

ABSTRACT

NING, BO. Bayesian Analysis of Dynamic Times Series and High-dimensional Models with Their Applications. (Under the direction of Peter Bloomfield and Subhashis Ghoshal.)

This dissertation uses time series and high-dimensional models to analyze large datasets which contain time series or high-dimension information. Bayesian methods are used for conducting analyses due to their flexibility in incorporating such information through prior distributions. The goal of this dissertation is to solve three major challenges tied to the two different datasets the author analyzed. The first challenge is to detect causality given the data that have a low signal-to-noise ratio. The second is to derive asymptotic results for the multivariate sparse linear regression when the covariance is unknown and its dimension is large. The third is to analyze time series datasets containing extreme values, the major problem is to make an inference for a highly nonlinear and non-Gaussian state-space model (SSM). To tackle the first challenge, a novel Bayesian method is proposed to detect causality through latent variables. The method is to compare two posterior distributions of a latent variable—one obtained by using the observed data from the observed data and the other one obtained by using the data from their counterfactual potential outcomes. To solve the second challenge, a test is constructed to apply general theory by bounding moments of likelihood ratio statistics around points in the alternative in order to derive the posterior contraction rate for the multivariate sparse linear regression with an unknown covariance. Also, a semi-parametric Bernstein-von Mises theorem is used to quantify the uncertainty for the regression coefficients with frequentist validity. For the third challenge, a new dynamic generalized extreme value (GEV) model is proposed. This model has Gumbel marginal distributions linked together with a Gaussian copula with order-one auto-regression (AR(1)) dependence. This model is a highly nonlinear SSM. A particle Gibbs with ancestor sampling (PGAS) algorithm is introduced to sample draws of the parameters in the model. The PGAS algorithm does not have degenerate issues and produces much more draws that are accepted during the sampling process compares to other sampling based methods.

© Copyright 2018 by Bo Ning

All Rights Reserved

Bayesian Analysis of Dynamic Times Series and High-dimensional Models with Their
Applications

by
Bo Ning

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2018

APPROVED BY:

Peter Bloomfield
Co-chair of Advisory Committee

Subhashis Ghoshal
Co-chair of Advisory Committee

Eric Chi

David Dickey

Sujit Ghosh

DEDICATION

To my parents.

BIOGRAPHY

Please allow me to use first person as the story I am going to tell is a story about my life.

For the past five years, I have devoted myself to conducting research on Bayesian analysis. Soon this journey will come to the end, and I am ready to encounter a new chapter in my life that is to continue doing research as a postdoctoral researcher in the Department of Statistics and Data Science at Yale University. However, an academic career was not always on the horizon for me; in fact, it's surprising that I am even pursued an advance degree.

When I was in the middle school, I certainly did not appear to have the makings of a career in academic. My teachers constantly humiliated me for my ignorance. My Chinese teacher once berated me before the entire class, “Bo, you are an idiot! I hear that there is a kind of medicine that can cure stupidity. But you are exceptionally stupid—I don't think it will work for you.” I was so embarrassed by her crushing words that I have remembered them to this day.

Teachers were not the only ones who criticized me when I was young; I was bullied and laughed at by my classmates and my extended family as well. As a result, I became isolated and developed a sense of self-loathing. My grades became steadily worse across the three years of middle school. During my last year, I unsuccessfully attempted suicide. Shortly afterwards, I was involved in a car accident. At the time, I wish I had simply been killed. Luckily, I survived.

Even after having survived two life-threatening events, I noticed that no one—except my parents—seemed really care about me. My Chinese teacher actually pretended she did not know that I had those experiences. I finally began to lose faith on trusting people.

However, not everything in my middle school life was negative. My parents, who genuinely cared for me and who did not require me to have perfect scores on my exams, were wonderful. They constantly encouraged me to read books. I was especially fascinated with the book *A Short History of Nearly Everything* by Bill Bryson. From it, I learned a number of stimulating facts drawn from many sciences, including astronomy, geography and biology—subjects that were not being taught in school. These early lessons sparked my abiding interest in science. My parents had the good sense to fuel my interests by buying me chemical equipment so that I could conduct chemistry experiments at home.

Without their supports and unflagging encouragement, I do not think that I would have survived that difficult season. Their aid helped me to develop my nature curiosity about science, a curiosity that I have maintained; it still drives my research today.

Unfortunately, it took me a long time to recover from my middle school experiences, and I was neither a top student in my high school nor did I attend a prestigious college. Nevertheless, my curiosity carried me along the arduous road of my life and lead me to pursue graduate studies in the United States. I planned to pursue a Ph.D. in econometrics or microeconomics; unfortunately, due to I lacked the appropriate mathematics background, studies did not go well at the beginning. Eventually, however, curiosity and determination helped me to get up to speed.

It was Dr. Atsushi Inoue, my master advisor in economics, who first piqued my interest in Bayesian analysis. From there, Dr. Peter Bloomfield, my first Ph.D. advisor in statistics, brought me fully into the world of Bayesian analysis. Under his guidance, I applied Bayesian methods for dealing with time series to financial datasets. As I used these methods, my curious led me to question why some Bayesian methods did well and others did not. I was finally able to find answers to these questions when I met with Dr. Subhashis Ghoshal, who is an expert on Bayesian statistics. Dr. Ghoshal became my second advisor, as I turned my research focus on the theoretical study of Bayesian methods.

I should admit that it was tough to begin work in the theoretical statistics, as I barely understood the papers I was reading. However, once again curiosity and determination ultimately lead to breakthrough. I can still remember the joy that came when I finally and fully understand a theorem and its proof from one of his papers. After years of training, I finished my first theoretical work in Bayesian analysis. I view this as the major achievement of my doctoral studies.

Although I devoted most of my time to study theoretical properties of Bayesian methods, I want to be an applied statistician in the future. I want to provide theoretical sound statistical methods for solving the data analysis problems of other fields. I know that my curiosity will continue to drive me to solve problems and glean new knowledge from these fields, in particular, astronomy.

If I could travel back 15 years ago, I would tell myself to have greater faith in myself and to fight against injustice instead of putting up with it. The past, however, is irretrievable. From the standing point of today, I am thankful for the hard experiences

to some degree—they gave me the ability to understand the pain that many minorities endure; they taught me how large an impact a teacher can have (for good or bad) on students.

I am lucky that to have survived and to have, furthermore, the good fortune of being able to pursue scientific research. This, I am sad to say, is rare for many people who have had experiences like mine. Having survived, I feel that it is my responsibility to broadcast my success to these victims and to motivate them so that they do not give up on their lives. When I become a teacher in the future, I want to make sure that no student is ignored, dismissed, or cruelly criticized and that those students who want to become future scientists are given the best help that I can provide.

I would like to teach at the university level understandably, but I feel that it is my responsibility to teach pre-college students as well. I hope to be involved in some organizations, like “Teaching for Taiwan”, who make it their aim to teach minority and marginalized students and so ensure that quality education is available to all.

ACKNOWLEDGEMENTS

No one earns a doctoral degree without gathering a large number of people to thank. It is therefore no surprise that a lot of people have helped me. While it is not even possible to mention all those I should acknowledge, there are some people whom I absolutely must thank.

In the first place, I must thank my parents, to whom the dissertation is dedicated. They gave me the freedom and support to pursue things that interested me from an early age. They encouraged me to read scientific books and bought me equipment to conduct chemical experiments inflamed my nascent curiosity in science. This curiosity is the driving force behind my passion for academics.

I must also thank Dr. Atsushi Inoue, my academic advisor of the master's program in economics. He introduced me to Bayesian analysis and encouraged me to choose statistics as my Ph.D. concentration. Dr. Inoue continues to have an impact on my work through his example of incredible dedication and analytic rigor.

Dr. René Moore deserves many thanks for her mentoring me in pedagogy. Without her, I never would have won the Outstanding Teaching Assistant Awards from both the Graduate School and the Department of Statistics of NCSU. She gave me opportunities to design course materials and provided feedback on my teaching which greatly improved my teaching skills. Most importantly, Dr. Moore inspired me to pursue a career in teaching. She once told me that I would become a great teacher, and that encouragement carried me through the long process of learning how to motivate students to study (even the most abstract concepts of) statistics.

Of course, the lion's share of my thanks must go to my two academic advisors: Drs. Peter Bloomfield and Subhashis Ghoshal. Dr. Bloomfield was kind enough to see the value in my immature attempts at research. I am grateful to him for reaching down to help me make the transition from being a student to taking my first steps as a researcher. I am so fortunate to have had him as an advisor: He supported my research interests and gave me the opportunity to be an independent scholar.

I owe an inestimable debt of gratitude to Dr. Ghoshal. Ultimately, he is the reason that I decided to pursue career in academia. Moreover, I must thank him for giving shape to my research interests in their current form. Under his guidance, I moved from being an apprentice practitioner of Bayesian methods to being a discerning journeyman

who has a deep theoretical understanding of Bayesian analysis. Although Dr. Ghoshal is an incredibly knowledgeable person, he is also very humble and approachable. He was patient and kind to me, even when correcting my mistakes. I learned many things from him, not only about statistics but also about being a decent person. Dr. Ghoshal is the professor I hope to become.

I would be remiss in my thanks if I failed to mention the contributions of Drs. Sujit Ghosh, Angie Wolfgang, and Ryan Martin. Drs. Ghosh and Wolfgang introduced me to the joys of astrostatistics. In particular, I want to thank Dr. Ghosh for treating me as his own student and bringing into the Statistical and Applied Mathematical Sciences Institute (SAMSI) Astrostatistics program.

Dr. Martin deserves thanks for the enormous impact that he has made on my understanding about the philosophy of statistics. His kindness in offering me career advice I found it to be insightful was also appreciated.

This is turning into a long acknowledgement, but there are still several more people that I must recognize. I want to thank Drs. Eric Chi and David Dickey for joining my dissertation committee: Dr. Chi, I really appreciate your suggestions on how to preparing for applying academic jobs. Dr. Dickey, it is an honor to meet you in person. Even now, I can hardly believe that you are on my committee.

I want to thank Dr. Howard Bondell, who helped me with my funding issues; Dr. Emily Griffith, who helped me when I was applying my postdoctoral positions; Dr. Donald Martin, who offered me the opportunity to conduct my first research in statistics; and Mrs. Alison McCoy, who invited me to play piano in commencement and always brightened my visits to the office.

I would also like to thank my advisors at Maxpoint Interactive Inc., Mr. Jewell Thomas and Mr. Mark Lowe who accepted me as an intern for two summers and supported my ideas.

Last, I must thank the Taiwanese Student Association, who adopted me as a member and introduced me to many dear friends; my Japanese teacher, Mrs. Mari Mantooth, who devoted many long hours to help me pass the JLPT exam; Dr. Yen-Chi Chen, who continually challenges me to be a better statistician; and my writing mentor, Mr. James Lanier, who has helped me make dramatic improvements in my writing skills.

I am lucky to have all of these amazing people in my life. Thank you for teaching me how to stay humble and to keep a sense of humor.

TABLE OF CONTENTS

List of Tables	x
List of Figures	xi
Chapter 1 Introduction	1
1.1 Research challenges	2
1.2 Chapter outline	3
1.3 Contributions	5
Chapter 2 Measuring the advertising effect for a multivariate dynamic time series sales data	7
2.1 Introduction	7
2.2 Causal assumptions and causal estimands	12
2.3 Model and prior	15
2.3.1 Model	15
2.3.2 Prior	16
2.4 Posterior computation	18
2.4.1 Estimating the sparse regression parameter	19
2.4.2 Sampling the time-varying parameters	21
2.4.3 Sampling the stationarity constraint parameters	22
2.4.4 Sampling the covariance matrices of the residuals	22
2.5 A new method to infer causality	24
2.6 Simulation study	26
2.6.1 Data generation and Bayesian estimation	26
2.6.2 Performance of the commonly used causal inference method	29
2.6.3 Performance of the new method to infer causality	33
2.7 Model checking	36
2.7.1 Convergence diagnostic	36
2.7.2 Sensitivity analysis	38
2.7.3 Using a threshold by choosing a different percentile	40
2.8 Application to a real dataset	41
2.9 Conclusion and discussion	44
Chapter 3 Bayesian Linear Regression for Multivariate Responses Under Group Sparsity	46
3.1 Introduction	46
3.2 Notation.	49
3.3 Prior specifications	50
3.3.1 Prior for regression coefficients	50

3.3.2	Prior for the covariance matrix	52
3.4	Main results	52
3.4.1	Posterior contraction rate	52
3.4.2	Dimensionality and recovery	56
3.4.3	Distributional approximation	57
3.4.4	Selection	59
3.5	Proofs	60
3.6	Auxiliary results	78
Chapter 4 Bayesian inference for generalized extreme value distribution with Gaussian copula dependence		83
4.1	Introduction	83
4.2	The dGEV model.	86
4.3	Posterior computation	87
4.3.1	Sampling μ, ψ, ξ	88
4.3.2	Sampling σ^2	89
4.3.3	Sampling ϕ	89
4.3.4	Sampling $\beta_{1:T}$	89
4.3.5	Markov Chain Monte Carlo	93
4.4	Seasonal dGEV model	93
4.5	Illustrative simulation study	94
4.6	Real data study	98
4.6.1	Water flow data	99
4.6.2	S&P 500 datasets	101
4.7	Conclusion and discussion	104
References		106
Appendices		114
Appendix A	The Kalman filter and backward smoothing algorithm	115
Appendix B	Deriving the EMVS algorithm	117
B.1	E-step	117
B.2	M-step	118
Appendix C	Plots of MCMC outputs for the water flow data and the S&P 500 data	121

LIST OF TABLES

Table 2.1	Parameters of the posterior densities of Σ^{-1} , Σ_u^{-1} , Σ_v^{-1} , Σ_w^{-1}	23
Table 2.2	Posterior medians and 95% credible intervals of average causal impacts for simulated datasets estimated using the multivariate models with a stationary and a nonstationary local linear trend.	32
Table 2.3	Posterior medians and 95% credible intervals of average causal impacts for simulated datasets estimated using the univariate model.	33
Table 2.4	Results of the one-sided KS distances and thresholds obtained by applying the new method to detect causal impacts in Dataset 1, . . . , Dataset 5 using the multivariate model with a stationary local linear trend. We only present the results at the dates March 22, 2016, March 31, 2016 and April. 9, 2016 which correspond to the 1st day, 10th day and 20th day during the causal period.	34
Table 2.5	Posterior medians and 95% credible intervals of average causal impacts for the model (2.7.1).	38
Table 2.6	Results of the one-sided KS distances and thresholds obtained by applying the new method to detect causal impacts in Dataset 1, . . . , Dataset 5 using the model (2.7.1) with the stationarity constraint. We only present the results at the dates March 22, 2016, March 31, 2016 and April. 9, 2016 which correspond to the 1st day, 10th day and 20th day during the causal period.	39
Table 2.7	Number of test stores that received significant causal impacts for each week of running the advertisement campaign by using the multivariate model with a stationary local linear trend.	43
Table 2.8	Number of test stores that received significant causal impacts for each week of running the advertisement campaign by using the univariate model.	43
Table 4.1	Posterior medians and 95% credible intervals (C.I.s) for parameters μ , ψ , ξ , ϕ , σ	95
Table 4.2	Posterior median and 95% credible intervals (C.I.s) for parameters μ , ψ , ξ , ϕ , a_1 , a_2 , σ	97
Table 4.3	Posterior medians, 95% credible intervals and inefficient factors for parameters μ , ψ , ξ , ϕ , σ	101
Table 4.4	Posterior medians with their 95% credible intervals and inefficient factor for parameters μ , ψ , ξ , ϕ , σ in the dGEV model and parameters μ , ψ , ξ , ϕ , σ , a_1 , a_2 in the seasonal dGEV model.	104

LIST OF FIGURES

Figure 2.1	An example of test and control store locations in the State of Texas (Google Maps, 2017). The red dots represent the locations of the test stores; the blue dots represent the locations of the control stores. . . .	8
Figure 2.2	EMVS (left) and DAEMVS (with $s = 0.1$) (right) estimation of β based on the simulated datasets. The dark blue lines are the parameters that have simulated values 2; the light blue lines are the parameters that have simulated values 1 and the black lines are the parameters that have simulated values 0. The red lines are the calculated β_i^{th} values, within the two red lines, the parameters should be considered as zero parameters. . . .	28
Figure 2.3	DAEMVS (with $s = 0.1$) estimation of β based on the simulated datasets using the nonstationary model (left) and the misspecified model (right). The dark blue lines are the parameters that have simulated values 2; the light blue lines are the parameters that have simulated values 1 and the black lines are the parameters that have simulated values 0. The red lines are the calculated β_i^{th} values, within the two red lines, the parameters should be considered as zero.	30
Figure 2.4	Plot of the causal impact in Dataset 4 using models with a stationary and a nonstationary local linear trend. (a) and (c) are the plots of estimation (before March 21, 2016) and prediction (after March 21, 2016) of Dataset 4 without stationarity constraint (left) and with stationarity constraint (right). The gray line is the simulated dataset, the blue line is the estimated posterior median of the dataset using the model, the dashed blue line is the corresponding 95% credible and prediction intervals. (b) and (d) are the plots of estimated causal impact by taking the difference between the observed data and Bayesian estimates using the model with a nonstationary local linear trend (left) and the model with a stationary local linear trend (right). The black line is the simulated true impact, the blue line is the estimated median of the impact, the dashed blue lines are the corresponding 95% credible and prediction intervals.	31
Figure 2.5	Results of applying the new method to detect causal impacts in Dataset 1, . . . , Dataset 5 using the multivariate model with a stationary local linear trend during the causal period from March, 22, 2016 to April, 9, 2016. In each subplot, the red line gives the one-sided KS distances between two posterior distributions with one is given the data of counterfactuals; the light blue line gives the corresponding thresholds.	35
Figure 2.6	Traceplots for μ_{11} , Σ_{11} , Σ_{u11} , Σ_{v11} , Σ_{w11} , Φ_{11}	37

Figure 2.7	Plot of the inefficient factors (IFs) for $\mu_{1:T}, \Sigma, \Sigma_u, \Sigma_v, \Sigma_w, D, \Phi$. The first 400 values are IFs for parameters in $\mu_{1:T}$ (5 datasets each with 80 time periods), the following 100 values are IFs for parameters in $\Sigma, \Sigma_u, \Sigma_v,$ and Σ_w , with each has 25 parameters, the next 5 values are IFs for parameters in D ; and the last 25 values are IFs for parameters in Φ . The red lines separate IFs result from different parameters.	37
Figure 2.8	Results of applying the new method for detecting causal impacts in Dataset 1, . . . , Dataset 5 with the data generated from model (2.7.1) during the causal period from March, 22, 2016 to April, 9, 2016. In each subplot, the red line gives the one-sided KS distances between two posterior distributions with one is given the observed data and the other given the data of counterfactuals; the light blue line gives the corresponding thresholds.	39
Figure 2.9	Results of applying the new method for detecting causal impacts in Dataset 1, . . . , Dataset 5 during the causal period from March, 22, 2016 to April, 9, 2016 with thresholds chosen as the 99% upper percentile from the one-sided KS distances obtained from 30 generated counterfactuals. In each subplot, the red line gives the one-sided KS distances between two posterior distributions with one is given the observed data and the other is given their corresponding counterfactuals; the light blue line gives the corresponding thresholds.	40
Figure 2.10	Plot of the causal impacts at test stores at end of the second week (a), the fifth week (b) and the last week (c) for an advertising campaign of a consumer product at a large national retail chain. The impacts below their thresholds are set to zero. The United States map is produced using Google Maps, 2017.	42
Figure 4.1	Plot of MCMC draws (1st row), autocorrelation function (ACF) (2nd row) and histogram with density (3rd row) for $\mu, \psi, \xi, \phi, \sigma$. The red line indicates the true values.	96
Figure 4.2	Plot of posterior median (blue line) and 95% credible intervals (grey dash lines) for $\beta_1, \dots, \beta_{1000}$; the yellow dashed lines indicates our simulated value for β s; the four dash red lines from up to down represent the 0.995, 0.975, 0.025, 0.005 quantiles of the standard normal distribution.	97
Figure 4.3	Plot of MCMC draws (1st row), autocorrelation function (ACF) (2nd row) and histogram with density (3rd row) for $\mu, \psi, \xi, \phi, a_1, a_2$. The red line indicates the true values.	98

Figure 4.4	Plot of posterior median (blue line) and 95% credible intervals (grey dash lines) for $\beta_1, \dots, \beta_{1000}$; the yellow dashed lines indicates our simulated value for β s; the four dash red lines from up to down represent the 0.995, 0.975, 0.025, .005 quantiles of the standard normal distribution.	99
Figure 4.5	Annual maximum water flow of French Broad River at Asheville, North Carolina.	100
Figure 4.6	Plot of posterior medians (blue line) and their 95% credible intervals (grey dash lines) for β_{t} s by fitting dGEV into water flow dataset. The four dash red lines from up to down represent the 0.995, 0.975, 0.025, 0.005 quantiles of the standard normal distribution.	100
Figure 4.7	Weekly minimum S&P 500 log-return dataset, adjust the dataset to be -1	102
Figure 4.8	(a) Plot of posterior medians (blue line) and their 95% credible intervals (grey dash lines) for β_{t} s by fitting the dGEV model into S&P 500 dataset. (b) Plot of posterior medians (blue line) and their 95% credible intervals (grey dash lines) for β_{t} s by fitting the seasonal dGEV model into S&P 500 dataset. The four dash red lines from up to down represent the 0.995, 0.975, 0.025, 0.005 quantiles of the standard normal distribution. dataset.	103
Figure C.1	Plot of MCMC draws (1st row), autocorrelation functions (ACF) (2nd row) and histograms along with densities (yellow lines) (3rd row) for parameters $\mu, \psi, \xi, \phi, \sigma^2$ in dGEV model of the water flow dataset.	121
Figure C.2	Plot of MCMC draws (1st row), autocorrelation functions (ACF) (2nd row) and histograms along with densities (yellow line) (3rd row) for parameters $\mu, \psi, \xi, \phi, \sigma^2$ in the dGEV model by fitting the S&P 500 dataset.	122
Figure C.3	Plot of MCMC draws (1st row), autocorrelation function (ACF) (2nd row) and histograms along with densities (yellow line) (3rd row) of parameters $\mu, \psi, \xi, \phi, a_1, a_2$ in a seasonal dGEV model by fitting the S&P 500 dataset.	123

Chapter 1

Introduction

In the era of big data, analyzing large datasets is challenging not only because computation is a time-consuming process, but also because some data structures are very complex. Bayesian approaches to inference especially for big datasets are attractive because it is flexible enough to handle complex data objects by incorporating the structural information through prior distributions. By using Markov chain Monte Carlo (MCMC) algorithms, which are sampling-based Bayesian methods, one can obtain sample draws of the desired distributions for parameters in the model. As a result, the posterior mean and credible intervals can be obtained with no additional cost. Bayesian methods also possess good frequentist properties. For example, in high-dimensional analysis, which is used to deal with the situation when the number of regression coefficient exceeds the number of observations (known as the $p > n$ problem), it has been shown that by choosing appropriate priors, such as in [Castillo and van der Vaart \(2012\)](#) and [Martin et al. \(2017\)](#), the resulting posteriors contract at the minimax (or near minimax) rate.

However, many challenges also present in Bayesian analysis. For instance, although Bayesian methods for the state-space model (SSM) have been developed, they are not suitable for dealing with highly non-linear SSMs (i.e., the dynamic generalized extreme value model in the later chapter). Thus, new Bayesian inference techniques are needed for computing those models. Another example is to study the theoretical properties of a multivariate sparse linear regression model, which is often used in many real data analyses. Due to the techniques used to derive the theoretical results of a univariate model cannot be applied to derive the theoretical properties of a multivariate model, there is a need to develop a new technique to derive those theoretical results. In this

dissertation, the author develops new Bayesian methods and techniques to tackle three challenges in time series and high-dimensional models, which are applied to analyze large datasets. The three challenges will be described in the next section.

1.1 Research challenges

Of the three major challenges this dissertation addresses, the first challenge emerged in a study conducted for *MaxPoint Interactive Inc. (MaxPoint)*, which is an online advertising company located in Morrisville, North Carolina. The goal of this study is to measure the sales lift caused by a running an advertising campaign, which is a causal inference problem. A commonly used Bayesian approach (Rubin, 2005; Brodersen et al., 2015) to estimate causal impacts is to take differences between the observed sales data and the estimated sales data obtained from counterfactual (also known as potential outcomes). The counterfactual is constructed from several control stores that the advertising campaign did not run on. The challenge is to detect causality from the dataset when its signal-to-noise ratio is so low that if using the approach described above, it often gives a misleading result that there does not have any causal impacts.

The second challenge emerged during efforts to provide a theoretical justification for using the model developed in the first study: to derive the asymptotic properties of a multivariate sparse linear regression model where the covariance matrix is also unknown and its dimension is large. To derive those asymptotic properties is a challenging task because the previous techniques (for example, Castillo et al. (2015)) used to study the univariate sparse linear regression models cannot be applied directly. This is due to the previous asymptotic results are derived directly from the corresponding posterior distribution—this is possible only when the variance is assumed to be known. However, the derivation does not go through for a multivariate model with a unknown covariance matrix. Thus, an alternative approach needs to be explored.

The third challenge is to develop a Bayesian inference algorithm for a new dynamic extreme value model, which is a highly nonlinear SSM. Bayesian inference for a linear Gaussian SSM is often carried out by using the Kalman filtering and backward smoothing (KFBS) algorithm. For a nonlinear SSM, a simple method is to apply the KFBS algorithm to an approximated linear SSM first, which is often obtained by finding a linear approximation for the nonlinear equations (for example, using a second order Taylor

expansion to the log-likelihood functions). Then use a Metropolis-Hasting (M-H) algorithm to keep useful draws (Shephard and Pitt, 1997; Durbin and Koopman, 2012; Niemi and West, 2010). An alternative approach is to use the sequential Monte Carlo (SMC) algorithm. This algorithm first generates samples (often called *particles*) from a chosen proposal density; and then calculates the weights between the proposal density and the true density. The generated particles and their weights give a discrete approximation to the true density. However, the difficulty occurs when the SSM is highly nonlinear, this makes it hard to find the approximated density or a proposal density so that the KFBS and the SMC algorithm can be applied easily. A poor approximating density or a proposal density will produce too many useless sample draws and make the computation time very long and the mixing rate of the chain very high.

To summarize, the three challenges are as follows:

1. How to detect causality when the data have a low signal-to-noise ratio?
2. How to derive asymptotic results for the multivariate sparse linear regression model with the covariance is unknown and its dimension is large?
3. How to make inference for the highly nonlinear and non-Gaussian SSM?

1.2 Chapter outline

Chapter 2 addresses the problem of detecting causality when data have a low signal-to-noise ratio. A novel Bayesian method is proposed. The new method is to compare two posterior distributions of a latent variable—one obtained by using the observed data from the test stores and the other one obtained by using the data from their corresponding counterfactuals. Each counterfactual is estimated from the data of synthetic controls, each of which is a linear combination of sales figures at many control stores over the causal period. Control stores are selected using a revised Expectation-Maximization variable selection (EMVS) method. A multivariate structural time series model is used to capture the spatial correlation between stores by placing a \mathcal{G} -Wishart prior (Letac and Massam, 2007) on the precision matrix. A two-stage algorithm is proposed to estimate the parameters of the model. To prevent the prediction intervals from being explosive, a stationarity constraint is imposed on the local linear trend of the model through a recently proposed prior. The benefit of using this prior is discussed in this chapter. A

detailed simulation study shows the effectiveness of using our proposed method to detect weaker causal impact. The new method is applied to measure the causal effect of an advertising campaign for a consumer product sold at stores of a large national retail chain.

To give a theoretical justification for using the model developed in Chapter 2, Chapter 3 explores the asymptotic properties of a Bayesian high-dimensional multivariate linear regression model with correlated responses. Thus addressing the second research challenge. This model has two unique features: (i) the covariance matrix is unknown and its dimensions can be high. (ii) group sparsity is imposed on the predictors. Sparsity on individual coefficients is considered as a special case. A product of independent spike-and-slab priors is chosen for the regression coefficients, each of which is a mixture of a point mass at zero and a multivariate density involving a ℓ_2/ℓ_1 -norm (Yuan and Lin, 2006). A Wishart prior with increasing dimension is placed on the inverse of the covariance matrix. Four main results are obtained: First, the posterior contraction rate is derived. Second, the bounds on the effective dimension of the model is obtained with high posterior probabilities. Third, it shows that the multivariate regression coefficients can be recovered under certain compatibility conditions. Last, the uncertainty for the regression coefficients is quantified with frequentist validity through a Bernstein-von Mises type theorem. This result also leads to selection consistency for the Bayesian method. The posterior contraction rate is derived using the general theory through constructing a suitable test from the first principle by bounding moments of likelihood ratio statistics around points in the alternative. This leads to the posterior concentrates around the truth with respect to the average negative log-affinity.

In Chapter 4, a new dependent generalized extreme value (dGEV) model is constructed. This model incorporates a dependent Gumbel process into a GEV model through using a variable transformation technique, which combines the marginal cumulative distribution function (CDF) of a Gumbel distribution with the standard normal copula. The model can be written as a highly nonlinear state space model in which the hidden process is a dependent Gaussian autoregressive with order one (AR(1)) process. A particle Gibbs with ancestor sampling (PGAS) algorithm is used to sample the elements of the state vector. This algorithm turns out to be very efficient in solving highly nonlinear state space models. This model is also flexible enough to incorporate seasonality. A simulation study and two real data analysis—a water flow dataset and a S&P 500 dataset

are conducted.

1.3 Contributions

This dissertation makes a number of original contributions that can best be considered in terms of its three driving research challenges.

As for the first challenge, a novel Bayesian causal inference method to detect causality is proposed. This idea of detecting causality at a latent variable level is not restricted to the specific structural time series model used in this thesis and can be applied to many other models in different applications. Furthermore, the posterior distributions of the latent variable are obtained without any extra cost because of using the MCMC algorithm. The second contribution made in Chapter 2 is a revised EMVS (expectation-maximization variable selection) algorithm being used to select variables faster for time series datasets. The calculation of this algorithm is provided. This algorithm can be easily extended to other scenarios in dynamic time series analysis. The third contribution is to use a recently proposed prior for the first order vector autoregressive (VAR(1)) process of a SSM. The need to use this prior is due to the fact that other commonly used priors, including the conjugate-prior, failed to produce samples of draws of the coefficient matrix to meet this Schur-stable constraint (Roy et al., 2016), then the VAR(1) process can be nonstationary. The new method is able to solve this problem. A simulation study suggests the effectiveness of using this method.

Regarding the second challenge, in order to derive the posterior contraction rate for the multivariate sparse linear regression with unknown covariance, this dissertation constructs a test to apply the general theory of posterior contraction by bounding moments of likelihood ratio statistics around points in the alternative. This test can be applied to many other multivariate analyses when the mean and the covariance matrix are not known. This study also shows good recovery and selection consistency properties for the regression coefficients. Furthermore, this study quantifies the uncertainty for the regression coefficients with frequentist validity through a semi-parametric Bernstein-von Mises type theorem. The third contribution is that we study the posterior that group sparsity is imposed on the predictors. This setting is more general as it contain the special case that the sparsity is put on individual coefficients.

Finally, concerning the third challenge, this dissertation proposes a new dynamic

GEV model by incorporating the dependence between extreme events through a dependent Gumbel distribution and then matching its distribution with a same type dependence standard normal copula. The advantage is that the marginal distribution of each parameter is sampled from an exact GEV distribution. Furthermore, a particle Gibbs with an ancestor sampling (PGAS) algorithm is used to sample the hidden process of the model. Unlike the KFBS and SMC algorithms, the PGAS algorithm produces more useful draws. This is because this algorithm first finds a pre-specified ancestor lineage or trajectory path that is used to guide draws from the invariant unconditional distribution, and then allows the pre-defined ancestor lineage to update during the forward sampling process, and drop some lineage with degenerate weights. As a result, the degenerate issue is resolved and a large amount of draws is accepted.

Chapter 2

Measuring the advertising effect for a multivariate dynamic time series sales data

2.1 Introduction

MaxPoint is interested in measuring the sales increases associated with running advertising campaigns for products distributed through brick-and-mortar retail stores¹. The dataset was obtained as follows: *MaxPoint* ran an advertising campaign at 627 test stores across the United States. An additional 318 stores were chosen as control stores. Control stores were not targeted in the advertising campaign. The company collected weekly sales data from all of these stores for 36 weeks before the campaign began and for the 10 weeks in which the campaign was conducted. The time during which the campaign was conducted is known. The test stores and the control stores were randomly selected from different economic regions across the U.S.. Figure 2.1 shows an example of the locations of stores in the State of Texas.²

To the best of our knowledge, the work of [Brodersen et al. \(2015\)](#) is the most related one to the present study. Their method can be described as follows. For each test store, they first split its time series data into two parts: before and during a causal impact (in our

¹The methodology developed and presented in this chapter is not connected to any commercial products currently sold by *MaxPoint*.

²Note: The locations of the stores shown in the figure are not associated with any real datasets collected by *MaxPoint*.

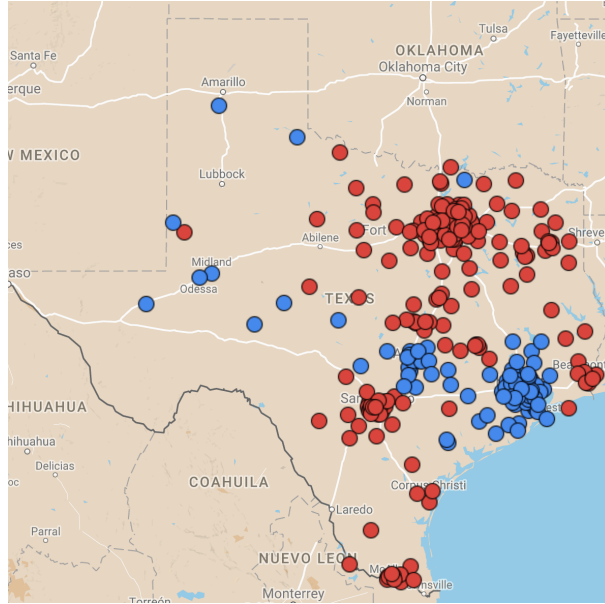


Figure 2.1: An example of test and control store locations in the State of Texas (Google Maps, 2017). The red dots represent the locations of the test stores; the blue dots represent the locations of the control stores.

case, the impact is the advertising campaign). Then they used the data collected before the impact to predict the values during the causal period. At the same time, they applied a stochastic search variable selection (SSVS) method to construct a synthetic control for that store. The counterfactual potential outcomes (Rubin, 2005) are the sum of the predicted values and the data from the synthetic control. Clearly, the potential outcomes of the store exposed to advertising were the observed data. Finally, they compared the difference between the two potential outcomes and took the average of differences across different time points. The averaged difference is a commonly used causal estimand that measures the temporal average treatment effect (Bojinov and Shephard, 2017).

The method proposed by Brodersen et al. (2015) is novel and attractive; however, it cannot directly apply to analyze our dataset due to the following three reasons: (1) Many causal impacts in our dataset are weak. The causal estimand that Brodersen et al. (2015) used often fails to detect them; (2) The test stores within an economic region are spatially correlated as they share similar demographic information. Using Brodersen et al. (2015)'s method would not allow to consider the spatial correlation between stores; (3) The SSVS method is computationally slow because it requires sampling from a large

model space consisting of 2^p possible combinations of p control stores. In the following, we will discuss our proposed method for addressing these three difficulties.

First, we propose a new method for detecting weaker causal impacts. The method compares two posterior distributions of the latent variables of the model, where one distribution is computed by conditioning the observed data and the other one is computed by conditioning the counterfactual potential outcomes. We use the one-sided Kolmogorov-Smirnov (KS) distance to quantify the distance between the two posterior distributions.

The new method can successfully detect weaker impacts because it compares two potential outcomes at the latent variable level; while the commonly used method compares them at the observation level. Since the observed data often contain “inconvenient” components—such as seasonality and random errors—which inflate the uncertainty of the estimated causal effect, the commonly used method may fail to detect weaker impacts. In the simulation study, we show that the new method outperforms the commonly used method even when the model is slightly misspecified.

The causal estimand in the new method is different from the one of the commonly used method. The former one measures the temporal average treatment effect using the KS distance between two posterior distributions and the latter measures that effect using the difference between two potential outcomes. Formal definitions of the two causal estimands are provided in Section 2.2.

Secondly, we use a multivariate version of a structural time series model (Harvey and Peters, 1990) to model the sales data of test stores by allowing pooling of information among those stores that locate in geographically contiguous economic regions. This model enjoys a few advantages that make it especially suitable for our causal inference framework. First, the model is flexible to adapt to different structures of the latent process. Secondly, it can be written as a linear Gaussian state-space model and exact posterior sampling methods can be carried out by applying the Kalman filter and simulation smoother algorithm proposed by Durbin and Koopman (2002, 2012). Thirdly, it is relatively easy to deal with missing data due to the use of the Kalman filter and backward smoothing (KFBS) algorithm. The imputing process can be naturally incorporated into the Markov chain Monte Carlo (MCMC) loop.

Since test stores are correlated, the number of parameters in the covariance matrix grows quadratically with the dimension. Consequently, there will not be enough data to estimate all these parameters. In our approach, we reduce the number of parameters by

imposing sparsity based on a given spatial structure (Smith and Fahrmeir, 2007; Barber and Drton, 2015; Li et al., 2015). We consider a graphical model structure for dependence based on geographical distances between stores. If the distance between two stores is very large, we treat them conditionally independent given other stores. In terms of a graphical model, this is equivalent to not put an edge between them. We denote the corresponding graph by \mathcal{G} . Note that \mathcal{G} is given in our setting and is completely determined by the chosen thresholding procedure. We use a graphical Wishart prior with respect to the given graph \mathcal{G} , in short a \mathcal{G} -Wishart prior (Roverato, 2002), to impose sparsity on the precision matrix. One advantage is that this prior is conjugate for a multivariate normal distribution. If \mathcal{G} is decomposable, sampling from a conjugate \mathcal{G} -Wishart posterior is relatively easy due to an available closed form expression for the normalizing constant in the density (Lauritzen, 1996; Roverato, 2000, 2002). However, if \mathcal{G} is non-decomposable, the normalizing constant does not usually have a simple closed form (see however; Uhler et al., 2017), and thus one cannot easily sample directly from its posterior. In such a situation, an approximation for the normalizing constant is commonly used (Atay-Kayis and Massam, 2005; Mitsakakis et al., 2011; Wang and Li, 2012; Khare et al., 2015). A recent method introduced by Mohammadi and Wit (2015) is a birth-death Markov chain Monte Carlo (BDMCMC) sampling method. It uses a trans-dimension MCMC algorithm that transforms sampling of a high-dimensional matrix to lower dimensional matrices, thus improving efficiency when working with large precision matrices.

In a multivariate state-space model, the time dynamics are described by a multivariate stochastic trend, usually an order-one vector autoregressive (VAR(1)) process (de Jong, 1991; de Jong and Chu-Chun-Lin, 1994; Koopman, 1997; Durbin and Koopman, 2002). To use a VARMA(p, q) (order p vector autoregression with order q moving average) with $p > 1, q \geq 0$ process is also possible and the choice of p, q can be made based on data (e.g., chosen by the Bayesian Information Criterion). However, the larger the p and q are, the larger the number of parameters that need to be estimated. For the sake of tractability, we treat the hidden process as a VAR(1) process throughout the chapter.

Putting stationarity constraints on the VAR(1) process is necessary to prevent the prediction intervals from becoming too wide to be useful. However, constructing an appropriate prior complying with the constraints is not straightforward. Gelfand et al. (1992) proposed a naive approach that puts a conjugate prior on the vector autoregressive parameter to generate samples and only keep the samples meeting the constraints.

However, it can be highly inefficient when many draws from the posterior correspond to nonstationary processes. A simple remedy is to project these nonstationary draws on the stationarity region to force them to meet the constraints (Gunn and Dunson, 2005). However, the projection method is somewhat unappealing from a Bayesian point of view because it would make the majority of the projected draws have eigenvalues lying on the boundary of the corresponding space (Galindo-Garre and Vermunt, 2006; Roy et al., 2016). We instead follow the recently proposed method of Roy et al. (2016) to decompose the matrix into several unrestricted parameters so that commonly used priors can be put on those parameters. While conjugacy will no longer be possible, efficient algorithms for drawing samples from the posterior distribution are available.

Thirdly, to accelerate the computational speed of selection control stores, we suggest using a revised version of the Expectation-Maximization variable selection (EMVS) method (Ročková and George, 2014). The model uses an Expectation-Maximization (EM) algorithm that is faster and does not need to search 2^p possible combinations.

It is worth mentioning that there are many other popular methods for constructing a synthetic control, such as the synthetic control method proposed by Abadie and Gardeazabal (2003), the difference-in-differences method (Abadie, 2005; Bonhomme and Sauder, 2011; Donald and Lang, 2007), and the matching method (Stuart, 2010). Moreover, Doudchenko and Imbens (2016) provided a nice discussion on the advantages and disadvantages of each method. Unlike these methods, there are two advantages of using our proposed method: 1) It does not need to have a prior knowledge about the relevant control stores, the process of selecting control stores is completely driven by data and can be easily incorporated into a Bayesian framework. 2) It provides a natural model-based causal inference by viewing counterfactual potential outcomes as missing values and generating predicting credible intervals from their posterior predictive distributions, and finally providing a quantitative measure for the strength of the causal effect (Rubin, 2005).

We apply our method on both simulated datasets and the real dataset provided by *MaxPoint*. In the simulation study, we compare the new method with the method proposed by Brodersen et al. (2015).

The rest of the chapter is organized as follows. Section 2.2 introduces causal assumptions and causal estimands. Section 2.3 describes the model and the priors. Section 2.4 describes posterior computation techniques. Section 2.5 introduces our proposed new ap-

proach to infer causal effects in times series models. Simulation studies are conducted in Section 2.6 and model diagnostics are performed in Section 2.7. In Section 2.8, the proposed method is applied on a real dataset from an advertising campaign conducted by *MaxPoint*. Finally, Section 2.9 concludes with a discussion.

2.2 Causal assumptions and causal estimands

This section includes three parts. First, we will introduce the potential outcomes framework. Secondly, we discuss three causal assumptions. Finally, we define two causal estimands with one of them is new.

The potential outcomes framework is widely used in causal inference literature (Rubin, 1974, 2005; Ding and Li, 2017). Potential outcomes are defined as the values of an outcome variable at a future point in time after treatment under two different treatment levels. Clearly, at most one of the potential outcomes for each unit can be observed, and the rest are missing (Holland, 1986; Rubin, 1977; Imbens and Rubin, 2015). The missing values can be predicted using statistical methods. In the study, we predict the values using the data from a synthetic control that is constructed from several control stores.

Based on the potential outcomes framework, we conduct the causal inference. There are three assumptions need to make for conducting the inference. They are,

1. The stable unit treatment value assumption (SUTVA);
2. The strong ignorability assumption on the assignment mechanism;
3. The trend stays stable in the absence of treatment for each test store.

The SUTVA contains two sub-assumptions: no interference between units and no different versions of a treatment (Rubin, 1974). The first assumption is reasonable because the stores did not interact with each other after the advertising was assigned. As Rosenbaum (2007) pointed out, “interference is distinct from statistical dependence produced by pretreatment clustering.” Since the spatial correlation between test stores is produced by pretreatment clustering, it is different from the interference between stores. The second assumption is also sensible because we assume that there are no multiple versions of the advertising campaign. For example, the advertising campaign is not launched across multiple channels.

The strong ignorability assumption also contains two parts: unconfoundedness and positivity (Ding and Li, 2017). Unconfoundedness means that the treatment is assigned randomly and positivity means that the probability for each store being assigned is positive. In our study, we assume the company randomly assigned advertising to stores and each store has an equal probability of being assigned.

The last assumption says that the counterfactual potential outcomes in the absence of the advertising in test stores are predictable.

