Nowcasting GDP using dynamic factor model: A Bayesian approach

by

Yixiao Zhang

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Cindy Yu, Major Professor
Huaiqing Wu
Sergio Lence
Ulrike Genschel
Wolfgang Kliemann

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2020

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this dissertation. First and foremost, Dr. Cindy Yu for her guidance, patience and support throughout this research and the writing of this dissertation. Her insights and words of encouragement have often inspired me and renewed my hopes for completing my graduate education. I would additionally like to thank Dr. Haitao Li for his guidance throughout the research.

ABSTRACT

Real-time nowcasting is an assessment of current economic conditions from timely released economic series (such as monthly macroeconomic data) before the direct measure (such as quarterly GDP figure) is disseminated. Dynamic factor models (DFMs) are widely used in econometrics to bridge series with different frequencies and achieve a reduction in dimensionality. However, most of the research using DFMs often assumes the number of factors is known. In this dissertation, we first develop a Bayesian approach to provide a way to deal with unbalanced feature of the data set and to estimate latent common factors when the number of factors is assumed to be fixed and known. Then we extend our method such that it can identify the unknown number of factors and estimate the latent dynamic factors of DFMs accurately in a real-time nowcasting framework. The proposed method can deal with the unbalanced data, which is typical of a real-time nowcasting analysis. We demonstrate the validity of our approach through simulation studies and explore the applicability of our approach through empirical studies in nowcasting China's GDP or US GDP using monthly data series of several categories in each country's market respectively. The simulation studies and empirical studies indicate that our Bayesian approach is a viable option to conduct real-time nowcasting for China's and US's quarterly GDP.

CHAPTER 1. INTRODUCTION

A real time nowcasting is a process to assess or reconstruct current-quarter GDP from timely released economic and financial series before the figure is disseminated in order to gauge the overall macroeconomic conditions in real time. This is of interest because most data are released with a lag and are subsequently released. In principle, any release, no matter at what frequency, may potentially affect current-quarter estimates and their precision. Both forecasting and nowcasting are important tasks for central banks due to the following, including but not limited to, two reasons. Firstly, many policy decisions, including monetary policy, need to be made in real time and are based on assessments of current and future economic conditions. Secondly, to central banks, estimated current-quarter GDP figures are often used as relevant inputs for model-based longer-term forecasting exercises in the banks.

Real time nowcasting faces some challenges. The first one is how to bridge information contained in monthly data with the quarterly GDP. Baffigi et al. (2004), Rünstler and Sédillot (2003), Kitchen and Monaco (2003) study the idea of bridge equations which use small models to "bridge" the information contained in one or a few key monthly data series with the quarterly growth rate of GDP. However, they involve judgmental nowcasts and only deal with a few monthly data series. Then, how to deal with a large number of monthly data series becomes the second challenge. The use of factor models (FMs) for macroeconomic forecasting is now standard at central banks and other institutions. Many authors, such as Boivin and Ng (2005), Forni et al. (2005), D'Agostino and Giannone (2006), have shown that these models are successful in this regard. But FMs have not been used specifically for the problem of nowcasting in real time. The third challenge is that a large number of monthly data series are released at alternative times and with different lags, causing unbalanced data at the end of the sample. In real time, some data are released at the beginning of the month, some are in the middle, and some are at the end. Consequently,

the underlying data sets are unbalanced at end of the sample (i.e. at a real time when a new release happens). Some authors, including Croushore and Stark (2001), Koenig et al. (2003) and Orphanides (2002), discussed about this issue, but focused on data revisions and its implications, instead of statistical estimation. Appropriately dealing with this "jagged edge" feature of the data is the key for producing a nowcast by exploiting information in the most recent releases, and it has a chance to compete with judgmental nowcasts.

Giannone et al. (2008) provided a frequentist inference framework for the parametric dynamic factor models, and takes advantages of different data releases throughout the month and updates the nowcast based on each new data release. They use dynamic factor models (DFMs) to bridge monthly information with quarterly GDP and achieve a reduction in the dimensionality of the monthly data. The framework also formalizes the updating of the GDP nowcast as monthly data are released throughout the quarter. They combined principal component analysis (PCA) together with modified Kalman Filter (KF) to deal with the jagged edge feature of the data. Hereafter, we call the method proposed in Giannone et al. (2008) the GRS approach.

Since invented, the GRS approach has been implemented in many applications. Yiu and Chow (2011) nowcasted Chinese GDP using the GRS approach and discovered that interest rate data is the single most important category of economic series in estimating current-quarter GDP in China. Chernis and Sekkel (2017) showed that in a pseudo-real-time setting, the DFM outperformed univariate benchmarks as well as other commonly used nowcasting models, such as mixed-data sampling and bridge regressions, when nowcasting Canada's GDP growth. For the US market, the Federal Reserve Bank of New York published a platform called the New York Fed Staff Nowcast which has been estimating the US's GDP growth for the current and subsequent quarter, based on data released over the course of each week since April 2016. The behind-the-scenes methodology of the platform is built on the GRS approach and details can be found in Aarons et al. (2016) and Bok et al. (2018).

Based on the DFMs in Giannone et al. (2008), we propose a Bayesian Markov Chain Monte Carlo (MCMC) based inference framework which provides a more natural way to deal with the

"jagged edge" feature of the data and generates timely nowcast results of quarterly GDP in Chapter 2. Hereafter, we refer our Bayesian approach as the BAY method. There are some differences between the BAY method and the GRS method. One, the GRS approach parameterizes the variance matrix of monthly information data series to be a diagonal matrix to offer convenience of solving the unbalanced data issue, while we consider non-zero cross-sectional correlations. Two, the GRS approach estimates parameters and latent factors in multi-steps, i.e. first using PCA to obtain parameter estimates and then using KF to obtain latent factor estimates, while we combine estimation of all parameters and factors. We integrate all into one single estimation framework so that uncertainty of parameters and latent variables can be taken into account simultaneously and inferences are readily made based on posterior draws after the burn-in period. Through simulation studies, we evaluate our BAY approach based on the accuracy of estimated latent factors and nowcasting results. We also investigate the applicability of our approach by applying it to nowcast annual growth rates of China's quarterly GDP using monthly released data series in several categories, including industrial production, fixed asset investment, external sector, money market and financial market in China.

All the above applications of the GRS approach, along with the BAY approach proposed in Chapter 2 has two fundamental assumptions. They assume the number of factors of DFMs is fixed and given, while in reality it is unknown to us. In practice, people determine the number of factors by looking at the cumulative proportions of variances explained by the first few principal components from PCA. It is a rather subjective choice. The second key assumption is that the volatilities of macroeconomic series are treated as constant over time. In the past decade, macro empirical literature has been paying close attention to time-varying parameters. Many papers, such as Primiceri (2005), Cogley and Sargent (2005), and Benati and Surico (2008), mentioned that characterizing macroeconomic data with constant parameter models is deficient, and the form of slow, continuous, and time-varying parameters is much more desirable. Negro and Otrok (2008) bridged the literature on factor models with the literature on parameter instability by considering DFMs with time-varying factor loadings and stochastic volatility (SV). Clark (2011) focused on

adding SV to density forecasts of US's GDP growth, unemployment, inflation, and the federal funds rate from Bayesian Vector Autoregression (BVAR) analyses and discovered material improvements in the real-time accuracy of density forecasts and the accuracy of point forecasts when SV was taken into consideration.

We relax the above two assumptions, allowing the DFM to have an unknown number of factors in Chapter 3, and the volatility of macroeconomic data to vary over time in Chapter 4. We make modifications and improvements to our BAY approach to handling these two changes. Time-varying factor loadings are beyond the scope of this dissertation. The improved BAY approach addresses the unknown number of latent factors and SV by consolidating ideas from prior work. Zhang et al. (2013) proposed a Bayesian method of estimating the covariance matrices in the form of a factor model with an unknown number of latent factors by introducing binary indicators for factor selections. We adopt the idea and introduce a binary indicator to a set of candidate latent factors. If the data suggests a certain factor to be selected, the binary indicator will return 1 for that factor, and 0 otherwise. By counting how many 1's, we get an estimate of the number of latent factors. Follett and Yu (2019) introduced a SV model in the framework of vector autoregression (VAR) that involves the comovement of the time-changing variances across series. The time-varying volatility is achieved by allowing a static correlation matrix generated from an Lewandowski, Kurowicka, and Joe (LKJ) prior proposed in Lewandowski et al. (2009), and a random walk process for the variance of each individual time series. They also proposed to use the particle Gibbs with backward simulation algorithm to estimate the time-varying volatility parameters effectively. The alogrithm is directly used in our MCMC algorithm to generate posterior samples for SV.

We are not the first one to consider estimating the number of factors. Bai and Ng (2002, 2006) proposed a class of information criteria and showed the number of factors could be consistently estimated using those criteria in a large panel of data setting in DFMs. Alternative methods that involve the calculation of eigenvalues of the sample covariance matrix are available, including Onatski (2009, 2010) and Ahn and Horenstein (2013) who considered the estimation of the number of factors in approximate factor models and generalized dynamic factor structure. Our BAY ap-

proach is fundamentally different from these approaches in the following aspects. First, we resolve the problem of unknown number of factors using a Bayesian framework, while they approached the question using a frequentist view. Second, we estimate the number of factors within a nowcasting set-up, while they isolated the question on its own and did not consider forecasting/nowcasting. Third, we consider time-varying volatility while they assumed constant volatility. Through simulation studies, we evaluate our BAY approach based its accuracy in estimating the unknown number of factors and the latent state variables (dynamic factors and SV), and its effectiveness in producing reliable nowcasts in real time. We also investigate the applicability of our approach by applying it to nowcast US's quarterly GDP growth using monthly released data series in several categories in the US market.

The rest of this dissertation is organized as follows. In Chapter 2, we present the results from the simulation and empirical studies using Chinese market data when the number of latent factors is assumed to be fixed and known. In Chapter 3, we show the results from the simulation and empirical studies using US's market data when the number of latent factors is unknown. In Chapter 4 we extend the model in pervious chapter to allow time-varying volatility and show results from the simulation and empirical studies using US's market data based on extended SV model. Chapter 5 concludes the dissertation.

CHAPTER 2. NOWCASTING GDP USING DYNAMIC FACTOR MODEL WITH KNOWN NUMBER OF FACTORS

In this chapter, we use dynamic factor models to bridge monthly information with quarterly GDP and achieve reduction in the dimensionality of the monthly data assuming the number of factors is fixed and known. We develop a Bayesian approach to provide a way to deal with unbalanced feature of the dataset and to estimate latent common factors. We demonstrate the validity of our approach through simulation studies, and explore the applicability of our approach through an empirical study in nowcasting China's GDP using 117 monthly data series of several categories in Chinese market.

2.1 Structure of Dataset and the GRS Approach

In this section, we will first describe the problem in a stylized way. The goal is to evaluate the current quarter nowcast of GDP based on the flow of information that becomes available during the quarter. Then we will review the frequentist framework proposed by Giannone et al. (2008) in order to build a foundation for our BAY approach.

2.1.1 Working With Unbalanced Data

In real time at a particular release date, some series have observations through the current month, whereas for others the most recent observations are from the previous month. Consequently, the underlying datasets are unbalanced. Appropriately dealing with this unbalanced feature of the data is key for nowcasting.

Let t be the index for month and k be the index for quarter. Let $\mathbf{x}_t = (x_{1,t}, ..., x_{n,t})'$ be a $n \times 1$ vector denoting n monthly data series at month t, and y_k be quarterly GDP at quarter k. Assume $\mathbf{x}_t$ are released at Q different dates of month t. Each releasing date is denoted as (q, t), representing

the q^{th} release date in month t, where $q = 1, ...Q$. The releasing set $v^*_{q,t}$ contains indexes of n^*_q monthly series that are released at the release date (q,t), where $n^*_q = ||v^*_{q,t}||$ denoting the number of newly released monthly series. Let $v_{q,t}$ denote the set collecting indexes of all $x_{i,t}$'s that have been released at or before the release date (q,t), that is $v_{q,t} = \bigcup_{i \leq q} v^*_{i,t}$, and $n_q = \sum_{i \leq q} n^*_i$ is the number of available series at (q,t). Thus $\mathbf{x}_{i \in v_{q,t}}$ represents monthly series that are available at the release date (q,t). Without loss of generality, we assume the release dates for all series are fixed across months.

Table 2.1 illustrates how the data is released in a given quarter using a toy example. Consider there are $n = 6$ monthly series, $\mathbf{x}_t = (x_{1,t}, x_{2,t}, x_{3,t}, x_{4,t}, x_{5,t}, x_{6,t})'$, released at three different dates (i.e. $Q = 3$). Without loss of generality, $\mathbf{x}_{it}$ can always be arranged such that index i follows the order of releasing dates. Suppose $\{x_{1,t}, x_{2,t}, x_{3,t}\}$, with the dark gray background, are released at the first releasing date $(1,t)$, so the releasing set $v^*_{1,t} = \{1,2,3\}$. Since this is the very first releasing date so the available set $v_{1,t}$ is also $\{1,2,3\}$ and the available monthly series $\mathbf{x}_{i \in v_{1,t}}$ is $\{x_{1,t}, x_{2,t}, x_{3,t}\}$. Then at the second releasing date $(2,t)$, $\{x_{4,t}, x_{5,t}\}$, with the gray background, are released, hence the releasing set $v^*_{2,t} = \{4,5\}$. The available set $v_{2,t}$ is $\{1,2,3,4,5\}$ so $\mathbf{x}_{i \in v_{2,t}} = \{x_{1,t}, x_{2,t}, x_{3,t}, x_{4,t}, x_{5,t}\}$. At the last releasing date $(3,t)$, the last series $x_{6,t}$, with the light gray background, becomes available, so $v^*_{3,t} = \{6\}$, $v_{3,t} = \{1,2,3,4,5,6\}$, and $\mathbf{x}_{i \in v_{3,t}} = \{x_{1,t}, x_{2,t}, x_{3,t}, x_{4,t}, x_{5,t}, x_{6,t}\}$.

Table 2.1: Data Releasing Format in Current Quarter. Dark gray cells mean $\mathbf{x}_{i \in v^*_{1,t}}$, gray cells mean $\mathbf{x}_{i \in v^*_{2,t}}$, light gray cell means $\mathbf{x}_{i \in v^*_{3,t}}$, white cells mean unreleased data series at (q,t).

(q,t)	$(1,t)$	$(2,t)$	$(3,t)$
$\mathbf{x}_{i \in v_{q,t}}$	$x_{1,t}$	$x_{1,t}$	$x_{1,t}$
	$x_{2,t}$	$x_{2,t}$	$x_{2,t}$
	$x_{3,t}$	$x_{3,t}$	$x_{3,t}$
	NA	$x_{4,t}$	$x_{4,t}$
	NA	$x_{5,t}$	$x_{5,t}$
	NA	NA	$x_{6,t}$

Table 2.2 gives an overall picture of the entire dataset available when doing nowcasting. Particularly, we consider the first release date of the second month in the current quarter as an example. Suppose we want to nowcast GDP in the current quarter y_{K+1} using all monthly information up through month T (the end of the sample), where $T = 3K + 1$ (means the first month nowcast), or $T = 3K + 2$ (means the second month nowcast), or $T = 3K + 3$ (means the third month nowcast). The observations available to use at the release date (q, t), highlighted in gray color in Table 2.2, include $\{y_1, y_2, ..., y_K\}$ and $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{T-1}, \mathbf{x}_{i \in v_{q,T}}\}$. The goal is to nowcast y_{K+1} using all information available at every release date (q, t) in the current quarter, i.e. real time nowcasting. Note that the factors $\{\mathbf{F}_1, ..., \mathbf{F}_T\}$ are not observed, and need to be estimated using the adjusted Kalman Filter (e.g. the GRS approach), or using the Bayesian smoothing techniques (e.g. our BAY approach). At every new release date (q, t) $(q = 1, ..., Q)$, model parameters and $\{\mathbf{F}_1, ..., \mathbf{F}_T\}$ are updated with this additional information from the new release, and nowcast of y_{K+1} is re-produced. Therefore there are $3Q$ nowcast results in the current quarter.

Table 2.2: Overall picture of the entire dataset available at the first release date of the second month in the current quarter. Cells highlighted in gray color indicates the corresponding variables in the cells are available.

k	1			2			$\cdots$	K			K+1		
t	1	2	3	4	5	6	$\cdots$	$T-4$	$T-3$	$T-2$	$T-1$	T	$T+1$
$\mathbf{x}_{i,t}$	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	$x_{1,5}$	$x_{1,6}$	$\cdots$	$x_{1,T-4}$	$x_{1,T-3}$	$x_{1,T-2}$	$x_{1,T-1}$	$x_{1,T}$	NA
	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	$x_{2,5}$	$x_{2,6}$	$\cdots$	$x_{2,T-4}$	$x_{2,T-3}$	$x_{2,T-2}$	$x_{2,T-1}$	$x_{2,T}$	NA
	$x_{3,1}$	$x_{3,2}$	$x_{3,3}$	$x_{3,4}$	$x_{3,5}$	$x_{3,6}$	$\cdots$	$x_{3,T-4}$	$x_{3,T-3}$	$x_{3,T-2}$	$x_{3,T-1}$	$x_{3,T}$	NA
	$x_{4,1}$	$x_{4,2}$	$x_{4,3}$	$x_{4,4}$	$x_{4,5}$	$x_{4,6}$	$\cdots$	$x_{4,T-4}$	$x_{4,T-3}$	$x_{4,T-2}$	$x_{4,T-1}$	NA	NA
	$x_{5,1}$	$x_{5,2}$	$x_{5,3}$	$x_{5,4}$	$x_{5,5}$	$x_{5,6}$	$\cdots$	$x_{5,T-4}$	$x_{5,T-3}$	$x_{5,T-2}$	$x_{5,T-1}$	NA	NA
	$x_{6,1}$	$x_{6,2}$	$x_{6,3}$	$x_{6,4}$	$x_{6,5}$	$x_{6,6}$	$\cdots$	$x_{6,T-4}$	$x_{6,T-3}$	$x_{6,T-2}$	$x_{6,T-1}$	NA	NA
y_k	y_1			y_2			$\cdots$	y_K			y_{K+1}		
$\mathbf{F}_t$	$\mathbf{F}_1$	$\mathbf{F}_2$	$\mathbf{F}_3$	$\mathbf{F}_4$	$\mathbf{F}_5$	$\mathbf{F}_6$	$\cdots$	$\mathbf{F}_{T-4}$	$\mathbf{F}_{T-3}$	$\mathbf{F}_{T-2}$	$\mathbf{F}_{T-1}$	$\mathbf{F}_T$	

How to handle the unbalanced feature of the monthly data series will be discussed in Section 2.1.2 for the GRS approach, and in Section 2.2.2 for our BAY approach.

2.1.2 The GRS Approach

Since there are numerous series in the information set, modeling y directly on all $\mathbf{x}$ would involve too many parameters, and hence the model performs poorly in nowcasting/forecasting because of the large uncertainty in parameters' estimation ("the curse of dimensionality"). The fundamental idea of Giannone et al. (2008) is to explore the collinearity of the series by summarizing all available information into a few common factors. Because of collinearity, a linear combination of the common factors is able to approximate the dynamic interactions among the series and to postulate a parsimonious model that works well in nowcasting/forecasting. The GRS approach formulates the model in the following ways.

First, they assume the monthly data series is a linear function of a few unobserved common factors $\mathbf{F}_t$,

$$\mathbf{x}_t = \boldsymbol{\mu} + \boldsymbol{\Theta}\mathbf{F}_t + \boldsymbol{\epsilon}_t, \tag{2.1}$$

where $\mathbf{x}_t = (x_{1,t}, ..., x_{n,t})'$ be a $n \times 1$ monthly data series at month t, for $t = 1, ..., T$, $\mathbf{F}_t = \{f_{1t}, \cdots, f_{rt}\}'$ be a $r \times 1$ monthly common factors at month t, $\boldsymbol{\Theta}$ is the $n \times r$ factor loading matrix, $\boldsymbol{\mu}$ is the mean vector, and $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Omega}_{n\times n})$. The number of latent factors, r, is assumed to be known and $r << n$. Then they further specify the dynamics of the common factors as follows,

$$\mathbf{F}_t = \mathbf{A}\mathbf{F}_{t-1} + \mathbf{u}_t, \tag{2.2}$$

where $\mathbf{A}$ is a $r \times r$ matrix and all roots of $det(\mathbf{I}_r - \mathbf{A}z)$ lie outside the unit circle, and $\mathbf{u}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{r\times r})$. Finally, the quarterly GDP is assumed to be a linear function of the common factors in the third month of the quarter,

$$y_k = \beta_0 + \boldsymbol{\beta}_1'\mathbf{F}_{3k} + \nu_k, \tag{2.3}$$

where β_0 is a scalar, $\boldsymbol{\beta}_1$ is a $r \times 1$ vector, $v_k \sim N(0, \eta^2)$, and $k = 1, ..., K$. We assume $3K + 1 \leq T \leq 3K + 3$. The dynamic factor model specified this way not only can bridge the information contained in monthly data series with the quarterly GDP, but also helps reduce the dimension of parameters, thus increasing the degrees of freedom.

If the complete set of monthly data series is observed, the unobserved common factors $\mathbf{F}_t$ $(t = 1, ..., T)$ could be consistently estimated by PCA as recently shown by several authors in literature. However, when doing real time nowcasting, the dataset is unbalanced and we want to exploit the additional information from a newly released set, which requires dealing with missing data at the end of the sample. To overcome this difficulty, the GRS estimates parameters and common factors $\mathbf{F}_t$ $(t = 1, ..., T)$ based on the following three stages. The first stage uses an OLS regression on principal components (PCs) extracted from a balanced panel that has truncated at the previous month $T - 1$, i.e. the balanced panel is $\{\mathbf{x}_1, ..., \mathbf{x}_{T-1}\}$ in Table 2.2. And the second stage adjusts the Kalman smoother based on the estimated parameters from the first stage in order to deal with the unbalanceness of data. More specifically, at the release date (q, T), the variance covariance matrix used in the Kalman filter is defined as $\tilde{\mathbf{\Omega}}_{v_{q,T}} = diag(\tilde{w}_{11}^2, ..., \tilde{w}_{nn}^2)$, where

$$\tilde{w}_{ii}^2 = \begin{cases} w_{ii}^2 \text{ if } i \in v_{q,T} \\ \infty \text{ if } i \notin v_{q,T} \end{cases}, \text{ for } i = 1, ..., n, \tag{2.4}$$

w_{ii}^2 is the i^{th} diagonal element of covariance matrix for $\boldsymbol{\epsilon}_t$, estimated from the balanced panel data in the first stage. And then the Kalman Filter algorithm is applied on the entire (unbalanced) data, i.e. $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{T-1}, \mathbf{x}_{i \in v_{q,T}}\}$, using $\tilde{\mathbf{\Omega}}_{v_{q,T}}$. In this way, Giannone et al. (2008) argues that the filter, through its implicit signal extraction process, will put no weight on missing observations in the computation of the factors. Then the third stage estimates the coefficients in equation (2.3) by OLS regression of GDP on the latent factors estimated by the Kalman filtering. Readers can refer to Giannone et al. (2008) for more details.

2.2 Alternative Approach to Nowcasting: Bayesian Markov Chain Monte Carlo Method

In this section, we discuss some restrictions imposed on the model to avoid the non-identifiable issue of common factors $\mathbf{F}_t$, introduce our BAY approach to estimate model parameters and latent factors $\mathbf{F}_t$, and provide formulas for nowcasting using both GRS and BAY approaches.

2.2.1 Model Modifications

It is known that dynamic factor models suffer from a non-identifiable issue. Following Stock and Watson (2002), two sets of assumptions, Assumption F and Assumption M are constructed (details are in Appendix A.1). Specifically in this chapter, some restrictions are imposed on matrices $\mathbf{A}$ and $\mathbf{\Sigma}$ in the dynamics of latent factors in equation (2.2) as follows, $\mathbf{A} = diag(a_1, a_2, ..., a_r)$ and $\mathbf{\Sigma} = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_r^2)$, where $|a_j| < 1$ $(j = 1, ..., r)$ and $\sigma_i^2/(1 - a_i^2) > \sigma_j^2/(1 - a_j^2)$, $\forall i < j$. These restrictions together with the prior specification for loading matrix $\mathbf{\Theta}$ (discussed later) satisfy the identification assumption F1 in Appendix A.1. Stock and Watson (2002) also shows that this assumption identifies the factors up to a change of sign.

For comparison with the random walk approach, we also consider allowing the quarterly GDP y_k to depend on lags of latent factors and GDP itself. The model considered in our article can be summarized by the following:

$$\begin{aligned} \mathbf{x}_t &= \boldsymbol{\mu} + \mathbf{\Theta}\mathbf{F}_t + \boldsymbol{\epsilon}_t, \text{ for } t = 1, ..., T \\ \mathbf{F}_t &= \mathbf{A}\mathbf{F}_{t-1} + \mathbf{u}_t, \text{ for } t = 1, ..., T \\ y_k &= \beta_0 + \boldsymbol{\beta}_1'\mathbf{F}_{3k} + \boldsymbol{\beta}_2'\mathbf{F}_{3k-1} + \boldsymbol{\beta}_3'\mathbf{F}_{3k-2} + \beta_4 y_{k-1} + \nu_k, \text{ for } k = 1, ..., K \end{aligned} \tag{2.5}$$

where $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \mathbf{\Omega}_{n\times n})$, $\mathbf{u}_t \sim N(\mathbf{0}, \mathbf{\Sigma}_{r\times r})$, $v_k \sim N(0, \eta^2)$, and matrices $\mathbf{A}$ and $\mathbf{\Sigma}$ have the restrictions discussed in the beginning of this subsection, which implies $f_{j,t} = a_j f_{j,t-1} + \sigma_j u_{j,t}$ for $j = 1, ..., r$ and $u_{j,t} \sim N(0, 1)$.

At the release date (q, T), we have observations $\mathbf{Y} = \{y_1, y_2, ..., y_K\}$ and $\mathbf{X}_{(q,T)} = \{\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_{T-1}, \mathbf{x}_{i\in v_{q,T}}\}$; latent variables $\mathbf{F} = \{\mathbf{F}_1, \mathbf{F}_2, ..., \mathbf{F}_T\}$; and model parameters $\mathbf{\Psi} = \{\boldsymbol{\mu}, \mathbf{\Theta}, \mathbf{\Omega}, \mathbf{A}, \mathbf{\Sigma}, \beta_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \beta_4, \eta^2\}$. The goals are to estimate $\mathbf{\Psi}$ and $\mathbf{F}$ using the observables and to nowcast the current quarter GDP y_{K+1} at every release date (q, T) for $q = 1, ..., Q$. The original GRS approach is adjusted to accommodate such model modifications.

2.2.2 Estimating Dynamic Factor Models via a Bayesian MCMC approach

We face a couple of challenges in estimating the above model. First, it is computationally infeasible to integrate out the high-dimensional latent variable $\mathbf{F}$ to obtain the likelihood based only on observables. Second, the observations in the panel of $\mathbf{X}_{(q,T)} = \{\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_{T-1}, \mathbf{x}_{i\in v_{q,T}}\}$ are not balanced. To overcome these difficulties, we develop a computational Bayesian Markov Chain Monte Carl (MCMC) approach for estimating the above dynamic factor model in real time. MCMC conducts inferences by simulating efficiently from (potentially complicated) posterior distributions of model parameters and latent variables given the observables. MCMC samples from the typically high-dimensional and complex posterior distributions by generating a Markov Chain over parameters and latent variables whose equilibrium distribution is the desired posterior distribution. The Monte Carlo method uses these samples for numerical integration for parameter and state estimation.

In order to facilitate the derivation of the joint posterior distribution, the dynamic of $\mathbf{x}_t$ in equation (2.5) is rewritten as

$$\mathbf{x}_t = \boldsymbol{\mu} + [\mathbf{I}_{n\times n} \otimes \mathbf{F}_t'] * \boldsymbol{\theta} + \boldsymbol{\epsilon}_t,$$

where $\boldsymbol{\theta} = vec(\boldsymbol{\Theta}) = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n)'$ if $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1', ..., \boldsymbol{\theta}_n')'$ such that $\boldsymbol{\theta}_i, (i = 1, ..., n)$ is a $1 \times r$ vector representing the i^{th} row of $\boldsymbol{\Theta}$. Thus the conditional density of $\mathbf{x}_t$ in equation (2.5) is

$$\mathbf{x}_t|\mathbf{F}_t, \boldsymbol{\Theta}, \boldsymbol{\Omega} \sim N(\boldsymbol{\mu} + [\mathbf{I}_{n\times n} \otimes \mathbf{F}_t'] * \boldsymbol{\theta}, \boldsymbol{\Omega}), \text{ for } t = 1, ..., T-1, \tag{2.6}$$

the conditional density of $\mathbf{F}_t$ in equation (2.5) is

$$\mathbf{F}_t|\mathbf{F}_{t-1}, \mathbf{A}, \boldsymbol{\Sigma} \sim N(\mathbf{A}\mathbf{F}_{t-1}, \boldsymbol{\Sigma}), \text{ for } t = 2, ..., T, \tag{2.7}$$

and the conditional density of y_k in equation (2.5) is

$$y_k|\boldsymbol{\beta}, \mathbf{F}_{3k}, \mathbf{F}_{3k-1}, \mathbf{F}_{3k-2}, y_{k-1}, \eta^2 \sim N(\beta_0 + \boldsymbol{\beta}_1'\mathbf{F}_{3k} + \boldsymbol{\beta}_2'\mathbf{F}_{3k-1} + \boldsymbol{\beta}_3'\mathbf{F}_{3k-2} + \beta_4 y_{k-1}, \eta^2), \tag{2.8}$$

for $k = 1, ..., K$. The joint posterior distribution, $p(\mathbf{\Psi}, \mathbf{F}|\mathbf{Y}, \mathbf{X}_{(q,T)})$, can be decomposed into products of individual conditionals,

$$\begin{aligned} & p(\mathbf{\Psi}, \mathbf{F}|\mathbf{Y}, \mathbf{X}_{(q,T)}) \propto p(\mathbf{Y}, \mathbf{X}_{(q,T)}, \mathbf{\Psi}, \mathbf{F}) \\ \propto & \prod_{t=1}^{T-1} p(\mathbf{x}_t|\mathbf{F}_t, \boldsymbol{\theta}, \mathbf{\Omega}) p(\mathbf{x}_{i\in v_{q,T}}|\mathbf{F}_T, \boldsymbol{\theta}, \mathbf{\Omega}) \times \prod_{t=2}^{T} p(\mathbf{F}_t|\mathbf{F}_{t-1}, \mathbf{A}, \mathbf{\Sigma}) \\ \times & \prod_{k=2}^{K} p(y_k|\boldsymbol{\beta}, \mathbf{F}_{3k}, \mathbf{F}_{3k-1}, \mathbf{F}_{3k-2}, y_{k-1}, \eta^2) \times \pi(\mathbf{\Psi}), \end{aligned} \tag{2.9}$$

where $p(\mathbf{x}_t|\mathbf{F}_t, \boldsymbol{\theta}, \mathbf{\Omega})$, $p(\mathbf{F}_t|\mathbf{F}_{t-1}, \mathbf{A}, \mathbf{\Sigma})$, and $p(y_k|\boldsymbol{\beta}, \mathbf{F}_{3k}, \mathbf{F}_{3k-1}, \mathbf{F}_{3k-2}, y_{k-1}, \eta^2)$ are given according to the distributions in equations (2.6), (2.7) and (2.8) respectively. Here $\pi(\mathbf{\Psi})$ is the prior distribution for $\mathbf{\Psi}$, which will be specified later.

To deal with the missing data in $\mathbf{x}_{i\in v_{q,T}}$ at the end of the sample, we define an indicator matrix $\mathbf{1}_{v_{q,T}}$ as a $n_q \times n$ matrix obtained by deleting the i^{th} row from the identity matrix $\mathbf{I}_{n\times n}$ if the corresponding $x_{i,T}$ is missing at time (q, T), where i could be any index from $i = 1, ..., n$. For the toy example discussed in Section 2.1.1, we have

$$\mathbf{1}_{v_{1,T}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{1}_{v_{2,T}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \quad \mathbf{1}_{v_{3,T}} = \mathbf{I}_{6\times 6}.$$

Then we can write $\mathbf{x}_{i\in v_{q,T}}$ as $\mathbf{x}_{i\in v_{q,T}} = \mathbf{1}_{v_{q,T}}\mathbf{x}_T$, thus the conditional density of $\mathbf{x}_{i\in v_{q,T}}$ in equation (2.9) is

$$\mathbf{x}_{i\in v_{q,T}} = \mathbf{1}_{v_{q,T}}\mathbf{x}_T|\mathbf{F}_T, \mathbf{\Theta}, \mathbf{\Omega} \sim N(\mathbf{1}_{v_{q,T}}(\boldsymbol{\mu} + [\mathbf{I}_{n\times n} \otimes \mathbf{F}_T'] * \boldsymbol{\theta}), \mathbf{1}_{v_{q,T}}\mathbf{\Omega}\mathbf{1}'_{v_{q,T}}). \tag{2.10}$$

Note this remedy for the unbalanced data issue is different from GRS which restricts matrix $\tilde{\mathbf{\Omega}}$ to be diagonal with ∞ variances for the missing series in order to implement the Kalman Filtering algorithm. We complete the model specification by assigning prior distributions for the parameter set $\mathbf{\Psi}$ in Bayesian framework.

We set the prior for $\mathbf{\Theta}$ as,

$$\mathbf{\Theta}_{n\times r} \sim Matrix\ Normal(\mathbf{0}_{n\times r}, \mathbf{I}_{n\times n}, \mathbf{I}_{r\times r}). \tag{2.11}$$

In this way, the identification assumption F1 in Appendix A.1 is satisfied by the law of large number.

The prior for $\mathbf{\Omega}$ is

$$\mathbf{\Omega}_{n\times n} \sim Inverse\ Wishart\Big(\frac{1}{n}\mathbf{I}_{n\times n}, \nu_\theta\Big), \tag{2.12}$$

where ν_θ is a pre-specified scaler. This prior allows $\boldsymbol{\epsilon}_t$ to be correlated across series, satisfying assumption M1 in Appendix A.1.

The prior for $\mathbf{A}$ is standard normal truncated at $[-1, 1]$, for $j = 1, ..., r$, that is

$$\pi(a_j) = \begin{cases} 0 & \text{if } a_j \leq -1 \\ \dfrac{\phi(a_j)}{\Phi(1) - \Phi(-1)} & \text{if } -1 < a_j < 1 \\ 0 & \text{if } a_j \geq 1 \end{cases}, \tag{2.13}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are PDF and CDF for standard normal distribution respectively. And $\pi(\mathbf{A}) = \prod_{j=1}^{r} \pi(a_j)$.

The prior for $\mathbf{\Sigma}$ is, for $j = 1, ..., r$,

$$\sigma_j^2 \overset{iid}{\sim} Inverse\ Gamma(\alpha_s, \beta_s), \tag{2.14}$$

where α_s, β_s are prespecified scalers. And $\pi(\mathbf{\Sigma}) = \prod_{j=1}^{r} \pi(\sigma_j)$.

The prior for $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \boldsymbol{\beta}_3', \beta_4)'$ is

$$\boldsymbol{\beta}_{(3r+2)\times 1} \sim N(\mathbf{0}_{(3r+2)\times 1}, \mathbf{I}_{(3r+2)\times(3r+2)}). \tag{2.15}$$

The prior for η^2 is

$$\eta^2 \sim Inverse\ Gamma(\alpha_h, \beta_h), \tag{2.16}$$

where α_h, β_h are pre-specified scalers.

We assume all priors are independent. Following standard MCMC procedure, we derive the complete conditional distributions for each parameter and latent variable, and obtain posterior samples

by simulating from these individual complete conditionals iteratively. More specifically, we obtain the posterior distribution $p(\Psi_i|\mathbf{\Psi}_{-i}, \mathbf{Y}, \mathbf{X}_{(q,T)}, \mathbf{F})$ where Ψ_i is the i^{th} element of $\mathbf{\Psi}$ and $\mathbf{\Psi}_{-i}$ contains all the parameters except for Ψ_i, and the posterior distribution for factors $p(\mathbf{F}_t|\mathbf{\Psi}, \mathbf{Y}, \mathbf{X}_{(q,T)})$ for all t. In estimation, we draw posterior samples from the above complete conditional distributions and use the means of the posterior samples as parameter estimates and the standard deviations of the posterior samples as standard errors of the parameter estimates. Appendix A.2 provides the posterior distributions for all model parameters and latent factors.

The GRS approach estimates parameters and latent factors in multi-stages, i.e. first using PCA to obtain parameter estimates, then using KF to obtain latent factor estimates and lastly using OLS to estimate coefficients $\boldsymbol{\beta}$'s in the dynamics of quarterly GDP. However, we integrate estimation of all parameters and factors into a single framework so that uncertainty of parameters and latent variables can be taken into account simultaneously and inferences are readily made based on posterior draws after the burn-in period. Also note that in the GRS approach, only information from $\mathbf{X}$ enters into estimation of $\mathbf{F}$ in its PCA and Kalman Filter stages, while our BAY approach uses both information from monthly series $\mathbf{X}$ and quarterly GDP $\mathbf{Y}$ to update $\mathbf{F}$ (see details in Appendix A.2).

2.2.3 Nowcasting GDP y_{K+1}

Recall that we have complete $\mathbf{x}_t$ for $t = 1, ..., T-1$, and associated y_k for $k = 1, ..., K$ are also available. Suppose we are at (q, T) in month T $(q = 1, ..., Q)$, the task is to nowcast y_{K+1} based on $\{\mathbf{x}_1, ..., \mathbf{x}_{T-1}, \mathbf{x}_{i\in v_{q,T}}\}$ and $\{y_1, ..., y_K\}$. Here T can be the first $(T = 3K+1)$, second $(T = 3K+2)$ or even third $(T = 3K+3)$ month of quarter $K+1$. Let $\hat{\beta}_0$, $\hat{\boldsymbol{\beta}}_i$ $(i = 1, 2, 3)$, $\hat{\beta}_4$ and $\hat{\mathbf{F}}_t$ $(t = 1, ..., T)$ are estimated parameters and latent common factors from the GRS approach, and let $\beta_0^{(g)}$, $\boldsymbol{\beta}_i^{(g)}$ $(i = 1, 2, 3)$, $\beta_4^{(g)}$ and $\mathbf{F}_t^{(g)}$ $(t = 1, ..., T)$ be the g^{th} posterior draws of parameters and latent factors after the burn-in period in the BAY approach, where $g = 1, ..., G$. The nowcast can be calculated as follows.

- In the first month, i.e. when $T = 3K + 1$, the nowcast of $K + 1$ quarterly GDP using BAY and GRS are given by:

$$\hat{y}_{K+1}^{BAY} = \frac{1}{G}\sum_{g=1}^{G}\Big[\beta_0^{(g)} + (\boldsymbol{\beta}_1^{(g)})'(\mathbf{A}^{(g)})^2\mathbf{F}_T^{(g)} + (\boldsymbol{\beta}_2^{(g)})'\mathbf{A}^{(g)}\mathbf{F}_T^{(g)} + (\boldsymbol{\beta}_3^{(g)})'\mathbf{F}_T^{(g)} + \beta_4^{(g)}y_K\Big], \quad (2.17a)$$

$$\hat{y}_{K+1}^{GRS} = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}_1'\hat{\mathbf{A}}^2\hat{\mathbf{F}}_T + \hat{\boldsymbol{\beta}}_2'\hat{\mathbf{A}}\hat{\mathbf{F}}_T + \hat{\boldsymbol{\beta}}_3'\hat{\mathbf{F}}_T + \hat{\beta}_4 y_K. \quad (2.17b)$$

- In the second month, i.e. when $T = 3K + 2$, the nowcast of $K + 1$ quarterly GDP using BAY and GRS are given by:

$$\hat{y}_{K+1}^{BAY} = \frac{1}{G}\sum_{g=1}^{G}\Big[\beta_0^{(g)} + (\boldsymbol{\beta}_1^{(g)})'\mathbf{A}^{(g)}\mathbf{F}_T^{(g)} + (\boldsymbol{\beta}_2^{(g)})'\mathbf{F}_T^{(g)} + (\boldsymbol{\beta}_3^{(g)})'\mathbf{F}_{T-1}^{(g)} + \beta_4^{(g)}y_K\Big], \quad (2.18a)$$

$$\hat{y}_{K+1}^{GRS} = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}_1'\hat{\mathbf{A}}\hat{\mathbf{F}}_T + \hat{\boldsymbol{\beta}}_2'\hat{\mathbf{F}}_T + \hat{\boldsymbol{\beta}}_3'\hat{\mathbf{F}}_{T-1} + \hat{\beta}_4 y_K. \quad (2.18b)$$

- In the third month, i.e. when $T = 3K + 3$, the nowcast of $K + 1$ quarterly GDP using BAY and GRS are given by:

$$\hat{y}_{K+1}^{BAY} = \frac{1}{G}\sum_{g=1}^{G}\Big[\beta_0^{(g)} + (\boldsymbol{\beta}_1^{(g)})'\mathbf{F}_T^{(g)} + (\boldsymbol{\beta}_2^{(g)})'\mathbf{F}_{T-1}^{(g)} + (\boldsymbol{\beta}_3^{(g)})'\mathbf{F}_{T-2}^{(g)} + \beta_4^{(g)}y_K\Big], \quad (2.19a)$$

$$\hat{y}_{K+1}^{GRS} = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}_1'\hat{\mathbf{F}}_T + \hat{\boldsymbol{\beta}}_2'\hat{\mathbf{F}}_{T-1} + \hat{\boldsymbol{\beta}}_3'\hat{\mathbf{F}}_{T-2} + \hat{\beta}_4 y_K. \quad (2.19b)$$

Note that all of these $\hat{\mathbf{F}}_t$'s for the GRS, or $\boldsymbol{\beta}_i^{(g)}$'s and $\mathbf{F}_t^{(g)}$ for the BAY, are updated in every release date within a month. And then $\hat{y}_{K+1}^{GRS}$ and $\hat{y}_{K+1}^{BAY}$ are re-produced for each additional release date. The superscript (q, T) in both $\hat{y}_{K+1}^{GRS}$ and $\hat{y}_{K+1}^{BAY}$ has been suppressed to simplify notations.

2.3 Bayesian Approach in Nowcasting: Simulation Evidence

In this section, through numerical simulations, we investigate two questions on Bayesian analysis of dynamic factor models. The first question is whether it can identify the latent factors $\mathbf{F}_t$ accurately. The second question is whether it can produce reliable nowcasting results. Specifically, we compare in-sample estimation of $\mathbf{F}_t$ and out-of-sample nowcasting performance of the BAY and the GRS approaches, when addressing these two questions.

In each of 2 simulation studies below, we generate data according to the model in (2.5) where $T = 210$ (months), $K = 70$ (quarters), and $n = 60$ (monthly data series) with $Q = 3$ releases in each month. We set the first, second and third release sets as $v^*_{1,t} = \{1, ..., 20\}$, $v^*_{2,t} = \{21, ..., 40\}$, and $v^*_{3,t} = \{41, ..., 60\}$, i.e. $n^*_1 = n^*_2 = n^*_3 = 20$.

