國立臺灣大學理學院數學系

碩士論文

Department of Mathematics

College of Science

National Taiwan University

Master's Thesis

自迴歸隨機波動的向量自迴歸模型之馬蹄鐵估計

Horseshoe Estimates for Vector Autoregression Model with Log Normal Stochastic Volatility

王柏崴

Po-Wei Wang

指導教授: 楊鈞澔 博士

Advisor: Chun-Hao Yang, Ph.D.

中華民國 113 年 7 月

July, 2024

doi:10.6342/NTU202402563

國立臺灣大學碩士學位論文
口試委員會審定書
MASTER'S THESIS ACCEPTANCE CERTIFICATE
NATIONAL TAIWAN UNIVERSITY

自迴歸隨機波動的向量自迴歸模型之馬蹄鐵估計

Horseshoe Estimates for Vector Autoregression Model with Log Normal Stochastic Volatility

本論文係王柏崴（R10221006）在國立臺灣大學數學系完成之碩士學位論文，於民國 113 年 7 月 11 日承下列考試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Department of Mathematics on 11th July 2024 have examined a Master's Thesis entitled above presented by Po-Wei Wang (R10221006) candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

楊鈞澔 ______ 蔡宗翰 ______ 李志煌 ______

（指導教授 Advisor）

陳宏庭 ______ ______ ______

系（所、學位學程）主管 Director: ______

致謝

謝謝指導教授楊鈞澔教授這三年以來的指導，也謝謝教授體諒我讓我能夠在碩三上時先將我人生規劃中重要的一環的教育實習完成，並接著在碩三下時給予我許多的協助來讓我完成我的論文。感謝我的父母願意讓我自己決定自己未來走的路，給予我一個很好的機會在數學所中進行研究、充實自我。最後更要感謝這一路走來在數學系、數學所上遇到的所有教授以及同儕，謝謝教授們在這共七年來的時間帶我走進數學的世界中，也謝謝所有的同儕們在這七年來攜手並進，一同共學共好。

doi:10.6342/NTU202402563

doi:10.6342/NTU202402563

摘要

在這篇研究中，我們會給出一個包含向量自迴歸模型、自迴歸隨機波動、馬蹄鐵先驗分佈的模型，再使用變分貝氏方法估計模型中的變數服從的分佈，像是常態分佈或是逆伽瑪分佈這些比較常見以及期望值容易計算的分佈，最後再使用迭代的方式找到每個變數的估計值。

關鍵字：向量自迴歸模型、隨機波動、馬蹄鐵先驗分佈、變分貝氏方法、牛頓-拉弗森方法

doi:10.6342/NTU202402563

doi:10.6342/NTU202402563

Abstract

In this thesis, we propose an approach to find an estimate of the variables in a model that combines vector autoregression (VAR), log-normal autoregressive stochastic volatility (ARSV) and the horseshoe prior. Using variational Bayes (VB) method, we can show an approximated distribution to each variable such as normal or inverse gamma distribution which are well-known and the expectation is easy to be obtained, which allows us to use iteration to estimate each variable.

Keywords: vector autoregression, stochastic volatility, horseshoe prior, variational Bayesian, Newton-Raphson

doi:10.6342/NTU202402563

doi:10.6342/NTU202402563

Contents

doi:10.6342/NTU202402563

doi:10.6342/NTU202402563

Chapter 1 Introduction

1.1 Large Bayesian VAR Model

VAR, first introduced in Sims (1980), is a statistical model which is used to describe a relationship among the multiple quantities that change over time. It is a generalization of the well-known autoregression (AR) where AR is univariate but VAR is in vector terms which allows multivariate time series. VAR models are commonly used in macroeconomic (such as Stock and Watson (2001)), deep learning (such as Choi et al. (2021)), finance (such as Sharma (2016)) and monetary policy (such as Miranda-Agrippino and Ricco (2021)) to describe the relation between the value that is going to be predicted and the observed values.

The large Bayesian VAR was introduced in Bańbura et al. (2010) for characterizing a large number of macroeconomic and financial variables. Since the size of the VARs typically used in empirical applications ranges from three to about ten variables and this potentially creates an omitted variable bias with adverse consequences both for structural analysis and for forecasting. There have been a large amount of research dealing with the topic of Large Bayesian VAR such as Carriero et al. (2012), Carriero et al. (2019), Kalli and Griffin (2018), Gefang (2014), Cuaresma et al. (2016), etc. Most of these papers used computation methods such as Markov Chain Monte Carlo (MCMC), and some others used natural conjugate priors to obtain an analytical result instead of using MCMC. The reason that some of them did not use MCMC as the computation method is because the number of the variables is too large resulting in the computation time might be too long. But there is still a limitation in using the natural conjugate priors while the distribution doesn't seem to have one such as the normal log-normal distribution while solving the case for log-

doi:10.6342/NTU202402563

normal SV in this thesis. So we are going to use another way, which is called variational Bayesian (VB), to deal with a VAR case with hierarchical shrinkage prior and multivariate stochastic volatility, which is the main topic in this thesis. The VB method will be further introduced in Section 2.3.

1.2 GARCH Versus SV

The VAR introduced in Section 1.1 has a constant volatility over time, but in economic applications we know that volatility usually changes with respect to time, so it might not make sense if the volatility is kept to be constant without a relation between time. To deal with this condition, there are two famous approaches to solve this problem: Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) and Stochastic Volatility (SV).

The Generalized AutoRegressive Conditional Heteroskedasticity model with orders q and s, GARCH(q, s), was introduced in Bollerslev (1986) to allow for past conditional variances in the current conditional. It differs from the usually used AutoRegressive Conditional Heteroskedasticity model with order q, ARCH(q), which assumes the time series of observations $\{\epsilon_t\}$ satisfies

$$\epsilon_t = \sqrt{h_t}\eta_t \tag{1.1}$$

$$h_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 \tag{1.2}$$

where $\{\eta_t\}$ is a sequence of independent and identically distributed variables with zero mean and unit variance, and $\{h_t\}$ is a varying variance process. Also, to ensure that the conditional variance is positive, α_0 is assumed to be positive and α_i is nonnegative for $i = 1, \cdots, q$. GARCH(q, s) assumes $\{\epsilon_t\}$ by replacing (1.2) with

$$h_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{s} \beta_j h_{t-j} \tag{1.3}$$

here the vector $\theta = (\alpha_0, (\alpha_i)_{1\le i\le q}, (\beta_j)_{1\le j\le s}) \in \Theta$ of parameters is to be estimated, the assumption of α_0 and α_i are the same as ARCH, α_0 is positive and α_i is nonnegative for

doi:10.6342/NTU202402563

$i = 1, \cdots, q$, and β_j is nonnegative for $j = 1, \cdots, s$, also the condition that $\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{s} \beta_j < 1$ is required to ensure that h_t is positive.

The Stochastic Volatility (SV) Model also allows for past conditional variances in the current conditional. The log-normal ARSV model with order p, which we denote as ARSV(p), is given by

$$\begin{aligned} \epsilon_t &= \sqrt{h_t}\eta_t \\ \log h_t &= \gamma_0 + \sum_{i=1}^{p} \gamma_i \log h_{t-i} + \sigma\epsilon_t^{h_t} \end{aligned} \tag{1.4}$$

where $\{\eta_t\}, \{\epsilon_t^{h_t}\}$ is a sequence of independent and identically distributed variables with zero mean and unit variance, and the vector $\theta = (\gamma_0, \gamma_i, \sigma) \in \Theta$ of parameters is to be estimated. Here σ is assumed to be positive in order to make sure the variance makes sense.

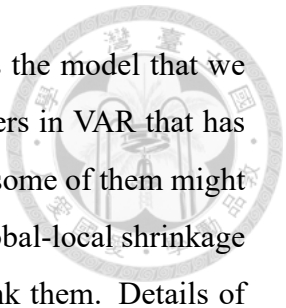
Both GARCH and ARSV are modeling the time varying volatility. Here we are going to talk about several differences between these two models and some comparison of GARCH and ARSV. We will analyze the difference and comparison according to some papers.

The main difference between GARCH and ARSV is that in GARCH, the h_t is the variance and the autoregressive term is linear, whereas in ARSV we assume log-normal autoregressive term, which means there is linear relation of $\log h_t$ in terms of $(\log h_{t-i})_{1 \le i \le p}$. This difference can be seen in (1.3) and (1.4), and it makes difference in the assumption of parameters that we need α_0 to be positive in GARCH in order to make sure that the variance will be positive, but the same assumption is not needed in ARSV.

1.3 Goal

In Section 1.1, we introduced that in macroeconomics, time series with time-varying variance is often used as a model. In this thesis, we focus on how to find out the approximated marginal density of each parameter, where the parameter are in the model that

doi:10.6342/NTU202402563



combines VAR, log-normal ARSV and the horseshoe prior, which is the model that we are interested in. Since we knew that there might be lots of parameters in VAR that has to be estimated, while not all of them are important, or we'll say that some of them might have large affect while some of them not. So we are going to use a global-local shrinkage prior to make the important ones to be remained and otherwise shrink them. Details of the model assumptions and the introduction of the global-local shrinkage prior, horseshoe prior, that we are going to use in thesis will be introduced in Chapter 2.

doi:10.6342/NTU202402563

Chapter 2 Theoretical Preliminary

In Section 2.1, we introduce a commonly used global-local shrinkage prior called Horseshoe prior. We'll first talk a little about the origin of Horseshoe prior, then give an alternative representation that will be used in this thesis as our main shrinkage prior of the large Bayesian VAR model.

In Section 2.2, we show how we set up the model that is the main topic of this thesis, which is a VAR model with stochastic volatility and Horseshoe prior. We will give the formula of our model and set up all priors for each parameter, so that we can get the approximated distribution and hence estimate each of the parameters.

In Section 2.3, we talk about the main method that how we will use the main method VB to find the approximated distribution for all the parameters in our model so that we may find the estimators for the model. It contains the introduction about a famous VB method which is called Kullback-Leibler (KL) divergence and a common way to deal with KL divergence which is called the evidence lower bound (ELBO).

2.1 Horseshoe prior

The Horseshoe prior was introduced in Carvalho et al. (2010), which is one of the global-local shrinkage prior. Assume that $\mathbf{y} \sim N(\theta, \sigma^2 I)$, a global-local shrinkage prior is the case that the assumption has hierarchical prior $\theta_i \sim N(0, \lambda_i^2 \tau^2)$, where θ_i denotes the i-th component of the vector θ. Here τ is a hyperparameter that globally controls all the θ_i, while λ_i is a hyperparameter that locally controls only θ_i to decide whether it has to be remained or be shrunk.

doi:10.6342/NTU202402563

There are some commonly used global-local shrinkage prior which are determined by the hierarchical prior for λ_i, the results can be summarized as a table such as:

prior distribution of λ_i	Name of marginal prior for θ_i
$\lambda_i^2 \sim \text{Exp}(2)$	Laplacian
$\lambda_i^2 \sim IG(\alpha, \beta)$	Student-t
$\lambda_i \sim C^+(0,1)$	Horseshoe
$p(\lambda_i) \propto \lambda_i(1+\lambda_i^2)^{1/2}$	Strawderman-Berger

Table 2.1: Global-Local Shrinkage Priors (Carvalho et al., 2009)

Next, in the expectation of θ_i, we first assume that $\sigma^2 = 1$ and denote $\dfrac{1}{1+\lambda_i^2\tau^2}$ as κ_i, then it can be shown that:

$$\mathbb{E}(\theta_i|y_i, \lambda_i^2, \tau^2) = \left(\frac{\lambda_i^2\tau^2}{1+\lambda_i^2\tau^2}\right) y_i + \left(\frac{1}{1+\lambda_i^2\tau^2}\right) 0 = (1-\kappa_i)y_i$$

By the definition of $\kappa_i = \dfrac{1}{1+\lambda_i^2\tau^2}$, we may observe that for $\kappa_i \in [0,1]$, $\kappa_i = 0$ means no shrinkage and $\kappa_i = 1$ means total shrinkage to zero for θ_i. Also, by the definition of κ_i and the hierarchical prior of λ_i in the above Table 2.1, we may find out the density of κ_i and drawn as graph shown as Figure 2.1 (Carvalho et al., 2009).

