ABSTRACT

BHATTACHARYA, INDRABATI. Bayesian Inference on Multivariate Median and Quantiles. (Under the direction of Subhashis Ghoshal.)

Median and quantiles are robust alternatives of moments-based estimators, like mean and covariance. In higher dimensions, there is no objective notion of ordering, and as a consequence, there is no universally accepted definition of quantiles. Various definitions of multivariate quantiles have been proposed over the years, and their properties have been investigated. In this dissertation, we study nonparametric Bayesian and Gibbs posterior approaches for inference on multivariate quantiles.

In Chapter 2, we consider Bayesian inference on a type of multivariate median and quantile functionals of a joint distribution using a Dirichlet process prior. Since, unlike univariate quantiles, the exact posterior distribution of multivariate median and quantiles are not obtainable exactly, we study these distributions asymptotically. We derive a Bernstein-von Mises theorem for the multivariate ℓ_1-median with respect to a general ℓ_p-norm, which, in particular, shows that its posterior concentrates around its true value at the $\sqrt{n}$-rate and its credible sets have asymptotically correct frequentist coverage. In particular, asymptotic normality for the empirical multivariate median with a general ℓ_p-norm is also derived in the course of the proof, which extends the results from $p = 2$ in the literature to a general p. The technique involves approximating the posterior Dirichlet process by a Bayesian bootstrap process and deriving a conditional Donsker theorem. We also obtain an analogous result for an affine equivariant version of the multivariate ℓ_1-median based on an adaptive transformation and re-transformation technique. The results are extended to the joint distribution of a set of multivariate quantiles. The accuracy of the asymptotic result is assessed by a simulation study. We also use the results to obtain a Bayesian credible region for multivariate medians for Fisher's iris data, which consists of four features measured for each of three plant species.

Chapter 3 explores the Gibbs posterior approach for inference on multivariate quantiles. The Gibbs posterior is a direct and model-free approach and is free from the issues that one may run into while using parametric Bayesian and other likelihood-based methods for inference on quantiles. The quantiles are not naturally described as parameters of a statistical model, and hence probabilistic inference on them are challenging. Being model-free means that inferences drawn from the Gibbs posterior are not subject to model misspecification bias, and being direct means that no priors for or marginalization over nuisance parameters are required. In this chapter, we show that the Gibbs posterior enjoys a $\sqrt{n}$- convergence

rate and a Bernstein–von Mises property, i.e., for large n, the Gibbs posterior distribution can be approximated by a normal distribution. Moreover, we present numerical results showing the validity and efficiency of credible sets derived from a suitably scaled Gibbs posterior.

In Chapter 4, we propose Bayesian non-parametric tests for one-sample and two-sample multivariate location problems. We model the underlying distributions using a Dirichlet process prior. For one-sample problem, we compute a Bayesian credible set of the multivariate spatial median and accept the null hypothesis if the credible set contains the null value. For the two-sample problem, we form a credible set of the difference of the spatial medians of the two samples and we accept the null hypothesis of equality if the credible set contains zero. We derive the local asymptotic power of the test under Pitman alternative, and also present a simulation study to compare the finite-sample performance of our testing procedures with existing parametric and non-parametric tests.

In Chapter 5, we consider a non-parametric Bayesian approach to multivariate quantile regression. We model the collection of conditional distributions of a response vector Y given the vector X using a Dependent Dirichlet Process (DDP) prior. This way, marginally for each value x of X, the random conditional distribution $P_{Y|x}$ follows a Dirichlet process. The DDP is used to introduce dependence across x. As the random distributions $P_{Y|x}$ are almost surely discrete, we need to convolve it with a kernel. To model the error distribution as flexibly as possible, we use a finite mixture of multidimensional normal distributions as our kernel. For posterior computations, we use a truncated stick-breaking representation of the DDP. This approximation enables us to deal with only a finitely number of parameters. We use a Block Gibbs sampler for estimating the model parameters. We illustrate our method with simulation studies and real data applications. Finally, we provide a theoretical justification for the proposed method through posterior consistency. Our proposed procedure is new even when the response is univariate.

Bayesian Inference on Multivariate Median and Quantiles

by
Indrabati Bhattacharya

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

2020

APPROVED BY:

Sujit Ghosh	Ryan Martin
Minh Tang	Subhashis Ghoshal Chair of Advisory Committee

DEDICATION

To my parents, and my husband, Ranadeep.

BIOGRAPHY

The author was born in Kolkata, India on September 21, 1992. She earned a bachelor's degree in Statistics from St. Xavier's College, Kolkata, India, under the affiliation of the University of Calcutta, India, in the year 2014. After that, she got a Masters of Statistics (M.Stat) degree from Indian Statistical Institute, Kolkata, India, in the year 2016. After graduating in June 2016, she came to the United states to pursue a Ph.D. degree in Statistics at North Carolina State University.

ACKNOWLEDGEMENTS

It has been quite an eventful journey over the last four years. This journey has been shaped by quite a few wonderful people, and here I thank all of them. Firstly, I express my sincere gratitude to my advisor Dr. Subhashis Ghoshal for his brilliant inspiration, patient guidence and kind support on my research. He has been incredibly patient with me at all times over the last three years, even at times when I kept making the same mistakes repeatedly! As he guided me through the last three years, I have realized that I could not have asked for a better advisor.

I was lucky to have another mentor in Dr. Ryan Martin. Apart from serving on my committee, he was kind enough to share one of his research ideas with me, which eventually became one of the chapters in this thesis. Whenever I reached out to him with complex research questions, he took time to write out crystal clear explanations. I had quite a few intellectually charged discussions with him, some of which had made me view my own research problems in a new light. Apart from research, Dr. Martin had also helped me out with my job search. Whenever I needed a piece of advice, he was literally just one email away!

I am thankful to the other members in my committee, Dr. Sujit Ghosh and Dr. Minh Tang. Dr. Ghosh had provided helpful comments during my prelims and also helped me prepare my CV for my job search. I am especially grateful to Dr. Minh Tang, who agreed to be on my committee at the very last minute. I am also thankful to Dr. Soumendra Lahiri, who served on my committee at the past, but I had to replace him due to unavoidable circumstances. Finally, I would like to thank Dr. Steven Hunter for serving as the graduate school representative on my committee.

I am grateful to all the faculty and staff members in the Department of Statistics at North Carolina State University, who provided me with a lot of help and support during the course of my graduate study. A special mention goes out to Dr. Charles Smith, whose cheerful words have been a constant source of support. He also deserves a mention for all the wonderful food he used to bring me when I used to work late at the department.

I found a family in my wonderful friends in Raleigh, to name a few, Suman, Rahul, Salil, Dhrubajyoti, Arnab, Indranil, Moumita, Sohini, Tuhin, Sukanya, and Akanksha. I am fortunate to still be in touch with some of my friends from my undergrad and Masters days in India who check on me regularly, namely, Abhijnan, Sneha, Sanhita, Vishal, Srijata and Animesh. All of them have been a part of this journey and been a positive influence on me.

I would like to acknowledge the constant mental support that I have received from my parents. They have always been my biggest source of inspiration and helped me to stay

positive, even in the darkest of times. Lastly, I owe thanks to my loving husband, Ranadeep, for his continued love and support. Even after living more than 900 miles away from me and having his own graduate studies to deal with, he has always been around at times when things became difficult, and helped me keep things in perspective. I greatly value his contribution and deeply appreciate his belief in me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

Inference on multivariate quantiles is an important and interesting problem, as they are popular robust alternatives to method-of-moments estimators, namely, multivariate mean vector and covariance matrix. If the data cloud does not conform to well-known probability distributions, and exhibits significant departure from Gaussianity, the moment based estimators may be unstable. Lack of an objective basis of ordering makes the problem of extending univariate quantiles to higher dimensions significantly challenging. There have been various attempts towards multidimensional generalization of univariate quantiles. We start by briefly discussing the historical overview of multivariate medians and quantiles.

A starting point of research on multivariate medians was the Twelfth Census of USA in 1900 (Small (1990)). Researchers wanted to study the flow of the population in USA and intended to find the geographical center of the population. Hayford (1992) recognized that the centroid of the geographical distribution is an inappropriate measure of location in this context because the death of an individual on an outer edge of the country affects the centroid more than the death of someone in the middle part of the country does. This property of the centroid created the need for a median-like measure. Hayford (1992) proposed the vector of coordinate-wise medians of the orthogonal coordinates as a possible solution but also mentioned that the solution is dependent on the choice of orthogonal coordinates. Scates (1933) reworked on this problem and used the newly developed idea

of spatial median as the solution which is obtained by minimizing the average distance. Scates (1933) found the geographical center to be at '15 miles northwest of Dayton, Ohio'.

The vector of coordinate-wise medians is equivariant under coordinate-wise locations and scale transformations, but not equivariant under arbitrary affine group of transformations. There are data-driven transformation techniques available in the literature to make the coordinate-wise median affine equivariant (Chakraborty and Chaudhuri (1996)). The asymptotic properties of the coordinate-wise medians have also been studied in the literature, and it has been established that its asymptotic distribution is Gaussian (Babu and Rao (1989)).

While the coordinate-wise medians do not generally provide much information about the joint distribution, a more natural version of the multivariate median is the ℓ_1-median, which is the minimizer of the average of distances of the data points. The most commonly used version is the spatial median, which minimizes the sum of the Euclidean distances. The spatial median has many attractive properties, including a high breakdown point. The concept of breakdown helps us to understand the robustness of an estimator. Suppose We have a random sample $X = X_1, X_2, \ldots, X_n$ and we wish to estimate θ by an estimator $\hat{\theta}(X)$. We can construct a contaminated sample X' by replacing m many points by arbitrary values and estimate θ by $\hat{\theta}(X')$. The maximum bias induced is given by

$$\text{Bias}\,(m;\hat{\theta}) = \sup_{X'} \|\hat{\theta}(X') - \hat{\theta}(X)\|_2, \tag{1.0.1}$$

where $\|\cdot\|_p$ denotes the ℓ_p-norm. The breakdown point is then given by

$$\text{BD}\,(\hat{\theta};X) = \inf\{\frac{m}{n} : \text{Bias}\,(m;\hat{\theta}) = \infty\}. \tag{1.0.2}$$

This thesis covers Bayesian inference on the ℓ_1-median, with a special focus on spatial medians. Just like the coordinate-wise medians, the ℓ_1-median also suffers from the drawback of not being affine equivariant, and the transformation-and-retransformation technique can be applied to ℓ_1-median as well (Chakraborty, Chaudhuri and Oja (1998)). The asymptotic normality of the spatial median is shown by Möttönen et al. (2010). There are some other notions of a multivariate median as well, based on the concept of depth of a data point, namely, Oja median (Oja (1983)), Tukey's halfspace median (Tukey (1975)), and Liu's simplicial median (Liu (1988)).

Univariate quantiles are obtained by inverting a cumulative distribution function (CDF), but such an approach is difficult in the multivariate situation because a multivariate CDF has multiple arguments. The most immediate version of a multivariate quantile is the

vector of coordinate-wise quantiles, which has similar properties as the coordinate-wise median. The most well-known version is called the geometric quantile (Chaudhuri (1996)), which is a generalization of the spatial median. Here we study Bayesian inferencial methods on geometric quantiles. The above versions of quantiles can be generalized to formulate multivariate quantile regression models as well.

The Bayesian literature on inference of multivariate quantiles is extremely limited. A Bayesian approach is attractive because it gives an immediate visual summary and uncertainty quantification, in the form of the posterior distribution. A major part of this thesis explores a non-parametric Bayesian approach, i.e., we consider the underlying random distribution with a Dirichlet Process (DP) prior. The posterior distribution on the quantile functional is induced from the posterior distribution of the random distribution, which also follows a Dirichlet process. The Dirichlet process is a conjugate prior and hence, the posterior for the quantiles is easy to sample from. The posterior is shown to have good asymptotic properties, i.e., the $\sqrt{n}$-convergence rate and the Bernstein-von Mises property. This approach is next extended to a regression setting, i.e., the collection of dependent conditional distributions of a k-variate response Y given a univariate predictor X is given a Dependent Dirichlet Process (DDP) prior, which is an extension of the DP prior. We also construct non-parametric Bayesian testing procedures for multivariate locations, i.e., we construct posterior credible sets using the DP model, and make our decision based on whether the credible sets contain the null value. The tests are asymptotically non-parametric and have been shown to perform good.

Chapter 3 of this thesis deals with a Gibbs posterior construction for multivariate quantiles. The Gibbs posterior has received considerable attention in the literature and is free from model misspecification issues encountered in parametric Bayesian methods. The theoretical properties of the Gibbs posterior have also been investigated. We also discuss how to scale the Gibbs posterior so that it achieves the nominal frequentist coverage. Overall, this thesis explores various theoretical and methodological aspects of Bayesian and Gibbs posterior constructions, and also confirms the theoretical findings through simulation study and real data applications. A detailed outline of the thesis is given in the next subsection.

1.1 Outline

Chapter 2 begins with a survey of multidimensional medians and quantiles that have been studied in the literature over the years. Then, we introduce our non-parametric Bayesian modeling framework for the estimation of a multivariate ℓ_1-median. Unfortunately, the

exact distribution of a multivariate median cannot be derived analytically, so we study the distributions asymptotically. We approximate the posterior Dirichlet process by a Bayesian Bootstrap process, and we study the asymptotic behavior of the posterior distribution of a multivariate ℓ_1-median functional using the theory of exchangeable bootstrap, which includes the Bayesian bootstrap as a special case. The asymptotic posterior distribution has been shown to be the same with the asymptotic distribution of the sample ℓ_1-median, thus proving a Bernstein-von Mises (BvM) theorem for the ℓ_1-median functional. The ℓ_1-median is not equivariant under arbitrary affine transformation, and there are transformation techniques available to make that affine equivariant. The most popular data-driven technique is called the transformation-and-retransformation technique (Chakraborty, Chaudhuri and Oja (1998)). We study this median in a non-parametric Bayesian framework, and show that the transformed-and retransformed median also enjoys a Bernstein-von Mises approximation properly. Next, we extend the theoretical results to a set of finitely many geometric quantiles, i.e., we show that the joint asymptotic posterior distribution of finitely many geometric quantiles can also be approximated by a Gaussian distribution in the Bernstein-von Mises sense. Finally, we carry out comprehensive numerical studies to demonstrate frequentist coverage of the posterior credible sets in finite sample situations. This chapter is based on Bhattacharya and Ghosal (2019).

We start Chapter 3 by giving a brief overview of the Gibbs posteriors, and explaining the advantages it offers over parametric likelihood-based methods. Then we show that a Gibbs posterior construction for a geometric quantile is immediate, as the quantiles are formulated as mimimizers of ℓ_1-objective functions. We derive the convergence rate and a Bernstein-von Mises theorem. Unfortunately, there is a "covariance mismatch" issue in the BvM theorem, i.e., the asymptotic covariance matrix in the Gaussian approximation of the posterior asymptotic distribution is not the same as that of the sample median. This results in the Gibbs posterior credible sets not having the nominal coverage. This issue is similar to what one encounters while constructing a posterior of a parameter in misspecified locally asymptotic normal models (see Kleijn and van der Vaart (2012)). The Gibbs posterior can be scaled with a user specific "learning rate" to "correct" this mismatch, and calibrate the credible set. We choose a bootstrap-based calibration algorithm (Syring and Martin (2019)) to choose the learning rate, and show that the scaled Gibbs posterior performs reasonably well. The materials in this chapter are based on Bhattacharya and Martin (2020).