Simulation 1 In this simulation, we assume there is no correlation among blocks of three releases. The following parameters are used to generate the data: $\mathbf{A} = diag(0.7, -0.65, 0.6)$; $\mathbf{\Sigma} = diag(5, 3.5, 2)$; $\beta_0 = 0.5$, $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}_3 = (4, 1, 0.5)'$, and $\beta_4 = 0.65$; $\eta^2 = 2$; $\mathbf{\Omega} = diag(\mathbf{\Omega}_{11}, \mathbf{\Omega}_{22}, \mathbf{\Omega}_{33})$ where $\mathbf{\Omega}_{qq}$ $(q = 1, 2, 3)$ is $n^*_q \times n^*_q$ covariance matrix for the q^{th} release dataset that is generated from an Inverse-Wishart distribution $IW(n^*_q, \mathbf{I}_{n^*_q \times n^*_q}/60)$; and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{60})'$, where $\boldsymbol{\theta}_i = (\boldsymbol{\theta}_{i1}, \boldsymbol{\theta}_{i2}, \boldsymbol{\theta}_{i3})$ is the i^{th} row of $\mathbf{\Theta}$, where $\boldsymbol{\theta}_{ir}$ is the coefficient for $f_{j,t}$. For $i = 1, ..., 20$, $\boldsymbol{\theta}_{i1}$ is simulated from $N(0, 1)$, while $\boldsymbol{\theta}_{i2}, \boldsymbol{\theta}_{i3}$ are simulated from $N(0, 0.0025)$. For $i = 21, ..., 40$, $\boldsymbol{\theta}_{i1}, \boldsymbol{\theta}_{i2}$ are simulated from $N(0, 1)$, while $\boldsymbol{\theta}_{i3}$ is simulated from $N(0, 0.0025)$. For $i = 41, ..., 60$, $\boldsymbol{\theta}_{i1}$, $\boldsymbol{\theta}_{i2}, \boldsymbol{\theta}_{i3}$ are simulated from $N(0, 1)$. This set-up assumes the first release mainly contains information from the first latent factor, the second release mainly contains information from the first two latent factors, and the last release contains information from all three latent factors.

Simulation 2 In this simulation, we allow no-zero correlation among blocks of three releases, that is, $\mathbf{\Omega}$ is simulated from $IW(60, \mathbf{I}_{60\times 60}/60)$. Other parameters have the same set-up with Simulation 1.

Figure 2.1 plots the simulated GDP y_k and three common factors $\mathbf{F}_t$ from Simulation 1 and Simulation 2. The first 38 quarters and the corresponding 114 monthly data series are used as in-sample data, and the out-of-sample nowcasting performance is assessed based on the rest of 32 quarters and 96 monthly data series.

2.3.1 In-sample Estimation of Latent Factors $\mathbf{F}_t$

We first focus on simulation evidence that the BAY method can accurately estimate the latent factors $\mathbf{F}_t$. Obtaining good estimates of latent factors is an important task if one is interested in using the same set of common factors to explain the movements of different economic or financial

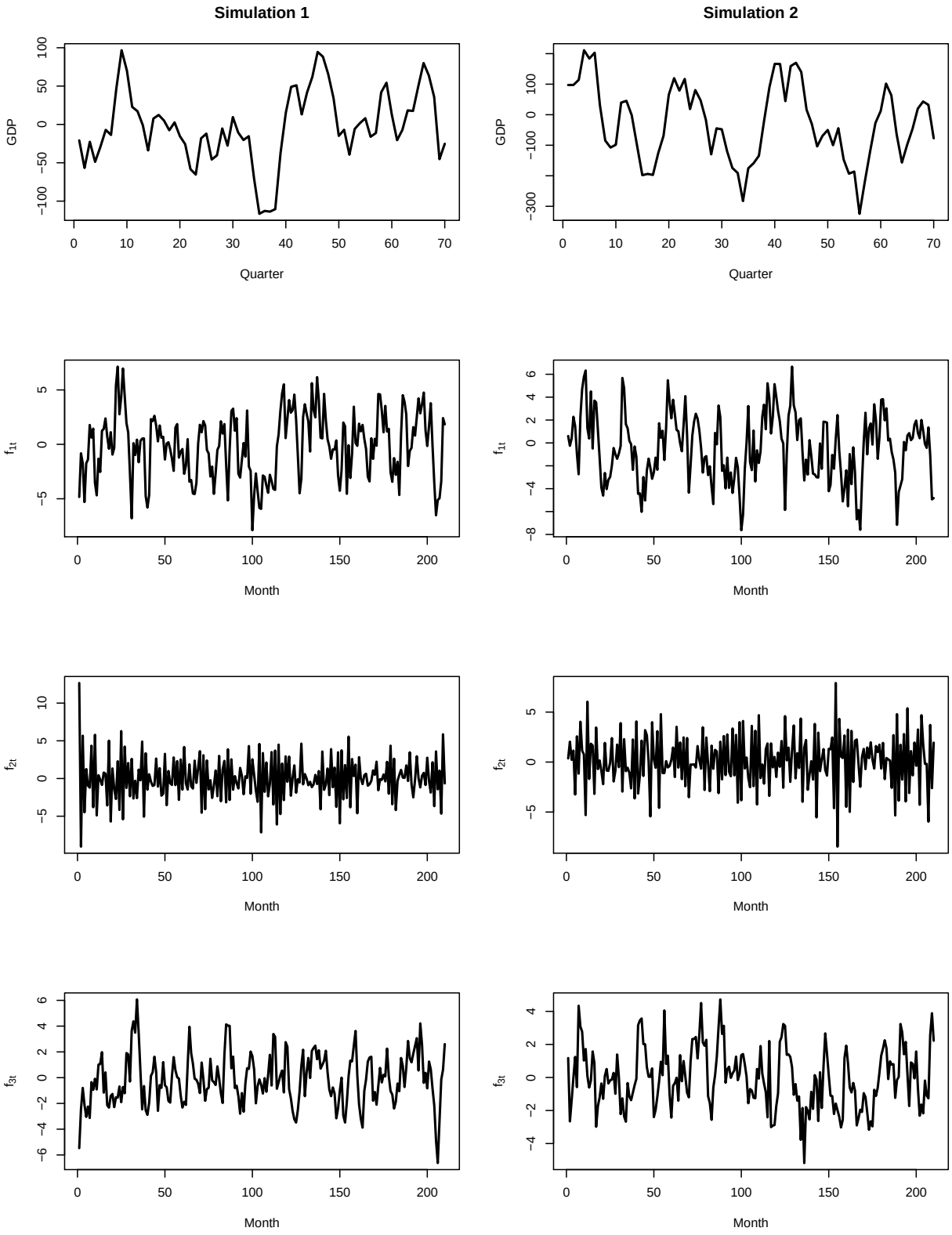

Figure 2.1: Simulated GDP and latent factors. The top row is the GDP, and the bottom three rows are simulated first, second, and third latent factors respectively. The left panel is for Simulation 1 and the right panel is for Simulation 2.

time series. For the BAY MCMC, we use the results from PCA and Kaman Filter in the GRS approach as the initial values of parameters and latent factors. The hyper-parameter values are set to be: $\nu_\theta = n + 2$ for the prior of $\mathbf{\Omega}$ in (2.12), $(\alpha_s, \beta_s) = (2, r + 2)$ for the prior of $\mathbf{\Sigma}$ in (2.14), and $(\alpha_h, \beta_h) = (2, 0.0001)$ for the prior of η^2 in (2.16).

We run our MCMC procedure for 7,000 iterations, discarding the first 5,000 and using the last 2,000 iterations for posterior summaries. Investigation (not reported here) is conducted to check how fast posterior draws of parameters and latent factors converge. It is found that the burn-in period with 5,000 iterations is enough. See more details in the supplemental file B.

Figure 2.2 plots the estimated latent factors from both BAY and GRS approaches, together with the true latent factors, for two simulations in the in-sample period $t = 1, ..., 114$. Absolute values are used in the pictures because the restrictions based on Stock and Watson (2002) can identify the factors up to a change of sign. Figure 2.2 shows that the estimated latent factors from both BAY and GRS approaches are close to the true factors, which confirms that by introducing Assumption F and Assumption M, the latent factors are identifiable for both approaches.

In addition to graphical illustrations, we also compare the in-sample fit errors of GDP and in-sample estimation error of latent factors between the two approaches. The relative in-sample fit error for GDP is calculated as $RISFE = \sum_{k=1}^{38} |\hat{y}_k^{fit} - y_k| / \sum_{k=1}^{38} |y_k|$, where $\hat{y}_k^{fit} = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}_1' \hat{\mathbf{F}}_{3k} + \hat{\boldsymbol{\beta}}_2' \hat{\mathbf{F}}_{3k-1} + \hat{\boldsymbol{\beta}}_3' \hat{\mathbf{F}}_{3k-2} + \hat{\beta}_4 y_{k-1}$, and $\hat{\boldsymbol{\beta}}$'s and $\hat{\mathbf{F}}_t$'s are estimated using the in-sample data either from GRS or BAY method. For BAY, posterior means are used as estimated parameters and latent factors. The relative in-sample estimation error for $\mathbf{F}_t$ is calculated as $RISEE(j) = \sum_{t=1}^{114} ||\hat{f}_{jt}| - |f_{jt}|| / \sum_{t=1}^{114} |f_{jt}|$, where $\hat{f}_{jt}$ $(j = 1, 2, 3)$ is the estimated j^{th} factor at time t from either GRS and BAY method. Table 2.3 reports both $RISFE$ and $RISEE(j)$ $(j = 1, 2, 3)$ for both methods and both simulation studies. Table 2.3 again shows that the averages of in-sample estimation errors of latent factors for both BAY and GRS method are small in all three factors. For Simulation 1, the BAY approach produces better RISEE for the second and third latent factors and close RISEE for the first latent factor. For Simulation 2, the BAY approach only outperforms GRS in terms of RISEE for the third latent factor and close RISEE for the first two latent factors.

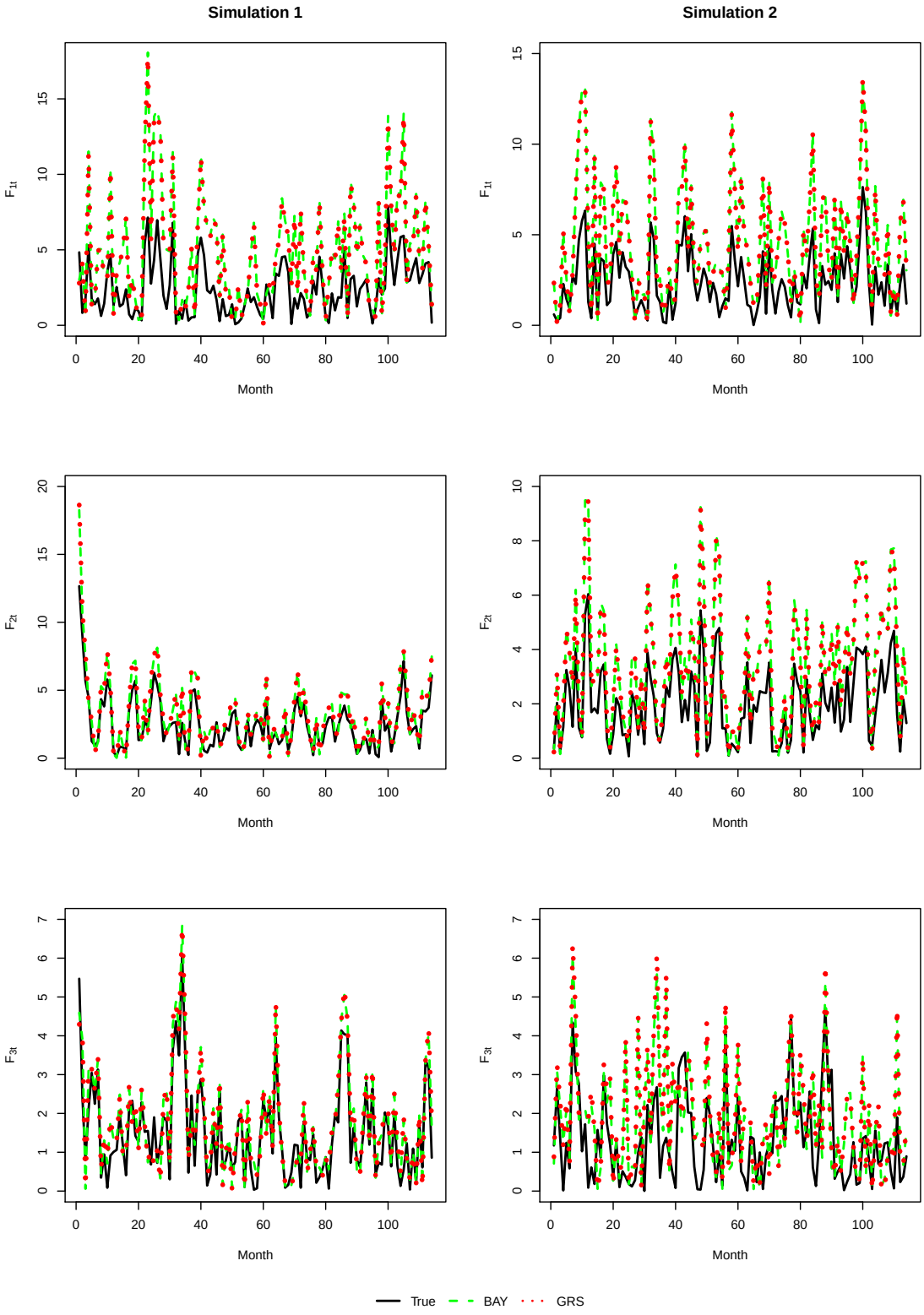

Figure 2.2: Absolute values of estimated latent factors from BAY (green dashed line) and GRS (red dotted line) approaches versus the truth (black solid line), with the left panel being Simulation 1 and the right panel being Simulation 2.

Table 2.3: Relative in-sample fit errors of GDP $RISFE$, and relative in-sample estimation error $RISEE(j)$ $(j = 1, 2, 3)$ in 3 factors, for both simulations.

Simulation	$RISFE$		$RISEE(1)$		$RISEE(2)$		$RISEE(3)$	
	GRS	BAY	GRS	BAY	GRS	BAY	GRS	BAY
1	0.448	0.446	1.098	1.124	0.382	0.360	0.309	0.277
2	0.336	0.325	0.923	0.934	0.684	0.707	0.804	0.767

The in-sample fit results in this subsection show that our MCMC method provides accurate identification of latent factors, which can help shed light on comovements of latent factors and some of the economic/financial series.

2.3.2 Out-of-sample Nowcasting Performance

Out-of-sample nowcasting performance of GRS and BAY methods is assessed based on 32 one-step-ahead nowcasting. For both methods, in each additionally added quarter, the model parameters and latent factors are updated for each release date within each of the 3 nowcasting months in the current quarter. Then the nowcast results are produced according to (2.17) - (2.19), and there are totally $32 \times 3 \times 3 = 288$ nowcast results.

Figure 2.3 and 2.4 show the nowcasting performance for Simulation 1 and 2. The left panel plots BAY nowcasts and the right panel plots GRS nowcasts. The first, second, and third row are nowcasts in the first, second, and third month of the given quarter respectively. In each cell, the curves colored in red, green and blue with different knot types represent the nowcast results based on the first, second, and third release dates in a given month of a given quarter respectively. Comparing GRS to BAY approaches, both methods give excellent nowcasts from the very first release in the first month all the way to the last release in the third month of a quarter. There are no distinguishable changes from releases to releases in the same month for both GRS and BAY approaches. But some improvement can be spotted when moving from nowcasts in the first month to nowcasts in the third month.

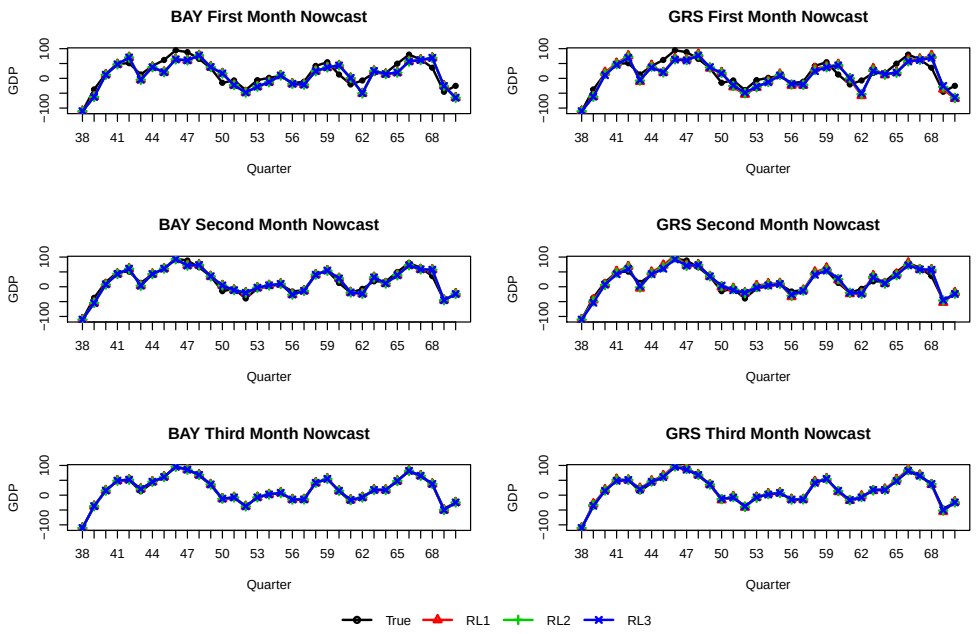

Figure 2.3: Nowcasting performance for Simulation 1, the left panel plots BAY nowcasts and the right panel plots GRS nowcasts. The first, second, and third row are nowcasts in the first, second, and third month of the given quarter respectively. In each cell, the curves colored in red, green and blue with different knot types represent the nowcast results based on the first, second, and third release dates in a given month of a given quarter respectively.

We use the mean absolute nowcasting error (MANE) as a measure of nowcasting accuracy. For either GRS or BAY, let $\hat{y}_{K+1}^{(q,T)}$ be the nowcast calculated according to (2.17) - (2.19) at q^{th} release date of month T, where $q = 1, 2, 3$ and $T = 3K + 1$ (first month nowcast), or $T = 3K + 2$ (second month nowcast), or $T = 3K + 3$ (third month nowcast) of the current quarter $K + 1$. Then $MANE(q, T) = 32^{-1} \sum_{K+1=39}^{70} |\hat{y}_{K+1}^{(q,T)} - y_{K+1}|$. In order to compare the nowcasting performance of the two methods with the random walk (RW) approach, which takes GDP from the previous quarter as the nowcast for the current quarter GDP, relative MANE's (relative to the MANE of RW) are used. Figure 2.5 and 2.6 show the ratios (in percentages) of MANE's of GRS or BAY to the MANE of RW in nine combinations of 3 releases and 3 nowcasting months for Simulation 1 and 2. The left panel is for BAY, the right panel is for GRS, and the horizontal line is 100% representing the baseline for RW, with first, second, and third release colored from dark to light. Bars shorter than the reference line indicates that nowcasts are better than RW, otherwise worse

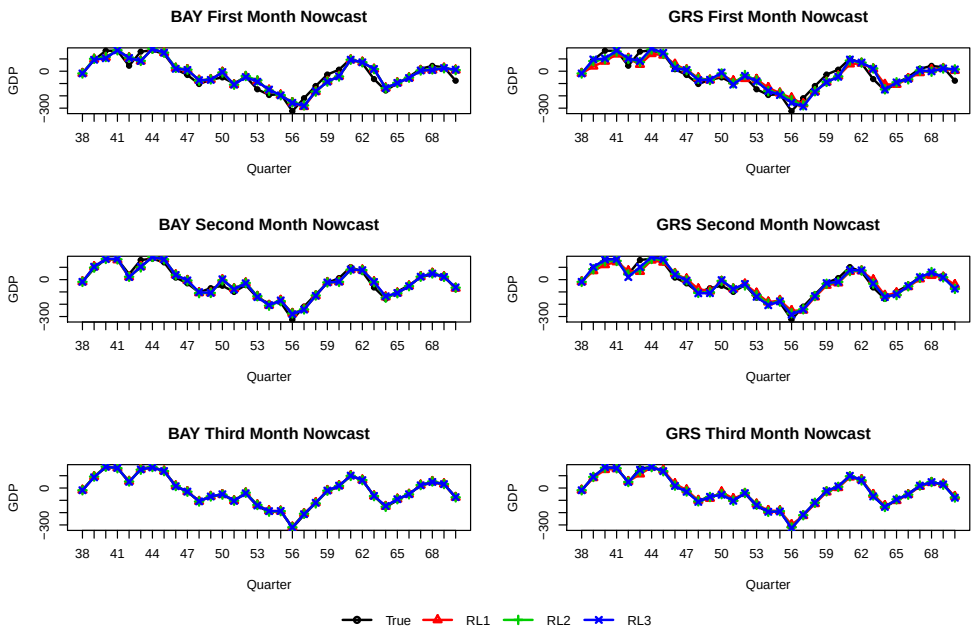

Figure 2.4: Nowcasting performance for Simulation 2, the left panel plots BAY nowcasts and the right panel plots GRS nowcasts. The first, second, and third row are nowcasts in the first, second, and third month of the given quarter respectively. In each cell, the curves colored in red, green and blue with different knot types represent the nowcast results based on the first, second, and third release dates in a given month of a given quarter respectively.

than RW. Figure 2.5 and 2.6 tell the same story as Figure 2.3 and 2.4, moving from the first month to the third month, there are significant reductions in terms of MANE's for both GRS and BAY approaches. Comparing the nowcasting errors between releases in the same month, GRS nowcasting slightly improves, but there is no significant change in MANE for BAY. Both methods beat RW in terms of nowcasting errors.

Table 2.4 reports the percentages of reduction in MANE's of both methods relative to RW, i.e. $(MANE^m - MANE^{RW})/MANE^{RW} \times 100$ (in %), where $m \in \{BAY, GRS\}$, for Simulation 1 and 2. The more negative the percentage is, the more reduction in nowcasting errors it represents. When comparing the averages of percentages over 3 releases for each month, we see BAY in general has more reduction in terms of nowcasting errors than GRS. Even though there is no noticeable change across releases in BAY, the percentages of reduction in MANE of BAY in the first release

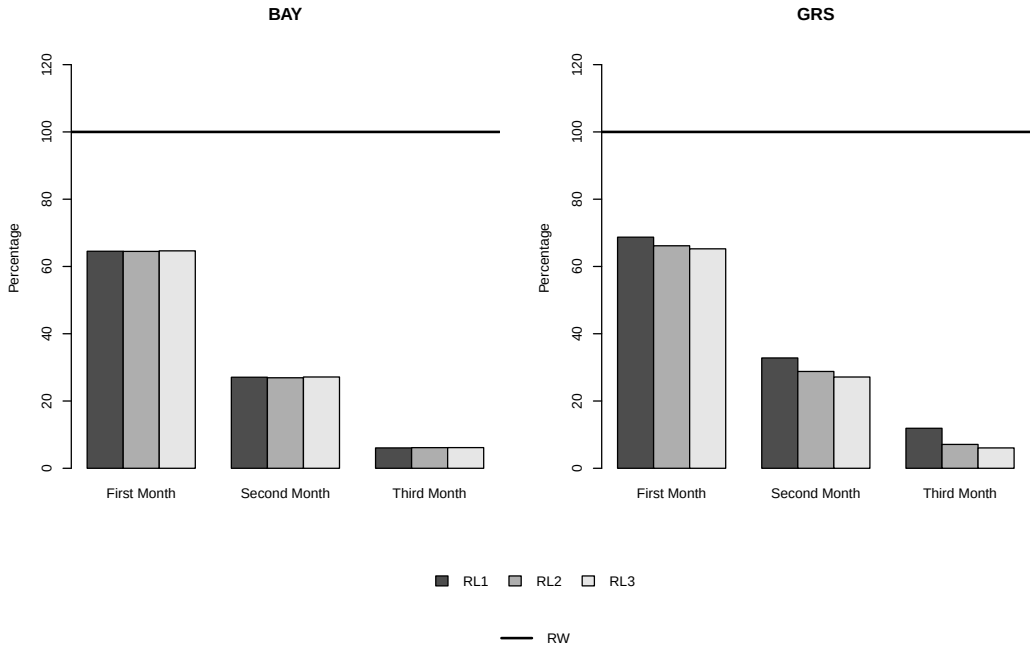

Figure 2.5: Mean absolute nowcasting error ratios for Simulation 1. The left panel is for BAY, the right panel is for GRS, and the horizontal line is 100% representing the baseline for RW. The first, second, and third release are colored from dark to light.

is higher than those of GRS in the third release for the same month in most of the places, except for the second and third months in Simulation 2.

We conduct some investigation to understand why there is not much change across different releases in BAY method. First, in the results reported in the supplemental file B, we find that there are no significant changes in estimated parameters and factors across different releases, based on some tests and graphical displays. It is actually not surprising given that little extra information becomes available at a new release date. For example, when nowcasting y_{39} based on the very first release in the first month, the number of data points in hand is $60 \times 114 + 20 = 6860$, and when the second set of release is available, the number of data points becomes $6860 + 20 = 6880$, which has only roughly 0.29% new information added. However, the estimated factors in GRS have relatively much bigger changes across different releases. Because the BAY method results in estimated factors that are already very accurate (even at the first release) as shown in Section

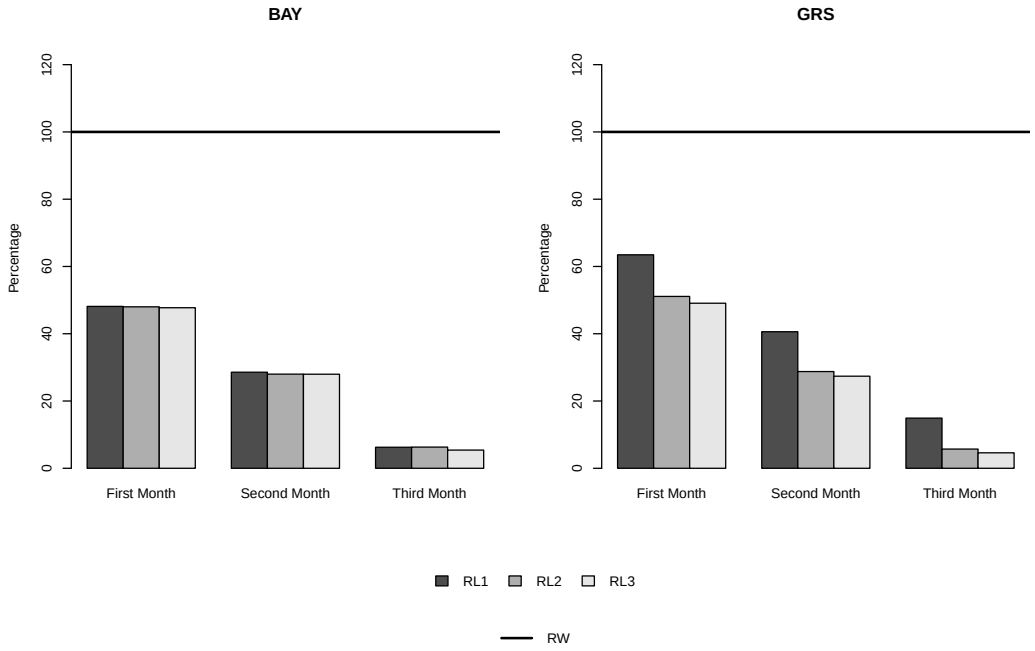

Figure 2.6: Mean absolute nowcasting error ratios for Simulation 2 (b). The left panel is for BAY, the right panel is for GRS, and the horizontal line is 100% representing the baseline for RW. The first, second, and third release are colored from dark to light.

2.3.1, it is much more difficult for BAY-estimated factors to improve further, compared to the GRS-estimated factors. Second, theoretically, adding more series on the left hand side of dynamics of $\mathbf{x}_t$ in (2.1) won't change the asymptotic consistency property of the estimator for $\mathbf{F}_t$, but will help gain efficiency. In order to see such efficiency gain effect, we calculate the posterior standard deviations of nowcasts for different releases. Define the standard deviation of nowcast for quarterly GDP at release q in month T for quarter $K+1$, where $q = 1, 2, 3$, $T = 1, 2, 3$, and $K + 1 = 39, ..., 70$, as $SD_{K+1}^{(q,T)} = \sqrt{(G-1)^{-1} \sum_{g=1}^{G} \left[(\hat{y}_{K+1}^{(q,T)})^{(g)} - \bar{y}_{K+1}^{(q,T)}\right]^2}$ where $\bar{y}_{K+1}^{(q,T)} = \frac{1}{G} \sum_{g=1}^{G} (\hat{y}_{K+1}^{(q,T)})^{(g)}$. Furthermore, define the aggregated level standard deviation for release q in month T of quarter $K+1$, as $SD^{(q,T)} = 32^{-1} \sum_{K+1=39}^{70} SD_{K+1}^{(q,T)}$. Figure 2.7 plots the side-by-side aggregated standard deviation for BAY nowcasts for Simulation 1 and 2. The left panel is for Simulation 1, the right panel is for Simulation2, with first, second, and third release colored from dark to light. A downward stepwise

Table 2.4: This table reports the percentages of reduction in MANE's of both methods relative to RW, i.e. $(MANE^m - MANE^{RW})/MANE^{RW} \times 100$ (in %), where $m \in \{BAY, GRS\}$, for both simulation studies.

(a) Percentage of reductions in MANE's relative to RW in Simulation 1

Release	BAY 1^{st} Month	BAY 2^{nd} Month	BAY 3^{rd} Month	GRS 1^{st} Month	GRS 2^{nd} Month	GRS 3^{rd} Month
1st	-35.473%	-72.928%	-93.946%	-31.294%	-67.197%	-88.106%
2nd	-35.519%	-73.089%	-93.860%	-33.847%	-71.211%	-92.905%
3rd	-35.355%	-72.855%	-93.850%	-34.753%	-72.865%	-93.946%
Average	-35.782%	-72.857%	-93.885%	-33.298%	-70.424%	-91.652%

(b) Percentage of reductions in MANE's relative to RW in Simulation 2

Release	BAY 1^{st} Month	BAY 2^{nd} Month	BAY 3^{rd} Month	GRS 1^{st} Month	GRS 2^{nd} Month	GRS 3^{rd} Month
1st	-51.869%	-71.435%	-93.742%	-36.531%	-59.402%	-85.088%
2nd	-51.997%	-72.006%	-93.699%	-48.905%	-71.240%	-94.285%
3rd	-52.264%	-72.034%	-94.590%	-50.923%	-72.620%	-95.400%
Average	-52.043%	-71.825%	-94.010%	-45.453%	-67.754%	-91.591%

trend can be detected. Therefore, for BAY approach, although additional releases does not improve nowcastings in terms of MANE, more monthly data series do help the precision of BAY nowcastings.

In the simulation studies considered in this article, we evaluate our BAY method based on its estimation accuracy in latent factors and nowcasting performance for the quarterly GDP. In terms of estimation accuracy in factors, both of the BAY and GRS approaches can produce accurate estimated factors. In terms of nowcasting performance, the two methods are comparable with BAY being slightly better in sense of resulting in smaller nowcasting errors. Our simulation results suggest that the BAY method has the potential to estimate the dynamic factor models well and produce reliable nowcasting results.

2.4 Bayesian Approach in Nowcasting: Empirical Evidence

Despite its advantages demonstrated in the numerical simulations, it is not immediately clear that the BAY method can outperform the RW or the GRS approach in empirical application. In

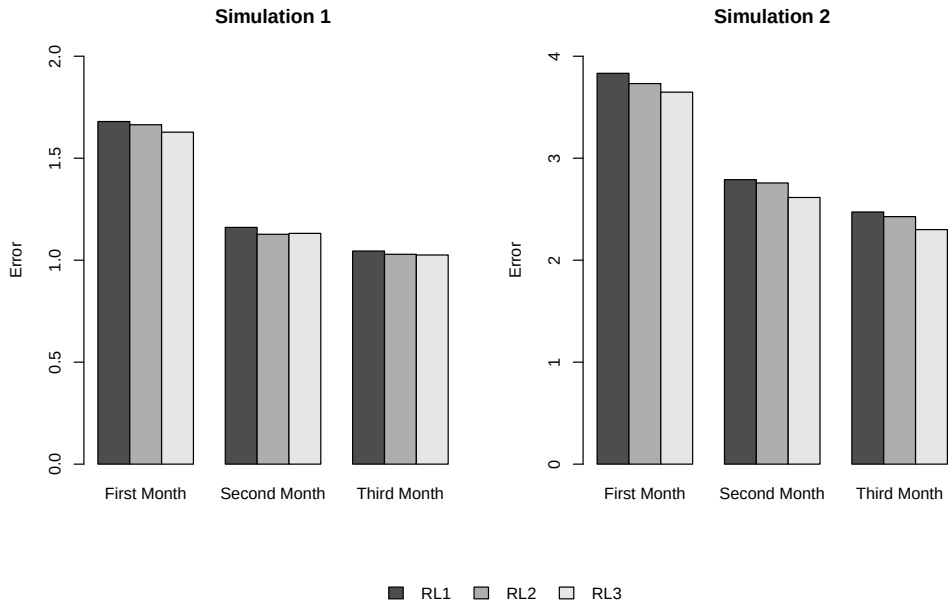

Figure 2.7: Aggregated standard deviations for Simulation 1 and 2. The left panel is for Simulation 1, the right panel is for Simulation2, with first, second, and third release colored from dark to light.

this section, we examine empirical performance of the BAY method, compared with the RW and the GRS, in nowcasting the quarterly China's GDP.

We follow Yiu and Chow (2011), which applies GRS method to nowcast annual growth rate of China's GDP, to select relevant monthly data series. The dataset in Yiu and Chow (2011) contains 189 indicator series of several categories, such as prices, industrial production, fixed asset investment, external sector, money market and financial market. The span of the panel is from January 1998 to June 2009. Most series are transformed to either a 12-month growth rate or a 12-month difference to induce stationarity, detailed transformation information can be found in the original paper. They divide the monthly series into 16 data blocks, which are released on eight dates throughout months. We extend the time span of the data a little longer until June of 2010. Therefore, there are totally 150 months and 50 quarters in the dataset considered in our empirical study. The GDP used in this chapter is China's quarterly nominal GDP growth rate (relative to nominal GDP in the same period of the previous year). For example, the GDP in the third quarter

of 2008 is defined as,

$$\left(\frac{\text{Nominal GDP of 2008Q3}}{\text{Nominal GDP of 2007Q3}}\right) \times 100\%.$$

Figure 2.8 plots the GDP data covering the period from $1998Q1$ to $2010Q2$.

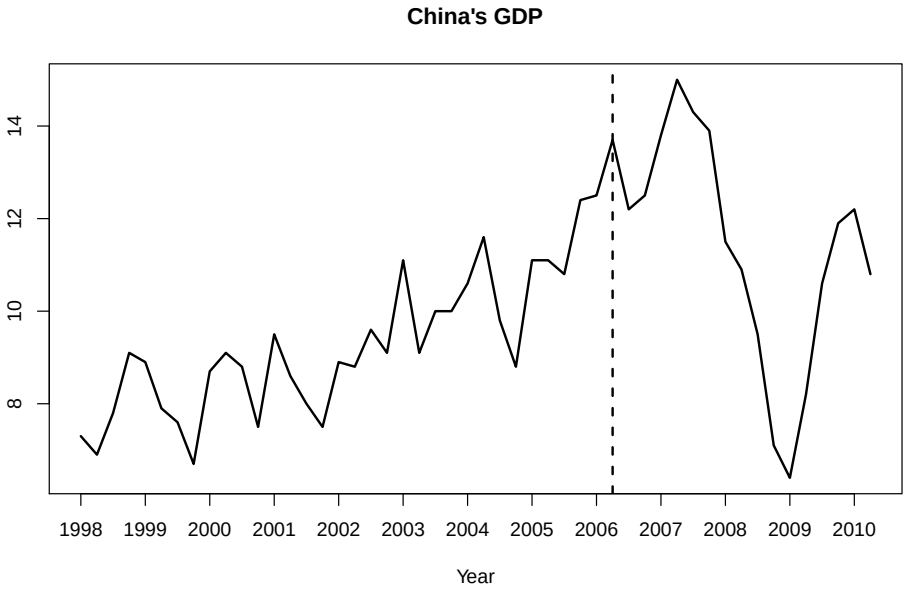

Figure 2.8: China's quarterly nominal GDP growth rate for 1999Q1 to 2010Q2. Quarters to the left of the dashed line are reserved as in-sample data, quarters to the right of the dashed line are out-sample nowcasting targets.

Due to missingness, we managed to obtain 159 out of the 189 monthly data series that have no missing information for all months in the time period considered. The monthly data are transformed using the same transformation in Yiu and Chow (2011). Moreover, it is not necessarily true that all the chosen monthly series are related to the GDP. Hence, some preliminary regression analyses are done to further remove irrelevant monthly series. Specifically, for each $x_{i,t}$ of the 159 collected series, we run a OLS regression, regressing y_k on $x_{i,3k}$ and its two lags $x_{i,3k-1}$ and $x_{i,3k-2}$ for $k = 1, ..., 50$. The monthly data series $x_{i,t}$ is kept if at least one of the three coefficients in the OLS regression is statistically significant. In the end, 117 monthly data series are kept. Table 2.5 provides the summary information of data series used. Specific names of the series can be found in the supplemental file B.

Table 2.5: Data Used For Nowcasting. Each column, from left to right, represents No. of release, name of data block, release date, No. of series released in each block, and total number of series for the corresponding release. RL stands for release.

Release	Block	Release Date	No. of Series	Total No. of Series
RL1	Interest Rate Stock Market Foreign Financial	1^{st}	9 6 13	28
RL2	Price Index External 1 Building Goverment	12^{th}	6 2 4 1	13
RL3	Investment	17^{th}	4	4
RL4	Exchange Rate Producer Price Index	19^{th}	9 9	18
RL5	World Commodity Price	21^{st}	10	10
RL6	External 2 Retail Sales	24^{th}	21 1	22
RL7	Economic Climate Indicator Industrial Production	25^{th}	2 14	16
RL8	Monetary	28^{th}	6	6
Total	16 Blocks	8 Dates	117 Series	

We reserve the first 34 quarters (1998Q1 to 2006Q2) as the in-sample data, and use the rest of 16 quarters (2006Q3 to 2010Q2) to assess one-step-ahead nowcasting performance. The vertical dashed line in Figure 2.8 represents this cut-off time. For each additional release in each additional month added, parameters and latent factors are updated and nowcast for the current quarter is re-calculated. Therefore, there are totally $16 \times 3 \times 8 = 384$ nowcast results reported. Yiu and Chow (2011) adopts the method in Bai and Ng (2002) to determine the number of common factors in the factor model, and the number of factors they chose is 2. A PCA is performed based on the 117 monthly data series after elimination. We see that the first three PCs contribute 83.44% of the total variation, while the fourth PC only adds additional 8% of the variation explained. Therefore, we chose the number of factors to be 3 instead of 2. We use the same hyper-parameter values for the priors, and the same way of choosing initials for parameters and factors, as discussed in Section

2.3.1. Again, we discard the first 5,000 runs as "burn-in" period and use posterior draws in the last 2,000 iterations to estimate model parameters and common factors.

Nowcasting results (focusing on the out-of-sample period) are plotted in Figure 2.9, with the left panel being BAY approach, the right panel being GRS approach, and the 3 rows being 3 nowcasting months. In each subplot, the solid curve is the real GDP, while 8 other curves in different colors and line types represent nowcasting results from 8 different release dates. In general, both GRS and BAY capture the trend of real GDP reasonably well. However, there is less variation in BAY nowcasts from release to release than in GRS nowcasts for all 3 nowcasting months. GRS nowcasts deviate further away from the real GDP for the first couple of releases, especially in the third nowcasting month.

The measure of nowcasting errors, MANE, can be computed in the same fashion as in simulation study. Figure 2.10 plots the relative ratios of MANEs for both GRS (1st row) and BAY (2nd row) approaches using RW as the baseline. Three columns are for 3 nowcasting months, and 8 bars in each subplot represent 8 different dates. The horizontal line located at 100% is the reference line for RW. Table 2.6 summarizes the percentages of reduction in MANEs of both methods relative to RW, defined in the same way as in Section 2.3.2. Each panel in Table 2.6 reports a nowcasting month. According to Figure 2.10, for the first 5 releases, BAY outperforms GRS in terms of having smaller nowcasting errors for all 3 nowcasting months. This is particularly true for RL1, RL4, and RL5 (RL stands for release). It is interesting to see that MANE's of GRS are bigger than RW, especially in the third month. This is the same story found in Yiu and Chow (2011). One conclusion drawn by Yiu and Chow is that the Producer Price Index only ruins the nowcasting, and this particular data block is contained in the 4th release. However, we find the 4th release actually results in smaller MANEs in BAY nowcasts than in RW nowcasts. For the last 3 releases, GRS performs better than BAY in the 1st and 3rd nowcasting months, but not in the 2nd nowcasting month. Consistent with these findings shown in Figure 2.10, Table 2.6 provides numerical comparisons of two methods, relative to RW. The last column in each panel reports the averages of percentages of reduction in MANEs relative to RW across all 8 releases. On average, the percentages of reduction in nowcasting

errors in BAY approach increases when moving from the 1st month (2.93% reduction), to the 2nd month (5.79% reduction), and to the 3rd month (14.76% reduction). However, we don't see this effect in the averages of reduction percentages for GRS approach. Across 3 nowcasting months, BAY results in the least amount of nowcasting errors in the 3rd month, ranging from 10% reduction to 18% reduction.

Table 2.6: Percentage of reduction in MANE's of both methods relative to RW, i.e. $(MANE^m - MANE^{RW})/MANE^{RW} \times 100$ (in %), where $m \in \{BAY, GRS\}$, for empirical studies. M1/M2/M3 stands for the first/second/third nowcasting month.

Method	M1 RL1	M1 RL2	M1 RL3	M1 RL4	M1 RL5	M1 RL6	M1 RL7	M1 RL8	Average
BAY	-0.94%	-3.47%	0.07%	-5.73%	-6.98%	-2.87%	-3.95%	0.40%	-2.93%
GRS	47.37%	-2.50%	0.78%	16.54%	-3.71%	-16.72%	-13.73%	-12.92%	1.89%
Method	M2 RL1	M2 RL2	M2 RL3	M2 RL4	M2 RL5	M2 RL6	M2 RL7	M2 RL8	Average
BAY	-0.21%	-7.58%	-9.40%	-4.91%	-5.61%	-10.09%	-2.14%	-6.41%	-5.79%
GRS	7.87%	1.17%	2.08%	5.36%	-3.42%	-3.12%	-2.74%	-2.97%	0.53%
Method	M3 RL1	M3 RL2	M3 RL3	M3 RL4	M3 RL5	M3 RL6	M3 RL7	M3 RL8	Average
BAY	-15.39%	-10.80%	-14.48%	-14.43%	-17.77%	-15.27%	-17.73%	-12.26%	-14.76%
GRS	44.70%	2.23%	0.18%	89.49%	26.85%	-24.70%	-16.47%	-14.24%	13.50%

Both methods can also estimate the latent common factors. Figure 2.11 plots the in-sample estimates for $\mathbf{F}_t$ by GRS and BAY approaches. Both methods estimate similar paths for the 1st and 2nd factors, but slightly differ on the 3rd factor.

Collectively, the empirical analysis in this section demonstrates the empirical relevance of the BAY approach in nowcasting China's GDP for the time period considered. These results, however, don't suggest that we should completely abandon GRS approach. In fact, (as we mention) for the last 3 releases in this dataset, GRS performs better than BAY in the 1st and 3rd nowcasting months.

