advantage of 3.9.8 is that its second row can be readily recognized as the density of a multivariate normal distribution for β. Therefore, when integrating with respect to β , this row will simply integrate to 1, greatly simplifying 3.9.8.

So, integrate 3.9.8 with respect to β to obtain:

$$
\begin{aligned}
m(y) &= \int f(y\,|\beta)\pi(\beta)d\beta \\
&= (2\pi)^{-nT/2}\left|\bar{\Sigma}\right|^{-1/2}|\Omega_0|^{-1/2}\left|\bar{\Omega}\right|^{1/2}\exp\left[-\frac{1}{2}\left(\beta_0'\Omega_0^{-1}\beta_0-\bar{\beta}'\bar{\Omega}^{-1}\bar{\beta}+y'\bar{\Sigma}^{-1}y\right)\right] \\
&\int (2\pi)^{-nk/2}\left|\bar{\Omega}\right|^{-1/2}\exp\left[-\frac{1}{2}\left(\beta-\bar{\beta}\right)'\bar{\Omega}^{-1}\left(\beta-\bar{\beta}\right)\right]d\beta
\end{aligned}
\tag{3.9.9}
$$

From this, one eventually concludes:

$$
m(y) = (2\pi)^{-nT/2}\left|\bar{\Sigma}\right|^{-1/2}|\Omega_0|^{-1/2}\left|\bar{\Omega}\right|^{1/2}\exp\left[-\frac{1}{2}\left(\beta_0'\Omega_0^{-1}\beta_0-\bar{\beta}'\bar{\Omega}^{-1}\bar{\beta}+y'\bar{\Sigma}^{-1}y\right)\right] \tag{3.9.10}
$$

Numerical issues with the marginal likelihood for the Minnesota prior

While 3.9.10 is a perfectly correct formula for the marginal likelihood, it may suffer from numerical instability. In other words, numerical softwares may not be able to compute it, and will return an error. This is mostly due to the fact that for large models, Ω_0 may become close to singular (some diagonal entries will become very close to zero), leading the software to conclude that $|\Omega_0|$=0. It is then preferable to transform 3.9.10, in order to obtain a formula which will be more stable, and computationally more efficient. After some manipulations, it can be rewritten as:

$$
\begin{aligned}
m(y) &= (2\pi)^{-nT/2}|\Sigma|^{-T/2}\left|I_{nk}+F_\Omega'\left(\Sigma^{-1}\otimes X'X\right)F_\Omega\right|^{-1/2} \\
&\times\exp\left[-\frac{1}{2}\left(\beta_0'\Omega_0^{-1}\beta_0-\bar{\beta}'\bar{\Omega}^{-1}\bar{\beta}+y'\left(\Sigma^{-1}\otimes I_T\right)y\right)\right]
\end{aligned}
\tag{3.9.11}
$$

where F_Ω denotes the square root matrix of Ω_0, that is, the matrix F_Ω such that $F_\Omega F_\Omega' = \Omega_0$. 3.9.11 is both numerically stable and faster to compute than 3.9.10. The determinant of the first row can be obtained from A.1.18, taking the product of 1 plus the eigenvalues of $F_\Omega'\left(\Sigma^{-1}\otimes X'X\right)F_\Omega$.

Finally, as it is typically easier to work with logs than with values in level, 3.9.11 can be equivalently reformulated as:

$$
\begin{aligned}
\log\left(m(y)\right) &= -\frac{nT}{2}\log\left(2\pi\right)-\frac{T}{2}\log\left(|\Sigma|\right)-\frac{1}{2}\log\left(\left|I_{nk}+F_\Omega'\left(\Sigma^{-1}\otimes X'X\right)F_\Omega\right|\right) \\
&-\frac{1}{2}\left(\beta_0'\Omega_0^{-1}\beta_0-\bar{\beta}'\bar{\Omega}^{-1}\bar{\beta}+y'\left(\Sigma^{-1}\otimes I_T\right)y\right)
\end{aligned}
\tag{3.9.12}
$$

Deriving the marginal likelihood for the normal Wishart prior

The strategy to derive the marginal likelihood for the normal Wishart prior is comparable to that of the Minnesota, though the computation are made more complex by the inclusion of two sets of parameters β and Σ. Hence, for the normal Wishart, the set of parameters is $\theta = \beta, \Sigma$, so that 3.9.2 writes:

$$m(y) = \iint f\left(y \,|\beta, \Sigma\right) \pi(\beta, \Sigma) d\beta d\Sigma \tag{3.9.13}$$

Assuming as usual independence between β and Σ, one obtains

$$m(y) = \iint f\left(y \,|\beta, \Sigma\right) \pi(\beta)\pi(\Sigma) d\beta d\Sigma \tag{3.9.14}$$

From 3.1.13, 3.1.14 and 3.3.1, the likelihood for the data is given by:

$$\begin{aligned} f(y \,|\beta, \Sigma) &= (2\pi)^{-nT/2} \\ &\times |\Sigma \otimes I_T|^{-1/2} \exp\left[-\frac{1}{2}(y - (I_n \otimes X)\,\beta)'(\Sigma \otimes I_T)^{-1}\,(y - (I_n \otimes X)\,\beta)\right] \end{aligned} \tag{3.9.15}$$

The second row of 3.9.15 is just A.4.1. Therefore, from A.4.8, the data density rewrites as:

$$\begin{aligned} f(y \,|\beta, \Sigma) &= (2\pi)^{-nT/2}|\Sigma|^{-T/2} \exp\left[-\frac{1}{2} tr\left\{\Sigma^{-1}\left(B - \hat{B}\right)'(X'X)\left(B - \hat{B}\right)\right\}\right] \\ &\times \exp\left[-\frac{1}{2} tr\left\{\Sigma^{-1}(Y - X\hat{B})'(Y - X\hat{B})\right\}\right] \end{aligned} \tag{3.9.16}$$

The prior distribution for β is given by:

$$\pi(\beta) = (2\pi)^{-nk/2}|\Sigma \otimes \Phi_0|^{-1/2} \exp\left(-\frac{1}{2}(\beta - \beta_0)'(\Sigma \otimes \Phi_0)^{-1}\,(\beta - \beta_0)\right) \tag{3.9.17}$$

GivenA.2.3.4, 3.9.17 reformulates as:

$$\pi(\beta) = (2\pi)^{-nk/2}|\Sigma|^{-k/2}|\Phi_0|^{-n/2} \exp\left(-\frac{1}{2} tr\left[\Sigma^{-1}(B - B_0)'\Phi_0^{-1}\,(B - B_0)\right]\right) \tag{3.9.18}$$

Finally, the prior distribution for Σ is given by:

$$\pi(\Sigma) = \frac{1}{2^{\alpha_0 n/2}\Gamma_n\left(\frac{\alpha_0}{2}\right)}|S_0|^{\alpha_0/2}|\Sigma|^{-(\alpha_0+n+1)/2} \exp\left(-\frac{1}{2} tr\left\{\Sigma^{-1}S_0\right\}\right) \tag{3.9.19}$$

Combining 3.9.16, 3.9.18 and 3.9.19:

$$
\begin{aligned}
f(y\,|\beta,\Sigma)\pi(\beta)\pi(\Sigma) &= (2\pi)^{-nT/2}(2\pi)^{-nk/2}|\Phi_0|^{-n/2}\frac{1}{2^{\alpha_0 n/2}\Gamma_n\left(\frac{\alpha_0}{2}\right)}|S_0|^{\alpha_0/2} \\
&\times |\Sigma|^{-T/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left(B-\hat{B}\right)'(X'X)\left(B-\hat{B}\right)\right\}\right] \\
&\times \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right\}\right] \\
&\times |\Sigma|^{-k/2}\exp\left(-\frac{1}{2}tr\left[\Sigma^{-1}(B-B_0)'\Phi_0^{-1}(B-B_0)\right]\right) \\
&\times |\Sigma|^{-(\alpha_0+n+1)/2}\exp\left(-\frac{1}{2}tr\left\{\Sigma^{-1}S_0\right\}\right)
\end{aligned}
\tag{3.9.20}
$$

The last four rows of 3.9.20 can be recognised as A.4.10. Therefore, from A.4.16, 3.9.20 rewrites as:

$$
\begin{aligned}
f(y\,|\beta,\Sigma)\pi(\beta)\pi(\Sigma) &= (2\pi)^{-nT/2}(2\pi)^{-nk/2}|\Phi_0|^{-n/2}\frac{1}{2^{\alpha_0 n/2}\Gamma_n\left(\frac{\alpha_0}{2}\right)}|S_0|^{\alpha_0/2} \\
&\times |\Sigma|^{-k/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left[(B-\bar{B})'\bar{\Phi}^{-1}(B-\bar{B})\right]\right\}\right] \\
&\times |\Sigma|^{-(\bar{\alpha}+n+1)/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\bar{S}\right\}\right]
\end{aligned}
\tag{3.9.21}
$$

where $\bar{B},\bar{\Phi},\bar{\alpha}$ and $\bar{S}$ are defined as in 3.9.17-3.9.20. It is then convenient to reformulate 3.9.21 as:

$$
\begin{aligned}
f(y\,|\beta,\Sigma)\pi(\beta)\pi(\Sigma) &= (2\pi)^{-nT/2}|\Phi_0|^{-n/2}|S_0|^{\alpha_0/2}\left|\bar{\Phi}\right|^{n/2}\left|\bar{S}\right|^{-\bar{\alpha}/2}\frac{2^{\bar{\alpha}n/2}\Gamma_n\left(\frac{\bar{\alpha}}{2}\right)}{2^{\alpha_0 n/2}\Gamma_n\left(\frac{\alpha_0}{2}\right)} \\
&\times (2\pi)^{-nk/2}|\Sigma|^{-k/2}\left|\bar{\Phi}\right|^{-n/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left[(B-\bar{B})'\bar{\Phi}^{-1}(B-\bar{B})\right]\right\}\right] \\
&\times \frac{1}{2^{\bar{\alpha}n/2}\Gamma_n\left(\frac{\bar{\alpha}}{2}\right)}\left|\bar{S}\right|^{\bar{\alpha}/2}|\Sigma|^{-(\bar{\alpha}+n+1)/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\bar{S}\right\}\right]
\end{aligned}
\tag{3.9.22}
$$

3.9.22 greatly facilitates the integration process in 3.9.14, since the second row of 3.9.22 can be recognised as the density of a matrix normal distribution, while the third row can be recognised as the density of an inverse Wishart distribution, both integrating to 1.

Hence, substitute 3.9.22 in 3.9.14 to eventually obtain:

$$
\begin{aligned}
m(y) &= \iint f\left(y\,|\beta,\Sigma\right)\pi(\beta)\pi(\Sigma)d\beta d\Sigma \\
&= \int (2\pi)^{-nT/2}|\Phi_0|^{-n/2}|S_0|^{\alpha_0/2}\left|\bar{\Phi}\right|^{n/2}\left|\bar{S}\right|^{-\bar{\alpha}/2}\frac{2^{\bar{\alpha}n/2}\Gamma_n\left(\frac{\bar{\alpha}}{2}\right)}{2^{\alpha_0 n/2}\Gamma_n\left(\frac{\alpha_0}{2}\right)} \\
&\times \int (2\pi)^{-nk/2}|\Sigma|^{-k/2}\left|\bar{\Phi}\right|^{-n/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left[(B-\bar{B})'\bar{\Phi}^{-1}(B-\bar{B})\right]\right\}\right]d\beta \\
&\times \frac{1}{2^{\bar{\alpha}n/2}\Gamma_n\left(\frac{\bar{\alpha}}{2}\right)}\left|\bar{S}\right|^{\bar{\alpha}/2}|\Sigma|^{-(\bar{\alpha}+n+1)/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\bar{S}\right\}\right]d\Sigma \\
&= (2\pi)^{-nT/2}|\Phi_0|^{-n/2}|S_0|^{\alpha_0/2}\left|\bar{\Phi}\right|^{n/2}\left|\bar{S}\right|^{-\bar{\alpha}/2}\frac{2^{\bar{\alpha}n/2}\Gamma_n\left(\frac{\bar{\alpha}}{2}\right)}{2^{\alpha_0 n/2}\Gamma_n\left(\frac{\alpha_0}{2}\right)} \\
&\times \int \frac{1}{2^{\bar{\alpha}n/2}\Gamma_n\left(\frac{\bar{\alpha}}{2}\right)}\left|\bar{S}\right|^{\bar{\alpha}/2}|\Sigma|^{-(\bar{\alpha}+n+1)/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\bar{S}\right\}\right]d\Sigma
\end{aligned}
\tag{3.9.23}
$$

Which yields:

$$
m(y) = (2\pi)^{-nT/2}|\Phi_0|^{-n/2}|S_0|^{\alpha_0/2}\left|\bar{\Phi}\right|^{n/2}\left|\bar{S}\right|^{-\bar{\alpha}/2}\frac{2^{\bar{\alpha}n/2}\Gamma_n\left(\frac{\bar{\alpha}}{2}\right)}{2^{\alpha_0 n/2}\Gamma_n\left(\frac{\alpha_0}{2}\right)} \tag{3.9.24}
$$

After some additional manipulations, one eventually obtains:

$$
m(y) = \pi^{-nT/2}|\Phi_0|^{-n/2}|S_0|^{\alpha_0/2}\left|\bar{\Phi}\right|^{n/2}\left|\bar{S}\right|^{-\bar{\alpha}/2}\frac{\Gamma_n\left(\frac{\bar{\alpha}}{2}\right)}{\Gamma_n\left(\frac{\alpha_0}{2}\right)} \tag{3.9.25}
$$

Numerical issues with the marginal likelihood for the normal-Wishart prior

3.9.25 is a valid formula for the marginal likelihood, but it tends to suffer from the same numerical instability as in the Minnesota case. To improve numerical stability and efficiency, some reformulation is required. After quite a bit of manipulations, it is possible to show that 3.9.25 reformulates as:

$$
m(y) = \pi^{-nT/2}\frac{\Gamma_n\left(\frac{\bar{\alpha}}{2}\right)}{\Gamma_n\left(\frac{\alpha_0}{2}\right)}|S_0|^{-T/2}\left|I_k + F_\Phi' X' X F_\Phi\right|^{-n/2}\left|I_n + F_S'\left[\left(\bar{S}-S_0\right)\right]F_S\right|^{-\bar{\alpha}/2} \tag{3.9.26}
$$

where F_Φ denotes the square root matrix of Φ_0 so that $F_\Phi F_\Phi' = \Phi_0$, and F_S denotes the inverse square root matrix of S_0 so that $F_S F_S' = S_0^{-1}$. Once again, the two determinant terms in 3.9.26 can be computed taking the products of 1 plus the eigenvalues of $F_\Phi' X' X F_\Phi$ and $F_S'\left[\left(\bar{S}-S_0\right)\right]F_S$.

Working from A.1.18, by once again with logs rather than levels, 3.9.26 becomes:

$$\log(m(y)) = -\frac{nT}{2}\log(\pi) + \log\left(\Gamma_n\left(\frac{\bar{\alpha}}{2}\right)\right) - \log\left(\Gamma_n\left(\frac{\alpha_0}{2}\right)\right) - \frac{T}{2}\log(|S_0|)$$
$$-\frac{n}{2}\log(|I_k + F_\Phi' X' X F_\Phi|) - \frac{\bar{\alpha}}{2}\log\left(\left|I_n + F_S'\left(\bar{S} - S_0\right)F_S\right|\right) \tag{3.9.27}$$

Deriving the marginal likelihood for the independent normal Wishart prior

While in the case of the Minnesota and normal Wishart priors, it was possible to derive the value of the marginal likelihood by direct application of 3.9.2, this is not possible anymore in the case of the independent normal Wishart prior. The reason is similar to that preventing to obtain an analytical formula for the posterior distributions for this prior: in the product term $f(y\,|\beta,\Sigma)\pi(\beta)\pi(\Sigma)$, the terms involving β and Σ are so interwoven that it is not possible to separate them into marginal distributions disappearing during the integration process.

There exists, fortunately, alternative ways to obtain the marginal likelihood. The one used for the independent normal-Wishart prior follows the methodology proposed by Chib (1995), who proposes a simple technique based on Gibbs sampling outputs. Start again from 3.9.2 and develop, using basic definitions of conditional probabilities:

$$m(y) = \int f(y\,|\theta)\pi(\theta)d\theta = \int \left[f(y,\theta)/\pi(\theta)\right]\pi(\theta)d\theta = \int f(y,\theta)d\theta = f(y) \tag{3.9.28}$$

That is, the marginal likelihood is equal to $f(y)$, the density of the data. The strategy of Chib (1995) consists in noticing that the inverse of $f(y)$ is used as the normative constant in Bayes rules 3.2.2, and use this rule to obtain the marginal likelihood. Hence, start from Bayes rule 3.2.2:

$$\pi(\theta\,|y) = \frac{f(y\,|\theta)\pi(\theta)}{f(y)} \tag{3.9.29}$$

Use 3.9.27 and rearrange to obtain:

$$m(y) = \frac{f(y\,|\theta)\pi(\theta)}{\pi(\theta\,|y)} \tag{3.9.30}$$

Given that $\theta = \beta, \Sigma$, and assuming as usual independence, 3.9.30 rewrites:

$$m(y) = \frac{f(y\,|\beta,\Sigma)\pi(\beta)\pi(\Sigma)}{\pi(\beta,\Sigma\,|y)} \tag{3.9.31}$$

Equation 3.9.30 is perfectly general and thus works for any chosen value for β and Σ. In practice, however, a point of high density such as the posterior median is chosen in order to increase numerical precision. Denoting respectively by $\tilde{\beta}$ and $\tilde{\Sigma}$ the posterior median for β and Σ, 3.9.31 rewrites as:

$$m(y) = \frac{f(y \,|\tilde{\beta}, \tilde{\Sigma})\pi(\tilde{\beta})\pi(\tilde{\Sigma})}{\pi(\tilde{\beta}, \tilde{\Sigma}\,|y)} \tag{3.9.32}$$

While the denominator of 3.9.32 can be readily computed (it is the product of the data likelihood with the priors distributions for β and Σ, estimated at $\tilde{\beta}$ and $\tilde{\Sigma}$), the numerator is unknown. Indeed, as shown by 3.2.3, only the kernel of the posterior distribution $\pi(\beta, \Sigma\,|y)$ is computed in practice, since the normalising term $f(y)$ in 3.2.2 is unknown (indeed, this term is the marginal likelihood, to be estimated). Therefore, it is not possible to use the kernel of the posterior density obtained from 3.2.3 in place of $\pi(\tilde{\beta}, \tilde{\Sigma}\,|y)$, which is the full posterior value. The solution proposed by Chib (1995) consists first in noticing that $\pi(\tilde{\beta}, \tilde{\Sigma}\,|y)$ can be manipulated to obtain:

$$\pi(\tilde{\beta}, \tilde{\Sigma}\,|y) = \frac{\pi(\tilde{\beta}, \tilde{\Sigma}, y)}{\pi(y)} = \frac{\pi(\tilde{\beta}, \tilde{\Sigma}, y)}{\pi(\tilde{\Sigma}, y)}\frac{\pi(\tilde{\Sigma}, y)}{\pi(y)} = \pi(\tilde{\beta}\,|\tilde{\Sigma}, y)\pi(\tilde{\Sigma}\,|y) \tag{3.9.33}$$

The first term on the right-hand side of 3.9.33 is the conditional posterior distribution of β. It is necessarily known since it is required to run the Gibbs sampling algorithm (in the case of the independent normal Wishart, it is simply the multivariate normal distribution). The second term is the marginal posterior distribution of Σ. It is unknown analytically, but can be approximated using the identity:

$$\pi(\tilde{\Sigma}\,|y) = \int \pi(\tilde{\Sigma}, \beta\,|y)d\beta = \int \frac{\pi(\tilde{\Sigma}, \beta, y)}{\pi(y)}d\beta = \int \frac{\pi(\tilde{\Sigma}, \beta, y)}{\pi(\beta, y)}\frac{\pi(\beta, y)}{\pi(y)}d\beta = \int \pi(\tilde{\Sigma}\,|\beta, y)\pi(\beta\,|y)d\beta \tag{3.9.34}$$

While the integral in 3.9.34 is not analytically computable, an approximation of this term can be estimated from:

$$\pi(\tilde{\Sigma}\,|y) = \int \pi(\tilde{\Sigma}\,|\beta, y)\pi(\beta\,|y)d\beta \approx \frac{1}{(It - Bu)}\sum_{n=1}^{It-Bu} \pi(\tilde{\Sigma}\,|\beta^{(n)}, y) \tag{3.9.35}$$

Then, from 3.9.33 and 3.9.35, it is possible to rewrite 3.9.32 as:

$$m(y) = \frac{f(y \,|\tilde{\beta}, \tilde{\Sigma})\pi(\tilde{\beta})\pi(\tilde{\Sigma})}{\pi(\tilde{\beta}\,|\tilde{\Sigma}, y)(It - Bu)^{-1}\sum_{n=1}^{It-Bu} \pi(\tilde{\Sigma}\,|\beta^{(n)}, y)} \tag{3.9.36}$$

Combining the likelihood function A.2.3.9 with the priors 3.3.13 and 3.9.19, and simplifying, one eventually obtains:

$$
\begin{aligned}
m(y) = & \frac{(2\pi)^{-nT/2}}{2^{\alpha_0 n/2}\Gamma_n\left(\frac{\alpha_0}{2}\right)}|S_0|^{\alpha_0/2}\left|\tilde{\Sigma}\right|^{-(T+\alpha_0+n+1)/2}\left(\frac{|\Omega_0|}{|\bar{\Omega}|}\right)^{-1/2} \\
& \times \exp\left(-\frac{1}{2}tr\left[\tilde{\Sigma}^{-1}\left\{(Y-X\tilde{B})'(Y-X\tilde{B})+S_0\right\}\right]\right) \\
& \times \exp\left(-\frac{1}{2}(\tilde{\beta}-\beta_0)'\Omega_0^{-1}\left(\tilde{\beta}-\beta_0\right)\right) \times \frac{1}{(It-Bu)^{-1}\sum_{n=1}^{It-Bu}\pi(\tilde{\Sigma}\,|\beta^{(n)},y)}
\end{aligned}
\tag{3.9.37}
$$

where $\bar{\Omega}$ is defined as in 3.5.9, and $\tilde{B}$ is simply $\tilde{\beta}$ reshaped to be of dimension $k \times n$.

Numerical issues with the marginal likelihood for the independent normal-Wishart prior

3.9.37 provides an accurate approximation of the marginal likelihood value, but it can also suffer from numerical instability. It is thus preferable to rewrite it to obtain a more stable and efficient formulation. After some manipulations, one obtains:

$$
\begin{aligned}
m(y) = & \frac{(2\pi)^{-nT/2}}{2^{\alpha_0 n/2}\Gamma_n\left(\frac{\alpha_0}{2}\right)}|S_0|^{\alpha_0/2}\left|\tilde{\Sigma}\right|^{-(T+\alpha_0+n+1)/2} \times \left|I_q + F_\Omega'\left(I_n \otimes X'\right)\left(\tilde{\Sigma}^{-1}\otimes X\right)F_\Omega\right|^{-1/2} \\
& \times \exp\left(-\frac{1}{2}tr\left[\tilde{\Sigma}^{-1}\left\{(Y-X\tilde{B})'(Y-X\tilde{B})+S_0\right\}\right]\right) \times \exp\left(-\frac{1}{2}(\tilde{\beta}-\beta_0)'\Omega_0^{-1}\left(\tilde{\beta}-\beta_0\right)\right) \\
& \times \frac{1}{(It-Bu)^{-1}\sum_{n=1}^{It-Bu}\pi(\tilde{\Sigma}\,|\beta^{(n)},y)}
\end{aligned}
\tag{3.9.38}
$$

where F_Ω denotes the square root matrix of Ω_0, that is, $F_\Omega F_\Omega' = \Omega_0$. The determinant term in the first row can be computed, as usual from A.1.18, by taking the products of 1 plus the eigenvalues of $F_\Omega'\left(I_n \otimes X'\right)\left(\tilde{\Sigma}^{-1}\otimes X\right)F_\Omega$.

Once again, it is more convenient to work with logs rather than with values in level. Therefore, 3.9.38 is rewritten as:

$$\log\left(m(y)\right) = -\frac{nT}{2}\log(2\pi) - \frac{\alpha_0 n}{2}\log(2) - \log\left(\Gamma_n\left(\frac{\alpha_0}{2}\right)\right) + \frac{\alpha_0}{2}\log\left(|S_0|\right) - \frac{T+\alpha_0+n+1}{2}\log\left(\left|\tilde{\Sigma}\right|\right)$$
$$- \frac{1}{2}\log\left(\left|I_q + F_\Omega^{,}\left(I_n \otimes X^{,}\right)\left(\tilde{\Sigma}^{-1}\otimes X\right)F_\Omega\right|\right) - \frac{1}{2}tr\left[\tilde{\Sigma}^{-1}\left\{(Y-X\tilde{B})^{,}(Y-X\tilde{B}) + S_0\right\}\right]$$
$$- \frac{1}{2}(\tilde{\beta}-\beta_0)^{,}\Omega_0^{-1}\left(\tilde{\beta}-\beta_0\right) - \log\left((It-Bu)^{-1}\sum_{n=1}^{It-Bu}\pi(\tilde{\Sigma}\left|\beta^{(n)},y\right.)\right) \tag{3.9.39}$$

Deriving the marginal likelihood for the normal diffuse and dummy priors

In the case of the normal diffuse prior, it is not possible to derive the marginal likelihood. This can be seen simply from 3.9.2:

$$m(y) = \int f(y\,|\theta)\pi(\theta)d\theta \tag{3.9.40}$$

Since for the normal-diffuse, $\theta = \beta, \Sigma$, and β and Σ are independent, this reformulates as:

$$m(y) = \iint f\left(y\,|\beta,\Sigma\right)\pi(\beta)\pi(\Sigma)d\beta d\Sigma \tag{3.9.41}$$

The problem, as shown by 3.6.3, is that $\pi(\Sigma)$ is an improper prior. Because it is diffuse and does not integrate to one, only its kernel is known:

$$\pi(\Sigma) \propto |\Sigma|^{-(n+1)/2} \tag{3.9.42}$$

This renders estimation of 3.9.40 impossible, as the full proper prior $\pi(\Sigma)$ is required, and not just the kernel. For the dummy prior the reasoning is exactly the same.

Deriving the marginal likelihood when dummy extensions are applied

When dummy observations are included in the estimation process, it is not correct to calculate naively the marginal likelihood over the total data set (comprising both actual and dummy observations), as the dummy observations are now included into the estimation process, and hence their contributions must be taken into account if one wants to obtain the marginal likelihood for actual data only.

Concretely, and following 3.7.44, denote by Y^* the total data set (comprising both actual and dummy observations) and by $Y^{dum} = \begin{pmatrix} Y_s \\ Y_o \end{pmatrix}$ the set of dummy observations. Denote their respective marginal likelihoods by $m(y^*)$ and $m(y^{dum})$. The object of interest is $m(y)$, the marginal likelihood for the actual data only. Notice that from 3.9.28, the marginal likelihood is just equal to the density

of the data. Therefore, the conventional properties of densities apply. Then, using the fact that Y and Y^{dum} are independent, one obtains:

$$m(y^*) = m(y, y^{dum}) = m(y) \times m(y^{dum}) \tag{3.9.43}$$

It follows that:

$$m(y) = \frac{m(y^*)}{m(y^{dum})} \tag{3.9.44}$$

That is, the marginal likelihood for the data can be recovered by dividing the marginal likelihood for the total data set by that of the dummy observations alone.

Optimizing hyperparameter values from a grid search:

In their seminal paper, Giannone et al. (2012) introduce a procedure allowing to optimise the values of the hyperparameters. Optimal here means that the estimated hyperparameters maximise the value of the marginal likelihood for the model. This implies that the hyperparameter values are not given anymore but have to be estimated within the Bayesian framework, adopting the hierarchical approach described in section 3.2. A similar but simpler approach consists into running a grid search. To do so, one defines for every hyperparameter of the model a minimum and a maximum value (hence defining a range) along with a step size defining the size of the increment within the range. Then the marginal likelihood can be estimated for the model for every possible combination of hyperparameter values within the specified ranges, and the optimal combination is retained as that which maximises this criterion.

4 Basic applications

4.1 Forecasts

Forecasts are a central issue for economists. In a traditional frequentist approach, obtaining forecasts is relatively straightforward, by using what is known as the chain rule of forecasting. Indeed, consider again the VAR model 3.1.2, or its vectorised form 3.1.12. With the frequentist approach, estimation of the model would produce a single value of $\hat{\beta}$, the vector of VAR coefficients, which would be used as the estimate for the true parameter value β. Equivalently, under its original form, estimation of the model would produce $\hat{A}_1, \hat{A}_2, ..., \hat{A}_p$ and $\hat{C}$, which would be used as estimates for the true parameter values $A_1, A_2, ..., A_p$ and C. Once these coefficients are estimated, the chain rule of forecasting simply consists in constructing forecasts for $\tilde{y}_{T+1}, \tilde{y}_{T+2}, ..., \tilde{y}_{T+h}$ sequentially by using conditional expectation in the following way [3]:

$$\begin{aligned}
\tilde{y}_{T+1} &= E_t(y_{T+1}\,|y_T) \\
&= E_t(\hat{A}_1 y_T + \hat{A}_2 y_{T-1} + ... + \hat{A}_p y_{T+1-p} + \hat{C}x_{T+1} + \varepsilon_{T+1}\,|y_T) \\
&= \hat{A}_1 E_t(y_T\,|y_T) + \hat{A}_2 E_t(y_{T-1}\,|y_T) + ... + \hat{A}_p E_t(y_{T+1-p}\,|y_T) + \hat{C}E_t(x_{T+1}\,|y_T) + E_t(\varepsilon_{T+1}\,|y_T) \\
&= \hat{A}_1 y_T + \hat{A}_2 y_{T-1} + ... + \hat{A}_p y_{T+1-p} + \hat{C}x_{T+1}
\end{aligned} \tag{4.1.1}$$

where $\tilde{y}_{T+1}$ denotes the predicted value for y_{T+1}, and $y_T = \{y_t\}_{t=1}^T$ is the information set containing data observed up to period T. Applying the same method for $\tilde{y}_{T+2}$ yields:

$$\begin{aligned}
\tilde{y}_{T+2} &= E_t(y_{T+2}\,|y_T) \\
&= E_t(\hat{A}_1 y_{T+1} + \hat{A}_2 y_T + ... + \hat{A}_p y_{T+2-p} + \hat{C}x_{T+2} + \varepsilon_{T+2}\,|y_T) \\
&= \hat{A}_1 E_t(y_{T+1}\,|y_T) + \hat{A}_2 E_t(y_T\,|y_T) + ... + \hat{A}_p E_t(y_{T+2-p}\,|y_T) + \hat{C}E_t(x_{T+2}\,|y_T) + E_t(\varepsilon_{T+2}\,|y_T) \\
&= \hat{A}_1 \tilde{y}_{T+1} + \hat{A}_2 y_T + ... + \hat{A}_p y_{T+2-p} + \hat{C}x_{T+2}
\end{aligned} \tag{4.1.2}$$

Going on:

$$\begin{aligned}
\tilde{y}_{T+3} &= \hat{A}_1 \tilde{y}_{T+2} + \hat{A}_2 \tilde{y}_{T+1} + ... + \hat{A}_p y_{T+3-p} + \hat{C}x_{T+2} \\
&\vdots \\
\tilde{y}_{T+h} &= \hat{A}_1 \tilde{y}_{T+h} + \hat{A}_2 \tilde{y}_{T+h-1} + ... + \hat{A}_p \tilde{y}_{T+h+1-p} + \hat{C}x_{T+h}
\end{aligned} \tag{4.1.3}$$

[3] It can be shown that this estimator minimizes the mean squared forecast error (see Luetkepohl (1993), chapter 2)

One would like to use this methodology in a Bayesian context as well. This is however not possible as this method ignores the parameter uncertainty proper to Bayesian analysis. In other words, in a Bayesian framework, the "correct" parameter values β or $A_1, A_2, ..., A_p, C$ do not exist, they are only characterised by a distribution from which realisations can be obtained. This implies that forecasts cannot be obtained either as a single value: they face the same uncertainty as the parameters, and are thus characterised also by a distribution rather than a single estimate. Such a distribution is called the posterior predictive distribution, and is a central object of interest for Bayesian practitioners. It is denoted by:

$$f(y_{T+1:T+h}\,|y_T\,) \tag{4.1.4}$$

where $f(y_{T+1:T+h}\,|y_T\,)$ denotes the distribution of future datapoints $y_{T+1}, ..., y_{T+h}$, conditional on the information set y_T. This posterior predictive distribution is the fundamental element used in Bayesian forecasting: once it is obtained, it is straightforward to produce confidence intervals and point estimates, using 3.2.16 and 3.2.17. The main difficulty thus consists in identifying the form of this posterior distribution. Although analytical solutions may exist, they may not produce tractable implementations. In practice, the following solution is thus favoured: rearrange 4.1.4 to obtain:

$$\begin{aligned} f(y_{T+1:T+h}|y_T) &= \int_\theta f(y_{T+1:T+h}, \theta|y_T)d\theta \qquad (4.1.5)\\ &= \int_\theta \frac{f(y_{T+1:T+h}, \theta, y_T)}{f(y_T)}d\theta \\ &= \int_\theta \frac{f(y_{T+1:T+h}, \theta, y_T)}{f(y_T, \theta)}\frac{f(y_T, \theta)}{f(y_T)}d\theta \\ &= \int_\theta f(y_{T+1:T+h}|\theta, y_T)f(\theta|y_T)d\theta \qquad (4.1.6) \end{aligned}$$

where θ denotes the parameters of interest for the VAR model (typically, β and Σ). 4.1.6 shows that the posterior predictive distribution rewrites as an (integrated) product of two distributions: the posterior distribution, and the distribution of future observations, conditional on data and parameter values. This way, 4.1.6 suggests a direct simulation method to obtain draws from $f(y_{T+1:T+h}\,|y_T\,)$. Suppose one can generate random draws from the posterior distribution $f(\theta\,|y_T\,)$, and then, from these drawn values and y_T, compute a sequence $y_{T+1}, y_{T+2}, ..., y_{T+h}$. This would form a random draw from $f(y_{T+1:T+h}\,|\theta, y_T\,)$, so that the product with the draw from $f(\theta\,|y_T\,)$ would give a draw from $f(y_{T+1:T+h}, \theta\,|y_T\,)$, as indicated by 4.1.6. Marginalizing, which simply implies to discard the values for θ, would then produces a draw from $f(y_{T+1:T+h}\,|y_T\,)$.

But doing so is easy. In the case of the independent normal-Wishart and normal-diffuse priors developed in subsections 3.5 and 3.6., draws from $f(\theta\,|y_T)$ are already obtained from the Gibbs sampling process of the BVAR estimation. For the Minnesota and normal-Wishart priors, suffice is to run a comparable algorithm. Once estimates for the Gibbs algorithm are obtained, the procedure is simple: for each iteration of the algorithm, a sequence $y_{T+1}, y_{T+2}, ..., y_{T+h}$ can be obtained by drawing h series of residuals, then constructing $(y_{T+1}\,|\theta, y_T)$, $(y_{T+2}\,|\theta, y_{T+1}, y_T)$, $(y_{T+3}\,|\theta, y_{T+2}, y_{T+1}, y_T)$ and so on, recursively using 3.1.2.

Karlsson (2012) hence proposes the following algorithm to derive the posterior predictive distribution. Note that the algorithm supposes that random draws from the posterior distributions of β and Σ are available.

Algorithm 2.1.1 (forecasts, all priors):

1. Define the number of iterations $(It - Bu)$ of the algorithm. Because the Gibbs algorithm previously run for the model estimation phase has already produced $(It - Bu)$ draws from the posterior distribution of β and Σ, an efficient strategy consists in recycling those draws, rather than running again the whole Gibbs sampler. Therefore, $(It - Bu)$ iterations are sufficient. Also, define the forecast horizon h.

2. At iteration n, draw $\Sigma_{(n)}$ and $\beta_{(n)}$ from their posterior distributions. Simply recycle draw n from the Gibbs sampler.

3. Draw simulated series of residuals $\tilde{\varepsilon}^{(n)}_{T+1}, \tilde{\varepsilon}^{(n)}_{T+2}, ..., \tilde{\varepsilon}^{(n)}_{T+h}$ from $N(0, \Sigma_{(n)})$.

4. Generate recursively the simulated values $\tilde{y}^{(n)}_{T+1}, \tilde{y}^{(n)}_{T+2}, ..., \tilde{y}^{(n)}_{T+h}$ from 3.1.2: $\tilde{y}^{(n)}_{T+1} = A_1 y_T + A_2 y_{T-1} + ... + A_p y_{T+1-p} + C x_{T+1} + \tilde{\varepsilon}^{(n)}_{T+1}$ Once $\tilde{y}_{T+1}$ is computed, use: $\tilde{y}^{(n)}_{T+2} = A_1 \tilde{y}^{(n)}_{T+1} + A_2 y_T + ... + A_p y_{T+2-p} + C x_{T+2} + \tilde{\varepsilon}^{(n)}_{T+2}$ And continue this way until $\tilde{y}^{(n)}_{T+h}$ is obtained. The values of $c, A_1', A_2' ... A_p'$ come from $\beta_{(n)}$. Note that each iteration may involve both actual data (regressors up to period T:y_T, y_{T-1}, y_{T-2}...) and simulated values (forecasts $\tilde{y}^{(n)}_{T+1}, \tilde{y}^{(n)}_{T+2}$... and simulated residual series $\tilde{\varepsilon}^{(n)}_{T+1}, \tilde{\varepsilon}^{(n)}_{T+2}, ...$). Notice also that the forecasts include the exogenous values $x_{T+1}, x_{T+2} ... x_{T+h}$, which have hence to be known and provided exogenously.

5. Discard $\Sigma_{(n)}$ and $\beta_{(n)}$ to obtain draws $\tilde{y}^{(n)}_{T+1}, \tilde{y}^{(n)}_{T+2} ... \tilde{y}^{(n)}_{T+h}$ from the predictive distribution $f(y_{T+1:T+h}\,|y_T)$.

6. Repeat until $(It - Bu)$ iterations are realised. This produces:
$\left\{ \tilde{y}^{(n)}_{T+1}\,|y_T\,, \tilde{y}^{(n)}_{T+2}\,|y_T\,, ..., \tilde{y}^{(n)}_{T+h}\,|y_T \right\}_{n=1}^{It-Bu}$,
a sample of independent draws from the joint predictive distribution which can be used for inference and computation of point estimates.

Because algorithm 3.2.1.1 requires the series $\Sigma_{(1)}, \Sigma_{(2)}, ..., \Sigma_{(It-Bu)}$ and $\beta_{(1)}, \beta_{(2)}, ..., \beta_{(It-Bu)}$, it is necessary to run a Gibbs sampling algorithm to obtain them, even if the retained prior (for instance the Minnesota or the normal-Wishart prior) yields analytical solutions. Therefore, to conclude this subsection, the Gibbs algorithm for these two priors is presented. They are very similar to algorithm 3.1.5.1, but are somewhat simpler as the priors admit analytical solutions. Indeed, no burn-in or transition sample is required since the unconditional posterior distribution is known, so that draws from the correct marginal distribution can be realised as soon as the algorithm starts. In the case of the Minnesota prior, this means drawing β directly from a $\mathcal{N}(\bar{\beta}, \bar{\Omega})$. In the normal-Wishart case, this means drawing B directly from $B \sim \mathcal{MT}(\tilde{\alpha}, \bar{B}, \bar{S}, \bar{\Phi})$, and Σ directly from $\mathcal{IW}(\bar{S}, \bar{\alpha})$.

Algorithm 2.1.2 (Gibbs sampling with a Minnesota prior):

1. Set the number of iterations of the algorithm as $(It - Bu)$. Since draws are directly realised form the posterior distribution, no burn-in sample is required.

2. Fix the value Σ for the BVAR model, in accordance with the prior.

3. At iteration n, draw the value $\beta_{(n)}$ conditional on Σ, from a multivariate normal with mean $\bar{\beta}$ and variance-covariance matrix $\bar{\Omega}$, as defined in 3.3.17 and 3.3.18: $\beta_{(n)} \sim \mathcal{N}(\bar{\beta}, \bar{\Omega})$ with: $\bar{\Omega} = \left[\Omega_0^{-1} + \Sigma^{-1} \otimes X`X\right]^{-1}$ and $\bar{\beta} = \bar{\Omega}\left[\Omega_0^{-1}\beta_0 + (\Sigma^{-1} \otimes X`)y\right]$

4. repeat until $(It - Bu)$ iterations are realized.

Algorithm 2.1.3 (Gibbs sampling with a Normal-Wishart prior):

1. Set the number of iterations of the algorithm as $(It - Bu)$. Since draws are directly realised form the posterior distribution, no burn-in sample is required.

2. At iteration n, draw the value $B_{(n)}$ from a matrix-variate student distribution, as defined by 3.4.16, 3.4.17, 3.4.19, 3.4.23 and 3.4.24: $B_{(n)} \sim \mathcal{MT}(\bar{B}, \bar{S}, \bar{\Phi}, \tilde{\alpha})$ with $\tilde{\alpha} = T + \alpha_0 - n + 1$, $\bar{\Phi} = \left[\Phi_0^{-1} + X`X\right]^{-1}$, $\bar{B} = \bar{\Phi}\left[\Phi_0^{-1}B_0 + X`Y\right]^{-1}$ and $\bar{S} = Y`Y + S_0 + B_0`\Phi_0^{-1}B_0 - \bar{B}`\bar{\Phi}^{-1}\bar{B}$

3. at iteration n, draw the value $\Sigma_{(n)}$ from an inverse Wishart distribution, as defined by 3.4.18 and 3.4.19: $\Sigma_{(n)} \sim \mathcal{IW}\left(\bar{\alpha}, \bar{S}\right)$ with $\bar{\alpha} = T + \alpha_0$ and $\bar{S} = \left(Y - X\hat{B}\right)`\left(Y - X\hat{B}\right) + S_0 + \hat{B}`X`X\hat{B} + B_0`\Phi_0^{-1}B_0 - \bar{B}`\bar{\Phi}^{-1}\bar{B}$

4. repeat until $(It - Bu)$ iterations are realized.

Finally, note that closely associated to the concept of forecast is that of forecast evaluation, namely how good is the estimated model at producing forecast values (or distributions) which are close to the realised values. There exist many criteria for this: some of them only compare the realised value

with the forecast value, while others compare the realised value with the whole posterior distribution for the forecasts. For more details on the different forecast evaluation criteria and their estimation, you may refer to Appendix A.10.

4.2 Impulse response functions

A second object of interest for economists is what is known as impulse response functions. The idea is related to the study of the dynamic effect of shocks. Imagine that the VAR model 3.1.2 is at its long-run value and that shocks (that is, the error terms or residuals) have zero values at every period: then the model is stable and remain at the same value over time[4]. Imagine now that a one-time shock occurs for a single variable at some period t, before all the shocks return to zero for all subsequent periods. Then the effect of this shock will propagate to all the variables over the subsequent periods $t+1, t+2, ...$

Formally, start again from model 3.1.2:

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + ... + A_p y_{t-p} + C x_t + \varepsilon_t \tag{4.2.1}$$

This model can be reformulated as:

$$\begin{aligned} & y_t = A_1 y_{t-1} + A_2 y_{t-2} + ... + A_p y_{t-p} + C x_t + \varepsilon_t \\ \Leftrightarrow\ & y_t = (A_1 L + A_2 L^2 ... + A_p L^p) y_t + C x_t + \varepsilon_t \\ \Leftrightarrow\ & (I - A_1 - A_2 L^2 ... - A_p L^p) y_t = C x_t + \varepsilon_t \\ \Leftrightarrow\ & A(L) y_t = C x_t + \varepsilon_t \end{aligned} \tag{4.2.2}$$

where $A(L)$ denotes the lag polynomial in y_t, defined as:

$$A(L) \equiv I - A_1 - A_2 L^2 ... - A_p L^p \tag{4.2.3}$$

Under the assumption of covariance stationarity, the Wold theorem implies that it is possible to "invert" this lag polynomial, that is, it is possible to reformulate the model as an infinite order moving average:

[4] That is, if the exogenous variables remain constant as well

$$
\begin{aligned}
& A(L)y_t = Cx_t + \varepsilon_t \\
& \Leftrightarrow y_t = A(L)^{-1}\left[Cx_t + \varepsilon_t\right] \\
& \Leftrightarrow y_t = A(L)^{-1}Cx_t + A(L)^{-1}\varepsilon_t \\
& \Leftrightarrow y_t = A(L)^{-1}Cx_t + \sum_{i=0}^{\infty}\Psi_i\varepsilon_{t-i} \\
& \Leftrightarrow y_t = A(L)^{-1}Cx_t + \Psi_0\varepsilon_t + \Psi_1\varepsilon_{t-1} + \Psi_2\varepsilon_{t-2}\ldots
\end{aligned} \tag{4.2.4}
$$

The moving average part is made of a sequence of $n \times n$ matrices Ψ_0, Ψ_1, Ψ_2..., where each matrix is of the form:

$$
\Psi_i = \begin{pmatrix} \phi_{i,11} & \phi_{i,12} & \cdots & \phi_{i,1n} \\ \phi_{i,21} & \phi_{i,22} & \cdots & \phi_{i,2n} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{i,n1} & \phi_{i,n2} & \cdots & \phi_{i,nn} \end{pmatrix} \tag{4.2.5}
$$

The sequence Ψ_0, Ψ_1, Ψ_2... actually represents the impulse response functions for the model 4.2.1. To see this, move model 4.2.4 forward by h periods to obtain:

$$
y_{t+h} = A(L)^{-1}Cx_{t+h} + \Psi_0\varepsilon_{t+h} + \Psi_1\varepsilon_{t+h-1} + ... + \Psi_{h-1}\varepsilon_{t+1} + \Psi_h\varepsilon_t + \Psi_{h+1}\varepsilon_{t-1} + \ldots \tag{4.2.6}
$$

Then, the impact of the shock ε_t on y_{t+h} can be obtained by:

$$
\frac{\partial y_{t+h}}{\partial \varepsilon_t} = \Psi_h \tag{4.2.7}
$$

And this shows that Ψ_h represents the impulse response function for the shock ε_t on y_{t+h}. More precisely, from 4.2.6 and 4.2.5, it is possible to obtain the impact of the shock in variable j on variable $y_{i,t+h}$ as:

$$
\frac{\partial y_{i,t+h}}{\partial \varepsilon_{j,t}} = \phi_{h,ij} \tag{4.2.8}
$$

While there exists elegant analytical methods to invert a lag polynomial and compute Ψ_0, Ψ_1, Ψ_2..., in actual applications it is simpler to rely on simulation methods. To see this, go back to model 4.2.1 and compute the impact of ε_t on y_t:

$$
\frac{\partial y_t}{\partial \varepsilon_t} = I = \Psi_0 \tag{4.2.9}
$$

Move the model forward by one period:

$$\begin{aligned} y_{t+1} &= A_1 y_t + A_2 y_{t-1} + ... + A_p y_{t+1-p} + C x_{t+1} + \varepsilon_{t+1} \\ &= A_1 \left(A_1 y_{t-1} + A_2 y_{t-2} + ... + A_p y_{t-p} + C x_t + \varepsilon_t \right) \\ &+ A_2 y_{t-1} + ... + A_p y_{t+1-p} + C x_{t+1} + \varepsilon_{t+1} \end{aligned} \tag{4.2.10}$$

or:

$$\begin{aligned} y_{t+1} &= A_1 A_1 y_{t-1} + A_1 A_2 y_{t-2} + ... + A_1 A_p y_{t-p} + A_1 C x_t + A_1 \varepsilon_t \\ &+ A_2 y_{t-1} + ... + A_p y_{t+1-p} + C x_{t+1} + \varepsilon_{t+1} \end{aligned} \tag{4.2.11}$$

From 4.2.11, it is straightforward to obtain:

$$\frac{\partial y_{t+1}}{\partial \varepsilon_t} = A_1 = \Psi_1 \tag{4.2.12}$$

Move again the model forward by one period and use 4.2.1 and 4.2.11:

$$\begin{aligned} y_{t+2} &= A_1 y_{t+1} + A_2 y_t + ... + A_p y_{t+2-p} + C x_{t+2} + \varepsilon_{t+2} \\ &= A_1 (A_1 A_1 y_{t-1} + A_1 A_2 y_{t-2} + ... + A_1 A_p y_{t-p} + A_1 C x_t + A_1 \varepsilon_t + A_2 y_{t-1} + ... + A_p y_{t+1-p} + C x_{t+1} + \varepsilon_{t+1}) \\ &+ A_2 (A_1 y_{t-1} + A_2 y_{t-2} + ... + A_p y_{t-p} + C x_t + \varepsilon_t) + ... + A_p y_{t+2-p} + C x_{t+2} + \varepsilon_{t+2} \end{aligned} \tag{4.2.13}$$

or:

$$\begin{aligned} y_{t+2} &= A_1 A_1 A_1 y_{t-1} + A_1 A_1 A_2 y_{t-2} + ... + A_1 A_1 A_p y_{t-p} + A_1 A_1 C x_t + A_1 A_1 \varepsilon_t + A_1 A_2 y_{t-1} \\ &+ ... + A_1 A_p y_{t+1-p} + A_1 C x_{t+1} + A_1 \varepsilon_{t+1} \\ &+ A_2 A_1 y_{t-1} + A_2 A_2 y_{t-2} + ... + A_2 A_p y_{t-p} + A_2 C x_t + A_2 \varepsilon_t + ... + A_p y_{t+2-p} + C x_{t+2} + \varepsilon_{t+2} \end{aligned} \tag{4.2.14}$$

Then, obtain from 4.2.14:

$$\frac{\partial y_{t+2}}{\partial \varepsilon_t} = A_1 A_1 + A_2 = \Psi_2 \tag{4.2.15}$$

Going on, it is possible to recover numerically the whole sequence $\Psi_0, \Psi_1, \Psi_2, \Psi_3, \Psi_4 ...$ Note that this method is in essence very similar to the chain rule of forecasts developed in the previous

subsection.

The problem with this method is that it implies the computation of partial derivatives, a task that numerical softwares are not able to perform. However, looking at the recursive sequence 4.2.1, 4.2.11, 4.2.14, and so on, it is clear that only the terms related to ε_t are of interest. Therefore, it is possible to recover numerically $\phi_{0,ij}, \phi_{1,ij}, \phi_{2,ij}, \ldots$, the response of variable i to shock j over periods $0, 1, 2, \ldots$ directly from these equations by setting a unit shock at period t, and switching off to 0 all the other values of the recursive sequence which are of no interest. Precisely, set

$$\varepsilon_t = \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \vdots \\ \varepsilon_{n,t} \end{pmatrix} \tag{4.2.16}$$

such that

$$\begin{cases} \varepsilon_{j,t} = 1 \\ \varepsilon_{\neq j,t} = 0 \end{cases} \tag{4.2.17}$$

and

$$y_{t-1}, y_{t-2}, ..., x_t, x_{t+1}, x_{t+2}, ..., \varepsilon_{t+1}, \varepsilon_{t+2}, \cdots = 0 \tag{4.2.18}$$

Then run the recursive sequence 4.2.1, 4.2.11, 4.2.14. Do this for all variables and all shocks, and recover finally Ψ_h from 4.2.5 by gathering all the $\phi_{h,ij}$ elements. Finally, it should be clear from 4.2.1 that assuming $\varepsilon_{i,t} = 1$ and $y_{t-1}, y_{t-2}, ..., x_t = 0$ is equivalent to assuming $y_{t-1}, y_{t-2}, ..., x_t, \varepsilon_t = 0$ and $y_{i,t} = 1$, a formulation that will prove more convenient for numerical computations.

As for the forecasts, it is not possible to apply directly this simulation method in a Bayesian framework, due to the additional parameter uncertainty that prevails in a Bayesian framework. Fortunately, it is easy to integrate the impulse response calculation into the Gibbs sampling framework previously developed. Note in particular that the simulation experiment described above really is equivalent to producing forecasts for $\tilde{y}_{t+1}, \ldots, \tilde{y}_{t+h}$ in the particular context of the simulation experiment. Therefore, intuitively, one may like to use the methodology introduced in subsubsection 4.1 for forecasts, and adapt it to impulse response functions. This is indeed what will be done: to calculate the impulse response functions, simply obtain the posterior predictive distribution 4.1.4:

$$f(y_{t+1:t+h} \,|y_t) \tag{4.2.19}$$

for $h = 1, 2, \ldots$

conditional on:

- $y_{t-1}, y_{t-2}, \cdots = 0$
- $y_{i,t} = 1$ for some $i \in 1, 2, \ldots, n$, and $y_{j,t} = 0$ for $j \neq i$
- $\varepsilon_{t+1}, \varepsilon_{t+2}, \cdots = 0$
- $x_t, x_{t+1}, x_{t+2} \cdots = 0$

Following, it is straightforward to adapt algorithm 2.1.1 to impulse response functions:

Algorithm 2.2.1 (impulse response functions, all priors):

1. Define the number of iterations $(It - Bu)$ of the algorithm, and the time horizon h.
2. Fix $i = 1$. Then set $y_{i,T} = 1$.
3. At iteration n, draw $\beta_{(n)}$ from its posterior distributions. Simply recycle draw n from the Gibbs sampler.
4. Generate recursively the simulated values $\tilde{y}_{T+1}^{(n)}, \tilde{y}_{T+2}^{(n)}, ..., \tilde{y}_{T+h}^{(n)}$ from 3.1.2: $\tilde{y}_{T+1}^{(n)} = A_1 y_T + A_2 y_{T-1} + ... + A_p y_{T+1-p}$ Once $\tilde{y}_{T+1}$ is computed, use: $\tilde{y}_{T+2}^{(n)} = A_1 \tilde{y}_{T+1}^{(n)} + A_2 y_T + ... + A_p y_{T+2-p}$ And continue this way until $\tilde{y}_{T+h}^{(n)}$ is obtained. Once again, both the exogenous terms and the shocks are ignored since they are assumed to take a value of 0 at all periods. The values of $A_1', A_2' ... A_p'$ come from $\beta_{(n)}$.
5. Discard $\beta_{(n)}$ to obtain draws $\tilde{y}_{T+1}^{(n)}, \tilde{y}_{T+2}^{(n)} ... \tilde{y}_{T+h}^{(n)}$ from the predictive distribution $f(y_{T+1:T+h} \,|y_T)$.
6. Repeat until $(It - Bu)$ iterations have been performed. This produces: $\left\{ \tilde{y}_{T+1}^{(n)} \,|y_T \,, \tilde{y}_{T+2}^{(n)} \,|y_T \,, ..., \tilde{y}_{T+h}^{(n)} \,|y_T \right\}$ a sample of independent draws from the joint predictive distribution in the case $y_{i,T} = 1$.
7. Go back to step 2, and fix $i = 2$. Then go through steps 3-6 all over again. Then repeat the process for $i = 3, ..., n$. This generates the impulse response functions for all the shocks in the model.

4.3 Structural VARs

An issue arising with conventional impulse response functions is that they arise from correlated shocks (that is, the residual covariance matrix Σ of the reduced form VAR is typically not diagonal). Therefore a statement such as "the impulse response function reflects the response of variable y_i following a shock in variable y_j, everything else being constant" is actually meaningless when shocks

typically arise together. A solution to such a problem can be found in structural VARs, which permit to obtain the responses of variables to orthogonal shocks.

Consider again model 3.1.2:

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + ... + A_p y_{t-p} + C x_t + \varepsilon_t \tag{4.3.1}$$

with $\varepsilon_t \sim \mathcal{N}(0, \Sigma)$. This is a reduced-form VAR model. An alternative specification of this model is the structural VAR model:

$$D_0 y_t = D_1 y_{t-1} + D_2 y_{t-2} + ... + D_p y_{t-p} + F x_t + \eta_t \tag{4.3.2}$$

with $\eta_t \sim \mathcal{N}(0, \Gamma)$ a vector of structural innovations with variance-covariance matrix Γ. Most of the time, one wants Γ to be diagonal. This will imply that the shocks in η_t are mutually orthogonal, which is what is required to obtain meaningful impulse response functions not suffering from the shock correlation issue. Also, from a more theoretical point of view, it makes also sense to assume that structural shocks (say e.g. supply shocks, demand shocks, monetary shocks, etc) are uncorrelated and arise independently.

For notation convenience, define:

$$D = D_0^{-1} \tag{4.3.3}$$

Then premultiplying both sides of 4.3.2 by D makes straightforward the relation between 4.3.1 and 4.3.2 :

$$A_i = D D_i \tag{4.3.4}$$

$$C = D F \tag{4.3.5}$$

$$\varepsilon_t = D \eta_t \tag{4.3.6}$$

4.3.6 shows that D can be interpreted as a structural matrix: it permits to recover structural innovations from the reduced-form VAR residuals. Note that 4.3.6 also implies:

$$\Sigma = E\left(\varepsilon_t \varepsilon_t^{'}\right) = E\left(D \eta_t \eta_t^{'} D^{'}\right) = D E\left(\eta_t \eta_t^{'}\right) D^{'} = D \Gamma D^{'} \tag{4.3.7}$$

or

$$\Sigma = D \Gamma D^{'} \tag{4.3.8}$$

Another interesting feature of structural VARs is the moving average representation of the model. Remember from 4.2.4 that 4.3.1 can rewrite as an infinite order moving average process:

$$y_t = A(L)^{-1}Cx_t + \Psi_0\varepsilon_t + \Psi_1\varepsilon_{t-1} + \Psi_2\varepsilon_{t-2}... \tag{4.3.9}$$

where the series $I, \Psi_1, \Psi_2, \Psi_3, ...$ represents the impulse response functions of the reduced form VAR. Rewrite then 4.3.9 as:

$$y_t = A(L)^{-1}Cx_t + DD^{-1}\varepsilon_t + \Psi_1 DD^{-1}\varepsilon_{t-1} + \Psi_2 DD^{-1}\varepsilon_{t-2} + ... \tag{4.3.10}$$

And note that this implies from 4.3.6:

$$y_t = A(L)^{-1}Cx_t + D\eta_t + (\Psi_1 D)\,\eta_{t-1} + (\Psi_2 D)\,\eta_{t-2} + ... \tag{4.3.11}$$

This can be reformulated as:

$$y_t = A(L)^{-1}Cx_t + D\eta_t + \tilde{\Psi}_1\eta_{t-1} + \tilde{\Psi}_2\eta_{t-2} + ... \tag{4.3.12}$$

or:

$$y_t = A(L)^{-1}Cx_t + \sum_{i=0}^{\infty}\tilde{\Psi}_i\eta_{t-i} \tag{4.3.13}$$

where

$$\tilde{\Psi}_0 \equiv D \quad \text{and} \quad \tilde{\Psi}_i \equiv \Psi_i D, \quad \text{for } i = 1, 2, 3, \ldots \tag{4.3.14}$$

The series $D, \tilde{\Psi}_1, \tilde{\Psi}_2, \tilde{\Psi}_3, ...$ represents the impulse response functions of the structural VAR, that is, the response of the VAR variables to structural innovations. Then, as long as Γ is diagonal, the series of responses $D, \tilde{\Psi}_1, \tilde{\Psi}_2, \tilde{\Psi}_3, ...$ will result from independent shocks, and can be given meaningful economic interpretation.

If D was known, one could use estimates of the reduced form VAR 4.3.1 to identify the structural disturbances (from 4.3.6), and to obtain the responses of the variables to structural innovations (from 4.3.14). However, 4.3.1 provides no information about D. Some information is provided by 4.3.8: if the VAR 4.3.1 comprises n variables, then D comprises n^2 elements to identify, and Γ comprises $n \times (n+1)/2$ distinct elements, which hence makes a total of $(n/2)\,(3n+1)$elements to identify. Since in 4.3.8 the known elements of Σ provide only $n \times (n+1)/2$ restrictions on D and Γ, n^2 additional restrictions have to be implemented to identify D and Γ.

4.4 Identification by Choleski factorisation

The most common identification scheme for D and Γ is Choleski factorization. First, simplify the problem by assuming that Γ is diagonal, and simplify further by normalizing, that is, by assuming unit variance for all shocks. Then $\Gamma = I$, and 4.3.8 simplifies to:

$$\Sigma = DD' \tag{4.4.1}$$

Given the n^2 elements of D to identify, and the $n \times (n+1)/2$ restrictions provided by Σ, there remains an additional $n \times (n-1)/2$ restrictions to impose to identify D. A simple way to obtain them is to implement contemporaneous restrictions on D. Ignoring the exogenous variables, 4.3.2 rewrites:

$$y_t = D\eta_t + \tilde{\Psi}_1\eta_{t-1} + \tilde{\Psi}_2\eta_{t-2} + \dots \tag{4.4.2}$$

Looking at 4.4.2, it is easy to see that the contemporaneous effect $\frac{\partial y_t}{\partial \eta_t}$ of η_t on y_t is given by D. It is then common practice to identify D by assuming that certain variables do not respond contemporaneously to certain structural shocks.

For example, assuming that the second variable in y_t does not respond immediately to the third structural shock will be translated by the fact that entry $(2,3)$ of D is equal to 0. By selecting carefully the order in which variables enter into y_t, one may impose a set of $n \times (n-1)/2$ contemporaneous restrictions that will yield D to be lower triangular. Then, the $n \times (n-1)/2$ zero constraints on D will yield exact identification of it.

Identification itself is obtained as follows: the identification scheme amounts to finding a lower triangular matrix D such that $DD' = \Sigma$, with Σ a symmetric matrix. But this is precisely the definition of the Choleski factor of Σ.

Hence the process can be summarised as:

- Order the variables so as to impose contemporaneous restrictions on D that will make it lower triangular.
- Compute D as the Choleski factor of Σ.
- Compute shocks and orthogonalized response functions from 4.3.6 and 4.3.14.

In a Bayesian framework, D is considered as a random variable, as any other parameter of the VAR. Therefore, in theory, one would have to compute its posterior distribution. In practice, however, because the Choleski identification exactly identifies D (that is, there is a one-to-one relationship between Σ and D), it is possible to estimate draws from the posterior of D directly from draws from the posterior distribution of Σ. If the latter have been already obtained for a Gibbs sampling process,

computing the posterior distribution of D merely amounts to recycling existing posterior draws for Σ.

The following identification algorithm is thus proposed:

Algorithm 2.4.1 (SVAR with Choleski orderering, all priors):

1. Define the number of iterations $(It - Bu)$ of the algorithm.
2. At iteration n, draw $\Sigma_{(n)}$ from its posterior distributions. Simply recycle draw n from the Gibbs sampler.
3. Obtain $D_{(n)}$ by computing the Choleski factor of $\Sigma_{(n)}$.
4. Obtain $\tilde{\Psi}_1^{(n)}, \tilde{\Psi}_2^{(n)}...$ from $\Psi_1^{(n)} D_{(n)}, \Psi_2^{(n)} D_{(n)}$, following 4.3.14. In practice, there is no need to calculate $\Psi_1^{(n)}, \Psi_2^{(n)}$ since they have already been obtained when the impulse functions have been computed. Suffice is thus to recycle those estimates.
5. Repeat until $(It - Bu)$ iterations have been achieved. This yields a sample of independent posterior draws: $\left\{ D_{(n)}, \tilde{\Psi}_1^{(n)}, \tilde{\Psi}_2^{(n)} \right\}_{n=1}^{It-Bu}$

These draws can then be used as usual for point estimates and confidence intervals.

4.5 Identification by triangular factorisation

Assuming that $\Gamma = I$ as it is the case with Choleski factorisation may constitute an excessively restrictive hypothesis. Indeed, this assumption implies that all the structural shocks have a similar unit variance, even though the variance may actually differ from unity, and different shocks may have very different sizes. A simple solution to this problem is to use what is known as a triangular factorisation. To do this, the following assumptions are formed:

- Γ is diagonal, but not identity. This way, the zeros below the diagonal impose $n(n-1)/2$ constraints over the n^2 required to identify Γ and D with 4.3.8.
- D is lower triangular, and its main diagonal is made of ones. This assumes contemporaneous restrictions in a way that is similar to a Cholesky decomposition, and additionally imposes a unit contemporaneous response of variables to their own shocks. Then, the zeros above the main diagonal combined with the diagonal of ones generate another $n(n+1)/2$ constraints.

Combining the $n(n-1)/2$ constraints on Γ with the $n(n+1)/2$ constraints on D results in n^2 constraints, which exactly identify Γ and D from 4.3.8. In practice, identification follows from the following result from Hamilton (1994) (result 4.4.1, p 87):

Any positive definite symmetric $n \times n$ matrix Σ has a unique representation of the form:

$$\Sigma = D\Gamma D' \tag{4.5.1}$$

where D is a lower triangular matrix with ones along the principal diagonal, and Γ is a diagonal matrix.

Suffice is thus to apply the result directly to satisfy 4.3.8. Also, uniqueness of the decomposition permits to integrate it into the Gibbs sampling process. Empirical computation of D and Γ follows from the result on Choleski factors provided in Hamilton (1994), p 92. This result basically states the following: because Σ is positive definitive, it has a unique Choleski factor H such that:

$$\Sigma = HH' \tag{4.5.2}$$

But then, from 4.5.1 and 4.5.2, it is straightforward to conclude that $D\Gamma D' = HH'$ or equivalently that $D\Gamma^{1/2} = H$, which implies:

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ d_{21} & 1 & 0 & \cdots & 0 \\ d_{31} & d_{32} & 1 & & 0 \\ \vdots & \vdots & & \ddots & \\ d_{n1} & d_{n2} & d_{n3} & \cdots & 1 \end{pmatrix} \begin{pmatrix} \sqrt{g_{11}} & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{g_{22}} & 0 & \cdots & 0 \\ 0 & 0 & \sqrt{g_{33}} & & 0 \\ \vdots & \vdots & & \ddots & \\ 0 & 0 & 0 & & \sqrt{g_{nn}} \end{pmatrix} = \begin{pmatrix} \sqrt{g_{11}} & 0 & 0 & \cdots & 0 \\ d_{21}\sqrt{g_{11}} & \sqrt{g_{22}} & 0 & \cdots & 0 \\ d_{31}\sqrt{g_{11}} & d_{32}\sqrt{g_{22}} & \sqrt{g_{33}} & & 0 \\ \vdots & \vdots & & \ddots & \\ d_{n1}\sqrt{g_{11}} & d_{n2}\sqrt{g_{22}} & d_{n3}\sqrt{g_{33}} & \cdots & \sqrt{g_{nn}} \end{pmatrix} \tag{4.5.3}$$

Any numerical software can compute the Choleski factor H (the right-hand side of 4.5.3) very easily. Once H is known, 4.5.3 makes it clear that it suffices to recover D to divide each column of H by its corresponding diagonal entry: divide column one by $\sqrt{g_{11}}$, column two by $\sqrt{g_{22}}$, and so on. As for Γ, 4.5.3 identifies $\Gamma^{1/2}$. To obtain Γ, the only step to take is to square all the diagonal entries of $\Gamma^{1/2}$. Then D and Γ are identified.

It is then possible to propose the following algorithm:

Algorithm 2.4.2 (SVAR with triangular factorisation, all priors):

1. Define the number of iterations $(It - Bu)$ of the algorithm.
2. At iteration n, draw $\Sigma_{(n)}$ from its posterior distributions. Simply recycle draw n from the Gibbs sampler.
3. Obtain $D_{(n)}$ and $\Gamma_{(n)}$ from triangular factorisation of $\Sigma_{(n)}$, using 4.5.3

4. Obtain $\tilde{\Psi}_1^{(n)}, \tilde{\Psi}_2^{(n)}...$ from $\Psi_1^{(n)} D_{(n)}, \Psi_2^{(n)} D_{(n)}...$, following 4.3.14. In practice, there is no need to calculate $\Psi_1^{(n)}, \Psi_2^{(n)}...$ since they have already been obtained when the impulse functions have been computed. Suffice is thus to recycle those estimates.

5. Repeat until $(It - Bu)$ iterations have been achieved. This yields a sample of independent posterior draws:
$\left\{ D_{(n)}, \Gamma_{(n)}, \tilde{\Psi}_1^{(n)}, \tilde{\Psi}_2^{(n)}... \right\}_{n=1}^{It-Bu}$

These draws can then be used as usual for point estimates and confidence intervals.

4.6 Identification by sign, magnitude and zero restrictions

Another popular application with Bayesian VAR models is the implementation of quantitative restrictions on impulse response functions. The structural VAR models with a Choleski or triangular factorization identification scheme constitute a very simple example of such quantitative restrictions, since they permit to assume that some variables have no immediate response to certain structural shocks. While these simple settings are already quite attractive, it is possible to get more from BVAR models. The methodology developed by Arias et al. (2014) makes it possible to set not only zero restrictions, but also sign and magnitude restrictions on the impulse response functions of a BVAR model. Also, while the Choleski/triangular scheme only allow generating constraints over the contemporaneous responses (that is, period 0 of the impulse response functions), the general restriction methodology proposed by Arias et al. (2014) permits to implement restrictions at any period of the impulse response functions.

To be precise, the methodology described in this subsection develops three types of restrictions. Sign restriction represents the action of constraining the response of a variable to a specific structural shock to be positive or negative. Magnitude restriction represents the fact of constraining the response of a variable to a specific structural shock to be included into some interval of values. Finally, zero restrictions represent the action of constraining the response of a variable to a specific structural shock to take the value of zero. Each such restriction set by the user of the VAR model can be defined over any number of impulse response function periods, which need not be the same across restrictions.

The analysis starts from the conventional SVAR model 4.3.2:

$$D_0 y_t = D_1 y_{t-1} + D_2 y_{t-2} + ... + D_p y_{t-p} + F x_t + \eta_t, \quad t = 1, 2, ..., T \tag{4.6.1}$$

For simplicity, assume that $\eta_t \sim \mathcal{N}(0, I)$, that is, the structural shocks are mutually orthogonal and have unit variance. The aim consists in finding a structural matrix $D = D_0^{-1}$ such that the structural impulse response functions $\tilde{\Psi}_0, \tilde{\Psi}_1, \tilde{\Psi}_2...$ produced from model 4.6.1 satisfy the restrictions specified by the user. To verify that the restrictions hold, it is convenient to stack the structural impulse response function matrices of all periods subject to a restriction into a single matrix denoted by $f(D, D_1, ..., D_p)$. For instance, if restrictions are implemented for periods $p_1, p_2, ..., p_n$, then $f(D, D_1, ..., D_p)$ is:

$$f(D, D_1, ..., D_p) = \begin{pmatrix} \tilde{\Psi}_{p_1} \\ \tilde{\Psi}_{p_2} \\ \vdots \\ \tilde{\Psi}_{p_n} \end{pmatrix} \tag{4.6.2}$$

Verification of the restrictions can then be realized by the way of selection matrices. For example, with sign restrictions, the matrix for sign restrictions with respect to structural shock j, for $j = 1, 2, ..., n$, will be the matrix S_j with a number of columns equal to the number of rows of $f(D, D_1, ..., D_p)$, and a number of rows equal to the number of sign restrictions on shock j. Each row of S_j represents one restriction and is made only of zeros, save for the entry representing the restriction which is a one (for a positive sign restriction), or a minus one (for negative sign restrictions). Then, the restrictions on shock j hold if:

$$S_j \times f_j(D, D_1, ..., D_p) > 0 \tag{4.6.3}$$

where $f_j(D, D_1, ..., D_p)$ represents column j of the matrix $f(D, D_1, ..., D_p)$. The sign restrictions hold if 4.6.3 holds for all shocks $j = 1, 2, ..., n$.

Magnitude restrictions on shock j can be represented by the way of the selection matrix M_j, and two vectors $M_{l,j}$ and $M_{u,j}$. M_j is a selection matrix with a number of columns equal to the number of rows of $f(D, D_1, ..., D_p)$, a number of rows equal to the number of magnitude restrictions on shock j, and selection entries equal to one. $M_{l,j}$ and $M_{u,j}$ are vectors with as many rows as M_j. Each row entry of $M_{l,j}$ is the lower bound of the corresponding restriction interval, while each row entry of $M_{u,j}$ contains the upper bound of the restriction interval. Then, the magnitude restrictions on structural shock j hold if :

$$(M_j \times f_j(D, D_1, ..., D_p) - M_{l,j}) . \times (M_{u,j} - M_j \times f_j(D, D_1, ..., D_p)) > 0 \tag{4.6.4}$$

where the operator $\times$ denotes element wise product rather than a standard matrix product. The idea is the following: if the magnitude restrictions are satisfied, then the considered impulse

response function values $M_j \times f_j(D, D_1, ..., D_p)$ are comprised between $M_{l,j}$ and $M_{u,j}$. Therefore, $M_j \times f_j(D, D_1, ..., D_p) - M_{l,j}$ should be a vector of positive values, and $M_{u,j} - M_j \times f_j(D, D_1, ..., D_p)$ should also be a vector of positive values. Element wise multiplication should thus result in a strictly positive vector. On the other hand, if any one restrictions is not satisfied, then either the corresponding row of $M_j \times f_j(D, D_1, ..., D_p) - M_{l,j}$ will be negative (if the impulse response function value is smaller than the lower bound) while the corresponding row of $M_{u,j} - M_j \times f_j(D, D_1, ..., D_p)$ will be positive, or the row of $M_j \times f_j(D, D_1, ..., D_p) - M_{l,j}$ will be negative (if the impulse response function value is larger than the upper bound) while the corresponding row of $M_j \times f_j(D, D_1, ..., D_p) - M_{l,j}$ will be positive. In both cases, the element wise product will be negative, indicating failure of the restriction. The magnitude restrictions hold if 4.6.4 holds for all shocks $j = 1, 2, ..., n$.

Eventually, zero restrictions with respect to structural shock j can be checked using a selection matrix Z_j, with a number of columns equal to the number of rows of $f(D, D_1, ..., D_p)$, a number of rows equal to the number of zero restrictions on shock j, and zero entries, except for the entries relative to the restrictions which take a value of one. Then, the zero restrictions on structural shock j hold if:

$$S_j \times f_j(D, D_1, ..., D_p) = 0 \tag{4.6.5}$$

The zero restrictions hold if 4.6.5 holds for all shocks $j = 1, 2, ..., n$. In the line of the Gibbs sampler methodology systematically used so far, a natural strategy would be to adopt the following algorithm:

Algorithm 2.6.1 (Gibbs sampling for sign, magnitude and zero restrictions):

1. Draw the SVAR coefficients $D_0, D_1, D_2, ..., D_p$ and F from the unrestricted posterior distribution.

2. Compute the structural impulse response functions $\tilde{\Psi}_0, \tilde{\Psi}_1, \tilde{\Psi}_2...$ from the coefficients.

3. Check if the restrictions are satisfied, using 4.6.3, 4.6.4 and 4.6.5. If yes, keep the draw, if not, discard the draw.

4. Repeat steps 1-3 until the desired number of iterations satisfying the restrictions is obtained.

Of course, the difficulty with algorithm 2.6.1 is that the traditional Gibbs sampler framework does not allow to draw directly from the posterior distribution of the SVAR coefficients $D_0, D_1, D_2, ..., D_p$ and F. Rather, the Gibbs sampling procedure produces draws from the posterior distribution of the reduced form VAR model 3.1.2:

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + ... + A_p y_{t-p} + C x_t + \varepsilon_t \tag{4.6.6}$$

with $\varepsilon_t \sim \mathcal{N}(0, \Sigma)$. That is, the conventional Gibbs sampler methodology only provides draws for $A_1, A_2, ..., A_p, C$ and Σ.

Intuitively, it would be tempting to apply the following strategy: draw the coefficients $A_1, A_2, \ldots, A_p, C$ and Σ of a reduced-form VAR model, from the posterior distribution. From these coefficients, obtain the impulse response functions $\Psi_0, \Psi_1, \Psi_2...$ Then apply a standard identification scheme such as a Choleski factorization of the residual covariance matrix Σ to recover the structural matrix D. With D at hand, finally, identify the SVAR impulse response functions $\tilde{\Psi}_0, \tilde{\Psi}_1, \tilde{\Psi}_2...$ Check if the restrictions are satisfied, as described in algorithm 2.6.1, and preserve the draw only if the requirement is met. The problem is that such an identification would not produce a posterior draw from the correct distribution for the SVAR coefficients $D_0, D_1, D_2, ..., D_p$ and F. Arias et al. (2014) show however (theorem 2) that it is possible to use draws from the posterior distributions of the reduced form model to obtain draws from the correct posterior distribution of the SVAR model. The only requirement is the implementation of an additional orthogonalisation step.

Because the procedure differs depending on the fact the zero restrictions are included or not in the setting, the two methodologies are developed in turn, starting with the simplest case of sign and magnitude restrictions only.

Identification with pure sign and magnitude restrictions: In the case of pure sign and magnitude restrictions, the procedure remains simple. First, draw a vector β (that is, a set of reduced form VAR coefficients $A_1, A_2, ..., A_p, C$), and a residual covariance matrix Σ from their posterior distributions. From this, recover the reduced form VAR model 3.1.2:

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + ... + A_p y_{t-p} + C x_t + \varepsilon_t \tag{4.6.7}$$

From this reduced VAR model, one can obtain the impulse response functions $\Psi_0, \Psi_1, \Psi_2...$Then, define a preliminary structural matrix $h(\Sigma)$, where $h()$ is any continuously differentiable function of symmetric positive definite matrices such that $h(\Sigma) \times h(\Sigma)' = \Sigma$. In practice, the usual Choleski factor is used for $h()$. From this preliminary structural matrix, obtain a first set of structural impulse response functions $\bar{\Psi}_0, \bar{\Psi}_1, \bar{\Psi}_2...$ from 4.3.14:

$$\bar{\Psi}_i = \Psi_i h(\Sigma) \tag{4.6.8}$$

These preliminary impulse response functions, however, are not drawn from the correct distri-

bution. To draw from the correct posterior distribution, an additional orthogonalisation step is required. To do so, one has to draw a random matrix Q from a uniform distribution, and define:

$$D = h(\Sigma)Q \tag{4.6.9}$$

The strategy is then to draw such a Q matrix which would be orthogonal, in order to preserve the SVAR property 4.3.8:

$$D\Gamma D' = DID' = DD' = h(\Sigma)QQ'h(\Sigma)' = h(\Sigma)Ih(\Sigma)' = h(\Sigma)h(\Sigma)' = \Sigma \tag{4.6.10}$$

To obtain an orthogonal matrix Q from the uniform distribution, the procedure is the following. First, draw a $n \times n$ random matrix X, for which each entry is drawn from an independent standard normal distribution. Then, use a QR decomposition of X, such that $X = QR$, with Q an orthogonal matrix and R an upper triangular matrix. It is then possible to obtain the definitive structural impulse response functions from 4.2.12 as:

$$\tilde{\Psi}_i = \Psi_i D = \Psi_i h(\Sigma)Q = \bar{\Psi}_i Q \tag{4.6.11}$$

With this, the stacked structural matrix 4.6.2 can write as:

$$f(D, D_1, ..., D_p) = \begin{pmatrix} \tilde{\Psi}_{p_1} \\ \tilde{\Psi}_{p_2} \\ \vdots \\ \tilde{\Psi}_{p_n} \end{pmatrix} = \begin{pmatrix} \bar{\Psi}_{p_1} \\ \bar{\Psi}_{p_2} \\ \vdots \\ \bar{\Psi}_{p_n} \end{pmatrix} Q = \bar{f}(D, D_1, ..., D_p) \times Q \tag{4.6.12}$$

with

$$\bar{f}(D, D_1, ..., D_p) = \begin{pmatrix} \bar{\Psi}_{p_1} \\ \bar{\Psi}_{p_2} \\ \vdots \\ \bar{\Psi}_{p_n} \end{pmatrix} \tag{4.6.13}$$

If the restrictions are respected, then 4.6.3 and 4.6.4 hold for all structural shocks $j = 1, 2, \ldots, n$. If it is not the case, then restart the whole process all over again. It is thus possible to propose the following Gibbs algorithm:

Algorithm 2.6.2 (Gibbs sampling for sign and magnitude restrictions):

1. Define the restriction matrices $S_j, M_j, M_{l,j}$ and $M_{u,j}$, for $j = 1, 2, ..., n$.

2. Define the number of successful iterations $It - Bu$ of the algorithm.

3. At iteration n, draw the reduced-form VAR coefficients $B_{(n)}$and $\Sigma_{(n)}$ from their posterior distributions, and recover model 4.6.6.

4. At iteration n, obtain $\Psi_0^{(n)}, \Psi_1^{(n)}, \Psi_2^{(n)}$... from $B_{(n)}$.

5. At iteration n, calculate $h(\Sigma_{(n)})$, and generate $\bar{\Psi}_0^{(n)}, \bar{\Psi}_1^{(n)}, \bar{\Psi}_2^{(n)}$... from 4.6.8. Create the preliminary stacked matrix $\bar{f}(D, D_1, ..., D_p) = \begin{pmatrix} \bar{\Psi}_{p_1}^{(n)} \\ \bar{\Psi}_{p_2}^{(n)} \\ \vdots \\ \bar{\Psi}_{p_n}^{(n)} \end{pmatrix}$ from 4.6.14.

6. At iteration n, draw a random matrix X, for which each entry is drawn from an independent standard normal distribution. Then, use a QR decomposition of X to obtain the structural matrix Q.

7. At iteration n, compute the candidate stacked structural impulse response function matrix $f(D, D_1, ..., D_p)$ from 4.6.13.

8. At iteration n, verify that the restrictions hold from 4.6.3 and 4.6.4. If yes, keep the matrix Q and go for the next iteration. If not, repeat steps 3 to 8 until a valid matrix Q is obtained. Then go for the next iterations.