Now, we shall introduce some notations before defining causal estimands. Let n be the total number of test stores to which the advertising were assigned. The i -th test store has p_i control stores (stores not assigned with the advertising), $i = 1, \dots, n$. The total number of control stores are denoted as p , $p = \sum_{i=1}^n p_i$. The length of the time series data is $T + P$. Let $1, \dots, T$ be the periods before running the advertising campaign and $T + 1, \dots, T + P$ be the periods during the campaign. Let $\mathbf{w}_t = (w_{1t}, \dots, w_{n+p,t})'$ be a vector of treatment at time $t = T + 1, \dots, T + P$, with each w_{it} being a binary variable. The treatment assignment is time-invariant, so $\mathbf{w}_t = \mathbf{w}$. For stores assigned with advertising, we denote the sales value for the i -th store at times t as y_{it} . Let y_{it}^{obs} be the observed data and y_{it}^{cf} be the counterfactual potential outcomes which are missing. We let $\mathbf{Y}_t^{\text{obs}} = (y_{1t}^{\text{obs}}, \dots, y_{nt}^{\text{obs}})'$ and $\mathbf{Y}_t^{\text{cf}} = (y_{1t}^{\text{cf}}, \dots, y_{nt}^{\text{cf}})'$ respectively be the observed and missing potential outcomes for n test stores at time t , $t = 1, \dots, T + P$. Clearly, $\mathbf{Y}_t^{\text{obs}} = \mathbf{Y}_t^{\text{cf}}$ when $t = 1, \dots, T$. We define $\mathbf{Y}_{T+1:T+P}^{\text{obs}} = (\mathbf{Y}_{T+1}^{\text{obs}}, \dots, \mathbf{Y}_{T+P}^{\text{obs}})'$ and $\mathbf{Y}_{T+1:T+P}^{\text{cf}} = (\mathbf{Y}_{T+1}^{\text{cf}}, \dots, \mathbf{Y}_{T+P}^{\text{cf}})'$.

We first define the causal estimand of a commonly used method. For the i -th test store, it is defined as

$$\frac{1}{P} \sum_{t=T+1}^{T+P} (y_{it}^{\text{obs}} - y_{it}^{\text{cf}}),$$

which is the temporal average treatment effects (Bojinov and Shephard, 2017) at P time points. In our setting, the treatment effects for n test stores are defined as

$$\frac{1}{P} \sum_{t=T+1}^{T+P} (\mathbf{Y}_t^{\text{obs}} - \mathbf{Y}_t^{\text{cf}}). \quad (2.2.1)$$

To introduce our new causal estimand, let x_{it} be the data for the synthetic control for the i -th test store at time t . Recall that the data of a synthetic control is a weighted sum of the sales from several control stores. Define $\mathbf{X}_{1:T+P} = (\mathbf{X}_1, \dots, \mathbf{X}_{T+P})$, where \mathbf{X}_t is an $n \times p$ matrix containing data from p control stores at time t . Let μ_{it} be a latent

variable of a model, which is of interest. Define $\boldsymbol{\mu}_t = (\mu_{1t}, \dots, \mu_{nt})$ which is an $n \times 1$ vector. We let

$$p\left(\sum_{t=T+1}^{T+P} \boldsymbol{\mu}_t \mid \mathbf{Y}_{1:T+P}^{\text{obs}}, \mathbf{X}_{1:T+P}\right) \quad (2.2.2)$$

be the posterior distribution of the latent variable conditional on $\mathbf{Y}_{1:T+P}^{\text{obs}}$ and $\mathbf{X}_{1:T+P}$, and

$$p\left(\sum_{t=T+1}^{T+P} \boldsymbol{\mu}_t \mid \mathbf{Y}_{1:T}^{\text{obs}}, \mathbf{Y}_{T+1:T+P}^{\text{cf}}, \mathbf{X}_{1:T+P}\right) \quad (2.2.3)$$

be the density conditional on $\mathbf{Y}_{1:T+P}^{\text{cf}}$ and $\mathbf{X}_{1:T+P}$.

The new causal estimand is defined as the one-sided Kolmogorov-Smirnov (KS) distance between the two distributions for i -th store, which can be expressed as

$$\begin{aligned} \sup_x \left[\mathcal{F}\left(\sum_{t=T+1}^{T+P} \mu_{it} \leq x \mid y_{i,1:T}^{\text{obs}}, y_{i,T+1:T+P}^{\text{cf}}, x_{i,1:T+P}\right) \right. \\ \left. - \mathcal{F}\left(\sum_{t=T+1}^{T+P} \mu_{it} \leq x \mid y_{i,1:T+P}^{\text{obs}}, x_{i,1:T+P}\right) \right], \end{aligned}$$

where $\mathcal{F}(\cdot)$ stands for the corresponding cumulative distribution function. In our setting, since test stores are spatially correlated, the causal effect of the i -th test store is defined as

$$\begin{aligned} \sup_x \left[\mathcal{F}\left(\sum_{t=T+1}^{T+P} \mu_{it} \leq x \mid \mathbf{Y}_{1:T}^{\text{obs}}, \mathbf{Y}_{T+1:T+P}^{\text{cf}}, \mathbf{X}_{1:T+P}\right) \right. \\ \left. - \mathcal{F}\left(\sum_{t=T+1}^{T+P} \mu_{it} \leq x \mid \mathbf{Y}_{1:T+P}^{\text{obs}}, \mathbf{X}_{1:T+P}\right) \right]. \end{aligned} \quad (2.2.4)$$

A larger value of the one-sided KS distance implies a potentially larger scale of causal impact. An impact is declared to be significant if the one-sided KS distance is larger than its corresponding threshold. The threshold is calculated based on several datasets that are randomly drawn from the posterior predictive distribution of (2.2.3) (See Section 2.5 for more details.)

We would like to mention that although the proposed method is applied to a multivariate time series model, even in the context of a univariate model, the idea of comparing

posterior distributions of latent variables appears to be new. Generally speaking, this idea can be adopted into many other applications with different Bayesian models as long as these models are described in terms of latent variables.

2.3 Model and prior

2.3.1 Model

We consider a multivariate structural time series model given by (to simplify the notation, we use \mathbf{Y}_t instead of $\mathbf{Y}_t^{\text{obs}}$ in the current and the following sections),

$$\mathbf{Y}_t = \boldsymbol{\mu}_t + \boldsymbol{\delta}_t + \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\epsilon}_t, \quad (2.3.1)$$

where \mathbf{Y}_t , $\boldsymbol{\mu}_t$, $\boldsymbol{\delta}_t$ and $\boldsymbol{\epsilon}_t$ are $n \times 1$ vectors standing for the response variable, trend, seasonality and measurement error respectively. n is the number of test stores, \mathbf{X}_t is an $n \times p$ matrix containing data from p control stores at time t and $\boldsymbol{\beta}$ is a sparse $p \times 1$ vector of regression coefficients, where p can be very large. We allow each response in \mathbf{Y}_t to have different number of control stores, and write

$$\mathbf{X}_t = \begin{pmatrix} x_{11,t} & \cdots & x_{1p_1,t} & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & x_{21,t} & \cdots & x_{2p_2,t} & \cdots & 0 & \cdots & 0 \\ & & \ddots & & & \ddots & & & & \ddots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & x_{n1,t} & \cdots & x_{np_n,t} \end{pmatrix},$$

with $\sum_{i=1}^n p_i = p$. Let $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)$ be the indicator variable such that $\gamma_j = 1$ if and only if $\beta_j \neq 0$. $\boldsymbol{\epsilon}_t$ is an independent and identically distributed (i.i.d) error process.

The trend of the time series is modeled as

$$\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t + \boldsymbol{\tau}_t + \mathbf{u}_t, \quad (2.3.2)$$

where $\boldsymbol{\tau}_t$ is viewed as a term replacing the slope of the linear trend at time t to allow for a general trend, and \mathbf{u}_t is an i.i.d. error process. The process $\boldsymbol{\tau}_t$ can be modeled as a stationary VAR(1) process, driven by the equation

$$\boldsymbol{\tau}_{t+1} = \mathbf{D} + \boldsymbol{\Phi}(\boldsymbol{\tau}_t - \mathbf{D}) + \mathbf{v}_t, \quad (2.3.3)$$

where \mathbf{D} is an $n \times 1$ vector and Φ is an $n \times n$ matrix of the coefficients of the VAR(1) process with eigenvalues having modulus less than 1. If no stationarity restriction is imposed on τ_t , we model it by

$$\tau_{t+1} = \tau_t + \mathbf{v}_t, \quad (2.3.4)$$

where \mathbf{v}_t is an i.i.d. error process.

The seasonal component δ_t in (2.3.1) is assumed to follow the evolution equation

$$\delta_{t+1} = - \sum_{j=0}^{S-2} \delta_{t-j} + \mathbf{w}_t, \quad (2.3.5)$$

where S is the total length of a cycle and \mathbf{w}_t is an i.i.d. error process. For example, for an annual dataset, $S = 12$ represents the monthly effect while $S = 4$ represents the quarterly effect. This equation ensures that the summation of S time periods of each variable has expectation zero.

We assume that the residuals of (2.3.1)–(2.3.5) are mutually independent and time invariant, and are distributed as multivariate normals with mean $\mathbf{0}_{n \times 1}$ and covariance matrices Σ , Σ_u , Σ_v and Σ_w respectively.

By denoting parameters $\alpha_t = (\mu'_t, \tau'_t, \delta'_t, \dots, \delta'_{t-S+2})'$ and $\eta_t = (\mathbf{u}'_t, \mathbf{v}'_t, \mathbf{w}'_t)'$, the model can be represented as a linear Gaussian state-space model

$$\mathbf{Y}_t = \mathbf{z}\alpha_t + \mathbf{X}_t\beta + \epsilon_t, \quad (2.3.6)$$

$$\alpha_{t+1} = \mathbf{c} + \mathbf{T}\alpha_t + \mathbf{R}\eta_t, \quad (2.3.7)$$

where \mathbf{z} , \mathbf{c} , \mathbf{T} and \mathbf{R} can be rearranged accordingly based on the model (2.3.1)–(2.3.5); and $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \Sigma)$, $\eta \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$, $\mathbf{Q} = \text{bdiag}(\Sigma_u, \Sigma_v, \Sigma_w)$ are mutually independent; here and below “bdiag” refers to a block-diagonal matrix with entries as specified. If τ_t is a nonstationary process in (2.3.3), then we set $\mathbf{c} = \mathbf{0}$.

2.3.2 Prior

We now discuss the priors for the parameters in the model. We separate the parameters into four blocks: the time varying parameter α_t , the stationarity constraint parameters \mathbf{D} and Φ , the covariance matrices of the error terms Σ , Σ_u , Σ_v and Σ_w , and the sparse regression parameter β .

For the time varying parameter, we give a prior $\boldsymbol{\alpha}_1 \sim \mathcal{N}(\boldsymbol{a}, \boldsymbol{P})$ with \boldsymbol{a} is the mean and \boldsymbol{P} is the covariance matrix. For the covariance matrices of the errors, we choose priors as follows:

$$\begin{aligned} \boldsymbol{\Sigma}^{-1} &\sim W_{\mathcal{G}}(\nu, \boldsymbol{H}), & \boldsymbol{\Sigma}_u^{-1} &\sim W_{\mathcal{G}}(\nu, k_1^2(n+1)\boldsymbol{H}), \\ \boldsymbol{\Sigma}_v^{-1} &\sim W_{\mathcal{G}}(\nu, k_2^2(n+1)\boldsymbol{H}), & \boldsymbol{\Sigma}_w^{-1} &\sim W_{\mathcal{G}}(\nu, k_3^2(n+1)\boldsymbol{H}), \end{aligned}$$

where $W_{\mathcal{G}}$ stands for a \mathcal{G} -Wishart distribution. For the stationarity constraint parameter \boldsymbol{D} , we choose a conjugate prior $\boldsymbol{D} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$.

Putting a prior on the stationarity constraint matrix of a univariate AR(1) process is straightforward. However, for the VAR(1) process in (2.3.3), the stationarity matrix $\boldsymbol{\Phi}$ has to meet the Schur-stability constraint (Roy et al., 2016), that is, it needs to satisfy $|\lambda_j(\boldsymbol{\Phi})| < 1$, $j = 1, \dots, n$, where λ_j stands for the j th eigenvalue. Thus the parameter space of $\boldsymbol{\Phi}$ is given by

$$\mathfrak{S}^n = \{\boldsymbol{\Phi} \in \mathbb{R}^{n \times n} : |\lambda_j(\boldsymbol{\Phi})| < 1, j = 1, \dots, n\}. \quad (2.3.8)$$

Clearly simply putting a conjugate matrix-normal prior on $\boldsymbol{\Phi}$ does not guarantee that all the sample draws are Schur-stable. We follow Roy et al. (2016)'s method of putting priors on $\boldsymbol{\Phi}$ through a representation as given below.

We first denote $\tilde{\boldsymbol{\tau}}_t = \boldsymbol{\tau}_t - \boldsymbol{D}$, then the Yule-Walker equation for $\tilde{\boldsymbol{\tau}}_t$ is

$$\boldsymbol{U} = \boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{\Phi}' + \boldsymbol{\Sigma}_v, \quad (2.3.9)$$

where $\boldsymbol{U} = \mathbb{E}(\tilde{\boldsymbol{\tau}}_t \tilde{\boldsymbol{\tau}}_t')$ is a symmetric matrix. Letting $f(\boldsymbol{\Phi}, \boldsymbol{U}) = \boldsymbol{U} - \boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{\Phi}'$, we have that $f(\boldsymbol{\Phi}, \boldsymbol{U})$ is a positive definite matrix if and only if $\boldsymbol{\Phi} \in \mathfrak{S}^n$ (Stein, 1952). Furthermore, we have the following proposition:

Proposition 2.3.1. [Roy et al. (2016)] *Given a positive definite matrix \boldsymbol{M} , there exists a positive matrix \boldsymbol{U} , and a square matrix $\boldsymbol{\Phi} \in \mathfrak{S}^n$ such that $f(\boldsymbol{\Phi}, \boldsymbol{U}) = \boldsymbol{M}$ if and only if $\boldsymbol{U} \geq \boldsymbol{M}$ and $\boldsymbol{\Phi} = (\boldsymbol{U} - \boldsymbol{M})^{1/2} \boldsymbol{O}\boldsymbol{U}^{-1/2}$ for an orthogonal matrix \boldsymbol{O} with rank $r = \text{rank}(\boldsymbol{U} - \boldsymbol{M})$, where $(\boldsymbol{U} - \boldsymbol{M})^{1/2}$ and $\boldsymbol{U}^{-1/2}$ are full column rank square root of matrices $(\boldsymbol{U} - \boldsymbol{M})$ and \boldsymbol{U}^{-1} .*

In view of Proposition 2.3.1, given $\boldsymbol{\Phi} \in \mathfrak{S}^n$ and an arbitrary value of \boldsymbol{M} , the solution

for \mathbf{U} in equation (2.3.9) is given by

$$\text{vec}(\mathbf{U}) = (\mathbf{I}_{n^2} - \mathbf{\Phi} \otimes \mathbf{\Phi})^{-1} \text{vec}(\mathbf{M}). \quad (2.3.10)$$

Letting $\mathbf{V} = \mathbf{U} - \mathbf{M}$, we have $\mathbf{\Phi} = \mathbf{V}^{1/2} \mathbf{O} \mathbf{U}^{-1/2}$, where \mathbf{V} is a positive definite matrix, and \mathbf{O} is an orthogonal matrix. The matrix \mathbf{V} can be represented by the Cholesky decomposition $\mathbf{V} = \mathbf{L} \mathbf{\Lambda} \mathbf{L}'$, where \mathbf{L} is a lower triangular matrix and $\mathbf{\Lambda}$ is a diagonal matrix with positive entries. Thus the number of unknown parameters in \mathbf{V} reduces to $n(n-1)/2 + n$. The parameter \mathbf{O} can be decomposed by using the Cayley representation

$$\mathbf{O} = \mathbf{E}_\iota \cdot [(\mathbf{I}_n - \mathbf{G})(\mathbf{I}_n + \mathbf{G})^{-1}]^2 \quad (2.3.11)$$

with $\mathbf{E}_\iota = \mathbf{I}_n - 2\iota \mathbf{e}_1 \mathbf{e}_1'$, $\iota \in \{0, 1\}$ and $\mathbf{e}_1 = (1, 0, \dots, 0)'$, where \mathbf{G} is a skew-symmetric matrix. Thus the number of parameters in \mathbf{O} is $n(n-1)/2 + 1$. By taking the log-transform, the parameters in $\mathbf{\Lambda}$ can be made free of restrictions. Therefore there are n^2 unrestricted parameters in $\mathbf{\Phi}$ plus one binary parameter. We put normal priors on the n^2 unrestricted parameters: the lower triangular elements of \mathbf{L} , the log-transformed diagonal elements of $\mathbf{\Lambda}$ and the lower triangular elements of \mathbf{G} . For convenience, we choose the same normal prior for those parameters and choose a binomial prior for the binary parameter ι .

For the sparse regression parameter $\boldsymbol{\beta}$, we chose a spike-and-slab prior with $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}_\gamma)$, $\mathbf{A}_\gamma = \text{diag}(a_1, \dots, a_p)$ with $a_i = v_0(1 - \gamma_i) + v_1\gamma_i$, where $0 \leq v_0 < v_1$, diag refers to a diagonal matrix with entries as specified; $\pi(\boldsymbol{\gamma}|\theta) = \theta^{|\boldsymbol{\gamma}|} (1 - \theta)^{p-|\boldsymbol{\gamma}|}$ with $|\boldsymbol{\gamma}| = \sum_{i=1}^p \gamma_i$; $\theta \sim \text{Beta}(\zeta_1, \zeta_2)$.

2.4 Posterior computation

In this section, we propose a two-stage estimation algorithm to estimate the parameters. In the first stage, we adopt a fast variable selection method to obtain a point estimator for $\boldsymbol{\beta}$. In the second stage, we plug-in its estimated value and sample the remaining parameters using an MCMC algorithm.

To conduct the variable selection on $\boldsymbol{\beta}$, a popular choice would be using a SSVS method (George and McCulloch, 1993). The algorithm searches for 2^p possible combinations of β_i in $\boldsymbol{\beta}$ using Gibbs sampling under $\gamma = 0$ and $\gamma = 1$, $i = 1, \dots, p$. In the

multivariate setting, this method is computationally very challenging when p is large. An alternative way is to use the EMVS method (Ročková and George, 2014). This method uses the EM algorithm to maximize the posterior of $\boldsymbol{\beta}$ and thus obtain the estimated model. It is computationally much faster than the SSVS method. Although SSVS gives a fully Bayesian method quantifying the uncertainty of variable selection through posterior distributions, the approach is not scalable for our application which involves a large sized data. Since quantifying uncertainty of variable selection is not an essential goal, as variable selection is only an auxiliary tool here to aid inference, the faster EMVS algorithm seems to be a pragmatic method to use in our application.

After obtaining $\hat{\boldsymbol{\beta}}$, we plug it into (2.3.6)–(2.3.7) and deduct $\mathbf{X}_t \hat{\boldsymbol{\beta}}$ from \mathbf{Y}_t . We denote the new data as $\tilde{\mathbf{Y}}_t$, and will work with the following model:

$$\begin{aligned}\tilde{\mathbf{Y}}_t &= \mathbf{z}\boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t, \\ \boldsymbol{\alpha}_{t+1} &= \mathbf{c} + \mathbf{T}\boldsymbol{\alpha}_t + \mathbf{R}\boldsymbol{\eta}_t.\end{aligned}\tag{2.4.1}$$

In the MCMC step, we sample the parameters in the Model (2.4.1) from their corresponding posteriors. Those parameters include: the time-varying parameters $\boldsymbol{\alpha}_{1:T}$, the stationarity constraint parameters \mathbf{D} and $\boldsymbol{\Phi}$, the covariance matrices of the residuals $\boldsymbol{\Sigma}^{-1}$, $\boldsymbol{\Sigma}_u^{-1}$, $\boldsymbol{\Sigma}_v^{-1}$, and $\boldsymbol{\Sigma}_w^{-1}$.

2.4.1 Estimating the sparse regression parameter

If we let $\mathbf{Y}_t^* = \mathbf{Y}_t - \mathbb{E}(\mathbf{Y}_t)$, $\mathbf{X}_t^* = \mathbf{X}_t - \mathbb{E}(\mathbf{X}_t)$ and $\boldsymbol{\alpha}_t^* = \boldsymbol{\alpha}_t - \mathbb{E}(\boldsymbol{\alpha}_t)$, then we have

$$\begin{aligned}\mathbf{Y}_t^* &= \mathbf{z}\boldsymbol{\alpha}_t^* + \mathbf{X}_t^* \boldsymbol{\beta} + \boldsymbol{\epsilon}_t, \\ \boldsymbol{\alpha}_{t+1}^* &= \mathbf{T}\boldsymbol{\alpha}_t^* + \mathbf{R}\boldsymbol{\eta}_t.\end{aligned}\tag{2.4.2}$$

Recall that $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$ are i.i.d. random errors.

The joint posterior distribution for parameters in this model can be written as

$$\begin{aligned}
& \pi(\boldsymbol{\alpha}_{1:T}^*, \boldsymbol{\beta}, \boldsymbol{\gamma}, \theta, \boldsymbol{\Phi}, \boldsymbol{\Sigma}, \boldsymbol{Q} \mid \boldsymbol{Y}_t^*, \boldsymbol{X}_t^*) \\
& \propto \prod_{t=1}^T f(\boldsymbol{Y}_t^* \mid \boldsymbol{X}_t^*, \boldsymbol{\alpha}_t^*, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \prod_{t=2}^T f(\boldsymbol{\alpha}_t^* \mid \boldsymbol{\alpha}_{t-1}^*, \boldsymbol{\Phi}, \boldsymbol{Q}) \pi(\boldsymbol{\alpha}_1^*) \\
& \quad \times \pi(\boldsymbol{\beta} \mid \boldsymbol{\gamma}) \pi(\boldsymbol{\gamma} \mid \theta) \pi(\theta) \pi(\text{vec}(\boldsymbol{\Phi})) \pi(\boldsymbol{\Sigma}^{-1} \mid \nu, \boldsymbol{H}) \\
& \quad \times \pi(\boldsymbol{\Sigma}_u^{-1} \mid k_1, \nu, \boldsymbol{H}) \pi(\boldsymbol{\Sigma}_v^{-1} \mid k_2, \nu, \boldsymbol{H}) \pi(\boldsymbol{\Sigma}_w^{-1} \mid k_3, \nu, \boldsymbol{H}).
\end{aligned} \tag{2.4.3}$$

To obtain $\hat{\boldsymbol{\beta}}$, we use a revised EMVS algorithm and the priors for almost all the parameters except for $\boldsymbol{\Phi}$. Since $\boldsymbol{\beta}$ is the only parameter we are interested in here, to reduce the complexity of deriving the expression appearing in the EMVS algorithm, we consider a conjugate prior for $\boldsymbol{\Phi}$: $\text{vec}(\boldsymbol{\Phi}) \sim \mathcal{N}(\mathbf{0}, 0.1 \times \boldsymbol{I}_{n^2})$. In the simulation study, we show that the choice of the prior for $\boldsymbol{\Phi}$ is not influential.

Our EMVS algorithm indirectly maximize the posterior by iteratively maximizing the object function:

$$\begin{aligned}
& \mathcal{Q}(\boldsymbol{\beta}, \theta, \boldsymbol{\Phi}, \boldsymbol{\Sigma}, \boldsymbol{Q} \mid \boldsymbol{\beta}^{(k)}, \theta^{(k)}, \boldsymbol{\Phi}^{(k)}, \boldsymbol{\Sigma}^{(k)}, \boldsymbol{Q}^{(k)}) \\
& = C + \mathcal{Q}_1(\boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\Sigma}, \boldsymbol{Q} \mid \boldsymbol{\beta}^{(k)}, \theta^{(k)}, \boldsymbol{\Phi}^{(k)}, \boldsymbol{\Sigma}^{(k)}, \boldsymbol{Q}^{(k)}) + \mathcal{Q}_2(\theta \mid \boldsymbol{\beta}^{(k)}, \theta^{(k)}),
\end{aligned}$$

where C is a constant,

$$\begin{aligned}
& \mathcal{Q}_1(\boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\Sigma}, \boldsymbol{Q} \mid \boldsymbol{\beta}^{(k)}, \theta^{(k)}, \boldsymbol{\Phi}^{(k)}, \boldsymbol{\Sigma}^{(k)}, \boldsymbol{Q}^{(k)}) \\
& = -\frac{1}{2} \sum_{t=1}^T \mathbb{E}_{\boldsymbol{\alpha}_{1:T}^*} \left[(\boldsymbol{Y}_t^* - \boldsymbol{X}_t^* \boldsymbol{\beta} - \boldsymbol{z} \boldsymbol{\alpha}_t^*)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y}_t^* - \boldsymbol{X}_t^* \boldsymbol{\beta} - \boldsymbol{z} \boldsymbol{\alpha}_t^*) \right] \\
& \quad - \frac{1}{2} \sum_{t=1}^{T-1} \mathbb{E}_{\boldsymbol{\alpha}_{1:T}^*} \left[(\boldsymbol{\alpha}_{t+1}^* - \boldsymbol{T} \boldsymbol{\alpha}_t^*)' \boldsymbol{R} \boldsymbol{Q}^{-1} \boldsymbol{R}' (\boldsymbol{\alpha}_{t+1}^* - \boldsymbol{T} \boldsymbol{\alpha}_t^*) \right] - \frac{1}{2} \boldsymbol{\alpha}_1^{*'} \boldsymbol{\alpha}_1^* \\
& \quad + \frac{T + \nu - 2}{2} \log |\boldsymbol{\Sigma}^{-1}| + \frac{T + \nu - 3}{2} \left(\log |\boldsymbol{\Sigma}_u^{-1}| + \log |\boldsymbol{\Sigma}_v^{-1}| + \log |\boldsymbol{\Sigma}_w^{-1}| \right) \\
& \quad - \frac{1}{2} \left(\text{Tr}(\boldsymbol{H} \boldsymbol{\Sigma}^{-1}) + \text{Tr}(k_1^2 (n+1) \boldsymbol{H} \boldsymbol{\Sigma}_u^{-1}) \right) \\
& \quad + \text{Tr}(k_2^2 (n+1) \boldsymbol{H} \boldsymbol{\Sigma}_v^{-1}) + \text{Tr}(k_3^2 (n+1) \boldsymbol{H} \boldsymbol{\Sigma}_w^{-1}) \\
& \quad - \frac{1}{2 \times 0.1} \text{vec}(\boldsymbol{\Phi})' \text{vec}(\boldsymbol{\Phi}) - \sum_{i=1}^p \mathbb{E}_{\boldsymbol{\gamma} \mid \cdot} \left[\log(v_0(1 - \gamma_i) + v_1 \gamma_i) \right]
\end{aligned}$$

$$-\frac{1}{2} \sum_{i=1}^p \beta_i^2 \mathbb{E}_{\gamma_i} \left[\frac{1}{v_0(1 - \gamma_i) + v_1 \gamma_i} \right],$$

and

$$\begin{aligned} \mathcal{Q}_2(\theta \mid \boldsymbol{\beta}^{(k)}, \theta^{(k)}) \\ = \sum_{i=1}^p \log \left(\frac{\theta}{1 - \theta} \right) \mathbb{E}_{\gamma_i}(\gamma_i) + (\zeta_1 - 1) \log \theta + (p + \zeta_2 - 1) \log(1 - \theta). \end{aligned}$$

A more detailed derivation of $\mathcal{Q}_1(\boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\Sigma}, \boldsymbol{Q} \mid \boldsymbol{\beta}^{(k)}, \theta^{(k)}, \boldsymbol{\Phi}^{(k)}, \boldsymbol{\Sigma}^{(k)}, \boldsymbol{Q}^{(k)})$ and $\mathcal{Q}_2(\theta \mid \boldsymbol{\beta}^{(k)}, \theta^{(k)})$ are given in Appendix B.

2.4.2 Sampling the time-varying parameters

From (2.4.1), the posterior distribution of $\boldsymbol{\alpha}_{1:T}$ can be expressed as

$$\pi(\boldsymbol{\alpha}_{1:T} \mid \boldsymbol{\vartheta}, \tilde{\boldsymbol{Y}}_t) \propto \prod_{t=1}^T p(\tilde{\boldsymbol{Y}}_t \mid \boldsymbol{\alpha}_t, \boldsymbol{\vartheta}) \prod_{t=2}^T p(\boldsymbol{\alpha}_t \mid \boldsymbol{\alpha}_{t-1}, \boldsymbol{\vartheta}) \pi(\boldsymbol{\alpha}_1) \quad (2.4.4)$$

with each density being a multivariate normal distribution; here $\boldsymbol{\vartheta}$ stands for all the parameters in the model excluding $\boldsymbol{\alpha}_{1:T}$. To sample $\boldsymbol{\alpha}_{1:T}$, we apply the Kalman-filter and simulation smoother algorithm (Durbin and Koopman, 2012, 2002). Recall $\boldsymbol{\alpha}_1 \sim \mathcal{N}(\boldsymbol{a}, \boldsymbol{P})$. To choose a definite prior for $\boldsymbol{\alpha}_1$, we proceed as follows: if $\boldsymbol{\alpha}_{1:T}$ follows a nonstationary stochastic process, we choose a diffuse prior for $\boldsymbol{\alpha}_1$, that is we let $\boldsymbol{a} = \mathbf{0}$ and \boldsymbol{P} be a diagonal matrix with large diagonal elements. If $\boldsymbol{\tau}_{1:T}$ is restricted to be stationary, then the component of $\boldsymbol{\alpha}_1$ corresponding to $\boldsymbol{\tau}_1$ needs to have a prior distribution with a smaller variance. Under some circumstances, one can estimate the values for the variances using classical methods by treating parameters as time-invariant in a given training datasets (Primiceri, 2005). But here $\boldsymbol{\tau}_t$ is a latent variable standing for the trend and thus is not estimable. However, we found that the resulting procedure is not sensitive to the choice of values of variance. For example, we can choose the covariance matrix of $\boldsymbol{\tau}_t$ to be the identity matrix.

2.4.3 Sampling the stationarity constraint parameters

Given the priors described in Section 2.3.2, we sample the parameters from their posterior distributions using a Metropolis-Hastings algorithm. We choose the proposal density for each parameter to be a normal distribution centered at the value of its lastest MCMC draw and variance to be a small number, say 0.1.

One can choose to update $n^2 + 1$ parameters one by one. However, when n is large, it is more efficient to update them in blocks. One may choose the block knots to be either fixed or stochastic. For stochastic knots, one may use the method introduced by Shephard and Pitt (1997).

Given a current draw for Φ , by using the prior $\mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, \mathbf{D} can be sampled from its posterior given by

$$\pi(\mathbf{D}|\Phi, \boldsymbol{\tau}_{1:T}) \sim \mathcal{N}(\boldsymbol{\mu}_D, \mathbf{V}_D), \quad (2.4.5)$$

with

$$\begin{aligned} \mathbf{V}_D &= [(T-1)(\mathbf{I}_n - \Phi)' \boldsymbol{\Sigma}_v^{-1} (\mathbf{I}_n - \Phi) + \mathbf{I}_n]^{-1}, \\ \boldsymbol{\mu}_D &= \mathbf{V}_D [(\mathbf{I}_n - \Phi)' \boldsymbol{\Sigma}_v^{-1} \sum_{t=1}^{T-1} (\boldsymbol{\tau}_{t+1} - \Phi \boldsymbol{\tau}_t)]. \end{aligned}$$

2.4.4 Sampling the covariance matrices of the residuals

The residual error terms are both multivariate normal with precision matrices having the same sparsity structure. In the following we derive the posterior of $\boldsymbol{\Sigma}^{-1}$; the posteriors for $\boldsymbol{\Sigma}_u^{-1}$, $\boldsymbol{\Sigma}_v^{-1}$ and $\boldsymbol{\Sigma}_w^{-1}$ can be derived in a similar way. We summarize these parameters in Table 2.1.

We impose sparsity on $\boldsymbol{\Sigma}^{-1}$ to reduce the number of parameters needed to be estimated. The sparsity structure is a pre-determined matrix of zeros and ones which assumes that two responses are conditionally independent if the corresponding entry is 0. In Bayesian analysis, sparsity on an inverse covariance matrix is often imposed through a \mathcal{G} -Wishart prior (Dawid and Lauritzen, 1993; Roverato, 2000, 2002). Given a sparsity structure \mathcal{G} , the prior $W_{\mathcal{G}}(\nu, \mathbf{H})$ has the expression

$$\pi(\boldsymbol{\Sigma}^{-1}|\mathcal{G}) = I_{\mathcal{G}}(\nu, \mathbf{H}) |\boldsymbol{\Sigma}^{-1}|^{(\nu-2)/2} \exp\left\{-\frac{1}{2}\text{tr}(\mathbf{H}\boldsymbol{\Sigma}^{-1})\right\} \mathbb{1}_{\{\boldsymbol{\Sigma}^{-1} \in M^+(\mathcal{G})\}}, \quad (2.4.6)$$

where the normalizing constant $I_{\mathcal{G}}(\nu, \mathbf{H})$ for a decomposable graph is given by

$$I_{\mathcal{G}}(\nu, \mathbf{H}) = \int |\boldsymbol{\Sigma}^{-1}|^{(\nu-2)/2} \exp\left\{-\frac{1}{2}\text{tr}(\mathbf{H}\boldsymbol{\Sigma}^{-1})\right\} \mathbb{1}_{\{\boldsymbol{\Sigma}^{-1} \in M^+(\mathcal{G})\}} d\boldsymbol{\Sigma}^{-1}; \quad (2.4.7)$$

here $\nu > n - 1$ is the degree of freedom, \mathbf{H} is an $n \times n$ symmetric positive definite matrix, $M^+(\mathcal{G})$ is the cone of the symmetric positive definite matrix $\boldsymbol{\Sigma}^{-1}$ based on the graphical structure \mathcal{G} . We allow \mathcal{G} to be either decomposable or non-decomposable and use the `BDgraph` package in the R library for computation, which uses explicit expression for the normalizing constant in (2.4.7) if the graph is decomposable and uses the BDMCMC algorithm (Mohammadi and Wit, 2015) if the graph is non-decomposable.

Table 2.1: Parameters of the posterior densities of $\boldsymbol{\Sigma}^{-1}$, $\boldsymbol{\Sigma}_u^{-1}$, $\boldsymbol{\Sigma}_v^{-1}$, $\boldsymbol{\Sigma}_w^{-1}$.

Parameters	DF	Scale matrix
$\boldsymbol{\Sigma}^{-1}$	$T + \nu$	$\sum_{t=1}^T (\tilde{\mathbf{Y}}_t - \mathbf{z}_t \boldsymbol{\alpha}_t)(\tilde{\mathbf{Y}}_t - \mathbf{z}_t \boldsymbol{\alpha}_t)' + \mathbf{H}$
$\boldsymbol{\Sigma}_u^{-1}$	$T + \nu - 1$	$\sum_{t=1}^{T-1} (\boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}_t - \boldsymbol{\tau}_t)(\boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}_t - \boldsymbol{\tau}_t)' + k_1^2(n+1)\mathbf{H}$
$\boldsymbol{\Sigma}_v^{-1}$	$T + \nu - 1$	$\sum_{t=1}^{T-1} (\tilde{\boldsymbol{\tau}}_{t+1} - \boldsymbol{\Phi} \tilde{\boldsymbol{\tau}}_t)(\tilde{\boldsymbol{\tau}}_{t+1} - \boldsymbol{\Phi} \tilde{\boldsymbol{\tau}}_t)' + k_2^2(n+1)\mathbf{H}$
$\boldsymbol{\Sigma}_w^{-1}$	$T + \nu - 1$	$\sum_{t=1}^{T-1} (\boldsymbol{\delta}_{t+1} + \sum_{j=0}^{S-2} \boldsymbol{\delta}_{t-j})(\boldsymbol{\delta}_{t+1} + \sum_{j=0}^{S-2} \boldsymbol{\delta}_{t-j})' + k_3^2(n+1)\mathbf{H}$

The proposed two-stage estimation algorithm is thus summarized as follows:

Stage 1: EMVS step. Choose initial values for $\boldsymbol{\beta}^{(0)}$, $\mathbf{a}_1^{*(0)}$ and $\mathbf{P}_1^{*(0)}$ using the revised EMVS algorithm to find the optimized value for $\boldsymbol{\beta}$.

The two-stage algorithm is proposed as follows. In Stage 1, we estimate the sparse regression parameter $\hat{\boldsymbol{\beta}}$. In Stage 2, we sample the parameters in the model (except $\boldsymbol{\beta}$) using the MCMC algorithm. In below, we provide the details of our algorithm.

Stage 2: MCMC step. Given $\tilde{\mathbf{Y}}_t$, we sample parameters using MCMC with the following steps:

- (a) Generate $\boldsymbol{\alpha}_t$ using the Kalman filter and simulation smoother method.
- (b) Generate $\boldsymbol{\Phi}$ using the Metropolis-Hastings algorithm.
- (c) Generate \mathbf{D} .

- (d) Generate covariance matrices from their respective \mathcal{G} -Wishart posterior densities.
- (e) Go to Step (a) and repeat until the chain converges.

Skip Step (b) and (c) if no stationarity restriction is imposed on $\boldsymbol{\tau}_t$.

2.5 A new method to infer causality

In this section, we will introduce our new method to infer causality (in short, “the new method”) along with a commonly used method.

Recall the treatment effects of the commonly used method is defined in (2.2.1). Since $\sum_{t=T+1}^{T+P} \mathbf{Y}_t^{\text{cf}}$ is an unobserved quantity, we replace it by its posterior samples from $p(\sum_{t=T+1}^{T+P} \mathbf{Y}_t^{\text{cf}} | \mathbf{Y}_{1:T}^{\text{obs}}, \mathbf{X}_{1:T+P})$.

The commonly used method may fail to detect even for a moderately sized impact for two reasons. First, the prediction intervals increase linearly as the time lag increases. Secondly, the trends are the only latent variables would give a response to an impact; including the random noise and the seasonality components would inflate the uncertainty of the estimated effect. For the data with a low signal-to-noise ratio, this method is which even harder to detect causal impacts.

We thus propose a new method by comparing only the posterior distributions of the latent trend in the model given the observations and the data from counterfactuals. The new method consists the following five steps:

Step 1: Applying the two-stage algorithm to obtain posterior samples for parameters in the model using the data from the period without causal impacts.

Step 2: Based on those posterior samples, obtaining sample draws of $\mathbf{Y}_{T+1:T+P}^{\text{cf}}$ from its predictive posterior distribution $p(\mathbf{Y}_{T+1:T+P}^{\text{cf}} | \mathbf{Y}_{1:T}^{\text{obs}}, \mathbf{X}_{1:T+P})$.

Step 3: Generating k different datasets from counterfactual potential outcomes (in short, “counterfactual datasets”) from the predictive posterior distribution, for the j -th dataset, $j \in \{1, \dots, k\}$, denoted by $\mathbf{Y}_{T+1:T+P}^{\text{cf}(j)}$. Then fitting each $\mathbf{Y}_{T+1:T+P}^{\text{cf}(j)}$ into the model to obtain sample draws of the trend from its posterior distribution, which is shown in (2.2.3) (here, we replace $\mathbf{Y}_{T+1:T+P}^{\text{cf}}$ with $\mathbf{Y}_{T+1:T+P}^{\text{cf}(j)}$). Also, fitting the observed data $\mathbf{Y}_{1:T+P}^{\text{obs}}$ into the model and sampling from (2.2.2).

Step 4: Using the one-sided Kolmogorov-Smirnov (KS) distance to quantify the difference between the posterior distributions of the trend given by the observed data and

the counterfactual datasets. The posterior distribution of the trend given by the counterfactual datasets is obtained by stacking the sample draws estimated from all the k simulated datasets, then calculating the KS distance between the two posterior distributions for each store as follows:

$$\sup_x \left[\frac{1}{k} \sum_{j=1}^k \left(\mathcal{F} \left(\sum_{t=T+1}^{T+P} \mu_{it} \leq x \mid \mathbf{Y}_{1:T}^{\text{obs}}, \mathbf{Y}_{T+1:T+P}^{\text{cf}(j)}, \mathbf{X}_{1:T+P} \right) \right) - \mathcal{F} \left(\sum_{t=T+1}^{T+P} \mu_{it} \leq x \mid \mathbf{Y}_{1:T+P}^{\text{obs}}, \mathbf{X}_{1:T+P} \right) \right], \quad (2.5.1)$$

where $i = 1, \dots, n$, and $\mathcal{F}(\cdot)$ stands for the empirical distribution function of the obtained MCMC samples.

Step 5: Calculating the $k \times (k - 1)$ pairwise one-sided KS distances between the posterior distributions of the trends given by the k simulated counterfactual datasets, that is to calculate the following expression

$$\sup_x \left[\mathcal{F} \left(\sum_{t=T+1}^{T+P} \mu_{it} \leq x \mid \mathbf{Y}_{1:T}^{\text{obs}}, \mathbf{Y}_{T+1:T+P}^{\text{cf}(j)}, \mathbf{X}_{1:T+P} \right) - \mathcal{F} \left(\sum_{t=T+1}^{T+P} \mu_{it} \leq x \mid \mathbf{Y}_{1:T}^{\text{obs}}, \mathbf{Y}_{T+1:T+P}^{\text{cf}(j')}, \mathbf{X}_{1:T+P} \right) \right], \quad (2.5.2)$$

where $j, j' = 1, \dots, k, j \neq j'$. Then, for each i , choosing the 95% upper percentile among those distances as a threshold to decide whether the KS distance calculated from (2.5.1) is significant or not. If the KS distance is smaller than this threshold, then the corresponding causal impact is declared not significant.

The use of a threshold is necessary, since the two posterior distributions of the trend obtained under observed data and the data from the counterfactual are not exactly equal even when there is no causal impact. Our method automatically selects a data-driven threshold through a limited repeated sampling as in multiple imputations.

So far we described the commonly used method and the new method in the setting where the period without a causal impact comes before that with the impact. However, the new method can be extended to allow datasets in more general situations when: 1) there are missing data from the period without causal impact; 2) the period without causal impact comes after the period with a impact; 3) there are more than one periods

without causal impact, both before and after the period with a impact. This is because the KFBS method is flexible to impute missing values at any positions in a times series dataset.

2.6 Simulation study

In this section, we conduct a simulation study to compare the two different methods introduced in the last section. To keep the analysis simple, we only consider the setting that the period with causal impact follows that without the impact.

2.6.1 Data generation and Bayesian estimation

We simulate five spatially correlated datasets, and assume the precision matrices in the model have the adjacency matrix as follows:

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}, \quad (2.6.1)$$

that is, we assume variables align in a line with each one correlated with only its nearest neighbors. We generate daily time series for an arbitrary date range from January 1, 2016 to April 9, 2016, with a perturbation beginning on March 21, 2016. We specify dates in the simulation to facilitate the description of the intervention period. We first generate five multivariate datasets for test stores with varying levels of impact and label them as Datasets 1–5.

For each Dataset i , $i = 1, \dots, 5$, the trend is generated from $\mu_{it} \sim \mathcal{N}(0.8\mu_{i,t-1}, 0.1^2)$ with $\mu_{i0} = 1$. The weekly components are generated from two sinusoids of the same frequency 7 as follows:

$$\delta_{it} = 0.1 \times \cos(2\pi t/7) + 0.1 \times \sin(2\pi t/7). \quad (2.6.2)$$

Additional datasets for 10 control stores are generated, each from an AR(1) process with coefficient 0.6 and standard error 1. We let the first and second datasets to have regression

coefficients $\beta_1 = 1, \beta_2 = 2$ and let the rest to be 0. We then generate the residuals ϵ_t in the observation equation from the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \Sigma)$ with precision matrix having sparsity structure given by (2.6.1). We set the diagonal elements for Σ^{-1} to 10, and its non-zero off-diagonal elements to 5. The simulated data for test stores are the sum of the simulated values of $\mu_t, \delta_t, \mathbf{X}_t\beta$ and ϵ_t . The causal impacts are generated as follows: for each Dataset $i, i = 1, \dots, 5$, we add an impact scale $\frac{(i-1)}{2} \times (\log 1, \dots, \log 20)$ from March 21, 2016 to April 9, 2016. Clearly no causal impact is added in Dataset 1.