Note that Laplacian shown in the figure has the feature that the density of κ_i at 0 is null, which means that most of the coefficients will be shrunk.

The Student-t prior and the Strawderman-Berger prior are both unbounded near $\kappa_i = 0$, reflecting their heavy tails, and represents that they are both good at remaining some coefficients from being shrunk. But both of these two priors are bounded near $\kappa_i = 1$, limiting these priors in their ability to make the noise in the model to be shrunk to zero.

As for the Horseshoe prior, we observed that the density of κ_i is higher in $\kappa_i = 0$ and $\kappa_i = 1$, which means that most of the coefficients, may result in aggressive shrinkage of small coefficients, which means the data that has small effect, and virtually no shrinkage of sufficiently large coefficients, which means the data that has large effect. This is because of the particular choice of the heavy-tailed half-Cauchy prior distribution over the global and local hyperparameters. This is in contrast to the well-known Laplacian where the shrinkage effect is uniform across all coefficients. Therefore, we will use the Horseshoe

doi:10.6342/NTU202402563

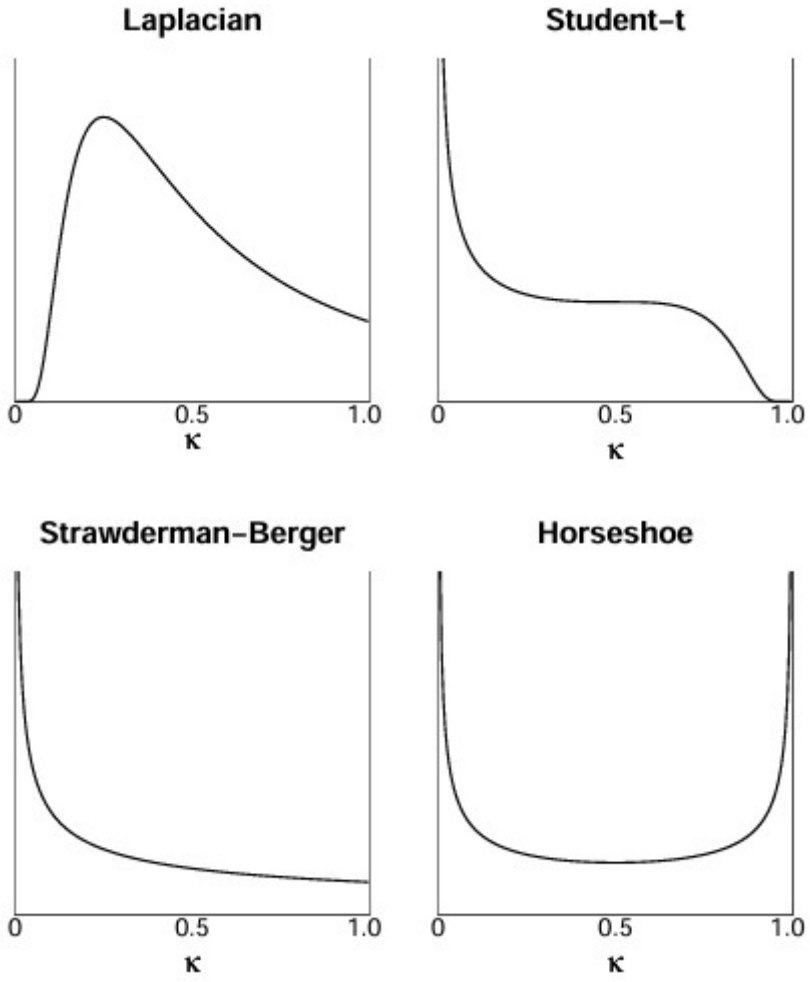


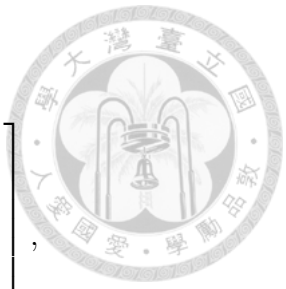
Figure 2.1: Densities for the shrinkage weights $\kappa_i \in [0, 1]$

prior as the global-local shrinkage prior in this thesis.

According to the definition of the Horseshoe prior above, we are going to use half-Cauchy distribution as the prior of λ_i. However, Makalic and Schmidt (2015) provided an alternative sampling scheme for all model parameters based on auxiliary varaiables that leads to conjugate conditional posterior distributions for all parameters, which changes the assumption of half-Cauchy into several Inverse Gamma distribution that is easier to be computed since we know that Inverse Gamma is a conjugate prior for the variance of

doi:10.6342/NTU202402563

normal distribution. Here the resulting alternative prior is shown as:



$$\theta \sim N(0, V),\ V = \begin{bmatrix} \lambda_0\tau & 0 & \cdots & 0 \\ 0 & \lambda_1\tau & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n\tau \end{bmatrix},$$

$$\lambda_j | v_j \sim IG(\frac{1}{2}, \frac{1}{v_j}),$$

$$\tau | \xi \sim IG(\frac{1}{2}, \frac{1}{\xi}),$$

$$v_1, \cdots, v_n, \xi \sim IG(\frac{1}{2}, 1),$$

here $\lambda_i\tau$ denotes the variance instead of standard deviation, and IG denotes the Inverse-Gamma distribution with probability density function

$$p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\frac{\beta}{x}},$$

and this is the term that will be mainly used in this thesis as the Horseshoe prior.

2.2 VAR with Stochastic Volatility and Horseshoe prior

In this thesis, we are going to introduce a model that integrates the concepts from Gefang et al. (2023), Makalic and Schmidt (2015), and Chan and Yu (2022). We will work with a VAR model as introduced in Gefang et al. (2023), as for the parameters in the model, we will use Horseshoe prior which is the alternative term introduced in Makalic and Schmidt (2015) as a shrinkage prior and finally deal with the log-normal ARSV model as Chan and Yu (2022). The detail will be given below.

The VAR(p) model we are using in this paper is as following:

$$\mathbf{B}_0\mathbf{y}_t = \mathbf{b} + \mathbf{B}_1\mathbf{y}_{t-1} + \cdots + \mathbf{B}_p\mathbf{y}_{t-p} + \epsilon_t^y, \epsilon_t^y \sim N(\mathbf{0}, \Sigma_t) \tag{2.1}$$

where $\Sigma_t = \text{diag}(e^{h_{1,t}}, \cdots, e^{h_{n,t}})$ is diagonal and $\mathbf{B}_0$ is a $n \times n$ lower triangular matrix with ones on the main diagonal. And each log-volatility $h_{i,t}$ evolves as an independent

doi:10.6342/NTU202402563

random walk ARSV(q), which is as we introduced in (1.4) and is now written as:

$$h_{i,t} = \sum_{m=1}^{q} h_{i,t-m} + \epsilon_{i,t}^{h},\ \epsilon_{i,t}^{h} \sim N(0, \sigma_{h,i}^2) \tag{2.2}$$

for $t = 1, \cdots, T$, here we use the condition that $\gamma_0 = 0$ and $\gamma_m = 1$ for $m = 1, \cdots, q$, and the initial condition $h_{i,0}$ is treated as an unknown parameter.

Let b_i denote the i-th element of $\mathbf{b}$ and let $\mathbf{b}_{j,i}$ denote the i-th row of $\mathbf{B}_j$. Then we may define $\beta_i = (b_i, \mathbf{b}_{1,i}, \cdots, \mathbf{b}_{p,i})'$ and α_i to be the i-th row of $\mathbf{B}_0$. By summarizing above, the i-th equation of the system in (2.1) can be rewritten as:

$$y_{i,t} = \tilde{\mathbf{w}}_{i,t}\alpha_{\mathbf{i}} + \tilde{\mathbf{x}}_t\beta_{\mathbf{i}} + \epsilon_{i,t}^{y},\ \epsilon_{i,t}^{y} \sim N(0, e^{h_{i,t}}) \tag{2.3}$$

where $\tilde{\mathbf{w}}_{i,t} = (-y_{1,t}, \cdots, -y_{i-1,t})$ contains the appropriate contemporaneous elements of $\mathbf{y}_t$, and $\tilde{\mathbf{x}}_t = (1, \mathbf{y}_{t-1}', \cdots, \mathbf{y}_{t-p}')$.

Next, let $\mathbf{x}_{i,t} = (\tilde{\mathbf{w}}_{i,t}, \tilde{\mathbf{x}}_t)$ and $\theta_i = (\alpha_i', \beta_i')'$, we can simplify (2.3) as:

$$y_{i,t} = \mathbf{x}_{i,t}\theta_i + \epsilon_{i,t}^{y},\ \epsilon_{i,t}^{y} \sim N(0, e^{h_{i,t}}) \tag{2.4}$$

which is now can be seen as an autoregressive model AR(k_i), where $k_i = np + i$.

Here we initially set the prior for the AR model:

$$\begin{aligned}
\sigma_{h,i}^2 &\sim IG(c_i, d_i) \\
h_{i,0} &\sim N(0, 1) \\
\theta_i \sim N(0, V_i),\ V_i &= \begin{bmatrix} \lambda_{i,1}\tau_i & 0 & \dots & 0 \\ 0 & \lambda_{i,2}\tau_i & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_{i,k_i}\tau_i \end{bmatrix} \\
\lambda_{i,j} &\sim IG(\frac{1}{2}, \frac{1}{v_{i,j}}),\ 1 \le j \le k_i \\
\tau_i &\sim IG(\frac{1}{2}, \frac{1}{\xi_i}) \\
v_{i,j}, \xi_i &\sim IG(\frac{1}{2}, 1),\ 1 \le j \le k_i
\end{aligned}$$

doi:10.6342/NTU202402563

The prior of θ_i is the alternating term of Horseshoe prior introduced in the bottom of Section 2.1.