9. Repeat steps 3-8 until $It - Bu$ successful iterations are obtained.

Identification with sign, magnitude, and zero restrictions: When zero restrictions are included into the setting, identification becomes more complicated. The reason is that if one tries to apply naively algorithm 2.6.2 to identify a matrix Q that would satisfy the zero restrictions, the algorithm would always fail since the set of matrices satisfying the zero restrictions has measure zero. In other words, the probability to draw by chance a Q matrix that would satisfy the zero restriction is null. The procedure has then to be adapted to first force the Q draw to produce an orthogonal Q matrix which will satisfy for sure the zero restrictions, and then run the normal checking process 4.6.3 and 4.6.4 of Q for the remaining sign and magnitude restrictions.

The algorithm to obtain a matrix Q satisfying the sign restrictions goes as follows:

Algorithm 2.6.3 (obtention of a matrix Q satisfying the zero restrictions):

1. Set $j = 1$ (that is, consider restrictions on structural shock 1).

2. Create $R_j = \begin{pmatrix} Z_j \bar{f}(D, D_1, \ldots, D_p) \\ Q'_{j-1} \end{pmatrix}$

   where $Q_{j-1} = [q1 \; q_2 \; \ldots \; q_{j-1}]$, with q_j the j^{th} column of matrix Q. Of course, when $j = 1$, no column of Q has yet been identified, so that $Q_0 = \phi$. If $Z_j = \phi$ (no zero restriction on structural shock j), set $Z_j \bar{f}(D, D_1, \ldots, D_p) = \phi$.

3. Find a matrix N_{j-1} whose columns form a non-zero orthonormal basis for the nullspace of R_j. Do it only if $R_j \neq \phi$. If $R_j = \phi$, there are no restrictions on shock j, nor any already existing element of Q. In this case, do not do anything.

4. Draw a random vector x_j from a standard normal distribution on $\mathbb{R}^n$.

5. If $j = 2$, define $q_j = N'_{j-1} \left(N_{j-1} x_j / ||N'_{j-1} x_j|| \right)$. If $R_j = \phi$, simply define $q_j = x_j / ||x_j||$.

6. Set $j = 2$ (that is, consider now restrictions on structural shock 2), and repeat steps 2 to 5.

7. Repeat for $j = 3, 4, \ldots, n$. This produces an orthonormal Q matrix that satisfies the zero restrictions.

Note: When trying to implement sign restrictions, one has to be careful with the number of restriction implemented on each structural shock, as not every zero restriction setting will be well-identified. The reason is the following: for $j = 1, 2, \ldots, n$, the solution q_j to the equation:

$$R_j q_j \tag{4.6.14}$$

will be a non-zero solution if and only if $z_j \leq n - j$, where z_j is the total number of zero restrictions on structural shock j. That is, a well identified orthonormal matrix can only be obtained if for each structural shock j, the number of zero restrictions on that shock is at most equal to $n - j$, with n denoting, as usual, the total number of variables in the model. If for any shock, $z_j > n - j$, the basis N_{j-1} of the nullspace of R_j will be only the trivial basis (a zero vector), so that $q_j = 0$, and Q cannot be an orthonormal matrix.

However, this issue does not make the exercise as restrictive as it may seem, since one can play on the order of the variables in the VAR model to achieve a certain setting. Indeed, unlike a conventional a Choleski/triangular factorization scheme where the ordering of the variable does matter, in a restriction setting, the variable ordering is perfectly arbitrary. Therefore assume for instance A VAR model with 4 variables, and two zero restrictions on shock 4. Then $z_4 = 2 > 0 = 4 - 4 = n - j$. So the zero restrictions are not identified in this case. However, a simple solution consists in simply inverting the order of variables 2 and 4 in the VAR model. Then, the two restrictions become restrictions on shock 2, so that $z_2 = 2 > 4 - 2 = n - j$, and the identification scheme will work.

It then becomes possible to adapt algorithm 2.6.2 for zero restrictions:

Algorithm 2.6.4 (Gibbs sampling for general zero, sign and magnitude restrictions):

1. Define the restriction matrices $S_j, M_j, M_{l,j}, M_{u,j}$ and Z_j , for $j = 1, 2, ..., n$.
2. Define the number of successful iterations $It - Bu$ of the algorithm.
3. At iteration n, draw the reduced-form VAR coefficients $B_{(n)}$and $\Sigma_{(n)}$ from their posterior distributions, and recover model 4.6.7.
4. At iteration n, obtain $\Psi_0^{(n)}, \Psi_1^{(n)}, \Psi_2^{(n)}$... from $B_{(n)}$.
5. At iteration n, calculate $h(\Sigma_{(n)})$, and generate $\bar{\Psi}_0^{(n)}, \bar{\Psi}_1^{(n)}, \bar{\Psi}_2^{(n)}$... from 4.6.8. Create the preliminary stacked matrix:

$$\begin{pmatrix} \bar{\Psi}_{p_1}^{(n)} \\ \bar{\Psi}_{p_2}^{(n)} \\ \vdots \\ \bar{\Psi}_{p_n}^{(n)} \end{pmatrix} \tag{4.6.15}$$

6. At iteration n, apply algorithm 2.6.3 to obtain an orthonormal Q matrix which satisfies the zero restrictions.
7. At iteration n, compute the candidate stacked structural impulse response function matrix $f(D, D_1, ..., D_p)$ from 4.6.13.
8. At iteration n, verify that the sign and magnitude restrictions hold from 4.6.3and 4.6.4. If yes, keep the matrix Q and go for the next iteration. If not, repeat steps 3 to 8 until a valid matrix Q is obtained. Then go for the next iterations.
9. Repeat steps 3 to 8 until $It - Bu$ successful iterations are obtained.

An example: Consider a 3-variable VAR model with two lags, and assume for simplicity[5] that there are no exogenous variables included in the model:

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \\ y_{3,t} \end{pmatrix} = \begin{pmatrix} a_{11}^1 & a_{12}^1 & a_{13}^1 \\ a_{21}^1 & a_{22}^1 & a_{23}^1 \\ a_{31}^1 & a_{32}^1 & a_{33}^1 \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \\ y_{3,t-1} \end{pmatrix} + \begin{pmatrix} a_{11}^2 & a_{12}^2 & a_{13}^2 \\ a_{21}^2 & a_{22}^2 & a_{23}^2 \\ a_{31}^2 & a_{32}^2 & a_{33}^2 \end{pmatrix} \begin{pmatrix} y_{1,t-2} \\ y_{2,t-2} \\ y_{3,t-2} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{pmatrix} \tag{4.6.16}$$

[5]Note that exogenous variables are irrelevant for the computation of impulses response functions, and thus for implementing restrictions on them. Therefore, they can be omitted without loss of generality.

The conventional assumption on residuals applies:

$$E\left(\varepsilon_t \varepsilon_t^{'}\right) = \Sigma \tag{4.6.17}$$

, or

$$E\begin{pmatrix}\varepsilon_{1,t}\\ \varepsilon_{2,t}\\ \varepsilon_{3,t}\end{pmatrix}\begin{pmatrix}\varepsilon_{1,t} & \varepsilon_{2,t} & \varepsilon_{3,t}\end{pmatrix} = \begin{pmatrix}\sigma_{11} & \sigma_{12} & \sigma_{13}\\ \sigma_{12} & \sigma_{22} & \sigma_{23}\\ \sigma_{13} & \sigma_{23} & \sigma_{33}\end{pmatrix} \tag{4.6.18}$$

The following restrictions are implemented:

- Sign restrictions: a positive sign restriction on the response of variable 1 to structural shock 2, to be implemented at period 2. And a negative sign restriction on the response of variable 1 to structural shock 3, to be implemented at period 3.
- A magnitude restriction of $\begin{bmatrix}-0.5 & 0.5\end{bmatrix}$ on the response of variable 2 to structural shock 2, and a magnitude restriction of $\begin{bmatrix}-0.3 & 0.7\end{bmatrix}$ on the response of variable 3 to structural shock 2, both to be implemented at period 1.
- Zero restrictions: a zero restriction to be implemented on the response of variable 1 and 3 to structural shock 1, to be implemented at period 0. Note that even though the general methodology makes it possible to develop zero restrictions at any period, zero restrictions typically make sense only at impact (period 0).

Case of pure sign and magnitude restrictions (algorithm 2.6.2):

1. Because there are restrictions over three periods (periods 1, 2 and 3), the stacked matrix for impulse response functions will be:

$$f(D, D_1, ..., D_p) = \begin{pmatrix}\tilde{\Psi}_1\\ \tilde{\Psi}_2\\ \tilde{\Psi}_3\end{pmatrix} \tag{4.6.19}$$

Because the model comprises three variables, each individual impulse response function matrix $\tilde{\Psi}$ will be of size 3×3. With three of them stacked, the matrix $f(D, D_1, ..., D_p)$ will comprise a total of $3 \times 3 = 9$ rows. Therefore, each restriction matrix $S_j$and M_j will comprise 9 columns. The restriction matrices are thus as follows:

- A positive sign restriction on the response of variable 1 to structural shock 2, to be implemented at period 2:

$$S_2 = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{4.6.20}$$

- A negative sign restriction on the response of variable 1 to structural shock 3, to be implemented at period 3:

$$S_3 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \end{pmatrix} \tag{4.6.21}$$

- A magnitude restriction of $\begin{bmatrix} -0.5 & 0.5 \end{bmatrix}$ on the response of variable 2 to structural shock 2, and a magnitude restriction of $\begin{bmatrix} -0.3 & 0.7 \end{bmatrix}$ on the response of variable 3 to structural shock 2, both to be implemented at period 1:

$$M_2 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{4.6.22}$$

,

$$M_{l,2} = \begin{pmatrix} -0.5 \\ -0.3 \end{pmatrix} \tag{4.6.23}$$

,

$$M_{u,2} = \begin{pmatrix} 0.5 \\ 0.7 \end{pmatrix} \tag{4.6.24}$$

2. Assume that the following draw is realized for the values of the VAR coefficients:

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \\ y_{3,t} \end{pmatrix} = \begin{pmatrix} 0.78 & 0.12 & -0.08 \\ 0.05 & 0.82 & 0.03 \\ -0.16 & 0.21 & 0.85 \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \\ y_{3,t-1} \end{pmatrix} + \begin{pmatrix} 0.02 & -0.04 & 0.07 \\ 0.11 & 0.06 & -0.16 \\ -0.03 & 0.08 & -0.09 \end{pmatrix} \begin{pmatrix} y_{1,t-2} \\ y_{2,t-2} \\ y_{3,t-2} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{pmatrix} \tag{4.6.25}$$

It can be shown that such a draw produces a stationary model. Assume that the draw for Σ is given by:

$$\Sigma = \begin{pmatrix} 0.028 & -0.013 & 0.034 \\ -0.013 & 0.042 & 0.006 \\ 0.034 & 0.006 & 0.125 \end{pmatrix} \tag{4.6.26}$$

3. Obtain Ψ_1, Ψ_2 and Ψ_3 as:

$$\Psi_1 = \begin{pmatrix} 0.78 & 0.12 & -0.08 \\ 0.05 & 0.82 & 0.03 \\ -0.16 & 0.21 & 0.85 \end{pmatrix} \tag{4.6.27}$$

$$\Psi_2 = \begin{pmatrix} 0.65 & 0.14 & -0.06 \\ 0.19 & 0.74 & -0.11 \\ -0.28 & 0.41 & 0.65 \end{pmatrix} \tag{4.6.28}$$

$$\Psi_3 = \begin{pmatrix} 0.55 & 0.15 & -0.05 \\ 0.29 & 0.66 & -0.22 \\ -0.31 & 0.53 & 0.47 \end{pmatrix} \tag{4.6.29}$$

4. Obtain $h(\Sigma)$, the lower Choleski factor of Σ, as:

$$h(\Sigma) = \begin{pmatrix} 0.1673 & 0 & 0 \\ -0.0777 & 0.1896 & 0 \\ 0.2032 & 0.1149 & 0.2656 \end{pmatrix} \tag{4.6.30}$$

Obtain the preliminary stacked matrix:

$$\begin{pmatrix} \bar{\Psi}_1 \\ \bar{\Psi}_2 \\ \bar{\Psi}_3 \end{pmatrix} = \begin{pmatrix} \Psi_1 \\ \Psi_2 \\ \Psi_3 \end{pmatrix} h(\Sigma) = \begin{pmatrix} 0.105 & 0.013 & -0.021 \\ -0.049 & 0.159 & 0.008 \\ 0.129 & 0.137 & 0.226 \\ 0.086 & 0.019 & -0.015 \\ -0.050 & 0.128 & -0.030 \\ 0.053 & 0.153 & 0.173 \\ 0.070 & 0.021 & -0.014 \\ -0.047 & 0.099 & -0.058 \\ 0.002 & 0.154 & 0.124 \end{pmatrix} \tag{4.6.31}$$

5. assume the random matrix X that was drawn is the following:

$$X = \begin{pmatrix} -0.3034 & 0.8884 & -0.8095 \\ 0.2939 & -1.1471 & -2.9443 \\ -0.7873 & 1.0689 & 1.4384 \end{pmatrix} \tag{4.6.32}$$

From the QR decomposition of X, the orthonormal matrix obtained from X is:

$$Q = \begin{pmatrix} -0.3396 & 0.5484 & 0.7642 \\ 0.3289 & -0.6919 & -0.6427 \\ -0.8812 & -0.4696 & 0.0546 \end{pmatrix} \tag{4.6.33}$$

6. compute the candidate stacked structural impulse response function matrix:

$$f(D, D_1, ..., D_p) = \begin{pmatrix} \tilde{\Psi}_{p_1} \\ \tilde{\Psi}_{p_2} \\ \vdots \\ \tilde{\Psi}_{p_n} \end{pmatrix} = \begin{pmatrix} \bar{\Psi}_{p_1} \\ \bar{\Psi}_{p_2} \\ \vdots \\ \bar{\Psi}_{p_n} \end{pmatrix} Q = \begin{pmatrix} -0.0125 & 0.0581 & -0.0901 \\ 0.0620 & -0.1407 & -0.0641 \\ -0.1977 & -0.1300 & -0.1751 \\ 0.0097 & 0.0412 & -0.0790 \\ 0.0858 & -0.1019 & -0.0458 \\ -0.1204 & -0.1577 & 0.1297 \\ -0.0042 & 0.0302 & -0.0682 \\ 0.1002 & -0.0675 & -0.0311 \\ -0.0596 & -0.1633 & -0.0939 \end{pmatrix} \tag{4.6.34}$$

7. verify that the restrictions hold from 4.6.3 and 4.6.4 :

   - Positive sign restriction on the response of variable 1 to structural shock 2, to be implemented at period 2:

$$S_2 \times f_2(D, D_1, ..., D_p) = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0.0581 \\ -0.1407 \\ -0.1300 \\ 0.0412 \\ -0.1019 \\ -0.1577 \\ 0.0302 \\ -0.0675 \\ -0.1633 \end{pmatrix} = (0.0412) > 0 \tag{4.6.35}$$

   - A negative sign restriction on the response of variable 1 to structural shock 3, to be implemented at period 3:

$$S_3 \times f_3(D, D_1, ..., D_p) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \end{pmatrix} \begin{pmatrix} -0.0901 \\ -0.0641 \\ -0.1751 \\ -0.0790 \\ -0.0458 \\ 0.1297 \\ -0.0682 \\ -0.0311 \\ -0.0939 \end{pmatrix} = (0.0682) > 0 \tag{4.6.36}$$

- A magnitude restriction of $\begin{bmatrix} -0.5 & 0.5 \end{bmatrix}$on the response of variable 2 to structural shock 2, and a magnitude restriction of $\begin{bmatrix} -0.3 & 0.7 \end{bmatrix}$ on the response of variable 3 to structural shock 2, both to be implemented at period 1:

$$
\begin{aligned}
&(M_2 \times f_2(D, D_1, ..., D_p) - M_{l,2}) \times (M_{u,2} - M_2 \times f_2(D, D_1, ..., D_p)) \\
&= \left(\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0.0581 \\ -0.1407 \\ -0.1300 \\ 0.0412 \\ -0.1019 \\ -0.1577 \\ 0.0302 \\ -0.0675 \\ -0.1633 \end{pmatrix} - \begin{pmatrix} -0.5 \\ -0.3 \end{pmatrix} \right) \times \\
&\left(\begin{pmatrix} 0.5 \\ 0.7 \end{pmatrix} - \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0.0581 \\ -0.1407 \\ -0.1300 \\ 0.0412 \\ -0.1019 \\ -0.1577 \\ 0.0302 \\ -0.0675 \\ -0.1633 \end{pmatrix} \right) \\
&= \left(\begin{pmatrix} -0.1407 \\ -0.1300 \end{pmatrix} - \begin{pmatrix} -0.5 \\ -0.3 \end{pmatrix} \right) \times \left(\begin{pmatrix} 0.5 \\ 0.7 \end{pmatrix} - \begin{pmatrix} -0.1407 \\ -0.1300 \end{pmatrix} \right) \\
&= \begin{pmatrix} 0.3593 \\ 0.1700 \end{pmatrix} \times \begin{pmatrix} 0.6407 \\ 0.8300 \end{pmatrix} \\
&= \begin{pmatrix} 0.2302 \\ 0.1411 \end{pmatrix} > 0
\end{aligned} \tag{4.6.37}
$$

Hence all the restrictions are satisfied. Keep this matrix Q since it is a valid candidate.

Case of general restrictions (algorithm 2.6.4):

1. There are now restrictions over four periods (periods 0, 1, 2 and 3, see below), so that the stacked matrix for impulse response functions will be:

$$f(D, D_1, ..., D_p) = \begin{pmatrix} \tilde{\Psi}_0 \\ \tilde{\Psi}_1 \\ \tilde{\Psi}_2 \\ \tilde{\Psi}_3 \end{pmatrix} \tag{4.6.38}$$

Because the model comprises three variables, each individual impulse response function matrix $\tilde{\Psi}$ will be of size 3×3. With four of them stacked, the matrix $f(D, D_1, ..., D_p)$ will comprise a total of $3 \times 4 = 12$ rows. Therefore, each restriction matrix S_j and M_j will comprise 12 columns.

The restriction matrices are thus as follows:

- A positive sign restriction on the response of variable 1 to structural shock 2, to be implemented at period 2:

$$S_2 = (0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 1 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0) \tag{4.6.39}$$

- A negative sign restriction on the response of variable 1 to structural shock 3, to be implemented at period 3:

$$S_3 = (0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; -1 \;\; 0 \;\; 0) \tag{4.6.40}$$

- A magnitude restriction of $\begin{bmatrix} -0.5 & 0.5 \end{bmatrix}$ on the response of variable 2 to structural shock 2, and a magnitude restriction of

$$\begin{bmatrix} -0.3 & 0.7 \end{bmatrix} \tag{4.6.41}$$

on the response of variable 3 to structural shock 2, both to be implemented at period 1:

$$M_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \; M_{l,2} = \begin{pmatrix} -0.5 \\ -0.3 \end{pmatrix}, \; M_{u,2} = \begin{pmatrix} 0.5 \\ 0.7 \end{pmatrix} \tag{4.6.42}$$

- A zero restriction to be implemented on the response of variable 1 and 3 to structural shock 1, to be implemented at period 0:

$$Z_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{4.6.43}$$

2. Assume that the following draws are realized for the values of the VAR coefficients and residual covariance matrix Σ (same as the pure sign and magnitude restriction case):

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \\ y_{3,t} \end{pmatrix} = \begin{pmatrix} 0.78 & 0.12 & -0.08 \\ 0.05 & 0.82 & 0.03 \\ -0.16 & 0.21 & 0.85 \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \\ y_{3,t-1} \end{pmatrix} + \begin{pmatrix} 0.02 & -0.04 & 0.07 \\ 0.11 & 0.06 & -0.16 \\ -0.03 & 0.08 & -0.09 \end{pmatrix} \begin{pmatrix} y_{1,t-2} \\ y_{2,t-2} \\ y_{3,t-2} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{pmatrix} \tag{4.6.44}$$

$$\Sigma = \begin{pmatrix} 0.028 & -0.013 & 0.034 \\ -0.013 & 0.042 & 0.006 \\ 0.034 & 0.006 & 0.125 \end{pmatrix} \tag{4.6.45}$$

3. Obtain Ψ_0, Ψ_1, Ψ_2 and Ψ_3 as:

$$\Psi_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{4.6.46}$$

$$\Psi_1 = \begin{pmatrix} 0.78 & 0.12 & -0.08 \\ 0.05 & 0.82 & 0.03 \\ -0.16 & 0.21 & 0.85 \end{pmatrix} \tag{4.6.47}$$

$$\Psi_2 = \begin{pmatrix} 0.65 & 0.14 & -0.06 \\ 0.19 & 0.74 & -0.11 \\ -0.28 & 0.41 & 0.65 \end{pmatrix} \tag{4.6.48}$$

$$\Psi_3 = \begin{pmatrix} 0.55 & 0.15 & -0.05 \\ 0.29 & 0.66 & -0.22 \\ -0.31 & 0.53 & 0.47 \end{pmatrix} \tag{4.6.49}$$

4. Obtain $h(\Sigma)$, the lower Choleski factor of Σ, as:

$$h(\Sigma) = \begin{pmatrix} 0.1673 & 0 & 0 \\ -0.0777 & 0.1896 & 0 \\ 0.2032 & 0.1149 & 0.2656 \end{pmatrix} \tag{4.6.50}$$

Obtain the preliminary stacked matrix:

$$\bar{f}(D, D_1, ..., D_p) = \begin{pmatrix} \bar{\Psi}_0 \\ \bar{\Psi}_1 \\ \bar{\Psi}_2 \\ \bar{\Psi}_3 \end{pmatrix} = \begin{pmatrix} \Psi_0 \\ \Psi_1 \\ \Psi_2 \\ \Psi_3 \end{pmatrix} h(\Sigma) = \begin{pmatrix} 0.1673 & 0 & 0 \\ -0.0777 & 0.1896 & 0 \\ 0.2032 & 0.1149 & 0.2656 \\ 0.105 & 0.013 & -0.021 \\ -0.049 & 0.159 & 0.008 \\ 0.129 & 0.137 & 0.226 \\ 0.086 & 0.019 & -0.015 \\ -0.050 & 0.128 & -0.030 \\ 0.053 & 0.153 & 0.173 \\ 0.070 & 0.021 & -0.014 \\ -0.047 & 0.099 & -0.058 \\ 0.002 & 0.154 & 0.124 \end{pmatrix} \tag{4.6.51}$$

5. Apply algorithm 2.6.3 to obtain an orthonormal Q matrix which satisfies the zero restrictions:

- Set $j = 1$
- Create

$$R_1 = \begin{bmatrix} Z_1 \bar{f}(D, D_1, ..., D_p) \\ Q_0^{'} \end{bmatrix} = \begin{bmatrix} Z_1 \bar{f}(D, D_1, ..., D_p) \\ \phi \end{bmatrix} = \begin{pmatrix} 0.1673 & 0 & 0 \\ 0.2032 & 0.1149 & 0.2656 \end{pmatrix} \tag{4.6.52}$$

- Obtain

$$N_0 = \begin{pmatrix} 0 \\ -0.9178 \\ 0.3970 \end{pmatrix} \tag{4.6.53}$$

- Draw

$$x_1 = \begin{pmatrix} 0.3252 \\ -0.7549 \\ 1.3703 \end{pmatrix} \tag{4.6.54}$$

- Obtain

$$q_1 = N_0 \left(N_0^{'} x_1 / \left\| N_0^{'} x_1 \right\| \right) = \begin{pmatrix} 0 \\ -0.9178 \\ 0.3970 \end{pmatrix} \tag{4.6.55}$$

- set $j = 2$

- Create

$$R_2 = \begin{bmatrix} Z_2\bar{f}(D, D_1, ..., D_p) \\ Q_1^{'} \end{bmatrix} = \begin{bmatrix} \phi \\ Q_1^{'} \end{bmatrix} = \begin{pmatrix} 0 & -0.9178 & 0.3970 \end{pmatrix} \tag{4.6.56}$$

- Obtain

$$N_1 = \begin{pmatrix} -0.9178 & 0.3970 \\ 0.1576 & 0.3644 \\ 0.3644 & 0.8424 \end{pmatrix} \tag{4.6.57}$$

- Draw

$$x_2 = \begin{pmatrix} 0.3192 \\ 0.3129 \\ -0.8649 \end{pmatrix} \tag{4.6.58}$$

- Obtain

$$q_2 = N_1 \left(N_1^{'} x_2 / \left\| N_1^{'} x_2 \right\| \right) = \begin{pmatrix} 0.4303 \\ -0.3584 \\ -0.8285 \end{pmatrix} \tag{4.6.59}$$

- Set $j = 3$
- Create

$$R_3 = \begin{bmatrix} Z_3\bar{f}(D, D_1, ..., D_p) \\ Q_2^{'} \end{bmatrix} = \begin{bmatrix} \phi \\ Q_2^{'} \end{bmatrix} = \begin{pmatrix} 0 & -0.9178 & 0.3970 \\ 0.4303 & -0.3584 & 0.8285 \end{pmatrix} \tag{4.6.60}$$

- Obtain

$$N_2 = \begin{pmatrix} 0.9027 \\ 0.1709 \\ 0.3950 \end{pmatrix} \tag{4.6.61}$$

- Draw

$$x_2 = \begin{pmatrix} -0.0301 \\ -0.1649 \\ 0.6277 \end{pmatrix} \tag{4.6.62}$$

- Obtain

$$q_3 = N_2 \left(N_2^{'} x_3 / \left\| N_2^{'} x_3 \right\| \right) = \begin{pmatrix} 0.9027 \\ 0.1709 \\ 0.3950 \end{pmatrix} \tag{4.6.63}$$

Obtain:

$$Q = \begin{pmatrix} 0 & 0.4303 & 0.9027 \\ -0.9178 & -0.3584 & 0.1709 \\ 0.3970 & -0.8285 & 0.3950 \end{pmatrix} \tag{4.6.64}$$

6. Compute the candidate stacked structural impulse response function matrix $f(D, D_1, ..., D_p)$:

$$f(D, D_1, ..., D_p) = \bar{f}(D, D_1, ..., D_p) \times Q = \begin{pmatrix} 0 & 0.0720 & 0.1511 \\ -0.1741 & -0.1014 & -0.0377 \\ 0 & -0.1737 & 0.3079 \\ -0.0209 & 0.0579 & 0.0887 \\ -0.1427 & -0.0848 & -0.0141 \\ -0.0366 & -0.1805 & 0.2297 \\ -0.0235 & 0.0428 & 0.0752 \\ -0.1296 & -0.0424 & -0.0352 \\ -0.0716 & -0.1751 & 0.1428 \\ -0.0254 & 0.0342 & 0.0614 \\ -0.1146 & -0.0077 & -0.0487 \\ -0.0918 & -0.1569 & 0.0775 \end{pmatrix} \tag{4.6.65}$$

7. Verify that the sign and magnitude restrictions hold from 4.6.3 and 4.6.4 . It is not necessary to check for the zero restrictions since Q has been constructed to satisfy them (and indeed, it can be readily verified from $f(D, D_1, ..., D_p)$ that the zero restrictions are satisfied).

   - Positive sign restriction on the response of variable 1 to structural shock 2, to be implemented at period 2:

$$S_2 \times f_2(D, D_1, ..., D_p) = \begin{pmatrix} 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0.0720 \\ -0.1014 \\ -0.1737 \\ 0.0579 \\ -0.0848 \\ -0.1805 \\ 0.0428 \\ -0.0424 \\ -0.1751 \\ 0.0342 \\ -0.0077 \\ -0.1569 \end{pmatrix} = (0.0428) > 0 \tag{4.6.66}$$

The first sign restriction is satisfied.

- A negative sign restriction on the response of variable 1 to structural shock 3, to be implemented at period 3:

$$S_3 \times f_3(D, D_1, ..., D_p) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0.1511 \\ -0.0377 \\ 0.3079 \\ 0.0887 \\ -0.0141 \\ 0.2297 \\ 0.0752 \\ -0.0352 \\ 0.1428 \\ 0.0614 \\ -0.0487 \\ 0.0775 \end{pmatrix} = (-0.0614) < 0 \tag{4.6.67}$$

Hence, the second sign restriction is not satisfied.

- A magnitude restriction of

$$\begin{bmatrix} -0.5 & 0.5 \end{bmatrix} \tag{4.6.68}$$

on the response of variable 2 to structural shock 2, and a magnitude restriction of

$$\begin{bmatrix} -0.3 & 0.7 \end{bmatrix} \tag{4.6.69}$$

on the response of variable 3 to structural shock 2, both to be implemented at period 1:

$$
\begin{aligned}
&(M_2 \times f_2(D, D_1, ..., D_p) - M_{l,2}) \,. \times (M_{u,2} - M_2 \times f_2(D, D_1, ..., D_p)) \\
&= \left(\begin{pmatrix} -0.0848 \\ -0.1805 \end{pmatrix} - \begin{pmatrix} -0.5 \\ -0.3 \end{pmatrix} \right) . \times \left(\begin{pmatrix} 0.5 \\ 0.7 \end{pmatrix} - \begin{pmatrix} -0.0848 \\ -0.1805 \end{pmatrix} \right) \\
&= \begin{pmatrix} 0.4152 \\ 0.1195 \end{pmatrix} . \times \begin{pmatrix} 0.5848 \\ 0.8805 \end{pmatrix} \\
&= \begin{pmatrix} 0.2428 \\ 0.1052 \end{pmatrix} > 0
\end{aligned} \tag{4.6.70}
$$

Therefore, the magnitude restrictions are satisfied. Only one restriction is not satisfied (the second sign restriction), but this is sufficient to conclude that the identified Q matrix is not a valid candidate. Go back to steps 3-8 and repeat until a valid candidate is obtained.

5 Advanced analysis

5.1 Forecast error variance decomposition

Consider again model 3.1.2:

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + ... + A_p y_{t-p} + C x_t + \varepsilon_t, \qquad t = 1, 2, ..., T \tag{5.1.1}$$

If it is invertible, it admits the infinite order moving average representation 4.2.4:

$$y_t = A(L)^{-1} C x_t + \varepsilon_t + \Psi_1 \varepsilon_{t-1} + \Psi_2 \varepsilon_{t-2} + ... \tag{5.1.2}$$

This is the moving average representation of the reduced-form VAR. As shown by 4.3.13, by using a structural decomposition, one may obtain a moving average representation of the VAR in terms of uncorrelated structural disturbances as:

$$y_t = A(L)^{-1} C x_t + D\eta_t + \tilde{\Psi}_1 \eta_{t-1} + \tilde{\Psi}_2 \eta_{t-2} + ... \tag{5.1.3}$$

with:

$$\tilde{\Psi}_k = \begin{pmatrix} \tilde{\phi}_{k,11} & \tilde{\phi}_{k,12} & \cdots & \tilde{\phi}_{k,1n} \\ \tilde{\phi}_{k,21} & \tilde{\phi}_{k,22} & \cdots & \tilde{\phi}_{k,2n} \\ \vdots & \vdots & & \vdots \\ \tilde{\phi}_{k,n1} & \tilde{\phi}_{k,n2} & \cdots & \tilde{\phi}_{k,nn} \end{pmatrix} \tag{5.1.4}$$

Consider forecasting the value y_{t+h} using 5.1.3:

$$y_{t+h} = A(L)^{-1} C x_{t+h} + \sum_{k=0}^{\infty} \tilde{\Psi}_i \eta_{t+h-k} \tag{5.1.5}$$

This can be separated into three components:

$$y_{t+h} = \underbrace{A(L)^{-1} C x_{t+h}}_{\substack{\text{future values of exogenous variables,} \\ \text{assumed to be known}}} + \underbrace{\sum_{k=0}^{h-1} \tilde{\Psi}_k \eta_{t+h-k}}_{\text{unknown, future shocks}} + \underbrace{\sum_{k=0}^{\infty} \tilde{\Psi}_k \eta_{t+h-k}}_{\text{known present and past shocks}} \tag{5.1.6}$$

Then, from 5.1.6, one obtains:

$$E_t\left(y_{t+h}\right) = \underbrace{A(L)^{-1}Cx_{t+h}}_{\substack{\text{future values of exogenous variables,}\\ \text{assumed to be known}}} + \underbrace{\sum_{k=0}^{\infty}\tilde{\Psi}_k\eta_{t+h-k}}_{\text{known present and past shocks}} \tag{5.1.7}$$

Therefore, the forecast error for y_{t+h} is given by:

$$y_{t+h} - E_t\left(y_{t+h}\right) = \underbrace{\sum_{k=0}^{h-1}\tilde{\Psi}_k\eta_{t+h-k}}_{\text{unknown, future shocks}} \tag{5.1.8}$$

Or, from 5.1.4:

$$\begin{pmatrix} y_{1,t+h} - E_t\left(y_{1,t+h}\right) \\ y_{2,t+h} - E_t\left(y_{2,t+h}\right) \\ \vdots \\ y_{n,t+h} - E_t\left(y_{n,t+h}\right) \end{pmatrix} = \begin{pmatrix} \tilde{\phi}_{0,11} & \tilde{\phi}_{0,12} & \cdots & \tilde{\phi}_{0,1n} \\ \tilde{\phi}_{0,21} & \tilde{\phi}_{0,22} & \cdots & \tilde{\phi}_{0,2n} \\ \vdots & \vdots & & \vdots \\ \tilde{\phi}_{0,n1} & \tilde{\phi}_{0,n2} & \cdots & \tilde{\phi}_{0,nn} \end{pmatrix} \begin{pmatrix} \eta_{1,t+h} \\ \eta_{2,t+h} \\ \vdots \\ \eta_{n,t+h} \end{pmatrix} + \begin{pmatrix} \tilde{\phi}_{1,11} & \tilde{\phi}_{1,12} & \cdots & \tilde{\phi}_{1,1n} \\ \tilde{\phi}_{1,21} & \tilde{\phi}_{1,22} & \cdots & \tilde{\phi}_{1,2n} \\ \vdots & \vdots & & \vdots \\ \tilde{\phi}_{1,n1} & \tilde{\phi}_{1,n2} & \cdots & \tilde{\phi}_{1,nn} \end{pmatrix} \begin{pmatrix} \eta_{1,t+h-1} \\ \eta_{2,t+h-1} \\ \vdots \\ \eta_{n,t+h-1} \end{pmatrix} \ldots \tag{5.1.9}$$

Therefore, considering variable i in the VAR ($i = 1, 2, ..., n$), and using 5.1.9, the forecast error 5.1.8 rewrites:

$$y_{i,t+h} - E_t\left(y_{i,t+h}\right) = \sum_{k=0}^{h-1}\left(\tilde{\phi}_{k,i1}\eta_{1,t+h-k} + \tilde{\phi}_{k,i2}\eta_{2,t+h-k} + ... + \tilde{\phi}_{k,in}\eta_{n,t+h-k}\right) \tag{5.1.10}$$

or

$$y_{i,t+h} - E_t\left(y_{i,t+h}\right) = \sum_{k=0}^{h-1}\left(\tilde{\phi}_{k,i1}\eta_{1,t+h-k}\right) + \sum_{k=0}^{h-1}\left(\tilde{\phi}_{k,i2}\eta_{2,t+h-k}\right) + ... + \sum_{k=0}^{h-1}\left(\tilde{\phi}_{k,in}\eta_{n,t+h-k}\right) \tag{5.1.11}$$

Denote the variance of this forecast error by $\sigma^2_{y,i}(h)$, and the constant variances of the structural innovation series found on the diagonal of Γ by $\sigma^2_{\eta,1}, \sigma^2_{\eta,2}, ..., \sigma^2_{\eta,n}$. Then, taking the variance of both sides of 5.1.11, it rewrites as:

$$\sigma^2_{y,i}(h) = \sigma^2_{\eta,1}\sum_{k=0}^{h-1}\left(\tilde{\phi}_{k,i1}\right)^2 + \sigma^2_{\eta,2}\sum_{k=0}^{h-1}\left(\tilde{\phi}_{k,i2}\right)^2 + ... + \sigma^2_{\eta,n}\sum_{k=0}^{h-1}\left(\tilde{\phi}_{k,in}\right)^2 \tag{5.1.12}$$

where use has been made of the fact that structural disturbances are uncorrelated, so that no covariance terms appear in the formula. Then, dividing both sides of 5.1.12 by $\sigma^2_{y,i}(h)$, one obtains:

$$1 = \underbrace{\frac{\sigma^2_{\eta,1}}{\sigma^2_{y,i}(h)} \sum_{k=0}^{h-1} \left(\tilde{\phi}_{k,i1}\right)^2}_{\substack{\text{proportion of } \sigma^2_{y,i}(h) \\ \text{due to shocks in the} \\ \eta_{1,t} \text{ sequence}}} + \underbrace{\frac{\sigma^2_{\eta,2}}{\sigma^2_{y,i}(h)} \sum_{k=0}^{h-1} \left(\tilde{\phi}_{k,i2}\right)^2}_{\substack{\text{proportion of } \sigma^2_{y,i}(h) \\ \text{due to shocks in the} \\ \eta_{2,t} \text{ sequence}}} + ... + \underbrace{\frac{\sigma^2_{\eta,n}}{\sigma^2_{y,i}(h)} \sum_{k=0}^{h-1} \left(\tilde{\phi}_{k,in}\right)^2}_{\substack{\text{proportion of } \sigma^2_{y,i}(h) \\ \text{due to shocks in the} \\ \eta_{n,t} \text{ sequence}}} \tag{5.1.13}$$

or

$$1 = \sigma_i(1,h) + \sigma_i(2,h) + ... + \sigma_i(n,h) \tag{5.1.14}$$

with

$$\sigma_i(j,h) = \frac{\sigma^2_{\eta,j}}{\sigma^2_{y,i}(h)} \sum_{k=0}^{h-1} \left(\tilde{\phi}_{k,ij}\right)^2 \tag{5.1.15}$$

$\sigma_i(j,h)$ represents the proportion of forecast error variance of variable i due to structural shock j at horizon $T+h$. Once again, this characterizes identification of variance decomposition in a frequentist approach. In a Bayesian framework, one must take into account the uncertainty related to the parameters. Hence, rather than computing a single estimate, one draws estimates for the variance decomposition directly from the posterior distribution of orthogonalised impulse response functions. Then, with these draws from the posterior distribution of variance decomposition, one may as usual compute point estimates and credibility intervals. The following algorithm is thus proposed:

Algorithm 3.1.1 (forecast error variance decomposition, all priors):

1. define the number of iterations $(It - Bu)$ of the algorithm, and h, the time horizon for the forecast error variance decomposition.

2. at iteration n, draw $\Gamma_{(n)}, \tilde{\Psi}_1^{(n)}, \tilde{\Psi}_2^{(n)}$... from their posterior distributions. Simply recycle draw n from the SVAR Gibbs algorithm previously run.

3. obtain $\sigma_i^{(n)}(j,h)$, the variance decomposition values from 5.1.15, for $i = 1,..,n$, $j = 1,..,n$, and $t = 1,..,h$.

4. repeat until $(It - Bu)$ iterations have been achieved. This yields a sample of independent draws from the posterior distribution of variance decomposition,

$$\left\{\sigma_i^{(n)}(j,h)\right\}_{n=1}^{It-Bu} \tag{5.1.16}$$

for $i = 1,..,n$, $j = 1,..,n$, and $t = 1,..,h$.

These draws can then be used for point estimates and confidence intervals.

5.2 Historical decomposition

A matter of interest with VAR models is to establish the contribution of each structural shock to the historical dynamics of the data series. Precisely, for every period of the sample, one may want to decompose the value of each variable into its different components, each components being due to one structural shock of the model. This identifies the historical contribution of each shock to the observed data sample.

Concretely, consider again model 3.1.2:

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + C x_t + \varepsilon_t, \qquad t = 1, 2, \ldots, T \tag{5.2.1}$$

Consider again for simplicity the case where the VAR model has only one lag:

$$y_t = A_1 y_{t-1} + C x_t + \varepsilon_t \tag{5.2.2}$$

By backward substitution, one obtains:

$$\begin{aligned} y_t &= A_1 y_{t-1} + C x_t + \varepsilon_t \\ &= A_1 \left(A_1 y_{t-2} + C x_{t-1} + \varepsilon_{t-1} \right) + C x_t + \varepsilon_t \\ &= A_1 A_1 y_{t-2} + C x_t + A_1 C x_{t-1} + \varepsilon_t + A_1 \varepsilon_{t-1} \end{aligned} \tag{5.2.3}$$

Going one step further:

$$\begin{aligned} y_t &= A_1 A_1 y_{t-2} + C x_t + A_1 C x_{t-1} + \varepsilon_t + A_1 \varepsilon_{t-1} \\ &= A_1 A_1 \left(A_1 y_{t-3} + C x_{t-2} + \varepsilon_{t-2} \right) + C x_t + A_1 C x_{t-1} + \varepsilon_t + A_1 \varepsilon_{t-1} \\ &= A_1 A_1 A_1 y_{t-3} + C x_t + A_1 C x_{t-1} + A_1 A_1 C x_{t-2} + \varepsilon_t + A_1 \varepsilon_{t-1} + A_1 A_1 \varepsilon_{t-2} \end{aligned} \tag{5.2.4}$$

Going on, one may go back to the beginning of the sample:

$$y_t = (A_1)^t y_0 + \sum_{j=0}^{t-1} (A_1)^j C x_{t-j} + \sum_{j=0}^{t-1} (A_1)^j \varepsilon_{t-j} \tag{5.2.5}$$

When there is more than one lag, the backward substitution process becomes more complex, but the logic remains the same. In general, for a model with p lags, one can rewrite y_t as:

$$y_t = \sum_{j=1}^{p} A_j^{(t)} y_{1-j} + \sum_{j=0}^{t-1} C_j x_{t-j} + \sum_{j=0}^{t-1} B_j \varepsilon_{t-j} \tag{5.2.6}$$

where the matrix series $A_j^{(t)}$, C_j and B_j are (potentially complicated) functions of $A_1, A_2, ..., A_p$. For instance, in 5.2.5, $A_1^{(t)} = (A_1)^t$, $C_j = (A_1)^j C$, and $B_j = (A_1)^j$. The t superscript on $A_j^{(t)}$ emphasizes the fact that the matrix $A_j^{(t)}$ depends on $t = 1, 2, ..., T$, while the C_j and $B_j$series are independent of it.

Note also that the matrices $B_1, B_2, ..., B_{t-1}$ provide the response of y_t to shocks occurring at periods $t, t-1, ..., 2, 1$. By definition, they are thus the series of impulse response function matrices:

$$B_j = \Psi^j \tag{5.2.7}$$

Therefore, one can rewrite 5.2.6 as:

$$y_t = \sum_{j=1}^{p} A_j^{(t)} y_{1-j} + \sum_{j=0}^{t-1} C_j x_{t-j} + \sum_{j=0}^{t-1} \Psi_j \varepsilon_{t-j} \tag{5.2.8}$$

Using 4.3.6 and 4.3.14, one may obtain a representation of the impulse response functions in terms of structural shocks:

$$\Psi_j \varepsilon_{t-j} = \Psi_j D D^{-1} \varepsilon_{t-j} = \tilde{\Psi}_j \eta_{t-j} \tag{5.2.9}$$

Then, 5.2.8 rewrites as:

$$y_t = \underbrace{\sum_{j=1}^{p} A_j^{(t)} y_{1-j} + \sum_{j=0}^{t-1} C_j x_{t-j}}_{\substack{\text{historical contribution of} \\ \text{deterministic variables}}} + \underbrace{\sum_{j=0}^{t-1} \tilde{\Psi}_j \eta_{t-j}}_{\substack{\text{historical contribution} \\ \text{of structural shocks}}} \tag{5.2.10}$$

5.2.10 makes it clear that y_t can be separated into two parts: one due to deterministic exogenous variables and initial conditions, and one due to the contribution of unpredictable structural disturbances affecting the dynamics of the model. Because it is only the latter part which is of interest in this exercise, one may simply rewrite 5.2.10 as:

$$y_t = d^{(t)} + \underbrace{\sum_{j=0}^{t-1} \tilde{\Psi}_j \eta_{t-j}}_{\substack{\text{historical contribution} \\ \text{of structural shocks}}} \tag{5.2.11}$$

where for and VAR with n variables, $d^{(t)}$ is a $n \times 1$ vector of contributions from deterministic

variables and initial conditions. Considering variable i of the model (for $i = 1, 2, ..., n$) , one can then express the value of $y_{i,t}$ as:

$$y_{i,t} = d_i^{(t)} + \sum_{j=0}^{t-1} \left(\tilde{\phi}_{j,i1} \eta_{1,t-j} + \tilde{\phi}_{j,i2} \eta_{2,t-j} + ... + \tilde{\phi}_{j,in} \eta_{n,t-j} \right) \tag{5.2.12}$$

where $\tilde{\phi}_{j,ik}$ denotes entry (i, k) of the structural impulse response matrix $\tilde{\Psi}_j$. Rearranging:

$$y_{i,t} = d_i^{(t)} + \underbrace{\sum_{j=0}^{t-1} \tilde{\phi}_{j,i1} \eta_{1,t-j}}_{\substack{\text{Historical contribution} \\ \text{of structural shock 1}}} + \underbrace{\sum_{j=0}^{t-1} \tilde{\phi}_{j,i2} \eta_{2,t-j}}_{\substack{\text{Historical contribution} \\ \text{of structural shock 2}}} + ... + \underbrace{\sum_{j=0}^{t-1} \tilde{\phi}_{j,in} \eta_{n,t-j}}_{\substack{\text{Historical contribution} \\ \text{of structural shock n}}} \tag{5.2.13}$$

Equation 5.2.13 finally expresses, for each variable in the model, the historical decomposition of this variable in terms of present and past structural shocks, along with its exogenous component.

Once again, this describes the historical decomposition in a traditional, frequentist context. Because the Bayesian framework implies uncertainty with respect to the VAR coefficients, this uncertainty must be integrated into the above framework, and one must compute the posterior distribution of the historical decomposition. As usual, this is done by integrating the historical decomposition framework into the Gibbs sampler, in order to obtain draws from the posterior distribution. The following Gibbs algorithm is proposed:

Algorithm 3.2.1 (Historical decomposition, all priors):

1. Define the number of iterations $(It - Bu)$ of the algorithm. Then run the algorithm:
2. At iteration n, obtain random draws $\beta_{(n)}$, $\Sigma_{(n)}$ and $D_{(n)}$ from their posterior distribution. Simply recycle draw n from the Gibbs sampler.
3. For $j = 1, 2..., T$, compute the impulse response function matrices Ψ_j and $\tilde{\Psi}_j$ from $\beta_{(n)}$, $\Sigma_{(n)}$ and $D_{(n)}$.
4. For $j = 1, 2..., T$, obtain the VAR residuals ε_t from $\beta_{(n)}$. Then, using $D_{(n)}$, obtain the structural disturbances η_t.
5. With these elements, compute for $j = 1, 2..., T$ the historical contribution of each shock, using 5.2.13.

6. For $j = 1, 2..., T$, obtain the contribution of the non-shock components by rearranging 5.2.13:

$$d_i^{(t)} = y_{i,t} - \sum_{j=0}^{t-1} \tilde{\phi}_{j,i1}\eta_{1,t-j} - \sum_{j=0}^{t-1} \tilde{\phi}_{j,i2}\eta_{2,t-j} - ... - \sum_{j=0}^{t-1} \tilde{\phi}_{j,in}\eta_{n,t-j} \tag{5.2.14}$$

7. Repeat until $(It - Bu)$ iterations are realised.

5.3 Conditional forecasts

Sometimes, one may also be interested into obtaining what is known as conditional forecasts. These are defined as forecasts obtained by constraining the path of certain variables to take specific values decided by the statistician. In other words, they are obtained conditional on given values for a subset of variables, over a subset of periods. This technique is very useful when one wants for instance to simulate a scenario for some specific variables and observe the outcome for the other variables, or compare the differences in outcomes obtained from different scenarios.

To derive conditional forecasts, consider again the VAR model 3.1.2:

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + C x_t + \varepsilon_t \tag{5.3.1}$$

Assume that one wants to use model 5.3.1 to produce forecasts, and to make things more concrete, consider the simple case of predicting y_{T+2} for a VAR with one lag. By recursive iteration similar to that applied in subsubsection 4.2, one obtains:

$$y_{T+1} = A_1 y_T + C x_{T+1} + \varepsilon_{T+1} \tag{5.3.2}$$

and

$$\begin{aligned} y_{T+2} &= A_1 y_{T+1} + C x_{T+2} + \varepsilon_{T+2} \\ &= A_1(A_1 y_T + C x_{T+1} + \varepsilon_{T+1}) + C x_{T+2} + \varepsilon_{T+2} \\ &= A_1 A_1 y_T + A_1 C x_{T+1} + C x_{T+2} + A_1 \varepsilon_{T+1} + \varepsilon_{T+2} \end{aligned} \tag{5.3.3}$$

The forecasts are functions of three terms: terms involving the present value of the endogenous variables y_t, terms involving future values of the exogenous variables x_t, and terms involving future values of the reduced form residuals ε_t. In general, for a forecast of period $T + h$ from a VAR with $p$lags, one obtains an expression of the form:

$$y_{T+h} = \sum_{j=1}^{p} A_j^{(h)} y_{T-j+1} + \sum_{j=1}^{h} C_j^{(h)} x_{T+j} + \sum_{j=1}^{h} B_j^{(h)} \varepsilon_{T+j} \tag{5.3.4}$$

where $A_j^{(h)}, B_j^{(h)}$ and $C_j^{(h)}$ denote the respective matrix coefficients on $y_{T-j+1}, \varepsilon_{T+j}$ and x_{T+j} for a forecast of y_{T+h}. In the preceding example for instance, $A_1^{(1)} = A_1$ and $A_1^{(2)} = A_1 A_1$. With more lags and longer forecast horizons, the series $A_j^{(h)}, B_j^{(h)}$ and $C_j^{(h)}$ can quickly become fairly complicated functions of $A_1, A_2, ..., A_p$ and C, but they are easy to recover numerically by successive iterations.

Note also that the matrices $B_1^{(h)}, B_2^{(h)}, ..., B_h^{(h)}$ provide the response of y_{T+h} to shocks in $y_{T+1}, y_{T+2} ... y_{T+h}$. By definition, as showned in subsection 4.2, they are thus the series of impulse response function matrices:

$$B_j^{(h)} = \Psi_{h-j} \tag{5.3.5}$$

Therefore, one may rewrite 5.3.4 as:

$$y_{T+h} = \sum_{j=1}^{p} A_j^{(h)} y_{T-j+1} + \sum_{j=1}^{h} C_j^{(h)} x_{T+j} + \sum_{j=1}^{h} \Psi_{h-j} \varepsilon_{T+j} \tag{5.3.6}$$

Now, using 4.3.6 and 4.3.14, one may rewrite

$$\Psi_{h-j} \varepsilon_{T+j} = \Psi_{h-j} D D^{-1} \varepsilon_{T+j} = \tilde{\Psi}_{h-j} \eta_{T+j} \tag{5.3.7}$$

5.3.7 then allows to rewrite the forecast function 5.3.6 in terms of structural shocks:

$$y_{T+h} = \underbrace{\sum_{j=1}^{p} A_j^{(h)} y_{T-j+1} + \sum_{j=1}^{h} C_j^{(h)} x_{T+j}}_{\text{Forecast in the absence of shocks}} + \underbrace{\sum_{j=1}^{h} \tilde{\Psi}_{h-j} \eta_{T+j}}_{\substack{\text{Dynamic impact of} \\ \text{future structural shocks}}} \tag{5.3.8}$$

In the right-hand side of 5.3.8, the first and second terms are known values (they are just the regular forecasts), while the third is unknown. Therefore, one may rewrite 5.3.8 as:

$$y_{T+h} = \tilde{y}_{T+h} + \sum_{j=1}^{h} \tilde{\Psi}_{h-j} \eta_{T+j} \tag{5.3.9}$$

with $\tilde{y}_{T+h}$ the unconditional forecast value for period $T+h$. Consider imposing the condition:

$$y_{i,T+h} = \bar{y} \tag{5.3.10}$$

Where $y_{i,T+h}$ is the value of variable i (over the n variables in the model) at period $T+h$, and $\bar{y}$ is any scalar value decided by the statistician. Then, considering row i in 5.3.9, one obtains:

$$y_{i,T+h} = \tilde{y}_{i,T+h} + \sum_{j=1}^{h} \tilde{\Psi}_{h-j,i}\eta_{T+j} = \bar{y} \tag{5.3.11}$$

where $\tilde{\Psi}_{h-j,i}$ denotes row i of the impulse response matrix $\tilde{\Psi}_{h-j}$. This can be rearranged to obtain:

$$\sum_{j=1}^{h} \tilde{\Psi}_{h-j,i}\eta_{T+j} = \bar{y} - \tilde{y}_{i,T+h} \tag{5.3.12}$$

5.3.12 shows that constraining y_{T+h} to take the fixed value $\bar{y}$ will imply restrictions on the value of future structural innovations, up to period $T+h$. 5.3.12 can be reformulated in linear form as:

$$\underbrace{(\tilde{\phi}_{h-1,i1} \quad \cdots \quad \tilde{\phi}_{h-1,in}}_{j=1} | \underbrace{\cdots \quad \cdots}_{j=2,\ldots,h-1} | \underbrace{\tilde{\phi}_{0,i1} \quad \cdots \quad \tilde{\phi}_{0,in}}_{j=h}) \begin{pmatrix} \eta_{1,T+1} \\ \vdots \\ \eta_{n,T+1} \\ \vdots \\ \vdots \\ \eta_{1,T+h} \\ \vdots \\ \eta_{n,T+h} \end{pmatrix} = \left(\bar{y} - \tilde{y}_{i,T+h}\right) \tag{5.3.13}$$

where the vectors of restriction and the vector of shocks are of size $1 \times s$, with $s = h \times n$ the total number of shocks to constrain (n structural shocks for a n-variable VAR, to be constrained over h periods). $\tilde{\phi}_{h-j,i1}, \ \ldots \ , \tilde{\phi}_{h-j,in}$ are the n elements of $\tilde{\Psi}_{h-j,i}$.

If one wants to impose v conditions of the type of 5.3.10, then these conditions can be all put in the linear form 5.3.13 and then stacked to form a linear system:

$$\underbrace{R}_{\nu \times s} \underbrace{\eta}_{s \times 1} = \underbrace{r}_{\nu \times 1} \quad \nu \leq s = nh \tag{5.3.14}$$

R is the $\nu \times s$ matrix of linear restrictions, η is the $s \times 1$ vector gathering all the structural shocks to be constrained for a forecast at horizon $T+h$, and r is the $\nu \times 1$ vector gathering the differences between the predicted and conditional values.

To make things more concrete, consider the case of a 2-variable VAR with variables $y_{1,t}$ and $y_{2,t}$ gathered in the vector y_t, and assume one wants to produce a forecast for y_{T+3} , so that the final forecast horizon $h = 3$. Assume that one wants to produce this forecast by constraining the first

variable to take specific values over the first two forecast periods, that is: $y_{1,T+1} = \bar{y}_1$ and $y_{1,T+2} = \bar{y}_2$. Then $n = 2$, $h = 3$ so that $s = h \times n = 2 \times 3 = 6$. Because two restrictions are imposed, $\nu = 2$. Hence R will be $\nu \times s = 2 \times 6$, η will be $s \times 1$ or 6×1, and r will be $\nu \times 1$ or 2×1.

Precisely, the stacked linear form system 5.3.14 will be given by:

$$\begin{pmatrix} \tilde{\phi}_{0,11} & \tilde{\phi}_{0,12} & 0 & 0 & 0 & 0 \\ \tilde{\phi}_{1,11} & \tilde{\phi}_{1,12} & \tilde{\phi}_{0,11} & \tilde{\phi}_{0,12} & 0 & 0 \end{pmatrix} \begin{pmatrix} \eta_{1,T+1} \\ \eta_{2,T+1} \\ \eta_{1,T+2} \\ \eta_{2,T+2} \\ \eta_{1,T+3} \\ \eta_{2,T+3} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 - y_{1,T+1} \\ \bar{y}_2 - y_{2,T+2} \end{pmatrix} \tag{5.3.15}$$

Once the system 5.3.14 is obtained, it remains to know how to draw the structural disturbances so that this system of constraint will be satisfied. The main contribution is due to Waggoner and Zha (1999) who derive a Gibbs sampling algorithm to construct the posterior predictive distribution of the conditional forecast. In particular they show that the distribution of the restricted future shocks is normal, and characterised by:

$$\eta \sim \mathcal{N}(\bar{\eta}, \bar{\Gamma}) \tag{5.3.16}$$

with

$$\bar{\eta} = R^{'}(RR^{'})^{-1}r \tag{5.3.17}$$

and

$$\bar{\Gamma} = I - R^{'}(RR^{'})^{-1}R \tag{5.3.18}$$

While in theory, one could draw shocks directly from $N(\bar{\eta}, \bar{\Gamma})$, in practice an equivalent but numerically more efficient solution has been proposed by Jarocinski (2010a). It consists in taking the singular value decomposition of R, that is, in identifying a $v \times v$ unitary matrix U, a $v \times s$ matrix $S$and $s \times s$ unitary matrix Vsuch that $R = USV$. Denoting by P the matrix made of the first s columns of S, by V_1 the matrix made of the first s columns of V, and by V_2 the matrix made of the remaining $v - s$ columns of V, Jarocinski (2010a) shows that drawing from 5.3.16 is equivalent to drawing from:

$$V_1 P^{-1} U^{,} r + V_2 \lambda, \text{ where } \lambda \sim \mathcal{N}(0, I_{s-v}) \tag{5.3.19}$$

Once these shocks are drawn, the system of constraints 5.3.14 will be satisfied by construction,

and forecast can be formed by direct application of 5.3.9.

It is then possible to use the traditional Gibbs sampler framework to produce the posterior distribution of conditional forecasts, as detailed in the incoming algorithm. A note of warning however: the methodology adopted here is conventional, but is not exactly similar to that proposed by Waggoner and Zha (1999). These authors suggest that in order to obtain a posterior distribution accounting for parameter uncertainty in finite samples, the Gibbs sampler algorithm should, at each iteration, augment the sample of data with the predicted values previously obtained. Doing so, the posterior distribution will shift away from the one that would be obtained by considering only the sample of actual data, which is labelled by the authors as a "shift in distribution" phenomenon. The algorithm developed hereafter does not adopt this specific methodology, as it can suffer from numerical instability, and produce shifts in distribution that are so large that the implied forecast values may not have reasonable economic interpretation anymore. Adopting thus a standard approach, the following algorithm is proposed:

Algorithm 3.3.1 (conditional forecasts, all priors):

1. define the total number of iterations of the algorithm $It - Bu$, the forecast horizon h, and the v conditions : $\bar{y}_1, \bar{y}_2, ..., \bar{y}_v$ to be imposed on the series.
2. at iteration n, draw $\beta_{(n)}$, $\Sigma_{(n)}$and $D_{(n)}$. In practice, recycle draws from the Gibbs sampler.
3. at iteration n, compute first the unconditional forecasts $\tilde{y}_{T+1}, \tilde{y}_{T+2}, ..., \tilde{y}_{T+h}$ from $\beta_{(n)}$, by iteratively using 5.3.4, but excluding shocks.
4. for $j = 1, 2..., h$, compute the impulse response function matrices Ψ_j and $\tilde{\Psi}_j$ from $\beta_{(n)}$, $\Sigma_{(n)}$and $D_{(n)}$.
5. build R and r, as defined in 5.3.14.
6. draw the constrained shocks, using 5.3.19.
7. calculate the conditional forecasts using 5.3.9, with the unconditional forecast values obtained in step 2 and the constrained shocks obtained in step 5.
8. repeat steps 2-7 until $It - Bu$ iterations are realised.

5.4 Conditional forecasts generated by specific shocks

In practice, when one works with conditional forecasts, assuming that the conditions are generated by all the structural shocks of the model may be undesirable or economically irrelevant. For instance, consider a simple two-variable VAR model in GDP and a central bank interest rate. Assume that the

condition to be implemented is an increase of the interest rate to over the next four quarters. There can be two stories behind this increase. A first possibility is that the central authorities wanted to implement this rise for some reason (for instance, attract more foreign investment), so that the decision has been taken regardless of GDP. In this case, the increase in the interest rate will be implemented through monetary shocks, and its expected effect is a negative impact on GDP in the short run.

The second possibility is that it is actually a fuelling in activity (rise in GDP) that motivated the rise in interest rate, pushing the central authorities to fight inflationary pressure. In this case the condition is due to GDP shocks, and one will observe an initial increase in GDP in the short run (the one generating the policy, before the policy impacts negatively real activity). It then appears that the observed effect on output will be opposite, depending on which shock originates the constraint.

This simple example suggests that one may want to select carefully the shocks originating the constraint, in order to produce meaningful economic results. Unfortunately, doing so complicates substantially the conditional forecast framework, and it is not always possible to implement any arbitrary choice of shocks generating the conditions.

So, consider again the conditional forecast framework 5.3.11, in which some variable i of the VAR model takes the value $\bar{y}$ at forecast horizon $T+h$:

$$y_{i,T+h} = \bar{y} = \tilde{y}_{i,T+h} + \sum_{j=1}^{h} \tilde{\Psi}_{h-j,i}\eta_{T+j} \tag{5.4.1}$$

Before one enters into the computational details, some definitions are stated. First, a shock in η_{T+h} (the vector of all structural shocks at period $T+h$) is said to be constructive for $y_{i,T+h}$ if it is used to generate the condition $y_{i,T+h} = \bar{y}$. If it is not used to generate this condition, the shock is labelled as non-constructive. Shocks can only be constructive for conditions on their own periods. Secondly, one defines a block of variables for period $T+h$ as the set of all variables on which there is a condition for this period, and for which the conditions are generated by the same shocks. A single variable may perfectly constitute a block by itself, if its condition is generated by shocks that are not constructive for any other variables.

Using a concrete example is probably the simplest way to illustrate the methodology and the concepts used in conditional forecasts generated by specific shocks. Consider for instance a 3-variable VAR model for which forecasts are produced up to horizon $T+3$, and for which conditions are set over the first two periods.

At period $T+1$, conditions are set on variables 1 and 2 : $y_{1,T+1} = \bar{y}_1$ and $y_{2,T+1} = \bar{y}_2$. It is assumed that these conditions are generated by structural shocks 1 and 2 for both variables.

They hence constitute the constructive shocks for $y_{1,T+1}$ and $y_{2,T+1}$, while structural shock 3 is non-constructive. Also, because the conditions on $y_{1,T+1}$ and $y_{2,T+1}$ are generated by the same shocks, variables 1 and 2 constitute a block (actually, the unique block for $T+1$). At period $T+2$, only one condition is set, on variable 3: $y_{3,T+2} = \bar{y}_3$. This condition is generated by shock 3. Variable 3 therefore constitutes the unique block for period $T+2$, associated with the unique constructive shock 3. Structural shocks 1 and 2 are non-constructive. Finally, at period 3, there are no conditions at all. Therefore, all the shocks are non-constructive and there are no blocks.

The next paragraph illustrates the procedure used to identify the values of the structural shocks, both constructive and non-constructive, that will fulfil the conditions placed on every forecast period. The general idea is to split the initial problem into separate problems (one for each forecast period), and then solve in a recursive way, starting with the first period, up to the final forecast period. To start the resolution, first recover the linear constraint system 5.3.14 corresponding to the complete conditional forecast problem:

$$\begin{pmatrix} \tilde{\phi}_{0,11} & \tilde{\phi}_{0,12} & \tilde{\phi}_{0,13} & 0 & 0 & 0 & 0 & 0 & 0 \\ \tilde{\phi}_{0,21} & \tilde{\phi}_{0,22} & \tilde{\phi}_{0,23} & 0 & 0 & 0 & 0 & 0 & 0 \\ \tilde{\phi}_{1,31} & \tilde{\phi}_{1,32} & \tilde{\phi}_{1,33} & \tilde{\phi}_{0,31} & \tilde{\phi}_{0,32} & \tilde{\phi}_{0,33} & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \eta_{1,T+1} \\ \eta_{2,T+1} \\ \eta_{3,T+1} \\ \eta_{1,T+2} \\ \eta_{2,T+2} \\ \eta_{3,T+2} \\ \eta_{1,T+3} \\ \eta_{2,T+3} \\ \eta_{3,T+3} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 - \tilde{y}_{1,T+1} \\ \bar{y}_2 - \tilde{y}_{2,T+1} \\ \bar{y}_3 - \tilde{y}_{3,T+2} \end{pmatrix} \tag{5.4.2}$$

Then start solving recursively, period by period:

Period T+1: First, because only period $T+1$ is of concern for now, one can retain only the rows of system 5.4.2 related to this period and obtain a modified system:

$$\begin{pmatrix} \tilde{\phi}_{0,11} & \tilde{\phi}_{0,12} & \tilde{\phi}_{0,13} & 0 & 0 & 0 & 0 & 0 & 0 \\ \tilde{\phi}_{0,21} & \tilde{\phi}_{0,22} & \tilde{\phi}_{0,23} & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \eta_{1,T+1} \\ \eta_{2,T+1} \\ \eta_{3,T+1} \\ \eta_{1,T+2} \\ \eta_{2,T+2} \\ \eta_{3,T+2} \\ \eta_{1,T+3} \\ \eta_{2,T+3} \\ \eta_{3,T+3} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 - \tilde{y}_{1,T+1} \\ \bar{y}_2 - \tilde{y}_{2,T+1} \end{pmatrix} \tag{5.4.3}$$

One can actually go further: because the shocks corresponding to future periods are irrelevant for period $T+1$ (see the 0 coefficients corresponding to shocks in periods $T+2$ and $T+3$), one may take them out of 5.4.3, and reformulate it as:

$$\begin{pmatrix} \tilde{\phi}_{0,11} & \tilde{\phi}_{0,12} & \tilde{\phi}_{0,13} \\ \tilde{\phi}_{0,21} & \tilde{\phi}_{0,22} & \tilde{\phi}_{0,23} \end{pmatrix} \begin{pmatrix} \eta_{1,T+1} \\ \eta_{2,T+1} \\ \eta_{3,T+1} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 - \tilde{y}_{1,T+1} \\ \bar{y}_2 - \tilde{y}_{2,T+1} \end{pmatrix} \tag{5.4.4}$$

The advantage of formulation 5.4.4 over 5.4.3 is that it is much faster to solve in computational applications.

- Consider then the non-constructive shocks. Here only structural shock 3 is non-constructive, so draw $\eta_{3,T+1}$ from its distribution. Once its value is known, note that it is possible to produce a modified but equivalent formulation of 5.4.4 which integrates the impact of this shock by transferring it on the right-hand side.

$$\begin{pmatrix} \tilde{\phi}_{0,11} & \tilde{\phi}_{0,12} & 0 \\ \tilde{\phi}_{0,21} & \tilde{\phi}_{0,22} & 0 \end{pmatrix} \begin{pmatrix} \eta_{1,T+1} \\ \eta_{2,T+1} \\ \eta_{3,T+1} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 - \tilde{y}_{1,T+1} - \tilde{\phi}_{0,13}\eta_{3,T+1} \\ \bar{y}_2 - \tilde{y}_{2,T+1} - \tilde{\phi}_{0,23}\eta_{3,T+1} \end{pmatrix} \tag{5.4.5}$$

  The interest of system 5.4.4 is that it integrates the value of the non-constructive shocks into the conditions on the right hand side. Hence, when one will draw the subsequent shocks, they will account for the already known value of $\eta_{3,T+1}$.

- Consider then the constructive shocks for block 1 in period $T+1$. These shocks must ensure that the conditions $y_{1,T+1} = \bar{y}_1$ and $y_{2,T+1} = \bar{y}_2$ hold. To do so, draw them for The Waggoner and Zha (1999) distribution 5.3.16-5.3.18, using the modified matrices R and r defined in 5.4.5. These shocks ensure that the two conditions will be met. Retain only the shocks $\eta_{1,T+1}$ and

$\eta_{2,T+1}$ since $\eta_{3,T+1}$ has already been drawn. Note that discarding the draw of $\eta_{3,T+1}$ is of no consequence on the constraint as it is a non-constructive shock.

There is then no need to go further for period $T+1$: all the shocks (constructive and non-constructive) have been drawn, and the constructive shocks have taken into account the values of the non-constructive shocks, so that the conditions $y_{1,T+1} = \bar{y}_1$ and $y_{2,T+1} = \bar{y}_2$ hold.

Period T+2:

- Similarly to period 1, because only period $T+2$ is now of interest, one may consider only the rows of 5.4.2 which are related to this period, ignore coefficients on shocks beyond $T+2$ and obtain a modified system:

$$\begin{pmatrix} \tilde{\phi}_{1,31} & \tilde{\phi}_{1,32} & \tilde{\phi}_{1,33} & \tilde{\phi}_{0,31} & \tilde{\phi}_{0,32} & \tilde{\phi}_{0,33} \end{pmatrix} \begin{pmatrix} \eta_{1,T+1} \\ \eta_{2,T+1} \\ \eta_{3,T+1} \\ \eta_{1,T+2} \\ \eta_{2,T+2} \\ \eta_{3,T+2} \end{pmatrix} = (\bar{y}_3 - \tilde{y}_{3,T+2}) \tag{5.4.6}$$

- Since the shocks $\eta_{1,T+1}$,$\eta_{2,T+1}$ and $\eta_{3,T+1}$ are already known, they become constant values that have to be integrated to the conditioning set (the right-hand side of 5.4.6). Therefore, reformulate 5.4.6 by transferring the impact of previous shocks to the right-hand side:

$$\begin{pmatrix} 0 & 0 & 0 & \tilde{\phi}_{0,31} & \tilde{\phi}_{0,32} & \tilde{\phi}_{0,33} \end{pmatrix} \begin{pmatrix} \eta_{1,T+1} \\ \eta_{2,T+1} \\ \eta_{3,T+1} \\ \eta_{1,T+2} \\ \eta_{2,T+2} \\ \eta_{3,T+2} \end{pmatrix} = \left(\bar{y}_3 - \tilde{y}_{3,T+2} - \tilde{\phi}_{1,31}\eta_{1,T+1} - \tilde{\phi}_{1,32}\eta_{2,T+1} - \tilde{\phi}_{1,33}\eta_{3,T+1} \right) \tag{5.4.7}$$

Again, for computational purposes, it is suitable to reduce the system further and take out of the system the zero columns corresponding to shocks of period $T+1$, now that they have been taken into account:

$$\begin{pmatrix} \tilde{\phi}_{0,31} & \tilde{\phi}_{0,32} & \tilde{\phi}_{0,33} \end{pmatrix} \begin{pmatrix} \eta_{1,T+2} \\ \eta_{2,T+2} \\ \eta_{3,T+2} \end{pmatrix} = \begin{pmatrix} \bar{y}_3 - \tilde{y}_{3,T+2} - \tilde{\phi}_{1,31}\eta_{1,T+1} - \tilde{\phi}_{1,32}\eta_{2,T+1} - \tilde{\phi}_{1,33}\eta_{3,T+1} \end{pmatrix} \tag{5.4.8}$$

- Then draw the non-constructive shocks. Here they are $\eta_{1,T+2}$ and $\eta_{2,T+2}$. Draw these shocks from their distributions, then, again, transfer the impact to the right-hand side to update the condition set:

$$\begin{pmatrix} 0 & 0 & \tilde{\phi}_{0,33} \end{pmatrix} \begin{pmatrix} \eta_{1,T+2} \\ \eta_{2,T+2} \\ \eta_{3,T+2} \end{pmatrix} = \begin{pmatrix} \bar{y}_3 - \tilde{y}_{3,T+2} - \tilde{\phi}_{1,31}\eta_{1,T+1} - \tilde{\phi}_{1,32}\eta_{2,T+1} - \tilde{\phi}_{1,33}\eta_{3,T+1} - \tilde{\phi}_{0,31}\eta_{1,T+2} - \tilde{\phi}_{0,32}\eta_{2,T} \end{pmatrix} \tag{5.4.9}$$

- Then draw the constructive shock $\eta_{3,T+2}$ for period $T+2$, from the Waggoner and Zha (1999) distribution, using the modified matrices R and r defined in 5.4.9. Discard all the shocks obtained from this draw except $\eta_{3,T+2}$, which is the only constructive shock. This shock guarantees that the condition is respected. All the shocks are drawn for period , and the unique condition holds. There is nothing more to be done.

Period T+3:

- Because there are no conditions at period $T+3$, there is no linear system to create. So the only thing to do is:

- Draw the three non-constructive shocks $\eta_{1,T+3}, \eta_{2,T+3}$ and $\eta_{3,T+3}$ from their distributions. Note once again that the values of these shocks is of no consequence for past conditions, since these shocks are non-impacting for past values (see all the 0 coefficients of in 5.4.2). This finishes the process. A vector of shocks η for all forecast periods has been drawn, which ensures that all the conditions hold, and that all the non-constructive shocks are just random values drawn from their distributions.

This example illustrates the methodology to follow. It is however a simplified exercise, since it features only one block per period. In general, the full methodology to determine the shocks for a conditional forecast exercise at horizon $T+h$, with an arbitrary number of blocks for each period is as follows:

Algorithm 3.4.1 (shock identification for conditional forecasts, all priors):

1. For each forecast period $T+i$, with $i = 1, 2, \ldots, h$, decide the set of conditions, the constructive shocks and the blocks.

2. Formulate the complete problem as a linear system of the form 5.3.14: $R\eta = r$.

3. At period $T + i$: consider only the rows of the linear system $R\eta = r$ which are related to $T + i$ and obtain a restricted system. Reformulate the system to make it smaller by ignoring columns of R and shocks related to periods beyond $T + i$.

4. Then, update the system by transferring the impact of all the previous shocks $\eta_{T+1}, \eta_{T+2}, \ldots, \eta_{T+(i-1)}$ on the right-hand side. This again makes the system smaller as it permits to supress all the shocks and columns of R related to periods preceding $T + i$.

   At this point, the system has been reduced so that it only accounts for the current shocks and conditions: in the system $R\eta = r$, R is of dimension $\nu_i \times n$, with ν_i the number of conditions for period $T+i$, η is of dimension n and comprises only current period shocks, and r is $\nu_i \times 1$.

5. Draw now the non-constructive shocks (if any) from their distributions. Transfer their effects to the right-hand-side of the system. This updates R and r in the linear system.

6. Draw the constructive shocks corresponding to the first block from the Waggoner and Zha (1999) distribution, using the updated matrices R and r. Discard all the shocks which are non-constructive for block 1. The retained constructive shocks ensure that the conditions in block 1 hold. Transfer their effects to the right-hand-side of the system. This updates R and r in the linear system.

7. Then draw the constructive shocks corresponding to the second block from the Wagonner-Zha distribution, using the updated matrices R and r. Discard all the shocks which are non-constructive for block 2. The retained constructive shocks ensure that the conditions in block 2 hold. Transfer their effects to the right-hand-side of the system. This updates R and r in the linear system.

8. At period $T + 1$: repeat with all the other blocks. When all the blocks have been treated, a vector of shocks η_{T+i} has been produced that allows for all the period conditions to hold. Update η with the values η_{T+i}, and go for the next period.

9. Repeat steps 3-8 for all periods, until η_{T+h} is calculated. This produces the complete vector of shocks η allowing for every condition at every period to hold.

One would wish that this is the end of the story. Unfortunately, it is not. Things do not necessarily go as smoothly as algorithm 3.4.1 suggests. In fact, algorithm 3.4.1 may often fail, and produce nonsense values. It will actually produce correct conditional forecasts only when the conditions, shocks and blocks have been designed so that conditional forecasts are well defined. Otherwise, conditional forecasts simply don't exist. To see why, it is once again better to use examples. Of the three following examples, one works, and two fail. This helps illustrating the potential pitfalls one may face when trying to use conditional forecasts.

For example 1, consider a simple VAR model with 3 variables, where conditional forecasts have only to be produced for $T+1$. There is a condition on variable 1, so that $y_{1,T+1} = \bar{y}_1$, and this condition is generated by shock 1. This constitutes block 1. Block 2 is made of variable 2 with $y_{2,T+1} = \bar{y}_2$, generated by shocks 2 and 3. Implement algorithm 3.4.1 for this example. Generate first the linear system:

$$\begin{pmatrix} \tilde{\phi}_{0,11} & \tilde{\phi}_{0,12} & \tilde{\phi}_{0,13} \\ \tilde{\phi}_{0,21} & \tilde{\phi}_{0,22} & \tilde{\phi}_{0,23} \end{pmatrix} \begin{pmatrix} \eta_{1,T+1} \\ \eta_{2,T+1} \\ \eta_{3,T+1} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 - \tilde{y}_{1,T+1} \\ \bar{y}_2 - \tilde{y}_{2,T+1} \end{pmatrix} \tag{5.4.10}$$

Because there are no non-constructive shocks, draw directly the constructive shock $\eta_{1,T+1}$ for block 1 from the Waggoner and Zha (1999) distribution, and transfer its impact on the right-hand side of 5.4.10:

$$\begin{pmatrix} 0 & \tilde{\phi}_{0,12} & \tilde{\phi}_{0,13} \\ 0 & \tilde{\phi}_{0,22} & \tilde{\phi}_{0,23} \end{pmatrix} \begin{pmatrix} \eta_{1,T+1} \\ \eta_{2,T+1} \\ \eta_{3,T+1} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 - \tilde{y}_{1,T+1} - \tilde{\phi}_{0,11}\eta_{1,T+1} \\ \bar{y}_2 - \tilde{y}_{2,T+1} - \tilde{\phi}_{0,11}\eta_{1,T+1} \end{pmatrix} \tag{5.4.11}$$

This shock ensures that the condition $y_{1,T+1} = \bar{y}_1$ holds (implying that row 1 on the right-hand side of 5.4.11 is now equal to 0). Then go for the second block: draw the constructive shocks $\eta_{2,T+1}$ and $\eta_{3,T+1}$ for block 2 from the Waggoner and Zha (1999) distribution, using the updated system 5.4.11, and transfer their impact on the right-hand side:

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \eta_{1,T+1} \\ \eta_{2,T+1} \\ \eta_{3,T+1} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 - \tilde{y}_{1,T+1} - \tilde{\phi}_{0,11}\eta_{1,T+1} - \tilde{\phi}_{0,12}\eta_{2,T+1} - \tilde{\phi}_{0,13}\eta_{3,T+1} \\ \bar{y}_2 - \tilde{y}_{2,T+1} - \tilde{\phi}_{0,11}\eta_{1,T+1} - \tilde{\phi}_{0,22}\eta_{1,T+1} - \tilde{\phi}_{0,23}\eta_{3,T+1} \end{pmatrix} \tag{5.4.12}$$

This draw ensures that the condition on block 2 $y_{2,T+1} = \bar{y}_2$ holds. However, the first condition $y_{1,T+1} = \bar{y}_1$ may not hold anymore. Indeed compare row 1 in the right-hand sides of 5.4.11 with that of 5.4.12: there is additional term $-\tilde{\phi}_{0,12}\eta_{2,T+1} - \tilde{\phi}_{0,13}\eta_{3,T+1}$ due to the impact of the subsequent draw

of shocks for block 2. Hence, $y_{1,T+1} = \bar{y}_1$ will not hold anymore (that is, row 1 in the right-hand side of 5.4.11 will not be equal to 0 anymore), except if the additional term $-\tilde{\phi}_{0,12}\eta_{2,T+1} - \tilde{\phi}_{0,13}\eta_{3,T+1}$ is equal to 0. By chance, it turns out that this is true when the structural matrix D is lower triangular (which is the case when D is identified by Choleski or triangular factorisation). Indeed, since $\tilde{\Psi}_0 = D$ (see 4.2.14), then $\tilde{\phi}_{0,12} = \tilde{\phi}_{0,13} = 0$.

However, there is no necessity for the system to yield such a favourable result. Consider, as a second example, the same setting, except that block 1 is now made of variable 1 with $y_{1,T+1} = \bar{y}_1$, generated by shock 3, while block 2 is made of variable 2 with $y_{2,T+1} = \bar{y}_2$, generated by shocks 1 and 2. The linear system is again 5.4.10. Start by drawing the constructive shock $\eta_{3,T+1}$ for block 1, then transfer its effect on the right-hand side:

$$\begin{pmatrix} \tilde{\phi}_{0,11} & \tilde{\phi}_{0,12} & 0 \\ \tilde{\phi}_{0,21} & \tilde{\phi}_{0,22} & 0 \end{pmatrix} \begin{pmatrix} \eta_{1,T+1} \\ \eta_{2,T+1} \\ \eta_{3,T+1} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 - \tilde{y}_{1,T+1} - \tilde{\phi}_{0,13}\eta_{3,T+1} \\ \bar{y}_2 - \tilde{y}_{2,T+1} - \tilde{\phi}_{0,23}\eta_{3,T+1} \end{pmatrix} \tag{5.4.13}$$

This guarantees that the condition for block 1 holds, so that row 1 in the right-hand side of 5.4.13 is equal to 0. Then, draw now the constructive shocks $\eta_{1,T+1}$ and $\eta_{2,T+1}$ for block 2 from the Waggoner and Zha (1999) distribution using the updated system 5.4.13, and transfer their impact on the right-hand side:

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \eta_{1,T+1} \\ \eta_{2,T+1} \\ \eta_{3,T+1} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 - \tilde{y}_{1,T+1} - \tilde{\phi}_{0,13}\eta_{3,T+1} - \tilde{\phi}_{0,11}\eta_{1,T+1} - \tilde{\phi}_{0,12}\eta_{2,T+1} \\ \bar{y}_2 - \tilde{y}_{2,T+1} - \tilde{\phi}_{0,23}\eta_{3,T+1} - \tilde{\phi}_{0,21}\eta_{1,T+1} - \tilde{\phi}_{0,22}\eta_{2,T+1} \end{pmatrix} \tag{5.4.14}$$

The condition on block 2 $y_{2,T+1} = \bar{y}_2$ now holds, but the condition $y_{1,T+1} = \bar{y}_1$ will not hold anymore, except if $-\tilde{\phi}_{0,11}\eta_{1,T+1} - \tilde{\phi}_{0,12}\eta_{2,T+1} = 0$. When the structural matrix D is lower triangular, $\tilde{\phi}_{0,11}$ will not be equal to 0: if one naively implements algorithm 3.4.1, the first condition will not hold: it is actually impossible to obtain conditional forecasts with this setting. The impossibility here is due to the fact that the two blocks are conflicting, and that the structural matrix $\tilde{\Psi}_0$ does not correct for the conflict.