We impose the graphical structure with adjacency matrix in (2.6.1) in both observed and hidden processes in the model and then apply the two-stage algorithm to estimate parameters. In Stage 1, we apply the revised EMVS algorithm. We choose the initial values $\beta^{(0)}$ and $\mathbf{a}_1^{*(0)}$ to be the zero vectors and the first 15×15 elements of $\mathbf{P}_1^{*(0)}$, which correspond to the covariances of the trend, local trend and seasonality components, to be a diagonal matrix. The remaining elements in $\mathbf{P}_1^{*(0)}$ are set to 0. We select 20 equally spaced relatively small values for v_0 from 10^{-6} to 0.02 and a relatively larger value for $v_1, 10$. For the prior of θ , we set $\zeta_1 = \zeta_2 = 1$. The maximum number of iterations of the EMVS algorithm is chosen to be 50. We calculate the threshold of non-zero value of β_i from the inequality: $p(\gamma_i = 1 | \beta_i, \mathbf{Y}_t^*, \mathbf{X}_t^*) > 0.5$ (See the detailed discussions in Ročková and George, 2014). Then the threshold can be expressed as

$$|\beta_i^{\text{th}}| \geq \sqrt{\frac{\log(v_0/v_1) + 2 \log(\hat{\theta}/(1 - \hat{\theta}))}{v_1^{-1} - v_0^{-1}}},$$

where $\hat{\theta}$ is the maximized value obtained from the EMVS algorithm. Ročková and George (2014) also suggested using a deterministic annealing variant of the EMVS (DAEMVS) algorithm which maximizes

$$\mathbb{E}_{(\alpha_{1:T}^*, \gamma)} \left[\frac{1}{s} \log \pi(\alpha_{1:T}^*, \beta, \gamma, \theta, \Phi, \Sigma, \mathbf{Q} | \mathbf{Y}_t^*, \mathbf{X}_t^*)^s \mid \beta^{(k)}, \theta^{(k)}, \Phi^{(k)}, \Sigma^{(k)}, \mathbf{Q}^{(k)} \right], \quad (2.6.3)$$

where $0 \leq s \leq 1$. The parameter $1/s$ is known as a temperature function (Ueda and Nakano, 1998). When the temperature is higher, that is when $s \rightarrow 0$, the DAEMVS algorithm has a higher chance to find a global mode and thus reduces the chance of getting trapped at a local maximum.

Figure 2.2 compares the results for using EMVS and DAEMVS with $s = 0.1$ algorithms. We plot $\hat{\beta}$ and their thresholds based on 20 different values of v_0 from 10^{-6} to

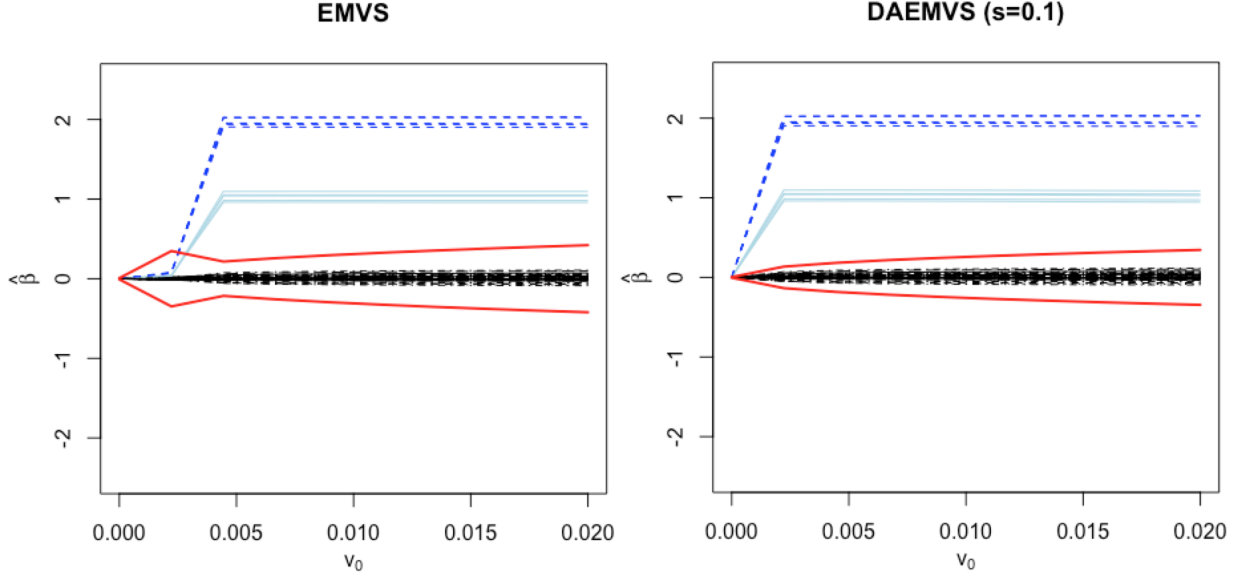


Figure 2.2: EMVS (left) and DAEMVS (with $s = 0.1$) (right) estimation of β based on the simulated datasets. The dark blue lines are the parameters that have simulated values 2; the light blue lines are the parameters that have simulated values 1 and the black lines are the parameters that have simulated values 0. The red lines are the calculated β_i^{th} values, within the two red lines, the parameters should be considered as zero parameters.

0.02. From the plot, the estimated values for β using both EMVS and DAEMVS methods are close to their true values.

The true zero coefficients are estimated to be very close to 0. However, we observe that the values of β_i^{th} is larger by using the EMVS method compared to the DAEMVS method. This is because in the region where v_0 is less than 0.005, the $\hat{\theta}$ estimated from EMVS is very close to 0, thus the negative value of $\log(\hat{\theta}/(1-\hat{\theta}))$ is very large and the threshold becomes larger. Based on the simulation results, we use DAEMVS with $s = 0.1$ throughout the rest of the chapter.

The DAEMVS gives a smaller value of β_i^{th} , yet the thresholds can distinguish the true zero and non-zero coefficients in this case. Nevertheless it may miss a non-zero coefficient if the coefficient is within the thresholds. In practice, since our goal is to identify significant control variables and use them to build counterfactuals for a causal inference, we may choose to include more variables than the threshold suggests provided that the total number of included variables is still manageable.

Recall that in the Stage 1, we used a conjugate prior for $\text{vec}(\Phi)$ instead of the origi-

nally proposed prior described in Section 2.4.3. Here, we want to make sure the change of prior would not affect the results of $\hat{\beta}$ too much. We conduct the analysis by choosing two different values of the covariance matrix of the prior: \mathbf{I}_5 and $0.01 \times \mathbf{I}_5$. We found the estimates $\hat{\beta}$ s are almost identical to the estimated values shown in Figure 2.2. We also consider using other two models: one ignores the stationarity constraint for τ_t (henceforth the “nonstationary model”); another ignores the time dependency of the model (henceforth the “misspecified model”). To be more explicit, for the nonstationary model, we let the local linear trend follow (2.3.4). The misspecified model is given by $\mathbf{Y}_t^* = \mathbf{X}_t^* \beta + \boldsymbol{\zeta}_t$, with $\boldsymbol{\zeta}_t$ s are i.i.d random errors with multivariate normally distributed and mean $\mathbf{0}$ by ignoring their dependency. We conduct DAEMVS with $s = 0.1$ for both of the two models. In the nonstationary model, we choose a diffuse prior for α_1^* and change the covariance corresponding to the local linear trend in $\mathbf{P}_1^{*(0)}$ to be $10^6 \times \mathbf{I}_5$. In the misspecified model, the M-step in Section B.1 can be simplified to only updates for β , θ and the covariance matrix of $\boldsymbol{\zeta}_t$. We plot the results into Figure 2.3. Comparing the results in Figure 2.3 with Figure 2.2, there are not much differences among the results obtained using the three different models for estimating β .

In Stage 2, we plug-in $\hat{\beta}$ and calculate $\tilde{\mathbf{Y}}_t$ in (2.4.1). We choose the prior for the rest of parameters as follows: we let $\alpha_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. If τ_t is a nonstationarity process, the initial condition is considered as a diffuse random variable with large variance (Durbin and Koopman, 2002). Then we let the covariance matrix of τ_t to be $10^6 \times \mathbf{I}_5$. We let $\nu = 1$, $k_1 = k_2 = k_3 = 0.1$. We choose $\mathbf{H} = \mathbf{I}_5$ and the priors for 25 parameters decomposed from Φ to be $\mathcal{N}(0, \sqrt{5}^2)$, and let $\iota \sim \text{Bernoulli}(0.5)$. We run total 10,000 MCMC iterations with the first 2,000 draws as burn-in.

2.6.2 Performance of the commonly used causal inference method

In this section, we study the performance of the commonly used method. The causal effect is estimated by taking the difference between observed data during causal period and the potential outcomes of counterfactuals during that period. In Stage 1, we use the DAEMVS ($s = 0.1$) algorithm to estimate $\hat{\beta}$ for the model (2.3.6)–(2.3.7). A stationarity constraint is added on the local linear trend τ_t . In Stage 2, we consider two different settings for τ_t : with and without adding the stationarity constraint. We choose Dataset 4 as an example and plot accuracy of the model based on the two different settings in Figure 2.4. There are four subplots: the left two subplots are the results for the model

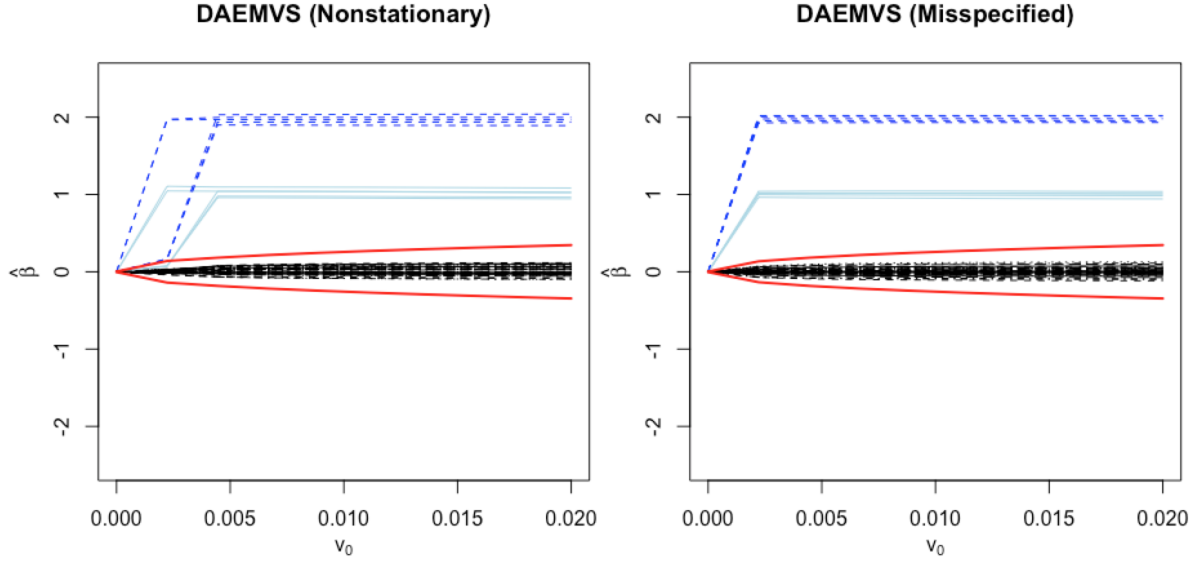


Figure 2.3: DAEMVS (with $s = 0.1$) estimation of β based on the simulated datasets using the nonstationary model (left) and the misspecified model (right). The dark blue lines are the parameters that have simulated values 2; the light blue lines are the parameters that have simulated values 1 and the black lines are the parameters that have simulated values 0. The red lines are the calculated β_i^{th} values, within the two red lines, the parameters should be considered as zero.

with a nonstationary local linear trend and the right two subplots are the results for the model with a stationary local linear trend. Before the period with a causal impact, which is March, 21, 2016, the estimated posterior medians and 95% credible intervals obtained from the two models are close (see plots (b) and (d) in Figure 2.4); but their prediction intervals during the period with a causal impact are quite different. In the model with a nonstationary local linear trend, the prediction intervals are much wider and expand more rapidly than those resulting from the model with a stationary local linear trend. In the former case, the observed data during the campaign are fully contained inside the prediction intervals and thus failed to detect a causal impact. However, the model with a stationary local linear trend gives only moderately increasing prediction intervals and thus can detect the causal impact. Plots (b) and (d) shown in the bottom of Figure 2.4 are the estimated causal impact in each model for Dataset 4 calculated by taking the difference between observed values and counterfactual potential outcomes. In each plot, the estimated causal impact medians are able to capture the shape of the simulated

causal impact. However, the prediction intervals in plot (b) contain the value 0 and thus negate the impact. The shorter prediction intervals in plot (d) do not contain the value 0, and thus indicate the existence of a impact.

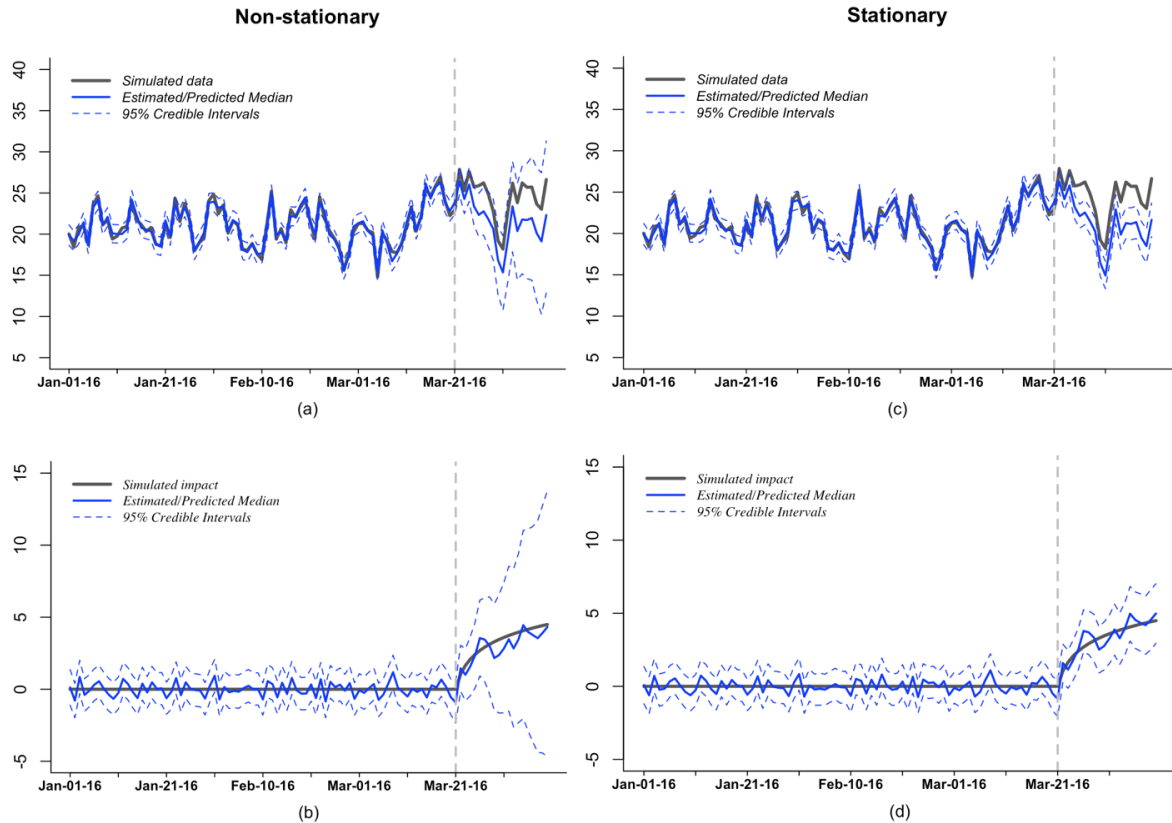


Figure 2.4: Plot of the causal impact in Dataset 4 using models with a stationary and a nonstationary local linear trend. (a) and (c) are the plots of estimation (before March 21, 2016) and prediction (after March 21, 2016) of Dataset 4 without stationarity constraint (left) and with stationarity constraint (right). The gray line is the simulated dataset, the blue line is the estimated posterior median of the dataset using the model, the dashed blue line is the corresponding 95% credible and prediction intervals. (b) and (d) are the plots of estimated causal impact by taking the difference between the observed data and Bayesian estimates using the model with a nonstationary local linear trend (left) and the model with a stationary local linear trend (right). The black line is the simulated true impact, the blue line is the estimated median of the impact, the dashed blue lines are the corresponding 95% credible and prediction intervals.

To give an overall picture of the model fitting for the five simulated datasets, we

Table 2.2: Posterior medians and 95% credible intervals of average causal impacts for simulated datasets estimated using the multivariate models with a stationary and a nonstationary local linear trend.

	Simulated impact	Nonstationary	Stationary
Dataset 1	0.00	0.00 [-4.419, 4.425]	0.29 [-1.440, 1.996]
Dataset 2	1.06	0.64 [-3.989, 5.298]	1.07 [-0.648, 2.780]
Dataset 3	2.12	1.22 [-3.758, 5.965]	2.27 [0.399, 4.014]
Dataset 4	3.18	2.83 [-1.793, 7.575]	3.16 [1.500, 4.862]
Dataset 5	4.23	4.25 [-0.249, 8.771]	4.25 [2.520, 5.904]

summarize the posterior medians and their 95% credible intervals of the estimated causal impact for all the datasets in Table 2.2. In the model with a nonstationary local linear trend, no impacts are detected for all the five datasets since their corresponding prediction intervals all contain the value 0. In the model with a stationary local linear trend on τ_t , the impacts are successfully detected for the last three datasets. For Dataset 2, it has a weaker impact. Its impact is not detected even after imposing the stationarity constraint. Also, when the stationarity constraint is imposed, including the intercept \mathbf{D} in (2.3.3) helps give a robust long run prediction. Thus, from Table 2.2, we find that the estimated medians using the model with a stationary local linear trend are closer to the true impact compared with that obtained from using the model with a nonstationary local linear trend.

In the setting where the sales in the test stores are spatially correlated, the use of the multivariate model with a stationary local linear trend is necessary for obtaining more accurate estimates for causal effects. We compare the results with a univariate model which ignores the correlation between the five simulated datasets. We fit the five datasets independently into that model. The model is the univariate version of the model (2.3.6)–(2.3.7). In the univariate model, the errors ϵ_t , \mathbf{u}_t , \mathbf{v}_t and \mathbf{w}_t become scalars. We denote σ^2 , σ_u^2 , σ_v^2 and σ_w^2 as their corresponding variances. We choose their priors as $\sigma^{-2} \sim \text{Gamma}(0.1/2, 0.1 \times \text{SS}/2)$, $\sigma_u^{-2}, \sigma_v^{-2}, \sigma_w^{-2} \sim \text{Gamma}(0.01, 0.01 \times \text{SS})$, where $\text{SS} = \sum_{t=1}^T (y_t - \bar{y})^2 / (T - 1)$ and $\bar{y} = \sum_{t=1}^T y_t / T$. The parameters \mathbf{D} and Φ in (2.3.3) also become scalars and to be denoted by d and ϕ respectively. We give them the priors $d \sim \mathcal{N}(0, 0.1^2)$ and $\phi \sim \mathcal{N}(0, 0.1^2) \mathbb{1}_{(-1, 1)}$.

In order to make the comparison between the multivariate model and the univariate model meaningful, we plug-in the same $\hat{\beta}$ obtained from Stage 1 for both models. We

Table 2.3: Posterior medians and 95% credible intervals of average causal impacts for simulated datasets estimated using the univariate model.

	Simulated impact	Stationary (univariate)
Dataset 1	0.00	0.17 [-2.197, 2.472]
Dataset 2	1.06	1.03 [-1.365, 3.473]
Dataset 3	2.12	2.16 [-0.370, 4.476]
Dataset 4	3.18	3.20 [0.821, 5.748]
Dataset 5	4.23	4.08 [1.564, 6.489]

conduct an MCMC algorithm for the five datasets separately using the univariate model by sequentially sampling draws from the corresponding posterior distributions of $\alpha_{1:T}$, d , ϕ , σ^2 , σ_u^2 , σ_v^2 and σ_w^2 . We run the MCMC algorithm for 10,000 iterations and treat the first 2,000 as burn-in. The estimated causal impacts are shown in Table 2.3. By comparing the results with the results in Table 2.2, the univariate model produces wider credible intervals for all of datasets even though their posterior medians are close to the truth. Thus the multivariate model with a stationary local linear trend is more accurate for detecting a causal impact.

We conduct additional independent 10 simulation studies by generating datasets using the same scheme which described above, but using different random number generators from the software. We conduct the same analysis for the 10 simulated studies using the multivariate model with stationarity constraints. All of these studies show that the commonly used method failed to detect causal effect for the second dataset, which is the one with the smallest amount of simulated causal impact.

2.6.3 Performance of the new method to infer causality

In this section, we study the performance of the new method. We use the same simulated data in Section 2.6.1. We calculate the one-sided KS distance in (2.5.1) and the threshold in (2.5.2) for each $i = 1, \dots, n$. We also calculate the one-sided KS distances

$$\sup_x \left[\frac{1}{k} \sum_{j=1}^k \left(\mathcal{F} \left(\sum_{t=T+1}^{T+m} \mu_{it} \leq x \mid \mathbf{Y}_{1:T}^{\text{obs}}, \mathbf{Y}_{T+1:T+m}^{\text{cf}(j)}, \mathbf{X}_{1:T+m} \right) \right) - \mathcal{F} \left(\sum_{t=T+1}^{T+m} \mu_{it} \leq x \mid \mathbf{Y}_{1:T+m}^{\text{obs}}, \mathbf{X}_{1:T+m} \right) \right],$$

and the corresponding thresholds for $m = 1$ to $m = P$. This allows to see how the KS distances grow over time.

We plot the results in Figure 2.5. There are five subplots in that figure with each represents one simulated dataset. For each subplot, the red line represents the one-sided KS distances between posteriors from a test store and its counterfactuals, and the light-blue line represents its corresponding thresholds. The threshold is calculated based on $k = 30$ simulated counterfactual datasets. In the plot, Dataset 1 is the only one with the one-sided KS distances completely below the thresholds and it is the dataset which does not receive any impacts. This suggests that our method has successfully distinguished between impact and no impact in these datasets. For Dataset 2, the impact at the early period is small, thus we observe the causal impact in the first three predicting periods are not significant; however, the new method can detect the impact after the fourth period.

We also summarized the results in Table 2.4. Compared with the results from the commonly used method (see Table 2.2), the new method shows a significant improvement in detecting causal impacts. From Dataset 3 to Dataset 5, the one-sided KS distances are all above their corresponding thresholds. Also, as the impact grows stronger, we observe that the distances becomes larger. The thresholds too increase along the time, since the predicting intervals for the trends become wider.

Table 2.4: Results of the one-sided KS distances and thresholds obtained by applying the new method to detect causal impacts in Dataset 1, . . . , Dataset 5 using the multivariate model with a stationary local linear trend. We only present the results at the dates March 22, 2016, March 31, 2016 and April. 9, 2016 which correspond to the 1st day, 10th day and 20th day during the causal period.

		Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
March, 22 (1st day)	KS distance	0.005	0.033	0.103	0.137	0.277
	Threshold	0.118	0.083	0.112	0.110	0.120
March, 31 (10th day)	KS distance	0.143	0.402	0.612	0.884	0.989
	Threshold	0.313	0.192	0.256	0.299	0.369
April, 9 (20th day)	KS distance	0.520	0.763	0.928	0.999	1.000
	Threshold	0.715	0.349	0.354	0.409	0.636

To check the performance of the new method, we conduct 10 more simulation studies

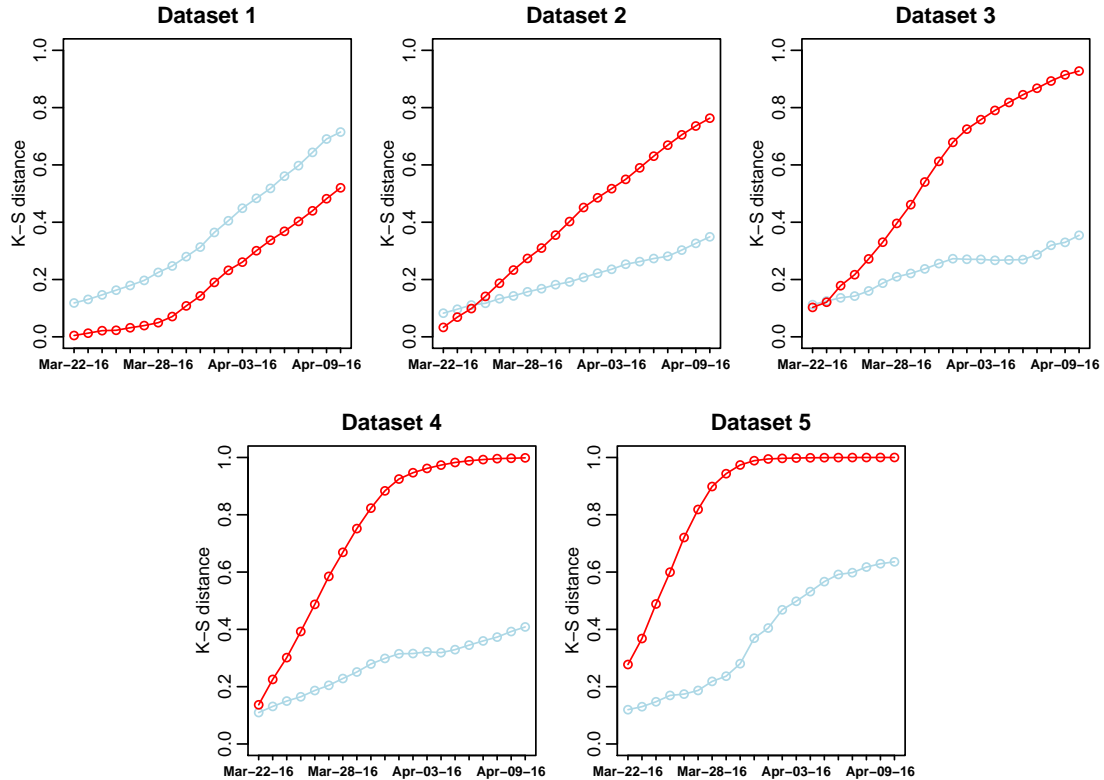


Figure 2.5: Results of applying the new method to detect causal impacts in Dataset 1, . . . , Dataset 5 using the multivariate model with a stationary local linear trend during the causal period from March, 22, 2016 to April, 9, 2016. In each subplot, the red line gives the one-sided KS distances between two posterior distributions with one is given the data of counterfactuals; the light blue line gives the corresponding thresholds.

using the data generated from the same model. Although the values of the one-sided KS distances and thresholds are not identical for each simulation, since the model is highly flexible and the estimated trend is sensitive to local changes of a dataset, the new method successfully detects the causal impacts in Dataset 2, . . . , Dataset 5.

We applied the new method to the univariate model, which is described in Section 2.6.2, using the same simulated dataset. The graphical and tabular representations of the results are presented in Section 3 of the supplementary material. We found that by comparing with the results obtained from the multivariate model (see Figure 2.5), the thresholds are much larger among all the datasets. Recall that from Table 2.3, the credible intervals estimated using the commonly used method are wider. Thus when we randomly

draw samples from a counterfactual with a larger variance, the posterior distributions for their trend are more apart. As a result, the pairwise one-sided KS distances between the posterior distributions of the trends are larger. Even though the thresholds are larger when using the univariate model, unlike the results obtained by using the commonly used method, the new method can still detect the causal impact for almost all the datasets which received an impact successfully, except for the very weak impacts in Dataset 2 during the first three periods and Dataset 3 during the first period.

2.7 Model checking

In this section, we present the convergence diagnostic results for MCMC chains. We also present the results for a sensitivity analysis. Furthermore, we discuss the choice of the threshold for declaring a significant impact.

2.7.1 Convergence diagnostic

We conduct an MCMC convergence diagnostic based on the simulated datasets we used above. Figure 2.6 gives the trace plots for several parameters including the trend for Dataset 1 at time $t = 1$, $\boldsymbol{\mu}_{1,1}$; the (1,1) coordinate element for covariance matrix $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_u$, $\boldsymbol{\Sigma}_v$, $\boldsymbol{\Sigma}_w$ and stationarity constraint matrix $\boldsymbol{\Phi}$. From Figure 2.6, we see that all the six chains mix well. For each chain, the burn-in periods are short and the chain remains stable after the burn-in period. Due the limitation of space, we do not show the trace plots for all the parameters of $\boldsymbol{\mu}_{1:T}$, $\boldsymbol{\tau}_{1:T}$, $\boldsymbol{\delta}_{1:T}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_u$, $\boldsymbol{\Sigma}_v$, $\boldsymbol{\Sigma}_w$ and $\boldsymbol{\Phi}$, but all of them have well-mixed chains.

We then use inefficient factors (IFs) to calculate the efficient draws for each MCMC chain. The formula is given as $1 + 2 \sum_{i=1}^M (1 - i/M) \rho_i$ (see Chib, 2001), where ρ_i is the estimated autocorrelation at lag i , M is the batch size which we take as 5000 in this simulation study. When we say that the inefficient factor has value M , this means that the number of effective draws is the total number of iterations without burn-in divided by M . Figure 2.7 plots the IFs for parameters in the model including all the parameters in the trend $\boldsymbol{\mu}_{1:T}$, the covariance matrices $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_u$, $\boldsymbol{\Sigma}_v$, $\boldsymbol{\Sigma}_w$, the intercept for the local linear trend \boldsymbol{D} and the parameters in the stationarity constraint matrix $\boldsymbol{\Phi}$. From the plots, we observe that all the IFs except for a few parameters in $\boldsymbol{\Phi}$ are very small which suggests that the correlations are low. The large inefficient factors for parameters in $\boldsymbol{\Phi}$

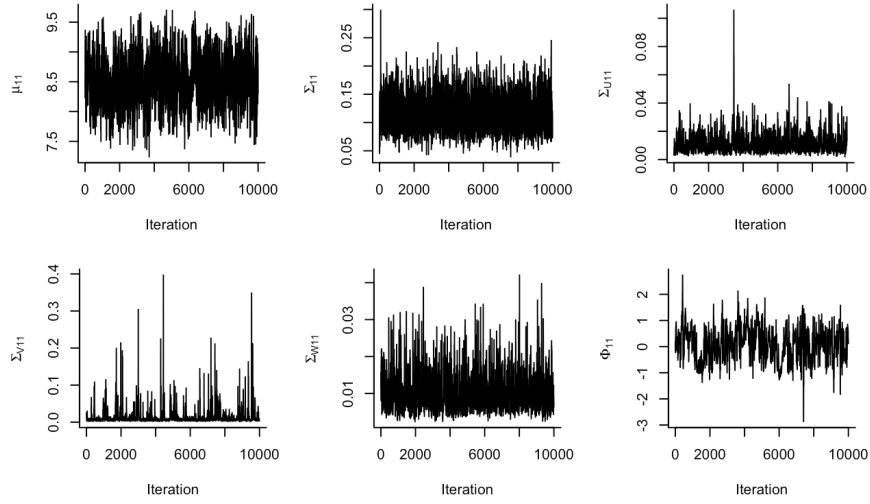


Figure 2.6: Traceplots for μ_{11} , Σ_{11} , Σ_{u11} , Σ_{v11} , Σ_{w11} , Φ_{11} .

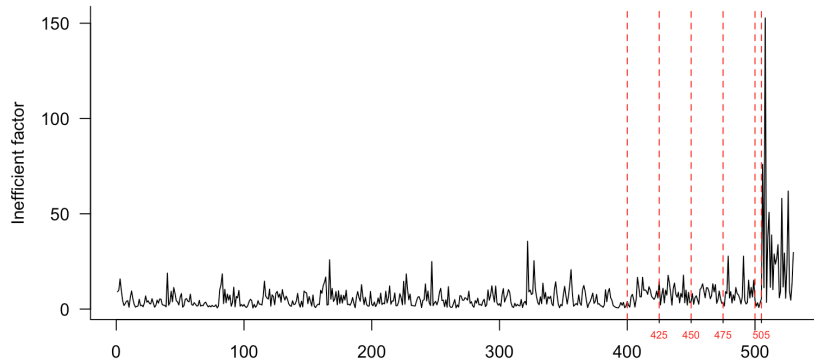


Figure 2.7: Plot of the inefficient factors (IFs) for $\mu_{1:T}$, Σ , Σ_u , Σ_v , Σ_w , D , Φ . The first 400 values are IFs for parameters in $\mu_{1:T}$ (5 datasets each with 80 time periods), the following 100 values are IFs for parameters in Σ , Σ_u , Σ_v , and Σ_w , with each has 25 parameters, the next 5 values are IFs for parameters in D ; and the last 25 values are IFs for parameters in Φ . The red lines separate IFs result from different parameters.

occur because the use of Metropolis-Hasting algorithm, which is expected to have a larger inefficient draws than using a Gibbs sampling algorithm.

2.7.2 Sensitivity analysis

Consider the following model,

$$\begin{aligned} \mathbf{Y}_t &= \mathbf{z}\boldsymbol{\alpha}_t + \mathbf{X}_t\boldsymbol{\beta} + \varrho_t\boldsymbol{\epsilon}_t, \\ \boldsymbol{\alpha}_{t+1} &= \mathbf{c} + \mathbf{T}\boldsymbol{\alpha}_t + \mathbf{R}\boldsymbol{\eta}_t. \end{aligned} \tag{2.7.1}$$

where ϱ_t s are random draws from an exponential distribution with mean equal to 1. Thus the error is heteroskedastic with a heavy tail. We generate the data from (2.7.1) but use the model (3.1) of our chapter to conduct the causal inference. The data generating process is similar to the one described in our chapter. We apply both the commonly used method and the new method to estimate the causal effects of the simulated datasets. The results for using the commonly used method are shown in Table 2.5 and the results for using the new method are shown in Table 2.6 and Figure 2.8.

From Table 2.5, the commonly used method fails to detect the causal impact for Dataset 2 and 4. Even though Dataset 4 has a relatively large impact. However, from Table 2.6 and Figure 2.8, we found that by using the new method, all the Datasets containing causal impact are successfully detected. Our method also does a good job on distinguishing the datasets containing causal impacts with the one not containing any causal impacts.

Table 2.5: Posterior medians and 95% credible intervals of average causal impacts for the model (2.7.1).

	Simulated impact	Posterior median	95% credible intervals
Dataset 1	0.00	-0.03	[-2.177, 2.233]
Dataset 2	1.06	1.81	[-0.421, 4.044]
Dataset 3	2.12	2.65	[0.520, 5.193]
Dataset 4	3.18	2.71	[-0.393, 5.102]
Dataset 5	4.23	4.28	[2.108, 6.000]

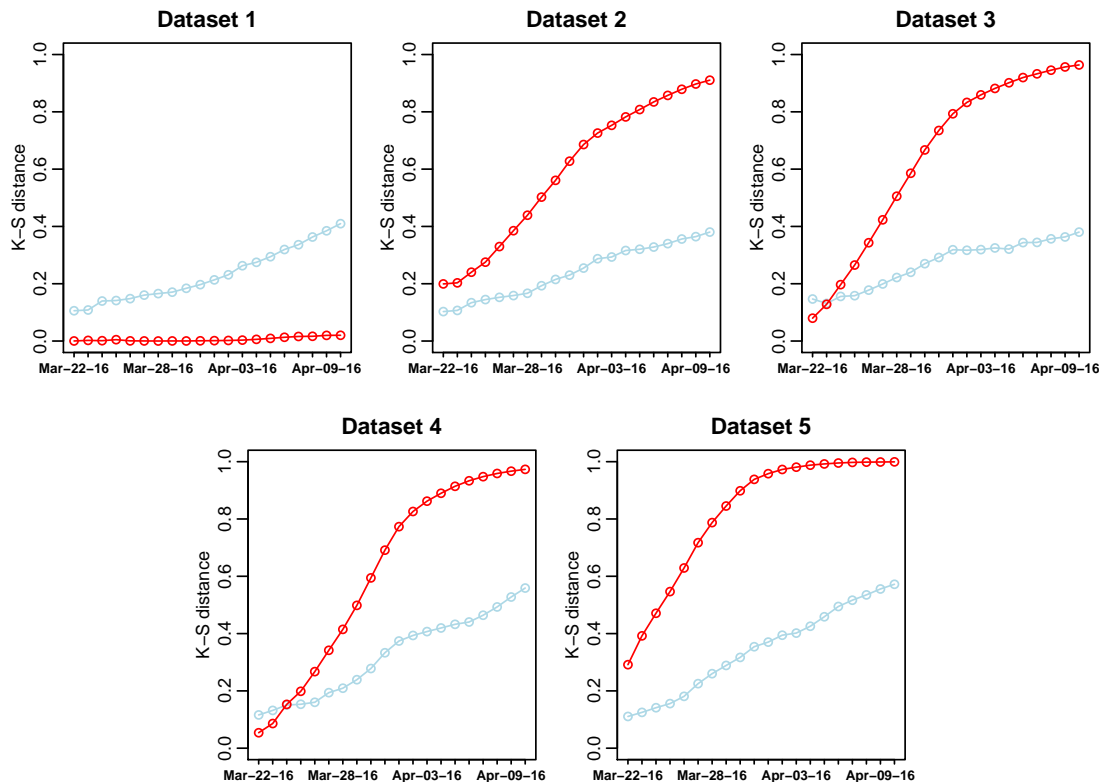


Figure 2.8: Results of applying the new method for detecting causal impacts in Dataset 1, \dots , Dataset 5 with the data generated from model (2.7.1) during the causal period from March, 22, 2016 to April, 9, 2016. In each subplot, the red line gives the one-sided KS distances between two posterior distributions with one is given the observed data and the other given the data of counterfactuals; the light blue line gives the corresponding thresholds.

Table 2.6: Results of the one-sided KS distances and thresholds obtained by applying the new method to detect causal impacts in Dataset 1, \dots , Dataset 5 using the model (2.7.1) with the stationarity constraint. We only present the results at the dates March 22, 2016, March 31, 2016 and April, 9, 2016 which correspond to the 1st day, 10th day and 20th day during the causal period.

		Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
March, 22 (1st day)	KS distance	0.000	0.199	0.080	0.054	0.292
	Threshold	0.106	0.103	0.147	0.116	0.111
March, 31 (10th day)	KS distance	0.001	0.628	0.735	0.691	0.938
	Threshold	0.197	0.230	0.292	0.333	0.354
April, 9 (20th day)	KS distance	0.020	0.911	0.964	0.973	0.999
	Threshold	0.410	0.380	0.380	0.559	0.572

2.7.3 Using a threshold by choosing a different percentile

In our study, we proposed using the 95% upper percentile to calculate the threshold for detecting a significant impact. If we choose a higher percentile, it becomes more difficult to detect an impact. On the other hand, it prevents false discoveries. We then conduct a study by using the 99% upper percentile for the same simulated datasets. We plot the results in Figure 2.9. Compared with Figure 2.5, the thresholds in Figure 2.9 are randomly larger for all the subplots. However, we can still successfully detect the causal impacts in all the datasets.

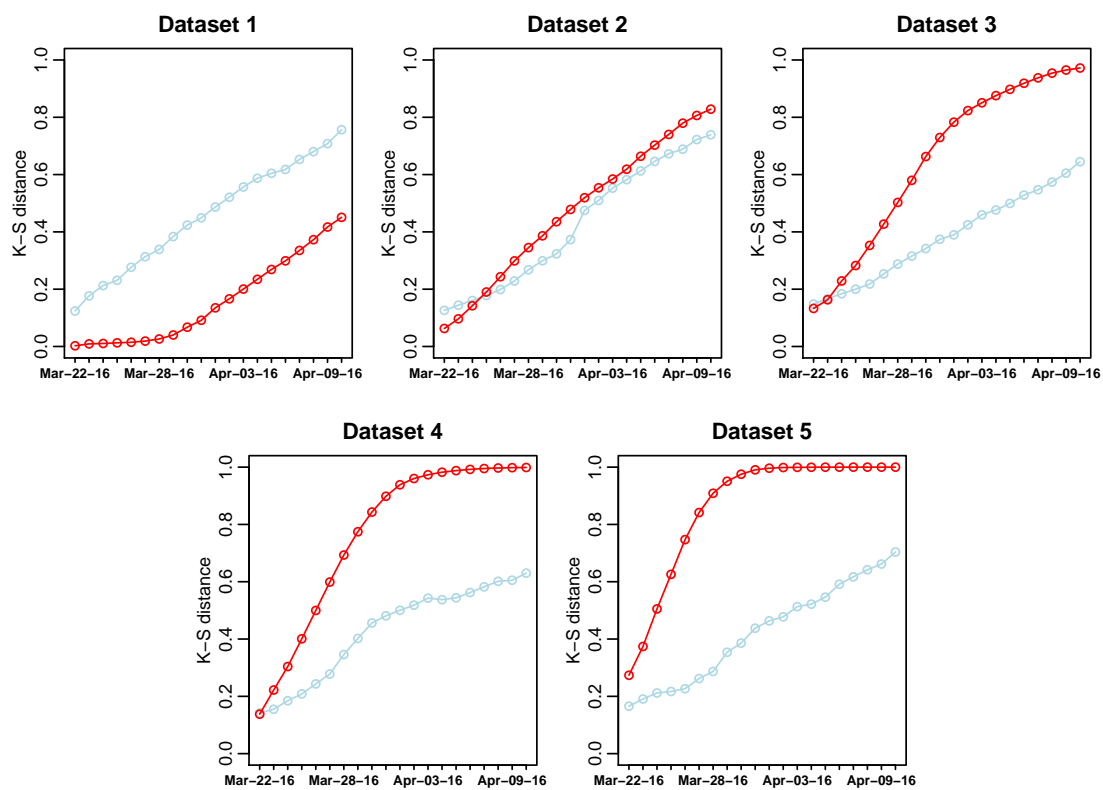


Figure 2.9: Results of applying the new method for detecting causal impacts in Dataset 1, . . . , Dataset 5 during the causal period from March, 22, 2016 to April, 9, 2016 with thresholds chosen as the 99% upper percentile from the one-sided KS distances obtained from 30 generated counterfactuals. In each subplot, the red line gives the one-sided KS distances between two posterior distributions with one is given the observed data and the other is given their corresponding counterfactuals; the light blue line gives the corresponding thresholds.

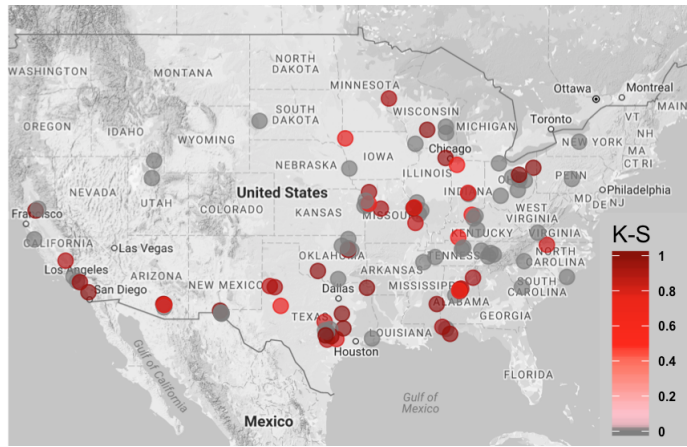
2.8 Application to a real dataset

In this section, we present the results of a real data analysis for measuring the causal impact of an online advertising campaign (run by *MaxPoint*) for a consumer product at a large national retail chain.

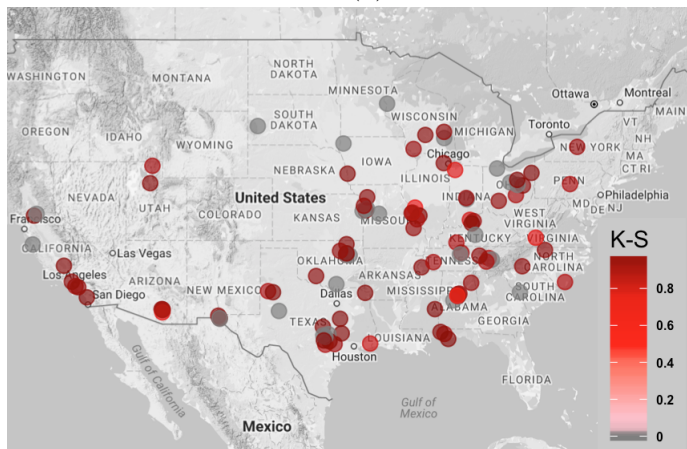
Due to commercial confidentiality, we do not show full details of the results, but the following description explains how our method works in this real dataset. *MaxPoint* targets this campaign at 627 test stores and 318 control stores spread out across the country and collects weekly data throughout the campaign. We choose all the control stores in the corresponding state for each dataset. If a state does not at all have control stores, we remove such data from the analysis. In Stage 1, we use the DAEMVS (with $s = 0.1$) algorithm to select the control stores for each test store. If for a test store, all the potential control stores are eliminated by the DAEMVS algorithm we also eliminate that store from the causal analysis, because without building a counterfactual, the causal inference cannot be conducted. After making the selection, we conduct the causal analysis on 323 test stores in total. For each dataset, there are 46 weekly observations in total with the last 10 observations occurring in the causal period. Since the length before the causal period is only 35 per dataset, we have to separate these 323 stores into smaller datasets and fit the model separately on them. As large national chain retailers organize promotional and operations activity differently in each state, we treat stores in different states as independent. State-wise splitting typically keeps the number of stores less than 15. If one state has more than 15 stores, we split further into subregions to meet the requirement. We further assume that the stores in two different subregions behave independently. The regions are separated based on city boundaries. Within each region, we assume that stores are connected with each other. This means that the inverse covariance matrix (equivalently, the covariance matrix) follows a block-diagonal structure with at most 15 nodes in a block.