In Chapter 4, we develop testing procedures for one-sample and two-sample multivariate location problems in a non-parametric Bayes framework. The most popular testing method for multivariate locations is the Hotelling's T^2-test, and more generally, the MANOVA (Multivariate Analysis of Variance), which is based on the assumption that

the underlying distributions are multivariate Gaussian. In the univariate case, some non-parametric testing methods are the sign test, Wilcoxon rank-sum test, Wilcoxon signed-rank test, etc. These tests have been generalized to higher dimensions, and the multivariate methods are called spatial sign test, spatial rank test and spatial signed-rank test (Oja and Randles (2004)). All these tests are asymptotically distribution-free. The corresponding test statistics are asymptotically Gaussian, and an asymptotic p-value can be constructed based on a chi-square distribution. For our Bayesian one-sample test, we put a Dirichlet process prior on the underlying distribution and form a posterior credible set for the spatial median, and we accept the null hypothesis if the credible set contains the null value. For the two-sample problem, we construct a credible set for the difference of the spatial medians, and we accept the null hypothesis of equality of the medians if the credible set contains the value zero. The advantage of using these Bayesian tests is that the testing procedure can be constructed directly from the posterior samples, and we do not have to make a decision based on a large-sample approximation. We demonstrate the performance of our tests in finite samples, and they perform better than the existing procedures, especially for heavy-tailed and non-symmetric distributions.

Finally in Chapter 5, we discuss a non-parametric Bayesian framework for multivariate quantile regression. Quantile regression is used as a way to discover predictive relationships between variables when there is little to no relationship between means of the variables. We begin Chapter 5 by reviewing the existing methods for univariate and multiple regression, and then describe the usefulness of the DDP prior in modelling a collection of dependent random measures. Next, we describe posterior computation in the model. We use a Block Gibbs sampler and describe the full posterior conditional distributions. We also derive a posterior consistency theorem for geometric quantiles of a smoothed conditional distribution of Y given $X = x$. First, we show that for every $x \in \mathfrak{X}$, the smoothed posterior distribution is weakly consistent at its truth, and next, we show the corresponding quantiles are close. We illustrate our method on simulated and real datasets, and our method seems to perform well, and produces better results than frequentist parametric methods, especially when the response is non-Gaussian.

CHAPTER

2

NONPARAMETRIC BAYESIAN INFERENCE ON MULTIVARIATE MEDIANS AND QUANTILES

2.1 Introduction

It is well known that for real valued data, the median is a more robust measure of location than mean. Similarly, in multivariate analysis, there are situations where the multivariate mean vector is not a good measure of location— for example, when the dataset has a wide spread, outliers etc., a multivariate median would be a much more robust measure. There is no universally accepted definition of a multivariate median, because there is no objective basis of ordering the data points in higher dimensions. Over the years, various definitions of multivariate medians and, more generally, multivariate quantiles have been proposed; see Small (1990) for a comprehensive review on multivariate medians.

One of the most popular versions of multivariate median is called the multivariate ℓ_1-median. For a set of sample points $X_1, X_2, \ldots, X_n \in \mathbb{R}^k$, $k \geq 2$, the sample ℓ_1-median is obtained by minimizing $n^{-1}\sum_{i=1}^{n} \|X_i - \theta\|$ with respect to θ, where $\|\cdot\|$ denotes some norm.

The most commonly used norm is the ℓ_p-norm $\|x\|_p = \left(\sum_{j=1}^k |x_j|^p\right)^{1/p}$, $1 \le p \le \infty$. The most popular version of the ℓ_1-median that uses the usual Euclidean norm $\|x\|_2 = \left(\sum_{j=1}^k x_j^2\right)^{1/2}$ is known as the spatial median. This corresponds to $p = 2$. Clearly the case $p = 1$ gives rise to the vector of coordinatewise medians. The sample ℓ_1-median with ℓ_p-norm is given by

$$\hat{\theta}_{n;p} = \underset{\theta \in \mathbb{R}^k}{\arg\min} \frac{1}{n} \sum_{i=1}^n \|X_i - \theta\|_p. \tag{2.1.1}$$

The spatial median has been widely studied in the literature. It is a highly robust estimator of the location and its breakdown point is 1/2 which is as high as that of the coordinatewise median (Lopuhaa and Rousseeuw (1991)). The ℓ_1-median functional of a probability distribution P based on the ℓ_p-norm is given by

$$\theta_p(P) = \underset{\theta}{\arg\min}\, P\left(\|X - \theta\|_p - \|X\|_p\right), \tag{2.1.2}$$

for $Pf = \int f \, dP$ and $1 \le p \le \infty$. It may be noted that this definition does not require any moment assumption on X, since $|\|X - \theta\|_p - \|X\|_p| \le \|\theta\|_p$. Henceforth, we fix $1 < p < \infty$ and drop p from the notations $\hat{\theta}_{n;p}$ and $\theta_p(P)$ and just write $\hat{\theta}_n$ and $\theta(P)$ respectively.

In statistical applications, the distribution P is unknown. An obvious strategy to estimate $\theta(P)$ is to replace P by the empirical measure $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$, where δ_x denotes the point-mass distribution at x, which gives rise to the sample ℓ_1-median in (2.1.1). The usual method for making inference on multivariate medians is the M-estimation framework, i.e., the median is estimated by minimizing a data-driven objective function, as in (2.1.1). Asymptotic distributional results for M-estimators can be used to construct confidence regions.

A Bayesian approach gives a nice visual summary of uncertainty, and the posterior credible regions can be directly used for inference, without any asymptotic approximations being required. Here we take a nonparametric Bayesian approach. We let the random distribution P be completely flexible and treat $\theta(P)$ as a functional of P. The most commonly used prior on an arbitrary distribution P is the Dirichlet process prior which we discuss in Section 2.2. In the univariate case, the exact posterior distribution of the median functional can be derived explicitly (see Chapter 4, Ghosal and van der Vaart (2017) for more details). Unfortunately, in the multivariate case, the exact posterior distribution can only be computed by simulations. The posterior distribution can be used to compute point estimates and credible sets. It is of interest to study the frequentist accuracy of the Bayesian estimator and frequentist coverage of posterior credible regions. In the parametric con-

text, the Bernstein-von Mises theorem ensures that the Bayes estimator converges at the parametric rate $n^{-1/2}$ and a Bayesian $100(1-\alpha)\%$ credible set has asymptotic frequentist coverage $(1-\alpha)$. Interestingly, a functional version of the Bernstein-von Mises theorem holds for the distribution under the Dirichlet process prior as shown by Lo (1983, 1986). A functional Bernstein-von Mises theorem can potentially establish Bernstein-von Mises theorem for certain functionals. We study posterior concentration properties of the multivariate ℓ_1-median $\theta(P)$ and show that the posterior distribution of $\theta(P)$ centered at the sample ℓ_1-median $\hat{\theta}_n$ is asymptotically normal. We also note that this asymptotic distribution matches with the asymptotic distribution of $\hat{\theta}_n$ centered at the true value $\theta^\star \equiv \theta(P^\star)$, where $P^\star$ is the truth of P, thus proving a Bernstein-von Mises theorem for the multivariate ℓ_1-median.

One possible shortcoming of the multivariate ℓ_1-median is that it lacks equivariance under affine transformation of the data. Chakraborty, Chaudhuri and Oja (1998) proposed an affine-equivariant modification of the sample spatial median using a data-driven transformation and re-transformation technique. There is no population analog of this modified median. We define a Bayesian analog of this modified ℓ_1-median in the following way. We put a Dirichlet process prior on the distribution of a transformed data depending on the observed data and induce the posterior distribution on $\theta(P)$ to make its distribution translation equivariant. We show that the asymptotic posterior distribution of $\theta(P)$ thus obtained centered at the affine-equivariant multivariate median estimate matches with the asymptotic distribution of the latter centered at $\theta^\star$, while both the limiting distributions are normal.

As we pointed out before, the lack of an objective basis of ordering observations in higher dimensions makes it harder to define a multivariate quantile as well. The most common version of a multivariate quantile is the co-ordinatewise quantile (see Abdous and Theodorescu (1992), Babu and Rao (1989)). As Chaudhuri (1996) pointed out, the co-ordinatewise quantiles lack some useful geometric properties (e.g., rotational invariance).

Chaudhuri (1996) introduced the notion of geometric quantile based on geometric configuration of multivariate data clouds. These quantiles are natural generalizations of the spatial median. For the univariate case it is easy to see that for $X_1, \ldots, X_n \in \mathbb{R}$ and $u = 2\alpha - 1$, the sample α-quantile $\hat{Q}_n(u)$ is obtained by minimizing $\sum_{i=1}^n \{|X_i - \xi| + u(X_i - \xi)\}$ with respect to ξ. Chaudhuri (1996) extended this idea and indexed the k-variate quantiles by points in the open unit ball $B_2^{(k)} = \{u : u \in \mathbb{R}^k, \|u\|_2 < 1\}$. For any $u \in B_2^{(k)}$, Chaudhuri (1996) obtained the sample geometric u-quantile by minimizing $\sum_{i=1}^n \{\|X_i - \xi\|_2 + \langle u, X_i - \xi\rangle\}$ with respect to ξ. Generalizing Chaudhuri's (1996) definition of multivariate quantile based on the ℓ_2-norm to the ℓ_p-norm with $1 < p < \infty$, we define the multivariate sample quantile

process as

$$\hat{Q}_n(u) = \arg\min_{\xi \in \mathbb{R}^k} \frac{1}{n} \sum_{i=1}^{n} \Phi_p(u, X_i - \xi), \tag{2.1.3}$$

where $\Phi_p(u,t) = \|t\|_p + \langle u, t \rangle$ with $u \in B_q^{(k)} = \{u : u \in \mathbb{R}^k, \|u\|_q < 1\}$ and q is the conjugate index of p, i.e., $p^{-1}+q^{-1} = 1$. It is easy to see that $\hat{Q}_n(0)$ coincides with the sample multivariate ℓ_1-median $\hat{\theta}_n$. Similarly, for $u \in B_q^{(k)}$, the multivariate quantile process of a probability measure P is given by

$$Q_P(u) = \arg\min_{\xi \in \mathbb{R}^k} P\{\Phi_p(u, X - \xi) - \Phi_p(u, X)\}. \tag{2.1.4}$$

with $Q^\star(u) \equiv Q_{P^\star}(u)$ being the multivariate quantile function for the true distribution $P^\star$.

The geometric features and the asymptotic properties of geometric quantiles have been investigated in the literature (See Chaudhuri (1996)). Here, we study geometric quantiles in the previously discussed non-parametric Bayes framework and study the posterior distributions asymptotically. We prove that, with P having a Dirichlet process prior and for finitely many $u_1, \ldots, u_m$, the joint distribution of $\{\sqrt{n}(Q_P(u_1) - \hat{Q}_n(u_1)), \ldots, \sqrt{n}(Q_P(u_m) - \hat{Q}_n(u_m))\}$ given the data, converges to a multivariate normal distribution. Moreover, it is also shown that the joint distribution of $\{\sqrt{n}(\hat{Q}_n(u_1) - Q^\star(u_1)), \ldots, \sqrt{n}(\hat{Q}_n(u_m) - Q^\star(u_m))\}$ converges to the same multivariate normal distribution. Thus, we prove a Bernstein-von Mises theorem for any finite set of geometric quantiles.

The rest of this chapter is organized as follows. In Section 2.2, we give the background needed to introduce the main results. In Section 2.3, we state the BvM theorem for the multivariate ℓ_1-median and the theorems we need to prove it. In Sections 2.4 and 2.5, we present BvM theorems for the affine-equivariant ℓ_1-median and multivariate quantiles, respectively. In Section 2.6, we investigate the finite sample performance of our approach through a simulation study and an analysis of Fisher's iris data. A few concluding remarks are given in Section 2.7 and all the proofs are given in Section 2.8.

2.2 Background and Preliminaries

Before giving the background, we introduce some notations that we follow in this chapter. Throughout this thesis, $\mathrm{N}_k(\mu, \Sigma)$ denotes a k-variate multivariate normal distribution with mean vector μ and covariance matrix Σ, and $\mathrm{Gamma}_k(s, r, V)$ denotes a k-dimensional gamma distribution with shape parameter s, rate parameter r and correlation matrix V, constructed using a Gaussian copula (Xue-Kun Song (2000)). To describe the construction briefly, let $Y_1, \ldots, Y_k$ be k many univariate gamma distributions $\mathrm{Ga}(s, r)$ with cumulative

distribution functions (CDF) and probability density functions (PDF) being denoted by F_j and f_j, $j = 1, \ldots, k$. Then the joint density of $Y = (Y_1, \ldots, Y_k)$ is then given by

$$g(y, s, r, V) = c_\phi\{F_1, \ldots, F_k | V\} \prod_{j=1}^{k} f_j(y_j, s, r),$$

where $c_\phi(\cdot|V)$ denotes the density of the k-dimensional Gaussian copula. Also, DP(α) denotes a Dirichlet process with centering measure α (see Chapter 4, Ghosal and van der Vaart 2017 for more details).

Let $\rightsquigarrow$ and $\xrightarrow{P}$ denote weak convergence i.e. convergence in distribution and convergence in probability respectively. For a sequence X_n, the notation $X_n = O_P(a_n)$ means that X_n/a_n is stochastically bounded. Also, $\|P - Q\|_{TV}$ denotes the total variation distance $\sup_A |P(A) - Q(A)|$ between measures P and Q. Moreover, diag$(a_1, \ldots, a_n)$ denotes a diagonal matrix with diagonal elements $a_1, \ldots, a_n$, and sign$(\cdot)$ denotes the signum fucntion

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ -1 & \text{if } x < 0. \end{cases}$$

Finally, 0_k denotes a vector of all 0's of length k, $\mathbf{1}_k$ denotes a vector of all 1's of length k, and I_k denotes an identity matrix of order $k \times k$.