The third example is even worse: it shows that the system may fail even if there is only one block and hence no possible issue of conflict between blocks. For example 3, consider again the same VAR model, but consider now that there is only one condition on variable 1, $y_{1,T+1} = \bar{y}_1$, generated by shock 2. The linear system is now:

$$\begin{pmatrix} \tilde{\phi}_{0,11} & \tilde{\phi}_{0,12} & \tilde{\phi}_{0,13} \end{pmatrix} \begin{pmatrix} \eta_{1,T+1} \\ \eta_{2,T+1} \\ \eta_{3,T+1} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 - \tilde{y}_{1,T+1} \end{pmatrix} \tag{5.4.15}$$

Applying algorithm 3.4.1, there are two non-constructive shocks, $\eta_{1,T+1}$ and $\eta_{3,T+1}$, so draw these shocks and transfer their impacts to the right-hand side:

$$\begin{pmatrix} 0 & \tilde{\phi}_{0,12} & 0 \end{pmatrix} \begin{pmatrix} \eta_{1,T+1} \\ \eta_{2,T+1} \\ \eta_{3,T+1} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 - \tilde{y}_{1,T+1} - \tilde{\phi}_{0,11}\eta_{1,T+1} - \tilde{\phi}_{0,13}\eta_{3,T+1} \end{pmatrix} \tag{5.4.16}$$

Now, one would like to draw the constructive shock $\eta_{1,T+2}$ from the Waggoner and Zha (1999) distribution, using the updated system (15.14). But with a conventional lower triangular structure on D, this is not possible. The reason is that the matrix R in 5.4.16 is entirely made of zeros. Indeed, $\tilde{\Psi}_0 = D$, and since D is lower triangular, $\tilde{\phi}_{0,12} = 0$. In this case, the Waggoner and Zha (1999) distribution is not even defined. Another way to see the issue is to say that whatever the value selected for $\eta_{1,T+2}$, since R is entirely made of zeros, the shock will have no impact on the system:

$$\begin{pmatrix} 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \eta_{1,T+1} \\ \eta_{2,T+1} \\ \eta_{3,T+1} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 - \tilde{y}_{1,T+1} - \tilde{\phi}_{0,11}\eta_{1,T+1} - \tilde{\phi}_{0,13}\eta_{3,T+1} - \underbrace{\tilde{\phi}_{0,13}\eta_{3,T+1}}_{=0} \end{pmatrix} \tag{5.4.17}$$

Then there is no way to select a shock value $\eta_{2,T+1}$ that will satisfy the condition. The impossibility here is due to the fact that the impact $\tilde{\phi}_{0,12}$ on the shock $\eta_{2,T+1}$ used to build the condition is null.

The general conclusion of these three examples is that there is no automatic guaranty for a conditional forecast setting to be identified. In general, a particular setting for conditional forecasts will be identified only if the design of shocks (constructive and non-constructive), blocks, and the structure of are consistent. And since there exists no automatic verification procedure, there is no choice but to verify the consistence of the setting manually before conditional forecasts are estimated.

Such verifications can be a hassle. Fortunately, there exist general conditions to identify settings that will always fail, and settings that will always work. The list is not exhaustive however, and for settings which do not comply with these categories, manual verification is still required. This is true in particular for risks of conflicting blocks, for which preliminary verification remains pretty much the only method available.

Outline of conditions for conditional forecasts genetared by specific shocks Failure conditions: these conditions describe settings that will always result in non-identification of the conditional forecast.

1. There are more variables in a block than shocks generating their conditions. *Example*: block 2 is made of variables 2,3 and 4, and the conditions on these variables are generated by shocks 3 and 4 only (3 variables, only 2 shocks). *Solution*: blocks should have at least as may shocks as variables. More shocks than variables is fine, but less is not possible. *Proof*: see Appendix A.11 for an example.

2. The same shock generate conditions on different blocks. *Example*: block 1 is made of variables 1 and 2 whose conditions are generated by shocks 1 and 2; block 2 is generated by variable 3, generated by shock 2 as well (shock 2 is shared by blocks 1 and 2).
   Solution: different blocks should have different shocks.
   Proof: see Appendix A.11 for an example.

3. For some block, the conditions on the variables are generated only by shocks further away in the ordering than the variables, and is lower triangular. *Example*: variables 1 and 3 constitutes block 1. Their conditions are generated by shock 4 and 5. D is defined by Choleski factorisation. *Solution*: choose either shocks preceding the variables in the ordering, or the variable's own shocks. *Proof*: direct consequence of example 3.

Success condition: satisfying these conditions will ensure that conditional forecasts are well identified, conditional of course on the fact that no failure conditions are met.

1. There is only one block, and for all variables in this block, the conditions are generated by their own shocks (and possibly other shocks); D is lower triangular. *Example*: Block 1 is the only block, made by variables 1, 2 and 3, and the conditions are generated by shocks 1, 2 and 3. D is defined by triangular factorisation. *Proof*: Because there is only block, no conflict between blocks may arise. Because each variable is determined (at least) by its own shock, the impact is given by $\tilde{\phi}_{0,11}$, which is non-zero since D is lower triangular.

2. There are several blocks, but each block is identified by its own shocks. In addition, the order of the blocks is consistent with the order of the variables, and D is lower triangular. *Example*: Block 1 is made of variables 1 and 2, defined by shocks 1 and 2; block 2 is made of variable 4, defined by shock 4. *Counter-example*: Block 1 is made of variable 4, defined by shock 4; block 2 is made of variables 1 and 2, defined by shocks 1 and 2; this will fail. *Proof*: Direct consequence of example 1: conflicts between shocks avoided or not avoided thanks to the lower triangular structure of D.

Note finally that the general conditional forecast framework introduced in subsection 3.3 corresponds to success condition 1 (there is only one block, in which the constraints are generated by all the shocks). Therefore, it is guaranteed to be always well identified. Now that the framework has been exhaustively developed, it is possible to introduce the Gibbs algorithm used for conditional forecasts when conditions are generated by specific shocks:

Algorithm 3.4.2 (conditional forecasts generated by specific shocks, all priors):

1. Define the total number of iterations $It - Bu$ of the algorithm, the forecast horizon h, and the ν conditions: $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_\nu$ to be imposed on the series. For each forecast period, decide which shocks generate which conditions (i.e., define the blocks). Check that conditional forecasts are identified, either manually, or using the conditions above.

2. At iteration n, draw $\beta_{(n)}$, $\Sigma_{(n)}$, D_n, and $\Gamma_{(n)}$ (obtained from $\Sigma_{(n)}$).

3. At iteration n, compute first the unconditional forecasts $\tilde{y}_{T+1}, \tilde{y}_{T+2}, \ldots, \tilde{y}_{T+h}$ by iteratively using 5.3.4, but excluding shocks.

4. For $j = 1, 2 \ldots, h$, compute the impulse response function matrices Ψ_j and $\tilde{\Psi}_j$ from $\beta_{(n)}$, $\Sigma_{(n)}$ and D_n.

5. Run algorithm 3.4.1 to identify the shocks that will satisfy the conditions.

6. Calculate the conditional forecasts, using 5.3.9 with the unconditional forecast values obtained in step 3 and the constrained shocks obtained in step 4.

7. Repeat until $It - Bu$ iterations are realised.

5.5 Relative entropy methodology - Tilting

The conditional forecast methodology developed by Waggoner and Zha (1999) produces what is known as "hard forecasts": the conditions set by the user always hold exactly. This may however not be very realistic: rather than an exact value, the researcher may want to obtain a distribution centered at the condition value, with some variability allowed around this value. This is known as "soft forecast". While Waggoner and Zha (1999) also propose a soft forecast methodology, it is interesting here to go for an alternative approach proposed by Robertson et al. (2005), called the relative entropy approach. This approach has two main advantages over the classical Waggoner and Zha (1999) methodology. First, while the Waggoner and Zha (1999) method only allows the user to set the condition value, or, in other words, the center of the predictive distribution, the relative entropy method allows the user to determine any moment associated with the distribution,

along with quantile values. Second, the central idea of the approach is to obtain a new predictive distribution compliant with the condition that is as close as possible to the initial unconditional forecast distribution. In this respect, the conditional forecasts obtained this way are as consistent as possible with the initial distribution, which may not be the case with the Waggoner and Zha (1999) approach.

The relative entropy methodology is now formally introduced. The presentation starts with the general method, applicable to any random variable. It will then be straightforward to show how the method adapts to the specific case of conditional forecasts. So, consider any random variable y, with an associated density function $f(y)$. Note that this random variable can be multivariate, in which case y will be a vector. If y contains n variables, then:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \tag{5.5.1}$$

The analytical form of the density function $f(y)$ may or may not be known, but it is assumed that it is possible to sample from the distribution numerically, using computer applications. Therefore, it is possible to obtain a series of N draws $\left\{y^{(i)}\right\}_{i=1}^{N}$ on y, together with a series of weights $(\pi_1, \pi_2, ...\pi_N)$ attributed to each draw. Naturally, when N draws are realized from a computer application, the weights are just $1/N$ for all π_i.

The idea underlying the relative entropy method is the following: imagine that there is some additional or new information available, which is not taken into account by the distribution $f(y)$. Ideally, if this information is relevant, one would like to account for it, but of course, that would result in a new distribution, say $f^*(y)$, which would necessarily be different from $f(y)$, as the latter ignores this information. The fundamental problem is: how to recover $f^*(y)$? Robertson et al. (2005) simply notice that if one keeps the same draws $\left\{y^{(i)}\right\}_{i=1}^{N}$, but alters the weights attributed to them, (that is, a new series of weights $(\pi_1^*, \pi_2^*, ...\pi_N^*)$ is now used for the draws), a modified, or "tilted" distribution will be obtained. The idea is then that it is possible to define those new weights $(\pi_1^*, \pi_2^*, ...\pi_N^*)$ so as to obtain a distribution that will be compliant with the new information. An obvious issue is that there may exist several (and potentially, an infinite number of) distributions compliant with this information, and hence several possible sets of new weights $(\pi_1^*, \pi_2^*, ...\pi_N^*)$. However, for consistency reasons, the new distribution should also be designed to be as close as possible to the original distribution. From this principle, as will be detailed hereafter, a unique set of new weights $(\pi_1^*, \pi_2^*, ...\pi_N^*)$ can be obtained.

The technical aspects of the methodology are now considered. Assume thus that some new information is available about the distribution of the random variable y, and that this information can be expressed as the expectation of some function $g(y)$ of the random variable y. Assume also that this expectation is equal to some known value $\bar{g}$. That is, the information takes the form:

$$E\left(g(y)\right) = \int_{-\infty}^{\infty} g(y) f^*(y) dy = \bar{g} \tag{5.5.2}$$

Note that the expectation is taken with respect to $f^*(y)$, the modified density, and not $f(y)$, the original density. In practice, because the analytical form of the density $f^*(y)$ may not be known, one uses the equivalent approximation, based on a sample of draws from $f^*(y)$:

$$E\left(g(y)\right) = \sum_{i=1}^{N} g\left(y^{(i)}\right) \pi_i^* = \bar{g} \tag{5.5.3}$$

While just any function $g(y)$ is permitted, in practice one typically wants to use a function $g(y)$ that defines quantities which are relevant for a distribution such as moments or quantiles. These two cases are now detailed.

Start with moments. To express the information that the r^{th} moment of some $y_j \in y$ is equal to $\bar{x}$, define:

$$g(y) = (y_j)^r \quad \text{and} \quad \bar{g} = \bar{x} \tag{5.5.4}$$

This yields:

$$\int_{-\infty}^{\infty} (y_j)^r f^*(y_j) dy = \bar{x} \tag{5.5.5}$$

which is the definition of the r^{th} moment. In practice, one uses:

$$\sum_{i=1}^{N} \left(y_j^{(i)}\right)^r \pi_i^* = \bar{x} \tag{5.5.6}$$

When r is equal to 1, one simply determines the mean of the distribution, and $g(y)$ is trivially defined as $g(y) = y_j$. Note however that 5.5.3 can only define basic moments: central moments (like the variance) cannot be directly obtained since $g\left(y^{(i)}\right)$ would then require the value $E^*\left(y_j\right)$ which is unknown.

The second principal object of interest for a distribution is quantiles. Defining quantiles can be done easily. To generate the information that the α^{th} quantile of y_j is equal to $\bar{x}$, define:

$$g(y) = 1\left(y_j \ \leq \ \bar{x}\right) \quad \text{and} \quad \bar{g} = \alpha \tag{5.5.7}$$

where $1\left(.\right)$ denotes the indicator function, which takes a value of 1 if the condition holds, and 0 otherwise. Then, 5.5.2 becomes:

$$\int_{-\infty}^{\bar{x}} f^*(y_j) dy_j = \alpha \tag{5.5.8}$$

This says that the distribution function of y_j at $\bar{x}$, $F^*(\bar{x})$, is equal to α , which is equivalent to saying that $\bar{x}$ is the α^{th} quantile of y_j. In practice, one uses:

$$\sum_{y_j^{(i)} \leq \bar{x}} \pi_i^* = \alpha \tag{5.5.9}$$

Note that there can be several elements of information to integrate. If there exists L such pieces of new information to integrate, then 5.5.3 becomes:

$$\begin{aligned} \sum_{i=1}^{N} g_1\left(y^{(i)}\right) \pi_i^* &= \bar{g}_1 \\ \sum_{i=1}^{N} g_2\left(y^{(i)}\right) \pi_i^* &= \bar{g}_2 \\ &\vdots \\ \sum_{i=1}^{N} g_L\left(y^{(i)}\right) \pi_i^* &= \bar{g}_L \end{aligned} \tag{5.5.10}$$

Once the functions $g_1(y), g_2(y), ..., g_L(y)$ and the values $\bar{g}_1, \bar{g}_2, ..., \bar{g}_L$ are defined, the next step in the procedure consists in determining how to derive the set of new weights $(\pi_1^*, \pi_2^*, ...\pi_N^*)$ that will define the new distribution. This is done by noticing that the objective of the new weights is twofold: first, they should result in a distribution satisfying the new information; and second, the new distribution should be as close as possible to the original one. To achieve these objectives, Robertson et al. (2005) use a criterion known as the Kullback-Leibler Information Criterion (or KLIC), which represents a measure of the distance between the initial weights and the new weights. The criterion is defined as:

$$K(\pi, \pi^*) = \sum_{i=1}^{N} \pi_i^* \log\left(\frac{\pi_i^*}{\pi_i}\right) \tag{5.5.11}$$

The greater the value of $K(\pi, \pi^*)$, the greater the overall distance between the series of initial weights π and the series of new weights π^*, and the further the new distribution from the initial one. It is now possible to formulate the problem completely: one wants to find a new set of weights π^* that

will minimize $K(\pi, \pi^*)$ and satisfy the new information. Formally, the program is:

$$\underset{\pi^*}{min} \;\; K(\pi, \pi^*) = \sum_{i=1}^{N} \pi_i^* \log\left(\frac{\pi_i^*}{\pi_i}\right) \tag{5.5.12}$$

subject to the constraints:

$$\pi_i^* \geq 0 \;\; , \forall i = 1, 2, ..., N \tag{5.5.13}$$

$$\sum_{i=1}^{N} \pi_i^* = 1 \tag{5.5.14}$$

and

$$\begin{aligned} \sum_{i=1}^{N} g_1\left(y^{(i)}\right) \pi_i^* &= \bar{g}_1 \\ \sum_{i=1}^{N} g_2\left(y^{(i)}\right) \pi_i^* &= \bar{g}_2 \\ &\vdots \\ \sum_{i=1}^{N} g_L\left(y^{(i)}\right) \pi_i^* &= \bar{g}_L \end{aligned} \tag{5.5.15}$$

The first two conditions are justified by the fact that the weights represent probabilities, while 5.5.15 states that the new distribution must be compliant with the additional information. Using the method of Lagrange, the solution can be written as:

$$\pi_i^* = \frac{\pi_i \exp\left(\lambda' g\left(y^{(i)}\right)\right)}{\sum_{i=1}^{N} \pi_i \exp\left(\lambda' g\left(y^{(i)}\right)\right)} \tag{5.5.16}$$

with

$$\lambda = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_L \end{pmatrix} \quad \text{and} \quad g\left(y^{(i)}\right) = \begin{pmatrix} g_1\left(y^{(i)}\right) \\ g_2\left(y^{(i)}\right) \\ \vdots \\ g_L\left(y^{(i)}\right) \end{pmatrix} \tag{5.5.17}$$

and where λ denotes the $L \times 1$ vector of Lagrange multipliers associated with the constraints in 5.5.15 . It can then be shown that it is possible to obtain the values of λ as the solution of the following minimization problem:

$$\lambda = \arg\min_{\tilde{\lambda}} \sum_{i=1}^{N} \pi_i \exp\left(\tilde{\lambda}, \left[g\left(y^{(i)}\right) - \bar{g}\right]\right) \tag{5.5.18}$$

with

$$\bar{g} = \begin{pmatrix} \bar{g}_1 \\ \bar{g}_2 \\ \vdots \\ \bar{g}_L \end{pmatrix} \tag{5.5.19}$$

Note that for numerical softwares, a reformulation of 5.5.16 and 5.5.18 may prove convenient. Start with 5.5.18 :

$$\begin{aligned}
&\sum_{i=1}^{N} \pi_i \exp\left(\tilde{\lambda}, \left[g\left(y^{(i)}\right) - \bar{g}\right]\right) = \sum_{i=1}^{N} \pi_i \exp\left(\left[g\left(y^{(i)}\right) - \bar{g}\right], \tilde{\lambda}\right) \\
&= \sum_{i=1}^{N} \pi_i \exp\left\{ \begin{bmatrix} g_1\left(y^{(i)}\right) - \bar{g}_1 & g_2\left(y^{(i)}\right) - \bar{g}_2 & \cdots & g_L\left(y^{(i)}\right) - \bar{g}_L \end{bmatrix} \begin{pmatrix} \tilde{\lambda}_1 \\ \tilde{\lambda}_2 \\ \vdots \\ \tilde{\lambda}_L \end{pmatrix} \right\} \\
&= \begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_N \end{pmatrix} \exp. \left\{ \begin{bmatrix} g_1\left(y^{(1)}\right) - \bar{g}_1 & g_2\left(y^{(1)}\right) - \bar{g}_2 & \cdots & g_L\left(y^{(1)}\right) - \bar{g}_L \\ g_1\left(y^{(2)}\right) - \bar{g}_1 & g_2\left(y^{(2)}\right) - \bar{g}_2 & \cdots & g_L\left(y^{(2)}\right) - \bar{g}_L \\ \vdots & \vdots & & \vdots \\ g_1\left(y^{(N)}\right) - \bar{g}_1 & g_2\left(y^{(N)}\right) - \bar{g}_2 & \cdots & g_L\left(y^{(N)}\right) - \bar{g}_L \end{bmatrix} \begin{pmatrix} \tilde{\lambda}_1 \\ \tilde{\lambda}_2 \\ \vdots \\ \tilde{\lambda}_L \end{pmatrix} \right\} \\
&= \pi, \times \exp. \left\{ G\tilde{\lambda} \right\}
\end{aligned} \tag{5.5.20}$$

with:

$$\pi = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_N \end{pmatrix} \quad \text{and} \quad G = \begin{bmatrix} g_1\left(y^{(1)}\right) - \bar{g}_1 & g_2\left(y^{(1)}\right) - \bar{g}_2 & \cdots & g_L\left(y^{(1)}\right) - \bar{g}_L \\ g_1\left(y^{(2)}\right) - \bar{g}_1 & g_2\left(y^{(2)}\right) - \bar{g}_2 & \cdots & g_L\left(y^{(2)}\right) - \bar{g}_L \\ \vdots & \vdots & & \vdots \\ g_1\left(y^{(N)}\right) - \bar{g}_1 & g_2\left(y^{(N)}\right) - \bar{g}_2 & \cdots & g_L\left(y^{(N)}\right) - \bar{g}_L \end{bmatrix} \tag{5.5.21}$$

The notation $''\exp.''$ expresses element-wise exponentiation. Proceed similarly for 5.5.16 :

$$\pi_i^* = \frac{\pi_i \exp\left(\lambda' g\left(y^{(i)}\right)\right)}{\sum_{i=1}^{N} \pi_i \exp\left(\lambda' g\left(y^{(i)}\right)\right)}$$
$$= \frac{\pi_i \exp\left(g\left(y^{(i)}\right)'\lambda\right)}{\sum_{i=1}^{N} \pi_i \exp\left(g(y^{(i)})'\lambda\right)}$$
$$= \left(\begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_N \end{pmatrix} \exp. \left\{ \begin{bmatrix} g_1\left(y^{(1)}\right) & g_2\left(y^{(1)}\right) & \cdots & g_L\left(y^{(1)}\right) \\ g_1\left(y^{(1)}\right) & g_2\left(y^{(1)}\right) & \cdots & g_L\left(y^{(1)}\right) \\ \vdots & \vdots & & \vdots \\ g_1\left(y^{(1)}\right) & g_2\left(y^{(1)}\right) & \cdots & g_L\left(y^{(1)}\right) \end{bmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_L \end{pmatrix} \right\} \right)^{-1}$$
$$\times \pi_i \exp \left\{ \begin{bmatrix} g_1\left(y^{(1)}\right) & g_2\left(y^{(1)}\right) & \cdots & g_L\left(y^{(1)}\right) \end{bmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_L \end{pmatrix} \right\}$$

This implies:

$$\begin{pmatrix} \pi_1^* \\ \pi_2^* \\ \vdots \\ \pi_N^* \end{pmatrix} = \left(\begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_N \end{pmatrix} \exp. \left\{ \begin{bmatrix} g_1\left(y^{(1)}\right) & g_2\left(y^{(1)}\right) & \cdots & g_L\left(y^{(1)}\right) \\ g_1\left(y^{(2)}\right) & g_2\left(y^{(2)}\right) & \cdots & g_L\left(y^{(2)}\right) \\ \vdots & \vdots & & \vdots \\ g_1\left(y^{(N)}\right) & g_2\left(y^{(N)}\right) & \cdots & g_L\left(y^{(N)}\right) \end{bmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_L \end{pmatrix} \right\} \right)^{-1}$$
$$\times \begin{pmatrix} \pi_1 & 0 & \cdots & 0 \\ 0 & \pi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pi_N \end{pmatrix} \times . \left(\exp. \left\{ \begin{bmatrix} g_1\left(y^{(1)}\right) & g_2\left(y^{(1)}\right) & \cdots & g_L\left(y^{(1)}\right) \\ g_1\left(y^{(2)}\right) & g_2\left(y^{(2)}\right) & \cdots & g_L\left(y^{(2)}\right) \\ \vdots & \vdots & & \vdots \\ g_1\left(y^{(N)}\right) & g_2\left(y^{(N)}\right) & \cdots & g_L\left(y^{(N)}\right) \end{bmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_L \end{pmatrix} \right\} \right)$$

or

$$\pi^* = (\pi' \times \exp. \{g\lambda\})^{-1} \times I_\pi \times (\exp. \{g\lambda\}) \tag{5.5.22}$$

with:

$$\pi^* = \begin{pmatrix} \pi_1^* \\ \pi_2^* \\ \vdots \\ \pi_N^* \end{pmatrix}, \; I_\pi = \begin{pmatrix} \pi_1 & 0 & \cdots & 0 \\ 0 & \pi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pi_N \end{pmatrix} \quad \text{and} \quad g = \begin{bmatrix} g_1\left(y^{(1)}\right) & g_2\left(y^{(1)}\right) & \cdots & g_L\left(y^{(1)}\right) \\ g_1\left(y^{(2)}\right) & g_2\left(y^{(2)}\right) & \cdots & g_L\left(y^{(2)}\right) \\ \vdots & \vdots & & \vdots \\ g_1\left(y^{(N)}\right) & g_2\left(y^{(N)}\right) & \cdots & g_L\left(y^{(N)}\right) \end{bmatrix} \tag{5.5.23}$$

The program then reduces to obtaining the vector of Lagrange multipliers λ by minimizing 5.5.20 , and recovering the vector of new weights π^* from 5.5.22 .

Once π^* is obtained, there still remains some work to do. Indeed the new weights are useful to define the updated distribution. But in practical applications, what one wants to obtain is not the new distribution itself, but rather a series of draws from this new distribution. To obtain those draws, one has to sample from $f(y)$ with weights π^* rather than with weights π.

This can be done using the multinomial resampling algorithm of Gordon et al. (1993), which is now introduced. The procedure is simple in essence: it amounts to noticing that the series of updated weights $\pi^* = (\pi_1^*, \pi_2^*, ...\pi_N^*)$ represent probabilities, and can thus be interpreted as the chance of drawing a given value of $\left\{y^{(i)}\right\}_{i=1}^N$ for each draw we want to realize. Let us say then that one wants to obtain N^* draws from the new distribution (setting $N^* > N$ helps to improve the precision of the process). It is then possible to consider the drawing process as a multinomial experiment with N^* independent draws, and probabilities $\pi^* = (\pi_1^*, \pi_2^*, ...\pi_N^*)$ assigned to each draw. The methodology is then straightforward:

Algorithm 3.5.1 (multinomial resampling algorithm):

1. Define the total number of draws N^* to be realized from the new distribution, preferably with $N^* > N$.
2. Realise a draw d from the multinomial distribution: $d = \begin{pmatrix} d_1 & d_2 & \cdots & d_N \end{pmatrix}' \sim mn\left(N^*, \pi_1^*, \pi_2^*, ...\pi_N^*\right)$ This defines N integer values $d_1, d_2, ..., d_N$, such that $d_1 + d_2 + \ldots + d_N = N^*$.
3. Generate the new sample of size N^* from the updated distribution by gathering d_1 copies of $y^{(1)}$, d_2 copies of $y^{(2)}$, $\ldots$, d_N copies of $y^{(N)}$.

It is now possible to present a general algorithm implementing the full procedure.

Algorithm 3.5.2 (distribution tilting for a general random variable):

1. Consider the random variable y, with an associated density function $f(y)$, possibly unknown. Obtain a sample of N draws $\left\{y^{(i)}\right\}_{i=1}^{N}$ on y, together with a series of weights $(\pi_1, \pi_2, ...\pi_N)$ attributed to each draw (typically: $\pi_i = 1/N$ for all i).

2. Define the functions $g_1(y), g_2(y), ..., g_L(y)$ and the values $\bar{g}_1, \bar{g}_2, ..., \bar{g}_L$ in order to integrate the new information in the form of 5.5.10.

3. Build the vector π and the matrix G, defined in 5.5.21. Then, obtain the series of Lagrange multipliers, from the minimization problem 5.5.18, reformulated as 5.5.20:
   $\lambda = \arg\min_{\tilde{\lambda}} \ \pi, \times \exp.\left\{G\tilde{\lambda}\right\}$

4. Build the matrices I_π and g, defined in 5.5.23. Recover the vector of new weights π^* from 5.5.22.

5. Obtain a sample of draws from the updated distribution by applying algorithm 3.5.1.

It is now possible to show how to adapt this procedure to the case of conditional forecasts. Algorithm 2.1.1 in subsubsection 4.1 provides a way to obtain draws from the posterior predictive distribution $f(y_{T+1:T+h}\,|y_T)$, that is, to obtain draws from the posterior distribution of unconditional forecasts. For a model with n endogenous variable, a horizon of h forecast periods, and $It - Bu$ draws obtained from the Gibbs sampler algorithm, each draw $y^{(i)}_{T+1:T+h}$ will be a vector of size $nh \times 1$:

$$y^{(i)}_{T+1:T+h} = \begin{pmatrix} y^{(i)}_{1,T+1} \\ \vdots \\ y^{(i)}_{1,T+h} \\ \\ \vdots \\ \\ y^{(i)}_{n,T+1} \\ \vdots \\ y^{(i)}_{n,T+h} \end{pmatrix} \quad i = 1, 2, \ldots, It - Bu \tag{5.5.24}$$

As usual, each draw from the Gibbs sampler is given a weight $\pi_i = 1/(It - Bu)$, for $i = 1, 2, ..., It - Bu$.

Now assume that one wants to conduct a conditional forecast exercise. In the tilting context, this will amount to use the posterior distribution of unconditional forecasts, and tilt it in order to obtain new distributions compliant with the conditions the user wants to set. The primary object of interest

for conditional forecasts is the median (point estimate) value of the distribution. The second object of interest are the lower and upper bounds of the confidence interval for this distribution: tight bands around the median value reflect high certainty about the condition, while looser bands express low confidence and allows for more variability. All these conditions are easily implemented by the way of quantiles. For instance, assume that one wants to set the conditions that the posterior distribution for the first variable at forecast horizon $T+2$ will take a median value of $\bar{x}_1$, and that the 95% probability bands will take their lower and upper bounds at respectively $\bar{x}_2$ and $\bar{x}_3$. Then, using 5.5.3 and 5.5.9 , this set of conditions can be expressed as:

$$\begin{aligned}
&\sum_{i=1}^{N} g_1\left(y_{1,T+2}^{(i)}\right)\pi_i^* = \bar{g}_1 \Rightarrow \sum^{y_{1,T+2}^{(i)} \leq \bar{x}_1} \pi_i^* = 0.5 \text{ , using } g_1(y_{T+1:T+h}) = 1\left(y_{1,T+2} \leq \bar{x}_1\right) \text{ and } \bar{g}_1 = 0.5 \\
&\sum_{i=1}^{N} g_2\left(y_{1,T+2}^{(i)}\right)\pi_i^* = \bar{g}_2 \Rightarrow \sum^{y_{1,T+2}^{(i)} \leq \bar{x}_2} \pi_i^* = 0.025 \text{ , using } g_2(y_{T+1:T+h}) = 1\left(y_{1,T+2} \leq \bar{x}_2\right) \text{ and } \bar{g}_2 = 0.025 \\
&\sum_{i=1}^{N} g_3\left(y_{1,T+2}^{(i)}\right)\pi_i^* = \bar{g}_2 \Rightarrow \sum^{y_{1,T+2}^{(i)} \leq \bar{x}_3} \pi_i^* = 0.975 \text{ , using } g_3(y_{T+1:T+h}) = 1\left(y_{1,T+2} \leq \bar{x}_3\right) \text{ and } \bar{g}_3 = 0.975
\end{aligned} \tag{5.5.25}$$

The same strategy can then be used for any condition on any variable of the model, at any forecast horizon. The remaining steps in the process consists then in a direct application of the general methodology. It is thus possible to obtain the following algorithm:

Algorithm 3.5.3 (computation of conditional forecasts with relative entropy):

1. Using the Gibbs sampler, obtain a series of $(It - Bu)$ draws from the posterior predictive distribution: $y_{T+1:T+h}^{(1)}, y_{T+1:T+h}^{(2)}, ..., y_{T+1:T+h}^{(It-Bu)}$. Define the weights π_i associated to these draws, typically $\pi_i = 1/(It - Bu)$ for all i .

2. Define the conditions to be set on forecast values. Generate the series of functions $g_1(y_{T+1:T+h}), g_2(y_{T+1:T+$ along with the series of values $\bar{g}_1, \bar{g}_2, ..., \bar{g}_L$, to expresses the conditions in the form of 5.5.25.

3. Build the vector π and the matrix G , defined in 5.5.21. Then, obtain the series of Lagrange multipliers, from the minimization problem5.5.18, reformulated as 5.5.20 :
   $\lambda = \arg\min_{\tilde{\lambda}} \ \pi' \times \exp\left.\cdot\right\{G\tilde{\lambda}\right\}$

4. Build the matrices I_π and g, defined in 5.5.25. Recover the vector of new weights π^* from 5.5.22 .

5. Obtain a sample of draws from the conditional forecast distribution by applying algorithm 3.5.1.

5.6 Mean-adjusted VAR models

The main advantage of Bayesian modelling consists in integrating prior information into the model, allowing the final estimates to reflect (partly) the belief of the researcher about which values the parameters should take. Yet, if Bayesian VAR models traditionally integrate prior information about the dynamic coefficients (see the work of Litterman (1986) and subsubsection 3.3), they remain most of the time uninformative on the deterministic components, in particular constant terms. In typical applications, deterministic terms are given a prior mean of 0 associated with a very large variance in order to obtain a diffuse prior distribution. This way, the posterior values entirely stem from the information contained in the data.

This approach presents two disadvantages, which both result from the fact that the long-run, or steady-state value of a VAR model depends on the deterministic coefficients. The first drawback is that if the researcher actually has some knowledge about the long run values of the model, not integrating these values into the model generates a loss of relevant prior information. The second disadvantage of not integrating this information into the model is that the deterministic coefficients and, hence, the steady-state will be entirely determined by the data. This may result in long term forecast grossly at odds with the prior opinion of the researcher.

To overcome this issue, Villani (2009)proposes a reformulation of the classical Bayesian VAR model allowing for explicit inclusion of prior information about steady-state values. To see this, start from the conventional VAR model 3.1.2:

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + C x_t + \varepsilon_t, \text{ where } t = 1, 2, ..., T \tag{5.6.1}$$

From 4.3.9, this model may rewrite as:

$$y_t = A(L)^{-1} C x_t + \Psi_0 \varepsilon_t + \Psi_1 \varepsilon_{t-1} + \Psi_2 \varepsilon_{t-2} ... \tag{5.6.2}$$

with $A(L) = I - A_1 L - A_2 L^2 \dots A_p L^p$ the matrix lag polynomial representation of 5.6.1. This is the usual VAR model, or VAR model on standard form. Villani (2009) proposes an alternative representation:

$$A(L)(y_t - F x_t) = \varepsilon_t \tag{5.6.3}$$

This representation is known as a VAR model on mean-adjusted form. In this representation,

$A(L)$ is a lag polynomial similar to that in model 5.6.2. The p matrices $A_1, A_2, \ldots, A_p$ are of dimension $n \times n$. In addition, F is a $n \times m$ matrix of coefficients with respect to the m exogenous variables. This structure implies that each equation comprises $k_1 = np$ coefficients to estimate with respect to y_t, and m coefficients with respect to x_t, leaving a total of $q_2 = nk_1 = n^2p$ coefficients to estimate for the full model with respect to y_t, and $q_2 = nm$ coefficients with respect to x_t.

The advantage of model 5.6.3 is clear in terms of long-term or steady-state values: taking expectations on both sides and rearranging, one obtains:

$$E(y_t) = Fx_t \tag{5.6.4}$$

That is, the long-run value of the VAR is simply the deterministic of exogenous component of the model. This representation is particularly convenient when the exogenous components comprise only constant terms, since then Fx_t reduces to a vector of constants, say μ, so that $E(y_t) = \mu$. In other words, the steady-state values for the data are just the constants in μ. This is in contrast with model 5.6.2, where C does not represent the long-term values, and has no direct interpretation. Hence, if the researcher has at his disposal information about the long run values for the model, this information can be integrated directly into the model through the prior mean of μ.

The main inconvenient of model 5.6.3, however, it that it is intractable as it is. While the variable of interest in the model is y_t, 5.6.3 is expressed as a complicated and non-linear function of it, so that y_t cannot be calculated directly. It is thus necessary to convert first the mean-adjusted model into a tractable standard form, which fortunately can be done easily. To do so, start from 5.6.3 and rearrange:

$$\begin{aligned}
&A(L)(y_t - Fx_t) - \varepsilon_t \\
&\Leftrightarrow A(L)y_t - A(L)Fx_t = \varepsilon_t \\
&\Leftrightarrow A(L)y_t = A(L)Fx_t = \varepsilon_t \\
&\Leftrightarrow y_t = A_1y_{t-1} + A_2y_{t-2} + \cdots + A_py_{t-p} + A(L)Fx_t + \varepsilon_t \\
&\Leftrightarrow y_t = A_1y_{t-1} + A_2y_{t-2} + \cdots + A_py_{t-p} + Fx_t - A_1Fx_{t-1} \cdots - A_pFx_{t-p} + \varepsilon_t
\end{aligned} \tag{5.6.5}$$

5.6.5 shows that a mean-adjusted VAR is simply a VAR in standard form integrating additional lagged values of exogenous variables in its deterministic component. In this modified model, there are still k_1 and q_1 parameters to estimate for y_t, but now the model comprises $k_3 = m(p+1)$ coefficients in each equation with respect to x_t, and thus $q_3 = nk_3 = nm(p+1)$ coefficients in total.

Actually, one can go further and convert 5.6.5 into more convenient matrix forms equivalent to 3.1.7 and 3.1.12. First, rewrite 5.6.5 in transpose form:

$$y_t^{'} = y_{t-1}^{'}A_1^{'} + y_{t-2}^{'}A_2^{'} \cdots + y_{t-p}^{'}A_p^{'} + x_t^{'}F^{'} - x_{t-1}^{'}F^{'}A_1^{'} - \cdots - x_{t-p}^{'}F^{'}A_p^{'} + \varepsilon^{'} \tag{5.6.6}$$

Then stack observations:

$$\begin{aligned}\underbrace{\begin{pmatrix} y_1^{'} \\ y_2^{'} \\ \vdots \\ y_T^{'} \end{pmatrix}}_{T\times n} &= \underbrace{\begin{pmatrix} y_0^{'} \\ y_1^{'} \\ \vdots \\ y_{T-1}^{'} \end{pmatrix}}_{T\times n} \underbrace{A_1^{'}}_{n\times n} + \underbrace{\begin{pmatrix} y_{-1}^{'} \\ y_0^{'} \\ \vdots \\ y_{T-2}^{'} \end{pmatrix}}_{T\times n} \underbrace{A_2^{'}}_{n\times n} + \cdots + \underbrace{\begin{pmatrix} y_{1-p}^{'} \\ y_{2-p}^{'} \\ \vdots \\ y_{T-p}^{'} \end{pmatrix}}_{T\times n} \underbrace{A_p^{'}}_{n\times n} \\ &+ \underbrace{\begin{pmatrix} x_1^{'} \\ x_2^{'} \\ \vdots \\ x_T^{'} \end{pmatrix}}_{T\times m} \underbrace{F^{'}}_{m\times n} - \underbrace{\begin{pmatrix} x_0^{'} \\ x_1^{'} \\ \vdots \\ x_{T-1}^{'} \end{pmatrix}}_{T\times m} \underbrace{F^{'}A_1^{'}}_{m\times n} - \cdots - \underbrace{\begin{pmatrix} x_{1-p}^{'} \\ x_{2-p}^{'} \\ \vdots \\ x_{T-p}^{'} \end{pmatrix}}_{T\times m} \underbrace{F^{'}A_p^{'}}_{m\times n} + \underbrace{\begin{pmatrix} \varepsilon_1^{'} \\ \varepsilon_2^{'} \\ \vdots \\ \varepsilon_T^{'} \end{pmatrix}}_{T\times n}\end{aligned} \tag{5.6.7}$$

Gather the regressors into matrices to obtain:

$$\underbrace{\begin{pmatrix} y_1^{'} \\ y_2^{'} \\ \vdots \\ y_T^{'} \end{pmatrix}}_{T\times n} = \underbrace{\begin{pmatrix} y_0^{'} & y_{-1}^{'} & \cdots & y_{1-p}^{'} \\ y_1^{'} & y_0^{'} & \cdots & y_{2-p}^{'} \\ \vdots & \vdots & \ddots & \vdots \\ y_{T-1}^{'} & y_{T-2}^{'} & \cdots & y_{T-p}^{'} \end{pmatrix}}_{T\times k_1} \underbrace{\begin{pmatrix} A_1^{'} \\ A_2^{'} \\ \vdots \\ A_p^{'} \end{pmatrix}}_{k_1\times n} + \underbrace{\begin{pmatrix} x_1^{'} & -x_0^{'} & \cdots & -x_{1-p}^{'} \\ x_2^{'} & -x_1^{'} & \cdots & -x_{2-p}^{'} \\ \vdots & \vdots & \ddots & \vdots \\ x_T^{'} & -x_{T-1}^{'} & \cdots & -x_{T-p}^{'} \end{pmatrix}}_{T\times k_1} \underbrace{\begin{pmatrix} F^{'} \\ F^{'}A_1^{'} \\ \vdots \\ F^{'}A_p^{'} \end{pmatrix}}_{k_3\times n} + \underbrace{\begin{pmatrix} \varepsilon_1^{'} \\ \varepsilon_2^{'} \\ \vdots \\ \varepsilon_T^{'} \end{pmatrix}}_{T\times n} \tag{5.6.8}$$

Or, in compact notation:

$$Y = XB + Z\Delta + \mathcal{E} \tag{5.6.9}$$

with

$$Y = \begin{pmatrix} y_1' \\ y_2' \\ \vdots \\ y_T' \end{pmatrix}, \; X = \begin{pmatrix} y_0' & y_{-1}' & \dots & y_{1-p}' \\ y_1' & y_0' & \dots & y_{2-p}' \\ \vdots & \vdots & \ddots & \vdots \\ y_{T-1}' & y_{T-2}' & \dots & y_{T-p}' \end{pmatrix}, \; Z = \begin{pmatrix} x_1' & -x_0' & \dots & -x_{1-p}' \\ x_2' & -x_1' & \dots & -x_{2-p}' \\ \vdots & \vdots & \ddots & \vdots \\ x_T' & -x_{T-1}' & \dots & -x_{T-p}' \end{pmatrix}$$

$$B = \begin{pmatrix} A_1' \\ A_2' \\ \vdots \\ A_p' \end{pmatrix}, \; \Delta = \begin{pmatrix} F' \\ F'A_1' \\ \vdots \\ F'A_p' \end{pmatrix}, \; \mathcal{E} = \begin{pmatrix} \varepsilon_1' \\ \varepsilon_2' \\ \vdots \\ \varepsilon_T' \end{pmatrix} \tag{5.6.10}$$

Vectorising 5.6.8, the model eventually rewrites:

$$\underbrace{\begin{pmatrix} y_{1,1} \\ \vdots \\ y_{1,T} \\ \vdots \\ y_{n,1} \\ \vdots \\ y_{n,T} \end{pmatrix}}_{nT\times 1} = \underbrace{\begin{pmatrix} y_0' & y_{-1}' & \dots & y_{1-p}' & & 0 & \dots & 0 \\ y_1' & y_0' & \dots & y_{2-p}' & & & \dots & \\ \vdots & \vdots & \ddots & \vdots & & & & \vdots \\ y_{T-1}' & y_{T-2}' & \dots & y_{T-p}' & & 0 & \dots & 0 \\ 0 & \dots & \dots & 0 & y_0' & y_{-1}' & \dots & y_{1-p}' \\ \vdots & \dots & \dots & & y_1' & y_0' & \dots & y_{2-p}' \\ \vdots & & \dots & & \vdots & \vdots & & \vdots \\ 0 & \dots & \dots & 0 & y_{T-1}' & y_{T-2}' & \dots & y_{T-p}' \end{pmatrix}}_{T\times k_1} \underbrace{\begin{pmatrix} A_1^{(1)} \\ \vdots \\ A_p^{(1)} \\ \vdots \\ A_1^{(n)} \\ \vdots \\ A_p^{(n)} \end{pmatrix}}_{k_1\times n}$$

$$\underbrace{\begin{pmatrix} x_0' & x_{-1}' & \dots & x_{1-p}' & & 0 & \dots & 0 \\ x_1' & x_0' & \dots & x_{2-p}' & & & \dots & \\ \vdots & \vdots & \ddots & \vdots & & & & \vdots \\ x_{T-1}' & x_{T-2}' & \dots & x_{T-p}' & & 0 & \dots & 0 \\ 0 & \dots & \dots & 0 & x_0' & y_{-1}' & \dots & x_{1-p}' \\ \vdots & \dots & \dots & & x_1' & x_0' & \dots & x_{2-p}' \\ \vdots & & \dots & & \vdots & \vdots & & \vdots \\ 0 & \dots & \dots & 0 & x_{T-1}' & x_{T-2}' & \dots & x_{T-p}' \end{pmatrix}}_{T\times k_1} \underbrace{\begin{pmatrix} F^{(1)} \\ \vdots \\ (FA_p)^{(1)} \\ \vdots \\ F^{(n)} \\ \vdots \\ (FA_p)^{(n)} \end{pmatrix}}_{q_2\times 1} + \underbrace{\begin{pmatrix} \varepsilon_{1,1} \\ \vdots \\ \varepsilon_{1,T} \\ \vdots \\ \varepsilon_{n,1} \\ \vdots \\ \varepsilon_{n,T} \end{pmatrix}}_{nT\times 1} \tag{5.6.11}$$

And 5.6.11 reformulates compactly as:

$$y = \bar{X}\beta + \bar{Z}\Delta + \varepsilon \tag{5.6.12}$$

with:

$$y = vec(Y),\ \bar{X} = I_n \otimes X,\ \bar{Z} = I_n \otimes Z,\ \beta = vec(B),\ \delta = vec(\Delta),\ \text{and}\ \varepsilon = vec(\mathcal{E}) \tag{5.6.13}$$

In practical applications, it is either the form 5.6.9 or the form 5.6.12 which will be used, rather than 5.6.3. A decomposition which proves useful to recover Δ in 5.6.12 is the following:

$$vec(\Delta') = vec(F\ A_1F\ \ldots\ A_pF) = \begin{pmatrix} I_{nm} \\ I_m \otimes A_1 \\ \vdots \\ I_m \otimes A_p \end{pmatrix} vec(F) = U\psi \tag{5.6.14}$$

with:

$$U = \begin{pmatrix} I_{nm} \\ I_m \otimes A_1 \\ \vdots \\ I_m \otimes A_p \end{pmatrix} \tag{5.6.15}$$

and

$$\psi = vec(F) \tag{5.6.16}$$

and where A.1.5 was used to obtain the third term.

It remains yet to determine how to estimate mean-adjusted model 5.6.3 with Bayesian methods. Villani (2009) only provides derivation in the case of the normal-diffuse prior distribution, so the incoming analysis will be restricted to this case. Note first that there are now three blocks to estimate, and not two anymore: β, defined in 5.6.12, which corresponds to the endogenous variables y_t ; ψ, defined in 5.6.16, which corresponds to the exogenous variables x_t; and Σ , the usual residual variance-covariance matrix.

The prior distributions for these parameters are as follows:

$$\beta \sim \mathcal{N}(\beta_0, \Omega_0) \tag{5.6.17}$$

$$\pi(\Sigma) \propto |\Sigma|^{-(n+1)/2} \tag{5.6.18}$$

$$\psi(\Sigma) \sim \mathcal{N}(\psi_0, \Lambda_0) \tag{5.6.19}$$

In practice, the prior parameters β_0 and Ω_0 are set just as for the Minnesota prior. For ψ, the prior is a bit more complicated. Setting a flat prior on ψ as would be the case for the exogenous variables in a Minnesota scheme is not possible, for two reasons. The first reason is that the very purpose of a mean-adjusted VAR is to explicitly integrate prior information about the exogenous variables into the estimation process. If no such information is available, or is lost anyway in a flat prior, there is no point into using a mean-adjusted model. The second reason is technical: Villani (2009) shows that when an uninformative prior is used for exogenous variables, the Gibbs sampler may behave badly and generate draws for ψ that are remote from the actual posterior. To avoid this, a prior that is at least moderately informative is required. The simplest solution is then for the researcher to specify a (subjective) 95% probability interval for the prior values, and to calculate the prior mean and variance retrospectively from this interval.

Turn now to the derivation of the posterior distribution. Similarly to the normal-diffuse prior, there exist no analytical posterior distributions for β, Σ and ψ. It is however possible to derive conditional posterior distributions, and integrate them to the usual Gibbs sampler process.

To derive the posterior distribution, define first the demeaned data vector $\hat{y}_t$ as :

$$\hat{y}_t = y_t - Fx_t \tag{5.6.20}$$

Then, the mean-adjusted model 5.6.3 may rewrite as:

$$A(L)\hat{y}_t = \varepsilon_t \tag{5.6.21}$$

Villani (2009) then remarks that 5.6.21 is just a standard form VAR for $\hat{y}_t = y_t - Fx_t$. Therefore, conditional on F (that is, on ψ), the conditional posterior distributions for β and Σ are simply those obtained with a normal-diffuse prior. Therefore, the conditional posteriors are similar to 3.6.7 and 3.6.9 and given by:

$$\pi(\beta|\Sigma, \psi, y) \sim \mathcal{N}(\bar{\beta}, \bar{\Omega}) \tag{5.6.22}$$

with

$$\bar{\Omega} = [\Omega_0^{-1} + \Sigma^{-1} \otimes \hat{X}'\hat{X}]^{-1} \tag{5.6.23}$$

and

$$\bar{\beta} = [\Omega_0^{-1}\beta_0 + (\Sigma^{-1} \otimes \hat{X}')\hat{y}] \tag{5.6.24}$$

where $\hat{X}$ and $\hat{y}$ are defined as X and y in 5.6.10 and 5.6.13, using $\hat{y}_t$ rather than y_t.

$$\pi(\Sigma|\beta,\psi,y) \sim \mathcal{IW}(\tilde{S},T) \tag{5.6.25}$$

with

$$\tilde{S} = (\hat{Y} - \hat{X}B)'(\hat{Y} - \hat{X}B) \tag{5.6.26}$$

where $\hat{Y}$ is defined in accordance with 5.6.10, using $\hat{y}_t$ rather than y_t.

Deriving the posterior distribution of ψ is more complex, but Villani (2009) shows (See Appendix A. in Villani (2005) for a complete derivation) that it is given by:

$$\pi(\psi|\beta,\Sigma,y) \sim \mathcal{N}(\bar{\psi},\bar{\Lambda}) \tag{5.6.27}$$

with

$$\bar{\Lambda} = [\Lambda_0^{-1} + U'(Z'Z \otimes \Sigma^{-1})U]^{-1} \tag{5.6.28}$$

and

$$\bar{\psi} = \bar{\Lambda}[\Lambda_0^{-1}\psi_0 + U'vec(\Sigma^{-1}(Y - XB)'Z)] \tag{5.6.29}$$

With these values, it is now possible to introduce the Gibbs algorithm used to estimate a mean-adjusted VAR:

Algorithm 3.5.1 (Gibbs algorithm for mean-adjusted VAR, normal-diffuse prior)

1. Define the number of iterations It of the algorithm, and the burn-in sample Bu .
2. Define initial values $\beta_{(0)}$, $B_{(0)}$ and $\Sigma_{(0)}$ for the algorithm. Obtain the initial value for U from $\beta_{(0)}$.
3. At iteration n , draw $\psi_(n)$ conditional on $\beta_{(n-1)}$ and $\Sigma_{(n-1)}$. Draw $\psi_{(n)}$ from a multivariate normal with mean $\bar{\psi}$ and covariance matrix $\bar{\Lambda}$:

   $\pi(\psi|\beta_{(n-1)}\Sigma_{(n-1)},y) \sim \mathcal{N}(\bar{\psi},\bar{\Lambda})$

   with

   $\bar{\Lambda} = [\Lambda_0^{-1} + U'(Z'Z \otimes \Sigma_{(n-1)}^{-1})U]^{-1}$

   and

   $\bar{\psi} = \bar{\Lambda}[\Lambda_0^{-1}\psi_0 + U'vec(\Sigma_{(n-1)}^{-1}(Y - XB'_{(n-1)})Z)]$

   Reshape $\psi_{(n)}$ to obtain $F_{(n)}$.
4. use $F_{(n)}$ to obtain $\hat{Y}$, $\hat{X}$ and $\hat{y}$.

5. Draw the value $\Sigma_{(n)}$, conditional on $B_{(n-1)}$ and $\psi_{(n)}$. Draw $\Sigma_{(n)}$ from an inverse Wishart distribution with scale matrix $\tilde{S}$ and degrees of freedom T:

   $\pi(\Sigma_{(n)}|B_{(n-1)}, \psi_{(n)}, y) \sim \mathcal{IW}(\tilde{S}, T)$

   with:

   $\tilde{S} = (\hat{Y} - \hat{X}B_{(n-1)})'(\hat{Y} - \hat{X}B_{(n-1)})$

6. Finally, draw $\beta_{(n)}$ conditional on $\Sigma_{(n)}$ and $\psi_{(n)}$, and reshape into $B_{(n)}$. Draw $\beta_{(n)}$ from a multivariate normal with mean $\bar{\beta}$ and covariance matrix $\bar{\Omega}$:

   $\pi(\beta_{(n)}|\Sigma_{,(n)}\, \psi_{(n)}, y) \sim \mathcal{N}(\bar{\beta}, \bar{\Omega})$

   with $\bar{\Omega} = [\Omega_0^{-1} + \Sigma_{(n)}^{-1} \otimes \hat{X}'\hat{X}]^{-1}$

   and

   $\bar{\beta} = [\Omega_0^{-1}\beta_0 + (\Sigma_{(n)}^{-1} \otimes \hat{X}')\hat{y}]$

   Update U from $B_{(n)}$.

7. Repeat until It iterations are realized, then discard the first Bu iterations.

ï»¿

6 Bayesian Panel VARs

6.1 Panel VAR models

VAR models are convenient tools to analyse the economic dynamics of economic entities such as countries, financial markets, trade areas or monetary unions. However, it may sometimes be desirable to push the analysis further and study the dynamic interactions of several entities at a time, rather than limit the analysis to a single entity. For instance, one may want to study the interactions existing between several countries (for instance, several Euro area countries, as they are characterised by the same monetary policy, or several emerging Asian economies if they trade intensively). In this case, the specific class of VAR models constituted by the panel VAR models, which considers the dynamics of several entities considered in parallel, are appropriate. These models are typically richer than simple VAR models because they do not only consider naively the interaction between variables as would a normal VAR model do, but they also add a cross-subsectional structure to the model. This allows to separate components which are common from components which are specific, be it in terms of countries, variables, time periods and so on, and then use this structural information to improve the quality of the estimation.

The terminology of Panel VAR models is now introduced. The approach followed in this subsection and the incoming ones owes much to Canova and Ciccarelli (2013), and additional references can be found in this survey paper. Formally, a panel VAR model comprises N entities or "units", which can be countries, economic sectors or industries, firms, and so on. As for a standard VAR, each unit includes n endogenous variables, and p lags, defined over T periods. Only balanced panels are considered, that is, panels for which the n variables are the same for each units, and defined over the same T time periods. The model also includes m exogenous variables, assumed to be common across units.

In its most general form, the panel VAR model for unit i (with $i = 1, 2, \ldots, N$) writes as:

$$\begin{aligned} y_{i,t} &= \sum_{j=1}^{N}\sum_{k=1}^{p} A_{ij,t}^{k} y_{j,t-k} + C_{i,t}x_t + \varepsilon_{i,t} \\ &= A_{i1,t}^{1} y_{1,t-1} + \ldots + A_{i1,t}^{p} y_{1,t-p} \\ &\quad + A_{i2,t}^{1} y_{2,t-1} + \ldots + A_{i2,t}^{p} y_{2,t-p} \\ &\quad + \ldots \\ &\quad + A_{iN,t}^{1} y_{N,t-1} + \ldots + A_{iN,t}^{p} y_{N,t-p} \\ &\quad + C_{i,t}x_t + \varepsilon_{i,t} \end{aligned} \tag{6.1.1}$$

with:
$$y_{i,t} = \underbrace{\begin{pmatrix} y_{i1,t} \\ y_{i2,t} \\ \vdots \\ y_{in,t} \end{pmatrix}}_{n\times 1} \qquad A^k_{ij,t} = \underbrace{\begin{pmatrix} a^k_{ij,11,t} & a^k_{ij,12,t} & \cdots & a^k_{ij,1n,t} \\ a^k_{ij,21,t} & a^k_{ij,22,t} & \cdots & a^k_{ij,2n,t} \\ \vdots & \vdots & \ddots & \vdots \\ a^k_{ij,n1,t} & a^k_{ij,n2,t} & \cdots & a^k_{ij,nn,t} \end{pmatrix}}_{n\times n}$$

$$C_{i,t} = \underbrace{\begin{pmatrix} c_{i1,1,t} & c_{i1,2,t} & \cdots & c_{i1,m,t} \\ c_{i2,1,t} & c_{i2,2,t} & \cdots & c_{i2,m,t} \\ \vdots & \vdots & \ddots & \vdots \\ c_{in,1,t} & c_{in,2,t} & \cdots & c_{in,m,t} \end{pmatrix}}_{n\times m} \qquad x_t = \underbrace{\begin{pmatrix} x_{1,t} \\ x_{2,t} \\ \vdots \\ x_{m,t} \end{pmatrix}}_{m\times 1} \qquad \varepsilon_{i,t} = \underbrace{\begin{pmatrix} \varepsilon_{i1,t} \\ \varepsilon_{i2,t} \\ \vdots \\ \varepsilon_{in,t} \end{pmatrix}}_{n\times 1} \tag{6.1.2}$$

$y_{i,t}$ denotes a $n \times 1$ vector comprising the n endogenous variables of unit i at time t, while $y_{ij,t}$ is the j^{th} endogenous variables of unit i. $A^k_{ij,t}$ is a $n \times n$ matrix of coefficients providing the response of unit i to the k^{th} lag of unit j at period t. For matrix $A^k_{ij,t}$, the coefficient $a^k_{ij,lm,t}$ gives the response of variable l of unit i to the k^{th} lag of variable m of unit j. x_t is the $m \times 1$ vector of exogenous variables, and $C_{i,t}$ is the $n \times m$ matrix relating the endogenous variables to these exogenous variables. For $C_{i,t}$, the coefficient $c_{ij,l,t}$ gives the response of endogenous variable j of unit i to the l^{th} exogenous variable. Finally, $\varepsilon_{i,t}$ denotes a $n \times 1$ vector of residuals for the variables of unit i, with the following properties:

$$\varepsilon_{i,t} \sim \mathcal{N}(0, \Sigma_{ii,t}) \tag{6.1.3}$$

with:

$$\Sigma_{ii,t} = \mathbb{E}(\varepsilon_{i,t}\varepsilon_{i,t}^{'}) = \mathbb{E}\begin{pmatrix} \varepsilon_{i,1,t} \\ \varepsilon_{i,2,t} \\ \vdots \\ \varepsilon_{i,n,t} \end{pmatrix}\begin{pmatrix} \varepsilon_{i,1,t}^{'} & \varepsilon_{i,2,t}^{'} & \cdots & \varepsilon_{i,n,t}^{'} \end{pmatrix} = \underbrace{\begin{pmatrix} \sigma_{ii,11,t} & \sigma_{ii,12,t} & \cdots & \sigma_{ii,1n,t} \\ \sigma_{ii,21,t} & \sigma_{ii,22,t} & \cdots & \sigma_{ii,2n,t} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{ii,n1,t} & \sigma_{ii,n2,t} & \cdots & \sigma_{ii,nn,t} \end{pmatrix}}_{n\times n} \tag{6.1.4}$$

$\varepsilon_{i,t}$ is assumed to be non-autocorrelated, so that $\mathbb{E}(\varepsilon_{i,t}\varepsilon_{i,t}^{'}) = \Sigma_{ii,t}$, while $\mathbb{E}(\varepsilon_{i,t}\varepsilon_{i,s}^{'}) = 0$ when $t \neq s$. Note that in this general setting the variance-covariance matrix for the VAR residuals is allowed to be period-specific, which implies a general form of heteroskedasticity.

For each variable in unit i, the dynamic equation at period t contains a total of $k = Nnp + m$

coefficients to estimate, implying $q = n(Nnp + m)$ coefficients to estimate for the whole unit. Stacking over the N units, the model reformulates as:

$$\begin{aligned} y_t &= \sum_{k=1}^{p} A_t^k y_{t-k} + C_t x_t + \varepsilon_t \\ &= A_t^1 y_{t-1} + \ldots + A_t^p y_{t-p} + C_t x_t + \varepsilon_t \end{aligned} \tag{6.1.5}$$

or:

$$\begin{aligned} \begin{pmatrix} y_{1,t} \\ y_{2,t} \\ \vdots \\ y_{N,t} \end{pmatrix} &= \begin{pmatrix} A^1_{11,t} & A^1_{12,t} & \cdots & A^1_{1N,t} \\ A^1_{21,t} & A^1_{22,t} & \cdots & A^1_{2N,t} \\ \vdots & \vdots & \ddots & \vdots \\ A^1_{N1,t} & A^1_{N2,t} & \cdots & A^1_{NN,t} \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \\ \vdots \\ y_{N,t-1} \end{pmatrix} + \cdots \\ &+ \begin{pmatrix} A^p_{11,t} & A^p_{12,t} & \cdots & A^p_{1N,t} \\ A^p_{21,t} & A^p_{22,t} & \cdots & A^p_{2N,t} \\ \vdots & \vdots & \ddots & \vdots \\ A^p_{N1,t} & A^p_{N2,t} & \cdots & A^p_{NN,t} \end{pmatrix} \begin{pmatrix} y_{1,t-p} \\ y_{2,t-p} \\ \vdots \\ y_{N,t-p} \end{pmatrix} + \begin{pmatrix} C_{1,t} \\ C_{2,t} \\ \vdots \\ C_{N,t} \end{pmatrix} x_t + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \vdots \\ \varepsilon_{N,t} \end{pmatrix} \end{aligned} \tag{6.1.6}$$

with:

$$y_t = \underbrace{\begin{pmatrix} y_{1,t-p} \\ y_{2,t-p} \\ \vdots \\ y_{N,t-p} \end{pmatrix}}_{Nn \times 1} \quad A_t^k = \underbrace{\begin{pmatrix} A^k_{11,t} & A^k_{12,t} & \cdots & A^k_{1N,t} \\ A^k_{21,t} & A^k_{22,t} & \cdots & A^k_{2N,t} \\ \vdots & \vdots & \ddots & \vdots \\ A^k_{N1,t} & A^k_{N2,t} & \cdots & A^k_{NN,t} \end{pmatrix}}_{Nn \times Nn} \quad C_t = \underbrace{\begin{pmatrix} C_{1,t} \\ C_{2,t} \\ \vdots \\ C_{N,t} \end{pmatrix}}_{Nn \times m} \quad \varepsilon_t = \underbrace{\begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \vdots \\ \varepsilon_{N,t} \end{pmatrix}}_{Nn \times 1} \tag{6.1.7}$$

The vector of residuals ε_t has the following properties:

$$\varepsilon_t \sim \mathcal{N}(0, \Sigma_t) \tag{6.1.8}$$

with:

$$\Sigma_t = \mathbb{E}(\varepsilon_t \varepsilon_t^{'}) = \mathbb{E} \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \vdots \\ \varepsilon_{N,t} \end{pmatrix} \begin{pmatrix} \varepsilon_{1,t}^{'} & \varepsilon_{2,t}^{'} & \cdots & \varepsilon_{N,t}^{'} \end{pmatrix} = \underbrace{\begin{pmatrix} \Sigma_{11,t} & \Sigma_{12,t} & \cdots & \Sigma_{1N,t} \\ \Sigma_{21,t} & \Sigma_{22,t} & \cdots & \Sigma_{2N,t} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{N1,t} & \Sigma_{N2,t} & \cdots & \Sigma_{NN,t} \end{pmatrix}}_{Nn \times Nn} \tag{6.1.9}$$

The assumption of absence of autocorrelation is then extended to the whole model: $\mathbb{E}(\varepsilon_t \varepsilon_t^{'}) = \Sigma_t$ while $\mathbb{E}(\varepsilon_t \varepsilon_s^{'}) = 0$ when $t \neq s$. Formulation (6.1.6) of the model now implies that there are $h = Nq = Nn(Nnp + m)$ coefficients to estimate.

This is the most general form of the panel VAR model. Under this form, it is characterised by four properties:

1. Dynamic interdependencies: the dynamic behaviour of each unit is determined by lagged values of itself, but also by lagged values of all the other endogenous variables of all other units. In other words, $A_{ij,t}^k \neq 0$ when $i \neq j$.

2. Static interdependencies: the $\varepsilon_{i,t}$ are allowed to be correlated across units. That is, in general, $\Sigma_{ij,t} \neq 0$ when $i \neq j$.

3. Cross-subsectional heterogeneity: the VAR coefficients and residual variances are allowed to be unit-specific. In other words, $A_{ik,t}^l \neq A_{jk,t}^l, C_{i,t} \neq C_{j,t}$ and $\Sigma_{ii,t} \neq \Sigma_{jj,t}$ when $i \neq j$.

4. Dynamic heterogeneity: the VAR coefficients and the residual variance-covariance matrix are allowed to be period-specific. In other words, $A_{ij,t}^k \neq A_{ij,s}^k$ and $\Sigma_{ij,t} \neq \Sigma_{ij,s}$ when $t \neq s$.

In practice, this general form may be too complex to produce accurate estimates. As it consumes many degrees of freedom, if one has legitimate reasons to assume that some of the properties will not hold, better estimates can be obtained by relaxing them and opt for less degrees-of-freedom consuming procedures. For instance, if one considers a group of countries that are very homogenous and tend to react in a similar way to structural economic shocks, it may be reasonable to relax property 3. For the sake of convenience, the incoming developments will thus present the models by increasing order of complexity, from the simplest one (all properties relaxed) to the most complex one (all properties satisfied).

Example

Before turning to the different estimation methodologies, it may be useful to provide a simple example. Assume a panel VAR model with 2 units so that $N = 2$, 2 endogenous variables per unit so that $n = 2$, one common exogenous variable so that $m = 1$, and $p = 1$ lag. This implies that for each period, every individual equation comprises $k = Nnp + m = 2 \times 2 \times 1 + 1 = 5$ coefficients to estimate. Each unit then implies $q = n(Nnp + m) = 2 \times (2 \times 2 \times 1 + 1) = 2 \times 5 = 10$ coefficients to estimate, and the full model comprises $h = Nk = 2 \times 10 = 20$ coefficients.

This gives the following model in the form of (6.1.1):

$$\begin{pmatrix} y_{11,t} \\ y_{12,t} \end{pmatrix} = \begin{pmatrix} a^1_{11,11,t} & a^1_{11,12,t} \\ a^1_{12,11,t} & a^1_{12,12,t} \end{pmatrix} \begin{pmatrix} y_{11,t-1} \\ y_{12,t-1} \end{pmatrix} + \begin{pmatrix} a^1_{11,21,t} & a^1_{11,22,t} \\ a^1_{12,21,t} & a^1_{12,22,t} \end{pmatrix} \begin{pmatrix} y_{21,t-1} \\ y_{22,t-1} \end{pmatrix} + \begin{pmatrix} c_{11,1,t} \\ c_{12,1,t} \end{pmatrix} \begin{pmatrix} x_{1,t} \end{pmatrix} + \begin{pmatrix} \varepsilon_{11,t} \\ \varepsilon_{12,t} \end{pmatrix}$$
$$\begin{pmatrix} y_{21,t} \\ y_{22,t} \end{pmatrix} = \begin{pmatrix} a^1_{21,11,t} & a^1_{21,12,t} \\ a^1_{22,11,t} & a^1_{22,12,t} \end{pmatrix} \begin{pmatrix} y_{11,t-1} \\ y_{12,t-1} \end{pmatrix} + \begin{pmatrix} a^1_{21,21,t} & a^1_{21,22,t} \\ a^1_{22,21,t} & a^1_{22,22,t} \end{pmatrix} \begin{pmatrix} y_{21,t-1} \\ y_{22,t-1} \end{pmatrix} + \begin{pmatrix} c_{21,1,t} \\ c_{22,1,t} \end{pmatrix} \begin{pmatrix} x_{1,t} \end{pmatrix} + \begin{pmatrix} \varepsilon_{21,t} \\ \varepsilon_{22,t} \end{pmatrix}$$

And the residuals are characterised by the following variance-covariance matrices:

$$\Sigma_{11,t} = \begin{pmatrix} \sigma_{11,11,t} & \sigma_{11,12,t} \\ \sigma_{11,21,t} & \sigma_{11,22,t} \end{pmatrix} \quad \text{and} \quad \Sigma_{22,t} = \begin{pmatrix} \sigma_{22,11,t} & \sigma_{22,12,t} \\ \sigma_{22,21,t} & \sigma_{22,22,t} \end{pmatrix}$$

The full model, under the form (6.1.5) is given by:

$$\begin{pmatrix} y_{11,t} \\ y_{12,t} \\ y_{21,t} \\ y_{22,t} \end{pmatrix} = \begin{pmatrix} a^1_{11,11,t} & a^1_{11,12,t} & a^1_{11,21,t} & a^1_{11,22,t} \\ a^1_{12,11,t} & a^1_{12,12,t} & a^1_{12,21,t} & a^1_{12,22,t} \\ a^1_{21,11,t} & a^1_{21,12,t} & a^1_{21,21,t} & a^1_{21,22,t} \\ a^1_{22,11,t} & a^1_{22,12,t} & a^1_{22,21,t} & a^1_{22,22,t} \end{pmatrix} \begin{pmatrix} y_{11,t-1} \\ y_{12,t-1} \\ y_{21,t-1} \\ y_{22,t-1} \end{pmatrix} + \begin{pmatrix} c_{11,1,t} \\ c_{12,1,t} \\ c_{21,1,t} \\ c_{22,1,t} \end{pmatrix} \begin{pmatrix} x_{1,t} \end{pmatrix} + \begin{pmatrix} \varepsilon_{11,t} \\ \varepsilon_{12,t} \\ \varepsilon_{21,t} \\ \varepsilon_{22,t} \end{pmatrix}$$

And (6.1.9) yields:

$$\Sigma_t = \begin{pmatrix} \sigma_{11,11,t} & \sigma_{11,12,t} & \sigma_{12,11,t} & \sigma_{12,12,t} \\ \sigma_{11,21,t} & \sigma_{11,22,t} & \sigma_{12,21,t} & \sigma_{12,22,t} \\ \sigma_{21,11,t} & \sigma_{21,12,t} & \sigma_{22,11,t} & \sigma_{22,12,t} \\ \sigma_{21,21,t} & \sigma_{21,22,t} & \sigma_{22,21,t} & \sigma_{22,22,t} \end{pmatrix}$$

6.2 A preliminary OLS model: the mean-group estimator

A standard way to estimate panel VAR models in a non-Bayesian way is to use the so-called mean-group estimator described in Pesaran and Smith (1995). These authors show that in a standard maximum likelihood framework, this estimation technique yields consistent estimates. By contrast,

other classical estimators such as the pooled estimators, aggregate estimators and cross-subsection estimators are either inconsistent (for the pooled and aggregate estimators), or consistent only under certain conditions (for the cross-subsection estimator). This makes the mean-group estimator preferable to most conventional estimators. It is fairly straightforward to integrate the Pesaran and Smith methodology into the general panel VAR framework developed so far.

In the Pesaran and Smith framework, it is assumed that the N units of the model are characterised by heterogenous VAR coefficients, but that these coefficients are random processes sharing a common mean. Therefore, the parameters of interest are the average, or mean effects of the group. If the same assumption is formed about the residual variance-covariance matrix, namely that it is heterogenous across units but is characterised by a common mean, then a single and homogenous VAR model is estimated for all the units. Hence, in this model, the four panel properties are relaxed. Start from the general formulation (6.1.6). Given these assumptions, one obtains:

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \\ \vdots \\ y_{N,t} \end{pmatrix} = \begin{pmatrix} A_1^1 & 0 & \cdots & 0 \\ 0 & A_2^1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_N^1 \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \\ \vdots \\ y_{N,t-1} \end{pmatrix} + \cdots + \begin{pmatrix} A_1^p & 0 & \cdots & 0 \\ 0 & A_2^p & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_N^p \end{pmatrix} \begin{pmatrix} y_{1,t-p} \\ y_{2,t-p} \\ \vdots \\ y_{N,t-p} \end{pmatrix} + \begin{pmatrix} C_{1,t} \\ C_{2,t} \\ \vdots \\ C_{N,t} \end{pmatrix} x_t + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \vdots \\ \varepsilon_{N,t} \end{pmatrix} \tag{6.2.1}$$

and

$$\begin{pmatrix} \Sigma_1 & 0 & \cdots & 0 \\ 0 & \Sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_N \end{pmatrix} \tag{6.2.2}$$

Because each unit only responds to itself, the double subscripts ii in A_{ii}^l and Σ_{ii} can be dropped without ambiguity in favour of A_i^l and Σ_i. Consider individual unit i. From (6.2.1), one obtains:

$$y_{i,t} = A_i^1 y_{i,t-1} + \cdots + A_i^p y_{i,t-p} + C_i x_t + \varepsilon_{i,t} \tag{6.2.3}$$

with:

$$\varepsilon_{i,t} \sim \mathcal{N}(0, \Sigma_i) \tag{6.2.4}$$

Transpose (6.2.3):

$$y_{i,t}^{'} = y_{i,t-1}^{'}(A_i^1)^{'} + \cdots + y_{i,t-p}^{'}(A_i^p)^{'} + x_t^{'}C_i^{'} + \varepsilon_{i,t}^{'} \tag{6.2.5}$$

In compact form:

$$y_{i,t}^{'} = \begin{pmatrix} y_{i,t-1}^{'} & \cdots & y_{i,t-p}^{'} & x_t^{'} \end{pmatrix} \begin{pmatrix} (A_i^1)^{'} \\ \vdots \\ (A_i^p)^{'} \\ C_i^{'} \end{pmatrix} + \varepsilon_{i,t}^{'} \tag{6.2.6}$$

Stack over the T sample periods:

$$\begin{pmatrix} y_{i,1}^{'} \\ y_{i,2}^{'} \\ \vdots \\ y_{i,T}^{'} \end{pmatrix} = \begin{pmatrix} y_{i,0}^{'} & \cdots & y_{i,1-p}^{'} & x_1^{'} \\ y_{i,1}^{'} & \cdots & y_{i,2-p}^{'} & x_2^{'} \\ \vdots & \ddots & \vdots & \vdots \\ y_{i,T-1}^{'} & \cdots & y_{i,T-p}^{'} & x_T^{'} \end{pmatrix} \begin{pmatrix} (A_i^1)^{'} \\ \vdots \\ (A_i^p)^{'} \\ C_i^{'} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i,1}^{'} \\ \varepsilon_{i,2}^{'} \\ \vdots \\ \varepsilon_{i,T}^{'} \end{pmatrix} \tag{6.2.7}$$

or:

$$Y_i = X_i B_i + \mathcal{E}_i \tag{6.2.8}$$

with:

$$Y_i = \underbrace{\begin{pmatrix} y_{i,1}^{'} \\ y_{i,2}^{'} \\ \vdots \\ y_{i,T}^{'} \end{pmatrix}}_{T \times n} \qquad X_i = \underbrace{\begin{pmatrix} y_{i,0}^{'} & \cdots & y_{i,1-p}^{'} & x_1^{'} \\ y_{i,1}^{'} & \cdots & y_{i,2-p}^{'} & x_2^{'} \\ \vdots & \ddots & \vdots & \vdots \\ y_{i,T-1}^{'} & \cdots & y_{i,T-p}^{'} & x_T^{'} \end{pmatrix}}_{T \times (np+m)} \qquad B_i = \underbrace{\begin{pmatrix} (A_i^1)^{'} \\ \vdots \\ (A_i^p)^{'} \\ C_i^{'} \end{pmatrix}}_{(np+m) \times n} \qquad \mathcal{E}_i = \underbrace{\begin{pmatrix} \varepsilon_{i,1}^{'} \\ \varepsilon_{i,2}^{'} \\ \vdots \\ \varepsilon_{i,T}^{'} \end{pmatrix}}_{T \times n} \tag{6.2.9}$$

Using A.1.5 and A.1.9, the model (6.2.8) reformulates in vectorised form as:

$$vec(Y_i) = (I_n \otimes X_i)\, vec(B_i) + vec(\mathcal{E}_i) \tag{6.2.10}$$

or:

$$\underbrace{\begin{pmatrix} y_{i1,1} \\ y_{i1,2} \\ \vdots \\ y_{i1,T} \\ \vdots \\ y_{in,1} \\ y_{in,2} \\ \vdots \\ y_{in,T} \end{pmatrix}}_{nT\times 1} = \underbrace{\begin{pmatrix} y_{i,0}^{'} & \cdots & y_{i,1-p}^{'} & x_1^{'} & & 0 & \cdots & \cdots & 0 \\ y_{i,1}^{'} & \cdots & y_{i,2-p}^{'} & x_2^{'} & & \vdots & \ddots & & \vdots \\ \vdots & \ddots & \vdots & \vdots & & \vdots & & \ddots & \vdots \\ y_{i,T-1}^{'} & \cdots & y_{i,T-p}^{'} & x_T^{'} & & 0 & \cdots & \cdots & 0 \\ & & & & \ddots & & & & \\ 0 & \cdots & \cdots & 0 & & y_{i,0}^{'} & \cdots & y_{i,1-p}^{'} & x_1^{'} \\ \vdots & \ddots & & \vdots & & y_{i,1}^{'} & \cdots & y_{i,2-p}^{'} & x_2^{'} \\ \vdots & & \ddots & \vdots & & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 & & y_{i,T-1}^{'} & \cdots & y_{i,T-p}^{'} & x_T^{'} \end{pmatrix}}_{nT\times n(np+m)} \underbrace{\begin{pmatrix} A_i^{1(1)} \\ \vdots \\ A_i^{p(1)} \\ C_i^{(1)} \\ \vdots \\ A_i^{1(n)} \\ \vdots \\ A_i^{p(n)} \\ C_i^{(n)} \end{pmatrix}}_{n(np+m)\times 1} + \underbrace{\begin{pmatrix} \varepsilon_{i,1,1} \\ \varepsilon_{i,1,2} \\ \vdots \\ \varepsilon_{i,1,T} \\ \vdots \\ \varepsilon_{i,n,1} \\ \varepsilon_{i,n,2} \\ \vdots \\ \varepsilon_{i,n,T} \end{pmatrix}}_{nT\times 1} \tag{6.2.11}$$

where $A_i^{k(j)}$ and $C_i^{(j)}$ respectively denote the transpose of row j of matrix A_i^k and C_i. (6.2.10) makes it clear that each unit comprises a total of $q = n(np + m)$ coefficients to estimate. (6.2.10) rewrites:

$$y_i = \bar{X}_i \beta_i + \varepsilon_i \tag{6.2.12}$$

with:

$$y_i = \underbrace{vec(Y_i)}_{nT\times 1} \qquad \bar{X}_i = \underbrace{(I_n \otimes X_i)}_{nT\times q} \qquad \beta_i = \underbrace{vec(B_i)}_{q\times 1} \qquad \varepsilon_i = \underbrace{vec(\mathcal{E}_i)}_{nT\times 1} \tag{6.2.13}$$

Also, from (6.2.4), it follows that:

$$\varepsilon_i \sim \mathcal{N}(0, \bar{\Sigma}_i) \qquad , \qquad \bar{\Sigma}_i = \underbrace{\Sigma_i \otimes I_T}_{nT\times nT} \tag{6.2.14}$$

Consider the VAR model written in the form (6.2.12). The mean-group estimator model assumes that for each unit i, β_i can be expressed as:

$$\beta_i = b + b_i \tag{6.2.15}$$

with b a $k \times 1$ vector of parameters and $b_i \sim \mathcal{N}(0, \Sigma_b)$. (6.2.15) implies that the coefficients of the VAR in different units will differ, but have similar means and variances. In the Pesaran and Smith approach, the main parameter of interest is the average or mean effect b. To obtain it, Pesaran and Smith propose the following strategy. First, obtain an estimate of β_i for each unit by standard OLS.

That is, obtain:

$$\hat{\beta}_i = \left(\bar{X}_i' \bar{X}_i\right)^{-1} \bar{X}_i' y_i \tag{6.2.16}$$

As for the standard OLS VAR, it may be more efficient to obtain a similar estimator by using (6.2.8) rather than (6.2.12). Then what is estimated is:

$$\hat{B}_i = (X_i' X_i)^{-1} X_i' Y_i \tag{6.2.17}$$

$\hat{\beta}_i$ can then be obtained by vectorising $\hat{B}_i$. Once the estimator $\hat{\beta}_i$ is obtained for all units, the mean-group estimator for b in (6.2.15) is simply obtained from the following formula:

$$\hat{b} = \frac{1}{N} \sum_{i=1}^{N} \hat{\beta}_i \tag{6.2.18}$$

The standard error for the mean-group estimator is then given by:

$$\hat{\Sigma}_b = \frac{1}{N(N-1)} \sum_{i=1}^{N} (\hat{\beta}_i - \hat{b})(\hat{\beta}_i - \hat{b})' \tag{6.2.19}$$

A similar strategy is followed for the mean-group estimate of the residual variance-covariance matrix Σ. For each unit i, an estimate of Σ_i is obtained from:

$$\hat{\Sigma}_i = \frac{1}{T-k-1} \mathcal{E}_i' \mathcal{E}_i \tag{6.2.20}$$

And the mean group estimator then obtains from:

$$\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} \hat{\Sigma}_i \tag{6.2.21}$$

This concludes the subsection dedicated to OLS panel VAR models. The incoming subsections will discuss panel VAR models from a Bayesian perspective only. Once again, treatement is done by increasing order of complexity: from the simplest model (all properties relaxed) to the most complex one (all properties satisfied).