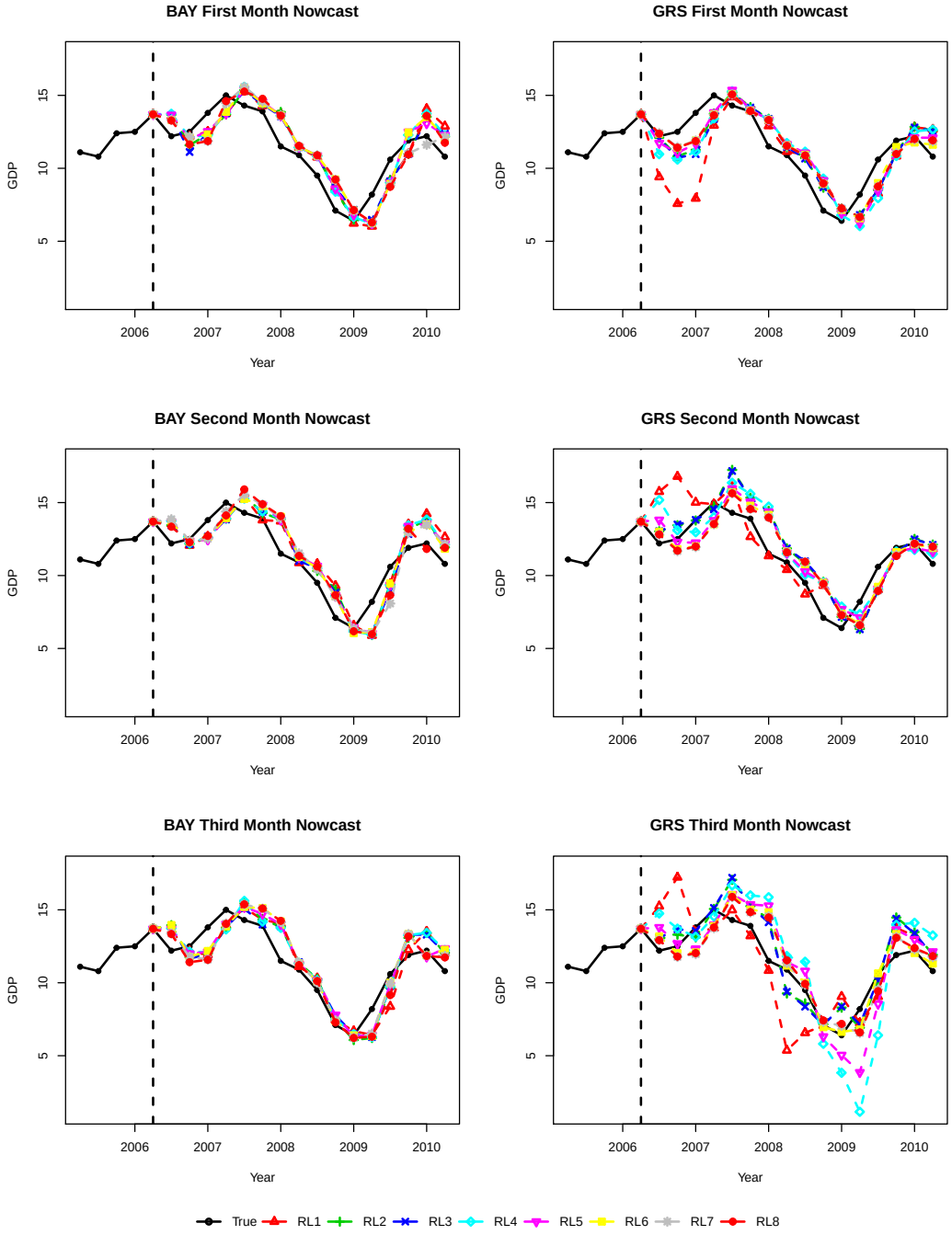

Figure 2.9: Nowcasting performance in the empirical study, the left panel being BAY approach, the right panel being GRS approach, and the 3 rows being 3 nowcasting months. In each subplot, the solid curve is the real GDP, while 8 other curves in different colors and line types represent nowcasting results from 8 different release dates.

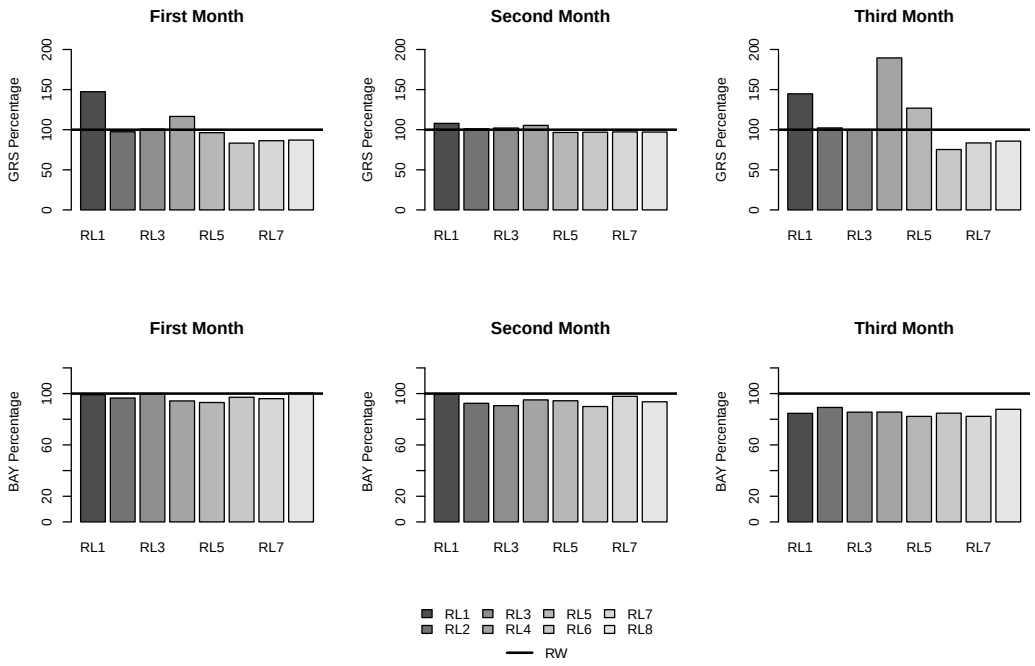

Figure 2.10: Relative ratios of MANE's for both GRS (1st row) and BAY (2nd row) approaches using RW as the baseline. Three columns are for 3 nowcasting months, and 8 bars in each subplot represent 8 different dates. RL stands for release.

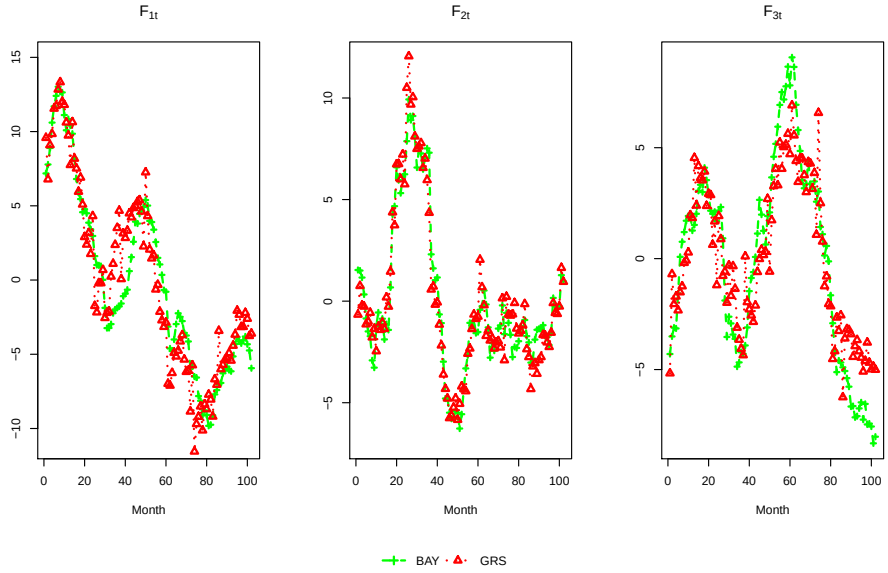

Figure 2.11: Estimates of 3 latent factors from both GRS and BAY methods in the empirical study.

CHAPTER 3. NOWCASTING GDP USING DYNAMIC FACTOR MODEL WITH UNKNOWN NUMBER OF FACTORS

In this Chapter, we develop a Bayesian approach to identify the unknown number of factors and estimate the latent dynamic factors of DFMs accurately in a real-time nowcasting framework. The proposed method can deal with the unbalanced data, which is typical of a real-time nowcasting analysis. We demonstrate the validity of our approach through simulation and explore the applicability of our approach through an empirical study in nowcasting US's quarterly GDP growth using monthly data series of several categories in the US market.

3.1 Structure of Dataset and Model Set-ups

In Section 3.1.1, we introduce the notations and model set-ups used in our BAY approach when volatility is constant. In Section 3.1.2 we formalize the unbalanced data structure, which is typical of a real time nowcasting analysis.

3.1.1 DFM with Constant Volatility

Since there are numerous series in the information set, modeling y (GDP) directly on all $\mathbf{x}$ (monthly series) would involve too many parameters, and hence the model performs poorly in nowcasting/forecasting because of the large uncertainty in parameters' estimation ("the curse of dimensionality"). The fundamental idea of DFMs is to explore the collinearity of the series by summarizing all available information into a few common factors. Because of collinearity, a linear combination of the common factors can approximate the dynamic interactions among the series and to can lead to a parsimonious model that works well in nowcasting/forecasting. The DFM considered in this chapter is specified in the following ways.

First, we assume the monthly data series is a linear function of a few unobserved common factors $\mathbf{F}_t$,

$$\mathbf{x}_t = \boldsymbol{\mu} + \boldsymbol{\Theta}\mathbf{F}_t + \boldsymbol{\epsilon}_t, \tag{3.1}$$

where $\mathbf{x}_t = (x_{1,t}, ..., x_{n,t})'$ be a $n \times 1$ monthly data series at month t, for $t = 1, ..., T$, $\mathbf{F}_t = (f_{1t}, \cdots, f_{rt})'$ be a $r \times 1$ monthly common factors at month t, $\boldsymbol{\Theta}$ is the $n \times r$ factor loading matrix, $\boldsymbol{\mu}$ is the $r \times 1$ mean vector, and $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Omega}_{n\times n})$. The number of latent factors, r, is assumed to be unknown and satisfies $r << n$. Then we further specify the dynamics of the common factors as follows,

$$\mathbf{F}_t = \mathbf{A}\mathbf{F}_{t-1} + \mathbf{u}_t, \tag{3.2}$$

where $\mathbf{A}$ is a $r \times r$ matrix and all roots of $|\mathbf{I}_r - \mathbf{A}z|$ lie outside the unit circle ($\mathbf{I}_r$ denotes $r \times r$ identity matrix), and $\mathbf{u}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{r\times r})$. Finally, the quarterly GDP is assumed to be a linear function of the common factors in each month of the quarter and its own lag,

$$y_k = \beta_0 + \boldsymbol{\beta}_1'\mathbf{F}_{3k} + \boldsymbol{\beta}_2'\mathbf{F}_{3k-1} + \boldsymbol{\beta}_3'\mathbf{F}_{3k-2} + \beta_4 y_{k-1} + \nu_k, \tag{3.3}$$

where β_0, β_4 are scalars, $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3$ are $r \times 1$ vectors, $v_k \sim N(0, \eta^2)$, and $k = 1, ..., K$. We assume $3K + 1 \leq T \leq 3K + 3$. The DFM specified this way not only can bridge the information contained in monthly data series with the quarterly GDP, but also helps reduce the dimension of parameters, thus increasing the degrees of freedom.

It is known that DFMs suffer from a non-identifiable issue. Following Stock and Watson (2002), two sets of assumptions (Assumption F and Assumption M) are constructed (details are in Appendix A). Specifically in this chapter, some restrictions are imposed on matrices $\mathbf{A}$ and $\boldsymbol{\Sigma}$ in the dynamics of latent factors in equation (3.2) as follows, $\mathbf{A} = diag(a_1, a_2, ..., a_r)$ and $\boldsymbol{\Sigma} = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_r^2)$, where $|a_j| < 1$ $(j = 1, ..., r)$ and $\sigma_i^2/(1 - a_i^2) > \sigma_j^2/(1 - a_j^2)$, $\forall i < j$ (for $i, j = 1, ..., r$). Under these specified $\mathbf{A}$ and $\boldsymbol{\Sigma}$, $f_{j,t} = a_j f_{j,t-1} + \sigma_j u_{j,t}$ for $j = 1, ..., r$ and $u_{j,t} \sim N(0, 1)$. These restrictions together with the prior specification for loading matrix $\boldsymbol{\Theta}$ (discussed later) satisfy the identification assumption F1 in Appendix A. Stock and Watson (2002) concluded that this assumption identified the factors up to a change of sign.

Stock and Watson (2002) also showed the first r principal components obtained from the covariance matrix of $\mathbf{x}_t$ are consistent estimators of the latent factors in DFMs if r is known. The consistency of the principal components estimator of $\mathbf{F}_t$ can also be found in Connor and Korajczyk (1986), Bai (2003) and others. Borrowing the idea in Zhang et al. (2013) we introduce a binary indicator matrix $\mathbf{Z}$ associated with the factors into our model,

$$\begin{aligned} \mathbf{x}_t &= \boldsymbol{\mu} + \boldsymbol{\Theta}\mathbf{Z}\mathbf{F}_t + \boldsymbol{\epsilon}_t, \text{ for } t = 1, ..., T \\ \mathbf{F}_t &= \mathbf{A}\mathbf{F}_{t-1} + \mathbf{u}_t, \text{ for } t = 1, ..., T \\ y_k &= \beta_0 + \boldsymbol{\beta}_1'\mathbf{Z}\mathbf{F}_{3k} + \boldsymbol{\beta}_2'\mathbf{Z}\mathbf{F}_{3k-1} + \boldsymbol{\beta}_3'\mathbf{Z}\mathbf{F}_{3k-2} + \beta_4 y_{k-1} + \nu_k, \text{ for } k = 1, ..., K \end{aligned} \tag{3.4}$$

where $\mathbf{Z} = diag(z_1, z_2, ..., z_R)$ is a $R \times R$ diagonal matrix with binary entries $z_j \in \{0, 1\}$ $(j = 1, ..., R)$ for some $R > r$. R denotes the maximum number of candidate latent factors, and is predetermined satisfying $R << n$. Now, $\mathbf{F}_t = (f_{1t}, \cdots, f_{Rt})'$ is a $R \times 1$ vector, collecting R candidate factors to be selected at month t, $\boldsymbol{\Theta}$ is the $n \times R$ factor loading matrix, $\mathbf{A} = diag(a_1, a_2, ..., a_R)$, $\boldsymbol{\Sigma} = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_R^2)$, and $\boldsymbol{\beta}_i$'s $(i = 2, 3, 4)$ are $R \times 1$ vectors. Matrices $\mathbf{A}, \boldsymbol{\Sigma}$ and $\boldsymbol{\Theta}$ have the same restrictions as stated previously. Furthermore, we model $\mathbf{Z}$ as

$$z_j \sim Bernoulli(p_j), \text{ where } 0 \leq p_j \leq 1 \text{ for } j = 1, ..., R. \tag{3.5}$$

Indicator z_j determines whether the j^{th} latent factor is included in the model or not. For example, assuming $R = 4$, if $\mathbf{Z} = diag(1, 1, 0, 0)$, then $\mathbf{Z}\mathbf{F}_t = diag(1, 1, 0, 0)(f_{1t}, f_{2t}, f_{3t}, f_{4t})' = (f_{1t}, f_{2t}, 0, 0)'$, impling the third and fourth latent factors do not contribute in explaining the dynamic interaction of $\mathbf{x}_t$ and y_k. Thus the true number of latent factors is $r = 2$ in this case. The number of latent factors is only determined by the number of 1's in the diagonal of $\mathbf{Z}$, i.e. $r = \sum_{j=1}^{R} z_j$. In this way, estimating r is now equivalent to selecting the true number of latent factors, or in other words, estimating $\mathbf{Z}$. Section 3.2.1 talks about the estimation procedure of $\mathbf{Z}$.

3.1.2 Working With Unbalanced Data

In real time, macroeconomic series are released with diverse lags, some series have observations through the current month, whereas for others the most recent observations are from the previous

months. For instance, measurements from survey such as the Empire State Manufacturing Survey and Manufacturing Business Outlook Survey, tend to have no lag when released. On the other hand, money stock measures, taking M1, M2, and M3 for example, usually have long release delay. Consequently, the underlying data sets are unbalanced. Appropriately dealing with this unbalanced feature of the data is crucial for nowcasting.

Suppose we are in month T of quarter $K+1$, we want to nowcast GDP in the current quarter, y_{K+1}, using all monthly information up through month T (the end of the sample), where T can be the first ($T = 3K+1$), second ($T = 3K+2$), or third ($T = 3K+3$) month of quarter $K+1$. To simplify life, we assume the release structures are exactly the same across all months. Let q denotes the release date within month T, and consider there are Q different release dates, i.e $q = 1, ..., Q$. Let n_q denote the number of series that are newly released at the release date q. When we say at q, we are referring to the time point when the release has already happend, i.e. at q, the n_q newly released series are available already. Due to diverse lags among monthly series, these n_q series do not necessarily refer to the same month. Let T_q denote the latest month in which all monthly data series are availabe at release date q, i.e. data is balanced for $t = 1, ..., T_q$, and for $t = T_q + 1, ..., T$ data is unbalanced. Let $v_{q,t}$ denotes the set of indexes of monthly series that are available at q for month $t = T_q + 1, ..., T$. So, the monthly series that are available at q for month $t = T_q + 1, ..., T$ are denoted as $\mathbf{x}_{i \in v_{q,t}}$. Let $n_{q,t}$ denote the number of monthly series in $\mathbf{x}_{i \in v_{q,t}}$ for $t = T_q + 1, ..., T$.

Consider a toy example of $n = 6$ monthly series, $\mathbf{x}_t = (x_{1,t}, x_{2,t}, x_{3,t}, x_{4,t}, x_{5,t}, x_{6,t})'$. Without loss of generality, x_{it} can always be arranged such that index i follows the order of release dates. Pretending we are at the release date q of month T of quarter $K+1$, Table 3.1 gives an overall picture of the entire dataset available when doing nowcast at q. The monthly data that is available are higlighted in gray cells in the table. At q, monthly data is balanced up to month $T-3$ therefore $T_q = T-3$. For month $t = T_q + 1 = T-2$, $\{x_{1,t}, x_{2,t}, x_{3,t}, x_{4,t}, x_{5,t}\}$ are available, so $v_{q,T-2} = \{1, 2, 3, 4, 5\}$ and $\mathbf{x}_{i \in v_{q,T-2}} = (x_{1,T-2}, x_{2,T-2}, x_{3,T-2}, x_{4,T-2}, x_{5,T-2})'$. Similarly, for month $t = T_q + 2 = T-1$, we have $v_{q,T-1} = \{1, 2, 3\}$ with $\mathbf{x}_{i \in v_{q,T-1}} = (x_{1,T-1}, x_{2,T-1}, x_{3,T-1})'$; for month $t = T$, we have $v_{q,T} = \{1, 2\}$ with $\mathbf{x}_{i \in v_{q,T}} = (x_{1,T}, x_{2,T})'$.

Table 3.1: Overall picture of the entire dataset available at the one release date of the second month in the current quarter. Cells highlighted in gray color indicates the corresponding variables in the cells are available.

k	1			2			$\cdots$	K			K+1		
t	1	2	3	4	5	6	$\cdots$	$T-4$	$T-3$	$T-2$	$T-1$	T	$T+1$
$\mathbf{x}_{i,t}$	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	$x_{1,5}$	$x_{1,6}$	$\cdots$	$x_{1,T-4}$	$x_{1,T-3}$	$x_{1,T-2}$	$x_{1,T-1}$	$x_{1,T}$	NA
	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	$x_{2,5}$	$x_{2,6}$	$\cdots$	$x_{2,T-4}$	$x_{2,T-3}$	$x_{2,T-2}$	$x_{2,T-1}$	$x_{2,T}$	NA
	$x_{3,1}$	$x_{3,2}$	$x_{3,3}$	$x_{3,4}$	$x_{3,5}$	$x_{3,6}$	$\cdots$	$x_{3,T-4}$	$x_{3,T-3}$	$x_{3,T-2}$	$x_{3,T-1}$	NA	NA
	$x_{4,1}$	$x_{4,2}$	$x_{4,3}$	$x_{4,4}$	$x_{4,5}$	$x_{4,6}$	$\cdots$	$x_{4,T-4}$	$x_{4,T-3}$	$x_{4,T-2}$	NA	NA	NA
	$x_{5,1}$	$x_{5,2}$	$x_{5,3}$	$x_{5,4}$	$x_{5,5}$	$x_{5,6}$	$\cdots$	$x_{5,T-4}$	$x_{5,T-3}$	$x_{5,T-2}$	NA	NA	NA
	$x_{6,1}$	$x_{6,2}$	$x_{6,3}$	$x_{6,4}$	$x_{6,5}$	$x_{6,6}$	$\cdots$	$x_{6,T-4}$	$x_{6,T-3}$	NA	NA	NA	NA
y_k	y_1			y_2			$\cdots$	y_K			y_{K+1}		
$\mathbf{F}_t$	$\mathbf{F}_1$	$\mathbf{F}_2$	$\mathbf{F}_3$	$\mathbf{F}_4$	$\mathbf{F}_5$	$\mathbf{F}_6$	$\cdots$	$\mathbf{F}_{T-4}$	$\mathbf{F}_{T-3}$	$\mathbf{F}_{T-2}$	$\mathbf{F}_{T-1}$	$\mathbf{F}_T$	

The observations, available to use at the release date q of month T, include $\mathbf{Y} = \{y_1, y_2, ..., y_K\}$ and $\mathbf{X}_q = \{\mathbf{x}_1, ..., \mathbf{x}_{T_q}, \mathbf{x}_{i \in v_{q,T_q+1}}, ..., \mathbf{x}_{i \in v_{q,T}}\}$. The goal is to nowcast y_{K+1} using all information available at every release date q in the current quarter, i.e. real-time nowcasting. Note that the factors $\mathbf{F} = \{\mathbf{F}_1, ..., \mathbf{F}_T\}$ are not observed, and need to be estimated using the adjusted Kalman Filter (e.g. the GRS approach), or using the Bayesian smoothing techniques (e.g. our BAY approach). At every new release date q ($q = 1, ..., Q$), model parameters and latent variables are updated with this additional information from the new release, and nowcast of y_{K+1} is re-produced. Therefore there are $3 \times Q$ nowcast results in the current quarter. How to handle the unbalanced feature of the monthly data series will be discussed in Section 3.2.1.

At any release date q, we have observations $\mathbf{X}_q$ and $\mathbf{Y}$; latent variables $\mathbf{Z}$ and $\mathbf{F}$; and model parameters $\mathbf{\Phi}^{cv} = \{\boldsymbol{\mu}, \mathbf{\Theta}, \mathbf{A}, \mathbf{\Sigma}, \beta_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \beta_4, \eta^2, p_1, ..., p_R, \mathbf{\Omega}\}$. The goals are to estimate model parameters and latent variables using the observables, and to nowcast the current quarter GDP y_{K+1} at every release date q for $q = 1, ..., Q$ in month T for $T = 3K+1, 3K+2, 3K+3$.

3.2 Bayesian MCMC Estimation Method and Nowcasting

In Section 3.2.1, we outline the MCMC algorithm to estimate model parameters including the number of factors and latent dynamic factors. Formulas for nowcasting are provided in Section 3.2.2.

3.2.1 Estimating DFM With Constant Volatility via MCMC

We face a couple of challenges in estimating the above model. First, it is computationally infeasible to integrate out the high-dimensional latent factors $\mathbf{F}_t$ to obtain the likelihood based only on observables. Second, the observations in the panel of $\mathbf{X}_q = \{\mathbf{x}_1, ...\mathbf{x}_{T_q}, \mathbf{x}_{i\in v_{q,T_{q+1}}}, ..., \mathbf{x}_{i\in v_{q,T}}\}$ are not balanced. To overcome these difficulties, we develop a Bayesian MCMC approach for estimating the above DFM in real time. MCMC conducts inferences by simulating efficiently from (potentially complicated) posterior distributions of model parameters and latent variables given the observables. MCMC samples from the typically high-dimensional and complex posterior distributions by generating a Markov Chain over parameters and latent variables whose equilibrium distribution is the desired posterior distribution. The Monte Carlo method uses these samples for numerical integration for parameter and state estimation.

In order to facilitate the derivation of the joint posterior distribution, the dynamic of $\mathbf{x}_t$ in equation (3.4) is rewritten as

$$\mathbf{x}_t = \boldsymbol{\mu} + [\mathbf{I}_{n\times n} \otimes (\mathbf{Z}\mathbf{F}_t)'] * \boldsymbol{\theta} + \boldsymbol{\epsilon}_t,$$

where $\boldsymbol{\theta} = vec(\boldsymbol{\Theta}) = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n)'$ if $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1', ..., \boldsymbol{\theta}_n')'$ such that $\boldsymbol{\theta}_i$ $(i = 1, ..., n)$ is a $1 \times R$ vector representing the i^{th} row of $\boldsymbol{\Theta}$. Thus the conditional density of $\mathbf{x}_t$ in equation (3.4) is

$$\mathbf{x}_t|\boldsymbol{\mu}, \mathbf{Z}, \mathbf{F}_t, \boldsymbol{\Theta}, \boldsymbol{\Omega} \sim N(\boldsymbol{\mu} + [\mathbf{I}_{n\times n} \otimes (\mathbf{Z}\mathbf{F}_t)'] * \boldsymbol{\theta}, \boldsymbol{\Omega}), \text{ for } t = 1, ..., T_q. \tag{3.6}$$

The conditional density of $\mathbf{F}_t$ in equation (3.4) is

$$\mathbf{F}_t|\mathbf{F}_{t-1}, \mathbf{A}, \boldsymbol{\Sigma} \sim N(\mathbf{A}\mathbf{F}_{t-1}, \boldsymbol{\Sigma}), \text{ for } t = 2, ..., T, \tag{3.7}$$

and the conditional density of y_k in equation (3.4) is

$$\begin{aligned} & y_k|\boldsymbol{\beta}, \mathbf{Z}, \mathbf{F}_{3k}, \mathbf{F}_{3k-1}, \mathbf{F}_{3k-2}, y_{k-1}, \eta^2 \\ \sim \; & N(\beta_0 + \boldsymbol{\beta}_1' \mathbf{Z}\mathbf{F}_{3k} + \boldsymbol{\beta}_2' \mathbf{Z}\mathbf{F}_{3k-1} + \boldsymbol{\beta}_3' \mathbf{Z}\mathbf{F}_{3k-2} + \beta_4 y_{k-1}, \eta^2), \end{aligned} \tag{3.8}$$

for $k = 2, ..., K$. The joint posterior distribution, $p(\boldsymbol{\Phi}^{cv}, \mathbf{F}, \mathbf{Z}|\mathbf{Y}, \mathbf{X}_q)$, can be decomposed into products of individual conditionals,

$$\begin{aligned} & p(\boldsymbol{\Phi}^{cv}, \mathbf{F}, \mathbf{Z}|\mathbf{Y}, \mathbf{X}_q) \propto p(\mathbf{Y}, \mathbf{X}_q, \boldsymbol{\Phi}^{cv}, \mathbf{F}, \mathbf{Z}) \\ \propto \; & \prod_{t=1}^{T_q} p(\mathbf{x}_t|\boldsymbol{\mu}, \mathbf{Z}, \mathbf{F}_t, \boldsymbol{\theta}, \boldsymbol{\Omega}) \prod_{t=T_q+1}^{T} p(\mathbf{x}_{i\in v_{q,t}}|\boldsymbol{\mu}, \mathbf{Z}, \mathbf{F}_t, \boldsymbol{\theta}, \boldsymbol{\Omega}) \\ \times \; & \prod_{k=2}^{K} p(y_k|\boldsymbol{\beta}, \mathbf{Z}, \mathbf{F}_{3k}, \mathbf{F}_{3k-1}, \mathbf{F}_{3k-2}, y_{k-1}, \eta^2) \times \prod_{t=2}^{T} p(\mathbf{F}_t|\mathbf{F}_{t-1}, \mathbf{A}, \boldsymbol{\Sigma}) \times \prod_{j=1}^{R} p(z_j|p_j) \\ \times \; & p(\boldsymbol{\Phi}^{cv}), \end{aligned} \tag{3.9}$$

where $p(\mathbf{x}_t|\boldsymbol{\mu}, \mathbf{Z}, \mathbf{F}_t, \boldsymbol{\theta}, \boldsymbol{\Omega})$, $p(\mathbf{F}_t|\mathbf{F}_{t-1}, \mathbf{A}, \boldsymbol{\Sigma})$, $p(y_k|\boldsymbol{\beta}, \mathbf{Z}, \mathbf{F}_{3k}, \mathbf{F}_{3k-1}, \mathbf{F}_{3k-2}, y_{k-1}, \eta^2)$, and $p(z_j|p_j)$ are given according to the distributions in equations (3.6), (3.7), (3.8), and (3.5) respectively. Here $p(\boldsymbol{\Phi}^{cv})$ is the prior distribution for $\boldsymbol{\Phi}^{cv}$, which will be specified later.

To deal with the missing data in $\mathbf{x}_{i\in v_{q,t}}$ at the end of the sample ($t = T_q + 1, ..., T$), we define an indicator matrix $\mathbf{1}_{v_{q,t}}$ as a $n_{q,t} \times n$ matrix obtained by deleting the i^{th} row from the identity matrix $\mathbf{I}_{n\times n}$ if the corresponding $x_{i,t}$ is missing at q, where i could be any index from $i = 1, ..., n$. For the toy example discussed in Table 3.1 Section 3.1.2, we have

$$\mathbf{1}_{v_{q,T-2}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \quad \mathbf{1}_{v_{q,T-1}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{1}_{v_{q,T}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Then we can write $\mathbf{x}_{i\in v_{q,t}}$ as $\mathbf{x}_{i\in v_{q,t}} = \mathbf{1}_{v_{q,t}}\mathbf{x}_t$, thus the conditional density of $\mathbf{x}_{i\in v_{q,t}}$ in equation (3.9) is, for $t = T_q + 1, ..., T$,

$$\mathbf{x}_{i\in v_{q,t}} = \mathbf{1}_{v_{q,t}}\mathbf{x}_t|\boldsymbol{\mu}, \mathbf{Z}, \mathbf{F}_t, \boldsymbol{\Theta}, \boldsymbol{\Omega} \sim N(\mathbf{1}_{v_{q,t}}(\boldsymbol{\mu} + [\mathbf{I}_{n\times n} \otimes (\mathbf{Z}\mathbf{F}_t)'] * \boldsymbol{\theta}), \mathbf{1}_{v_{q,t}}\boldsymbol{\Omega}\mathbf{1}'_{v_{q,t}}). \tag{3.10}$$

We complete the model specification by assigning prior distributions for $\mathbf{\Phi}^{cv}$ and $\mathbf{Z}$ in Bayesian framework.

We set the prior for $\boldsymbol{\mu}$ as,

$$\boldsymbol{\mu}_{n\times 1} \sim N(\mathbf{0}_{n\times 1}, \mathbf{I}_{n\times n}). \tag{3.11}$$

The prior for $\mathbf{\Theta}$ is

$$\mathbf{\Theta}_{n\times R} \sim Matrix\ Normal(\mathbf{0}_{n\times R}, \mathbf{I}_{n\times n}, \mathbf{I}_{R\times R}). \tag{3.12}$$

In this way, the identification assumption F1 in Appendix A is satisfied by the law of large number.

The prior for $\mathbf{\Omega}$ is

$$\mathbf{\Omega}_{n\times n} \sim Inverse\ Wishart\Big(\frac{1}{n}\mathbf{I}_{n\times n}, \nu_\theta\Big), \tag{3.13}$$

where ν_θ is a pre-specified scalar. This prior allows $\boldsymbol{\epsilon}_t$ to be correlated across series, satisfying assumption M1 in Appendix A.

The prior for $\mathbf{A}$ is standard normal truncated at $[-1, 1]$, for $j = 1, ..., R$, that is

$$p(a_j) = \begin{cases} 0 & \text{if } a_j \le -1 \\ \dfrac{\phi(a_j)}{\Phi(1) - \Phi(-1)} & \text{if } -1 < a_j < 1 \\ 0 & \text{if } a_j \ge 1 \end{cases}, \tag{3.14}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are PDF and CDF for standard normal distribution respectively. And $p(\mathbf{A}) = \prod_{j=1}^{R} p(a_j)$.

The prior for $\mathbf{\Sigma}$ is, for $j = 1, ..., R$,

$$\sigma_j^2 \overset{iid}{\sim} Inverse\ Gamma(\alpha_s, \beta_s), \tag{3.15}$$

where $\alpha_s > 0, \beta_s > 0$ are prespecified scalars. And $p(\mathbf{\Sigma}) = \prod_{j=1}^{R} p(\sigma_j)$.

The prior for $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \boldsymbol{\beta}_3', \beta_4)'$ is

$$\boldsymbol{\beta}_{(3R+2)\times 1} \sim N(\mathbf{0}_{(3R+2)\times 1}, \mathbf{I}_{(3R+2)\times(3R+2)}). \tag{3.16}$$

The prior for η^2 is

$$\eta^2 \sim Inverse\ Gamma(\alpha_h, \beta_h), \tag{3.17}$$

where $\alpha_h > 0, \beta_h > 0$ are pre-specified scalars.

The prior for p_j is, for $j = 2, ..., R$,

$$p_j \sim (1 - \pi^j)\mathbf{1}_{\{p_j=0\}} + \pi^j Beta(\alpha_p, \beta_p), \tag{3.18}$$

where $\alpha_p > 0, \beta_p > 0$ are pre-specified scalars and π is a hyperparameter satisfying $0 < \pi < 1$. Notice that the prior of p_j start at $j = 2$, because we assume $z_1 = 1$ with probability 1 (i.e. $p(z_1 = 1) = 1$). We believe there is at least one important factor and the importance of factors decreases as j increases based on the identifiable conditions stated in Section 2.2. The prior of p_j looks like a zero-inflated Beta distribution with unequal weights depending on π^j $(j = 2, ..., R)$. The zero point mass has a weight of $(1 - \pi^j)$, implying a higher probability of obtaining exact zero for p_j as j increases.

The prior for the hyperparameter π is

$$\pi \sim Beta(\alpha_\pi, \beta_\pi), \tag{3.19}$$

where $\alpha_\pi > 0, \beta_\pi > 0$ are pre-specified scalars. Equations (3.18)-(3.19) imply the joint prior for $(p_1, ..., p_R, \pi)$ is $p(p_1, ..., p_R, \pi) \propto \prod_{j=2}^{R} p(p_j|\pi)p(\pi)$. We assume other priors are independent.

Following standard MCMC procedure, we derive the complete conditional distributions for each parameter and latent variables, and obtain posterior samples by simulating from these individual complete conditionals iteratively. More specifically, we obtain the posterior distribution $p(\Phi_i|\mathbf{\Phi}^{cv}_{-i}, \mathbf{Y}, \mathbf{X}_q, \mathbf{F}, \mathbf{Z})$ where Φ_i is the i^{th} element of $\mathbf{\Phi}^{cv}$ and $\mathbf{\Phi}^{cv}_{-i}$ contains all the parameters except for Φ_i, and the posterior distribution for latent variables including $p(\mathbf{F}_t|\mathbf{\Phi}^{cv}, \mathbf{Y}, \mathbf{X}_q, \mathbf{Z})$ for all t, and $p(\mathbf{Z}|\mathbf{\Phi}^{cv}, \mathbf{Y}, \mathbf{X}_q, \mathbf{F})$. In a Markov chain, the current state only depends on the most recent previous state. MCMC is a conditional simulation methodology that generates random samples from a given target distribution, i.e. the posterior distribution. Assume after a paticular iteration g within the MCMC algorithm, we obtain posterior samples $(\mathbf{\Phi}^{cv})^{(g)}, \mathbf{F}^{(g)}, \mathbf{Z}^{(g)}$. In the next iteration $g + 1$, we first generate $\Phi_1^{(g+1)}$ from $p(\Phi_1|(\mathbf{\Phi}^{cv}_{-1})^{(g)}, \mathbf{Y}, \mathbf{X}_q, \mathbf{F}^{(g)}, \mathbf{Z}^{(g)})$, then we generate $\Phi_2^{(g+1)}$ from $p(\Phi_2|\Phi_1^{(g+1)}, (\mathbf{\Phi}^{cv}_{\{-1,-2\}})^{(g)}, \mathbf{Y}, \mathbf{X}_q, \mathbf{F}^{(g)}, \mathbf{Z}^{(g)})$. We keep the process until we have all $(\Phi_i)^{(g+1)}$ for all i. Next, we generate $\mathbf{F}_t^{(g+1)}$ from $p(\mathbf{F}_t|(\mathbf{\Phi}^{cv})^{(g+1)}, \mathbf{Y}, \mathbf{X}_q, \mathbf{Z}^{(g)})$ for all t. Lastly we generate $\mathbf{Z}^{(g+1)}$

from $p(\mathbf{Z}|(\mathbf{\Phi}^{cv})^{(g+1)}, \mathbf{Y}, \mathbf{X}_q, \mathbf{F}^{(g+1)})$. This sequence of posterior samples forms a Markov Chain whose distribution converges to the target posterior distribution.

In estimation, we draw posterior samples from the above complete conditional distributions and use the means of the posterior samples as parameter estimates, i.e. $\hat{\Phi}_i^{cv} = \frac{1}{G}\sum_{g=1}^{G}(\Phi_i^{cv})^{(g)}$, where $g = 1, ..., G$ denotes the MCMC iterations after burn-in and $(\Phi_i^{cv})^{(g)}$ is the g^{th} posterior draw of parameter Φ_i^{cv}. Latent variables $\mathbf{F}_t$ is also esimated using posterior means, i.e. $\hat{\mathbf{F}}_t = \frac{1}{G}\sum_{g=1}^{G}\mathbf{F}_t^{(g)}$ for $t = 1, ..., T$. In each iteration, the number of latent factors is simply estimated by counting how many 1's in the diagonal of $\hat{\mathbf{Z}}^{(g)}$. Thus the estimated number of factors is obtained using $\hat{r} = \frac{1}{G}\sum_{g=1}^{G}\hat{r}^{(g)}$ where $\hat{r}^{(g)} = \sum_{j=1}^{R} z_j^{(g)}$. $\hat{r}$ is further rounded to the nearest integer. Appendix C provides the posterior distributions for all model parameters and latent variables $\mathbf{F}_t$ and $\mathbf{Z}$.

3.2.2 Nowcasting GDP y_{K+1} and Estimating the Number of Factors

After MCMC converges, we obtain sequences of posterior draws for parameters and latent variables, which can now be used to construct nowcasts. Recall that we have complete $\mathbf{x}_t$ for $t = 1, ..., T_q$, unbalanced $\mathbf{x}_{i\in v_{q,t}}$ for $t = T_q + 1, ..., T$, and associated y_k for $k = 1, ..., K$ available. Suppose we are at q ($q = 1, ..., Q$) in month T ($T = 3K+1, 3K+2, 3K+3$) of quarter K, the task is to nowcast y_{K+1} based on $\{\mathbf{x}_1, ..., \mathbf{x}_{T_q}, \mathbf{x}_{i\in v_{q,T_q+1}}, ..., \mathbf{x}_{i\in v_{q,T}}\}$ and $\{y_1, ..., y_K\}$.

Let $\beta_0^{(g)}$, $\boldsymbol{\beta}_i^{(g)}$ ($i = 1, 2, 3$), $\beta_4^{(g)}$, $\mathbf{A}^{(g)}$, $\mathbf{F}_t^{(g)}$ ($t = 1, ..., T$), and $\mathbf{Z}^{(g)}$ be the g^{th} posterior draws of parameters and latent variables after the burn-in period, where $g = 1, ..., G$. The nowcast can be calculated as follows, for model with constant volatility.

- Quarter $K+1$'s nowcast by BAY in the first month ($T = 3K+1$) is given by:

$$\hat{y}_{K+1}^{q} = \frac{1}{G}\sum_{g=1}^{G}\Big[\beta_0^{(g)} + (\boldsymbol{\beta}_1^{(g)})'\mathbf{Z}^{(g)}(\mathbf{A}^{(g)})^2\mathbf{F}_T^{(g)} + (\boldsymbol{\beta}_2^{(g)})'\mathbf{Z}^{(g)}\mathbf{A}^{(g)}\mathbf{F}_T^{(g)} + (\boldsymbol{\beta}_3^{(g)})'\mathbf{Z}^{(g)}\mathbf{F}_T^{(g)} + \beta_4^{(g)}y_K\Big], \quad (3.20)$$

- Quarter $K+1$'s nowcast by BAY in the second month ($T = 3K+2$) is given by:

$$\hat{y}_{K+1}^{q} = \frac{1}{G}\sum_{g=1}^{G}\Big[\beta_0^{(g)} + (\boldsymbol{\beta}_1^{(g)})'\mathbf{A}^{(g)}\mathbf{F}_T^{(g)} + (\boldsymbol{\beta}_2^{(g)})'\mathbf{F}_T^{(g)} + (\boldsymbol{\beta}_3^{(g)})'\mathbf{F}_{T-1}^{(g)} + \beta_4^{(g)}y_K\Big], \quad (3.21)$$

- Quarter $K+1$'s nowcast by BAY in the third month ($T = 3K+3$) is given by:

$$\hat{y}^q_{K+1} = \frac{1}{G}\sum_{g=1}^{G}\Big[\beta_0^{(g)} + (\boldsymbol{\beta}_1^{(g)})'\mathbf{F}_T^{(g)} + (\boldsymbol{\beta}_2^{(g)})'\mathbf{F}_{T-1}^{(g)} + (\boldsymbol{\beta}_3^{(g)})'\mathbf{F}_{T-2}^{(g)} + \beta_4^{(g)} y_K\Big]. \tag{3.22}$$

Above nowcasting equations imply that at any q within month T, we have posterior samples of $\mathbf{F}_T^{(g)}$, which may not be the case when we deal with a specific release structure/pattern in real time. Modifications are necessary depending on a specific release pattern. In the next section when we introduce the release structure used in simulations, we give an example of how we modify the nowcasting equations. Note that all of these $\beta_0^{(g)}$, $\boldsymbol{\beta}_i^{(g)}$'s, $\beta_4^{(g)}$, $\mathbf{A}^{(g)}$, $\mathbf{F}_t^{(g)}$, and $\mathbf{Z}^{(g)}$, are updated in every release date within a month. Therefore, nowcasts are re-produced for each $q = 1, ..., Q$.

3.3 Bayesian Approach in Nowcasting: Simulation Evidence

In this section, through numerical simulation, we investigate three questions on our BAY approach in nowcasting setting. The first question is whether it can identify the number of latent dynamic factors r correctly. The second question is whether it can produce reliable nowcasting results. And the third question is whether it can estimate the latent factors $\mathbf{F}_t$. We answer these three questions by evaluating estimation of the number of latent factors r, out-of-sample nowcasting performance, and in-sample estimation of $\mathbf{F}_t$.