Let $X_i \in \mathbb{R}^k$, $i = 1, \ldots, n$, be independently and identically distributed (i.i.d.) observations from a k-variate distribution P and let P have the DP(α) prior. The parameter space is taken to be $\mathbb{R}^k$. The Bayesian model is then formulated as

$$X_1, X_2, \ldots, X_n | P \overset{\text{iid}}{\sim} P, \quad P \sim \text{DP}(\alpha). \tag{2.2.1}$$

The posterior distribution of P given $X_1, X_2, \ldots, X_n$ is DP($\alpha + n\mathbb{P}_n$), (see Theorem 4.6, Ghosal and van der Vaart (2017) for more details). As stated in Ghosal and van der Vaart (2017), $\sqrt{n}(P - \mathbb{P}_n)$ with $P \sim \text{DP}(\alpha + n\mathbb{P}_n)$ converges conditionally in distribution to a Brownian bridge process. But this result cannot be used to find the posterior asymptotic distribution of $\theta(P)$, because $\theta(P)$ is not a smooth functional of P. To deal with this, we use the following fact stated in Chapter 12, Ghosal and van der Vaart (2017). The posterior distribution DP($\alpha + n\mathbb{P}_n$) can be expressed as $V_n Q + (1 - V_n)\mathbb{B}_n$, where the processes $Q \sim \text{DP}(\alpha)$, $\mathbb{B}_n \sim \text{DP}(n\mathbb{P}_n)$ and $V_n \sim \text{Be}(|\alpha|, n)$ are independent and Be(a, b) denotes a beta distribution with parameters a and b. The process $\mathbb{B}_n$ is also known as the Bayesian bootstrap distribution and can be defined by the linear operator $\mathbb{B}_n f = \sum_{i=1}^n B_{ni} f(X_i)$, where $(B_{n1}, B_{n2}, \ldots, B_{nn})$ is

a random vector following the Dirichlet distribution Dir$(n;1,1,\ldots,1)$. We approximate the posterior Dirichlet process by the Bayesian bootstrap process and show that given $X_1,\ldots,X_n$, the posterior distribution of $\sqrt{n}(\theta(P)-\hat{\theta}_n)$ is asymptotically the same as the conditional distribution of $\sqrt{n}(\theta(\mathbb{B}_n)-\hat{\theta}_n)$ (Lemma 2.8.1), where $\theta(\mathbb{B}_n)=\arg\min_{\theta\in\mathbb{R}^k}\mathbb{B}_n\|X-\theta\|_p$.

With the approximation in Lemma 2.8.1, we are just left to show that given $X_1,\ldots,X_n$, $\sqrt{n}(\theta(\mathbb{B}_n)-\hat{\theta}_n)$ is asymptotically normal. In order to show that, we use the fact that $\hat{\theta}_n$ can be viewed as a Z-estimator (van der Vaart (1995)), because it satisfies the system of equations $\Psi_n(\theta)=\mathbb{P}_n\psi(\cdot,\theta)=0$, where $\psi(\cdot,\theta)=(\psi_1(\cdot,\theta),\ldots,\psi_k(\cdot,\theta))^T$ is a $k\times 1$ vector of functions from $\mathbb{R}^k\times\mathbb{R}^k$ to $\mathbb{R}$ with

$$\psi_j(x,\theta)=\frac{|x_j-\theta_j|^{p-1}}{\|x-\theta\|_p^{p-1}}\operatorname{sign}(\theta_j-x_j),\quad j=1,\ldots,k, \tag{2.2.2}$$

In addition, we view $\theta(\mathbb{B}_n)$ as a bootstrapped analog of the Z-estimator $\hat{\theta}_n$ (more details are given in Subsection 2.3.1). Next, we use the asymptotic theory of Z-estimators to find the asymptotic distributions of $\hat{\theta}_n$ and $\theta(\mathbb{B}_n)$. In the next section, we state the Bernstein-von Mises theorem for the ℓ_1-median, and discuss how to derive it with the help of the asymptotic theory of Z-estimators..

2.3 Bernstein-von Mises theorem for ℓ_1-median

Before stating the theorem, we introduce a few more notations that will be used in the theorem. Define $V_{\theta^\star}=\int\dot{\psi}_{x,\theta^\star}\,dP^\star$, where

$$\dot{\psi}_{x,\theta^\star}=\Big[\frac{\partial\psi(x,\theta)}{\partial\theta}\Big]_{\theta=\theta^\star}.$$

The matrix $\dot{\psi}_{x,\theta^\star}$ is given by

$$\dot{\psi}_{x,\theta^\star}=\frac{p-1}{\|x-\theta^\star\|_p}\Big[\operatorname{diag}\left(\frac{|x_1-\theta_1^\star|^{p-2}}{\|x-\theta^\star\|_p^{p-2}},\ldots,\frac{|x_k-\theta_k^\star|^{p-2}}{\|x-\theta^\star\|_p^{p-2}}\right)-\psi(x,\theta^\star)\psi(x,\theta^\star)^T\Big]. \tag{2.3.1}$$

Also, we denote $U_{\theta^\star}=P^\star(\psi(\cdot,\theta^\star)\psi(\cdot,\theta^\star)^T$.

Theorem 2.3.1. *Let $p\geq 2$ be a fixed integer. Suppose that the following conditions hold for $k\geq 2$.*

C1. *The true probability distribution of $X\in\mathbb{R}^k$, $P^\star$ has a probability density that is bounded on compact subsets of $\mathbb{R}^k$.*

C2. *The ℓ_1-median of $P^\star$, given by $\theta^\star = \theta(P^\star)$, is unique.*

Then

(i) $\sqrt{n}(\hat{\theta}_n - \theta^\star) \rightsquigarrow \mathrm{N}_k(0, V_{\theta^\star}^{-1} U_{\theta^\star} V_{\theta^\star}^{-1})$,

(ii) *given $X_1, \ldots, X_n$, $\sqrt{n}(\theta(P) - \hat{\theta}_n) \rightsquigarrow \mathrm{N}_k(0, V_{\theta^\star}^{-1} U_{\theta^\star} V_{\theta^\star}^{-1})$ in $P^{\star n}$-probability.*

Further if $k = 2$, (i) *and* (ii) *hold for any $1 < p < \infty$.*

The uniqueness holds unless $P^\star$ is completely supported on a straight line in $\mathbb{R}^k$, for $k \geq 2$ (section 3, Chaudhuri (1996)). As we have pointed out before, finding the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta^\star)$ can be viewed as an application of the problem of finding the asymptotic distribution of a Z-estimator centered at its true value. The asymptotic theory of the Z-estimators has been studied extensively in the literature. Huber (1967) proved the asymptotic normality of Z-estimators when the parameter space is finite-dimensional. Van der Vaart (1995) extended Huber's (1967) theorem to the infinite-dimensional case.

We mentioned that $\theta(\mathbb{B}_n)$ is a bootstrapped version of the estimator $\hat{\theta}_n$, where the bootstrap weights are drawn from a $\mathrm{Dir}(n; 1, 1, \ldots, 1)$ distribution. In other words, $\theta(\mathbb{B}_n)$ satisfies the system of equations $\hat{\Psi}_n(\theta) = \mathbb{B}_n \psi(\cdot, \theta) = 0$. Wellner and Zhan (1996) extended van der Vaart's (1995) infinite-dimensional Z-estimator theorem by showing that for any exchangeable vector of non-negative bootstrap weights, the bootstrap analog of a Z-estimator conditional on the observations is also asymptotically normal. We use Wellner and Zhan's (1996) theorem to prove the asymptotic normality of $\theta(\mathbb{B}_n)$. Wellner and Zhan's (1996) theorem ensures that both $\sqrt{n}(\hat{\theta}_n - \theta^\star)$, and $\sqrt{n}(\theta(\mathbb{B}_n) - \hat{\theta}_n)$ given the data, converge in distribution to the same normal limit, which, together with Lemma 2.8.1 proves Theorem 2.3.1. In Section 2.8, we provide a detailed verification of the conditions of Wellner and Zhan's (1996) theorem in our situation.

2.3.1 Bootstrapping a Z-estimator

In this subsection, we state Wellner and Zhan's (1996) bootstrap theorem for Z-estimators. Let $W_n = (W_{n1}, W_{n2}, \ldots, W_{nn})$ be a set of bootstrap weights. The bootstrap empirical measure is defined as $\hat{\mathbb{P}}_n = n^{-1} \sum_{i=1}^n W_{ni} \delta_{X_i}$. Wellner and Zhan (1996) assumed that the bootstrap weights $W = \{W_{ni},\ i = 1, 2, \ldots, n,\ n = 1, 2, \ldots\}$ form a triangular array defined on a probability space $(\mathfrak{Z}, \mathscr{E}, \hat{P})$. Thus $\hat{P}$ refers to the distribution of the bootstrap weights. According to Wellner and Zhan (1996), the following conditions are imposed on the bootstrap weights:

(i) The vectors $W_n = (W_{n1}, W_{n2}, \ldots, W_{nn})^T$ are exchangeable for every n, i.e., for any permutation $\pi = (\pi_1, \ldots, \pi_n)$ of $\{1, 2, \ldots, n\}$, the joint distribution of the permuted weights $\pi(W_n) = (W_{n\pi_1}, W_{n\pi_2}, \ldots, W_{n\pi_n})^T$ is same as that of W_n.

(ii) The weights $W_{ni} \geq 0$ for every $n,\ i$ and $\sum_{i=1}^{n} W_{ni} = n$ for all n.

(iii) The $L_{2,1}$ norm of W_{n1} is uniformly bounded: for some $0 < K < \infty$

$$\|W_{n1}\|_{2,1} = \int_0^\infty \sqrt{\hat{P}(W_{n1} \geq u)}\, \mathrm{d}u \leq K. \tag{2.3.2}$$

(iv) $\lim_{\lambda\to\infty} \limsup_{n\to\infty} \sup_{t\geq\lambda}(t^2 \hat{P}\{W_{n1} \geq t)\} = 0$.

(v) $n^{-1}\sum_{i=1}^{n}(W_{ni}-1)^2 \to c^2 > 0$ in $\hat{P}$-probability for some constant $c > 0$.

Van der Vaart and Wellner (1995) noted that if $Y_1, \ldots, Y_n$ are exponential random variables with mean 1, then the weights $W_{ni} = Y_i/\bar{Y}_n,\ i = 1, \ldots, n$, satisfy conditions (i)–(v), resulting in the Bayesian bootstrap scheme with $c = 1$ because the left hand side in (v) is given by $n^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2/\bar{Y}_n^2 \xrightarrow{P} \mathrm{Var}(Y)/\{\mathrm{E}(Y)\}^2 = 1$. To apply the bootstrap theorem, we also need to assume that the function class

$$\mathscr{F}_R = \{\psi_j(\cdot,\theta) : \|\theta - \theta^\star\|_2 \leq R,\ j = 1, 2, \ldots, k\} \tag{2.3.3}$$

has "enough measurability" for randomization with independently and identically distributed multipliers to be possible and Fubini's theorem can be used freely. We call a function class $\mathscr{F} \in \mathfrak{m}(P)$ if $\mathscr{F}$ is countable, or if the empirical process $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$ is stochastically separable (the definition of a separable stochastic process is provided in Section 2.8), or $\mathscr{F}$ is image admissible Suslin (see Chapter 5, Dudly (2014) for a definition). Now we formally state Wellner and Zhan's (1996) theorem for a sequence of consistent asymptotic bootstrap Z-estimators $\hat{\theta}_n$ of $\theta \in \mathbb{R}^k$, which satisfies the system of equations $\hat{\Psi}_n(\theta) = \hat{\mathbb{P}}_n\psi(\cdot,\theta) = \sum_{i=1}^{n} W_{ni}\psi(X_i,\theta) = 0$.

Theorem 2.3.2 (Wellner and Zhan)**.** *Assume that the class of functions $\mathscr{F} \in \mathfrak{m}(P^\star)$ and the following conditions hold.*

1. *There exists a $\theta^\star \equiv \theta(P^\star)$ such that*

$$\Psi(\theta^\star) = P^\star\psi(X,\theta^\star) = 0. \tag{2.3.4}$$

The function $\Psi(\theta) = P^\star\psi(X,\theta)$ is differentiable at $\theta^\star$ with nonsingular derivative matrix $\dot{\Psi}_0$:

$$\dot{\Psi}_0 = \Big[\frac{\partial\Psi}{\partial\theta}\Big]_{\theta=\theta^\star}. \tag{2.3.5}$$

2. *For any* $\delta_n \to 0$,

$$\sup\Big\{\frac{\|\mathbb{G}_n(\psi(\cdot,\theta)-\psi(\cdot,\theta^\star))\|_2}{1+\sqrt{n}\|\theta-\theta^\star\|_2} : \|\theta-\theta^\star\|_2 \le \delta_n\Big\} = o_{P^\star}(1). \tag{2.3.6}$$

3. *The* k*-vector of functions* ψ *is square-integrable at* $\theta^\star$ *with covariance matrix*

$$\Sigma_0 = P^\star\psi(X,\theta^\star)\psi^T(X,\theta^\star) < \infty. \tag{2.3.7}$$

For any $\delta_n \to 0$, *the envelope functions*

$$D_n(x) = \sup\Big\{\frac{|\psi_j(x,\theta)-\psi_j(x,\theta^\star)|}{1+\sqrt{n}\|\theta-\theta^\star\|_2} : \|\theta-\theta^\star\|_2 \le \delta_n,\ j=1,2,\ldots,k\Big\} \tag{2.3.8}$$

satisfy

$$\lim_{\lambda\to\infty}\limsup_{n\to\infty}\sup_{t\ge\lambda} t^2 P^\star(D_n(X_1) > t) = 0. \tag{2.3.9}$$

4. *The estimators* $\hat\theta_n$ *and* $\hat\theta_n$ *are consistent for* $\theta^\star$, *i.e.,* $\|\hat\theta_n-\theta^\star\|_2 \xrightarrow{P^\star} 0$ *and* $\|\hat\theta_n-\hat\theta_n\|_2 \xrightarrow{\hat P} 0$ *in* $P^{\star n}$*-probability.*

5. *The bootstrap weights satisfy conditions* (i)–(v).

Then

(i) $\sqrt{n}(\hat\theta_n-\theta^\star) \rightsquigarrow \mathrm{N}_k(0,\dot\Psi_0^{-1}\Sigma_0\dot\Psi_0^{-1})$;

(ii) $\sqrt{n}(\hat\theta_n-\hat\theta_n) \rightsquigarrow \mathrm{N}_k(0,c^2\dot\Psi_0^{-1}\Sigma_0\dot\Psi_0^{-1})$ *in* $P^{\star n}$*- probability.*

It has already been mentioned that, for the Bayesian bootstrap weights, the value of the constant c is 1. Thus if $\psi(\cdot,\theta)$ defined in (2.2.2) satisfies the conditions in Theorem 2.3.2, then Theorem 2.3.1 holds.

It may be mentioned that Cheng and Huang (2010) also studied asymptotic theory for bootstrap Z-estimators, and developed consistency and asymptotic normality results. We could have also considered an M-estimator framework and used their results to prove our theorems.

2.4 Affine-equivariant Multivariate ℓ_1-median

We start this section by describing the transformation and retransformation technique that has been used in the literature to obtain an affine equivariant version of a multivariate

median. Here we consider a nonparametric Bayesian framework for an affine equivariant version of the ℓ_1-median. Although the sample multivariate ℓ_1-median is equivariant under location transformation and orthogonal transformation of the data, it is not equivariant under arbitrary affine transformation of the data. Chakraborty and Chaudhuri (1996, 1998) used a data-driven transformation-and-retransformation technique to convert the non-equivariant coordinatewise median to an affine-equivariant one. Chakraborty, Chaudhuri and Oja (1998) applied the same idea to the sample spatial median.