6.3 The simplest case: a pooled estimator (relaxing all the properties)

In the simplest case, all the properties are relaxed. The only panel feature in this model is that the data set as a whole comes from multiple units. In this case, the estimator is simply a pooled estimator. Start from (6.1.6). Relaxing properties 1, 2, 3 and 4, it rewrites:

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \\ \vdots \\ y_{N,t} \end{pmatrix} = \begin{pmatrix} A^1 & 0 & \dots & 0 \\ 0 & A^1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A^1 \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \\ \vdots \\ y_{N,t-1} \end{pmatrix} + \dots + \begin{pmatrix} A^p & 0 & \dots & 0 \\ 0 & A^p & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A^p \end{pmatrix} \begin{pmatrix} y_{1,t-p} \\ y_{2,t-p} \\ \vdots \\ y_{N,t-p} \end{pmatrix} + \begin{pmatrix} C \\ C \\ \vdots \\ C \end{pmatrix} x_t + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \vdots \\ \varepsilon_{N,t} \end{pmatrix} \tag{6.3.1}$$

where both units and time subscripts have been dropped from the $A^k_{ij,t}$ coefficient matrices since the dynamic coefficients are homogenous across units, and coefficients are time-invariant. The zero entries reflect the fact that each unit is determined only by its own variables, and is independent from the other units. Also, relaxing property 2 and 3 implies that:

$$\Sigma_{ii,t} = \mathbb{E}(\varepsilon_{i,t}\varepsilon_{i,t}^{'}) = \Sigma_c \ \forall i, \text{ while } \mathbb{E}(\varepsilon_{i,t}\varepsilon_{j,t}^{'}) = 0, \text{ for } i \neq j \tag{6.3.2}$$

The c subscript in Σ_c emphasises the fact that that the value is both time invariant and common to all units. Then, from (6.1.9):

$$\Sigma_t = \begin{pmatrix} \Sigma_{11_t} & \Sigma_{12_t} & \cdots & \Sigma_{1N_t} \\ \Sigma_{21_t} & \Sigma_{22_t} & \cdots & \Sigma_{2N_t} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{N1_t} & \Sigma_{N2_t} & \cdots & \Sigma_{NN_t} \end{pmatrix} = \begin{pmatrix} \Sigma_c & 0 & \cdots & 0 \\ 0 & \Sigma_c & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_c \end{pmatrix} = I_N \otimes \Sigma_c \tag{6.3.3}$$

Suppressing the unnecessary zero entries in (6.3.1), it reformulates as:

$$\begin{aligned} y_{1,t} &= A^1 y_{1,t-1} + \cdots + A^p y_{1,t-p} + C x_t + \varepsilon_{1,t} \\ y_{2,t} &= A^1 y_{2,t-1} + \cdots + A^p y_{2,t-p} + C x_t + \varepsilon_{2,t} \\ &\vdots \\ y_{N,t} &= A^1 y_{N,t-1} + \cdots + A^p y_{N,t-p} + C x_t + \varepsilon_{N,t} \end{aligned} \tag{6.3.4}$$

Take transposes:

$$\begin{aligned} y_{1,t}^{'} &= y_{1,t-1}^{'}(A^1)^{'} + \cdots + y_{1,t-p}^{'}(A^p)^{'} + x_t^{'} C^{'} + \varepsilon_{1,t}^{'} \\ y_{2,t}^{'} &= y_{2,t-1}^{'}(A^1)^{'} + \cdots + y_{2,t-p}^{'}(A^p)^{'} + x_t^{'} C^{'} + \varepsilon_{2,t}^{'} \\ &\vdots \\ y_{N,t}^{'} &= y_{N,t-1}^{'}(A^1)^{'} + \cdots + y_{N,t-p}^{'}(A^p)^{'} + x_t^{'} C^{'} + \varepsilon_{N,t}^{'} \end{aligned} \tag{6.3.5}$$

Reformulated in compact form:

$$\underbrace{\begin{pmatrix} y_{1,t}^{'} \\ y_{2,t}^{'} \\ \vdots \\ y_{N,t}^{'} \end{pmatrix}}_{N \times n} = \underbrace{\begin{pmatrix} y_{1,t-1}^{'} & \cdots & y_{1,t-p}^{'} & x_t^{'} \\ y_{2,t-1}^{'} & \cdots & y_{2,t-p}^{'} & x_t^{'} \\ \vdots & \ddots & \vdots & \vdots \\ y_{N,t-1}^{'} & \cdots & y_{N,t-p}^{'} & x_t^{'} \end{pmatrix}}_{N \times (np+m)} \underbrace{\begin{pmatrix} (A^1)^{'} \\ \vdots \\ (A^p)^{'} \\ C^{'} \end{pmatrix}}_{(np+m) \times n} + \underbrace{\begin{pmatrix} \varepsilon_{1,t}^{'} \\ \varepsilon_{2,t}^{'} \\ \vdots \\ \varepsilon_{N,t}^{'} \end{pmatrix}}_{N \times n} \tag{6.3.6}$$

or:

$$Y_t = X_t B + \mathcal{E}_t \tag{6.3.7}$$

with:

$$Y_t = \begin{pmatrix} y_{1,t}^{'} \\ y_{2,t}^{'} \\ \vdots \\ y_{N,t}^{'} \end{pmatrix} \quad X_t = \begin{pmatrix} y_{1,t-1}^{'} & \cdots & y_{1,t-p}^{'} & x_t^{'} \\ y_{2,t-1}^{'} & \cdots & y_{2,t-p}^{'} & x_t^{'} \\ \vdots & \ddots & \vdots & \vdots \\ y_{N,t-1}^{'} & \cdots & y_{N,t-p}^{'} & x_t^{'} \end{pmatrix} \quad B = \begin{pmatrix} (A^1)^{'} \\ \vdots \\ (A^p)^{'} \\ C^{'} \end{pmatrix} \quad \mathcal{E}_t = \begin{pmatrix} \varepsilon_{1,t}^{'} \\ \varepsilon_{2,t}^{'} \\ \vdots \\ \varepsilon_{N,t}^{'} \end{pmatrix} \tag{6.3.8}$$

Stacking over the T time periods:

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_T \end{pmatrix}}_{NT\times n} = \underbrace{\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_T \end{pmatrix}}_{NT\times(np+m)} \underbrace{B}_{(np+m)\times n} + \underbrace{\begin{pmatrix} \mathcal{E}_1 \\ \mathcal{E}_2 \\ \vdots \\ \mathcal{E}_T \end{pmatrix}}_{NT\times n} \tag{6.3.9}$$

or:

$$Y = XB + \mathcal{E} \tag{6.3.10}$$

As usual, using A.1.5 , it is possible to reformulate the model in vectorised form as:

$$\underbrace{vec(Y)}_{NnT\times 1} = \underbrace{(I_n \otimes X)}_{NnT\times n(np+m)} \underbrace{vec(B)}_{n(np+m)\times 1} + \underbrace{vec(\mathcal{E})}_{NnT\times 1} \tag{6.3.11}$$

or:

$$y = \bar{X}\beta + \varepsilon \tag{6.3.12}$$

It also follows from (6.3.2) that:

$$\varepsilon \sim \mathcal{N}(0, \bar{\Sigma}), \text{ with } \bar{\Sigma} = \Sigma_c \otimes I_{NT} \tag{6.3.13}$$

This model is smaller in dimension than the general model. As a single VAR is estimated for the whole set of units, each equation in the model only implies $k = np + m$ coefficients to estimate, implying $h = q = n(np+m)$ coefficients to estimate at the unit scale (and thus for the entire model). In this model, there are two object of interest to identify: B, defined in (6.3.7) (or its vectorised equivalent in (6.3.12)), and Σ_c, defined in (6.3.2).

Obtaining a Bayesian estimator for the pooled model

In essence, the model described by (6.3.10) and (6.3.12) is just a conventional VAR model. As such, standard estimation techniques for the derivation of the posterior apply. Here, a traditional normal-Wishart identification strategy is adopted.

Start by the likelihood function. Given (6.3.12) and (6.3.13), it is given by:

$$f(y\left|\bar{\Sigma}\right.) \propto \left|\bar{\Sigma}\right|^{-1/2} \exp\left(-\frac{1}{2}(y-\bar{X}\beta)'\bar{\Sigma}^{-1}(y-\bar{X}\beta)\right) \tag{6.3.14}$$

As for the normal-Wishart, the prior for β is assumed to be multivariate normal:

$\beta \sim \mathcal{N}(\beta_0, \Sigma_c \otimes \Phi_0)$

Φ_0 is defined similarly to 3.4.7, except that the residual variance terms $\sigma_1^2, \sigma_2^2, ..., \sigma_n^2$ are now defined over pooled sample of variables. That is, σ_1^2 is estimated by pooling the samples for variable 1 over units $1, 2, \cdots, N$, and then estimating an autoregressive model over this pooled series. $\sigma_2^2, ..., \sigma_n^2$ are defined similarly. The prior density is given by:

$$\pi(\beta) \propto |\Sigma_c|^{-k/2} \exp\left[-\frac{1}{2}(\beta-\beta_0)'(\Sigma_c \otimes \Phi_0)^{-1}(\beta-\beta_0)\right] \tag{6.3.15}$$

The prior for Σ_c is inverse Wishart:

$\Sigma_c \sim \ IW(S_0, \alpha_0)$

Similarly to Φ_0, S_0 is defined as in 3.4.11, but with residual variance terms $\sigma_1^2, \sigma_2^2, ..., \sigma_n^2$ defined over pooled sample of variables. The prior density is given by:

$$\pi(\Sigma_c) \propto |\Sigma_c|^{-(\alpha_0+n+1)/2} \exp\left[-\frac{1}{2}tr\left\{\Sigma_c^{-1}S_0\right\}\right] \tag{6.3.16}$$

Using Baye's rule 3.2.5, one combines the likelihood function with the prior distributions and rearrange to obtain the posterior distribution:

$$\begin{aligned}\pi(\beta, \Sigma_c\left|y\right.) \propto |\Sigma_c|^{-k/2} \exp\left[-\frac{1}{2}tr\left\{\Sigma_c^{-1}\left[(B-\bar{B})'\bar{\Phi}^{-1}(B-\bar{B})\right]\right\}\right] \\ \times |\Sigma_c|^{-(\bar{\alpha}+n+1)/2} \exp\left[-\frac{1}{2}tr\left\{\Sigma_c^{-1}\bar{S}\right\}\right]\end{aligned} \tag{6.3.17}$$

with:

$$\bar{\Phi} = \left[\Phi_0^{-1} + X'X\right]^{-1} \tag{6.3.18}$$

$$\bar{B} = \bar{\Phi}\left[\Phi_0^{-1}B_0 + X'Y\right] \tag{6.3.19}$$

$$\bar{\alpha} = NT + \alpha_0 \tag{6.3.20}$$

$$\bar{S} = Y'Y + S_0 + B_0'\Phi_0^{-1}B_0 - \bar{B}'\bar{\Phi}^{-1}\bar{B} \tag{6.3.21}$$

Marginalising for β and Σ, one obtains:

$$\pi(\Sigma_c\,|y) \sim \ IW\left(\bar{\alpha}, \bar{S}\right) \tag{6.3.22}$$

and:

$$\pi(B\,|y) \sim \ MT(\bar{B}, \bar{S}, \bar{\Phi}, \tilde{\alpha}) \tag{6.3.23}$$

with:

$$\tilde{\alpha} = \bar{\alpha} - n + 1 = NT + \alpha_0 - n + 1 \tag{6.3.24}$$

6.4 A richer model: the random effect model (introducing cross-subsectional heterogeneity)

With the introduction of cross-subsectional heterogeneity, one now obtains a domestic VAR for each unit. Start from (6.1.5). Relaxing properties 1, 2 and 4, but preserving property 3, one obtains:

$$\begin{aligned}\begin{pmatrix} y_{1,t} \\ y_{2,t} \\ \vdots \\ y_{N,t} \end{pmatrix} &= \begin{pmatrix} A_1^1 & 0 & \cdots & 0 \\ 0 & A_2^1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_N^1 \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \\ \vdots \\ y_{N,t-1} \end{pmatrix} + \cdots \\ &+ \begin{pmatrix} A_1^p & 0 & \cdots & 0 \\ 0 & A_2^p & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_N^p \end{pmatrix} \begin{pmatrix} y_{1,t-p} \\ y_{2,t-p} \\ \vdots \\ y_{N,t-p} \end{pmatrix} + \begin{pmatrix} C_1 \\ C_2 \\ \vdots \\ C_N \end{pmatrix} x_t + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \vdots \\ \varepsilon_{N,t} \end{pmatrix}\end{aligned} \tag{6.4.1}$$

and

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 & \cdots & 0 \\ 0 & \Sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_N \end{pmatrix} \tag{6.4.2}$$

Because each unit only responds to itself, a single subscript i can be used without ambiguity in A_{ii}^l

and Σ_i. Consider individual unit i. From (6.4.1), one obtains:

$$y_{i,t} = A_i^1 y_{i,t-1} + \cdots + A_i^p y_{i,t-p} + C_i x_t + \varepsilon_{i,t} \tag{6.4.3}$$

with:

$$\varepsilon_{i,t} \sim \mathcal{N}(0, \Sigma_i) \tag{6.4.4}$$

(6.4.3) implies that each individual equation comprises $k = np + m$ coefficients to estimate.

Transpose:

$$y_{i,t}^{'} = y_{i,t-1}^{'}(A_i^1)^{'} + \cdots + y_{i,t-p}^{'}(A_i^p)^{'} + x_t^{'}C_i^{'} + \varepsilon_{i,t}^{'} \tag{6.4.5}$$

In compact form:

$$y_{i,t}^{'} = \begin{pmatrix} y_{i,t-1}^{'} & \cdots & y_{i,t-p}^{'} & x_t^{'} \end{pmatrix} \begin{pmatrix} (A_i^1)^{'} \\ \vdots \\ (A_i^p)^{'} \\ C_i^{'} \end{pmatrix} + \varepsilon_{i,t}^{'} \tag{6.4.6}$$

Stack over the T time periods:

$$\begin{pmatrix} y_{i,1}^{'} \\ y_{i,2}^{'} \\ \vdots \\ y_{i,T}^{'} \end{pmatrix} = \begin{pmatrix} y_{i,0}^{'} & \cdots & y_{i,1-p}^{'} & x_0^{'} \\ y_{i,1}^{'} & \cdots & y_{i,2-p}^{'} & x_1^{'} \\ \vdots & \ddots & \vdots & \vdots \\ y_{i,T-1}^{'} & \cdots & y_{i,T-p}^{'} & x_T^{'} \end{pmatrix} \begin{pmatrix} (A_i^1)^{'} \\ \vdots \\ (A_i^p)^{'} \\ C_i^{'} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i,1}^{'} \\ \varepsilon_{i,2}^{'} \\ \vdots \\ \varepsilon_{i,T}^{'} \end{pmatrix} \tag{6.4.7}$$

or:

$$Y_i = X_i B_i + \mathcal{E}_i \tag{6.4.8}$$

with:

$$\underbrace{Y_i = \begin{pmatrix} y_{i,1}^{'} \\ y_{i,2}^{'} \\ \vdots \\ y_{i,T}^{'} \end{pmatrix}}_{T \times n} \quad \underbrace{X_i = \begin{pmatrix} y_{i,0}^{'} & \cdots & y_{i,1-p}^{'} & x_0^{'} \\ y_{i,1}^{'} & \cdots & y_{i,2-p}^{'} & x_1^{'} \\ \vdots & \ddots & \vdots & \vdots \\ y_{i,T-1}^{'} & \cdots & y_{i,T-p}^{'} & x_T^{'} \end{pmatrix}}_{T \times k} \quad \underbrace{B_i = \begin{pmatrix} (A_i^1)^{'} \\ \vdots \\ (A_i^p)^{'} \\ C_i^{'} \end{pmatrix}}_{k \times n} \quad \underbrace{\mathcal{E}_i = \begin{pmatrix} \varepsilon_{i,1}^{'} \\ \varepsilon_{i,2}^{'} \\ \vdots \\ \varepsilon_{i,T}^{'} \end{pmatrix}}_{T \times n} \tag{6.4.9}$$

Using A.1.5 and A.1.9, model (6.4.8) reformulates in vectorised form as:

$$vec(Y_i) = (I_n \otimes X_i)\, vec(B_i) + vec(\mathcal{E}_i) \tag{6.4.10}$$

or:

$$\underbrace{\begin{pmatrix} y_{i1,1} \\ y_{i1,2} \\ \vdots \\ y_{i1,T} \\ \vdots \\ y_{in,1} \\ y_{in,2} \\ \vdots \\ y_{in,T} \end{pmatrix}}_{nT\times 1} = \underbrace{\begin{pmatrix} y_{i,0}^{'} & \cdots & y_{i,1-p}^{'} & x_1^{'} & & 0 & \cdots & \cdots & 0 \\ y_{i,1}^{'} & \cdots & y_{i,2-p}^{'} & x_2^{'} & & \vdots & \ddots & & \vdots \\ \vdots & \ddots & \vdots & \vdots & & \vdots & & \ddots & \vdots \\ y_{i,T-1}^{'} & \cdots & y_{i,T-p}^{'} & x_T^{'} & & 0 & \cdots & \cdots & 0 \\ & & & & \ddots & & & & \\ 0 & \cdots & \cdots & 0 & & y_{i,0}^{'} & \cdots & y_{i,1-p}^{'} & x_1^{'} \\ \vdots & \ddots & & \vdots & & y_{i,1}^{'} & \cdots & y_{i,2-p}^{'} & x_2^{'} \\ \vdots & & \ddots & \vdots & & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 & & y_{i,T-1}^{'} & \cdots & y_{i,T-p}^{'} & x_T^{'} \end{pmatrix}}_{nT\times n(np+m)} \underbrace{\begin{pmatrix} A_i^{1(1)} \\ \vdots \\ A_i^{p(1)} \\ C_i^{(1)} \\ \vdots \\ A_i^{1(n)} \\ \vdots \\ A_i^{p(n)} \\ C_i^{(n)} \end{pmatrix}}_{n(np+m)\times 1} + \underbrace{\begin{pmatrix} \varepsilon_{i1,1} \\ \varepsilon_{i1,2} \\ \vdots \\ \varepsilon_{i1,T} \\ \vdots \\ \varepsilon_{in,1} \\ \varepsilon_{in,2} \\ \vdots \\ \varepsilon_{in,T} \end{pmatrix}}_{nT\times 1} \tag{6.4.11}$$

where $A_i^{k(j)}$ and $C_i^{(j)}$ respectively denote the transpose of row j of matrices A_i^k and C_i. (6.4.11) shows that each unit comprises $q = n(np + m)$ coefficients to estimate, implying h=Nq=Nn(np+m) coefficients to estimate for the whole model. (6.4.11) can reformulate as:

$$y_i = \bar{X}_i \beta_i + \varepsilon_i \tag{6.4.12}$$

with:

$$y_i = \underbrace{vec(Y_i)}_{nT\times 1}, \qquad \bar{X}_i = \underbrace{(I_n \otimes X_i)}_{nT\times q}, \qquad \beta_i = \underbrace{vec(B_i)}_{q\times 1}, \qquad \varepsilon_i = \underbrace{vec(\mathcal{E}_i)}_{nT\times 1} \tag{6.4.13}$$

Also, from (6.4.4), it follows that:

$$\varepsilon_i \sim \mathcal{N}(0, \bar{\Sigma}_i), \qquad \text{with } \bar{\Sigma}_i = \underbrace{\Sigma_i \otimes I_T}_{nT\times nT} \tag{6.4.14}$$

Consider the VAR model written in the form (6.4.12). The random coefficient model assumes that for each unit i, β_i can be expressed as:

$$\beta_i = b + b_i \tag{6.4.15}$$

with b a $k \times 1$ vector of parameters and $b_i \sim \mathcal{N}(0, \Sigma_b)$. It follows immediately that:

$$\beta_i \sim \mathcal{N}(b, \Sigma_b) \tag{6.4.16}$$

(6.4.16) implies that the coefficients of the VAR will differ across units, but are drawn from a distribution with similar mean and variance. From this setting, different identification strategies are

possible, typically treating (6.4.16) as an exchangeable prior in order to derive the posterior distribution. Two of these strategies are now described: the Zellner and Hong (1989) approach, and the hierarchical prior approach developed by Jarocinski (2010b).

6.5 The Zellner and Hong prior

Zellner and Hong (1989) propose a specific prior that results in a posterior distribution combining unit specific and average sample information. To derive the posterior, stack first the model for its N units in order to estimate simultaneously the h coefficients of the model. Define:

$$y = \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}}_{NnT\times 1} \qquad \bar{X} = \underbrace{\begin{pmatrix} \bar{X}_1 & 0 & \cdots & 0 \\ 0 & \bar{X}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \bar{X}_N \end{pmatrix}}_{NnT\times h} \qquad \beta = \underbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_N \end{pmatrix}}_{h\times 1}$$

$$\varepsilon = \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix}}_{NnT\times 1} \qquad \bar{\Sigma} = E\left(\varepsilon\varepsilon'\right) = \underbrace{\begin{pmatrix} \bar{\Sigma}_1 & 0 & \cdots & 0 \\ 0 & \bar{\Sigma}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \bar{\Sigma}_N \end{pmatrix}}_{NnT\times NnT} \tag{6.5.1}$$

and:

$$\bar{b} = \mathbf{1}_N \otimes b = \underbrace{\begin{pmatrix} b \\ b \\ \vdots \\ b \end{pmatrix}}_{h\times 1} \qquad \bar{\Sigma}_b = I_N \otimes \Sigma_b = \underbrace{\begin{pmatrix} \Sigma_b & 0 & \cdots & 0 \\ 0 & \Sigma_b & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \Sigma_b \end{pmatrix}}_{h\times h} \tag{6.5.2}$$

It follows directly from (6.5.1) and (6.5.2) that the model as a whole may rewrite as:

$$y = \bar{X}\beta + \varepsilon \tag{6.5.3}$$

with:

$$\varepsilon \sim \mathcal{N}\left(0, \bar{\Sigma}\right) \tag{6.5.4}$$

Also, assuming independence between the β_i's implies that:

$$\beta \sim N\left(\bar{b}, \bar{\Sigma}_b\right) \tag{6.5.5}$$

Zellner and Hong assume a simple form for Σ_b and the series of Σ_i:

$$\Sigma_b = \lambda_1 \sigma_\varepsilon^2 I_q \tag{6.5.6}$$

and

$$\Sigma_i = \sigma_\varepsilon^2 I_n \qquad \forall i \tag{6.5.7}$$

σ_ε^2 is a residual variance term, assumed to be similar across units and endogenous variables, and λ_1 represents an overall tightness parameter. It then follows from (6.4.14), (6.5.1) and (6.5.2) that:

$$\bar{\Sigma} = \sigma_\varepsilon^2 I_{NnT} \tag{6.5.8}$$

and:

$$\bar{\Sigma}_b = \lambda_1 \sigma_\varepsilon^2 I_h \tag{6.5.9}$$

Zellner and Hong then derive a posterior distribution for the model by adopting a Minnesota framework. That is, they assume a fixed and known value for the residual covariance matrix $\bar{\Sigma}$, obtained directly from (6.5.8). Thus, only β remains to be estimated by the model, and (6.5.5) is used as a prior distribution, conditional on known values for $\bar{b}$ and $\bar{\Sigma}_b$. $\bar{\Sigma}_b$ is defined from (6.5.9). Then, conditional on $\bar{\Sigma}$, the likelihood function for the data is given by:

$$f(y\,|\beta\,) \propto \exp\left[-\frac{1}{2}\left(y - \bar{X}\beta\right)'\bar{\Sigma}^{-1}\left(y - \bar{X}\beta\right)\right] \tag{6.5.10}$$

The prior for β is given by:

$$\pi(\beta) \propto \exp\left(-\frac{1}{2}(\beta - \bar{b})'\bar{\Sigma}_b^{-1}(\beta - \bar{b})\right) \tag{6.5.11}$$

Then, using Bayes rule 3.2.3 and rearranging, one obtains the posterior for β as:

$$\pi(\beta\,|y\,) \propto \exp\left[-\frac{1}{2}(\beta - \bar{\beta})'\bar{\Omega}_b^{-1}(\beta - \bar{\beta})\right] \tag{6.5.12}$$

with:

$$\bar{\Omega}_b = \sigma_\varepsilon^2\left(\lambda_1^{-1} I_h + \bar{X}'\bar{X}\right)^{-1} \tag{6.5.13}$$

and:

$$\bar{\beta} = \left(\lambda_1^{-1} I_h + \bar{X}' \bar{X}\right)^{-1} \left(\bar{X}' y + \lambda_1^{-1} \bar{b}\right) \tag{6.5.14}$$

This is the kernel of a multivariate normal distribution with mean $\bar{\beta}$ and covariance matrix $\bar{\Omega}_b$. The marginal posterior distribution for each β_i can then be obtained from (6.5.1), by marginalising over the (β_i)'s. From A.2.2.4, each β_i then follows a multivariate normal distribution with mean $\bar{\beta}_i$ and covariance matrix $\bar{\Omega}_{bi}$, where $\bar{\beta}_i$ and $\bar{\Omega}_{bi}$ are respectively $q \times 1$ partitions of $\bar{\beta}$ and $q \times q$ partitions of $\bar{\Omega}_b$.

The only remaining question is how to define b and σ_ε^2, which are respectively required to obtain $\bar{b}$ from (6.5.2), and $\bar{\Omega}_b$ from (6.5.13). For b, Zellner and Hong use a pooled estimator, which allows to integrate average sample information into the prior distribution. It is defined as:

$$b = \left(\sum_{i=1}^{N} \bar{X}_i' \bar{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \bar{X}_i' y_i\right) \tag{6.5.15}$$

For σ_ε^2, a simple solution is to substitute $\bar{b}$ (as defined in (6.5.2)) into (6.5.3) to obtain:

$$\varepsilon = y - \bar{X}\bar{b} \tag{6.5.16}$$

σ_ε^2 can then be obtained by computing the variance of ε.

6.6 A hierarchical prior

A more sophisticated alternative to the strategy proposed by Zellner and Hong is to rely on a hierarchical prior identification scheme. The identification methodology proposed in this subsection essentially follows that of Jarocinski (2010b). In the simple approach of Zellner and Hong, the only parameter estimated was β, that is, the set of vectors β_i, $(i = 1, 2, \cdots, N)$. The other underlying parameters, that is, the set of residual covariance matrices Σ_i $(i = 1, 2, \cdots, N)$ and the common mean and covariance of the VAR coefficients b and Σ_b were assumed to be known. In the hierarchical prior identification strategy, the model is made richer by also treating these parameters as random variables and including them in the estimation process. This implies in particular that the series of Σ_i's defined in (6.4.2) is now endogenously estimated by the model. β_i is still characterized by (6.4.16), but the hyperparameters b and Σ_b are now also treated as random variables, with a hyperprior distribution applying to them. Before turning to the specific forms of the likelihood functions and prior distributions, it is useful to derive formally the version of Bayes rule used for this specific

problem. First, for notation convenience, denote respectively the sets of coefficients β_i and Σ_i by β and Σ. That is, define:

$$\beta = \{\beta_1, \beta_2, \cdots, \beta_N\} \qquad \text{and} \qquad \Sigma = \{\Sigma_1, \Sigma_2, ..., \Sigma_N\} \tag{6.6.1}$$

The complete posterior distribution for the model is then:

$$\pi\left(\beta, \Sigma, b, \Sigma_b \,|y\right) \tag{6.6.2}$$

Using an approach similar to that described in subsection 3.2, it is straightforward to show that it is given by:

$$\pi\left(\beta, b, \Sigma_b, \Sigma \,|y\right) \propto \pi\left(y \,|\beta, \Sigma\right) \pi\left(\beta \,|b, \Sigma_b\right) \pi\left(b\right) \pi\left(\Sigma_b\right) \pi\left(\Sigma\right) \tag{6.6.3}$$

In other words, the full posterior distribution is equal to the product of the data likelihood function $\pi\left(y \,|\beta, \Sigma\right)$ with the conditional prior distribution $\pi\left(\beta \,|b, \Sigma_b\right)$ for β and the prior $\pi\left(\Sigma\right)$ for Σ, along with the two hyperpriors $\pi\left(b\right)$ and $\pi\left(\Sigma_b\right)$.

The specific forms selected for the likelihood and the priors are now detailed. Unlike the Zellner and Hong approach, it proves more convenient here not to aggregate the data across units. Start with the likelihood function. Given (6.4.12) and (6.4.14), it obtains as:

$$\pi\left(y \,|\beta, \Sigma\right) \propto \prod_{i=1}^{N} \left|\bar{\Sigma}_i\right|^{-1/2} \exp\left(-\frac{1}{2}\left(y_i - \bar{X}_i\beta_i\right)'\left(\bar{\Sigma}_i\right)^{-1}\left(y_i - \bar{X}_i\beta_i\right)\right) \tag{6.6.4}$$

Following (6.4.16), the vectors of coefficients β_i follow a normal distribution, with common mean b and common variance Σ_b:

$$\beta_i \sim \mathcal{N}\left(b, \Sigma_b\right) \tag{6.6.5}$$

This implies that the prior density for β is given by:

$$\pi\left(\beta \,|b, \Sigma_b\right) \propto \prod_{i=1}^{N} \left|\Sigma_b\right|^{-1/2} \exp\left(-\frac{1}{2}(\beta_i - b)'(\Sigma_b)^{-1}\left(\beta_i - b\right)\right) \tag{6.6.6}$$

For the hyperparameters b and Σ_b, the assumed hyperpriors are the following. For b, the selected form is simply a diffuse (improper) prior:

$$\pi\left(b\right) \propto 1 \tag{6.6.7}$$

For Σ_b, the adopted functional form is designed to replicate the VAR coefficient covariance matrix of the Minnesota prior. It relies on a covariance matrix Ω_b, which is a diagonal matrix of dimension

$q \times q$, defined as follows:

1. For parameters in β relating endogenous variables to their own lags, the variance is given by:

$$\sigma_{a_{ii}}^2 = \left(\frac{1}{l^{\lambda_3}}\right)^2 \tag{6.6.8}$$

2. For parameters in β related to cross-lag coefficients, the variance is given by:

$$\sigma_{a_{ij}}^2 = \left(\frac{\sigma_i^2}{\sigma_j^2}\right)\left(\frac{\lambda_2}{l^{\lambda_3}}\right)^2 \tag{6.6.9}$$

As for the Minnesota prior, σ_i^2 and σ_j^2 represent scaling parameters controlling for the relative coefficient sizes on variables i and j. They are obtained by fitting autoregressive models by OLS for the n endogenous variables of the model, and computing their standard deviations. Because the variance is assumed to be common across units, the autoregressive models are computed by pooling the data of all the units, for each endogenous variable.

3. For exogenous variables (including constants), the variance is given by:

$$\sigma_{c_i}^2 = \sigma_i^2(\lambda_4)^2 \tag{6.6.10}$$

λ_2, λ_3 and λ_4 have an interpretation which is similar to that of the Minnesota prior. Comparing (6.6.8)-(6.6.10) with (3.3.5)-(3.3.7), one can see that the terms are similar, save for the overall tightness parameter λ_1. The full covariance matrix is then defined as:

$$\Sigma_b = (\lambda_1 \otimes I_q)\,\Omega_b \tag{6.6.11}$$

$(\lambda_1 \otimes I_q)$ is a $q \times q$ diagonal matrix with all its diagonal entries being equal to λ_1. This way, Σ_b in (6.6.11) corresponds to the Minnesota prior covariance matrix 3.3.8, except that the value of λ_1 in the present case corresponds to $(\lambda_1)^2$ in the Minnesota prior[6]. Considering Ω_b as fixed and known, but treating λ_1 as a random variable conveniently reduces the determination of the full prior for Σ_b to the determination of the prior for the single parameter λ_1.

[6]It turns out that it is more practical to work with λ_1 rather than $(\lambda_1)^2$ in order to derive the conditional posterior distribution for λ_1. Nevertheless, it is straightforward to establish the equivalence with the actual Minnesota prior: to obtain the λ_1 overall tightness value of the Minnesota, simply take the square root of λ_1 in this model.

Note the implications of the value of λ_1. When $\lambda_1 = 0$, prior variance is null and (6.4.16) implies that all the β_i's will take the identical value b: data is fully pooled and the obtained estimate is simply the pooled estimator. As λ_1 is growing larger, coefficients are allowed to differ more and more across units, and get closer to the respective single unit estimates. In the limit case $\lambda_1 \to \infty$, the prior becomes uninformative on b and no sharing of information is applied between units, so that the coefficients for each unit become their own individual estimates. In-between values for λ_1 imply some degree of information sharing between units, and ideally λ_1 should be designed so as to provide a good balance between individual and pooled estimates. To achieve this result, a traditional choice for the prior distribution of λ_1 is an inverse Gamma distribution with shape $s_0/2$ and scale $v_0/2$:

$$\lambda_1 \sim \ IG\left(s_0/2, v_0/2\right) \tag{6.6.12}$$

This implies:

$$\pi\left(\lambda_1 \left| s_0/2, v_0/2\right.\right) \propto \lambda^{-\frac{s_0}{2}-1} \exp\left(-\frac{v_0}{2\lambda_1}\right) \tag{6.6.13}$$

However, Jarocinski (2010b) and Gelman (2006) show that such a prior can be problematic as the results can be quite sensitive to the choice of values for $s_0/2$ and $v_0/2$. When the number of units is greater than 5, those authors advocate either the use of the uniform uninformative prior $\pi\left(\lambda_1\right) \propto \lambda_1^{-1/2}$, or to make the prior a weakly informative prior by using low values for s_0 and v_0, such as $s_0, v_0 \leq 0.001$. The latter solution will be retained for the present model.

Finally, the prior distribution for Σ_i is simply the classical diffuse prior given by:

$$\pi\left(\Sigma_i\right) \propto \left|\Sigma_i\right|^{-(n+1)/2} \tag{6.6.14}$$

And this implies that the prior full density for Σ is given by:

$$\pi\left(\Sigma\right) \propto \prod_{i=1}^{N} \left|\Sigma_i\right|^{-(n+1)/2} \tag{6.6.15}$$

This concludes the description of the model. The likelihood function is given by (6.6.4), while the prior distributions for the 4 set of parameters of the model (β, b, λ and Σ) are respectively given by (6.6.6), (6.6.7), (6.6.13) and (6.6.15). Substituting for these expressions in (6.6.3), one may obtain the full posterior distribution. This posterior distribution however does not allow for any analytical derivations of the marginal posteriors as the parameters are too much interwoven. One has then to rely on the numerical methods provided by the Gibbs sampler framework. In this respect, it is

necessary to obtain the conditional posterior distributions for each parameter. The simplest way to do so is to start from (6.6.3) and marginalise.

Start with the full conditional distribution for β_i. Note first that the conditional posterior is proportional to the joint posterior (6.6.3). Thus, any term in the product which does not involve β_i can be relegated to the proportionality constant. Hence this produces:

$$\pi\left(\beta_i \left| \beta_{-i}, y, b, \Sigma_b, \Sigma\right.\right) \propto \pi\left(y \left| \beta_i, \Sigma\right.\right) \pi\left(\beta_i \left| b, \Sigma_b\right.\right) \tag{6.6.16}$$

where β_{-i} is used to denote the set of all β coefficients less β_i. Combining (6.6.4) and (6.6.6) in (6.6.16) and rearranging, it is straightforward to show that the posterior for β_i is conditionally multivariate normal:

$$\pi\left(\beta_i \left| \beta_{-i}, y, b, \Sigma_b, \Sigma\right.\right) \sim \mathcal{N}(\bar{\beta}_i, \bar{\Omega}_i) \tag{6.6.17}$$

with:

$$\bar{\Omega}_i = \left[\Sigma_i^{-1} \otimes X_i' X_i + \Sigma_b^{-1}\right]^{-1} \tag{6.6.18}$$

and:

$$\bar{\beta}_i = \bar{\Omega}_i \left[\left(\Sigma_i^{-1} \otimes X_i'\right) y_i + \Sigma_b^{-1} b\right] \tag{6.6.19}$$

Because of conditional independence, it is possible to draw each β_i in turn by sampling from the corresponding conditional posterior.

Turn now to the conditional distribution of b. Starting again from (6.6.3) and relegating any term not involving b to the proportionality constant yields:

$$\pi\left(b \left| y, \beta, \Sigma_b, \Sigma\right.\right) \propto \pi\left(\beta \left| b, \Sigma_b\right.\right) \pi\left(b\right) \tag{6.6.20}$$

Using (6.6.6) and (6.6.7), and rearranging, this yields:

$$\pi\left(b \left| y, \beta, \Sigma_b, \Sigma\right.\right) \propto \exp\left(-\frac{1}{2}(b-\beta_m)' \left(N^{-1}\Sigma_b\right)^{-1} (b-\beta_m)\right) \tag{6.6.21}$$

with $\beta_m = N^{-1}\sum_{i=1}^{N} \beta_i$ denoting the arithmetic mean over the β_i vectors. This is the kernel of a multivariate normal distribution with mean β_m and covariance matrix $N^{-1}\Sigma_b$:

$$\pi\left(b \left| y, \beta, \Sigma_b, \Sigma\right.\right) \sim \mathcal{N}\left(\beta_m, N^{-1}\Sigma_b\right) \tag{6.6.22}$$

Obtain now the conditional posterior for Σ_b. Using once again (6.6.3) and relegating to the normal-

ising constant any term not involving Σ_b, one obtains:

$$\pi\left(\Sigma_b \left| y, \beta, b, \Sigma\right.\right) \propto \pi\left(\beta \left| b, \Sigma_b\right.\right) \pi\left(\Sigma_b\right) \tag{6.6.23}$$

Substituting (6.6.6) and (6.6.13) into (6.6.23) and rearranging, one may eventually obtain:

$$\pi\left(\Sigma_b \left| y, \beta, b, \Sigma\right.\right) \propto \lambda_1^{-\frac{\bar{s}}{2}-1} \exp\left(-\frac{\bar{v}}{2}\frac{1}{\lambda_1}\right) \tag{6.6.24}$$

with:

$$\bar{s} = h + s_0 \tag{6.6.25}$$

and

$$\bar{v} = v_0 + \sum_{i=1}^{N} \left\{\left(\beta_i - b\right)' \Omega_b^{-1} \left(\beta_i - b\right)\right\} \tag{6.6.26}$$

This is the kernel of a inverse Gamma distribution with shape $\frac{\bar{s}}{2}$ and scale $\frac{\bar{v}}{2}$:

$$\pi\left(\Sigma_b \left| y, \beta, b, \Sigma\right.\right) \sim IG\left(\frac{\bar{s}}{2}, \frac{\bar{v}}{2}\right) \tag{6.6.27}$$

Eventually, obtain the conditional posterior distribution for the set of residual covariance matrices $\Sigma = \{\Sigma_1, \Sigma_2, ..., \Sigma_N\}$. Once again relegating to the proportionality constant any term not involving Σ_i in (6.6.3), one obtains:

$$\pi\left(\Sigma_i \left| \Sigma_{-i}, y, \beta, b, \Sigma_b\right.\right) \propto \pi\left(y \left| \beta, \Sigma_i\right.\right) \pi\left(\Sigma_i\right) \tag{6.6.28}$$

Using (6.6.4), (6.6.14) and rearranging, one eventually obtains:

$$\pi\left(\Sigma_i \left| \Sigma_{-i}, y, \beta, b, \Sigma_b\right.\right) \propto \left|\Sigma_i\right|^{-(T+n+1)/2} \exp\left(-\frac{1}{2} tr\left[\Sigma_i^{-1} \tilde{S}_i\right]\right) \tag{6.6.29}$$

with:

$$\tilde{S}_i = \left(Y_i - X_i B_i\right)' \left(Y_i - X_i B_i\right) \tag{6.6.30}$$

This is the kernel of an inverse Wishart distribution with scale $\tilde{S}_i$ and degrees of freedom T:

$$\pi\left(\Sigma_i \left| \Sigma_{-i}, y, \beta, b, \Sigma_b\right.\right) \sim \ IW\left(\tilde{S}_i, T\right) \tag{6.6.31}$$

Because of conditional independence, it is possible to draw each Σ_i in turn by sampling from the corresponding conditional posterior.

With these elements, it is eventually possible to define the Gibbs sampler procedure allowing to derive the posterior distribution for the model:

Algorithm 4.6.1 (Gibbs sampler for the hierarchical prior):

1. Define initial values for β, b , Σ_b and Σ. For β, use OLS estimates: $\beta^{(0)} = \left\{\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_N\right\}$, where $\hat{\beta}_i$ denotes the OLS estimate for β_i. For b, set $b^{(0)} = N^{-1}\sum_{i=1}^{N}\hat{\beta}_i$. For Σ_b, set $\lambda_1^{(0)} = 0.01$, which implies that $\sqrt{\lambda_1^{(0)}} = 0.1$, so that $\Sigma_b^{(0)}$ corresponds to the Ω_0 matrix from the Minnesota prior. Finally, for Σ, use also OLS values: $\Sigma^{(0)} = \left\{\hat{\Sigma}_1, \hat{\Sigma}_2, ..., \hat{\Sigma}_N\right\}$, with $\hat{\Sigma}_i$ defined as in 3.1.10.

2. At iteration n, draw $b^{(n)}$ from a multivariate normal distribution:
   $b^{(n)} \sim \mathcal{N}\left(\beta_m^{(n-1)}, N^{-1}\Sigma_b^{(n-1)}\right)$
   with:
   $\beta_m^{(n)} = N^{-1}\sum_{i=1}^{N}\beta_i^{(n-1)}$

3. At iteration n, draw $\Sigma_b^{(n)}$. To do so, draw $\lambda_1^{(n)}$ from an inverse Gamma distribution:
   $\lambda_1^{(n)} \sim IG\left(\frac{\bar{s}}{2}, \frac{\bar{v}}{2}\right)$
   with:
   $\bar{s} = h + s_0$
   and:
   $\bar{v} = v_0 + \sum_{i=1}^{N}\left\{\left(\beta_i^{(n-1)} - b^{(n)}\right)'\left(\Omega_b^{-1}\right)\left(\beta_i^{(n-1)} - b^{(n)}\right)\right\}$
   Then obtain $\Sigma_b^{(n)}$ from:
   $\Sigma_b^{(n)} = \left(\lambda_1^{(n)} \otimes I_q\right)\Omega_b$

4. At iteration n, draw $\beta^{(n)} = \left\{\beta_1^{(n)}, \beta_2^{(n)}, ..., \beta_N^{(n)}\right\}$ from a multivariate normal distribution:
   $\beta_i^{(n)} \sim \mathcal{N}\left(\bar{\beta}_i, \bar{\Omega}_i\right)$
   with:
   $\bar{\Omega}_i = \left[\left(\Sigma_i^{(n-1)}\right)^{-1} \otimes X_i'X_i + \left(\Sigma_b^{(n)}\right)^{-1}\right]^{-1}$
   and:
   $\bar{\beta}_i = \bar{\Omega}_i\left[\left(\left(\Sigma_i^{(n-1)}\right)^{-1} \otimes X_i'\right)y_i + \left(\Sigma_b^{(n)}\right)^{-1}b^{(n)}\right]$

5. At iteration n, draw $\Sigma^{(n)} = \left\{\Sigma_1^{(n)}, \Sigma_2^{(n)}, ..., \Sigma_N^{(n)}\right\}$ from an inverse Wishart distribution:
   $\Sigma_i^{(n)} \sim IW\left(\tilde{S}_i, T\right)$
   with:
   $\tilde{S}_i = \left(Y_i - X_iB_i^{(n)}\right)'\left(Y_i - X_iB_i^{(n)}\right)$

This concludes the process.

6.7 Reintroducing static and dynamic interdependencies: a structural factor approach

While panel VAR models offer the convenient possibility to share information across units and estimate pooled estimators, limiting the estimation to these features is sub-optimal. Ideally, one may want to estimate not only independent models benefiting from some degree of information sharing, but also to allow for direct dynamic interactions between units. In other words, while cross-subsectional heterogeneity is a nice property, a good panel VAR model should also allow for static and dynamic interdependencies. This is what separates single VAR models estimated with panel data from actual panel VAR models where a single model allowing for cross-unit interactions is estimated. The methodology developed in this subsection essentially follows the factor approach proposed by Canova and Ciccarelli (2006) and Canova and Ciccarelli (2013).

Start from the general formulation (6.1.5). Allowing for cross-subsectional heterogeneity, static interdependency and dynamic interdependency, but ignoring dynamic heterogeneity, one obtains the dynamic equation for the full model at period t as:

$$y_t = A^1 y_{t-1} + \cdots + A^p y_{t-p} + Cx_t + \varepsilon_t \tag{6.7.1}$$

Take transpose:

$$y_t' = y_{t-1}'(A^1)' + \cdots + y_{t-p}'(A^p)' + x_t' C' + \varepsilon_t' \tag{6.7.2}$$

Reformulate in compact form:

$$y_t' = \begin{pmatrix} y_{t-1}' & \cdots & y_{t-p}' & x_t' \end{pmatrix} \begin{pmatrix} (A^1)' \\ \vdots \\ (A^p)' \\ C' \end{pmatrix} + \varepsilon_t' \tag{6.7.3}$$

or:

$$y_t' = X_t B + \varepsilon_t' \tag{6.7.4}$$

with:

$$X_t = \underbrace{\begin{pmatrix} y_{t-1}' & \cdots & y_{t-p}' & x_t' \end{pmatrix}}_{1\times k} \qquad B = \underbrace{\begin{pmatrix} (A^1)' \\ \vdots \\ (A^p)' \\ C' \end{pmatrix}}_{k\times Nn} \tag{6.7.5}$$

Next, obtain a vectorised form by using A.1.5:

$$y_t = (I_{Nn} \otimes X_t)\, vec(B) + \varepsilon_t \tag{6.7.6}$$

or:

$$y_t = \bar{X}_t \beta + \varepsilon_t \tag{6.7.7}$$

with:

$$\bar{X}_t = \underbrace{(I_{Nn} \otimes X_t)}_{Nn\times h} \qquad \beta = \underbrace{vec(B)}_{h\times 1} \tag{6.7.8}$$

Because one allows for static interdependency in the model, the variance-covariance matrix of the residual term ε_t does not have to be block diagonal anymore. In addition, a higher degree of flexibility is permitted by assuming that the error term ε_t follows the following distribution:

$$\varepsilon_t \sim\ N(0,\Sigma) \qquad \Sigma = \sigma\tilde{\Sigma} = \underbrace{\begin{pmatrix} \sigma\underbrace{\tilde{\Sigma}_{11}}_{n\times n} & \sigma\tilde{\Sigma}_{12} & \cdots & \sigma\tilde{\Sigma}_{1N} \\ \sigma\tilde{\Sigma}_{21} & \sigma\tilde{\Sigma}_{22} & \cdots & \sigma\tilde{\Sigma}_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma\tilde{\Sigma}_{N1} & \sigma\tilde{\Sigma}_{N2} & \cdots & \sigma\tilde{\Sigma}_{NN} \end{pmatrix}}_{Nn\times Nn} = \begin{pmatrix} \underbrace{\Sigma_{11}}_{n\times n} & \Sigma_{12} & \cdots & \Sigma_{1N} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{N1} & \Sigma_{N2} & \cdots & \Sigma_{NN} \end{pmatrix} \tag{6.7.9}$$

where σ is a scaling random variable following an inverse Gamma distribution:

$$\sigma \sim\ IG\left(\frac{\alpha_0}{2}, \frac{\delta_0}{2}\right) \tag{6.7.10}$$

Therefore, the error term follows a distribution which is a mixture of normal and inverse Gamma, and this can be shown to be actually a Student-t distribution. This formulation thus allows for fat tail distributions for the error terms, which makes it more flexible than the usual normal assumption.

(6.7.7) takes the form of a standard linear model and could in principle be estimated by any standard OLS or Bayesian methods. However, it suffers from the curse of dimensionality: the number of

coefficients is likely to be larger than the size of the data set, which renders estimation untractable. For instance, a complete data set corresponding to model (6.7.7) would contain NnT elements, while the number of coefficients to estimate is $h = Nn(Np + m)$. Estimation using standard methods is then possible only if $T > Np + m$, which may not be easily satisfied. Even if the condition is satisfied, with only few degrees of freedom left, estimation is likely to be of poor quality.

For this reason, it is necessary to find an alternative approach where the dimensionality of the problem is reduced. Canova and Ciccarelli (2006) propose to simplify the problem by assuming that the h elements of the vector of coefficients β can be expressed as a linear function of a much lower number r of structural factors:

$$\beta = \Xi_1\theta_1 + \Xi_2\theta_2 + ... + \Xi_r\theta_r = \sum_{i=1}^{r} \Xi_i\theta_i \tag{6.7.11}$$

$\theta_1, \theta_2, ..., \theta_r$ are vectors of dimension $d_1 \times 1, d_2 \times 1, \cdots, d_r \times 1$ containing the structural factors, while $\Xi_1, \Xi_2, \cdots, \Xi_r$ are selection matrices of dimension $h \times d_1$, $h \times d_2$, $\cdots$, $h \times d_r$ with all their entries being either 0 or 1 picking the relevant elements in $\theta_1, \theta_2, ..., \theta_r$. (6.7.11) can then be rewritten in compact form. Define:

$$\Xi = \underbrace{\begin{pmatrix} \Xi_1 & \Xi_2 & \cdots & \Xi_r \end{pmatrix}}_{h \times d} \qquad \theta = \underbrace{\begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_r \end{pmatrix}}_{d \times 1} \tag{6.7.12}$$

Then (6.7.11) rewrites:

$$\beta = \Xi\theta \tag{6.7.13}$$

Substitute in (6.7.7) to obtain a reformulated model:

$$\begin{aligned} y_t &= \bar{X}_t\beta + \varepsilon_t \\ &= \bar{X}_t\Xi\theta + \varepsilon_t \\ &= (\bar{X}_t\Xi)\theta + \varepsilon_t \end{aligned}$$

or:

$$y_t = \tilde{X}_t\theta + \varepsilon_t \tag{6.7.14}$$

with:

$$\tilde{X}_t = \underbrace{\bar{X}_t\Xi}_{Nn \times d} \tag{6.7.15}$$

To identify the model, one may then define the following structural factors:

- a factor θ_1 is used to capture a common component. It thus always comprises $d_1 = 1$ coefficient.
- a factor θ_2 is used to capture components which are specific to the unit to which belongs the explained variable. As there are N units in the model, θ_2 comprises $d_2 = N$ coefficients.
- a factor θ_3 is used to capture components which are specific to the explained variable itself. As there are n endogenous variables in the model, θ_3 comprises $d_3 = n$ coefficients.
- a factor θ_4 is used to capture lag-specific components. Each equation includes p lags, but in order to avoid colinearity issues with the common component, θ_4 can comprise $d_4 = p - 1$ coefficients at most (in the incoming applications, the last lag is the one which will be omitted).
- a factor θ_5 finally is used to capture exogenous variables components. As there are m exogenous variables in the model, θ_5 comprises $d_5 = m$ coefficients.

This parsimonious approach then conveniently reduces the number of coefficients to estimate from $h = Nn(Nnp + m)$ for a traditional VAR to $d = d_1 + d_2 + d_3 + d_4 + d_5 = 1 + N + n + (p - 1) + m = N + n + p + m$ with the factor approach. Using the example of a moderate size panel VAR model with $N = 8$ units, $n = 6$ endogenous variables, $p = 5$ lags and $m = 1$ exogenous variable, a standard VAR formulation 3.1.12 would involve $h = Nn(Nnp + m) = 11568$ coefficients to be estimated. On the other hand, the retained factor approach only requires $d = N + n + p + m = 20$ elements to estimate, a much smaller number indeed. Even with additional factors, the number of elements to estimate would remain significantly lower than with a traditional approach.

The remaining question is then how one should define the series of matrices $\Xi_1, \Xi_2, \Xi_3, \Xi_4$ and Ξ_5. The easiest way to do so is probably to use an example. Hence, consider the case of a panel VAR

model with $N = 2$ units, $n = 2$ endogenous variables, $p = 2$ lags and $m = 1$ exogenous variable:

$$\begin{pmatrix} y_{11,t} \\ y_{12,t} \\ y_{21,t} \\ y_{22,t} \end{pmatrix} = \begin{pmatrix} a^1_{11,11} & a^1_{11,12} & a^1_{12,11} & a^1_{12,12} \\ a^1_{11,21} & a^1_{11,22} & a^1_{12,21} & a^1_{12,22} \\ a^1_{21,11} & a^1_{21,12} & a^1_{22,11} & a^1_{22,12} \\ a^1_{21,21} & a^1_{21,22} & a^1_{22,21} & a^1_{22,22} \end{pmatrix} \begin{pmatrix} y_{11,t-1} \\ y_{12,t-1} \\ y_{21,t-1} \\ y_{22,t-1} \end{pmatrix} + \begin{pmatrix} a^2_{11,11} & a^2_{11,12} & a^2_{12,11} & a^2_{12,12} \\ a^1_{11,21} & a^1_{11,22} & a^1_{12,21} & a^1_{12,22} \\ a^1_{21,11} & a^1_{21,12} & a^1_{22,11} & a^1_{22,12} \\ a^1_{21,21} & a^1_{21,22} & a^1_{22,21} & a^1_{22,22} \end{pmatrix} \begin{pmatrix} y_{11,t-2} \\ y_{12,t-2} \\ y_{21,t-2} \\ y_{22,t-2} \end{pmatrix} + \begin{pmatrix} c_{1,11} \\ c_{1,21} \\ c_{2,11} \\ c_{2,21} \end{pmatrix} (x_{1,t}) + \begin{pmatrix} \varepsilon_{11,t} \\ \varepsilon_{12,t} \\ \varepsilon_{21,t} \\ \varepsilon_{22,t} \end{pmatrix} \tag{6.7.16}$$

The single common component factor implies that $\theta_1 = (\theta_{11})$. There are 2 units in this model so that $\theta_2 = \begin{pmatrix} \theta_{21} \\ \theta_{22} \end{pmatrix}$. There are 2 endogenous variables so that $\theta_3 = \begin{pmatrix} \theta_{31} \\ \theta_{32} \end{pmatrix}$, 2 lags so that $\theta_4 = (\theta_{41})$ (the final lag is omitted to avoid colinearity), and one exogenous variable so that $\theta_5 = \begin{pmatrix} \theta_{51} \end{pmatrix}$. Also, the model comprises $h = Nn(Nnp + m) = 36$ coefficients. Following, the dimensions of the selection matrices are: Ξ_1 is $h \times d_1$ or 36×1, Ξ_2 is $h \times d_2$ or 36×2, Ξ_3 is $h \times d_3$ or 36×2, Ξ_4 is $h \times d_4$ or 36×1, and Ξ_5 is $h \times d_5$ or 36×1.

Consider the explained variable $y_{11,t}$. As it belongs to unit 1 and represents endogenous variable 1, a representation consistent with the adopted factor structure would be to express its value as:

$$\begin{aligned} y_{11,t} = {} & (y_{11,t-1} + y_{12,t-1} + y_{21,t-1} + y_{22,t-1} + y_{11,t-2} + y_{12,t-2} + y_{21,t-2} + y_{22,t-2} + x_{1,t})\theta_{11} \\ & + (y_{11,t-1} + y_{12,t-1} + y_{11,t-2} + y_{12,t-2})\theta_{21} \\ & + (y_{11,t-1} + y_{21,t-1} + y_{11,t-2} + y_{21,t-2})\theta_{31} \\ & + (y_{11,t-1} + y_{12,t-1} + y_{21,t-1} + y_{22,t-1})\theta_{41} \\ & + (x_{1,t})\theta_{51} + \varepsilon_{11,t} \end{aligned} \tag{6.7.17}$$

or:

$$y_{11,t} = \mathcal{Z}_{11,t}\theta_{11} + \mathcal{Z}_{21,t}\theta_{21} + \mathcal{Z}_{31,t}\theta_{31} + \mathcal{Z}_{41,t}\theta_{41} + \mathcal{Z}_{51,t}\theta_{51} + \varepsilon_{11,t} \tag{6.7.18}$$

This formulation is motivated as follows:

- $\mathcal{Z}_{11,t} = y_{11,t-1} + y_{12,t-1} + y_{21,t-1} + y_{22,t-1} + y_{11,t-2} + y_{12,t-2} + y_{21,t-2} + y_{22,t-2} + x_{1,t}$ represents the common component of the model. It hence includes all the explanatory variables.

- $\mathcal{Z}_{21,t} = y_{11,t-1} + y_{12,t-1} + y_{11,t-2} + y_{12,t-2}$ represents the component specific to unit 1. It hence includes all the values corresponding to this unit.

- $\mathcal{Z}_{31,t} = y_{11,t-1} + y_{21,t-1} + y_{11,t-2} + y_{21,t-2}$ represents the component specific to variable 1. It hence

includes all the values corresponding to this variable.
- $\mathcal{Z}_{41,t} = y_{11,t-1} + y_{12,t-1} + y_{21,t-1} + y_{22,t-1}$ represents the component specific to lag 1. It hence includes all the values corresponding to this lag.
- $\mathcal{Z}_{51,t} = x_{1,t}$ represents the contribution of the first (and here, only) exogenous variable of the model. It hence includes the value of this variable.

Pursuing the same way with the other equations, one obtains:

$$\begin{pmatrix} y_{11,t} \\ y_{12,t} \\ y_{21,t} \\ y_{22,t} \end{pmatrix} = \begin{pmatrix} \mathcal{Z}_{11,t} \\ \mathcal{Z}_{11,t} \\ \mathcal{Z}_{11,t} \\ \mathcal{Z}_{11,t} \end{pmatrix} (\theta_{11}) + \begin{pmatrix} \mathcal{Z}_{21,t} & 0 \\ \mathcal{Z}_{21,t} & 0 \\ 0 & \mathcal{Z}_{22,t} \\ 0 & \mathcal{Z}_{22,t} \end{pmatrix} \begin{pmatrix} \theta_{21} \\ \theta_{22} \end{pmatrix} + \begin{pmatrix} \mathcal{Z}_{31,t} & 0 \\ 0 & \mathcal{Z}_{32,t} \\ \mathcal{Z}_{31,t} & 0 \\ 0 & \mathcal{Z}_{32,t} \end{pmatrix} \begin{pmatrix} \theta_{31} \\ \theta_{32} \end{pmatrix} + \begin{pmatrix} \mathcal{Z}_{41,t} \\ \mathcal{Z}_{41,t} \\ \mathcal{Z}_{41,t} \\ \mathcal{Z}_{41,t} \end{pmatrix} (\theta_{41}) + \begin{pmatrix} \mathcal{Z}_{51,t} \\ \mathcal{Z}_{51,t} \\ \mathcal{Z}_{51,t} \\ \mathcal{Z}_{51,t} \end{pmatrix} \begin{pmatrix} \theta_{51} \end{pmatrix} + \begin{pmatrix} \varepsilon_{11,t} \\ \varepsilon_{12,t} \\ \varepsilon_{21,t} \\ \varepsilon_{22,t} \end{pmatrix} \tag{6.7.19}$$

with:
$\mathcal{Z}_{22,t} = y_{21,t-1} + y_{22,t-1} + y_{21,t-2} + y_{22,t-2}$
$\mathcal{Z}_{32,t} = y_{12,t-1} + y_{22,t-1} + y_{12,t-2} + y_{22,t-2}$

(6.7.19) can rewrite in compact form as:

$$\begin{pmatrix} y_{11,t} \\ y_{12,t} \\ y_{21,t} \\ y_{22,t} \end{pmatrix} = \begin{pmatrix} \mathcal{Z}_{11,t} & \mathcal{Z}_{21,t} & 0 & \mathcal{Z}_{31,t} & 0 & \mathcal{Z}_{41,t} & \mathcal{Z}_{51,t} \\ \mathcal{Z}_{11,t} & \mathcal{Z}_{21,t} & 0 & 0 & \mathcal{Z}_{32,t} & \mathcal{Z}_{41,t} & \mathcal{Z}_{51,t} \\ \mathcal{Z}_{11,t} & 0 & \mathcal{Z}_{22,t} & \mathcal{Z}_{31,t} & 0 & \mathcal{Z}_{41,t} & \mathcal{Z}_{51,t} \\ \mathcal{Z}_{11,t} & 0 & \mathcal{Z}_{22,t} & 0 & \mathcal{Z}_{32,t} & \mathcal{Z}_{41,t} & \mathcal{Z}_{51,t} \end{pmatrix} \begin{pmatrix} \theta_{11} \\ \theta_{21} \\ \theta_{22} \\ \theta_{31} \\ \theta_{32} \\ \theta_{41} \\ \theta_{51} \end{pmatrix} + \begin{pmatrix} \varepsilon_{11,t} \\ \varepsilon_{12,t} \\ \varepsilon_{21,t} \\ \varepsilon_{22,t} \end{pmatrix} \tag{6.7.20}$$

or:

$$y_t = \mathcal{Z}_t \theta + \varepsilon_t \tag{6.7.21}$$

with:

$$\mathcal{Z}_t = \begin{pmatrix} \mathcal{Z}_{11,t} & \mathcal{Z}_{21,t} & 0 & \mathcal{Z}_{31,t} & 0 & \mathcal{Z}_{41,t} & \mathcal{Z}_{51,t} \\ \mathcal{Z}_{11,t} & \mathcal{Z}_{21,t} & 0 & 0 & \mathcal{Z}_{32,t} & \mathcal{Z}_{41,t} & \mathcal{Z}_{51,t} \\ \mathcal{Z}_{11,t} & 0 & \mathcal{Z}_{22,t} & \mathcal{Z}_{31,t} & 0 & \mathcal{Z}_{41,t} & \mathcal{Z}_{51,t} \\ \mathcal{Z}_{11,t} & 0 & \mathcal{Z}_{22,t} & 0 & \mathcal{Z}_{32,t} & \mathcal{Z}_{41,t} & \mathcal{Z}_{51,t} \end{pmatrix} \tag{6.7.22}$$

Comparing (6.7.14) with (6.7.21), one obtains:

$$\bar{X}_t \Xi = \mathcal{Z}_t \tag{6.7.23}$$

Therefore, the series of matrices Ξ must be defined so that (6.7.23) holds. For the VAR model used as an example, the series of matrices $\Xi_1, \Xi_2, \Xi_3, \Xi_4$ and Ξ_5 must be defined as:

$$\Xi_1 = \begin{pmatrix} \mathbf{1}_9 \\ \mathbf{1}_9 \\ \mathbf{1}_9 \\ \mathbf{1}_9 \end{pmatrix} \quad \Xi_2 = \begin{pmatrix} v_1 & \mathbf{0}_9 \\ v_1 & \mathbf{0}_9 \\ \mathbf{0}_9 & v_2 \\ \mathbf{0}_9 & v_2 \end{pmatrix} \quad \Xi_3 = \begin{pmatrix} v_3 & \mathbf{0}_9 \\ \mathbf{0}_9 & v_4 \\ v_3 & \mathbf{0}_9 \\ \mathbf{0}_9 & v_4 \end{pmatrix} \quad \Xi_4 = \begin{pmatrix} v_5 \\ v_5 \\ v_5 \\ v_5 \end{pmatrix} \quad \Xi_5 = \begin{pmatrix} v_6 \\ v_6 \\ v_6 \\ v_6 \end{pmatrix} \tag{6.7.24}$$

Where $\mathbf{1}_n$ and $\mathbf{0}_n$ respectively denote a $n \times 1$ vector of ones and zeros, and v_1, v_2, v_3 and v_4 are defined as:

$$v_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad v_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix} \quad v_3 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \quad v_4 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \quad v_5 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad v_6 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \tag{6.7.25}$$

Hence:

$$\Xi = \begin{pmatrix} \mathbf{1}_9 & v_1 & \mathbf{0}_9 & v_3 & \mathbf{0}_9 & v_5 & v_6 \\ \mathbf{1}_9 & v_1 & \mathbf{0}_9 & \mathbf{0}_9 & v_4 & v_5 & v_6 \\ \mathbf{1}_9 & \mathbf{0}_9 & v_2 & v_3 & \mathbf{0}_9 & v_5 & v_6 \\ \mathbf{1}_9 & \mathbf{0}_9 & v_2 & \mathbf{0}_9 & v_4 & v_5 & v_6 \end{pmatrix} \tag{6.7.26}$$

Using (6.7.8), It can be readily verified that:

$$
\begin{aligned}
&\bar{X}_t \Xi \\
&= (I_4 \otimes X_t)\, \Xi \\
&= \left(I_4 \otimes \begin{pmatrix} y_{11,t-1} & y_{12,t-1} & y_{21,t-1} & y_{22,t-1} & y_{11,t-2} & y_{12,t-2} & y_{21,t-2} & y_{22,t-2} & x_{1,t} \end{pmatrix}\right) \\
&\quad \times \begin{pmatrix} \mathbf{1}_9 & v_1 & \mathbf{0}_9 & v_3 & \mathbf{0}_9 & v_5 & v_6 \\ \mathbf{1}_9 & v_1 & \mathbf{0}_9 & \mathbf{0}_9 & v_4 & v_5 & v_6 \\ \mathbf{1}_9 & \mathbf{0}_9 & v_2 & v_3 & \mathbf{0}_9 & v_5 & v_6 \\ \mathbf{1}_9 & \mathbf{0}_9 & v_2 & \mathbf{0}_9 & v_4 & v_5 & v_6 \end{pmatrix} \\
&= \begin{pmatrix} \mathcal{Z}_{11,t} & \mathcal{Z}_{21,t} & 0 & \mathcal{Z}_{31,t} & 0 & \mathcal{Z}_{41,t} & \mathcal{Z}_{51,t} \\ \mathcal{Z}_{11,t} & \mathcal{Z}_{21,t} & 0 & 0 & \mathcal{Z}_{32,t} & \mathcal{Z}_{41,t} & \mathcal{Z}_{51,t} \\ \mathcal{Z}_{11,t} & 0 & \mathcal{Z}_{22,t} & \mathcal{Z}_{31,t} & 0 & \mathcal{Z}_{41,t} & \mathcal{Z}_{51,t} \\ \mathcal{Z}_{11,t} & 0 & \mathcal{Z}_{22,t} & 0 & \mathcal{Z}_{32,t} & \mathcal{Z}_{41,t} & \mathcal{Z}_{51,t} \end{pmatrix} \\
&= \mathcal{Z}_t
\end{aligned}
\tag{6.7.27}
$$

The procedure can then be extended to any VAR model with an arbitrary number of units, lags, endogenous and exogenous variables.

This concludes the description of the factor approach. Because $\tilde{X}_t$ can be computed for each period t once Ξ is defined, it is possible to stack (6.7.14) over the T periods and estimate the model directly by OLS methods. From a Bayesian perspective, the objective is to recover the posterior distribution for the three parameters of interest: $\theta, \tilde{\Sigma}$ and σ. Once this is done, it is possible to draw values for θ, and thus to recover draws for β from (6.7.13). Also, combining a draw for $\tilde{\Sigma}$ with a draw for σ, one may recover a draw for Σ. This allows to recover draws from the original model (6.7.7).

To compute the posterior distribution, obtain first an expression for Bayes rule. Relying on a standard independence assumption between $\theta, \tilde{\Sigma}$ and σ, 3.2.5 yields:

$$
\pi(\theta, \tilde{\Sigma}, \sigma \,|y) \propto f(y \,|\theta, \sigma, \tilde{\Sigma}) \pi(\theta) \pi(\tilde{\Sigma}) \pi(\sigma) \tag{6.7.28}
$$

This is a classical Bayes rule, stating that the posterior distribution is obtained by combining the data likelihood $f(y \,|\theta, \sigma, \tilde{\Sigma})$ with the respective prior distributions for θ, σ and $\tilde{\Sigma}$, respectively given by $\pi(\theta)\pi(\sigma)$ and $\pi(\tilde{\Sigma})$. Start with the likelihood function. Given (6.7.9), it is given by:

$$
f(y \,|\theta, \tilde{\Sigma}, \sigma) \propto (\sigma)^{-TNn/2} \left|\tilde{\Sigma}\right|^{-T/2} \prod_{t=1}^{T} \left\{ \exp\left(-\frac{1}{2} \sigma^{-1} (y_t - \tilde{X}_t \theta)' \tilde{\Sigma}^{-1} (y_t - \tilde{X}_t \theta) \right) \right\} \tag{6.7.29}
$$

Consider next the prior distributions for $\theta, \tilde{\Sigma}$ and σ. The prior for θ is a multivariate normal distribution with mean θ_0 and covariance Θ_0:

$$\pi(\theta \,|\theta_0, \Theta_0) \propto \exp\left(-\frac{1}{2}(\theta-\theta_0)'\Theta_0^{-1}(\theta-\theta_0)\right) \tag{6.7.30}$$

Because there is no obvious identification strategies for θ_0 and Θ_0, one simply sets θ_0 as a vector of zeros while Θ_0 is set as a diagonal matrix with large values to produce an uninformative prior. For $\tilde{\Sigma}$, an uninformative (diffuse) prior is used:

$$\pi(\tilde{\Sigma}) \propto \left|\tilde{\Sigma}\right|^{-(Nn+1)/2} \tag{6.7.31}$$

Finally, as already stated,σ follows an inverse Gamma distribution with shape $\frac{\alpha_0}{2}$ and scale $\frac{\delta_0}{2}$:

$$\pi(\sigma) \propto \sigma^{-\frac{\alpha_0}{2}-1} \exp\left(\frac{-\delta_0}{2\sigma}\right) \tag{6.7.32}$$

Using Bayes rule (6.7.28), and combining the likelihood (6.7.29) with the priors (6.7.30), (6.7.31) and (6.7.32), the joint posterior obtains as:

$$\begin{aligned} f(\theta, \tilde{\Sigma}, \sigma \,|y) \propto \prod_{t=1}^{T}\left\{\exp\left(-\frac{1}{2}\sigma^{-1}(y_t-\tilde{X}_t\theta)'\tilde{\Sigma}^{-1}(y_t-\tilde{X}_t\theta)\right)\right\} \times \exp\left(\frac{-\delta_0}{2\sigma}\right) \\ \times(\sigma)^{-(NnT+\alpha_0)/2-1} \times \left|\tilde{\Sigma}\right|^{-(T+Nn+1)/2} \times \exp\left(-\frac{1}{2}(\theta-\theta_0)'\Theta_0^{-1}(\theta-\theta_0)\right) \end{aligned} \tag{6.7.33}$$

As often, the parameters are so interwoven that it is not possible to integrate out the posterior distributions analytically. One has then to turn to numerical methods. Obtain first the conditional posterior distributions for $\theta, \tilde{\Sigma}$ and σ. For the conditional posterior of θ, start from (6.7.33) and relegate any term not involving θ to the proportionality constant. Then, rearranging, one obtains:

$$\pi(\theta \left|y, \sigma, \tilde{\Sigma}\right.) \propto \exp\left(-\frac{1}{2}(\theta-\bar{\theta})'\bar{\Theta}^{-1}(\theta-\bar{\theta})\right) \tag{6.7.34}$$

with:

$$\bar{\Theta} = \left(\tilde{X} I_{\Sigma} \tilde{X}' + \Theta_0^{-1}\right)^{-1} \tag{6.7.35}$$

and

$$\bar{\theta} = \bar{\Theta}\left(\tilde{X} I_{\Sigma} y + \Theta_0^{-1}\theta_0\right) \tag{6.7.36}$$

with:

$$\tilde{X} = \underbrace{\begin{pmatrix} \tilde{X}_1' & \tilde{X}_2' & \cdots & \tilde{X}_T' \end{pmatrix}}_{d \times NnT} \qquad I_\Sigma = (I_T \otimes \Sigma^{-1}) = \underbrace{\begin{pmatrix} \Sigma^{-1} & 0 & \cdots & 0 \\ 0 & \Sigma^{-1} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma^{-1} \end{pmatrix}}_{NnT \times NnT} \qquad y = \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix}}_{NnT \times 1} \tag{6.7.37}$$

This is the kernel of a multivariate normal distribution with mean $\bar{\theta}$ and covariance matrix $\bar{\Theta}$:

$$\pi(\theta \left| y, \sigma, \tilde{\Sigma} \right.) \sim \mathcal{N}(\bar{\theta}, \bar{\Theta}) \tag{6.7.38}$$

Then obtain the conditional posterior for $\tilde{\Sigma}$. Relegating to the proportionality constant any term not involving $\tilde{\Sigma}$ in (6.7.33) and rearranging, one obtains:

$$\pi(\tilde{\Sigma} \left| y, \theta, \sigma \right.) \propto \left| \tilde{\Sigma} \right|^{-(T+Nn+1)/2} \exp\left(-\frac{1}{2} tr\left\{ \tilde{\Sigma}^{-1} \bar{S} \right\} \right) \tag{6.7.39}$$

with:

$$\bar{S} = \sigma^{-1} \left(Y - \ddot{X} I_\theta \right) \left(Y - \ddot{X} I_\theta \right)' \tag{6.7.40}$$

where:

$$Y = \underbrace{\begin{pmatrix} y_1 & y_2 & \cdots & y_T \end{pmatrix}}_{Nn \times T} \qquad \ddot{X} = \underbrace{\begin{pmatrix} \tilde{X}_1 & \tilde{X}_2 & \cdots & \tilde{X}_T \end{pmatrix}}_{Nn \times Td} \qquad I_\theta = (I_T \otimes \theta) = \underbrace{\begin{pmatrix} \theta & 0 & \cdots & 0 \\ 0 & \theta & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \theta \end{pmatrix}}_{Td \times T} \tag{6.7.41}$$

This is the kernel of an inverse Wishart distribution with scale $\bar{S}$ and T degrees of freedom:

$$\pi(\theta, \tilde{\Sigma}, \sigma \left| y \right.) \sim \ IW(\bar{S}, T) \tag{6.7.42}$$

Finally, obtain the conditional posterior for σ. Relegating to the proportionality constant any term not involving σ in (6.7.33), then rearranging, one obtains:

$$\pi(\sigma \left| y, \theta, \tilde{\Sigma} \right.) \propto (\sigma)^{-\frac{\bar{\alpha}}{2}-1} \exp\left(-\frac{\bar{\delta}}{2\sigma} \right) \tag{6.7.43}$$

with:

$$\bar{\alpha} = NnT + \alpha_0 \tag{6.7.44}$$

and:

$$\bar{\delta} = \left[tr\left((Y - \ddot{X} I_\theta)(Y - \ddot{X} I_\theta)' \tilde{\Sigma}^{-1} \right) + \delta_0 \right] \tag{6.7.45}$$

This is the kernel of an inverse Gamma distribution with shape $\frac{\bar{\alpha}}{2}$ and scale $\frac{\bar{\delta}}{2}$:

$$\pi(\sigma_t \left| y, \theta, \tilde{\Sigma} \right.) \sim \ IG\left(\frac{\bar{\alpha}}{2}, \frac{\bar{\delta}}{2} \right) \tag{6.7.46}$$

With these elements at hand, it is eventually possible to derive the Gibbs sampling algorithm for the posterior distribution of the full model:

Algorithm 4.7.1 (Gibbs sampling algorithm for a panel VAR model with a factor approach):

1. Define starting values $\theta^{(0)}, \tilde{\Sigma}^{(0)}$ and $\sigma^{(0)}$. For $\theta^{(0)}$, use the OLS value $\hat{\theta}$. For $\tilde{\Sigma}^{(0)}$, use (6.7.14) to obtain $\varepsilon_t = y_t - \tilde{X}_t \hat{\theta}$, and obtain $\tilde{\Sigma}^{(0)} = 1/T \sum_{t=1}^{T} \varepsilon_t \varepsilon_t'$. For $\sigma^{(0)}$, the value is set to 1 , which implies $\Sigma^{(0)} = \tilde{\Sigma}^{(0)}$. Then compute $I_\Sigma^{(0)} = I_T \otimes (\Sigma^{(0)})^{-1}$ and $I_\theta^{(0)} = I_T \otimes \theta^{(0)}$.
2. At iteration n, draw $\tilde{\Sigma}^{(n)}$ from $\pi\left(\tilde{\Sigma}^{(n)} \left| y, \theta^{(n-1)}, \sigma^{(n-1)} \right. \right) \sim \ IW(\bar{S}, T)$, with:
   $\bar{S} = (\sigma^{(n-1)})^{-1} \left(Y - \ddot{X} I_\theta^{(n-1)} \right) \left(Y - \ddot{X} I_\theta^{(n-1)} \right)'$
3. At iteration n, draw $\sigma^{(n)}$ from $\pi\left(\sigma^{(n)} \left| y, \theta^{(n-1)}, \tilde{\Sigma}^{(n)} \right. \right) \sim \ IG\left(\frac{\bar{\alpha}}{2}, \frac{\bar{\delta}}{2} \right)$, with:
   $\bar{\alpha} = NnT + \alpha_0$
   and:
   $\bar{\delta} = \left[tr\left((Y - \ddot{X} I_\theta^{(n-1)})(Y - \ddot{X} I_\theta^{(n-1)})' (\tilde{\Sigma}^{(n)})^{-1} \right) + \delta_0 \right]$
4. At iteration n, compute $\Sigma^{(n)} = \sigma^{(n)} \tilde{\Sigma}^{(n)}$, and use it to compute $I_\Sigma^{(n)} = I_T \otimes (\Sigma^{(n)})^{-1}$.
5. At iteration n, draw $\theta^{(n)}$ from $\pi\left(\theta^{(n)} \left| y, \sigma^{(n)}, \tilde{\Sigma}^{(n)} \right. \right) \sim \mathcal{N}(\bar{\theta}, \bar{\Theta})$, with:
   $\bar{\Theta} = \left(\tilde{X} I_\Sigma^{(n)} \tilde{X}' + \Theta_0^{-1} \right)^{-1}$
   and
   $\bar{\theta} = \bar{\Theta}\left(\tilde{X} I_\Sigma^{(n)} y + \Theta_0^{-1} \theta_0 \right)^{-1}$
6. At iteration n, compute $I_\theta^{(n)} = I_T \otimes \theta^{(n)}$.
7. Repeat until $(It - Bu)$ iterations are realised.

6.8 Adding dynamic heterogeneity to the model: a dynamic factor approach

The most flexible version of the Bayesian panel VAR model is eventually introduced. It integrates all the four possible panel properties: dynamic and static interdependencies, cross-subsectional heterogeneity, and also dynamic heterogeneity. This latter aspect is important as modern macroeconomic methodologies consider seriously the possibility that dynamic coefficients may evolve over time. The approach developed in this subsection essentially follows Canova and Ciccarelli (2013) and Ciccarelli et al. (2012), and builds on the structural factor approach previously introduced for static coefficients. Some modifications have been applied to the original methodology of Canova and Ciccarelli (2013) in order to implement a proper dynamic heteroskedasticity scheme, and to replace the Kalman smoother used by these authors by a faster sparse matrix approach. While this dynamic approach is numerically heavier than the static one, adaptation to a time-varying context is relatively straightforward.

Start again from the general formulation 6.1.5, but do not relax dynamic heterogeneity. The dynamic equation for the model at period t is then given by:

$$y_t = A_t^1 y_{t-1} + \cdots + A_t^p y_{t-p} + C_t x_t + \varepsilon_t \tag{6.8.1}$$

Take transpose:

$$y_t' = (A_t^1)' y_{t-1}' + \cdots + (A_t^p)' y_{t-p}' + (C_t)' x_t' + \varepsilon_t' \tag{6.8.2}$$

Reformulate in compact form:

$$y_t' = \begin{pmatrix} y_{t-1}' & \cdots & y_{t-p}' & x_t' \end{pmatrix} \begin{pmatrix} (A_t^1)' \\ \vdots \\ (A_t^p)' \\ (C_t)' \end{pmatrix} + \varepsilon_t' \tag{6.8.3}$$

or:

$$y_t' = X_t B_t + \varepsilon_t' \tag{6.8.4}$$

with:

$$X_t = \underbrace{\begin{pmatrix} y_{t-1}' & \cdots & y_{t-p}' & x_t' \end{pmatrix}}_{1 \times k} \qquad B_t = \underbrace{\begin{pmatrix} (A_t^1)' \\ \vdots \\ (A_t^p)' \\ (C_t)' \end{pmatrix}}_{k \times Nn} \tag{6.8.5}$$

Obtain a vectorised form by using A.1.5:

$$y_t = (I_{Nn} \otimes X_t)\, vec(B_t) + \varepsilon_t \tag{6.8.6}$$

or:

$$y_t = \bar{X}_t \beta_t + \varepsilon_t \tag{6.8.7}$$

with:

$$\bar{X}_t = \underbrace{(I_{Nn} \otimes X_t)}_{Nn \times h} \qquad \beta_t = \underbrace{vec(B_t)}_{h \times 1} \tag{6.8.8}$$

A general form of heteroskedasticity is introduced for the error. Precisely, it is assumed that the error term is independently distributed across periods according to:

$$\varepsilon_t \sim \mathcal{N}(0, \Sigma_t) \qquad \text{with} \qquad \Sigma_t = \exp(\zeta_t)\tilde{\Sigma} \tag{6.8.9}$$

Hence:

$$\Sigma_t = E\left(\varepsilon_t \varepsilon_t^{'}\right) = \exp(\zeta_t)\tilde{\Sigma} = \underbrace{\exp(\zeta_t)}_{1 \times 1} \underbrace{\begin{pmatrix} \underbrace{\tilde{\Sigma}_{11}}_{n \times n} & \tilde{\Sigma}_{12} & \cdots & \tilde{\Sigma}_{1N} \\ \tilde{\Sigma}_{21} & \tilde{\Sigma}_{22} & \cdots & \tilde{\Sigma}_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\Sigma}_{N1} & \tilde{\Sigma}_{N2} & \cdots & \tilde{\Sigma}_{NN} \end{pmatrix}}_{Nn \times Nn} \tag{6.8.10}$$

ζ_t is a dynamic coefficient whose law of motion is defined as:

$$\zeta_t = \gamma \zeta_{t-1} + \upsilon_t \tag{6.8.11}$$

The initial value of the process ζ_0 is set to be 0, which implies initial homoscedasticity. υ_t is a disturbance following a normal distribution:

$$\upsilon_t \sim \mathcal{N}(0, \varphi) \tag{6.8.12}$$

A final layer of uncertainty is integrated by assuming that φ is also a random variable, which follows an inverse Gamma distribution:

$$\varphi \sim \ IG\left(\frac{\alpha_0}{2}, \frac{\delta_0}{2}\right) \tag{6.8.13}$$

Note the implications of the setting. Thanks to (6.8.11), heteroskedasticity is modelled in a flexible fashion: when $\gamma = 0$, ζ_t is determined by a white noise process, so that the residuals are heteroskedastic in a purely random way. When $0 < \gamma < 1$, the model allows for inertia in heteroskedasticity which then resembles that of a traditional ARCH model. Finally, when $\gamma = 1$ the residual follow a random

walk form of heteroskedasticity, implying permanent jumps on the volatility of the disturbances. The random variable φ determines if there exists heteroskedasticity altogether: a value of 0 implies that all υ_t shocks endorse the mean value of 0, so that $\exp(\zeta_t) = 1$ for all t and the model is homoscedastic. Then the larger φ, the larger the disturbance volatility implied by the setting.