Next, by the ARSV assumption with order q and some computation, one can show that for $1 \leq l \leq q$:

$$h_{i,l} = 2^{l-1} h_{i,0} + \sum_{k=1}^{l-2} 2^{l-k-1} \epsilon_{i,k}^{h} + \epsilon_{i,l-1}^{h} + \epsilon_{i,l}^{h} \tag{2.5}$$

Now, let $\mathbf{u}_{i,t} = (h_{i,t-q+1}, \cdots, h_{i,t})'$, we will get the i-th equation of our final VAR model we are going to work with in this thesis:

$$y_{i,t} = \mathbf{x}_{i,t}\theta_i + \epsilon_{i,t}^{y},\ \epsilon_{i,t}^{y} \sim N(0, e^{\mathbf{u}_{i,t}{}'\phi}), \tag{2.6}$$

$$\mathbf{u}_{i,t} = A\mathbf{u}_{i,t-1} + \epsilon_{i,t}^{u},\ \epsilon_{i,t}^{u} \sim N(0, \Sigma_{i,t}^{u}), \tag{2.7}$$

Where

$$\phi = (0, \cdots, 0, 1)',$$

$$A = \left[\begin{array}{c|ccc} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \hline 1 & 1 & 1 & \cdots & 1 \end{array}\right],$$

$$\Sigma_{i,t}^{u} = \left[\begin{array}{ccc|c} 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \\ \hline 0 & \cdots & 0 & \sigma_{h,i}^{2} \end{array}\right].$$

Then we can change the prior of $h_{i,0}$ into the one of $u_{i,q}$ by substitution, which is set to be the new initial case in our model by (2.5):

$$\mathbf{u}_{i,q} | h_{i,0} \sim N(h_{i,0}\psi, \sigma_{h,i}^{2}\Sigma_{i,q}) \tag{2.8}$$

doi:10.6342/NTU202402563

where

$$\psi = (2^0, 2^1, \cdots, 2^{q-1})'$$

$$\Sigma_{i,q}(r,s) = \text{Cov}(h_{i,r}, h_{i,s}) = \begin{cases} \frac{2^{r+s-2}+5\times 2^{\max(r,s)-\min(r,s)}}{3}, & \text{if } r \neq s \\ \frac{2^{2r-2}+2}{3}, & \text{if } r = s \end{cases}$$

Since we know that $\mathbb{E}(h_{i,0}) = 0$, by the result of substitution into $\mathbf{u}_{i,q}$, we may obtain that:

$$\mathbf{u}_{i,q}|h_{i,0} \sim N(0, \sigma_{h,i}^2 \Sigma_{i,q})$$

Hence the model and the corresponding prior that we are going to work with can be summarized as below:

$$y_{i,t} = \mathbf{x}_{i,t}\theta_i + \epsilon_{i,t}^y,\ \epsilon_{i,t}^y \sim N(0, e^{\mathbf{u}_{i,t}{}'\phi})$$
$$\mathbf{u}_{i,t} = A\mathbf{u}_{i,t-1} + \epsilon_{i,t}^u,\ \epsilon_{i,t}^u \sim N(0, \Sigma_{i,t}^u)$$

where

$$\begin{aligned} \mathbf{u}_{i,q}|h_{i,0} &\sim N(0, \sigma_{h,i}^2 \Sigma_{i,q}) \\ \sigma_{h,i}^2 &\sim IG(c_i, d_i) \\ \theta_i \sim N(0, V_i),\ V_i &= \begin{bmatrix} \lambda_{i,1}\tau_i & 0 & \dots & 0 \\ 0 & \lambda_{i,2}\tau_i & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_{i,k_i}\tau_i \end{bmatrix} \\ \lambda_{i,j} &\sim IG(\frac{1}{2}, \frac{1}{v_{i,j}}),\ 1 \leq j \leq k_i \\ \tau_i &\sim IG(\frac{1}{2}, \frac{1}{\xi_i}) \\ v_{i,j}, \xi_i &\sim IG(\frac{1}{2}, 1),\ 1 \leq j \leq k_i \end{aligned}$$

Next, we are going to introduce our main method for obtaining the variables in ap-

doi:10.6342/NTU202402563

proximated optimal density for each of our parameters, the VB method. Different from Prüser (2021) which gives the conditional density to each parameter and use Gipps sampling to estimate each parameter, we'll provide the approximated marginal density and use the iterative algorithm to estimate each of the parameters.

2.3 Variational Bayesian: KL and ELBO

VB is a way to find out the closest analytical approximation to the posterior probability of the variables that are not observed, which is the so-called latent variable, so the question is that how should we measure the closeness of two distributions, with one being the density we got analytically and the other being the approximation of the analytical one.

Divergence is the way we measure the closeness between two distributions, which is defined in Amari (2016). A function $D : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ is called a divergence if it satisfies

(i) $D(P\|Q) \geq 0$ for all $P, Q \in \mathcal{M}$,

(ii) $D(P\|Q) = 0$ if and only if $P = Q$, and

(iii) When P and Q are sufficiently close, by denoting their coordinates by δ_P and $\delta_Q = \delta_P + d\delta$, the Taylor expansion of D is written as

$$D(P\|Q) = \frac{1}{2}\sum g_{ij}(\delta_P)d\delta_i d\delta_j + O(|d\delta|^3),$$

and matrix $\mathbf{G} = (g_{ij})$ is positive-definite, depending on δ_P.

The way we measure the closeness of the two distribution is by applying the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951). The KL divergence is defined to be:

$$D_{KL}(q||p) = \int_\theta q(\theta)\log\frac{q(\theta)}{p(\theta|y)} = \mathbb{E}_q\left[\log\frac{q(\theta)}{p(\theta|y)}\right] \tag{2.9}$$

doi:10.6342/NTU202402563

in the definition of KL divergence we may observe that here $D_{KL}(q||p) \neq D_{KL}(p||q)$, which differs from the definition of "distance" that it has to be symmetric.

Next, we are going to introduce the evidence lower bound (ELBO), which is derived by the Jensen's inequality to the log probability of the observations:

$$\begin{aligned}
\log p(y) &= \log \int_{\theta} p(y, \theta) \\
&= \log \int_{\theta} p(y, \theta) \frac{q(\theta)}{q(\theta)} \\
&= \log \left(\mathbb{E}_q \left[\frac{p(y, \theta)}{q(\theta)} \right] \right) \\
&\geq \mathbb{E}_q \left[\log p(y, \theta) \right] - \mathbb{E}_q \left[\log q(\theta) \right] \qquad (2.10)
\end{aligned}$$

ELBO is defined to be as the (2.10), which is $\mathbb{E}_q \left[\log p(y, \theta) \right] - \mathbb{E}_q \left[\log q(\theta) \right]$. It can be shown that minimizing the KL divergence is equivalent to maximizing the ELBO. Using the definition of KL divergence above we may obtain that:

$$\begin{aligned}
D_{KL}(q||p) &= \mathbb{E}_q \left[\log \frac{q(\theta)}{p(\theta|y)} \right] \\
&= \mathbb{E}_q[\log q(\theta)] - \mathbb{E}_q[\log p(\theta|y)] \\
&= \mathbb{E}_q[\log q(\theta)] - \mathbb{E}_q[\log p(\theta, y)] + \log p(y) \\
&= - \left(\mathbb{E}_q \left[\log p(y, \theta) \right] - \mathbb{E}_q \left[\log q(\theta) \right] \right) + \log p(y)
\end{aligned}$$

Here we may observe that KL divergence to the posterior is equal to the negative ELBO plus a constant. More specifically, for the case that the class of the approximating densities is the mean field variational family:

$$q(\theta) = \prod_{m=1}^{M} q_m(\theta_m) \qquad (2.11)$$

In this case, Ormerod and Wand (2010) has proved that the maximizer of the ELBO is:

$$q_m^*(\theta_m) \propto \exp\left\{ \mathbb{E}_{q_{-m}} \left[\log p(y, \theta_m, \theta_{-m}) \right] \right\} \qquad (2.12)$$

where $q_m^*(\theta_m)$ is the approximated optimal density of θ_m, θ_{-m} denotes the other parameters other than θ_m, and the expectation is taken all over $q(\theta_{-m})$. Thus, we will assume that the

doi:10.6342/NTU202402563

parameters in our model to be mean field variational family so that we may use this result to find the approximated distribution for each parameter.

doi:10.6342/NTU202402563

Chapter 3 Main Results

Here we will first summarize the approximated distributions of each parameter and the corresponding estimates of each parameter.

The distribution of each parameter in VAR and SV part are:

$$
\begin{aligned}
q^*(\theta_i) &\sim N(\widehat{\theta_i}, \widehat{\Sigma}_{\theta_i}) \\
q^*(\mathbf{u}_{i,q}) &\sim N\left(\widehat{\mathbf{u}}_{i,q}, \widehat{\Sigma}_{\mathbf{u}_{i,q}}\right) \\
q^*(\sigma^2_{h,i}) &\sim IG(\widehat{c_i}, \widehat{d_i}) \\
q^*(\mathbf{u}_i) &\sim N(\widehat{\mathbf{u}}_i, \widehat{\Sigma}_{\mathbf{u}_i})
\end{aligned}
$$

where

$$
\begin{aligned}
\widehat{\Sigma}_{\theta_i} &= \left(V_i{}^{-1} + \mathbf{X}_i' \widehat{C}_{\mathbf{u}_i} \mathbf{X}_i\right)^{-1} \\
\widehat{\theta_i} &= \widehat{\Sigma}_{\theta_i} \mathbf{X}_i' \widehat{C}_{\mathbf{u}_i} \mathbf{y}_i \\
\widehat{\Sigma}_{\mathbf{u}_{i,q}} &= \left[\widehat{\sigma^2_{h,i}{}^{-1}} \left(\Sigma_{i,q}^{-1} + A' D A\right)\right]^{-1} \\
\widehat{\mathbf{u}}_{i,q} &= \widehat{\Sigma}_{\mathbf{u}_{i,q}} \left(A' D \widehat{\mathbf{u}}_{i,q+1}\right) \\
\widehat{c_i} &= c_i + \frac{T-q}{2} + \frac{1}{2} \\
\widehat{d_i} &= d_i + \frac{1}{2}\Bigg[\left(\hat{\mathbf{u}}_i - \widetilde{A}(\mathbf{1}_{T-q} \otimes \widehat{\mathbf{u}}_{i,q})\right)' \mathbf{A}' \mathbf{A} \left(\hat{\mathbf{u}}_i - \widetilde{A}(\mathbf{1}_{T-q} \otimes \widehat{\mathbf{u}}_{i,q})\right) + \text{tr}(\mathbf{A}' \mathbf{A} \widehat{\Sigma}_{\mathbf{u}_i}) \\
&\qquad + \text{tr}\left[(\mathbf{A}\widetilde{A})' \left(\mathbf{A}\widetilde{A}(I \otimes \widehat{\Sigma}_{\mathbf{u}_{i,q}})\right)\right] \widehat{\mathbf{u}}_{i,q}' \Sigma_{i,q}^{-1} \widehat{\mathbf{u}}_{i,q}\Bigg] \\
\widehat{\sigma^2_{h,i}{}^{-1}} &= \frac{\widehat{c_i}}{\widehat{d_i}}
\end{aligned}
$$

doi:10.6342/NTU202402563

and $\widehat{\Sigma}_{\mathbf{u}_i}$ is set to be the inverse of negative Hessian of $\log q^*_{\mathbf{u}_i}(\mathbf{u}_i)$ that is evaluated at the mode of $\log q^*_{\mathbf{u}_i}(\mathbf{u}_i)$. $\widehat{\mathbf{u}}_i$ is obtained by applying Newton-Raphson method with gradient and Hessian matrix shown as:

$$\begin{aligned}\text{Gradient} = &\frac{1}{2}\left\{(\mathbf{1}_{T-q}\otimes\phi) - \left[\widehat{\mathbf{s}}^2 \odot \exp\left(-(I\otimes\phi)'\boldsymbol{\mu} + (I\otimes\phi)'\widehat{\Sigma}_{\mathbf{u}_i}(\mathbf{1}_{T-q}\otimes\phi)\right)\right]\otimes\phi\right\} \\ &+ \widehat{\sigma^2_{h,i}{}^{-1}}\mathbf{A}'\mathbf{A}\left(\boldsymbol{\mu} - \widetilde{A}(\mathbf{1}_{T-q}\otimes\widehat{\mathbf{u}}_{i,q})\right) \\ \text{Hessian} = &\frac{1}{2}\text{diag}\left[\widehat{\mathbf{s}}^2 \odot \exp\left(-(I\otimes\phi)'\boldsymbol{\mu} + (I\otimes\phi)'\widehat{\Sigma}_{\mathbf{u}_i}(\mathbf{1}_{T-q}\otimes\phi)\right)\right]\otimes(\phi\phi') \\ &+ \widehat{\sigma^2_{h,i}{}^{-1}}\mathbf{A}'\mathbf{A}\end{aligned}$$