We use the transformation-and-retransformation technique to construct an affine equivariant version of the multivariate ℓ_1-median. Suppose that we have n sample points $X_1, X_2, \ldots, X_n \in \mathbb{R}^k$, with $n > k+1$. We consider the points $X_{i_0}, X_{i_1}, \ldots, X_{i_k}$, where $\alpha = \{i_0, i_1, \ldots, i_k\}$ is a subset of $\{1, 2, \ldots, n\}$. The matrix $X(\alpha)$ consisting the columns $X_{i_1} - X_{i_0}, X_{i_2} - X_{i_0}, \ldots, X_{i_k} - X_{i_0}$ is the data-driven transformation matrix. The transformed data points are $Z_j^{(\alpha)} = \{X(\alpha)\}^{-1} X_j,\ j \notin \alpha$. The matrix $X(\alpha)$ is invertible with probability 1 if $X_i,\ i = 1, \ldots, n$, are i.i.d. samples from a distribution that is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^k$. The sample ℓ_1-median based on the transformed observations is then given by

$$\hat{\phi}_n^{(\alpha)} = \arg\min_{\phi} \sum_{j \notin \alpha} \| Z_j^{(\alpha)} - \phi \|_p. \tag{2.4.1}$$

We transform it back in terms of the original coordinate system as

$$\hat{\theta}_n^{(\alpha)} = X(\alpha) \hat{\phi}_n^{(\alpha)}. \tag{2.4.2}$$

It can be shown that $\hat{\theta}_n^{(\alpha)}$ is affine equivariant. Chakraborty, Chaudhuri and Oja (1998) suggested that $X(\alpha)$ should be chosen in such a way that the matrix $\{X(\alpha)\}^T \Sigma^{-1} X(\alpha)$ is as close as possible to a matrix of the form λI_k where Σ is the covariance matrix of X. Chakraborty, Chaudhuri and Oja (1998) proved that conditional on $X(\alpha)$, the asymptotic distribution of the transformed-and-retransformed spatial median is normal.

2.4.1 Bernstein-von Mises theorem for the affine-equivariant multivariate median

Here, we develop a non-parametric Bayesian framework for studying the affine-equivariant ℓ_1-median. Let $X_1, X_2, \ldots, X_n \in \mathbb{R}^k$ be a random sample from a distribution P that is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^k$. Let $X(\alpha)$ be the transformation matrix and $Z_j^{(\alpha)} = \{X(\alpha)\}^{-1} X_j,\ j \notin \alpha$, be the transformed observations. The sample median of $X_1, \ldots, X_n$ is denoted by $\hat{\theta}_n$.

Let the distribution of $Z_j^{(\alpha)},\ j \notin \alpha$, be denoted by P_Z. We equip P_Z with a DP(β) prior.

The true value of P_Z is denoted by $P_Z^\star$, i.e., the distribution of Z when $X \sim P^\star$. Hence the Bayesian model can be described as

$$Z_j^{(\alpha)} | P_Z \overset{iid}{\sim} P_Z, \quad P_Z \sim \mathrm{DP}(\beta),\ j \notin \alpha, \tag{2.4.3}$$

which implies that

$$P_Z | \{Z_j^{(\alpha)} : j \notin \alpha\} \sim \mathrm{DP}(\beta + \sum_{j \notin \alpha} \delta_{Z_j}). \tag{2.4.4}$$

Following the same arguments used in Section 2.2, we can approximate the posterior Dirichlet process P_Z by the Bayesian bootstrap process $\mathbb{B}_{n-k-1}$, since we are excluding the $(k+1)$ observations that have been used to construct the transformation matrix $X(\alpha)$. Note that, this exclusion will not have any effect on the asymptotic study. Define

$$\phi^{(\alpha)}(\mathbb{B}_n) = \arg\min_{\phi} \mathbb{B}_{n-k-1} \| Z^{(\alpha)} - \phi \|_p, \tag{2.4.5}$$

$$\phi^{(\alpha)}(P_Z) = \arg\min_{\phi} \left\{ P_Z(\| Z^{(\alpha)} - \phi \|_p - \| Z^{(\alpha)} \|_p) \right\}. \tag{2.4.6}$$

Thus the transformed-and-retransformed medians are given by

$$\hat{\theta}_n^{(\alpha)} = X(\alpha) \hat{\phi}_n^{(\alpha)}, \quad \theta^{(\alpha)}(\mathbb{B}_n) = X(\alpha) \phi^{(\alpha)}(\mathbb{B}_n). \tag{2.4.7}$$

Also define $\theta^{(\alpha)}(P) = X(\alpha)\phi^{(\alpha)}(P_Z)$. We view $\hat{\phi}_n^{(\alpha)}$ as a Z-estimator satisfying $\Psi_{Z_n}(\phi) = \mathbb{P}_n \psi_Z(\cdot, \phi) = 0$. The "population version" of $\Psi_{Z_n}(\phi)$ is denoted by $\Psi_Z(\phi) = P_Z \psi_Z(\cdot, \phi)$. The real-valued elements of the vector $\psi_Z(z, \phi)$ are then given by

$$\psi_{Z;j}(z, \phi) = \frac{|z_j - \phi_j|^{p-1}}{\| z - \phi \|_p^{p-1}} \operatorname{sign}(\phi_j - z_j), \quad j = 1, \ldots, k. \tag{2.4.8}$$

Let $\phi^{\star(\alpha)} \equiv \phi^{(\alpha)}(P_Z^\star)$ satisfy $\Psi_Z^\star(\phi^{(\alpha)}) = P_Z^\star \psi_Z(\cdot, \phi^{(\alpha)}) = 0$. In the following, we denote $V_{\phi^{\star(\alpha)}} = \left[\partial \Psi_Z^\star / \partial \phi \right]_{\phi = \phi^{\star(\alpha)}}$ and $U_{\phi^{\star(\alpha)}} = P_Z^\star \psi_Z(\cdot, \phi^{\star(\alpha)}) \psi_Z^T(\cdot, \phi^{\star(\alpha)})$. The matrix $V_{\phi^{\star(\alpha)}}$ can be written as $V_{\phi^{\star(\alpha)}} = \int \dot{\psi}_{Z, \phi^{\star(\alpha)}} \, \mathrm{d}P_Z^\star$, where

$$\dot{\psi}_{Z, \phi^{\star(\alpha)}} = \Big[\frac{\partial \psi_Z(z, \phi)}{\partial \phi} \Big]_{\phi = \phi^{\star(\alpha)}}. \tag{2.4.9}$$

The matrix $\dot{\psi}_{Z, \phi^{\star(\alpha)}}$ is given by

$$\dot{\psi}_{Z, \phi^{\star(\alpha)}} = \frac{p-1}{\| z - \phi^{\star(\alpha)} \|_p} \Big[\operatorname{diag} \left(\frac{|z_1 - \phi_1^{\star(\alpha)}|^{p-2}}{\| z - \phi^{\star(\alpha)} \|_p^{p-2}}, \ldots, \frac{|z_k - \phi_k^{\star(\alpha)}|^{p-2}}{\| z - \phi^{\star(\alpha)} \|_p^{p-2}} \right) - \psi_Z(z, \phi) \psi_Z(z, \phi)^T \Big].$$

Theorem 2.4.1. *Let $p \geq 2$ be a fixed integer. For $k \geq 2$ and a given subset $\alpha = \{i_0, i_1, \ldots, i_k\}$ of $\{1, 2, \ldots, n\}$ with size $k+1$, suppose that the following conditions hold.*

C1. *The true distribution of $Z^{(\alpha)}$, $P_Z^\star$ has a density which is bounded on compact subsets of $\mathbb{R}^k$.*

C2. *The ℓ_1-median of $P_Z^\star$, denoted by $\phi^{\star(\alpha)} = \phi^{(\alpha)}(P_Z^\star)$, is unique.*

Then

(i) $\sqrt{n}(\hat{\theta}_n^{(\alpha)} - \theta^{(\alpha)}(P^\star)|\{X_i : i \in \alpha\} \rightsquigarrow \mathrm{N}_k(0, X(\alpha)\{V_{\phi^{\star(\alpha)}} U_{\phi^{\star(\alpha)}} \{V_{\phi^{\star(\alpha)}}\}^{-1}\ \{X(\alpha)\}^T)$;

(ii) *given* $X_1, \ldots, X_n$, $\sqrt{n}(\theta^{(\alpha)}(P) - \hat{\theta}_n^{(\alpha)}) \rightsquigarrow \mathrm{N}_k(0, X(\alpha)\{V_{\phi^{\star(\alpha)}}\}^{-1} U_{\phi^{\star(\alpha)}} \{V_{\phi^{\star(\alpha)}}\}^{-1}\ \{X(\alpha)\}^T)$ *in* $P^{\star n}$*-probability.*

Further if $k = 2$, (i) *and* (ii) *hold for any $1 < p < \infty$.*

The uniqueness holds unless $P_Z^\star$ is completely supported on a straight line in $\mathbb{R}^k$, for $k \geq 2$, (Section 3, Chaudhuri 1996). It can be noted that the $\mathrm{DP}(\beta)$ prior on P_Z induces the $\mathrm{DP}(\beta \circ \psi^{-1})$ prior on $P \equiv P_Z \circ \psi^{-1}$, where $\psi(Y) = X(\alpha)Y$ with $Y \in \mathbb{R}^k$. Then the proof of the preceding theorem directly follows from Theorem 2.3.1. Apart from Theorem 2.3.1, this theorem uses the affine equivariance of the normal family: if a random vector $X \sim \mathrm{N}(\mu, \Sigma)$, then $Y = AX + b \sim \mathrm{N}(A\mu + b,\ A\Sigma A^T)$.

2.5 Bernstein-von Mises theorem for multivariate quantiles

The asymptotic results for the multivariate ℓ_1-medians almost directly translate to multivariate quantiles. Let X_i, $i = 1, \ldots, n$, be i.i.d. observations from a k-variate distribution P on $\mathbb{R}^k$ and P is given the $\mathrm{DP}(\alpha)$ prior. We study the posterior distributions asymptotically, and for every fixed $u_1, \ldots, u_m \in B_q^{(k)}$, Theorem 2.5.1 gives the joint posterior asymptotic distribution of the centered quantiles $\{\sqrt{n}(Q_P(u_1) - \hat{Q}_n(u_1)), \ldots, \sqrt{n}(Q_P(u_m) - \hat{Q}_n(u_m))\}$.

Firstly we introduce some notations. For each u, the sample u-quantile is viewed as a Z-estimator that satisfies the system of equations $\Psi_n^{(u)}(\xi) = \mathbb{P}_n \psi^{(u)}(\cdot, \xi) = 0$. We denote the population version of $\Psi_n^{(u)}(\xi)$ by $\Psi^{(u)}(\xi) = P\psi^{(u)}(\cdot, \xi)$. The true value of $Q_P(u)$ is denoted by $Q^\star(u) \equiv Q_{P^\star}(u)$ and it satisfies the system of equations $\Psi^{\star(u)}(\cdot, \xi) = P^\star \psi^{(u)}(\cdot, \xi) = 0$. The real-valued components of $\psi^{(u)}(\cdot, \xi)$ are then given by

$$\psi_j^{(u)}(x, \xi) = \frac{|x_j - \xi_j|^{p-1}}{\|x - \xi\|_p^{p-1}} \operatorname{sign}(\xi_j - x_j) + u_j, \quad j = 1, \ldots, k. \tag{2.5.1}$$

Define $V_{Q^\star(u)} = \int \dot{\psi}^{(u)}_{x,Q^\star(u)}\, dP^\star$, where

$$\dot{\psi}^{(u)}_{x,Q^\star(u)} = \Big[\frac{\partial \psi^{(u)}(x,\xi)}{\partial \xi}\Big]_{\xi=Q^\star(u)}. \tag{2.5.2}$$

The matrix $\dot{\psi}^{(u)}_{x,Q^\star(u)}$ is given by

$$\dot{\psi}^{(u)}_{x,Q^\star(u)} = \frac{p-1}{\|x-Q^\star(u)\|_p}\Big[\operatorname{diag}\left(\frac{|x_1-Q_1^\star(u)|^{p-2}}{\|x-Q^\star(u)\|_p^{p-2}},\ldots,\frac{|x_k-Q_k^\star(u)|^{p-2}}{\|x-Q^\star(u)\|_p^{p-2}}\right) - \psi^{(u)}(x,Q^\star(u))\psi^{(u)}(x,Q^\star(u))^T\Big], \tag{2.5.3}$$

In the above, $Q_j^\star(u)$, $j=1,\ldots,k$ denotes the jth component of the vector $Q^\star(u)$. We also define $U_{Q^\star(u),\,Q^\star(v)} = P^\star\psi^{(u)}(x,Q^\star(u))\{\psi^{(v)}(x,Q^\star(v))\}^T$.

Theorem 2.5.1. *Let $p \geq 2$ be a fixed integer. Suppose that the following conditions hold for $k \geq 2$.*

C1. *The true distribution of X, $P^\star$ has a density that is bounded on compact subsets of $\mathbb{R}^k$.*

C2. *For every $u_1,\ldots,u_m \in B_q^{(k)}$, the $u_1,\ldots,u_m$-quantiles of $P^\star$, denoted by $Q^\star(u_1),\ldots,Q^\star(u_m)$, are unique.*

Then

(i) *the joint distribution of $\big(\sqrt{n}(\hat{Q}_n(u_1)-Q^\star(u_1)),\ldots,\sqrt{n}(\hat{Q}_n(u_m)-Q^\star(u_m))$ converges to a km-dimensional normal distribution with mean zero, and the (j,l)th block of the covariance matrix is given by $V^{-1}_{Q^\star(u_j)}U_{Q^\star(u_j),\,Q^\star(u_l)}V^{-1}_{Q^\star(u_l)}$, $1 \leq j,\ l \leq m$;*

(ii) *given $X_1,\ldots,X_n$, the posterior joint distribution of $\{\sqrt{n}(Q_P(u_1)-\hat{Q}_n(u_1)),\ldots,\sqrt{n}(Q_P(u_m)-\hat{Q}_n(u_m))\}$ converges to km-dimensional normal distribution with mean zero, and the (j,l)th block of the covariance matrix is given by $V^{-1}_{Q^\star(u_j)}U_{Q^\star(u_j),\,Q^\star(u_l)}V^{-1}_{Q^\star(u_l)}$, $1 \leq j,\ l \leq m$.*

Further if $k=2$, (i) *and* (ii) *hold for any $1 < p < \infty$.*

2.6 Numerical Results

2.6.1 Simulation Study

Here, we demonstrate the finite sample performance of the non-parametric Bayesian credible sets for the multivariate ℓ_1-median. The data is generated from the following

mixture distribution $P^\star = 0.5\mathrm{N}_k(\mathbf{1}_k, I_k) + 0.5\mathrm{Gamma}_k(1,1,V)$, with cases $k=2$ and $k=3$, and the sample size being 100. All the diagonal elements of V have been chosen to be 1, and the off-diagonal elements are 0.7. The prior considered here is a Dirichlet process with centering measure $2 \times N_k(0_k, 10I_k)$, and a 95% credible ellipsoid is constructed as

$$\{\vartheta : (\vartheta - \bar{\theta})^\top S^{-1}(\vartheta - \bar{\theta}) \le r_{0.95}\},$$

where $\bar{\theta}$ and S are the Monte Carlo sample mean and covariance matrix respectively, and $r_{1-\gamma}$ is the $100(1-\gamma)\%$ percentile of $\{(\vartheta_b - \bar{\theta})^\top S^{-1}(\vartheta_b - \bar{\theta}),\ b = 1, \ldots, B\}$, where $\vartheta_1, \ldots, \vartheta_B$ are the posterior samples, with $B = 5000$. The coverage probability is defined as usual and, as a measure of the credible set's size, we use $r_{0.95}$. For comparison, we use a parametric Bayesian model as follows:

$$(X_1, \ldots, X_n) \mid \theta \overset{iid}{\sim} \mathrm{N}_k(\theta, \sigma^2 I_k), \quad \theta \sim \mathrm{N}_k(0_k, 10I_k), \quad \sigma^{-2} \sim \mathrm{Gamma}(1,1).$$

A simple Gibbs sampler can be used for posterior inference from the above model, and a 95% credible set is constructed in the same way. However, the above model suffers from the model misspecification bias, which our non-parametric Bayes model is free from.