We assume the three causal assumptions in Section 2.2 hold. The following table summarizes the number of stores with significant causal effects from the advertising campaign. From the table we found that the number of stores are increasing from the first week to the last week. During the first five weeks, the number of stores that received causal impact increased rapidly compared with that in the last five weeks.

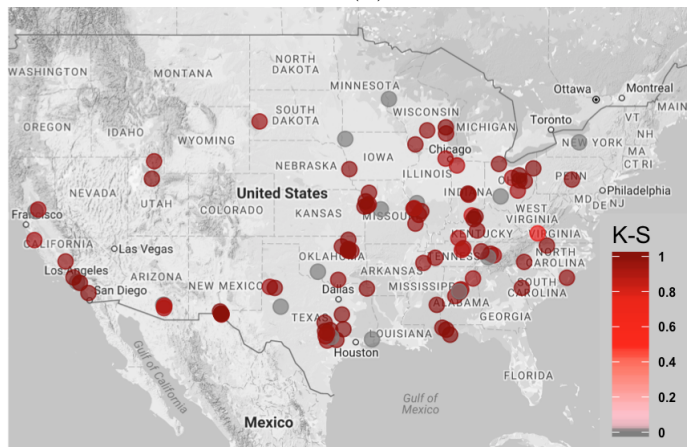
Not only the number of impacted stores increased during the advertising campaign period (shown in Table 2.7), the magnitudes of the impacts in those stores also increased.



(a)



(b)



(c)

Figure 2.10: Plot of the causal impacts at test stores at end of the second week (a), the fifth week (b) and the last week (c) for an advertising campaign of a consumer product at a large national retail chain. The impacts below their thresholds are set to zero. The United States map is produced using Google Maps, 2017.

Table 2.7: Number of test stores that received significant causal impacts for each week of running the advertisement campaign by using the multivariate model with a stationary local linear trend.

	1st week	2nd week	3rd week	4th week	5th week
Number of stores	23	44	55	62	73
	6th week	7th week	8th week	9th week	10th week
Number of stores	72	77	78	82	84

In Figure 2.10, we plot the estimated one-sided KS distances for stores along with their locations at Weeks 2, 5 and 10. In each figure, we plot only the stores with significant causal effects. The red dots represent the stores with the one-sided KS distances larger than their corresponding thresholds, which suggests that those stores received significant causal effects. The grey dots represent the stores that do not show significant causal effects. We find that the magnitudes of the impacts for most of the stores have a larger increase from the first five weeks compared with the last five weeks. Comparing the plots of the fifth week and the tenth week, we find that only a few stores in California, South Dakota, Ohio and Texas got increased causal effects.

We also conduct an analysis by assuming that the test stores are independent and thus ignoring their spatial correlation. Table 2.8 lists the number of stores that received significant causal effects. The numbers are smaller than those obtained using the multivariate model. This suggests most of the impacts are weak and the spatial correlation between sales in different stores help detect the weaker impacts.

Table 2.8: Number of test stores that received significant causal impacts for each week of running the advertisement campaign by using the univariate model.

	1st week	2nd week	3rd week	4th week	5th week
Number of stores	25	19	22	23	18
	6th week	7th week	8th week	9th week	10th week
Number of stores	17	15	15	13	14

2.9 Conclusion and discussion

In this chapter, we proposed a novel causal inference method which compares the posterior distributions of the latent trend conditional on two different sets of data: one is the observed data which contain a causal effect; the other one is the data from a synthetic control. We calculated the one-sided KS test statistics between the two posterior distributions. A threshold was used to decide whether a causal impact is significant or not. In the simulation study, we showed that our method can detect a smaller sized causal impact more efficiently compared with the commonly used method even when the model is slightly misspecified. The new causal inference method is not restricted to the specific structural time series model used in this chapter and can be applied to many other models in different applications.

We used a multivariate structural time series model to estimate the causal impact of a stimulus on subjects such as an advertising campaign for each individual store. Sales in those stores are spatially correlated. A Bayesian analysis was used to estimate parameters in this model. We imposed sparsity on the precision matrix based on the distance between each pair of stores. The sparsity was imposed through a \mathcal{G} -Wishart prior, where the graph \mathcal{G} can be either decomposable or non-decomposable. We restricted the hidden process τ_t to be stationary in order to stabilize the prediction intervals. To sample its time-varying variables, we used the Kalman filter and simulation smoother algorithm. This algorithm can be used to impute missing values inside the MCMC loops.

We used the revised EMVS algorithm to select control stores. We also discussed the advantage of using the DAEMVS algorithm which is a modified version of the EMVS algorithm. Compared to the EMVS algorithm, the DAEMVS algorithm reduces the chance of getting trapped at a local maximum. Both the EMVS and DAEMVS algorithms are computationally much faster than the sampling based method like SSVS. Since the EMVS algorithms cannot be incorporated into MCMC loops, we proposed a two-stage algorithm to estimate parameters. In Stage 1, we used the DAEMVS to obtain $\hat{\beta}$; in Stage 2, we plugged-in $\hat{\beta}$ and used an MCMC algorithm to obtain posterior distributions of the remaining parameters.

We compared the multivariate model with the univariate model which assumes independence between responses based on simulated datasets. The results indicate that the univariate model gives wider credible intervals (if using the commonly used method) and larger threshold (if using the new method) than the multivariate model. Thus incorpo-

rating of the spatial relationships between test stores is beneficial.

Finally, we analyzed a real dataset on sales data of products distributed through brick and mortar retail stores for an advertising campaign run by *MaxPoint*. Even though, due to commercial confidentiality, we did not provide the full details of the results, the summarization tables of the number of stores that received significant impact suggests the effectiveness of using the new causal inference method.

Chapter 3

Bayesian Linear Regression for Multivariate Responses Under Group Sparsity

3.1 Introduction

Asymptotic behavior of variable selection methods, such as the lasso, have been extensively studied (Bühlmann and van der Geer, 2011). However, theoretical studies on Bayesian variable selection methods are limited to relatively simple settings (Castillo et al., 2015; Chae et al., 2016; Martin et al., 2017; Ročková, 2018; Belitser and Ghosal, 2017; Song and Liang, 2017). For example, Castillo et al. (2015) studied a sparse linear regression model in which the response variable is one-dimensional and the variance is known. However, it is not straightforward to extend those results to study the multivariate linear regression models with unknown covariance matrix (or even the univariate case with unknown variance).

In many applications, predictors are naturally clustered in groups. Below, we give three examples.

1. *Cancer genomics study*. It is important for biologists to understand the relationship between clinical phenotypes and DNA mutations, which are detected by DNA sequencing. Since these mutations are spaced linearly along the DNA sequence, it is often assumed that the adjacent DNA mutations on the chromosome have a similar genetic effect and thus should be grouped together (Li and Zhan, 2010).

2. *Multi-task learning.* When information for multiple tasks is shared, it is preferable to solve these tasks at the same time to improve learning efficiency and prediction accuracy. Relevant information is preserved across different equations by grouping them together (Lounici et al., 2009).
3. *Causal inference in advertising.* Measuring the effectiveness of an advertising campaign running on stores is an important task for advertising companies. Counterfactuals, which are constructed using the sales data of a few stores, chosen by a variable selection method, from a large number of control stores not subject to the advertising campaign, are needed to conduct a causal analysis (see Chapter 2). Control stores within the same geographical region—as they share the same demographic information—can be grouped together and selected or not selected at the same time.

Driven by these applications, new variable selection methods designed to select or not select variables as groups, through imposing *group sparsity* on the regression coefficients, have been developed. For example, the group lasso method is proposed (Yuan and Lin, 2006). It replaces the ℓ_1 -norm in the lasso with the ℓ_2/ℓ_1 -norm, where the ℓ_2 -norm is put on the predictors within each group and the ℓ_1 -norm is put across the groups. Theoretical properties of the group lasso have been studied (Nardi and Rinaldo, 2008) and its benefits over the lasso in the group selection problem have been demonstrated (Lounici et al., 2009, 2011; Huang and Zhang, 2010). Recently, various Bayesian methods for selecting variables as groups were also proposed (Li and Zhan, 2010; Curtis et al., 2014; Ročková and George, 2014; Xu and Ghosh, 2015; Chen et al., 2016; Greenlaw et al., 2017; Lique et al., 2017). However, large-sample frequentist properties of these Bayesian methods have not been studied yet.

In this chapter, we study a Bayesian method for the multivariate linear regression model with two distinct features: group sparsity that is imposed on the regression coefficients and an unknown covariance matrix. To the best of our knowledge, even in a simpler setting without assuming group sparsity, convergence and selection properties of methods for high-dimensional regression with a multivariate response having an unknown covariance matrix have not been studied in either the frequentist or the Bayesian literature. However, it is important to understand the theoretical properties of those models because correlated responses arise in many applications. For example, in the study of the causal effect of an advertising campaign, sales in different stores are often spatially

correlated (Ning et al., 2018). Furthermore, when the dimension of the covariance matrix is large, it would affect the quality of the estimation of the regression coefficients.

When the covariance matrix is unknown and high-dimensional, the techniques that were developed for deriving posterior concentration rates (Castillo et al., 2015; Martin et al., 2017; Belitser and Ghosal, 2017) cannot be applied. Also, The general theory of posterior concentration (Ghosal and van der Vaart, 2017) in its basic form is not appropriate to use because it typically deals with the average Hellinger distance which is not sufficient for our analysis. Thus in order to apply the general theory to derive a rate, we shall construct required tests directly by controlling the moments of likelihood ratios in small pieces.

In this study, we consider a multivariate linear regression model

$$Y_i = \sum_{j=1}^G X_{ij}\beta_j + \varepsilon_i, \quad i = 1, \dots, n, \quad (3.1.1)$$

where Y_i is a $1 \times d$ response variable, $i = 1, \dots, n$, X_{ij} is a $1 \times p_j$ predictor variable, $j = 1, \dots, G$, β_j is a $p_j \times d$ matrix containing the regression coefficients, and $\varepsilon_1, \dots, \varepsilon_n$ are independent identically distributed (i.i.d) as $\mathcal{N}(0, \Sigma)$ with Σ being a $d \times d$ unknown covariance matrix. In other words, in the regression model, there are $G > 1$ non-overlapping groups of predictor variables with the group structure being pre-determined. When $G = p$, it reduces to the setting that the sparsity is imposed on the individual coordinates. Thus the results derived in our chapter are applicable to the ungrouped setting as well. The model can be rewritten in the vector form as

$$Y_i = X_i\beta + \varepsilon_i, \quad (3.1.2)$$

where $\beta = (\beta'_1, \dots, \beta'_G)$ is a $p \times d$ matrix, where $p = \sum_{j=1}^G p_j$, and $X_i = (X_{i1}, \dots, X_{iG})$ is a $1 \times p$ vector. The dimension p can be very large. The dimension d can be large as well to a lesser extent when the sample size is large. The number of total groups G is clearly bounded by p . We denote the groups which contain at least a non-zero coordinate as *non-zero groups* and the remaining groups as *zero groups*.

To allow derivation of asymptotic properties of estimation and selection, certain conditions on the growth of p , G , d and p_1, \dots, p_G need to be imposed. We allow $p \gg n$ (which means that $n/p \rightarrow 0$) but require that the total number of the coefficients in all

non-zero groups together are less than n in order. We further assume that the number of coordinates in any single group must be of order less than p and that $\log G \ll n$. Finally, to make the covariance matrix is consistently estimable, we assume that the dimension d of the covariance matrix satisfies the condition that $d^2 \log n \ll n$.

As for the priors, we choose a product of d independent spike-and-slab priors for β and a Wishart prior for Σ^{-1} , the precision matrix. The spike-and-slab prior is a mixture of point mass for the zero coordinates and a density for non-zero coordinates. In the ungrouped setting, commonly used densities for non-zero coordinates are a Laplace density (Castillo et al., 2015), a Cauchy density (Castillo and Misner, 2018) and a normal density with mean chosen by empirical Bayes methods (Martin et al., 2017; Belitser and Ghosal, 2017). In this chapter, we choose a special density for the non-zero coordinates (see (3.3.1)). This density involves the ℓ_2/ℓ_1 -norm, which is used as a penalty to obtain the group lasso in a non-Bayesian setting.

The remainder of the chapter is organized as follows. Section 3.2 introduces notations that will be used in this chapter. Section 3.3 describes the priors, along with the necessary assumptions. Section 3.4 provides the main results. The proofs of the main results are given in Section 3.5. Auxiliary lemmas are provided in Section 3.6.

3.2 Notation.

We assume that $\mathcal{G}_1, \dots, \mathcal{G}_G$ are G disjoint groups such that $\cup_{j=1}^G \mathcal{G}_j = \{1, \dots, p\}$. Since these groups are given and will be kept the same throughout, their notations will be dropped from subscription notations. Clearly, p_j is the number of elements in \mathcal{G}_j . Let $p_{\max} = \max_{1 \leq j \leq G} p_j$. For each $k = 1, \dots, d$, let $S_k \subseteq \{1, \dots, G\}$ stand for the set which contains the indices of the non-zero groups for the k -th component and $s_k = |S_k|$ be the cardinality of the set S_k . Also, define $S = \cup_{k=1}^d S_k$ and $s = \sum_{k=1}^d s_k$. Let $S_{0,k}$ be the set containing the indices of the true non-zero groups, where $S_{0,k} \subseteq \{1, \dots, G\}$. Define $p_{S_{0,k}} = \sum_{j \in S_{0,k}} p_j$, and $p_{S_0} = \sum_{k=1}^d p_{S_{0,k}}$.

For a vector A , let $\|A\|_1$, $\|A\|_{2,1}$ and $\|A\|$ be the ℓ_1 -, ℓ_2/ℓ_1 - and ℓ_2 -norm of A respectively, where $\|A\|_{2,1} = \sum_{j=1}^G \|A_j\|$ with A_j is the submatrix of A consisting of $k \in \mathcal{G}_j$ coordinates. For a $d \times p$ matrix \mathbf{B} , we denote B_k as the k -th column of \mathbf{B} and $\|\mathbf{B}\|_F = \sqrt{\text{Tr}(\mathbf{B}^T \mathbf{B})}$ as the Frobenius norm of \mathbf{B} . For a $d \times d$ symmetric positive definite matrix \mathbf{C} , let $\text{eig}_1(\mathbf{C}), \dots, \text{eig}_d(\mathbf{C})$ denote the eigenvalues of \mathbf{C} ordered from the

smallest to the largest and $\det(\mathbf{C})$ stand for the determinant of \mathbf{C} . For a scalar c , we denote $|c|$ to be the absolute value of c .

Let $\rho(f, g) = -\log(\int f^{1/2}g^{1/2}d\nu)$ be the negative log-affinity between densities f and g and $h^2(f, g) = \int (f^{1/2} - g^{1/2})^2 d\nu$ be their squared Hellinger distance. The Kullback-Leibler divergence between f and g is given by $K(f, g) = \int f \log(f/g)$ and the Kullback-Leibler variation between f and g is denoted by $V(f, g) = \int f(\log(f/g) - K(f, g))^2$. The symbol f_0 stands for the density of f with the parameters at their true values. The notation $\|\mu - \nu\|_{TV}$ denotes the total variation distance between two probability measures μ and ν .

We let $N(\epsilon, \mathcal{F}, \rho)$ stand for the ϵ -covering number of a set \mathcal{F} with respect to ρ , which is the minimal number of ϵ -balls needed to cover the set \mathcal{F} . Let \mathbf{I}_d stand for the d dimensional identity matrix and $\mathbb{1}$ stand for the indicator function.

The symbols \lesssim and \gtrsim will be used to denote inequality up and down to a constant while $a \asymp b$ stand for $C_1 a \leq b \leq C_2 a$ for two constants C_1 and C_2 . The notations $a \ll b$ and $a \vee b$ stand for $a/b \rightarrow 0$ and $\max\{a, b\}$ respectively. The symbol $\delta_0(\cdot)$ stands for a Dirac measure.

3.3 Prior specifications

In this section, we introduce the priors used in this study. We place two independent priors on $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ as they are both unknown. We place d products of independent spike-and-slab priors on $\boldsymbol{\beta}$ and a Wishart prior on $\boldsymbol{\Sigma}^{-1}$, which is known as the precision matrix.

3.3.1 Prior for regression coefficients

We denote the k -th column of $\boldsymbol{\beta}$ as β_k and the notations β_{S_k} and $\beta_{S_k^c}$ as collections of the regression coordinates in the non-zero groups and the zero groups respectively. Each spike-and-slab prior is constructed as follows. First, a dimension s_k is chosen from a prior π_G on the set $\{0, 1, \dots, G\}$. Next, a subset S_k of cardinality s_k is randomly chosen from the set $\{1, \dots, G\}$. Finally, A vector $\beta_{S_k} = \{\beta_j \mathbb{1}(j \in S_k)\}$ is chosen from a probability density g_{S_k} on \mathbb{R}^{s_k} given by (3.3.3). The remaining coordinates $\beta_{S_k^c}$ set to 0. To summarize, the

prior for β is

$$\pi(S_1, \dots, S_d, \beta) = \prod_{k=1}^d \pi(S_k, \beta_k), \quad (3.3.1)$$

where $\pi(S_k, \beta_k) = \pi_G(|S_k|) \frac{1}{\binom{G}{|S_k|}} g_{s_k}(\beta_{S_k}) \delta_0(\beta_{S_k^c})$ and the density $\pi_G(|S_k|)$ is the prior for the dimension $s_k = |S_k|$.

Assumption 1 (Prior on dimension). *There are positive constants A_1, A_2, A_3, A_4 with*

$$A_1 G^{-A_3} \pi_G(s_k - 1) \leq \pi_G(s_k) \leq A_2 G^{-A_4} \pi_G(s_k - 1), \quad (3.3.2)$$

for $s_k = 1, \dots, G, k = 1, \dots, d$.

For example, the complexity prior given by [Castillo et al. \(2015\)](#) by replacing p by G shall satisfy the above assumption.

The Laplace density ([Castillo et al., 2015](#)) or the Cauchy density ([Castillo and Misner, 2018](#)) are generally chosen as g , since the normal density has too sharp tail that overshrinks the non-zero coefficients, although some empirical Bayes modifications of the mean can overcome the issue (see [Martin et al., 2017](#); [Belitser and Ghosal, 2017](#)). However, in our setting, as sparsity is imposed at the group level, a more natural choice of the prior is the density which incorporates the ℓ_2/ℓ_1 -norm. We thus consider the prior

$$g(\beta_k) = \left(\prod_{j=1}^G \left(\frac{\lambda_k}{a_j} \right)^{p_j} \right) \exp(-\lambda_k \|\beta_k\|_{2,1}), \quad (3.3.3)$$

where $a_j = \sqrt{\pi} \left(\frac{\Gamma(p_j + 1)}{\Gamma(p_j/2 + 1)} \right)^{1/p_j} \geq 2$ (see [Lemma 3.6.1](#)). This density has its tail lighter than the corresponding Laplace density. From Stirling's approximation, it follows that $a_j = O(p_j^{1/2})$. We would like to mention that [Xu and Ghosh \(2015\)](#) developed a posterior computational strategy for a similar prior. They also incorporated the ℓ_2/ℓ_1 -norm into their prior, except for that they did not provide the explicit expression of the normalizing constant for that prior.

The tuning parameter λ_k needs to be bounded both from above and below. The value of λ_k cannot be too large or it will shrink the non-zero coordinates too much towards 0. It should not be too small because a very small value will be unable to prevent many false signals appear in the model and hence making the posterior to contract slower. The

upper and lower bounds for the permissible limits are stated in below.

Assumption 2. For each $k = 1, \dots, d$, $\underline{\lambda}_k \leq \lambda_k \leq \bar{\lambda}$, where

$$\bar{p}_{S_{0,k}} = \frac{\sum_{j \in S_{0,k}} p_j}{s_{0,k}}, \quad \underline{\lambda}_k = G^{-1/\bar{p}_{S_{0,k}}} \sqrt{\sum_{i=1}^n \|X_i\|^2}, \quad \bar{\lambda} = 3 \sqrt{\sum_{i=1}^n \|X_i\|^2 \log G}. \quad (3.3.4)$$

The lower bound of λ_k is derived from (3.5.12). Suppose that $G = p$ (when each group has only one element), the lower bound reduces to $\underline{\lambda}_k = \sqrt{\sum_{i=1}^n \|X_i\|^2}/p$, which is analogous to the lower bound displayed in Castillo et al. (2015).

The upper bound of λ_k is motivated from the following lemma.

Lemma 3.3.1. Under Assumption 2,

$$\mathbb{P}_0 \left(\sum_{i=1}^n \|X_i'(Y_i - X_i \beta_0) \Sigma_0^{-1/2}\|_F^2 \geq d \bar{\lambda}^2 \right) \rightarrow 0. \quad (3.3.5)$$

3.3.2 Prior for the covariance matrix

We put a Wishart prior on the precision matrix: $\Sigma^{-1} \sim \mathcal{W}_d(\nu, \Phi)$, where \mathcal{W}_d stands for a d -dimensional Wishart distribution, Φ is a symmetric positive definite matrix, $\nu > d - 1$, $\nu \asymp d$, and $d \rightarrow \infty$.

Although, other priors can be used, the Wishart prior is the most commonly used prior in practice as it is conjugate for the multivariate normal likelihood.

3.4 Main results

3.4.1 Posterior contraction rate

We study the posterior contraction rate for the model (3.1.1) and the priors given in Section 3.3. We write β_0 and Σ_0 for the true values for β and Σ respectively. Recall that $S_{0,k}$ is the set which includes the index of the true non-zero groups of β_k , $s_{0,k} = |S_{0,k}|$ is the cardinality of that set $S_{0,k}$. Let $S_0 = \cup_{k=1}^d S_{0,k}$ and $s_0 = \sum_{k=1}^d s_{0,k}$. Define the set $\mathcal{S} = \{S : |S_k| = s_{0,k}, k = 1, \dots, d\}$.

The general theory of posterior contraction for independent non-identically distributed observations (see Theorem 8.23 of Ghosal and van der Vaart, 2017) is often used to derive

a posterior contraction rate, which is based on the average squared Hellinger distance. Since the average squared Hellinger distance between multivariate normal densities with an unknown covariance is small does not necessarily imply that the parameters in the two densities are also close on average, we work directly with on the average negative log-affinity which is still very tractable in the multivariate normal setting.

To derive a posterior contraction rate, we construct a suitable test from the first principle by breaking up the effective parameter space given by a sequence of appropriate sieves under the alternative hypothesis into several pieces. For each piece sufficiently separated from the truth, we pick up a representative and obtain a most powerful test (i.e., the Neymann-Pearson test) for the truth against that alternative. We bound the moments of the likelihood ratio of an arbitrary density in the piece to the density of the representative of that piece to show that the most powerful test for the truth against the representative has adequate power for any alternatives in the corresponding piece.

By using this approach, we require the true values of β_0 and Σ_0 to be restricted into certain regions to ensure that the prior concentration around the true point is not too small so that the posterior contraction rate is sufficiently fast. This is unlike [Castillo et al. \(2015\)](#), who obtained results uniformly over the whole space as their case (univariate with known variance and Laplace prior) allows explicit expressions for a direct treatment. More precisely, we require $\beta_0 \in \mathcal{B}_0$ and $\Sigma_0 \in \mathcal{H}_0$, where \mathcal{B}_0 and \mathcal{H}_0 are shown in the following assumption.

Assumption 3. *The true values of $\beta_0 \in \mathcal{B}_0$ and $\Sigma_0 \in \mathcal{H}_0$, where*

$$\mathcal{B}_0 = \{\beta : \max_{1 \leq k \leq d} \|\beta_k\|_{2,1} \leq \bar{\beta}\}, \quad \mathcal{H}_0 = \{\Sigma : b_1 \mathbf{I}_d \leq \Sigma \leq b_2 \mathbf{I}_d\}, \quad (3.4.1)$$

b_1, b_2 are two fixed positive values, $\bar{\beta} = \frac{s_0(\log G \vee p_{\max} \log n)}{d \max_{1 \leq k \leq d} \lambda_k}$, ϵ_n is given in [\(3.4.4\)](#) and λ_k satisfies [Assumption 2](#).

The largest value of $\bar{\beta}$ is obtained is by taking $\lambda_k = \underline{\lambda}_k$ for all k and $G = p$, which is the ungrouped case. Then the upper bound becomes $\bar{\beta} = \frac{ps_0 \log p}{d \sqrt{\sum_{i=1}^n \|X_i\|^2}}$. For the sake of simplicity, we assume that $n^{-1} \sum_{i=1}^n \|X_i\|^2$ is bounded above by a large constant. Then that upper bound increases to infinity very quickly since $p \gg n$. When $G \ll p$, the upper bound increases at a slower rate than the bound when $G = p$, as p increases.

Theorem 3.4.1. For the model (3.1.1) and the priors given in Section 3.3, suppose that $\sum_{i=1}^n \|X_i\|^2 \leq nb_3$ for a fixed positive number b_3 , $d^2 \log n \ll n \ll p$, $s_0 p_{\max} \log n \ll n$, where $p_{\max} = \max(p_1, \dots, p_G)$, and that Assumptions 1–3 hold. Then, for $M_1 > 0$ sufficiently large,

$$\Pi\left(\boldsymbol{\beta} : \sum_{i=1}^n \|X_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|^2 \geq M_1 n \epsilon_n^2 \middle| Y_1, \dots, Y_n\right) \rightarrow 0, \quad (3.4.2)$$

$$\Pi\left(\boldsymbol{\Sigma} : \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_0\|_F^2 \geq M_1 \epsilon_n^2 \middle| Y_1, \dots, Y_n\right) \rightarrow 0, \quad (3.4.3)$$

where

$$\epsilon_n = \left\{ \sqrt{\frac{s_0 \log G}{n}} \vee \sqrt{\frac{s_0 p_{\max} \log n}{n}} \vee \sqrt{\frac{d^2 \log n}{n}} \right\}. \quad (3.4.4)$$

Remark 1. A major contribution we make to prove this theorem is the construction of exponentially powerful tests for the truth against the complement of a ball by splitting the complement in suitable pieces (not necessarily balls) where we can control a moment of the likelihood ratio for two points within each piece. This gives a general technique of construction of tests required for the application of the general theory, which can be useful in many other problems.

Remark 2. Instead of using the prior given in (3.3.3), one can also choose a Laplace density for the coordinates in the non-zero groups. Then the ℓ_2/ℓ_1 -norm of $\beta_{0,k}$, $\|\beta_{0,k}\|_{2,1}$, in the set \mathcal{B}_0 should be replaced by $\|\beta_{0,k}\|_1$. Clearly, $\|\beta_{0,k}\|_{2,1} \leq \|\beta_{0,k}\|_1$, hence in the latter case the set \mathcal{B}_0 will be smaller. In fact, one can replace the ℓ_2/ℓ_1 -norm with any other ℓ_q/ℓ_1 -norms, for $1 \leq q \leq \infty$. Then the norm of $\beta_{0,k}$ in \mathcal{B}_0 needs to be adjusted accordingly.

Remark 3. When $G = p$, the rate reduces to $\epsilon_n = \{ \sqrt{(s_0 \log p)/n} \vee \sqrt{(d^2 \log n)/n} \}$. The first part of the rate is the same as the rate obtained when the sparsity is imposed at the individual level, such as in Bühlmann and van der Geer (2011) and Castillo et al. (2015). When $G \ll p$, the first rate can be obtained when the ratio $s_0 p_{\max} \log n = O(s_0 \log G)$ (i.e., when the number of coordinates in each group takes a fixed number) and d is sufficiently slowly growing.

Remark 4. The second rate in (3.4.4) reveals that in some situations, by imposing group sparsity, the posterior will contract at a slower rate than imposing sparsity at the

individual level. This happens when too many zeros are put into non-zero groups (often known as *weakly group-sparse* (Huang and Zhang, 2010)).

From Theorem 3.4.1, if the dimension of the covariance is too large, then the posterior contraction rate can be much slower. Under such a situation, the rate can be improved if we know any special structures for the precision matrix. Here we give two examples.

Example 1 (Independent responses). If the responses are independent across components, then the model (3.1.1) can be written as d independent model with each one is

$$\frac{1}{\sigma_k} Y_{ik} = \frac{1}{\sigma_k} X_i \beta_k + \xi_{ik}, \quad \xi_{ik} \sim \mathcal{N}(0, 1).$$

Then one can estimate the parameters in the d models separately. The posterior concentration rate for each corresponding posterior becomes $\epsilon_n = \sqrt{\sum_{k=1}^d \epsilon_{n,k}^2}$, where $\epsilon_{n,k} = \left\{ \sqrt{\frac{s_{0,k} \log G}{n}} \vee \sqrt{\frac{s_{0,k} p_{\max} \log n}{n}} \right\}$, $k = 1, \dots, d$.

Example 2 (Sparse precision matrix). The third rate in ϵ_n may be improved if the precision matrix is appropriately sparse. Banerjee and Ghosal (2014) showed that when the matrix has an exact banding structure with banding size k , then using an appropriate \mathcal{G} -Wishart prior, the posterior for the precision matrix Σ^{-1} contracts at the rate $k^{5/2}(\log n/n)^{1/2}$ with respect to the spectral norm. When the sparsity does not possess a specific structure, Banerjee and Ghosal (2015) showed that the rate reduces from d/\sqrt{n} to $\sqrt{(d+m) \log d/n}$ with respect to the Frobenius norm, where m is the number of non-zero off-diagonal elements.

As a consequence of posterior contraction near the truth, the following estimate is easily obtained (see page 200 of Ghosal and van der Vaart, 2017).

Lemma 3.4.2. For positive constants C_1 and C_2 , and the rate ϵ_n^2 in (3.4.4), define the event

$$E_n = \left\{ \int \int \prod_{i=1}^n \frac{f_i}{f_{0,i}}(Y_i) d\Pi(\boldsymbol{\beta}) d\Pi(\boldsymbol{\Sigma}) \geq e^{-C_1 n \epsilon_n^2} \right\},$$

then

$$\mathbb{P}_0(E_n^c) \leq \exp\left(- (1 + C_2) n \epsilon_n^2\right). \quad (3.4.5)$$

3.4.2 Dimensionality and recovery

In this section, we study the dimensionality and recovery properties of the the marginal posterior of β .

Lemma 3.4.3 (Dimension). *Let a prior $\pi_G(s_k)$ satisfying (3.3.2) for all $k = 1, \dots, d$ be given. Assume that $s_0 p_{\max} \log n \ll n$, $s_k^* = \max\{s_{0,k}, s_{0,k} p_{\max} \log n / \log G, d \log n / \log G\}$, and $\log d < A_4 \log G$. Then for a sufficiently large number $M_2 \geq 2(1 + C_2)/A_4 + 1$,*

$$\sup_{\beta_0 \in \mathcal{B}_0, \Sigma_0 \in \mathcal{H}_0} \mathbb{E}_0 \Pi \left(\beta_k : |S_k| < M_2 s_k^*, k = 1, \dots, d \mid Y_1, \dots, Y_n \right) \rightarrow 1. \quad (3.4.6)$$

Lemma 3.4.3 also implies that the sum of the cardinalities of the non-zero groups in d different columns will not exceed $s^* = \sum_{k=1}^d s_k^*$. We state this result in the following corollary.

Corollary 3.4.4. *Under the setup of Lemma 3.4.3, with $s^* = n \epsilon_n^2 / \log G$,*

$$\sup_{\beta_0 \in \mathcal{B}_0, \Sigma_0 \in \mathcal{H}_0} \mathbb{E}_0 \Pi \left(\beta : |S| \geq M_2 s^* \mid Y_1, \dots, Y_n \right) \rightarrow 0, \quad (3.4.7)$$

From Corollary 3.4.4, $s^* > s_0$ when either $s_0 p_{\max} \log n / \log G \gg s_0$ or $d^2 \log n / \log G \gg s_0$. This means that the support of the posterior can substantially overshoot the true dimension s_0 . In the next corollary, we show that the posterior is still able to recover β_0 even when $s^* > s_0$;

Corollary 3.4.5 (Recovery). *Under Assumption 2, if $s_0 p_{\max} \log n \ll n$, then for a sufficiently large constant $M_3 > 0$,*

$$\sup_{\beta_0 \in \mathcal{B}_0, \Sigma_0 \in \mathcal{H}_0} \mathbb{E}_0 \Pi \left(\|\beta - \beta_0\|_F^2 \geq \frac{M_3 n \epsilon_n^2}{\sum_{i=1}^n \|X_i\|^2 \phi_{\ell_2}^2(s^*)} \mid Y_1, \dots, Y_n \right) \rightarrow 0, \quad (3.4.8)$$

where $\phi_{\ell_2}^2(s^*)$ is the restricted eigenvalue (see Definition 3.4.6 below).

Definition 3.4.6 (Restricted eigenvalue). *The smallest scaled singular value of dimension \tilde{s} is defined as*

$$\phi_{\ell_2}^2(\tilde{s}) = \inf \left\{ \frac{\sum_{i=1}^n \|X_i \beta\|^2}{\sum_{i=1}^n \|X_i\|^2 \|\beta\|_F^2}, \quad 0 \leq s \leq \tilde{s} \right\}. \quad (3.4.9)$$

As $p \gg n$, the smallest eigenvalue of the design matrix must be 0. The restricted eigenvalue condition assumes that the smallest eigenvalue for the sub-matrix of the design matrix, which corresponds to the coefficients within non-zero groups, is not 0.

The results for other norms for the difference between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$ can be also derived by assuming different assumptions on the smallest eigenvalue for the sub-matrix of the design matrix. For example, using the uniform compatibility condition (in Definition 3.4.7 below), we can conclude that for a sufficiently large number $M_4 > 0$,

$$\sup_{\boldsymbol{\beta}_0 \in \mathcal{B}_0, \boldsymbol{\Sigma}_0 \in \mathcal{H}_0} \mathbb{E}_0 \Pi \left(\sum_{k=1}^d \|\beta_k - \beta_{0,k}\|_{2,1}^2 \geq \frac{M_4 n \epsilon_n^2 s^*}{\sum_{i=1}^n \|X_i\|^2 \phi_{\ell_{2,1}}^2(s^*)} \middle| Y_1, \dots, Y_n \right) \rightarrow 0.$$

The proof is almost identical to that of Corollary 3.4.5.

Definition 3.4.7 (Uniform compatibility, ℓ_2/ℓ_1 -norm). *The $\ell_{2,1}$ -compatibility number in vectors of dimension \tilde{s} is defined as*

$$\phi_{\ell_{2,1}}^2(\tilde{s}) = \inf \left\{ \frac{s \sum_{i=1}^n \|X_i \boldsymbol{\beta}\|^2}{\sum_{i=1}^n \|X_i\|^2 \sum_{k=1}^d \|\beta_k\|_{2,1}^2}, \quad 0 \leq s \leq \tilde{s} \right\}. \quad (3.4.10)$$

By the Cauchy-Schwarz inequality, $\|\beta_k\|^2 \geq \|\beta_k\|_{2,1}^2 / s_k$, it follows that $\phi_{\ell_2}(\tilde{s}) \leq \phi_{\ell_{2,1}}(\tilde{s})$ for any $\tilde{s} \ll G$.

3.4.3 Distributional approximation

In this section, we show that the posterior distribution can be approximated by a mixture of multivariate normal densities.

We first rewrite the model (3.1.1) as

$$Y_i = \text{Vec}(\boldsymbol{\beta}) \mathbf{X}_i + \varepsilon_i, \quad i = 1, \dots, n, \quad (3.4.11)$$

where $\text{Vec}(\boldsymbol{\beta})$ is a $1 \times pd$ vector by stacking all the columns of $\boldsymbol{\beta}$ into a row vector, $\mathbf{X}_i = \mathbf{I}_d \otimes X_i$ is a $pd \times d$ block diagonal matrix. The above model can be also written as

$$Y_i = \text{Vec}(\boldsymbol{\beta}_S) \mathbf{X}_{i,S} + \varepsilon_i, \quad i = 1, \dots, n,$$

where $\text{Vec}(\boldsymbol{\beta}_S)$ is a $1 \times (\sum_{k=1}^d \sum_{j \in S_k} p_j)$ vector, which consists of coordinates of $\boldsymbol{\beta}$ from

the set S , and $\mathbf{X}_{i,S}$ is a $(\sum_{k=1}^d \sum_{j \in S_k} p_j) \times d$ matrix, which is the submatrix of \mathbf{X}_i .

Then log-likelihood function is given by

$$\begin{aligned} \ell_n(\text{Vec}(\boldsymbol{\beta}_S), \boldsymbol{\Sigma}) &= \sum_{i=1}^n \log f(Y_i | \text{Vec}(\boldsymbol{\beta}_S) \mathbf{X}_{i,S}, \boldsymbol{\Sigma}) \\ &= -\frac{nd}{2} \log(2\pi) - \frac{n}{2} \log(\det(\boldsymbol{\Sigma})) \\ &\quad - \frac{1}{2} \sum_{i=1}^n (Y_i - \text{Vec}(\boldsymbol{\beta}_S) \mathbf{X}_{i,S}) \boldsymbol{\Sigma}^{-1} (Y_i - \text{Vec}(\boldsymbol{\beta}_S) \mathbf{X}_{i,S})'. \end{aligned} \quad (3.4.12)$$

If $\sum_{k=1}^d \sum_{j \in S_k} p_j \ll n$, then the maximum likelihood estimator (MLE) of $\text{Vec}(\boldsymbol{\beta}_S)$ is unique. We denote the MLE as $\text{Vec}(\hat{\boldsymbol{\beta}}_S)$ and the Fisher information matrix as $\mathbb{I}_{n,S}$. From (3.4.12), we can obtain that $\text{Vec}(\hat{\boldsymbol{\beta}}_S) = (\sum_{i=1}^n \mathbf{X}_{i,S} \mathbf{X}_{i,S}')^{-1} (\sum_{i=1}^n \mathbf{X}_{i,S} Y_i')$ and $\mathbb{I}_{n,S} = n^{-1} \sum_{i=1}^n \mathbf{X}_{i,S} \boldsymbol{\Sigma}_0^{-1} \mathbf{X}_{i,S}'$.

Based on the model (3.4.11), the marginal posterior distribution of $\boldsymbol{\beta}$ is

$$\begin{aligned} \Pi(B|Y_1, \dots, Y_n) &= \frac{\int \int_B \exp(\ell(\text{Vec}(\boldsymbol{\beta}), \boldsymbol{\Sigma}) - \ell(\text{Vec}(\boldsymbol{\beta}_0), \boldsymbol{\Sigma}_0)) d\Pi(\text{Vec}(\boldsymbol{\beta})) d\Pi(\boldsymbol{\Sigma})}{\int \int \exp(\ell(\text{Vec}(\boldsymbol{\beta}), \boldsymbol{\Sigma}) - \ell(\text{Vec}(\boldsymbol{\beta}_0), \boldsymbol{\Sigma}_0)) d\Pi(\text{Vec}(\boldsymbol{\beta})) d\Pi(\boldsymbol{\Sigma})}, \end{aligned} \quad (3.4.13)$$

with

$$d\Pi(\text{Vec}(\boldsymbol{\beta})) = \prod_{k=1}^d \left(\sum_{S_k \subseteq \{1, \dots, G\}} \prod_{j \in S_k} \left(\frac{\lambda_k}{a_j} \right)^{p_j} \exp(-\lambda_k \|\beta_{S_k}\|_{2,1}) d\beta_{S_k} \otimes \delta_{S_k^c} \right).$$

It is clear that the posterior distribution is a mixture density over different subsets S_1, \dots, S_d .

Let $\mathcal{S}_k^* = \{S_k \subseteq \{1, \dots, G\} : |S_k| \leq M_2 s_k^*, k = 1, \dots, d$. In the next theorem, we show that the posterior $\Pi(B|Y_1, \dots, Y_n)$ can be approximated by a mixture of multivariate normal densities given by

$$\Pi^\infty(B|Y_1, \dots, Y_n) \propto \prod_{k=1}^d \sum_{S_k \in \mathcal{S}_k^*} w_{S_k}^\infty \mathcal{N}(\text{Vec}(\hat{\boldsymbol{\beta}}_{S_k}), \mathbb{I}_{n,S_k}^{-1}) \otimes \delta_{S_k^c}, \quad (3.4.14)$$

where

$$\begin{aligned}
w_{S_k}^\infty &\propto \pi_G(s_k) \frac{1}{\binom{G}{s_k}} \prod_{j \in S_k} \left(\frac{\lambda_k}{a_j} \right)^{p_j} (2\pi)^{\sum_{j \in S_k} p_j / 2} \left(\det \left(\sum_{i=1}^n \mathbf{X}'_{i,S_k} \boldsymbol{\Sigma}_0^{-1} \mathbf{X}_{i,S_k} \right) \right)^{-1/2} \\
&\times \exp \left\{ \frac{1}{2} \sum_{i=1}^n \left\| \text{Vec}(\hat{\boldsymbol{\beta}}_{S_k}) \mathbf{X}_{i,S_k} \boldsymbol{\Sigma}_0^{-1/2} \right\|^2 \right\} \mathbb{1}\{S_k \in \mathcal{S}_k^*\},
\end{aligned} \tag{3.4.15}$$

with $\sum_{S_k} w_{S_k}^\infty = 1$, for all $k = 1, \dots, d$.

Before we state the theorem, one more terminology should be introduced. We recall the notion of the *small λ region* (see [Castillo et al., 2015](#)). In our setting, each λ_k belongs to the small λ region if $\frac{\max_{1 \leq k \leq d} \lambda_k s_k^* \sqrt{\log G}}{\sqrt{\sum_{i=1}^n \|X_i\|^2}} \rightarrow 0$. When λ_k belongs to this region, the MLE, $\text{Vec}(\hat{\boldsymbol{\beta}}_S)$, is an asymptotically unbiased estimator and does not depend on the choice of different values of λ_k . When choosing the value of λ_k outside the small λ region, this MLE is no longer asymptotically unbiased and will depend on the choice of λ_k (cf. see Theorem 11 of the supplementary material of [Castillo et al., 2015](#)). As a result, the posterior will concentrate near a distribution with center differing a lot with different values of λ_k .

Theorem 3.4.8 (Distributional approximation). *For $k = 1, \dots, d$, if $\pi_G(s_k)$ satisfies (3.3.2) and $\frac{\max_k \lambda_k s_k^* \sqrt{\log G}}{\sqrt{\sum_{i=1}^n \|X_i\|^2}} \rightarrow 0$, then for a positive constant c ,*

$$\sup_{\substack{\boldsymbol{\beta}_0 \in \mathcal{B}_0: \phi_{\ell_2}(s^*) > c, \\ \boldsymbol{\Sigma}_0 \in \mathcal{H}_0}} \left\| \Pi(\cdot | Y_1, \dots, Y_n) - \Pi^\infty(\cdot | Y_1, \dots, Y_n) \right\|_{TV} \rightarrow 0. \tag{3.4.16}$$

Note that the above theorem does not require that the cardinality of the set S to be close to s_0 . The result still holds when $s^* \gg s_0$.

3.4.4 Selection

In the previous two sections, we have shown that even if $s^* > s_0$, the marginal posterior of $\boldsymbol{\beta}$ can recover the truth and can be approximated by a mixture of multivariate normal densities. In this section, we derive conditions for selection consistency. Since selection consistency requires $s^* = s_0$, we need to assume the dimension of the covariance and the coordinates in the non-zero groups are sufficiently small. We also need to assume that

the smallest signal cannot be too small, which is a group sparse version of the *Beta-min condition*. This condition is stated in below:

$$\tilde{\mathcal{B}} = \left\{ \min_{\substack{j \in S_{0,k} \\ 1 \leq k \leq d}} \|\beta_{jk}\| \geq \sqrt{\frac{M_3 n \epsilon_n^2}{\sum_{i=1}^n \|X_i\|^2 \phi_{\ell_2}^2(s_0)}} \right\}. \quad (3.4.17)$$

The lower bound displayed in the condition is derived from (3.4.8). Unlike the *Beta-min condition* in Castillo et al. (2015) which the individual components are bounded away from 0, our condition allows a zero to be included in a non-zero group.