The distributions of parameters in the Horseshoe prior are shown to be:

$$\begin{aligned} q^*_{\lambda_{i,j}}(\lambda_{i,j}) &\sim IG\left(1, \frac{1}{2}\left(\widehat{\theta}_i(j)^2 + \widehat{\Sigma}_{\theta_i}(j,j)\widehat{\tau_i^{-1}}\right) + \widehat{v_{i,j}^{-1}}\right) \\ q^*_{\tau_i}(\tau_i) &\sim IG\left(\frac{k_i+1}{2}, \frac{1}{2}\left[\sum_{j=1}^{k_i}\left(\widehat{\theta}_i(j)^2 + \widehat{\Sigma}_{\theta_i}(j,j)\right)\widehat{\lambda_{i,j}^{-1}}\right] + \widehat{\xi_i^{-1}}\right) \\ q^*_{v_{i,j}}(v_{i,j}) &\sim IG\left(1, 1 + \widehat{\lambda_{i,j}^{-1}}\right) \\ q^*_{\xi_i}(\xi_i) &\sim IG\left(1, 1 + \widehat{\tau_i^{-1}}\right) \end{aligned}$$

and our estimators are the expectation of each distribution. Note that for the inverse gamma we will just find the expectation of their inverse, which will be expressed as below:

$$\begin{aligned} \widehat{\lambda_{i,j}{}^{-1}} &= \frac{1}{\frac{1}{2}\left(\widehat{\theta}_i(j)^2 + \widehat{\Sigma}_{\theta_i}(j,j)\widehat{\tau_i^{-1}}\right) + \widehat{v_{i,j}^{-1}}} \\ \widehat{\tau_i{}^{-1}} &= \frac{k_i+1}{\left[\sum_{j=1}^{k_i}\left(\widehat{\theta}_i(j)^2 + \widehat{\Sigma}_{\theta_i}(j,j)\right)\widehat{\lambda_{i,j}^{-1}}\right] + 2\widehat{\xi_i^{-1}}} \\ \widehat{v_{i,j}{}^{-1}} &= \frac{1}{1+\widehat{\lambda_{i,j}^{-1}}} \\ \widehat{\xi_i{}^{-1}} &= \frac{1}{1+\widehat{\tau_i^{-1}}} \end{aligned}$$

and the estimators for parameters that distributed in normal distribution are mean.

To find the posterior of each parameter, we first find out the joint density of each

doi:10.6342/NTU202402563

parameters:

$$
\begin{aligned}
p(\cdot) \propto & \prod_{t=q+1}^{T} \frac{1}{\sqrt{e^{\mathbf{u}_{i,t}{}'\phi}}} \cdot \exp\left\{ -\frac{(y_{i,t} - \mathbf{x}_{i,t}\theta_i)^2}{2e^{\mathbf{u}_{i,t}{}'\phi}} \right\} \\
& \times \prod_{t=q+1}^{T} \left\{ \left[\det{}^*(\Sigma_{i,t}^u)\right]^{-\frac{1}{2}} \cdot \exp\left\{ -\frac{1}{2} \left[\left(\mathbf{u}_{i,t} - A\mathbf{u}_{i,t-1}\right)' \Sigma_{i,t}^u{}^{\dagger} \left(\mathbf{u}_{i,t} - A\mathbf{u}_{i,t-1}\right) \right] \right\} \right\} \\
& \times \det\left(\sigma_{h,i}^2 \Sigma_{i,q}\right)^{-\frac{1}{2}} \cdot \exp\left\{ -\frac{1}{2} \left[\mathbf{u}_{i,q}' \left(\sigma_{h,i}^2 \Sigma_{i,q}\right)^{-1} \mathbf{u}_{i,q} \right] \right\} \times \left(\frac{1}{\sigma_{h,i}^2} \right)^{c_i+1} \cdot e^{-\frac{d_i}{\sigma_{h,i}^2}} \\
& \times \det(V_i)^{-\frac{1}{2}} \cdot e^{-\frac{1}{2}\left(\theta_i' V_i^{-1} \theta_i\right)} \times \prod_{j=1}^{k_i} \left[\left(\frac{1}{v_{i,j}} \right)^{\frac{1}{2}} \cdot \left(\frac{1}{\lambda_{i,j}} \right)^{\frac{3}{2}} \cdot e^{-\frac{1}{v_{i,j}\lambda_{i,j}}} \right] \\
& \times \left(\frac{1}{\xi_i} \right)^{\frac{1}{2}} \cdot \left(\frac{1}{\tau_i} \right)^{\frac{3}{2}} \cdot e^{-\frac{1}{\xi_i \tau_i}} \times \prod_{j=1}^{k_i} \left[\left(\frac{1}{v_{i,j}} \right)^{\frac{3}{2}} \cdot e^{\frac{1}{v_{i,j}}} \right] \times \left(\frac{1}{\xi_i} \right)^{\frac{3}{2}} \cdot e^{-\frac{1}{\xi_i}}
\end{aligned} \tag{3.1}
$$

note that since by our definition $\Sigma_{i,t}^u$ is not full rank, here we use pseudo inverse to deal with $\det^*(\Sigma_{i,t}^u)$, which denotes the pseudo deteminant, and $\Sigma_{i,t}^u{}^{\dagger}$, where $\det^*(\Sigma_{i,t}^u) = \frac{1}{\sigma_{h,i}^2}$ and $\Sigma_{i,t}^u{}^{\dagger}$ is a $q \times q$ matrix with $\frac{1}{\sigma_{h,i}^2}$ as the (q,q) element and else zero. Also, we define $\mathbf{u}_i = (\mathbf{u}_{i,q+1}{}', \ldots, \mathbf{u}_{i,T}{}')'$.

With joint distribution and the observed value $\mathbf{y}_i = (y_{i,q+1}, \ldots, y_{i,T})'$, we may find out the corresponding approximated posterior of the following 8 kinds of parameters, $(\theta_i, \mathbf{u}_{i,q}, \sigma_{h,i}^2, \mathbf{u}_i, \lambda_{i,j}, \tau_i, v_{i,j}, \xi_i)$, using VB method that is introduced in Section 2.3.

In Section 3.1 we will show the detail that how we find the approximated distribution for each parameter in VAR and SV terms, which are $\theta_i, \mathbf{u}_{i,q}, \sigma_{h,i}^2$ and $\mathbf{u}_i$. Also we will further discuss why Newton-Raphson is needed in the case of $\mathbf{u}_i$ and how we find it out.

In Section 3.2 we will show the detail that how the approximated distribution of parameters in the Horseshoe prior, which are $\lambda_{i,j}, \tau_i, v_{i,j}$ and ξ_i, are found.

In Section 3.3 we will introduce the iterative algorithm we are going to use to find the approximated distributions and hence the estimators for each parameter in our model.

doi:10.6342/NTU202402563

3.1 The Optimal Density for Parameters in VAR and SV

Here we will derive the result of the approximated marginal density of the parameters in VAR and SV, $\theta_i, \mathbf{u}_{i,q}, \sigma^2_{h,i}$ and $\mathbf{u}_i$. First, consider the optimal density of θ_i. Using the result of VB we may obtain $q^*_{\theta_i}$ by:

$$q^*_{\theta_i}(\theta_i) \propto \exp\{\mathbb{E}_{-\theta_i}\left[\log p(\theta_i|\mathbf{y}_i, \mathbf{u}_{i,q}, \sigma^2_{h,i}, \mathbf{u}_i, \lambda_{i,j}, \tau_i, v_{i,j}, \xi_i)\right]\}$$

here the log-density is given by (3.1):

$$\log p(\theta_i|\cdot) = c_{\theta_i} - \frac{1}{2}\sum_{t=q+1}^{T} e^{-\mathbf{u}_{i,t}{}'\phi}(y_{i,t} - \mathbf{x}_{i,t}\theta_i)^2 - \frac{1}{2}\left(\theta_i' V_i^{-1}\theta_i\right) \tag{3.2}$$

Taking the expectation over the parameters without θ_i, we obtain:

$$\mathbb{E}_{-\theta_i}\left[\log p(\theta_i|\cdot)\right] = c_{\theta_i} - \frac{1}{2}\sum_{t=q+1}^{T} \exp\left(-\widehat{\mathbf{u}}'_{i,t}\phi + \frac{1}{2}\phi'\widehat{\Sigma}_{\mathbf{u}_{i,t}}\phi\right)(y_{i,t} - \mathbf{x}_{i,t}\theta_i)^2 - \frac{1}{2}\left(\theta_i'\widehat{V_i^{-1}}\theta_i\right) \tag{3.3}$$

here $\widehat{\mathbf{u}}_{i,t}$ and $\widehat{\Sigma}_{\mathbf{u}_{i,t}}$ are the estimated expectation and variance for $\mathbf{u}_{i,t}$, which will be shown in the case for $\mathbf{u}_i$. Note that since $\mathbf{u}_i$ is a vector contains $T-q$ vectors with size q, so it can be shown that $\widehat{\mathbf{u}}_{i,t}$ is the $[t-(q+1)\times q+1]$-th to the $[t-(q+1)\times q+q]$-th element of the expectation for $\mathbf{u}_i$, and $\widehat{\Sigma}_{\mathbf{u}_{i,t}}$ is actually the $(t-q)$-th block in the diagonal of the covariance matrix for $\mathbf{u}_i$ which will be shown below in the $\mathbf{u}_i$ case. By the horseshoe prior assumption for V_i, $\widehat{V_i^{-1}}$ can be expressed as the estimators for $\widehat{\lambda_{i,j}^{-1}}$ and $\widehat{\tau_i^{-1}}$, which will be shown in the bottom of Section 3.2. Also, (3.3) can be rewritten into the matrix form as:

$$\mathbb{E}_{-\theta_i}\left[\log p(\theta_i|\cdot)\right] = c_{\theta_i} - \frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\theta_i)'\widehat{C}_{\mathbf{u}_i}(\mathbf{y}_i - \mathbf{X}_i\theta_i) - \frac{1}{2}\left(\theta_i'\widehat{V_i^{-1}}\theta_i\right) \tag{3.4}$$

where

$$\widehat{C}_{\mathbf{u}_i} = \text{diag}\left(\exp\left(-\widehat{\mathbf{u}}'_{i,q+1}\phi + \frac{1}{2}\phi'\widehat{\Sigma}_{\mathbf{u}_{i,q+1}}\phi\right), \ldots, \exp\left(-\widehat{\mathbf{u}}'_{i,T}\phi + \frac{1}{2}\phi'\widehat{\Sigma}_{\mathbf{u}_{i,T}}\phi\right)\right)$$

$$\mathbf{X}_i = (\mathbf{x}_{i,q+1}, \ldots, \mathbf{x}_{i,T})'$$

doi:10.6342/NTU202402563

Then we will combine the two terms for θ_i and the optimal density of θ is in the form of multivariate normal distribution $N(\widehat{\theta}_i, \widehat{\Sigma}_{\theta_i})$, with

$$\widehat{\Sigma}_{\theta_i} = \left(V_i^{-1} + \mathbf{X}_i' \widehat{C}_{\mathbf{u}_i} \mathbf{X}_i\right)^{-1} \tag{3.5}$$

$$\widehat{\theta}_i = \widehat{\Sigma}_{\theta_i} \mathbf{X}_i' \widehat{C}_{\mathbf{u}_i} \mathbf{y}_i \tag{3.6}$$