We generate data according to the model in (3.4) where $T = 180$ (months), $K = 60$ (quarters), $r = 4$ (number of latent factors), and $Q = 3$ (releases in each month). Table 3.2 visualizes the data release pattern when nowcasting quarter $K+1$'s GDP in month $3K+1$, $3K+2$, and $3K+3$. For any $T = 3K+1, 3K+2, 3K+3$, when $q = 1$, RL1 cell becomes available; when $q = 2$, RL2 cell becomes available; when $q = 3$, RL3 cell becomes available. Notice that when $q = 1, 2$, we have $\mathbf{x}_T = \emptyset$ (an empty set), we cannot directly generate posterior samples $\mathbf{F}_T^{(g)}$. As a solution, we use $\tilde{\mathbf{F}}_T^{(g)} = A^{(g)}\mathbf{F}_{T-1}^{(g)}$ to replace $\mathbf{F}_T^{(g)}$ in (3.20)-(3.22). When $q = 3$, since $\mathbf{x}_T \neq \emptyset$, we can directly generate posterior samples $\mathbf{F}_T^{(g)}$.

Simulation In this simulation, we let $n = 50$ (monthly data series), and $(n_1, n_2, n_3) = (20, 20, 10)$ meaning 20 monthly series are released in the first and second releases, and 10 series become

Table 3.2: Data Releasing Format in Simulation study when nowcasting quarter $K+1$'s GDP in month $3K+1$, $3K+2$, and $3K+3$. RL stands for release. Cells in red represents data released in the first month $3K+1$. Cells in green represents data released on the second month $3K+2$. Cells in blue represents data released in the third month $3K+3$.

Quarter	K			$K+1$		
Month	$3K-2$	$3K-1$	$3K$	$3K+1$	$3K+2$	$3K+3$
n_3 series	Known	Known	Known	RL3	RL3	RL3
n_2 series	Known	Known	RL2	RL2	RL2	
n_1 series	Known	RL1	RL1	RL1		

available in the last release. The following parameters are used to generate the data: $\mathbf{A} = diag(0.9, -0.8, 0.75, 0.7)$; $\mathbf{\Sigma} = diag(5.5, 3, 0.5, 0.25)$; $\beta_0 = 0.5$, $\boldsymbol{\beta}_1 = (2, 1.4, 1, 0.35)'$; $\boldsymbol{\beta}_2 = (1, 0.7, 0.5, 0.175)'$; $\boldsymbol{\beta}_3 = (0.5, 0.35, 0.25, 0.0875)'$; $\beta_4 = 0.15$; $\eta^2 = 2$; each element of $\boldsymbol{\mu}$ is simulated from $N(10, 1)$; $\mathbf{\Omega}$ is simulated from $Inverse\ Wishart(50, \mathbf{I}_{50\times 50}/50)$; and $\mathbf{\Theta}$ is simulated from $Matrix\ Normal(\mathbf{0}, \mathbf{I}_{50\times 50}, \mathbf{I}_{4\times 4})$. This set-up assumes the importance of latent factors to monthly series decreases when moving from the first latent factor to the last latent factor. Also, the third month's factor has the greatest impact when generating GDP, the second month's factor has less contribution to the GDP, and the first month's factor is the least significant.

100 Monte Carlo sample paths are generated. For each data set, we nowcast the last 20 quarters' GDP's after each release during the 3 months in the quarter. Model parameters and latent variables are estimated based on the most recent past 10 years of data up to the nowcasting month of that quarter, i.e. $T = 121, 122, 123$ for each $K+1$. The hyper-parameter values are set as follows: $\nu_\theta = n + 2$ for the Inverse Wishart prior of $\mathbf{\Omega}$ in (3.13) such that the expectation of $\mathbf{\Omega}$ is $\frac{1}{n}\mathbf{I}_{n\times n}$; $(\alpha_s, \beta_s) = (2, R+2)$ for the Inverse Gamma prior of $\mathbf{\Sigma}$ in (3.15) such that the prior has infinite variance; $(\alpha_h, \beta_h) = (2, 0.0001)$ for the Inverse Gamma prior of η^2 in (3.17) such that the prior has infinite variance; $(\alpha_p, \beta_p) = (1, 3)$ for the Beta prior of p_j in (3.18) such that the prior places higher mass on smaller value; $(\alpha_\pi, \beta_\pi) = (2, 2)$ for the Beta prior of π in (3.19) such that the prior is symmetric at 0.5.

3.3.1 Estimating the Number of Latent Factors

Our algorithm requires a predetermined cap R for the number of factors r. R can be chosen to be the same as the number of monthly series n. However from a practical standpoint, we prefer to start with a relatively small value since a large R can result in heavy computational burden. Then depending on the results, we either increase R by a small step and re-run the nowcasting exercises or stop and conclude the current R is sufficient enough to estimate the number of factors correctly. We mimic these two scenarios in practice (under and over specify the cap) by running two parallel nowcasting exercises on all of the 100 data sets generated with $R = 3$ (under specified) and $R = 6$ (over specified) respectively, in hopes of providing practicle guidance of how to choose R.

Recall that we conduct nowcasting afer each release of 3 months for each of the last 20 quarters. Thus there are $20(quarters) \times 3(releases) \times 100(data\ sets) = 6,000$ estimates for the number of factors in each nowcasting month. Figure 3.1 shows the distribution of the 6,000 estimated number of factors in each month. The results for $R = 3$ are in the first row, and $R = 6$ are in the second row. Figure 3.1 suggests that the distribution of the estimated number of factors remains similar no matter which month we are in when doing nowcasts. Another interesting observation (not shown in this plot) is that the different releases have no effect on the estimation of the number of factors, or in other words, given a particular month of a quarter, the additional new released data series does not change the estimated number of factors. This is as expected since each additional new release only adds 20 new data points compared to the already existed 10 years full history of monthly series, hence the number of factors can be sufficiently estimated by the balanced portion of the data. When we under specify the cap for the number of factors (i.e. the 1st row when $R = 3$) we see a momotone increase in the bars and the peaks stops at the chosen cap $R = 3$. If we chose a larger $R > r$ (i.e. the 2nd row when $R = 6$), the bars peaks at the true number of factors, which is $r = 4$ in this case, and quickly dies off when exceeding r. The findings in Figure 3.1 offer a useful guidance for choosing R in practice. That is we can start with a relatively small R. If the results look similar to what is shown in the 1st row, it suggests that R is chosen too small. In this case one should increase R and re-run the nowcasting exercises until we obtain results similar to what is

shown in the 2nd or 3rd row. Then we can safely conclude the peak of the distribution represents the true number of factors, and the results are reliable.

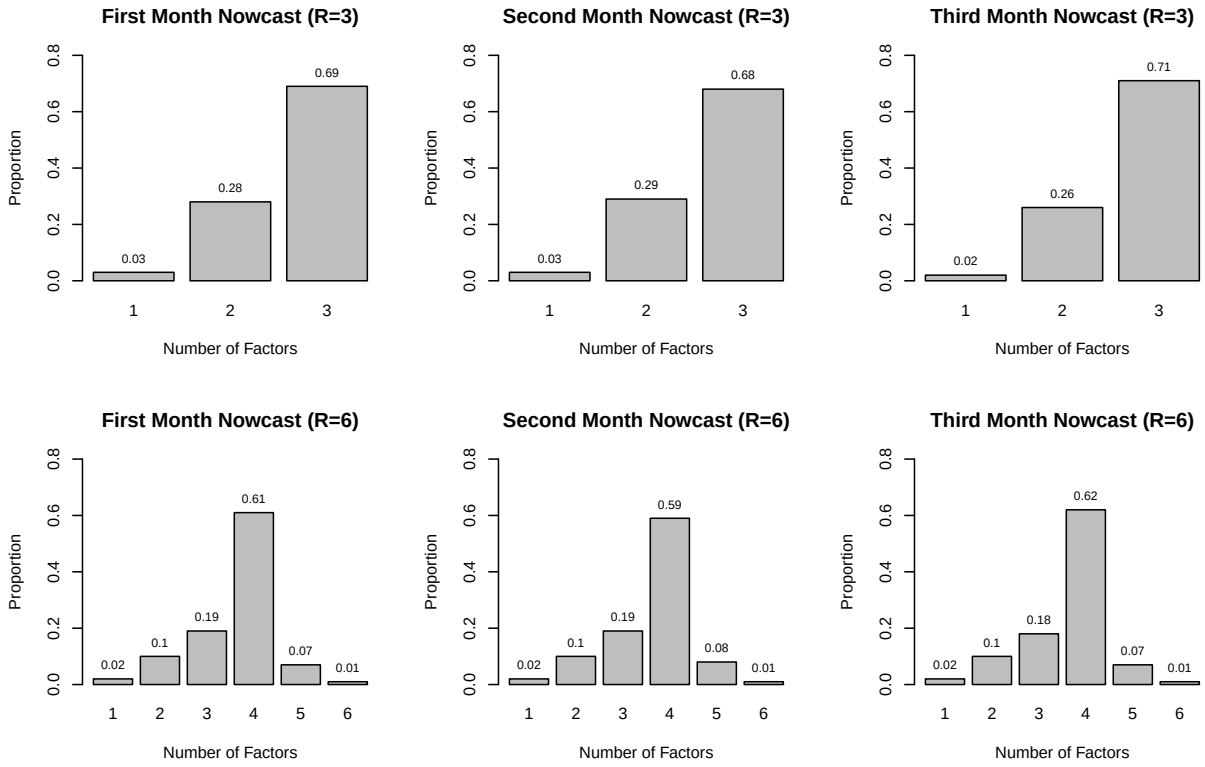

Figure 3.1: Distribution of estimated number of latent factors by month when the true number of factors $r = 4$. First row is the results for $R = 3$, second row is the results for $R = 6$. The first, second, and third columns represent nowcasting in the first, second, and third month of the quarter respectively. In each subplot, on the x-axis is the estimated number of factors, the height of the bars represent the porpotion of the estimated number of factors.

Recall for each of the last 20 quarters, there are $3(releases) \times 100(data\ sets) = 300$ estimates for the number of factors in each month. Since there is no difference in estimates among releases within the same month, we pick one estimate out of 3 releases in a month to represent the estimate of number of factors in that month. So there are 100 $\hat{r}$'s used in each subplot of Figure 3.2, where we present the 95% naive confidence interval based on these 100 estimates of the number of latent factors over 20 quarters period by month. The results for $R = 3$ are in the first column, and the

results for $R = 6$ are in the second column. The naive confidence intervals are calculated from normal approximation based on 100 nowcasts for each quarter. The flat black line represent the true number of factors, the red dashed line is the mean of the 100 estimates. The CIs do not capture the true number of latent factors in the first column since $R = 3$ is chosen too small. For $R = 6$, the CIs capture the true number of factors across all months in all quarters. We conclude jointly from Figure 3.1 and Figure 3.2, our BAY approach can adequately estimate the number of latent factors of DFMs in the context of nowcasting.

3.3.2 Out-of-sample Nowcasting Performance

Out-of-sample nowcasting performance is assessed based on 20 one-step-ahead nowcastings averaged across the 100 data sets. We use the mean absolute nowcasting error (MANE) as a measure of nowcasting accuracy. Let $\hat{y}^q_{K+1}$ be the nowcast calculated according to (3.20) - (3.22) at q^{th} release date of month T for a given data set, where $q = 1, 2, 3$ and $T = 3K + 1$ (first month nowcast), or $T = 3K + 2$ (second month nowcast), or $T = 3K + 3$ (third month nowcast) of the current quarter $K + 1$. Then $MANE(q, T) = 20^{-1} \sum_{K+1=41}^{60} |\hat{y}^q_{K+1} - y_{K+1}|$. Lastly, MANEs are averaged across the 100 data sets. To compare the nowcasting performance with the random walk (RW) approach, which takes GDP from the previous quarter as the nowcast for the current quarter GDP, relative MANEs (relative to the MANE of RW) are used. Figure 3.6 shows the ratios (in percentages) of MANE's of our BAY approach to the MANE of RW in 9 combinations of 3 releases and 3 nowcasting months, averaged across 100 data sets. Panel (a) is for $R = 3$, panel (b) is for $R = 6$. The horizontal line is 100% representing the baseline for RW, with first, second, and third release colored in dark gray, gray, and light gray respectively. Bars shorter than the reference line indicates that nowcasts are better than RW, otherwise worse than RW. Figure 3.6 suggests, moving from the first month to the third month, there are significant reductions in terms of MANE's. Comparing MANEs between releases in the same month, there is no significant change in MANEs from the first release to the second release, however, the third release sees a significant decrease in MANEs comparing to the second release.

This is due to the design of the release structure. Refer back to Table 3.2, take $T = 3K + 1$ as an example, when $q = 1$, $\mathbf{F}_T^{(g)}$ is projected using $\mathbf{F}_{T-1}^{(g)}$ which is estimated using balanced data from $t = 1, ..., T - 3$ $(3K - 2)$, $n_2 + n_3$ data points at $t = T - 2$ $(3K - 1)$, and n_3 data points at $t = T - 1$ $(3K)$; when $q = 2$, $\mathbf{F}_T^{(g)}$ is again projected using $\mathbf{F}_{T-1}^{(g)}$ which is now estimated using balanced data from $t = 1, ..., T-2$ $(3K-1)$, and n_2+n_3 data points at $t = T-1$ $(3K)$; when $q = 3$, since now $\mathbf{x}_T$ is not empty anymore, comparing to the first and second release, $\mathbf{F}_T^{(g)}$ can be directly estimated using balanced data from $t = 1, ..., T-2$ $(3K-1)$, n_2+n_3 data points at $t = T-1$ $(3K)$, and n_3 data points at $t = T$ $(3K + 1)$. Therefore, if a new release just add additional data points to month that already had some data information prior to the new release, the improvement in MANEs is negligible. If the new release adds new information of a month for the first time, the MANEs drop noticeably. This is true for $T = 3K + 2$ and $T = 3K + 3$. MANEs of $R = 6$ are uniformly better than MANEs of $R = 3$ for Simulation 1 since the true number of factors is 4. Comparing MANEs of Simulation 2 to Simulation 1 with $R = 6$, the latter has slightly better MANEs across all nine combinations of 3 releases and 3 nowcasting months due to the reason that the model is more parsimonious than the model with SV, thus retain higher nowcasting power. Table 3.3, the numerical summary of MANE reduction relative to RW for both simulations confirm the above observations. The BAY approach beats RW in terms of nowcasting errors.

To see how nowcasts results capture the trends of GDP, we choose one data set at random and look at its nowcasting performance across last 20 quarters. Figure 3.7 shows the nowcasting performance. The first, second, and third row are trend plots of nowcasts over the last 20 quarters in the first, second, and third month respectively. In each subplot, the curves colored in red, green and blue with different line types and knot types represent the nowcast results based on the first, second, and third release dates in a given month of a given quarter respectively. Our BAY approach gives reasonable nowcasting trend from the very first release in the first month all the way to the last release in the third month of a quarter. Nowcastings based on the 3rd release are slightly closer to the true trend comparing to nowcasts based on the 1st and 2nd release in the same month. Also

Table 3.3: This table reports the percentages of reduction in MANE's relative to RW, i.e. $(MANE^R - MANE^{RW})/MANE^{RW} \times 100$ (in %). Panel (a) is for $R = 3$, panel (b) is for $R = 6$

(a) Percentage of reductions in MANE's relative to RW for $R = 3$

Release	1^{st} Month	2^{nd} Month	3^{rd} Month
1st	-11.30%	-30.10%	-50.70%
2nd	-11.39%	-30.67%	-52.14%
3rd	-31.67%	-51.13%	-72.65%
Average	-18.12%	-37.30%	-58.50%

(b) Percentage of reductions in MANE's relative to RW for for $R = 6$

Release	1^{st} Month	2^{nd} Month	3^{rd} Month
1st	-11.67%	-32.58%	-53.37%
2nd	-11.93%	-33.08%	-53.72%
3rd	-32.92%	-54.39%	-77.01%
Average	-18.84%	-40.02%	-61.37%

improvement can be spotted when moving from nowcasts in the first month to nowcasts in the third month.

3.3.3 Estimation of Latent Variables

To assess the precision in the estimated latent state variables, we further look at the in-sample estimation of latent dynamic factors for the same randomly chosen data set the previous section. The in-sample analysis uses balanced information of the first 100 months of the data ($t = 1, ..., 100$).

Figure 3.8 plots the estimated in-sample fit of the first four latent factors versus the true value. The absolute value is plotted against each other since as mentioned previously, the factors are identifiable up to a change of sign. Figure 3.8 shows that the estimated latent factors are close to the true factors, which confirms that by introducing Assumption F and Assumption M, the latent factors are identifiable.

In summary, our simulation results suggest that the BAY method can estimate the number of latent factors correctly, has the ability to estimate the latent dynamic factors well and produce reliable nowcasting results.

3.4 Bayesian Approach in Nowcasting: Empirical Evidence

In this section, we examine the empirical performance of our BAY method in nowcasting the US's quarterly GDP growth rate.

The New York Fed Staff Nowcast platform updates its estimates of GDP growth for the current and subsequent quarter every Friday at 11:15 a.m., which can be found on the Federal Reserve Bank of New Yorks public website (https://www.newyorkfed.org/research/policy/nowcast). The underlying data set is also available on Github (https://github.com/FRBNY-TimeSeriesAnalysis/Nowcasting). We decide to borrow the data for our empirical study. We choose the data from 1993Q1 to 2016Q4 in our nowcasting exercise, the nowcasting horizon covers 2003Q1 to 2016Q4 and in each quarter we nowcast the current quarter's GDP. GDP (Y_k) enters the model as annualized quarter-over-quarter percentage change, which is calculated as

$$Y_k = \{(1 + \frac{GDP_k - GDP_{k-1}}{GDP_{k-1}})^4 - 1\} \times 100, \tag{3.23}$$

where GDP_k is the real GDP for quarter k. Figure 3.9 plots the real GDP growth rate computed by (3.23), and to the right of the dashed line is the nowcasting horizon of 2003Q1 to 2016Q4.

The latent common factors are distilled from 30 monthly data series, which are released by both government agencies and private institutions. The monthly series are assigned into 9 different categories/blocks (ADP National Employment Report, housing and construction, manufacturing, surveys, retail and consumption, income, labor, and international trade) based on economic insight. The real chronological release dates are hard to mimic since they vary from month to month, quarter to quarter, year to year. However, there is a roughly consistent pattern that can be used as a good proxy. Based on the approximate release date of each individual series and to align with simulation set up, we group the 30 monthly series into 3 release dates: 10th of the month, 20th of the month,

and 30th of the month (assume 30 days in each month for simplicity). The monthly series that released between release is treated as if they are released together on the release date. Three transformations are performed on the monthly series in the interest of inducing stationarity. Table 3.4, 3.5, and 3.6 provide release structure, data transformation, and details of each individual series. We nowcast 56 quarters' GDP using the BAY approach using $R = 6$.

Table 3.4: Data Releasing Format of month T for empirical data. RL stands for Release, with release 1 colored in red, release 2 colored in green, and releae 3 colored in blue. The number in parentheses is the number of series for that particular release. Set 1 to 6 are for notation puporse.

Month	$T-3$	$T-2$	$T-1$	T
Set 1	Known	Known	Known	RL2(2)
Set 2	Known	Known	RL1(3)	
Set 3	Known	Known	RL2(10)	
Set 4	Known	Known	RL3(5)	
Set 5	Known	RL1(7)		
Set 6	Known	RL2(3)		

Table 3.5: Data Transformation. x^*_{it} denotes the raw data and x_{it} denotes the transformed data.

Code	Transformation	Description
1	$x_{it} = x^*_{it}$	no transformation
2	$x_{it} = x^*_{it} - x^*_{i,t-1}$	level change
3	$x_{it} = \frac{x^*_{it} - x^*_{i,t-1}}{x^*_{i,t-1}} \times 100$	MoM % change

In each month, there are 56(quarters)×3(releases)=168 estimates of the number of factors. Figure 3.10 shows the distribution of the estimated number of latent factors by month. Consistently over the 3 nowcasting months, the estimated number of latent factors is 1 approximately 90% of the time, with the rest 10% all concentrated on 2 and completely dies off at 3 for both. This suggests that for these particular 30 US market macroeconomic series, 1 factor is sufficient to summarise the information. In fact, PCA on these 30 monthly series over 1993Q1 to 2016Q4 tells that the

first PC itself explains 60.29% of the variation among series, the second and third PC add 26.85% and 10.00% of the variation respectively, and the fourth PC only explains 2.56% of the variation.

Figure 3.11 plots the evolution of the out-of-sample GDP nowcasts over the last 56 quarters in each month based on 3 releases. In each subplot, the solid curve is the real GDP, while 3 other curves in different colors, line types and knot types represent nowcasting results from 3 different release dates. In general, our approach capture the trend of real GDP reasonably well. For the period prior to the 2008 to 2009 recession, nowcasts do not reflect much difference. During the recession, nowcasts have a clear lag in predicting the economy's downfall, but successfully capture the bouncing back trend when the economy recovers from the stress period.

The measure of nowcasting errors, MANE, can be computed in the same fashion as in simulation study. Figure 3.12 plots the relative ratios of MANEs. Three columns are for 3 nowcasting months, and 3 bars in each month represent 3 different release dates. The horizontal line located at 100% is the reference line for RW. Table 3.7 provides numerical comparisons relative to RW. The last row of Table 3.7 reports the averages of percentages of reduction in MANEs relative to RW across all 3 releases for that month. On average, BAY nowcasts see the largest reduction in nowcasting errors in the 1st month (23.31%), then drops significantly in the 2nd month (18.30%), and increase slightly in the 3rd month (20.08%). In general, unlike simulation study, Figure 3.12 and Table 3.7 indicate that as more data become available through the quarter, only some releases improve the nowcasting in terms of MANEs. Boivin and Ng (2006) studied how the size and the composition of the data affect the factor estimates in DFMs and concluded that sample size alone does not determine the properties of the estimates, the quality of the data must be taken into account. Therefore, the MANEs are not ensured to decrease as more data become available. Bok et. al. (2018) also found similar phenomenon. Another potential reason may be related to the notion of lagging and leading indicators in economics. A lagging indicator is an economic statistic that tends to have a delayed reaction to a change in the economic cycle. A leading indicator, on the other hand, is an economic statistic that tends to predict future changes in the economic cycle. Well-known lagging indicators such as unemployment, consumer price index, leading indicators

such as consumer confidence, construction sector are mixed in the data set used for our empirical study. Assuming the GDP function in (3.4) is appropriate, our model specification on the monthly data series explicitly assumes the data in the same month all reflect the economic condition, i.e. latent factors, in the same month, or in in other words, $\mathbf{x}_t$ is generated by $\mathbf{F}_t$. Lagging and leading indicators suggest $x_{i,t}$ is generated by $\mathbf{F}_{t-h}$ for some lagging indicators i with lag h, and $x_{j,t}$ is generated by $\mathbf{F}_{t+h'}$ for some leading indicators j with lead h'. If this is true, estimating $\mathbf{F}_t$ using data information from $\mathbf{x}_t$ is actually not ideal thus new release is not guaranteed to bring more information than noise.

Figure 3.13 plots the absolute estimated latent factors for the in-sample analysis. Since our BAY approach selects 1 latent factor for the in-sample period, only the first estimated latent factor is plotted.

Collectively, the empirical analysis in this section demonstrates the empirical relevance of the BAY approach in nowcasting the US's GDP for the time period considered. The results suggest one factor is sufficient to summarise the dynamics within the US market.

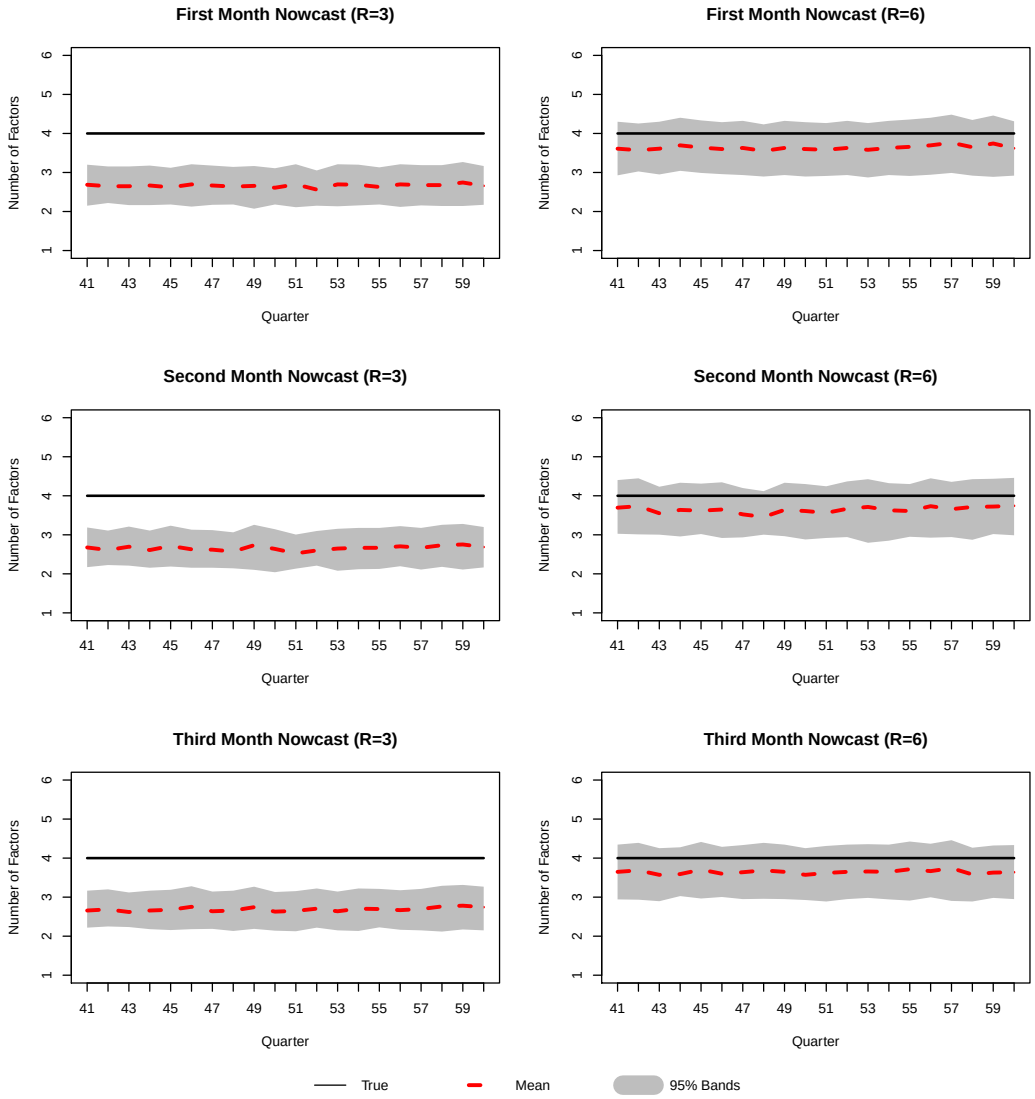

Figure 3.2: 95% confidence intervals of estimated number of latent factors by month, eolving over the last 20 quarters. Left column is the results for $R = 3$, right column is the results for $R = 6$. The first, second, and third column are nowcasting in the first, second, and third month of the quarter respectively. The solid flat line represents the true number of factors, the red dashed line represent the mean of estimated number of factors for that quarter. The gray shaded area is the 95% confidence intervals calculated using normal approximation based on 100 estimates.

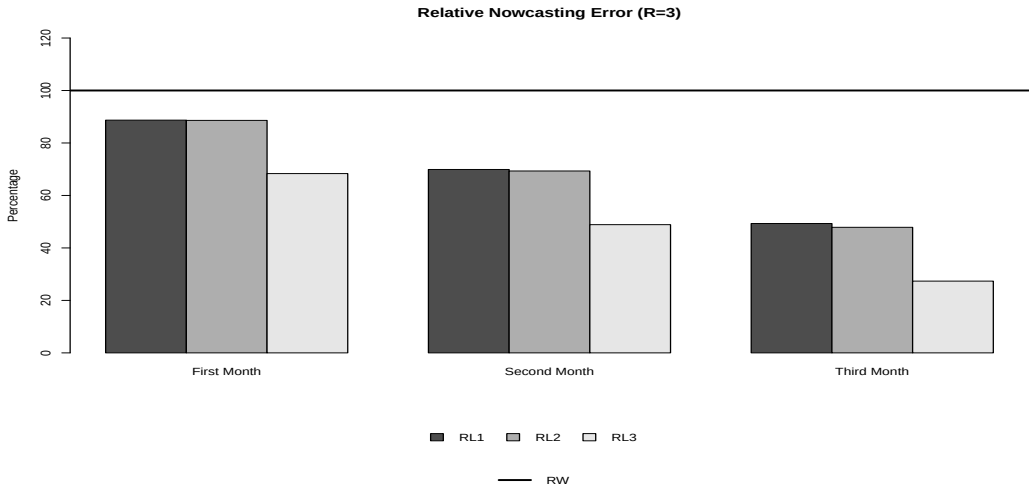

Figure 3.4 Averaged MANE ratios for $R = 3$

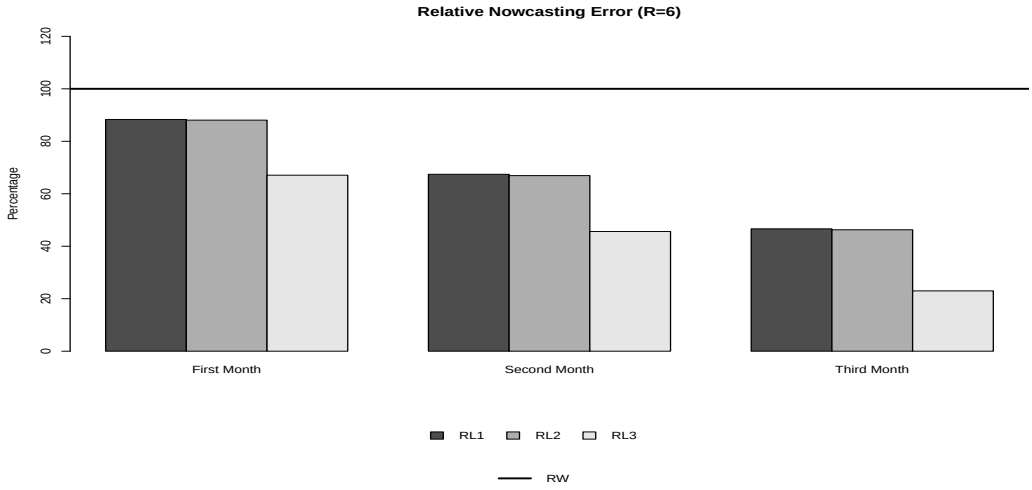

Figure 3.5 Averaged MANE ratios for $R = 6$

Figure 3.6: Averages mean absolute nowcasting error ratios (relative to RW). Panel (a) is for $R = 3$, panel (b) is for $R = 6$. The first, second, and third release are colored as dark gray, gray, and light gray.

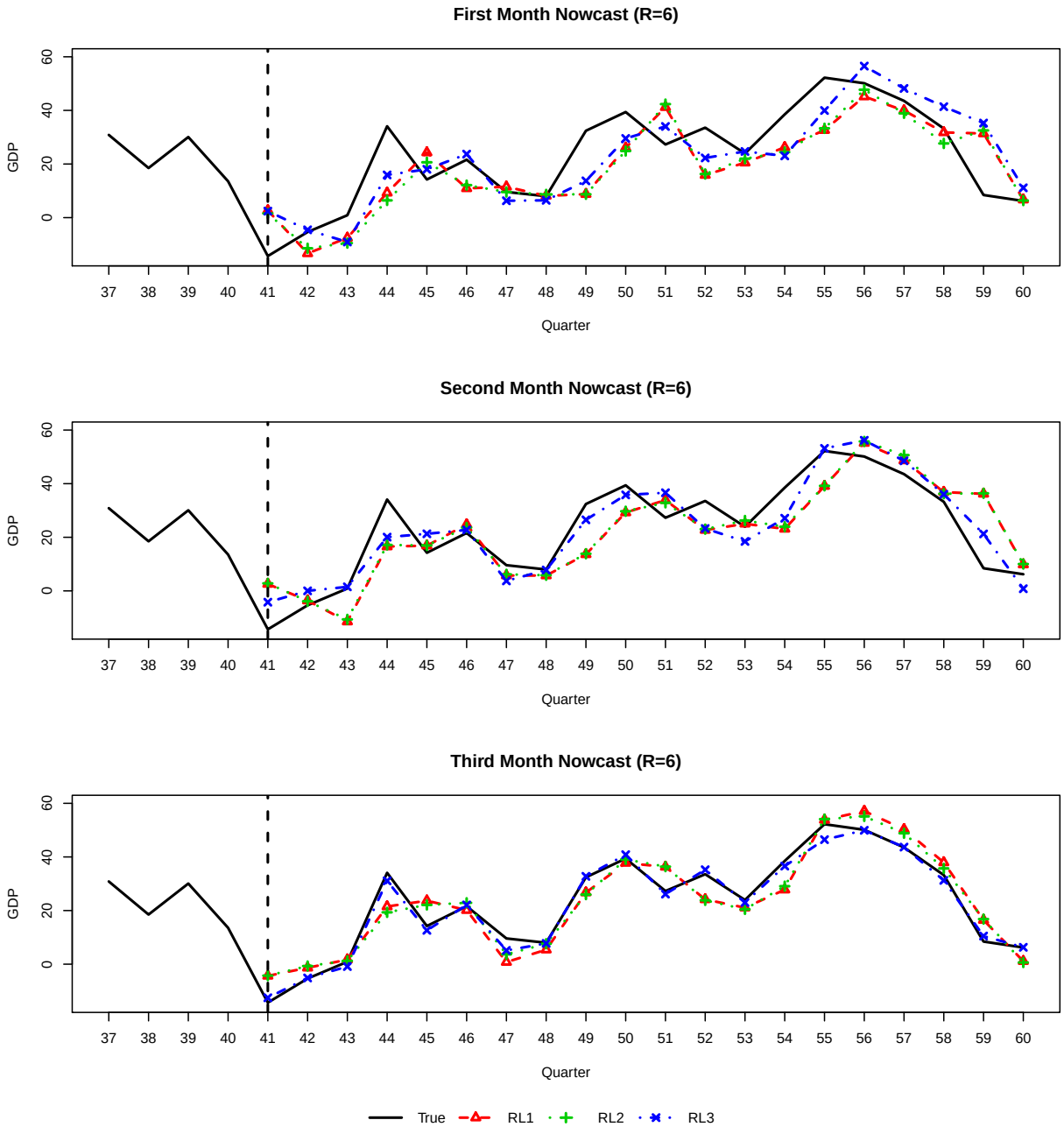

Figure 3.7: Nowcasting performance over the last 20 quarters by 3 releases in each month. Each row represents nowcasting in each month of the quarter. Black solid line represents the true GDP value, and dashed lines with different knot types represent the GDP nowcasts with red, green, and blue as release 1, 2, and 3.

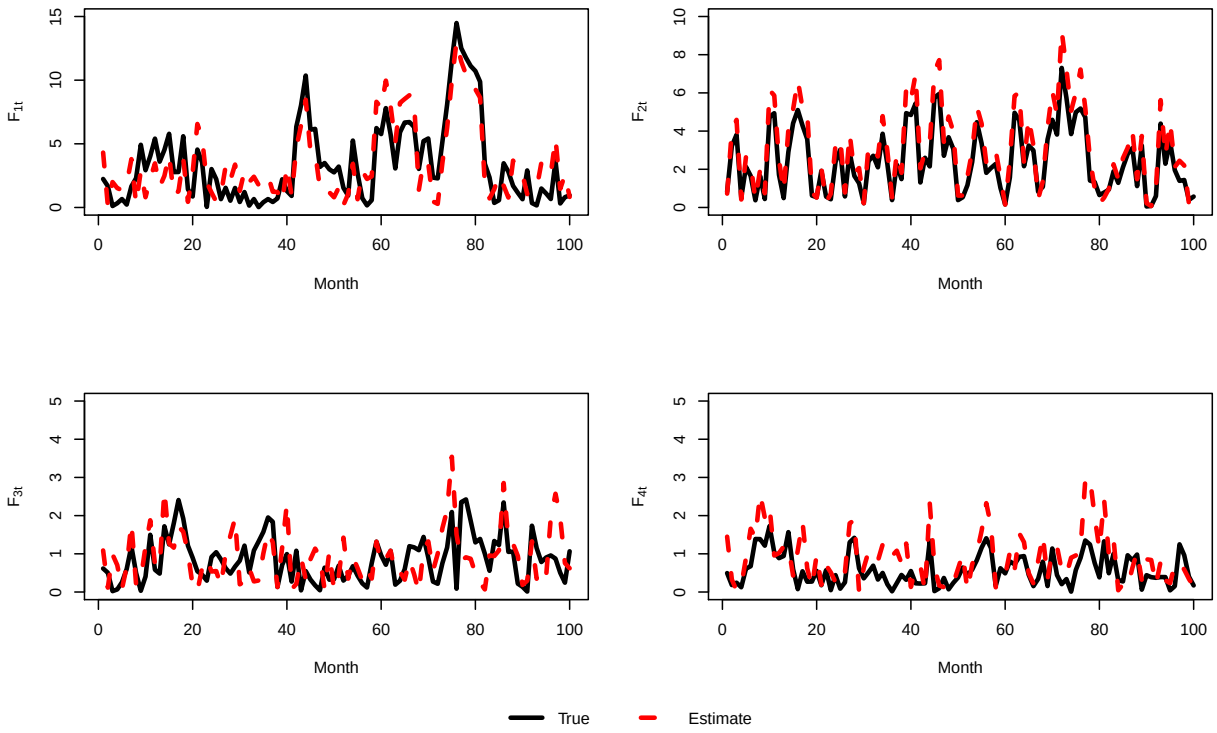

Figure 3.8: In-sample fit of latent factors for $R = 6$. Absolute value is used for the true latent factors and in-sample fits. Black solid line represents the true value and red dashed line represents the in-sample fitted value.

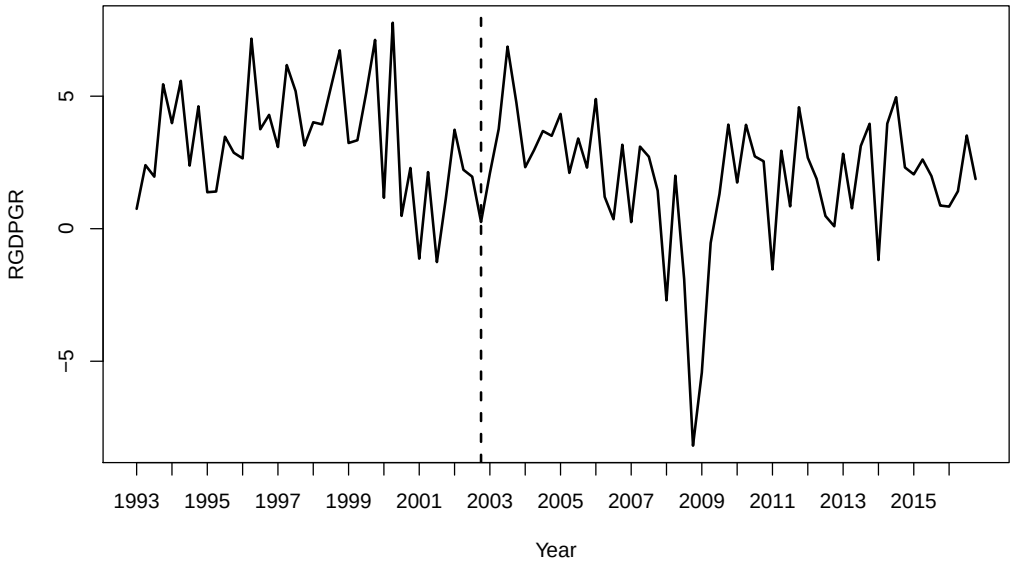

Figure 3.9: Quarterly real GDP growth rate in US. Right to the dashed line is the nowcasting horizon.

Table 3.6: Detailed information of monthly series used in empirical study. Series names are adopted from Federal Reserve Bank of St. Louis.

Release	Block	Names	Transformation	Lag
1st	Housing and Construction	TTLCONS	3	2
	International trade	BOPTEXP	3	2
		BOPTIMP	3	2
	Manufacturing	DGORDER	3	2
		BUSINV	3	2
		AMDMVS	3	2
		AMTMUO	3	2
	ADP national employment report	ADP nonfarm	2	1
	Labor	PAYEMS	2	1
		JTSJOL	2	1
2nd	Labor	UNRATE	2	2
	International trade	IR	3	1
		IQ	3	1
	Retail and Consumption	RSAFS	3	1
	Survey	GACDISA066MSFRBNY	1	0
		GACDFSA066MSFRBPHI	1	0
	Manufacturing	INDPRO	3	1
		WHLSLRIMSA	3	2
		TCU	2	1
		AMDMTI	3	2
	Other	CPIAUCSL	3	1
		PPIFIS	3	1
		CPILFESL	3	1
	Housing and Construction	HOUST	3	1
		PERMIT	2	1
3rd	Housing and Construction	HSN1F	3	1
	Income	DSPIC96	3	1
	Other	PCEPI	3	1
		GDPC1	3	1
	Retail and Consumption	PCEC96	3	1

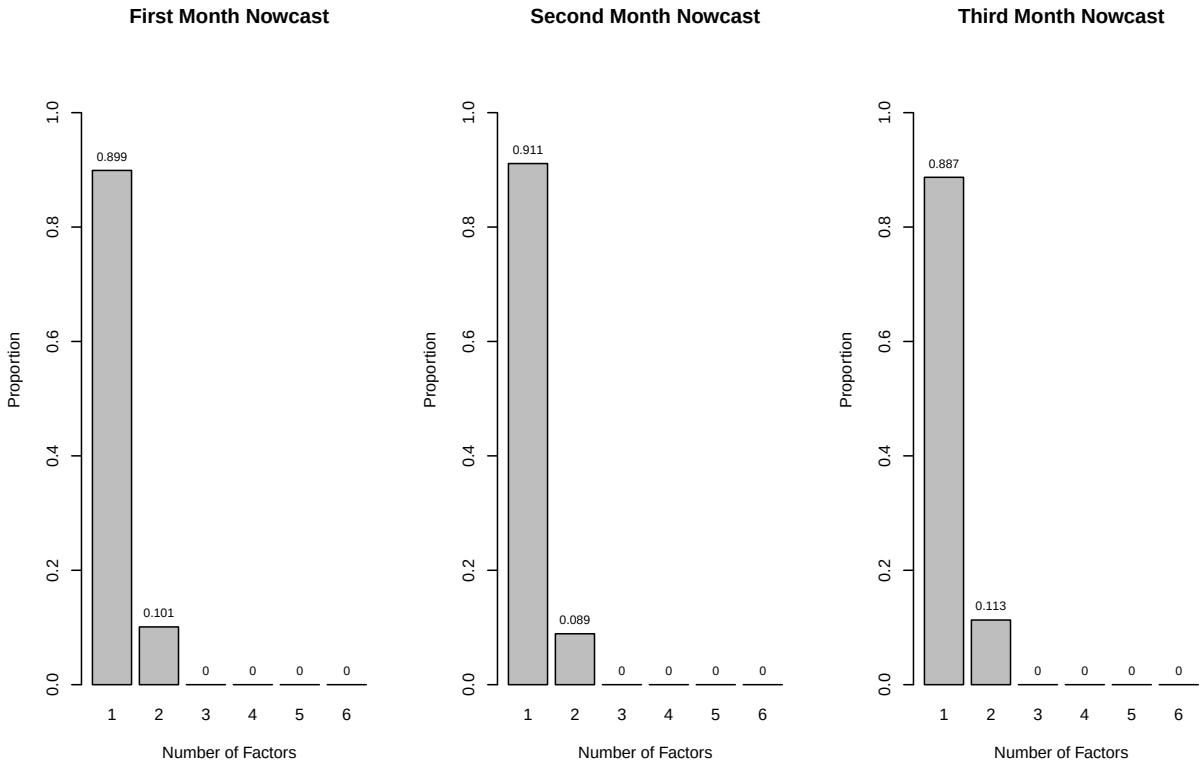

Figure 3.10: Distribution of estimated number of latent factors by month for the US market data. The left, middle, and right columns are nowcasting in the first, second, and third month of the quarter respectively.