Theorem 3.4.9 (Selection consistency). *If $\pi_G(s_k)$ satisfies Assumption 1 for all $k = 1, \dots, d$, $\frac{\max_k \lambda_k s_{0,k} \sqrt{n \log G}}{\sqrt{\sum_{i=1}^n \|X_i\|^2}} \rightarrow 0$, $d^2 \log n \lesssim s_0 \log G$, and $\log d \leq A_4 \log G$, then for $s_n \leq G^{A_4-1}$ with $A_4 \geq \max(1, 2b_2)$, where b_2 is defined in (3.4.1), and a positive number c ,*

$$\sup_{\substack{\beta_0 \in \mathcal{B}_0 \cap \tilde{\mathcal{B}}: |S_{0,k}| \leq s_n, k=1, \dots, d, \\ \phi_{\ell_2}(s_0) \geq c, \Sigma_0 \in \mathcal{H}_0}} \mathbb{E}_0 \Pi(\beta : S_\beta = S_{\beta_0} | Y_1, \dots, Y_n) \rightarrow 0.$$

If the conditions of Theorem 3.4.9 are satisfied, then the marginal posterior distribution of β in non-zero groups can be approximated by a multivariate normal distribution with mean $\text{Vec}(\hat{\beta}_{S_0})$ and the covariance matrix $\mathbb{I}_{0,S_0}^{-1} = n(\sum_{i=1}^n \mathbf{X}_{i,S_0} \Sigma_0^{-1} \mathbf{X}'_{i,S_0})^{-1}$. Therefore, credible intervals for β can be obtained directly from the approximating multivariate normal density.

3.5 Proofs

Proof of Lemma 3.3.1. Let $\xi_i = (Y_i - X_i \beta_0) \Sigma_0^{-1/2}$, then

$$\sum_{i=1}^n \|X'_i (Y_i - X_i \beta_0) \Sigma_0^{-1/2}\|_F^2 = \sum_{i=1}^n \|X_i\|^2 \|\xi_i\|^2 = \sum_{i=1}^n \|X_i\|^2 \sum_{k=1}^d \xi_{ik}^2.$$

Therefore, the probability in (3.3.5) can be also written as

$$\mathbb{P}_0 \left(\frac{\sum_{i=1}^n \|X_i\|^2 \sum_{k=1}^d (\xi_{ik}^2 - 1)}{\sqrt{2d \sum_{i=1}^n \|X_i\|^4}} \geq t \right), \quad (3.5.1)$$

where $t = \left(\sum_{k=1}^d \bar{\lambda}_k^2 - d \sum_{i=1}^n \|X_i\|^2 \right) / \left(2d \sum_{i=1}^n \|X_i\|^4 \right)^{1/2}$. By (3.6.4), the probability (3.5.1) is bounded above by

$$2 \exp \left(-t^2 / \left(2(1 + \sqrt{2/d} t m(X)) \right) \right),$$

where $m(X) = \max_i \|X_i\|^2 / \left(\sum_{i=1}^n \|X_i\|^4 \right)^{1/2}$. We plug-in the expression for t in (3.6.2) and $m(X)$, then the last display is bounded below by $2G^{-q}$ if

$$\begin{aligned} \sum_{k=1}^d \bar{\lambda}_k^2 &\geq 2 \left(dq \log G \sum_{i=1}^n \|X_i\|^4 + q^2 \log^2 G \max_i \|X_i\|^4 \right)^{1/2} \\ &\quad + 2q \log G \max_i \|X_i\|^2 + d \sum_{i=1}^n \|X_i\|^2. \end{aligned} \quad (3.5.2)$$

By choosing $q = d$ and $\bar{\lambda} = \bar{\lambda}_k = 3\sqrt{\sum_{i=1}^n \|X_i\|^2 \log G}$, (3.5.1) is bounded above by G^{-d} , which goes to 0 as $G \rightarrow \infty$ or $d \rightarrow \infty$. \square

Proof of Theorem 3.4.1. The proof contains two parts. In the first part, we quantify prior concentration around the truth in the sense of Kullback-Leibler divergence from the true density. In the second part, using the results obtained from the first part, we derive (3.4.2) and (3.4.3).

Part I. The method we use to obtain the posterior contraction rate is described as follows. We construct a test from the first principle by breaking up the effective parameter space into several pieces which are sufficiently separated from the truth. Then for each piece, we consider the likelihood ratio test for the truth against a representative in the piece. We show that this test works for the entire piece by controlling the likelihood ratio. Finally, we consider the maximum of these tests and control its size by estimating the total number of pieces.

Let \mathcal{F}_n be a suitable ‘‘sieve’’. We shall verify that

$$\Pi(K(f, f_0) \leq \epsilon_n^2, V(f, f_0) \leq \epsilon_n^2) \geq \exp(-C_1 n \epsilon_n^2), \quad (3.5.3)$$

$$\Pi(\mathcal{F}_n^c) \leq \exp(-C_2 n \epsilon_n^2), \quad (3.5.4)$$

for positive constants C_1 and $C_2 > C_1 + 2$ and the following condition.

Let $f_0 = \prod_{i=1}^n f_{0,i}$ and $f = \prod_{i=1}^n f_i$, where $f_{0,i} = \mathcal{N}(X_i \boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$ and $f_i = \mathcal{N}(X_i \boldsymbol{\beta}, \boldsymbol{\Sigma})$. Then there exists a test ϕ_n such that

$$\mathbb{E}_{f_0} \phi_n \leq e^{-n\epsilon_n^2}, \quad \sup_{\substack{f \in \mathcal{F}_n: \sum_{i=1}^n \|X_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|^2 \geq nM_1\epsilon_n^2, \\ \text{or } \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_0\|_F^2 \geq M_1\epsilon_n^2}} \mathbb{E}_f(1 - \phi_n) \leq e^{-C_3 n\epsilon_n^2/4}, \quad (3.5.5)$$

where C_3 is a positive constant.

Consider the sieve

$$\mathcal{F}_n = \left\{ f : \max_{1 \leq k \leq d} |S_k| < \bar{s}_n, \max_{\substack{1 \leq j \leq G, \\ 1 \leq k \leq d}} \|\beta_{jk}\| < p_{\max} H_n, \right. \\ \left. \exp(-d \log n) < \text{eig}_1(\boldsymbol{\Sigma}^{-1}), \text{eig}_d(\boldsymbol{\Sigma}^{-1}) < n \right\},$$

where

$$\bar{s}_n = \max \left\{ \frac{s_0}{d}, \frac{s_0 p_{\max} \log n}{d \log G}, \frac{d \log n}{\log G} \right\}, \quad H_n = \frac{3\bar{s}_n (\log G \vee p_{\max} \log n)}{p_{\max} (\min_k \underline{\lambda}_k)}. \quad (3.5.6)$$

Recall that the expression of $\underline{\lambda}_k$ is shown in (3.3.4).

First, we check (3.5.3). The Kullback-Leibler divergence between f and f_0 is

$$K(f, f_0) = \frac{1}{2} \left(\text{Tr}(\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\Sigma}) - d - \log(\det(\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\Sigma})) + \frac{1}{n} \sum_{i=1}^n X_i (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)' X_i' \right),$$

and the Kullback-Leibler variation between f and f_0 is

$$V(f, f_0) = \frac{1}{2} \left(\text{Tr}(\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\Sigma} \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\Sigma}) - 2\text{Tr}(\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\Sigma}) + d \right) \\ + \frac{1}{n} \sum_{i=1}^n X_i (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\Sigma} \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)' X_i'.$$

We define the two events \mathcal{A}_1 and \mathcal{A}_2 as follows:

$$\mathcal{A}_1 = \left\{ \boldsymbol{\Sigma} : \text{Tr}(\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\Sigma}) - d - \log(\det(\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\Sigma})) \leq \epsilon_n^2, \right. \\ \left. \text{Tr}(\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\Sigma} \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\Sigma}) - 2\text{Tr}(\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\Sigma}) + d \leq \epsilon_n^2 \right\}, \\ \mathcal{A}_2 = \left\{ (\boldsymbol{\beta}, \boldsymbol{\Sigma}) : \sum_{i=1}^n X_i (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)' X_i' \leq n\epsilon_n^2 \right\},$$

$$\sum_{i=1}^n X_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'X_i' \leq n\epsilon_n^2/2\}.$$

Writing $\Pi(K(f, f_0) \leq \epsilon_n^2, V(f, f_0) \leq \epsilon_n^2) = \Pi(\mathcal{A}_2|\mathcal{A}_1)\Pi(\mathcal{A}_1)$, we shall derive lower bounds for $\Pi(\mathcal{A}_1)$ and $\Pi(\mathcal{A}_2|\mathcal{A}_1)$ separately.

Define $\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1/2}$ and note that $\boldsymbol{\Sigma}^{*-1} \sim \mathcal{W}_d(\nu, \boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\Phi}\boldsymbol{\Sigma}_0^{1/2})$ as $\boldsymbol{\Sigma}^{-1} \sim \mathcal{W}_d(\nu, \boldsymbol{\Phi})$. Then

$$\mathcal{A}_1 = \left\{ \boldsymbol{\Sigma} : \sum_{k=1}^d (\text{eig}_k(\boldsymbol{\Sigma}^*) - 1 - \log(\text{eig}_k(\boldsymbol{\Sigma}^*))) \leq \epsilon_n^2, \sum_{k=1}^d (\text{eig}_k(\boldsymbol{\Sigma}^*) - 1)^2 \leq \epsilon_n^2 \right\}. \quad (3.5.7)$$

Furthermore, we define $\mathcal{A}_1^* = \bigcap_{k=1}^d \{ \boldsymbol{\Sigma} : 1 \leq \text{eig}_k(\boldsymbol{\Sigma}^*) \leq 1 + d^{-1/2}\epsilon_n \}$. It is easy to verify that $\mathcal{A}_1 \supset \mathcal{A}_1^*$. By (3.6.10), we obtain that

$$\begin{aligned} \Pi(\mathcal{A}_1) &\geq \Pi(\mathcal{A}_1^*) \\ &\geq \exp\left(-c_{11}d \log d - c_{12}d^2 \log d - d(d+1) \log(1/\epsilon_n)/2 - c_{13}d\right), \end{aligned} \quad (3.5.8)$$

where c_{11}, c_{12} and c_{13} are positive constants.

To derive a lower bound for $\Pi(\mathcal{A}_2|\mathcal{A}_1)$, we need the following two results. First, by $\boldsymbol{\Sigma}_0 \geq b_1\mathbf{I}_d$,

$$\frac{1}{n} \sum_{i=1}^n X_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'X_i' \leq \frac{1}{nb_1} \sum_{i=1}^n \|X_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|^2.$$

Second, we have

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n X_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'X_i' \\ &\leq \frac{1}{n} \sum_{i=1}^n \|X_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\boldsymbol{\Sigma}_0^{-1/2}\|^2 \|\boldsymbol{\Sigma}^* - \mathbf{I}_d\|_F + \frac{1}{n} \sum_{i=1}^n \|X_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\boldsymbol{\Sigma}_0^{-1/2}\|^2 \\ &\leq \frac{1}{n} \sum_{i=1}^n \|X_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\boldsymbol{\Sigma}_0^{-1/2}\|^2 \left(1 + \sqrt{\sum_{k=1}^d (\text{eig}_k(\boldsymbol{\Sigma}^*) - 1)^2} \right). \end{aligned}$$

Conditional on \mathcal{A}_1 and again, by $\boldsymbol{\Sigma}_0 \geq b_1\mathbf{I}_d$, the last expression can be further bounded

above by

$$\frac{1 + \epsilon_n}{n} \sum_{i=1}^n \|X_i(\beta - \beta_0)\Sigma_0^{-1/2}\|^2 < \frac{2}{nb_1} \sum_{i=1}^n \|X_i(\beta - \beta_0)\|^2,$$

provided that $\epsilon_n < 1$. Thus $\Pi(\mathcal{A}_2|\mathcal{A}_1)$ is bounded below by

$$\begin{aligned} \Pi\left(\frac{1}{n} \sum_{i=1}^n \|X_i(\beta - \beta_0)\|^2 \leq \frac{b_1\epsilon_n^2}{4}\right) &\geq \Pi\left(\max_{1 \leq k \leq d} \left(\frac{1}{n} \sum_{i=1}^n |X_i(\beta_k - \beta_{0,k})|^2\right) \leq \frac{b_1\epsilon_n^2}{4d}\right) \\ &\geq \Pi\left(\max_{1 \leq k \leq d} \|\beta_k - \beta_{0,k}\|_{2,1} \leq \frac{r_n\sqrt{b_1}}{2}\right), \end{aligned} \quad (3.5.9)$$

where $r_n = \sqrt{\frac{n\epsilon_n^2}{d \sum_{i=1}^n \|X_i\|^2}}$. By (3.3.1), (3.5.9) can be further bounded below by

$$\prod_{k=1}^d \left(\pi_G(s_{0,k}) \frac{1}{\binom{G}{s_{0,k}}} \int_{\|\beta_k - \beta_{0,k}\|_{2,1} \leq \frac{r_n\sqrt{b_1}}{2}} g_{s_k}(\beta_{S_k}) d\beta_{S_k} \right). \quad (3.5.10)$$

By changing the variable $\beta_{S_k} - \beta_{0,S_k}$ to $\check{\beta}_{S_k}$ and using the fact that $\|x\| \leq \|x\|_1$ for any vector x , each integral in (3.5.10) is bounded below by

$$\begin{aligned} e^{-\lambda_k \|\beta_{0,k}\|_{2,1}} \prod_{j \in S_{0,k}} \int_{\|\check{\beta}_j\|_1 \leq \frac{r_n\sqrt{b_1}}{2s_{0,k}}} \left(\frac{\lambda_k}{a_j}\right)^{p_j} e^{-\lambda_k \|\check{\beta}_j\|_1} d\check{\beta}_j \\ \geq e^{-\lambda_k \|\beta_{0,k}\|_{2,1}} \prod_{j \in S_{0,k}} \left(e^{-\lambda_k r_n \sqrt{b_1}/(2s_{0,k})} \frac{1}{p_j!} \left(\frac{\lambda_k r_n \sqrt{b_1}}{s_{0,k} a_j}\right)^{p_j} \right). \end{aligned}$$

The lower bound in the last inequality is obtained by using the result that the integrand equals to the probability of the first p_j events of a Poisson process happen before time $r_n \sqrt{b_1}/(2s_{0,k})$ (similar to the argument used to derive (6.2) in Castillo et al., 2015).

Now, (3.5.10) is bounded below by

$$\prod_{k=1}^d \left(\pi_G(s_{0,k}) \frac{1}{\binom{G}{s_{0,k}}} e^{-\lambda_k \|\beta_{0,k}\|_{2,1}} \prod_{j \in S_{0,k}} \left(e^{-\lambda_k r_n \sqrt{b_1}/(2s_{0,k})} \frac{1}{p_j!} \left(\frac{\lambda_k r_n \sqrt{b_1}}{s_{0,k} a_j}\right)^{p_j} \right) \right).$$

By Assumption 1, the last display can be further bounded below by

$$\begin{aligned}
& A_1^{s_0} G^{-(1+A_3)s_0} \exp\left(-\sum_{k=1}^d \lambda_k \|\beta_{0,k}\|_{2,1} - \sum_{k=1}^d \lambda_k r_n \sqrt{b_1}/2\right) \\
& \times \prod_{k=1}^d \prod_{j \in S_{0,k}} \frac{1}{p_j!} \left(\frac{\lambda_k r_n \sqrt{b_1}}{s_{0,k} a_j}\right)^{p_j}. \tag{3.5.11}
\end{aligned}$$

Combining the lower bounds (3.5.8) and (3.5.11), $\log \Pi(K(f, f_0) \leq \epsilon_n^2, V(f, f_0) \leq \epsilon_n^2)$ is bounded below by

$$\begin{aligned}
& -c_{11}d \log d - c_{12}d^2 \log d - d(d+1) \log(1/\epsilon_n)/2 - c_{13}d + s_0 \log A_1 \\
& -c_{14}s_0 \log G - \sum_{k=1}^d \lambda_k \|\beta_{0,k}\|_{2,1} - \sum_{k=1}^d \frac{\lambda_k r_n \sqrt{b_1}}{2} + \sum_{k=1}^d \sum_{j \in S_{0,k}} p_j \log(\lambda_k r_n \sqrt{b_1}) \\
& - \sum_{k=1}^d \sum_{j \in S_{0,k}} p_j \log(s_{0,k} a_j) - \sum_{k=1}^d \sum_{j \in S_{0,k}} \log p_j!. \tag{3.5.12}
\end{aligned}$$

Let ϵ_n^2 be as in (3.4.4). Since $\epsilon_n^2 \geq (d^2 \log n)/n$, the sum of the first four terms in (3.5.12) is bounded below by a multiple of $-n\epsilon_n^2$. By Assumption 2,

$$\sum_{k=1}^d \lambda_k r_n \sqrt{b_1}/2 - \sum_{k=1}^d \sum_{j \in S_{0,k}} p_j \log(\lambda_k r_n \sqrt{b_1}) \lesssim s_0 \log G \leq n\epsilon_n^2,$$

as $\epsilon_n^2 \geq s_0(\log G)/n$. Also, since $\max_k \|\beta_{0,k}\|_{2,1} \leq \bar{\beta}$ with the expression of $\bar{\beta}$ is displayed in (3.4.1), then $\sum_{k=1}^d \lambda_k \|\beta_{0,k}\| \leq n\epsilon_n^2$. Furthermore, since $\log(p_j!) \leq p_j \log p_j$ and $a_j = O(p_j^{1/2})$, we obtain that

$$\sum_{k=1}^d \sum_{j \in S_{0,k}} p_j \log(s_{0,k} a_j) + \sum_{k=1}^d \sum_{j \in S_{0,k}} \log(p_j!) \leq 3 \sum_{k=1}^d \sum_{j \in S_{0,k}} p_j \log p_j \lesssim n\epsilon_n^2,$$

as $\epsilon_n^2 \geq s_0 p_{\max} \log n/n$. This completes the verification of (3.5.3).

Next, we verify (3.5.4). We obtain that

$$\begin{aligned} \Pi(\mathcal{F}_n^c) &\leq \sum_{k=1}^d \Pi(|S_k| \geq \bar{s}_n) + \sum_{k=1}^d \left(\sum_{S_k \in \mathcal{S}_n} \sum_{j \in S_k} \Pi(\|\beta_{j,k}\| \geq p_{\max} H_n) \right) \\ &\quad + \Pi(\text{eig}_1(\Sigma^{*-1}) \leq \exp(-d \log n)) + \Pi(\text{eig}_d(\Sigma^{*-1}) \geq n), \end{aligned} \quad (3.5.13)$$

where $\mathcal{S}_n = \{S \subseteq \{1, \dots, G\} : |S| \leq \bar{s}_n\}$. By (3.3.2), the first term in (3.5.13) is bounded above by

$$\sum_{k=1}^d \sum_{s_k \geq \bar{s}_n} \pi(s_k) \leq d\pi(\bar{s}_n) \sum_{j=0}^{\infty} \left(\frac{A_2}{G^{A_4}} \right)^j \leq 2dA_2^{\bar{s}_n} G^{-A_4 \bar{s}_n}.$$

To derive an upper bound for the second term in (3.5.13), we apply the the upper bound of the tail of a gamma density in page 29 of [Boucheron et al. \(2013\)](#) and the inequality $1 + x - \sqrt{1 + 2x} \geq x^2/(2(1 + x))$, for any $x > 0$, to obtain that

$$\begin{aligned} \Pi(\|\beta_{jk}\| > p_{\max} H_n) &= \int_{r > p_{\max} H_n} \frac{\lambda_k^{p_{\max}}}{\Gamma(p_{\max})} r^{p_{\max}-1} e^{-\lambda_k r} dr \\ &\leq \exp\left(-p_{\max}(1 + \lambda_k H_n - \sqrt{1 + 2\lambda_k H_n})\right) \\ &\leq \exp\left(-\frac{p_{\max} \lambda_k^2 H_n^2}{2(1 + \lambda_k H_n)}\right), \end{aligned} \quad (3.5.14)$$

where $j = 1, \dots, G$, $k = 1, \dots, d$. We then plug-in (3.5.14) to obtain the upper bound, which is

$$\exp\left(\log d + 2d \log \bar{s}_n + \bar{s}_n \log G - \sum_{k=1}^d \frac{p_{\max} \lambda_k^2 H_n^2}{2(1 + \lambda_k H_n)}\right).$$

For the summation of the third and the fourth terms in (3.5.13), we apply (3.6.7) and (3.6.8), then it is bounded above by

$$\exp(c_{25} d^2 \log d - c_{26} d^2 \log n) + \exp(-d^2 \log d + c_{23} d^2 \log n + d^2/2 - c_{24} n),$$

where $c_{21}, c_{22}, \dots, c_{26}$ are positive constants.

Now we combine the upper bounds for each term in (3.5.13), choose H_n in (3.5.6) and $C_2 \geq C_1 + 2$. Then (3.5.13) is bounded above by $\exp(-C_2 n \epsilon_n^2)$. We thus obtain (3.5.4).

Last, we verify (3.5.5). Let $\phi_n = \mathbb{1}\{f_1/f_0 \geq 1\}$, the most powerful Neyman-Pearson test. If the average negative log-affinity $-n^{-1} \log \int f_0^{1/2} f_1^{1/2}$ between f_0 and f_1 is bigger than ϵ_n^2 , then

$$\mathbb{E}_{f_0} \phi_n = \mathbb{E}_{f_0} \left(\sqrt{f_1/f_0} \geq 1 \right) \leq \int \sqrt{f_0 f_1} \leq e^{-n\epsilon_n^2}.$$

This gives the first inequality in (3.5.5). For the second inequality in (3.5.5), observe that by the Cauchy-Schwarz inequality,

$$\mathbb{E}_f(1 - \phi_n) \leq \left\{ \mathbb{E}_{f_1}(1 - \phi_n) \right\}^{1/2} \left\{ \mathbb{E}_{f_1} \left(\frac{f}{f_1} \right)^2 \right\}^{1/2}.$$

By following the similar arguments used in proving the first inequality in (3.5.5), we obtain that

$$\mathbb{E}_{f_1}(1 - \phi_n) = \mathbb{E}_{f_1} \left(\sqrt{f_0/f_1} \geq 1 \right) \leq \int \sqrt{f_0 f_1} \leq e^{-n\epsilon_n^2}.$$

For $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}\|_F \leq \delta_n = \sqrt{\epsilon_n^2/(6nb_3)}$ and $\|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}\| \leq \delta'_n = 1/(n^2d)$, we have $\mathbb{E}_{f_1}(f/f_1)^2 \leq e^{n\epsilon_n^2/2}$. Recall that b_3 is the upper bound for $n^{-1} \sum_{i=1}^n \|X_i\|^2$. To end this, observe that

$$\begin{aligned} \mathbb{E}_{f_1}(f/f_1)^2 &= (\det(\boldsymbol{\Sigma}^*))^{n/2} (\det(2\mathbf{I} - \boldsymbol{\Sigma}^{*-1}))^{-n/2} \\ &\quad \times \exp \left(\sum_{i=1}^n X_i \boldsymbol{\beta}^* \boldsymbol{\Sigma}^{-1/2} (2\boldsymbol{\Sigma}^* - \mathbf{I})^{-1} \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\beta}^{*'} X_i' \right). \end{aligned} \quad (3.5.15)$$

Now $\|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}\| \leq \delta'_n = 1/(n^2d)$ implies that

$$\|\boldsymbol{\Sigma}^* - \mathbf{I}\| \leq \|\boldsymbol{\Sigma}^{-1}\| \|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}\| \leq n \|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}\|_F \leq n\delta'_n.$$

Therefore, $1 - n\delta'_n \leq \|\boldsymbol{\Sigma}^*\| \leq 1 + n\delta'_n$. Writing $\text{eig}_k(\boldsymbol{\Sigma}^*)$ for the k -th eigenvalue of $\boldsymbol{\Sigma}^*$, we obtain that

$$\begin{aligned} &(\det(\boldsymbol{\Sigma}^*))^{n/2} (\det(2\mathbf{I} - \boldsymbol{\Sigma}^{*-1}))^{-n/2} \\ &= \exp \left(\frac{n}{2} \sum_{k=1}^d \log(\text{eig}_k(\boldsymbol{\Sigma}^*)) - \frac{n}{2} \sum_{k=1}^d \log \left(2 - \frac{1}{\text{eig}_k(\boldsymbol{\Sigma}^*)} \right) \right) \end{aligned}$$

$$\begin{aligned}
&\leq \exp\left(\frac{n}{2} \sum_{k=1}^d \log(1 + n\delta'_n) - \frac{n}{2} \sum_{k=1}^d \log\left(1 + \frac{n\delta'_n}{1 + n\delta'_n}\right)\right) \\
&\leq \exp(n^2 d \delta'_n / 4) \leq \exp(1/4),
\end{aligned}$$

by the choice of δ'_n ; here the second line is obtained by applying the inequalities $1 - x^{-1} \leq \log x \leq x - 1$ for $x > 0$ and $1 + n\delta'_n < 2$. Using the inequalities $\|(2\mathbf{\Sigma}^* - \mathbf{I})^{-1}\| \leq (2(1 - n\delta'_n) - 1)^{-1} \leq 2$ and $\|\mathbf{\Sigma}^{-1}\| = \text{eig}_d(\mathbf{\Sigma}^{-1}) \leq n$, we bound the exponential term of (3.5.15) by

$$\sum_{i=1}^n \|X_i\|^2 \|\beta_1 - \beta\|_F^2 \|\mathbf{\Sigma}^{-1}\| \|(2\mathbf{\Sigma}^* - \mathbf{I})^{-1}\| \leq 6n^2 b_3 \delta_n^2 \leq n\epsilon_n^2 / 2.$$

Hence, (3.5.15) is bounded above by $\exp(1/4 + n\epsilon_n^2/2)$. This verifies the second inequality in (3.5.5).

Now, with $\|\beta_1 - \beta\|_F \leq \delta_n$ and $\|\mathbf{\Sigma}_1 - \mathbf{\Sigma}\|_F \leq n\delta'_n$, the metric entropy can be calculated as follows:

$$\begin{aligned}
&\log N(\epsilon_n, \mathcal{F}_n, \rho) \\
&\leq \log\left(\prod_{k=1}^d \sum_{S_k \in \mathcal{S}_n} \binom{G}{|S_k|} \left(\frac{6p_{\max} H_n}{\delta_n}\right)^{\sum_{j \in S_k} p_j}\right) + d^2 \log(6d^{3/2} \log n / (n\delta'_n)) \\
&\leq d \log \bar{s}_n + d \bar{s}_n \log G + d \bar{s}_n p_{\max} \log\left(\frac{6p_{\max} H_n \sqrt{6nb_3}}{\epsilon_n}\right) + d^2 \log(6nd^{5/2} \log n) \\
&\lesssim d \bar{s}_n \log G + d \bar{s}_n p_{\max} \log(p_{\max} H_n) + d \bar{s}_n p_{\max} \log n + d^2 \log d + d^2 \log n \\
&\lesssim n\epsilon_n^2.
\end{aligned}$$

This shows that the posterior $\Pi(\sum_{i=1}^n \rho(f_i, f_{0,i}) \gtrsim n\epsilon_n^2 | Y_1, \dots, Y_n) \rightarrow 0$ in \mathbb{P}_0 -probability.

Part II. Observing that

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \rho(f_i, f_{0,i}) &= -\log\left(\frac{(\det(\mathbf{\Sigma}))^{1/4} (\det(\mathbf{\Sigma}_0))^{1/4}}{(\det(\frac{\mathbf{\Sigma} + \mathbf{\Sigma}_0}{2}))^{1/2}}\right) \\
&\quad + \frac{1}{8n} \sum_{i=1}^n X_i (\beta - \beta_0) \left(\frac{\mathbf{\Sigma} + \mathbf{\Sigma}_0}{2}\right)^{-1} (\beta - \beta_0)' X_i'.
\end{aligned}$$

Then $\sum_{i=1}^n \rho(f_i, f_{0,i}) \leq n\epsilon_n^2$ implies both

$$-\log \left(\frac{(\det(\boldsymbol{\Sigma}))^{1/4} (\det(\boldsymbol{\Sigma}_0))^{1/4}}{\left(\det\left(\frac{\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_0}{2}\right)\right)^{1/2}} \right) \leq \epsilon_n^2, \quad (3.5.16)$$

and

$$\frac{1}{n} \sum_{i=1}^n X_i (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \left(\frac{\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_0}{2} \right)^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)' X_i' \leq \epsilon_n^2. \quad (3.5.17)$$

We first show the probability of (3.5.16) goes to 1 implies (3.4.3). Let

$$d^2(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_0) = h^2(\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_0)) = 1 - \frac{(\det(\boldsymbol{\Sigma}))^{1/4} (\det(\boldsymbol{\Sigma}_0))^{1/4}}{\left(\det\left(\frac{\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_0}{2}\right)\right)^{1/2}}.$$

Because $\boldsymbol{\Sigma}_0$ has eigenvalues bounded away from zero and infinity, by Lemma 2 of [Suarez and Ghosal \(2017\)](#), we obtain that

$$d^2(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_0) \leq \|\boldsymbol{\Sigma}_0^{-1/2}(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_0)\boldsymbol{\Sigma}_0^{-1/2}\|_F^2 \lesssim d^2(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_0). \quad (3.5.18)$$

Since

$$-\log \left(\frac{(\det(\boldsymbol{\Sigma}))^{1/4} (\det(\boldsymbol{\Sigma}_0))^{1/4}}{\left(\det\left(\frac{\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_0}{2}\right)\right)^{1/2}} \right) = -\log(1 - d^2(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_0)) \asymp d^2(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_0),$$

we obtain that $\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_0\|_F^2 \lesssim \epsilon_n^2$.

We now show that the probability (3.5.17) goes to 1 implies (3.4.2). Given (3.4.3) and by Assumption 3, we obtain that

$$\|\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_0\|^2 = \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_0 + 2\boldsymbol{\Sigma}_0\|^2 \leq 2\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_0\|_F^2 + 8\|\boldsymbol{\Sigma}_0\|^2 \leq 2\epsilon_n^2 + 8b_2^2.$$

Hence, by

$$\text{eig}_1\left(\left(\frac{\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_0}{2}\right)^{-1}\right) = \left(\text{eig}_d\left(\frac{\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_0}{2}\right)\right)^{-1} = \left\|\frac{\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_0}{2}\right\|^{-1},$$

where $\text{eig}_1(\mathbf{A})$ and $\text{eig}_d(\mathbf{A})$ are the smallest and the largest eigenvalues of a matrix \mathbf{A}

respectively. Then (3.5.17) implies that

$$\epsilon_n^2 \geq \frac{1}{n} \sum_{i=1}^n \|X_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|^2 \left\| \frac{\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_0}{2} \right\|^{-1} \geq \frac{1}{n} \sum_{i=1}^n \|X_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|^2 / \sqrt{\epsilon_n^2/2 + 2b_2^2}.$$

Combining with (3.4.3), we obtain (3.4.2). \square

Proof of Lemma 3.4.3. Let $\mathcal{S} = \{S : s_1 < r_1, \dots, s_d < r_d\}$, where $r_k \geq M_2 s_k^*$ for $k = 1, \dots, d$, we need to show that $\mathbb{E}_0 \Pi(S : S \in \mathcal{S}^c | Y_1, \dots, Y_n) \rightarrow 0$ as $n \rightarrow \infty$. The posterior probability $\Pi(\mathcal{S}^c | Y_1, \dots, Y_n)$ is given by

$$\Pi(\mathcal{S}^c | Y_1, \dots, Y_n) = \frac{\int \int_{\mathcal{S}^c} \prod_{i=1}^n \frac{f_i}{f_{0,i}}(Y_i) d\Pi(\boldsymbol{\beta}) d\Pi(\boldsymbol{\Sigma})}{\int \int \prod_{i=1}^n \frac{f_i}{f_{0,i}}(Y_i) d\Pi(\boldsymbol{\beta}) d\Pi(\boldsymbol{\Sigma})}. \quad (3.5.19)$$

By Lemma 3.4.2, the denominator of (3.5.19) is bounded below by $e^{-(1+C_2)n\epsilon_n^2}$ with a large probability. For the numerator of the posterior (3.5.19), we have

$$\mathbb{E}_0 \left(\int \int_{\mathcal{S}^c} \prod_{i=1}^n \frac{f_i}{f_{0,i}}(Y_i) d\Pi(\boldsymbol{\beta}) d\Pi(\boldsymbol{\Sigma}) \right) \leq \int_{\mathcal{S}^c} d\Pi(\boldsymbol{\beta}) \leq \sum_{k=1}^d \Pi(s_k \geq r_k). \quad (3.5.20)$$

By Assumption 1 and $A_2/G^{A_4} \leq 1/2$ as $G \rightarrow \infty$, for each k ,

$$\Pi(s_k \geq r_k) = \sum_{s_k=r_k}^{\infty} \pi_G(s_k) \leq \pi_G(s_{0,k}) \left(\frac{A_2}{G^{A_4}} \right)^{r_k - s_{0,k}} \sum_{j=0}^{\infty} \left(\frac{A_2}{G^{A_4}} \right)^j \leq 2 \left(\frac{A_2}{G^{A_4}} \right)^{r_k}.$$

Therefore, $\sum_{k=1}^d \Pi(s_k > r_k) \leq 2 \sum_{k=1}^d (A_2/G^{A_4})^{r_k}$ and we obtain that

$$\begin{aligned} & \mathbb{E}_0 \Pi(\mathcal{S}^c | Y_1, \dots, Y_n) \\ & \leq \mathbb{E}_0 \Pi(\mathcal{S}^c | Y_1, \dots, Y_n) \mathbb{1}_{E_n} + \mathbb{P}_0(E_n^c) \\ & \leq \exp \left((1 + C_2)n\epsilon_n^2 + \log 2 + \log \left(\sum_{k=1}^d (A_2/G^{A_4})^{r_k} \right) \right) + o(1). \end{aligned} \quad (3.5.21)$$

Now we shall show that (3.5.21) goes to 0 as $G \rightarrow \infty$. We write $n\epsilon_n^2 = \sum_{k=1}^d s_k^* \log G$.

Then the expression in the exponential function of (3.5.21) equals to

$$\log 2 + \sum_{k=1}^d \log \left(\sum_{k=1}^d G^{(1+C_2)s_k^* - A_4 r_k} A_2^{r_k} \right). \quad (3.5.22)$$

For $r_k = M_2 s_k^*$,

$$\sum_{k=1}^d G^{(1+C_2)s_k^* - A_4 M_2 s_k^*} A_2^{M_2 s_k^*} \leq d \max_k (G^{(1+C_2 - A_4 M_2)s_k^*} A_2^{M_2 s_k^*}) \rightarrow 0,$$

as $G \rightarrow \infty$ if $\log d \leq A_4 \log G$ and $M_2 \geq 2(1 + C_2)/A_4 + 1$. Therefore, (3.5.22) goes to ∞ and thus (3.5.21) goes to 0. We then complete the proof. \square

Proof of Corollary 3.4.5. By Lemma 3.4.3 and Definition 3.4.6, we have

$$\sum_{i=1}^n \|X_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|^2 \geq \phi_{\ell_2}^2(s^*) \sum_{i=1}^n \|X_i\|^2 \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_F^2.$$

Plugging-in the inequality into (3.4.2), we obtain (3.4.8). \square

Proof of Theorem 3.4.8. Let

$$\Theta_n = \left\{ \boldsymbol{\beta} : |S| \leq M_2 s^*, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_F^2 \leq (M_3 n \epsilon_n^2) / \left(\sum_{i=1}^n \|X_i\|^2 \phi_{\ell_2}^2(s^*) \right) \right\},$$

$$\mathcal{H}_n = \left\{ \boldsymbol{\Sigma} : \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_0\|_F^2 \leq M_1 \epsilon_n^2 \right\}.$$

The proof contains two parts. In the first part, we show that the total variation distance between two measures $\Pi(\cdot|Y_1, \dots, Y_n)$ and $\Pi_{\Theta_n}(\cdot|Y_1, \dots, Y_n)$ is small, where the second measure is the renormalized measure of $\Pi(\cdot|Y_1, \dots, Y_n)$ restricted to the set Θ_n . We also show that the total variation distance between $\Pi^\infty(\cdot|Y_1, \dots, Y_n)$ and $\Pi_{\Theta_n}^\infty(\cdot|Y_1, \dots, Y_n)$ is also small, where the second measure is the renormalized measure of $\Pi^\infty(\cdot|Y_1, \dots, Y_n)$ that is restricted to the same set. In the second part, we show that the total variation distance between $\Pi(\cdot|Y_1, \dots, Y_n)$ and $\Pi^\infty(\cdot|Y_1, \dots, Y_n)$ is small.

Part I. For any set A , let $\Pi_A(\cdot)$ be the renormalized measure which restricted to the set A . Then $\|\Pi(\cdot) - \Pi_A(\cdot)\| \leq 2\Pi(A^c)$. Clearly,

$$\|\Pi(\cdot|Y_1, \dots, Y_n) - \Pi_{\Theta_n}(\cdot|Y_1, \dots, Y_n)\|_{TV} \leq 2\Pi(\boldsymbol{\beta} \notin \Theta_n|Y_1, \dots, Y_n)$$

$$\leq 2\Pi(|S| \geq M_2 s^* |Y_1, \dots, Y_n) + 2\Pi\left(\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_F^2 \geq \frac{M_3 n \epsilon_n^2}{\sum_{i=1}^n \|X_i\|^2 \phi_{\ell_2}^2(s^*)} \middle| Y_1, \dots, Y_n\right) \\ \rightarrow 0,$$

by (3.4.2) and (3.4.7).

Now, to show that

$$\|\Pi^\infty(\cdot | Y_1, \dots, Y_n) - \Pi_{\Theta_n}^\infty(\cdot | Y_1, \dots, Y_n)\|_{TV} \leq 2\Pi^\infty(\boldsymbol{\beta} \in \Theta_n^c | Y_1, \dots, Y_n),$$

we write

$$\begin{aligned} & \Pi^\infty(\boldsymbol{\beta} \in \Theta_n^c | Y_1, \dots, Y_n) \\ &= \frac{\int_{\Theta_n^c} \exp\{\ell(\text{Vec}(\boldsymbol{\beta}), \boldsymbol{\Sigma}_0) - \ell(\text{Vec}(\boldsymbol{\beta}_0), \boldsymbol{\Sigma}_0)\} dU(\text{Vec}(\boldsymbol{\beta}))}{\int \exp\{\ell(\text{Vec}(\boldsymbol{\beta}), \boldsymbol{\Sigma}_0) - \ell(\text{Vec}(\boldsymbol{\beta}_0), \boldsymbol{\Sigma}_0)\} dU(\text{Vec}(\boldsymbol{\beta}))}, \end{aligned} \quad (3.5.23)$$

where

$$dU(\text{Vec}(\boldsymbol{\beta})) = \prod_{k=1}^d \sum_{S_k \in \mathcal{S}_k^*} \frac{\pi_G(|S_k|)}{\binom{G}{|S_k|}} \prod_{j \in S_k} \left(\frac{\lambda_k}{a_j}\right)^{p_j} d\Pi(\text{Vec}(\boldsymbol{\beta}_S)) \otimes \delta_{S^c}.$$

Recall that $\mathcal{S}_k^* = \{S_k : |S_k| \leq M_2 s_k^*\}$.

By (3.4.12), $\ell(\text{Vec}(\boldsymbol{\beta}), \boldsymbol{\Sigma}_0) - \ell(\text{Vec}(\boldsymbol{\beta}_0), \boldsymbol{\Sigma}_0)$ equals to

$$\sum_{k=1}^d \left(-\frac{1}{2} \sum_{i=1}^n \|(\beta_k - \beta_{0,k}) \mathbf{X}_{i,k} \boldsymbol{\Sigma}_0^{-1/2}\|^2 + \sum_{i=1}^n (Y_i - \text{Vec}(\boldsymbol{\beta}_0) \mathbf{X}_i) \boldsymbol{\Sigma}_0^{-1} \mathbf{X}'_{i,k} (\beta_k - \beta_{0,k})' \right).$$

Note that except for the elements of the first row of $\mathbf{X}_{i,k}$, the rest are 0.

By plugging-in the last display into (3.5.23), the denominator can be bounded below by

$$\begin{aligned} & \prod_{k=1}^d \frac{\pi_G(s_{0,k})}{\binom{G}{s_{0,k}}} \prod_{j \in S_{0,k}} \left(\frac{\lambda_k}{a_j}\right)^{p_j} \int \exp\left(-\frac{1}{2} \sum_{i=1}^n \|(\beta_{S_k} - \beta_{0,k}) \mathbf{X}_{i,S_{0,k}} \boldsymbol{\Sigma}_0^{-1/2}\|^2\right) \\ & \times \exp\left(\sum_{i=1}^n (Y_i - \text{Vec}(\boldsymbol{\beta}_0) \mathbf{X}_{i,S_{0,k}}) \boldsymbol{\Sigma}_0^{-1} \mathbf{X}'_{i,S_{0,k}} (\beta_{S_k} - \beta_{0,k})'\right) d\beta_{S_k}. \end{aligned}$$

By changing the variables $\beta_{S_k} - \beta_{0,k}$ to $\tilde{\beta}_{S_k}$ and applying Jensen's inequality, the last

display is bounded below by

$$\begin{aligned}
& \prod_{k=1}^d \frac{\pi_G(s_{0,k})}{\binom{G}{s_{0,k}}} \prod_{j \in S_{0,k}} \left(\frac{\lambda_k}{a_j} \right)^{p_j} \int \exp \left(-\frac{1}{2} \sum_{i=1}^n \|\check{\beta}_{S_k} \mathbf{X}_{i,S_{0,k}} \Sigma_0^{-1/2}\|^2 \right) d\check{\beta}_{S_k} \\
&= \prod_{k=1}^d \frac{\pi_G(s_{0,k})}{\binom{G}{s_{0,k}}} \prod_{j \in S_{0,k}} \left(\frac{\lambda_k}{a_j} \right)^{p_j} \frac{(2\pi)^{\sum_{j \in S_{0,k}} p_j/2}}{(\det(\sum_{i=1}^n \mathbf{X}_{i,S_{0,k}} \Sigma_0^{-1} \mathbf{X}'_{i,S_{0,k}}))^{1/2}}. \tag{3.5.24}
\end{aligned}$$

We thus obtain a lower bound for the denominator.