Next, the optimal density of $\mathbf{u}_{i,q}$ is similarly given by VB, hence we will just write down the conditional log-density of $\mathbf{u}_{i,q}$, which by (3.1) can be shown to be:

$$\begin{aligned}\log p(\mathbf{u}_{i,q}|\cdot) = c_{\mathbf{u}_{i,q}} &- \frac{1}{2}\left[(\mathbf{u}_{i,q+1} - A\mathbf{u}_{i,q})' {\Sigma_{i,t}^{u}}^{\dagger} (\mathbf{u}_{i,q+1} - A\mathbf{u}_{i,q})\right] \\ &- \frac{1}{2}\left[\mathbf{u}_{i,q}' \left(\sigma_{h,i}^2 \Sigma_{i,q}\right)^{-1} \mathbf{u}_{i,q}\right]\end{aligned}$$

Taking the expectation of above over the parameters without $\mathbf{u}_{i,q}$, we will have:

$$\begin{aligned}\mathbb{E}_{-\mathbf{u}_{i,q}}\left[\log p(\mathbf{u}_{i,q}|\cdot)\right] = c_{\mathbf{u}_{i,q}} - \frac{1}{2}\widehat{\sigma_{h,i}^2}^{-1}\Big[&(\widehat{\mathbf{u}}_{i,q+1} - A\mathbf{u}_{i,q})' D (\widehat{\mathbf{u}}_{i,q+1} - A\mathbf{u}_{i,q}) \\ &+ \mathrm{tr}(D\widehat{\Sigma}_{\mathbf{u}_{i,q+1}}) + \mathbf{u}_{i,q}' \Sigma_{i,q}^{-1} \mathbf{u}_{i,q}\Big]\end{aligned} \tag{3.7}$$

where

$$D = \left[\begin{array}{ccc|c} 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \\ \hline 0 & \cdots & 0 & 1 \end{array}\right] \tag{3.8}$$

and here $\widehat{\Sigma}_{\mathbf{u}_{i,q+1}}$ is the most top-left $q \times q$ block of the covariance matrix $\widehat{\Sigma}_{\mathbf{u}_{i,q+1}}$, $\widehat{(\sigma_{h,i}^2)^{-1}}$ is the expectation of the optimal density for $(\sigma_{h,i}^2)^{-1}$, which will be shown next. $\widehat{\mathbf{u}}_{i,q+1}$ is as same as the one in the case we shown in θ_i, which is the first to the q-th element of the expectation for $\mathbf{u}_i$. Similar to what we have done in the case for θ_i, the two terms of $\mathbf{u}_{i,q}$ can be combined and the optimal density for $\mathbf{u}_{i,q}$ is in the form of multivariate normal

doi:10.6342/NTU202402563

distribution $N\left(\widehat{\mathbf{u}}_{i,q}, \widehat{\Sigma}_{\mathbf{u}_{i,q}}\right)$, where

$$\widehat{\Sigma}_{\mathbf{u}_{i,q}} = \left[\widehat{\sigma_{h,i}^{2\ -1}}\left(\Sigma_{i,q}^{-1} + A^{'}DA\right)\right]^{-1} \tag{3.9}$$

$$\widehat{\mathbf{u}}_{i,q} = \widehat{\Sigma}_{\mathbf{u}_{i,q}}\left(A^{'}D\widehat{\mathbf{u}}_{i,q+1}\right) \tag{3.10}$$

The conditional log-density of $\sigma_{h,i}^2$ can be expressed as the form:

$$\begin{aligned}\log p(\sigma_{h,i}^2|\cdot) =& c_{\sigma_{h,i}^2} - \frac{T-q}{2}\log\sigma_{h,i}^2 - \frac{1}{2\sigma_{h,i}^2}\left(\mathbf{u}_i - \widetilde{A}(\mathbf{1}_{T-q}\otimes\mathbf{u}_{i,q})\right)^{'}\mathbf{A}^{'}\mathbf{A}\left(\mathbf{u}_i - \widetilde{A}(\mathbf{1}_{T-q}\otimes\mathbf{u}_{i,q})\right) \\ & - \frac{1}{2}\log\sigma_{h,i}^2 - \frac{1}{2}\left[\mathbf{u}_{i,q}^{'}\left(\sigma_{h,i}^2\Sigma_{i,q}\right)^{-1}\mathbf{u}_{i,q}\right] - (c_i+1)\log\sigma_{h,i}^2 - \frac{d_i}{\sigma_{h,i}^2}\end{aligned}$$

where

$$\widetilde{A} = \begin{pmatrix} I & O & \cdots & O \\ O & A & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & A^{T-q} \end{pmatrix}$$

$$\mathbf{1}_{T-q} = (1, \cdots, 1)^{'}$$

$$\mathbf{A}^{'}\mathbf{A} = \begin{pmatrix} I & I & \cdots & I \\ I & A+I & \cdots & A+I \\ \vdots & \vdots & \ddots & \vdots \\ I & A+I & \cdots & A^{T-q-1}+\cdots+A+I \end{pmatrix}$$

and $(\mathbf{1}_{T-q}\otimes\mathbf{u}_{i,q})$ denotes the Kronecker product of the two vectors. Taking the expectation of above over theparameters without $\sigma_{h,i}^2$, we will have:

$$\begin{aligned}\mathbb{E}_{-\sigma_{h,i}^2}\left[\log p(\sigma_{h,i}^2|\cdot)\right] = c_{\sigma_{h,i}^2} & - \frac{T-q+2c_i+3}{2}\log\sigma_{h,i}^2 \\ & - \frac{1}{2\sigma_{h,i}^2}\Bigg[\left(\widehat{\mathbf{u}}_i - \widetilde{A}(\mathbf{1}_{T-q}\otimes\widehat{\mathbf{u}}_{i,q})\right)^{'}\mathbf{A}^{'}\mathbf{A}\left(\widehat{\mathbf{u}}_i - \widetilde{A}(\mathbf{1}_{T-q}\otimes\widehat{\mathbf{u}}_{i,q})\right) \\ & + \mathrm{tr}(\mathbf{A}^{'}\mathbf{A}\widehat{\Sigma}_{\mathbf{u}_i}) + \mathrm{tr}\left[(\mathbf{A}\widetilde{A})^{'}\left(\mathbf{A}\widetilde{A}(I\otimes\widehat{\Sigma}_{\mathbf{u}_{i,q}})\right)\right] + \widehat{\mathbf{u}}_{i,q}^{'}\Sigma_{i,q}^{-1}\widehat{\mathbf{u}}_{i,q} + 2d_i\Bigg]\end{aligned}$$

Here the expectation is in the form of the log-density of an inverse gamma distribu-

doi:10.6342/NTU202402563

tion, and hence we obtain that $\sigma^2_{h,i} \sim IG(\widehat{c}_i, \widehat{d}_i)$, where

$$\begin{aligned}
\widehat{c}_i &= c_i + \frac{T-q}{2} + \frac{1}{2} \\
\widehat{d}_i &= d_i + \frac{1}{2}\Bigg[\left(\hat{\mathbf{u}}_i - \widetilde{A}(\mathbf{1}_{T-q} \otimes \widehat{\mathbf{u}}_{i,q})\right)' \mathbf{A}'\mathbf{A} \left(\hat{\mathbf{u}}_i - \widetilde{A}(\mathbf{1}_{T-q} \otimes \widehat{\mathbf{u}}_{i,q})\right) + \text{tr}(\mathbf{A}'\mathbf{A}\widehat{\Sigma}_{\mathbf{u}_i}) \\
&\quad + \text{tr}\left[(\mathbf{A}\widetilde{A})' \left(\mathbf{A}\widetilde{A}(I \otimes \widehat{\Sigma}_{\mathbf{u}_{i,q}})\right)\right] \widehat{\mathbf{u}}'_{i,q}\Sigma^{-1}_{i,q}\widehat{\mathbf{u}}_{i,q} \Bigg]
\end{aligned}$$

and we will set $\widehat{(\sigma^2_{h,i})^{-1}}$ as $\frac{\widehat{c}_i}{\widehat{d}_i}$ in the sense of the expectation of inverse gamma distribution.

Lastly, the conditional log-density of $\mathbf{u}_i$ has the form:

$$\begin{aligned}
\log p(\mathbf{u}_i|\cdot) =& c_{\mathbf{u}_i} - \frac{1}{2}\sum_{t=q+1}^{T} {\mathbf{u}_{i,t}}'\phi - \frac{1}{2}\sum_{t=q+1}^{T} \frac{(y_{i,t} - \mathbf{x}_{i,t}\theta_i)^2}{e^{{\mathbf{u}_{i,t}}'\phi}} \\
& - \frac{1}{2}\sum_{t=q+1}^{T} (\mathbf{u}_{i,t} - A\mathbf{u}_{i,t-1})'{\Sigma^u_{i,t}}^{\dagger}(\mathbf{u}_{i,t} - A\mathbf{u}_{i,t-1})
\end{aligned}$$

Taking the expectation with respect to the parameters without $\mathbf{u}_i$, we obtain that:

$$\begin{aligned}
\mathbb{E}_{-\mathbf{u}_i}\left[\log p(\mathbf{u}_i|\cdot)\right] = c_{\mathbf{u}_i} &- \frac{1}{2}\sum_{t=q+1}^{T} {\mathbf{u}_{i,t}}'\phi - \frac{1}{2}\sum_{t=q+1}^{T} e^{-{\mathbf{u}_{i,t}}'\phi}\left[(y_{i,t} - \mathbf{x}_{i,t}\widehat{\theta}_i)^2 + \text{tr}(\mathbf{x}'_{i,t}\mathbf{x}_{i,t}\widehat{\Sigma}_{\theta_i})\right] \\
&- \frac{1}{2}\widehat{(\sigma^2_{h,i})^{-1}}\Bigg[\sum_{t=q+2}^{T} (\mathbf{u}_{i,t} - A\mathbf{u}_{i,t-1})'D(\mathbf{u}_{i,t} - A\mathbf{u}_{i,t-1}) \\
&+ (\mathbf{u}_{i,q+1} - A\mathbf{u}_{i,q})'D(\mathbf{u}_{i,q+1} - A\mathbf{u}_{i,q}) + \text{tr}(D\widehat{\Sigma}_{\mathbf{u}_{i,q}})\Bigg]
\end{aligned}$$

and hence by the result of ELBO,

$$\begin{aligned}
\log q^*_{\mathbf{u}_i}(\mathbf{u}_i) = \widetilde{c}_{\mathbf{u}_i} &- \frac{1}{2}\sum_{t=q+1}^{T} {\mathbf{u}_{i,t}}'\phi - \frac{1}{2}\sum_{t=q+1}^{T} e^{-{\mathbf{u}_{i,t}}'\phi}\left[(y_{i,t} - \mathbf{x}_{i,t}\widehat{\theta}_i)^2 + \text{tr}(\mathbf{x}'_{i,t}\mathbf{x}_{i,t}\widehat{\Sigma}_{\theta_i})\right] \\
&- \frac{1}{2}\widehat{(\sigma^2_{h,i})^{-1}}\Bigg[\sum_{t=q+2}^{T} (\mathbf{u}_{i,t} - A\mathbf{u}_{i,t-1})'D(\mathbf{u}_{i,t} - A\mathbf{u}_{i,t-1}) \\
&+ (\mathbf{u}_{i,q+1} - A\mathbf{u}_{i,q})'D(\mathbf{u}_{i,q+1} - A\mathbf{u}_{i,q})\Bigg]
\end{aligned}$$

where D was defined in (3.8). Note that here we want to find an approximation for $\mathbf{u}_i$, but there is a term of $\mathbf{u}_{i,t}$ in exponential term, where we can not find a known or common distribution for it. Hence we apply the KL-divergence in VB again to obtain an ideal

doi:10.6342/NTU202402563

distribution for $\mathbf{u}_i$. Since in prior setting we assumed that $\mathbf{u}_{i,t}$ follows normal distribution, it might be better to assume the approximated distribution be a normal distribution too. So it suffices to find appropriate $\widehat{\mathbf{u}}_i$ and $\widehat{\Sigma}_{\mathbf{u}_i}$ as the mean and variance of an approximated normal distribution $N(\widehat{\mathbf{u}}_i, \widehat{\Sigma}_{\mathbf{u}_i})$, and note that $\mathbb{E}(\mathbf{u}_{i,t}) = \widehat{\mathbf{u}}_{i,t},\ q+1 \leq t \leq T$.