For inferring about the affine equivariant median, we choose $X(\alpha)$ as suggested in Chakraborty, Chaudhuri and Oja (1998). The parametric Bayesian model gets the form

$$Z_j^{(\alpha)} | \phi \overset{iid}{\sim} N_k(\phi, \sigma^2 I_k), \quad \phi \sim \mathrm{N}_k(0_k, 10I_k), \quad \sigma^{-2} \sim \mathrm{Gamma}(1,1).$$

Table 2.1 and Table 2.2 summarize the size and coverage probability over 2000 replications for both models, for $k=2$ and 3, respectively. It can be noticed that the non-parametric Bayes method gives a smaller credible set with nominal coverage probability, thus protecting from the model misspecification bias in the paremetric Bayesian approach.

2.6.2 Application to Fisher's iris data

We also analyze Fisher's iris data which consists of three plant species, namely, Setosa, Virginica and Versicolor and four features, namely, sepal length, sepal width, petal length and petal width. The object of interest is the 4-dimensional spatial median of the above mentioned features. We take the non-parametric Bayesian approach with a $\mathrm{DP}(\alpha)$ prior with $\alpha = 2 \times \mathrm{N}_4(0_4, 10I_4)$. We construct the 95% Bayesian credible ellipsoid of the 4-dimensional spatial median and report its four principal axes in Table 2.3. Also, for the purpose of illustration, we plot 6 pairs of features for each species and the credible ellipsoids for the

Table 2.1 Estimated coverage probability, mean size of the 95% credible ellipsoids (in parentheses) of the non-affine equivariant (Non AE) and affine equivariant (AE) ℓ_1-medians for both parametric (PBayes) and non-parametric Bayes (NPBayes) models, when for $k = 2$.

	p	Coverage (Size) (NPBayes)	Coverage (Size) (PBayes)
Non AE	2	0.950 (5.94)	0.925 (6.37)
	3	0.942 (5.54)	0.925 (6.37)
AE	2	0.977 (6.09)	0.980 (6.19)
	3	0.955 (5.97)	0.979 (6.19)

Table 2.2 Estimated coverage probability, mean size of the 95% credible ellipsoids (in parentheses) of the non-affine equivariant (Non AE) and affine equivariant (AE) ℓ_1-medians for both parametric (PBayes) and non-parametric Bayes (NPBayes) models, when for $k = 3$.

	p	Coverage (Size) (NPBayes)	Coverage (Size) (PBayes)
Non AE	2	0.955 (5.81)	0.945 (5.99)
	3	0.948 (5.88)	0.945 (5.99)
AE	2	0.972 (5.91)	0.950 (6.11)
	3	0.961 (5.99)	0.950 (6.11)

corresponding two dimensional spatial medians (Figures 2.1, 2.2 and 2.3).

Table 2.3 Principal axes of 95% credible ellipsoid of spatial median for iris data

1st axis	2nd axis	3rd axis	4th axis
0.0580	0.3129	0.6747	0.6629
0.1461	0.2193	0.6143	0.7437
0.2965	0.8626	0.4089	0.0252
0.9420	0.3252	0.0081	0.0824

2.7 Concluding Remarks

• This work is the first to study the asymptotic behavior of posterior distributions of multivariate median and quantiles. Multivariate quantiles can be the object of interest in various types of study, for example, network analysis, genetic experiments and image analysis, where the datasets do not fit into well-known distributions and exhibit non-normality,

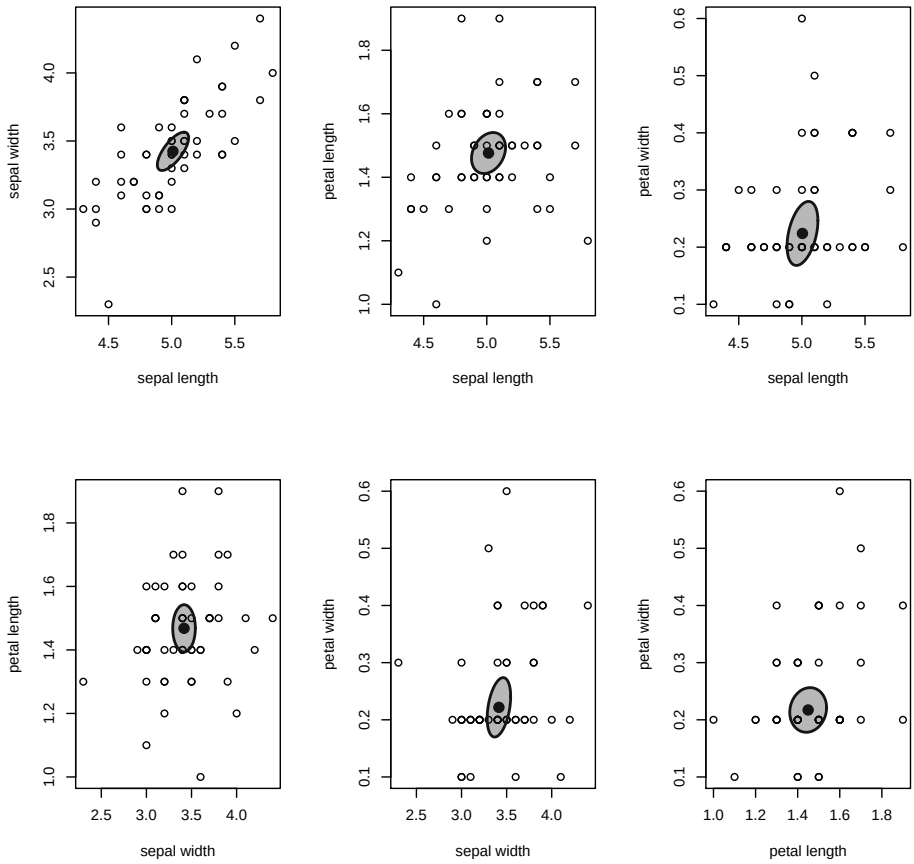


Figure 2.1 95% Credible ellipsoids of two-dimensional spatial medians for the species Setosa, in iris data

skewness and outliers. The Bayesian approach gives us automatic uncertainty quantification through the posterior distributions without requiring any large-sample approximations. The nonparametric Bayesian approach discussed here is appealing because it does not need any distributional assumptions.

- It would be interesting to explore the high dimensional setting, i.e., when $k \to \infty$. We can modify the objective function by incorporating a Lasso-like penalty. Then a k-dimensional u-quantile for $u \in B_q^{(k)}$ can be obtained by minimizing $P\{\Phi_p(u, X-\xi) - \Phi_p(u, X) + \lambda\|\xi\|_p\}$, with respect to ξ, where λ is a tuning parameter. A non-parametric Bayesian framework can be formulated by putting a Dirichlet process prior on P, and asymptotic properties of the posterior distributions can be explored as before.

- The asymptotic results for multivariate quantiles translate to multivariate L-estimates

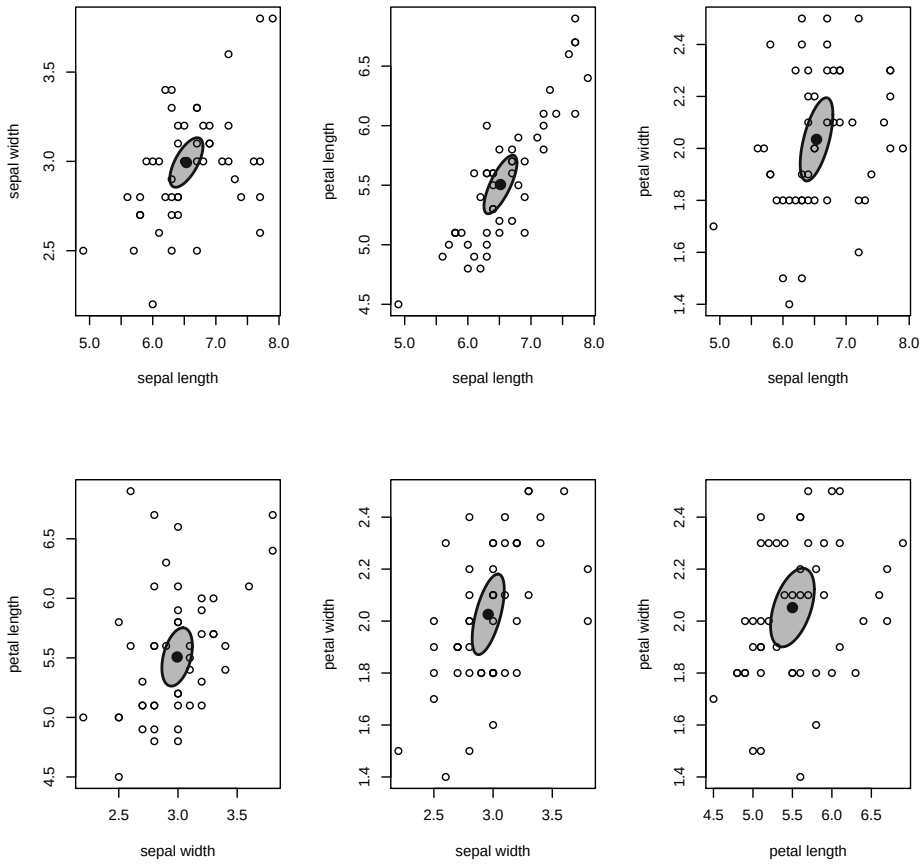


Figure 2.2 95% Credible ellipsoids of two-dimensional spatial medians for the species Virginica, in iris data

(see Chaudhuri (1996)). An L-estimator is a weighted average of order statistics. Chaudhuri (1996) defined an L-estimator of multivariate location of the form $\int_S \hat{Q}_n(u)\mu(\mathrm{d}u)$, where μ is an appropriately chosen probability measure supported on a subset S of $B_2^{(k)}$. We propose a non-parametric Bayesian analog of the form $\int_S Q_P(u)\mu(\mathrm{d}u)$, and put a DP($\alpha$) prior on P. If S is a finite set $\{u_1,\ldots,u_s\}$, then the integral is of the form $\sum_{i=1}^{s} Q_P(u_i)\mu(\{u_i\})$, whose posterior asymptotic distribution can directly be obtained from Theorem 2.5.1.

• Our approach has a deep connection with the bootstrap, as we are essentially doing a bootstrap approximation to the posterior Dirichlet process. The Bayesian bootstrap is a smoother version of Efron's bootstrap. For Efron's bootstrap, the weights $(W_{n1},\ldots,W_{nn})$ are multinomial with probabilities $(1/n,\ldots,1/n)$, and they satisfy conditions (i)–(v) in Subsection 2.3.1, with $c = 1$. Thus, credible sets obtained from Efron's bootstrap will be asymptotically equivalent with the credible sets we have obtained here.

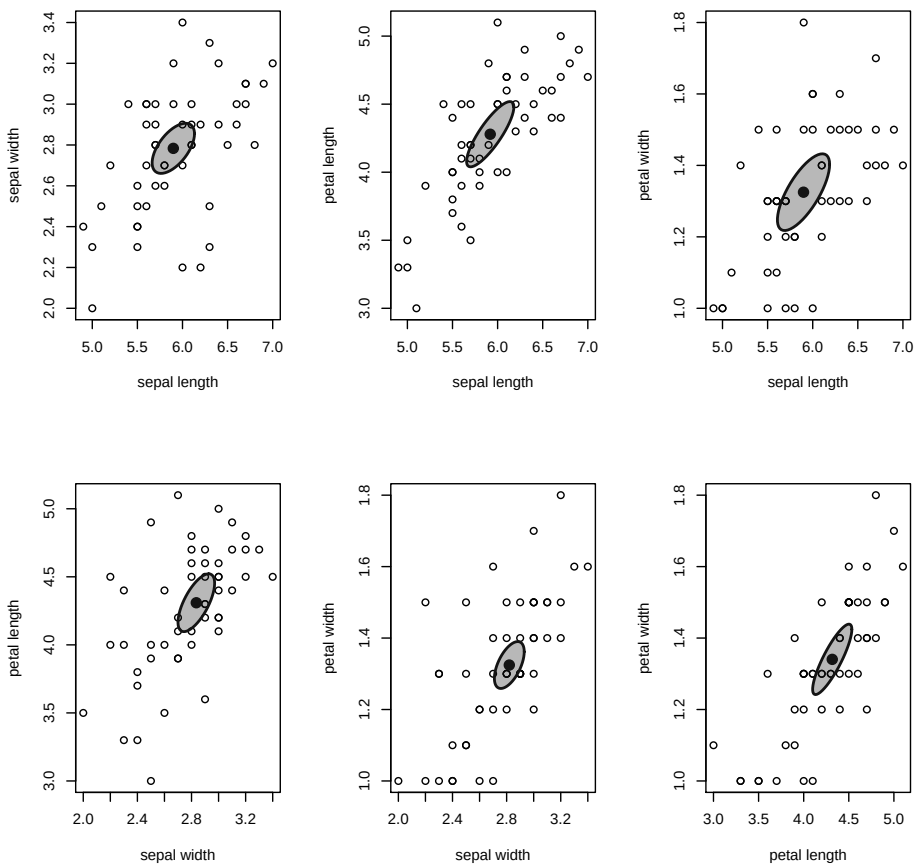


Figure 2.3 95% Credible ellipsoids of two-dimensional spatial medians for the species Versicolor, in iris data

2.8 Proof

2.8.1 Technical preliminaries

Before we proceed to the proof, we introduce some notations and definitions that we will need in the proof. For the rest of this chapter, $L_r(Q)$ denotes the norm $\|f\|_{Q,r} = \left(\int |f|^r dQ\right)^{1/r}$.

Definition 2.8.1 (Covering Numbers and Uniform Entropy)**.** The covering number $N(\epsilon, \mathscr{F}, \|\cdot\|)$ is the minimal number of balls $\{g : \|g - f\| < \epsilon\}$ of radius ϵ needed to cover $\mathscr{F}$.