This basically concludes the description of the model. (6.8.7) is a standard linear model which suffers however from the same curse of dimensionality issue than its static counterpart. The problem is actually made even worse in a time-varying context: because the model is period-specific, there are $h = Nn(Nnp + m)$ coefficients to estimate, but only Nn data points available. This renders estimation impossible with classical methods. Canova and Ciccarelli (2013) thus propose to adopt a structural factor approach similar to that developed in the previous subsection. Concretely, the vector of coefficients β_t is decomposed into r structural factors:

$$\beta_t = \Xi_1\theta_{1,t} + \Xi_2\theta_{2,t} + ... + \Xi_r\theta_{r,t} = \sum_{i=1}^{r} \Xi_i\theta_{i,t} \tag{6.8.14}$$

Once again, $\theta_{1,t}, \theta_{2,t}, ..., \theta_{r,t}$ are vectors of dimension $d_1 \times 1, d_2 \times 1, \cdots, d_r \times 1$ containing the structural factors, while $\Xi_1, \Xi_2, \cdots, \Xi_r$ are selection matrices of dimension $h \times d_1, h \times d_2, \cdots, h \times d_r$ with all their entries being either 0 or 1 picking the relevant elements in $\theta_{1,t}, \theta_{2,t}, \cdots, \theta_{r,t}$. Unlike the static version of the model, all the factors are now allowed to be time-varying. A compact form of the factor decomposition can be obtained as:

$$\beta_t = \Xi\theta_t \tag{6.8.15}$$

with:

$$\Xi = \underbrace{\begin{pmatrix} \Xi_1 & \Xi_2 & \cdots & \Xi_r \end{pmatrix}}_{h \times d} \qquad \text{and} \qquad \theta_t = \underbrace{\begin{pmatrix} \theta_{1,t} \\ \theta_{2,t} \\ \vdots \\ \theta_{r,t} \end{pmatrix}}_{d \times 1} \tag{6.8.16}$$

Substitute (6.8.14) in (6.8.7) to obtain a reformulated model:

$$\begin{aligned} y_t &= \bar{X}_t\beta_t + \varepsilon_t \\ &= \bar{X}_t\left(\sum_{i=1}^{r}\Xi_i\theta_{i,t}\right) + \varepsilon_t \\ &= \left(\sum_{i=1}^{r}\bar{X}_t\Xi_i\theta_{i,t}\right) + \varepsilon_t \\ &= \sum_{i=1}^{r}\left(\bar{X}_t\Xi_i\right)\theta_{i,t} + \varepsilon_t \end{aligned}$$

or:

$$y_t = \tilde{X}_{1,t}\theta_{1,t} + \tilde{X}_{2,t}\theta_{2,t} + \cdots + \tilde{X}_{r,t}\theta_{r,t} + \varepsilon_t \tag{6.8.17}$$

with:

$$\tilde{X}_{i,t} = \bar{X}_t\Xi_i, i = 1, 2, \cdots, r \tag{6.8.18}$$

Similarly, a compact form can be recovered using (6.8.15) in (6.8.7):

$$y_t = \tilde{X}_t\theta_t + \varepsilon_t \tag{6.8.19}$$

with:

$$\tilde{X}_t = \bar{X}_t\Xi \tag{6.8.20}$$

Finally, define the law of motion of θ_t. A simple form is the general autoregressive process:

$$\theta_t = (1-\rho)\bar{\theta} + \rho\theta_{t-1} + \eta_t \tag{6.8.21}$$

with:

$$\eta_t \sim \mathcal{N}(0, B) \tag{6.8.22}$$

$0 \leq \rho \leq 1$ determines the persistence of the process, while the constant term $\bar{\theta}$ represents the long-run value of the model. This form is flexible and allows for many particular representations . For instance, a common special case nested in this general specification is the random walk, corresponding to $\rho = 1$:

$$\theta_t = \theta_{t-1} + \eta_t \tag{6.8.23}$$

B is a block diagonal matrix for which each block $i = 1, 2, \cdots, r$ corresponds to one of the r structural factors and is of dimension d_i. It is defined as:

$$B = \begin{pmatrix} B_1 & 0 & \cdots & 0 \\ 0 & B_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & B_r \end{pmatrix} = \begin{pmatrix} b_1 I_{d_1} & 0 & \cdots & 0 \\ 0 & b_2 I_{d_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & b_r I_{d_r} \end{pmatrix} \tag{6.8.24}$$

This concludes the time varying factor approach. The models comprises 5 sets of parameters of interest to be estimated: the factor coefficients $\theta = \{\theta_t\}_{t=1}^T$, the set of VAR coefficient variances $b = \{b_i\}_{i=1}^r$, the homoskedastic residual covariance matrix $\tilde{\Sigma}$, the set of dynamic coefficients $\zeta = \{\zeta_t\}_{t=1}^T$, and the heteroskedasticity variance coefficient φ. The considered setting is hierarchical, since θ is obtained conditional on b, and ζ obtains conditional on φ.

Obtaining the prior for $\theta = \{\theta_t\}_{t=1}^T$ is troublesome since (6.8.21) implies that each term depends on the previous period value θ_{t-1}. The strategy to obtain a simple joint formulation relies on the sparse matrix approach by Chan and Jeliazkov (2009)[7]. It essentially consists in noticing that every value θ_t ultimately depends on the initial condition θ_0, the long-run value $\bar{\theta}$, and the set of shocks $\eta_t, t = 1, 2, \cdots, T$. (6.8.21) can be reformulated for all the periods simultaneously as:

$$H\Theta = \tilde{\Theta} + \eta \tag{6.8.25}$$

with:

$$H = \underbrace{\begin{pmatrix} I_d & 0 & 0 & \cdots & 0 \\ -\rho I_d & I_d & 0 & \cdots & 0 \\ 0 & -\rho I_d & I_d & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -\rho I_d & I_d \end{pmatrix}}_{Td \times Td} \qquad \Theta = \underbrace{\begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \vdots \\ \theta_T \end{pmatrix}}_{Td \times 1}$$

[7]The authors are grateful to Aubrey Poon (Australian National University) and Matteo Ciccarelli (European Central Bank) for suggesting this approach.

$$\tilde{\Theta} = \underbrace{\begin{pmatrix} (1-\rho)\bar{\theta} + \rho\theta_0 \\ (1-\rho)\bar{\theta} \\ (1-\rho)\bar{\theta} \\ \vdots \\ (1-\rho)\bar{\theta} \end{pmatrix}}_{Td\times 1} \qquad \eta = \underbrace{\begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \vdots \\ \eta_T \end{pmatrix}}_{Td\times 1} \tag{6.8.26}$$

Assuming independence between the errors η_t, it follows from (6.8.22) that:

$$\eta \sim \mathcal{N}(0, \tilde{B}) \qquad \text{with} \qquad \tilde{B} = I_T \otimes B \tag{6.8.27}$$

Then (6.8.25) implies that:

$$\Theta = H^{-1}\tilde{\Theta} + H^{-1}\eta \tag{6.8.28}$$

And from A.2.2.6, (6.8.27) and (6.8.28), one can eventually conclude that:

$$\Theta \sim \mathcal{N}(H^{-1}\tilde{\Theta}, H^{-1}\tilde{B}(H^{-1})') \qquad \text{or} \qquad \Theta \sim \mathcal{N}(\Theta_0, B_0) \tag{6.8.29}$$

with:

$$\Theta_0 = H^{-1}\tilde{\Theta} \qquad \text{and} \qquad B_0 = H^{-1}\tilde{B}(H^{-1})' \tag{6.8.30}$$

Thus the prior distribution for the whole series $\theta = \{\theta_t\}_{t=1}^T$ can be expressed in stacked form as:

$$\pi(\theta\,|b) \propto |B_0| \exp\left(-\frac{1}{2}(\Theta - \Theta_0)' B_0^{-1}(\Theta - \Theta_0)\right) \tag{6.8.31}$$

It remains to determine the values for θ_0 and $\bar{\theta}$. These two values are set as the OLS estimate of the static version of (6.8.19).

The prior distribution for each b_i is inverse Gamma with scale $\frac{a_0}{2}$ and shape $\frac{b_0}{2}$:

$$\pi\left(b_i\,|a_0, b_0\right) \propto b_i^{-(\frac{a_0}{2})-1} \exp\left(\frac{-b_0}{2b_i}\right) \tag{6.8.32}$$

The prior for $\tilde{\Sigma}$ is a standard diffuse prior:

$$\pi(\tilde{\Sigma}) \propto \left|\tilde{\Sigma}\right|^{-(Nn+1)/2} \tag{6.8.33}$$

The prior for $\zeta = \{\zeta_t\}_{t=1}^T$ faces the same dynamic dependence as θ, so that the same identification strategy is retained. Reformulate (6.8.11) simultaneously for all periods as:

$$GZ = \upsilon \tag{6.8.34}$$

with:

$$G = \underbrace{\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\gamma & 1 & 0 & \cdots & 0 \\ 0 & -\gamma & 1 & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -\gamma & 1 \end{pmatrix}}_{T\times T} \qquad Z = \underbrace{\begin{pmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \\ \vdots \\ \zeta_T \end{pmatrix}}_{T\times 1} \qquad \upsilon = \underbrace{\begin{pmatrix} \upsilon_1 \\ \upsilon_2 \\ \upsilon_3 \\ \vdots \\ \upsilon_T \end{pmatrix}}_{T\times 1} \tag{6.8.35}$$

where use has been made of the fact that $\zeta_0 = 0$. Assuming independence between the errors υ_t, it follows from (6.8.12) that:

$$\upsilon \sim \mathcal{N}(0, \varphi I_T) \tag{6.8.36}$$

Then (6.8.34) implies that $Z = G^{-1}\upsilon$, and from (A.2.5) one concludes that:

$$Z \sim \mathcal{N}(0, G^{-1}\varphi I_T (G^{-1})') \tag{6.8.37}$$

or:

$$Z \sim \mathcal{N}(0, \Phi_0) \qquad \text{with} \qquad \Phi_0 = \varphi (G'G)^{-1} \tag{6.8.38}$$

The prior distribution for the whole series $\zeta = \{\zeta_t\}_{t=1}^T$ can thus be expressed in stacked form as:

$$\pi(Z\,|\varphi) \propto |\Phi_0|^{-1/2} \exp\left(-\frac{1}{2} Z' \Phi_0^{-1} Z\right) \tag{6.8.39}$$

Finally, the prior for φ is given by (6.8.13), so that:

$$\pi(\varphi) \propto \varphi^{-\frac{\alpha_0}{2}-1} \exp\left(\frac{-\delta_0}{2\varphi}\right) \tag{6.8.40}$$

To obtain the posterior distribution, obtain first an expression for Bayes rule. Relying on a standard independence assumption between θ, Σ and ζ, and following the hierarchical setting described in subsection 3.2 one obtains:

$$\pi(\theta, b, \tilde{\Sigma}, \zeta, \varphi\,|y) \propto f(y\,\big|\,\theta, \tilde{\Sigma}, \zeta)\pi(\theta\,|b)\pi(b)\pi(\tilde{\Sigma})\pi(\zeta\,|\varphi)\pi(\varphi) \tag{6.8.41}$$

Given (6.8.9) and (6.8.19), the likelihood function $\pi(y\,|\theta,\Sigma,\zeta)$ is given by:

$$f(y\,\big|\theta,\tilde{\Sigma},\zeta) \propto \left|\tilde{\Sigma}\right|^{-T/2} \exp\left(-\frac{1}{2}\sum_{t=1}^{T}\left\{\exp(-\zeta_t)(y_t-\tilde{X}_t\theta_t)'\tilde{\Sigma}^{-1}(y_t-\tilde{X}_t\theta_t)+Nn\zeta_t\right\}\right) \tag{6.8.42}$$

From Bayes rule (6.8.41), the likelihood (6.8.42) and the priors (6.8.31), (6.8.32), (6.8.33), (6.8.39) and (6.8.40), one can derive the full posterior distribution:

$$\begin{aligned}\pi(\theta,b,\tilde{\Sigma},\zeta,\varphi\,|y) \propto & \left|\tilde{\Sigma}\right|^{-T/2} \exp\left(-\frac{1}{2}\sum_{t=1}^{T}\left\{\exp(-\zeta_t)(y_t-\tilde{X}_t\theta_t)'\tilde{\Sigma}^{-1}(y_t-\tilde{X}_t\theta_t)+Nn\zeta_t\right\}\right)\\ & \times |B_0| \exp\left(-\frac{1}{2}(\Theta-\Theta_0)'B_0^{-1}(\Theta-\Theta_0)\right)\\ & \times \prod_{i=1}^{r} b_i^{-(a_0/2)-1}\exp\left(\frac{-b_0}{2b_i}\right)\\ & \times \left|\tilde{\Sigma}\right|^{-(Nn+1)/2}\\ & \times |\Phi_0|^{-1/2}\exp\left(-\frac{1}{2}Z'\Phi_0^{-1}Z\right)\\ & \times \varphi^{-\alpha_0/2-1}\exp\left(\frac{-\delta_0}{2\varphi}\right)\end{aligned} \tag{6.8.43}$$

The complexity of the formula makes any analytical marginalisation intractable, so that one relies as usual on the numerical framework provided by the Gibbs sampler.

Obtain the full set of conditional posterior distributions. Derive first the conditional posterior for $\theta=\{\theta_t\}_{t=1}^{T}$. Given Bayes rule (6.8.41), relegate to the proportionality constant any term that does not involve θ to obtain:

$$\pi(\theta\,\big|y,b,\tilde{\Sigma},\zeta,\varphi) \propto f(y\,\big|\theta,\tilde{\Sigma},\zeta)\pi(\theta\,|b) \tag{6.8.44}$$

However, because the prior for θ is formulated in terms of Θ (defined in (6.8.26)), the likelihood must also be reformulated in terms of Θ in order to derive the posterior. This can be done easily by considering a stacked form of (6.8.19):

$$y=\tilde{X}\Theta+\varepsilon \tag{6.8.45}$$

with:

$$\underbrace{y=\begin{pmatrix}y_1\\y_2\\\vdots\\y_T\end{pmatrix}}_{NnT\times 1} \qquad \underbrace{\tilde{X}=\begin{pmatrix}\tilde{X}_1 & 0 & \cdots & 0\\ 0 & \tilde{X}_2 & & \\ \vdots & & \ddots & \vdots\\ 0 & & \cdots & \tilde{X}_T\end{pmatrix}}_{NnT\times Td} \qquad \underbrace{\varepsilon=\begin{pmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_T\end{pmatrix}}_{NnT\times 1} \tag{6.8.46}$$

Also, from (6.8.9), the error terms ε_t are independently distributed so that:

$$\varepsilon \sim \mathcal{N}(0, \Sigma) \qquad \text{with} \qquad \Sigma = \underbrace{\begin{pmatrix} \Sigma_1 & 0 & \cdots & 0 \\ 0 & \Sigma_2 & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & \Sigma_T \end{pmatrix}}_{NnT \times NnT} \tag{6.8.47}$$

The likelihood function for the full model may then rewrite as:

$$f(y \,|\Theta, \tilde{\Sigma}, \zeta) \propto |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(y - \tilde{X}\Theta)'\Sigma^{-1}(y - \tilde{X}\Theta)\right) \tag{6.8.48}$$

From (6.8.44), one combines the likelihood (6.8.48) with the prior (6.8.31), and rearranges to obtain:

$$\pi(\theta \,|y, b, \tilde{\Sigma}, \zeta, \varphi) \propto \exp\left(-\frac{1}{2}(\Theta - \bar{\Theta})'\bar{B}^{-1}(\Theta - \bar{\Theta})\right) \tag{6.8.49}$$

with:

$$\bar{B} = \left(\tilde{X}'\Sigma^{-1}\tilde{X} + B_0^{-1}\right)^{-1} \tag{6.8.50}$$

and:

$$\bar{\Theta} = \bar{B}\left(\tilde{X}'\Sigma^{-1}y + B_0^{-1}\Theta_0\right) \tag{6.8.51}$$

This is the kernel of a multivariate normal distribution with mean $\bar{\Theta}$ and covariance $\bar{B}$:

$$\Theta \sim \mathcal{N}(\bar{\Theta}, \bar{B}) \tag{6.8.52}$$

Obtain then the conditional posterior for $b = \{b_i\}_{i=1}^r$. Because the b_is are conditionally independent, the posteriors can be derived individually. Using Bayes rule (6.8.41) and relegating to the proportionality constant any term not involving b_i yields:

$$\pi(b_i \,|y, \theta, \; b_{-i}, \tilde{\Sigma}, \zeta, \varphi) \propto \pi(\theta_i \,|b_i)\pi(b_i) \tag{6.8.53}$$

Using (6.8.53) to combine (6.8.22)-(6.8.23) with (6.8.32) and then rearranging eventually yields:

$$\pi(b_i \,|y, \theta, \; b_{-i}, \tilde{\Sigma}, \zeta, \varphi) \propto b_i^{-\frac{\bar{a}_i}{2}-1} \exp\left(-\frac{\bar{b}_i}{2b_i}\right) \tag{6.8.54}$$

with:

$$\bar{a}_i = Td_i + a_0 \tag{6.8.55}$$

and:

$$\bar{b}_i = \sum_{t=1}^{T} (\theta_{i,t} - \theta_{i,t-1})'(\theta_{i,t} - \theta_{i,t-1}) + b_0 \tag{6.8.56}$$

This is the kernel of an inverse Gamma distribution with shape $\frac{\bar{a}}{2}$ and scale $\frac{\bar{b}}{2}$:

$$\pi(b_i \,|y, \theta,\ b_{-i}, \tilde{\Sigma}, \zeta, \varphi) \sim\ IG\left(\frac{\bar{a}_i}{2}, \frac{\bar{b}_i}{2}\right) \tag{6.8.57}$$

Obtain the conditional posterior for $\tilde{\Sigma}$. Starting from Bayes rule (6.8.41) and relegating to the proportionality constant any term not involving $\tilde{\Sigma}$, one obtains:

$$\pi(\Sigma \,|y\ , \theta, b, \zeta, \varphi) \propto f(y \,|\theta, \Sigma, \zeta)\pi(\Sigma) \tag{6.8.58}$$

Using (6.8.42) and (6.8.33) and rearranging, one obtains:

$$\pi(\tilde{\Sigma} \,|y, \theta, b,\ \zeta, \varphi) \propto \left|\tilde{\Sigma}\right|^{-(T+Nn+1)/2} \exp\left(-\frac{1}{2} tr\left\{\tilde{\Sigma}^{-1}\bar{S}\right\}\right) \tag{6.8.59}$$

with:

$$\bar{S} = \sum_{t=1}^{T} (y_t - \tilde{X}_t\theta_t) \exp(-\zeta_t)(y_t - \tilde{X}_t\theta_t)' \tag{6.8.60}$$

This the kernel of an inverse Wishart distribution with scale $\bar{S}$ and degrees of freedom T:

$$\pi(\tilde{\Sigma} \,|y, \theta, b,\ \zeta, \varphi) \sim\ IW(\bar{S}, T) \tag{6.8.61}$$

Compute the conditional posterior for φ. Starting from Bayes rule (6.8.41) and relegating to the proportionality constant any term not involving φ yields:

$$\pi(\varphi \,\Big|y, \theta, b, \tilde{\Sigma}, \zeta\,) \propto \pi(\zeta \,|\varphi)\pi(\varphi) \tag{6.8.62}$$

From (6.8.39), (6.8.40) and rearranging, one eventually obtains:

$$\pi(\theta, b, \tilde{\Sigma}, \zeta, \varphi \,|y\,) \propto \varphi^{-\frac{\bar{\alpha}}{2}-1} \exp\left(-\frac{\bar{\delta}}{2\varphi}\right) \tag{6.8.63}$$

with:

$$\bar{\alpha} = T + \alpha_0 \tag{6.8.64}$$

and:

$$\bar{\delta} = Z'G'GZ + \delta_0 \tag{6.8.65}$$

This is the kernel of an inverse Gamma distribution with scale $\frac{\bar{\alpha}}{2}$ and shape $\frac{\bar{\delta}}{2}$:

$$\pi(\varphi \left| y, \theta, b, \tilde{\Sigma}, \zeta \right.) \sim \ IG\left(\frac{\bar{\alpha}}{2}, \frac{\bar{\delta}}{2}\right) \tag{6.8.66}$$

Obtain finally the conditional posterior for $\zeta = \{\zeta_t\}_{t=1}^T$. Consider Bayes rule (6.8.41) and relegate to the proportionality constant any term not involving ζ:

$$\pi(\zeta \left| y, \theta, b, \tilde{\Sigma}, \varphi \right.) \propto f(y \left| \theta, \tilde{\Sigma}, \zeta \right.)\pi(\zeta \left| \varphi \right.) \tag{6.8.67}$$

Using (6.8.42) and (6.8.39) and rearranging, one obtains:

$$\begin{aligned} &\pi(\zeta \left| y, \theta, b, \tilde{\Sigma}, \varphi \right.) \\ &\propto \exp\left(-\frac{1}{2}\left[\sum_{t=1}^{T}\left\{\exp(-\zeta_t)(y_t - \tilde{X}_t\theta_t)'\tilde{\Sigma}^{-1}(y_t - \tilde{X}_t\theta_t) + Nn\zeta_t\right\} + Z'\Phi_0^{-1}Z\right]\right) \end{aligned} \tag{6.8.68}$$

(6.8.68) is problematic: it does not correspond to a known distribution. Because its form is non-standard, it is not possible to sample directly from it, which renders the Gibbs sampler methodology inapplicable. There exists a solution to solve this issue, but this comes at the cost of additional technical complications. The methodology to be used in this case is known as the Metropolis-Hastings algorithm, and it can be shown to be a generalisation of the Gibbs sampler, applicable even in cases where the posterior distribution takes an unknown form. The technical details behind the methodology are beyond the scope of this manual so that only the principles will be introduced[8].

The idea of the Metropolis-Hastings methodology is the following: while for the Gibbs sampler a new draw from the conditional distribution was obtained and accepted at every iteration of the algorithm, with the Metropolis-Hastings algorithm a new draw will only be accepted with a certain probability. If the draw is rejected, the previous iteration value is retained.

Concretely, the algorithm works as follows: consider any parameter θ for which one wants to obtain draws from the posterior distribution. The analytical formula $\pi(\theta)$ corresponding to the posterior distribution is identified so that one can calculate the density value (as in the case of (6.8.68)), but this formula does not correspond to a known distribution, so that it is not possible to sample directly from $\pi(\theta)$. One has then to define what is known as a transition kernel $q(\theta^{(n-1)}, \theta^{(n)})$, which is a distribution establishing how to obtain at iteration n a value $\theta^{(n)}$ from the previous iteration value $\theta^{(n-1)}$. Since many distributions can be used as a possible transition kernel, the choice has to

[8] Readers interested in a more formal treatment may find valuable content in Greenberg (2008), chapters 6 and 7, and Chib and Greenberg (1995).

be made based on convenience. Finally, given $\pi(\theta)$ and $q(\theta^{(n-1)}, \theta^{(n)})$, one has to define a function $\alpha(\theta^{(n-1)}, \theta^{(n)})$ which determines the probability that the draw obtained in the current iteration will be accepted. This function is always the same and given by:

$$\alpha(\theta^{(n-1)}, \theta^{(n)}) = \begin{cases} \min\left\{1, \frac{\pi(\theta^{(n)})q(\theta^{(n)}, \theta^{(n-1)})}{\pi(\theta^{(n-1)})q(\theta^{(n-1)}, \theta^{(n)})}\right\} & \text{if } \pi(\theta^{(n)})q(\theta^{(n)}, \theta^{(n-1)}) \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{6.8.69}$$

It can then be shown that given the chosen transition kernel $q(\theta^{(n-1)}, \theta^{(n)})$, the distribution obtained from $\alpha(\theta^{(n-1)}, \theta^{(n)})$ corresponds to the unconditional posterior distribution for θ. The Gibbs sampler can thus be seen as a special case of the Metropolis-Hastings algorithm where the acceptance probability $\alpha(\theta^{(n-1)}, \theta^{(n)})$ for the current draw is always equal to 1. The general algorithm then goes as follows:

Algorithm 4.8.1 (Metropolis-Hastings algorithm for a generic parameter θ):

1. Obtain the posterior density $\pi(\theta)$.
2. Define a transition kernel $q(\theta^{(n-1)}, \theta^{(n)})$.
3. Start iterating over θ values: at iteration n, obtain a candidate value $\tilde{\theta}$ from the transition kernel $q(\theta^{(n-1)}, \theta^{(n)})$.
4. At iteration n, obtain an acceptance probability $\alpha(\theta^{(n-1)}, \theta^{(n)})$ from (6.8.69).
5. At iteration n, draw a random number x from a uniform distribution $x \sim U(0, 1)$.
6. If $x \leq \alpha(\theta^{(n-1)}, \theta^{(n)})$, then the draw is accepted: define $\theta^{(n)} = \tilde{\theta}$. If $x > \alpha(\theta^{(n-1)}, \theta^{(n)})$, the draw is rejected, keep the former value: define $\theta^{(n)} = \theta^{(n-1)}$.
7. Return to step 3 and repeat until the desired number of iterations is realised.

The Metropolis-Hastings algorithm can be integrated to a wider setting. If there are several blocks of parameters (for instance, in the present model: $\theta, b, \tilde{\Sigma}$ and φ in addition to ζ), it is possible to run the Metropolis-Hastings only for the blocks characterised by posterior distributions with unknown forms. For the blocks with known forms, a standard Gibbs sampling approach can be applied. All the draws, from the Gibbs sampler or from the Metropolis-Hastings algorithm, have to be realised conditional on the other block values.

The final question is then the determination of the transition kernel $q(\theta^{(n-1)}, \theta^{(n)})$. Ideally, a good kernel should allow for sufficient variability in the value of θ between two iterations. This ensures that

a large part of the support of $\pi(\theta)$ will be covered by the iterations of the algorithm, which improves the mixing between iterations and the quality of the posterior. However, larger differences between $\theta^{(n)}$ and $\theta^{(n-1)}$ typically imply larger differences between $\pi(\theta^{(n)})$ and $\pi(\theta^{(n-1)})$, which increases the probability of rejection from (6.8.69). Then some values may be repeated often and the algorithm may perform poorly. The kernel must thus be chosen to generate the most efficient compromise between these two aspects.

A common choice is the random walk kernel defined as:

$$\theta^{(n)} = \theta^{(n-1)} + \omega \tag{6.8.70}$$

with ω an error term with a known distribution. Typically, ω is defined as a (multivariate) normal random variable:

$$\omega \sim \mathcal{N}(0, \Omega) \tag{6.8.71}$$

Ω is of particular importance as it determines how much variability is permitted by the transition kernel, and thus sets the compromise between variability and acceptance. (6.8.70) then implies that the transition kernel is given by:

$$q(\theta^{(n-1)}, \theta^{(n)}) \sim \mathcal{N}(\theta^{(n-1)}, \Omega) \tag{6.8.72}$$

This definition of the kernel is particularly convenient as one can use the symmetry of the normal distribution for ω to simplify (6.8.69). Indeed, from (6.8.70), one obtains the following result for the transition kernel:

$$\begin{aligned} q(\theta^{(n-1)}, \theta^{(n)}) &= \pi(\theta^{(n)} \left| \theta^{(n-1)} \right.) = \pi(\omega = \theta^{(n)} - \theta^{(n-1)}) \\ &= \pi(\omega = \theta^{(n-1)} - \theta^{(n)}) \text{ (symmetry)} && (6.8.73) \\ &= \pi(\theta^{(n-1)} \left| \theta^{(n)} \right.) = q(\theta^{(n)}, \theta^{(n-1)}) && (6.8.74) \end{aligned}$$

Then (6.8.69) simplifies to:

$$\alpha(\theta^{(n-1)}, \theta^{(n)}) = \begin{cases} \min\left\{1, \frac{\pi(\theta^{(n)})}{\pi(\theta^{(n-1)})}\right\} & \text{if } \pi(\theta^{(n)}) q(\theta^{(n)}, \theta^{(n-1)}) \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{6.8.75}$$

This choice is retained for the transition kernel, which is then defined as:

$$Z^{(n)} = Z^{(n-1)} + \omega \tag{6.8.76}$$

The covariance matrix Ω is set as $\Omega = \psi I_T$, where ψ is some parameter determining the variance of the draw. As previously discussed, the value of ψ must be chosen to achieve a good compromise between the variability of the draw and its acceptance probability. In practice, it is calibrated so as to obtain an acceptance rate of 20-30%, with the adequate value determined by trial and error.

Following, (6.8.68) and (6.8.76) imply that the acceptance probability for the Metropolis-Hastings algorithm is given by:

$$\begin{aligned}
&\alpha(Z^{(n-1)}, Z^{(n)}) = \frac{\pi(Z^{(n)})}{\pi(Z^{(n-1)})} \\
&= \exp\left(-\frac{1}{2}\sum_{t=1}^{T}(y_t - \tilde{X}_t\theta_t)'\tilde{\Sigma}^{-1}(y_t - \tilde{X}_t\theta_t)\left\{\exp(-\zeta_t^{(n)}) - \exp(-\zeta_t^{(n-1)})\right\}\right) \\
&\times \exp\left(-\frac{Nn}{2}\sum_{t=1}^{T}\left\{\zeta_t^{(n)} - \zeta_t^{(n-1)}\right\}\right) \\
&\times \exp\left(-\frac{1}{2}\left\{(Z^{(n)})'\Phi_0^{-1}Z^{(n)} - (Z^{(n-1)})'\Phi_0^{-1}Z^{(n-1)}\right\}\right) \qquad (6.8.77)
\end{aligned}$$

This completes the derivations of the model. It is finally possible to derive the full algorithm for the posterior distribution:

Algorithm 4.8.2 (Gibbs sampling/Metropolis-Hastings algorithm for a time-varying panel VAR model):

1. Define starting values $\theta^{(0)} = \left\{\theta_t^{(0)}\right\}_{t=1}^{T}, b^{(0)} = \left\{b_i^{(0)}\right\}_{i=1}^{r}, \tilde{\Sigma}^{(0)}, \zeta^{(0)} = \left\{\zeta_t^{(0)}\right\}_{t=1}^{T}$ and $\varphi^{(0)}$. For $\theta^{(0)}$, use the long run value $\bar{\theta}$ for all t. For $b^{(0)}$, the value is set to $b_i^{(0)} = 10^5$ for all factors $i = 1, 2, \cdots, r$, which amounts to setting a diffuse prior on θ. For $\tilde{\Sigma}^{(0)}$, use (6.8.19) to obtain $\varepsilon_t = y_t - \tilde{X}_t\theta_t$, then obtain $\tilde{\Sigma}^{(0)} = 1/T\sum_{i=1}^{T}\varepsilon_t\varepsilon_t'$. For $\zeta^{(0)}$, the value is set to $\zeta_t^{(0)} = 0$ for all t, which implies no heteroskedasticity at the initiation of the algorithm. Finally, $\varphi^{(0)}$ is set to 1, which corresponds to a standard normal distribution for the heteroskedasticity disturbance.

2. At iteration n, draw $\tilde{\Sigma}^{(n)}$ from $\pi(\tilde{\Sigma}^{(n)}|y\ , \theta^{(n-1)}, b^{(n-1)}, \zeta^{(n-1)}, \varphi^{(n-1)}) \sim\ IW(\bar{S}, T)$, with: $\bar{S} = \sum_{t=1}^{T}(y_t - \tilde{X}_t\theta_t^{(n-1)})\exp(-\zeta_t^{(n-1)})(y_t - \tilde{X}_t\theta_t^{(n-1)})'$

3. At iteration n, draw $\zeta^{(n)} = \left\{\zeta_t^{(n)}\right\}_{t=1}^{T}$ by using the Metropolis-Hastings algorithm (XXX). The candidate is drawn from: $Z^{(n)} = Z^{(n-1)} + \omega \qquad \text{with} \qquad \omega \sim \mathcal{N}(0, \psi I_T)$

And the acceptance function is:

$$\begin{aligned}
&\alpha(Z^{(n-1)}, Z^{(n)}) \\
&= \exp\left(-\frac{1}{2}\sum_{t=1}^{T}(y_t - \tilde{X}_t\theta_t^{(n-1)})'(\tilde{\Sigma}^{(n)})^{-1}(y_t - \tilde{X}_t\theta_t^{(n-1)})\left\{\exp(-\zeta_t^{(n)}) - \exp(-\zeta_t^{(n-1)})\right\}\right) \\
&\times \exp\left(-\frac{Nn}{2}\sum_{t=1}^{T}\left\{\zeta_t^{(n)} - \zeta_t^{(n-1)}\right\}\right) \\
&\times \exp\left(-\frac{1}{2}\left\{(Z^{(n)})'\Phi_0^{-1}Z^{(n)} - (Z^{(n-1)})'\Phi_0^{-1}Z^{(n-1)}\right\}\right)
\end{aligned}$$

4. At iteration n, draw $\varphi^{(n)}$ from $\pi(\varphi^{(n)} \left| y, \theta^{(n-1)}, b^{(n-1)}, \tilde{\Sigma}^{(n)}, \zeta^{(n)}\right.) \sim\ IG(\frac{\bar{\alpha}}{2}, \frac{\bar{\delta}}{2})$, with:
   $\bar{\alpha} = T + \alpha_0$
   and:
   $\bar{\delta} = (Z^{(n)})'G'GZ^{(n)} + \delta_0$

5. At iteration n, draw $b^{(n)} = \left\{b_i^{(n)}\right\}_{i=1}^{r}$ from $\pi(b_i \left| y, \theta^{(n-1)},\ \tilde{\Sigma}^{(n)}, \zeta^{(n)}, \varphi^{(n)}\right.) \sim\ IG(\frac{\bar{a}_i}{2}, \frac{\bar{b}_i}{2})$, with:
   $\bar{a}_i = Td_i + a_0$
   and:
   $\bar{b}_i = \sum_{t=1}^{T}(\theta_{i,t}^{(n-1)} - \theta_{i,t-1}^{(n-1)})'(\theta_{i,t}^{(n-1)} - \theta_{i,t-1}^{(n-1)}) + b_0$

6. At iteration n, draw $\theta^{(n)} = \left\{\theta_t^{(n)}\right\}_{t=1}^{T}$ from $\pi(\theta \left| y, b^{(n)}, \tilde{\Sigma}^{(n)}, \zeta^{(n)}, \varphi^{(n)}\right.) \sim \mathcal{N}(\bar{\Theta}, \bar{B})$, with:
   $\bar{B} = \left(\tilde{X}'\Sigma^{-1}\tilde{X} + B_0^{-1}\right)^{-1}$
   and:
   $\bar{\Theta} = \bar{B}\left(\tilde{X}'\Sigma^{-1}y + B_0^{-1}\Theta_0\right)$
   where:
   $$\Sigma = \begin{pmatrix} \exp(\zeta_1^{(n)})\tilde{\Sigma}^{(n)} & 0 & \cdots & 0 \\ 0 & \exp(\zeta_2^{(n)})\tilde{\Sigma}^{(n)} & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & \exp(\zeta_T^{(n)})\tilde{\Sigma}^{(n)} \end{pmatrix}$$
   and
   $B_0 = H^{-1}\left(I_T \otimes B^{(n)}\right)(H^{-1})'$

7. At iteration n, recover $\Sigma^{(n)}$ from:
   $\Sigma_t^{(n)} = \exp(\zeta_t^{(n)})\tilde{\Sigma}^{(n)}$

This concludes the model.

6.9 Applications with panel VARs: forecasts, impulse response functions, forecast error variance decomposition

Producing forecasts with Bayesian panel VARs is straightforward. Indeed, any panel VAR ultimately results in the estimation of a (set of) standard VAR models. In the case of the OLS mean-group estimator and Bayesian pooled estimator, a single homogenous model is estimated for all the units. In the case of the random effect model (Zellner and Hong or hierarchical), a set of N independent VAR models are obtained. Finally, for the static factor model, a single large VAR model in which all units are interacting is estimated. As all these models are of standard form, producing forecasts can be realised by direct application of the forecast algorithm 2.1.1 for regular Bayesian VAR models, and the presentation will not enter into further details.

The case of the dynamic factor model however is more complicated. In this case, it is not possible to apply naively the methodology proposed for a static model since the usual methodology ignores the possibility that the VAR coefficients and residual covariance matrix may evolve across time. It is however possible to extend directly the static approach to time-varying models by integrating the laws of motion for the parameters into the sampling process. The equations of interest for the time-varying factor approach are stated again for the sake of convenience:

$$y_t = \bar{X}_t \beta_t + \varepsilon_t \tag{6.9.1}$$

$$\varepsilon_t \sim \mathcal{N}(0, \Sigma_t) \qquad \text{with} \qquad \Sigma_t = \exp(\zeta_t)\tilde{\Sigma} \tag{6.9.2}$$

$$\zeta_t = \gamma \zeta_{t-1} + \upsilon_t \qquad \text{with} \qquad \upsilon_t \sim \mathcal{N}(0, \varphi) \tag{6.9.3}$$

$$\beta_t = \Xi \theta_t \tag{6.9.4}$$

$$\theta_t = (1-\rho)\bar{\theta} + \rho\theta_{t-1} + \eta_t \qquad \text{with} \qquad \eta_t \sim \mathcal{N}(0, B) \tag{6.9.5}$$

$\tilde{\Sigma}, \varphi$ and B are static parameters for which posterior draws have been obtained from the Gibbs sampling process. The trick then consists in noticing that thanks to φ and B it is possible to draw series of disturbances υ_t and η_t for any forecast period, and then to use the laws of motion (6.9.3) and (6.9.5) to evaluate sequentially $\zeta_{T+1}, ..., \zeta_{T+h}$ and $\theta_{T+1}, ..., \theta_{T+h}$. Concretely, the following adaptation of algorithm 2.1.1 to dynamic models is developed:

Algorithm 4.9.1 (forecasts with a panel VAR model (dynamic factor approach):

1. define the number of iterations $(It - Bu)$ of the algorithm, and the forecast horizon h.

2. set the period to $T + 1$.

3. at iteration n, draw $\Sigma_{(n)}$ from its posterior distributions. To do so, recycle draw n from the Gibbs sampler.

4. at iteration n, draw $\varphi_{(n)}$ from its posterior distributions. To do so, recycle draw n from the Gibbs sampler. Then draw $\upsilon_{T+1}^{(n)}$ from $\upsilon_{T+1} \sim \mathcal{N}(0, \varphi_{(n)})$. Finally, obtain $\zeta_{T+1}^{(n)}$ from $\zeta_{T+1}^{(n)} = \gamma\zeta_{T}^{(n)} + \upsilon_{T+1}^{(n)}$.

5. obtain $\Sigma_{T+1}^{(n)} = \exp(\zeta_{T+1}^{(n)})\tilde{\Sigma}_{(n)}$. Draw the simulated residual $\tilde{\varepsilon}_{T+1}^{(n)}$.

6. at iteration n, draw $B_{(n)}$ from its posterior distributions. To do so, recycle draw n from the Gibbs sampler. Then draw $\eta_{T+1}^{(n)}$ from $\eta_{T+1}^{(n)} \sim \mathcal{N}(0, B_{(n)})$. Finally, obtain $\theta_{T+1}^{(n)}$ from $\theta_{T+1}^{(n)} = (1 - \rho)\bar{\theta} + \rho\theta_{T}^{(n)} + \eta_{T+1}^{(n)}$.

7. at iteration n, obtain $\beta_{T+1}^{(n)} = \Xi\theta_{T+1}^{(n)}$.

8. at iteration n, obtain $y_{T+1}^{(n)} = \bar{X}_{T+1}\beta_{T+1}^{(n)} + \varepsilon_{T+1}^{(n)}$

9. repeat steps 3-9 for $T + 2, \cdots, T + h$.

10. repeat steps 2-10 until $(It - Bu)$ iterations are realised. This produces:
$\left\{\tilde{y}_{T+1}^{(n)} | y_T, \tilde{y}_{T+2}^{(n)} | y_T, ..., \tilde{y}_{T+h}^{(n)} | y_T\right\}_{n=1}^{It-Bu}$
a sample of independent draws from the joint predictive distribution which can be used for inference and computation of point estimates.

The logic for impulse response functions and forecast error variance decomposition is rigorously similar to that of forecasts and is not developed further. In the case of impulse response functions standard identification schemes apply. Two standard choices are the usual Choleski and triangular factorisation.

6.10 Historical decomposition

Similarly to the previous applications, historical decomposition is straightforward to estimate for the first five panel VAR models developed so far, simply following the standard Bayesian VAR methodology. For the dynamic factor model, however, the time-varying properties of the impulse

response functions involve some adaptation of the traditional methodology. Start again from the general panel VAR model (6.1.5):

$$y_t = A_t^1 y_{t-1} + \cdots + A_t^p y_{t-p} + C_t x_t + \varepsilon_t \tag{6.10.1}$$

For the present purpose, it is convenient to reformulate the model in companion form, that is, to reformulate it as a VAR(1) model:

$$\underbrace{\begin{pmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t+2-p} \\ y_{t+1-p} \end{pmatrix}}_{Nnp\times 1} = \underbrace{\begin{pmatrix} A_t^1 & A_t^2 & \cdots & A_t^{p-1} & A_t^p \\ I_{Nn} & 0 & \cdots & 0 & 0 \\ 0 & I_{Nn} & & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I_{Nn} & 0 \end{pmatrix}}_{Nnp\times Nnp} \underbrace{\begin{pmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t+1-p} \\ y_{t-p} \end{pmatrix}}_{Nnp\times 1} + \underbrace{\begin{pmatrix} C_t \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}}_{Nnp\times m} \underbrace{(x_t)}_{m\times 1} + \underbrace{\begin{pmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}}_{Nnp\times 1} \tag{6.10.2}$$

or:

$$\bar{y}_t = A_t \bar{y}_{t-1} + \bar{C}_t \bar{x}_t + \bar{\varepsilon}_t \tag{6.10.3}$$

The advantage of formulation (6.10.3) over formulation (6.10.1) is that an AR(1) model makes it a lot easier to solve for past values by backward substitution. Thus, use (6.10.3) to obtain a general formulation for $\bar{y}_t$:

$$\begin{aligned} \bar{y}_t &= A_t \bar{y}_{t-1} + \bar{C}_t \bar{x}_t + \bar{\varepsilon}_t \\ &= A_t (A_{t-1} \bar{y}_{t-2} + \bar{C}_{t-1} \bar{x}_{t-1} + \bar{\varepsilon}_{t-1}) + \bar{C}_t \bar{x}_t + \bar{\varepsilon}_t \\ &= A_t A_{t-1} \bar{y}_{t-2} + A_t \bar{C}_{t-1} \bar{x}_{t-1} + \bar{C}_t \bar{x}_t + A_t \bar{\varepsilon}_{t-1} + \bar{\varepsilon}_t \end{aligned}$$

Go one step further:

$$\begin{aligned} \bar{y}_t &= A_t A_{t-1} \bar{y}_{t-2} + A_t \bar{C}_{t-1} \bar{x}_{t-1} + \bar{C}_t \bar{x}_t + A_t \bar{\varepsilon}_{t-1} + \bar{\varepsilon}_t \\ &= A_t A_{t-1} (A_{t-2} \bar{y}_{t-3} + \bar{C}_{t-2} \bar{x}_{t-2} + \bar{\varepsilon}_{t-2}) + A_t \bar{C}_{t-1} \bar{x}_{t-1} + \bar{C}_t \bar{x}_t + A_t \bar{\varepsilon}_{t-1} + \bar{\varepsilon}_t \\ &= A_t A_{t-1} A_{t-2} \bar{y}_{t-3} + A_t A_{t-1} \bar{C}_{t-2} \bar{x}_{t-2} + A_t \bar{C}_{t-1} \bar{x}_{t-1} + \bar{C}_t \bar{x}_t + A_t A_{t-1} \bar{\varepsilon}_{t-2} + A_t \bar{\varepsilon}_{t-1} + \bar{\varepsilon}_t \end{aligned}$$

Going on this way, one recovers the general formulation:

$$\bar{y}_t = \left(\prod_{i=0}^{t-1} \bar{A}_{t-i} \right) \bar{y}_0 + \sum_{i=0}^{t-1} \left(\prod_{j=1}^{i} \bar{A}_{t+1-j} \right) \bar{C}_{t-i} \bar{x}_{t-i} + \sum_{i=0}^{t-1} \left(\prod_{j=1}^{i} \bar{A}_{t+1-j} \right) \bar{\varepsilon}_{t-i} \tag{6.10.4}$$

Note that similarly to 5.2.10, the first two terms represent the contribution of deterministic variables (exogenous variables and initial conditions), while the final term represents the contribution of the residuals:

$$\bar{y}_t = \underbrace{\left(\prod_{i=0}^{t-1} A_{t-i}\right)\bar{y}_0 + \sum_{i=0}^{t-1}\left(\prod_{j=1}^{i} A_{t+1-j}\right)\bar{C}_{t-i}\bar{x}_{t-i}}_{\text{historical contribution of deterministic variables}} + \underbrace{\sum_{i=0}^{t-1}\left(\prod_{j=1}^{i} A_{t+1-j}\right)\bar{\varepsilon}_{t-i}}_{\substack{\text{historical contribution} \\ \text{of residuals}}} \tag{6.10.5}$$

It is thus convenient to simplify (6.10.5) by defining the whole deterministic contribution as $\bar{d}^{(t)}$, and then rewrite as:

$$\bar{y}_t = \bar{d}^{(t)} + \sum_{i=0}^{t-1} \bar{\Psi}_{t,i}\bar{\varepsilon}_{t-i} \tag{6.10.6}$$

However, $\bar{y}_t, \bar{d}^{(t)}$ and $\bar{\Psi}_{t,i}$ represents elements of model (6.10.3), while what is of interest is the original model (6.10.1). In order to recover the original elements, use the following selection matrix:

$$J = \underbrace{\begin{pmatrix} I_{Nn} & 0 & 0 & \cdots & 0 \end{pmatrix}}_{Nn \times Nnp} \tag{6.10.7}$$

Then note that the original elements can be recovered from:

$$y_t = J\bar{y}_t \qquad d^{(t)} = J\bar{d}^{(t)} \qquad \Psi_{t,i} = J\bar{\Psi}_{t,i}J' \qquad \varepsilon_{t-i} = J\bar{\varepsilon}_{t-i} \tag{6.10.8}$$

Therefore, premultiplying both sides of (6.10.6) by J and manipulating, this rewrites as:

$$\bar{y}_t = \bar{d}^{(t)} + \sum_{i=0}^{t-1} \bar{\Psi}_{t,i}\bar{\varepsilon}_{t-i} \Leftrightarrow J\bar{y}_t = J\bar{d}^{(t)} + \sum_{i=0}^{t-1} J\bar{\Psi}_{t,i}JJ'\bar{\varepsilon}_{t-i} \Leftrightarrow y_t = d^{(t)} + \sum_{i=0}^{t-1} \Psi_{t,i}\varepsilon_{t-i} \tag{6.10.9}$$

with the series of period-specific impulse response functions recovered from:

$$\Psi_{t,i} = J\left(\prod_{j=1}^{i} A_{t+1-j}\right)J' \tag{6.10.10}$$

Also, using the period-specific structural matrix D_t, recovered from the period-specific residual covariance matrix Σ_t, one obtains:

$$\Psi_{t,i}\varepsilon_{t-i} = \Psi_{t,i}D_tD_t^{-1}\varepsilon_{t-i} = \tilde{\Psi}_{t,i}\eta_{t-i} \tag{6.10.11}$$

And (6.10.9) reformulates as:

$$y_t = d^{(t)} + \sum_{i=0}^{t-1} \tilde{\Psi}_{t,i} \eta_{t-i} \tag{6.10.12}$$

This formulation is similar to 5.2.11, except for the facts that the impulse response functions are now period-specific. (6.10.12) can thus be used to recover the historical decomposition in a standard way, simply accounting for time-specificity. The following algorithm is thus proposed:

Algorithm 4.10.1 (historical decomposition for the dynamic factor panel model):

1. define the number of iterations $(It - Bu)$ of the algorithm. Then run the algorithm:
2. At iteration n, draw $\theta_t^{(n)}$ and $\Sigma_t^{(n)}$ from their posterior distributions, for $t = 1, 2, \cdots, T$. Simply recycle draws from the Gibbs sampler.
3. Recover $\beta_t^{(n)}$ from $\beta_t^{(n)} = \Xi \theta_t^{(n)}$, and use it to obtain A_t from (6.10.2), for $t = 1, 2, \cdots, T$.
4. Calculate $\Psi_{t,i} = J \left(\prod_{j=1}^{i} A_{t+1-j} \right) J'$, for $t = 1, 2, \cdots, T$.
5. Obtain the structural matrix D_t from Σ_t, and compute $\tilde{\Psi}_{t,i} = \Psi_{t,i} D_t$, for $t = 1, 2, \cdots, T$.
6. Obtain the historical decomposition in a regular way, using 6.10.12.

6.11 Conditional forecasts

Estimating conditional forecasts in the context of panel VAR models results in a situation which is similar to that of historical decomposition: conditional forecasts can be computed in a regular way for the first five models, since the latter are nothing but regular VAR models (or groups thereof). For the last panel model, however, the dynamic heterogeneity property implies that the coefficients are time varying so that impulse response functions become period-specific. This leads to apply some modifications to the traditional methodology.

Start again from the general panel VAR model (6.1.5):

$$y_t = A_t^1 y_{t-1} + \cdots + A_t^p y_{t-p} + C_t x_t + \varepsilon_t \tag{6.11.1}$$

Similarly to the estimation of historical decomposition, it is convenient to work with a model in companion form. Thus, reformulate (6.11.1) as a VAR(1) model:

$$\underbrace{\begin{pmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t+2-p} \\ y_{t+1-p} \end{pmatrix}}_{Nnp\times 1} = \underbrace{\begin{pmatrix} A_t^1 & A_t^2 & \cdots & A_t^{p-1} & A_t^p \\ I_{Nn} & 0 & \cdots & 0 & 0 \\ 0 & I_{Nn} & & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I_{Nn} & 0 \end{pmatrix}}_{Nnp\times Nnp} \underbrace{\begin{pmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t+1-p} \\ y_{t-p} \end{pmatrix}}_{Nnp\times 1} + \underbrace{\begin{pmatrix} C_t \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}}_{Nnp\times m} \underbrace{(x_t)}_{m\times 1} + \underbrace{\begin{pmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}}_{Nnp\times 1} \tag{6.11.2}$$

or:

$$\bar{y}_t = A_t \bar{y}_{t-1} + \bar{C}_t \bar{x}_t + \bar{\varepsilon}_t \tag{6.11.3}$$

Then consider forecasting for period $T+h$. Rather than iterating backward as in the case of historical decomposition, forward iteration is used. Consider period $T+1$:

$$\bar{y}_{T+1} = A_{T+1}\bar{y}_t + \bar{C}_{T+1}\bar{x}_{T+1} + \bar{\varepsilon}_{T+1} \tag{6.11.4}$$

Iterate one step forward to obtain the value at period $T+2$:

$$\begin{aligned} \bar{y}_{T+2} &= A_{T+2}\bar{y}_{T+1} + \bar{C}_{T+2}\bar{x}_{T+2} + \bar{\varepsilon}_{T+2} \\ &= A_{T+2}\left(A_{T+1}\bar{y}_T + \bar{C}_{T+1}\bar{x}_{T+1} + \bar{\varepsilon}_{T+1}\right) + \bar{C}_{T+2}\bar{x}_{T+2} + \bar{\varepsilon}_{T+2} \\ &= A_{T+2}A_{T+1}\bar{y}_T + A_{T+2}\bar{C}_{T+1}\bar{x}_{T+1} + \bar{C}_{T+2}\bar{x}_{T+2} + A_{T+2}\bar{\varepsilon}_{T+1} + \bar{\varepsilon}_{T+2} \end{aligned} \tag{6.11.5}$$

Iterate again one step forward to obtain an expression for period $T+3$:

$$\begin{aligned} \bar{y}_{T+3} &= A_{T+3}\bar{y}_{T+2} + \bar{C}_{T+3}\bar{x}_{T+3} + \bar{\varepsilon}_{T+3} \\ &= A_{T+3}\left(A_{T+2}A_{T+1}\bar{y}_T + A_{T+2}\bar{C}_{T+1}\bar{x}_{T+1} + \bar{C}_{T+2}\bar{x}_{T+2} + A_{T+2}\bar{\varepsilon}_{T+1} + \bar{\varepsilon}_{T+2}\right) \\ &+ \bar{C}_{T+3}\bar{x}_{T+3} + \bar{\varepsilon}_{T+3} \\ &= A_{T+3}A_{T+2}A_{T+1}\bar{y}_T + A_{T+3}A_{T+2}\bar{C}_{T+1}\bar{x}_{T+1} + A_{T+3}\bar{C}_{T+2}\bar{x}_{T+2} + \bar{C}_{T+3}\bar{x}_{T+3} \\ &+ A_{T+3}A_{T+2}\bar{\varepsilon}_{T+1} + A_{T+3}\bar{\varepsilon}_{T+2} + \bar{\varepsilon}_{T+3} \end{aligned} \tag{6.11.6}$$

Going on this way, one obtains a general formula:

$$\bar{y}_{T+h} = \left(\prod_{i=1}^{h} A_{t+1+h-i}\right)\bar{y}_T + \sum_{i=1}^{h}\left(\prod_{j=i}^{h-1} A_{T+h+i-j}\right)\bar{C}_{T+i}\bar{x}_{T+i} + \sum_{i=1}^{h}\left(\prod_{j=i}^{h-1} A_{T+h+i-j}\right)\bar{\varepsilon}_{T+i} \tag{6.11.7}$$

The first two terms give the predicted values in the absence of shocks and hence represent the regular forecasts. The final term represents the dynamic impact of past residuals at period $T+h$. It then represents the series of impulse response functions specific to this period. Therefore, (6.11.7) can be written as:

$$\bar{y}_{T+h} = \underbrace{\left(\prod_{i=1}^{h} A_{t+1+h-i}\right)\bar{y}_T + \sum_{i=1}^{h}\left(\prod_{j=i}^{h-1} A_{T+h+i-j}\right)\bar{C}_{T+i}\bar{x}_{T+i}}_{\text{Forecast in the absence of shocks}} + \underbrace{\sum_{i=1}^{h}\left(\prod_{j=i}^{h-1} A_{T+h+i-j}\right)\bar{\varepsilon}_{T+i}}_{\text{Dynamic impact of residuals}} \tag{6.11.8}$$

or:

$$\bar{y}_{T+h} = \tilde{\bar{y}}_{T+h} + \sum_{i=1}^{h} \bar{\Psi}_{T+h,h-i}\bar{\varepsilon}_{T+i} \tag{6.11.9}$$

Because this formula provides values for $\bar{y}_{T+h}$, while only y_{T+h} is of interest, define the selection matrix:

$$J = \underbrace{\begin{pmatrix} I_{Nn} & 0 & 0 & \cdots & 0 \end{pmatrix}}_{Nn \times Nnp} \tag{6.11.10}$$

Then note that:

$$\begin{aligned} \bar{y}_{T+h} &= \tilde{\bar{y}}_{T+h} + \sum_{i=1}^{h} \bar{\Psi}_{T+h,h-i}\bar{\varepsilon}_{T+i} \\ \Leftrightarrow J\bar{y}_{T+h} &= J\tilde{\bar{y}}_{T+h} + \sum_{i=1}^{h} J\bar{\Psi}_{T+h,h-i}J'J\bar{\varepsilon}_{T+i} \\ \Leftrightarrow y_{T+h} &= \tilde{y}_{T+h} + \sum_{i=1}^{h} \Psi_{T+h,h-i}\varepsilon_{T+i} \end{aligned} \tag{6.11.11}$$

Therefore, the regular forecasts can be recovered from:

$$\tilde{y}_{T+h} = J\left[\left(\prod_{i=1}^{h} A_{t+1+h-i}\right)\bar{y}_T + \sum_{i=1}^{h}\left(\prod_{j=i}^{h-1} A_{T+h+i-j}\right)\bar{C}_{T+i}\bar{x}_{T+i}\right] \tag{6.11.12}$$

And the series of period-specific impulse response functions obtains from:

$$\Psi_{T+h,h-i} = J\left(\prod_{j=i}^{h-1} A_{T+h+i-j}\right)J' \tag{6.11.13}$$

Also, using the period-specific structural matrix D_{T+h}, recovered from the period-specific residual covariance matrix Σ_{T+h}, one obtains:

$$\Psi_{T+h,h-i}\varepsilon_{T+i} = \Psi_{T+h,h-i}D_{T+h}D_{T+h}^{-1}\varepsilon_{T+i} = \tilde{\Psi}_{T+h,h-i}\eta_{T+i} \tag{6.11.14}$$

And (6.11.11) reformulates as:

$$y_{T+h} = \tilde{y}_{T+h} + \sum_{i=1}^{h} \tilde{\Psi}_{T+h,h-i}\eta_{T+i} \tag{6.11.15}$$

It can be seen that (6.11.15) is similar to 5.4.1, save for the fact that the impulse response functions are period-specific. Therefore, (6.11.15) can be used to recover conditional forecasts in a standard way, simply accounting for time-specificity. The following algorithm is thus proposed:

Algorithm 4.11.1 (conditional forecasts for the dynamic factor panel model):

1. define the number of iterations $(It - Bu)$ of the algorithm. Then run the algorithm:
2. At iteration n, draw $\theta_T^{(n)}, \zeta_T^{(n)}, \tilde{\Sigma}^{(n)}$ and $B^{(n)}$ from their posterior distributions. Simply recycle draws from the Gibbs sampler.
3. At iteration n, obtain recursively $\theta_{T+i}^{(n)}$ from:
   $\theta_{T+i}^{(n)} = (1-\rho)\bar{\theta} + \rho\theta_{T+i-1}^{(n)} + \eta_{T+i}^{(n)}$ with: $\eta_{T+i}^{(n)} \sim \mathcal{N}(0, B^{(n)})$, for $i = 1, 2, \ldots, h$.
4. At iteration n, obtain $\beta_{T+i}^{(n)} = \Xi\theta_{T+i}^{(n)}$ and use it to obtain $A_{T+i}^{(n)}$ and $\bar{C}_{T+i}^{(n)}$ from (6.11.2), for $i = 1, 2, \cdots, h$.
5. At iteration n, obtain the unconditional forecast $\tilde{y}_{T+h}^{(n)}$ from:
   $$\tilde{y}_{T+h}^{(n)} = J\left[\left(\prod_{i=1}^{h} A_{t+1+h-i}^{(n)}\right)\bar{y}_T + \sum_{i=1}^{h}\left(\prod_{j=i}^{h-1} A_{T+h+i-j}^{(n)}\right)\bar{C}_{T+i}^{(n)}\bar{x}_{T+i}\right]$$
6. At iteration n, obtain the series of impulse response functions $\Psi_{T+h,h-i}$, from:
   $$\Psi_{T+h,h-i}^{(n)} = J\left(\prod_{j=i}^{h-1} A_{T+h+i-j}^{(n)}\right)J' \text{ , for } i = 1, 2, \ldots, h$$
7. At iteration n, obtain recursively $\zeta_{T+i}^{(n)}$ from:
   $\zeta_{T+i}^{(n)} = \gamma\zeta_{T+i-1}^{(n)} + \upsilon_{T+i}^{(n)}$ with $\upsilon_{T+i}^{(n)} \sim \mathcal{N}(0, \varphi^{(n)})$, for $i = 1, 2, \ldots, h$
8. At iteration n, obtain $\Sigma_{T+h}^{(n)} = \exp(\zeta_{T+h}^{(n)})\tilde{\Sigma}^{(n)}$, then obtain the structural matrix $D_{T+h}^{(n)}$ from $\Sigma_{T+h}^{(n)}$.

9. At iteration n, obtain the structural impulse response functions $\tilde{\Psi}_{T+h,h-i}$ from: $\tilde{\Psi}_{T+h,h-i} = \Psi_{T+h,h-i} D_{T+h}$, for $i = 1, 2, \ldots, h$

10. Obtain the conditional forecasts in a regular way, using (6.11.15).

7 Summary and Conclusions

Bayesian econometrics has now become a very dynamic field, with innovative research as well as promising applications being released on a regular basis. We have created BEAR by believing that it would be a dynamic tool and could always remain at the frontier of current econometric research. While BEAR offers already a wide range of applications, we believe there is still room for improvement. Significant contributions have been recently produced in terms of Bayesian VAR modeling, and these developments should be integrated to BEAR at some point.

Two fields seem particularly attractive. The first one is the category of mixed frequency models which has been recently developing at a fast pace, mainly under the motivation of improved forecast performances. Mixed frequency models are indeed constructed to integrate high frequency data within lower frequency frameworks, allowing to update forecasts as soon as new data is released. This may represent a significant advance for economists for whom now-casting and short-run forecasting represent a central concern. The approach is now getting more firmly established, see for instance Schorfheide and Song (2016). The second field is that of time-varying models. Economists now consider seriously the possibility that the dynamic process of a model may change over time, which motivates the development of such approaches. Research is still on-going and alternative Bayesian methodologies currently coexist: see for instance Primiceri (2005), the state-space approach of Durbin and Koopman (2002), or the sparse matrix approach of Chan and Eisenstat (2015). Other methodologies are worth mentioning and could be profitably integrated to BEAR. Error correction models remain relevant, all the more since it is often customary to use macroeconomic data in log levels. Threshold models are useful to account for non-linearities in the data, and this issue has become prominent since the financial crisis. Markov-switching models could also represent an option to model data characterised by non-linear behaviour.

Keeping in mind these possible developments, the main objectives of BEAR remain unchanged: providing an easy access to innovative Bayesian applications to the widest possible audience, from expert practitioners to newcomers in Bayesian methods.

References

Arias, J. E., Rubio-Ramirez, J. F., and Waggoner, D. F. (2014). Inference Based on SVARs Identified with Sign and Zero Restrictions: Theory and Applications. Dynare Working Papers 30, CEPREMAP.

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

Bernstein, D. (2005). *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton University Press.

Canova, F. and Ciccarelli, M. (2006). Estimating multi-country VAR models. Working Paper Series 603, European Central Bank.

Canova, F. and Ciccarelli, M. (2013). Panel vector autoregressive models: a survey. Working Paper Series 1507, European Central Bank.

Casella, G. and Berger, R. (2001). *Statistical Inference*. Duxbury Press, 2 edition.

Chan, J. and Eisenstat, E. (2015). Bayesian model comparison for time-varying parameter vars with stochastic volatility. Working Paper 32/2015, CAMA.

Chan, J. and Jeliazkov, I. (2009). Efficient Simulation and Integrated Likelihood Estimation in State Space Models . *International Journal of Mathematical Modelling and Numerical Optimisation*, 1(1-2):101–120.

Chib, S. (1995). Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association*, 90(432):1313–1321.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49(4):327–335.

Christiano, L. J., Eichenbaum, M., and Evans, C. L. (1999). Monetary policy shocks: What have we learned and to what end? In Taylor, J. B. and Woodford, M., editors, *Handbook of Macroeconomics*, volume 1 of *Handbook of Macroeconomics*, chapter 2, pages 65–148. Elsevier.

Ciccarelli, M., Ortega, E., and Valderrama, M. T. (2012). Heterogeneity and cross-country spillovers in macroeconomic-financial linkages. Working Papers 1241, Banco de Espana;Working Papers Homepage.

Doan, T., Litterman, R., and Sims, C. (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews*, 3(1):1–100.

Durbin, J. and Koopman, S. J. (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, 89(3):603–616.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–534.

Georgiadis, G. (2015). To Bi, or not to Bi? Differences in Spillover Estimates from Bilateral and Multilateral Multi-country Models. Working Paper 256, Federal Reserve Bank of Dallas.

Giannone, D., Lenza, M., and Primiceri, G. E. (2012). Prior Selection for Vector Autoregressions. Working Papers ECARES ECARES 2012-002, ULB – Universite Libre de Bruxelles.

Giannone, D., Lenza, M., and Primiceri, G. E. (2015). Prior selection for vector autoregressions. *Review of Economics and Statistics*, 27(2):436–451.

Gneiting, T. and Raftery, A. (2007). Strict Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477).

Gordon, N., Salmond, D., and Smith, A. (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEEE Proceedings F, Radar and Signal Processing*, 140(2):107–113.

Greenberg, E. (2008). *Introduction to Bayesian econometrics*. Cambridge University Press.

Gupta, A. K. and Nagar, D. K. (1999). *Matrix Variate Distributions*. Chapman & Hall/CRC.

Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press.

Jarocinski, M. (2010a). Conditional forecasts and uncertainty about forecast revisions in vector autoregressions. *Economics Letters*, 108(3):257–259.

Jarocinski, M. (2010b). Responses to monetary policy shocks in the east and the west of Europe: a comparison. *Journal of Applied Econometrics*, 25(5):833–868.

Jeffreys, H. (1961). *Theory of Probability*. Clarendon Press, Oxford, 3 edition.

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

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

Litterman, R. (1986). Forecasting with Bayesian Vector Autoregressions – Five years of experience : Robert b. Litterman, Journal of Business and Economic Statistics 4 (1986) 25-38. *International Journal of Forecasting*, 2(4):497–498.

Luetkepohl, H. (1993). *Introduction to Multiple Time Series Analysis*. Springer - Verlag, 2st ed. 2006. corr. 2nd printing edition.

Mao, W. (2010). Bayesian multivariate predictions. Phd thesis, University of Iowa.

Marsaglia, G. and Tsang, W. W. (2000). A simple method for generating gamma variables. *ACM Transactions on Mathematical Software*, 26(3):363–372.

Matheson, J. and Winkler, R. (1976). Scoring Rules for Continuous Probability Distributions. *Management Science*, 22:1087–1096.

Pesaran, H. and Smith, R. (1995). Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics*, 1:79–113.

Pozrikidis, C. (2014). *An Introduction to Grids, Graphs, and Networks*. Oxford University Press.

Primiceri, G. (2005). Time-varying structural vector autoregressions and monetary policy. *Review of Economic Studies*, 72:821–852.

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

Schorfheide, F. and Song, D. (2016). Real-time forecasting with a mixed-frequency var. *Journal of Business and Economic Statistics*, pages 1–30.

Simon, C. and Blume, L. (1994). *Mathematics for economists*. W. W. Norton & Company.

Sims, C. and Zha, T. (1997). Bayesian Methods for Dynamic Multivariate Models. Frb atlanta working paper, Federal Reserve Bank of Atlanta.

Sims, C. A. (1992). A Nine Variable Probabilistic Macroeconomic Forecasting Model. Cowles Foundation Discussion Papers 1034, Cowles Foundation for Research in Economics, Yale University.

Villani, M. (2005). Inference in Vector Autoregressive Models with an Informative Prior on the Steady-State. Technical report, Sveriges Riksbank.

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

Waggoner, D. F. and Zha, T. (1999). Conditional Forecasts In Dynamic Multivariate Models. *The Review of Economics and Statistics*, 81(4):639–651.

Warne, A., Coenen, G., and Christoffel, K. (2013). Marginalised Predictive Likelihood Comparisons with Application to DSGE, DSGE-VAR, and VAR Models. Technical report, Center for Financial Studies.

Zellner, A. and Hong, C. (1989). Forecasting international growth rates using bayesian shrinkage and other procedures. *Journal of Econometrics*, 40(1):183–202.

A Appendix

A.1 Preliminary mathematical results

While the derivations of the different posterior distributions are often not difficult intrinsically, they rely heavily on certain results related to linear algebra. This appendix hence states these results, in order to make the incoming derivations easier to read. Most results are standard, so that their proofs are omitted. For further details on the derivations of these results, one may consult mathematics textbooks such as Bernstein (2005) or Simon and Blume (1994). Proofs are developed for non standard results.

Results related to Kronecker products:

$$(A \otimes B)' = A' \otimes B' \tag{A.1.1}$$

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1} \tag{A.1.2}$$

For matrices A, B, C and D, such that AC and BD are defined:

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD) \tag{A.1.3}$$

For a matrix A of dimension $n \times n$, and a matrix B of dimension $p \times p$, one has:

$$|A \otimes B| = |A|^p |B|^n \tag{A.1.4}$$

$$vec(ABC) = (C' \otimes A)vec(B) \tag{A.1.5}$$

Proof: See Hamilton (1994), proposition 10.4 and proof p 289.

Results related to the trace of a square matrix:

$$tr(A + B) = tr(A) + tr(B) \tag{A.1.6}$$

$$tr(ABCD) = tr(BCDA) = tr(CDAB) = tr(DABC) \quad \text{(cyclical property)} \tag{A.1.7}$$

$$tr(A'B) = vec(A)' \times vec(B) \tag{A.1.8}$$

Results related to vectorisation:

$$vec(A+B) = vec(A) + vec(B) \tag{A.1.9}$$

For matrices V, X, M, U, Y and N, such that V is $n \times n$ and symmetric, U is $k \times k$, and X, M, Y and N are $k \times n$, one has:

$$tr\left\{V^{-1}(X-M)'U^{-1}(Y-N)\right\} = (vec(X) - vec(M))'(V \otimes U)^{-1}(vec(Y) - vec(N)) \tag{A.1.10}$$

Proof :

$$\begin{aligned} tr\left\{V^{-1}(X-M)'U^{-1}(Y-N)\right\} &= tr\left\{(X-M)'U^{-1}(Y-N)V^{-1}\right\} \text{ A.17.15} \\ &= vec(X-M)' \times vec\left(U^{-1}(Y-N)V^{-1}\right) \text{ A.1.8} \\ &= vec(X-M)' \times (V^{-1} \otimes U^{-1}) vec(Y-N) \text{ A.1.5 and symmetry} \\ &= (vec(X) - vec(M))'(V^{-1} \otimes U^{-1})(vec(Y) - vec(N)) \text{ A.1.9} \\ &= (vec(X) - vec(M))'(V \otimes U)^{-1}(vec(Y) - vec(N)) \text{ A.1.2} \end{aligned}$$

Results related to determinants and eigenvalues:

Let A be a triangular matrix. Then the determinant of A is equal to the product of its diagonal terms. (A.1.11)

Corollary:

Let A be a diagonal matrix. Then the determinant of A is equal to the product of its diagonal terms. (A.1.12)

Let A and B be $k \times k$ matrices. Then:

$$|AB| = |A|\,|B| \tag{A.1.13}$$

Let c be any scalar and A be a $k \times k$ matrix. Then

$$|cA| = c^k\,|A| \tag{A.1.14}$$

Let A be a $k \times k$ matrix. Then

$$\left|A^{-1}\right| = |A|^{-1} \tag{A.1.15}$$

Let B be a $k \times m$ matrix, and C be a $m \times k$ matrix. Then:

$$|I_k + BC| = |I_m + CB| \quad \text{by Sylvesters determinant theorem.} \tag{A.1.16}$$

generalisation:
Let A be any invertible $k \times k$ matrix, B be a $k \times m$ matrix, and C be a $m \times k$ matrix. Then: $|A + BC| = |A| \, . \, |I_m + CA^{-1}B|$
proofs: see Pozrikidis (2014) p 271.

Let A be a $k \times k$ matrix. If λ is an eigenvalue of A , then $\lambda + t$ is an eigenvalue of

$$A + tI_k, \tag{A.1.17}$$

where t is some scalar. When $t = 1$, this says that if λ is an eigenvalue of A , then $\lambda + 1$ is an eigenvalue of $A + I_k$.
proof: because λ is an eigenvalue of A, there exists some eigenvector u such that $Au = \lambda u$. Then: $Au = \lambda u \Rightarrow Au + tu = \lambda u + tu \Rightarrow Au + tI_k u = \lambda u + tu \Rightarrow (A + tI_k)u = (\lambda + t)u$. Therefore, by definition, $\lambda + t$ is an eigenvalue of $A + tI_k$.

- let A be a $k \times k$ symmetric positive definite matrix. Then:

$$|I_k + A| = \prod_{i=1}^{k} (1 + \lambda_i(A)) \tag{A.1.18}$$

where $\lambda_i(A)$ denotes the i^{th} eigenvalue of A. In other words, the determinant of $I_k + A$ can be obtained from the product of 1 plus the eigenvalues of A. proof: because the determinant of a matrix is equal to the product of its eigenvalues, one can write:
$|I_k + A| = \prod_{i=1}^{k} \lambda_i \left(I_k + A\right)$ where $\lambda_i \left(I_k + A\right)$ denotes the i^{th} eigenvalue of $I_k + A$
$= \prod_{i=1}^{k} \left(1 + \lambda_i \left(A\right)\right)$ by direct application of A.1.17

A.2 Statistical distributions

Bayesian analysis in general may rely on a very large number of different statistical distributions. This section does not aim at being exhaustive and only provides a brief introduction for the families of distribution used at some point in this guide. Proofs for the results do not constitute the main object of this guide and are thus omitted. For a more formal presentation, textbooks such as Casella and Berger (2001) can be consulted.

A.2.1 Uniform distribution

The scalar random variable x follows a continuous uniform distribution over the interval $[a, b]$:

$$x \sim U(a, b) \tag{A.2.1.1}$$

if its density is given by:

$$f(x\,|a, b) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases} \tag{A.2.1.2}$$

It has the following properties:

$$\mathbb{E}(x) = \frac{a+b}{2} \qquad \text{and} \qquad \text{var}(x) = \frac{(b-a)^2}{12} \tag{A.2.1.3}$$

A.2.2 Multivariate normal distribution

A k-dimensional random vector x is said to follow a multivariate normal distribution with location μ and covariance Σ:

$$x \sim \mathcal{N}_k(\mu, \Sigma)$$

if its density is given by:

$$f(x\,|\mu, \Sigma) = (2\pi)^{-k/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right) \tag{A.2.2.1}$$

where μ is $k \times 1$ vector and Σ is $k \times k$ symmetric positive definite matrix. Its kernel is given by:

$$f(x\,|\mu, \Sigma) \propto \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right) \tag{A.2.2.2}$$

It has the following properties:

$$E(x) = \mu \text{ and } Var(x) = \Sigma \tag{A.2.2.3}$$

If $x \sim \mathcal{N}_{r_1+r_2}(\mu, \Sigma)$, and x, μ and Σ are partitioned in the following way:

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \begin{matrix} r_1 \\ r_2 \end{matrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \begin{matrix} r_1 \\ r_2 \end{matrix} \text{ and } \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \\ r_1 & r_2 \end{pmatrix} \begin{matrix} r_1 \\ r_2 \end{matrix}$$

Then:

$$x_1 \sim \mathcal{N}_{r_1}(\mu_1, \Sigma_{11}) \text{ and } x_2 \sim \mathcal{N}_{r_2}(\mu_2, \Sigma_{22}) \tag{A.2.2.4}$$

In other words, a subset of multivariate normal distribution is itself a multivariate normal distribution, with mean and covariance matrices defined as the corresponding subset of the original distribution.

The converse property holds in case of independence: let x_1 be a $k_1 \times 1$ random vector following a multivariate normal distribution with mean μ_1 and covariance Σ_1, and x_2 be a $k_2 \times 1$ random vector following a multivariate normal distribution with mean μ_2 and covariance Σ_2. If x_1 and x_2 are independent, then the $k \times 1$ random vector x follows a multivariate normal distribution with mean μ and covariance Σ with:

$$k = k_1 + k_2 \qquad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \tag{A.2.2.5}$$

A final important property is the affine or linear property of the multivariate normal distribution: let $x \sim \mathcal{N}_k(\mu, \Sigma)$ and let A and b respectively denote a $m \times k$ matrix and a $m \times 1$ vector of coefficients. Then the random vector $y = Ax + b$ also follows a multivariate normal distribution with mean $A\mu + b$ and covariance matrix $A\Sigma A'$. That is:

$$x \sim \mathcal{N}_k(\mu, \Sigma) \Rightarrow y = Ax + b \sim \mathcal{N}_m(A\mu + b, A\Sigma A') \tag{A.2.2.6}$$

Therefore, an affine combination of a multivariate normal random variable is itself multivariate normal, with mean and covariance matrices defined in accordance with the affine function.

This property can be used to generate easily multivariate random numbers with an arbitrary mean and covariance from a multivariate standard normal draw:

Algorithm a.2.2.1 (random number generator for the multivariate normal distribution): in order to generate a $k \times 1$ random vector y from a multivariate normal distribution with mean μ and covariance matrix Σ:

1. first draw a $k \times 1$ random vector x from a multivariate standard normal distribution, that is, draw x from $x \sim \mathcal{N}_k(0, I_k)$.

2. estimate A, where A is any matrix such that $AA' = \Sigma$. Typically, A will be chosen as the (lower triangular) Choleski factor of Σ, but other choices are possible. For instance, A could be chosen as the square root matrix of Σ).

3. eventually compute $y = \mu + Ax$. Then from the affine property $y = \mu + Ax$ is a random draw from $\mathcal{N}_k(\mu, \Sigma)$.

A.2.3 Matrix Normal Distribution

A $k \times n$ random matrix X is said to follow a matrix normal distribution with location M, and scale matrices Σ and Φ:

$$X \sim \ M\mathcal{N}_{k,n}(M, \Sigma, \Phi)$$

if its density is given by:

$$f(X\,|M, \Phi, \Sigma) = (2\pi)^{-nk/2}|\Sigma|^{-k/2}|\Phi|^{-n/2} \exp\left(-\frac{1}{2} tr\left[\Sigma^{-1}(X - M)'\Phi^{-1}(X - M)\right]\right) \tag{A.2.3.1}$$

where M is $k \times n$ matrix, and Σ is $n \times n$ symmetric positive definite matrix, and Φ is $k \times k$ symmetric positive definite matrix.

Its kernel is given by:

$$f(X\,|M, \Phi, \Sigma) \propto \exp\left(-\frac{1}{2} tr\left[\Sigma^{-1}(X - M)'\Phi^{-1}(X - M)\right]\right) \tag{A.2.3.2}$$

It has the following properties:

$$E(X) = M \qquad Var(X) = \begin{cases} \Phi \text{ (among rows)} \\ \Sigma \text{ (among columns)} \end{cases} \tag{A.2.3.3}$$

Equivalence of the matrix normal distribution with the multivariate normal distribution:

$$X \sim \ MN_{k,n}(M, \Phi, \Sigma) \text{ if and only if } vec(X) \sim \ \mathcal{N}_{kn}(vec(M), \Sigma \otimes \Phi). \tag{A.2.3.4}$$

Proof: the proof consists in showing that the density of the two distributions are equivalent, and it follows directly from A.1.4 and A.1.10.

To illustrate this relation, it is now shown how the data density can be conveniently rewritten in terms of a matrix normal distribution, rather than a multivariate normal. Start from the data density 3.3.1:

$$f(y\,|\beta, \bar{\Sigma}) = (2\pi)^{-nT/2}\left|\bar{\Sigma}\right|^{-1/2} \exp\left[-\frac{1}{2}\left(y - \bar{X}\beta\right)'\bar{\Sigma}^{-1}\left(y - \bar{X}\beta\right)\right] \tag{A.2.3.5}$$

From 3.1.13, $y = vec(Y)$, $\quad \bar{X} = I_n \otimes X$, and $\beta = vec(B)$, so that it is possible to rewrite:

$$y - \bar{X}\beta = vec(Y - XB) \tag{A.2.3.6}$$

Also, from 3.1.14, $\bar{\Sigma} = \Sigma \otimes I_T$. With these elements, it is possible to reformulate the data multivariate normal density as:

$$f(y \,|\beta, \bar{\Sigma}) \sim \; N_{Tn}\left(vec(Y - XB), \Sigma \otimes I_T\right) \tag{A.2.3.7}$$

Now, apply A.2.3.4, and conclude that equivalently, the data density may be expressed a matrix normal distribution:

$$f(y \,|B, \Sigma) \sim \; MN_{T,n}\left(Y - XB, \Sigma, I_T\right) \tag{A.2.3.8}$$

From A.2.3.1, this implies the following density function:

$$f(y \,|B, \Sigma) = (2\pi)^{-nT/2}|\Sigma|^{-T/2}|I_T|^{-n/2} \exp\left(-\frac{1}{2} tr\left[\Sigma^{-1}(Y - XB)' {I_T}^{-1}(Y - XB)\right]\right)$$

And this simplifies to:

$$f(y \,|B, \Sigma) = (2\pi)^{-nT/2}|\Sigma|^{-T/2} \exp\left(-\frac{1}{2} tr\left[\Sigma^{-1}(Y - XB)'(Y - XB)\right]\right) \tag{A.2.3.9}$$

The advantage of A.2.3.9 over A.2.3.5 is that it is computationally more efficient. In practical applications, such as in the calculation of the data density for the marginal likelihood 3.9.36, it is thus this formulation which will be retained.