To find such $\widehat{\mathbf{u}}_i$ and $\widehat{\Sigma}_{\mathbf{u}_i}$, first we set $\widehat{\Sigma}_{\mathbf{u}_i}$ to be the inverse of negative Hessian of $\log q^*_{\mathbf{u}_i}(\mathbf{u}_i)$ that is evaluated at the mode of $\log q^*_{\mathbf{u}_i}(\mathbf{u}_i)$, since the mean of a normal distribution is the number that appears most frequently and the fisher information matrix is used as the inverse of the maximum likelihood estimator of the variance, while Fisher information matrix is the negative Hessian matrix.

Next, we apply KL-divergence to find the estimator of the mean $\widehat{\mathbf{u}}_i$. Let $f_{\boldsymbol{\mu}_i}(\cdot)$ denotes the density function of multivariate normal distribution with mean $\boldsymbol{\mu}_i$ and variance $\widehat{\Sigma}_{\mathbf{u}_i}$. Then we are going to solve the following problem:

$$\min_{\boldsymbol{\mu}_i \in \mathbb{R}^m} \mathbb{E}\left[\log \frac{f_{\boldsymbol{\mu}_i}(\mathbf{u}_i)}{q^*_{\mathbf{u}_i}(\mathbf{u}_i)}\right]$$

here $m = q \times (T-q)$ the expectation is taken according to the assumption of the multivariate normal distribution $f_{\boldsymbol{\mu}_i}(\mathbf{u}_i)$. Notice that the problem we aim to solve is a convex optimization problem with unique minimum, we are going to find out the critical point that makes the gradient of $\mathbb{E}\left[\log \frac{f_{\boldsymbol{\mu}_i}(\mathbf{u}_i)}{q^*_{\mathbf{u}_i}(\mathbf{u}_i)}\right]$ a zero vector. Since the gradient and Hessian are solvable, we may apply Newton-Raphson method to find out the critical point and hence solve the minimum of $\mathbb{E}\left[\log \frac{f_{\boldsymbol{\mu}_i}(\mathbf{u}_i)}{q^*_{\mathbf{u}_i}(\mathbf{u}_i)}\right]$. Note that Newton-Raphson method is doing iteration with following formula:

$$\mathbf{x}_{n+1} = \mathbf{x}_n - H^{-1}(\mathbf{x}_n)G(\mathbf{x}_n)$$

here H denotes the Hessian matrix and G denotes the gradient.

doi:10.6342/NTU202402563

First, we derive the expression of $\log \frac{f_{\boldsymbol{\mu}_i}(\mathbf{u}_i)}{q^*_{\mathbf{u}_i}(\mathbf{u}_i)}$ as below:

$$\begin{aligned}\log \frac{f_{\boldsymbol{\mu}_i}(\mathbf{u}_i)}{q^*_{\mathbf{u}_i}(\mathbf{u}_i)} = c - \frac{1}{2}(\mathbf{u}_i - \boldsymbol{\mu})' \widehat{\Sigma}^{-1}_{\mathbf{u}_i}(\mathbf{u}_i - \boldsymbol{\mu}) \\ + \frac{1}{2}\Bigg[(\mathbf{1}_{T-q} \otimes \phi)' \mathbf{u}_i + (\widehat{\mathbf{s}}^2)' e^{-(I\otimes\phi)'\mathbf{u}_i} \\ + \widehat{(\sigma^2_{h,i})^{-1}} \left(\mathbf{u}_i - \widetilde{A}(\mathbf{1}_{T-q} \otimes \widehat{\mathbf{u}}_{i,q})\right)' \mathbf{A}'\mathbf{A} \left(\mathbf{u}_i - \widetilde{A}(\mathbf{1}_{T-q} \otimes \widehat{\mathbf{u}}_{i,q})\right) \Bigg]\end{aligned}$$

where c is a constant of $\mathbf{u}_i$, the last term is same as the one given in $\sigma^2_{h,i}$ case, and $\widehat{\mathbf{s}}^2$ is defined to be $(y_{i,t} - \mathbf{x}_{i,t}\widehat{\theta}_i)^2 + \text{tr}(\mathbf{x}'_{i,t}\mathbf{x}_{i,t}\widehat{\Sigma}_{\theta_i})$.

Next, we take the expectation with respect to the ideal normal density $f_{\boldsymbol{\mu}_i}$, which gives that $\mathbb{E}(\mathbf{u}_i) = \boldsymbol{\mu}_i$, then we may obtain:

$$\begin{aligned}\mathbb{E}\left[\log \frac{f_{\boldsymbol{\mu}}(\mathbf{u}_i)}{q^*_{\mathbf{u}_i}(\mathbf{u}_i)}\right] = c' + \frac{1}{2}\Bigg\{ (\mathbf{1}_{T-q} \otimes \phi)' \boldsymbol{\mu} + (\widehat{\mathbf{s}}^2)' \exp\Big[-(I\otimes\phi)'\boldsymbol{\mu} + (I\otimes\phi)'\widehat{\Sigma}_{\mathbf{u}_i}(\mathbf{1}_{T-q}\otimes\phi)\Big] \\ + \widehat{\sigma^2_{h,i}{}^{-1}} \left(\boldsymbol{\mu} - \widetilde{A}(\mathbf{1}_{T-q} \otimes \widehat{\mathbf{u}}_{i,q})\right)' \mathbf{A}'\mathbf{A} \left(\boldsymbol{\mu} - \widetilde{A}(\mathbf{1}_{T-q} \otimes \widehat{\mathbf{u}}_{i,q})\right) \Bigg\}\end{aligned}$$

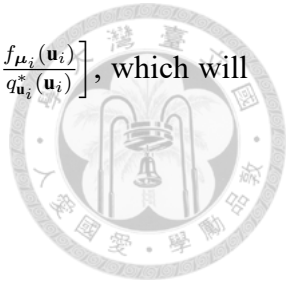
here c' is a constant that is irrelevant to $\boldsymbol{\mu}_i$. In the formula, we can first notice that it is convex in $\widehat{\mathbf{u}}_i$, which means the minimum exists. Then we may apply Newton-Raphson to the gradient of the function of expectation in order to find the critical point and hence find the point which satisfies the KL-divergence to get $\widehat{\mathbf{u}}_i$. To apply Newton-Raphson to the gradient, we'll first calculate the gradient and Hessian of the expectation $\mathbb{E}\left[\log \frac{f_{\boldsymbol{\mu}_i}(\mathbf{u}_i)}{q^*_{\mathbf{u}_i}(\mathbf{u}_i)}\right]$ with respect to parameter $\widehat{\mathbf{u}}_i$, which is given by:

$$\begin{aligned}\text{Gradient} = \frac{1}{2}\Big\{ (\mathbf{1}_{T-q}\otimes\phi) - \Big[\widehat{\mathbf{s}}^2 \odot \exp\Big(-(I\otimes\phi)'\boldsymbol{\mu}_i + (I\otimes\phi)'\widehat{\Sigma}_{\mathbf{u}_i}(\mathbf{1}_{T-q}\otimes\phi)\Big)\Big] \otimes \phi \Big\} \\ + \widehat{(\sigma^2_{h,i})^{-1}}\mathbf{A}'\mathbf{A}\left(\boldsymbol{\mu}_i - \widetilde{A}(\mathbf{1}_{T-q}\otimes\widehat{\mathbf{u}}_{i,q})\right) \\ \text{Hessian} = \frac{1}{2}\text{diag}\left[\widehat{\mathbf{s}}^2 \odot \exp\Big(-(I\otimes\phi)'\boldsymbol{\mu} + (I\otimes\phi)'\widehat{\Sigma}_{\mathbf{u}_i}(\mathbf{1}_{T-q}\otimes\phi)\Big)\right] \otimes (\phi\phi') \\ + \widehat{(\sigma^2_{h,i})^{-1}}\mathbf{A}'\mathbf{A}\end{aligned}$$

Note that here by the definition of ϕ, the Hessian matrix is not full rank, hence we'll deal with it by using pseudo inverse. At last by applying Newton-Raphson method to $\boldsymbol{\mu}_i$,

doi:10.6342/NTU202402563

we might find the value for $\boldsymbol{\mu}_i$ that achieves the minimum of $\mathbb{E}\left[\log \frac{f_{\mu_i}(\mathbf{u}_i)}{q^*_{\mathbf{u}_i}(\mathbf{u}_i)}\right]$, which will be set to be $\widehat{\mathbf{u}}_i$.



3.2 The Optimal Density for Horseshoe Prior Parameters

Here we will derive the approximated marginal density of the parameters in the Horseshoe prior, $\lambda_{i,j}, \tau_i, v_{i,j}$ and ξ_i. Note that in the assumption of the prior of these four parameters, they follow inverse gamma distributions, also they all appears as inverse in the joint density of all parameters, hence the expectation might be taken in the sense of $\lambda_{i,j}^{-1}, \tau_i^{-1}, v_{i,j}^{-1}$ and ξ_i^{-1}, as we have done in Section 3.1.