Table 3.7: This table reports the percentages of reduction in MANE's of both methods relative to RW, i.e. $(MANE - MANE^{RW})/MANE^{RW} \times 100$ (in %) using the US market data.

Release	1^{st} Month	2^{nd} Month	3^{rd} Month
1st	-24.15%	-18.17%	-25.51%
2nd	-23.27%	-18.55%	-14.06%
3rd	-22.52%	-18.18%	-20.68%
Average	-23.31%	-18.30%	-20.08%

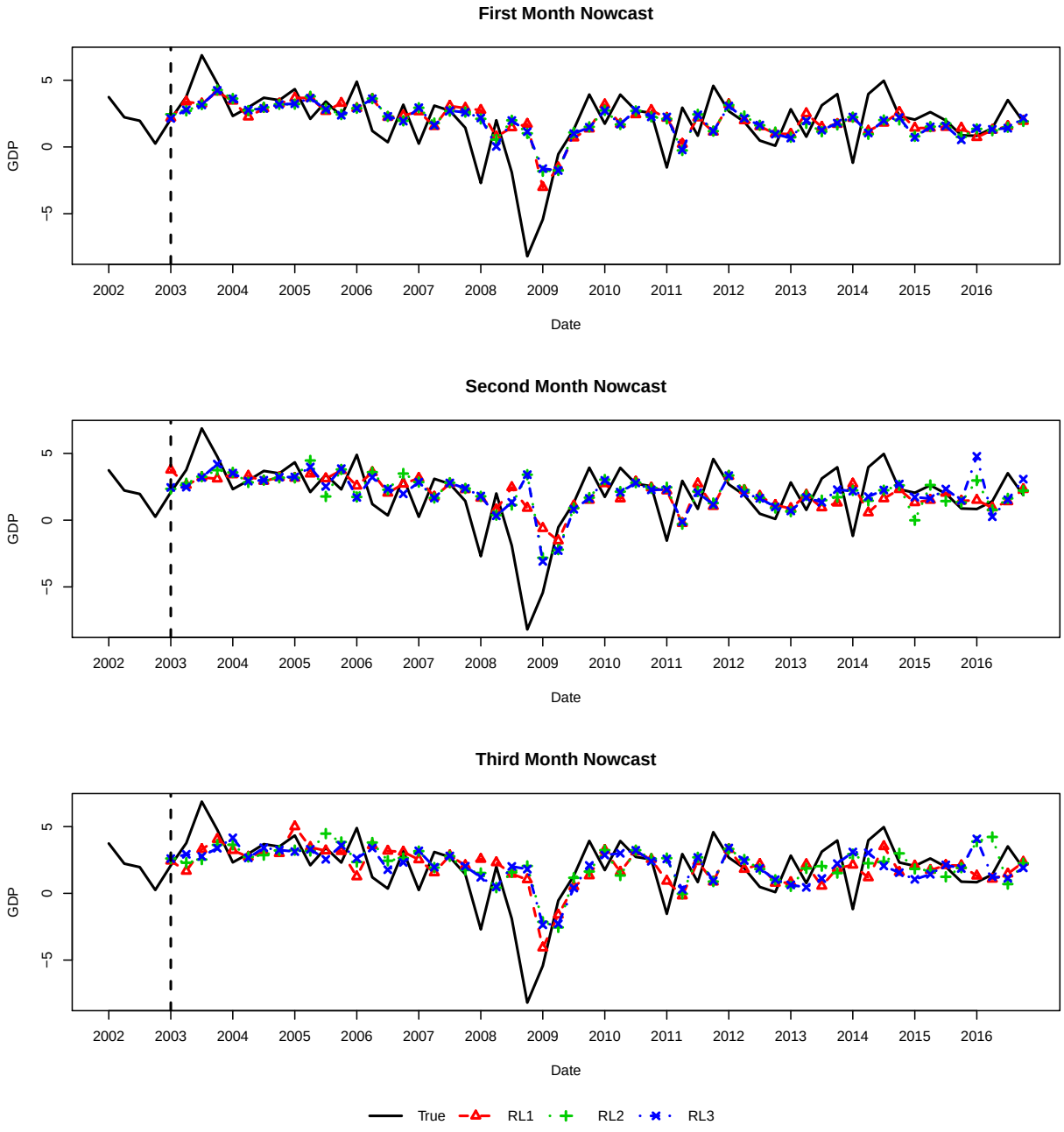

Figure 3.11: Nowcasting over 2003Q1 to 2016Q4 by 3 releases in each month for the US market data. Black solid line represents the true GDP value, dashed lines with different knot types represent the GDP nowcasts with red, green, and blue as release 1, 2, and 3.

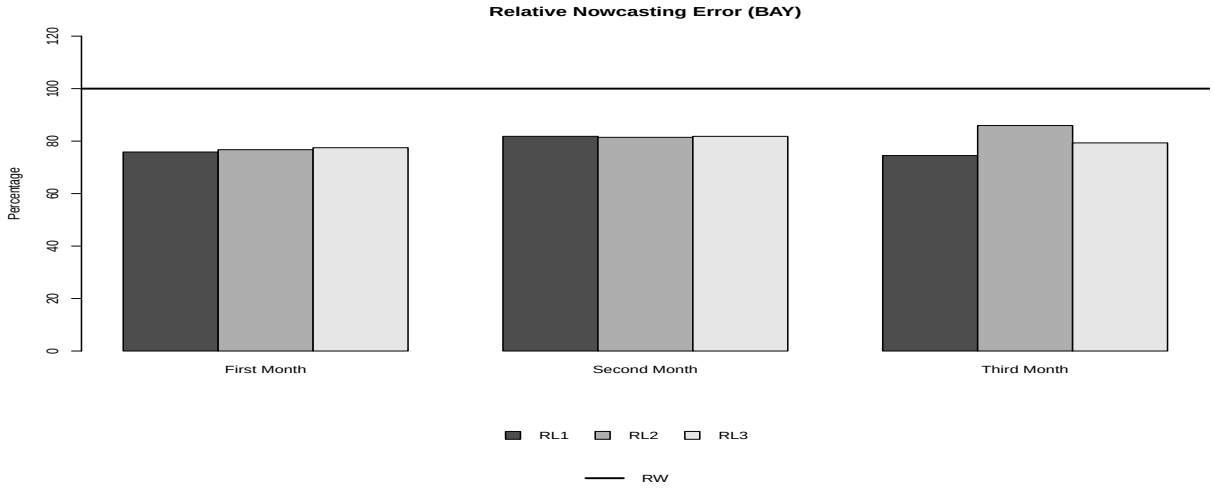

Figure 3.12: Averaged mean absolute nowcasting error ratios (relative to RW). The horizontal line is 100% representing the baseline for RW. The first, second, and third release are colored as dark gray, gray, and light gray for each month.

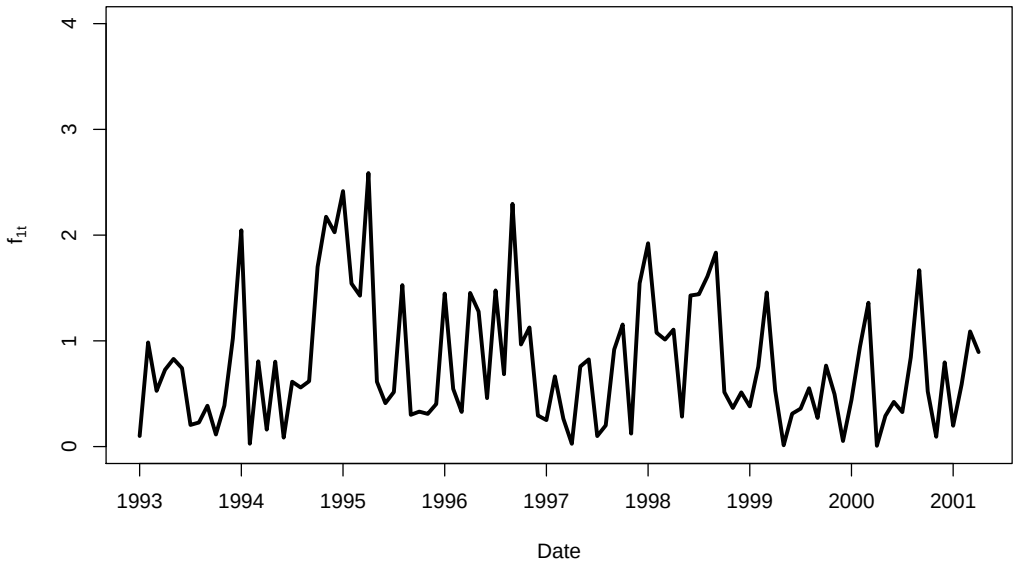

Figure 3.13: Absolute values of estimated first latent factors of in-sample analysis in the US market data.

CHAPTER 4. NOWCASTING GDP USING DYNAMIC FACTOR MODEL WITH UNKNOWN NUMBER OF FACTORS AND STOCHASTIC VOLATILITY

In this Chapter, we further extend athe Bayesian approach proposed in the previous chapter to allow time-varying volatilities in monthly series. We consider the particle Gibbs with backward simulation algorithm for obtaining potential stochastic volatility (SV) in monthly series effciently. We demonstrate the validity of our approach through simulation and explore the applicability of our approach through an empirical study using the same US's market data from previous chapter.

4.1 DFM with Stochastic Volatility

In this section, we extend the model in pervious chapter to allow the time-varying volatilities within monthly series.

Likely to be presented in macroeconomic data, SV has potential to result in more efficient estimates of coefficients and thus more accurate nowcasting if modeled accordingly. So we further consider a SV model in this chapter. Following Follett and Yu (2019), we allow time-varying variances across monthly series $\mathbf{x}_t$, by assuming its random noise $\boldsymbol{\epsilon}_t$ in (3.1) has time dependent covariance matrix $\boldsymbol{\Omega}_t$, i.e. $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Omega}_t)$ where $\boldsymbol{\Omega}_t$ is specified as

$$\boldsymbol{\Omega}_t = diag(e^{\omega_{1t}}, ..., e^{\omega_{nt}})\boldsymbol{\Psi} diag(e^{\omega_{1t}}, ..., e^{\omega_{nt}}). \tag{4.1}$$

Here the correlation matrix between the errors $\boldsymbol{\Psi}$ is assumed to be constant over time. The elements of the $n \times n$ matrix $\boldsymbol{\Psi}$ are denoted as ψ_{ij} for $i, j \in \{1, ..., n\}$ with $\psi_{ij} = 1$ when $i = j$ and $-1 \leq \psi_{ij} \leq 1$ when $i \neq j$. Thus, the variance of the i^{th} series is $\boldsymbol{\Omega}_{t[i,i]} = exp(2\omega_{it})$ and the covariance between i^{th} and j^{th} series $(i \neq j)$ is $\boldsymbol{\Omega}_{t[i,j]} = \psi_{ij} exp(\omega_{it} + \omega_{jt})$.

Since a form of slow, continuous, and time-varying parameters is much more desirable as mentioned in Primiceri (2005), Cogley and Sargent (2005), and Benati and Surico (2008), a random

walk process is postulated to model the log time-varying standard deviations:

$$\omega_{it} = \omega_{i,t-1} + e_{it}, \text{ for } i = 1, ..., n, \text{ and } t = 1, ..., T, \tag{4.2}$$

where $e_{it} \overset{iid}{\sim} N(0, \tau^2)$. Section 3.2 talks about how to estimate $\{\boldsymbol{\omega}_t = (\omega_{1t}, ..., \omega_{nt})' : t = 1, ..., T\}$ and the correlation matrix $\boldsymbol{\Psi}$.

At any release date q, we have observations $\mathbf{X}_q = \{\mathbf{x}_1, ..., \mathbf{x}_{T_q}, \mathbf{x}_{i \in v_{q,T_q+1}}, ..., \mathbf{x}_{i \in v_{q,T}}\}$ and $\mathbf{Y} = \{y_1, y_2, ..., y_K\}$; latent variables $\mathbf{Z}$, $\mathbf{F}$, and $\mathbf{W} = \{\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_T\}$; and model parameters $\boldsymbol{\Phi}^{sv} = \{\boldsymbol{\mu}, \boldsymbol{\Theta}, \mathbf{A}, \boldsymbol{\Sigma}, \beta_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \beta_4, \eta^2, p_1, ..., p_R, \boldsymbol{\Psi}\}$. The goals are to estimate model parameters and latent variables using the observables, and to nowcast the current quarter GDP y_{K+1} at every release date q for $q = 1, ..., Q$ in month T for $T = 3K + 1, 3K + 2, 3K + 3$.

4.2 Estimating Stochastic Volatility via MCMC

In this section, we discuss estimation of latent time-varying volatility for DFMs with SV.

We complete our Bayesian framework for DFMs with SV specified in (4.1) by assigning prior distributions to the SV components as follows. For model parameters that are in common (i.e. $\Phi_i \in \boldsymbol{\Phi}^{sv} \cap \boldsymbol{\Phi}^{cv}$), the same priors as the ones in previous chapter are assigned. The prior for the initial state ω_{i1} (log standard deviation at $t = 1$) is, for $i = 1, ..., n$,

$$\omega_{i1} \sim N(a_\omega, b_\omega), \tag{4.3}$$

where $\alpha_\omega > 0, \beta_\omega > 0$ are pre-specified scalars.

The prior for τ^2, which is the variance of the random process for log time-varying standard deviation in (4.2), is

$$\tau^2 \sim Inverse\ Gamma(\alpha_l, \beta_l), \tag{4.4}$$

where α_l, β_l are pre-specified scalars.

The prior for the correlation matrix $\boldsymbol{\Psi}$ in (4.1) is

$$\boldsymbol{\Psi} \sim LKJ(m) \tag{4.5}$$

where LKJ denotes the Lewandowski, Kurowicka and Joe (LKJ) distribution, proposed in Lewandowski et al. (2009). Random matrices that follow $LKJ(m)$ distribution have 1 on the diagonal elements with off-diagonal elements between -1 and 1. $LKJ(m)$ is a distribution over all possible correlation matrices. If $\mathbf{\Psi} \sim LKJ(m)$, then $p(\mathbf{\Psi}) \propto |\mathbf{\Psi}|^{m-1}$, where $m > 0$ is a shape parameter that controls the placement of the mass of the distribution. When $m = 1$, it is a uninfomative prior over all possible correlation matrices. When $0 < m < 1$, it has higher probabilities placed on non-zero correlations. When $m \to \infty$, it converges to the identity matrix. We choose $m = 1$ in our application. It is worth noting that $\mathbf{\Psi}$ generated from $LKJ(m)$ is automatically positive definite, thus leading to guaranteed positive definiteness of $\mathbf{\Omega}_t$ based on the sandwitch form in (4.1).

The joint posterior distribution in this case is

$$\begin{aligned}
& p(\mathbf{\Phi}^{sv}, \mathbf{F}, \mathbf{Z}, \mathbf{W} | \mathbf{Y}, \mathbf{X}_q) \propto p(\mathbf{Y}, \mathbf{X}_q, \mathbf{\Phi}^{sv}, \mathbf{F}, \mathbf{Z}, \mathbf{W}) \\
\propto & \prod_{t=1}^{T_q} p(\mathbf{x}_t | \boldsymbol{\mu}, \mathbf{Z}, \mathbf{F}_t, \boldsymbol{\theta}, \boldsymbol{\omega}_t, \mathbf{\Psi}) \prod_{t=T_q+1}^{T} p(\mathbf{x}_{i \in v_{q,t}} | \boldsymbol{\mu}, \mathbf{Z}, \mathbf{F}_t, \boldsymbol{\theta}, \boldsymbol{\omega}_t, \mathbf{\Psi}) \\
\times & \prod_{k=2}^{K} p(y_k | \boldsymbol{\beta}, \mathbf{Z}, \mathbf{F}_{3k}, \mathbf{F}_{3k-1}, \mathbf{F}_{3k-2}, y_{k-1}, \eta^2) \times \prod_{t=2}^{T} p(\mathbf{F}_t | \mathbf{F}_{t-1}, \mathbf{A}, \mathbf{\Sigma}) \times \prod_{j=1}^{R} p(z_j | p_j) \\
\times & \prod_{t=1}^{T} p(\boldsymbol{\omega}_t | \tau^2) \times p(\mathbf{\Phi}^{sv}),
\end{aligned} \tag{4.6}$$

where the first two pieces associated with $\mathbf{X}_q$ can be easily modeified from (3.6) and (3.10) by replacing $\mathbf{\Omega}$ with $\mathbf{\Omega}_t$ according to (4.1). The middle three pieces associated with $\mathbf{Y}$, $\mathbf{F}$, and $\mathbf{Z}$ follow exactly as before. Equation (4.2) together with (4.3) give the joint distribution of $\boldsymbol{\omega}_t$, i.e. $\prod_{t=1}^{T} p(\boldsymbol{\omega}_t) \propto \prod_{i=1}^{n} p(\omega_{i1}) \prod_{i=1}^{n} \prod_{t=2}^{T} p(\omega_{it} | \omega_{it-1}, \tau^2) p(\tau^2)$, where $\omega_{it} \sim N(\omega_{i,t-1}, \tau^2)$. $p(\mathbf{\Phi}^{sv})$ is the prior distribution for $\mathbf{\Phi}^{sv}$. The additional SV components do not change the estimation procedure mentioned in the previous section other than a few more complete conditional distributions for the SV components.

When estimating stochastic volatility in a Bayesian framework, two Metropolis procedures stand out. The first one is an accept/reject algorithm proposed by Kim et al. (1998), which addressed the problem of non-gaussian and non-linear when sampling the volatility parameters by transformations and approximations using an offset mixture of normals. Applications include Clark and Ravazzolo

(2014), Carriero et al. (2016), and Karlsson (2013). The other method can be found in Clark (2011), Cogley and Sargent (2005) in their use of the algorithm of Jacquier et al. (2002), which broke down the joint posterior of the vector of volatilities by considering the series of univariate conditional densities, and constructed a Metropolis chain by allowing repeats of points in the sample sequence in order to deal with costly-to-compute normalizing constants of the accept/reject density.

However, standard Metropolis procedure is no longer the preferred choice as pointed out in Andrieu et al. (2010) due to the unreliable performance caused by poorly chosen proposal distributions and/or updating highly correlated variables independently. Instead, they proposed the Particle Gibbs (PG) to sample from a potentially high dimensional posterior distribution and argued it was well suited for large nonlinear/non-Gaussian state-space models, such as a SV model. Whiteley (2010) mentioned the PG sampler can have poorly mixing when the particle filter inheres severe degeneracy but can be improved by adding backward simulation step to the sampler. Lindsten and Schön (2012) further derived an explicit PG sampler with backward simulation and validated its legitimacy of being an MCMC method. They also showed that the mixing can be increased considerably through backward simulation especially when a few particles are used and/or the number of observations is large. Follett and Yu (2019) employed this idea to sample the SV parameters $\{\boldsymbol{\omega}_t = (\omega_{1t}, ..., \omega_{nt})' : t = 1, ..., T\}$ in the context of VAR. Because its efficiency in filtering out high dimensional latent state variables, this particle Gibbs with backward simulation algorithm is also implemented in our Bayesian framework. Estimate of log standard deviation is computed as $\hat{\omega_{it}} = \frac{1}{G}\sum_{g=1}^{G} \omega_{it}^{(g)}$ for $i = 1, ..., n$ and $t = 1, ..., T$. Appendix D provides the posterior distributions for the SV components and details of the algorithm used to generate posterior samples of volatility following Follett and Yu (2019).

We assume all priors are independent. Following standard MCMC procedure, we derive the complete conditional distributions for each parameter and latent variables, and obtain posterior samples by simulating from these individual complete conditionals iteratively. More specifically, we obtain the posterior distribution $p(\Phi_i|\mathbf{\Phi}_{-i}^{sv}, \mathbf{Y}, \mathbf{X}_q, \mathbf{F}, \mathbf{Z}, \mathbf{W})$ where Φ_i is the i^{th} element of $\mathbf{\Phi}^{sv}$ and $\mathbf{\Phi}_{-i}^{sv}$ contains all the parameters except for Φ_i, and the posterior distribution for latent variables

including $p(\mathbf{F}_t|\mathbf{\Phi}^{sv}, \mathbf{Y}, \mathbf{X}_q, \mathbf{Z}, \mathbf{W})$ for all t, $p(\mathbf{Z}|\mathbf{\Phi}^{sv}, \mathbf{Y}, \mathbf{X}_q, \mathbf{F}, \mathbf{W})$, and $p(\mathbf{W}|\mathbf{\Phi}^{sv}, \mathbf{Y}, \mathbf{X}_q, \mathbf{F}, \mathbf{Z})$. The additional SV components are estimated by posterior means, by $\hat{\mathbf{\Phi}} = \frac{1}{G}\sum_{g=1}^{G} \mathbf{\Phi}^{(g)}$ and $\hat{\omega_{it}} = \frac{1}{G}\sum_{g=1}^{G} \omega_{it}^{(g)}$ for $i = 1, ..., n$ and $t = 1, ..., T$. Appendix D provides the posterior distributions for all model parameters and latent variables.

For model with SV components, the nowcasts follows exactly as equations in Section 3.2.2 since the SV components are not directly involved in the nowcasting equations.

4.3 Bayesian Approach in Nowcasting: Simulation Evidence

In this section, through numerical simulation, we investigate three questions on our BAY approach in nowcasting setting when SV is considered. The first question is, whether it can identify the number of latent factors r correctly. The second question is, whether it can produce reliable nowcasting results. And the third question is, whether it can estimate the latent factors $\mathbf{F}_t$. We answer these three questions by evaluating estimation of the number of latent factors r, out-of-sample nowcasting performance, and in-sample estimation of $\mathbf{F}_t$ and SV. The simulation has following set-ups:

Simulation In this simulation, we add SV as specified in Section 4.1. We shrink n to 30 to compromise the additional burden of Particle Gibbs with backward simulation algorithm and let $(n_1, n_2, n_3) = (10, 10, 10)$ in this case. The following parameters are used to generate SV process: $\mathbf{\Psi}$ is simulated from $LKJ(1)$; ω_{i1} is simulated from $N(0, 1)$; and $\tau^2 = 0.1$.

Again, 100 Monte Carlo sample paths are generated. For each data set, we nowcast the last 20 quarters' GDP's after each release during the 3 months in the quarter. Model parameters and latent variables are estimated based on the most recent past 10 years of data up to the nowcasting month of that quarter, i.e. $T = 121, 122, 123$ for each $K+1$. The hyper-parameter values are set as follows: $(\alpha_\omega, \beta_\omega) = (0, 1)$ for the Normal prior of ω_{i1} in (4.3) for simplicity; and $(\alpha_l, \beta_l) = (2, 1/n)$ for the Inverse Gamma prior of τ^2 in (4.4) such that the prior has infinite variance. The number of particle P is chosen to be 5. For parameters and hyper-parameters that are not mentioned here

will have the same set-up with simulation study in previous chapter. The data release structure also follows previous chapter.

4.3.1 Estimating the Number of Latent Factors

The primary task for the simulation is to investigate if our BAY approach can estimate the number of factors when there is additional estimation burden for latent SV components. Therefore, we focus on nowcasting using $R = 6$ only.

Figure 4.1 shows the distribution of the 6,000 estimated number of latent factors in each month for the simulation study. Consistent with the findings of simulation study in the previous chapter, the distribution of the estimated number of latent factors remains similar no matter which month we are in when doing nowcasts The distribution of the estimated number of factors remains the same as in the previous chapter when $R = 6$, except that the propotions of the peaks are slightly lower. This is due to the additional estimation burden for SV latent variables.

In Figure 4.2, present the 95% naive confidence interval based on these 100 estimates of the number of latent factors over 20 quarters period by month. The true $r = 4$ line falls into the CIs most of the time, but not as promising as than the previous chapter when $R = 6$. This indicates that we lose some power of identifying the correct number of factors when carrying out additional task of estimating SV. We conclude jointly from Figure 4.1 and Figure 4.2, our BAY approach can adquately estimate the number of latent factors in a nowcasting framework when R is appropriately chosen.

4.3.2 Out-of-sample Nowcasting Performance

Out-of-sample nowcasting performance is assessed based on 20 one-step-ahead nowcastings averaged across the 100 data sets. MANEs defined in the previous chapter are used as a measure of nowcasting accuracy. Figure 3.6 shows the ratios (in percentages) of MANE's of our BAY approach to the MANE of RW in 9 combinations of 3 releases and 3 nowcasting months, averaged across 100 data sets. The horizontal line is 100% representing the baseline for RW, with first, second,

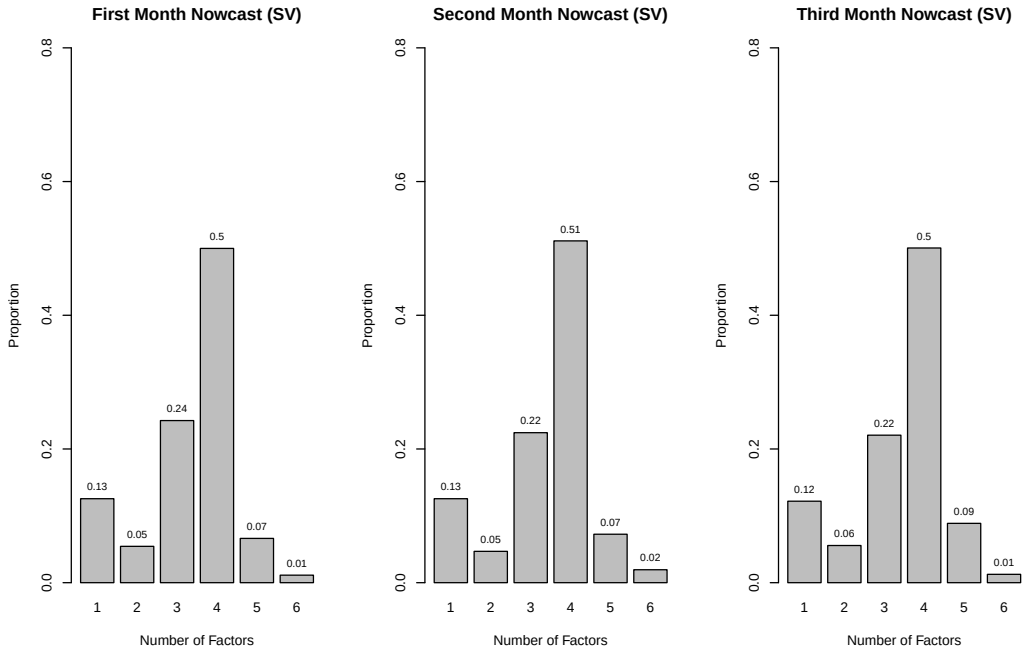

Figure 4.1: Distribution of estimated number of latent factors by month when the true number of factors $r = 4$. The first, second, and third columns represent nowcasting in the first, second, and third month of the quarter respectively. In each subplot, on the x-axis is the estimated number of factors, the height of the bars represent the porpotion of the estimated number of factors.

and third release colored in dark gray, gray, and light gray respectively. Bars shorter than the reference line indicates that nowcasts are better than RW, otherwise worse than RW. Figure 4.3 suggests, moving from the first month to the third month, there are significant reductions in terms of MANE's. Comparing MANEs between releases in the same month, there is no significant change in MANEs from the first release to the second release, however, the third release sees a significant decrease in MANEs comparing to the second release.

Table 4.1, the numerical summary of MANE reduction relative to RW for both simulations confirm the above observations. The BAY approach beats RW in terms of nowcasting errors. Comparing MANEs to the simulation in the previous chapter with $R = 6$, the latter has slightly better MANEs across all nine combinations of 3 releases and 3 nowcasting months due to the reason that the model is more parsimonious than the model with SV, thus retain higher nowcasting power.

Table 4.1: This table reports the percentages of reduction in MANE's relative to RW, i.e. $(MANE - MANE^{RW})/MANE^{RW} \times 100$ (in %).

Release	1^{st} Month	2^{nd} Month	3^{rd} Month
1st	-10.25%	-26.23%	-47.90%
2nd	-10.02%	-27.87%	-49.80%
3rd	-27.08%	-48.13%	-67.87%
Average	-15.78%	-34.08%	-55.19%

To see how nowcasts results capture the trends of GDP, we choose one data set at random and look at its nowcasting performance across last 20 quarters. Figure 4.4 shows the nowcasting performance. The first, second, and third row are trend plots of nowcasts over the last 20 quarters in the first, second, and third month respectively. In each subplot, the curves colored in red, green and blue with different line types and knot types represent the nowcast results based on the first, second, and third release dates in a given month of a given quarter respectively. Our BAY approach gives reasonable nowcasting trend from the very first release in the first month all the way to the last release in the third month of a quarter. Nowcastings based on the 3rd release are slightly closer to the true trend comparing to nowcasts based on the 1st and 2nd release in the same month. Also improvement can be spotted when moving from nowcasts in the first month to nowcasts in the third month.

4.3.3 Estimation of Latent Variables

To assess the precision in the estimated latent state variables, we further look at the in-sample estimation of latent dynamic factors and SV for the same randomly chosen data set the previous section. The in-sample analysis uses balanced information of the first 100 months of the data $(t = 1, ..., 100)$.

Figure 4.5 plots the estimated in-sample fit of the first four latent factors versus the true value. The absolute value is plotted against each other since as mentioned previously, the factors are identifiable up to a change of sign. Figure 4.5 shows that the estimated latent factors are close to

the true factors, which confirms that by introducing Assumption F and Assumption M, the latent factors are identifiable.

To see if the the Particle Gibbs with backward simulation algorithm can effectively capture the time-varying volatility structure, we summarize in Figure 4.6 and Figure 4.7 the estimated SV components over the in-sample period. Figure 4.6 is the heat map of the difference between estimated correlation matrix and the true value, i.e. $(\hat{\mathbf{\Psi}} - \mathbf{\Psi})$ where $\hat{\mathbf{\Psi}}$ is the posterior mean of the correlation matrix. Positive differences are colored in blue, negative differences are colored in red with lighter color represents a smaller difference. A majority of the area colored in either light color or white in the heat map indicates we have a reasonable estimation of correlation matrix. Figure 4.7 is the trend plots of the estimated log time-varying standard deviations versus the true value for each individual monthly series, i.e. $\{\hat{\omega}_{i1}, ..., \hat{\omega}_{iT}\}$ for all i where $\hat{\omega}_{it}$ the posterior meam of log standard deviation of i^{th} series at month t. We see that for most of the 30 series, the estimated log time-varying standard deviation (red dashed line) is very close to the true value (black line), indicating the Particle Gibbs with backward simulation algorithm indeed does a decent job of estimating the stochastic volatility well.

In summary, our simulation results suggest that the BAY method can estimate the number of latent factors correctly, has the ability to estimate the latent dynamic factors and SV well and produce reliable nowcasting results.

4.4 Bayesian Approach in Nowcasting: Empirical Evidence

In this section, we examine the empirical performance of our BAY method in nowcasting the US's quarterly GDP growth rate when we assume time-varying volatilities within monthly series. Information about empirical data can be found in the pervious chapter.

In each month, there are 56(quarters)×3(releases)=168 estimates of the number of factors. Figure 4.8 shows the distribution of the estimated number of latent factors by month. Consistently over the 3 nowcasting months, the estimated number of latent factors is 1 approximately 90% of the time, with the rest 10% all concentrated on 2 and completely dies off at 3 for both. This suggests

that for these particular 30 US market macroeconomic series, 1 factor is sufficient to summarise the information.

Figure 4.9 plots the evolution of the out-of-sample GDP nowcasts over the last 56 quarters in each month based on 3 releases. In each subplot, the solid curve is the real GDP, while 3 other curves in different colors, line types and knot types represent nowcasting results from 3 different release dates. In general, our approach capture the trend of real GDP reasonably well. For the period prior to the 2008 to 2009 recession, nowcasts do not reflect much difference. During the recession, nowcasts have a clear lag in predicting the economy's downfall, but successfully capture the bouncing back trend when the economy recovers from the stress period.

The measure of nowcasting errors, MANE, can be computed in the same fashion as in simulation study. Figure 4.10 plots the relative ratios of MANEs. Three columns are for 3 nowcasting months, and 3 bars in each month represent 3 different release dates. The horizontal line located at 100% is the reference line for RW. Table 4.2 provides numerical comparisons relative to RW. The last row of Table 4.2 reports the averages of percentages of reduction in MANEs relative to RW across all 3 releases for that month. On average, the percentages of reduction in nowcasting errors increases noticeably when moving from the 1st month (21.56% reduction), to the 2nd month (26.02% reduction), then decreases slightly in the 3rd month (25.18% reduction). In general, unlike simulation study, Figure 4.10 and Table 4.2 indicate that as more data become available through the quarter, only some releases improve the nowcasting in terms of MANEs.

Table 4.2: This table reports the percentages of reduction in MANE's of both methods relative to RW, i.e. $(MANE - MANE^{RW})/MANE^{RW} \times 100$ (in %) using the US market data.

Release	1^{st} Month	2^{nd} Month	3^{rd} Month
1st	-20.92%	-30.96%	-28.45%
2nd	-22.84%	-24.79%	-26.39%
3rd	-20.91%	-22.30%	-20.69%
Average	-21.56%	-26.02%	-25.18%

Figure 4.11 plots the absolute estimated latent factors for the in-sample analysis. Since our BAY approach selects 1 latent factor for the in-sample period, only the first estimated latent factor is plotted. Figure 4.12 shows the estimated volatility (i.e. $\mathbf{\Omega}_{t[i,i]}$ for all i) for each individual monthly series over the in-sample period. It is clear that the volatility is time-varying for all monthly series considered in this chapter.

Collectively, the empirical analysis in this section demonstrates the empirical relevance of the BAY approach in nowcasting the US's GDP for the time period considered. The results suggest one factor is sufficient to summarise the dynamics within the US market.

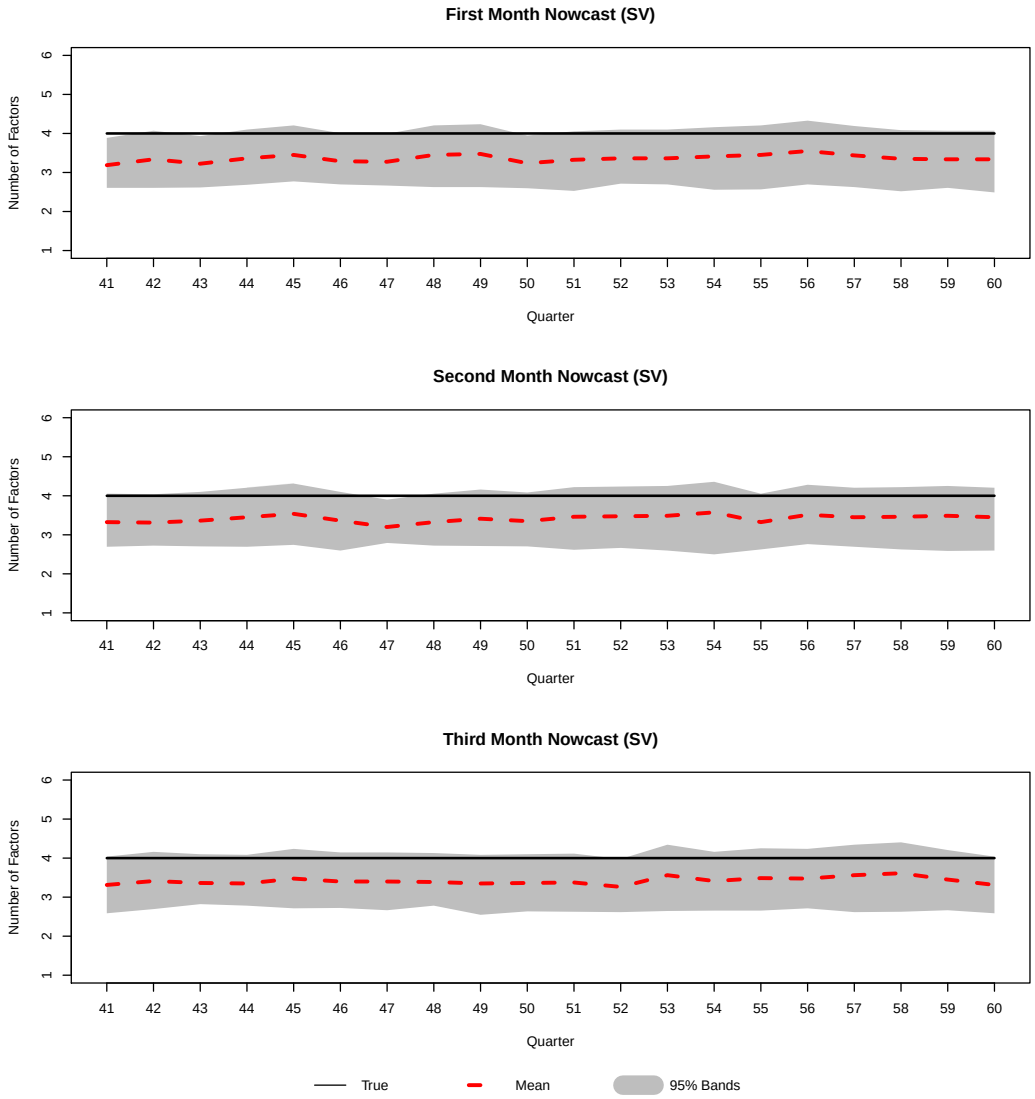

Figure 4.2: 95% confidence intervals of estimated number of latent factors by month, eolving over the last 20 quarters. The first, second, and third column are nowcasting in the first, second, and third month of the quarter respectively. The solid flat line represents the true number of factors, the red dashed line represent the mean of estimated number of factors for that quarter. The gray shaded area is the 95% confidence intervals calculated using normal approximation based on 100 estimates.

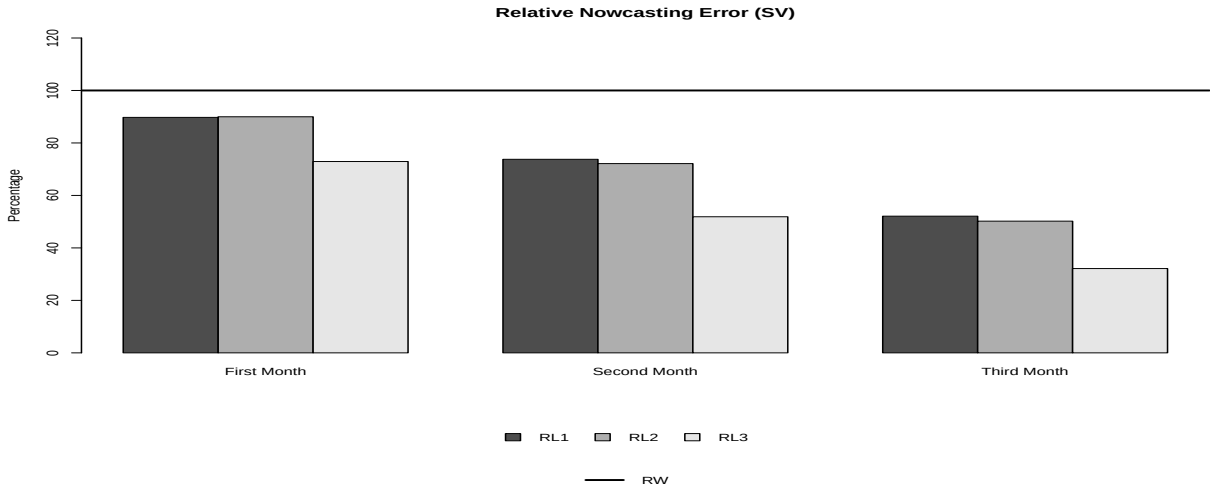

Figure 4.3: Averages mean absolute nowcasting error ratios (relative to RW). The first, second, and third release are colored as dark gray, gray, and light gray.

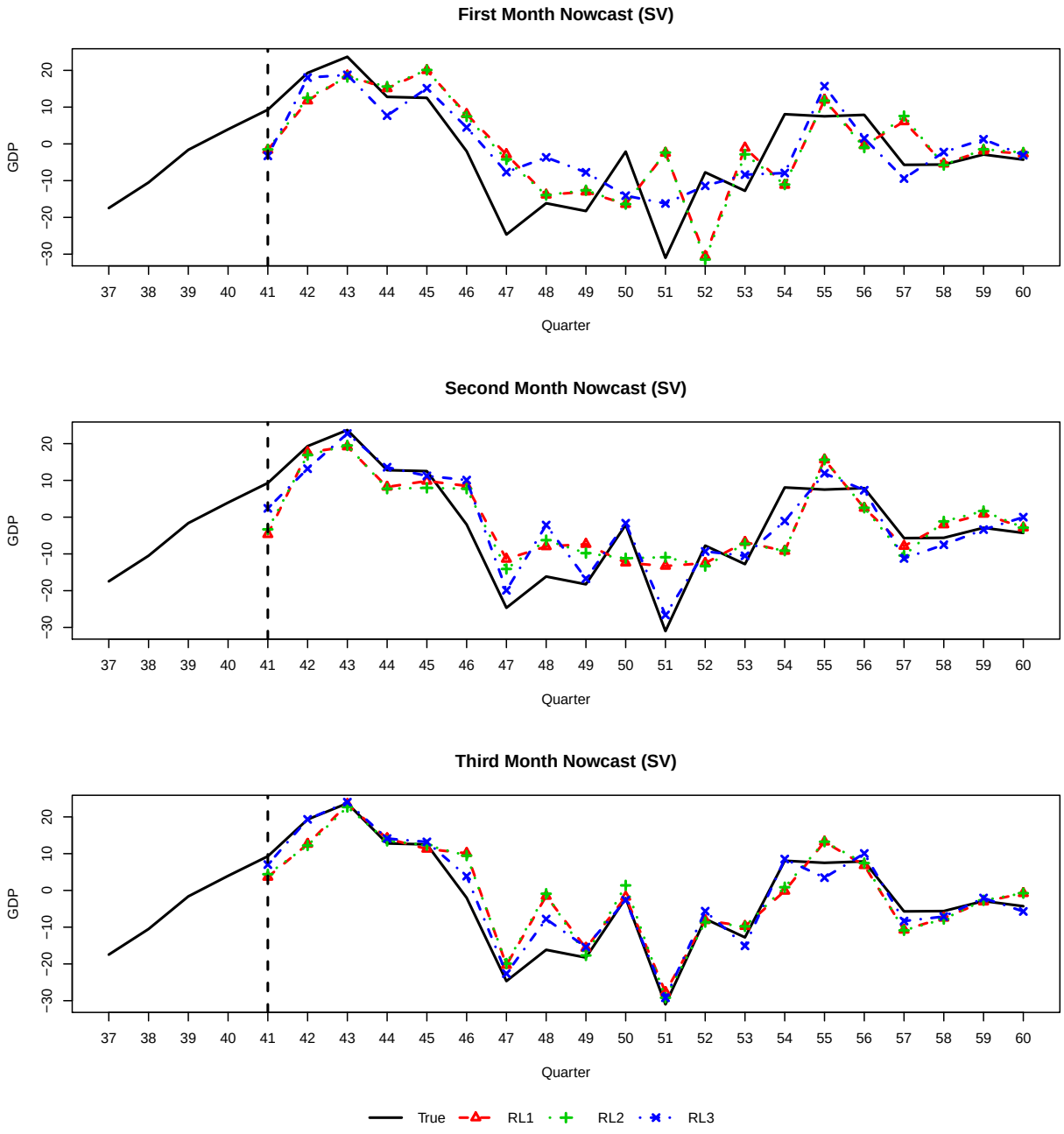

Figure 4.4: Nowcasting performance over the last 20 quarters by 3 releases in each month. Each row represents nowcasting in each month of the quarter. Black solid line represents the true GDP value, and dashed lines with different knot types represent the GDP nowcasts with red, green, and blue as release 1, 2, and 3.