The numerator of (3.5.23) can be written as follows:

$$\begin{aligned}
& \int_{\Theta_n^c} \exp \left(-\frac{1}{2} \sum_{i=1}^n \|\text{Vec}(\beta_S - \beta_0) \mathbf{X}_{i,S} \Sigma_0^{-1/2}\|^2 \right) \\
& \quad \times \exp \left(\sum_{i=1}^n (Y_i - \text{Vec}(\beta_0) \mathbf{X}_{i,S}) \Sigma_0^{-1} \mathbf{X}'_{i,S} \text{Vec}(\beta_S - \beta_0)' \right) dU(\text{Vec}(\beta_S)).
\end{aligned}$$

By applying the tail bound for a standard multivariate normal distribution,

$$\mathbb{P} \left(\max_k \left\| \sum_{i=1}^n (Y_i - \text{Vec}(\beta_0) \mathbf{X}_i) \Sigma_0^{-1} \mathbf{X}_{i,k} \right\| \geq 2 \sqrt{b_1 \log n \sum_{i=1}^n \|X_i\|^2} \right) \leq \frac{2d}{n} \rightarrow 0,$$

as $d \ll n$. Note that $2\sqrt{b_1 \log n \sum_{i=1}^n \|X_i\|^2} = \bar{\lambda} \sqrt{b_1 \log n / \log G}$. By the Cauchy-Schwartz inequality, with probability tending to one, we obtain that

$$\begin{aligned}
& \sum_{i=1}^n (Y_i - \text{Vec}(\beta_0) \mathbf{X}_i) \Sigma_0^{-1} \mathbf{X}'_i (\text{Vec}(\beta - \beta_0))' \\
&= \sum_{k=1}^d \sum_{i=1}^n (Y_i - \text{Vec}(\beta_0) \mathbf{X}_i) \Sigma_0^{-1} \mathbf{X}'_{i,k} (\beta_k - \beta_{0,k})' \leq \frac{\bar{\lambda} \sqrt{b_1 \log n}}{\sqrt{\log G}} \sum_{k=1}^d \|\beta_k - \beta_{0,k}\|.
\end{aligned}$$

By applying the inequality $\sum_{i=1}^d \|A_i\| \leq \sqrt{d \sum_{i=1}^d \|A_i\|^2}$ for vectors A_1, \dots, A_d and writing $x = 3x/2 - x/2$ for any x , the upper bound in the last display is bounded above by

$$\frac{3\bar{\lambda} \sqrt{db_1 \log n}}{2\sqrt{\log G}} \sqrt{\sum_{k=1}^d \|\beta_k - \beta_{0,k}\|^2} - \frac{\bar{\lambda} \sqrt{b_1 \log n}}{2\sqrt{\log G}} \sum_{k=1}^d \|\beta_k - \beta_{0,k}\|. \tag{3.5.25}$$

Furthermore, by writing $3x/2 = 2x - x/2$ for any x and using the restricted eigenvalue condition in (3.4.6), the first term in (3.5.25) can be further bounded above by

$$\frac{2\bar{\lambda}\sqrt{db_1 \log n} \sum_{i=1}^n \|\text{Vec}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \mathbf{X}_i\|^2}{\sqrt{\log G} \sum_{i=1}^n \|X_i\|^2 \phi_{\ell_2}^2(s^*)} - \frac{\bar{\lambda}\sqrt{db_1 \log n}}{2\sqrt{\log G}} \sqrt{\sum_{k=1}^d \|\beta_k - \beta_{0,k}\|^2}.$$

We apply the inequality $2ab \leq a^2 + b^2$ for any two numbers a and b to the first term in the last display. As $\boldsymbol{\beta} \in \Theta_n^c$, recall that the posterior is concentrated on the set $\mathcal{S}^* = \{S : |S| \leq M_2 s^*\}$, then the last display can be further bounded above by

$$\frac{1}{2} \sum_{i=1}^n \|\text{Vec}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \mathbf{X}_i\|^2 + \frac{2\bar{\lambda}^2 db_1 \log n}{\log G \sum_{i=1}^n \|X_i\|^2 \phi_{\ell_2}^2(s^*)} - \frac{\bar{\lambda} \epsilon_n \sqrt{M_3 n d b_1 \log n}}{2\sqrt{\log G} \sum_{i=1}^n \|X_i\|^2 \phi_{\ell_2}^2(s^*)}. \quad (3.5.26)$$

Now we combine the upper bounds (3.5.25) and (3.5.26), the numerator of the posterior (3.5.23) is bounded above by

$$\begin{aligned} & \exp \left(\frac{2\bar{\lambda}^2 db_1 \log n}{\log G \sum_{i=1}^n \|X_i\|^2 \phi_{\ell_2}^2(s^*)} - \frac{\bar{\lambda} \epsilon_n \sqrt{M_5 n d b_1 \log n}}{\sqrt{\log G} \sum_{i=1}^n \|X_i\|^2 \phi_{\ell_2}^2(s^*)} \right) \\ & \times \int \exp \left(- \frac{\sqrt{b_1 \log n}}{2\sqrt{\log G}} \sum_{k=1}^d \lambda_k \|\beta_k - \beta_{0,k}\| \right) dU(\text{Vec}(\boldsymbol{\beta}_S)). \end{aligned}$$

The integral part in the last display can be further bounded above by

$$\begin{aligned} & \prod_{k=1}^d \int \prod_{j \in S_k} \left(\frac{\lambda_k}{a_j} \right)^{p_j} \exp \left(- \frac{\lambda_k \sqrt{b_1 \log n}}{2\sqrt{\log G}} \|\beta_k - \beta_{0,k}\| \right) d\beta_k \sum_{s_k=0}^G \pi_G(s_k) \mathbb{1}\{S_k \in \mathcal{S}_k^*\} \\ & = \prod_{k=1}^d \prod_{j \in S_k} \left(\frac{\log G}{b_1 \log n} \right)^{p_j/2} \sum_{s_k=0}^G \pi_G(s_k) \mathbb{1}\{S_k \in \mathcal{S}_k^*\}. \end{aligned}$$

Finally, we obtain the upper bound for the numerator, which is

$$\exp \left(\frac{2\bar{\lambda}^2 db_1 \log n}{\log G \sum_{i=1}^n \|X_i\|^2 \phi_{\ell_2}^2(s^*)} - \frac{\bar{\lambda} \epsilon_n \sqrt{M_5 n d b_1 \log n}}{\sqrt{\log G} \sum_{i=1}^n \|X_i\|^2 \phi_{\ell_2}^2(s^*)} \right)$$

$$\begin{aligned}
& \times \exp\left(-\frac{\sum_{k=1}^d \sum_{j \in S_k} p_j}{2} \log\left(\frac{b_1 \log n}{\log G}\right)\right) \prod_{k=1}^d \sum_{s_k=0}^G \pi_G(s_k) \mathbb{1}\{S_k \in \mathcal{S}_k^*\} \\
& = \exp\left(\frac{18db_1 \log n}{\phi_{\ell_2}^2(s^*)} - \frac{3\epsilon_n \sqrt{M_5 n d b_1 \log n}}{\phi_{\ell_2}(s^*)} - \frac{\sum_{k=1}^d \sum_{j \in S_k} p_j}{2} \log\left(\frac{b_1 \log n}{\log G}\right)\right) \\
& \quad \times \prod_{k=1}^d \sum_{s_k=0}^G \pi_G(s_k) \mathbb{1}\{S_k \in \mathcal{S}_k^*\}. \tag{3.5.27}
\end{aligned}$$

Now, we combine the lower bound of the denominator (3.5.24) and the upper bound of the numerator (3.5.27). Then the posterior (3.5.23) is bounded above by

$$\begin{aligned}
& \prod_{k=1}^d \frac{\binom{G}{s_{0,k}}}{\pi_G(s_{0,k})} \prod_{j \in S_{0,k}} \left(\frac{a_j}{\lambda_k}\right)^{p_j} \frac{(\det(\sum_{i=1}^n \mathbf{X}_{i,S_{0,k}} \Sigma_0^{-1} \mathbf{X}'_{i,S_{0,k}}))^{1/2}}{(2\pi)^{\sum_{j \in S_{0,k}} p_j / 2}} \\
& \quad \times \exp\left(\frac{18db_1 \log n}{\phi_{\ell_2}^2(s^*)} - \frac{3\epsilon_n \sqrt{M_3 n d b_1 \log n}}{\phi_{\ell_2}(s^*)} - \frac{\sum_{k=1}^d \sum_{j \in S_k} p_j}{2} \log\left(\frac{b_1 \log n}{\log G}\right)\right) \\
& \quad \times \sum_{s_k=0}^G \pi_G(s_k) \mathbb{1}\{S_k \in \mathcal{S}_k^*\}. \tag{3.5.28}
\end{aligned}$$

Note that $\Sigma_0 \geq b_1 \mathbf{I}_d$, by letting $\Gamma_{S_{0,k}} = \sum_{i=1}^n \mathbf{X}_{i,S_{0,k}} \Sigma_0^{-1} \mathbf{X}'_{i,S_{0,k}}$ and applying Jensen's inequality,

$$\begin{aligned}
\det(\Gamma_{S_{0,k}}) & \leq \left(\frac{1}{\sum_{j \in S_{0,k}} p_j} \text{Tr}(\Gamma_{S_{0,k}})\right)^{\sum_{j \in S_{0,k}} p_j} \leq \left(\frac{1}{b_1} \sum_{i=1}^n \|\mathbf{X}_{i,S_{0,k}}\|^2\right)^{\sum_{j \in S_{0,k}} p_j} \\
& \leq \left(\frac{1}{b_1} \sum_{i=1}^n \|X_i\|^2\right)^{\sum_{j \in S_{0,k}} p_j}.
\end{aligned}$$

Thus

$$\prod_{j \in S_{0,k}} \left(\frac{a_j}{\lambda_k}\right)^{p_j} \det(\Gamma_{S_{0,k}})^{1/2} \leq \left(\frac{p_j}{\sqrt{b_1}}\right)^{\sum_{j \in S_{0,k}} p_j} G^{s_{0,k}} = \exp\left(\sum_{j \in S_{0,k}} p_j \log(p_j / \sqrt{b_1})\right) G^{s_{0,k}}.$$

By putting the term $\exp\left(\sum_{j \in S_{0,k}} p_j \log(p_j / \sqrt{b_1})\right)$ together with the exponential expression in (3.5.28) and choosing a large enough value of M_3 , then the exponential term goes to 0 as $n \rightarrow \infty$ (note that we assume that $\phi_{\ell_2}(s^*)$ is bounded below by a constant c). Also, the prior mass $\pi_G(s_k)$ can be bounded below by G^{-s_k} . As a result, the product of

the rest terms also goes to 0. Therefore, the posterior goes to 0 as $n \rightarrow \infty$.

Part II. For a generic set B ,

$$\begin{aligned} \Pi_{\Theta_n}(B|Y_1, \dots, Y_n) &\propto \prod_{k=1}^d \sum_{S_k \in \mathcal{S}^*} \frac{\pi_G(S_k)}{\binom{G}{s_k}} \prod_{j \in S_k} \left(\frac{\lambda_k}{a_j} \right)^{p_j} \\ &\times \int \int_{(B \cap \Theta_n)} \exp\left(-\frac{1}{2} \sum_{i=1}^n \|(Y_i - \text{Vec}(\boldsymbol{\beta}_S) \mathbf{X}_{i,S}) \boldsymbol{\Sigma}^{-1/2}\|^2\right) \exp(-\lambda_k \|\beta_k\|_{2,1}) d\boldsymbol{\beta}_S d\Pi(\boldsymbol{\Sigma}), \end{aligned}$$

and

$$\begin{aligned} \Pi_{\Theta_n}^\infty(B|Y_1, \dots, Y_n) &\propto \prod_{k=1}^d \sum_{S_k \in \mathcal{S}^*} \frac{\pi_G(S_k)}{\binom{G}{s_k}} \prod_{j \in S_k} \left(\sum_{k=1}^d \frac{\lambda_k}{a_j} \right)^{p_j} \\ &\times \int_{(B \cap \Theta_n)} \exp\left(-\frac{1}{2} \sum_{i=1}^n \|(Y_i - \text{Vec}(\boldsymbol{\beta}_S) \mathbf{X}_{i,S}) \boldsymbol{\Sigma}_0^{-1/2}\|^2\right) \exp\left(-\sum_{k=1}^d \lambda_k \|\beta_{0,k}\|_{2,1}\right) d\boldsymbol{\beta}_S. \end{aligned}$$

In the following, we shall show that the total variation distance between the above two measures goes to 0 as $n \rightarrow \infty$.

We take an S_k from the corresponding set \mathcal{S}_k^* for all k and denote the corresponding measure by $\Pi_{\Theta_n}^S(B|Y_1, \dots, Y_n)$ and $\Pi_{\Theta_n}^{S,\infty}(B|Y_1, \dots, Y_n)$. We first show that the total variation distance between the two measures goes to 0. To this end, we use the Bernstein von-Mises theorem (see Theorem 1 of [Castillo, 2010](#)) in a semiparametric model. The main difference with [Castillo \(2010\)](#)'s setup is that is the nuisance part in our model is parametric but high-dimensional. This theorem requires calculating the remainder term in the local asymptotic normality (LAN) expansion of the log-likelihood, which is denoted as $\text{Rem}_n(\text{Vec}(\boldsymbol{\beta}_t), \boldsymbol{\Sigma}_t)$, and showing that

$$\sup_{\boldsymbol{\beta} \in \Theta_n, \boldsymbol{\Sigma} \in \mathcal{H}_m} \frac{|\text{Rem}_n(\text{Vec}(\boldsymbol{\beta}_0), \boldsymbol{\Sigma}) - \text{Rem}_n(\text{Vec}(\boldsymbol{\beta}), \boldsymbol{\Sigma})|}{1 + n \|\text{Vec}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|^2} \rightarrow 0. \quad (3.5.29)$$

For simplicity, we denote $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ and define two local paths,

$$\text{Vec}(\boldsymbol{\beta}_t) = \text{Vec}(\boldsymbol{\beta}_0) + tb, \quad \boldsymbol{\Omega}_t = \boldsymbol{\Omega}_0 + \frac{t\boldsymbol{\Phi}}{\|\boldsymbol{\Sigma}_0^{1/2} \boldsymbol{\Phi} \boldsymbol{\Sigma}_0^{1/2}\|_F},$$

where b is a $1 \times pd$ vector, $\boldsymbol{\Phi}$ is a $d \times d$ symmetric matrix and $t = n^{-1/2}$. Then by Lemma

3.6.5, the remainder term in the LAN expansion on the log-likelihood is given by

$$\text{Rem}_n(\text{Vec}(\boldsymbol{\beta}_t), \boldsymbol{\Sigma}_t) = \left| \frac{n}{2} \sum_{k=1}^d \int_0^{\rho_k} \frac{(\rho_k - s)^2}{(1-s)^3} ds \right|,$$

where ρ_k is the k -th eigenvalue of $\boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{\Omega}_t - \boldsymbol{\Omega}_0)\boldsymbol{\Sigma}_0^{1/2}$. Since the remainder term does not depend on $\boldsymbol{\beta}$, $|\text{Rem}_n(\text{Vec}(\boldsymbol{\beta}_0), \boldsymbol{\Sigma}) - \text{Rem}_n(\text{Vec}(\boldsymbol{\beta}), \boldsymbol{\Sigma})| = 0$. Thus we verified (3.5.29). We now conclude that the total variation distance between the two measures $\Pi_{\Theta_n}^S(B|Y_1, \dots, Y_n)$ and $\Pi_{\Theta_n}^{S, \infty}(B|Y_1, \dots, Y_n)$ goes to 0. Then the total variation distance between the two measures $\Pi_{\Theta_n}(B|Y_1, \dots, Y_n)$ and $\Pi_{\Theta_n}^\infty(B|Y_1, \dots, Y_n)$ also goes to 0. \square

Proof of Theorem 3.4.9. The proof is similar to the proof of Theorem 4 of Castillo et al. (2015), thus we will simplify our proofs when the results can be obtained easily by following their arguments. Let Ξ be a collection of all sets for S such that $S \in \mathcal{S}_0$, $S_k \supset S_{0,k}$ and $S_k \neq S_{0,k}$, where $\mathcal{S}_0 = \cup_{k=1}^d \{S_k : |S_k| \leq M_2 s_{0,k}, \beta_{0,S_k} \in \tilde{\mathcal{B}}\}$ and let $\Gamma_{S_k} = \sum_{i=1}^n \mathbf{X}_{i,S_k} \boldsymbol{\Sigma}_0^{-1} \mathbf{X}'_{i,S_k}$. We shall show that $\Pi^\infty(\boldsymbol{\beta} : S \in \Xi | Y_1, \dots, Y_n) \rightarrow 0$ in probability.

By (3.4.14), we obtain that

$$\begin{aligned} & \Pi^\infty(\boldsymbol{\beta} : S \in \mathcal{S}_0 | Y_1, \dots, Y_n) \\ & \leq \prod_{k=1}^d \sum_{S_k \in \mathcal{S}_0} w_{S_k}^\infty \\ & \leq \prod_{k=1}^d \sum_{s_k = s_{0,k} + 1}^{M_2 s_{0,k}} \frac{\pi_G(s_k) \binom{G}{s_{0,k}} \binom{G-s_{0,k}}{s_k - s_{0,k}}}{\pi_G(s_{0,k}) \binom{G}{s_k}} \max_{\substack{S_k \in \Xi, \\ |S_k| = s_k}} \frac{\left(\prod_{j \in S_k} (\lambda_k / a_j)^{p_j} \right) (2\pi)^{\sum_{j \in S_k} p_j / 2}}{\left(\prod_{j' \in S_{0,k}} (\lambda_k / a_{j'})^{p_{j'}} \right) (2\pi)^{\sum_{j' \in S_{0,k}} p_{j'} / 2}} \\ & \quad \times \left(\frac{\det(\Gamma_{S_{0,k}})}{\det(\Gamma_{S_k})} \right)^{1/2} \exp \left\{ \frac{1}{2} \sum_{i=1}^n \|\text{Vec}(\hat{\boldsymbol{\beta}}_{S_k}) \mathbf{X}_{i,S_k} \boldsymbol{\Sigma}_0^{-1/2}\|^2 \right\} \\ & \quad \times \exp \left\{ -\frac{1}{2} \sum_{i=1}^n \|\text{Vec}(\hat{\boldsymbol{\beta}}_{S_{0,k}}) \mathbf{X}_{i,S_{0,k}} \boldsymbol{\Sigma}_0^{-1/2}\|^2 \right\}, \end{aligned} \tag{3.5.30}$$

where $M_2 = 2(1 + C_2)/A_4 + 1$ (see Lemma 3.4.3).

By using the same techniques as those used to prove (6.11) and (6.12) of Castillo

et al. (2015), we can verify that

$$\frac{\prod_{j \in S_k} (\lambda_k / a_j)^{p_j} (\det(\Gamma_{S_{0,k}}))^{1/2}}{\prod_{j' \in S_{0,k}} (\lambda_k / a_{j'})^{p_{j'}} (\det(\Gamma_{S_k}))^{1/2}} \leq \left(\frac{9\sqrt{\log G}}{b_1 \phi_{\ell_2}(s_0)} \right)^{\sum_{j \in S_{0,k}} p_j - \sum_{j' \in S_k} p_{j'}},$$

and

$$\begin{aligned} & \mathbb{P} \left(\sum_{i=1}^n \|\text{Vec}(\hat{\beta}_{S_k}) \mathbf{X}_{S_k} \Sigma_0^{-1/2}\|^2 - \sum_{i=1}^n \|\text{Vec}(\hat{\beta}_{S_{0,k}}) \mathbf{X}_{S_{0,k}} \Sigma_0^{-1/2}\|^2 \right. \\ & \quad \left. \leq t_k \left(\sum_{j \in S_k} p_j - \sum_{j' \in S_{0,k}} p_{j'} \right) (\log G) / b_2 \right) \rightarrow 1, \end{aligned}$$

where $t_k > 2b_2(s_k - s_{0,k}) / (\sum_{j \in S_k} p_j - \sum_{j' \in S_{0,k}} p_{j'})$. Then (3.5.30) is bounded above by

$$\begin{aligned} & \prod_{k=1}^d \sum_{s_k = s_{0,k} + 1}^{M_2 s_{0,k}} (A_1 G^{-A_4})^{s_k - s_{0,k}} \max_{\substack{S_k \in \Xi, \\ |S_k| = s_k}} \frac{\binom{s_k}{s_{0,k}} \prod_{j' \in S_{0,k}} a_j^{p_{j'}}}{\prod_{j \in S_k} a_j^{p_j}} \\ & \quad \times \left(\frac{3\sqrt{2\pi \log G}}{b_1 \phi_{\ell_2}(s_0)} \right)^{\sum_{j \in S_k} p_j - \sum_{j' \in S_{0,k}} p_{j'}} G^{s_k - s_{0,k}}, \end{aligned}$$

with probability tending to 1. The last display goes to 0 as $n \rightarrow \infty$ since $s_{0,k} \leq G^{A_4 - 1}$ and $\binom{s_k}{s_{0,k}} \leq (M_2 G^{A_4 - 1})^{s_k - s_{0,k}}$. \square

3.6 Auxiliary results

Lemma 3.6.1. For $a = \sqrt{\pi} \left(\frac{\Gamma(m+1)}{\Gamma(\frac{m}{2}+1)} \right)^{1/m}$,

$$\int_{\mathbb{R}^m} \left(\frac{\lambda}{a} \right)^m \exp(-\lambda \|(X_1, \dots, X_m)\|) dX_1 \cdots dX_m = 1. \quad (3.6.1)$$

Also as $m \rightarrow \infty$, $a \asymp m^{1/2}$. Expressing X_i in the spherical polar coordinates by a radius r , a base angle $\theta_{m-1} \in (0, 2\pi)$, and m_2 angles $\theta_1, \dots, \theta_{m-2}$ ranging over $(-\pi/2, \pi/2)$, then the density of r is given by

$$f(r|\lambda) = \frac{\lambda^m}{\Gamma(m)} r^{m-1} \exp(-\lambda r), \quad (3.6.2)$$

which is a gamma density with the shape parameter m and rate parameter λ .

Proof. Applying the polar transformation, evaluating the Jacobian, and applying the results shown in Chapter 1.5.1 of [Scott \(2015\)](#), the integral in (3.6.1) equals to

$$\begin{aligned} & \int_0^{2\pi} \int_{-\pi/2}^{\pi/2} \cdots \int_{-\pi/2}^{\pi/2} \int \left(\frac{\lambda}{a}\right)^m r^{m-1} \exp(-\lambda r) \prod_{i=1}^{m-2} \cos^i \theta_{m-i-1} dr d\theta_1 \cdots d\theta_{m-2} d\theta_{m-1} \\ &= \int \frac{2\pi^{m/2} \lambda^m}{\Gamma(m/2) a^m} r^{m-1} \exp(-\lambda r) dr. \end{aligned} \quad (3.6.3)$$

The second line of the last display is obtained by using the results in Chapter 1.5.2 of [Scott \(2015\)](#). Since $\int r^{m-1} e^{-\lambda r} dr = \Gamma(m)/\lambda^m$, the choice

$$a = \sqrt{\pi} \left(\frac{2\Gamma(m)}{\Gamma(m/2)} \right)^{1/m} = \sqrt{\pi} \left(\frac{\Gamma(m+1)}{\Gamma(m/2+1)} \right)^{1/m}$$

makes $(\lambda/a)^m \exp(-\lambda \|(X_1, \dots, X_m)\|)$ a probability density function. Now by Stirling's approximation to the gamma functions, we obtain that $\frac{\sqrt{2\pi}}{e} \left(\frac{2m}{e}\right)^{1/2} \leq a \leq \frac{e}{\sqrt{2}} \left(\frac{2m}{e}\right)^{1/2}$, which implies that $a \asymp m^{1/2}$. (3.6.2) is self-evident from (3.6.3). \square

Lemma 3.6.2 ([Lounici et al., 2011](#)). *Let $\xi_1, \dots, \xi_k \sim \mathcal{N}(0, 1)$ be i.i.d random variables and $v = (v_1, \dots, v_k)$ be a non-zero vector. Define $\eta_v = \sum_{i=1}^k (\xi_i^2 - 1)v_i / (\sqrt{2}\|v\|)$ and $m(v) = \|v\|_\infty / \|v\|$. Then for $t > 0$,*

$$\mathbb{P}(|\eta_v| > t) \leq 2 \exp\left(-\frac{t^2}{2(1 + \sqrt{2}tm(v))}\right). \quad (3.6.4)$$

Lemma 3.6.3 ([Muirhead, 1982](#)). *If $\mathbf{A} \sim \mathcal{W}_d(\nu, \Phi)$ with $\nu > d-1$, $\rho_1 < \dots < \rho_d$ are the eigenvalues for \mathbf{A} , then the joint distribution for (ρ_1, \dots, ρ_d) is*

$$\begin{aligned} & \frac{\pi^{d^2/2} 2^{-d\nu/2} (\det(\Phi))^{-\nu/2}}{\Gamma_d(d/2) \Gamma_d(\nu/2)} \prod_{i=1}^d \rho_i^{(\nu-d-1)/2} \prod_{i<j}^d (\rho_j - \rho_i) \\ & \times \int_{\mathcal{O}(d)} \exp\left(-\frac{1}{2} \text{Tr}(\Phi^{-1} \mathbf{H} \Delta(\rho) \mathbf{H}^T)\right) d\mathbf{H}, \end{aligned} \quad (3.6.5)$$

where $\mathcal{O}(d)$ denotes to the space of orthogonal matrices, $\Delta(\rho) = \text{diag}(\rho_1, \dots, \rho_d)$, and Γ_d

denotes to the d -dimensional multivariate gamma function, which is defined as

$$\Gamma_d(a) = \int_{\mathbf{B} > 0} \exp(\text{Tr}(-\mathbf{B})) (\det(\mathbf{B}))^{a-(d+1)/2} d\mathbf{B},$$

where $a > (d-1)/2$ and $\mathbf{B} > 0$ means that \mathbf{B} is positive definite.

Further if $\Phi = \mathbf{I}_d$, then (3.6.5) reduces to

$$\frac{\pi^{d^2/2} 2^{-d\nu/2}}{\Gamma_d(d/2) \Gamma_d(\nu/2)} \exp\left(-\frac{1}{2} \sum_{i=1}^d \rho_i\right) \prod_{i=1}^d \rho_i^{(\nu-d-1)/2} \prod_{i < j}^d (\rho_j - \rho_i). \quad (3.6.6)$$

Lemma 3.6.4. If $\Sigma^{-1} \sim \mathcal{W}_d(\nu, \Phi)$, where $\nu > d-1$ and $\nu \asymp d$, then for positive constants $a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4$, with $t_1 > \nu d$, $t_2 > 0$ and $0 \leq t_3 \leq 1$,

$$\mathbb{P}(\text{eig}_d(\Sigma^{-1}) > t_1) \lesssim ((b_1 t_1)/d^2)^{d^2/2} \exp(d^2/2 - a_1 t_1), \quad (3.6.7)$$

$$\mathbb{P}(\text{eig}_1(\Sigma^{-1}) \leq t_2) \lesssim (b_2 d)^{b_3 d^2} t_2^{a_2 d}, \quad (3.6.8)$$

$$\mathbb{P}\left(\bigcap_{i=1}^d \{1 \leq \text{eig}_i(\Sigma^{-1}) \leq 1 + t_3\}\right) \gtrsim (b_4 d)^{-d} d^{-d^2/2} t_3^{d(d+1)/2} \exp(-a_3 d(1 + t_3)) \quad (3.6.9)$$

$$\mathbb{P}\left(\bigcap_{i=1}^d \{1 \leq \text{eig}_i(\Sigma) \leq 1 + t_3\}\right) \gtrsim (b_4 d)^{-d} d^{-d^2/2} t_3^{d(d+1)/2} \exp(-a_4 d). \quad (3.6.10)$$

Proof. The proofs of (3.6.7)–(3.6.9) follow from the proof of Lemma 9.16 of Ghosal and van der Vaart (2017), except that here we need to express factors involving d explicitly.

To prove (3.6.10), we need the following inequality:

$$\mathbb{P}\left(\bigcap_{i=1}^d \{\Sigma : 1 \leq \text{eig}_i(\Sigma) \leq 1 + t_3\}\right) \geq \mathbb{P}\left(\bigcap_{i=1}^d \{\Sigma : 1 - t_3 \leq \rho_i \leq 1\}\right),$$

for $0 \leq t_3 \leq 1$ and $\rho_i = \text{eig}_i(\Sigma^{-1})$, which is the i -th smallest eigenvalue of Σ^{-1} , $i = 1, \dots, d$. Consider the set $I_i = \{1 - (d-i+1)t_3/d, 1 - (d-i+1/2)t_3/d\}$ for each i . It is

easy to verify that if $\rho_i \in I_i$, then $\rho_i \in [1 - t_3, 1]$. By (3.6.5), we obtain that

$$\begin{aligned}
& \mathbb{P}(1 - t_3 \leq \rho_i \leq 1, i = 1, \dots, d) \\
& \geq \frac{\pi^{d^2/2} 2^{-d\nu/2} (\det(\Phi))^{-\nu/2}}{\Gamma_d(d/2) \Gamma_d(\nu/2)} \int_{I_d} \dots \int_{I_1} \prod_{i=1}^d \rho_i^{(\nu-d-1)/2} \prod_{i < j}^d (\rho_j - \rho_i) \\
& \quad \times \int_{\mathcal{O}(d)} \exp\left(-\frac{1}{2} \text{Tr}(\Phi^{-1} \mathbf{H} \Delta(\rho) \mathbf{H}^T)\right) d\mathbf{H} d\rho_1 \dots d\rho_d \\
& \geq \frac{\pi^{d^2/2} 2^{-d\nu/2} (\det(\Phi))^{-\nu/2}}{\Gamma_d(d/2) \Gamma_d(\nu/2)} \left(\frac{t}{2d}\right)^{d(d+1)/2} \exp\left(-\frac{1}{2} \text{Tr}(\Phi^{-1})\right).
\end{aligned}$$

for $j > i$, $l_j - l_i > t/(2d)$. The lower bound in the third line of the last display is obtained by noticing that $-\Delta(\rho) > -\rho_d \mathbf{I}_d > -\mathbf{I}_d$, and $\mathbf{H} \mathbf{H}^T = \mathbf{I}_d$. Then (3.6.10) is obtained by applying Stirling's approximation to the gamma functions in that lower bound. \square

Lemma 3.6.5. *For a multivariate normal density $\mathcal{N}(Y_i | \text{Vec}(\beta) \mathbf{X}_i, \Sigma)$, $i = 1, \dots, n$, let*

$$\text{Vec}(\beta_t) = \text{Vec}(\beta_0) + tb, \quad \Omega_t = \Omega_0 + \frac{t\Phi}{\|\Sigma_0^{1/2} \Phi \Sigma_0^{1/2}\|_F},$$

be local paths for $\text{Vec}(\beta)$ and Ω respectively, where $\Omega = \Sigma^{-1}$, b is a $pd \times 1$ vector, Φ is a $d \times d$ symmetric matrix and $0 < t < 1$. Then

$$\begin{aligned}
& \ell_n(\text{Vec}(\beta_t), \Sigma_t) - \ell_n(\text{Vec}(\beta_0), \Sigma_0) \\
& = -\frac{\sqrt{n} \text{Tr}((\mathbf{Q}_0 - \Sigma_0)\Phi)}{2\|\Sigma_0^{1/2} \Phi \Sigma_0^{1/2}\|_F} + \sqrt{n} \sum_{i=1}^n b \mathbf{X}_i \Omega_t (Y_i - \text{Vec}(\beta_0) \mathbf{X}_i)' \\
& \quad - \frac{1}{2n} \sum_{i=1}^n b \mathbf{X}_i \Omega_t \mathbf{X}_i' b' - \frac{1}{2} - \frac{n}{2} \sum_{i=1}^d \int_0^{\rho_i} \frac{(\rho_i - s)^2}{(1-s)^3} ds, \tag{3.6.11}
\end{aligned}$$

where ρ_i is the i -th eigenvalue of the matrix $\Sigma_0^{1/2}(\Omega_t - \Omega_0)\Sigma_0^{1/2}$.

Proof. The proof is similar to that of Lemma A.1 in Gao and Zhou (2016). Let $\mathbf{Q}_t = \frac{1}{n} \sum_{i=1}^n (Y_i - \text{Vec}(\beta_t) \mathbf{X}_i)' (Y_i - \text{Vec}(\beta_t) \mathbf{X}_i)$ and $\mathbf{Q}_0 = \frac{1}{n} \sum_{i=1}^n (Y_i - \text{Vec}(\beta_0) \mathbf{X}_i)' (Y_i - \text{Vec}(\beta_0) \mathbf{X}_i)$. From (3.4.12), we have

$$\begin{aligned}
& \ell_n(\text{Vec}(\beta_t), \Sigma_t) - \ell_n(\text{Vec}(\beta_0), \Sigma_0) \\
& = \frac{n}{2} (\log(\det(\Sigma_0)) - \log(\det(\Sigma))) - \frac{n}{2} \text{Tr}(\mathbf{Q}_t \Omega_t) + \frac{n}{2} \text{Tr}(\mathbf{Q}_0 \Omega_0)
\end{aligned}$$

$$\begin{aligned}
&= \frac{n}{2} \log \left(\det(\mathbf{I}_n - \boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{\Omega}_0 - \boldsymbol{\Omega}_t)\boldsymbol{\Sigma}_0^{1/2}) \right) - \frac{n}{2} \text{Tr} \left((\mathbf{Q}_0 - \boldsymbol{\Sigma}_0)(\boldsymbol{\Omega}_t - \boldsymbol{\Omega}_0) \right) \\
&\quad - \frac{n}{2} \text{Tr} \left(\boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{\Omega}_t - \boldsymbol{\Omega}_0)\boldsymbol{\Sigma}_0^{1/2} \right) - \frac{n}{2} \text{Tr} \left((\mathbf{Q}_t - \mathbf{Q}_0)\boldsymbol{\Omega}_t \right).
\end{aligned}$$

To obtain (3.6.11), first we plug-in the expressions of \mathbf{Q}_t and \mathbf{Q}_0 into $\text{Tr} \left((\mathbf{Q}_t - \mathbf{Q}_0)\boldsymbol{\Omega}_t \right)$ to obtain that

$$\frac{n}{2} \text{Tr} \left((\mathbf{Q}_t - \mathbf{Q}_0)\boldsymbol{\Omega}_t \right) = \sum_{i=1}^n t b \mathbf{X}_i \boldsymbol{\Omega}_t (Y_i - \text{Vec}(\boldsymbol{\beta}_0) \mathbf{X}_i)' - \frac{t^2}{2} \sum_{i=1}^n b \mathbf{X}_i \boldsymbol{\Omega}_t \mathbf{X}_i' b'.$$

Next, we apply Taylor's expansion with the integral form of the remainder to the log function to obtain that

$$\begin{aligned}
&\log \left(\det(\mathbf{I}_n - \boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{\Omega}_0 - \boldsymbol{\Omega}_t)\boldsymbol{\Sigma}_0^{1/2}) \right) + \text{Tr} \left(\boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{\Omega}_0 - \boldsymbol{\Omega}_t)\boldsymbol{\Sigma}_0^{1/2} \right) \\
&= \sum_{i=1}^d (\rho_i + \log(1 - \rho_i)) \\
&= -\frac{1}{2} \sum_{i=1}^d \rho_i^2 - \frac{1}{2} \sum_{i=1}^d \int_0^{\rho_i} \frac{(\rho_i - s)^2}{(1 - s)^3} ds \\
&= -\frac{1}{2} \|\boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{\Omega}_t - \boldsymbol{\Omega}_0)\boldsymbol{\Sigma}_0^{1/2}\|_F^2 - \frac{1}{2} \sum_{i=1}^d \int_0^{\rho_i} \frac{(\rho_i - s)^2}{(1 - s)^3} ds.
\end{aligned}$$

Finally, we plug-in the expression of $\boldsymbol{\Omega}_t - \boldsymbol{\Omega}_0 = \frac{t\boldsymbol{\Phi}}{\sqrt{n}\|\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\Phi}\boldsymbol{\Sigma}_0^{1/2}\|_F}$ into the last line of the last display to complete the proof. \square

Chapter 4

Bayesian inference for generalized extreme value distribution with Gaussian copula dependence

4.1 Introduction

Generalized extremely value models have been used extensively to estimate and predict extreme events in many applications, including environment, engineering, economics and finance studies. Both frequentist and Bayesian approaches have been developed to estimate those models (Coles, 2001; Coles and Powell, 1996). Recently, dependent GEV models (for short, dGEV models) have attracted much attention. Those models are useful in applications when extreme events are time-correlated. For example, higher temperature in one day may be followed by a higher temperature in another day.

A GEV distribution is shown as follows:

$$\mathbb{P}(Y \leq y) = F(y) = \exp \left\{ - \left(1 + \xi \frac{y - \mu}{\psi} \right)_+^{-\frac{1}{\xi}} \right\}, \quad (4.1.1)$$

where μ , ψ and ξ are the location, scale and shape parameters respectively. To extend the GEV model to a dGEV model, it is common to introduce dependence for some of its parameters. Previous studies have focused on letting μ follow an AR(1) process with other parameters time invariant (Huerta and Sansó, 2007); and letting μ , ψ and ξ be time-dependent and follow ARMA-type process (Wei and Huerta, 2016). However, since

we consider the dependency of extreme events, a more general way to introduce the time-dependent structure is to insert a copula. This idea is originally proposed by Nakajima et al. (2011, 2017). To be explicit, their model is constructed as follows.

If we define

$$\alpha_t \equiv \log \left\{ \left(1 + \xi \frac{Y_t - \mu}{\psi} \right)_+^{\frac{1}{\xi}} \right\}, \quad (4.1.2)$$

then α_t follows the Gumbel distribution. Solving (4.1.2) for Y_t , we obtain the following expression,

$$Y_t = \mu + \psi \frac{\exp(\xi \alpha_t) - 1}{\xi}, \quad (4.1.3)$$

and if α_t is a hidden AR(1) process,

$$\alpha_{t+1} = \phi \alpha_t + \eta_t, \quad \eta_t \sim \text{Gumbel}, \quad (4.1.4)$$

then the model becomes a nonlinear non-Gaussian state space model. To estimate the parameters in the model, they first show that the model can be approximated by a linear Gaussian state space model, by finding a linear approximation for nonlinear observation equations and using mixed normal distributions to approximate the Gumbel errors in the state equation. However, from (4.1.4) the marginal distribution of α_t does not follow a Gumbel distribution, because the sum of two Gumbel-distributed random variables in general does not follow a Gumbel distribution. We thus propose a method to allow α_t to follow exactly the Gumbel distribution. The model has Gumbel marginal distributions linked together with a Gaussian copula with AR(1) dependence. By this construction, the marginal distribution of α_t is exactly from a Gumbel distribution. The details are shown in Section 4.2.

Because the number of variables in the dGEV model is large, we use Bayesian analysis to draw samples from the joint distributions. The Markov chain Monte Carlo (MCMC) algorithm allows us to sample from the conditional distribution for each individual parameter, and thus we only need to take care of the conditional distribution of each parameter at one time instead of the joint distribution for all parameters. A drawback of using the MCMC algorithm is that the mixing can be slow due to the high dimensional state vector in the model. As a result, many iterations are needed and methods to reduce the correlation between chains such as marginalization, permutation and trimming (Dyk and Park, 2008; Liu et al., 1994; Liu, 1994) must be considered.

In Bayesian analysis, a common technique is to use Kalman filtering and backward smoothing (KFBS) (West and Harrison, 1997) for sampling draws from linear Gaussian state space models. For nonlinear non-Gaussian state space models, the KFBS cannot be applied directly. One way of sampling those models is to find a linear approximation for the nonlinear equations (Shephard and Pitt, 1997); common techniques include using a first-order Taylor series expansion based on the observation equations, or a second-order Taylor expansion based on the log-likelihood functions, and finding a mixed normal approximation for the non-Gaussian errors. The sampling approach is to apply the KFBS algorithm on the approximated linear Gaussian state space model, and then use a Metropolis-Hastings (M-H) algorithm to accept or reject the draws. However, when the model is highly nonlinear and non-Gaussian, these approximation techniques may perform very poorly; as a result, the M-H step will reject most of the draws, thus increasing the computation time and reducing the mixing rate of the chain. There are many extensions for the KFBS for adapting to a nonlinear state space model including the extended Kalman filter, the robust extended Kalman filter (Einicke and White, 1999) and using mixed normal densities to approximate non-Gaussian densities (Niemi and West, 2010).

An alternative way to make inferences in a nonlinear state space model is to use a sequential Monte Carlo (SMC) algorithm. This approach is as follows: it first generates particles from a chosen proposal density; and then calculates the weights between the proposal density and the true density. The generated particles and their weights give a discrete approximation to the true density. However, one issue with this method is the degeneration of the weights, thus a resampling step is needed when calculating the weights. Recently, much effort has been put into incorporating SMC within MCMC; some of the methods include particle Metropolis-Hastings and the particle Gibbs sampler (Andrieu et al., 2010), as using the MCMC approach is more convenient when dealing with a model with many variables.

The particle Gibbs sampler is the method we shall use in this study; this method uses a collapsed Gibbs sampler (Liu, 1994), which allows the joint distribution to be invariant in MCMC sampling. The essential idea is to combine a conditional SMC (cSMC) (Liu, 2001) with the Gibbs sampler. The cSMC method needs to keep a path known as an ancestor lineage to survive in each resampling step. However, this algorithm has two weaknesses: a poor mixing rate and particle degeneration (Doucet and Johansen, 2011; Chopin and Singh, 2015). An approach to address these weaknesses is to do backward smoothing,

which known as forward sampling and backward smoothing (FFBS) algorithm (Whiteley, 2010; Lindsten and Schön, 2013; Kantas et al., 2015). A drawback for the FFBS algorithm is that adding the backward step increases the computation time. A more recent method which is known as Particle Gibbs with ancestor sampling (PGAS) (Lindsten et al., 2014) is proposed to address the computation issues of FFBS. The PGAS approach is as follows: for each forward sampling step, instead of fixing the ancestor lineage, the algorithm resamples the ancestor lineage. In this case, some particles with small weights in the ancestor lineage will drop out.

In the rest of the chapter, we will introduce our novel dGEV model, which incorporates the dependence between extreme events through a dependent Gumbel distribution; the dependent Gumbel distribution is constructed through a normal copula. Then in the final model the observed variable is sampled from an exact GEV distribution. Our model can be expressed as a nonlinear Gaussian state space model; we use PGAS to sample the hidden process in the nonlinear state space model. We also show that our model may be extended to incorporate seasonal components, which are needed for some datasets; we name this model as a seasonal dGEV model.

In the simulation study, we show our estimated posterior medians for parameters are close to the true values, and the true values are contained in the 95% credible intervals. Because the latent variables marginally follow a standard normal distribution, we could use standard normal distribution tails to characterize the extremeness of the values.

We fit the model to two real datasets: one is an annual maximum water flow dataset and the other is a weekly minimum return for the S&P 500 index.

The remaining sections are arranged as follows: Section 4.2 introduces our dGEV model; Section 4.3 describes a Bayesian computation steps; Section 4.4 adds seasonality on the dGEV model; Section 4.5 illustrates the dGEV and seasonal dGEV models using simulated data; Section 4.6 conducts a real data study for a water flow dataset and a financial dataset; Section 4.7 concludes.

4.2 The dGEV model.

Our dGEV model is constructed as follows: Suppose that β_t follows the standard normal distribution, so that $\Phi(\beta_t)$ follows the uniform $(0, 1)$ distribution. Let $G(\alpha)$ be the CDF

of the Gumbel distribution: $G(\alpha) = \exp(-\exp(-\alpha))$. If

$$\alpha_t = G^{-1}(\Phi(\beta_t)) = -\log(-\log(\beta_t)),$$

then α_t follows the Gumbel distribution.

Now suppose that β_t follows an AR(1) process:

$$\beta_{t+1} = \phi\beta_t + \eta_t, \tag{4.2.1}$$

for $t = 1, \dots, T - 1$, where $\eta_t \sim \mathcal{N}(0, 1 - \phi^2)$ and $\beta_1 \sim \mathcal{N}(0, 1)$, and let

$$\begin{aligned} Y_t &= \mu + \psi \frac{\exp(\xi\alpha_t) - 1}{\xi} + \epsilon_t, \\ &= \mu + \psi \frac{(-\log(\Phi(\beta_t)))^{-\xi} - 1}{\xi} + \epsilon_t, \end{aligned} \tag{4.2.2}$$

for $t = 1, \dots, T$, where $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ and $\beta_1 \sim \mathcal{N}(0, 1)$. Adding the normal error ϵ_t in the observation equation makes the model have a nonlinear state space presentation. The variance of the error should be small and negligible.

The parameters in our model are μ , ψ , ξ , θ , σ^2 and the latent variables $\beta_{1:T}$. In Bayesian analysis, we adopt a similar priors for these parameters to those suggested by [Coles and Tawn \(1996\)](#); [Chavez-Demoulin and Davison \(2012\)](#). That is $\mu \sim \mathcal{N}(\nu_\mu, \sigma_\mu^2)$, $\psi \sim \text{Gamma}(a_\psi, b_\psi)$, $\xi \sim \mathcal{N}(\nu_\xi, \sigma_\xi^2)$. The prior for $\phi \sim U(-1, 1)$, where U stands for a uniform distribution. Last, the prior for $\sigma^{-2} \sim \text{InvGamma}(a, b)$, which we later abbreviate $\text{IG}(a, b)$.

4.3 Posterior computation

In this section, we describe the MCMC algorithm that will be used for sampling draws of the parameters in the dGEV model. Because of the number of parameters is large, the MCMC chains will converge very slowly. To ameliorate this problem, we consider using trimming and blocking techniques ([Dyk and Park, 2008](#)).

In the following, to simplify the notation, let $\boldsymbol{\theta}$ to be any parameter(s) in the model and $-\boldsymbol{\theta}$ to be the parameters in the model except $\boldsymbol{\theta}$. The outline of the MCMC algorithm is designed as follows: we sample $\pi_{-\boldsymbol{\theta}}(\boldsymbol{\vartheta}|\mathbf{Y})$ from its posterior distribution by grouping μ ,

ψ , and ξ together and treating them as a block; and then sample $\pi_{-\sigma^2}(\sigma^2|\mathbf{Y})$, $\pi_{-\phi}(\phi|\mathbf{Y})$ from the conditional posterior distribution using Gibbs and Metropolis-Hastings algorithm; last, sample $\pi_{-\beta}(\beta|\mathbf{Y})$ using the particle Gibbs with ancestor sampler (PGAS) algorithm, which will be introduced later.