Similar as we done in Section 3.1, using the result of VB we can first derive the optimal density for $\lambda_{i,j}$:

$$q^*_{\lambda_{i,j}}(\lambda_{i,j}) \propto \exp\left\{\mathbb{E}_{-\lambda_{i,j}}\left[\log p(\lambda_{i,j}|\mathbf{y}_i, \theta_i, \mathbf{u}_{i,q}, \mathbf{u}_i, \sigma^2_{h,i}, \tau_i, v_{i,j}, \xi_i)\right]\right\}$$

here the log-density is given by (3.1):

$$\log p(\lambda_{i,j}|\cdot) = c_{\lambda_{i,j}} - \frac{1}{2}\log \lambda_{i,j} - \frac{1}{2}\frac{\theta_i(j)^2}{\lambda_{i,j}\tau_i} - \frac{3}{2}\log \lambda_{i,j} - \frac{1}{\lambda_{i,j} v_{i,j}}$$

where $\theta_i(j)$ denotes the i-th element of vector θ_i. Then we take the expectation over parameters except for $\lambda_{i,j}$, we obtain:

$$\mathbb{E}_{-\lambda_{i,j}}\left[\log p(\lambda_{i,j}|\cdot)\right] = c_{\lambda_{i,j}} - 2\log \lambda_{i,j} - \left[\frac{1}{2}\left(\widehat{\theta}_i(j)^2 + \widehat{\Sigma}_{\theta_i}(j,j)\widehat{\tau_i^{-1}}\right) + \widehat{v_{i,j}^{-1}}\right]\frac{1}{\lambda_{i,j}}$$

which is the log-density of an inverse gamma distribution. Hence we have the optimal density $q^*_{\lambda_{i,j}}(\lambda_{i,j}) \sim IG\left(1, \frac{1}{2}\left(\widehat{\theta}_i(j)^2 + \widehat{\Sigma}_{\theta_i}(j,j)\widehat{\tau_i^{-1}}\right) + \widehat{v_{i,j}^{-1}}\right)$.

The log-density of τ_i is given by (3.1):

$$\log p(\tau_i|\cdot) = c_{\tau_i} - \frac{k_i}{2}\log \tau_i - \frac{1}{2}\sum_{j=1}^{k_i}\frac{\theta_i(j)^2}{\lambda_{i,j}\tau_i} - \frac{3}{2}\log \tau_i - \frac{1}{\tau_i \xi_i}$$

doi:10.6342/NTU202402563

Taking the expectation over parameters except τ_i, we obtain that:

$$\mathbb{E}_{-\tau_i}\left[\log p(\tau_i|\cdot)\right] = c_{\tau_i} - \frac{k_i+3}{2}\log\tau_i - \left[\frac{1}{2}\sum_{j=1}^{k_i}\left(\widehat{\theta}_i(j)^2 + \widehat{\Sigma}_{\theta_i}(j,j)\right)\widehat{\lambda_{i,j}^{-1}} + \widehat{\xi_i^{-1}}\right]\frac{1}{\tau_i}$$

which is the log-density of an inverse gamma distribution. Hence we have the optimal density $q^*_{\tau_i}(\tau_i) \sim IG\left(\frac{k_i+1}{2}, \frac{1}{2}\left[\sum_{j=1}^{k_i}\left(\widehat{\theta}_i(j)^2 + \widehat{\Sigma}_{\theta_i}(j,j)\right)\widehat{\lambda_{i,j}^{-1}}\right] + \widehat{\xi_i^{-1}}\right)$.

The log-density of $v_{i,j}$ is given by (3.1):

$$\log p(v_{i,j}|\cdot) = c_{v_{i,j}} - \frac{1}{2}\log v_{i,j} - \frac{1}{\lambda_{i,j}v_{i,j}} - \frac{3}{2}\log v_{i,j} - \frac{1}{v_{i,j}}$$

Taking the expectation over parameters except $v_{i,j}$, we obtain that:

$$\mathbb{E}_{-v_{i,j}}\left[\log p(v_{i,j}|\cdot)\right] = c_{v_{i,j}} - 2\log v_{i,j} - \left(1 + \widehat{\lambda_{i,j}^{-1}}\right)\frac{1}{v_{i,j}}$$

which is the log-density of an inverse gamma distribution. Hence we have the optimal density $q^*_{v_{i,j}}(v_{i,j}) \sim IG\left(1, 1 + \widehat{\lambda_{i,j}^{-1}}\right)$.

Lastly, the log-density of ξ_i is given by (3.1):

$$\log p(\xi_i|\cdot) = c_{\xi_i} - \frac{1}{2}\log\xi_i - \frac{1}{\tau_i\xi_i} - \frac{3}{2}\log\xi_i - \frac{1}{\xi_i}$$

Taking the expectation over parameters except $v_{i,j}$, we obtain that:

$$\mathbb{E}_{-\xi_i}\left[\log p(\xi_i|\cdot)\right] = c_{\xi_i} - 2\log\xi_i - \left(1 + \widehat{\tau_i^{-1}}\right)\frac{1}{\xi_i}$$

which is the log-density of an inverse gamma distribution, hence we have the optimal density $q^*_{\xi_i}(\xi_i) \sim IG\left(1, 1 + \widehat{\tau_i^{-1}}\right)$.

Here, by the result that each of the approximated distributions of parameters $\lambda_{i,j}, \tau_i, v_{i,j}$

doi:10.6342/NTU202402563

and ξ_i is inverse gamma, we know that:

$$\widehat{\lambda_{i,j}^{-1}} = \frac{1}{\frac{1}{2}\left(\widehat{\theta}_i(j)^2 + \widehat{\Sigma}_{\theta_i}(j,j)\widehat{\tau_i^{-1}}\right) + \widehat{v_{i,j}^{-1}}}$$

$$\widehat{\tau_i^{-1}} = \frac{k_i + 1}{\left[\sum_{j=1}^{k_i}\left(\widehat{\theta}_i(j)^2 + \widehat{\Sigma}_{\theta_i}(j,j)\right)\widehat{\lambda_{i,j}^{-1}}\right] + 2\widehat{\xi_i^{-1}}}$$

$$\widehat{v_{i,j}^{-1}} = \frac{1}{1 + \widehat{\lambda_{i,j}^{-1}}}$$

$$\widehat{\xi_i^{-1}} = \frac{1}{1 + \widehat{\tau_i^{-1}}}$$

which allows us to find out the estimated parameters through the iterative algorithm.

3.3 Iterative Algorithm

Here we provide an approach to use the iterative algorithm to find out the estimates of each variable in the approximated distribution that we obtained by VB. The algorithm is shown in Algorithm 3.3.

[htb] Intialize: $\widehat{\Sigma}_{\theta_i}, \widehat{\theta}_i, \widehat{\Sigma}_{\mathbf{u}_{i,q}}, \widehat{\mathbf{u}}_{i,q}, \widehat{c}_i, \widehat{d}_i, \widehat{\sigma_{h,i}^{2\,-1}}, \widehat{\lambda_{i,j}^{-1}}, \widehat{\tau_i^{-1}}, \widehat{v_{i,j}^{-1}}, \widehat{\xi_i^{-1}}$. Set $\widehat{c}_i = c_i +$

doi:10.6342/NTU202402563

$\frac{T-q}{2}+\frac{1}{2}$. Cycle:

$$
\begin{aligned}
\widehat{\mathbf{u}}_i, \widehat{\Sigma}_{\mathbf{u}_i} \leftarrow & \text{obtained by Newton-Raphson method} \\
\widehat{\Sigma}_{\theta_i} &\leftarrow \left(V_i^{-1}+\mathbf{X}_i^{'}\widehat{C}_{\mathbf{u}_i}\mathbf{X}_i\right)^{-1} \\
\widehat{\theta}_i &\leftarrow \widehat{\Sigma}_{\theta_i}\mathbf{X}_i^{'}\widehat{C}_{\mathbf{u}_i}\mathbf{y}_i \\
\widehat{\Sigma}_{\mathbf{u}_{i,q}} &\leftarrow \left[\widehat{(\sigma_{h,i}^2)^{-1}}\left(\Sigma_{i,q}^{-1}+A^{'}DA\right)\right]^{-1} \\
\widehat{\mathbf{u}}_{i,q} &\leftarrow \widehat{\Sigma}_{\mathbf{u}_{i,q}}\left(A^{'}D\widehat{\mathbf{u}}_{i,q+1}\right) \\
\widehat{c_i} &\leftarrow c_i+\frac{T-q}{2}+\frac{1}{2} \\
\widehat{d_i} &\leftarrow d_i+\frac{1}{2}\Bigg[\left(\hat{\mathbf{u}}_i-\widetilde{A}(\mathbf{1}_{T-q}\otimes\widehat{\mathbf{u}}_{i,q})\right)^{'}\mathbf{A}^{'}\mathbf{A}\left(\hat{\mathbf{u}}_i-\widetilde{A}(\mathbf{1}_{T-q}\otimes\widehat{\mathbf{u}}_{i,q})\right)+\text{tr}(\mathbf{A}^{'}\mathbf{A}\widehat{\Sigma}_{\mathbf{u}_i}) \\
&\qquad +\text{tr}\left[(\mathbf{A}\widetilde{A})^{'}\left(\mathbf{A}\widetilde{A}(I\otimes\widehat{\Sigma}_{\mathbf{u}_{i,q}})\right)\right]\widehat{\mathbf{u}}_{i,q}^{'}\Sigma_{i,q}^{-1}\widehat{\mathbf{u}}_{i,q}\Bigg] \\
\widehat{(\sigma_{h,i}^2)^{-1}} &\leftarrow \frac{\widehat{c_i}}{\widehat{d_i}} \\
\widehat{\lambda_{i,j}^{-1}} &\leftarrow \frac{1}{\frac{1}{2}\left(\widehat{\theta}_i(j)^2+\widehat{\Sigma}_{\theta_i}(j,j)\widehat{\tau_i^{-1}}\right)+\widehat{v_{i,j}^{-1}}} \\
\widehat{\tau_i^{-1}} &\leftarrow \frac{k_i+1}{\left[\sum_{j=1}^{k_i}\left(\widehat{\theta}_i(j)^2+\widehat{\Sigma}_{\theta_i}(j,j)\right)\widehat{\lambda_{i,j}^{-1}}\right]+2\widehat{\xi_i^{-1}}} \\
\widehat{v_{i,j}^{-1}} &\leftarrow \frac{1}{1+\widehat{\lambda_{i,j}^{-1}}} \\
\widehat{\xi_i^{-1}} &\leftarrow \frac{1}{1+\widehat{\tau_i^{-1}}}
\end{aligned}
$$

until the increase in $\widetilde{p}(\mathbf{y}_i;q)$ is negligible.