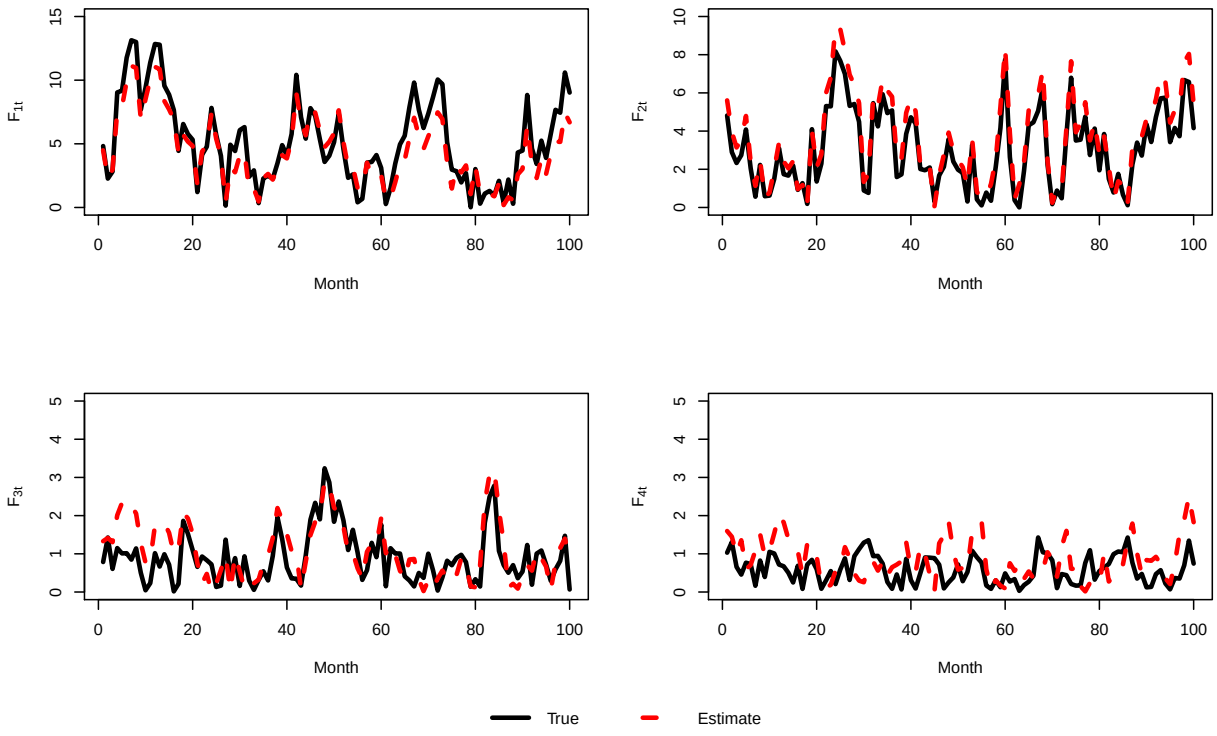

Figure 4.5: In-sample fit of latent factors for $R = 6$. Absolute value is used for the true latent factors and in-sample fits. Black solid line represents the true value and red dashed line represents the in-sample fitted value.

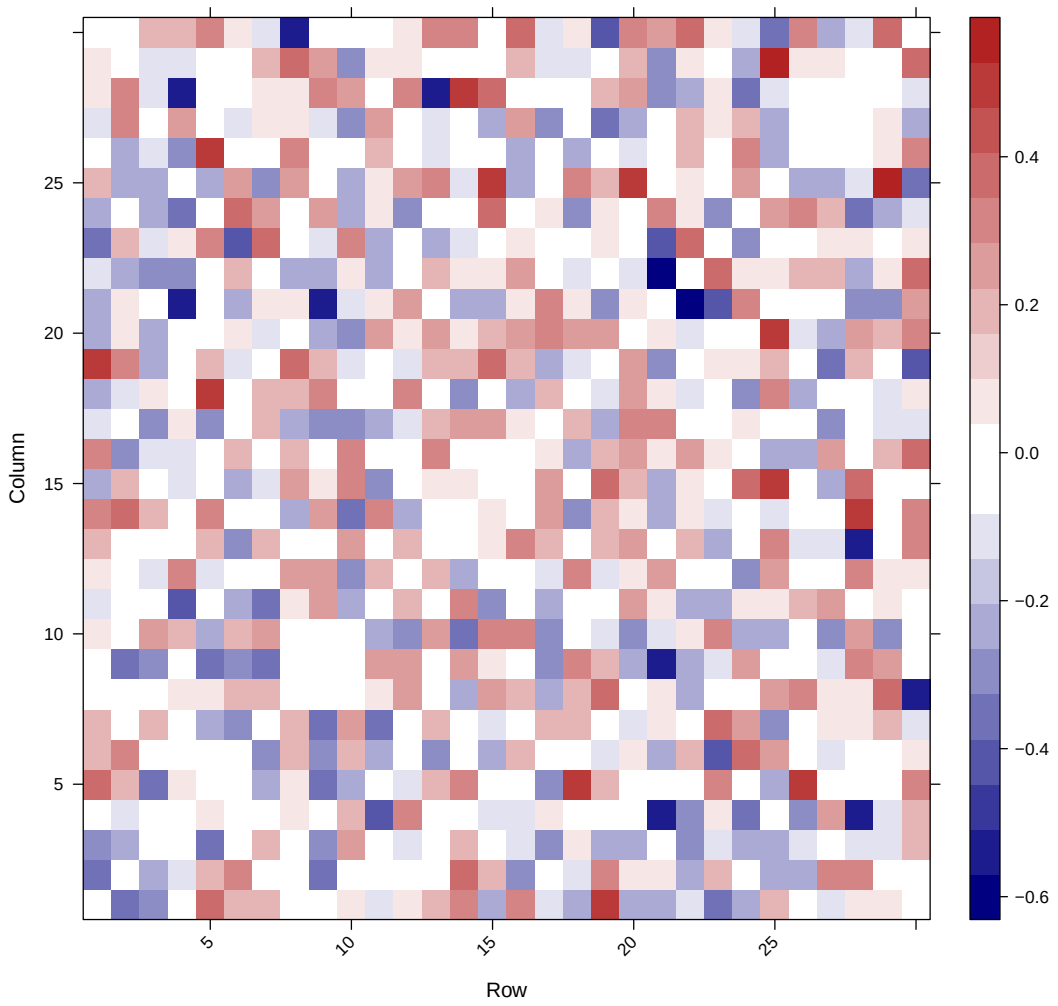

Figure 4.6: Heat map of the difference between estimated correlation matrix and the true value for in-sample analysis, i.e. $(\hat{\mathbf{\Psi}} - \mathbf{\Psi})$ where $\hat{\mathbf{\Psi}}$ is the posterior mean of correlation matrix. On the x-axis is the row index, on the y-axis is the column index. Positive differences are colored in red, negative differences are colored in blue with lighter color represents smaller difference.

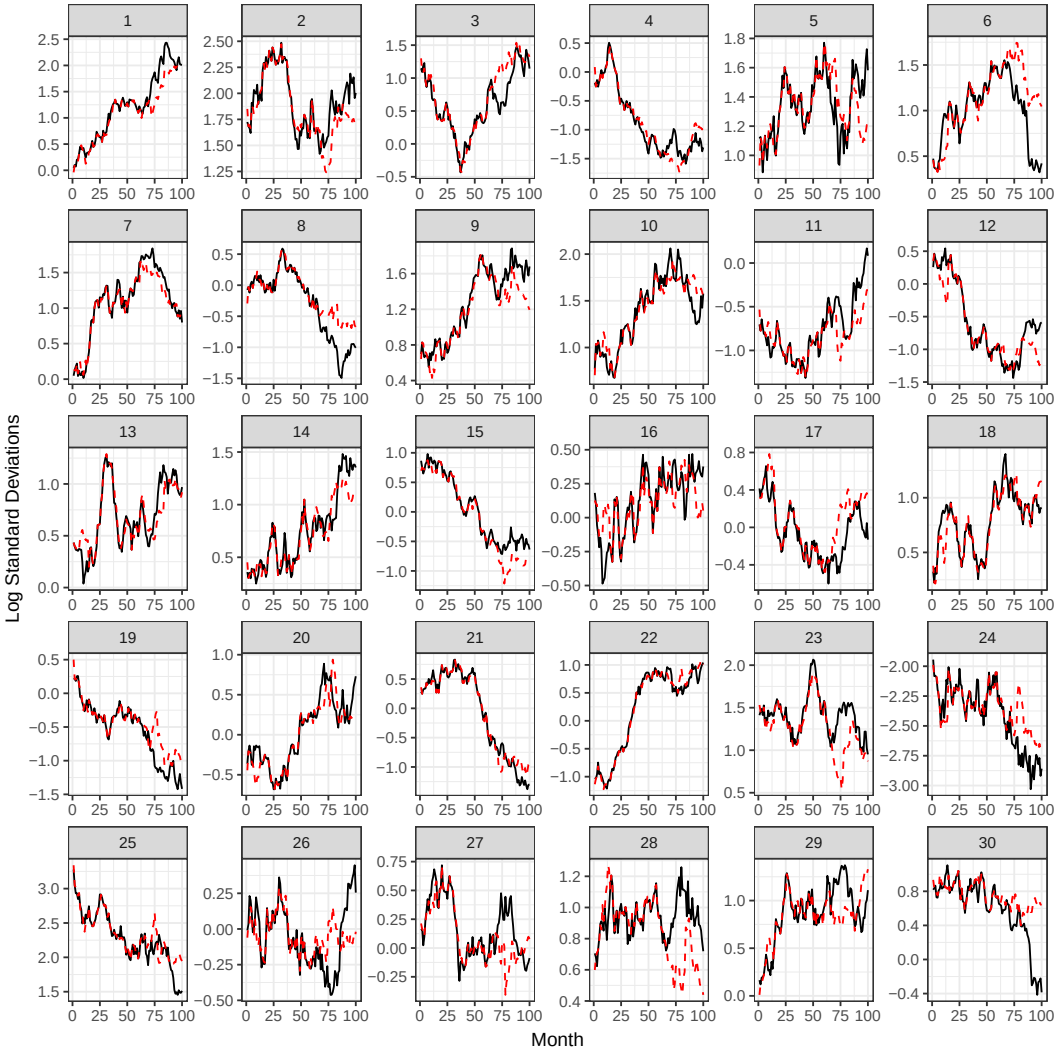

Figure 4.7: Estimated stochastic volatility ($\hat{\omega}_{it}$) for each individual series of in-sample analysis. In each subplot, the x-axis is the month of the in-sample period, on the y-axis is the value of $\hat{\omega_{it}}$. Black line represents the true value with red dashed line represents estimate. Index of the monthly series in on top of each subplot.

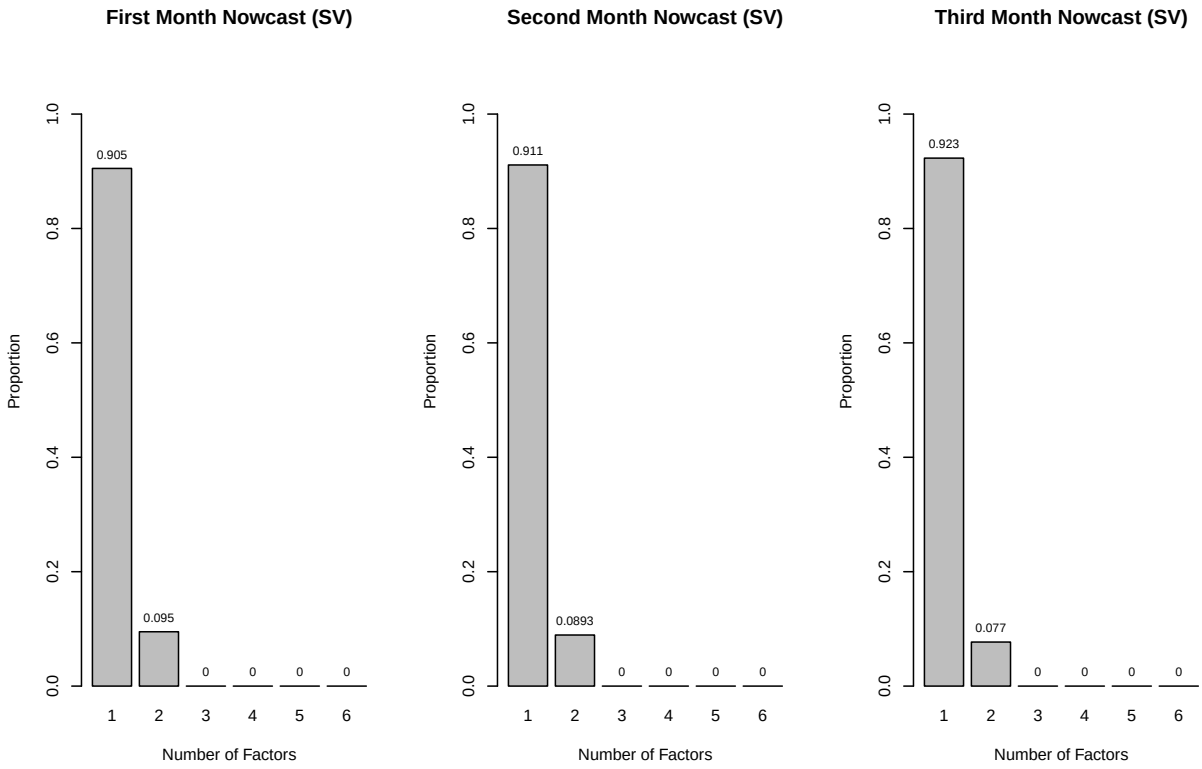

Figure 4.8: Distribution of estimated number of latent factors by month for the US market data.The left, middle, and right columns are nowcasting in the first, second, and third month of the quarter respectively.

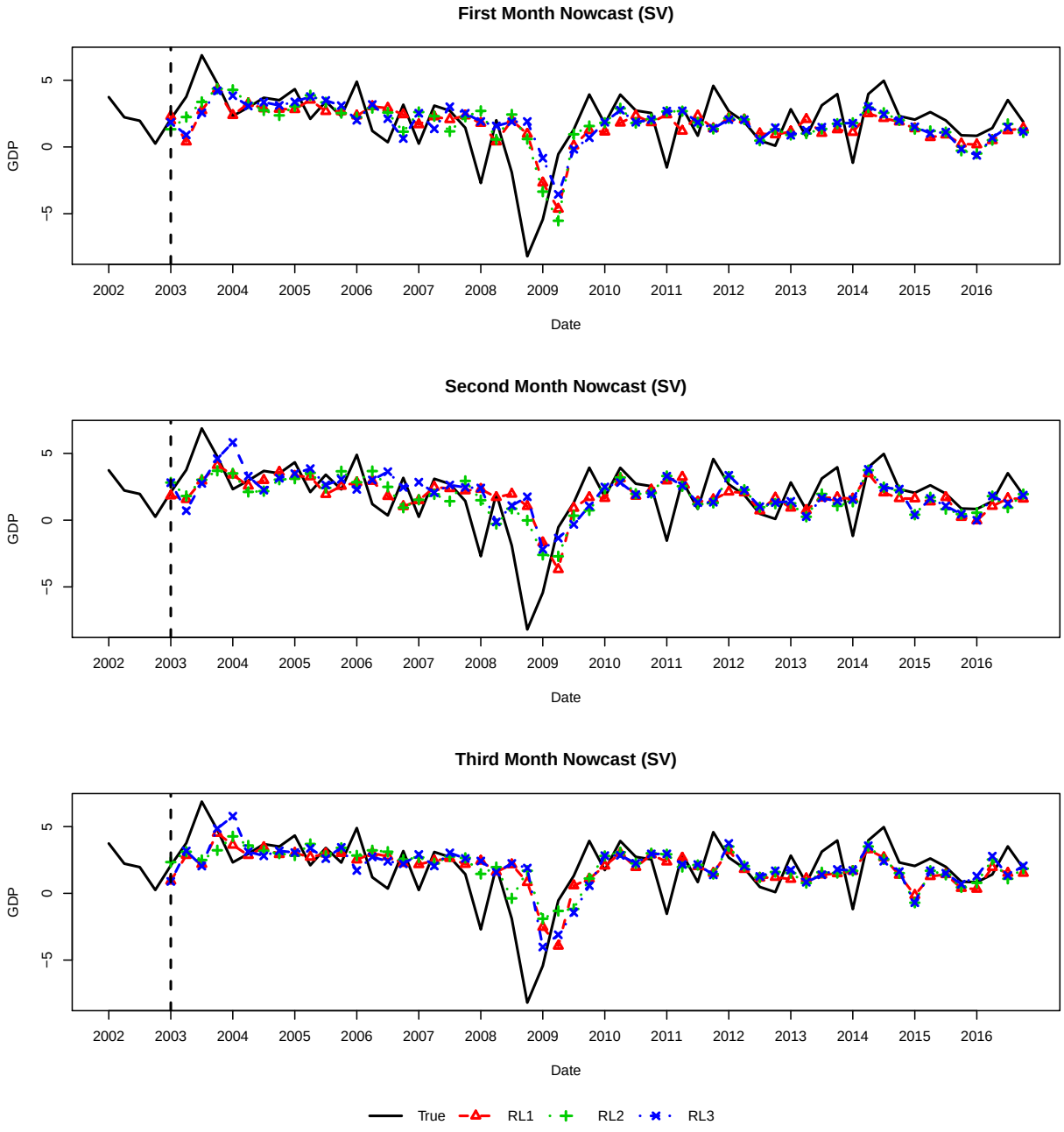

Figure 4.9: Nowcasting over 2003Q1 to 2016Q4 by 3 releases in each month for the US market data. Black solid line represents the true GDP value, dashed lines with different knot types represent the GDP nowcasts with red, green, and blue as release 1, 2, and 3.

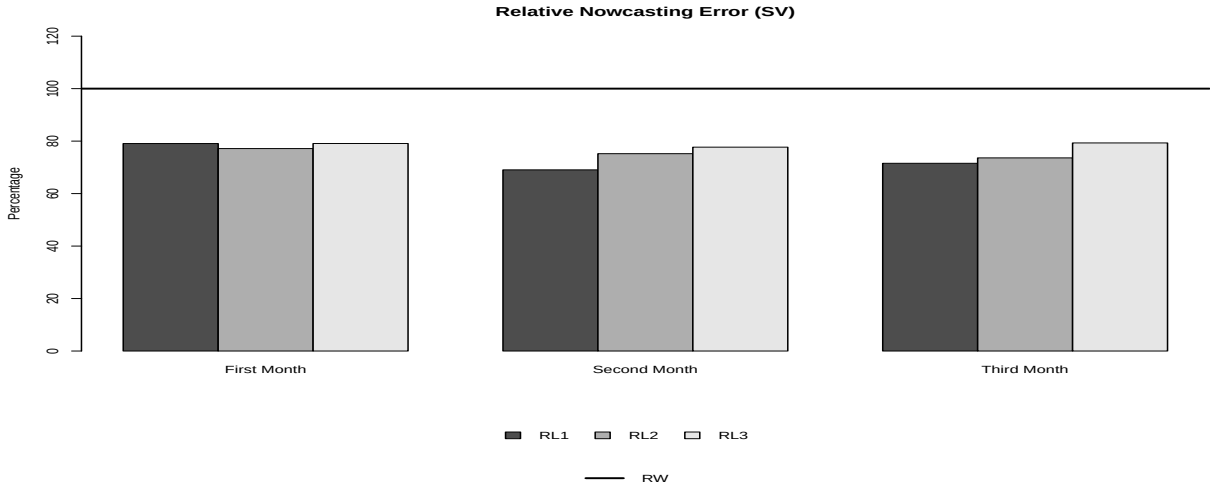

Figure 4.10: Averaged mean absolute nowcasting error ratios (relative to RW). The horizontal line is 100% representing the baseline for RW. The first, second, and third release are colored as dark gray, gray, and light gray for each month.

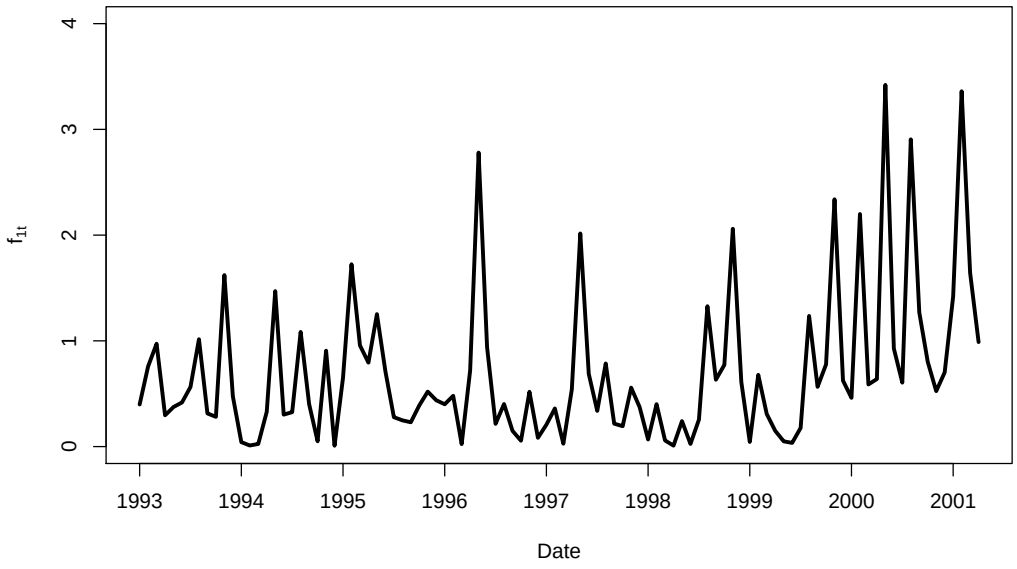

Figure 4.11: Absolute values of estimated first latent factors of in-sample analysis in the US market data.

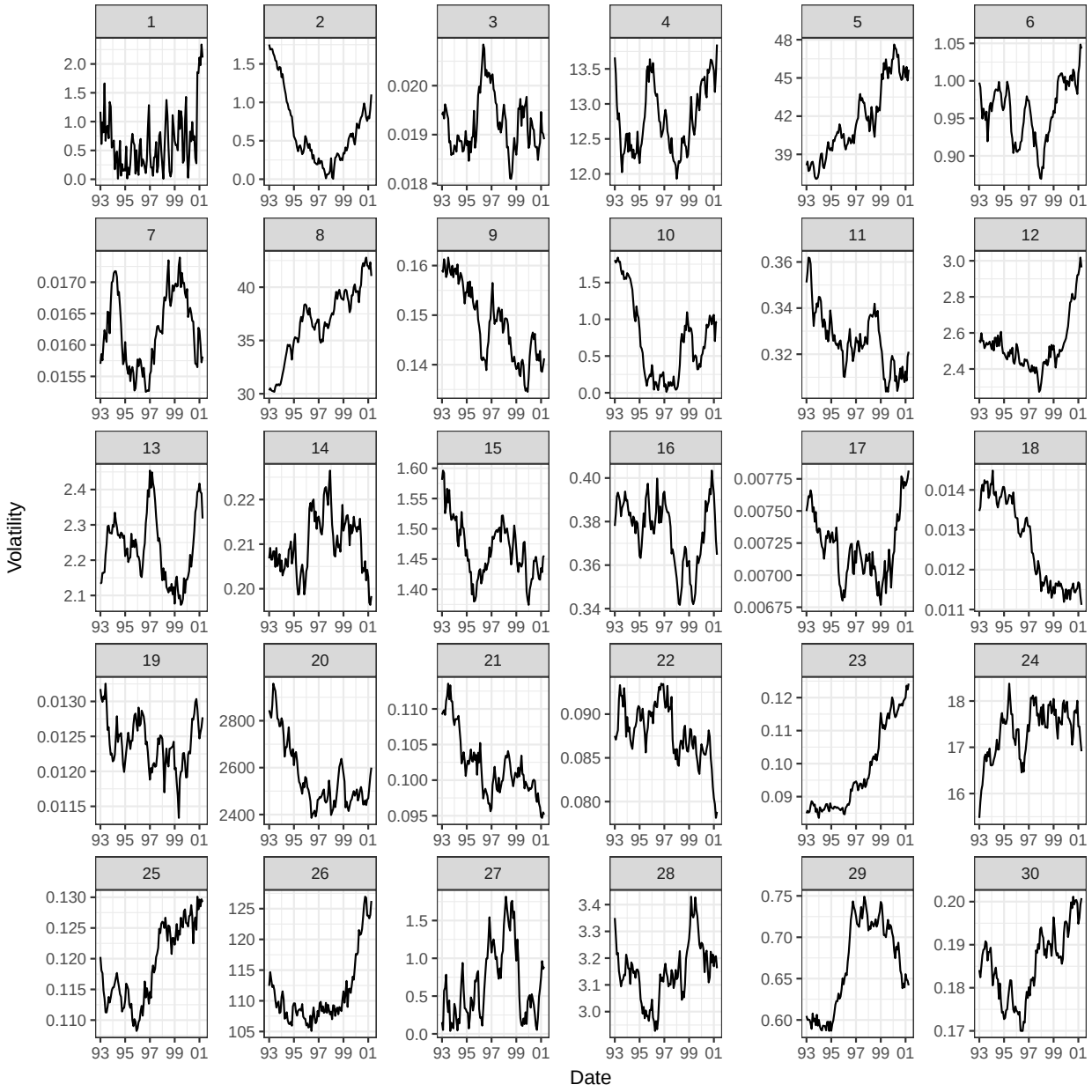

Figure 4.12: Estimated stochastic volatility ($\hat{\mathbf{\Omega}}_{t[i,i]} = exp(2\hat{\omega}_{it})$) for each individual series of in-sample analysis in the US market data. Index of the monthly series in on top of each subplot.

CHAPTER 5. SUMMARY AND DISCUSSION

In this dissertation, we develop a Bayesian approach to provide a way to deal with the unbalanced feature of the data set in a real-time nowcasting framework. In Chapter 2, we consider the number of latent factors is fixed and known and assume a relatively simple release pattern. We evaluate our BAY method based on its estimation accuracy in latent factors and nowcasting performance for the quarterly GDP and compare the nowcasting performance to the GRS approach. The simulation study suggests that in terms of estimation accuracy in factors, both of the BAY and GRS approaches can produce accurate estimated factors. In terms of nowcasting performance, the two methods are comparable with BAY being slightly better in sense of resulting in smaller nowcasting errors. Our simulation results suggest that the BAY method has the potential to estimate the dynamic factor models well and produce reliable nowcasting results. We exam the viability in empirical analysis using Chinese market data and demonstrates the empirical relevance of the BAY approach in nowcasting China's GDP for the time period considered. These results, however, don't suggest that we should completely abandon the GRS approach. In fact, (as we mentioned) for the last 3 releases in this dataset, GRS performs better than BAY in the 1st and 3rd nowcasting months.

In Chapter 3, we consider the number of factors is unknown and modify our BAY approach such that it can estimate the number of latent factors and other model parameters in a single statistical framework. In Chapter 4, we further extend our model to allow stochastic volatilities within monthly series. We evaluate our approach in a pseudo-real-time nowcasting set-up by mimic a realistic release pattern in simulation studies and analyze the US's market data. In the simulation studies considered in Chapter 3 and 4, we evaluate our BAY method based on its estimation accuracy of the number of latent factors and factors themselves, and nowcasting performance for the quarterly GDP, with constant volatility or SV. In terms of estimation accuracy of the number of

factors, the BAY approach can estimate the true number of factors. In terms of estimation accuracy in factors, the BAY approach can produce accurate estimated factors for both constant volatility and SV. In terms of nowcasting performance, the BAY approach produces better nowcasts in the sense of resulting in smaller nowcasting errors comparing to RW. Our simulation results suggest that the BAY method can estimate the number of latent factors, the dynamic factor models well and produce reliable nowcasting results. The empirical analyses demonstrate the empirical relevance of the BAY approach in nowcasting the US's GDP for the time period considered. The results suggest one factor is sufficient to summarise the dynamics within the US market. The slight improvement in MANEs and obvious dynamics in estimated volatility suggest SV is better to be considered into the DFM when nowcasting US's GDP.

Real-time nowcasting plays a crucial role in terms of policy-making and long term forecasting. In this dissertation, we develop a Bayesian MCMC-based method for inferences of DFMs for real time nowcasting. We demonstrate, through simulation studies and empirical analysis, the proposed Bayesian framework can (i) identify the number of factors correctly, (ii) estimate latent dynamic factors and stochastic volatility accurately, (iii) provide reliable real-time nowcasting results. The MCMC techniques developed here are generic that they can be applied to nowcast a wide range of economic indicators of interest, such as inflation rates, trade exports (or imports), and exchange rates, etc.

BIBLIOGRAPHY

[1] Aarons, G., Caratelli, D., Cocci, M., Giannone, D., Sbordone, A. M., and Tambalotti, A., 2016. Just released: introducing the New York Fed Staff Nowcast. *Liberty Street Economics*, Apr.12.

[2] Ahn, S., and Horenstein, A., 2013. Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203-1227.

[3] Andrieu, C. , Doucet, A. , and Holenstein, R. , 2010. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Statistical Methodology*, Series B, 72(3):269-342.

[4] Baffigi, A., Golinelli, R., and Parigi, G., 2004. Bridge models to forecast the euro area GDP. *International Journal of Forecasting*, 20(3), 447-460.

[5] Bai, J., 2003. Inferential Theory for Factor Models of Large Dimensions. *Econometrica*, 71:135-172.

[6] Bai, J., and Ng, S., 2002. Determining the Number of Factors in Approximate Factor Models. *Econometrica*, 70:191-221.

[7] Benati, L, and Surico, P., 2009. VAR Analysis and the Great Moderation. *American Economic Review*, 99(4): 1636-52.

[8] Bai, J., and Ng, S., 2006. Evaluating latent and observed factors in macroeconomics and finance. *Journal of Econometrics*, 131:507-537.

[9] Bates, B. J., Plagborg-Møller M., Stock, J. H., and Watson, M. W., 2013. Consistent factor estimation in dynamic factor models with structural instability. *Journal of Econometrics*, 117(2):289-304.

[10] Benati, L., Surico, P., 2008. Evolving u.s. monetary policy and the decline of inflation predictability. *Journal of the European Economic Association*, 6, 634646.

[11] Boivin, J., and Ng, S., 2006. Are more data always better for factor analysis? *Journal of Econometrics*, 132:169-194.

[12] Bok, B., Caratelli, D., Giannone, D., Sbordone, A. M., and Tambalotti, A., 2018. Macroeconomic Nowcasting and Forecasting with Big Data. *Annual Review of Economics*, 10:615-643.

[13] Carriero, A., Clark, T. E., and Marcellino, M., 2016. Common drifting volatility in large bayesian VARs. *Journal of Business & Economic Statistics*, 34(3):375-390.

[14] Chernis, T., and Sekkel, R., 2017. A dynamic factor model for nowcasting Canadian GDP growth. *Empirical Economics*, 53(4):21734.

[15] Clark, T. E., 2011. Real-time density forecasts from bayesian vector autoregressions with stochastic volatility. *Journal of Business & Economic Statistics*, 29:327-341.

[16] Clark, T. E., and Ravazzolo, F., 2014. Macroeconomic forecasting performance under alternative specifications of time-varying volatility. *Journal of Applied Econometrics*, 30(4): 551-575.

[17] Connor, G., and Korajczyk, R. A. , 1986. Performance Measurement with the Arbitrage Pricing Theory. *Journal of Financial Economics*, 15: 373-394.

[18] Cogley, T., and Sargent, T. J., 2005. Drift and volatilities: Monetary policies and outcomes in the post wwii u.s.. *Review of Economic Dynamics*, 8:262-302.

[19] Croushore, D., Stark, T., 2001. A real-time data set for macroeconomists. *Journal of Econometrics*, 105, 111-130.

[20] D'Agostino, A., Giannone, D., 2006. Comparing Alternative Predictors Based on Large-Panel Factor Models. *Working Paper Series*, 680, European Central Bank, Frankfurt, Germany.

[21] Follett, L., and Yu, C., 2019. Achieving parsimony in Bayesian vector autoregressions with the horseshoe prior. *Econometrics and Statistics*, 11:130-144.

[22] Forni, M., Hallin, M., Lippi, M., and Reichlin, L., 2005. The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting. *Journal of the American Statistical Association*, 100, 830-840.

[23] Giannone, D., Reichlin, L., and Small, D., 2008. Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55:665-676.

[24] Jacquier, E., Polson, N. G., and Rossi, P. E., 2002. Bayesian Analysis of Stochastic Volatility Models. *Journal of Business & Economic Statistics*, 20(1): 69-87.

[25] Karlsson, S., 2013. Forecasting with bayesian vector autoregressions. *Handbook of Economic Forecasting*, Vol 2, Part B, 791-897.

[26] Kim, S., Shephard, N., and Chib, S. 1998. Stochastic volatility: likelihood inference and comparison with arch models. *The Review of Economic Studies*, 65(3):361-393.

[27] Kitchen, J., Monaco, R., 2003. Real-Time Forecasting in Practice: The U.S. Treasury Staff's Real-Time GDP Forecast System. *Business Economics*, 38(4), 10-19.

[28] Koenig, E. F., Dolmas, S., and Piger, J., 2003. The Use and Abuse of Real-Time Data in Economic Forecasting. *Review of Economics and Statistics*, 85, 618-628.

[29] Lewandowski, D, Kurowicka, D., and Joe, H., 2009. Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9):1989-2001 .

[30] Lindsten, F., and Schön, T. B., 2012. On the use of backward simulation in the particle Gibbs sampler. *2012 IEEE International Conference on Acoustics*, Speech and Signal Processing (ICASSP), Kyoto, Japan.

[31] Negro, M. D., and Otrok, C., 2008. Dynamic factor models with time-varying parameters: measuring changes in international business cycles. *Staff Reports*, 326, Federal Reserve Bank of New York.

[32] Onatski, A., 2009. A formal statistical test for the number of factors in the approximate factor models. *Econometrica*, 77(5):1447-1479.

[33] Onatski, A., 2010. Determining the number of factors from the empirical distribution of eigenvalues. *The Review of Economics and Statistics*, 92:1004-1016.

[34] Orphanides, A., 2002. Monetary policy rules and the Great Inflation. *Journal of the American Economic Association*, 92, 115-120.

[35] Primiceri, G., 2005. Time varying structural vector autoregressions and monetary policy. *The Review of Economic Studies*, 72:821-852.

[36] Rünstler, G., Sédillot, F. (2003). Short-term Estimates of Euro Area Real GDP by Means of Monthly Data. *Working Paper Series*, 276, European Central Bank, Frankfurt, Germany.

[37] Stock, J. H., and Watson, M. W., 2002. Forecasting Using Principal Components From a Large Number of Predictors. *Journal of the American Statistical Association*, 97:1167-1179.

[38] Whiteley, N., 2010. Discussion on Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society*, Series B, 72:306-307.

[39] Yiu, M. S., and Chow, K. K. (2011). Nowcasting Chinese GDP: Information Content of Economic and Financial Data. *China Economic Journal*, 3(3):223-240.

[40] Zhang, L., Sarkar, A., and Mallick, B., 2013. Bayesian low rank and sparse covariance matrix decomposition. *Manuscript*, Texas A&M University.

APPENDIX A. CHAPTER 2 APPENDIX

A.1 Identification Assumptions

In this appendix, we provide Assumption F and Assumption M adapted from Stock and Watson (2002) to deal with identification issue. The two sets of assumptions are as follows.

Assumption F on factors and factor loadings, are

- F1. $\dfrac{\mathbf{\Theta}'\mathbf{\Theta}}{n} \to \mathbf{I}_{r\times r}$,
- F2. $E(\mathbf{F}_t'\mathbf{F}_t) = \mathbf{\Sigma}_{ff}$ where $\mathbf{\Sigma}_{ff}$ is a diagonal matrix with diagonal elements $\sigma_{ii}^2 > \sigma_{jj}^2 \; \forall \; i < j$.

Assumption F1 inhibits $\mathbf{\Theta}$ to be orthonormal. In additional with the assumption F2, the factor loadings are identifiable up to a change of sign.

Assumption M on moments of errors $\boldsymbol{\epsilon}_t = (e_{1t}, ..., e_{nt})'$ are

- M1. $E(e_{it}e_{jt}) = \tau_{ij}$, $\lim\limits_{n\to\infty} n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} |\tau_{ij}| < \infty$.
- M2. $\lim\limits_{n\to\infty} n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} |cov(e_{it}e_{is}, e_{jt}e_{js})| < \infty$.

Assumption M1 allows $\boldsymbol{\epsilon}_t$ to be weakly correlated across series. Normality is not assumed, but Assumption M2 limits the size of fourth moments.

A.2 Posterior Distributions

In this appendix, we present the full conditional posterior distributions and a MCMC algorithm to draw posterior samples. Most parameters have closed form full conditional distribution, allowing for a straightforward Gibbs sampling method. However, $\mathbf{\Omega}$ does not have closed form posterior distribution. In this case, we insert an independent Metropolis-Hastings (MH) sampler within Gibbs sampler to generate posterior samples for $\mathbf{\Omega}$.

A.2.1 Sampling the mean of monthly series $\boldsymbol{\mu}$

The full conditional of mean of monthly series, $\boldsymbol{\mu}$, is a multivariate normal distribution:

$$\boldsymbol{\mu}|\mathbf{X}_{(q,T)}, \mathbf{F}, \boldsymbol{\Theta}, \boldsymbol{\Omega} \sim N(\mathbf{U}_{\boldsymbol{\mu}}, \mathbf{W}_{\boldsymbol{\mu}}^{-1}), \tag{A.1}$$

where

$$\begin{aligned}
\mathbf{W}_{\boldsymbol{\mu}} &= (T-1)\boldsymbol{\Omega}^{-1} + \mathbf{1}'_{v_{q,T}}(\mathbf{1}_{v_{q,T}}\boldsymbol{\Omega}\mathbf{1}'_{v_{q,T}})^{-1}\mathbf{1}_{v_{q,T}} + \mathbf{I}_{n\times n}, \\
\mathbf{U}_{\boldsymbol{\mu}} &= \mathbf{W}_{\boldsymbol{\mu}}^{-1}\Big[\sum_{t=1}^{T-1}\boldsymbol{\Omega}^{-1}(\mathbf{x}_t - \boldsymbol{\Theta}\mathbf{F}_t) + (\mathbf{1}'_{v_{q,T}}(\mathbf{1}_{v_{q,T}}\boldsymbol{\Omega}\mathbf{1}'_{v_{q,T}})^{-1}\mathbf{1}_{v_{q,T}})(\mathbf{x}_T - \boldsymbol{\Theta}\mathbf{F}_T)\Big].
\end{aligned}$$

A.2.2 Sampling the factor loadings matrix $\boldsymbol{\theta}$

The full conditional of factor loadings, $\boldsymbol{\theta}$, is a multivariate normal distribution:

$$\boldsymbol{\theta}|\mathbf{X}_{(q,T)}, \mathbf{F}, \boldsymbol{\Omega} \sim N(\mathbf{U}_{\boldsymbol{\theta}}, \mathbf{W}_{\boldsymbol{\theta}}^{-1}), \tag{A.2}$$

where $\mathbf{C}_t = \mathbf{I}_{n\times n} \otimes \mathbf{F}'_t$, and

$$\begin{aligned}
\mathbf{W}_{\boldsymbol{\theta}} &= \sum_{t=1}^{T-1}\mathbf{C}'_t\boldsymbol{\Omega}^{-1}\mathbf{C}_t + \mathbf{I}_{nr\times nr} + \mathbf{C}'_T\mathbf{1}'_{v_{q,t}}(\mathbf{1}_{v_{q,T}}\boldsymbol{\Omega}\mathbf{1}'_{v_{q,T}})^{-1}\mathbf{1}_{v_{q,T}}\mathbf{C}_T, \\
\mathbf{U}_{\boldsymbol{\theta}} &= \mathbf{W}_{\boldsymbol{\theta}}^{-1}\Big[\sum_{t=1}^{T-1}\mathbf{C}'_t\boldsymbol{\Omega}^{-1}(\mathbf{x}_t - \boldsymbol{\mu}) + \mathbf{C}'_T\mathbf{1}'_{v_{q,T}}(\mathbf{1}_{v_{q,T}}\boldsymbol{\Omega}\mathbf{1}'_{v_{q,T}})^{-1}\mathbf{1}_{v_{q,T}}(\mathbf{x}_T - \boldsymbol{\mu})\Big].
\end{aligned}$$

A.2.3 Sampling the covariance in monthly series $\boldsymbol{\Omega}$

The posterior for $\boldsymbol{\Omega}$ is

$$\begin{aligned}
P(\boldsymbol{\Omega}|\cdot) &\propto \prod_{t=1}^{T-1} P(\mathbf{x}_t|\mathbf{F}_t, \boldsymbol{\theta}, \boldsymbol{\Omega}) \times P(\mathbf{x}_{i\in v_{q,T}}|\mathbf{F}_T, \boldsymbol{\theta}, \boldsymbol{\Omega}) \times P(\boldsymbol{\Omega}) \\
&\propto |\boldsymbol{\Omega}|^{-(T-1)/2} e^{-\frac{1}{2}tr\left[\sum_{t=1}^{T-1}(\mathbf{x}_t - \boldsymbol{\mu} - \mathbf{C}_t\boldsymbol{\theta})(\mathbf{x}_t - \boldsymbol{\mu} - \mathbf{C}_t\boldsymbol{\theta})'\boldsymbol{\Omega}^{-1}\right]} \\
&\times |\boldsymbol{\Omega}|^{-(\nu_\theta + n + 1)/2} e^{-\frac{1}{2}tr\left[\frac{1}{n}\boldsymbol{\Omega}^{-1}\mathbf{I}_{n\times n}\right]} \\
&\times |\mathbf{1}_{v_{q,T}}\boldsymbol{\Omega}\mathbf{1}'_{v_{q,T}}|^{-1/2} e^{-\frac{1}{2}tr\left[\mathbf{1}_{v_{q,T}}(\mathbf{x}_T - \boldsymbol{\mu} - \mathbf{C}_T\boldsymbol{\theta})\mathbf{1}'_{v_{q,T}}(\mathbf{x}_T - \boldsymbol{\mu} - \mathbf{C}_T\boldsymbol{\theta})'(\mathbf{1}_{v_{q,T}}\boldsymbol{\Omega}\mathbf{1}'_{v_{q,T}})^{-1}\right]},
\end{aligned}$$

which is not in closed form. Note that the first two pieces to yields $\mathbf{\Omega}$ has an inverse Wishart distribution $iW(\Psi_{\mathbf{\Omega}}, \nu_{\mathbf{\Omega}})$, where $\Psi_{\mathbf{\Omega}} = \sum_{t=1}^{T-1}(\mathbf{x}_t - \boldsymbol{\mu} - \mathbf{C}_t\boldsymbol{\theta})(\mathbf{x}_t - \boldsymbol{\mu} - \mathbf{C}_t\boldsymbol{\theta})' + \frac{1}{n}\mathbf{I}_{n\times n}$, and $\nu_{\mathbf{\Omega}} = T - 1 + \nu_{\boldsymbol{\theta}}$. Therefore, we can propose $\mathbf{\Omega}^*$ using $iW(\Psi_{\mathbf{\Omega}}, \nu_{\mathbf{\Omega}})$, and use the last piece in the full conditional, denote as $q(\mathbf{\Omega})$, to construct the acceptance-rejection rate. That is the proposal $\mathbf{\Omega}^*$ is accepted with probablity

$$A(\mathbf{\Omega}^*|\mathbf{\Omega}^c) = min\Big\{1, \frac{q(\mathbf{\Omega}^*)}{q(\mathbf{\Omega}^c)}\Big\},$$

where $\mathbf{\Omega}^c$ denotes the current state of $\mathbf{\Omega}$.