A class of functions $\mathscr{F}$ with the envelope function F is said to satisfy the uniform entropy condition if

$$\int_0^\infty \sup_Q \sqrt{\log N(\epsilon \|F\|_{Q,2}, \mathscr{F}, L_2(Q))}\, \mathrm{d}\epsilon < \infty, \tag{2.8.1}$$

where the supremum has been taken over all finite discrete probability measures with $\|F\|_{Q,2}^2 = \int F^2 \mathrm{d}Q > 0$.

Definition 2.8.2 (Bracketing Numbers)**.** For two functions l and u, the bracket $[l,u]$ is defined to be the set of all functions f with $l \le f \le u$. An ϵ-bracket in $L_r(P)$ is a bracket $[l,u]$ with $\|u-l\| \le \epsilon$.

The bracketing number $N_{[]}(\epsilon, \mathscr{F}, L_r(P))$ is the minimum number of ϵ-brackets needed to cover $\mathscr{F}$.

Definition 2.8.3 (VC Class of Sets)**.** Let $\mathscr{C}$ be a collection of subsets of a set $\mathfrak{X}$. We say that an arbitrary subset $S = \{x_1, x_2, \ldots, x_n\}$ of $\mathfrak{X}$ is shattered by $\mathscr{C}$ if for every subset $S' \subseteq S$, there exists $C \in \mathscr{C}$ such that $S' = S \cap C$.

The VC-index of the class $\mathscr{C}$ is the smallest n for which no set of size n is shattered by $\mathscr{C}$ i.e.

$$V(\mathscr{C}) = \inf\{n : \max_{x_1,\ldots,x_n} \Delta_n(\mathscr{C}, x_1, \ldots, x_n) < 2^n\}, \tag{2.8.2}$$

where $\Delta_n(\mathscr{C}, x_1, \ldots, x_n) = \#\{C \cap \{x_1, \ldots, x_n\} : C \in \mathscr{C}\}$. A collection of measurable sets is called a VC class of sets if its VC-index is finite.

Definition 2.8.4 (Glivenko-Cantelli Class)**.** A function class $\mathscr{F}$ for which

$$\|\mathbb{P}_n - P\|_{\mathscr{F}} = \sup_{f \in \mathscr{F}} |\mathbb{P}_n f - P f| \to 0,$$

is called a P-Glivenko-Cantelli class, where the convergence can be in probability or almost surely.

Definition 2.8.5 (Donsker Class)**.** For a function class $\mathscr{F}$ and a probability measure P, suppose that

$$\sup_{f \in \mathscr{F}} |f(x) - P f| < \infty. \tag{2.8.3}$$

Let $\mathscr{L}^\infty(T)$ be the set of all functions $f : T \mapsto R$ such that

$$\sup_{t \in T} |f(t)| < \infty.$$

By viewing the empirical process $\{\mathbb{G}_n f : f \in \mathscr{F}\}$ as a map into $\mathscr{L}^\infty(\mathscr{F})$, if

$$\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P) \rightsquigarrow \mathbb{G}, \qquad \text{in } \mathscr{L}^\infty(\mathscr{F}), \tag{2.8.4}$$

for a tight Borel measurable element $\mathbb{G}$ in $\mathscr{L}^\infty(\mathscr{F})$, then we say that $\mathscr{F}$ is called a P-Donsker class.

Definition 2.8.6 (Separability of a Stochastic Process)**.** A stochastic process $\{X(t),\ t \in T\}$, where (T,ρ) is a separable metric space, is separable if there exists a countable subset $S \in T$ and a null set N such that for each $\omega \notin N$ and $t \in T$, there exists a sequence $\{s_m\} \in S$ with $\rho(s_m, t) \to 0$, and

$$|X(s_m,\omega) - X(t,\omega)| \to 0.$$

2.8.2 Proof of Theorem 2.3.1

We give the proof in two steps. In the first step, we state and prove Lemma 2.8.1, i.e., we show that the asymptotic posterior distribution of $\sqrt{n}(\theta(P) - \hat{\theta}_n)$ is the same as the asymptotic conditional distribution of $\sqrt{n}(\theta(\mathbb{B}_n) - \hat{\theta}_n)$. Next, we verify the conditions of Theorem 2.3.2 in our situation and show that the asymptotic conditional distribution of $\sqrt{n}(\theta(\mathbb{B}_n) - \hat{\theta}_n)$ is $\mathrm{N}_k(0, V_{\theta^\star}^{-1} U_{\theta^\star} V_{\theta^\star}^{-1})$.

Lemma 2.8.1. *The asymptotic posterior distribution of $\sqrt{n}(\theta(P) - \hat{\theta}_n)$ is the same as the asymptotic conditional distribution of $\sqrt{n}(\theta(\mathbb{B}_n) - \hat{\theta}_n)$.*

Proof of Lemma 2.8.1. We know $\theta(\mathbb{B}_n)$ satisfies $\Psi'(\theta(\mathbb{B}_n)) = \mathbb{B}_n \psi(\cdot, \theta) = 0$ and $\theta(P)$ satisfies $\Psi(\theta(P)) = P\psi(\cdot,\theta) = 0$.

The posterior distribution of P given $X_1, \ldots, X_n$ is $\mathrm{DP}(\alpha + n\mathbb{P}_n)$. From the fact that $\|P - \mathbb{B}_n\|_{TV} = o_{P'}(n^{-1/2})$ a.s. $[P^{\star n}]$, where $P' = P^n \times \mathbb{B}_n$,

$$\left\| P\psi(X,\theta) - \mathbb{B}_n\psi(X,\theta) \right\| \le \left\|\psi\right\|_\infty \left\| P - \mathbb{B}_n \right\|_{\mathrm{TV}} \le \left\| P - \mathbb{B}_n \right\|_{\mathrm{TV}},$$

since $\|\psi\|_\infty = \sup_x |\psi(x,\theta)| = 1$. In view of this result, given $X_1, \ldots, X_n$,

$$\left\| \Psi'(\theta(P)) - \Psi(\theta(P)) \right\|_2 = \left\| \Psi'(\theta(P)) \right\|_2 = o_{P'}(n^{-1/2}). \tag{2.8.5}$$

Hence, for given $X_1, \ldots, X_n$, $\theta(P)$ makes the bootstrap scores $\Psi'(\theta)$ approximately zero in probability. Therefore, given the observations $X_1, \ldots, X_n$, $\theta(P)$ qualifies to be a sequence of bootstrap asymptotic Z-estimators. Theorem 3.1 in Wellner and Zhan (1996) (Theorem 2.3.2 in this paper) holds for any sequence of bootstrap asymptotic Z-estimators $\hat{\theta}_n$ that satisfies

$$\left\| \Psi'(\hat{\theta}_n) \right\| = o_{P'}(n^{-1/2}). \tag{2.8.6}$$

Thus, the asymptotic posterior distribution of $\sqrt{n}(\theta(P) - \hat{\theta}_n)$ is same as the asymptotic conditional distribution of $\sqrt{n}(\theta(\mathbb{B}_n) - \hat{\theta}_n)$. □

Next, we show that $\psi(\cdot,\theta)$ defined in (2.2.2) satisfies the conditions in Theorem 2.3.2. Firstly, we need to show that the function class $\mathscr{F}_R \in \mathfrak{m}(P^\star)$ where $\mathscr{F}_R$ is defined in (2.3.3).

To achieve this, we prove that the empirical process $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P^\star)$ indexed by $\mathscr{F}_R$ is stochastically separable. It can be noted that $\psi_j(x,\theta),\ j = 1,\ldots,k$, are left-continuous at each x for every θ such that $\|\theta - \theta^\star\|_2 \le R$. Hence there exists a null set N and a countable $\mathscr{G} \subset \mathscr{F}_R$ such that, for every $\omega \notin N$ and $f \in \mathscr{F}_R$, we have a sequence $g_m \in \mathscr{G}$ with $g_m \to f$ and $\mathbb{G}_n(g_m,\omega) \to \mathbb{G}_n(f,\omega)$. For more details, see Chapter 2.3, van der Vaart and Wellner (1996).

Verification of Condition 1 *in Theorem* 2.3.2. By Condition C2 in Theorem 2.3.1, the ℓ_1- median of $P^\star$ exists and is unique. Hence there exists a $\theta^\star \equiv \theta(P^\star) \in \mathbb{R}^k$ such that (2.3.4) is satisfied. Also, $\Psi^\star(\theta) = P^\star\psi(X,\theta)$ is differentiable from Condition C1. This follows from the fact that for a fixed $\theta \in \mathbb{R}^k$ and a density f bounded on compact subsets of $\mathbb{R}^k$, $P^\star(\|X-\theta\|_2^{-1})$ is finite, which in turn implies that $P^\star(\|X-\theta\|_p^{-1})$ is finite for every $p > 1$. This can be verified by using k-dimensional polar transformation for which the determinant of the Jacobian matrix contains $(k-1)$th power of the radius vector (Chaudhuri (1996)). □

Verification of Condition 2 *in Theorem* 2.3.2. From Wellner and Zhan (1996), Condition 2 is satisfied if, $\mathscr{F}_R$ in (2.3.3) is $P^\star$-Donsker for some $R > 0$ and

$$\max_{1\le j\le k} P^\star(\psi_j(\cdot,\theta) - \psi_j(\cdot,\theta^\star))^2 \to 0, \tag{2.8.7}$$

as $\theta \to \theta^\star$. In order to prove that $\mathscr{F}_R$ is $P^\star$-Donsker, we define the following two function classes:

$$\mathscr{F}_{1R} = \Big\{ \frac{|x_j - \theta_j|^{p-1}}{\|x-\theta\|_p^{p-1}} : j = 1,2,\ldots,k, \|\theta-\theta^\star\|_2 \le R \Big\}, \tag{2.8.8}$$

$$\mathscr{F}_{2R} = \Big\{ \operatorname{sign}(\theta_j - x_j) : j = 1,2,\ldots,k, \|\theta-\theta^\star\|_2 \le R \Big\}. \tag{2.8.9}$$

From Example 2.10.23 of van der Vaart and Wellner (1996), if $\mathscr{F}_{1R}$ and $\mathscr{F}_{2R}$ satisfy the uniform entropy condition and are suitably measurable, then $\mathscr{F}_R = \mathscr{F}_{1R}\mathscr{F}_{2R}$ is $P^\star$-Donsker provided their envelopes F_{1R} and F_{2R} satisfy $P^\star F_{1R}^2 F_{2R}^2 < \infty$. □

Lemma 2.8.2. *For $k > 2$, $\mathscr{F}_{1R}$ and $\mathscr{F}_{2R}$ are $P^\star$-Donsker classes for some fixed integer p, and hence they satisfy the uniform entropy condition.*

Proof of Lemma 2. Recall that, $\mathscr{F}_R = \mathscr{F}_{1R}\mathscr{F}_{2R}$, where

$$\mathscr{F}_{1R} = \Big\{ \frac{|x_j - \theta_j|^{p-1}}{\|x-\theta\|_p^{p-1}} : j = 1,2,\ldots,k, \|\theta-\theta^\star\|_2 \le R \Big\}, \tag{2.8.10}$$

$$\mathscr{F}_{2R} = \Big\{ \operatorname{sign}(\theta_j - x_j) : j = 1,2,\ldots,k, \|\theta-\theta^\star\|_2 \le R \Big\}. \tag{2.8.11}$$

Every $f \in \mathscr{F}_{1R}$ is continuous at each x, hence $\mathscr{F}_{1R}$ has a countable subset $\mathscr{G}$ such that for every $f \in \mathscr{F}_{1R}$ there exists a sequence $g_m \in \mathscr{G}$ such that $g_m(x) \to f(x)$ for every x. Then by Example 2.3.4 of van der Vaart and Wellner (1996), $\mathscr{F}_{1R}$ is P-measurable for every P. Since every $f \in \mathscr{F}_{2R}$ is left-continuous at each x, same conclusion holds for $\mathscr{F}_{2R}$ as well.

A class of functions $\mathscr{F}$ is called a VC-major class of functions if the sets $\{x : f(x) > t\}$ with f ranging over $\mathscr{F}$ and t over $\mathbb{R}$ form a VC-class of sets. By Corollary 2.6.12 of van der Vaart and Wellner (1996), if $\mathscr{F}_{1R}$ is a bounded VC-major class of functions, then it satisfies the uniform entropy condition. It is easy to see that $\mathscr{F}_{1R}$ is bounded. We now show that $\mathscr{F}_{1R}$ is a VC-major class of sets, that is, the sets $\{x : f(x) > t\}$ with f varying over $\mathscr{F}_{1R}$ and t over $\mathbb{R}$ form a VC class of sets. Define the collection of sets $\mathscr{S} = \{S_{\theta,t} : \|\theta - \theta^\star\|_2 \le R,\ t \in \mathbb{R}\}$, where $S_{\theta,t}$ is defined as

$$S_{\theta,t} = \Big\{x : \frac{|x_j - \theta_j|^{p-1}}{\|x-\theta\|_p^{p-1}} > t,\ j = 1,2,\ldots,k\Big\}. \tag{2.8.12}$$

We need to show $\mathscr{S}$ is a VC-class of sets. Note that $S_{\theta,t} = \cap_{j=1}^k S_{\theta,t}^j$, where $S_{\theta,t}^j$ is defined as

$$S_{\theta,t}^j = \Big\{x : \frac{|x_j - \theta_j|^{p-1}}{\|x-\theta\|_p^{p-1}} > t\Big\}. \tag{2.8.13}$$

In view of Lemma 2.6.17 of van der Vaart and Wellner (1996), it is enough to show

$$\mathscr{S}^j = \{S_{\theta,t}^j : \|\theta - \theta^\star\|_2 \le R,\ t \in \mathbb{R}\} \tag{2.8.14}$$

is a VC-class for every j; because if every $\mathscr{S}^j$ is a VC-class of sets, $\mathscr{S} = \sqcap_{j=1}^k \mathscr{S}^j = \{\cap_{j=1}^k S^j : S^j \in \mathscr{S}^j\}$ is also a VC-class of sets. Hence we only show that $\mathscr{S}^1 = \{S_{\theta,t}^1 : \|\theta-\theta^\star\|_2 \le R,\ t \in \mathbb{R}\}$ is a VC-class of sets. We can write $S_{\theta,t}^1$ as

$$S_{\theta,t}^1 = \Big\{x : |x_1 - \theta_1|^p > \frac{t^{p/(p-1)}}{1 - t^{p/(p-1)}} \sum_{j=2}^k |x_j - \theta_j|^p)\Big\}.$$