Eventually, a feature of central interest is to be able to obtain random draws from a matrix normal distribution. An algorithm to draw from a matrix normal distribution is provided by Karlsson (2012) (see algorithm 19).

algorithm a.2.3.1 (random number generator for the matrix normal distribution): In order to obtain a $k \times n$ random draw X from a matrix normal distribution with location matrix M, and scale matrices Σ and Φ:

1. first compute the Choleski factors C and P of Σ and Φ, so that $CC' = \Sigma$ and $PP' = \Sigma$.

2. draw a $kn \times 1$ random vector w from a multivariate standard normal distribution, and redimension it to transform it into a $k \times n$ random draw W from a standard matrix normal distribution, using A.2.3.4.

3. finally, obtain $X = M + PWC'$. Then X is a random draw from $MN_{k,n}(M, \Sigma, \Phi)$.

A.2.4 Inverse Wishart distribution

The $n \times n$ symmetric positive definite matrix Σ follows an inverse Wishart distribution with scale matrix S and degrees of freedom α:

$$\Sigma \sim \mathcal{IW}(S, \alpha)$$

If its density is given by:

$$f(\Sigma \,|S, \alpha) = \frac{1}{2^{\alpha n/2}\Gamma_n\left(\frac{\alpha}{2}\right)}|S|^{\alpha/2}|\Sigma|^{-(\alpha+n+1)/2}\exp\left(-\frac{1}{2}tr\left\{\Sigma^{-1}S\right\}\right) \tag{A.2.4.1}$$

where S is $n \times n$ symmetric positive definite matrix, and α is an integer value. Γ_n is the multivariate Gamma function, defined as:

$$\Gamma_n(x) = \pi^{p(p-1)/4}\prod_{i=1}^{n}\Gamma\left(x + (1-i)/2\right) \tag{A.2.4.2}$$

with $\Gamma(x)$ the (univariate) Gamma function defined as:

$$\Gamma\left(x\right) = \int_0^\infty t^{x-1}e^{-t}dt$$

The kernel of the inverse Wishart distribution is given by:

$$f(\Sigma \,|S, \alpha) \propto |\Sigma|^{-(\alpha+n+1)/2}\exp\left(-\frac{1}{2}tr\left\{\Sigma^{-1}S\right\}\right) \tag{A.2.4.3}$$

It has the following properties:

$$E(\Sigma) = \frac{S}{\alpha - n - 1} \text{ for } \alpha > n + 1 \tag{A.2.4.4}$$

For $\sigma_{i,j} \in \Sigma$:

$$Var(\sigma_{i,j}) = \frac{(\alpha - n + 1)s_{i,j}^2 + (\alpha - n - 1)s_{i,i}s_{j,j}}{(\alpha - n)(\alpha - n - 1)(\alpha - n - 3)} \tag{A.2.4.5}$$

where $s_{i,j}$ is the entry of row i and column j of the matrix S.

Similarly to the matrix normal distribution, an important feature in practical applications is how to obtain a random draw from the inverse Wishart distribution. A simple algorithm is proposed by Karlsson (2012) (see algorithms 20 and 21).

Algorithm a.2.4.1 (random number generator for the inverse Wishart distribution): In order to obtain a $n \times n$ random draw Σ from an inverse Wishart distribution with scale matrix S and degrees of freedom α:

1. compute the lower triangular Choleski factor C of the scale matrix S.
2. draw α random vectors $z_1, z_2, \ldots, z_\alpha$ from a multivariate standard normal distribution: $Z \sim \mathcal{N}_n(0, I_n)$.
3. arrange those vectors into a $\alpha \times n$ matrix Z, where row i of Z is the transpose of z_i.
4. eventually estimate $X = C(Z'Z)^{-1}C'$. Then X is a random draw from $\mathcal{IW}_n(S, \alpha)$.

A.2.5 Matrix variate student distribution

This distribution may have several different definitions. The definition retained here is that of Gupta and Nagar (1999). A $k \times n$ random matrix X is said to follow a matrix normal distribution with location M, scale matrices Σ and Φ, and degrees of freedom α :

$$X \sim \ Mt_{k,n}(M, \Sigma, \Phi, \alpha) \tag{A.2.5.1}$$

if its density is given by:

$$f(X\,|M, \Sigma, \Phi, \alpha) = \frac{\Gamma_k\left([\alpha + k + n - 1]/2\right)}{\pi^{kn/2}\Gamma_k\left([\alpha + k - 1]/2\right)}|\Sigma|^{-k/2}|\Phi|^{-n/2}\left|I_n + \Sigma^{-1}(X - M)'\Phi^{-1}(X - M)\right|^{-(\alpha+n+k-1)/2} \tag{A.2.5.2}$$

where M is $k \times n$ matrix, and Σ is $n \times n$ symmetric positive definite matrix, Φ is $k \times k$ symmetric positive definite matrix, and α is an integer value. Γ_k is the multivariate Gamma function.

Its kernel is given by:

$$f(X\,|M, \Sigma, \Phi, \alpha) \propto \left|I_n + \Sigma^{-1}(X - M)'\Phi^{-1}(X - M)\right|^{-(\alpha+n+k-1)/2} \tag{A.2.5.3}$$

It has the following properties:

$$E(X) = M \text{ and } Var(vec(X)) = \frac{1}{\alpha - 2}\left(\Sigma \otimes \Phi\right) \text{ for } \alpha > 2 \tag{A.2.5.4}$$

Another important property of the matrix-variate student is the following theorem (theorem 4.3.9 in Gupta and Nagar (1999)):

Let $X \sim \ Mt_{k,n}(M, \Sigma, \Phi, \alpha)$, and partition X, M, Σ and Φ as:

$$X = \begin{pmatrix} X_{r1} \\ X_{r2} \end{pmatrix} \begin{matrix} r_1 \\ r_2 \end{matrix} = \begin{matrix} (X_{c1} & X_{c2}) \\ c_1 & c_2 \end{matrix}$$

$$M = \begin{pmatrix} M_{r1} \\ M_{r2} \end{pmatrix} \begin{matrix} r_1 \\ r_2 \end{matrix} = \begin{matrix} (M_{c1} & M_{c2}) \\ c_1 & c_2 \end{matrix}$$

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{matrix} c_1 \\ c_2 \end{matrix}$$
$$\begin{matrix} c_1 & c_2 \end{matrix}$$

and

$$\Phi = \begin{pmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{pmatrix} \begin{matrix} r_1 \\ r_2 \end{matrix}$$
$$\begin{matrix} r_1 & r_2 \end{matrix}$$

then:

$$X_{r1} \sim \ Mt_{r_1,n}(M_{r1}, \Sigma, \Phi_{11}, \alpha)$$

and

$$X_{c1} \sim \ Mt_{k,c_1}(M_{c1}, \Sigma_{11}, \Phi, \alpha)$$

In other words, if one partitions a matrix-student distribution with α degrees of freedom, then the sub-partition remains a matrix-student distribution with α degrees of freedom, with mean and covariance defined in accordance with the partitions. Going on with partitioning up to individual entries of X, this result implies that:

$$X_{ij} \sim \ t(M_{ij}, \Phi_{ii} \times \Sigma_{jj}, \alpha) \tag{A.2.5.5}$$

That is, each individual element X_{ij} of X follows a univariate student distribution with mean M_{ij}, scale parameter $\Phi_{ii} \times \Sigma_{jj}$ and degrees of freedom α. This provides the mean and variance X_{ij}, but no direct way to compute confidence intervals. Because the definition of the matrix student distribution provided by Gupta and Nagar (1999) differs from the classical student distribution with respect to the scale parameter, some translation is required to compute correct results. Hence, to obtain confidence intervals, consider a classical univariate location-scale student distribution with location parameter μ and scale parameter σ, such that:

$$X = \mu + \sigma T$$

where T follows a classical univariate student distribution. Then this distribution has the following

properties:

$$E(X) = \mu \text{ and } Var(X) = \sigma^2 \frac{\alpha}{\alpha - 2}$$

with α the degrees of freedom of the distribution. This implies that the univariate scale parameter σ can be expressed as a function of the variance and degrees of freedom as:

$$\sigma = \sqrt{\frac{\alpha - 2}{\alpha} Var(X)} \tag{A.2.5.6}$$

Because $Var(X)$ can be directly obtained for the matrix student distribution from A.2.5.4, a γ confidence interval for the distribution can then be obtained in a classical way from:

$$M_{ij} \pm t_{(\gamma/2,\alpha)}\sigma \tag{A.2.5.7}$$

where $t_{(\gamma,\alpha)}$ is the is the γ^{th} quantile of the standard student distribution with α degrees of freedom.

Eventually, in a way similar to the matrix normal and inverse Wishart distributions, an algorithm to draw from the matrix student distribution is now introduced (see Karlsson (2012), algorithm 22).

algorithm a.2.5.1 (random number generator for the matrix variate student distribution):
In order to obtain a $k \times n$ random draw X from a matrix variate student distribution with location M, scale matrices Σ and Φ, and degrees of freedom α:

1. first draw a $n \times n$ random matrix V from an inverse Wishart distribution with scale matrix Σ and degrees of freedom α : $V \sim IW_n(\Sigma, \alpha)$, using algorithm a.2.2.

2. draw a $k \times n$ random matrix X from a matrix normal distribution: $Z \sim MN_{k,n}(M, V, \Phi)$, using algorithm a.2.1. Then X is a random draw from $Mt_{k,n}(M, \Sigma, \Phi, \alpha)$.

A.2.6 Gamma distribution

The scalar random variable x follows a Gamma distribution with shape parameter a and scale parameter b:

$$x \sim G(a, b) \tag{A.2.6.1}$$

If its density is given by:

$$f(x \,|a, b) = \frac{b^{-a}}{\Gamma(a)} x^{a-1} \exp\left(-\frac{x}{b}\right) \tag{A.2.6.2}$$

$\Gamma(a)$ denotes as usual the (univariate) Gamma function. The kernel of the distribution is given by:

$$f(x \,|a, b) \propto x^{a-1} \exp\left(-\frac{x}{b}\right) \tag{A.2.6.3}$$

It has the following properties:

$$\mathbb{E}(x) = ab \qquad \text{and} \qquad \text{var}(x) = ab^2 \tag{A.2.6.4}$$

To sample from $G(a, b)$, it is possible to use the following algorithm proposed by Marsaglia and Tsang (2000):

algorithm a.2.6.1 (random number generator for the Gamma distribution):
step 1. generate a random number from $G(a, 1)$:
If $a \geq 1$:

1. set $d = a - 1/3$ and $c = 1/\sqrt{9d}$.
2. generate $z \sim \mathcal{N}(0, 1)$ and $u \sim U(0, 1)$ independently.
3. generate $v = (1 + cz)^3$.
4. if $v > 0$ and $\log(u) < 0.5z^2 + d - dv + d \times \log(v)$, then set $x = dv$.
5. otherwise, go back to 2.

If $0 < a < 1$:

1. generate a random number $\tilde{x}$ from $G(a + 1, 1)$, using the above algorithm.
2. generate $u \sim U(0, 1)$
3. define $x = \tilde{x}u^{1/a}$; then $x \sim G(a, 1)$

step 2. transform into a random number from $G(a, b)$:

1. generate a random number $\bar{x}$ from $G(a, 1)$, using step 1.
2. define $x = \bar{x}b$; then $x \sim G(a, b)$

A.2.7 Inverse Gamma distribution

The scalar random variable x follows an inverse Gamma distribution with shape parameter a and scale parameter b:

$$x \sim IG(a, b) \tag{A.2.7.1}$$

if its density is given by:

$$f(x|a, b) = \frac{b^a}{\Gamma(a)} x^{-a-1} \exp\left(\frac{-b}{x}\right) \tag{A.2.7.2}$$

$\Gamma(a)$ denotes as usual the (univariate) Gamma function. The kernel of the distribution is given by:

$$f\left(x\,|a,b\right) \propto x^{-a-1}\exp\left(\frac{-b}{x}\right) \tag{A.2.7.3}$$

It has the following properties:

$$\mathbb{E}(x) = \frac{b}{a-1} \text{ for a>1} \qquad \text{and} \qquad \text{var}(x) = \frac{b^2}{(a-1)^2(a-2)} \text{ for a>2} \tag{A.2.7.4}$$

Another important property of the inverse Gamma is the following: if x follows a Gamma distribution with shape a and scale b, then $1/x$ follows an inverse Gamma distribution with shape a and scale $1/b$. That is:

$$x \sim G(a,b) \qquad \Rightarrow \qquad 1/x \sim IG(a,1/b) \tag{A.2.7.5}$$

It is then simple to propose the following algorithm to draw from $IG(a,b)$:

algorithm a.2.7.1 (random number generator for the Inverse Gamma distribution):

1. draw a random number $\tilde{x}$ from $G(a,1/b)$.
2. set $x = 1/\tilde{x}$; then x is a random draw from $IG(a,b)$.

Also, it turns out that the inverse Gamma distribution is a special univariate case of the inverse Wishart distribution. To see this, consider the inverse Wishart density A.2.4.1 when $n = 1$ (univariate case) so that $\Sigma = x$, $S = 2b$, $\alpha = 2a$ and compare with (A.2.7.2).

A.3 Derivations of the posterior distribution with a Minnesota prior

The derivation of the posterior distribution with a Minnesota prior remains relatively simple. It starts with equation 3.3.15:

$$\pi(\beta\,|y) \propto \exp\left[-\frac{1}{2}\left\{\left(y-\bar{X}\beta\right)'\bar{\Sigma}^{-1}\left(y-\bar{X}\beta\right)+(\beta-\beta_0)'\Omega_0^{-1}\left(\beta-\beta_0\right)\right\}\right] \tag{A.3.1}$$

To transform this expression, consider only the exponential part in the curly brackets and develop it:

$$\begin{aligned}
&\left(y-\bar{X}\beta\right)'\bar{\Sigma}^{-1}\left(y-\bar{X}\beta\right)+(\beta-\beta_0)'\Omega_0^{-1}\left(\beta-\beta_0\right)\\
&= y'\bar{\Sigma}^{-1}y+\beta'\bar{X}'\bar{\Sigma}^{-1}\bar{X}\beta-2\beta'\bar{X}'\bar{\Sigma}^{-1}y+\beta'\Omega_0^{-1}\beta+\beta_0'\Omega_0^{-1}\beta_0-2\beta'\left(\Omega_0^{-1}\right)'\beta_0\\
&= y'\bar{\Sigma}^{-1}y+\beta'\left(\Omega_0^{-1}+\bar{X}'\bar{\Sigma}^{-1}\bar{X}\right)\beta-2\beta'\left(\Omega_0^{-1}\beta_0+\bar{X}'\bar{\Sigma}^{-1}y\right)+\beta_0'\Omega_0^{-1}\beta_0
\end{aligned} \tag{A.3.2}$$

Notice that A.3.1 resembles the kernel of a normal distribution, but with a sum of squares rather than a single square term within the exponential part. It would therefore be nice to replace this sum by a single square, to obtain the kernel of a normal distribution. This can be done by applying the manipulations known as "completing the square". This most of the time amounts to adding and subtracting an additional matrix term, and inserting the product of a matrix with its inverse. After factoring, this will eventually lead to a single squared form, while the additional terms created will be independent of β and will hence be relegated to the proportionality constant. Hence, complete the squares in A.3.2):

$$= y'\bar{\Sigma}^{-1}y + \beta'\left(\Omega_0^{-1} + \bar{X}'\bar{\Sigma}^{-1}\bar{X}\right)\beta - 2\beta'\bar{\Omega}^{-1}\bar{\Omega}\left(\Omega_0^{-1}\beta_0 + \bar{X}'\bar{\Sigma}^{-1}y\right) + \beta_0'\Omega_0^{-1}\beta_0 + \bar{\beta}'\bar{\Omega}^{-1}\bar{\beta} - \bar{\beta}'\bar{\Omega}^{-1}\bar{\beta} \tag{A.3.3}$$

Note that A.3.3 holds whatever the definition of $\bar{\Omega}$ and $\bar{\beta}$ (as long as dimensions agree). Nevertheless, one may obtain the desired squared form from A.3.3 by defining:

$$\bar{\Omega} = \left(\Omega_0^{-1} + \bar{X}'\bar{\Sigma}^{-1}\bar{X}\right)^{-1}$$

$$\textit{and} \tag{A.3.4}$$

$$\bar{\beta} = \bar{\Omega}\left(\Omega_0^{-1}\beta_0 + \bar{X}'\bar{\Sigma}^{-1}y\right) \tag{A.3.5}$$

For then, A.3.3 rewrites:

$$\begin{aligned} &= y'\bar{\Sigma}^{-1}y + \beta'\bar{\Omega}^{-1}\beta - 2\beta'\bar{\Omega}^{-1}\bar{\beta} + \beta_0'\Omega_0^{-1}\beta_0 + \bar{\beta}'\bar{\Omega}^{-1}\bar{\beta} - \bar{\beta}'\bar{\Omega}^{-1}\bar{\beta} \\ &= \left(\beta'\bar{\Omega}^{-1}\beta - 2\beta'\bar{\Omega}^{-1}\bar{\beta} + \bar{\beta}'\bar{\Omega}^{-1}\bar{\beta}\right) + \left(\beta_0'\Omega_0^{-1}\beta_0 - \bar{\beta}'\bar{\Omega}^{-1}\bar{\beta} + y'\bar{\Sigma}^{-1}y\right) \\ &= \left(\beta - \bar{\beta}\right)'\bar{\Omega}^{-1}\left(\beta - \bar{\beta}\right) + \left(\beta_0'\Omega_0^{-1}\beta_0 - \bar{\beta}'\bar{\Omega}^{-1}\bar{\beta} + y'\bar{\Sigma}^{-1}y\right) \end{aligned} \tag{A.3.6}$$

Substituting back A.3.6 in A.3.1), one obtains:

$$\begin{aligned} \pi(\beta\,|y) &\propto \exp\left[-\frac{1}{2}\left\{\left(\beta - \bar{\beta}\right)'\bar{\Omega}^{-1}\left(\beta - \bar{\beta}\right) + \left(\beta_0'\Omega_0^{-1}\beta_0 - \bar{\beta}'\bar{\Omega}^{-1}\bar{\beta} + y'\bar{\Sigma}^{-1}y\right)\right\}\right] \\ &\propto \exp\left[-\frac{1}{2}\left(\beta - \bar{\beta}\right)'\bar{\Omega}^{-1}\left(\beta - \bar{\beta}\right)\right] \times \exp\left[-\frac{1}{2}\left(\beta_0'\Omega_0^{-1}\beta_0 - \bar{\beta}'\bar{\Omega}^{-1}\bar{\beta} + y'\bar{\Sigma}^{-1}y\right)\right] \\ &\propto \exp\left[-\frac{1}{2}\left(\beta - \bar{\beta}\right)'\bar{\Omega}^{-1}\left(\beta - \bar{\beta}\right)\right] \end{aligned} \tag{A.3.7}$$

where the last line obtains by noting that the second term in row 2 does not involve β and can

hence be relegated to the proportionality constant. Therefore, one finally obtains:

$$\pi(\beta\,|y\,) \propto \exp\left[-\frac{1}{2}\big(\beta-\bar{\beta}\big)'\bar{\Omega}^{-1}\left(\beta-\bar{\beta}\right)\right] \tag{A.3.8}$$

Which is 3.3.16 in the text. Finally, one may obtain a rewriting of A.3.5 and A.3.6.

Consider $\bar{X}'\bar{\Sigma}^{-1}\bar{X}$ in A.3.5:

$$\begin{aligned}\bar{X}'\bar{\Sigma}^{-1}\bar{X} =&(I_n\otimes X)'(\Sigma\otimes I_T)^{-1}(I_n\otimes X)\\ =&(I_n\otimes X')(\Sigma^{-1}\otimes I_T)(I_n\otimes X)A.1.1A.1.2\\ =&(\Sigma^{-1}\otimes X')(I_n\otimes X)A.1.3\\ =&\Sigma^{-1}\otimes X'XA.1.3\end{aligned} \tag{A.3.9}$$

Hence, A.3.5 rewrites:

$$\bar{\Omega} = \left[\Omega_0^{-1}+\Sigma^{-1}\otimes X'X\right]^{-1} \tag{A.3.10}$$

Similarly, consider the part $\bar{X}'\bar{\Sigma}^{-1}y$ in A.3.6:

$$\begin{aligned}\bar{X}'\bar{\Sigma}^{-1}y =&(I_n\otimes X)'(\Sigma\otimes I_T)^{-1}y\\ &(I_n\otimes X')(\Sigma^{-1}\otimes I_T)y \qquad A.1.1, A.1.2\\ =&(\Sigma^{-1}\otimes X')y \qquad A.1.3\end{aligned} \tag{A.3.11}$$

Therefore, A.1.5 rewrites:

$$\bar{\beta} = \bar{\Omega}\left[\Omega_0^{-1}\beta_0+(\Sigma^{-1}\otimes X')y\right] \tag{A.3.12}$$

The advantage of A.3.9 and A.3.10 over A.3.5 and A.3.6 is that they are numerically much faster to compute. One can then eventually recognise A.3.9 and A.3.10 as 3.3.17 and 3.3.18 in the text.

A.4 Derivations of the posterior distribution with a normal-Wishart prior

Computing the posterior distributions for β and Σ from 3.2.5 requires the identification of a likelihood function for the data, and of prior distributions for β and Σ. Start with the likelihood. One may first want to show how to rewrite the likelihood 3.6.2 in the form of 3.6.3. Start from 3.6.2:

$$f(y\,|\beta,\Sigma)\propto\left|\bar{\Sigma}\right|^{-1/2}\exp\left[-\frac{1}{2}\left(y-\bar{X}\beta\right)'\bar{\Sigma}^{-1}\left(y-\bar{X}\beta\right)\right]$$

For our purpose, it will be easier to use 3.1.13 and 3.1.14 to reformulate the likelihood as:

$$f(y\,|\beta,\Sigma)\propto|\Sigma\otimes I_T|^{-1/2}\exp\left[-\frac{1}{2}\left\{(y-(I_n\otimes X)\beta)'(\Sigma\otimes I_T)^{-1}(y-(I_n\otimes X)\beta)\right\}\right] \tag{A.4.1}$$

Consider only the part within the curly brackets and develop it:

$$\begin{aligned}
&(y-(I_n\otimes X)\beta)'(\Sigma\otimes I_T)^{-1}(y-(I_n\otimes X)\beta)\\
&=(y'-\beta'(I_n\otimes X)')\left(\Sigma^{-1}\otimes I_T\right)(y-(I_n\otimes X)\beta)\ A.1.2\\
&=y'\left(\Sigma^{-1}\otimes I_T\right)y-2\beta'(I_n\otimes X)'\left(\Sigma^{-1}\otimes I_T\right)y+\beta'(I_n\otimes X)'\left(\Sigma^{-1}\otimes I_T\right)(I_n\otimes X)\beta\\
&=y'\left(\Sigma^{-1}\otimes I_T\right)y-2\beta'(\Sigma^{-1}\otimes X')y+\beta'(\Sigma^{-1}\otimes X'X)\beta\ A.1.1, A.1.3
\end{aligned}$$

Now complete the squares:

$$\begin{aligned}
&=y'\left(\Sigma^{-1}\otimes I_T\right)y-2\hat{\beta}'\left(\Sigma^{-1}\otimes X'\right)y+2\hat{\beta}'\left(\Sigma^{-1}\otimes X'\right)y-2\beta'\left(\Sigma^{-1}\otimes X'\right)y+\beta'\left(\Sigma^{-1}\otimes X'X\right)\beta\\
&=y'\left(\Sigma^{-1}\otimes I_T\right)y-2\hat{\beta}'\left(\Sigma^{-1}\otimes X'\right)y+2\hat{\beta}'\left(\Sigma^{-1}\otimes X'X\right)\left(\Sigma^{-1}\otimes X'X\right)^{-1}\left(\Sigma^{-1}\otimes X'\right)y\\
&-2\beta'\left(\Sigma^{-1}\otimes X'X\right)\left(\Sigma^{-1}\otimes X'X\right)^{-1}\left(\Sigma^{-1}\otimes X'\right)y+\beta'\left(\Sigma^{-1}\otimes X'X\right)\beta
\end{aligned} \tag{A.4.2}$$

Define $\hat{\beta}$, the OLS estimate of β, as:

$$\hat{\beta}=\left(\Sigma^{-1}\otimes X'X\right)^{-1}\left(\Sigma^{-1}\otimes X'\right)y. \tag{A.4.3}$$

Then, A.4.2 rewrites:

$$
\begin{aligned}
&= y^{'}\left(\Sigma^{-1}\otimes I_T\right)y - 2\hat{\beta}^{'}\left(\Sigma^{-1}\otimes X^{'}\right)y + 2\hat{\beta}^{'}\left(\Sigma^{-1}\otimes X^{'}X\right)\hat{\beta} \\
&- 2\beta^{'}\left(\Sigma^{-1}\otimes X^{'}X\right)\hat{\beta} + \beta^{'}\left(\Sigma^{-1}\otimes X^{'}X\right)\beta \\
&= y^{'}\left(\Sigma^{-1}\otimes I_T\right)y - 2\hat{\beta}^{'}\left(\Sigma^{-1}\otimes X^{'}\right)y + \hat{\beta}^{'}\left(\Sigma^{-1}\otimes X^{'}X\right)\hat{\beta} \\
&+ \hat{\beta}^{'}\left(\Sigma^{-1}\otimes X^{'}X\right)\hat{\beta} - 2\beta^{'}\left(\Sigma^{-1}\otimes X^{'}X\right)\hat{\beta} + \beta^{'}\left(\Sigma^{-1}\otimes X^{'}X\right)\beta \\
&= y^{'}\left(\Sigma^{-1}\otimes I_T\right)y - 2\hat{\beta}^{'}\left(\Sigma^{-1}\otimes X^{'}\right)y + \hat{\beta}^{'}\left(\Sigma^{-1}\otimes X^{'}X\right)\hat{\beta} \\
&+ (\beta-\hat{\beta})^{'}\left(\Sigma^{-1}\otimes X^{'}X\right)(\beta-\hat{\beta}) \\
&= y^{'}\left(\Sigma^{-1}\otimes I_T\right)y - 2\hat{\beta}^{'}\left(\Sigma^{-1}\otimes X^{'}\right)y + \hat{\beta}^{'}\left(\Sigma^{-1}\otimes X^{'}X\right)\hat{\beta} \\
&+ (\beta-\hat{\beta})^{'}\left(\Sigma\otimes(X^{'}X)^{-1}\right)^{-1}(\beta-\hat{\beta}) \quad A.1.2
\end{aligned}
\tag{A.4.4}
$$

The second row of A.4.4 has a nice squared form, while the first row requires some additional work. Hence, focus on the first row only:

$$
\begin{aligned}
&y^{'}\left(\Sigma^{-1}\otimes I_T\right)y - 2\hat{\beta}^{'}\left(\Sigma^{-1}\otimes X^{'}\right)y + \hat{\beta}^{'}\left(\Sigma^{-1}\otimes X^{'}X\right)\hat{\beta} \\
&= y^{'}\left(\Sigma^{-1}\otimes I_T\right)y - 2\hat{\beta}^{'}(I_n\otimes X)^{'}\left(\Sigma^{-1}\otimes I_T\right)y + \hat{\beta}^{'}\left(\Sigma^{-1}\otimes X^{'}\right)(I_n\otimes X)\hat{\beta} \quad A.1.3 \\
&= y^{'}\left(\Sigma^{-1}\otimes I_T\right)y - 2\Big((I_n\otimes X)\hat{\beta}\Big)^{'}\left(\Sigma^{-1}\otimes I_T\right)y + \hat{\beta}^{'}(I_n\otimes X)^{'}\left(\Sigma^{-1}\otimes I_T\right)(I_n\otimes X)\hat{\beta} \quad A.1.3 \\
&= y^{'}(\Sigma\otimes I_T)^{-1}y - 2\Big((I_n\otimes X)\hat{\beta}\Big)^{'}(\Sigma\otimes I_T)^{-1}y + \hat{\beta}^{'}(I_n\otimes X)^{'}(\Sigma\otimes I_T)^{-1}(I_n\otimes X)\hat{\beta} \quad A.1.2 \\
&= tr\left\{\Sigma^{-1}Y^{'}I_TY\right\} - 2tr\left\{\Sigma^{-1}(X\hat{B})^{'}I_TY\right\} + tr\left\{\Sigma^{-1}\hat{B}^{'}X^{'}I_TX\hat{B}\right\} \quad A.1.10 \\
&= tr\left\{Y\Sigma^{-1}Y^{'}\right\} - 2tr\left\{Y\Sigma^{-1}\hat{B}^{'}X^{'}\right\} + tr\left\{X\hat{B}\Sigma^{-1}\hat{B}^{'}X^{'}\right\} \quad A.17.15 \\
&= tr\left\{Y\Sigma^{-1}Y^{'} - 2Y\Sigma^{-1}\hat{B}^{'}X^{'} + X\hat{B}\Sigma^{-1}\hat{B}^{'}X^{'}\right\} \quad A.17.15 \\
&= tr\left\{(Y-X\hat{B})\Sigma^{-1}(Y-X\hat{B})^{'}\right\} \\
&= tr\left\{\Sigma^{-1}(Y-X\hat{B})^{'}(Y-X\hat{B})\right\} \quad A.1.1
\end{aligned}
$$

Replace the obtained expression in A.4.4 to obtain finally:

$$
\begin{aligned}
&(y-(I_n\otimes X)\beta)^{'}(\Sigma\otimes I_T)^{-1}(y-(I_n\otimes X)\beta) = \\
&= tr\left\{\Sigma^{-1}(Y-X\hat{B})^{'}(Y-X\hat{B})\right\} + (\beta-\hat{\beta})^{'}\left(\Sigma\otimes(X^{'}X)^{-1}\right)^{-1}(\beta-\hat{\beta})
\end{aligned}
\tag{A.4.5}
$$

Before turning back to A.4.1, note also that Kronecker property A.1.4 implies that the determinant part of A.4.1 can rewrite as:

$$|\Sigma \otimes I_T|^{-1/2} = \left(|\Sigma|^T |I_T|^n\right)^{-1/2} = |\Sigma|^{-T/2} = |\Sigma|^{-k/2}|\Sigma|^{-(T-k)/2} = |\Sigma|^{-k/2}|\Sigma|^{-[(T-k-n-1)+n+1]/2} \tag{A.4.6}$$

Substituting A.4.5 and A.4.6 in A.4.1, one eventually obtains:

$$\begin{aligned} f(y\,|\beta,\Sigma) &\propto |\Sigma|^{-k/2}|\Sigma|^{-[(T-k-n-1)+n+1]/2} \\ &\times \exp\left[-\frac{1}{2}\left\{tr\left\{\Sigma^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right\} + (\beta-\hat{\beta})'\left(\Sigma\otimes(X'X)^{-1}\right)^{-1}(\beta-\hat{\beta})\right\}\right] \\ f(y\,|\beta,\Sigma) &\propto |\Sigma|^{-k/2}\exp\left[-\frac{1}{2}(\beta-\hat{\beta})'\left(\Sigma\otimes(X'X)^{-1}\right)^{-1}(\beta-\hat{\beta})\right]\times \\ &\times |\Sigma|^{-[(T-k-n-1)+n+1]/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right\}\right] \end{aligned} \tag{A.4.7}$$

A.4.7 is just 3.6.3 in the text. It can be recognised as the product of two kernels: the kernel of a multivariate normal distribution for β (given Σ), with mean $\hat{\beta}$ and covariance $\left(\Sigma\otimes(X'X)^{-1}\right)^{-1}$ and the kernel of an inverse Wishart distribution for Σ, with scale matrix $(Y-X\hat{B})'(Y-X\hat{B})$ and degrees of freedom $T-k-n-1$.

With the likelihood determined it is now possible to estimate the posterior distribution for β and Σ. As it will be more convenient to work with all equations expressed in terms of trace operators, start by reshaping the first row of A.4.7. This is easily done by noting that:

$$(\beta-\hat{\beta})'\left(\Sigma\otimes(X'X)^{-1}\right)^{-1}(\beta-\hat{\beta}) = tr\left\{\Sigma^{-1}\left(B-\hat{B}\right)'(X'X)\left(B-\hat{B}\right)\right\}\ A.1.10$$

Hence, using the previous expression, and collecting powers on $|\Sigma|$, A.4.7 rewrites as:

$$\begin{aligned} f(y\,|\beta,\Sigma) &\propto |\Sigma|^{-T/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left(B-\hat{B}\right)'(X'X)\left(B-\hat{B}\right)\right\}\right] \\ &\propto \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right\}\right] \end{aligned} \tag{A.4.8}$$

Also, using once again A.1.10, the prior density for β A.1.9 rewrites as:

$$\begin{aligned} \pi(\beta) &\propto |\Sigma|^{-k/2}\exp\left[-\frac{1}{2}(\beta-\beta_0)'\left(\Sigma\otimes\Phi_0\right)^{-1}(\beta-\beta_0)\right] \\ &\propto |\Sigma|^{-k/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(B-B_0)'\Phi_0^{-1}(B-B_0)\right\}\right] \end{aligned} \tag{A.4.9}$$

Applying Bayes rule 3.2.5 on the likelihood A.4.8 and the priors A.4.9 and 3.4.14, one obtains:

$$\begin{aligned}\pi(\beta,\Sigma\,|y) &\propto f(y\,|\beta,\Sigma)\pi(\beta)\pi(\Sigma)\\ &\propto |\Sigma|^{-T/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left(B-\hat{B}\right)'(X'X)\left(B-\hat{B}\right)\right\}\right]\\ &\times\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right\}\right]\\ &\times|\Sigma|^{-k/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(B-B_0)'\Phi_0^{-1}(B-B_0)\right\}\right]\\ &\times|\Sigma|^{-(\alpha_0+n+1)/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}S_0\right\}\right] \qquad \text{(A.4.10)}\end{aligned}$$

Rearrange:

$$\begin{aligned}\pi(\beta,\Sigma\,|y) \propto |\Sigma|^{-(T+k+\alpha_0+n+1)/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left[\left(B-\hat{B}\right)'(X'X)\left(B-\hat{B}\right)+(B-B_0)'\Phi_0^{-1}(B-B_0)\right]\right\}\right]\\ \times\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left[S_0+(Y-X\hat{B})'(Y-X\hat{B})\right]\right\}\right] \qquad \text{(A.4.11)}\end{aligned}$$

Focus first on the term in the curly brace in the first row of A.4.11 :

$$\begin{aligned}&=\Sigma^{-1}\left[\left(B-\hat{B}\right)'(X'X)\left(B-\hat{B}\right)+(B-B_0)'\Phi_0^{-1}(B-B_0)\right]\\ &=\Sigma^{-1}\left[B'X'XB+\hat{B}'X'X\hat{B}-2B'X'X\hat{B}+B'\Phi_0^{-1}B+B_0'\Phi_0^{-1}B_0-2B'\Phi_0^{-1}B_0\right]\\ &=\Sigma^{-1}\left[B'\left(X'X+\Phi_0^{-1}\right)B-2B'\left(X'X\hat{B}+\Phi_0^{-1}B_0\right)+\hat{B}'X'X\hat{B}+B_0'\Phi_0^{-1}B_0\right]\end{aligned}$$

Complete the squares:

$$=\Sigma^{-1}[B'\left(X'X+\Phi_0^{-1}\right)B-2B'\bar{\Phi}^{-1}\bar{\Phi}\left(X'X\hat{B}+\Phi_0^{-1}B_0\right)+\bar{B}'\bar{\Phi}^{-1}\bar{B}-\bar{B}'\bar{\Phi}^{-1}\bar{B}+\hat{B}'X'X\hat{B}+B_0'\Phi_0^{-1}B_0]$$

Now, define:

$$\bar{\Phi}=\left[\Phi_0^{-1}+X'X\right]^{-1} \qquad \text{(A.4.12)}$$

and

$$\bar{B}=\bar{\Phi}\left[\Phi_0^{-1}B_0+X'X\hat{B}\right] \qquad \text{(A.4.13)}$$

Then, the previous expression rewrites:

$$
\begin{aligned}
&= \Sigma^{-1}\left[B'\bar{\Phi}^{-1}B - 2B'\bar{\Phi}^{-1}\bar{B} + \bar{B}'\bar{\Phi}^{-1}\bar{B} - \bar{B}'\bar{\Phi}^{-1}\bar{B} + \hat{B}'X'X\hat{B} + B_0'\Phi_0^{-1}B_0\right] \\
&= \Sigma^{-1}\left[(B-\bar{B})'\bar{\Phi}^{-1}(B-\bar{B}) - \bar{B}'\bar{\Phi}^{-1}\bar{B} + \hat{B}'X'X\hat{B} + B_0'\Phi_0^{-1}B_0\right] \\
&= \Sigma^{-1}\left[(B-\bar{B})'\bar{\Phi}^{-1}(B-\bar{B})\right] + \Sigma^{-1}\left[\hat{B}'X'X\hat{B} + B_0'\Phi_0^{-1}B_0 - \bar{B}'\bar{\Phi}^{-1}\bar{B}\right]
\end{aligned}
$$

Substituting back in A.4.11, the posterior becomes:

$$
\begin{aligned}
&\pi(\beta,\Sigma\,|y) \propto \\
&|\Sigma|^{-(T+k+\alpha_0+n+1)/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left[(B-\bar{B})'\bar{\Phi}^{-1}(B-\bar{B})\right] + \Sigma^{-1}\left[\hat{B}'X'X\hat{B} + B_0'\Phi_0^{-1}B_0 - \bar{B}'\bar{\Phi}^{-1}\bar{B}\right]\right\}\right] \\
&\times \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left[S_0 + (Y-X\hat{B})'(Y-X\hat{B})\right]\right\}\right] \\
&= |\Sigma|^{-(T+k+\alpha_0+n+1)/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left[(B-\bar{B})'\bar{\Phi}^{-1}(B-\bar{B})\right]\right\}\right] \\
&\times \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left[S_0 + (Y-X\hat{B})'(Y-X\hat{B}) + \hat{B}'X'X\hat{B} + B_0'\Phi_0^{-1}B_0 - \bar{B}'\bar{\Phi}^{-1}\bar{B}\right]\right\}\right] \\
&= |\Sigma|^{-k/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left[(B-\bar{B})'\bar{\Phi}^{-1}(B-\bar{B})\right]\right\}\right] \\
&\times |\Sigma|^{-(T+\alpha_0+n+1)/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left[S_0 + (Y-X\hat{B})'(Y-X\hat{B}) + \hat{B}'X'X\hat{B} + B_0'\Phi_0^{-1}B_0 - \bar{B}'\bar{\Phi}^{-1}\bar{B}\right]\right\}\right]
\end{aligned}
$$

Define:

$$\bar{\alpha} = T + \alpha_0 \tag{A.4.14}$$

and

$$\bar{S} = \left(Y - X\hat{B}\right)'\left(Y - X\hat{B}\right) + S_0 + \hat{B}'X'X\hat{B} + B_0'\Phi_0^{-1}B_0 - \bar{B}'\bar{\Phi}^{-1}\bar{B} \tag{A.4.15}$$

This allows to rewrite the previous expression as:

$$
\begin{aligned}
\pi(\beta,\Sigma\,|y) \propto |\Sigma|^{-k/2}&\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left[(B-\bar{B})'\bar{\Phi}^{-1}(B-\bar{B})\right]\right\}\right] \\
&\times |\Sigma|^{-(\bar{\alpha}+n+1)/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\bar{S}\right\}\right]
\end{aligned} \tag{A.4.16}
$$

This can be recognised as the product of a matrix-variate normal distribution with mean $\bar{B}$ and

variance matrices Σ and $\bar{\Phi}$, and an inverse Wishart distribution with scale matrix $\bar{S}$ and degrees of freedom $\bar{\alpha}$. Alternatively, using A.1.10, one can rewrite A.4.16 as:

$$\pi(\beta, \Sigma \,|y\,) \propto |\Sigma|^{-k/2} \exp\left[-\frac{1}{2}(\beta - \bar{\beta})'\left(\Sigma \otimes \bar{\Phi}\right)^{-1}(\beta - \bar{\beta})\right] \times |\Sigma|^{-(\bar{\alpha}+n+1)/2} \exp\left[-\frac{1}{2} tr\left\{\Sigma^{-1}\bar{S}\right\}\right] \tag{A.4.17}$$

and recognise for β the kernel of a multivariate normal distribution with mean $\bar{\beta} = vec(\bar{B})$ and covariance matrix $\Sigma \otimes \bar{\Phi}$.

Note finally that this is a joint distribution, while the objects of interest are the marginal distributions for β and Σ. Deriving the marginal for Σ using 3.2.7 is quite trivial: it is easy to integrate out β as it only appears in the first term of A.4.17 as a multivariate normal variable. Doing so leaves us only with the second term. The details are as follows:

$$\begin{aligned}
\pi(\Sigma \,|y\,) &= \int_{\beta} \pi(\beta, \Sigma \,|y\,) d\beta \\
&\propto \int_{\beta} |\Sigma|^{-k/2} \exp\left[-\frac{1}{2}(\beta - \bar{\beta})'\left(\Sigma \otimes \bar{\Phi}\right)^{-1}(\beta - \bar{\beta})\right] \times |\Sigma|^{-(\bar{\alpha}+n+1)/2} \exp\left[-\frac{1}{2} tr\left\{\Sigma^{-1}\bar{S}\right\}\right] d\beta \\
&= |\Sigma|^{-(\bar{\alpha}+n+1)/2} \exp\left[-\frac{1}{2} tr\left\{\Sigma^{-1}\bar{S}\right\}\right] \times \int_{\beta} |\Sigma|^{-k/2} \exp\left[-\frac{1}{2}(\beta - \bar{\beta})'\left(\Sigma \otimes \bar{\Phi}\right)^{-1}(\beta - \bar{\beta})\right] d\beta \\
&= |\Sigma|^{-(\bar{\alpha}+n+1)/2} \exp\left[-\frac{1}{2} tr\left\{\Sigma^{-1}\bar{S}\right\}\right]
\end{aligned} \tag{A.4.18}$$

Which is once again immediately recognised as the kernel of an inverse Wishart distribution :

$$\pi(\Sigma \,|y\,) \sim \; IW\left(\bar{\alpha}, \bar{S}\right) \tag{A.4.19}$$

Deriving the posterior for β is trickier. Start by regrouping terms in A.4.16, then integrate with respect to Σ :

$$\begin{aligned}
\pi(\beta \,|y\,) &= \int_{\Sigma} \pi(\beta, \Sigma \,|y\,) d\Sigma \\
&\propto \int_{\Sigma} |\Sigma|^{-(k+\bar{\alpha}+n+1)/2} \exp\left[-\frac{1}{2} tr\left\{\Sigma^{-1}\left[(B - \bar{B})'\bar{\Phi}^{-1}(B - \bar{B}) + \bar{S}\right]\right\}\right] d\Sigma
\end{aligned}$$

Since the integrand has the form of an inverse Wishart distribution with scale matrix $(B - \bar{B})'\bar{\Phi}^{-1}(B-$

$\bar{B})+\bar{S}$ and degrees of freedom $k+\bar{\alpha}$, integration yields the reciprocal of the constant of that function [9]:

$$\pi(\beta\,|y)\propto 2^{\frac{(k+\bar{\alpha})n}{2}}\Gamma_n\left(\frac{k+\bar{\alpha}}{2}\right)\times\left|\bar{S}+(B-\bar{B})'\bar{\Phi}^{-1}(B-\bar{B})\right|^{-\frac{k+\bar{\alpha}}{2}} \tag{A.4.20}$$

But only the final term contains B. Accordingly, eliminating from A.4.20 the terms belonging to the proportionality constant, one obtains:

$$\begin{aligned}\pi(\beta\,|y)&\propto\left|\bar{S}+(B-\bar{B})'\bar{\Phi}^{-1}(B-\bar{B})\right|^{-\frac{k+\bar{\alpha}}{2}}\\&=\left|\bar{S}\left\{I_n+\bar{S}^{-1}(B-\bar{B})'\bar{\Phi}^{-1}(B-\bar{B})\right\}\right|^{-\frac{k+T+\alpha_0}{2}}\\&=\left|\bar{S}\right|^{-\frac{k+T+\alpha_0}{2}}\left|I_n+\bar{S}^{-1}(B-\bar{B})'\bar{\Phi}^{-1}(B-\bar{B})\right|^{-\frac{k+T+\alpha_0}{2}}\\&\propto\left|I_n+\bar{S}^{-1}(B-\bar{B})'\bar{\Phi}^{-1}(B-\bar{B})\right|^{-\frac{k+T+\alpha_0}{2}}\\&=\left|I_n+\bar{S}^{-1}(B-\bar{B})'\bar{\Phi}^{-1}(B-\bar{B})\right|^{-\frac{[T+\alpha_0-n+1]+n+k-1}{2}}\end{aligned}$$

Hence:

$$\pi(\beta\,|y)\propto\left|I_n+\bar{S}^{-1}(B-\bar{B})'\bar{\Phi}^{-1}(B-\bar{B})\right|^{-\frac{[T+\alpha_0-n+1]+n+k-1}{2}} \tag{A.4.21}$$

This is the kernel of a matrix-variate student distribution with mean $\bar{B}$, scale matrices $\bar{S}$ and $\bar{\Phi}$, and degrees of freedom $\alpha=T+\alpha_0-n+1$.

$$B\sim\ MT(\bar{B},\bar{S},\bar{\Phi},\widetilde{\alpha}) \tag{A.4.22}$$

A.2.5.5 then implies that each individual element $B_{i,j}$ of B follows a univariate student distribution with mean $\bar{B}_{i,j}$, scale parameter $\bar{\Phi}_{i,i}\times\bar{S}_{j,j}$ and degrees of freedom α.

$$B_{i,j}\sim\ t(\bar{B}_{i,j},\bar{\Phi}_{i,i}\times\bar{S}_{j,j},\widetilde{\alpha}) \tag{A.4.23}$$

Therefore, conditional on prior choices for β_0,Φ_0,S_0 and α_0, point estimates and inference can be realised on β and Σ from A.4.19 and A.4.22 , A.4.23 .

It should also be clear now why the Kronecker structure $\Sigma\otimes\Phi_0$ is imposed on the variance matrix of the prior distribution for β. Without this structure, it would be impossible to transform the prior distribution $\pi(\beta)$ into a trace argument as done in A.4.9. It would then not be possible to complete the squares and reformulate the posterior as in A.4.16, which is necessary to derive the unconditional

[9] Indeed, if a random variable X has density $f(X)=c\times g(X)$, with c the proportionality constant and $g(X)$ the variable part, then from the definition of a density, one obtains $1=\int_X f(X)dX=c\int_X g(X)dX$, so that $\int_X g(X)dX=\frac{1}{c}$.

marginal posterior of B as a student distribution.

Finally, notice that it is possible to simplify some of the expressions derived in this section. The first term that can be simplified is $\bar{S}$, defined in A.4.15. Keeping in mind definition 3.7.2 of the OLS estimate $\hat{B}$, notice that:

$$
\begin{aligned}
&(Y - X\hat{B})'(Y - X\hat{B}) + \hat{B}'X'X\hat{B} \\
&= Y'Y - Y'X\hat{B} - (X\hat{B})'Y + (X\hat{B})'(X\hat{B}) + \hat{B}'X'X\hat{B} \\
&= Y'Y - Y'X\hat{B} - \hat{B}'X'Y + \hat{B}'X'X\hat{B} + \hat{B}'X'X\hat{B} \\
&= Y'Y - Y'X\hat{B} - \hat{B}'X'Y + 2\hat{B}'X'X\hat{B} \\
&= Y'Y - Y'X\left\{(X'X)^{-1}X'Y\right\} - \left\{(X'X)^{-1}X'Y\right\}'X'Y + 2\left\{(X'X)^{-1}X'Y\right\}'X'X\left\{(X'X)^{-1}X'Y\right\} \\
&= Y'Y - Y'X(X'X)^{-1}X'Y - Y'X(X'X)^{-1}X'Y + 2Y'X(X'X)^{-1}X'X(X'X)^{-1}X'Y \\
&= Y'Y - 2Y'X(X'X)^{-1}X'Y + 2Y'X(X'X)^{-1}X'Y \\
&= Y'Y
\end{aligned}
$$

Therefore, $\bar{S}$ can rewrite as:

$$
\begin{aligned}
\bar{S} &= (Y - X\hat{B})'(Y - X\hat{B}) + S_0 + \hat{B}'X'X\hat{B} + B_0'\Phi_0^{-1}B_0 - \bar{B}'\bar{\Phi}^{-1}\bar{B} \\
&= Y'Y + S_0 + B_0'\Phi_0^{-1}B_0 - \bar{B}'\bar{\Phi}^{-1}\bar{B}
\end{aligned} \tag{A.4.24}
$$

From a similar reasoning, $\bar{B}$ defined in A.4.13 can be reformulated as:

$$
\bar{B} = \bar{\Phi}\left[\Phi_0^{-1}B_0 + X'X\hat{B}\right] = \bar{\Phi}\left[\Phi_0^{-1}B_0 + X'X\left\{(X'X)^{-1}X'Y\right\}\right] = \bar{\Phi}\left[\Phi_0^{-1}B_0 + X'Y\right]
$$

A.5 Derivations of the posterior distribution with an independent normal-Wishart prior

Obtaining the posterior distribution for β and Σ requires as usual a likelihood function for the data, and a prior for β and Σ.

The likelihood is similar as that of a normal Wishart and is thus given by 3.4.3, rewritten here with simplified powers on $|\Sigma|$:

$$f(y\,|\beta,\Sigma) \propto |\Sigma|^{-T/2} \exp\left[-\frac{1}{2}\left(\beta-\hat{\beta}\right)'\left(\Sigma\otimes(X'X)^{-1}\right)^{-1}\left(\beta-\hat{\beta}\right)\right]$$
$$\times \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right\}\right] \tag{A.5.1}$$

The prior for β is given by 3.5.3, and that for Σ is given by 3.5.5. Using Bayes rule 3.2.5, the posterior is then given by:

$$\pi(\beta,\Sigma\,|y) \propto |\Sigma|^{-T/2} \exp\left[-\frac{1}{2}\left(\beta-\hat{\beta}\right)'\left(\Sigma\otimes(X'X)^{-1}\right)^{-1}\left(\beta-\hat{\beta}\right)\right]$$
$$\times \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right\}\right]$$
$$\times \exp\left[-\frac{1}{2}(\beta-\beta_0)'\Omega_0^{-1}(\beta-\beta_0)\right]$$
$$\times |\Sigma|^{-(\alpha_0+n+1)/2} \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}S_0\right\}\right] \tag{A.5.2}$$

Or, rearranging and using A.1.2 and A.17.15:

$$\pi(\beta,\Sigma\,|y) \propto |\Sigma|^{-(T+\alpha_0+n+1)/2} \exp\left[-\frac{1}{2}\left\{\left(\beta-\hat{\beta}\right)'\left(\Sigma^{-1}\otimes X'X\right)\left(\beta-\hat{\beta}\right)+(\beta-\beta_0)'\Omega_0^{-1}(\beta-\beta_0)\right\}\right]$$
$$\times \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left[S_0+(Y-X\hat{B})'(Y-X\hat{B})\right]\right\}\right] \tag{A.5.3}$$

Consider only the term in the curly brackets in the first row:

$$\left(\beta-\hat{\beta}\right)'\left(\Sigma^{-1}\otimes X'X\right)\left(\beta-\hat{\beta}\right)+(\beta-\beta_0)'\Omega_0^{-1}(\beta-\beta_0)$$
$$=\beta'\left(\Sigma^{-1}\otimes X'X\right)\beta+\hat{\beta}'\left(\Sigma^{-1}\otimes X'X\right)\hat{\beta}-2\beta'\left(\Sigma^{-1}\otimes X'X\right)\hat{\beta}$$
$$+\beta'\Omega_0^{-1}\beta+\beta_0'\Omega_0^{-1}\beta_0-2\beta'\Omega_0^{-1}\beta_0$$
$$=\beta'\left[\Omega_0^{-1}+\Sigma^{-1}\otimes X'X\right]\beta-2\beta'\left[\Omega_0^{-1}\beta_0+\left(\Sigma^{-1}\otimes X'X\right)\hat{\beta}\right]+\hat{\beta}'\left(\Sigma^{-1}\otimes X'X\right)\hat{\beta}+\beta_0'\Omega_0^{-1}\beta_0$$

Complete the squares:

$$= \beta' \left[\Omega_0^{-1} + \Sigma^{-1} \otimes X'X\right] \beta - 2\beta'\bar{\Omega}^{-1}\bar{\Omega}\left[\Omega_0^{-1}\beta_0 + \left(\Sigma^{-1} \otimes X'X\right)\hat{\beta}\right] + \bar{\beta}'\bar{\Omega}^{-1}\bar{\beta} - \bar{\beta}'\bar{\Omega}^{-1}\bar{\beta}$$
$$+ \hat{\beta}' \left(\Sigma^{-1} \otimes X'X\right)\hat{\beta} + \beta_0'\Omega_0^{-1}\beta_0$$

Define:

$$\bar{\Omega} = \left[\Omega_0^{-1} + \Sigma^{-1} \otimes X'X\right]^{-1} \tag{A.5.4}$$

$$\bar{\beta} = \bar{\Omega}\left[\Omega_0^{-1}\beta_0 + \left(\Sigma^{-1} \otimes X'X\right)\hat{\beta}\right] \tag{A.5.5}$$

Then the above rewrites:

$$= \beta'\bar{\Omega}^{-1}\beta - 2\beta'\bar{\Omega}^{-1}\bar{\beta} + \bar{\beta}'\bar{\Omega}^{-1}\bar{\beta} - \bar{\beta}'\bar{\Omega}^{-1}\bar{\beta} + \hat{\beta}' \left(\Sigma^{-1} \otimes X'X\right)\hat{\beta} + \beta_0'\Omega_0^{-1}\beta_0$$
$$= (\beta - \bar{\beta})'\bar{\Omega}^{-1}(\beta - \bar{\beta}) - \bar{\beta}'\bar{\Omega}^{-1}\bar{\beta} + \hat{\beta}' \left(\Sigma^{-1} \otimes X'X\right)\hat{\beta} + \beta_0'\Omega_0^{-1}\beta_0$$

Substituting back in A.5.3:

$$\begin{aligned}\pi(\beta, \Sigma\,|y) \propto |\Sigma|^{-(T+\alpha_0+n+1)/2} \exp\left[-\frac{1}{2}(\beta - \bar{\beta})'\bar{\Omega}^{-1}(\beta - \bar{\beta})\right] \\ \times \exp\left[-\frac{1}{2}\left\{\hat{\beta}' \left(\Sigma^{-1} \otimes X'X\right)\hat{\beta} + \beta_0'\Omega_0^{-1}\beta_0 - \bar{\beta}'\bar{\Omega}^{-1}\bar{\beta}\right\}\right] \\ \times \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left[S_0 + (Y - X\hat{B})'(Y - X\hat{B})\right]\right\}\right]\end{aligned} \tag{A.5.6}$$

A.5.6 is recognised as 3.5.8 in the text. Note also that A.5.5 can be simplified. Consider the part $\left(\Sigma^{-1} \otimes X'X\right)\hat{\beta}$, and use definition A.4.3 of $\hat{\beta}$ to obtain:

$$\begin{aligned}&\left(\Sigma^{-1} \otimes X'X\right)\hat{\beta} \\ &= \left(\Sigma^{-1} \otimes X'X\right)\left(\Sigma^{-1} \otimes X'X\right)^{-1}\left(\Sigma^{-1} \otimes X'\right)y \\ &= \left(\Sigma^{-1} \otimes X'\right)y\end{aligned}$$

Hence, A.5.5 rewrites as:

$$\bar{\beta} = \bar{\Omega}\left[\Omega_0^{-1}\beta_0 + (\Sigma^{-1} \otimes X')y\right] \tag{A.5.7}$$

A.5.4 and A.5.7 are then recognised as 3.5.9 and 3.5.10 in the text. One may also note that A.5.7

and 3.3.18 are similar, implying that the independent normal Wishart prior yields a (conditional) posterior comparable to the Minnesota posterior with respect to β.

Now compute the conditional distributions. Relegating any term not involving β in A.5.6 to a proportionality constant, one obtains the distribution of β, conditional on Σ:

$$\pi(\beta\,|\Sigma, y\,) \propto |\Sigma|^{-(T+\alpha_0+n+1)/2} \exp\left[-\frac{1}{2}(\beta-\bar{\beta})'\bar{\Omega}^{-1}(\beta-\bar{\beta})\right] \tag{A.5.8}$$

This is recognised as the kernel of a multivariate normal distribution:

$$\pi(\beta\,|\Sigma, y) \sim\ N(\bar{\beta}, \bar{\Omega}) \tag{A.5.9}$$

To obtain the posterior for Σ, conditional on β, it is easier to work directly on A.5.3, since A.5.6 contains $\bar{\beta}$ and $\bar{\Omega}$, which are complicated functions of Σ and do not allow transformations into trace arguments. Hence consider the posterior A.5.3, ignore terms not involving Σ, and notice that the remaining terms rewrite as:

$$\begin{aligned}
\pi(\Sigma\,|\beta, y\,) \propto\ & |\Sigma|^{-(T+\alpha_0+n+1)/2} \exp\left[-\frac{1}{2}\left\{(\beta-\hat{\beta})'\left(\Sigma\otimes(X'X)^{-1}\right)^{-1}(\beta-\hat{\beta})\right\}\right] \quad A.1.2 \\
& \times \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left[S_0+(Y-X\hat{B})'(Y-X\hat{B})\right]\right\}\right] \\
=\ & |\Sigma|^{-(T+\alpha_0+n+1)/2} \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(B-\hat{B})'(X'X)(B-\hat{B})\right\}\right] \quad A.1.2 \\
& \times \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left[S_0+(Y-X\hat{B})'(Y-X\hat{B})\right]\right\}\right] \\
=\ & |\Sigma|^{-[(T+\alpha_0)+n+1]/2} \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left[(B-\hat{B})'(X'X)(B-\hat{B})+S_0+(Y-X\hat{B})'(Y-X\hat{B})\right]\right\}\right]
\end{aligned} \tag{A.5.10}$$

One then recognises in A.5.10 the kernel of an inverse Wishart distribution

$$\pi(\Sigma\,|\beta, y\,) \sim\ IW(\hat{S}, \hat{\alpha}) \tag{A.5.11}$$

with scale matrix:

$$\hat{S} = (B-\hat{B})'(X'X)(B-\hat{B}) + S_0 + (Y-X\hat{B})'(Y-X\hat{B}) \tag{A.5.12}$$

and degrees of freedom:

$$\hat{\alpha} = T + \alpha_0 \tag{A.5.13}$$

Eventually, note that similarly to the normal-Wishart prior, it is possible to apply some simplifications. Consider $\hat{S}$, defined in A.5.12. Using once again definition 3.7.2 of the OLS estimate $\hat{B}$, simplify:

$$
\begin{aligned}
&(B-\hat{B})'(X'X)(B-\hat{B})+(Y-X\hat{B})'(Y-X\hat{B})\\
&=B'(X'X)B-B'(X'X)\hat{B}-\hat{B}'(X'X)B+\hat{B}'(X'X)\hat{B}\\
&+Y'Y-Y'X\hat{B}-(X\hat{B})'Y+(X\hat{B})'(X\hat{B})\\
&=B'X'XB-B'X'X\hat{B}-\hat{B}'X'XB+\hat{B}'X'X\hat{B}\\
&+Y'Y-Y'X\hat{B}-\hat{B}'X'Y+\hat{B}'X'X\hat{B}\\
&=B'X'XB-B'X'X\left\{(X'X)^{-1}X'Y\right\}-\left\{(X'X)^{-1}X'Y\right\}'X'XB+\left\{(X'X)^{-1}X'Y\right\}'X'X\left\{(X'X)^{-1}X'Y\right\}\\
&+Y'Y-Y'X\left\{(X'X)^{-1}X'Y\right\}-\left\{(X'X)^{-1}X'Y\right\}'X'Y+\left\{(X'X)^{-1}X'Y\right\}'X'X\left\{(X'X)^{-1}X'Y\right\}\\
&=B'X'XB-B'X'X(X'X)^{-1}X'Y-Y'X(X'X)^{-1}X'XB+Y'X(X'X)^{-1}X'X(X'X)^{-1}X'Y\\
&+Y'Y-Y'X(X'X)^{-1}X'Y-Y'X(X'X)^{-1}X'Y+Y'X(X'X)^{-1}X'X(X'X)^{-1}X'Y\\
&=B'X'XB-B'X'Y-Y'XB+Y'X(X'X)^{-1}X'Y\\
&+Y'Y-Y'X(X'X)^{-1}X'Y-Y'X(X'X)^{-1}X'Y+Y'X(X'X)^{-1}X'Y\\
&=Y'Y-Y'XB-B'X'Y+B'X'XB\\
&=(Y-XB)'(Y-XB)
\end{aligned}
\tag{A.5.14}
$$

Therefore, one can rewrite:

$$
\begin{aligned}
\hat{S}&=(B-\hat{B})'(X'X)(B-\hat{B})+S_0+(Y-X\hat{B})'(Y-X\hat{B})\\
&=(Y-XB)'(Y-XB)+S_0
\end{aligned}
\tag{A.5.15}
$$

And A.5.10 can rewrite:

$$
\pi(\Sigma\,|\beta,y)\propto|\Sigma|^{-[(T+\alpha_0)+n+1]/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left[(Y-XB)'(Y-XB)+S_0\right]\right\}\right] \tag{A.5.16}
$$

A.6 Derivations of the posterior distribution with a normal-diffuse prior

Obtaining the posterior distribution for β and Σ requires the usual likelihood function for the data, and a prior for β and Σ. The likelihood is given by 3.6.1:

$$f(y\,|\beta,\Sigma) \propto |\Sigma|^{-T/2} \exp\left[-\frac{1}{2}\left(\beta-\hat{\beta}\right)'\left(\Sigma\otimes(X'X)^{-1}\right)^{-1}\left(\beta-\hat{\beta}\right)\right]$$
$$\times \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right\}\right] \tag{A.6.1}$$

The priors for β and Σ are respectively given by 3.6.2 and 3.6.3:

$$\pi(\beta) \propto \exp\left[-\frac{1}{2}(\beta-\beta_0)'\Omega_0^{-1}(\beta-\beta_0)\right] \tag{A.6.2}$$

and

$$\pi(\Sigma) \propto |\Sigma|^{-(n+1)/2} \tag{A.6.3}$$

Using Bayes rule 3.2.5, the posterior is then given by:

$$\pi(\beta,\Sigma\,|y) \propto |\Sigma|^{-T/2} \exp\left[-\frac{1}{2}\left(\beta-\hat{\beta}\right)'\left(\Sigma\otimes(X'X)^{-1}\right)^{-1}\left(\beta-\hat{\beta}\right)\right]$$
$$\times \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right\}\right]$$
$$\times \exp\left[-\frac{1}{2}(\beta-\beta_0)'\Omega_0^{-1}(\beta-\beta_0)\right]$$
$$\times |\Sigma|^{-(n+1)/2} \tag{A.6.4}$$

Or, rearranging and using A.1.2:

$$\pi(\beta,\Sigma\,|y) \propto |\Sigma|^{-T/2} \exp\left[-\frac{1}{2}\left(\beta-\hat{\beta}\right)'\left(\Sigma^{-1}\otimes(X'X)^{-1}\right)\left(\beta-\hat{\beta}\right)+(\beta-\beta_0)'\Omega_0^{-1}(\beta-\beta_0)\right]$$
$$\times \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right\}\right] \tag{A.6.5}$$

The expression in the curly brackets appearing in the first term is similar to that in A.5.3. Hence, using the same process as in Appendix A.5, one shows that it rewrites as:

$$\left(\beta-\hat{\beta}\right)'\left(\Sigma^{-1}\otimes X'X\right)\left(\beta-\hat{\beta}\right)+(\beta-\beta_0)'\Omega_0^{-1}(\beta-\beta_0)$$
$$=(\beta-\bar{\beta})'\bar{\Omega}^{-1}(\beta-\bar{\beta})-\bar{\beta}'\bar{\Omega}^{-1}\bar{\beta}+\hat{\beta}'\left(\Sigma^{-1}\otimes X'X\right)\hat{\beta}+\beta_0'\Omega_0^{-1}\beta_0 \tag{A.6.6}$$

$\bar{\beta}$ and $\bar{\Omega}$ are defined as in Appendix A.5 as:

$$\bar{\Omega}=\left[\Omega_0^{-1}+\Sigma^{-1}\otimes X'X\right]^{-1} \tag{A.6.7}$$

and

$$\bar{\beta}=\bar{\Omega}\left[\Omega_0^{-1}\beta_0+(\Sigma^{-1}\otimes X')y\right] \tag{A.6.8}$$

Therefore, A.6.5 rewrites as:

$$\pi(\beta,\Sigma\,|y)\propto|\Sigma|^{-(T+n+1)/2}\exp\left[-\frac{1}{2}(\beta-\bar{\beta})'\bar{\Omega}^{-1}(\beta-\bar{\beta})\right]$$
$$\times\exp\left[-\frac{1}{2}\left\{\hat{\beta}'\left(\Sigma^{-1}\otimes X'X\right)\hat{\beta}+\beta_0'\Omega_0^{-1}\beta_0-\bar{\beta}'\bar{\Omega}^{-1}\bar{\beta}\right\}\right]$$
$$\times\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right\}\right] \tag{A.6.9}$$

Now compute the conditional distributions. Relegating any term not involving β in A.6.9 to a proportionality constant, one obtains the distribution of β, conditional on Σ:

$$\pi(\beta\,|\Sigma,y)\propto\exp\left[-\frac{1}{2}(\beta-\bar{\beta})'\bar{\Omega}^{-1}(\beta-\bar{\beta})\right] \tag{A.6.10}$$

This is recognised as the kernel of a multivariate normal distribution:

$$\pi(\beta\,|\Sigma,y)\sim\mathcal{N}(\bar{\beta},\bar{\Omega}) \tag{A.6.11}$$

To obtain the posterior for Σ, conditional on β, it is once again easier to work directly on the primary posterior expression A.6.5, rather than with the modified expression A.6.9. Hence consider the posterior A.6.5, ignore terms not involving Σ, and notice that the remaining terms rewrite as:

$$\pi(\Sigma\,|\beta,y\,) \propto |\Sigma|^{-(T+n+1)/2} \exp\left[-\frac{1}{2}\left(\beta-\hat{\beta}\right)^{'}\left(\Sigma^{-1}\otimes X^{'}X\right)\left(\beta-\hat{\beta}\right)\right]$$
$$\times \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(Y-X\hat{B})^{'}(Y-X\hat{B})\right\}\right]$$
$$= |\Sigma|^{-(T+n+1)/2} \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(B-\hat{B})^{'}(X^{'}X)(B-\hat{B})\right\}\right]$$
$$\times \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(Y-X\hat{B})^{'}(Y-X\hat{B})\right\}\right]$$
$$= |\Sigma|^{-(T+n+1)/2} \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left[(B-\hat{B})^{'}(X^{'}X)(B-\hat{B})+(Y-X\hat{B})^{'}(Y-X\hat{B})\right]\right\}\right] \quad \text{(A.6.12)}$$

From A.5.14, A.6.12 simplifies to:

$$\pi(\Sigma\,|\beta,y\,) \propto |\Sigma|^{-[(T+\alpha_0)+n+1]/2} \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left[(Y-XB)^{'}(Y-XB)\right]\right\}\right]$$

One then recognises in A.6.12 the kernel of an inverse Wishart distribution

$$\pi(\Sigma\,|\beta,y\,) \sim \mathcal{IW}(\tilde{S},T) \quad \text{(A.6.13)}$$

with $\tilde{S}$ the scale matrix defined as:

$$\tilde{S} = (Y-XB)^{'}(Y-XB) \quad \text{(A.6.14)}$$

and degrees of freedom equal to T.

A.7 Derivations for the dummy observation prior

The kernel of the posterior distribution is given by:

$$f(\beta,\Sigma\,|y\,) \propto |\Sigma|^{-(T+n+1)/2} \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(B-\hat{B})^{'}(X^{'}X)(B-\hat{B})\right\}\right]$$
$$\times \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(Y-X\hat{B})^{'}(Y-X\hat{B})\right\}\right] \quad \text{(A.7.1)}$$

Reformulate first to obtain a form that will be easier to integrate:

$$f(\beta,\Sigma\,|y)\propto|\Sigma|^{-k/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(B-\hat{B})'(X'X)(B-\hat{B})\right\}\right]$$
$$\times|\Sigma|^{-(T-k+n+1)/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right\}\right] \tag{A.7.2}$$

Integrate first with respect to β to obtain the posterior for Σ :

$$f(\Sigma\,|y)\propto\int_B|\Sigma|^{-k/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(B-\hat{B})'(X'X)(B-\hat{B})\right\}\right]dB$$
$$\times|\Sigma|^{-(T-k+n+1)/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right\}\right]$$
$$=|\Sigma|^{-(T-k+n+1)/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right\}\right] \tag{A.7.3}$$

This is the kernel of an inverse Whishart distribution: $\Sigma\sim IW\left(\hat{S},\hat{\alpha}\right)$, with scale matrix :

$$\hat{S}=(Y-X\hat{B})'(Y-X\hat{B}) \tag{A.7.4}$$

and degrees of freedom

$$\hat{\alpha}=T-k \tag{A.7.5}$$

Now integrate with respect to Σ to obtain the posterior for β. To make things easier, gather first terms in A.7.1:

$$f(\beta,\Sigma\,|y)\propto|\Sigma|^{-(T+n+1)/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left[\hat{S}+(B-\hat{B})'(X'X)(B-\hat{B})\right]\right\}\right] \tag{A.7.6}$$

This can be recognized as the kernel of an inverse-Wishart density with scale matrix $\hat{S}+(B-\hat{B})'(X'X)(B-\hat{B})$ and degrees of freedom T. To integrate, note that similarly to appendix 4, integration will yield the reciprocal of the normalizing constant of the inverse-Wishart density:

$$\pi(\beta\,|y)\propto\int_\Sigma|\Sigma|^{-(T+n+1)/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left[\hat{S}+(B-\hat{B})'(X'X)(B-\hat{B})\right]\right\}\right]d\Sigma$$
$$\propto 2^{\frac{Tn}{2}}\Gamma_n\left(\frac{T}{2}\right)\left|\hat{S}+(B-\hat{B})'(X'X)(B-\hat{B})\right|^{-\frac{T}{2}}$$

Since only the determinant term comprises B, eliminate the other parts:

$$
\begin{aligned}
\pi(\beta\,|y) &\propto \left|\hat{S} + (B-\hat{B})'(X'X)(B-\hat{B})\right|^{-\frac{T}{2}} \\
&= \left|\hat{S}\left\{I + \hat{S}^{-1}(B-\hat{B})'(X'X)(B-\hat{B})\right\}\right|^{-\frac{T}{2}} \\
&= \left|\hat{S}\right|^{-\frac{T}{2}}\left|\left\{I + \hat{S}^{-1}(B-\hat{B})'(X'X)(B-\hat{B})\right\}\right|^{-\frac{T}{2}} \\
&\propto \left|\left\{I_n + \hat{S}^{-1}(B-\hat{B})'(X'X)(B-\hat{B})\right\}\right|^{-\frac{T}{2}} \\
&\propto \left|\left\{I_n + \hat{S}^{-1}(B-\hat{B})'((X'X)^{-1})^{-1}(B-\hat{B})\right\}\right|^{-\frac{(T-n-k+1)+n+k-1}{2}}
\end{aligned}
$$

This is recognized as the kernel of a matrix student distribution: $B\tilde{}MT(\hat{B},\hat{S},\hat{\Phi},\widehat{\alpha})$, with $\hat{B}$ and $\hat{S}$ defined by 3.7.2 and 3.7.6, and:

$$\hat{\Phi} = (X'X)^{-1} \tag{A.7.7}$$

And

$$\widehat{\alpha} = T - n - k + 1 \tag{A.7.8}$$

It is now shown how to recover a VAR in error correction form from a standard VAR formulation. In general, a reduced-form VAR with p lags writes as:

$y_t = A_1 y_{t-1} + A_2 y_{t-2} + ... + A_p y_{t-p} + Cx_t + \varepsilon_t$

From this formulation, manipulate to obtain :

$$
\begin{aligned}
y_t &= \sum_{i=1}^{p} A_i y_{t-i} + Cx_t + \varepsilon_t \\
\Rightarrow y_t - y_{t-1} &= \sum_{i=1}^{p} A_i y_{t-i} - y_{t-1} + Cx_t + \varepsilon_t \\
\Rightarrow y_t - y_{t-1} &= \sum_{i=1}^{p} \left(A_i y_{t-i} + \sum_{j=i+1}^{p} A_j y_{t-j} - \sum_{j=i+1}^{p} A_j y_{t-j} \right) - y_{t-1} + Cx_t + \varepsilon_t \\
\Rightarrow y_t - y_{t-1} &= -\left(I - \sum_{i=1}^{p} A_i \right) y_{t-1} - \sum_{i=1}^{p-1} \left(\sum_{j=i+1}^{p} A_j \right) (y_{t-i} - y_{t-i-1}) + Cx_t + \varepsilon_t \\
\Rightarrow \Delta y_t &= -\left(I - \sum_{i=1}^{p} A_i \right) y_{t-1} - \sum_{i=1}^{p-1} \left(\sum_{j=i+1}^{p} A_j \right) \Delta y_{t-i} + Cx_t + \varepsilon_t \\
\Rightarrow \Delta y_t &= -\left(I - \sum_{i=1}^{p} A_i \right) y_{t-1} + \sum_{i=1}^{p-1} B_i \Delta y_{t-i} + Cx_t + \varepsilon_t
\end{aligned}
$$

$$
\Rightarrow \Delta y_t = -(I - A_1 - A_2 ... - A_p) y_{t-1} + B_1 \Delta y_{t-1} + B_2 \Delta y_{t-2} + ... + B_{p-1} \Delta y_{t-(p-1)} + Cx_t + \varepsilon_t \quad \text{(A.7.9)}
$$

with:
$B_i = -\sum_{j=i+1}^{p} A_j$
This is the error correction form 3.7.28.

A.8 Derivation of the marginal likelihood

Deriving the marginal likelihood for the Minnesota prior Because 3.9.10 may suffer from numerical instability, reformulate it to obtain a more stable equation. Start from 3.9.10 and manipulate:

$$
\begin{aligned}
m(y) &= (2\pi)^{-nT/2}\left|\bar{\Sigma}\right|^{-1/2}|\Omega_0|^{-1/2}\left|\bar{\Omega}\right|^{1/2}\exp\left[-\frac{1}{2}\left(\beta_0'\Omega_0^{-1}\beta_0-\bar{\beta}'\bar{\Omega}^{-1}\bar{\beta}+y'\bar{\Sigma}^{-1}y\right)\right] \\
&= (2\pi)^{-nT/2}|\Sigma\otimes I_T|^{-1/2}|\Omega_0|^{-1/2}\left|\left[\Omega_0^{-1}+\Sigma^{-1}\otimes X'X\right]^{-1}\right|^{1/2} \\
&\times\exp\left[-\frac{1}{2}\left(\beta_0'\Omega_0^{-1}\beta_0-\bar{\beta}'\bar{\Omega}^{-1}\bar{\beta}+y'(\Sigma\otimes I_T)^{-1}y\right)\right] \\
&= (2\pi)^{-nT/2}|\Sigma|^{-T/2}|I_T|^{-n/2}|\Omega_0|^{-1/2}\left|\Omega_0^{-1}+\Sigma^{-1}\otimes X'X\right|^{-1/2} \\
&\times\exp\left[-\frac{1}{2}\left(\beta_0'\Omega_0^{-1}\beta_0-\bar{\beta}'\bar{\Omega}^{-1}\bar{\beta}+y'\left(\Sigma^{-1}\otimes I_T\right)y\right)\right] \\
&= (2\pi)^{-nT/2}|\Sigma|^{-T/2}\left(|\Omega_0|\left|\Omega_0^{-1}+\Sigma^{-1}\otimes X'X\right|\right)^{-1/2} \\
&\times\exp\left[-\frac{1}{2}\left(\beta_0'\Omega_0^{-1}\beta_0-\bar{\beta}'\bar{\Omega}^{-1}\bar{\beta}+y'\left(\Sigma^{-1}\otimes I_T\right)y\right)\right] \\
&= (2\pi)^{-nT/2}|\Sigma|^{-T/2}\left|\Omega_0\left(\Omega_0^{-1}+\Sigma^{-1}\otimes X'X\right)\right|^{-1/2} \\
&\times\exp\left[-\frac{1}{2}\left(\beta_0'\Omega_0^{-1}\beta_0-\bar{\beta}'\bar{\Omega}^{-1}\bar{\beta}+y'\left(\Sigma^{-1}\otimes I_T\right)y\right)\right] \\
&= (2\pi)^{-nT/2}|\Sigma|^{-T/2}\left|I_{nk}+\Omega_0\left(\Sigma^{-1}\otimes X'X\right)\right|^{-1/2} \\
&\times\exp\left[-\frac{1}{2}\left(\beta_0'\Omega_0^{-1}\beta_0-\bar{\beta}'\bar{\Omega}^{-1}\bar{\beta}+y'\left(\Sigma^{-1}\otimes I_T\right)y\right)\right] \\
&= (2\pi)^{-nT/2}|\Sigma|^{-T/2}\left|I_{nk}+F_\Omega F_\Omega'\left(\Sigma^{-1}\otimes X'X\right)\right|^{-1/2} \\
&\times\exp\left[-\frac{1}{2}\left(\beta_0'\Omega_0^{-1}\beta_0-\bar{\beta}'\bar{\Omega}^{-1}\bar{\beta}+y'\left(\Sigma^{-1}\otimes I_T\right)y\right)\right]
\end{aligned}
$$

where F_Ω denotes the square root matrix of Ω_0, such that $F_\Omega F_\Omega' = \Omega_0$

$$
\begin{aligned}
&= (2\pi)^{-nT/2}|\Sigma|^{-T/2}\left|I_{nk}+F_\Omega'\left(\Sigma^{-1}\otimes X'X\right)F_\Omega\right|^{-1/2} \\
&\times\exp\left[-\frac{1}{2}\left(\beta_0'\Omega_0^{-1}\beta_0-\bar{\beta}'\bar{\Omega}^{-1}\bar{\beta}+y'\left(\Sigma^{-1}\otimes I_T\right)y\right)\right] \quad A.1.16
\end{aligned}
$$

From A.1.18, the determinant $|I_{nk}+F_\Omega'(\Sigma^{-1}\otimes X'X)F_\Omega|$ can be obtained from the product of 1 plus the eigenvalues of $F_\Omega'(\Sigma^{-1}\otimes X'X)F_\Omega$.

Deriving the marginal likelihood for the normal-Wishart prior

First show how to obtain 3.9.25 from 3.9.24:

$$
\begin{aligned}
m(y) &= (2\pi)^{-nT/2}|\Phi_0|^{-n/2}|S_0|^{\alpha_0/2}\left|\bar{\Phi}\right|^{n/2}\left|\bar{S}\right|^{-\bar{\alpha}/2}\frac{2^{\bar{\alpha}n/2}\Gamma_n\left(\frac{\bar{\alpha}}{2}\right)}{2^{\alpha_0 n/2}\Gamma_n\left(\frac{\alpha_0}{2}\right)} \\
&= 2^{-nT/2}\pi^{-nT/2}|\Phi_0|^{-n/2}|S_0|^{\alpha_0/2}\left|\bar{\Phi}\right|^{n/2}\left|\bar{S}\right|^{-\bar{\alpha}/2}\frac{2^{(T+\alpha_0)n/2}\Gamma_n\left(\frac{\bar{\alpha}}{2}\right)}{2^{\alpha_0 n/2}\Gamma_n\left(\frac{\alpha_0}{2}\right)} \\
&= 2^{-nT/2}2^{nT/2}\pi^{-nT/2}|\Phi_0|^{-n/2}|S_0|^{\alpha_0/2}\left|\bar{\Phi}\right|^{n/2}\left|\bar{S}\right|^{-\bar{\alpha}/2}\frac{2^{\alpha_0 n/2}\Gamma_n\left(\frac{\bar{\alpha}}{2}\right)}{2^{\alpha_0 n/2}\Gamma_n\left(\frac{\alpha_0}{2}\right)} \\
&= \pi^{-nT/2}|\Phi_0|^{-n/2}|S_0|^{\alpha_0/2}\left|\bar{\Phi}\right|^{n/2}\left|\bar{S}\right|^{-\bar{\alpha}/2}\frac{\Gamma_n\left(\frac{\bar{\alpha}}{2}\right)}{\Gamma_n\left(\frac{\alpha_0}{2}\right)}
\end{aligned}
$$

which is 3.9.25.