A.2.4 Sampling the AR(1) coefficients a_j

For $j = 1, ..., r$, the full conditional of each AR(1) coefficient, a_j, is a truncated normal distribution:

$$P(a_j|f_{j,1}, ..., f_{j,T}, \sigma_j^2) = \begin{cases} 0 & \text{if } a_j \leq -1 \\ \dfrac{\phi(\dfrac{a_j - \mu_{a_j}}{\sigma_{a_j}})}{\Phi(\dfrac{1-\mu_{a_j}}{\sigma_{a_j}}) - \Phi(\dfrac{-1-\mu_{a_j}}{\sigma_{a_j}})} & \text{if } -1 < a_j < 1 \\ 0 & \text{if } a_j \geq 1 \end{cases} \tag{A.3}$$

where $\mu_{a_j} = \dfrac{\sum_{t=2}^{T} f_{j,t}f_{j,t-1}}{\sigma_j^2 + \sum_{t=2}^{T} f_{j,t-1}^2}$ and $\sigma_{a_j}^2 = \dfrac{\sigma_j^2}{\sigma_j^2 + \sum_{t=2}^{T} f_{j,t-1}^2}$.

A.2.5 Sampling the variance in factor equations σ_j^2

For $j = 1, ..., r$, the full conditional of σ_j^2 is an inverse Gamma distribution:

$$\sigma_j^2|f_{j,1}, ..., f_{j,T}, a_j \sim iG(\alpha_j, \beta_j), \tag{A.4}$$

where $\alpha_j = \alpha_s + (T-1)/2$ and $\beta_j = \beta_s + \sum_{t=2}^{T}(f_{j,t} - a_j f_{j,t-1})^2/2$.

A.2.6 Sampling the coefficients associated with factors $\boldsymbol{\beta}$

The full conditional of coefficients associated with factors in GDP equation, $\boldsymbol{\beta}$, is a multivariate normal distribution:

$$\boldsymbol{\beta}|\mathbf{F}, \mathbf{Y}, \eta^2 \sim N(\mathbf{U}_{\boldsymbol{\beta}}, \mathbf{W}_{\boldsymbol{\beta}}^{-1}), \tag{A.5}$$

where $\tilde{\mathbf{F}}_{3k} = [1, \mathbf{F}'_{3k}, \mathbf{F}'_{3k-1}, \mathbf{F}'_{3k-2}, y_{k-1}]'$, and

$$\begin{aligned} \mathbf{W}_{\boldsymbol{\beta}} &= \sum_{k=2}^{K} \tilde{\mathbf{F}}_{3k}\tilde{\mathbf{F}}'_{3k}/\eta^2 + \mathbf{I}_{(3r+2)\times(3r+2)}, \\ \mathbf{U}_{\boldsymbol{\beta}} &= \mathbf{W}_{\boldsymbol{\beta}}^{-1} \sum_{k=2}^{K} \tilde{\mathbf{F}}_{3k} y_k/\eta^2. \end{aligned}$$

A.2.7 Sampling the variance in GDP equation η^2

The full conditional of the variance in GDP equation, η^2, is a inverse Gamma distribution:

$$\eta^2|\mathbf{F}, \mathbf{Y}, \boldsymbol{\beta} \sim iG(\alpha_\eta, \beta_\eta), \tag{A.6}$$

where $\alpha_\eta = \alpha_h + (K-1)/2$ and $\beta_\eta = \beta_h + \sum_{k=2}^{K}(y_k - \beta_0 - \boldsymbol{\beta}'_1\mathbf{F}_{3k} - \boldsymbol{\beta}'_2\mathbf{F}_{3k-1} - \boldsymbol{\beta}'_3\mathbf{F}_{3k-2} - \beta_4 y_{k-1})^2/2$.

A.2.8 Sampling the latent factors $\mathbf{F}_t$

This section provides a way to understand the derivation of posterior distribution of $\mathbf{F}_t$ from the point of GLS's view.

The posterior distribution for $\mathbf{F}_t$ will have different form depending on t. The most general cases are, for $3 < t < 3K$, $k = 2, ..., K$. t can be written as $t = 3k - i$, for some k and $i = 0, 1, 2$, depending on which month of quarter k we are in. If $i = 0$, we are in the third month of quarter k, if $i = 1$, we are in the second month of quarter k, and, if $i = 2$, we are in the first month of quarter k. Then, at $t = 3k - i$, $\mathbf{F}_t$ enters the joint likelihood through four parts, $\mathbf{x}_t$, $\mathbf{F}_{t+1}$, $\mathbf{F}_{t-1}$, and y_k. Therefore, we can write

$$\begin{pmatrix} \mathbf{x}_t - \boldsymbol{\mu} \\ \mathbf{F}_{t+1} \\ \mathbf{F}_{t-1} \\ f_{y_k}(i) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Theta} \\ \mathbf{A} \\ \mathbf{A}^{-1} \\ \boldsymbol{\beta}'_{i+1} \end{pmatrix} \mathbf{F}_t + \begin{pmatrix} \epsilon_t \\ u_{t+1} \\ -\mathbf{A}^{-1}u_t \\ \nu_k \end{pmatrix} \equiv \tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\mathbf{F}_t + \tilde{\boldsymbol{\epsilon}},$$

where $f_{y_k}(i)$ is a function of i depending on y_k, defined as

$$f_{y_k}(i) = \begin{cases} y_k - \beta_0 - \boldsymbol{\beta}'_2\mathbf{F}_{t-1} - \boldsymbol{\beta}'_3\mathbf{F}_{t-2} - \beta_4 y_{k-1} & \text{if } i = 0 \\ y_k - \beta_0 - \boldsymbol{\beta}'_1\mathbf{F}_{t+1} - \boldsymbol{\beta}'_3\mathbf{F}_{t-1} - \beta_4 y_{k-1} & \text{if } i = 1 \\ y_k - \beta_0 - \boldsymbol{\beta}'_1\mathbf{F}_{t+2} - \boldsymbol{\beta}'_2\mathbf{F}_{t+1} - \beta_4 y_{k-1} & \text{if } i = 2 \end{cases} \tag{A.7}$$

Therefore, we have

$$\tilde{\mathbf{Y}} = \begin{bmatrix} \mathbf{x}_t - \boldsymbol{\mu} \\ \mathbf{F}_{t+1} \\ \mathbf{F}_{t-1} \\ f_{y_k}(i) \end{bmatrix}, \quad \tilde{\mathbf{X}} = \begin{bmatrix} \boldsymbol{\Theta} \\ \mathbf{A} \\ \mathbf{A}^{-1} \\ \boldsymbol{\beta}'_{i+1} \end{bmatrix}, \quad \tilde{\boldsymbol{\epsilon}} = \begin{bmatrix} \epsilon_t \\ u_{t+1} \\ -\mathbf{A}^{-1}u_t \\ \nu_k \end{bmatrix},$$

and

$$var(\tilde{\boldsymbol{\epsilon}}) = \tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \boldsymbol{\Omega} & 0 & 0 & 0 \\ 0 & \boldsymbol{\Sigma} & 0 & 0 \\ 0 & 0 & (\mathbf{A}'\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1} & 0 \\ 0 & 0 & 0 & \eta^2 \end{bmatrix}.$$

By weighted regression, $3 < t < 3K$, $k = 2, ..., K$, $t = 3k - i$, draw

$$\mathbf{F}_t|\tilde{\mathbf{Y}}, \tilde{\mathbf{X}}, \tilde{\boldsymbol{\Sigma}} \sim N((\tilde{\mathbf{X}}'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{Y}}, (\tilde{\mathbf{X}}'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{X}})^{-1}). \tag{A.8}$$

For other t, the posterior distribution for $\mathbf{F}_t$ is of the same form, with some deletion done to $\tilde{\mathbf{Y}}$, $\tilde{\mathbf{X}}$, and $\tilde{\boldsymbol{\Sigma}}$. For example, if $t = 1$, since we don't have $\mathbf{F}_0$ and y_0, delete entries corresponding to $\mathbf{F}_{t-1}$ and $f_{y_k}(i)$ in $\tilde{\mathbf{Y}}$, $\tilde{\mathbf{X}}$, and $\tilde{\boldsymbol{\Sigma}}$. If $t = 2, 3$, since we don't have y_0, delete entry corresponding to $f_{y_k}(i)$ in $\tilde{\mathbf{Y}}$, $\tilde{\mathbf{X}}$, and $\tilde{\boldsymbol{\Sigma}}$. If $3K < t < T$, since we don't have y_{K+1}, delete entry corresponding to $f_{y_K}(i)$ in $\tilde{\mathbf{Y}}$, $\tilde{\mathbf{X}}$, and $\tilde{\boldsymbol{\Sigma}}$. If $t = T$, since we don't have $\mathbf{F}_{T+1}$ and y_{K+1}, first delete entries

corresponding to $\mathbf{F}_{T+1}$and $f_{y_{K+1}}(i)$ in $\tilde{\mathbf{Y}}$, $\tilde{\mathbf{X}}$, and $\tilde{\mathbf{\Sigma}}$. Secondly, assume we are at (q, T), change entry corresponding to $\mathbf{x}_{i \in v_{q,T}}$ in $\tilde{\mathbf{Y}}$, $\tilde{\mathbf{X}}$, and $\tilde{\mathbf{\Sigma}}$.

APPENDIX B. CHAPTER 2 SUPPLEMENTAL MATERIAL

In this supplemental document, we provide some additional information that is not covered in Chapter 2 due to space limit. In Section B.1, we report the investigation for burn-in period for simulation study. In Section B.2, we compare the parameter estimation from release to release for simulation study. In Section B.3, we provide further inference for empirical study. In Section B.4, we provide the details for monthly data series used in empirical study.

B.1 Burn-in Period For Simulation Study

In this section, we report the investigation for burn-in period that is needed for MCMC chain to converge in simulation study. We save the posterior samples that are used in nowcasting calculations, which are $\mathbf{A}$, $\boldsymbol{\beta}$, and $\mathbf{F}_t$ from the first iteration to the $5000th$ iteration for Simulation 1. In Figure B.1, posterior draws for a_j are against iteration. In Figure B.2, posterior draws for $\beta_0, \boldsymbol{\beta}_4$ against iteration. In Figure B.3, posterior draws for $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3$ against iteration. In Figure B.4, posterior draws for F_t for the $1^{st}, 50^{th}, 500^{th}, 5000^{th}$ are plotted overlaid on each other.

Figure B.1, B.2, and B.3 suggest that from the very first iteration, the posterior draws for $\mathbf{A}$ and $\boldsymbol{\beta}$, stay close to the true value and the variation in posterior draws stay stable. Figure B.4 also suggests there is negligible difference between posterior draws for $\mathbf{F}_t$ from the first iteration and later iterations. Therefore, we conclude 5,000 iterations is enough for burn-in period.

B.2 Inference for Simulation Study

In this section, we compare the parameter estimation from release to release, in order to explain the MANE behaviors for GRS and BAY approach that we discovered in Chapter 2.

For BAY approach, let $\{(\boldsymbol{\alpha}_{K+1}^{(q,T)})^{(g)}\}_{(g)}$ be the set of posterior samples for parameter α based on release q of month T in quarter $K+1$, $\boldsymbol{\alpha}$ can be $\mathbf{A}$ and $\boldsymbol{\beta}$. Naive Kolmogorov−Smirnov tests

are performed to compare $\{(\boldsymbol{\alpha}_{K+1}^{(1,T)})^{(g)}\}_{(g)}$ with $\{(\boldsymbol{\alpha}_{K+1}^{(2,T)})^{(g)}\}_{(g)}$ and $\{(\boldsymbol{\alpha}_{K+1}^{(3,T)})^{(g)}\}_{(g)}$, for all months in a quarter, and all nowcasting quarters. None of the test fail to reject the null hypothesis, which suggests there is no statistically significant difference between posterior samples for $\mathbf{A}$ or $\boldsymbol{\beta}$ based on the first, second, or third release, no matter which month of the quarter we are in for all quarters of nowcast.

Another key set of parameters that is used in nowcasting equation is the latent factors for the quarter of nowcast. The time-series for factors are constructed as follows. For example, when nowcasting y_{39} at all three releases of month 115, $\mathbf{F}_{115}$ is estimated directly by $\hat{\mathbf{F}}_{115}$, $\mathbf{F}_{116}$ is estimated by $\hat{\mathbf{A}}\hat{\mathbf{F}}_{115}$, and $\mathbf{F}_{117}$ is estimated by $\hat{\mathbf{A}}^2\hat{\mathbf{F}}_{115}$. Similarly, when nowcasting y_{39} at all three releases of month 116, $\mathbf{F}_{115}$ is estimated directly by $\hat{\mathbf{F}}_{115}$, $\mathbf{F}_{116}$ is directly estimated by $\hat{\mathbf{F}}_{116}$, and $\mathbf{F}_{117}$ is estimated by $\hat{\mathbf{A}}\hat{\mathbf{F}}_{116}$. When nowcasting y_{39} at all three releases of month 117, $\mathbf{F}_{115}$ is estimated directly by $\hat{\mathbf{F}}_{115}$, $\mathbf{F}_{116}$ is estimated by $\hat{\mathbf{F}}_{116}$, and $\mathbf{F}_{117}$ is directly estimated by $\hat{\mathbf{F}}_{117}$. Then similar things are done for y_{40}, y_{41}, and so on. Therefore, factors in the first month are $(\hat{\mathbf{F}}_{115}, \hat{\mathbf{A}}\hat{\mathbf{F}}_{115}, \hat{\mathbf{A}}^2\hat{\mathbf{F}}_{115}, \hat{\mathbf{F}}_{118}, \hat{\mathbf{A}}\hat{\mathbf{F}}_{118}, \hat{\mathbf{A}}^2\hat{\mathbf{F}}_{118}, ...)$, factors in the second month are $(\hat{\mathbf{F}}_{115}, \hat{\mathbf{F}}_{116}, \hat{\mathbf{A}}\hat{\mathbf{F}}_{116}, \hat{\mathbf{F}}_{118}, \hat{\mathbf{F}}_{119}, \hat{\mathbf{A}}\hat{\mathbf{F}}_{119}, ...)$, and factors in the third month are $(\hat{\mathbf{F}}_{115}, \hat{\mathbf{F}}_{116}, \hat{\mathbf{F}}_{117}, \hat{\mathbf{F}}_{118}, \hat{\mathbf{F}}_{119}, \hat{\mathbf{F}}_{120}, ...)$, posterior means are used as parameter estimation. In Figure B.5 and B.6, factors based on all three releases when nowcasting in the first, second, and third month are presented in left, middle, and right column respectively, with first row being the first factor, second row being second factor, and third row being third factor. Estimations based on 3 releases are colored in red, green, and blue, with different node type.

There are no noticeable changes from release to release in all the three factor across all three months for BAY approach. Meanwhile, for GRS approach, no noticeable changes from release to release in the first factor across all three months. Some changes can be spotted for the second factors. From release to release, the estimations for third factor changes significantly in all three months. Therefore, given the fact that estimations for $\mathbf{A}$, $\boldsymbol{\beta}$, and $\mathbf{F}_t$ does not change much from release to release in the same month for BAY approach, the MANE does not change much from release to release in the same month. And, given the fact that estimations for $\mathbf{A}$, $\boldsymbol{\beta}$ stay same from

release to release in the same month, estimations for $\mathbf{F}_t$ changes from release to release in the same month, the MANE changes accordingly.

B.3 Inference for Empirical Study

In this section, we provide the reason why there is not much variation in BAY nowcasts from release to release, in all the three month for the empirical study.

Using posterior means as sample mean and posterior standard deviations as estimated sample standard deviation, 95% confidence intervals for $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3$ based on each release in each month are calculated and plotted in Figure B.7, B.8, and B.9. We see that all the confidence intervals contains 0, which means all the $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3$ are not statistically significant different from 0. Meanwhile, the confidence interval for β_4 based on each release in each month is contained in Table B.1. All CIs in table 1 do not contain 0 but contain 1 implies β_4 is statistically significant different from 0 but not statistically different from 1 for all releases in all month. Therefore, the nowcasts are dominated by the previous quarter's GDP value, and the coefficient associated with lagged GDP does not change much from release to release, hence, little difference can be found for nowcasts between releases.

Table B.1: 95% CIs for $\boldsymbol{\beta}_4$ based on all releases in all three month of the quarter.

Rleases	First Month	Second Month	Third Month
RL1	(0.691,1.046)	(0.686,1.050)	(0.685,1.040)
RL2	(0.689,1.047)	(0.689,1.047)	(0.686,1.042)
RL3	(0.697,1.047)	(0.696,1.054)	(0.691,1.040)
RL4	(0.690,1.040)	(0.693,1.039)	(0.684,1.042)
RL5	(0.695,1.052)	(0.692,1.045)	(0.687,1.039)
RL6	(0.700,1.049)	(0.696,1.044)	(0.681,1.049)
RL7	(0.710,1.056)	(0.695,1.054)	(0.704,1.056)
RL8	(0.701,1.047)	(0.712,1.058)	(0.707,1.052)

B.4 Information of Monthly Series in Empirical Study

In this section, we provide the information of the 117 monthly data series that are used in empirical study, including name of the series and the transformation used in order to achieve stationary of the series. Transformation 1, 2, 3 stand for 12-month growth rate, 12-month difference and no transformation respectively.

Table B.2: Information of the monthly data series. Transformation 1, 2, 3 stand for 12-month growth rate, 12-month difference and no transformation respectively.

Name of the Series	Transformation
National Interbank Offered Rate: Weighted Avg: NIBFC: 7 Days	2
National Interbank Offered Rate: Weighted Avg: NIBFC: 30 Days	2
National Interbank Offered Rate: Weighted Avg: NIBFC: 60 Days	2
National Interbank Offered Rate: Weighted Avg: NIBFC: 90 Days	2
Central Bank Base Interest Rate: Required Reserve	2
CN: Central Bank Benchmark Interest Rate: Reserve Requirement: Excess	2
Savings Deposits Rate	2
Time Deposits Rate: 1 Year	2
Base Lending Rate: Capital Construction: Less than 10 Year	2
Index: Shanghai Stock Exchange: Composite	1
Index: Shanghai Stock Exchange: Industrial	1
Index: Shanghai Stock Exchange: Utilities	1
Index: Shenzhen Stock Exchange: Composite	1
PE Ratio: Shanghai Stock Ex: All Shares: Wgt Avg by Issued Volume	2
Market Capitalization: Shanghai Stock Exchange: Stocks	1
LIBOR: BBA: Euro: 3 Months : Frequency Transform	2
FTSE 100 INDEX	1
The Bombay Stock Exchange Sensitive Index (Sensex)	1
The KOSPI Index	1
Kuala Lumpur Stock Exchange Composite Index	1
The Bangkok SET Index	1
Jakarta Stock Price Index	1
The Philippine Stock Exchange PSEi Index	1
TWSE Index	1
Hang Seng Index	1
Dividend Yield: Hang Seng Index	2
Dividend Yield: The Australian All Ordinaries Index	2
Dividend Yield: The New Zealand All Ordinaries Index	2

Table B.2 continued

Name of the Series	Transformation
CPI: Food (sa)	1
CPI: Clothing (sa)	1
CPI: Traffic, Communications (sa)	1
CPI: Residence (sa)	1
CPI: Aggregate (sa)	1
Retail Price Index, Aggregate_sa	1
Exports fob	1
Imports cif	1
Commodity Building Sold: Total	1
CN: Commodity Bldg Selling Price	1
CN: Floor Space Started: Total	1
CN: Floor Space under Construction: Commodity Bldg (CB)	1
Government Surplus or Deficit	1
Fixed Asset Investment: Levels, Total (revised) S.A.	1
Real Estate Investment, Total (SA)	1
Foreign Direct Investment Utilized	1
FDI: Utilized: Foreign Enterprises	1
Spot Exchange Rate: Period Avg: SAFE: RMB to US Dollar	1
Spot Exchange Rate: Period Avg: SAFE: RMB to Hong Kong Dollar	1
Spot Exchange Rate: Period Avg: SAFE: RMB to British Pound	1
Spot Exchange Rate: Period Avg: SAFE: RMB to Canadian Dollar	1
Spot Exchange Rate: Period Avg: SAFE: RMB to Swedish Krone	1
Spot Exchange Rate: Period Avg: SAFE: RMB to Singapore Dollar	1
Spot Exchange Rate: Period Avg: SAFE: RMB to Australian Dollar	1
Nominal effective exchange rate	1
Real effective exchange rate, sa	1
Producer Price Index: Industrial Products (yoy%)	3
Producer Price Index: Industrial Products: Producer Goods (yoy%)	3
Producer Price Index: Industrial Products: Producer Goods: Excavation (yoy%)	3
Producer Price Index: Industrial Products: Producer Goods: Raw Material (yoy%)	3
Producer Price Index: Industrial Products: Producer Goods: Manufacturing (yoy%)	3
Producer Price Index: Industrial Products: Consumer Goods (yoy%)	3
Producer Price Index, Consumer Goods: Food (yoy%)	3
Producer Price Index, Consumer Goods: Clothing, (yoy%)	3
Producer Price Index, Consumer Goods: Daily Use Articles (yoy%)	3

Table B.2 continued

Name of the Series	Transformation
Index of Market Prices: All Primary Commodities Index	1
Index of Market Prices: Non Fuel Primary Commodities Index	1
Index of Market Prices: Food	1
Index of Market Prices: Beverages	1
Index of Market Prices: Agricultural Raw Materials	1
Index of Market Prices: Metals	1
Index of Market Prices: Energy Index	1
Index of Market Prices: Petroleum, average crude price	1
Index of Market Prices: Sugar Caribbean (N.Y.) Free Market	1
Index of Market Prices: Uranium Index	1
Exports: Primary Products	1
Exports: Manufactures	1
Exports: Animal and Vegetable Oils, Fats and Waxes (AVFW)	1
Exports: Chemicals and Related Products (CRP)	1
Exports: Crude Materials, Inedible, Except Fuels (CM)	1
Exports: Food and Live Animals Chiefly For Food (FLA)	1
Exports: Machinery and Transport Equipment (MTE)	1
Exports: Manufactured Goods Chiefly by Materials (MG)	1
Exports: Mineral Fuels, Lubricants and Related Materials (MFLM)	1
Exports: Miscellaneous Manufactured Articles (MMA)	1
Imports: Primary Products	1
Imports: Manufactures	1
Imports: Animal and Vegetable Oils, Fats and Waxes (AVFW)	1
Imports: Beverages and Tobacco (BT)	1
Imports: Chemicals and Related Products (CRP)	1
Imports: Crude Materials, Inedible , Except Fuels (CM)	1
Imports: Food and Live Animals Chiefly For Food (FLA)	1
Imports: Machinery and Transport Equipment (MTE)	1
Imports: Manufactured Goods Chiefly by Materials (MG)	1
Imports: Mineral Fuels, Lubricants and Related Materials (MFLM)	1
Imports: Miscellaneous Manufactured Articles (MMA)	1

Table B.2 continued

Name of the Series	Transformation
Retail Sales of Consumer Goods, Total (revised) S.A.	1
CN: Coincident Index	1
CN: Business Cycle Signal	1
Industrial Production: Household Washing Machines	1
Industrial Production: Household Refrigerator	1
Industrial Production: Cloth	1
Industrial Production: Processed Crude Oil	1
Industrial Production: Diesel Oil	1
Industrial Production: Power Generated	1
Industrial Production: Steel	1
Industrial Production: Steel Products	1
Industrial Production: Iron Alloy	1
Industrial Production: Cement	1
Industrial Production: Automobiles	1
Industrial Production: Automobiles: Cars	1
Industrial Production: Automobiles: Loading Vehicles	1
Financial Institution Loans	1
Financial Institution Deposits	1
Money Supply M0	1
Money Supply M1	1
Money Supply M2	1
CN: Required Reserve Ratio	2
Foreign Reserves	1

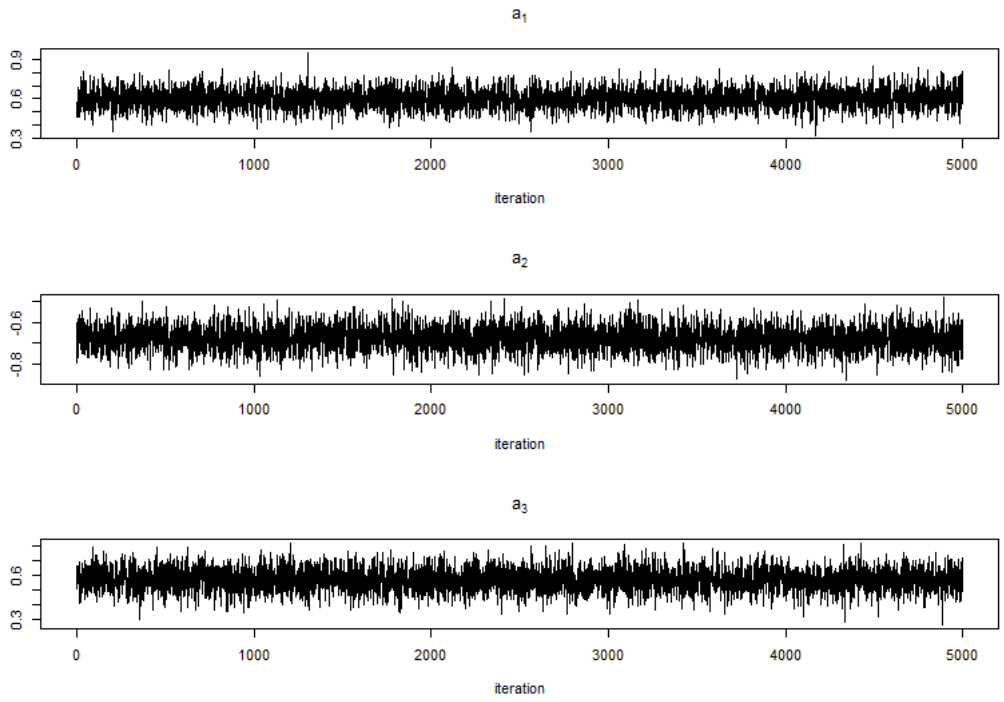

Figure B.1: Time-series plot for posterior samples of **A** for Simulation 1.

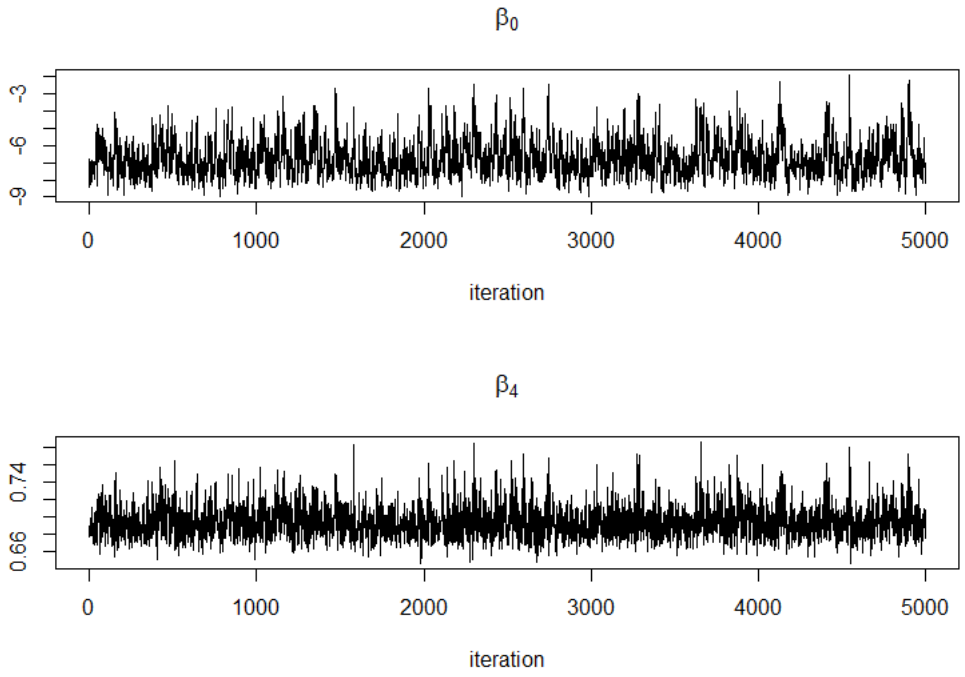

Figure B.2: Time-series plot for posterior samples of β_0, β_4 for Simulation 1.

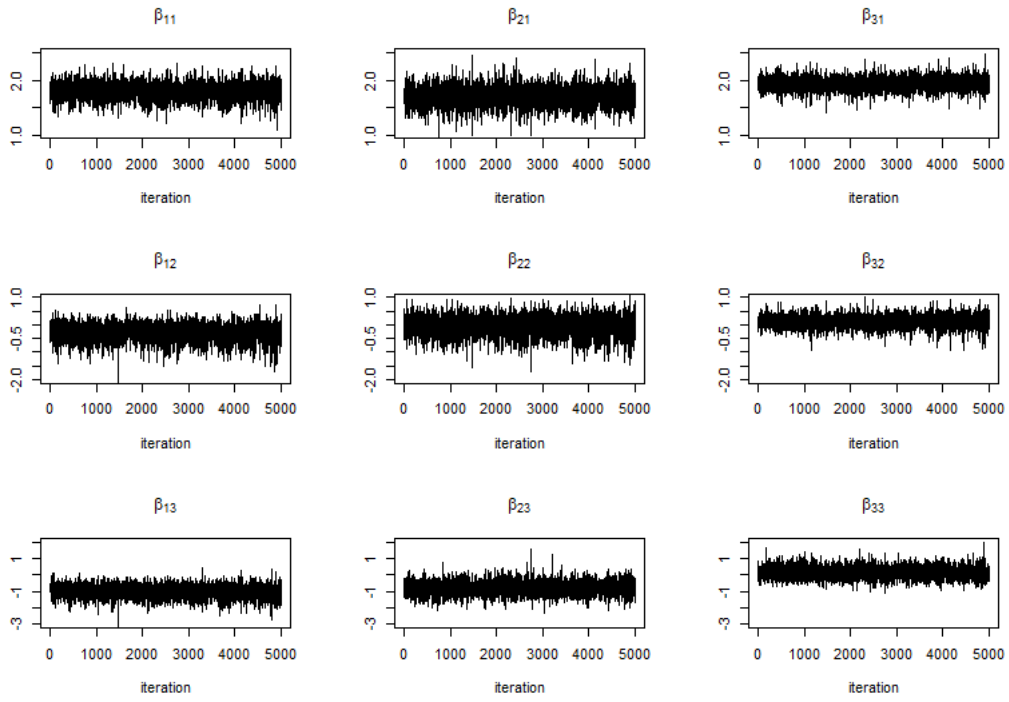

Figure B.3: Time-series plot for posterior samples of $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3$ for Simulation 1.

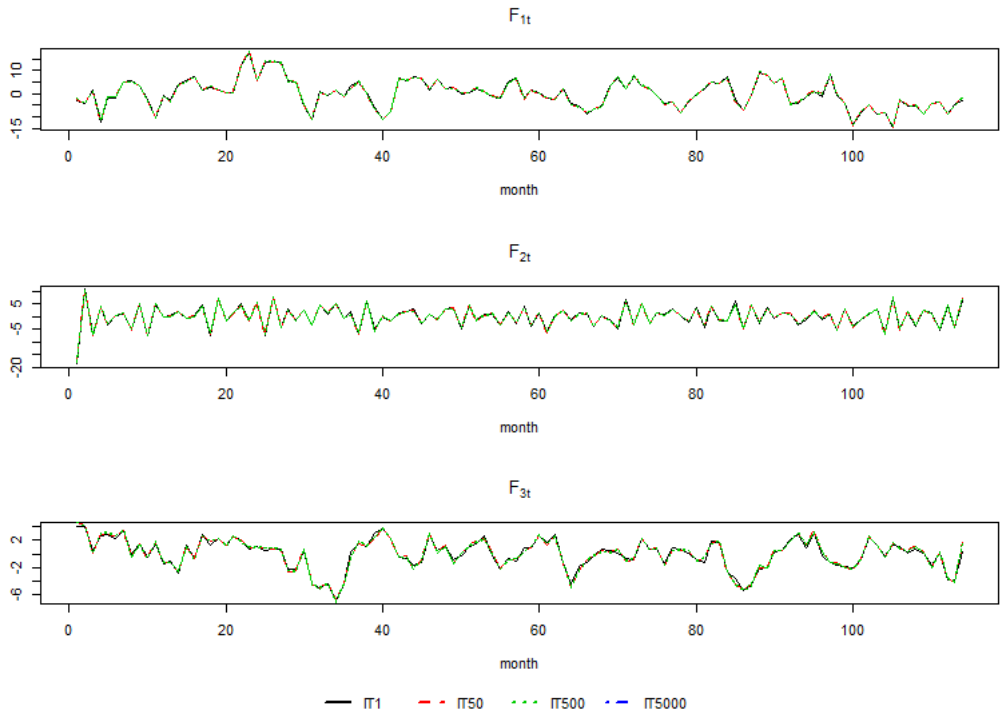

Figure B.4: Posterior samples of $\mathbf{F}_t$ from $1^{st}, 50^{th}, 500^{th}, 5000^{th}$ iteration for Simulation 1.

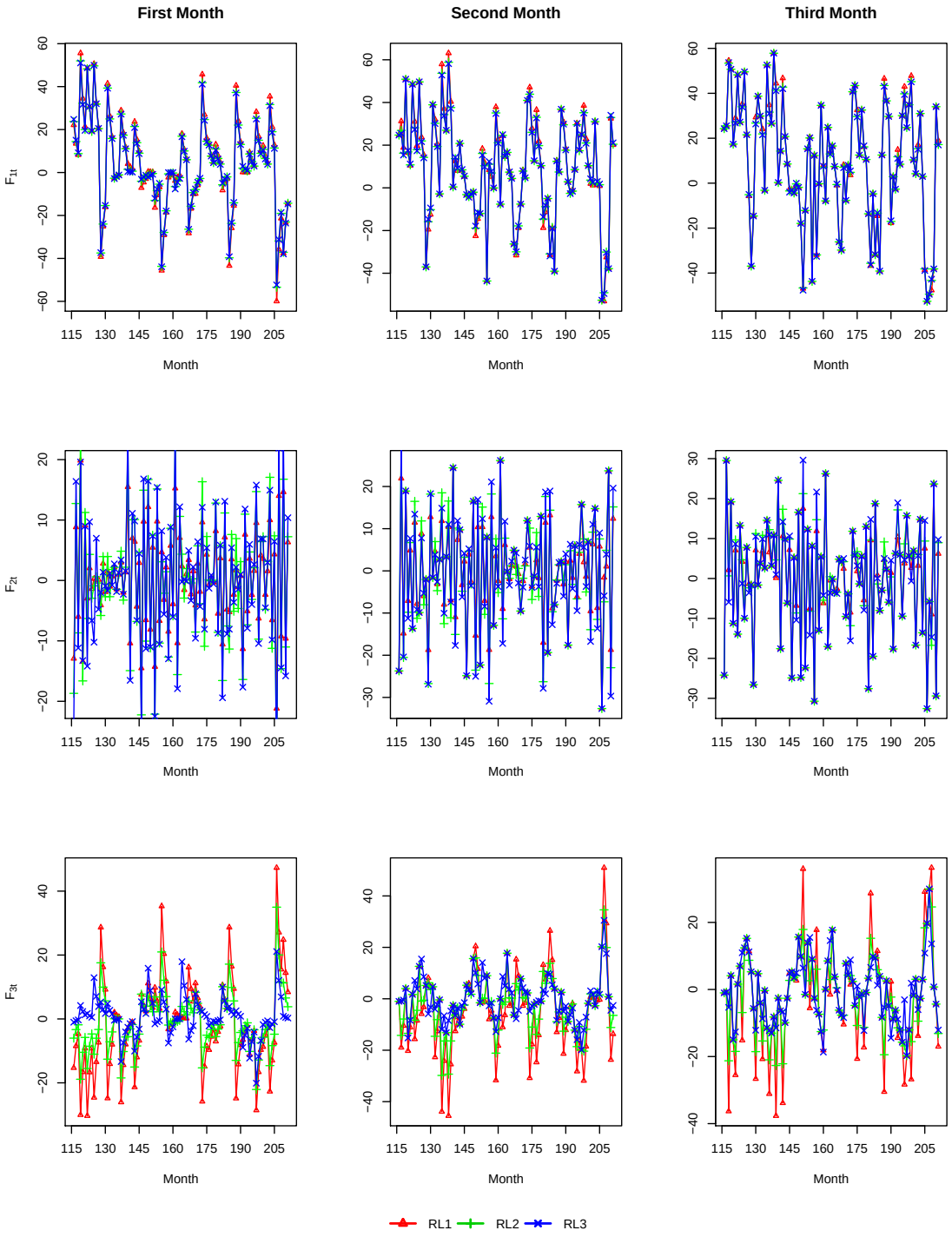

Figure B.5: $\mathbf{F}_t$ estimation based on three releases in all month of the quarter for GRS approach. From top to bottom, are estimations for the first, second and third latent factors. From left to right, are estimations in first, second, and third month of the quarter. Three releases are colored in red, green, abd blue with different node type.

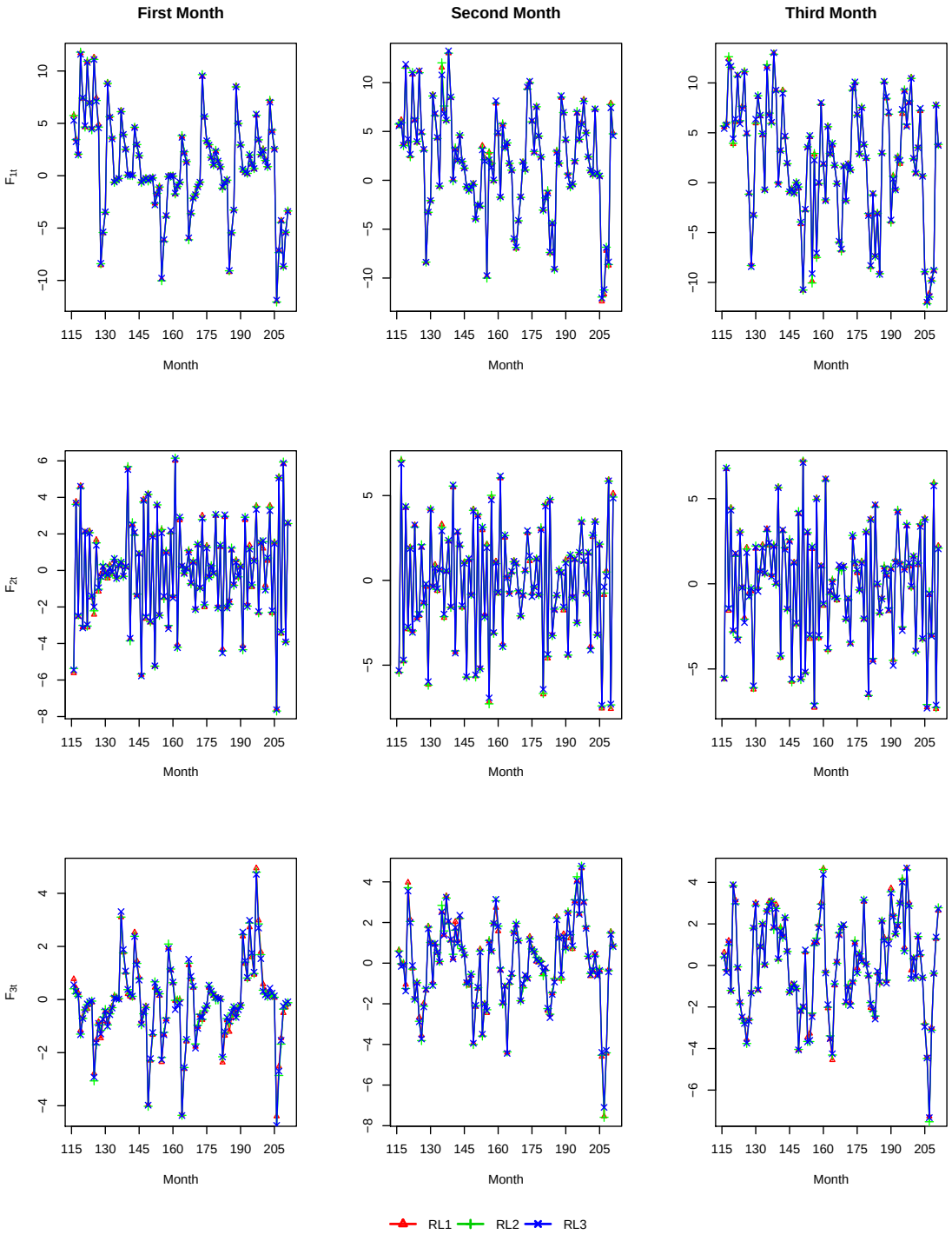

Figure B.6: $\mathbf{F}_t$ estimation based on three releases in all month of the quarter for BAY approach. From top to bottom, are estimations for the first, second and third latent factors. From left to right, are estimations in first, second, and third month of the quarter. Three releases are colored in red, green, abd blue with different node type.