In the algorithm, $\widetilde{p}(\mathbf{y}_i;q)$ denotes the variational lower bound, which can be obtained by first computing the log ratio of the joint posterior density (3.1) and the variational

doi:10.6342/NTU202402563

approximation:

$$
\begin{aligned}
\log\left[\frac{p(\mathbf{y}_i,\Theta)}{q(\Theta)}\right] =& c - \frac{1}{2}(\mathbf{1}_{T-q}\otimes\phi)'\mathbf{u}_i - \frac{1}{2}(\mathbf{y}_i-\mathbf{X}_i\theta_i)'C_{\mathbf{u}_i}(\mathbf{y}_i-\mathbf{X}_i\theta_i) \\
&- \frac{1}{2}\sum_{j=1}^{k_i}\log\left(\lambda_{i,j}^{-1}\tau_i^{-1}\right) - \frac{1}{2}\left(\theta_i' V_i^{-1}\theta_i\right) \\
&- \frac{1}{2\sigma_{h,i}^2}\left(\mathbf{u}_i - \widetilde{A}(\mathbf{1}_{T-q}\otimes\mathbf{u}_{i,q})\right)' \mathbf{A}'\mathbf{A}\left(\mathbf{u}_i - \widetilde{A}(\mathbf{1}_{T-q}\otimes\mathbf{u}_{i,q})\right) \\
&- \frac{T-q}{2}\log\sigma_{h,i}^2 - \frac{1}{2}\log\sigma_{h,i}^2 - \frac{1}{2}\left[\mathbf{u}_{i,q}'\left(\sigma_{h,i}^2\Sigma_{i,q}\right)^{-1}\mathbf{u}_{i,q}\right] \\
&- (c_i+1)\log\sigma_{h,i}^2 - \frac{d_i}{\sigma_{h,i}^2} - \sum_{j=1}^{k_i}\left[\frac{1}{2}\log v_{i,j} + \frac{3}{2}\log\lambda_{i,j} + \frac{1}{v_{i,j}\lambda_{i,j}}\right] \\
&- \frac{1}{2}\log\xi_i - \frac{3}{2}\log\tau_i - \frac{1}{\xi_i\tau_i} - \sum_{j=1}^{k_i}\left[\frac{3}{2}\log v_{i,j} + \frac{1}{v_{i,j}}\right] - \frac{3}{2}\log\xi_i - \frac{1}{\xi_i} \\
&+ \frac{1}{2}(\theta_i-\widehat{\theta})'\widehat{\Sigma}_{\theta_i}^{-1}(\theta_i-\widehat{\theta}) + \frac{1}{2}(\mathbf{u}_{i,q}-\widehat{\mathbf{u}}_{i,q})'\widehat{\Sigma}_{\mathbf{u}_{i,q}}^{-1}(\mathbf{u}_{i,q}-\widehat{\mathbf{u}}_{i,q}) \\
&+ \frac{1}{2}(\mathbf{u}_i-\widehat{\mathbf{u}}_i)'\widehat{\Sigma}_{\mathbf{u}_i}^{-1}(\mathbf{u}_i-\widehat{\mathbf{u}}_i) + (\widehat{c}_i+1)\log\sigma_{h,i}^2 + \frac{\widehat{d}_i}{\sigma_{h,i}^2} \\
&+ \sum_{j=1}^{k_i}\left\{2\log\lambda_{i,j} + \left[\frac{1}{2}\left(\widehat{\theta}_i(j)^2 + \widehat{\Sigma}_{\theta_i}(j,j)\widehat{\tau_i^{-1}}\right) + \widehat{v_{i,j}^{-1}}\right]{\lambda_{i,j}}^{-1}\right\} \\
&+ \frac{k_i+3}{2}\log\tau_i + \left\{\frac{1}{2}\left[\sum_{j=1}^{k_i}\left(\widehat{\theta}_i(j)^2 + \widehat{\Sigma}_{\theta_i}(j,j)\right)\widehat{\lambda_{i,j}^{-1}}\right] + \widehat{\xi_i^{-1}}\right\}\tau_i^{-1} \\
&+ \sum_{j=1}^{k_i}\left[2\log v_{i,j} + \left(1+\widehat{\lambda_{i,j}^{-1}}\right){v_{i,j}}^{-1}\right] + 2\log\xi_i + \left(1+\widehat{\tau_i^{-1}}\right)\xi_i^{-1}
\end{aligned}
$$

where Θ denotes all the parameters $(\theta_i, \mathbf{u}_{i,q}, \sigma_{h,i}^2, \mathbf{u}_i, \lambda_{i,j}, \tau_i, v_{i,j}, \xi_i)$, and c here denotes:

$$
\begin{aligned}
c = c' &+ \frac{1}{2}\log\left|\widehat{\Sigma}_{\theta_i}\right| + \frac{1}{2}\log\left|\widehat{\Sigma}_{\mathbf{u}_{i,q}}\right| + \frac{1}{2}\log\left|\widehat{\Sigma}_{\mathbf{u}_i}\right| - \widehat{c}_i\log\widehat{d}_i + \log\Gamma(\widehat{c}_i) \\
&- \log\left(\frac{1}{2}\left(\widehat{\theta}_i(j)^2 + \widehat{\Sigma}_{\theta_i}(j,j)\widehat{\tau_i^{-1}}\right) + \widehat{v_{i,j}^{-1}}\right) \\
&- \frac{k_i+1}{2}\log\left(\frac{1}{2}\left[\sum_{j=1}^{k_i}\left(\widehat{\theta}_i(j)^2 + \widehat{\Sigma}_{\theta_i}(j,j)\right)\widehat{\lambda_{i,j}^{-1}}\right] + \widehat{\xi_i^{-1}}\right) \\
&- \log\left(1+\widehat{\lambda_{i,j}^{-1}}\right) - \log\left(1+\widehat{\tau_i^{-1}}\right)
\end{aligned}
$$

is the constant term while taking expectation and c' denotes the constant terms and the known values. The notation $C_{\mathbf{u}_i} = \text{diag}(\exp\{-\mathbf{u}_{i,q+1}'\phi\}, \ldots, \exp\{-\mathbf{u}_{i,T}'\phi\})$.

doi:10.6342/NTU202402563

The variational lower bound is defined to be the expectation with respect to q of the log ratio of the joint posterior density and the variational approximation above, which can be computed by:

$$\begin{aligned}\widetilde{p}(\mathbf{y}_i; q) &= \mathbb{E}\left\{\log\left[\frac{p(\mathbf{y}_i, \Theta)}{q(\Theta)}\right]\right\} \\ &= c + c'' - \frac{1}{2}(\mathbf{1}_{T-q} \otimes \phi)' \widehat{\mathbf{u}}_i - \frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\widehat{\theta}_i)' \widehat{C}_{\mathbf{u}_i}(\mathbf{y}_i - \mathbf{X}_i\widehat{\theta}_i) - \frac{1}{2}\text{tr}\left(\mathbf{X}_i' \widehat{C}_{\mathbf{u}_i} \mathbf{X}_i \widehat{\Sigma}_{\theta_i}\right) \\ &\quad - \frac{1}{2}\left(\widehat{\theta}_i' \widehat{V}_i^{-1} \widehat{\theta}_i\right) - \frac{1}{2}\text{tr}(\widehat{V}_i^{-1} \widehat{\Sigma}_{\theta_i}) - \sum_{j=1}^{k_i} \widehat{v_{i,j}^{-1}} \widehat{\lambda_{i,j}^{-1}} \\ &\quad - \widehat{\tau_i^{-1}} \widehat{\xi_i^{-1}} - \sum_{j=1}^{k_i} \widehat{v_{i,j}^{-1}} - \widehat{\xi_i^{-1}}\end{aligned}$$

where c'' are some constant consists of some known values.

doi:10.6342/NTU202402563

doi:10.6342/NTU202402563

Chapter 4 Conclusion and Outlooks

4.1 Conclusion

In this thesis, we provided one approach to estimate the parameters in the model which combines the VAR model and log-normal ARSV(q) as a random walk of the variance in VAR. Here we deal with the case that assumes ARSV model to have easy γ_i to simplify the computation, since ARSV(q) was set to satisfy $\log h_t = \gamma_0 + \sum_{i=1}^{p} \gamma_i \log h_{t-i} + \sigma\epsilon_t^{h_t}$, where we set $\gamma_0 = 0$ and $\gamma_i = 1$ for $1 \leq i \leq q$. But in fact, for any proper γ_i we can actually deal with it in similar way as long as we use substitution for $h_{i,0}, \ldots, h_{i,q}$ to $\mathbf{u}_{i,q}$. So we have given an approach for using ARSV(q) as our time-varying volatility instead of just ARSV(1), although it might have to be done in pseudo inverse sense. Also, it contains the horseshoe prior as a global-local shrinkage prior to perform the shrinkage for the coefficients of the model which has small effect and no shrinkage of the coefficients with large effect. Note that for each parameter, we find out a simple distribution that is often used and well-known such as normal distribution or inverse gamma distribution. As a result we might obtain the estimation of the 8 parameters in this thesis by using the variational Bayesian iteration.

4.2 Outlooks

In fact, GARCH is also popular recently as a way to describe the time-varying volatility, so it will be interesting to find out a similar result as we change the assumption of ARSV to GARCH, while some parameters might be different and maybe the prior should

doi:10.6342/NTU202402563

be changed too. Also, there are still many kinds of shrinkage prior such as Laplace, Student-t, Strawderman-Berger, Minnesota prior, adaptive LASSO, etc. Each of them provides different shrinkage effects and provides distinct parameters and prior to θ_i. In future we can work on different shrinkage prior and find out which approximation is more effective, which one deals better with the macroeconomic issue or will have lower computation cost, and which shrinkage prior can give the best explanation to the data.

doi:10.6342/NTU202402563

References

Amari, S. i. (2016), *Information geometry and its applications*, Vol. 194, Springer.

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

Bollerslev, T. (1986), 'Generalized autoregressive conditional heteroskedasticity', *Journal of econometrics* **31**(3), 307–327.

Carriero, A., Clark, T. E. and Marcellino, M. (2019), 'Large bayesian vector autoregressions with stochastic volatility and non-conjugate priors', *Journal of Econometrics* **212**(1), 137–154.

Carriero, A., Kapetanios, G. and Marcellino, M. (2012), 'Forecasting government bond yields with large bayesian vector autoregressions', *Journal of Banking & Finance* **36**(7), 2026–2047.

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2009), Handling sparsity via the horseshoe, *in* 'Artificial intelligence and statistics', PMLR, pp. 73–80.

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010), 'The horseshoe estimator for sparse signals', *Biometrika* **97**(2), 465–480.

Chan, J. C. and Yu, X. (2022), 'Fast and accurate variational inference for large bayesian vars with stochastic volatility', *Journal of Economic Dynamics and Control* **143**, 104505.

Choi, K., Yi, J., Park, C. and Yoon, S. (2021), 'Deep learning for anomaly detection in time-series data: Review, analysis, and guidelines', *IEEE access* **9**, 120043–120065.

doi:10.6342/NTU202402563

Cuaresma, J. C., Feldkircher, M. and Huber, F. (2016), 'Forecasting with global vector autoregressive models: A bayesian approach', *Journal of Applied Econometrics* **31**(7), 1371–1391.

Gefang, D. (2014), 'Bayesian doubly adaptive elastic-net lasso for var shrinkage', *International Journal of Forecasting* **30**(1), 1–11.

Gefang, D., Koop, G. and Poon, A. (2023), 'Forecasting using variational bayesian inference in large vector autoregressions with hierarchical shrinkage', *International Journal of Forecasting* **39**(1), 346–363.

Kalli, M. and Griffin, J. E. (2018), 'Bayesian nonparametric vector autoregressive models', *Journal of econometrics* **203**(2), 267–282.

Kullback, S. and Leibler, R. A. (1951), 'On information and sufficiency', *The annals of mathematical statistics* **22**(1), 79–86.

Makalic, E. and Schmidt, D. F. (2015), 'A simple sampler for the horseshoe estimator', *IEEE Signal Processing Letters* **23**(1), 179–182.

Miranda-Agrippino, S. and Ricco, G. (2021), 'The transmission of monetary policy shocks', *American Economic Journal: Macroeconomics* **13**(3), 74–107.

Ormerod, J. T. and Wand, M. P. (2010), 'Explaining variational approximations', *The American Statistician* **64**(2), 140–153.

Prüser, J. (2021), 'The horseshoe prior for time-varying parameter vars and monetary policy', *Journal of Economic Dynamics and Control* **129**, 104188.

Sharma, D. (2016), 'Nexus between financial inclusion and economic growth: Evidence from the emerging indian economy', *Journal of financial economic policy* **8**(1), 13–36.

Sims, C. A. (1980), 'Macroeconomics and reality', *Econometrica: journal of the Econometric Society* pp. 1–48.

Stock, J. H. and Watson, M. W. (2001), 'Vector autoregressions', *Journal of Economic perspectives* **15**(4), 101–115.

doi:10.6342/NTU202402563