Now show how to obtain 3.9.26, for improved numerical accuracy. Start from 3.9.25:

$$
\begin{aligned}
m(y) &= \pi^{-nT/2}|\Phi_0|^{-n/2}|S_0|^{\alpha_0/2}\left|\bar{\Phi}\right|^{n/2}\left|\bar{S}\right|^{-\bar{\alpha}/2}\frac{\Gamma_n\left(\frac{\bar{\alpha}}{2}\right)}{\Gamma_n\left(\frac{\alpha_0}{2}\right)} \\
&= \pi^{-nT/2}|\Phi_0|^{-n/2}|S_0|^{\alpha_0/2}\left|\Phi_0^{-1}+X^{\cdot}X\right|^{-n/2}\left|\bar{S}\right|^{-\bar{\alpha}/2}\frac{\Gamma_n\left(\frac{\bar{\alpha}}{2}\right)}{\Gamma_n\left(\frac{\alpha_0}{2}\right)}
\end{aligned}
$$

Consider only the part $|\Phi_0|^{-n/2}\left|\Phi_0^{-1}+X^{\cdot}X\right|^{-n/2}$:

$$
\begin{aligned}
|\Phi_0|^{-n/2}\left|\Phi_0^{-1}+X^{\cdot}X\right|^{-n/2} &= |\Phi_0|^{-n/2}\left(\left|\Phi_0^{-1}\right|.|I_T+X\Phi_0X^{\cdot}|\right)^{-n/2} \quad A.1.13 \\
&= |\Phi_0|^{-n/2}|\Phi_0|^{n/2}|I_T+X\Phi_0X^{\cdot}|^{-n/2} \quad A.1.15 \\
&= |I_T+X\Phi_0X^{\cdot}|^{-n/2} \\
&= |I_T+XF_{\Phi}F_{\Phi}^{\cdot}X^{\cdot}|^{-n/2} \\
&= |I_k+F_{\Phi}^{\cdot}X^{\cdot}XF_{\Phi}|^{-n/2} \quad A.1.16
\end{aligned}
$$

where F_{Φ} denotes the square root matrix of Φ_0, that is, $F_{\Phi}F_{\Phi}^{\cdot} = \Phi_0$. Then, substituting, one eventually obtains:

$$
m(y) = \pi^{-nT/2}|S_0|^{\alpha_0/2}|I_k+F_{\Phi}^{\cdot}X^{\cdot}XF_{\Phi}|^{-n/2}\left|\bar{S}\right|^{-\bar{\alpha}/2}\frac{\Gamma_n\left(\frac{\bar{\alpha}}{2}\right)}{\Gamma_n\left(\frac{\alpha_0}{2}\right)}
$$

Now consider the part $|S_0|^{\alpha_0/2}\left|\bar{S}\right|^{-\bar{\alpha}/2}$:

$$
\begin{aligned}
|S_0|^{\alpha_0/2}\left|\bar{S}\right|^{-\bar{\alpha}/2} &= |S_0|^{\alpha_0/2}\left|S_0 + Y'Y + B_0'\Phi_0^{-1}B_0 - \bar{B}'\bar{\Phi}^{-1}\bar{B}\right|^{-\bar{\alpha}/2} \\
&= |S_0|^{\alpha_0/2}\left\{|S_0|\left|I_n + S_0^{-1}\left[Y'Y + B_0'\Phi_0^{-1}B_0 - \bar{B}'\bar{\Phi}^{-1}\bar{B}\right]\right|\right\}^{-\bar{\alpha}/2} \quad A.1.13 \\
&= |S_0|^{\alpha_0/2}|S_0|^{-\bar{\alpha}/2}\left|I_n + S_0^{-1}\left[\bar{S} - S_0\right]\right|^{-\bar{\alpha}/2} \\
&= |S_0|^{\alpha_0/2}|S_0|^{-(T+\alpha_0)/2}\left|I_n + S_0^{-1}\left[\bar{S} - S_0\right]\right|^{-\bar{\alpha}/2} \\
&= |S_0|^{-T}\left|I_n + F_S F_S'\left[\bar{S} - S_0\right]\right|^{-\bar{\alpha}/2}
\end{aligned}
$$

F_S denotes the inverse square root matrix of S_0 so that $F_S F_S' = S_0^{-1}$

$$
= |S_0|^{-T}\left|I_n + F_S'\left[\bar{S} - S_0\right]F_S\right|^{-\bar{\alpha}/2} A.1.16
$$

Substituting back:
$m(y) = \pi^{-nT/2}\frac{\Gamma_n\left(\frac{\bar{\alpha}}{2}\right)}{\Gamma_n\left(\frac{\alpha_0}{2}\right)}|I_k + F_\Phi' X'XF_\Phi|^{-n/2}|S_0|^{-T/2}\left|I_n + F_S'\left[\left(\bar{S} - S_0\right)\right]F_S\right|^{-\bar{\alpha}/2}$

which is 3.9.26.

Deriving the marginal likelihood for the independent normal-Wishart prior

First, show how to obtain 3.9.37. Start from 3.9.36, and substitute the likelihood function A.2.3.9 with the priors 3.3.13 and 3.9.19, all evaluated at $\tilde{\beta}$ and $\tilde{\Sigma}$:

$$
\begin{aligned}
m(y) &= (2\pi)^{-nT/2}\left|\tilde{\Sigma}\right|^{-T/2}\exp\left(-\frac{1}{2}tr\left[\tilde{\Sigma}^{-1}(Y - X\tilde{B})'(Y - X\tilde{B})\right]\right) \\
&\times \frac{(2\pi)^{-q/2}|\Omega_0|^{-1/2}\exp\left(-\frac{1}{2}(\tilde{\beta} - \beta_0)'\Omega_0^{-1}\left(\tilde{\beta} - \beta_0\right)\right)}{(2\pi)^{-q/2}\left|\bar{\Omega}\right|^{-1/2}\exp\left(-\frac{1}{2}(\tilde{\beta} - \bar{\beta})'\bar{\Omega}^{-1}(\tilde{\beta} - \bar{\beta})\right)} \\
&\times \frac{1}{2^{\alpha_0 n/2}\Gamma_n\left(\frac{\alpha_0}{2}\right)}|S_0|^{\alpha_0/2}\left|\tilde{\Sigma}\right|^{-(\alpha+n+1)/2}\exp\left(-\frac{1}{2}tr\left\{\tilde{\Sigma}^{-1}S_0\right\}\right) \times \frac{1}{\pi(\tilde{\Sigma}\,|y)} \\
&= (2\pi)^{-nT/2}\left|\tilde{\Sigma}\right|^{-T/2}\exp\left(-\frac{1}{2}tr\left[\tilde{\Sigma}^{-1}(Y - X\tilde{B})'(Y - X\tilde{B})\right]\right) \\
&\times \frac{|\Omega_0|^{-1/2}\exp\left(-\frac{1}{2}(\tilde{\beta} - \beta_0)'\Omega_0^{-1}\left(\tilde{\beta} - \beta_0\right)\right)}{\left|\bar{\Omega}\right|^{-1/2}} \\
&\times \frac{1}{2^{\alpha_0 n/2}\Gamma_n\left(\frac{\alpha_0}{2}\right)}|S_0|^{\alpha_0/2}\left|\tilde{\Sigma}\right|^{-(\alpha_0+n+1)/2}\exp\left(-\frac{1}{2}tr\left\{\tilde{\Sigma}^{-1}S_0\right\}\right) \times \frac{1}{\pi(\tilde{\Sigma}\,|y)}
\end{aligned}
$$

(using the fact that $\tilde{\beta} = \bar{\beta}$)

$$= \frac{(2\pi)^{-nT/2}}{2^{\alpha_0 n/2}\Gamma_n\left(\frac{\alpha_0}{2}\right)}|S_0|^{\alpha_0/2}\left|\tilde{\Sigma}\right|^{-(T+\alpha_0+n+1)/2}\exp\left(-\frac{1}{2}tr\left[\tilde{\Sigma}^{-1}\left\{(Y-X\tilde{B})'(Y-X\tilde{B})+S_0\right\}\right]\right)$$
$$\times\left(\frac{|\Omega_0|}{|\bar{\Omega}|}\right)^{-1/2}\exp\left(-\frac{1}{2}(\tilde{\beta}-\beta_0)'\Omega_0^{-1}\left(\tilde{\beta}-\beta_0\right)\right)\times\frac{1}{\pi(\tilde{\Sigma}\,|y)}$$

Then, use 3.9.35 to obtain the final approximation:

$$m(y) = \frac{(2\pi)^{-nT/2}}{2^{\alpha_0 n/2}\Gamma_n\left(\frac{\alpha_0}{2}\right)}|S_0|^{\alpha_0/2}\left|\tilde{\Sigma}\right|^{-(T+\alpha_0+n+1)/2}\left(\frac{|\Omega_0|}{|\bar{\Omega}|}\right)^{-1/2}$$
$$\times\exp\left(-\frac{1}{2}tr\left[\tilde{\Sigma}^{-1}\left\{(Y-X\tilde{B})'(Y-X\tilde{B})+S_0\right\}\right]\right)$$
$$\times\exp\left(-\frac{1}{2}(\tilde{\beta}-\beta_0)'\Omega_0^{-1}\left(\tilde{\beta}-\beta_0\right)\right)\times\frac{1}{(It-Bu)^{-1}\sum_{n=1}^{It-Bu}\pi(\tilde{\Sigma}\,|\beta^{(n)},y)}$$

which is 3.9.37.

Now develop to obtain 3.9.38. First, consider the term $\left(\frac{|\Omega_0|}{|\bar{\Omega}|}\right)^{-1/2}$:

$$
\begin{aligned}
\left(\frac{|\Omega_0|}{|\bar{\Omega}|}\right)^{-1/2} &= |\Omega_0|^{-1/2}|\bar{\Omega}|^{1/2} \\
&= |\Omega_0|^{-1/2}\left|\left[\Omega_0^{-1} + \tilde{\Sigma}^{-1} \otimes X^{,}X\right]^{-1}\right|^{1/2} \\
&= |\Omega_0|^{-1/2}\left|\Omega_0^{-1} + \tilde{\Sigma}^{-1} \otimes X^{,}X\right|^{-1/2} \\
&= |\Omega_0|^{-1/2}\left|\Omega_0^{-1} + (I_n \otimes X^{,})\left(\tilde{\Sigma}^{-1} \otimes X\right)\right|^{-1/2} \quad A.1.16 \\
&= |\Omega_0|^{-1/2}\left(\left|\Omega_0^{-1}\right|\left|I_q + \left(\tilde{\Sigma}^{-1} \otimes X\right)\Omega_0\left(I_n \otimes X^{,}\right)\right|\right)^{-1/2} \quad A.1.13 \\
&= |\Omega_0|^{-1/2}|\Omega_0|^{1/2}\left(\left|I_q + \left(\tilde{\Sigma}^{-1} \otimes X\right)\Omega_0\left(I_n \otimes X^{,}\right)\right|\right)^{-1/2} \quad A.1.15 \\
&= \left|I_q + \left(\tilde{\Sigma}^{-1} \otimes X\right)\Omega_0\left(I_n \otimes X^{,}\right)\right|^{-1/2} \\
&= \left|I_q + \left(\tilde{\Sigma}^{-1} \otimes X\right)F_\Omega F_\Omega^{,}\left(I_n \otimes X^{,}\right)\right|^{-1/2}
\end{aligned}
$$

where F_Ω denotes the square root matrix of Ω_0, that is, $F_\Omega F_\Omega^{,} = \Omega_0$

$$
= \left|I_q + F_\Omega^{,}\left(I_n \otimes X^{,}\right)\left(\tilde{\Sigma}^{-1} \otimes X\right)F_\Omega\right|^{-1/2} \quad A.1.16
$$

Substituting back:

$$
\begin{aligned}
m(y) = {} & \frac{(2\pi)^{-nT/2}}{2^{\alpha_0 n/2}\Gamma_n\left(\frac{\alpha_0}{2}\right)}|S_0|^{\alpha_0/2}\left|\tilde{\Sigma}\right|^{-(T+\alpha_0+n+1)/2}\left|I_q + F_\Omega^{,}\left(I_n \otimes X^{,}\right)\left(\tilde{\Sigma}^{-1} \otimes X\right)F_\Omega\right|^{-1/2} \\
& \times \exp\left(-\frac{1}{2}tr\left[\tilde{\Sigma}^{-1}\left\{(Y - X\tilde{B})^{,}(Y - X\tilde{B}) + S_0\right\}\right]\right) \times \exp\left(-\frac{1}{2}(\tilde{\beta} - \beta_0)^{,}\Omega_0^{-1}\left(\tilde{\beta} - \beta_0\right)\right) \\
& \times \frac{1}{(It - Bu)^{-1}\sum_{n=1}^{It-Bu}\pi(\tilde{\Sigma}\,|\beta^{(n)}, y)}
\end{aligned}
$$

Which is 3.9.38.

A.9 Derivation of the steady-state

It may be of interest to determine the long-run, or steady-state value of a model. Indeed, a good model should not produce long-run values grossly at odd with economic theory, or with the researcher belief. In this sense, calculating the steady-state of a model constitutes a way to verify the model relevance and adequacy. This is all the more important in a Bayesian context where the researcher can input personal information into the model in order to affect this long run value. And, of course, it is even more important in the case of a mean-adjusted model, which is especially designed to produce estimates based on the researcher belief about these long-run values.

Note first that the steady-state of a model will be meaningful only if the model is guaranteed to ultimately return to its long-run value. Otherwise, the model may just wander away of this value and never approach it again. This is related to the notion of weak or covariance stationarity, which implies that the expectation of the variables in the model should be independent of the period t. That is:

$$E(y_t) = \mu \quad \forall t \tag{A.9.1}$$

with μ a vector of constants. Therefore, before one turns to the computation of the steady-state, it is necessary to check that the model is covariance stationary. The usual issue of parameter uncertainty arises with Bayesian VAR models: talking about 'the' estimated model is not meaningful as the estimation process produces a posterior distribution for each coefficient, and not a single value. The difficulty is overcome by assuming that 'the' model to check for stationarity is that defined by the point estimates (typically the median) for each coefficient. This guarantees in no way, however, that *every* draw from the posterior distribution would produce a stationary model. It only checks whether a typical model satisfy stationarity. Assume hence that the point estimate $\tilde{\beta}$ is retained for β in 3.1.12, permitting to recover $\tilde{A}_1, \tilde{A}_2, ..., \tilde{A}_p$ in 3.1.2. Hamilton (1994) shows (see proposition 10.1, p 259) that the VAR model will be covariance stationary if all the eigenvalues of the matrix F are smaller than 1 in modulus, with F the companion matrix of the VAR defined as:

$$F = \begin{pmatrix} \tilde{A}_1 & \tilde{A}_2 & \cdots & \tilde{A}_{p-1} & \tilde{A}_p \\ I_n & 0 & \cdots & 0 & 0 \\ 0 & I_n & & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I_n & 0 \end{pmatrix} \tag{A.9.2}$$

Once covariance stationarity has been checked, it is possible to evaluate the steady-state of the model. The analysis starts with the classical VAR model 3.1.2:

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + ... + A_p y_{t-p} + C x_t + \varepsilon_t \tag{A.9.3}$$

The steady-state, or long-term value μ of the model at period t is simply the expectation $E(y_t)$ for model A.9.3 at this period. Hence, to derive the steady-state value of the model, take expectation on both sides of A.9.3:

$$\begin{aligned} E(y_t) =& E(A_1 y_{t-1} + A_2 y_{t-2} + ... + A_p y_{t-p} + C x_t + \varepsilon_t) \\ =& A_1 E(y_{t-1}) + A_2 E(y_{t-2}) + ... + A_p E(y_{t-p}) + C E(x_t) + E(\varepsilon_t) \end{aligned} \tag{A.9.4}$$

Under the assumption of covariance stationarity, $E(y_t) = \mu$. Note also that as x_t is an exogenous process with known values, $E(x_t) = x_t$. Finally, assumption 3.1.3 implies that $E(\varepsilon_t) = 0$. Then, A.9.4 rewrites as:

$$\mu = A_1 \mu + A_2 \mu + ... + A_p \mu + C x_t \tag{A.9.5}$$

Rearranging:

$$(I - A_1 - A_2 - ... - A_p)\mu = C x_t$$

This finally yields:

$$\mu = (I - A_1 - A_2 - ... - A_p)^{-1} C x_t \tag{A.9.6}$$

In the case of a mean-adjusted VAR, the procedure is a bit different. Start from the mean-adjusted model 5.6.3:

$$A(L)\left(y_t - F x_t\right) = \varepsilon_t \tag{A.9.7}$$

With this formulation, the simplest approach consists in adopting a Wold representation strategy. Under the assumption of covariance stationarity, the model A.9.7 can be inverted and reformulated as:

$$\begin{aligned} & A(L)\left(y_t - F x_t\right) = \varepsilon_t \\ \Leftrightarrow\ & \left(y_t - F x_t\right) = A(L)^{-1} \varepsilon_t \\ \Leftrightarrow\ & y_t - F x_t = \sum_{i=0}^{\infty} \Psi_i \varepsilon_{t-i} \end{aligned} \tag{A.9.8}$$

Take expectation on both sides, and use the facts that $E(x_t) = x_t$ and $E(\varepsilon_t) = 0, \forall t$ to obtain:

$$\begin{aligned} & E(y_t) - FE(x_t) = \sum_{i=0}^{\infty} \Psi_i E(\varepsilon_{t-i}) \\ & \Leftrightarrow E(y_t) - Fx_t = 0 \\ & \Leftrightarrow \mu = Fx_t \end{aligned} \tag{A.9.9}$$

A.10 Forecast evaluation

Forecast evaluation is a matter of central interest for model selection. To be useful, a model should be able to produce good forecasts. This means that the forecasts produce by the model should be as close as possible to the actual realised values taken by the data. Two sets of evaluation measures are developed in this section. The first set is fairly standard and is not proper to Bayesian techniques: it is the usual set of in-sample and out of sample fit measures. The second set is constituted by the family of statistics known as log predictive scores, and is more specific to Bayesian approaches.

Start thus with first set, and consider first the in-sample evaluation measures. A rough and preliminary measure of the goodness of fit of the model is given by the sum of squared residuals. Intuitively, the smaller this sum, the smaller the amplitude of the residuals, and thus the better the fit of the model to the data. For a VAR model with n variables, the sum of squared residuals for variable i, with $i = 1, 2, ..., n$, is defined as:

$$RSS_i = \varepsilon_i' \varepsilon_i \tag{A.10.1}$$

where $\varepsilon_i' = (\varepsilon_{i1} \quad \varepsilon_{i2} \quad \dots \quad \varepsilon_{iT})$ denotes the residual series for variable i over the whole sample. Using $\tilde{B}$ as a point estimate for B in 3.1.7, recovering the sum of squared residuals is straightforward. First obtain the predicted values $\tilde{Y}$ from the model, using 3.1.7:

$$\tilde{Y} = E(Y \,|X\,) = X\tilde{B} \tag{A.10.2}$$

Then, still from 3.1.7, the matrix of residuals obtains as:

$$\tilde{\mathcal{E}} = Y - X\tilde{B} = Y - \tilde{Y} \tag{A.10.3}$$

And it is possible to compute the full RSS matrix as:

$$RSS = \tilde{\mathcal{E}}' \tilde{\mathcal{E}} \tag{A.10.4}$$

Finally, obtain RSS_i from the i^{th} diagonal element of RSS.

A second common measure of fit, related to the sum of squared residuals, is the coefficient of determination, or R^2. This value measures the share of total variation in y that can be attributed to the model. The larger R^2, the more efficient the model to explain data variation. For equation i of the VAR model, R^2 is defined as:

$$R_i^2 = 1 - \frac{RSS_i}{TSS_i} \tag{A.10.5}$$

where TSS_i is the total sum of squares for equation i, defined as:

$$TSS_i = \sum_{j=1}^{T} (y_{i,j} - \bar{y}_i)^2 \tag{A.10.6}$$

with $\bar{y}_i$ the mean of y_i. A convenient way to compute TSS_i is to use a demeaning matrix $\bar{M} = I_T - (1/T)1_{T \times T}$, where $1_{T \times T}$ denotes a $T \times T$ matrix for which all entries are 1. $\bar{M}$ is idempotent and has the property that $\bar{M}x = x - \bar{x}$, for any vector x. Therefore, it is straightforward to define the full TSS matrix as:

$$TSS = Y' \bar{M} Y \tag{A.10.7}$$

One can then obtain a full matrix of R^2 as:

$$R^2 = I_n - \frac{RSS}{TSS} \tag{A.10.8}$$

Where the division of RSS with TSS has to be element-wise. R_i^2 can then be read from the i^{th}diagonal element of the R^2 matrix.

A limit of the coefficient of determination is that it increases mechanically as the number of variables integrated to the model increases. Hence, adding new variables into the model will result in higher R^2 values, which may fallaciously suggest an improvement in the model, even if it is actually poorer due to the additional loss of degrees of freedom during the estimation. Hence, a degrees of freedom corrected R^2, known as the adjusted R^2 is often estimated along with the traditional R^2. It is defined as:

$$\bar{R}_i^2 = 1 - \frac{T-1}{T-k}(1 - R_i^2) \tag{A.10.9}$$

It is possible to recycle the R^2 matrix in A.10.9 to obtain a full matrix of adjusted R^2:

$$\bar{R}^2 = I_n - \frac{T-1}{T-k}(I_n - R^2) \tag{A.10.10}$$

$\bar{R}_i^2$ can then be read from the i^{th}diagonal element of the $\bar{R}^2$ matrix.

Two evaluation criteria commonly used to discriminate among OLS models are the Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC). These two criteria are respectively defined as:

$$AIC = -2(L/T) + 2(q/T) \tag{A.10.11}$$

and

$$BIC = -2(L/T) + q\log(T)/T \tag{A.10.12}$$

L denotes the full system log-likelihood. Assuming as usual normal disturbances, it is defined as:

$$L = \frac{-Tn}{2}(1 + \log(2\pi)) - \frac{T}{2}\log\left|\frac{\tilde{\mathcal{E}}'\tilde{\mathcal{E}}}{T}\right| \tag{A.10.13}$$

Though the in-sample measures are informative, the most important measures evaluate the quality of out-of-sample predictions. Four standard out-of-sample measures are presented here: the root mean squared error, the mean absolute error, the mean absolute percentage error, and the Theil inequality coefficient. All these measures provide a way to assess the size of the difference between the predicted value and the value that actually occurred in the data set. Note that because these measures compare the forecast provided by the model with realised data values, computing them requires that at least one actual data point is available over the forecast period.

The first forecast evaluation measure is the root mean squared error or RMSE. Assume that forecasts are produced over h periods, and that in addition of the forecasts, the actual data values are known for these periods. Then the root mean squared error of this forecast for variable i in the VAR is defined as:

$$RMSE_i = \sqrt{(1/h)\sum_{i=1}^{h}(y_{T+i} - \tilde{y}_{T+i})^2} \tag{A.10.14}$$

where $\tilde{y}_{T+i}$ denotes as usual the predicted value of y_{T+i}. An alternative measure of forecast

efficiency is given by the mean absolute error:

$$MAE_i = (1/h)\sum_{i=1}^{h} |y_{T+i} - \hat{y}_{T+i}| \tag{A.10.15}$$

For these two measures, a lower value indicates a better fit. The squared terms indicate however that the RMSE place a greater penalty on large errors than does the MAE. A third measure is the mean absolute percentage error, defined for variable i as:

$$MAPE_i = (100/h)\sum_{i=1}^{h} \left| \frac{y_{T+i} - \hat{y}_{T+i}}{y_{T+i}} \right| \tag{A.10.16}$$

Unlike the RMSE and the MAE which measure the absolute size of the forecast error, the MAPE I a scaled measure, providing the size of the error relative to value of the variable. A final measure of forecast accuracy is provided by the Theil inequality coefficient or Theil U statistics:

$$U_i = \frac{\sqrt{(1/h)\sum_{i=1}^{h} (y_{T+i} - \hat{y}_{T+i})^2}}{\sqrt{(1/h)\sum_{i=1}^{h} (y_{T+i})^2} + \sqrt{(1/h)\sum_{i=1}^{h} (\hat{y}_{T+i})^2}} \tag{A.10.17}$$

This coefficient is always comprised between 0 and 1, a lower value indicating a better forecast. A value of 0 indicates a perfect fit, while a value of 1 says that the forecast is no better than a naive guess.

The second set of forecast measures, specific to Bayesian analysis, is the family of log predictive scores for forecasts. While the measures previously described compare the point estimates of the forecast with the actual values, in a pure frequentist style, the log scores compare the realised values with the whole posterior predictive density, in a more Bayesian fashion. The idea behind log scores is that a good model should produce a forecast distribution that makes it likely that the forecast occurs close to the realised values. In other words, the predictive distribution should be such that it takes a high density at the actual data value.

The procedure for computing log predictive scores that is now presented is due to Warne et al. (2013), as detailed by Mao (2010). Assume that some data set is used to estimate BVAR model M_1. The estimation sample runs until period T of the data set, with $T < T^f$, the final period of the data set. In other words, there remains some observed data values after the final sample period. Also, one wants to consider predictions for this model up to h periods after the sample end, with $T + h \leq T^f$. That is, predictions are formed over periods for which actual data is observed. Consider then any set

of predicted values for periods $T+1, T+2, ..., T+h$ for the BVAR model M_1. Recycling notations from Section 4.1, the vector of h-step ahead forecasts can be denoted as:

$$\tilde{y}_{T+1:T+h} = \begin{pmatrix} \tilde{y}_{T+1} \\ \tilde{y}_{T+2} \\ \vdots \\ \tilde{y}_{T+h} \end{pmatrix} \tag{A.10.18}$$

Following, it is possible to denote the conditional predictive density of these predicted values by $f(\tilde{y}_{T+1:T+h} \,|y_T^o\,, \beta, \Sigma, M_1)$, where y_T^o denotes data observed until period T, β and Σ denote respectively the VAR coefficient and residual covariance matrix drawn from the posterior distribution of model M_1, and M_1 denotes the candidate model. Under the assumption of normal disturbances, the conditional distribution of $\tilde{y}_{T+1:T+h}$ is multivariate normal:

$$\tilde{y}_{T+1:T+h} \,|y_T^o, \beta, \Sigma, M_1 \;\sim\; N(\underset{nh\times 1}{\mu}\;,\; \underset{nh\times nh}{\Upsilon}) \tag{A.10.19}$$

with

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_h \end{pmatrix} \; and \; \Upsilon = \begin{pmatrix} \Upsilon_{1,1} & \Upsilon_{1,2} & \cdots & \Upsilon_{1,h} \\ \Upsilon_{2,1} & \Upsilon_{2,2} & \cdots & \Upsilon_{2,h} \\ \vdots & \vdots & \ddots & \vdots \\ \Upsilon_{h,1} & \Upsilon_{h,2} & \cdots & \Upsilon_{h,h} \end{pmatrix} \tag{A.10.20}$$

μ is the mean vector for the h-step ahead predictions, in which each μ_i element represents the $n \times 1$ mean for prediction at period $T+i$ $(i = 1, 2, ..., h)$. On the other hand, Υ is the variance-covariance matrix for these predictions, and each $\Upsilon_{i,j}$ is the $n \times n$ covariance matrix between predictions at period $T+i$ and $T+j$. The issue resides in computing μ and Υ.

For μ, the i^{th}-step ahead expectation μ_i is given by:

$$\mu_i = \begin{cases} E(y_{T+i} \,|y_T^o, \beta, \Sigma, M_1\,) \text{ if } i > 0 \\ y_{T+i} \text{ if } i < 0 \end{cases} \tag{A.10.21}$$

μ_i is straightforward to obtain from the usual chain rule of forecasting. For instance given a VAR model in the form of 3.1.2, one has:

$$
\begin{aligned}
\mu_1 &= E(y_{T+1}\,|y_T^o, \beta, \Sigma, M_1\,) \\
&= A_1 E(y_T\,|y_T^o, \beta, \Sigma, M_1\,) + A_2 E(y_{T-1}\,|y_T^o, \beta, \Sigma, M_1\,) + ... + A_p E(y_{T+1-p}\,|y_T^o, \beta, \Sigma, M_1\,) \\
&+ CE(x_{T+1}\,|y_T^o, \beta, \Sigma, M_1\,) + E(\varepsilon_{T+1}\,|y_T^o, \beta, \Sigma, M_1\,) \\
&= A_1\mu_0 + A_2\mu_{-1} + ... + A_p\mu_{1-p} + CE(x_{T+1})
\end{aligned}
$$

Following:

$$
\begin{aligned}
\mu_2 &= E(y_{T+2}\,|y_T^o, \beta, \Sigma, M_1\,) \\
&= A_1 E(y_{T+1}\,|y_T^o, \beta, \Sigma, M_1\,) + A_2 E(y_T\,|y_T^o, \beta, \Sigma, M_1\,) + ... + A_p E(y_{T+2-p}\,|y_T^o, \beta, \Sigma, M_1\,) \\
&+ CE(x_{T+2}\,|y_T^o, \beta, \Sigma, M_1\,) + E(\varepsilon_{T+2}\,|y_T^o, \beta, \Sigma, M_1\,) \\
&= A_1\mu_1 + A_2\mu_0 + ... + A_p\mu_{2-p} + CE(x_{T+2})
\end{aligned}
$$

In general, one can use recursive substitution and obtain:

$$
\mu_i = \sum_{k=1}^{p} A_k\mu_{i-k} + CE(x_{T+i}) \tag{A.10.22}
$$

Now, for the covariance matrix Υ, define:

$$
\begin{aligned}
\Upsilon_{i,j} &= \mathrm{cov}(y_{T+i}, y_{T+j}\,|y_T^o, \beta, \Sigma, M_1\,) \\
&= E\left[(y_{T+i} - \mu_{y,i})(y_{T+j} - \mu_{y,j})'\,|y_T^o, \beta, \Sigma, M_1\right]
\end{aligned} \tag{A.10.23}
$$

Notice that from 3.1.2 and A.10.21, one can deduce:

$$
\begin{aligned}
y_{T+i} - \mu_i &= \sum_{k=1}^{p} A_k y_{T+i-k} + Cx_{T+i} + \varepsilon_{T+i} - \sum_{k=1}^{p} A_k\mu_{i-k} - CE(x_{T+i}) \\
&= \sum_{k=1}^{\min(i-1,p)} A_k(y_{T+i-k} - \mu_{i-k}) + C\left[x_{T+i} - E(x_{T+i})\right] + \varepsilon_{T+i}
\end{aligned} \tag{A.10.24}
$$

The $\min(i-1,p)$justifies by the fact that for $k \geq i$, the period becomes $t < T$, implying from A.10.21 that $(y_{T+i-k} - \mu_{i-k}) = (\mu_{i-k} - \mu_{i-k}) = 0$.

Substituting back in A.10.23, one obtains:

$$
\begin{aligned}
\Upsilon_{i,j} =& E\left[\left(\sum_{k=1}^{\min(i-1,p)} A_k(y_{T+i-k}-\mu_{i-k})+C\left[x_{T+i}-E(x_{T+i})\right]+\varepsilon_{T+i}\right)(y_{T+j}-\mu_j)' \,|y_T^o,\beta,\Sigma,M_1\right] \\
=& \sum_{k=1}^{\min(i-1,p)} A_k E\left[(y_{T+i-k}-\mu_{i-k})(y_{T+j}-\mu_j)' \,|y_T^o,\beta,\Sigma,M_1\right] \\
&+ C\left[E(x_{T+i}\,|y_T)-E(E(x_{T+i})\,|y_T^o,\beta,\Sigma,A)\right](y_{T+j}-\mu_j)' + E\left[\varepsilon_{T+i}(y_{T+j}-\mu_j)' \,|y_T^o,\beta,\Sigma,M_1\right] \\
=& \sum_{k=1}^{\min(i-1,p)} A_k E\left[(y_{T+i-k}-\mu_{i-k})(y_{T+j}-\mu_j)' \,|y_T^o,\beta,\Sigma,M_1\right] \\
&+ C\left[E(x_{T+i})-E(x_{T+i})\right](y_{T+j}-\mu_j)' + E\left[\varepsilon_{T+i}(y_{T+j}-\mu_j)' \,|y_T^o,\beta,\Sigma,M_1\right] \\
=& \sum_{k=1}^{\min(i-1,p)} A_k E\left[(y_{T+i-k}-\mu_{i-k})(y_{T+j}-\mu_j)' \,|y_T^o,\beta,\Sigma,M_1\right] + E\left[\varepsilon_{T+i}(y_{T+j}-\mu_j)' \,|y_T^o,\beta,\Sigma,M_1\right]
\end{aligned}
\tag{A.10.25}
$$

Without loss of generality, assume $i > j$ (the symmetry of Υ implies that $\Upsilon_{i,j} = \Upsilon_{j,i}'$, so that the case $i < j$ can simply be treated as the transpose of the case $i > j$). Then from A.10.25:

$$
\begin{aligned}
\Upsilon_{i,j} =& \sum_{k=1}^{\min(i-1,p)} A_k E\left[(y_{T+i-k}-\mu_{i-k})(y_{T+j}-\mu_j)' \,|y_T^o,\beta,\Sigma,M_1\right] + E\left[\varepsilon_{T+i}(y_{T+j}-\mu_j)' \,|y_T^o,\beta,\Sigma,M_1\right] \\
=& \sum_{k=1}^{\min(i-1,p)} A_k \Upsilon_{i-k,j}
\end{aligned}
\tag{A.10.26}
$$

where use has been made of the fact that $E\left[\varepsilon_{T+i}(y_{T+j}-\mu_j)' \,|y_T^o,\beta,\Sigma,M_1\right] = 0$ since $i > j$. In the case $i = j$, using again A.10.24 to substitute for $y_{T+j}-\mu_j$ in A.10.26, one obtains:

$$
\begin{aligned}
\Upsilon_{i,i} =& \sum_{k=1}^{\min(i-1,p)} A_k E\left[(y_{T+i-k}-\mu_{i-k})(y_{T+i}-\mu_i)' \,|y_T^o,\beta,\Sigma,M_1\right] \\
&+ E\left[\varepsilon_{T+i}\left(\sum_{k=1}^{\min(i-1,p)} A_k(y_{T+i-k}-\mu_{i-k})+C\left[x_{T+i}-E(x_{T+i})\right]+\varepsilon_{T+i}\right)' \,|y_T^o,\beta,\Sigma,M_1\right] \\
=& \sum_{k=1}^{\min(i-1,p)} A_k \Upsilon_{i-k,i} + \Sigma
\end{aligned}
\tag{A.10.27}
$$

where Σ denotes as usual the residual covariance matrix, and where use has been made of the

fact that $E\left[\varepsilon_{T+i}A_k(y_{T+i-k}-\mu_{i-k})'\,|y_T^o,\beta,\Sigma,M_1\right]=0$ since $T+i>T+i-k$.

A.10.22, A.10.26 and A.10.27 allow to fully identify μ and Υ, and, following, to compute $f(\tilde{y}_{T+1:T+h}\,|y_T^o\,,\beta,\Sigma,M_1)$ from A.10.19. Furthermore, it is possible to use property A.2.2.6 of the multivariate normal distribution to obtain:

$$R\tilde{y}_{T+1:T+h}\,|y_T^o,\beta,\Sigma,M_1 \;\sim\; N(\,\mathrm{R}\mu\;,\;\mathrm{R}\Upsilon R') \tag{A.10.28}$$

where R is any matrix comfortable with $\tilde{y}_{T+1:T+h}$. By selecting carefully R, one can obtain the properties of a subset only of $\tilde{y}_{T+1:T+h}$. For instance, considering the case $n=2$ and $h=3$ (two variables and three forecast periods):

- defining $R=I_{nh}$ selects $\tilde{y}_{T+1:T+h}$ as a whole
- defining $R=\begin{pmatrix}0&0&0&0&1&0\\0&0&0&0&0&1\end{pmatrix}$ selects only $\tilde{y}_{T+h}$
- defining $R=\begin{pmatrix}1&0&0&0&0&0\\0&0&0&0&0&0\\0&0&1&0&0&0\\0&0&0&0&0&0\\0&0&0&0&1&0\\0&0&0&0&0&0\end{pmatrix}$ selects only the first variable over all the periods.

And so on.

Note however that what has been so far defined in A.10.19 is only a conditional predictive density, that is, the density obtained for given draws of β and Σ from the posterior distribution of model A. However, what is required for the log predictive score is the *unconditional* predictive density. Fortunately, it is straightforward to derive the latter by relying on almost sure convergence properties. Precisely, one has:

$$\frac{1}{N}\sum_{n=1}^{N} f(Ry_{T+1:T+h}^o\,|y_T^o\,,\beta^{(n)},\Sigma^{(n)},M_1)\overset{a.s}{\to} f(Ry_{T+1:T+h}^o\,|y_T^o\,,M_1) \tag{A.10.29}$$

where $\beta^{(n)}$ and $\Sigma^{(n)}$ denote draw n from the posterior distribution of model A. Of course, in practice, one follows the usual strategy consisting in recycling the Gibbs sampler draws to obtain the sequences of $\beta^{(n)}$ and $\Sigma^{(n)}$ values. Also, it is important to note that the predictive density is evaluated from the observed values $y_{T+1:T+h}^o$ in order to evaluate the prediction performance of model M_1.

With this predictive density, it is possible to estimate the log predictive score of model A. The general definition of the log predictive score for model A is [10]:

$$S(y_T^o, M_1) = \sum_{t=T}^{T^f-h} \log f(Ry_{t+1:t+h}^o \,|y_t^o\,, M_1) \tag{A.10.30}$$

It is convenient to consider only the case $T + h = T^f$, for which A.10.30 becomes:

$$S(y_T^o, M_1) = \log f(Ry_{T+1:T+h}^o \,|y_t^o\,, M_1) \tag{A.10.31}$$

Two base forecasting scenarios are considered, each time for variable i of the model, $i = 1, 2, ..., n$. In the first scenario, R is sequentially defined as:

$$\underset{1\times nh}{R} = \begin{pmatrix} 0_{i,n} & 0_n & 0_n & \cdots & 0_n \end{pmatrix}, \begin{pmatrix} 0_n & 0_{i,n} & 0_n & \cdots & 0_n \end{pmatrix}, ..., \begin{pmatrix} 0_n & 0_n & 0_n & \cdots & 0_{i,n} \end{pmatrix} \tag{A.10.32}$$

where $0_{i,n}$ denotes a $1 \times n$ matrix of zeros, save for a unique 1, located at the i^{th}entry, and 0_n denotes a $1 \times n$ matrix of zeros. In this case, what is evaluated is the performance of the forecasts produced by model A for variable i at respective periods $T+1, T+2, ..., T+h$. In the second scenario, R is sequentially defined as:

$$R = \begin{pmatrix} 0_{i,n} & 0_n & 0_n & \cdots & 0_n \end{pmatrix}, \begin{pmatrix} 0_{i,n} & 0_n & 0_n & \cdots & 0_n \\ 0_n & 0_{i,n} & 0_n & \cdots & 0_n \end{pmatrix}, ..., \begin{pmatrix} 0_{i,n} & 0_n & 0_n & \cdots & 0_n \\ 0_n & 0_{i,n} & 0_n & \cdots & 0_n \\ \vdots & & & \ddots & \vdots \\ 0_n & \cdots & 0_n & 0_n & 0_{i,n} \end{pmatrix} \tag{A.10.33}$$

Where R is sequentially of dimension $1 \times nh, 2 \times nh, ..., h \times nh$. In this case, what is evaluated is the overall performance of the forecast produced by model A for variable i from period $T + 1$ up to periods $T + 2$, $T + 3$, ..., $T + h$.

To summarize, it is possible to propose the following procedure to compute the log predictive scores:

Algorithm A.10.1 (log predictive score, all priors):

1. Store observed values for post-sample data y_{T+1}^o,y_{T+2}^o,...,y_{T+h}^o.

2. Initiate the Gibbs sampler phase. At iteration n, draw $\beta^{(n)}$ and $\Sigma^{(n)}$.

3. At iteration n, obtain μ from A.10.22.

[10] There are nearly as many different definitions of log predictive scores than papers using them as an evaluation criterion. The definition adopted here is that of Mao (2010), which has the virtue of being fairly general.

4. At iteration n, obtain Υ from A.10.26 and A.10.27.

5. At iteration n, obtain $f(Ry^o_{T+1:T+h}\,|y^o_T\,,\beta^{(n)},\Sigma^{(n)},M_1)$ from A.10.19, for every value of R in the sequences A.10.32 and A.10.33.

6. Once It iterations are realised, compute $f(Ry^o_{T+1:T+h}\,|y^o_T\,,M_1)$ from A.10.29, once again for each value of R.

7. Compute the log predictive scores from A.10.31, for each value of R.

An alternative to the log predictive score is the continuous ranked probability score, first introduced by Matheson and Winkler (1976), and which has recently regained interest Gneiting and Raftery (2007). Similarly to the log predictive score, it evaluates the quality of the forecast produced by the model by assessing the adequacy between the posterior predictive distribution and the actual, realized value. The idea, once again, is that a good posterior predictive distribution should be at the same time characterized by an accurate location (close to the realized value), and a sharp density (predictive values should be tightly concentrated around the realized value).

Consider the cumulative distribution function F corresponding to the marginal predictive density f for the forecast at period $T+h$, along with the realized value y^o_{T+h} for this period. Then, the continuous ranked predictive score (CRPS) is then defined as:

$$CRPS(F,y^o_{T+h}) = \int_{-\infty}^{\infty} \left(F(x) - 1(x > y^o_{T+h})\right)^2 dx \tag{A.10.34}$$

where 1(.) denotes the indicator function, taking a value of 1 if the condition is verified, and zero otherwise. The CRPS value in A.10.34 can be conceived as a penalty function sanctioning the overall distance between the distribution points and the realized value. In the case of a perfect predictive distribution (a mass point of density 1 at $x = y^o_{T+h}$, so that $F(x) = 0$ for $x < y^o_{T+h}$, and $F(x) = 1$ for $x \geq y^o_{T+h}$), the value of $CRPS(F,y^o_{T+h})$ is 0. In any other case, there will be a positive penalty stemming from the possible deviations of the distribution value from the observed value, with greater penalty applied on values far away from the realized value. Hence, the larger the value of the $CRPS$, the poorer the performance of the predictive distribution f for the forecast at period $T+h$.

In practical applications, A.10.34 is not applicable because the analytical form of the cumulative distribution function F is not known. However, Gneiting and Raftery (2007) show that the $CRPS$ A.10.34 can be evaluated in closed form as:

$$CRPS(F,y^o_{T+h}) = E\left|x - y^o_{T+h}\right| - \frac{1}{2}E\left|x-y\right| \tag{A.10.35}$$

where expectation is to be taken with respect to the distribution function F, and x and y are independent random draws from the density f. As usual, the strategy then consists into approxi-

mating the expectation terms in A.10.35 by recycling the Gibbs sampler draws from f. This yields the following equivalent formula:

$$CRPS(F, y^o_{T+h}) = \frac{1}{(It - Bu)} \sum_{i=1}^{(It-Bu)} \left| \tilde{y}^{(i)}_{T+h} - y^o_{T+h} \right| - \frac{1}{2(It - Bu)^2} \sum_{i=1}^{(It-Bu)} \sum_{j=1}^{(It-Bu)} \left| \tilde{y}^{(i)}_{T+h} - \tilde{y}^{(j)}_{T+h} \right| \tag{A.10.36}$$

A.11 Derivation of confidence intervals for a standard OLS VAR model

So far, the analysis developed in this guide focused on Bayesian estimates. In a Bayesian framework, deriving confidence intervals (or more properly, credibility intervals) is extremely straightforward: suffice is to consider quantiles of the posterior distribution, be it analytical or empirical as a result of a Gibbs sampler process. Yet, such confidence intervals can also be provided for a standard OLS VAR model, though the procedure is typically more complex. This appendix presents the results used to derive confidence intervals for the VAR coefficients, the forecasts, and the impulse response functions.

The confidence intervals for the VAR coefficients are derived by using proposition 11.1 (p 298-299) in Hamilton (1994). This proposition states the following: consider the standard VAR model 3.1.2, and its vectorised reformulation 3.1.11. The two parameters of interest are the residual variance-covariance matrix Σ and the vectorised VAR coefficients β, and their OLS counterparts $\hat{\Sigma}$ and $\hat{\beta}$ are respectively given by 3.1.10 and 3.1.15. Then, under fairly standard conditions, the following holds:

1. $(1/T)(X^{\cdot}X) \xrightarrow{P} \Theta, \quad$ where $\Theta = E(X^{\cdot}X)$
2. $\hat{\beta} \xrightarrow{P} \beta$
3. $\hat{\Sigma} \xrightarrow{P} \Sigma$
4. $\sqrt{T}\left(\hat{\beta} - \beta\right) \xrightarrow{L} N\left(0, \Sigma \otimes \Theta^{-1}\right)$

1.) states that $(1/T)(X^{\cdot}X)$ is a consistent estimate for Θ. That is, for large enough T, $(1/T)(X^{\cdot}X) \approx \Theta$, which implies that $\Theta^{-1} \approx T(X^{\cdot}X)^{-1}$. Then, 2.) and 3.) tell us that $\hat{\beta}$ and $\hat{\Sigma}$ are consistent estimators for β and Σ. Finally, 4.) implies that $\hat{\beta}$ converges in distribution to a normal law: $\hat{\beta} \xrightarrow{L} \mathcal{N}\left(\beta, T^{-1}\left(\Sigma \otimes \Theta^{-1}\right)\right)$. Combining 1.) and 4.), it is possible to conclude that $\hat{\beta}$ approximately follows a normal distribution characterised by $\hat{\beta} \sim \mathcal{N}\left(\beta, \Sigma \otimes \left(XX\right)^{-1}\right)$. Finally, combining with 3.) to obtain a consistent estimate of Σ, one obtains the approximate distribution:

$$\hat{\beta} \sim \ N\left(\beta, \hat{\Sigma} \otimes (XX)^{-1}\right) \tag{A.11.1}$$

From this, and the property A.2.2.4 of the multivariate normal distribution, one may conclude that the i^{th} element of the vector $\hat{\beta}$ follows a normal distribution:

$$\hat{\beta}_i \sim \ \mathcal{N}\left(\beta_i, \hat{\sigma}_i^2 \otimes (XX)^{-1}\right) \tag{A.11.2}$$

where $\hat{\sigma}_i^2$ is the i^{th} diagonal element of $\hat{\Sigma}$. From this, it is straightforward to derive the standard deviation and confidence intervals of each coefficient.

For impulse response functions, the easiest approach is probably to rely on Monte Carlo simulation methods, as described in Hamilton (1994), p 337. Because from A.10.1, the distribution of $\hat{\beta}$ is known, it is possible to randomly generate vectors $\hat{\beta}^{(1)}, \hat{\beta}^{(2)}, ..., \hat{\beta}^{(N)}$, where N is some large number. Then, for each such vector, compute the series of impulse functions, using the methodology developed in section 4.2. For a confidence level of α , simply trim the $\alpha/2$ percent smallest and largest values, to obtain an empirical $\alpha\%$ confidence interval.

Finally, to derive the confidence interval of forecasts, one may use the Gaussian properties of y_t . This implies (see Luetkepohl (1993), equation 3.5.15 p 89) that an approximate $(1-\alpha)$ % confidence interval for an h-periods ahead forecast for variable $y_{i,t}$ obtains as:

$$\tilde{y}_{i,t+h} \pm z_{(\alpha/2)}\sigma_{i,h} \tag{A.11.3}$$

where $\tilde{y}_{i,t+h}$ denotes the predicted value for y_i at period $t+h$, $z_{(\alpha)}$ is the α^{th} quantile of the standard normal distribution, and $\sigma_{i,h}$ is the square root of the i^{th} diagonal element of $\tilde{\Sigma}_h$, the forecast error covariance matrix. The latter is defined as:

$$\tilde{\Sigma}_h = \sum_{i=0}^{h-1} \Psi_i \hat{\Sigma} \Psi_i^{'} = \tilde{\Sigma}_{h-1} + \Psi_{h-1} \hat{\Sigma} \Psi_{h-1}^{'} \tag{A.11.4}$$

and Ψ_i denotes the impulse response function matrix for period i (see Luetkepohl (1993), equation 2.2.11 p 32).

A.12 Examples on conditional forecasts

1. Case of conditional forecasts when there are more variables in a block than shocks generating them.

   Consider the case of a 3-variable VAR model for which forecasts are produced for $T+1$. There

is one condition on variable 1 so that $y_{1,T+1} = \bar{y}_1$, and one condition on variable 2 so that $y_{2,T+1} = \bar{y}_2$. Both conditions are generated by shock 1. There is thus one block made of two variables, generated only by one shock. The system 5.3.12 is given by:

$$\begin{pmatrix} \tilde{\phi}_{0,11} & \tilde{\phi}_{0,12} & \tilde{\phi}_{0,13} \\ \tilde{\phi}_{0,21} & \tilde{\phi}_{0,22} & \tilde{\phi}_{0,23} \end{pmatrix} \begin{pmatrix} \eta_{1,T+1} \\ \eta_{2,T+1} \\ \eta_{3,T+1} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 - \tilde{y}_{1,T+1} \\ \bar{y}_2 - \tilde{y}_{2,T+1} \end{pmatrix} \tag{A.12.1}$$

- Draw first the non-constructive shocks $\eta_{2,T+1}$ and $\eta_{3,T+1}$ from their distribution, and transfer their impacts on the right-hand side:

$$\begin{pmatrix} \tilde{\phi}_{0,11} & 0 & 0 \\ \tilde{\phi}_{0,21} & 0 & 0 \end{pmatrix} \begin{pmatrix} \eta_{1,T+1} \\ \eta_{2,T+1} \\ \eta_{3,T+1} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 - \tilde{y}_{1,T+1} - \tilde{\phi}_{0,12}\eta_{2,T+1} - \tilde{\phi}_{0,13}\eta_{3,T+1} \\ \bar{y}_2 - \tilde{y}_{2,T+1} - \tilde{\phi}_{0,22}\eta_{2,T+1} - \tilde{\phi}_{0,23}\eta_{3,T+1} \end{pmatrix} \tag{A.12.2}$$

  For clarity, note that this system can be equivalently rewritten as:

$$\begin{pmatrix} \tilde{\phi}_{0,11} \\ \tilde{\phi}_{0,21} \end{pmatrix} (\eta_{1,T+1}) = \begin{pmatrix} \bar{y}_1 - \tilde{y}_{1,T+1} - \tilde{\phi}_{0,12}\eta_{2,T+1} - \tilde{\phi}_{0,13}\eta_{3,T+1} \\ \bar{y}_2 - \tilde{y}_{2,T+1} - \tilde{\phi}_{0,22}\eta_{2,T+1} - \tilde{\phi}_{0,23}\eta_{3,T+1} \end{pmatrix} \tag{A.12.3}$$

  This is a system of two equations in only one unknown: $\eta_{1,T+1}$. The system is overdetermined and has no solution.

  Conclusion: conditions cannot hold for blocks where there are more variables than shocks generating the conditions. The reason is that this generates over-determined systems with no solutions.

2. Case of conditional forecasts when shocks are shared among blocks

   Consider the case of a 3-variable VAR model for which forecasts are produced for $T+1$. There is one condition on variable 1 so that $y_{1,T+1} = \bar{y}_1$, and one condition on variable 2 so that $y_{2,T+1} = \bar{y}_2$. Both conditions are generated by shock 1 and 2. There is thus one block made of two variables, generated by two shocks. There is also one condition on variable 3 so that $y_{3,T+1} = \bar{y}_3$, and this condition is generated by shock 2. This constitutes block 2.

   The system 5.3.12 is given by:

$$\begin{pmatrix} \tilde{\phi}_{0,11} & \tilde{\phi}_{0,12} & \tilde{\phi}_{0,13} \\ \tilde{\phi}_{0,21} & \tilde{\phi}_{0,22} & \tilde{\phi}_{0,23} \\ \tilde{\phi}_{0,31} & \tilde{\phi}_{0,32} & \tilde{\phi}_{0,33} \end{pmatrix} \begin{pmatrix} \eta_{1,T+1} \\ \eta_{2,T+1} \\ \eta_{3,T+1} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 - \tilde{y}_{1,T+1} \\ \bar{y}_2 - \tilde{y}_{2,T+1} \\ \bar{y}_3 - \tilde{y}_{3,T+1} \end{pmatrix} \tag{A.12.4}$$

- Draw first the non-constructive shock $\eta_{3,T+1}$ from its distribution, and transfer its impact on the right-hand side:

$$\begin{pmatrix} \tilde{\phi}_{0,11} & \tilde{\phi}_{0,12} & 0 \\ \tilde{\phi}_{0,21} & \tilde{\phi}_{0,22} & 0 \\ \tilde{\phi}_{0,31} & \tilde{\phi}_{0,32} & 0 \end{pmatrix} \begin{pmatrix} \eta_{1,T+1} \\ \eta_{2,T+1} \\ \eta_{3,T+1} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 - \tilde{y}_{1,T+1} - \tilde{\phi}_{0,13}\eta_{3,T+1} \\ \bar{y}_2 - \tilde{y}_{2,T+1} - \tilde{\phi}_{0,23}\eta_{3,T+1} \\ \bar{y}_3 - \tilde{y}_{3,T+1} - \tilde{\phi}_{0,33}\eta_{3,T+1} \end{pmatrix} \tag{A.12.5}$$

- Consider block 1: draw first the constructive shock $\eta_{1,T+1}$ and $\eta_{2,T+1}$ from the Waggoner-Zha distribution, and transfer their impacts on the right-hand side:

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \eta_{1,T+1} \\ \eta_{2,T+1} \\ \eta_{3,T+1} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 - \tilde{y}_{1,T+1} - \tilde{\phi}_{0,13}\eta_{3,T+1} - \tilde{\phi}_{0,11}\eta_{1,T+1} - \tilde{\phi}_{0,12}\eta_{2,T+1} \\ \bar{y}_2 - \tilde{y}_{2,T+1} - \tilde{\phi}_{0,23}\eta_{3,T+1} - \tilde{\phi}_{0,21}\eta_{1,T+1} - \tilde{\phi}_{0,22}\eta_{2,T+1} \\ \bar{y}_3 - \tilde{y}_{3,T+1} - \tilde{\phi}_{0,33}\eta_{3,T+1} - \tilde{\phi}_{0,31}\eta_{1,T+1} - \tilde{\phi}_{0,32}\eta_{2,T+1} \end{pmatrix} \tag{A.12.6}$$

All the shocks are now determined, but the draws have been realised to satisfy the conditions in block 1 only. There is thus no reason that the condition in block 2 is also satisfied. This is because shocks are actually not shared: shock $\eta_{2,T+1}$ has been used by block 1, and is not available anymore to determine block 2. Block 2 could be only identified if it was generated by at least one additional shock. But then, it would be more accurate to say that it is generated only by this other shock.

Conclusion: shocks cannot be shared among blocks. Different blocks must be generated by different shocks.

A.13 Derivations for the pooled estimator

Start by the likelihood function. Given (6.3.12) and (6.3.13), it is given by:

$$f(y\,|\bar{\Sigma}) = (2\pi)^{-NnT/2}\left|\bar{\Sigma}\right|^{-1/2}\exp\left(-\tfrac{1}{2}(y-\bar{X}\beta)'\bar{\Sigma}^{-1}(y-\bar{X}\beta)\right)$$

or:

$$f(y\,|\bar{\Sigma}) \propto \left|\bar{\Sigma}\right|^{-1/2} \exp\left(-\tfrac{1}{2}(y-\bar{X}\beta)'\bar{\Sigma}^{-1}(y-\bar{X}\beta)\right)$$

Using (6.3.11) and (6.3.13), the kernel reformulates as:

$$f(y\,|\Sigma_c) \propto |\Sigma_c \otimes I_{NT}|^{-1/2} \exp\left(-\frac{1}{2}(y-(I_n\otimes X)\,\beta)'(\Sigma_c\otimes I_{NT})^{-1}\,(y-(I_n\otimes X)\,\beta)\right) \qquad \text{(A.13.1)}$$

Consider only the part within the curly brackets of (A.13.1) and develop:

$$\begin{aligned}
&(y-(I_n\otimes X)\beta)'(\Sigma_c\otimes I_{NT})^{-1}\,(y-(I_n\otimes X)\beta)\\
&=(y'-\beta'(I_n\otimes X)')\,(\Sigma_c^{-1}\otimes I_{NT})\,(y-(I_n\otimes X)\beta) \qquad \text{A.1.2}\\
&=y'\,(\Sigma_c^{-1}\otimes I_{NT})\,y-2\beta'(I_n\otimes X)'\,(\Sigma_c^{-1}\otimes I_{NT})\,y+\beta'(I_n\otimes X)'\,(\Sigma_c^{-1}\otimes I_{NT})\,(I_n\otimes X)\beta\\
&=y'\,(\Sigma_c^{-1}\otimes I_{NT})\,y-2\beta'(\Sigma_c^{-1}\otimes X')y+\beta'(\Sigma_c^{-1}\otimes X'X)\beta \qquad \text{A.1.1,A.1.3}
\end{aligned}$$

Completing the squares:

$$\begin{aligned}
&=y'\left(\Sigma_c^{-1}\otimes I_{NT}\right)y-2\hat{\beta}'\left(\Sigma_c^{-1}\otimes X'\right)y+2\hat{\beta}'\left(\Sigma_c^{-1}\otimes X'\right)y\\
&\quad -2\beta'\left(\Sigma_c^{-1}\otimes X'\right)y+\beta'\left(\Sigma_c^{-1}\otimes X'X\right)\beta\\
&=y'\left(\Sigma_c^{-1}\otimes I_{NT}\right)y-2\hat{\beta}'\left(\Sigma_c^{-1}\otimes X'\right)y+2\hat{\beta}'\left(\Sigma_c^{-1}\otimes X'X\right)\left(\Sigma_c^{-1}\otimes X'X\right)^{-1}\left(\Sigma_c^{-1}\otimes X'\right)y\\
&\quad -2\beta'\left(\Sigma_c^{-1}\otimes X'X\right)\left(\Sigma_c^{-1}\otimes X'X\right)^{-1}\left(\Sigma_c^{-1}\otimes X'\right)y+\beta'\left(\Sigma_c^{-1}\otimes X'X\right)\beta \qquad \text{(A.13.2)}
\end{aligned}$$

Define $\hat{\beta}$, the OLS estimate of β, as:

$$\hat{\beta}=\left(\Sigma_c^{-1}\otimes(X'X)\right)^{-1}\left(\Sigma_c^{-1}\otimes X'\right)y \qquad \text{(A.13.3)}$$

Indeed, one has:

$$\begin{aligned}
\hat{\beta} &= vec(\hat{B}) \\
&= vec(\hat{B} I_n) \\
&= vec\left((X^{,}X)^{-1}X^{,}yI_n\right) \\
&= \left(I_n \otimes (X^{,}X)^{-1}X^{,}\right) vec(Y) \qquad A.1.5 \\
&= \left(I_n \otimes (X^{,}X)^{-1}X^{,}\right) y \\
&= \left(\Sigma_c \otimes (X^{,}X)^{-1}\right)\left(\Sigma_c^{-1} \otimes X^{,}\right) y \qquad A.1.3 \\
&= \left(\Sigma_c^{-1} \otimes (X^{,}X)\right)^{-1}\left(\Sigma_c^{-1} \otimes X^{,}\right) y \qquad A.1.2
\end{aligned}$$

This is indeed (A.13.3). Then, (A.13.2) rewrites:

$$\begin{aligned}
&= y^{,}\left(\Sigma_c^{-1} \otimes I_{NT}\right) y - 2\hat{\beta}^{,}\left(\Sigma_c^{-1} \otimes X^{,}\right) y + 2\hat{\beta}^{,}\left(\Sigma_c^{-1} \otimes X^{,}X\right)\hat{\beta} \\
&\quad -2\beta^{,}\left(\Sigma_c^{-1} \otimes X^{,}X\right)\hat{\beta} + \beta^{,}\left(\Sigma_c^{-1} \otimes X^{,}X\right)\beta \\
&= y^{,}\left(\Sigma_c^{-1} \otimes I_{NT}\right) y - 2\hat{\beta}^{,}\left(\Sigma_c^{-1} \otimes X^{,}\right) y + \hat{\beta}^{,}\left(\Sigma_c^{-1} \otimes X^{,}X\right)\hat{\beta} \\
&\quad +\hat{\beta}^{,}\left(\Sigma_c^{-1} \otimes X^{,}X\right)\hat{\beta} - 2\beta^{,}\left(\Sigma_c^{-1} \otimes X^{,}X\right)\hat{\beta} + \beta^{,}\left(\Sigma_c^{-1} \otimes X^{,}X\right)\beta \\
&= y^{,}\left(\Sigma_c^{-1} \otimes I_{NT}\right) y - 2\hat{\beta}^{,}\left(\Sigma_c^{-1} \otimes X^{,}\right) y + \hat{\beta}^{,}\left(\Sigma_c^{-1} \otimes X^{,}X\right)\hat{\beta} \\
&\quad +(\beta - \hat{\beta})^{,}\left(\Sigma_c^{-1} \otimes X^{,}X\right)(\beta - \hat{\beta}) \\
&= y^{,}\left(\Sigma_c^{-1} \otimes I_{NT}\right) y - 2\hat{\beta}^{,}\left(\Sigma_c^{-1} \otimes X^{,}\right) y + \hat{\beta}^{,}\left(\Sigma_c^{-1} \otimes X^{,}X\right)\hat{\beta} \\
&\quad +(\beta - \hat{\beta})^{,}\left(\Sigma_c \otimes (X^{,}X)^{-1}\right)^{-1}(\beta - \hat{\beta}) \qquad A.1.2
\end{aligned} \tag{A.13.4}$$

Reshape the first row of (A.13.4):

$$
\begin{aligned}
& y^{,}\left(\Sigma_c^{-1}\otimes I_{NT}\right)y-2\hat{\beta}^{,}\left(\Sigma_c^{-1}\otimes X^{,}\right)y+\hat{\beta}^{,}\left(\Sigma_c^{-1}\otimes X^{,}X\right)\hat{\beta} \\
&= y^{,}\left(\Sigma_c^{-1}\otimes I_{NT}\right)y-2\hat{\beta}^{,}(I_n\otimes X)^{,}\left(\Sigma_c^{-1}\otimes I_{NT}\right)y \\
&\quad +\hat{\beta}^{,}\left(\Sigma_c^{-1}\otimes X^{,}\right)(I_n\otimes X)\hat{\beta} \qquad A.1.1, A.1.3 \\
&= y^{,}\left(\Sigma_c^{-1}\otimes I_{NT}\right)y-2\Big((I_n\otimes X)\hat{\beta}\Big)^{,}\left(\Sigma_c^{-1}\otimes I_{NT}\right)y \\
&\quad +\hat{\beta}^{,}(I_n\otimes X)^{,}\left(\Sigma_c^{-1}\otimes I_{NT}\right)(I_n\otimes X)\hat{\beta} \qquad A.1.3 \\
&= y^{,}(\Sigma_c\otimes I_{NT})^{-1}y-2\Big((I_n\otimes X)\hat{\beta}\Big)^{,}(\Sigma_c\otimes I_{NT})^{-1}y \\
&\quad +\hat{\beta}^{,}(I_n\otimes X)^{,}(\Sigma_c\otimes I_{NT})^{-1}(I_n\otimes X)\hat{\beta} \qquad A.1.2 \\
&= tr\left\{\Sigma_c^{-1}Y^{,}I_{NT}Y\right\}-2tr\left\{\Sigma_c^{-1}(X\hat{B})^{,}I_{NT}Y\right\}+tr\left\{\Sigma_c^{-1}\hat{B}^{,}X^{,}I_{NT}X\hat{B}\right\} \qquad A.1.5, A.1.10 \\
&= tr\left\{Y\Sigma_c^{-1}Y^{,}\right\}-2tr\left\{Y\Sigma_c^{-1}\hat{B}^{,}X^{,}\right\}+tr\left\{X\hat{B}\Sigma_c^{-1}\hat{B}^{,}X^{,}\right\} \qquad A.17.15 \\
&= tr\left\{Y\Sigma_c^{-1}Y^{,}-2Y\Sigma_c^{-1}\hat{B}^{,}X^{,}+X\hat{B}\Sigma_c^{-1}\hat{B}^{,}X^{,}\right\} \qquad A.17.15 \\
&= tr\left\{(Y-X\hat{B})\Sigma_c^{-1}(Y-X\hat{B})^{,}\right\} \\
&= tr\left\{\Sigma_c^{-1}(Y-X\hat{B})^{,}(Y-X\hat{B})\right\} \qquad A.17.15
\end{aligned}
\tag{A.13.5}
$$

Reshape then the second row of (A.13.4):

$$
\begin{aligned}
& (\beta-\hat{\beta})^{,}\left(\Sigma_c\otimes(X^{,}X)^{-1}\right)^{-1}(\beta-\hat{\beta}) \\
&= tr\left\{\Sigma_c^{-1}\Big(B-\hat{B}\Big)^{,}(X^{,}X)\Big(B-\hat{B}\Big)\right\} \qquad A.1.10
\end{aligned}
\tag{A.13.6}
$$

Substitute finally (A.13.5) and (A.13.6) in (A.13.4) to obtain:

$$
\begin{aligned}
& (y-(I_n\otimes X)\beta)^{,}(\Sigma_c\otimes I_{NT})^{-1}(y-(I_n\otimes X)\beta) \\
&= tr\left\{\Sigma_c^{-1}(Y-X\hat{B})^{,}(Y-X\hat{B})\right\}+tr\left\{\Sigma_c^{-1}\Big(B-\hat{B}\Big)^{,}(X^{,}X)\Big(B-\hat{B}\Big)\right\}
\end{aligned}
\tag{A.13.7}
$$

Before turning back to (A.13.1), note also that Kronecker property A.1.4 implies that the deter-

minant part of (A.13.1) can rewrite as:

$$|\Sigma_c \otimes I_{NT}|^{-1/2} = \left(|\Sigma_c|^{NT}|I_T|^n\right)^{-1/2} = |\Sigma_c|^{-NT/2} \tag{A.13.8}$$

Substituting (A.13.7) and (A.13.8) in (A.13.1), one eventually obtains:

$f(y\,|\beta,\Sigma_c) \propto |\Sigma_c|^{-NT/2}$

$\exp\left(-\frac{1}{2}\left[tr\left\{\Sigma_c^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right\} + tr\left\{\Sigma_c^{-1}\left(B-\hat{B}\right)'(X'X)\left(B-\hat{B}\right)\right\}\right]\right)$

Or, rearranging:

$$\begin{aligned} f(y\,|\beta,\Sigma_c) \propto |\Sigma_c|^{-NT/2} \exp\left[-\frac{1}{2}tr\left\{\Sigma_c^{-1}\left(B-\hat{B}\right)'(X'X)\left(B-\hat{B}\right)\right\}\right] \\ \times \exp\left[-\frac{1}{2}tr\left\{\Sigma_c^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right\}\right] \end{aligned} \tag{A.13.9}$$

The prior distribution for β is multivariate normal:

$\beta \sim \mathcal{N}(\beta_0, \Sigma_c \otimes \Phi_0)$

Therefore, its density is given by:

$$\pi(\beta) = (2\pi)^{-q/2}|\Sigma_c \otimes \Phi_0|^{-1/2} \exp\left(-\frac{1}{2}(\beta-\beta_0)'(\Sigma_c \otimes \Phi_0)^{-1}(\beta-\beta_0)\right) \tag{A.13.10}$$

or, using A.1.4 :

$\pi(\beta) = (2\pi)^{-q/2}|\Sigma_c|^{-k/2}|\Phi_0|^{-n/2} \exp\left(-\frac{1}{2}(\beta-\beta_0)'(\Sigma_c \otimes \Phi_0)^{-1}(\beta-\beta_0)\right)$

The kernel is given by:

$\pi(\beta) \propto |\Sigma_c|^{-k/2} \exp\left[-\frac{1}{2}(\beta-\beta_0)'(\Sigma_c \otimes \Phi_0)^{-1}(\beta-\beta_0)\right]$

Using A.1.10, this rewrites as:

$$\pi(\beta) \propto |\Sigma_c|^{-k/2} \exp\left[-\frac{1}{2} tr\left\{\Sigma_c^{-1}(B-B_0)' \Phi_0^{-1}(B-B_0)\right\}\right] \tag{A.13.11}$$

The prior for Σ_c is inverse Wishart:

$\Sigma_c \sim \ IW(S_0, \alpha_0)$

The prior density is given by:

$\pi(\Sigma_c|S_0, \alpha_0) = \frac{1}{2^{\alpha_0 n/2}\Gamma_n\left(\frac{\alpha_0}{2}\right)}|S_0|^{\alpha_0/2}|\Sigma_c|^{-(\alpha_0+n+1)/2} \exp\left(-\frac{1}{2} tr\left\{\Sigma_c^{-1} S_0\right\}\right)$

The kernel is given by:

$$\pi(\Sigma_c) \propto |\Sigma_c|^{-(\alpha_0+n+1)/2} \exp\left[-\frac{1}{2} tr\left\{\Sigma_c^{-1} S_0\right\}\right] \tag{A.13.12}$$

Combining the likelihood (A.13.9) with the prior distributions (A.13.11) and (A.13.12), one obtains the posterior distribution as:

$$\begin{aligned}
\pi(\beta, \Sigma\,|y) &\propto f(y\,|\beta, \Sigma)\pi(\beta)\pi(\Sigma) \\
&= |\Sigma_c|^{-NT/2} \exp\left[-\frac{1}{2} tr\left\{\Sigma_c^{-1}\left(B-\hat{B}\right)'(X'X)\left(B-\hat{B}\right)\right\}\right] \\
&\quad \times \exp\left[-\frac{1}{2} tr\left\{\Sigma_c^{-1}(Y-X\hat{B})'(Y-X\hat{B})\right\}\right] \\
&\quad \times |\Sigma_c|^{-k/2} \exp\left[-\frac{1}{2} tr\left\{\Sigma_c^{-1}(B-B_0)' \Phi_0^{-1}(B-B_0)\right\}\right] \\
&\quad \times |\Sigma|^{-(\alpha_0+n+1)/2} \exp\left[-\frac{1}{2} tr\left\{\Sigma_c^{-1} S_0\right\}\right]
\end{aligned} \tag{A.13.13}$$

Rearranging:

$$\pi(\beta,\Sigma\,|y\,)\propto|\Sigma_c|^{-(NT+k+\alpha_0+n+1)/2}$$
$$\times\exp\left[-\frac{1}{2}tr\left\{\Sigma_c^{-1}\left[\left(B-\hat{B}\right)'(X'X)\left(B-\hat{B}\right)+(B-B_0)'\Phi_0^{-1}(B-B_0)\right]\right\}\right]$$
$$\times\exp\left[-\frac{1}{2}tr\left\{\Sigma_c^{-1}\left[S_0+(Y-X\hat{B})'(Y-X\hat{B})\right]\right\}\right] \tag{A.13.14}$$

Focusing on the term in the curly brace in the second row:

$$\begin{aligned}&\Sigma_c^{-1}\left[\left(B-\hat{B}\right)'(X'X)\left(B-\hat{B}\right)+(B-B_0)'\Phi_0^{-1}(B-B_0)\right]\\&=\Sigma_c^{-1}\left[B'X'XB+\hat{B}'X'X\hat{B}-2B'X'X\hat{B}+B'\Phi_0^{-1}B+B_0'\Phi_0^{-1}B_0-2B'\Phi_0^{-1}B_0\right]\\&=\Sigma_c^{-1}\left[B'\left(X'X+\Phi_0^{-1}\right)B-2B'\left(X'X\hat{B}+\Phi_0^{-1}B_0\right)+\hat{B}'X'X\hat{B}+B_0'\Phi_0^{-1}B_0\right]\end{aligned}$$

Completing the squares:

$$\begin{aligned}&=\Sigma_c^{-1}\Bigg[B'\left(X'X+\Phi_0^{-1}\right)B-2B'\bar{\Phi}^{-1}\bar{\Phi}\left(X'X\hat{B}+\Phi_0^{-1}B_0\right)+\bar{B}'\bar{\Phi}^{-1}\bar{B}-\bar{B}'\bar{\Phi}^{-1}\bar{B}\\&\quad+\hat{B}'X'X\hat{B}+B_0'\Phi_0^{-1}B_0\Bigg]\end{aligned}$$

Defining:

$$\bar{\Phi}=\left[\Phi_0^{-1}+X'X\right]^{-1} \tag{A.13.15}$$

and

$$\bar{B}=\bar{\Phi}\left[\Phi_0^{-1}B_0+X'X\hat{B}\right] \tag{A.13.16}$$

The previous expression may rewrite:

$$
\begin{aligned}
&= \Sigma_c^{-1}\left[B'\bar{\Phi}^{-1}B - 2B'\bar{\Phi}^{-1}\bar{B} + \bar{B}'\bar{\Phi}^{-1}\bar{B} - \bar{B}'\bar{\Phi}^{-1}\bar{B} + \hat{B}'X'X\hat{B} + B_0'\Phi_0^{-1}B_0\right] \\
&= \Sigma_c^{-1}\left[(B-\bar{B})'\bar{\Phi}^{-1}(B-\bar{B}) - \bar{B}'\bar{\Phi}^{-1}\bar{B} + \hat{B}'X'X\hat{B} + B_0'\Phi_0^{-1}B_0\right] \\
&= \Sigma_c^{-1}\left[(B-\bar{B})'\bar{\Phi}^{-1}(B-\bar{B})\right] + \Sigma_c^{-1}\left[\hat{B}'X'X\hat{B} + B_0'\Phi_0^{-1}B_0 - \bar{B}'\bar{\Phi}^{-1}\bar{B}\right]
\end{aligned}
$$

Substituting back in the posterior (A.13.14), it becomes:

$$
\begin{aligned}
\pi(\beta,\Sigma_c\,|y\,) \propto\; & |\Sigma_c|^{-(NT+k+\alpha_0+n+1)/2} \\
& \times \exp\left[-\frac{1}{2}tr\left\{\Sigma_c^{-1}\left[(B-\bar{B})'\bar{\Phi}^{-1}(B-\bar{B})\right] + \Sigma_c^{-1}\left[\hat{B}'X'X\hat{B} + B_0'\Phi_0^{-1}B_0 - \bar{B}'\bar{\Phi}^{-1}\bar{B}\right]\right\}\right] \\
& \times \exp\left[-\frac{1}{2}tr\left\{\Sigma_c^{-1}\left[S_0 + (Y-X\hat{B})'(Y-X\hat{B})\right]\right\}\right] \\
=\; & |\Sigma_c|^{-(NT+k+\alpha_0+n+1)/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma_c^{-1}\left[(B-\bar{B})'\bar{\Phi}^{-1}(B-\bar{B})\right]\right\}\right] \qquad A.1.2 \\
& \times \exp\left[-\frac{1}{2}tr\left\{\Sigma_c^{-1}\left[S_0 + (Y-X\hat{B})'(Y-X\hat{B}) + \hat{B}'X'X\hat{B} + B_0'\Phi_0^{-1}B_0 - \bar{B}'\bar{\Phi}^{-1}\bar{B}\right]\right\}\right] \\
=\; & |\Sigma_c|^{-k/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma_c^{-1}\left[(B-\bar{B})'\bar{\Phi}^{-1}(B-\bar{B})\right]\right\}\right] \\
& \times|\Sigma_c|^{-(NT+\alpha_0+n+1)/2} \\
& \times \exp\left[-\frac{1}{2}tr\left\{\Sigma_c^{-1}\left[S_0 + (Y-X\hat{B})'(Y-X\hat{B}) + \hat{B}'X'X\hat{B} + B_0'\Phi_0^{-1}B_0 - \bar{B}'\bar{\Phi}^{-1}\bar{B}\right]\right\}\right]
\end{aligned}
$$

Define:

$$\bar{\alpha} = NT + \alpha_0 \tag{A.13.17}$$

and

$$\bar{S} = S_0 + (Y-X\hat{B})'(Y-X\hat{B}) + \hat{B}'X'X\hat{B} + B_0'\Phi_0^{-1}B_0 - \bar{B}'\bar{\Phi}^{-1}\bar{B} \tag{A.13.18}$$

Then the previous equation rewrites:

$$
\begin{aligned}
\pi(\beta,\Sigma_c\,|y\,) \propto\; & |\Sigma_c|^{-k/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma_c^{-1}\left[(B-\bar{B})'\bar{\Phi}^{-1}(B-\bar{B})\right]\right\}\right] \\
& \times|\Sigma_c|^{-(\bar{\alpha}+n+1)/2}\exp\left[-\frac{1}{2}tr\left\{\Sigma_c^{-1}\bar{S}\right\}\right]
\end{aligned} \tag{A.13.19}
$$

This is exactly similar to A.4.16. Therefore, following a reasoning identical to that of A.5, it follows immediately that the marginal posteriors are given by:

$$\pi(\Sigma_c \,|y) \sim \; IW\left(\bar{\alpha}, \bar{S}\right) \tag{A.13.20}$$

and

$$\pi(B \,|y) \sim \; MT(\bar{B}, \bar{S}, \bar{\Phi}, \tilde{\alpha}) \tag{A.13.21}$$

with:

$$\tilde{\alpha} = \bar{\alpha} - n + 1 = NT + \alpha_0 - n + 1 \tag{A.13.22}$$

Finally, from (a.4.24) and (a.4.25), $\bar{S}$ and $\bar{B}$ can simplify to:

$$\bar{S} = Y`Y + S_0 + B_0`\Phi_0^{-1}B_0 - \bar{B}`\bar{\Phi}^{-1}\bar{B} \tag{A.13.23}$$

and

$$\bar{B} = \bar{\Phi}\left[\Phi_0^{-1}B_0 + X`Y\right] \tag{A.13.24}$$

A.14 Derivations for the Zellner and Hong prior

The likelihood function for the data is given by:

$$f(y\,|\beta, \bar{\Sigma}) = (2\pi)^{-nT/2}\left|\bar{\Sigma}\right|^{-1/2}\exp\left[-\frac{1}{2}\left(y - \bar{X}\beta\right)`\bar{\Sigma}^{-1}\left(y - \bar{X}\beta\right)\right] \tag{A.14.1}$$

Or, getting rid of the proportionality terms:

$$f(y\,|\beta) \propto \exp\left[-\frac{1}{2}\left(y - \bar{X}\beta\right)`\bar{\Sigma}^{-1}\left(y - \bar{X}\beta\right)\right] \tag{A.14.2}$$

Given (6.5.5), the prior density for β is given by:

$$\pi(\beta) = (2\pi)^{-h/2}\left|\bar{\Sigma}_b\right|^{-1/2}\exp\left(-\frac{1}{2}(\beta - \bar{b})`\bar{\Sigma}_b^{-1}(\beta - \bar{b})\right) \tag{A.14.3}$$

Getting rid of the proportionality terms:

$$\pi(\beta) \propto \exp\left(-\frac{1}{2}(\beta-\bar{b})'\bar{\Sigma}_b^{-1}(\beta-\bar{b})\right) \tag{A.14.4}$$

Then, using Bayes rule 3.2.3, one combines the likelihood and the prior to obtain:

$$f(\beta\,|y) \propto \exp\left[-\frac{1}{2}\left\{\left(y-\bar{X}\beta\right)'\bar{\Sigma}^{-1}\left(y-\bar{X}\beta\right)+\left(\beta-\bar{b}\right)'\bar{\Sigma}_b^{-1}\left(\beta-\bar{b}\right)\right\}\right] \tag{A.14.5}$$

Consider the term in the curly bracket of (A.14.5), and use (6.5.8) and (6.5.9) to develop:

$$\begin{aligned}
&\left(y-\bar{X}\beta\right)'\bar{\Sigma}^{-1}\left(y-\bar{X}\beta\right)+\left(\beta-\bar{b}\right)'\bar{\Sigma}_b^{-1}\left(\beta-\bar{b}\right)\\
&=\left(y-\bar{X}\beta\right)'\left(\sigma_\varepsilon^2 I_{NnT}\right)^{-1}\left(y-\bar{X}\beta\right)+\left(\beta-\bar{b}\right)'\left(\lambda_1\sigma_\varepsilon^2 I_q\right)^{-1}\left(\beta-\bar{b}\right)\\
&=\left(y-\bar{X}\beta\right)'\sigma_\varepsilon^{-2}\left(y-\bar{X}\beta\right)+\lambda_1^{-1}\left(\beta-\bar{b}\right)'\sigma_\varepsilon^{-2}\left(\beta-\bar{b}\right)\\
&=y'\sigma_\varepsilon^{-2}y+\beta'\bar{X}'\sigma_\varepsilon^{-2}\bar{X}\beta-2\beta'\bar{X}'\sigma_\varepsilon^{-2}y+\lambda_1^{-1}\beta'\sigma_\varepsilon^{-2}\beta+\lambda_1^{-1}\bar{b}'\sigma_\varepsilon^{-2}\bar{b}-2\lambda_1^{-1}\beta'\sigma_\varepsilon^{-2}\bar{b}\\
&=y'\sigma_\varepsilon^{-2}y+\beta'\left(\lambda_1^{-1}\sigma_\varepsilon^{-2}I_h+\sigma_\varepsilon^{-2}\bar{X}'\bar{X}\right)\beta-2\beta'\left(\bar{X}'\sigma_\varepsilon^{-2}y+\lambda_1^{-1}\sigma_\varepsilon^{-2}\bar{b}\right)+\lambda_1^{-1}\bar{b}'\sigma_\varepsilon^{-2}\bar{b}
\end{aligned}$$