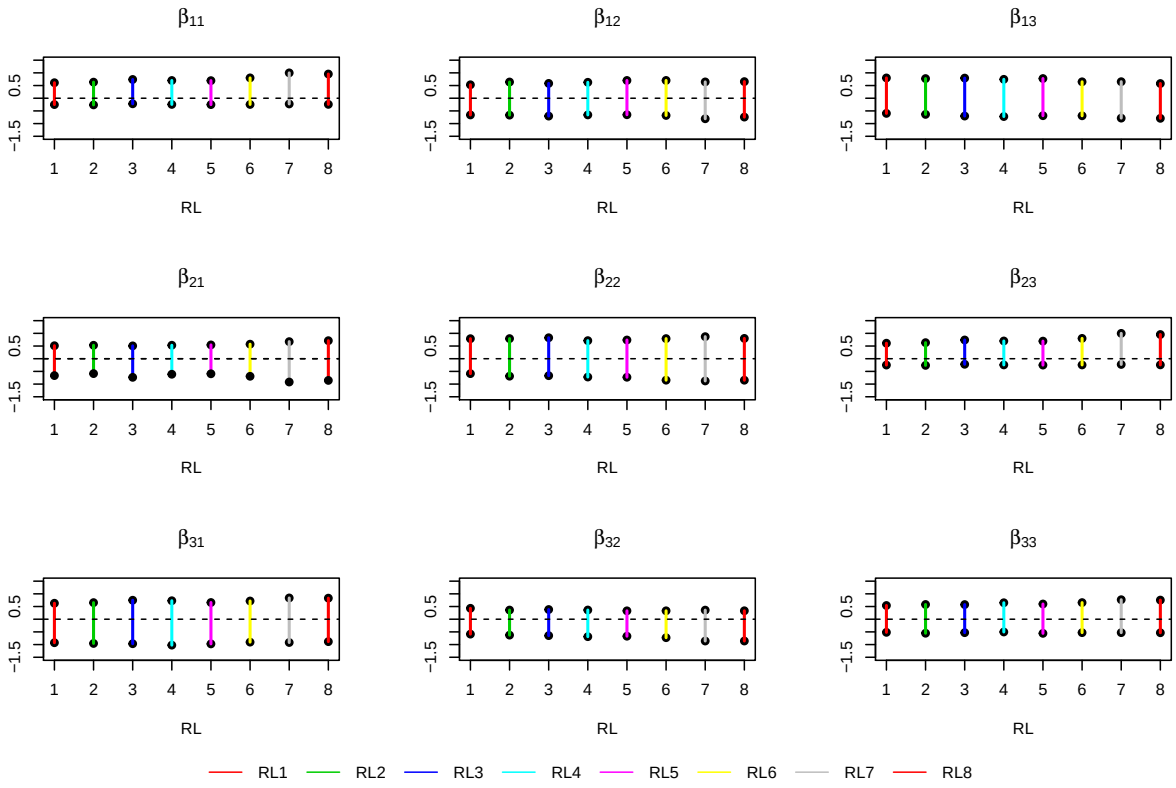

Figure B.7: 95% CIs for $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3$ estimation based on different releases in the first month. CIs based on different releases are color coded. RL stands for release, dotted line represents 0.

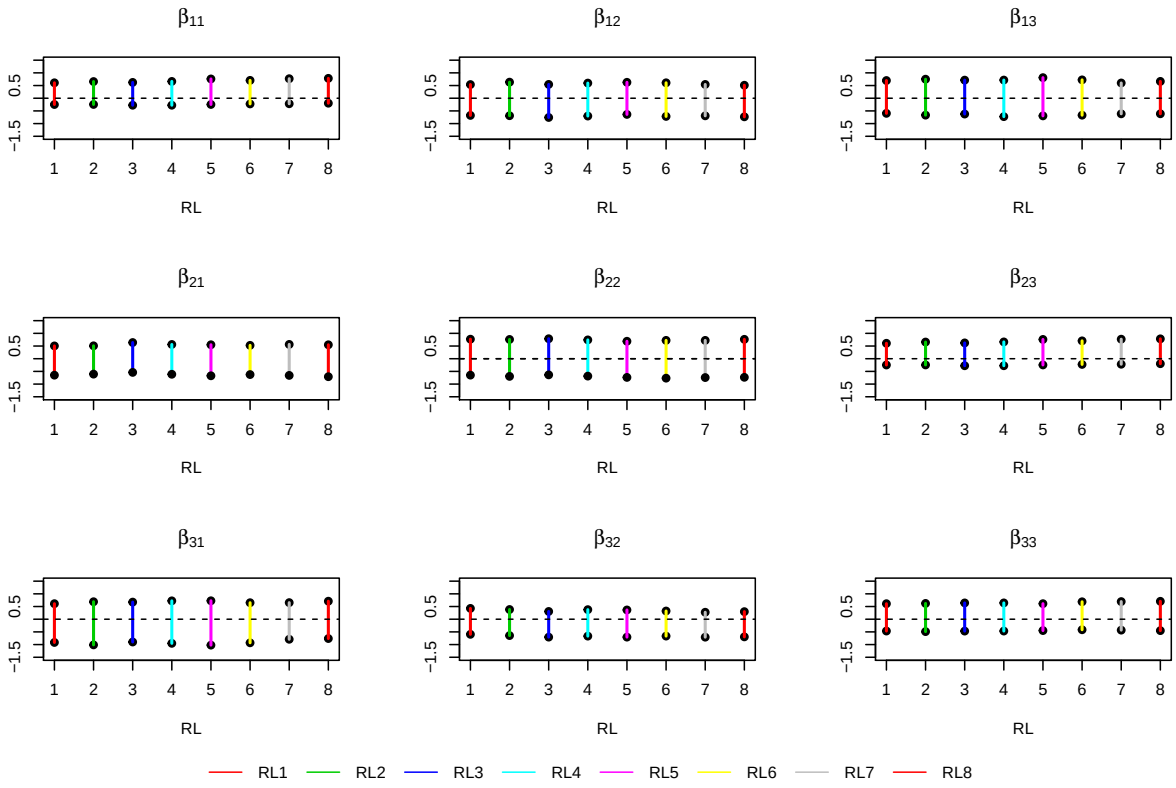

Figure B.8: 95% CIs for $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3$ estimation based on different releases in the second month. CIs based on different releases are color coded. RL stands for release, dotted line represents 0.

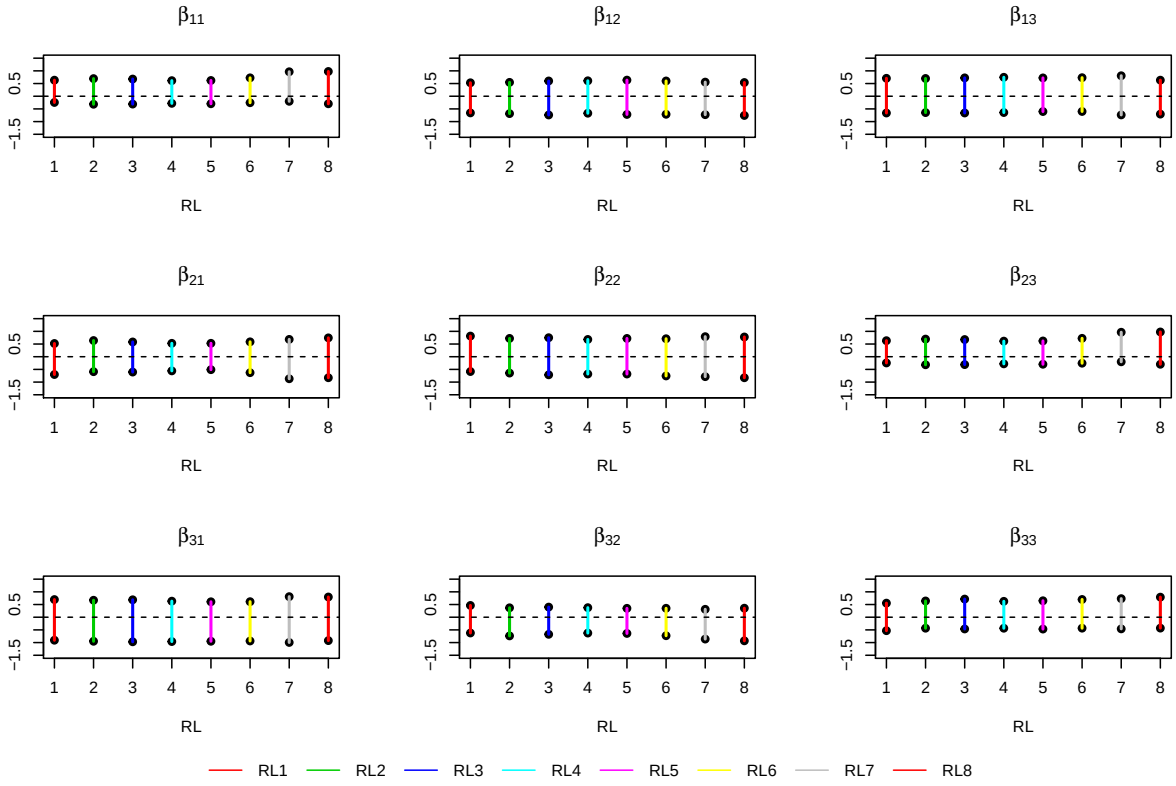

Figure B.9: 95% CIs for $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3$ estimation based on different releases in the third month. CIs based on different releases are color coded. RL stands for release, dotted line represents 0.

APPENDIX C. CHAPTER 3 APPENDIX

In this appendix, we present the full conditional posterior distributions. Most parameters have closed form full conditional distribution, allowing for a straightforward Gibbs sampling method. However, $\mathbf{\Omega}$ does not have closed form posterior distribution. In this case, we insert an independent Metropolis-Hastings (MH) sampler within Gibbs sampler to generate posterior samples for $\mathbf{\Omega}$.

C.1 Sampling the mean of monthly series $\boldsymbol{\mu}$

The full conditional of mean of monthly series, $\boldsymbol{\mu}$, is a multivariate normal distribution:

$$\boldsymbol{\mu}|\mathbf{X}_{(q,T)}, \mathbf{Z}, \mathbf{F}, \mathbf{\Theta}, \mathbf{\Omega} \sim N(\mathbf{U}_{\boldsymbol{\mu}}, \mathbf{W}_{\boldsymbol{\mu}}^{-1}), \tag{C.1}$$

where

$$\begin{aligned}
\mathbf{W}_{\boldsymbol{\mu}} &= T_q\mathbf{\Omega}^{-1} + \sum_{t=T_q+1}^{T} (\mathbf{1}'_{v_{q,t}}(\mathbf{1}_{v_{q,t}}\mathbf{\Omega}\mathbf{1}'_{v_{q,t}})^{-1}\mathbf{1}_{v_{q,t}}) + \mathbf{I}_{n\times n}, \\
\mathbf{U}_{\boldsymbol{\mu}} &= \mathbf{W}_{\boldsymbol{\mu}}^{-1}\Big[\sum_{t=1}^{T_q}\mathbf{\Omega}^{-1}(\mathbf{x}_t - \mathbf{\Theta}\mathbf{Z}\mathbf{F}_t) + \sum_{t=T_q+1}^{T} (\mathbf{1}'_{v_{q,t}}(\mathbf{1}_{v_{q,t}}\mathbf{\Omega}\mathbf{1}'_{v_{q,t}})^{-1}\mathbf{1}_{v_{q,t}})(\mathbf{x}_t - \mathbf{\Theta}\mathbf{Z}\mathbf{F}_t)\Big].
\end{aligned}$$

C.2 Sampling the factor loadings matrix $\boldsymbol{\theta}$

The full conditional of factor loadings, $\boldsymbol{\theta}$, is a multivariate normal distribution:

$$\boldsymbol{\theta}|\mathbf{X}_{(q,T)}, \boldsymbol{\mu}, \mathbf{Z}, \mathbf{F}, \mathbf{\Omega} \sim N(\mathbf{U}_{\boldsymbol{\theta}}, \mathbf{W}_{\boldsymbol{\theta}}^{-1}), \tag{C.2}$$

where $\mathbf{C}_t = \mathbf{I}_{n\times n} \otimes (\mathbf{Z}\mathbf{F}_t)'$, and

$$\begin{aligned}
\mathbf{W}_{\boldsymbol{\theta}} &= \sum_{t=1}^{T_q}\mathbf{C}'_t\mathbf{\Omega}^{-1}\mathbf{C}_t + \sum_{t=T_q+1}^{T}\mathbf{C}'_t\mathbf{1}'_{v_{q,t}}(\mathbf{1}_{v_{q,t}}\mathbf{\Omega}\mathbf{1}'_{v_{q,t}})^{-1}\mathbf{1}_{v_{q,t}}\mathbf{C}_t + \mathbf{I}_{nR\times nR}, \\
\mathbf{U}_{\boldsymbol{\theta}} &= \mathbf{W}_{\boldsymbol{\theta}}^{-1}\Big[\sum_{t=1}^{T_q}\mathbf{C}'_t\mathbf{\Omega}^{-1}(\mathbf{x}_t - \boldsymbol{\mu}) + \sum_{t=T_q+1}^{T}\mathbf{C}'_t\mathbf{1}'_{v_{q,t}}(\mathbf{1}_{v_{q,t}}\mathbf{\Omega}\mathbf{1}'_{v_{q,t}})^{-1}\mathbf{1}_{v_{q,t}}(\mathbf{x}_t - \boldsymbol{\mu})\Big].
\end{aligned}$$

C.3 Sampling the covariance in monthly series $\mathbf{\Omega}$

The posterior for $\mathbf{\Omega}$ is

$$\begin{aligned} P(\mathbf{\Omega}|\cdot) &\propto \prod_{t=1}^{T_q} P(\mathbf{x}_t|\boldsymbol{\mu},\mathbf{\Theta},\mathbf{Z},\mathbf{F}_t,\mathbf{\Omega}) \times \prod_{t=T_q+1}^{T} P(\mathbf{x}_{i\in v_{q,t}}|\boldsymbol{\mu},\mathbf{\Theta},\mathbf{Z},\mathbf{F}_t,\mathbf{\Omega}) \times P(\mathbf{\Omega}) \\ &\propto |\mathbf{\Omega}|^{-T_q/2} e^{-\frac{1}{2}tr\left[\sum_{t=1}^{T_q}(\mathbf{x}_t-\boldsymbol{\mu}-\mathbf{C}_t\boldsymbol{\theta})(\mathbf{x}_t-\boldsymbol{\mu}-\mathbf{C}_t\boldsymbol{\theta})'\mathbf{\Omega}^{-1}\right]} \\ &\times \prod_{t=T_q+1}^{T} |\mathbf{1}_{v_{q,t}}\mathbf{\Omega}\mathbf{1}'_{v_{q,t}}|^{-1/2} e^{-\frac{1}{2}tr\left[\mathbf{1}_{v_{q,t}}(\mathbf{x}_t-\boldsymbol{\mu}-\mathbf{C}_t\boldsymbol{\theta})\mathbf{1}'_{v_{q,t}}(\mathbf{x}_t-\boldsymbol{\mu}-\mathbf{C}_t\boldsymbol{\theta})'(\mathbf{1}_{v_{q,t}}\mathbf{\Omega}\mathbf{1}'_{v_{q,t}})^{-1}\right]} \\ &\times |\mathbf{\Omega}|^{-(\nu_\theta+n+1)/2} e^{-\frac{1}{2}tr\left[\frac{1}{n}\mathbf{\Omega}^{-1}\mathbf{I}_{n\times n}\right]}, \end{aligned}$$

which is not in closed form. Note that the first and third pieces to yields $\mathbf{\Omega}$ has an inverse Wishart distribution $iW(\Psi_{\mathbf{\Omega}},\nu_{\mathbf{\Omega}})$, where $\Psi_{\mathbf{\Omega}} = \sum_{t=1}^{T_q}(\mathbf{x}_t-\boldsymbol{\mu}-\mathbf{C}_t\boldsymbol{\theta})(\mathbf{x}_t-\boldsymbol{\mu}-\mathbf{C}_t\boldsymbol{\theta})' + \frac{1}{n}\mathbf{I}_{n\times n}$, and $\nu_{\mathbf{\Omega}} = T_q + \nu_{\boldsymbol{\theta}}$. Therefore, we can propose $\mathbf{\Omega}^*$ using $iW(\Psi_{\mathbf{\Omega}},\nu_{\mathbf{\Omega}})$, and use the middle piece in the full conditional, denote as $q(\mathbf{\Omega})$, to construct the acceptance-rejection rate. That is the proposal $\mathbf{\Omega}^*$ is accepted with probablity

$$A(\mathbf{\Omega}^*|\mathbf{\Omega}^c) = min\Big\{1, \frac{q(\mathbf{\Omega}^*)}{q(\mathbf{\Omega}^c)}\Big\},$$

where $\mathbf{\Omega}^c$ denotes the current state of $\mathbf{\Omega}$.

C.4 Sampling the AR(1) coefficients a_j

For $j = 1, ..., R$, the full conditional of each AR(1) coefficient, a_j, is a truncated normal distribution:

$$P(a_j|f_{j,1},...,f_{j,T},\sigma_j^2) = \begin{cases} 0 & \text{if } a_j \le -1 \\ \dfrac{\phi(\frac{a_j-\mu_{a_j}}{\sigma_{a_j}})}{\Phi(\frac{1-\mu_{a_j}}{\sigma_{a_j}}) - \Phi(\frac{-1-\mu_{a_j}}{\sigma_{a_j}})} & \text{if } -1 < a_j < 1 \\ 0 & \text{if } a_j \ge 1 \end{cases} \tag{C.3}$$

where $\mu_{a_j} = \dfrac{\sum_{t=2}^{T} f_{j,t} f_{j,t-1}}{\sigma_j^2 + \sum_{t=2}^{T} f_{j,t-1}^2}$ and $\sigma_{a_j}^2 = \dfrac{\sigma_j^2}{\sigma_j^2 + \sum_{t=2}^{T} f_{j,t-1}^2}$.

C.5 Sampling the variance in factor equations σ_j^2

For $j = 1, ..., R$, the full conditional of σ_j^2 is an inverse Gamma distribution:

$$\sigma_j^2 | f_{j,1}, ..., f_{j,T}, a_j \sim iG(\alpha_j, \beta_j), \tag{C.4}$$

where $\alpha_j = \alpha_s + (T-1)/2$ and $\beta_j = \beta_s + \sum_{t=2}^{T} (f_{j,t} - a_j f_{j,t-1})^2/2$.

C.6 Sampling the coefficients associated with factors $\boldsymbol{\beta}$

The full conditional of coefficients associated with factors in GDP equation, $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \boldsymbol{\beta}_3', \beta_4)'$, is a multivariate normal distribution:

$$\boldsymbol{\beta} | \mathbf{Z}, \mathbf{F}, \mathbf{Y}, \eta^2 \sim N(\mathbf{U}_{\boldsymbol{\beta}}, \mathbf{W}_{\boldsymbol{\beta}}^{-1}), \tag{C.5}$$

where $\tilde{\mathbf{F}}_{3k} = [1, (\mathbf{Z}\mathbf{F}_{3k})', (\mathbf{Z}\mathbf{F}_{3k-1})', (\mathbf{Z}\mathbf{F}_{3k-2})', y_{k-1}]'$, and

$$\begin{aligned} \mathbf{W}_{\boldsymbol{\beta}} &= \sum_{k=2}^{K} \tilde{\mathbf{F}}_{3k} \tilde{\mathbf{F}}_{3k}'/\eta^2 + \mathbf{I}_{(3R+2)\times(3R+2)}, \\ \mathbf{U}_{\boldsymbol{\beta}} &= \mathbf{W}_{\boldsymbol{\beta}}^{-1} \sum_{k=2}^{K} \tilde{\mathbf{F}}_{3k} y_k/\eta^2. \end{aligned}$$

C.7 Sampling the variance in GDP equation η^2

The full conditional of the variance in GDP equation, η^2, is a inverse Gamma distribution:

$$\eta^2 | \mathbf{Z}, \mathbf{F}, \mathbf{Y}, \boldsymbol{\beta} \sim iG(\alpha_\eta, \beta_\eta), \tag{C.6}$$

where $\alpha_\eta = \alpha_h + (K-1)/2$ and $\beta_\eta = \beta_h + \sum_{k=2}^{K} (y_k - \boldsymbol{\beta}' \tilde{\mathbf{F}}_{3k})^2/2$.

C.8 Sampling the latent factors $\mathbf{F}_t$

This section provides a way to understand the derivation of posterior distribution of $\mathbf{F}_t$ from the point of GLS's view.

The posterior distribution for $\mathbf{F}_t$ will have different form depending on t. The most general cases are, for $3 < t < 3K$, $k = 2, ..., K$. t can be written as $t = 3k - i$, for some k and $i = 0, 1, 2$, depending on which month of quarter k we are in. If $i = 0$, we are in the third month of quarter k, if $i = 1$, we are in the second month of quarter k, and, if $i = 2$, we are in the first month of quarter k. Then, at $t = 3k - i$, $\mathbf{F}_t$ enters the joint likelihood through four parts, $\mathbf{x}_t$, $\mathbf{F}_{t+1}$, $\mathbf{F}_{t-1}$, and y_k. Therefore, we can write

$$\begin{pmatrix} \mathbf{x}_t - \boldsymbol{\mu} \\ \mathbf{F}_{t+1} \\ \mathbf{F}_{t-1} \\ f_{y_k}(i) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Theta}\mathbf{Z} \\ \mathbf{A} \\ \mathbf{A}^{-1} \\ \boldsymbol{\beta}'_{i+1}\mathbf{Z} \end{pmatrix} \mathbf{F}_t + \begin{pmatrix} \epsilon_t \\ u_{t+1} \\ -\mathbf{A}^{-1}u_t \\ \nu_k \end{pmatrix} \equiv \tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\mathbf{F}_t + \tilde{\boldsymbol{\epsilon}},$$

where $f_{y_k}(i)$ is a function of i depending on y_k, defined as

$$f_{y_k}(i) = \begin{cases} y_k - \beta_0 - \boldsymbol{\beta}'_2\mathbf{Z}\mathbf{F}_{t-1} - \boldsymbol{\beta}'_3\mathbf{Z}\mathbf{F}_{t-2} - \beta_4 y_{k-1} & \text{if } i = 0 \\ y_k - \beta_0 - \boldsymbol{\beta}'_1\mathbf{Z}\mathbf{F}_{t+1} - \boldsymbol{\beta}'_3\mathbf{Z}\mathbf{F}_{t-1} - \beta_4 y_{k-1} & \text{if } i = 1 \\ y_k - \beta_0 - \boldsymbol{\beta}'_1\mathbf{Z}\mathbf{F}_{t+2} - \boldsymbol{\beta}'_2\mathbf{Z}\mathbf{F}_{t+1} - \beta_4 y_{k-1} & \text{if } i = 2 \end{cases} \tag{C.7}$$

Therefore, we have

$$\tilde{\mathbf{Y}} = \begin{bmatrix} \mathbf{x}_t - \boldsymbol{\mu} \\ \mathbf{F}_{t+1} \\ \mathbf{F}_{t-1} \\ f_{y_k}(i) \end{bmatrix}, \quad \tilde{\mathbf{X}} = \begin{bmatrix} \boldsymbol{\Theta}\mathbf{Z} \\ \mathbf{A} \\ \mathbf{A}^{-1} \\ \boldsymbol{\beta}'_{i+1}\mathbf{Z} \end{bmatrix}, \quad \tilde{\boldsymbol{\epsilon}} = \begin{bmatrix} \epsilon_t \\ u_{t+1} \\ -\mathbf{A}^{-1}u_t \\ \nu_k \end{bmatrix},$$

and

$$var(\tilde{\boldsymbol{\epsilon}}) = \tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \boldsymbol{\Omega} & 0 & 0 & 0 \\ 0 & \boldsymbol{\Sigma} & 0 & 0 \\ 0 & 0 & (\mathbf{A}'\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1} & 0 \\ 0 & 0 & 0 & \eta^2 \end{bmatrix}.$$

By weighted regression, $3 < t < 3K$, $k = 2, ..., K$, $t = 3k - i$, draw

$$\mathbf{F}_t | \tilde{\mathbf{Y}}, \tilde{\mathbf{X}}, \tilde{\boldsymbol{\Sigma}} \sim N((\tilde{\mathbf{X}}'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{Y}}, (\tilde{\mathbf{X}}'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{X}})^{-1}). \tag{C.8}$$

For other t, the posterior distribution for $\mathbf{F}_t$ is of the same form, with some deletion done to $\tilde{\mathbf{Y}}$, $\tilde{\mathbf{X}}$, and $\tilde{\mathbf{\Sigma}}$. For example, if $t = 1$, since we don't have $\mathbf{F}_0$ and y_0, delete entries corresponding to $\mathbf{F}_{t-1}$ and $f_{y_k}(i)$ in $\tilde{\mathbf{Y}}$, $\tilde{\mathbf{X}}$, and $\tilde{\mathbf{\Sigma}}$. If $t = 2, 3$, since we don't have y_0, delete entry corresponding to $f_{y_k}(i)$ in $\tilde{\mathbf{Y}}$, $\tilde{\mathbf{X}}$, and $\tilde{\mathbf{\Sigma}}$. If $3K < t < T$, since we don't have y_{K+1}, delete entry corresponding to $f_{y_K}(i)$ in $\tilde{\mathbf{Y}}$, $\tilde{\mathbf{X}}$, and $\tilde{\mathbf{\Sigma}}$. If $t = T$, since we don't have $\mathbf{F}_{T+1}$ and y_{K+1}, delete entries corresponding to $\mathbf{F}_{T+1}$and $f_{y_{K+1}}(i)$ in $\tilde{\mathbf{Y}}$, $\tilde{\mathbf{X}}$, and $\tilde{\mathbf{\Sigma}}$. Also we need to change entries corresponding to $\mathbf{x}_{i \in v_{q,t}}$ in $\tilde{\mathbf{Y}}$, $\tilde{\mathbf{X}}$, and $\tilde{\mathbf{\Sigma}}$, for $t = T_q + 1, ..., T$.

C.9 Sampling the binary indicator z_j

The full conditional of z_j for $j = 2, ..., R$ is of the form

$$\begin{aligned} & P(z_j|\cdot) \\ \propto & \prod_{t=1}^{T_q} P(\mathbf{x}_t|\boldsymbol{\mu}, \mathbf{Z}, \mathbf{F}_t, \mathbf{\Theta}, \mathbf{\Omega}) \prod_{t=T_q+1}^{T} P(\mathbf{x}_{i \in v_{q,t}}|\boldsymbol{\mu}, \mathbf{Z}, \mathbf{F}_t, \mathbf{\Theta}, \mathbf{\Omega}) \\ \times & \prod_{t=2}^{K} P(y_k|\mathbf{Z}, \tilde{\mathbf{F}}_{3k}, \boldsymbol{\beta}, \eta^2) P(z_j|p_j), \end{aligned}$$

Notice that z_j only takes value of 0 or 1, so let

$$\frac{P(z_{j=1}|\cdot)}{P(z_{j=0}|\cdot)} = \gamma_1 \times \gamma_2 \times \gamma_3$$

where

$$\begin{aligned} \gamma_1 &= \gamma_1^b \gamma_1^u, \\ \gamma_2 &= \frac{exp\{-\frac{1}{2\eta^2}\sum_{k=2}^{K}(y_k - \beta_0 - \boldsymbol{\beta}_1'\mathbf{Z}_1\mathbf{F}_{3k} - \boldsymbol{\beta}_2'\mathbf{Z}_1\mathbf{F}_{3k-1} - \boldsymbol{\beta}_3'\mathbf{Z}_1\mathbf{F}_{3k-2} - \beta_4 y_{k-1})^2\}}{exp\{-\frac{1}{2\eta^2}\sum_{k=2}^{K}(y_k - \beta_0 - \boldsymbol{\beta}_1'\mathbf{Z}_0\mathbf{F}_{3k} - \boldsymbol{\beta}_2'\mathbf{Z}_0\mathbf{F}_{3k-1} - \boldsymbol{\beta}_3'\mathbf{Z}_0\mathbf{F}_{3k-2} - \beta_4 y_{k-1})^2\}}, \\ \gamma_3 &= \frac{p_j}{1 - p_j}, \end{aligned}$$

and

$$\gamma_1^b = \frac{exp\{-\frac{1}{2}\sum_{t=1}^{T_q}(\mathbf{x}_t-\boldsymbol{\mu}-\boldsymbol{\Theta}\mathbf{Z}_1\mathbf{F}_t)'\boldsymbol{\Omega}^{-1}(\mathbf{x}_t-\boldsymbol{\mu}-\boldsymbol{\Theta}\mathbf{Z}_1\mathbf{F}_t)\}}{exp\{-\frac{1}{2}\sum_{t=1}^{T_q}(\mathbf{x}_t-\boldsymbol{\mu}-\boldsymbol{\Theta}\mathbf{Z}_0\mathbf{F}_t)'\boldsymbol{\Omega}^{-1}(\mathbf{x}_t-\boldsymbol{\mu}-\boldsymbol{\Theta}\mathbf{Z}_0\mathbf{F}_t)\}},$$

$$\gamma_1^u = \frac{exp\{-\frac{1}{2}\sum_{t=T_q+1}^{T}(\mathbf{1}_{v_{q,t}}(\mathbf{x}_t-\boldsymbol{\mu}-\boldsymbol{\Theta}\mathbf{Z}_1\mathbf{F}_t))'(\mathbf{1}_{v_{q,t}}\boldsymbol{\Omega}\mathbf{1}'_{v_{q,t}})^{-1}(\mathbf{1}_{v_{q,t}}(\mathbf{x}_t-\boldsymbol{\mu}-\boldsymbol{\Theta}\mathbf{Z}_1\mathbf{F}_t))\}}{exp\{-\frac{1}{2}\sum_{t=T_q+1}^{T}(\mathbf{1}_{v_{q,t}}(\mathbf{x}_t-\boldsymbol{\mu}-\boldsymbol{\Theta}\mathbf{Z}_0\mathbf{F}_t))'(\mathbf{1}_{v_{q,t}}\boldsymbol{\Omega}\mathbf{1}'_{v_{q,t}})^{-1}(\mathbf{1}_{v_{q,t}}(\mathbf{x}_t-\boldsymbol{\mu}-\boldsymbol{\Theta}\mathbf{Z}_0\mathbf{F}_t))\}}.$$

where $\mathbf{Z}_1$ is evaluated at $z_j = 1$, $\mathbf{Z}_0$ is evaluated at $z_j = 0$. Then it follows that

$$z_j|\cdot \sim \ Bernoulli(p_j^*) \tag{C.9}$$

where $p_j^* = \frac{\gamma}{1+\gamma}$, and $\gamma = \gamma_1 \times \gamma_2 \times \gamma_3$.

C.10 Sampling the binary probability p_j

The full conditional of p_j for $j = 2, ..., R$ is of the form

$$\begin{aligned} p_j|z_j = 1 &\sim Beta(\alpha_p + 1, \beta_p) \\ p_j|z_j = 0, \pi &\sim (1-\pi^*)\mathbf{1}_{(p_j=0)} + \pi^* Beta(\alpha_p, \beta_p + 1) \end{aligned} \tag{C.10}$$

where $\pi^* = \frac{\pi^j\beta_p}{\alpha_p+\beta_p-\pi^j\alpha_p}$.

C.11 Sampling the zero inflation probablity π

The full conditional of π is given by

$$\begin{aligned} P(\pi|\cdot) &\propto \prod_{j:p_j=0}(1-\pi^j)\prod_{j:p_j\neq 0}\pi^j\frac{p_j^{\alpha_p-1}(1-p_j)^{\beta_p-1}}{\mathbf{B}(\alpha_p,\beta_p)} \times \frac{\pi^{\alpha_\pi-1}(1-\pi)^{\beta_\pi-1}}{\mathbf{B}(\alpha_\pi,\beta_\pi)} \\ &\propto \prod_{j:p_j=0}(1-\pi^j) \times [\pi^{\alpha_\pi+\sum_{j=2}^{R} j\mathbf{1}_{\{p_j\neq 0\}}-1}(1-\pi)^{\beta_\pi-1}] \end{aligned}$$

which is not in closed form. Note that the last two pieces to yields π has a Beta distribution $Beta(\alpha_\pi + \sum_{j=2}^{R} j\mathbf{1}_{\{p_j\neq 0\}}, \beta_\pi)$. Therefore, we can propose π^* using this Beta distribution, and use the first piece in the full conditional, denote as $q(\pi)$, to construct the acceptance-rejection rate. That is the proposal π^* is accepted with probablity

$$A(\pi^*|\pi^c) = min\Big\{1, \frac{q(\pi^*)}{q(\pi^c)}\Big\},$$

where π^c denotes the current state of π.

APPENDIX D. CHAPTER 4 APPENDIX

In this appendix, we present the full conditional posterior distributions for SV components and the Particle Gibbs with backward simulation algorithm that is used to generate posterior samples for the volatility parameters $\{\boldsymbol{\omega}_t = (\omega_{1t}, ..., \omega_{nt})' : t = 1, ..., T\}$. The full conditional posterior distributions for other parameters can be easily obtained by simpliy replace $\boldsymbol{\Omega}$ with $\boldsymbol{\Omega}_t = diag(\omega_{1t}, ..., \omega_{nt})\mathbf{P}diag(\omega_{1t}, ..., \omega_{nt})$ in the Appendix C.

D.1 Sampling the correlation matrix $\boldsymbol{\Psi}$

The posterior of $\boldsymbol{\Psi}$ is

$$p(\boldsymbol{\Psi}|\cdot) \propto \prod_{t=1}^{T_q} p(\mathbf{x}_t|\boldsymbol{\mu}, \boldsymbol{\Theta}, \mathbf{Z}, \mathbf{F}_t, \boldsymbol{\omega}_t, \boldsymbol{\Psi}) \prod_{t=T_q+1}^{T} p(\mathbf{x}_{i\in v_{q,t}}|\boldsymbol{\mu}, \boldsymbol{\Theta}, \mathbf{Z}, \mathbf{F}_t, \boldsymbol{\omega}_t, \boldsymbol{\Psi}). \tag{D.1}$$

which is not in closed form, so $\boldsymbol{\Psi}$ can be sampled from its conditional posterior using a random walk Metropolis step. That is sampling $n(n-1)/2$ elements from $N(upper\ tri(\boldsymbol{\Psi}^c), \boldsymbol{\Sigma}_\psi)$ truncated at $[-1, 1]$ and organizing into $n \times n$ matrix $\boldsymbol{\Psi}^*$. Accept $\boldsymbol{\Psi}^*$ with probability

$$A(\boldsymbol{\Psi}^*|\boldsymbol{\Psi}^c) = min\Big(1, \frac{\prod_{t=1}^{T_q} f(\boldsymbol{\Psi}^*|\boldsymbol{\mu}, \mathbf{x}_t, \boldsymbol{\Theta}, \mathbf{Z}, \mathbf{F}_t, \omega_t) \prod_{t=T_q+1}^{T} f(\boldsymbol{\Psi}^*|\boldsymbol{\mu}, \mathbf{x}_{i\in v_{q,t}}, \boldsymbol{\Theta}, \mathbf{Z}, \mathbf{F}_t, \omega_t)}{\prod_{t=1}^{T_q} f(\boldsymbol{\Psi}^c|\boldsymbol{\mu}, \mathbf{x}_t, \boldsymbol{\Theta}, \mathbf{Z}, \mathbf{F}_t, \omega_t) \prod_{t=T_q+1}^{T} f(\boldsymbol{\Psi}^c|\boldsymbol{\mu}, \mathbf{x}_{i\in v_{q,t}}, \boldsymbol{\Theta}, \mathbf{Z}, \mathbf{F}_t, \omega_t)}\Big)$$

where $\boldsymbol{\Psi}^c$ means the current state. $\boldsymbol{\Sigma}_\psi$ is a tuning parameter that can be updated using the first B iterations of the chain and then held fixed.

D.2 Sampling the variance in volatility equation τ^2

The full conditional of the variance in volatility equation, τ^2, is a inverse Gamma distribution:

$$\tau^2|\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_t \sim iG(\alpha_\tau, \beta_\tau), \tag{D.2}$$

where $\alpha_\tau = n\alpha_l + n(t-1)/2$ and $\beta_\tau = n\beta_l + \sum_{i=1}^{n}\sum_{t=2}^{T}(\omega_{it} - \omega_{i,t-1})^2/2$.

D.3 Sampling the volatility parameters $\boldsymbol{\omega}_t$ using Particle Gibbs with backward simulation

The Particle Gibbs with backward simulation algorithm used to sample the volatility parameters details as follows. Let $g = 1, ..., G$ represent the MCMC iteration, $p = 1, ..., P$ represent the particle index. Assume we are at iteration $g+1$, we have $(\boldsymbol{\omega}_t^p)^{(g)}$ for all $p \in \{1, ..., P\}$ and $t = 1, ..., T$, and backward trajectory indices $(B_t)^{(g)}$ for all $t = 1, ..., T$. $\boldsymbol{\Phi}^{(g)}$ denotes all other parameters and latent variables in the pervious iteration.

1. For $t = 1$

   (a) Draw $(\boldsymbol{\omega}_1^p)^{(g+1)}$ from proposal distribution, $g(\boldsymbol{\omega}_1)$ for $p \in \{1, ..., P\}\backslash B_1^{(g)}$. For $p = B_1^{(g)}$, set $(\boldsymbol{\omega}_1^{B_1^{(g)}})^{(g+1)}$=$(\boldsymbol{\omega}_1^{B_1^{(g)}})^{(g)}$.

   (b) For $p \in \{1, ..., P\}$, calculate weights $\tilde{w}_1^p = \frac{f(\mathbf{x}_1|\boldsymbol{\omega}_1^{(g)}, \boldsymbol{\Psi}^{(g)}) f(\boldsymbol{\omega}_1^{(g)})}{g(\boldsymbol{\omega}_1^{(g)})}$.

   (c) Normalize the weights, for all $p = 1, ..., P$, $w_1^p = \tilde{w}_1^p / \sum_{p=1}^{P} \tilde{w}_1^p$.

2. For $t = 2, ..., T$

   (a) For $p \in \{1, ..., P\}\backslash B_t^{(g)}$, draw j_t^p with $P(j_t^p = i) \propto w_{t-1}^i f(\mathbf{x}_t|(\hat{\boldsymbol{\omega}}_t^i)^{(g+1)}, \boldsymbol{\Psi}^{(g)})$ where $(\hat{\boldsymbol{\omega}}_t^i)^{(g+1)} = E(\boldsymbol{\omega}_t|\boldsymbol{\omega}_{t-1} = (\boldsymbol{\omega}_{t-1}^i)^{(g+1)})$. If $t = T_q + 1, ..., T$, use $\mathbf{x}_{i \in v_{q,t}}$.

   (b) Draw $(\boldsymbol{\omega}_t^p)^{(g+1)}$ from $f(\boldsymbol{\omega}|\boldsymbol{\omega}_{t-1} = (\boldsymbol{\omega}_{t-1}^{j_t^p})^{(g+1)}, (\tau^2)^{(g)})$ for $p \in \{1, ..., P\}\backslash B_t^{(g)}$. For $p = B_t^{(g)}$, set $(\boldsymbol{\omega}_t^{B_t^{(g)}})^{(g+1)}$=$(\boldsymbol{\omega}_t^{B_t^{(g)}})^{(g)}$.

   (c) Draw $j_t^{B_t^{(g)}}$ with $P(j_t^{B_t^{(g)}} = i) \propto w_{t-1}^i f((\boldsymbol{\omega}^{B_t^{(g)}})^{(g+1)}|\boldsymbol{\omega}_{t-1} = (\boldsymbol{\omega}_{t-1}^i)^{(g+1)}, (\tau^2)^{(g)})$.

   (d) For $p \in \{1, ..., P\}$, calculate weights $\tilde{w}_t^p = \frac{f(\mathbf{x}_t|(\boldsymbol{\omega}_t^p)^{(g+1)}, \boldsymbol{\Psi}^{(g)})}{f(\mathbf{x}_t|(\boldsymbol{\omega}_t^{j_t^p})^{(g+1)}, \boldsymbol{\Psi}^{(g)})}$.

   (e) Normalize the weights,for all $p = 1, ..., P$, $w_t^p = \tilde{w}_t^p / \sum_{p=1}^{P} \tilde{w}_t^p$.

3. Perform backward simulation:

   (a) Sample $B_T^{(g+1)}$ with $P(B_T^{(g+1)} = p) = (w_T^p)^{(g+1)}$, set $(\boldsymbol{\omega}_T)^{(g+1)} = (\boldsymbol{\omega}_T^{B_T^{(g+1)}})^{(g+1)}$.

(b) For $t = T-1, ..., 1$:

i. Calculate $(w^p_{t|t+1})^{(g+1)} \propto (w^p_t)^{(g+1)} f((\boldsymbol{\omega}_{t+1})^{(g+1)}|(\boldsymbol{\omega}^p_t)^{(g+1)}, \tau^{(g)})$ for $p = 1, ..., P$.

ii. Draw $B_t^{(g+1)}$ with $P(B_t^{(g+1)} = p) = (w^p_{t|t+1})^{(g+1)}$, set $(\boldsymbol{\omega}_t)^{(g+1)} = (\boldsymbol{\omega}_t^{B_t^{(g+1)}})^{(g+1)}$.

4. Collect $\{(\boldsymbol{\omega}_1)^{(g+1)}, ..., (\boldsymbol{\omega}_T)^{(g+1)}\}$ and $\{B_1^{(g+1)}, ..., B_T^{(g+1)}\}$.

The proposal distribution, g, from which to draw proposals for the first volatility parameter, $\boldsymbol{\omega}_1$ is chosen to be a multivariate normal distribution where the mean is the sample variance of the residuals corresponding to a PCA on the balanced panel using the first R PCs. The covariance matrix is taken to be the identity matrix. The specification of the proposal distribution should be problem-specific

D.4 A illustration of conditional auxiliary particle filter with backwards simulation

In this section, we illustrate a run of the conditional auxiliary particle filter with backwards simulation for $t = 6$ and $P = 5$.

A conditional auxiliary particle filter step involves prespecifying a trajectory $\boldsymbol{\omega}_1^{B_1}$,..., $\boldsymbol{\omega}_T^{B_T}$ with backward trajectory indices $B_1, ..., B_T$ which is guaranteed to survive the sequential Monte Carlo (SMC) step. Suppose, here, that the conditioning particle can be identified by the trajectory $B_1 = 3, B_2 = 2, B_3 = 4, B_4 = 4, B_5 = 4, B_6 = 2$, i.e. trajectory $\boldsymbol{\omega}_1^3$, $\boldsymbol{\omega}_2^2$, $\boldsymbol{\omega}_3^4$, $\boldsymbol{\omega}_4^4$, $\boldsymbol{\omega}_5^4$, $\boldsymbol{\omega}_6^2$ is prespecified and is guaranteed to survive the particle filter SMC step. At $t = 1$, we draw $P - 1 = 4$ (i.e. for $p = 1, 2, 4, 5$) proposals from prespecified proposal distribution g for the first volatility parameter $\boldsymbol{\omega}_1$. $\boldsymbol{\omega}_1^3$ is prespecified (i.e. take value from previous iteration in MCMC). Next we assign importance weight $\{w_1^1, ..., w_1^5\}$ to all particles at $t = 1$ in an important sampling faction and thus we have a set of particles and corresponding weights $\{\boldsymbol{\omega}_1^p; w_1^p\}_{p=1}^P$. This is called a weighted particle system. Given a weighted particle system $\{\boldsymbol{\omega}^p_{1:t-1}; w^p_{t-1}\}^P_{p=1}$ at $t-1$, we first sample $P-1$ parent index j_t^p, and a index $j_t^{B_t}$ corresponding to the prespecified partile $\boldsymbol{\omega}_t^{B_t}$. Particles at time t

are then generated given all parents. Weights are then updated to account for the mismatch between the likelihood at the proposals and the prespecified partile. Thus, we obtain $\{\boldsymbol{\omega}_{1:T}^p; w_T^p\}_{p=1}^P$ at the end of the auxiliary particle filter.

In Table D.1, black arrows represent the corresponding ancestral lineage for each w_T^p. One would sample $\{\boldsymbol{\omega}_6^p : p = 1, ..., 5\}$ based on calculated weights and would only have 5 unique trajectories available for sampling at the end of a plain conditional auxiliary particle filter. But at $t = 1$ and $t = 2$ due to degeneracy, there is only 1 possibility to be sampled. The backwards simulator works through the particles from $t = T$ to $t = 1$ and provides indices corresponding to a backward trajectory. The backwards simulator in Table D.1 happened to sample one element from the conditioning particle at $t = 5$.

Table D.1: Graphical representation of a run of the conditional auxiliary particle filter with backwards simulation when $t = 6$ and $P = 5$. The left figure represents all ancestral trajectories with grey lines. In the right figure, the black lines represent the ancestral lineage for $\{\boldsymbol{\omega}_6^p : p = 1, ..., 5\}$. The blue path represents a sample that could be taken by the backwards simulator.