4.3.1 Sampling μ , ψ , ξ

We group μ , ψ , ξ together and sample from their joint distribution to reduce the correlation between MCMC chains (Nakajima et al., 2011). Thus we consider the joint conditional posterior distribution for $\boldsymbol{\vartheta}$:

$$\pi_{-\boldsymbol{\vartheta}}(\boldsymbol{\vartheta}|\mathbf{Y}) = \prod_{t=1}^T \pi_{-\boldsymbol{\vartheta}}(\boldsymbol{\vartheta}|Y_t)\pi(\mu)\pi(\psi)\pi(\xi). \quad (4.3.1)$$

Although the priors for $\boldsymbol{\mu}$ and $\boldsymbol{\psi}$ are conjugate with their likelihoods, the prior for $\boldsymbol{\xi}$ does not. When we group them as a trivariate parameter, the joint likelihood is not conjugacy with the priors, thus we use an Metropolis-Hastings (M-H) algorithm to obtain draws from the posterior. The M-H algorithm requires to find a proposal density, which is derived from the second-order Taylor expansion of $\log \pi_{-\boldsymbol{\vartheta}}(\boldsymbol{\vartheta}|\mathbf{Y})$. To be more specific, the proposal density is $\boldsymbol{\vartheta}^* \sim \text{MVN}(\boldsymbol{\nu}^*, \boldsymbol{\Sigma}^*)$, where

$$\begin{aligned} \boldsymbol{\Sigma}^{*-1} &= -\frac{\partial^2 \log \pi_{-\boldsymbol{\vartheta}}(\boldsymbol{\vartheta}|\mathbf{Y})}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}^T} \Big|_{\boldsymbol{\vartheta}=\hat{\boldsymbol{\vartheta}}}, \\ \boldsymbol{\nu}^* &= \hat{\boldsymbol{\vartheta}} + \boldsymbol{\Sigma}^* \cdot \frac{\partial \log \pi_{-\boldsymbol{\vartheta}}(\boldsymbol{\vartheta}|\mathbf{Y})}{\partial \boldsymbol{\vartheta}} \Big|_{\boldsymbol{\vartheta}=\hat{\boldsymbol{\vartheta}}}, \end{aligned}$$

and $\hat{\boldsymbol{\vartheta}}$ is a value that maximizes the posterior of $\pi_{-\boldsymbol{\vartheta}}(\boldsymbol{\vartheta}|\mathbf{Y})$; this value can be calculated by using `optim` in R. After sampling a draw $\boldsymbol{\vartheta}^*$ from this proposal density, we use the M-H algorithm to accept or reject $\boldsymbol{\vartheta}^*$ with ratio

$$\alpha(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*) = \min \left\{ 1, \frac{\pi_{-\boldsymbol{\vartheta}}(\boldsymbol{\vartheta}^*|\mathbf{Y})f(\boldsymbol{\vartheta}|\boldsymbol{\nu}^*, \boldsymbol{\Sigma}^*)}{\pi_{-\boldsymbol{\vartheta}}(\boldsymbol{\vartheta}|\mathbf{Y})f(\boldsymbol{\vartheta}^*|\boldsymbol{\nu}^*, \boldsymbol{\Sigma}^*)} \right\}.$$

4.3.2 Sampling σ^2

To sample σ^2 , we choose a conjugate prior inverse gamma prior as follows,

$$\sigma^2 \sim \text{IG}(a, b).$$

Since ϵ_t should be small, we choose a to be large and b to be small so that the prior will have a smaller variance. Then the posterior of σ^2 can be written as follows:

$$\pi_{-\sigma^2}(\sigma^2|\mathbf{Y}) = \text{IG} \left(a + \frac{T}{2}, b + \frac{1}{2} \sum_{t=1}^T \left(y_t - \mu - \frac{\psi}{\xi} (-\log(\Phi(\beta_t))^{-\epsilon} - 1) \right)^2 \right). \quad (4.3.2)$$

4.3.3 Sampling ϕ

The conditional posterior of ϕ can be expressed as follows:

$$\pi_{-\phi}(\phi|\mathbf{Y}) = (2\pi)^{-T/2} (1 - \phi^2)^{-T/2} \cdot \exp \left\{ - \frac{\sum_{t=2}^T (\beta_t - \phi\beta_{t-1})^2}{2(1 - \phi^2)} \right\} \cdot \pi(\phi), \quad (4.3.3)$$

where $\pi(\phi)$ is the prior of ϕ . Because the posterior density is not a well known distribution, similar to dealing with parameter $\boldsymbol{\vartheta}$, we find a proposal density for $\pi_{-\phi}(\phi|\mathbf{Y})$ as $q_{-\phi}(\phi|\mathbf{Y}) \sim \mathcal{N}(\nu_\phi, \sigma_\phi^2)$ with

$$\begin{aligned} \sigma_\phi^{-2} &= - \left. \frac{\partial^2 \log \pi_{-\phi}(\phi|\mathbf{Y})}{\partial \phi^2} \right|_{\phi=\hat{\phi}}, \\ \nu_\phi &= \hat{\phi} + \sigma_\phi^2 \cdot \left. \frac{\partial \log \pi_{-\phi}(\phi|\mathbf{Y})}{\partial \phi} \right|_{\phi=\hat{\phi}}. \end{aligned}$$

After we draw $\phi^* \sim q_{-\phi}(\phi|\mathbf{Y}) \sim \mathcal{N}(\nu_\phi, \sigma_\phi^2)$, ϕ^* is accepted with probability

$$\alpha(\phi, \phi^*) = \min \left\{ 1, \frac{\pi_{-\phi}(\phi^*|\mathbf{Y})q_{-\phi}(\phi|\mathbf{Y})}{\pi_{-\phi}(\phi|\mathbf{Y})q_{-\phi}(\phi^*|\mathbf{Y})} \right\}.$$

4.3.4 Sampling $\beta_{1:T}$

The posterior distribution for $\pi_{-\beta}(\boldsymbol{\beta}|\mathbf{Y})$ can be expressed as the following,

$$\pi_{-\beta}(\boldsymbol{\beta}|\mathbf{Y}) = \prod_{t=1}^T f(Y_t|\beta_t, \boldsymbol{\vartheta}, \sigma^2) \prod_{t=2}^T f(\beta_t|\beta_{t-1}, \phi)\pi(\beta_1),$$

with

$$\begin{aligned}
f(Y_t|\beta_t, \boldsymbol{\vartheta}, \sigma^2) &= \left(\frac{1}{2\pi\sigma^2}\right)^{-\frac{T}{2}} \exp\left\{-\frac{Y_t - \mu - \frac{\psi}{\xi}((-\log(\Phi(\beta_t)))^{-\xi} - 1)}{2\sigma^2}\right\}, \\
f(\beta_t|\beta_{t-1}, \phi) &= \left(\frac{1}{2\pi(1-\phi^2)}\right)^{-\frac{T-1}{2}} \exp\left\{-\frac{\sum_{t=2}^T (\beta_t - \phi\beta_{t-1})^2}{2(1-\phi^2)}\right\}, \\
\pi(\beta_1) &= -\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{\beta_1^2}{2}\right\}.
\end{aligned}$$

Due to the highly nonlinear expression of the likelihood $f(Y_t|\beta_t, \boldsymbol{\vartheta}, \sigma^2)$, the KFBS algorithm developed based on linear state space models cannot be applied directly. That is because the linear approximation using Taylor expansion has very poor performance. As a result, the acceptance probability is almost 0 for our model.

The particle Gibbs sampler proposed by [Andrieu et al. \(2010\)](#) provides an alternative to sampling the nonlinear state space model. It is based on the conditional sequential Monte Carlo (cSMC) ([Liu, 2001](#)) algorithm which is similar to the SMC algorithm except for that it assigns an ancestor lineage keep out of the resample stage. The particle Gibbs sampler algorithm treats the particles and weights generated from cSMC as a discrete distribution approximating the true density; a draw is randomly sampled from these particles according to their weights. Since the draws are obtained from an approximated distribution and not the real one, a pre-specified ancestor lineage or trajectory path is used to guide draws from the invariant unconditional distribution. However, this method is known to have degenerate issues and a poor mixing rate. A finer algorithm is to allow the pre-defined ancestor lineage to update during the forward sampling process, and drop some lineage with degenerating weights. The method is known as the PGAS algorithm. The method uses a trimming technique ([Dyk and Park, 2008](#)) to improve the mixing rate of a Gibbs sampler; also, the degenerate weight issue is ameliorated.

In the remainder of this section, we shall introduce this method. We summarize the notations to be used later for convenience: Let $q_{-\boldsymbol{\beta}}(\beta_t|Y_t)$ be a proposal density, $\beta_t^1, \dots, \beta_t^N$ be N particles drawn from the proposal density, k is a index in $\{1, \dots, N\}$ and $\mathbf{k} = (1, \dots, N)$; let $\boldsymbol{\beta}^k = (\beta_1^k, \dots, \beta_T^k)$ stand for the samples generated from the same trajectory path k ; let $\bar{\boldsymbol{\beta}}_t^k = (\beta_1^k, \dots, \beta_{t-1}^k, \beta_{t+1}^k, \dots, \beta_T^k)$, and $\bar{\mathbf{k}} = (1, \dots, k-1, k+1, \dots, N)$; let β'_t be the previous draws or the starting value for β_t and β_t^* be the current draws from $\beta_t^{\mathbf{k}}$; let $w_t(\beta_t^k)$ be the unnormalized weight for the k^{th} particle of β_t sampled from $q(\beta_t | -\boldsymbol{\beta}, Y_t, \beta_{t-1})$, and $W_t(\beta_t^k)$ be its normalized weight.

We consider two methods to choose a proposal density: The first method is to use Taylor expansion on the mean of the observation equation. For simplicity we write (4.2.2) as the following:

$$\begin{aligned} Y_t &= f_{\boldsymbol{\vartheta}}(\beta_t) + \epsilon_t, \\ \beta_t &= g_{\phi}(\beta_{t-1}) + \eta_t, \end{aligned}$$

with $f_{\boldsymbol{\vartheta}}(\beta_t) = \mu + \psi((-\log(\Phi(\beta_t)))^{-\xi} - 1)/\xi$ and $g_{\phi}(\beta_{t-1}) = \phi\beta_{t-1}$; then we approximate $f_{\boldsymbol{\vartheta}}(\beta_t)$ with

$$f_{\boldsymbol{\vartheta}}(\beta_t) = f_{\boldsymbol{\vartheta}}(g_{\phi}(\beta_{t-1}^{\mathbf{k}})) + f'_{\boldsymbol{\vartheta}}(g_{\phi}(\beta_{t-1}^{\mathbf{k}})) (\beta_t - g_{\phi}(\beta_{t-1}^{\mathbf{k}})) / c.$$

Because in a nonlinear equation, the second term in the Taylor expansion can be relatively large, the constant c is used to control the approximated density. We adjust the value c to keep the weights from degenerating. Then the proposal density then can be transformed into a Gaussian distribution,

$$q_{-\beta}(\beta_t | \mathbf{Y}) = \mathcal{N}(\nu_{\beta_t}, \sigma_{\beta_t}^2),$$

with

$$\begin{aligned} \sigma_{\beta_t}^{-2} &= \frac{1}{(1 - \phi^2)} + f'_{\boldsymbol{\vartheta}}(g_{\phi}(\beta_{t-1}^{\mathbf{k}}))^2 / \sigma^2, \\ \nu_{\beta_t} &= \frac{\sigma_{\beta_t}^2}{(1 - \phi^2)} g_{\phi}(\beta_{t-1}^{\mathbf{k}}) + \frac{f'_{\boldsymbol{\vartheta}}(g_{\phi}(\beta_{t-1}^{\mathbf{k}}))^2}{\sigma^2} (Y_t - f_{\boldsymbol{\vartheta}}(g_{\phi}(\beta_{t-1}^{\mathbf{k}})) \times g_{\phi}(\beta_{t-1}^{\mathbf{k}})). \end{aligned}$$

The second method is to ignore the ϵ_t in (4.2.2), then we write β_t as a function of y_t as follows,

$$\beta_t = \Phi^{-1} \left(\exp \left(- \left(1 + \xi \frac{Y_t - \mu}{\psi} \right)_+^{-1/\xi} \right) \right).$$

where Φ^{-1} denotes the inverse standard normal distribution function. Then we choose a proposal density to be the t -distribution with mean β_t and a small value of degrees of freedom, namely 5. This method does not apply to general state space models, but can perform very well in our model; this is because the ϵ_t in our model is small.

After specifying the proposal density, the PGAS algorithm is as follows,

- At current iteration i , given β' .
- For $t = 1$,
 - draw $\beta_1^{1:N-1} \sim q_{-\beta}(\beta_1|Y_1)$;
 - let $\beta_1^N = \beta'_1$;
 - calculate weight for $l = 1, \dots, N$ as

$$w_1^l := \frac{f(Y_1|\beta_1^l, \boldsymbol{\vartheta}, \sigma^2)\pi(\beta_1^l)}{q_{-\beta}(\beta_1^l|Y_1)},$$

$$W_1^l = \frac{w_1^l}{\sum_{l=1}^N w_1^l}.$$

- For $t = 2, \dots, T$,
 - resample β_{t-1}^k according to the weight W^k for $k = 1, \dots, N-1$, denote as $\tilde{\beta}_{t-1}^k$;
 - sample $\tilde{\beta}_{t-1}^N$ from β_{t-1}^k with the associated weights:

$$\frac{w_t^l f(\beta_t^l|\beta_{t-1}^l, \phi)}{\sum_{l=1}^N w_t^l f(\beta_t^l|\beta_{t-1}^l, \phi)};$$

- draw $\beta_t^{1,\dots,N-1} \sim q_{-\beta}(\beta_t|y_t)$;
- let $\beta_t^N = \beta'_t$;
- let $\beta_{1:t}^l := (\tilde{\beta}_{1:t-1}^l, \beta_t^l)$, for $l = 1, \dots, N$;
- calculate weight for $l = 1, \dots, N$ as

$$w_t^l = \frac{f(Y_t|\beta_t^l, \boldsymbol{\vartheta}, \sigma^2)p(\beta_t^{il}|\tilde{\beta}_{t-1}^{il}, \phi)}{q_{-\beta}(\beta_t^{il}|Y_t)},$$

$$W_t^l = \frac{w_t^l}{\sum_{l=1}^N w_t^l}.$$

- Draw k from $1, \dots, N$ with the corresponding probability $W_T^{1,\dots,N}$.
- Let $\beta = \beta^k$.
- Set $\beta' = \beta$, and go to the next iteration.

Then from the algorithm, we obtain draws for β .

4.3.5 Markov Chain Monte Carlo

In summary, the MCMC algorithm is conducted as follows:

1. initialize all the parameters values $\boldsymbol{\vartheta}$, $\boldsymbol{\beta}$, ϕ , σ^2 ;
2. sample $\boldsymbol{\vartheta} = (\mu, \psi, \xi)$ from (4.3.1), and using the M-H algorithm;
3. sample σ^2 from (4.3.2);
4. sample ϕ from (4.3.3), and using the M-H algorithm;
5. sample $\boldsymbol{\beta}$ using PGAS;
6. repeat steps 2–5 until sufficient draws have been obtained.

4.4 Seasonal dGEV model

Seasonality can be easily incorporated into our model by adding sinusoids to any of the location, shape or scale parameters. This method is convenient since it has flexibility in the choice of the number of sinusoids and it does not greatly increase the difficulty of making inferences about the model.

We start by adding two sinusoids to the location parameter; the model is shown as follows:

$$\begin{aligned}
 Y_t &= \mu + a_1 \cos(\omega t) + a_2 \sin(\omega t) \\
 &\quad + \psi \frac{(-\log(\Phi(\beta_t)))^{-\xi} - 1}{\xi} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2), \\
 \beta_{t+1} &= \phi \beta_t + \eta_t, \quad \eta_t \sim \mathcal{N}(0, 1 - \phi^2),
 \end{aligned} \tag{4.4.1}$$

where $\omega = 2\pi f$ is the angular frequency, f is the number of cycles that occur for each period of time. For a given dataset, f is typically known; for example, it has value $1/365.25$ for annual variability in a daily dataset, or $1/4$ for a annual variability in a quarterly dataset. In equation (4.4.1), $t = 1, \dots, T$, a_1 and a_2 are the coefficients for the two components of the sinusoid, we denote $\mathbf{a} = (a_1, a_2)$.

To estimate the parameters in (4.4.1), we use similar priors to those for the parameters described in Section 4.3. For the added parameters in the seasonal component in this

model, we treat ω as a known parameter. We observe that \mathbf{a} has a multivariate normal likelihood in the equation, we choose the prior for each element a_i to be $\mathcal{N}(0, \sigma_{a_i}^2)$. Then the posterior for \mathbf{a} is as follows,

$$\pi_{-\mathbf{a}}(\mathbf{a}|\mathbf{Y}) \sim \mathcal{N}(\boldsymbol{\nu}_{\mathbf{a}}, \boldsymbol{\Sigma}_{\mathbf{a}}), \quad (4.4.2)$$

with

$$\boldsymbol{\Sigma}_{\mathbf{a}}^{-1} = \sigma^{-2} \left(\sum_{t=1}^T \mathbf{p}_t \mathbf{p}_t' + \sigma^2 \boldsymbol{\Omega}^{-1} \right), \quad \boldsymbol{\nu}_{\mathbf{a}} = \boldsymbol{\Sigma}_{\mathbf{a}} \left(\sum_{t=1}^T \mathbf{p}_t \tilde{Y}_t \right) / \sigma^2,$$

where $\mathbf{p}_t = (\cos(\omega t), \sin(\omega t))^T$, $\boldsymbol{\Omega} = \text{diag}(\sigma_{a_1}^2, \sigma_{a_2}^2)$, $\tilde{Y}_t = Y_t - \mu - \psi \times ((-\log(\Phi(\beta_t)))^{-\xi} - 1)/\xi$. To conduct an MCMC algorithm for this model, we could adapt the algorithm described in Section 4.3.5, adding an extra step to sample \mathbf{a} .

Sometimes, a dataset may show seasonality in the scale parameter, such as in a weather dataset, the variation in the winter may be larger than in the summer. Our model has the flexibility to deal with a dataset like this by adding two similar sinusoids to ψ , i.e. $\psi + a_3 \cos(\omega t) + a_4 \sin(\omega t)$. It is possible to add sinusoids to the ξ parameter or adding more sinusoidal components to each parameter; however, more parameters in the model will reduce the mixing rate between chains in the MCMC algorithm; as we will shown in the simulation study, the dGEV model already exhibit a large autocorrelation between samples in the chain. For simplicity, we add seasonality only to the location parameter. In the later context, we will refer this model to be the seasonal dGEV model; however, it should be kept in mind that seasonality could be added to any of the parameters.

4.5 Illustrative simulation study

In this section, we illustrate the dGEV and seasonal dGEV models (4.4.1) using simulated data. To generate datasets, we start by generating time-varying parameters $\beta_{1:T}$, and then use $\beta_{1:T}$ to generate observations $Y_{1:T}$. To be more specific, for the dGEV model, we generate 1000 observations. The choice of the sample size is arbitrary; however, more observations are required for a model with more parameters, for example in the case where we add seasonality to both the location and scale parameters. In the simulations, we choose $\phi = 0.8$, $\mu = 0.5$, $\psi = 0.3$, $\xi = 0.05$, and $\sigma = 0.1$. We first simulate $\beta_1 \sim \mathcal{N}(0, 1)$,

Table 4.1: Posterior medians and 95% credible intervals (C.I.s) for parameters μ , ψ , ξ , ϕ , σ .

Parameter	True value	Posterior median	95% C.I.	Inefficiency factor
μ	0.5	0.492	[0.4332, 0.5642]	222.26
ψ	0.3	0.315	[0.2818, 0.3599]	233.84
ξ	0.05	0.101	[0.035, 0.1843]	186.26
ϕ	0.8	0.823	[0.7781, 0.8636]	173.69
σ	0.1	0.092	[0.0753, 0.1109]	61.61

then β_2, \dots, β_T can be generated according to the AR(1) process in (4.2.1). Based on the generated values of $\beta_{1:T}$, we then generate dataset $Y_{1:T}$ from (4.2.2). For the seasonal dGEV model 4.4.1, we choose $a_1 = 1$, $a_2 = 2$, and suppose we have an annual dataset, $f = 1/365.25$ and $\omega = 2\pi/365.25$.

After generating observations, we chose priors to conduct the Bayesian analysis. The priors for those parameters are as follows: $\mu \sim \mathcal{N}(0, 2^2)$, $\psi \sim \text{Gamma}(2, 2)$, $\xi \sim \mathcal{N}(0, 2^2)$, $\phi \sim U(-1, 1)$, and $\sigma^2 \sim \text{IG}(1, 0.01)$. We choose the total iteration for MCMC to be 20,000, when running the algorithm, with the first 5,000 as burnin. In the PGAS step, we choose the number of particles to be 1000. Lindsten et al. (2014) shows that when this number is chosen, the PGAS algorithm has update frequency very close to 0.999, the ideal rate.

We then fit the model to the simulated datasets. The first row in Figure 4.1 plots the MCMC draws for parameters except for β s for the dGEV model. The draws converge quickly for each parameter; the red lines that indicate the true values are all contained in the draws. Table 4.1 gives summary statistics for the posterior medians and credible intervals. The results show that the true values lie within the 95% credible intervals, and the posterior medians are close to their true values, except for the parameter ξ . The issue with ξ , the shape parameter in the GEV distribution, is that the proposed distribution is a less accurate approximation to the true posterior distribution than in the case of the location and scale parameters; thus its the posterior median has larger bias compared with the posterior medians of μ and ψ .

Table 4.1 also provides the inefficiency factor for each parameter; this statistic gives a diagnostic about how well the chains in MCMC mixed. It is calculated as $1 + 2 \sum_{s=1}^M (1 - s/M)\rho_s$ (see Chib, 2001), where ρ_s is the estimated autocorrelation at lag s , and M is

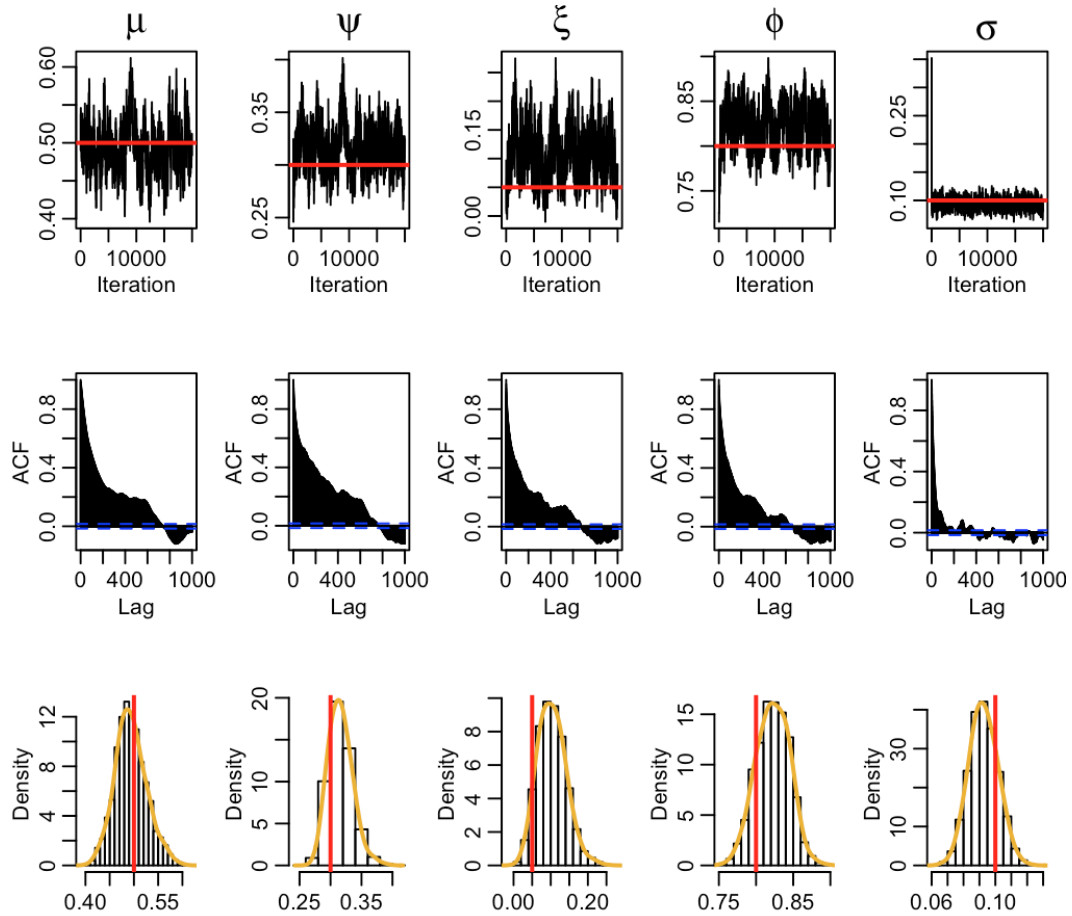


Figure 4.1: Plot of MCMC draws (1st row), autocorrelation function (ACF) (2nd row) and histogram with density (3rd row) for μ , ψ , ξ , ϕ , σ . The red line indicates the true values.

the batch size, which we take as 500. An inefficiency factor of m means that the effective number of draws is the total number of iterations, after deleting burn-in, divided by m . The results in the table show that those inefficiency factors are large, which corresponds to the slow decay in the ACF plots in Figure 4.1.

Figure 4.2 plots the posterior median and 95% credible intervals for $\beta_1, \dots, \beta_{1000}$. Recall that the time-varying parameters are in a standard normal copula; thus each β_t has a marginal standard normal distribution. Examining the posterior medians of β s shows that the majority of them fall in the 95% standard normal interval. The credible intervals are very small around each posterior median, as a result of the small value of

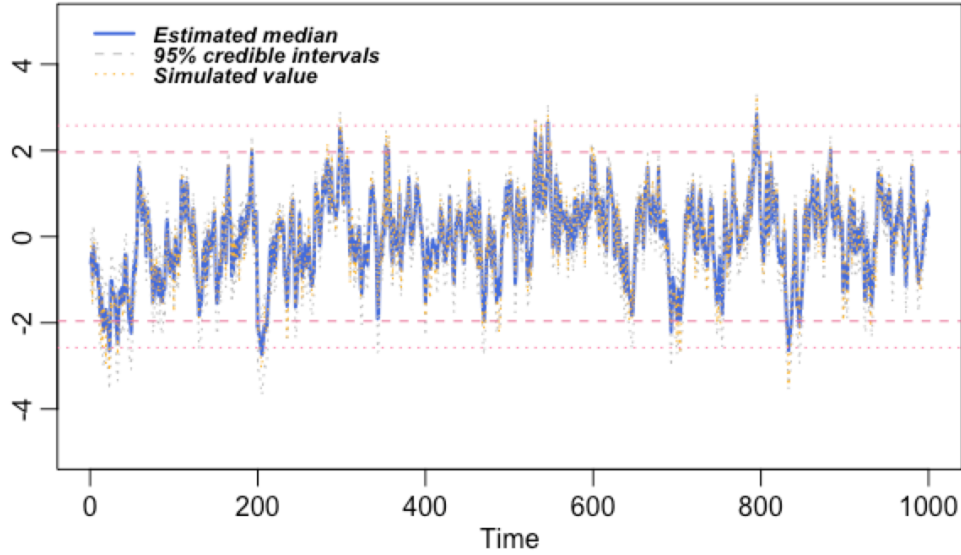


Figure 4.2: Plot of posterior median (blue line) and 95% credible intervals (grey dash lines) for $\beta_1, \dots, \beta_{1000}$; the yellow dashed lines indicates our simulated value for β_s ; the four dash red lines from up to down represent the 0.995, 0.975, 0.025, 0.005 quantiles of the standard normal distribution.

Table 4.2: Posterior median and 95% credible intervals (C.I.s) for parameters $\mu, \psi, \xi, \phi, a_1, a_2, \sigma$.

Parameter	True value	Posterior median	95% C.I.	Inefficiency factor
μ	0.5	0.472	[0.4172, 0.5485]	219.05
ψ	0.3	0.300	[0.2681, 0.3396]	214.55
ξ	0.05	0.014	[-0.0432, 0.08136]	92.22
ϕ	0.8	0.797	[0.7486, 0.8380]	103.36
a_1	0.1	0.984	[0.9212, 1.0409]	115.30
a_2	0.1	1.000	[0.9453, 1.0641]	129.67
σ	0.1	0.092	[0.0775, 0.1064]	101.07

measurement error ϵ_t .

Figure 4.3 shows that the simulation results of selected parameter draws based on the seasonal dGEV model. The estimated posterior median and 95% credible intervals of those parameters are summarized in Table 4.2. The 95% credible intervals all covered the true value, and posterior medians are close to their true value. We also found the inefficiency factor for parameters in the seasonal dGEV model to be on average higher

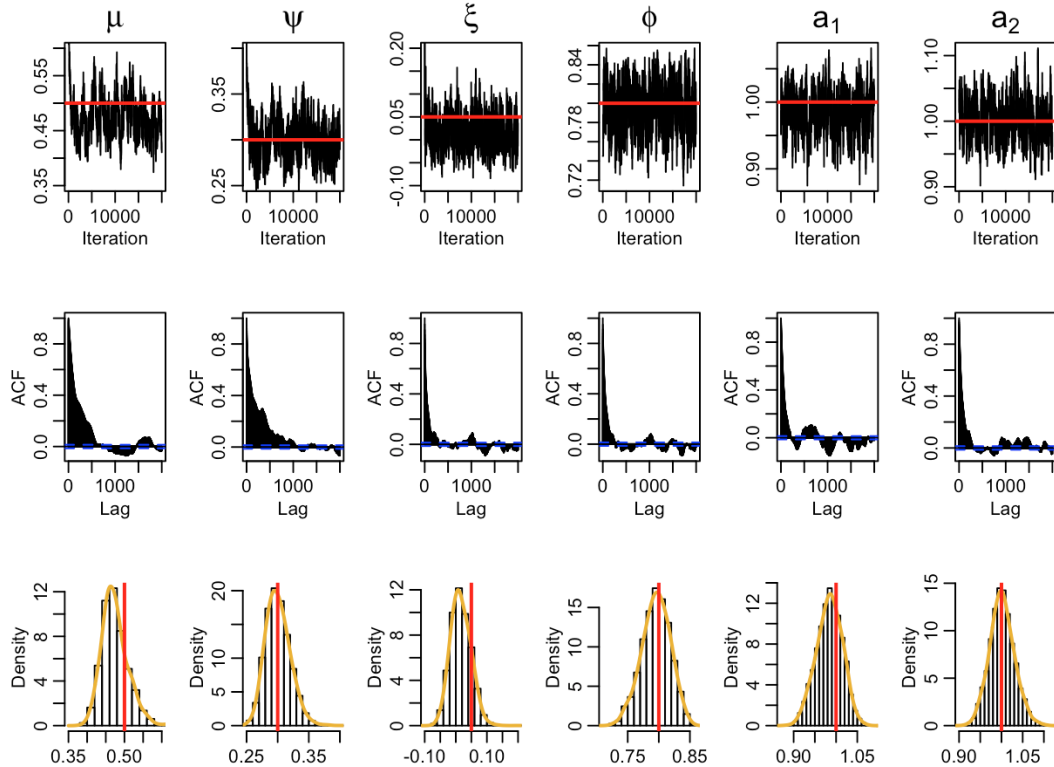


Figure 4.3: Plot of MCMC draws (1st row), autocorrelation function (ACF) (2nd row) and histogram with density (3rd row) for μ , ψ , ξ , ϕ , a_1 , a_2 . The red line indicates the true values.

than the non-seasonal model; this is because having more parameters in the model increases the correlations between chains. Figure 4.4 gives a plot of posterior medians and their credible intervals for $\beta_1, \dots, \beta_{100}$. The figures show that the majority of draws fall in the 95% region of the standard normal distribution. Also we plot the simulated true data of $\beta_1, \dots, \beta_{1000}$ in Figure 4.2; all the true values are covered by their estimated 95% credible intervals.

4.6 Real data study

We conduct a study of two real datasets in this section. The first dataset is an annual maximum water flow dataset and the second is a minimum log-return dataset for the S&P 500 stock index. We fit the dGEV model to both datasets, and we fit the seasonal

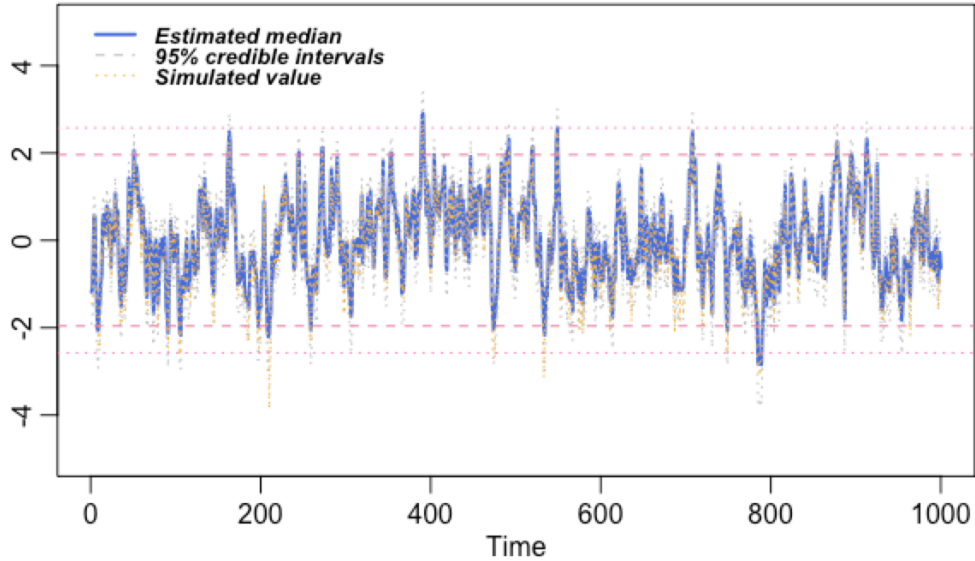


Figure 4.4: Plot of posterior median (blue line) and 95% credible intervals (grey dash lines) for $\beta_1, \dots, \beta_{1000}$; the yellow dashed lines indicates our simulated value for β_s ; the four dash red lines from up to down represent the 0.995, 0.975, 0.025, .005 quantiles of the standard normal distribution.

dGEV model to the S&P 500. The plots of MCMC outputs (including MCMC draws, autocorrelation functions (ACF) and histograms along with densities for parameters) obtained from the Bayesian inference for the two datasets are placed in Appendix C.

4.6.1 Water flow data

A water flow dataset is collected from French Broad River at Asheville in North Carolina. The datasets contains annual maximum water flow level from 1941 to 2009. A plot of this dataset is shown in Figure 4.5, the plot shows that there are two spikes in the year 1964 and 2004 which may consider to be unusual years. Our goal is to investigate how extreme those two values could be by fitting into the dGEV model.

We run the chain for 20,000 draws and treat the first 5,000 as burnin, the number of particles is chosen to be 1000. Before running the analysis, we standardize the dataset to avoid computation problem. The plot of draws and summary statistics is presented in Figure C.1 and Table 4.3. The ACF plots and inefficient factors for each parameters shows the chain mixed well.

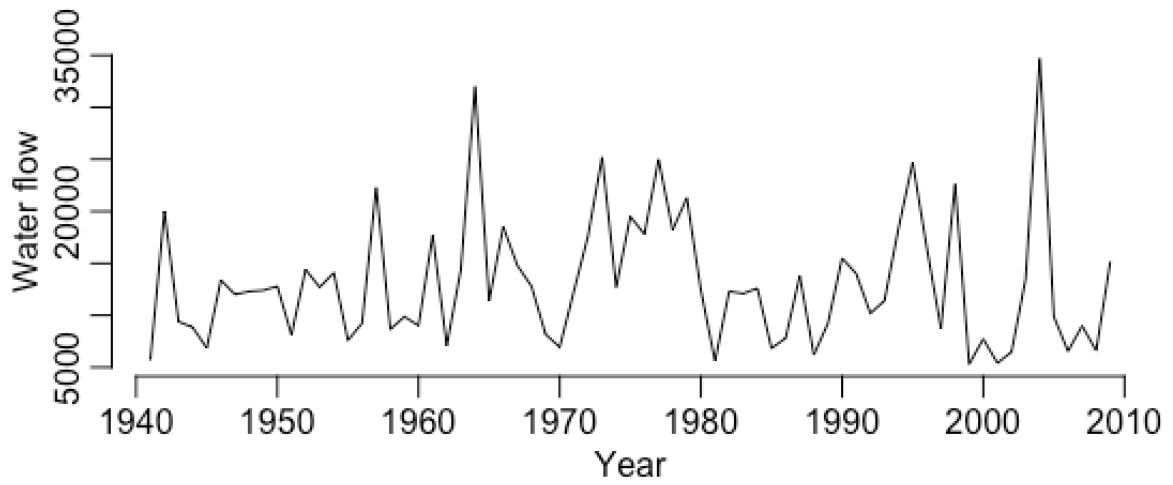


Figure 4.5: Annual maximum water flow of French Broad River at Asheville, North Carolina.

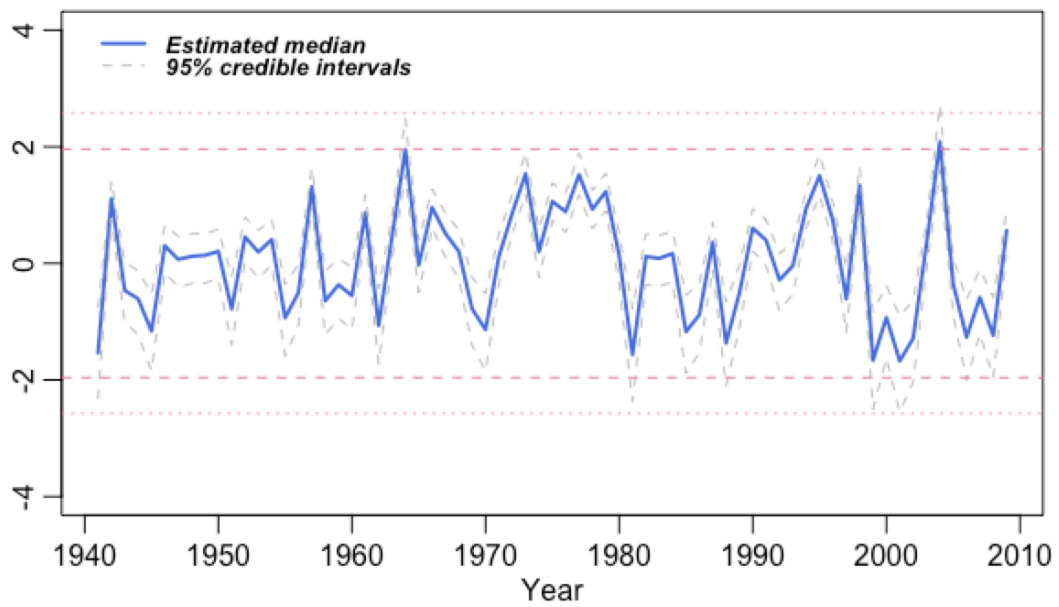


Figure 4.6: Plot of posterior medians (blue line) and their 95% credible intervals (grey dash lines) for β_t s by fitting dGEV into water flow dataset. The four dash red lines from up to down represent the 0.995, 0.975, 0.025, 0.005 quantiles of the standard normal distribution.

Table 4.3: Posterior medians, 95% credible intervals and inefficient factors for parameters μ , ψ , ξ , ϕ , σ .

Parameter	Posterior median	95% Credible interval	Inefficiency factor
μ	10144	[8566.1, 118645.0]	170.58
ψ	4351	[3148.0, 6303.0]	223.22
ξ	0.22	[-0.031, 0.664]	223.22
ϕ	0.21	[-0.075, 0.523]	74.16
σ	10.27	[6.468, 19.664]	34.59

From Table 4.3, the time varying parameter ϕ has posterior median 0.021, the 95% credible intervals cover with 0, the result suggests that the dependency between extreme values are low or possibly does not exist.

In Figure 4.6, we plot the posterior median for β_t s. Since each β_t s has marginal distribution from standard normal, it is convenient to compare these values with standard normal quantiles. If the values is larger than 2.326 in absolute value, then it is beyond 99% quantiles of normal distribution, which may seems unusual. From the plot, both of the two events in 1964 and 2004 we are interested in lies on the boundary of 95% quantile, and their credible intervals are below 99% quantile, which may shows the two events are not extreme as it looks like from the data plot.

4.6.2 S&P 500 datasets

Another dataset is from S&P 500 minimum weekly log-return, the data is collected from September, 25, 2006 to September, 12, 2016. A plot of this data is shown in Figure 4.7. The data is plotted in a -1 scale, so the largest value in the plots stands for the minimum return. Our interests is focus on finding the most unusual happened events in the dataset, especially during 2008 to 2010. Sometimes a time series shows seasonality, although the seasonality is not so obviously to appear in this model, we are interested to fit model to find out if seasonal effects exists. We fit the dataset in both of the dGEV model and the seasonal dGEV model, the plot of draws are shown in Figure C.2 and Figure C.3. The posterior medians with their 95% credible intervals for parameters in the model are summarized in Table 4.4. The inefficient factors for those parameters are also provided. In the MCMC algorithm, we run 20,000 iterations and treat the 5000 draws as burnin;

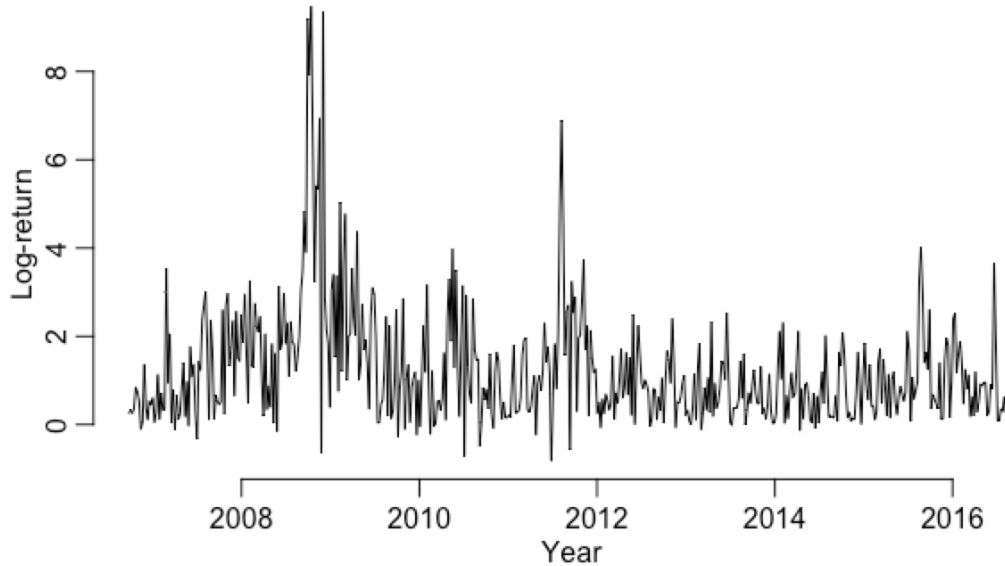
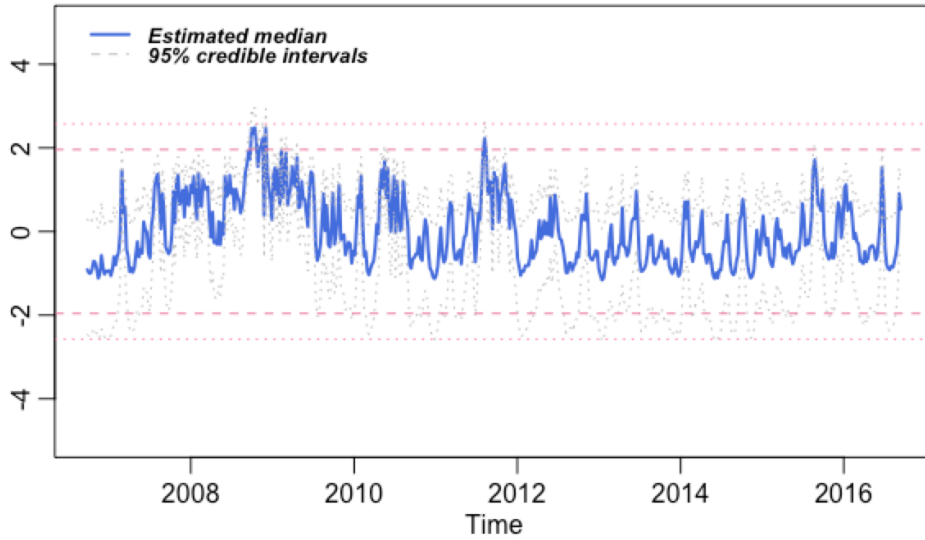


Figure 4.7: Weekly minimum S&P 500 log-return dataset, adjust the dataset to be -1 .