Completing the squares:

$$\begin{aligned}
&=y'\sigma_\varepsilon^{-2}y+\beta'\left(\lambda_1^{-1}\sigma_\varepsilon^{-2}I_h+\sigma_\varepsilon^{-2}\bar{X}'\bar{X}\right)\beta-2\beta'\bar{\Omega}_b^{-1}\bar{\Omega}_b\left(\bar{X}'\sigma_\varepsilon^{-2}y+\lambda_1^{-1}\sigma_\varepsilon^{-2}\bar{b}\right)\\
&\quad+\lambda_1^{-1}\bar{b}'\sigma_\varepsilon^{-2}\bar{b}+\bar{\beta}'\bar{\Omega}_b^{-1}\bar{\beta}-\bar{\beta}'\bar{\Omega}_b^{-1}\bar{\beta}
\end{aligned}$$

Define:

$$\bar{\Omega}_b=\left(\lambda_1^{-1}\sigma_\varepsilon^{-2}I_h+\sigma_\varepsilon^{-2}\bar{X}'\bar{X}\right)^{-1} \tag{A.14.6}$$

and

$$\bar{\beta}=\bar{\Omega}_b\left(\bar{X}'\sigma_\varepsilon^{-2}y+\lambda_1^{-1}\sigma_\varepsilon^{-2}\bar{b}\right) \tag{A.14.7}$$

Then the previous expression becomes:

$$\begin{aligned}
&=y'\sigma_\varepsilon^{-2}y+\beta'\bar{\Omega}_b^{-1}\beta-2\beta'\bar{\Omega}_b^{-1}\bar{\beta}+\lambda_1^{-1}\bar{b}'\sigma_\varepsilon^{-2}\bar{b}+\bar{\beta}'\bar{\Omega}_b^{-1}\bar{\beta}-\bar{\beta}'\bar{\Omega}_b^{-1}\bar{\beta}\\
&=(\beta-\bar{\beta})'\bar{\Omega}_b^{-1}(\beta-\bar{\beta})+y'\sigma_\varepsilon^{-2}y+\lambda_1^{-1}\bar{b}'\sigma_\varepsilon^{-2}\bar{b}-\bar{\beta}'\bar{\Omega}_b^{-1}\bar{\beta}
\end{aligned} \tag{A.14.8}$$

Substitute eventually (A.14.8) in (A.14.5) to obtain:

$$\begin{aligned} f(\beta\,|y) &\propto \exp\left[-\frac{1}{2}\left\{(\beta-\bar{\beta})'\bar{\Omega}_b^{-1}(\beta-\bar{\beta})+y'\sigma_\varepsilon^{-2}y+\lambda_1^{-1}\bar{b}'\sigma_\varepsilon^{-2}\bar{b}-\bar{\beta}'\bar{\Omega}_b^{-1}\bar{\beta}\right\}\right] \\ &\propto \exp\left[-\frac{1}{2}(\beta-\bar{\beta})'\bar{\Omega}_b^{-1}(\beta-\bar{\beta})\right]\exp\left[-\frac{1}{2}\left(y'\sigma_\varepsilon^{-2}y+\lambda_1^{-1}\bar{b}'\sigma_\varepsilon^{-2}\bar{b}-\bar{\beta}'\bar{\Omega}_b^{-1}\bar{\beta}\right)\right] \\ &\propto \exp\left[-\frac{1}{2}(\beta-\bar{\beta})'\bar{\Omega}_b^{-1}(\beta-\bar{\beta})\right] \end{aligned} \tag{A.14.9}$$

This is the kernel of a multivariate normal distribution with mean $\bar{\beta}$ and covariance matrix $\bar{\Omega}_b$. Finally, simplify the expressions. First, it is immediate from (A.14.6) that $\bar{\Omega}_b$ rewrites:

$$\bar{\Omega}_b = \sigma_\varepsilon^2\left(\lambda_1^{-1}I_h+\bar{X}'\bar{X}\right)^{-1} \tag{A.14.10}$$

Then, combining (A.14.7) and (A.14.10):

$$\begin{aligned} \bar{\beta} &= \sigma_\varepsilon^2\left(\lambda_1^{-1}I_h+\bar{X}'\bar{X}\right)^{-1}\left(\bar{X}'\sigma_\varepsilon^{-2}y+\lambda_1^{-1}\sigma_\varepsilon^{-2}\bar{b}\right) \\ &= \sigma_\varepsilon^2\left(\lambda_1^{-1}I_h+\bar{X}'\bar{X}\right)^{-1}\sigma_\varepsilon^{-2}\left(\bar{X}'y+\lambda_1^{-1}\bar{b}\right) \\ &= \left(\lambda_1^{-1}I_h+\bar{X}'\bar{X}\right)^{-1}\left(\bar{X}'y+\lambda_1^{-1}\bar{b}\right) \end{aligned} \tag{A.14.11}$$

A.15 Derivations for the hierarchical prior

First obtain Bayes rule for this specific hierarchical prior model. Start from the definition of the posterior distribution and develop:

$$\begin{aligned} \pi\left(\beta,b,\Sigma_b,\Sigma\,|y\right) &= \frac{\pi\left(\beta,b,\Sigma_b,\Sigma,y\right)}{\pi\left(y\right)} \\ &\propto \pi\left(\beta,b,\Sigma_b,\Sigma,y\right) \\ &= \frac{\pi\left(y,\beta,b,\Sigma_b,\Sigma\right)}{\pi\left(\beta,b,\Sigma_b,\Sigma\right)}\pi\left(\beta,b,\Sigma_b,\Sigma\right) \\ &= \pi\left(y\,|\beta,b,\Sigma_b,\Sigma\right)\pi\left(\beta,b,\Sigma_b,\Sigma\right) \end{aligned}$$

Now assuming as usual independence between β and Σ, this may rewrite:

$$\begin{aligned}
&= \pi\left(y\,|\beta, b, \Sigma_b, \Sigma\right)\pi\left(\beta, b, \Sigma_b\right)\pi\left(\Sigma\right) \\
&= \pi\left(y\,|\beta, b, \Sigma_b, \Sigma\right)\frac{\pi\left(\beta, b, \Sigma_b\right)}{\pi\left(b, \Sigma_b\right)}\pi\left(b, \Sigma_b\right)\pi\left(\Sigma\right) \\
&= \pi\left(y\,|\beta, b, \Sigma_b, \Sigma\right)\pi\left(\beta\,|b, \Sigma_b\right)\pi\left(b, \Sigma_b\right)\pi\left(\Sigma\right)
\end{aligned}$$

And also assuming independence between b and Σ_b, this yields:

$$= \pi\left(y\,|\beta, b, \Sigma_b, \Sigma\right)\pi\left(\beta\,|b, \Sigma_b\right)\pi\left(b\right)\pi\left(\Sigma_b\right)\pi\left(\Sigma\right)$$

b and Σ_b are relevant only inasmuch as they allow to determine β. In other words, any expression conditioning on β, b and Σ_b can be simplified as an expression conditioning on β only, since once a value for β is drawn from $\pi\left(\beta\,|b, \Sigma_b\right)$, b and Σ_b can have no further impact. This is the case for the likelihood function $\pi\left(y\,|\beta, b, \Sigma_b, \Sigma\right)$, which can hence rewrite simply as $\pi\left(y\,|\beta, \Sigma\right)$. From this, one eventually obtains:

$$\pi\left(\beta, b, \Sigma_b, \Sigma\,|y\right) \propto \pi\left(y\,|\beta, \Sigma\right)\pi\left(\beta\,|b, \Sigma_b\right)\pi\left(b\right)\pi\left(\Sigma_b\right)\pi\left(\Sigma\right) \tag{A.15.1}$$

Now turn to the derivation of the likelihood function. Given (6.4.12) and (6.4.14), the likelihood function for unit i writes as:

$$\begin{aligned}
\pi\left(y_i\,|\beta_i, \Sigma_i\right) &= (2\pi)^{-nT/2}\left|\bar{\Sigma}_i\right|^{-1/2}\exp\left(-\frac{1}{2}\left(y_i - \bar{X}_i\beta_i\right)'\bar{\Sigma}_i^{-1}\left(y_i - \bar{X}_i\beta_i\right)\right) \\
&\propto \left|\bar{\Sigma}_i\right|^{-1/2}\exp\left(-\frac{1}{2}\left(y_i - \bar{X}_i\beta_i\right)'\bar{\Sigma}_i^{-1}\left(y_i - \bar{X}_i\beta_i\right)\right)
\end{aligned} \tag{A.15.2}$$

Then, because static interdependencies does not apply, the residual series ε_i are independent across units, which allows to obtain the likelihood for the full data set from (A.15.2) as:

$$\begin{aligned}
\pi\left(y\,|\beta, \Sigma\right) &= \prod_{i=1}^{N}\pi\left(y_i\,|\beta_i, \Sigma_i\right) \\
&= \prod_{i=1}^{N}\left|\bar{\Sigma}_i\right|^{-1/2}\exp\left(-\frac{1}{2}\left(y_i - \bar{X}_i\beta_i\right)'\bar{\Sigma}_i^{-1}\left(y_i - \bar{X}_i\beta_i\right)\right)
\end{aligned} \tag{A.15.3}$$

Before starting the derivation of the conditional posterior distributions, obtain first the full prior distributions for β and Σ. Start with the prior distribution for β. From (6.4.16), the prior distribution

for β_i is multivariate normal: $\beta_i \sim \mathcal{N}(b, \Sigma_b)$. Therefore, its density is given by:

$$\begin{aligned}\pi(\beta_i | b, \Sigma_b) &= (2\pi)^{-q/2} |\Sigma_b|^{-1/2} \exp\left(-\frac{1}{2}(\beta_i - b)' \Sigma_b^{-1} (\beta_i - b)\right) \\ &\propto |\Sigma_b|^{-1/2} \exp\left(-\frac{1}{2}(\beta_i - b)' \Sigma_b^{-1} (\beta_i - b)\right)\end{aligned}$$

Then, assuming independence between the β_is, one obtains:

$$\begin{aligned}\pi(\beta | b, \Sigma_b) &= \prod_{i=1}^{N} \pi(\beta_i | b, \Sigma_b) \\ &\propto \prod_{i=1}^{N} |\Sigma_b|^{-1/2} \exp\left(-\frac{1}{2}(\beta_i - b)' \Sigma_b^{-1} (\beta_i - b)\right)\end{aligned} \tag{A.15.4}$$

Similarly, assuming independence between the Σ_is, one obtains from (6.6.14):

$$\begin{aligned}\pi(\Sigma_i) &\propto \prod_{i=1}^{N} \pi(\Sigma_i) \\ &= \prod_{i=1}^{N} |\Sigma_i|^{-(n+1)/2}\end{aligned} \tag{A.15.5}$$

Now derive the conditional posteriors. Obtain first the conditional posterior for β. Start from (6.6.16), and substitute for (6.6.4) and (6.6.6), for unit i only:

$$\begin{aligned}\pi(\beta_i | \beta_{-i}, y, b, \Sigma_b, \Sigma) &\propto \left|\bar{\Sigma}_i\right|^{-1/2} \exp\left(-\frac{1}{2}\left(y_i - \bar{X}_i\beta_i\right)' \left(\bar{\Sigma}_i\right)^{-1} \left(y_i - \bar{X}_i\beta_i\right)\right) \\ &\quad \times |\Sigma_b|^{-1/2} \exp\left(-\frac{1}{2}(\beta_i - b)' (\Sigma_b)^{-1} (\beta_i - b)\right) \\ &= \exp\left(-\frac{1}{2}\left\{\left(y_i - \bar{X}_i\beta_i\right)' \left(\bar{\Sigma}_i\right)^{-1} \left(y_i - \bar{X}_i\beta_i\right) + (\beta_i - b)' (\Sigma_b)^{-1} (\beta_i - b)\right\}\right)\end{aligned} \tag{A.15.6}$$

Following the usual strategy, consider the curly bracket term and develop:

$$\begin{aligned}&\left(y_i - \bar{X}_i\beta_i\right)' \bar{\Sigma}_i^{-1} \left(y_i - \bar{X}_i\beta_i\right) + (\beta_i - b)' \Sigma_b^{-1} (\beta_i - b) \\ &= y_i' \bar{\Sigma}_i^{-1} y_i + \beta_i' \bar{X}_i' \bar{\Sigma}_i^{-1} \bar{X}_i \beta_i - 2\beta_i' \bar{X}_i' \bar{\Sigma}_i^{-1} y_i + \beta_i' \Sigma_b^{-1} \beta_i + b' \Sigma_b^{-1} b - 2\beta_i' \Sigma_b^{-1} b \\ &= y_i' \bar{\Sigma}_i^{-1} y_i + \beta_i' \left(\bar{X}_i' \bar{\Sigma}_i^{-1} \bar{X}_i + \Sigma_b^{-1}\right) \beta_i - 2\beta_i' \left(\bar{X}_i' \bar{\Sigma}_i^{-1} y_i + \Sigma_b^{-1} b\right) + b' \Sigma_b^{-1} b\end{aligned}$$

Complete the squares:

$$\begin{aligned} & y_i'\bar{\Sigma}_i^{-1}y_i + \beta_i'\left(\bar{X}_i'\bar{\Sigma}_i^{-1}\bar{X}_i + \Sigma_b^{-1}\right)\beta_i - 2\beta_i'\left(\bar{X}_i'\bar{\Sigma}_i^{-1}y_i + \Sigma_b^{-1}b\right) + b'\Sigma_b^{-1}b \\ = & \, y_i'\bar{\Sigma}_i^{-1}y_i + \beta_i'\left(\bar{X}_i'\bar{\Sigma}_i^{-1}\bar{X}_i + \Sigma_b^{-1}\right)\beta_i - 2\beta_i'\bar{\Omega}_i^{-1}\bar{\Omega}_i\left(\bar{X}_i'\bar{\Sigma}_i^{-1}y_i + \Sigma_b^{-1}b\right) \\ + & \, b'\Sigma_b^{-1}b + \bar{\beta}_i'\bar{\Omega}_i^{-1}\bar{\beta}_i - \bar{\beta}_i'\bar{\Omega}_i^{-1}\bar{\beta}_i \end{aligned} \tag{A.15.7}$$

Define:

$$\bar{\Omega}_i = \left(\bar{X}_i'\bar{\Sigma}_i^{-1}\bar{X}_i + \Sigma_b^{-1}\right)^{-1} \tag{A.15.8}$$

and:

$$\bar{\beta}_i = \bar{\Omega}_i\left(\bar{X}_i'\bar{\Sigma}_i^{-1}y_i + \Sigma_b^{-1}b\right) \tag{A.15.9}$$

Then (A.15.7) rewrites:

$$\begin{aligned} & = y_i'\bar{\Sigma}_i^{-1}y_i + \beta_i'\bar{\Omega}_i^{-1}\beta_i - 2\beta_i'\bar{\Omega}_i^{-1}\bar{\beta} + b'\Sigma_b^{-1}b + \bar{\beta}_i'\bar{\Omega}_i^{-1}\bar{\beta}_i - \bar{\beta}_i'\bar{\Omega}_i^{-1}\bar{\beta}_i \\ & = \left(\beta_i'\bar{\Omega}_i^{-1}\beta_i - 2\beta_i'\bar{\Omega}_i^{-1}\bar{\beta}_i + \bar{\beta}_i'\bar{\Omega}_i^{-1}\bar{\beta}_i\right) + \left(b'\Sigma_b^{-1}b - \bar{\beta}_i'\bar{\Omega}_i^{-1}\bar{\beta}_i + y_i'\bar{\Sigma}_i^{-1}y_i\right) \\ & = \left(\beta_i - \bar{\beta}_i\right)'\bar{\Omega}_i^{-1}\left(\beta_i - \bar{\beta}_i\right) + \left(b'\Sigma_b^{-1}b - \bar{\beta}_i'\bar{\Omega}_i^{-1}\bar{\beta}_i + y_i'\bar{\Sigma}_i^{-1}y_i\right) \end{aligned} \tag{A.15.10}$$

Substitute back in (A.15.6):

$$\begin{aligned} \pi\left(\beta_i\,|\beta_{-i}, y, b, \Sigma_b, \Sigma\right) & \propto \exp\left[-\frac{1}{2}\left\{\left(\beta_i - \bar{\beta}\right)'\bar{\Omega}_i^{-1}\left(\beta_i - \bar{\beta}\right) + \left(b'\Sigma_b^{-1}b - \bar{\beta}_i'\bar{\Omega}_i^{-1}\bar{\beta}_i + y_i'\bar{\Sigma}_i^{-1}y_i\right)\right\}\right] \\ & = \exp\left[-\frac{1}{2}\left(\beta_i - \bar{\beta}\right)'\bar{\Omega}_i^{-1}\left(\beta_i - \bar{\beta}\right)\right]\exp\left[-\frac{1}{2}\left(b'\Sigma_b^{-1}b - \bar{\beta}_i'\bar{\Omega}_i^{-1}\bar{\beta}_i + y_i'\bar{\Sigma}_i^{-1}y_i\right)\right] \\ & \propto \exp\left[-\frac{1}{2}\left(\beta_i - \bar{\beta}\right)'\bar{\Omega}_i^{-1}\left(\beta_i - \bar{\beta}\right)\right] \end{aligned} \tag{A.15.11}$$

One eventually obtains:

$$\pi\left(\beta_i\,|\beta_{-i}, y, b, \Sigma_b, \Sigma\right) \propto \exp\left[-\frac{1}{2}\left(\beta_i - \bar{\beta}_i\right)'\bar{\Omega}_i^{-1}\left(\beta_i - \bar{\beta}_i\right)\right] \tag{A.15.12}$$

Therefore, the posterior for β_i is conditionally normal, with mean $\bar{\beta}_i$ and covariance matrix $\bar{\Omega}_i$. Considering (A.15.8) and (A.15.9), and making use of A.3.9 and A.3.11, it is possible to simplify $\bar{\beta}_i$ and $\bar{\Omega}_i$ as:

$$\bar{\Omega}_i = \left[\Sigma_i^{-1} \otimes X_i'X_i + \Sigma_b^{-1}\right]^{-1} \tag{A.15.13}$$

and:

$$\bar{\beta}_i = \bar{\Omega}_i \left[\left(\Sigma_i^{-1} \otimes X_i \right) y_i + \Sigma_b^{-1} b \right] \tag{A.15.14}$$

Obtaining the conditional posterior for b turns out to be a bit trickier. Start from the conditional posterior obtained from Bayes rule (6.6.20):

$$\pi \left(b \left| y, \beta, \Sigma_b, \Sigma \right. \right) \propto \pi \left(\beta \left| b, \Sigma_b \right. \right) \pi \left(b \right)$$

Using (6.6.6) and (6.6.7), this yields:

$$\begin{aligned} \pi \left(b \left| y, \beta, \Sigma_b, \Sigma \right. \right) &\propto \pi \left(\beta \left| b, \Sigma_b \right. \right) \pi \left(b \right) \\ &\propto \prod_{i=1}^{N} \exp \left(-\frac{1}{2} (\beta_i - b)' \left(\Sigma_b \right)^{-1} \left(\beta_i - b \right) \right) \times 1 \\ &= \prod_{i=1}^{N} \exp \left(-\frac{1}{2} (\beta_i - b)' \left(\Sigma_b \right)^{-1} \left(\beta_i - b \right) \right) \end{aligned} \tag{A.15.15}$$

This is the product of N independent multivariate normal distributions. Intuitively, this should lead to a final result in the form of a normal distribution, but this expression cannot be used as such, for two reasons. First, it is a product of densities, while what is required is a single density. Secondly, this is a distribution in β, while what is needed is the posterior distribution for b. Therefore, some additional work is required.

First, using A.2.2.5, it is possible to express the product of multivariate normal distributions (A.15.15) as a single multivariate normal distribution. Define:

$$\tilde{\beta} = \underbrace{\begin{pmatrix} \beta \\ \beta \\ \vdots \\ \beta \end{pmatrix}}_{h \times 1} \qquad \tilde{\Sigma}_b = I_N \otimes \Sigma_b = \underbrace{\begin{pmatrix} \Sigma_b & 0 & \cdots & 0 \\ 0 & \Sigma_b & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \Sigma_b \end{pmatrix}}_{h \times h} \qquad \tilde{b} = 1_N \otimes b = \underbrace{\begin{pmatrix} b \\ b \\ \vdots \\ b \end{pmatrix}}_{h \times 1} \tag{A.15.16}$$

Then $\tilde{\beta}$ follows a multivariate normal distribution with mean $\tilde{b}$ and covariance matrix $\tilde{\Sigma}_b$:

$$\tilde{\beta} \sim \ N \left(\tilde{b}, \tilde{\Sigma}_b \right) \tag{A.15.17}$$

Now, the trick consists in using the affine properties of the multivariate normal distribution. Define:

$$\begin{aligned} M &= N^{-1}\left(1_N^{\prime} \otimes I_q\right) \\ &= \underbrace{N^{-1}\begin{pmatrix} I_q & I_q & \cdots & I_q \end{pmatrix}}_{q \times h} \end{aligned} \tag{A.15.18}$$

Then, from A.2.2.6, $M\tilde{\beta}$ follows a multivariate normal distribution with mean $M\tilde{b}$ and covariance matrix $M\tilde{\Sigma}_b M^{\prime}$. Note that M acts as an averaging matrix. Indeed, from (A.15.16) and (A.15.18), one obtains:

$$M\tilde{\beta} = N^{-1}\begin{pmatrix} I_q & I_q & \cdots & I_q \end{pmatrix}\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_N \end{pmatrix} = N^{-1}\sum_{i=1}^{N}\beta_i = \beta_m \tag{A.15.19}$$

where β_m denotes the arithmetic mean of the β_is. Also, concerning $M\tilde{\Sigma}_b M^{\prime}$:

$$\begin{aligned} M\tilde{\Sigma}_b M^{\prime} &= N^{-1}\begin{pmatrix} I_q & I_q & \cdots & I_q \end{pmatrix}\begin{pmatrix} \Sigma_b & 0 & \cdots & 0 \\ 0 & \Sigma_b & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \Sigma_b \end{pmatrix} N^{-1}\begin{pmatrix} I_q \\ I_q \\ \vdots \\ I_q \end{pmatrix} \\ &= N^{-2}\begin{pmatrix} \Sigma_b & \Sigma_b & \cdots & \Sigma_b \end{pmatrix}\begin{pmatrix} I_q \\ I_q \\ \vdots \\ I_q \end{pmatrix} \\ &= N^{-2}N\Sigma_b \\ &= N^{-1}\Sigma_b \end{aligned}$$

Finally, it is straightforward to obtain $M\tilde{b} = b$. Therefore, one eventually concludes that:

$$\beta_m \sim \mathcal{N}\left(b, N^{-1}\Sigma_b\right) \tag{A.15.20}$$

From this, it is possible to rewrite (A.15.15) as:

$$\pi\left(b\,|y, \beta, \Sigma_b, \Sigma\right) \propto \exp\left(-\frac{1}{2}(\beta_m - b)^{\prime}\left(N^{-1}\Sigma_b\right)^{-1}(\beta_m - b)\right) \tag{A.15.21}$$

Then, notice that the term into brackets can be equivalently rewritten as:

$$(\beta_m - b)' \left(N^{-1}\Sigma_b\right)^{-1} (\beta_m - b) = (b - \beta_m)' \left(N^{-1}\Sigma_b\right)^{-1} (b - \beta_m)$$

Substituting in (A.15.21), it eventually rewrites:

$$\pi\left(b \,|y, \beta, \Sigma_b, \Sigma\right) \propto \exp\left(-\frac{1}{2}(b - \beta_m)' \left(N^{-1}\Sigma_b\right)^{-1} (b - \beta_m)\right) \tag{A.15.22}$$

This is the kernel of a multivariate normal distribution with mean β_m and covariance matrix $N^{-1}\Sigma_b$.

Obtain now the conditional posterior for Σ_b. Start from (6.6.23), and use (6.6.6) and (6.6.13) to obtain:

$$\begin{aligned}
\pi\left(\Sigma_b \,|y, \beta, b, \Sigma\right) &\propto \pi\left(\beta \,|b, \Sigma_b\right) \pi\left(\Sigma_b\right) \\
&\propto \prod_{i=1}^{N} |\Sigma_b|^{-1/2} \exp\left(-\frac{1}{2}(\beta_i - b)'(\Sigma_b)^{-1} (\beta_i - b)\right) \\
&\quad \times \lambda^{-\frac{s_0}{2}-1} \exp\left(-\frac{v_0}{2\lambda}\right)
\end{aligned}$$

Use (6.6.11) to substitute and rearrange:

$$\begin{aligned}
&\pi\left(\Sigma_b \,|y, \beta, b, \Sigma\right) \\
&\propto \prod_{i=1}^{N} |\Sigma_b|^{-1/2} \exp\left(-\frac{1}{2}(\beta_i - b)'(\Sigma_b)^{-1} (\beta_i - b)\right) \times \lambda_1^{-\frac{s_0+2}{2}} \exp\left(-\frac{v_0}{2\lambda_1}\right) \\
&= \left|\left(\lambda_1 \otimes I_q\right) \Omega_b\right|^{-N/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{N} \left\{(\beta_i - b)'\left[\left(\lambda_1 \otimes I_q\right) \Omega_b\right]^{-1} (\beta_i - b)\right\}\right) \times \lambda_1^{-\frac{s_0}{2}-1} \exp\left(-\frac{v_0}{2\lambda_1}\right) \\
&= |\lambda_1 \otimes I_q|^{-N/2} |\Omega_b|^{-N/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{N} \left\{(\beta_i - b)'\Omega_b^{-1} (\lambda_1 \otimes I_q)^{-1} (\beta_i - b)\right\}\right) \times \lambda_1^{-\frac{s_0}{2}-1} \exp\left(-\frac{v_0}{2\lambda_1}\right) \quad \text{(a.1.11)} \\
&\propto \lambda_1^{-\frac{qN}{2}} \exp\left(-\frac{1}{2} \sum_{i=1}^{N} \left\{(\beta_i - b)'\Omega_b^{-1}\lambda_1^{-1} (\beta_i - b)\right\}\right) \times \lambda_1^{-\frac{s_0}{2}-1} \exp\left(-\frac{v_0}{2\lambda_1}\right) \qquad A.1.2, A.1.4 \\
&= \lambda_1^{-\frac{h}{2}} \exp\left(-\frac{1}{2\lambda_1} \sum_{i=1}^{N} \left\{(\beta_i - b)'\Omega_b^{-1} (\beta_i - b)\right\}\right) \times \lambda_1^{-\frac{s_0}{2}-1} \exp\left(-\frac{v_0}{2\lambda_1}\right) \\
&= \lambda_1^{-\frac{h+s_0}{2}-1} \exp\left(-\frac{v_0 + \sum_{i=1}^{N} \left\{(\beta_i - b)'\Omega_b^{-1} (\beta_i - b)\right\}}{2\lambda_1}\right) \\
&= \lambda_1^{-\frac{\bar{s}}{2}-1} \exp\left(-\frac{\bar{v}}{2}\frac{1}{\lambda_1}\right)
\end{aligned} \tag{A.15.23}$$

with:

$$\bar{s} = h + s_0 \tag{A.15.24}$$

and:

$$\bar{v} = v_0 + \sum_{i=1}^{N} \left\{ (\beta_i - b)' \Omega_b^{-1} (\beta_i - b) \right\} \tag{A.15.25}$$

One concludes:

$$\pi (\Sigma_b | y, \beta, b, \Sigma) \propto \lambda_1^{-\frac{\bar{s}}{2}-1} \exp\left(-\frac{\bar{v}}{2} \frac{1}{\lambda_1} \right) \tag{A.15.26}$$

This the kernel of an inverse Gamma distribution with shape $\frac{\bar{s}}{2}$ and scale $\frac{\bar{v}}{2}$.

Obtain finally the conditional posteriors for the Σ_is. Start from (6.6.28), substitute (6.6.4) and (6.6.15) for unit i only, and rearrange:

$$\begin{aligned}
\pi (\Sigma_i | \Sigma_{-i}, y, \beta, b, \Sigma_b) &\propto \pi (y | \beta, \Sigma_i) \pi (\Sigma_i) \\
&= \left| \bar{\Sigma}_i \right|^{-1/2} \exp\left(-\frac{1}{2} (y_i - \bar{X}_i \beta_i)' \left(\bar{\Sigma}_i \right)^{-1} (y_i - \bar{X}_i \beta_i) \right) \times |\Sigma_i|^{-(n+1)/2} \\
&= |\Sigma_i|^{-T/2} \exp\left(-\frac{1}{2} tr \left[\Sigma_i^{-1} (Y_i - X_i B_i)' (Y_i - X_i B_i) \right] \right) \times |\Sigma_i|^{-(n+1)/2} \text{A.1.10} \\
&= |\Sigma_i|^{-(T+n+1)/2} \exp\left(-\frac{1}{2} tr \left[\Sigma_i^{-1} (Y_i - X_i B_i)' (Y_i - X_i B_i) \right] \right)
\end{aligned}$$

so:

$$\pi (\Sigma_i | \Sigma_{-i}, y, \beta, b, \Sigma_b) \propto |\Sigma_i|^{-(T+n+1)/2} \exp\left(-\frac{1}{2} tr \left[\Sigma_i^{-1} (Y_i - X_i B_i)' (Y_i - X_i B_i) \right] \right) \tag{A.15.27}$$

This is the kernel of an inverse Wishart distribution with scale $\tilde{S}_i = (Y_i - X_i B_i)' (Y_i - X_i B_i)$ and degrees of freedom T.

A.16 Derivations for the static factor model

Obtain the data likelihood. While it would be possible to formulate it as a joint density for all the periods, such a formulation would prevent convenient derivation of the conditional posteriors. Therefore, it is preferable to express it as the product of individual period densities. From (6.7.7),

the full density obtains as:

$$
\begin{aligned}
f(y\,|\theta,\tilde{\Sigma},\sigma) &= \prod_{t=1}^{T}\left\{(2\pi)^{-Nn/2}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}(y_t-\tilde{X}_t\theta)'\Sigma^{-1}(y_t-\tilde{X}_t\theta)\right)\right\}\\
&\propto \prod_{t=1}^{T}\left\{|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}(y_t-\tilde{X}_t\theta)'\Sigma^{-1}(y_t-\tilde{X}_t\theta)\right)\right\}\\
&= \prod_{t=1}^{T}\left\{\left|\sigma\tilde{\Sigma}\right|^{-1/2}\exp\left(-\frac{1}{2}(y_t-\tilde{X}_t\theta)'\left(\sigma\tilde{\Sigma}\right)^{-1}(y_t-\tilde{X}_t\theta)\right)\right\}\\
&= \prod_{t=1}^{T}\left\{(\sigma)^{-Nn/2}\left|\tilde{\Sigma}\right|^{-1/2}\exp\left(-\frac{1}{2}\sigma^{-1}(y_t-\tilde{X}_t\theta)'\tilde{\Sigma}^{-1}(y_t-\tilde{X}_t\theta)\right)\right\} \qquad A.1.14\\
&= (\sigma)^{-NnT/2}\left|\tilde{\Sigma}\right|^{-T/2}\prod_{t=1}^{T}\left\{\exp\left(-\frac{1}{2}\sigma^{-1}(y_t-\tilde{X}_t\theta)'\tilde{\Sigma}^{-1}(y_t-\tilde{X}_t\theta)\right)\right\} \qquad (A.16.1)
\end{aligned}
$$

Derive the full posterior distribution. Combine the likelihood function (6.7.29) with the priors (6.7.30), (6.7.31) and (6.7.32) to obtain the posterior as:

$$
\begin{aligned}
\pi(\theta,\tilde{\Sigma},\sigma\,|y) \propto{}& (\sigma)^{-NnT/2}\left|\tilde{\Sigma}\right|^{-T/2}\prod_{t=1}^{T}\left\{\exp\left(-\frac{1}{2}\sigma^{-1}(y_t-\tilde{X}_t\theta)'\tilde{\Sigma}^{-1}(y_t-\tilde{X}_t\theta)\right)\right\}\\
&\times\exp\left(-\frac{1}{2}(\theta-\theta_0)'\Theta_0^{-1}(\theta-\theta_0)\right)\times\left|\tilde{\Sigma}\right|^{-(Nn+1)/2}\times\sigma^{-\frac{\alpha_0}{2}-1}\exp\left(\frac{-\delta_0}{2\sigma}\right)\\
={}&\prod_{t=1}^{T}\left\{\exp\left(-\frac{1}{2}\sigma^{-1}(y_t-\tilde{X}_t\theta)'\tilde{\Sigma}^{-1}(y_t-\tilde{X}_t\theta)\right)\right\}\times\exp\left(\frac{-\delta_0}{2\sigma}\right)\\
&\times(\sigma)^{-(NnT+\alpha_0)/2-1}\times\left|\tilde{\Sigma}\right|^{-(T+Nn+1)/2}\times\exp\left(-\frac{1}{2}(\theta-\theta_0)'\Theta_0^{-1}(\theta-\theta_0)\right)
\end{aligned}
\qquad (A.16.2)
$$

Derive now the conditional distributions. Start with θ. Relegate to the proportionality constant any term not involving θ in (6.7.33):

$$
\begin{aligned}
\pi(\theta\,|y,\tilde{\Sigma},\sigma\,) &\propto \prod_{t=1}^{T}\left\{\exp\left(-\frac{1}{2}\sigma^{-1}(y_t-\tilde{X}_t\theta)'\tilde{\Sigma}^{-1}(y_t-\tilde{X}_t\theta)\right)\right\}\times\exp\left(-\frac{1}{2}(\theta-\theta_0)'\Theta_0^{-1}(\theta-\theta_0)\right)\\
&= \exp\left(-\frac{1}{2}\left\{\sum_{t=1}^{T}\sigma^{-1}(y_t-\tilde{X}_t\theta)'\tilde{\Sigma}^{-1}(y_t-\tilde{X}_t\theta)+(\theta-\theta_0)'\Theta_0^{-1}(\theta-\theta_0)\right\}\right)\\
&= \exp\left(-\frac{1}{2}\left\{\sum_{t=1}^{T}(y_t-\tilde{X}_t\theta)'\Sigma^{-1}(y_t-\tilde{X}_t\theta)+(\theta-\theta_0)'\Theta_0^{-1}(\theta-\theta_0)\right\}\right)
\end{aligned}
\qquad (A.16.3)
$$

Consider only the term in the curly brackets, develop and complete the squares:

$$
\begin{aligned}
&\sum_{t=1}^{T}(y_t-\tilde{X}_t\theta)'\Sigma^{-1}(y_t-\tilde{X}_t\theta)+(\theta-\theta_0)'\Theta_0^{-1}(\theta-\theta_0)\\
&=\sum_{t=1}^{T}\left(y_t'\Sigma^{-1}y_t+\theta'\tilde{X}_t'\Sigma^{-1}\tilde{X}_t\theta-2\theta'\tilde{X}_t'\Sigma^{-1}y_t\right)+\theta'\Theta_0^{-1}\theta+\theta_0'\Theta_0^{-1}\theta_0-2\theta'\Theta_0^{-1}\theta_0\\
&=\sum_{t=1}^{T}\left(y_t'\Sigma^{-1}y_t\right)+\theta'\left(\sum_{t=1}^{T}\tilde{X}_t'\Sigma^{-1}\tilde{X}_t\right)\theta-2\theta'\left(\sum_{t=1}^{T}\tilde{X}_t'\Sigma^{-1}y_t\right)+\theta'\Theta_0^{-1}\theta+\theta_0'\Theta_0^{-1}\theta_0-2\theta'\Theta_0^{-1}\theta_0\\
&=\sum_{t=1}^{T}\left(y_t'\Sigma^{-1}y_t\right)+\theta'\left(\sum_{t=1}^{T}\tilde{X}_t'\Sigma^{-1}\tilde{X}_t+\Theta_0^{-1}\right)\theta-2\theta'\left(\sum_{t=1}^{T}\tilde{X}_t'\Sigma^{-1}y_t+\Theta_0^{-1}\theta_0\right)+\theta_0'\Theta_0^{-1}\theta_0\\
&=\sum_{t=1}^{T}\left(y_t'\Sigma^{-1}y_t\right)+\theta'\left(\sum_{t=1}^{T}\tilde{X}_t'\Sigma^{-1}\tilde{X}_t+\Theta_0^{-1}\right)\theta-2\theta'\bar{\Theta}^{-1}\bar{\Theta}\left(\sum_{t=1}^{T}\tilde{X}_t'\Sigma^{-1}y_t+\Theta_0^{-1}\theta_0\right)\\
&\quad+\theta_0'\Theta_0^{-1}\theta_0+\bar{\theta}'\bar{\Theta}^{-1}\bar{\theta}-\bar{\theta}'\bar{\Theta}^{-1}\bar{\theta}
\end{aligned}
$$

Define:

$$\bar{\Theta}=\left(\sum_{t=1}^{T}\tilde{X}_t'\Sigma^{-1}\tilde{X}_t+\Theta_0^{-1}\right)^{-1} \tag{A.16.4}$$

and:

$$\bar{\theta}=\bar{\Theta}\left(\sum_{t=1}^{T}\tilde{X}_t'\Sigma^{-1}y_t+\Theta_0^{-1}\theta_0\right) \tag{A.16.5}$$

Then the expression rewrites:

$$
\begin{aligned}
&\sum_{t=1}^{T}\left(y_t'\Sigma^{-1}y_t\right)+\theta'\left(\sum_{t=1}^{T}\tilde{X}_t'\Sigma^{-1}\tilde{X}_t+\Theta_0^{-1}\right)\theta-2\theta'\bar{\Theta}^{-1}\bar{\Theta}\left(\sum_{t=1}^{T}\tilde{X}_t'\Sigma^{-1}y_t+\Theta_0^{-1}\theta_0\right)\\
&\quad+\theta_0'\Theta_0^{-1}\theta_0+\bar{\theta}'\bar{\Theta}^{-1}\bar{\theta}-\bar{\theta}'\bar{\Theta}^{-1}\bar{\theta}\\
&=\sum_{t=1}^{T}\left(y_t'\Sigma^{-1}y_t\right)+\theta'\bar{\Theta}^{-1}\theta-2\theta'\bar{\Theta}^{-1}\bar{\theta}+\theta_0'\Theta_0^{-1}\theta_0+\bar{\theta}'\bar{\Theta}^{-1}\bar{\theta}-\bar{\theta}'\bar{\Theta}^{-1}\bar{\theta}\\
&=\left(\theta'\bar{\Theta}^{-1}\theta+\bar{\theta}'\bar{\Theta}^{-1}\bar{\theta}-2\theta'\bar{\Theta}^{-1}\bar{\theta}\right)+\left(\sum_{t=1}^{T}\left(y_t'\Sigma^{-1}y_t\right)+\theta_0'\Theta_0^{-1}\theta_0-\bar{\theta}'\bar{\Theta}^{-1}\bar{\theta}\right)\\
&=(\theta-\bar{\theta})'\bar{\Theta}^{-1}(\theta-\bar{\theta})+\left(\sum_{t=1}^{T}\left(y_t'\Sigma^{-1}y_t\right)+\theta_0'\Theta_0^{-1}\theta_0-\bar{\theta}'\bar{\Theta}^{-1}\bar{\theta}\right)
\end{aligned}
$$

Substitute back in (A.16.3):

$$\begin{aligned}\pi(\theta\left|y,\tilde{\Sigma},\sigma\right.)&\propto\exp\left(-\frac{1}{2}\left\{(\theta-\bar{\theta})'\bar{\Theta}^{-1}(\theta-\bar{\theta})+\left(\sum_{t=1}^{T}\left(y_t'\Sigma^{-1}y_t\right)+\theta_0'\Theta_0^{-1}\theta_0-\bar{\theta}'\bar{\Theta}^{-1}\bar{\theta}\right)\right\}\right)\\&=\exp\left(-\frac{1}{2}(\theta-\bar{\theta})'\bar{\Theta}^{-1}(\theta-\bar{\theta})\right)\exp\left(-\frac{1}{2}\left(\sum_{t=1}^{T}\left(y_t'\Sigma^{-1}y_t\right)+\theta_0'\Theta_0^{-1}\theta_0-\bar{\theta}'\bar{\Theta}^{-1}\bar{\theta}\right)\right)\\&\propto\exp\left(-\frac{1}{2}(\theta-\bar{\theta})'\bar{\Theta}^{-1}(\theta-\bar{\theta})\right)\end{aligned}$$

Hence:

$$\pi(\theta\left|y,\sigma,\tilde{\Sigma}\right.)\propto\exp\left(-\frac{1}{2}(\theta-\bar{\theta})'\bar{\Theta}^{-1}(\theta-\bar{\theta})\right)\tag{A.16.6}$$

This is the kernel of a multivariate normal distribution $\pi(\theta\left|y,\sigma,\tilde{\Sigma}\right.)\sim\mathcal{N}(\bar{\theta},\bar{\Theta})$.

Now obtain the conditional posterior for $\tilde{\Sigma}$. Start from (6.7.33) and relegate to the proportionality constant any term not involving $\tilde{\Sigma}$:

$$\pi(\tilde{\Sigma}\left|y,\theta,\sigma\right.)\propto\prod_{t=1}^{T}\left\{\exp\left(-\frac{1}{2}\sigma^{-1}(y_t-\tilde{X}_t\theta)'\tilde{\Sigma}^{-1}(y_t-\tilde{X}_t\theta)\right)\right\}\times\left|\tilde{\Sigma}\right|^{-(T+Nn+1)/2}\tag{A.16.7}$$

Rearrange:

$$\begin{aligned}\pi(\tilde{\Sigma}\left|y,\theta,\sigma\right.)&\propto\prod_{t=1}^{T}\left\{\exp\left(-\frac{1}{2}\sigma^{-1}(y_t-\tilde{X}_t\theta)'\tilde{\Sigma}^{-1}(y_t-\tilde{X}_t\theta)\right)\right\}\times\left|\tilde{\Sigma}\right|^{-(T+Nn+1)/2}\\&=\left|\tilde{\Sigma}\right|^{-(T+Nn+1)/2}\exp\left(-\frac{1}{2}\sum_{t=1}^{T}\sigma^{-1}(y_t-\tilde{X}_t\theta)'\tilde{\Sigma}^{-1}(y_t-\tilde{X}_t\theta)\right)\\&=\left|\tilde{\Sigma}\right|^{-(T+Nn+1)/2}\exp\left(-\frac{1}{2}\sum_{t=1}^{T}tr\left\{\sigma^{-1}(y_t-\tilde{X}_t\theta)'\tilde{\Sigma}^{-1}(y_t-\tilde{X}_t\theta)\right\}\right)\\&=\left|\tilde{\Sigma}\right|^{-(T+Nn+1)/2}\exp\left(-\frac{1}{2}\sum_{t=1}^{T}tr\left\{\tilde{\Sigma}^{-1}(y_t-\tilde{X}_t\theta)\sigma^{-1}(y_t-\tilde{X}_t\theta)'\right\}\right)\quad\text{(a.1.7)}\\&=\left|\tilde{\Sigma}\right|^{-(T+Nn+1)/2}\exp\left(-\frac{1}{2}tr\left\{\sum_{t=1}^{T}\tilde{\Sigma}^{-1}(y_t-\tilde{X}_t\theta)\sigma^{-1}(y_t-\tilde{X}_t\theta)'\right\}\right)\qquad\text{(a.1.6)}\\&=\left|\tilde{\Sigma}\right|^{-(T+Nn+1)/2}\exp\left(-\frac{1}{2}tr\left\{\tilde{\Sigma}^{-1}\sum_{t=1}^{T}(y_t-\tilde{X}_t\theta)\sigma^{-1}(y_t-\tilde{X}_t\theta)'\right\}\right)\end{aligned}$$

Define:

$$\bar{S} = \sum_{t=1}^{T} (y_t - \tilde{X}_t\theta)\sigma^{-1}(y_t - \tilde{X}_t\theta)' \tag{A.16.8}$$

Then one eventually obtains:

$$\pi(\tilde{\Sigma}\,|y,\theta,\sigma\,) \propto \left|\tilde{\Sigma}\right|^{-(T+Nn+1)/2} \exp\left(-\frac{1}{2}tr\left\{\tilde{\Sigma}^{-1}\bar{S}\right\}\right) \tag{A.16.9}$$

This is the kernel of an inverse Wishart distribution $\pi(\tilde{\Sigma}\,|y,\theta,\sigma\,) \sim\ IW(\bar{S},T)$.

Finally, obtain the conditional posterior distribution for σ. Relegate to the proportionality constant any term not involving σ in (6.7.33):

$$\begin{aligned}\pi(\sigma\,\Big|y,\theta,\tilde{\Sigma}) &\propto \prod_{t=1}^{T}\left\{\exp\left(-\frac{1}{2}\sigma^{-1}(y_t - \tilde{X}_t\theta)'\tilde{\Sigma}^{-1}(y_t - \tilde{X}_t\theta)\right)\right\} \times \exp\left(\frac{-\delta_0}{2\sigma}\right) \times \sigma^{-(NnT+\alpha_0)/2-1} \\ &= \sigma^{-(NnT+\alpha_0)/2-1}\exp\left(-\frac{1}{2\sigma}\left\{\sum_{t=1}^{T}(y_t - \tilde{X}_t\theta)'\tilde{\Sigma}^{-1}(y_t - \tilde{X}_t\theta) + \delta_0\right\}\right)\end{aligned}$$

Define:

$$\bar{\alpha} = NnT + \alpha_0 \tag{A.16.10}$$

and:

$$\bar{\delta} = \left(\sum_{t=1}^{T}(y_t - \tilde{X}_t\theta)'\tilde{\Sigma}^{-1}(y_t - \tilde{X}_t\theta) + \delta_0\right) \tag{A.16.11}$$

Then the conditional posterior rewrites:

$$\pi(\sigma\,\Big|y,\theta,\tilde{\Sigma}) \propto (\sigma)^{-\frac{\bar{\alpha}}{2}-1}\exp\left(-\frac{\bar{\delta}}{2\sigma}\right) \tag{A.16.12}$$

This is the kernel of an inverse Gamma distribution: $\pi(\sigma\,\Big|y,\theta,\tilde{\Sigma}) \sim\ IG\left(\frac{\bar{\alpha}}{2},\frac{\bar{\delta}}{2}\right)$.

Note that it is possible to reformulate some of the formulas obtained for the conditional posterior distributions. Those reformulated formulas are simpler as they involve direct matrix products rather than large summations terms, and a a consequence they are also computationally faster, which matters when a very large number of replications is implemented. Start with $\bar{\Theta}$ and $\bar{\theta}$, respectively

defined by (A.16.4) and (A.16.5). The formula for $\bar{\Theta}$ can reformulate as:

$$\begin{aligned}
\bar{\Theta} &= \left(\sum_{t=1}^{T} \tilde{X}_t' \Sigma^{-1} \tilde{X}_t + \Theta_0^{-1}\right)^{-1} \\
&= \left(\begin{pmatrix} \tilde{X}_1' \Sigma^{-1} & \tilde{X}_2' \Sigma^{-1} & \cdots & \tilde{X}_T' \Sigma^{-1} \end{pmatrix} \begin{pmatrix} \tilde{X}_1 \\ \tilde{X}_2 \\ \vdots \\ \tilde{X}_T \end{pmatrix} + \Theta_0^{-1} \right)^{-1} \\
&= \left(\begin{pmatrix} \tilde{X}_1' & \tilde{X}_2' & \cdots & \tilde{X}_T' \end{pmatrix} \begin{pmatrix} \Sigma^{-1} & 0 & \cdots & 0 \\ 0 & \Sigma^{-1} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma^{-1} \end{pmatrix} \begin{pmatrix} \tilde{X}_1 \\ \tilde{X}_2 \\ \vdots \\ \tilde{X}_T \end{pmatrix} + \Theta_0^{-1} \right)^{-1} \\
&= \left(\tilde{X} I_\Sigma \tilde{X}' + \Theta_0^{-1}\right)^{-1}
\end{aligned} \tag{A.16.13}$$

with:

$$\tilde{X} = \underbrace{\begin{pmatrix} \tilde{X}_1' & \tilde{X}_2' & \cdots & \tilde{X}_T' \end{pmatrix}}_{d \times NnT} \qquad I_\Sigma = (I_T \otimes \Sigma^{-1}) = \underbrace{\begin{pmatrix} \Sigma^{-1} & 0 & \cdots & 0 \\ 0 & \Sigma^{-1} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma^{-1} \end{pmatrix}}_{NnT \times NnT} \tag{A.16.14}$$

Also:

$$\begin{aligned}
\bar{\theta} &= \bar{\Theta}\left(\sum_{t=1}^{T} \tilde{X}_t' \Sigma^{-1} y_t + \Theta_0^{-1}\theta_0\right) \\
&= \bar{\Theta}\left(\begin{pmatrix} \tilde{X}_1' \Sigma^{-1} & \tilde{X}_2' \Sigma^{-1} & \cdots & \tilde{X}_T' \Sigma^{-1} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix} + \Theta_0^{-1}\theta_0 \right) \\
&= \bar{\Theta}\left(\begin{pmatrix} \tilde{X}_1' & \tilde{X}_2' & \cdots & \tilde{X}_T' \end{pmatrix} \begin{pmatrix} \Sigma^{-1} & 0 & \cdots & 0 \\ 0 & \Sigma^{-1} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma^{-1} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix} + \Theta_0^{-1}\theta_0 \right) \\
&= \bar{\Theta}\left(\tilde{X} I_\Sigma y + \Theta_0^{-1}\theta_0\right)
\end{aligned} \tag{A.16.15}$$

with:

$$y = \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix}}_{NnT \times 1} \tag{A.16.16}$$

Similarly, reformulate the formula for $\bar{S}$, defined in (A.16.8):

$$\begin{aligned} \bar{S} &= \sum_{t=1}^{T} (y_t - \tilde{X}_t\theta)\sigma^{-1}(y_t - \tilde{X}_t\theta)' \\ &= \sigma^{-1} \sum_{t=1}^{T} (y_t - \tilde{X}_t\theta)(y_t - \tilde{X}_t\theta)' \\ &= \sigma^{-1} \begin{pmatrix} (y_1 - \tilde{X}_1\theta) & (y_2 - \tilde{X}_2\theta) & \cdots & (y_T - \tilde{X}_T\theta) \end{pmatrix} \begin{pmatrix} (y_1 - \tilde{X}_1\theta)' \\ (y_2 - \tilde{X}_2\theta)' \\ \vdots \\ (y_T - \tilde{X}_T\theta)' \end{pmatrix} \end{aligned} \tag{A.16.17}$$

Then notice that:

$$\begin{aligned} & \begin{pmatrix} (y_1 - \tilde{X}_1\theta) & (y_2 - \tilde{X}_2\theta) & \cdots & (y_T - \tilde{X}_T\theta) \end{pmatrix} \\ &= \begin{pmatrix} y_1 & y_2 & \cdots & y_T \end{pmatrix} - \begin{pmatrix} \tilde{X}_1\theta & \tilde{X}_2\theta & \cdots & \tilde{X}_T\theta \end{pmatrix} \\ &= \begin{pmatrix} y_1 & y_2 & \cdots & y_T \end{pmatrix} - \begin{pmatrix} \tilde{X}_1 & \tilde{X}_2 & \cdots & \tilde{X}_T \end{pmatrix} \begin{pmatrix} \theta & 0 & \cdots & 0 \\ 0 & \theta & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \theta \end{pmatrix} \\ &= Y - \ddot{X} I_\theta \end{aligned} \tag{A.16.18}$$

with:

$$Y = \underbrace{\begin{pmatrix} y_1 & y_2 & \cdots & y_T \end{pmatrix}}_{Nn \times T} \qquad \ddot{X} = \underbrace{\begin{pmatrix} \tilde{X}_1 & \tilde{X}_2 & \cdots & \tilde{X}_T \end{pmatrix}}_{Nn \times Td} \qquad I_\theta = (I_T \otimes \theta) = \underbrace{\begin{pmatrix} \theta & 0 & \cdots & 0 \\ 0 & \theta & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \theta \end{pmatrix}}_{Td \times T} \tag{A.16.19}$$

Then, (A.16.17) can rewrite as:

$$\bar{S} = \sigma^{-1}\begin{pmatrix}(y_1 - \tilde{X}_1\theta) & (y_2 - \tilde{X}_2\theta) & \cdots & (y_T - \tilde{X}_T\theta)\end{pmatrix}\begin{pmatrix}(y_1 - \tilde{X}_1\theta)' \\ (y_2 - \tilde{X}_2\theta)' \\ \vdots \\ (y_T - \tilde{X}_T\theta)'\end{pmatrix}$$
$$= \sigma^{-1}\left(Y - \ddot{X}I_\theta\right)\left(Y - \ddot{X}I_\theta\right)' \tag{A.16.20}$$

Similarly, it is possible to reformulate the equation for $\bar{\delta}$ defined in (A.16.11):

$$\begin{aligned}
\bar{\delta} &= \frac{1}{2}\left[\sum_{t=1}^{T}(y_t - \tilde{X}_t\theta)'\tilde{\Sigma}^{-1}(y_t - \tilde{X}_t\theta) + \delta_0\right] \\
&= \frac{1}{2}\left[\sum_{t=1}^{T} tr\left((y_t - \tilde{X}_t\theta)'\tilde{\Sigma}^{-1}(y_t - \tilde{X}_t\theta)\right) + \delta_0\right] \\
&= \frac{1}{2}\left[\sum_{t=1}^{T} tr\left((y_t - \tilde{X}_t\theta)(y_t - \tilde{X}_t\theta)'\tilde{\Sigma}^{-1}\right) + \delta_0\right] \qquad (A.1.7) \\
&= \frac{1}{2}\left[tr\left(\sum_{t=1}^{T}(y_t - \tilde{X}_t\theta)(y_t - \tilde{X}_t\theta)'\tilde{\Sigma}^{-1}\right) + \delta_0\right] \qquad (A.1.6) \\
&= \frac{1}{2}\left[tr\left(\left[\sum_{t=1}^{T}(y_t - \tilde{X}_t\theta)(y_t - \tilde{X}_t\theta)'\right]\tilde{\Sigma}^{-1}\right) + \delta_0\right] \\
&= \frac{1}{2}\left[tr\left(\begin{pmatrix}y_1 - \tilde{X}_1\theta & y_2 - \tilde{X}_2\theta & \cdots & y_T - \tilde{X}_T\theta\end{pmatrix}\begin{pmatrix}(y_1 - \tilde{X}_1\theta)' \\ (y_2 - \tilde{X}_2\theta)' \\ \vdots \\ (y_T - \tilde{X}_T\theta)'\end{pmatrix}\tilde{\Sigma}^{-1}\right) + \delta_0\right] \\
&= \frac{1}{2}\left[tr\left((Y - \ddot{X}I_\theta)(Y - \ddot{X}I_\theta)'\tilde{\Sigma}^{-1}\right) + \delta_0\right] \qquad (A.16.21)
\end{aligned}$$

where the last line obtains from (A.16.18).

A.17 Derivations for the dynamic factor approach

First derive the version of Bayes rule obtained from the hierarchical prior. Using basic rules of conditional probabilities, one obtains:

$$\begin{aligned}
\pi(\theta, b, \tilde{\Sigma}, \zeta, \varphi \,|y\,) &= \frac{\pi(\theta, b, \tilde{\Sigma}, \zeta, \varphi, y)}{\pi(y)} \\
&\propto \pi(\theta, b, \tilde{\Sigma}, \zeta, \varphi, y) \\
&= \frac{\pi(\theta, b, \tilde{\Sigma}, \zeta, \varphi, y)}{\pi(\theta, b, \tilde{\Sigma}, \zeta, \varphi)} \pi(\theta, b, \tilde{\Sigma}, \zeta, \varphi) \\
&= \pi(y \,\big|\, \theta, b, \tilde{\Sigma}, \zeta, \varphi) \pi(\theta, b, \tilde{\Sigma}, \zeta, \varphi) \\
&= \pi(y \,\big|\, \theta, b, \tilde{\Sigma}, \zeta, \varphi) \pi(\theta, b) \pi(\tilde{\Sigma}) \pi(\zeta, \varphi) \\
&= \pi(y \,\big|\, \theta, b, \tilde{\Sigma}, \zeta, \varphi) \frac{\pi(\theta, b)}{\pi(b)} \pi(b) \pi(\tilde{\Sigma}) \frac{\pi(\zeta, \varphi)}{\pi(\varphi)} \pi(\varphi) \\
&= \pi(y \,\big|\, \theta, b, \tilde{\Sigma}, \zeta, \varphi) \pi(\theta \,|b) \pi(b) \pi(\tilde{\Sigma}) \pi(\zeta \,|\varphi) \pi(\varphi) \\
&= \pi(y \,\big|\, \theta, \tilde{\Sigma}, \zeta) \pi(\theta \,|b) \pi(b) \pi(\tilde{\Sigma}) \pi(\zeta \,|\varphi) \pi(\varphi) \qquad \text{(A.17.1)}
\end{aligned}$$

where the last line obtains by noting that b and φ are of no relevance to compute the likelihood function once θ and ζ are known.

Then obtain the likelihood function. Given (6.8.9) and (6.8.19), it is given by:

$$
\begin{aligned}
f(y \,|\theta, \tilde{\Sigma}, \zeta) &= \prod_{t=1}^{T} f(y_t \,|\theta_t, \tilde{\Sigma}, \zeta_t) \\
&= \prod_{t=1}^{T} \left\{ (2\pi)^{-k/2} |\Sigma_t|^{-1/2} \exp\left(-\frac{1}{2}(y_t - \tilde{X}_t\theta_t)' \Sigma_t^{-1} (y_t - \tilde{X}_t\theta_t) \right) \right\} \\
&\propto \prod_{t=1}^{T} \left\{ |\Sigma_t|^{-1/2} \exp\left(-\frac{1}{2}(y_t - \tilde{X}_t\theta_t)' \Sigma_t^{-1} (y_t - \tilde{X}_t\theta_t) \right) \right\} \\
&= \prod_{t=1}^{T} \left\{ \left| \exp(\zeta_t)\tilde{\Sigma} \right|^{-1/2} \exp\left(-\frac{1}{2}(y_t - \tilde{X}_t\theta_t)' \left(\exp(\zeta_t)\tilde{\Sigma} \right)^{-1} (y_t - \tilde{X}_t\theta_t) \right) \right\} \\
&= \prod_{t=1}^{T} \left\{ \exp(\zeta_t)^{-Nn/2} \left| \tilde{\Sigma} \right|^{-1/2} \exp\left(-\frac{1}{2} \exp(\zeta_t)^{-1} (y_t - \tilde{X}_t\theta_t)' \tilde{\Sigma}^{-1} (y_t - \tilde{X}_t\theta_t) \right) \right\} \qquad A.1.14 \\
&= \prod_{t=1}^{T} \left\{ \exp\left(-\frac{Nn\zeta_t}{2} \right) \left| \tilde{\Sigma} \right|^{-1/2} \exp\left(-\frac{1}{2} \exp(-\zeta_t)(y_t - \tilde{X}_t\theta_t)' \tilde{\Sigma}^{-1} (y_t - \tilde{X}_t\theta_t) \right) \right\} \\
&= \left| \tilde{\Sigma} \right|^{-T/2} \prod_{t=1}^{T} \left\{ \exp\left(-\frac{Nn\zeta_t}{2} \right) \exp\left(-\frac{1}{2} \exp(-\zeta_t)(y_t - \tilde{X}_t\theta_t)' \tilde{\Sigma}^{-1} (y_t - \tilde{X}_t\theta_t) \right) \right\} \\
&= \left| \tilde{\Sigma} \right|^{-T/2} \prod_{t=1}^{T} \left\{ \exp\left(-\frac{Nn\zeta_t}{2} \right) \right\} \prod_{t=1}^{T} \left\{ \exp\left(-\frac{1}{2} \exp(-\zeta_t)(y_t - \tilde{X}_t\theta_t)' \tilde{\Sigma}^{-1} (y_t - \tilde{X}_t\theta_t) \right) \right\} \\
&= \left| \tilde{\Sigma} \right|^{-T/2} \exp\left(-\frac{Nn}{2} \sum_{t=1}^{T} \zeta_t \right) \exp\left(-\frac{1}{2} \sum_{t=1}^{T} \exp(-\zeta_t)(y_t - \tilde{X}_t\theta_t)' \tilde{\Sigma}^{-1} (y_t - \tilde{X}_t\theta_t) \right) \\
&= \left| \tilde{\Sigma} \right|^{-T/2} \exp\left(-\frac{1}{2} \sum_{t=1}^{T} \left\{ \exp(-\zeta_t)(y_t - \tilde{X}_t\theta_t)' \tilde{\Sigma}^{-1} (y_t - \tilde{X}_t\theta_t) + Nn\zeta_t \right\} \right) \qquad \text{(A.17.2)}
\end{aligned}
$$

Obtain first the conditional posterior for $\theta = \{\theta_t\}_{t=1}^{T}$. Use Bayes rule (6.8.44), and combine the prior (6.8.31) with the reformulated likelihood (6.8.48) to obtain:

$$
\begin{aligned}
\pi(\theta \,|y, b, \tilde{\Sigma}, \zeta, \varphi) &\propto f(y \,|\theta, \tilde{\Sigma}, \zeta) \pi(\theta \,|b) \\
&= |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(y - \tilde{X}\Theta)' \tilde{\Sigma}^{-1} (y - \tilde{X}\Theta) \right) \\
&\quad \times |B_0| \exp\left(-\frac{1}{2}(\Theta - \Theta_0)' B_0^{-1} (\Theta - \Theta_0) \right) \\
&\propto \exp\left(-\frac{1}{2}(y - \tilde{X}\Theta)' \Sigma^{-1} (y - \tilde{X}\Theta) \right) \exp\left(-\frac{1}{2}(\Theta - \Theta_0)' B_0^{-1} (\Theta - \Theta_0) \right) \\
&= \exp\left(-\frac{1}{2} \left\{ (y - \tilde{X}\Theta)' \Sigma^{-1} (y - \tilde{X}\Theta) + (\Theta - \Theta_0)' B_0^{-1} (\Theta - \Theta_0) \right\} \right) \qquad \text{(A.17.3)}
\end{aligned}
$$

Consider only the term in the curly brackets, develop and complete the squares:

$$\begin{aligned}
&(y-\tilde{X}\Theta)'\Sigma^{-1}(y-\tilde{X}\Theta)+(\Theta-\Theta_0)'B_0^{-1}(\Theta-\Theta_0)\\
&=y'\Sigma^{-1}y+\Theta'\tilde{X}'\Sigma^{-1}\tilde{X}\Theta-2\Theta'\tilde{X}'\Sigma^{-1}y+\Theta'B_0^{-1}\Theta+\Theta_0'B_0^{-1}\Theta_0-2\Theta B_0^{-1}\Theta_0\\
&=y'\Sigma^{-1}y+\Theta'\left(\tilde{X}'\Sigma^{-1}\tilde{X}+B_0^{-1}\right)\Theta-2\Theta'\left(\tilde{X}'\Sigma^{-1}y+B_0^{-1}\Theta_0\right)+\Theta_0'B_0^{-1}\Theta_0\\
&=y'\Sigma^{-1}y+\Theta'\left(\tilde{X}'\Sigma^{-1}\tilde{X}+B_0^{-1}\right)\Theta-2\Theta'\bar{B}^{-1}\bar{B}\left(\tilde{X}'\Sigma^{-1}y+B_0^{-1}\Theta_0\right)\\
&\quad+\Theta_0'B_0^{-1}\Theta_0+\bar{\Theta}'\bar{B}^{-1}\bar{\Theta}-\bar{\Theta}'\bar{B}^{-1}\bar{\Theta}
\end{aligned}$$

Define:

$$\bar{B}=\left(\tilde{X}'\Sigma^{-1}\tilde{X}+B_0^{-1}\right)^{-1} \tag{A.17.4}$$

and:

$$\bar{\Theta}=\bar{B}\left(\tilde{X}'\Sigma^{-1}y+B_0^{-1}\Theta_0\right) \tag{A.17.5}$$

Then the expression rewrites:

$$\begin{aligned}
&=y'\Sigma^{-1}y+\Theta'\bar{B}^{-1}\Theta-2\Theta'\bar{B}^{-1}\bar{\Theta}+\Theta_0'B_0^{-1}\Theta_0+\bar{\Theta}'\bar{B}^{-1}\bar{\Theta}-\bar{\Theta}'\bar{B}^{-1}\bar{\Theta}\\
&=\Theta'\bar{B}^{-1}\Theta+\bar{\Theta}'\bar{B}^{-1}\bar{\Theta}-2\Theta'\bar{B}^{-1}\bar{\Theta}+\Theta_0'B_0^{-1}\Theta_0-\bar{\Theta}'\bar{B}^{-1}\bar{\Theta}+y'\Sigma^{-1}y\\
&=(\Theta-\bar{\Theta})'\bar{B}^{-1}(\Theta-\bar{\Theta})+\Theta_0'B_0^{-1}\Theta_0-\bar{\Theta}'\bar{B}^{-1}\bar{\Theta}+y'\Sigma^{-1}y
\end{aligned}$$

Substitute back in (A.17.3):

$$\begin{aligned}
\pi(\theta\,|y,b,\Sigma,\tilde{\sigma},a)&\propto\exp\left(-\frac{1}{2}\left\{(\Theta-\bar{\Theta})'\bar{B}^{-1}(\Theta-\bar{\Theta})+\Theta_0'B_0^{-1}\Theta_0-\bar{\Theta}'\bar{B}^{-1}\bar{\Theta}+y'\Sigma^{-1}y\right\}\right)\\
&=\exp\left(-\frac{1}{2}(\Theta-\bar{\Theta})'\bar{B}^{-1}(\Theta-\bar{\Theta})\right)\exp\left(-\frac{1}{2}\left\{\Theta_0'B_0^{-1}\Theta_0-\bar{\Theta}'\bar{B}^{-1}\bar{\Theta}+y'\Sigma^{-1}y\right\}\right)\\
&\propto\exp\left(-\frac{1}{2}(\Theta-\bar{\Theta})'\bar{B}^{-1}(\Theta-\bar{\Theta})\right)
\end{aligned} \tag{A.17.6}$$

This is the kernel of a multivariate normal distribution: $\Theta\sim\mathcal{N}(\bar{\Theta},\bar{B})$.

Obtain the conditional posterior for $b=\{b_i\}_{i=1}^r$. First, noting that from (6.8.16) and (6.8.24), the priors for the θ_is and the b_is are independent across the r factors, one may rewrite:

$$\pi(\theta\,|b)=\prod_{i=1}^{r}\pi(\theta_i\,|b_i)\qquad\text{and}\qquad\pi(b)=\prod_{i=1}^{r}\pi(b_i)$$

Therefore, Bayes rule (6.8.41) can be rewritten as:

$$\pi(\theta, b, \tilde{\Sigma}, \zeta, \varphi \,|y) \propto f(y \,\Big|\theta, \tilde{\Sigma}, \sigma) \left(\prod_{i=1}^{r} \pi(\theta_i \,|b_i)\right) \left(\prod_{i=1}^{r} \pi(b_i)\right) \pi(\tilde{\Sigma})\pi(\zeta \,|\varphi)\pi(\varphi) \tag{A.17.7}$$

Then consider the conditional posterior for b_i. Consider the rewritten Bayes rule (A.17.7) and relegate to the proportionality constant any term not involving b_i:

$$\pi(b_i \,|y, \theta, \ b_{-i}, \tilde{\Sigma}, \sigma, a) \propto \pi(\theta_i \,|b_i)\pi(b_i) \tag{A.17.8}$$

To obtain $\pi(\theta_i \,|b_i)$, note that (6.8.23) implies:

$$\theta_{i,t} = \theta_{i,t-1} + \eta_{i,t} \qquad \text{with} \qquad \eta_{i,t} \sim \mathcal{N}(0, b_i I_{d_i}) \tag{A.17.9}$$

Hence:

$$\theta_{i,t} \,|\theta \sim \mathcal{N}(\theta_{i,t-1}, b_i I_{d_i})$$

So that:

$$\pi(\theta_i \,|b_i) \propto \left\{\prod_{t=1}^{T} |b_i I_{d_i}|^{-1/2} \exp\left(-\frac{1}{2}(\theta_{i,t} - \theta_{i,t-1})'(b_i I_{d_i})^{-1}(\theta_{i,t} - \theta_{i,t-1})\right)\right\} \tag{A.17.10}$$

Use (A.17.8) to combine (A.17.10) with (6.8.32) and obtain:

$$\begin{aligned}
\pi(b_i \,|y, \theta, \ b_{-i}, \tilde{\Sigma}, \zeta, \varphi) &\propto \pi(\theta_i \,|b_i)\pi(b_i) \\
&= \left\{\prod_{t=1}^{T} |b_i I_{d_i}|^{-1/2} \exp\left(-\frac{1}{2}(\theta_{i,t} - \theta_{i,t-1})'(b_i I_{d_i})^{-1}(\theta_{i,t} - \theta_{i,t-1})\right)\right\} \\
&\quad \times b_i^{-(a_0/2)-1} \exp\left(\frac{-b_0}{2b_i}\right) \\
&= \left\{\prod_{t=1}^{T} b_i^{-d_i/2} \exp\left(-\frac{1}{2b_i}(\theta_{i,t} - \theta_{i,t-1})'(\theta_{i,t} - \theta_{i,t-1})\right)\right\} \qquad A.1.14 \\
&\quad \times b_i^{-(a_0/2)-1} \exp\left(\frac{-b_0}{2b_i}\right) \\
&= b_i^{-Td_i/2} \left\{\prod_{t=1}^{T} \exp\left(-\frac{1}{2b_i}(\theta_{i,t} - \theta_{i,t-1})'(\theta_{i,t} - \theta_{i,t-1})\right)\right\} \\
&\quad \times b_i^{-(a_0/2)-1} \exp\left(\frac{-b_0}{2b_i}\right) \\
&= b_i^{-(Td_i+a_0)/2-1} \exp\left(-\frac{1}{2b_i}\left\{\sum_{t=1}^{T} (\theta_{i,t} - \theta_{i,t-1})'(\theta_{i,t} - \theta_{i,t-1}) + b_0\right\}\right)
\end{aligned}$$

Define:

$$\bar{a}_i = Td_i + a_0 \tag{A.17.11}$$

and:

$$\bar{b}_i = \sum_{t=1}^{T} (\theta_{i,t} - \theta_{i,t-1})'(\theta_{i,t} - \theta_{i,t-1}) + b_0 \tag{A.17.12}$$

Then the expression rewrites:

$$\pi(b_i \,|y, \theta,\ b_{-i}, \Sigma, \sigma, a) \propto b_i^{-\frac{\bar{a}_i}{2}-1} \exp\left(-\frac{\bar{b}_i}{2b_i}\right) \tag{A.17.13}$$

This is the kernel of an inverse Gamma distribution with shape $\frac{\bar{a}_i}{2}$ and scale $\frac{\bar{b}_i}{2}$.

Obtain the conditional posterior for $\tilde{\Sigma}$. Consider Bayes rule (6.8.41) and relegate to the proportionality constant any term not involving $\tilde{\Sigma}$ to obtain:

$$\pi(\tilde{\Sigma} \,|y, \theta, b,\ \zeta, \varphi) \propto f(y \,|\theta, \tilde{\Sigma}, \zeta)\pi(\tilde{\Sigma}) \tag{A.17.14}$$

Given (6.8.42) and (6.8.33), this gives:

$$\begin{aligned}
\pi(\tilde{\Sigma} \,|y, \theta, b,\ \zeta, \varphi) &\propto f(y \,|\theta, \tilde{\Sigma}, \zeta)\pi(\tilde{\Sigma}) \\
&= \left|\tilde{\Sigma}\right|^{-T/2} \exp\left(-\frac{1}{2}\sum_{t=1}^{T}\left\{\exp(-\zeta_t)(y_t - \tilde{X}_t\theta_t)'\tilde{\Sigma}^{-1}(y_t - \tilde{X}_t\theta_t) + Nn\zeta_t\right\}\right)\left|\tilde{\Sigma}\right|^{-(Nn+1)/2} \\
&= \left|\tilde{\Sigma}\right|^{-(T+Nn+1)/2} \exp\left(-\frac{1}{2}\sum_{t=1}^{T}\left\{\exp(-\zeta_t)(y_t - \tilde{X}_t\theta_t)'\tilde{\Sigma}^{-1}(y_t - \tilde{X}_t\theta_t) + Nn\zeta_t\right\}\right) \\
&\propto \left|\tilde{\Sigma}\right|^{-(T+Nn+1)/2} \exp\left(-\frac{1}{2}\sum_{t=1}^{T}\left\{\exp(-\zeta_t)(y_t - \tilde{X}_t\theta_t)'\tilde{\Sigma}^{-1}(y_t - \tilde{X}_t\theta_t)\right\}\right) \\
&= \left|\tilde{\Sigma}\right|^{-(T+Nn+1)/2} \exp\left(-\frac{1}{2}\sum_{t=1}^{T} tr\left\{\exp(-\zeta_t)(y_t - \tilde{X}_t\theta_t)'\tilde{\Sigma}^{-1}(y_t - \tilde{X}_t\theta_t)\right\}\right) \\
&= \left|\tilde{\Sigma}\right|^{-(T+Nn+1)/2} \exp\left(-\frac{1}{2}\sum_{t=1}^{T} tr\left\{\tilde{\Sigma}^{-1}(y_t - \tilde{X}_t\theta_t)\exp(-\zeta_t)(y_t - \tilde{X}_t\theta_t)'\right\}\right) \\
&= \left|\tilde{\Sigma}\right|^{-(T+Nn+1)/2} \exp\left(-\frac{1}{2} tr\left\{\sum_{t=1}^{T}\tilde{\Sigma}^{-1}(y_t - \tilde{X}_t\theta_t)\exp(-\zeta_t)(y_t - \tilde{X}_t\theta_t)'\right\}\right) \\
&= \left|\tilde{\Sigma}\right|^{-(T+Nn+1)/2} \exp\left(-\frac{1}{2} tr\left\{\tilde{\Sigma}^{-1}\sum_{t=1}^{T}(y_t - \tilde{X}_t\theta_t)\exp(-\zeta_t)(y_t - \tilde{X}_t\theta_t)'\right\}\right)
\end{aligned}$$

Define:

$$\bar{S} = \sum_{t=1}^{T} (y_t - \tilde{X}_t\theta_t)\exp(-\zeta_t)(y_t - \tilde{X}_t\theta_t)' \tag{A.17.15}$$

Then this rewrites:

$$\pi(\tilde{\Sigma}\,|y, \theta, b,\ \zeta, \varphi) \propto \left|\tilde{\Sigma}\right|^{-(T+Nn+1)/2} \exp\left(-\frac{1}{2} tr\left\{\tilde{\Sigma}^{-1}\bar{S}\right\}\right) \tag{A.17.16}$$

This is the kernel of an inverse Wishart distribution: $\pi(\tilde{\Sigma}\,|y, \theta, b,\ \zeta, \varphi) \sim\ IW(\bar{S}, T)$.

Obtain the conditional posterior for φ. Start from Bayes rule (6.8.41) and relegate to the proportionality constant any term not involving φ:

$$\pi(\theta, b, \tilde{\Sigma}, \zeta, \varphi\,|y) \propto \pi(\zeta\,|\varphi)\pi(\varphi) \tag{A.17.17}$$

Following, use (6.8.39) and (6.8.40) to obtain:

$$\begin{aligned}
\pi(\theta, b, \tilde{\Sigma}, \zeta, \varphi\,|y) &\propto \pi(\zeta\,|\varphi)\pi(\varphi) \\
&= |\Phi_0|^{-1/2} \exp\left(-\frac{1}{2} Z'\Phi_0^{-1} Z\right) \times \varphi^{-\alpha_0/2-1} \exp\left(\frac{-\delta_0}{2\varphi}\right) \\
&= \left|\varphi(G'G)^{-1}\right|^{-1/2} \exp\left(-\frac{1}{2} Z'\left(\varphi(G'G)^{-1}\right)^{-1} Z\right) \times \varphi^{-\alpha_0/2-1} \exp\left(\frac{-\delta_0}{2\varphi}\right) \\
&= \varphi^{-T/2}\left|(G'G)^{-1}\right|^{-1/2} \exp\left(-\frac{1}{2\varphi} Z'G'GZ\right) \times \varphi^{-\alpha_0/2-1} \exp\left(\frac{-\delta_0}{2\varphi}\right) \qquad \text{A.1.14} \\
&\propto \varphi^{-(T+\alpha_0)/2-1} \exp\left(-\frac{1}{\varphi}\left\{\frac{Z'G'GZ + \delta_0}{2}\right\}\right)
\end{aligned}$$

Define:

$$\bar{\alpha} = T + \alpha_0 \tag{A.17.18}$$

and:

$$\bar{\delta} = Z'G'GZ + \delta_0 \tag{A.17.19}$$

Then this rewrites:

$$\pi(\theta, b, \tilde{\Sigma}, \zeta, \varphi\,|y) \propto \varphi^{-\frac{\bar{\alpha}}{2}-1} \exp\left(-\frac{\bar{\delta}}{2\varphi}\right) \tag{A.17.20}$$

This is the kernel of an inverse Gamma distribution with shape $\frac{\bar{\alpha}}{2}$ and scale $\frac{\bar{\delta}}{2}$.

Finally, obtain the conditional posterior for $\zeta = \{\zeta_t\}_{t=1}^{T}$. Consider Bayes rule (6.8.41) and rele-

gate to the proportionality constant any term not involving ζ:

$$\pi(\zeta \left| y, \theta, b, \tilde{\Sigma}, \varphi \right.) \propto f(y \left| \theta, \tilde{\Sigma}, \zeta \right.) \pi(\zeta \left| \varphi \right.) \tag{A.17.21}$$

Following, combine the likelihood (6.8.42) with the prior (6.8.39) to obtain:

$$\begin{aligned}
\pi(\zeta \left| y, \theta, b, \tilde{\Sigma}, \varphi \right.) &\propto f(y \left| \theta, \tilde{\Sigma}, \zeta \right.) \pi(\zeta \left| \varphi \right.) \\
&= \left| \tilde{\Sigma} \right|^{-T/2} \exp\left(-\frac{1}{2} \sum_{t=1}^{T} \left\{ \exp(-\zeta_t)(y_t - \tilde{X}_t \theta_t)' \tilde{\Sigma}^{-1} (y_t - \tilde{X}_t \theta_t) + N n \zeta_t \right\} \right) \\
&\quad \times |\Phi_0|^{-1/2} \exp\left(-\frac{1}{2} Z' \Phi_0^{-1} Z \right) \\
&\propto \exp\left(-\frac{1}{2} \left[\sum_{t=1}^{T} \left\{ \exp(-\zeta_t)(y_t - \tilde{X}_t \theta_t)' \tilde{\Sigma}^{-1} (y_t - \tilde{X}_t \theta_t) + N n \zeta_t \right\} + Z' \Phi_0^{-1} Z \right] \right)
\end{aligned}$$

This is not a standard formula, so that a Metropolis-Hastings step is required to sample from the conditional posterior.

Obtain the value of the acceptance probability for the Metropolis-Hastings algorithm for the random walk kernel:

$$\begin{aligned}
&\alpha(Z^{(n-1)}, Z^{(n)}) \\
&= \frac{\pi(Z^{(n)})}{\pi(Z^{(n-1)})} \\
&= \frac{\exp\left(-\frac{1}{2} \left[\sum_{t=1}^{T} \left\{ \exp(-\zeta_t^{(n)})(y_t - \tilde{X}_t \theta_t)' \tilde{\Sigma}^{-1} (y_t - \tilde{X}_t \theta_t) + N n \zeta_t^{(n)} \right\} + (Z^{(n)})' \Phi_0^{-1} Z^{(n)} \right] \right)}{\exp\left(-\frac{1}{2} \left[\sum_{t=1}^{T} \left\{ \exp(-\zeta_t^{(n-1)})(y_t - \tilde{X}_t \theta_t)' \tilde{\Sigma}^{-1} (y_t - \tilde{X}_t \theta_t) + N n \zeta_t^{(n-1)} \right\} + (Z^{(n-1)})' \Phi_0^{-1} Z^{(n-1)} \right] \right)} \\
&= \exp\left(-\frac{1}{2} \sum_{t=1}^{T} (y_t - \tilde{X}_t \theta_t)' \tilde{\Sigma}^{-1} (y_t - \tilde{X}_t \theta_t) \left\{ \exp(-\zeta_t^{(n)}) - \exp(-\zeta_t^{(n-1)}) \right\} \right) \\
&\quad \times \exp\left(-\frac{Nn}{2} \sum_{t=1}^{T} \left\{ \zeta_t^{(n)} - \zeta_t^{(n-1)} \right\} \right) \\
&\quad \times \exp\left(-\frac{1}{2} \left\{ (Z^{(n)})' \Phi_0^{-1} Z^{(n)} - (Z^{(n-1)})' \Phi_0^{-1} Z^{(n-1)} \right\} \right)
\end{aligned} \tag{A.17.22}$$

Acknowledgements
We would like to thank Marta Banbura, Elena Bobeica, Fabio Canova, Matteo Ciccarelli, Marek Jarocinski, Michele Lenza, Carlos Montes-Galdon, Chiara Osbat, and Giorgio Primiceri for valuable input and advice as well as feedback received following presentations at various conferences.

Alistair Dieppe
European Central Bank, Frankfurt, Germany; email: alistair.dieppe@ecb.europa.eu
Romain Legrand
ESSEC Business School, Paris, France; email: romain.legrand@essec.edu

Björn van Roye
European Central Bank, Frankfurt, Germany; email: bjorn.van roye@ecb.europa.eu

Postal address	60640 Frankfurt am Main, Germany
Telephone	+49 69 1344 0
Website	www.ecb.europa.eu

ISSN	1725-2806 (online)
ISBN	978-92-899-2182-4
DOI	10.2866/292952
EU catalogue No	QB-AR-16-051-EN-N