Define $R_{\theta,c}^1 = \Big\{x : |x_1 - \theta_1|^p > c \sum_{j=2}^k |x_j - \theta_j|^p\Big\}$ and $\mathscr{R}^1 = \{R_{\theta,c}^1 : \theta \in \mathbb{R}^k, c \in \mathbb{R}\}$. It is enough to show that $\mathscr{R}^1$ is a VC-class, since $\mathscr{R}^1$ contains $\mathscr{S}^1$. For $i,\ j = 2,\ldots,k,\ i \ne j$, we define

$$A_{0;\theta,c} = \{x : (x_1 - \theta_1)^p > c \sum_{j=2}^k (x_j - \theta_j)^p\} \bigcap_{j-1}^k \{x :\ x_j - \theta_j \ge 0\},$$

$$A'_{0;\theta,c} = \{x : (\theta_1 - x_1)^p > c\sum_{j=2}^{k}(x_j - \theta_j)^p,\ \theta_1 - x_1 \le 0\} \bigcap_{j=2}^{k} \{x : x_j - \theta_j \ge 0\},$$

$$A_{i;\theta,c} = \{x : (x_1 - \theta_1)^p > c(\theta_i - x_i)^p + c\sum_{\substack{j=2 \\ j\neq i}}^{k}(x_j - \theta_j)^p,$$

$$x_1 - \theta_1 \ge 0,\ \theta_i - x_i \ge 0,\ x_j - \theta_j \ge 0,\ j = 2,\ldots,k,\ j \neq i\},$$

$$A'_{i;\theta,c} = \{x : (\theta_1 - x_1)^p > c(\theta_i - x_i)^p + c\sum_{\substack{j=2 \\ j\neq i}}^{k}(x_j - \theta_j)^p,$$

$$\theta_1 - x_1 \ge 0,\ \theta_i - x_i \ge 0,\ x_j - \theta_j \ge 0,\ j = 2,\ldots,k,\ j \neq i\},$$

$$A_{ij;\theta,c} = \{x : (x_1 - \theta_1)^p > c(\theta_i - x_i)^p + c(\theta_j - x_j)^p + c\sum_{\substack{l=2 \\ l\neq i,j}}^{k}(x_l - \theta_l)^p,$$

$$x_1 - \theta_1 \ge 0,\ \theta_i - x_i \ge 0,\ \theta_j - x_j \ge 0,\ x_l - \theta_l \ge 0,\ l = 2,\ldots,k,$$
$$l \neq i, j\},$$

$$A'_{ij;\theta,c} = \{x : (\theta_1 - x_1)^p > c(\theta_i - x_i)^p + c(\theta_j - x_j)^p + c\sum_{\substack{l=2 \\ l\neq i,j}}^{k}(x_l - \theta_l)^p,$$

$$\theta_1 - x_1 \ge 0,\ \theta_i - x_i \ge 0,\ \theta_j - x_j \ge 0,\ x_l - \theta_l \ge 0,\ l = 2,\ldots,k,$$
$$l \neq i, j\}.$$

Continuing this pattern, finally

$$A_{n-1;\theta,c} = \{x : (x_1 - \theta_1)^p > c\sum_{j=2}^{k}(\theta_j - x_j)^p,\ x_1 - \theta_1 \ge 0,\ \theta_j - x_j \ge 0,$$

$$j = 2,\ldots,k\},$$

$$A'_{n-1;\theta,c} = \{x : (\theta_1 - x_1)^p > c\sum_{j=2}^{k}(\theta_j - x_j)^p,\ \theta_1 - x_1 \ge 0,\ \theta_j - x_j \ge 0,$$

$$j = 2,\ldots,k\}.$$

Using the preceding notations, $R^1_{\theta,c}$ can be written as

$$R^1_{\theta,c} = A_{0;\theta,c} \cup A'_{0;\theta,c} \cup \{\bigcup_{l=1}^{n-2} B_{l;\theta,c}\} \cup \{\bigcup_{l=1}^{n-2} B'_{l;\theta,c}\} \cup A_{n-1;\theta,c} \cup A'_{n-1;\theta,c}, \tag{2.8.15}$$

where

$$B_{1;\theta,c} = \bigcup_{i=2}^{k} A_{i;\theta,c};\ B'_{1;\theta,c} = \bigcup_{i=2}^{k} A'_{i;\theta,c};\ B_{2;\theta,c} = \bigcup_{\substack{i=2 \\ i<j}}^{k} \bigcup_{j=2}^{k} A_{ij;\theta,c},$$

and so on. Since all the sets on the right hand side of (2.8.15) are in the same form, if $\mathscr{C} = \{A_{0;\theta,c} : \theta \in \mathbb{R}^k, c \in \mathbb{R}\}$ forms a VC-class of sets, then $\mathscr{R}^1$ also forms a VC-class of sets. This follows from Lemma 2.6.17 of van der Vaart and Wellner (1996), which says that if $\mathscr{F}$ and $\mathscr{G}$ are VC-classes, then $\mathscr{F} \sqcup \mathscr{G} = \{F \cup G : F \in \mathscr{F},\ G \in \mathscr{G}\}$ also forms a VC-class. We can write $A_{0;\theta,c}$ as

$$A_{0;\theta,c} = \{x : (x_1 - \theta_1)^p > c \sum_{j=2}^{k} (x_j - \theta_j)^p\} \bigcap_{j=1}^{k} \{x : x_j - \theta_j \geq 0\}. \tag{2.8.16}$$

Since p is a positive integer greater than 1, we can write

$$\begin{aligned}
&\{x : (x_1 - \theta_1)^p > c \sum_{j=2}^{k} (x_j - \theta_j)^p\} \\
&= \{x : \textstyle\sum_{r=0}^{p} \binom{p}{r} x_1^{p-r} (-1)^r \theta_1^r > c \sum_{j=2}^{k} \sum_{r=0}^{p} \binom{p}{r} x_j^{p-r} (-1)^r \theta_j^r\} \\
&= \{x : x_1^p - p x_1^{p-1} \theta_1 + \cdots + (-1)^p \theta_1^p - c \textstyle\sum_{j=2}^{k} (x_j^p - p x_j^{p-1} \theta_j \\
&\qquad + \cdots + (-1)^p \theta_j^p) > 0\}.
\end{aligned}$$

Consider the map $x \mapsto \phi(x)$, where

$$\phi(x) = \Big\{x_1^p, x_1^{p-1}, x_1^{p-2}, \ldots, x_1, \sum_{j=2}^{k} x_j^p, \sum_{j=2}^{k} x_j^{p-1}, \ldots, \sum_{j=2}^{k} x_j, 1\Big\}. \tag{2.8.17}$$

Note that the class of functions $\{g_a(x) = a^T \phi(x)\},\ a \in \mathbb{R}^{2p+1}$ is a finite dimensional vector space. The collection of sets

$$\{x : (x_1 - \theta_1)^p > c \sum_{j=2}^{k} (x_j - \theta_j)^p,\ \theta \in \mathbb{R}^k,\ c \in \mathbb{R}\}$$

is the same as $\mathscr{C}_1 = \{x : g_a(x) > 0\}, a \in \mathbb{R}^{2p+1}$ and $\mathscr{C}_1$ is a VC-class of sets by Lemma 2.6.15 of van der Vaart and Wellner (1996). Each of the classes $\{x : x_j - \theta_j \geq 0\}$ for $j = 1, 2, \ldots, k$, is a sub-collection of VC classes of sets $\mathscr{C}_2 = \{x : a^T x + b \geq 0\},\ a \in \mathbb{R}^k,\ b \in \mathbb{R}$. Hence by Lemma 2.6.15 of van der Vaart and Wellner (1996), $\mathscr{C}$ forms a VC-class of sets.

Thus we proved that $\mathscr{F}_{1R}$ is a bounded VC major class of functions. Hence $\mathscr{F}_{1R}$ satisfies

the uniform entropy condition.

For $\mathscr{F}_{2R}$, we see that the class of functions $\{x \mapsto \theta_j - x_j : \|\theta - \theta^\star\|_2 \le R\}$ belongs to a finite dimensional vector space and hence is a VC class. From the stability properties of VC classes (Example 3.3.9, van der Vaart and Wellner (1996)), the class of functions $\{x \mapsto \text{sign}(\theta_j - x_j) : \|\theta - \theta^\star\|_2 \le R\}$ is also VC. Hence from Lemma 2.6.15 of van der Vaart and Wellner (1996), $\mathscr{F}_{2R}$ is a bounded VC major class of functions and satisfies the uniform entropy condition. Therefore $\mathscr{F}_R = \mathscr{F}_{1R}\mathscr{F}_{2R}$ is $P^\star$-Donsker. □

In view of Lemma 2.8.2, next we need to prove (2.8.7), that is, $\max_{1\le j\le k} P^\star(\psi_j(\cdot,\theta) - \psi_j(\cdot,\theta^\star))^2 \to 0$ as $\theta \to \theta^\star$. Note that $\psi_j(x,\theta) \to \psi_j(x,\theta^\star)$ for every x as $\theta \to \theta^\star$ for $j \in \{1,\ldots,k\}$. Also $(\psi_j(x,\theta) - \psi_j(x,\theta^\star))^2 \le 4$ for every x and every θ. Hence by the Dominated Convergence Theorem (DCT), $P^\star(\psi_j(\cdot,\theta) - \psi_j(\cdot,\theta^\star))^2 \to 0$ as $\theta \to \theta^\star$ for $j \in \{1,\ldots,k\}$. Thus (2.8.7) is established.

Verification of Condition 3 *in Theorem* 2.3.2. For every $j \in \{1,2,\ldots,k\}$ and $\theta \in \mathbb{R}^k$, $\psi_j(x,\theta)$ is bounded by 1 and hence is square-integrable. The (i,j)th element of $U_{\theta^\star} = P^\star\psi(x,\theta^\star)\psi^T(x,\theta^\star)$ is given by

$$\begin{aligned}\sigma_{ij} &= \int \frac{|x_i - \theta_i^\star|^{p-1}|x_j - \theta_j^\star|^{p-1}}{\|x - \theta^\star\|_p^{2(p-1)}} \text{sign}(\theta_i^\star - x_i)\,\text{sign}(\theta_j^\star - x_j)\,\mathrm{d}P^\star \qquad (2.8.18)\\ &\le \int 1\,\mathrm{d}P^\star < \infty.\end{aligned}$$

The class of functions $\{\psi_j(x,\theta) : j = 1,2,\ldots,k,\ \|\theta - \theta^\star\|_2 \le R\}$ has a constant envelope 1. Hence $D_n(x)$ defined in (2.3.8) is equal to 2 and it satisfies (2.3.9). □

Verification of Condition 4. First we prove $\|\hat\theta_n - \theta^\star\|_2 \xrightarrow{P^\star} 0$. Note that $\hat\theta_n$ can be written as

$$\hat\theta_n = \arg\max_\theta \mathbb{P}_n m_\theta, \qquad (2.8.19)$$

where $m_\theta(x) = -\|x - \theta\|_p + \|x\|_p$. Naturally the population analog of $\hat\theta_n$ is given by

$$\theta(P) = \arg\max_\theta P m_\theta. \qquad (2.8.20)$$

From Corollary 3.2.3 of van der Vaart and Wellner (1996), we need to establish two conditions as follows:

(a) $\sup_\theta |\mathbb{P}_n m_\theta - P^\star m_\theta| \to 0$ in probability;

(b) there exists a $\theta^\star$ such that $P^\star m_{\theta^\star} > \sup_{\theta \notin G} P^\star m_\theta$ for every open set G containing $\theta^\star$.

The first condition can be proved by showing that the class of functions $\{m_\theta : \theta \in \mathbb{R}^k\}$ forms a $P^\star$-Glivenko-Cantelli class. From Theorem 19.4 of van der Vaart (2000), the class $\mathscr{M} = \{m_\theta : \theta \in \Theta \subset \mathbb{R}^k\}$ will be $P^\star$-Glivenko-Cantelli if $N_{[]}(\epsilon, \mathscr{M}, L_1(P^\star)) < \infty$ for every $\epsilon > 0$.

By Example 19.7 of van der Vaart (2000), for a class of measurable functions $\mathscr{F} = \{f_\theta : \theta \in \Theta \subset \mathbb{R}^k\}$, if there exists a measurable function m such that

$$|f_1(x) - f_2(x)| \leq m(x)\|\theta_1 - \theta_2\|_2, \tag{2.8.21}$$

for every θ_1, θ_2 and $P^\star|m|^r < \infty$, then there exists a constant K, depending on Θ and k only, such that the bracketing numbers satisfy

$$N_{[]}(\epsilon \|m\|_{P^\star,r}, \mathscr{F}, L_r(P^\star)) \leq K\left(\frac{\operatorname{diam}\Theta}{\epsilon}\right)^k, \tag{2.8.22}$$

for every $0 < \epsilon < \operatorname{diam}\Theta$. To use this example, we need to restrict the parameter space to a compact subset of $\mathbb{R}^k$. The next lemma shows that this can be avoided in our case by asserting that the parameter space can be restricted to a sufficiently large compact set with high probability.

Lemma 2.8.3. *For some $0 < \epsilon < 1/4$ and $K > 0$ such that $P^\star(\|X\|_p \leq K) > 1-\epsilon$, $\|\theta(\mathbb{B}_n)\|_p \leq 3K$ with high joint probability $P' = P^{\star n} \times \mathbb{B}_n$.*

Proof of Lemma 2.8.3. Define $M(P^\star, \theta) = P^\star m_\theta = P^\star(\|X - \theta\|_p - \|X\|_p)$. We show that for $0 < \epsilon < 1/4$, there exists $K > 0$ such that $\|\theta\|_p \geq 3K$ implies $M(P^\star, \theta) > 0$. If $\|X\|_p \leq K$ and $\|\theta\|_p \geq 3K$, then

$$\|X - \theta\|_p \geq \|\theta\|_p - \|X\|_p \geq \frac{2\|\theta\|_p}{3} + K - \|X\|_p \geq \frac{2\|\theta\|_p}{3},$$

Hence as $\|X\|_p \leq K \leq \|\theta\|_p/3$,

$$\|X - \theta\|_p - \|X\|_p \geq \frac{2\|\theta\|_p}{3} - \frac{\|\theta\|_p}{3} = \frac{\|\theta\|_p}{3}.$$

Now since always $\left|\|X-\theta\|_p-\|X\|_p\right|\le\|\theta\|_p$, we can write

$$\begin{aligned}M(P^\star,\theta)&=\int_{\|X\|_p\le K}(\|X-\theta\|_p-\|X\|_p)\mathrm{d}P^\star+\int_{\|X\|_p>K}(\|X-\theta\|_p-\|X\|_p)\mathrm{d}P^\star\\&\ge\|\theta\|_p\Big(\frac{1}{3}P^\star(\|X\|_p\le K)-P^\star(\|X\|_p>K)\Big)\\&=\|\theta\|_p\Big(\frac{1}{3}-\frac{4}{3}P^\star(\|X\|_p>K)\Big)\\&\ge\|\theta\|_p\Big(\frac{1}{3}-\frac{4}{3}\epsilon\Big)>0.\end{aligned}$$

We assume that $0<\epsilon<1/4$ and $K>0$ have been chosen so that $P^\star$ satisfies $P^\star(\|X\|_p\le K)>1-\epsilon$. Hence $\|\theta(P^\star)\|_p\le 3K$. Also, since $\mathbb{P}_n\rightsquigarrow P^\star$, for some $0<\epsilon<1/4$ and $K>0$, $\mathbb{P}_n$ satisfies $\mathbb{P}_n(\|X\|_p\le K)>1-\epsilon$ with high probability. Hence $\|\hat{\theta}_n\|_p\le 3K$ with high probability. Similarly, we know that $\mathbb{B}_n-\mathbb{P}_n\overset{P^{\star n}\times\mathbb{B}_n}{\to}0$ in the weak topology. Hence with high joint probability, $\mathbb{B}_n$ satisfies $\mathbb{B}_n(\|X\|_p\le K)>1-\epsilon$, leading to $\|\theta(\mathbb{B}_n)\|_p\le 3K$ with high joint probability. □

Because of Lemma 2.8.3, it suffices to establish (2.8.21). Using Minkowski's inequality,

$$|m_\theta(x)-m_{\theta'}(x)|=|\|x-\theta'\|_p-\|x-\theta\|_p|\le\|\theta-\theta'\|_p.$$

This expression is bounded by $\|\theta-\theta'\|_2$ for $p\ge 2$, by the fact that $\|z\|_{p+a}\le\|z\|_p$ for any vector z and real numbers $a\ge 0$ and $p\ge 1$. For $1<p<2$, the expression is bounded by $2^{(1/p)-(1/2)}\|\theta-\theta'\|_2$. Hence we choose $m(x)=1$ for every x and therefore $P^\star|m|=1$. This ensures that $N_{[]}(\epsilon,\mathscr{M},L_1(P^\star))<\infty$ and hence Condition (a) is satisfied. From Condition (C2) in Theorem 2.3.1, Condition (b) holds. Therefore $\hat{\theta}_n\to\theta^\star$ in $P^\star$-probability.