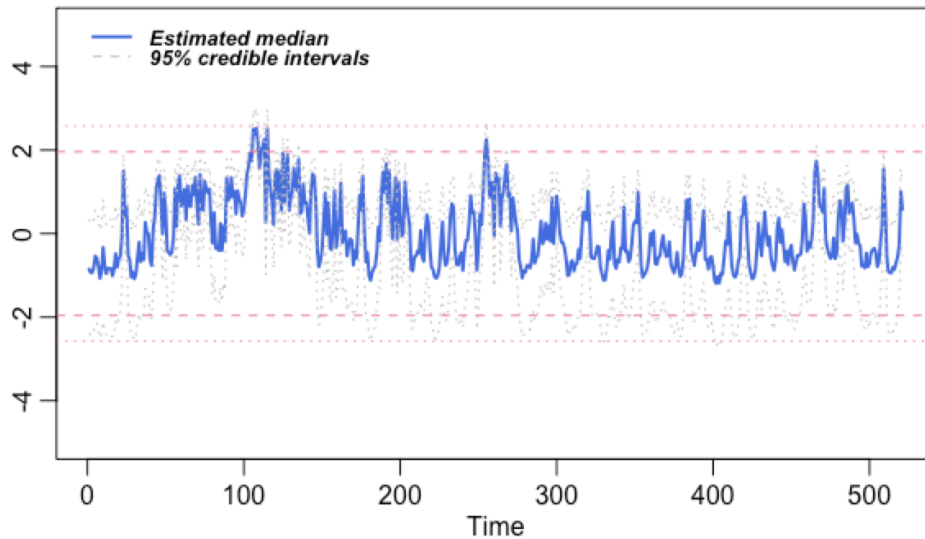
we choose the number of particles to be 1000 for both of the two models. For the seasonal dGEV model, we choose $f = 7/365.25$.

From the result shown in Table 4.4, the estimation for location, scale and shape parameters for both dGEV and seasonal dGEV model are close, the inefficient factors for seasonal dGEV model are larger than the dGEV model. In the output of the estimation of seasonality parameters a_1, a_2 , their value are very close to 0, and their 95% credible intervals contains 0. This suggests seasonality may not exist in this model.

Figure 4.8 gives plot of estimated β_t s posterior medians and their 95% credible intervals for fitting both models. The two models give a very similar estimate for those β_t s. Both plots indicate the period during the 2008 crisis is much more unusual than other periods. The highest point has its posterior median close to the 99% quantile of the standard normal, and its credible interval bounds contain the area which exceeds the 99% quantile.



(a)



(b)

Figure 4.8: (a) Plot of posterior medians (blue line) and their 95% credible intervals (grey dash lines) for β_{tS} by fitting the dGEV model into S&P 500 dataset. (b) Plot of posterior medians (blue line) and their 95% credible intervals (grey dash lines) for β_{tS} by fitting the seasonal dGEV model into S&P 500 dataset. The four dash red lines from up to down represent the 0.995, 0.975, 0.025, 0.005 quantiles of the standard normal distribution. dataset.

Table 4.4: Posterior medians with their 95% credible intervals and inefficient factor for parameters $\mu, \psi, \xi, \phi, \sigma$ in the dGEV model and parameters $\mu, \psi, \xi, \phi, \sigma, a_1, a_2$ in the seasonal dGEV model.

Parameter estimates for dGEV model			
Parameter	Posterior median	95% Credible interval	Inefficient factor
μ	0.696	[0.5847, 0.8307]	37.83
ψ	0.348	[0.2171, 0.4645]	80.29
ξ	0.449	[0.2638, 0.7264]	179.25
ϕ	0.713	[0.5347, 0.8555]	145.46
σ	0.967	[0.7425, 1.1817]	96.37

Parameter estimates for seasonal dGEV model			
Parameter	Posterior median	95% Credible interval	Inefficient factor
μ	0.683	[0.5691, 0.8101]	39.16
ψ	0.476	[0.3132, 0.6440]	91.94
ξ	0.425	[0.2294, 0.6912]	220.46
ϕ	0.680	[0.4547, 0.8277]	201.65
a_1	-0.034	[-0.1712, 0.1027]	7.75
a_2	-0.042	[-0.1794, 0.0890]	7.76
σ	0.531	[0.3729, 0.6480]	154.70

4.7 Conclusion and discussion

In this chapter, we proposed a novel dependent GEV model. The model can be expressed as nonlinear Gaussian state space model. Due to the observation equation is highly nonlinear, we use the PGAS algorithm to sample time-varying parameters. This algorithm can be incorporated into an MCMC algorithm. The simulation study based on the MCMC algorithm shows that the sampled parameter distribution could cover the true value and the posterior median is close to the truth. The shape parameter ξ is turned to be harder to estimate compare to the location and scale parameters. We also showed that our dGEV model can easily to incorporate seasonal components. Seasonality can be added to both the location, scale and shape parameters.

We did two case studies using the model: one is the annual maximum water flow dataset and another is the weekly minimum return of S&P 500. The estimated dependent parameter in the water flow dataset contains value 0 which suggests the correlation

between extreme values are low. The two unusual events in 1964 and 2004 does not turn out to be very unusual after plotting the corresponding estimated value of β . In the S&P 500 dataset, the correlation is very high. By fitting this dataset into the seasonal dGEV model which suggests there does not exist any seasonality effect. The draws of β_t s shows that extreme values during the 2008 economics crisis is very unusual compares to other values.

Our model turns out to require large computation time when the sample size is larger. However, the sample size usually is 10,000 when analyzing extremely daily values for temperature. In order to analyze those datasets, an improvement of current algorithm is need. On the other side, for datasets contains seasonality on the scale or shape parameters, the number of parameters will increase and cause the inefficient factors become larger. Thus a method to redesign the MCMC algorithm is needed as well. Both of these issues need to be address in the future study but is beyond the scope of this study.

REFERENCES

- Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies* 72, 1–19. 11
- Abadie, A. and J. Gardeazabal (2003). The economic costs of conflict: A case study of the basque country. *American Economics Review* 105, 113–132. 11
- Andrieu, C., A. Doucet, and R. Holenstein (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B* 72, 269–342. 85, 90
- Atay-Kayis, A. and H. Massam (2005). A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. *Biometrika* 92, 317–335. 10
- Banerjee, S. and S. Ghosal (2014). Posterior convergence rates for estimating large precision matrices using graphical models. *Electronic Journal of Statistics* 8, 2111–2137. 55
- Banerjee, S. and S. Ghosal (2015). Bayesian structure learning in graphical models. *Journal of Multivariate Analysis* 136, 147–162. 55
- Barber, R. F. and M. Drton (2015). High-dimensional Ising model selection with Bayesian information criteria. *Electronic Journal of Statistics* 9, 249–275. 10
- Belitser, E. and S. Ghosal (2017). Empirical Bayes oracle uncertainty quantification for regression. *Preprint*. https://www4.stat.ncsu.edu/~sghosal/papers/oracle_regression.pdf. 46, 48, 49, 51
- Bojinov, I. and N. Shephard (2017). Time series experiments and causal estimands: exact randomization tests and trading. *Arxiv: 1706.07840v2*. 8, 13
- Bonhomme, S. and U. Sauder (2011). Recovering distributions in difference-in-differences models: A comparison of selective and comprehensive schooling. *The Review of Economics and Statistics* 93(2), 479–494. 11
- Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press. 66
- Brodersen, K. H., F. Gaullusser, J. Koehler, N. Remy, and S. L. Scott (2015). Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics* 9, 247–274. 2, 7, 8, 11
- Bühlmann, P. and S. van der Geer (2011). *Statistics for High-dimensional Data: Methods, Theory and Applications*. Springer-Verlag. 46, 54

- Castillo, I. (2010). A semiparametric Bernstein-von Mises theorem for Gaussian process priors. *Probability Theory Related Fields* 152, 53–99. 76
- Castillo, I. and R. Mismar (2018). Empirical Bayes analysis of spike and slab posterior distributions. *arXiv:1801.01696*. 49, 51
- Castillo, I., J. Schmidt-Hieber, and A. van der Vaart (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics* 43, 1986–2018. 2, 46, 48, 49, 51, 52, 53, 54, 59, 60, 64, 77
- Castillo, I. and A. van der Vaart (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics* 40, 2069–2101. 1
- Chae, M., L. Lin, and D. B. Dunson (2016). Bayesian sparse linear regression with unknown symmetric error. *arXiv:1608.02143*. 46
- Chavez-Demoulin, V. and A. Davison (2012). Modeling time series extremes. *REVSTAT – Statistical Journal* 10(1), 109–133. 87
- Chen, R.-B., C.-H. Chu, S. Yuan, and Y. N. Wu (2016). Bayesian sparse group selection. *Journal of Computational and Graphical Statistics* 25, 665–683. 47
- Chib, S. (2001). *Markov Chain Monte Carlo Methods: Computation and Inference in Handbook of Econometrics, vol. 5*. North-Holland, Amsterdam: Elsevier. 36, 95
- Chopin, N. and S. S. Singh (2015, 08). On particle Gibbs sampling. *Bernoulli* 21(3), 1855–1883. 85
- Coles, S. (2001). *Introduction to Statistical Modeling of Extreme Values*. London, UK: Springer-Verlag. 83
- Coles, S. G. and E. A. Powell (1996). Bayesian methods in extreme value modelling: A review and new developments. *International Statistical Review / Revue Internationale de Statistique* 64(1), 119–136. 83
- Coles, S. G. and J. A. Tawn (1996). A Bayesian analysis of extreme rainfall data. *Journal of the Royal Statistical Society. Series C* 45, 329–347. 87
- Curtis, S. M., S. Banerjee, and S. Ghosal (2014). Fast Bayesian model assessment for nonparametric additive regression. *Computational Statistics and Data Analysis* 71, 347–358. 47
- Dawid, A. P. and S. L. Lauritzen (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics* 21, 1272–1317. 22

- de Jong, P. (1991). The diffuse Kalman filter. *The Annals of Statistics* 19, 1073–1083. [10](#)
- de Jong, P. and S. Chu-Chun-Lin (1994). Stationary and non-stationary state space models. *Journal of Time Series Analysis* 15, 151–166. [10](#)
- Ding, P. and F. Li (2017). Causal inference: A missing data perspective. *Statistical Science (to appear)*. [12](#), [13](#)
- Donald, S. G. and K. Lang (2007). Inference with difference-in-differences and other panel data. *The Review of Economics and Statistics* 89, 221–233. [11](#)
- Doucet, A. and A. M. Johansen (2011). A tutorial on particle filtering and smoothing: Fifteen years later. *Unpublished manuscript*. [85](#)
- Doudchenko, N. and G. W. Imbens (2016). Balancing, regression, difference-in-differences and synthetic control method: A synthesis. *NBER Working Paper No. 22791*. [11](#)
- Durbin, J. and S. J. Koopman (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika* 89, 603–615. [9](#), [10](#), [21](#), [29](#), [115](#)
- Durbin, J. and S. J. Koopman (2012). *Time Series Analysis by State Space Methods: Second Edition*. Great Clarendon Street, Oxford OX2 6DP: Oxford University Press. [3](#), [9](#), [21](#)
- Dyk, D. A. V. and T. Park (2008). Partially collapsed Gibbs samplers: Theory and methods. *Journal of the American Statistics Association* 103(482), 790–796. [84](#), [87](#), [90](#)
- Einicke, G. A. and L. B. White (1999). Robust extended Kalman filtering. *IEEE Transactions on signal processing* 47. [85](#)
- Galindo-Garre, F. and J. K. Vermunt (2006). Avoiding boundary estimates in latent class analysis by Bayesian posterior mode estimation. *Behaviormetrika* 33, 43–59. [11](#)
- Gao, C. and H. H. Zhou (2016). Bernstein-von Mises theorems for functionals of the covariance matrix. *Electronic Journal of Statistics* 10, 1751–1806. [81](#)
- Gelfand, A. E., A. F. M. Smith, and T.-M. Lee (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistics Association* 87, 523–532. [10](#)
- George, E. I. and R. E. McCulloch (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* 88(423), 881–889. [18](#)

- Ghosal, S. and A. van der Vaart (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press. 48, 52, 55, 80
- Greenlaw, K., E. Szefer, J. Graham, M. Lesperance, F. S. Nathoo, and F. the Alzheimer’s Disease Neuroimaging Initiative (2017). A Bayesian group sparse multi-task regression model for imaging genetics. *Bioinformatics* 33, 2513–2522. 47
- Gunn, L. H. and D. B. Dunson (2005). A transformation approach for incorporating monotone or unimodal constraints. *Biostatistics* 6, 434–449. 11
- Harvey, A. C. and S. Peters (1990). Estimation procedures for structural time series models. *Journal of Forecasting* 9, 89–108. 9
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association* 81, 945–960. 12
- Huang, J. and T. Zhang (2010). The benefit of group sparsity. *The Annals of Statistics* 38, 1978–2004. 47, 55
- Huerta, G. and B. Sansó (2007). Time-varying models for extreme values. *Environmental and Ecological Statistics* 14, 285–299. 83
- Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York, NY, USA: Cambridge University Press. 12
- Kantas, N., A. Doucet, S. S. Singh, J. Maciejowski, and N. Chopin (2015). On particle methods for parameter estimation in state-space models. *Statistical Science* 30(3), 328–351. 86
- Khare, K., B. Rajaratnam, and A. Saha (2015). Bayesian inference for Gaussian graphical models beyond decomposable graphs. *Arxiv:1505.00703*. 10
- Koopman, S. J. (1997). Exact initial Kalman filtering and smoothing for nonstationary time series models. *Journal of the American Statistical Association* 92, 1630–1638. 10
- Lauritzen, S. L. (1996). *Graphical Models*. New York, USA: Oxford University Press Inc. 10
- Letac, G. and H. Massam (2007). Wishart distributions for decomposable graphs. *The Annals of Statistics* 35, 1278–1323. 3
- Li, F. and N. R. Zhan (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association* 105(491), 1202–1214. 46, 47

- Li, F., T. Zhang, Q. Wang, M. Z. Gonzalez, E. L. Maresh, and J. A. Coan (2015). Spatial Bayesian variable selection and grouping for high-dimensional scalar-on-image regression. *The Annals of Applied Statistics* 9, 687–713. [10](#)
- Lindsten, F., M. I. Jordan, and T. B. Schön (2014). Particle Gibbs with ancestor sampling. *Journal of Machine Learning Research* 15, 2145–2184. [86](#), [95](#)
- Lindsten, F. and T. B. Schön (2013). Backward simulation methods for Monte Carlo statistical inference. *Foundations and Trends in Machine Learning* 6(1), 1–143. [86](#)
- Liquet, B., K. Mengersen, A. N. Pettitt, and M. Sutton (2017, 12). Bayesian variable selection regression of multivariate responses for group data. *Bayesian Analysis* 12, 1039–1067. [47](#)
- Liu, J. S. (1994). The collapsed Gibbs sampler in Bayesian computation with applications to a gene regulation problem. *Journal of the American Statistical Association* 89(427), 958–996. [84](#), [85](#)
- Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. New York, NY, USA: Springer-Verlag New York, Inc. [85](#), [90](#)
- Liu, J. S., W. H. Wong, and A. Kong (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* 81(1), 27–40. [84](#)
- Lounici, K., M. Pontil, A. B. Tsybakov, and S. van de Geer (2009). Taking advantage of sparsity in multi-task learning. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT-2009)*, 73–82. [47](#)
- Lounici, K., M. Pontil, S. van de Geer, and A. B. Tsybakov (2011). Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics* 39, 2164–2204. [47](#), [79](#)
- Martin, R., R. Mess, and S. G. Walker (2017). Empirical Bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli* 23, 1822–1857. [1](#), [46](#), [48](#), [49](#), [51](#)
- Mitsakakis, N., H. Massam, and M. D. Escobar (2011). A Metropolis-Hastings based method for sampling from the G-Wishart distribution in Gaussian graphical models. *Electronic Journal of Statistics* 5, 18–30. [10](#)
- Mohammadi, A. and E. C. Wit (2015). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Analysis* 10, 109–138. [10](#), [23](#)
- Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, Inc. [79](#)

- Nakajima, J., T. Kuniyama, and Y. Omori (2017). Bayesian modeling of dynamic extreme values: Extension of generalized extreme value distributions with latent stochastic processes. *Journal of Applied Statistics* 44(7), 1248–1268. 84
- Nakajima, J., T. Kuniyama, Y. Omori, and S. Frühwirth-Schnatter (2011). Generalized extreme value distribution with time-dependence using the AR and MA models in state space form. *Computational Statistics and Data Analysis* 56(11), 3241–3259. 84, 88
- Nardi, Y. and A. Rinaldo (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics* 2, 605–633. 47
- Niemi, J. and M. West (2010). Adaptive mixture modeling Metropolis methods for Bayesian analysis of nonlinear state-space models. *Journal of Computational and Graphical Statistics* 19(2), 260–280. 3, 85
- Ning, B., S. Ghosal, and J. Thomas (2018). Bayesian method for causal inference in spatially-correlated multivariate time series. *Bayesian Analysis (to appear)*. 48
- Primiceri, G. E. (2005). Time varying structural vector autoregressions and monetary policy. *The Review of Economic Studies* 72, 821–852. 21
- Ročková, V. and E. I. George (2014). EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association* 109(506), 828–846. 11, 19, 27, 47
- Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association* 102, 191–200. 12
- Ročková, V. (2018). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *The Annals of Statistics* 46, 401–437. 46
- Roverato, A. (2000). Cholesky decomposition of a hyper inverse Wishart matrix. *Biometrika* 87, 99–112. 10, 22
- Roverato, A. (2002). Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics* 29, 391–411. 10, 22
- Roy, A., T. S. McElroy, and P. Linton (2016). Constrained estimation of causal invertible VARMA models. *Statistica Sinica (to appear)*. 5, 11, 17
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66, 688–701. 12

- Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics* 2, 1–26. 12
- Rubin, D. B. (2005). Causal inference using potential outcomes: design, modeling, decisions. *Journal of the American Statistical Association* 100, 322–331. 2, 8, 11, 12
- Scott, D. W. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, Inc. 79
- Shephard, N. and M. K. Pitt (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika* 84, 653–667. 3, 22, 85
- Smith, M. and L. Fahrmeir (2007). Spatial Bayesian variable selection with application to functional magnetic resonance imaging. *Journal of the American Statistical Association* 102, 417–431. 10
- Song, Q. and F. Liang (2017). Nearly optimal Bayesian shrinkage for high-dimensional regression. *arXiv:1712.08964*. 46
- Stein, P. (1952). Some general theorems on iterants. *Journal of Research of the National Bureau of Standards* 48, 82–83. 17
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science* 25, 1–21. 11
- Suarez, A. J. and S. Ghosal (2017). Bayesian estimation of principal components for functional data. *Bayesian Analysis* 12, 311–333. 69
- Ueda, N. and R. Nakano (1998). Deterministic annealing EM algorithm. *Neural Networks* 11, 271–282. 27
- Uhler, C., A. Lenkoski, and D. Richardsz (2017). Exact formulas for the normalizing constants of Wishart distributions for graphical models. *The Annals of Statistics* 46, 90–118. 10
- Wang, H. and S. Z. Li (2012). Efficient Gaussian graphical model determination under G-Wishart prior distributions. *Electronic Journal of Statistics* 6, 168–198. 10
- Wei, Y. and G. Huerta (2016). Dynamic generalized extreme value modeling via particle filters. *Communications in Statistics* 46(8). 83
- West, M. and J. Harrison (1997). *Bayesian Forecasting and Dynamic Models (2nd Ed.)*. New York, NY, USA: Springer-Verlag New York, Inc. 85
- Whiteley, N. (2010). Discussion on the paper “particle Markov chain Monte Carlo methods”. *Journal of the Royal Statistical Society: Series B* 72(3). 86

- Xu, X. and M. Ghosh (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Analysis* 10, 909–936. 47, 51
- Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of Royal Statistical Society: Series B* 68, 49–67. 4, 47

APPENDICES

Appendix A

The Kalman filter and backward smoothing algorithm

For the following state-space model

$$\begin{aligned}\mathbf{Y}_t &= \mathbf{z}\boldsymbol{\alpha}_t + \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\epsilon}_t, \\ \boldsymbol{\alpha}_{t+1} &= \mathbf{c} + \mathbf{T}\boldsymbol{\alpha}_t + \mathbf{R}\boldsymbol{\eta}_t,\end{aligned}$$

where $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$, $\mathbf{Q} = \text{bdiag}(\boldsymbol{\Sigma}_u, \boldsymbol{\Sigma}_v, \boldsymbol{\Sigma}_w)$ are mutually independent, the [Durbin and Koopman \(2002\)](#)'s KFBS algorithm is described as follows: for a dataset with no missing values, the Kalman filter for updating $\boldsymbol{\alpha}_t$, where $\boldsymbol{\alpha}_t \sim \mathcal{N}(\mathbf{a}_t, \mathbf{P}_t)$, $t = 1, \dots, T$, is given by

$$\begin{aligned}\mathbf{v}_t &= \mathbf{Y}_t - \mathbf{z}\mathbf{a}_t, & \mathbf{F}_t &= \mathbf{z}\mathbf{P}_t\mathbf{z}' + \boldsymbol{\Sigma}, \\ \mathbf{K}_t &= \mathbf{T}\mathbf{P}_t\mathbf{z}'\mathbf{F}_t^{-1}, & \mathbf{L}_t &= \mathbf{T} - \mathbf{K}_t\mathbf{z}, \\ \mathbf{a}_{t+1} &= \mathbf{c} + \mathbf{T}\mathbf{a}_t + \mathbf{K}_t\mathbf{v}_t, & \mathbf{P}_{t+1} &= \mathbf{T}\mathbf{P}_t\mathbf{L}_t' + \mathbf{R}\mathbf{Q}\mathbf{R}'.\end{aligned}$$

For updating the missing values in the dataset, the Kalman filter algorithm becomes

$$\mathbf{a}_{t+1} = \mathbf{c} + \mathbf{T}\mathbf{a}_t, \quad \mathbf{P}_{t+1} = \mathbf{T}\mathbf{P}_t\mathbf{T}' + \mathbf{R}\mathbf{Q}\mathbf{R}'.$$

From the above, we obtain \mathbf{a}_{t+1} and \mathbf{P}_{t+1} depending on the data $\mathbf{Y}_1, \dots, \mathbf{Y}_t$. The backward smoothing algorithm will give the conditional expectation of \mathbf{a}_t and \mathbf{P}_t given the full dataset $\mathbf{Y}_1, \dots, \mathbf{Y}_T$. To distinguish between those two values, we denote the

latter one as $\mathbf{a}_{t|T}$ and $\mathbf{P}_{t|T}$.

The backward smoothing algorithm is given by

$$\begin{aligned}\mathbf{r}_{t-1} &= \mathbf{z}'\mathbf{F}_t^{-1}\mathbf{v}_t + \mathbf{L}'_t\mathbf{r}_t, & \mathbf{N}_{t-1} &= \mathbf{z}'\mathbf{F}_t^{-1}\mathbf{z} + \mathbf{L}'_t\mathbf{N}_t\mathbf{L}_t, \\ \mathbf{a}_{t|T} &= \mathbf{a}_t + \mathbf{P}_t\mathbf{r}_{t-1}, & \mathbf{P}_{t|T} &= \mathbf{P}_t - \mathbf{P}_t\mathbf{N}_{t-1}\mathbf{P}_t.\end{aligned}$$

The covariance between $\boldsymbol{\alpha}_t$ and $\boldsymbol{\alpha}_{t-1}$ is given by

$$\mathbf{P}_{t-1,t|T} = \text{cov}(\boldsymbol{\alpha}_{t-1}, \boldsymbol{\alpha}_t | \mathbf{Y}_1, \dots, \mathbf{Y}_T) = \mathbf{P}_{t-1}\mathbf{L}'_{t-1}(\mathbf{I} - \mathbf{N}_{t-1}\mathbf{P}_t).$$

Appendix B

Deriving the EMVS algorithm

In this chapter, we provide the details on deriving the EMVS algorithm used in Chapter 2. The algorithm contains two steps: the E-step (the expectation step) and the M-step (the maximization step).

B.1 E-step

The expectation of α_t^* , $\alpha_t^* \alpha_t^{*'}$ and $\alpha_t^* \alpha_{t-1}^{*'}$ respect to $\alpha_{1:T}^* | \cdot$ can be estimated using the KFBS algorithm (see details in Chapter A).

The term $\mathbb{E}_{\gamma | \cdot} [\log (v_0(1-\gamma_i) + v_1\gamma_i)]$ is a constant when conditioned on the parameters $\beta^{(k)}, v_0, v_1$.

Let $a_i^{*(k)} = \mathbb{E}_{\gamma | \cdot} \left[\frac{1}{v_0(1-\gamma_i) + v_1\gamma_i} \right]$, because γ_i is a binary variable,

$$a_i^{*(k)} = \mathbb{E}_{\gamma | \cdot} \left[\frac{1}{v_0(1-\gamma_i) + v_1\gamma_i} \right] = \frac{\mathbb{E}_{\gamma | \cdot} (1-\gamma_i)}{v_0} + \frac{\mathbb{E}_{\gamma | \cdot} (\gamma_i)}{v_1}.$$

Denote $w_i^{(k)} = \mathbb{E}_{\gamma | \cdot} (\gamma_i)$, then

$$w_i^{(k)} = \mathbb{E}_{\gamma | \cdot} (\gamma_i) = P(\gamma_i = 1 | \beta^{(k)}, \theta^{(k)}) = \frac{g_{1i}^{(k)}}{g_{1i}^{(k)} + g_{2i}^{(k)}},$$

where $g_{1i}^{(k)} = \pi(\beta_i^{(k)} | v_0, v_1, \gamma_i = 1)P(\gamma_i = 1 | \theta^{(k)})$ and $g_{2i}^{(k)} = \pi(\beta_i^{(k)} | v_0, v_1, \gamma_i = 0)P(\gamma_i = 0 | \theta^{(k)})$. Based on the prior in Section 3, $g_{1i}^{(k)}$ and $g_{2i}^{(k)}$ can be written as $g_{1i}^{(k)} = \pi(\beta_i^{(k)} | v_0, v_1, \gamma_i = 1)\theta^{(k)}$ and $g_{2i}^{(k)} = \pi(\beta_i^{(k)} | v_0, v_1, \gamma_i = 0)(1 - \theta^{(k)})$.

If using DAEMVS algorithm with parameter s , then

$$w_i^{(k)} = \frac{(g_{1i}^{(k)})^s}{(g_{1i}^{(k)})^s + (g_{2i}^{(k)})^s}.$$

To summarize, in the k -th E-step, we find

$$\begin{aligned} & \mathcal{Q}(\boldsymbol{\beta}, \theta, \boldsymbol{\Phi}, \boldsymbol{\Sigma}, \mathbf{Q} \mid \boldsymbol{\beta}^{(k)}, \theta^{(k)}, \boldsymbol{\Phi}^{(k)}, \boldsymbol{\Sigma}^{(k)}, \mathbf{Q}^{(k)}) \\ &= \mathbb{E}_{(\boldsymbol{\alpha}_{1:T}^*, \gamma) \cdot} [\log \pi(\boldsymbol{\alpha}_{1:T}^*, \boldsymbol{\beta}, \gamma, \theta, \boldsymbol{\Phi}, \boldsymbol{\Sigma}, \mathbf{Q} \mid \mathbf{Y}_t^*, \mathbf{X}_t^*) \mid \boldsymbol{\beta}^{(k)}, \theta^{(k)}, \boldsymbol{\Phi}^{(k)}, \boldsymbol{\Sigma}^{(k)}, \mathbf{Q}^{(k)}], \end{aligned} \quad (\text{B.1.1})$$

where $\mathbb{E}_{(\boldsymbol{\alpha}_{1:T}^*, \gamma) \cdot}(\cdot)$ denotes the conditional expectation $\mathbb{E}_{(\boldsymbol{\alpha}_{1:T}^*, \gamma) \mid \boldsymbol{\beta}^{(k)}, \theta^{(k)}, \boldsymbol{\Phi}^{(k)}, \boldsymbol{\Sigma}^{(k)}, \mathbf{Q}^{(k)}}$.

B.2 M-step

We first denote

$$\begin{aligned} \mathbf{a}_{t|T}^{*(k+1)} &= \mathbb{E}_{\boldsymbol{\alpha}_{1:T}^* \cdot}(\boldsymbol{\alpha}_t^* \mid \mathbf{Y}_{1:T}^*, \mathbf{X}_{1:T}^*, \boldsymbol{\beta}^{(k)}, \theta^{(k)}, \boldsymbol{\Phi}^{(k)}, \boldsymbol{\Sigma}^{(k)}, \mathbf{Q}^{(k)}), \\ \mathbf{V}_{t|T}^{*(k+1)} &= \mathbb{E}_{\boldsymbol{\alpha}_{1:T}^* \cdot}(\boldsymbol{\alpha}_t^* \boldsymbol{\alpha}_t^{*'} \mid \mathbf{Y}_{1:T}^*, \mathbf{X}_{1:T}^*, \boldsymbol{\beta}^{(k)}, \theta^{(k)}, \boldsymbol{\Phi}^{(k)}, \boldsymbol{\Sigma}^{(k)}, \mathbf{Q}^{(k)}), \\ \mathbf{V}_{t,t-1|T}^{*(k+1)} &= \mathbb{E}_{\boldsymbol{\alpha}_{1:T}^* \cdot}(\boldsymbol{\alpha}_t^* \boldsymbol{\alpha}_{t-1}^{*'} \mid \mathbf{Y}_{1:T}^*, \mathbf{X}_{1:T}^*, \boldsymbol{\beta}^{(k)}, \theta^{(k)}, \boldsymbol{\Phi}^{(k)}, \boldsymbol{\Sigma}^{(k)}, \mathbf{Q}^{(k)}), \\ \mathbf{P}_{t,t-1|T}^{*(k+1)} &= \mathbf{V}_{t,t-1|T}^{*(k+1)} - \mathbf{a}_{t|T}^{*(k+1)} (\mathbf{a}_{t-1|T}^{*(k+1)})', \\ \mathbf{P}_{t|T}^{*(k+1)} &= \mathbf{V}_{t|T}^{*(k+1)} - \mathbf{a}_{t|T}^{*(k+1)} (\mathbf{a}_{t|T}^{*(k+1)})', \end{aligned}$$

and obtain $\mathbf{a}_{t|T}^{*(k+1)}$, $\mathbf{P}_{t|T}^{*(k+1)}$ and $\mathbf{P}_{t,t-1|T}^{*(k+1)}$ from the KFBS algorithm for each $t = 1, \dots, T$.

We define $\mathbf{A}_{\boldsymbol{\gamma}}^{*(k)} = \text{diag}(a_1^{*(k)}, \dots, a_p^{*(k)})$, recall a_i is the i -th diagonal element in $\mathbf{A}_{\boldsymbol{\gamma}}$. For each $a_i^{*(k)}$, we have

$$a_i^{*(k)} = \mathbb{E}_{\boldsymbol{\gamma} \cdot} \left[\frac{1}{a_i} \right] = \mathbb{E}_{\boldsymbol{\gamma} \cdot} \left[\frac{1}{v_0(1 - \gamma_i) + v_1 \gamma_i} \right] = \frac{1 - w_i^{(k)}}{v_0} + \frac{w_i^{(k)}}{v_1},$$

where $w_i^{(k)} = \mathbb{E}_{\boldsymbol{\gamma} \cdot}(\gamma_i) = \mathbb{P}(\gamma_i = 1 \mid \boldsymbol{\beta}^{(k)}, \theta^{(k)}) = \frac{g_{1i}^{(k)}}{g_{1i}^{(k)} + g_{2i}^{(k)}}$ with

$$\begin{aligned} g_{1i}^{(k)} &= \pi(\beta_i^{(k)} \mid v_0, v_1, \gamma_i = 1) \theta^{(k)}, \\ g_{2i}^{(k)} &= \pi(\beta_i^{(k)} \mid v_0, v_1, \gamma_i = 0) (1 - \theta^{(k)}). \end{aligned}$$

Then in the k -th M-step, we maximize (B.1.1) with respect to θ , β , Φ , Σ^{-1} and \mathbf{Q} . Their optimized values can be obtained by executing the following steps:

$$\begin{aligned}
\theta^{(k+1)} &= \frac{\sum_{i=1}^p w_i^{(k)} + \zeta_1 - 1}{p + \zeta_1 + \zeta_2 - 2}, \\
\beta^{(k+1)} &= \left(\sum_{t=1}^T \mathbf{X}_t^{*'} (\Sigma^{(k)})^{-1} \mathbf{X}_t^* + \mathbf{A}_{\gamma}^{*(k+1)} \right)^{-1} \left(\sum_{t=1}^T \mathbf{X}_t^{*'} (\Sigma^{(k)})^{-1} (\mathbf{Y}_t^* - \mathbf{z} \mathbf{a}_{t|T}^{*(k+1)}) \right), \\
\text{vec}(\Phi^{(k+1)}) &= \left(\sum_{t=1}^{T-1} \left((\mathbf{V}(\boldsymbol{\tau})_{t|T}^{*(k+1)})' \otimes (\Sigma_v^{(k)})^{-1} \right) + 10\mathbf{I}_{n^2} \right)^{-1} \\
&\quad \times \left(\sum_{t=1}^{T-1} \left((\mathbf{V}(\boldsymbol{\tau})_{t,t-1|T}^{*(k+1)})' \otimes (\Sigma_v^{(k)})^{-1} \right) \text{vec}(\mathbf{I}_n) \right), \\
\Sigma^{(k+1)} &= \frac{1}{T + \nu - 2} \left[\sum_{t=1}^T (\mathbf{Y}_t^* - \mathbf{X}_t^* \beta^{(k+1)}) (\mathbf{Y}_t^* - \mathbf{X}_t^* \beta^{(k+1)})' \right. \\
&\quad - \sum_{t=1}^T \mathbf{z} \mathbf{a}_{t|T}^{*(k+1)} (\mathbf{Y}_t^* - \mathbf{X}_t^* \beta^{(k+1)})' - \sum_{t=1}^T (\mathbf{Y}_t^* - \mathbf{X}_t^* \beta^{(k+1)}) (\mathbf{a}_{t|T}^{*(k+1)})' \mathbf{z}' \\
&\quad \left. + \sum_{t=1}^T \mathbf{z} \mathbf{V}_{t|T}^{*(k+1)} \mathbf{z}' + \mathbf{H} \right] \mathbb{1}_{\{\Sigma^{-1} \in M^+(\mathcal{G})\}}, \\
\mathbf{Q}^{(k+1)} &= \frac{1}{T + \nu - 3} \left[\left(\sum_{t=1}^{T-1} \left(\mathbf{R}' \mathbf{V}_{t+1|T}^{*(k+1)} - \mathbf{T}^{(k+1)} \mathbf{V}_{t,t+1|T}^{*(k+1)} - (\mathbf{V}_{t,t+1|T}^{*(k+1)})' (\mathbf{T}^{(k+1)})' \right. \right. \right. \\
&\quad \left. \left. + \mathbf{T}^{(k+1)} \mathbf{V}_{t|T}^{*(k+1)} (\mathbf{T}^{(k+1)})' \right) \mathbf{R} \right) \circ \begin{pmatrix} \mathbf{1}_n \mathbf{1}_n' & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_n \mathbf{1}_n' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1}_n \mathbf{1}_n' \end{pmatrix} \\
&\quad \left. + \begin{pmatrix} k_1^2(n+1)\mathbf{H} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & k_2^2(n+1)\mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & k_3^2(n+1)\mathbf{H} \end{pmatrix} \right] \mathbb{1}_{\{\Sigma_u^{-1}, \Sigma_v^{-1}, \Sigma_w^{-1} \in M^+(\mathcal{G})\}},
\end{aligned} \tag{B.2.1}$$

where $\mathbb{1}_A$ stands for the indicator function of a set A , $\mathbf{1}_n$ is an $n \times 1$ vector with its elements are all 1s, $E \circ F$ is the Hadamard product between matrices E and F , $\mathbf{V}(\boldsymbol{\tau})_{t|T}^{*(k+1)}$ is the covariance matrix for $\boldsymbol{\tau}_t$ which is inside of $\mathbf{V}_{t|T}^{*(k+1)}$. The similar notation is used for $\mathbf{V}(\boldsymbol{\tau})_{t,t-1|T}^{*(k+1)}$. The E- and M-steps are repeated until the values of $\mathcal{Q}(\beta, \theta, \Phi, \Sigma, \mathbf{Q} \mid \beta^{(k)}, \theta^{(k)}, \Phi^{(k)}, \Sigma^{(k)}, \mathbf{Q}^{(k)})$ stop increasing significantly. Since $p > n$,

we can apply the Sherman-Morrison-Woodbury formula to rewrite $(\mathbf{X}_t^{*'}(\boldsymbol{\Sigma}^{(k)})^{-1}\mathbf{X}_t^* + \mathbf{A}_\gamma^{*(k+1)})^{-1}$ in $\boldsymbol{\beta}^{(k+1)}$ as

$$(\mathbf{A}_\gamma^{*(k+1)})^{-1} - (\mathbf{A}_\gamma^{*(k+1)})^{-1}\mathbf{X}_t^{*'}[\boldsymbol{\Sigma}^{(k)} + \mathbf{X}_t^*(\mathbf{A}_\gamma^{*(k+1)})^{-1}\mathbf{X}_t^{*'}]^{-1}\mathbf{X}_t^*(\mathbf{A}_\gamma^{*(k+1)})^{-1},$$

to only invert an $n \times n$ matrix instead of a $p \times p$ matrix.

Appendix C

Plots of MCMC outputs for the water flow data and the S&P 500 data

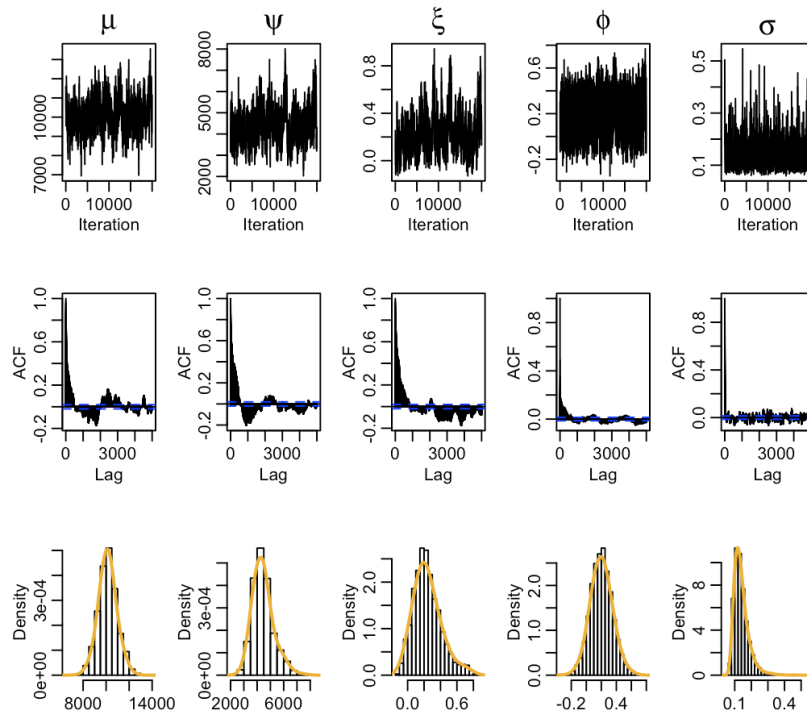


Figure C.1: Plot of MCMC draws (1st row), autocorrelation functions (ACF) (2nd row) and histograms along with densities (yellow lines) (3rd row) for parameters μ , ψ , ξ , ϕ , σ^2 in dGEV model of the water flow dataset.

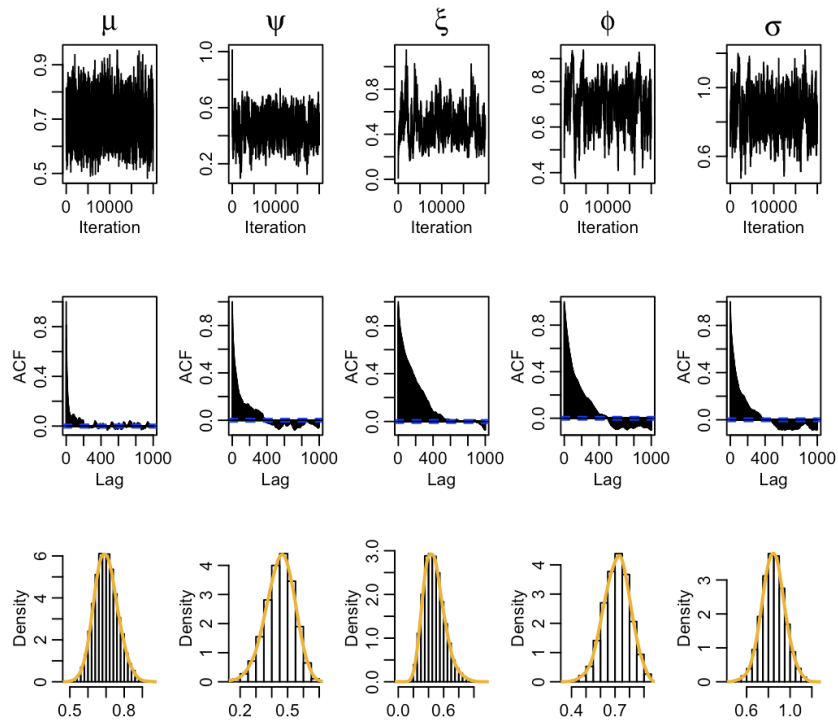


Figure C.2: Plot of MCMC draws (1st row), autocorrelation functions (ACF) (2nd row) and histograms along with densities (yellow line) (3rd row) for parameters μ , ψ , ξ , ϕ , σ^2 in the dGEV model by fitting the S&P 500 dataset.

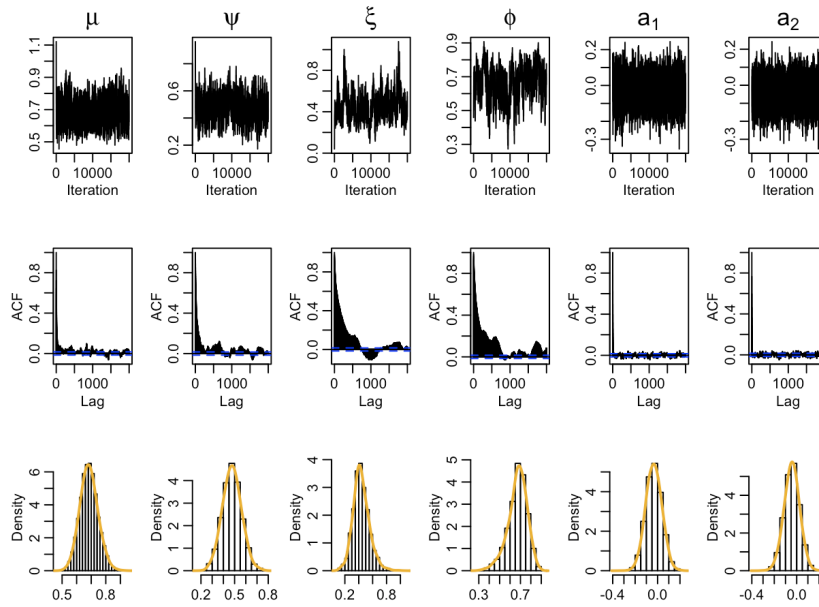


Figure C.3: Plot of MCMC draws (1st row), autocorrelation function (ACF) (2nd row) and histograms along with densities (yellow line) (3rd row) of parameters μ , ψ , ξ , ϕ , a_1 , a_2 in a seasonal dGEV model by fitting the S&P 500 dataset.