Now to prove the consistency of $\theta(\mathbb{B}_n)$, which is viewed as a "bootstrap estimator", we use Corollary 3.2.3 in van der Vaart and Wellner (1996). Two conditions are needed for proving this. The first condition is $\sup_\theta|\mathbb{B}_n m_\theta-P^\star m_\theta|\overset{P^\star\times\mathbb{B}_n}{\to}0$. We verify this condition using the multiplier Glivenko-Cantelli theorem which is given in Corollary 3.6.16 of van der Vaart and Wellner (1996). By the representation $\mathbb{B}_n=\sum_{i=1}^n B_{ni}\delta_{X_i}$, where $(B_{n1},\ldots,B_{nn})\sim\mathrm{Dir}(n;1,\ldots,1)$, it follows that $B_{ni}\ge 0$, $\sum_{i=1}^n B_{ni}=1$ and $B_{ni}\sim\mathrm{Be}(1,n-1)$. Therefore, for every $\epsilon>0$, as $n\to\infty$

$$P\Big(\max_{1\le i\le n}|B_{ni}|<\epsilon\Big)=\left(\int_0^\epsilon\frac{(1-y)^{n-2}}{B(1,n-1)}dy\right)^n=(1-(1-\epsilon)^{n-1})^n\to 1.$$

Thus the first condition is proved.

The second condition is the same as the "well-separatedness" condition (b) which we already verified. So, we have $\hat{\theta}_n \xrightarrow{P^\star} \theta^\star$ and $\theta(\mathbb{B}_n) \xrightarrow{P^\star \times \mathbb{B}_n} \theta^\star$. Hence by an application of the triangle inequality, $\theta(\mathbb{B}_n) \xrightarrow{\mathbb{B}_n} \hat{\theta}_n$ in $P^\star$-probability. □

Verification of Condition 5. It has already been mentioned that the Bayesian bootstrap weights satisfy the bootstrap weights (i)–(υ). □

2.8.3 Proof of Theorem 2.5.1

Just like before, the sample geometric quantiles $\hat{Q}_n(u_1), \ldots, \hat{Q}_n(u_m)$ are viewed as a Z-estimator satisfying the system of equations $\mathbb{P}_n \psi(\cdot, \xi) = 0$, where $\psi(\cdot, \xi) = \{\psi_{lj}(\cdot, \xi_{lj}) : l = 1, \ldots, m,\ j = 1, \ldots, k\}$ is the score vector with its real-valued elements being

$$\psi_{lj}(x, \xi_{lj}) = \frac{|x_j - \xi_{lj}|^{p-1}}{\|x - \xi_l\|_p^{p-1}} \operatorname{sign}(\xi_{lj} - x_j) + u_{lj}. \tag{2.8.23}$$

We define $Q_{\mathbb{B}_n}(u_1), \ldots, Q_{\mathbb{B}_n}(u_m)$ as the corresponding "Bayesian bbootstrapped" versions of the Z-estimators $\hat{Q}_n(u_1), \ldots, \hat{Q}_n(u_m)$, i.e., they satisfy the system of equations $\mathbb{B}_n \psi(\cdot, \xi) = 0$. We use the same technique of approximating the posterior distribution of P by a Bayesian bootstrap distribution, and we state the following lemma, which is the extension of Lemma 2.8.1 to the quantile case.

Lemma 2.8.4. *For every fixed $u_1, \ldots, u_m \in B_q^{(k)}$, the joint asymptotic posterior distribution of $\sqrt{n}(Q_P(u_1) - \hat{Q}_n(u_1)), \ldots, \sqrt{n}(Q_P(u_m) - \hat{Q}_n(u_m))$ is the same as the asymptotic conditional distribution of $\sqrt{n}(Q_{\mathbb{B}_n}(u_1) - \hat{Q}_n(u_1)), \ldots, \sqrt{n}(Q_{\mathbb{B}_n}(u_m) - \hat{Q}_n(u_m))$.*

The proof of Lemma 2.8.4 is very similar to that of Lemma 2.8.1, hence is omitted. The rest of the proof of Theorem 2.5.1 is along the lines of that of Theorem 2.3.1 as well, hence that is skipped too.

CHAPTER

3

GIBBS POSTERIOR INFERENCE ON MULTIVARIATE QUANTILES

3.1 Introduction

The previous chapter considers Bayesian inference on multivariate quantiles in an *indirect* approach, i.e., we first model the underlying infinite-dimensional random distribution P with a DP(α) prior first, and then marginalize to the quantile functional $Q_P(u)$. This approach works reasonably well, but a possible downside is that incorporating suitable prior information for the quantiles can be tricky.

Beyond estimation, if the goal is inference on multivariate quantiles, the main difficulty is that they are not naturally described as parameters in a statistical model. That is, no standard or otherwise "reasonable" model for multivariate data will include a quantile in its parametrization, so some potentially dangerous non-linear marginalization (Fraser (2011), Martin (2019)) would typically be required. More importantly, with the specification of a parametric Bayesian model comes the risk of model misspecification bias, and since our quantity of interest is well-defined without a model, it is not clear what can be gained by working in a model-based framework to balance out the risk of model misspecification bias.

Consequently, the go-to methods for making inference about multivariate quantiles make use of distributional approximations for M-estimators like in Abdous and Theodorescu (1992). As mentioned in the previous chapter, there are numerous advantages of having a posterior distribution on which inferences can be based. For example, a posterior distribution provides a nice visual summary of uncertainty, credible regions can be immediately read off of it, no asymptotic approximations required, and at least, in principle, genuine prior information about the quantity of interest can be incorporated whenever it is available. In the non-parametric Bayesian approach, the best option is to choose a Dirichlet process prior with base measure to have quantile equal to the prior guess, but it is not clear how this (and other features of the specified base measure) affect the marginal posterior for the quantile.

In this chapter, we investigate the construction of a *Gibbs posterior* for a multivariate quantile, which is a more direct way (without marginalization). On one hand, like M-estimation, this approach uses a suitable loss function, rather than a likelihood, to connect the quantity of interest to the observed data, which eliminates the risk of model misspecification bias; on the other hand, like Bayesian inference, it produces a genuine posterior distribution and allows for the direct incorporation of prior information, thus offering better comprehensibility than the non-parametric Bayesian approach.

After some background about multivariate quantiles and Gibbs posteriors in Section 3.2, we define our object of interest, namely, the Gibbs posterior distribution for a multivariate quantile, and investigate its properties. In particular, in Section 3.3.2, we first establish that the Gibbs posterior concentrates around the true quantile at the usual $\sqrt{n}$-rate and, second, that it has an asymptotic Gaussian approximation in the Bernstein–von Mises sense. Unfortunately, the covariance matrix in this Gaussian approximation is "wrong" in the sense that it does not match that of the M-estimator around which it is centered. Fortunately, the Gibbs posterior depends on a user-specified *learning rate* (Bissiri et al. (2016), Grünwald and van Ommen (2017), Syring and Martin (2019)) which can be tuned to at least partially correct for the covariance matrix mismatch. We use a bootstrap based calibration algorithm proposed by Syring and Martin (2019) for choosing the learning rate, which we describe in Section 3.3.3. Numerical results in Section 3.4.1 compare the finite-sample performance of our proposed Gibbs posterior inference to that based on existing Bayesian approaches demonstrates that the Gibbs posterior gives a better performance than model-based parametric Bayesian approach in misspecified situations. We also apply the Gibbs posterior approach for inferring about the spatial median on an real data set in Section 3.4.2 for which the assumption of normality is not unreasonable. We show that the Gibbs posterior outperforms a normality-based Bayesian solution in terms of out-of-

sample risk, implying that our Gibbs solution indeed avoids some inherent bias coming from the normality assumption. Some concluding remarks are given in Section 3.5 and proofs of the two main theorems are presented in Section 3.6.

3.2 Background

3.2.1 Multivariate quantiles

As defined in Chapter 2, for $u \in B_q^{(k)}$, the k-dimensional sample u-quantile is then defined as

$$\hat{Q}_n(u) = \arg\min_{\xi \in \mathbb{R}^k} \frac{1}{n} \sum_{i=1}^{n} \Phi_p(u, X_i - \xi), \tag{3.2.1}$$

where $\Phi_p(u,t) = \|t\|_p + \langle u, t \rangle$, with $\langle \cdot, \cdot \rangle$ being the usual inner product, with $p^{-1} + q^{-1} = 1$. It is easy to see that $\hat{Q}_n(0)$ is the same as the ℓ_1-median; $\hat{\theta}_n$. The population analog of $\hat{Q}_n(u)$ is given by

$$Q_P(u) = \arg\min_{\xi \in \mathbb{R}^k} P\{\Phi_p(u, X - \xi) - \Phi_p(u, X)\}. \tag{3.2.2}$$

Chaudhuri (1996) showed that the geometric quantiles are both equivariant under location transformation and homogeneous scale transformation of the individual coordinates.

Chaudhuri's (1996) approach has received considerable attention in the literature, and has also been extended to regression contexts, for example, in Chakraborty (1999). One other notable approach to generalizing univariate quantiles to multivariate case is the directional quantile approach developed by Hallin et. al (2010). A directional quantile τ is a function of two components, namely, a direction vector u and a depth $\gamma \in (0,1)$. Then the $\tau = u\gamma$ directional quantile, denoted by λ_τ is a hyperplane through $\mathbb{R}^k$.

In Chapter 2, we have considered the use of a Dirichlet process prior on the underlying distribution P, and explored properties of the corresponding marginal posterior distribution of $Q_P(u)$. We studied the posterior distribution asymptotically and established a Bernstein-von Mises theorem. The asymptotic covariance matrix in their Bernstein-von Mises theorem is the same as that for the M-estimator and hence does not match with that in our Gibbs posterior approach. After scaling suitably with the *learning rate*, our approach seems to work as good as the non-parametric Bayesian approach, as revealed by the numerical results.

3.2.2 Gibbs posterior distributions

The Gibbs measure has its origins in statistical physics but a version of it has received attention in the statistics, machine learning, and econometrics literature; see, e.g., Bissiri et. al (2016), Zhang (2006; 2006), and Chernozhukov and Hong (2003). Some recent statistical applications include data mining (Jiang and Tanner (2008)), clinical trials (Syring and Martin (2017)), image analysis (Syring and Martin (2016)), actuarial science (Syring and Martin (2019)), and classifier performance assessment (Wang and Martin (2020)). Below we define the Gibbs posterior and some features that will be relevant in what follows.

Let $X^n = (X_1, \ldots, X_n)$ be an i.i.d. sample from some distribution P. Suppose there is some functional $\theta = \theta(P)$ that we are interested in estimating and making inference about. By the way this problem has been stated, it should be clear that θ generally cannot be understood as a model parameter, so we cannot expect that there is a likelihood function that can be used to connect the data to the quantity of interest. Instead, the setup assumes that the functional is defined via an optimization problem. That is, there exists a function $m_\theta(x)$ such that the true value $\theta^\star$ of $\theta(P)$ is the minimizer of the function $R(\theta) = P m_\theta$; here, note that, as is customary in the literature, we denote the quantity of interest and a generic value of it with the same symbol, θ, and distinguish the true value $\theta^\star$ where necessary. The function m_θ is called the *loss* and $R(\theta)$ the corresponding *risk*. Since we do not know P, we also do not know the risk, so inference on θ requires that we replace P with the observed data in some way. In particular, define the empirical risk as $R_n(\theta) = \mathbb{P}_n m_\theta$, where $\mathbb{P}_n$ is the empirical distribution of the data X^n. The estimator $\hat{\theta}_n$ derived by minimizing $R_n(\theta)$ is often called an *M-estimator* (Huber (1981)).

Empirical risk minimization is a common task in machine learning and can be challenging because the data-dependent objective function R_n is not always well-behaved. As an alternative to optimization, the *PAC-Bayes* literature (McAllester (1999), Alquier (2008))—where PAC stands for probability approximately correct—proposed to construct a distribution that concentrates on θ values for which $R_n(\theta)$ is small. That distribution is the Gibbs posterior and is given by

$$\Pi_n(B) = \frac{\int_B e^{-\omega n R_n(\theta)} \Pi(d\theta)}{\int_{\mathbb{R}^k} e^{-\omega n R_n(\theta)} \Pi(d\theta)}, \quad B \subseteq \mathbb{R}^k, \tag{3.2.3}$$

where Π is a prior distribution and $\omega > 0$ is called the *learning rate*. Like the prior, the learning rate is not determined by the problem at hand, so must be specified by the user. A number of learning rate selection methods have been proposed in the literature, e.g., Grünwald (2012), Homes and Walker (2017), Homes and Walker (2019), and Syring and

Martin (2019). In Section 3.3.3 we discuss our method of choice in detail.

3.3 Gibbs posteriors for multivariate quantiles

3.3.1 Definition

Suppose we have a iid sample $X_1, \ldots, X_n$ from a distribution P on $\mathbb{R}^k$. Since the ℓ_1-median $\theta(P)$ is same as $Q_P(0)$, we will discuss the Gibbs posterior construction for a geometric quantile $Q(u)$ with ℓ_p-norm for some fixed $p \in (1, \infty)$ and fixed $u \in B_q^{(k)}$. For simplicity, we will denote $\hat{\theta}_n = \hat{Q}_n(u)$ and $\theta^\star = Q_P(u)$ from now on.

Since the quantity of interest is the minimizer of a function defined by an expectation, in (3.2.2), it makes sense to define the loss as the loss $\psi(\cdot, \theta)$ as that function inside the expectation, namely, we take $m_\theta(x) = \Phi_p(u, x - \theta)$, where, again, u and p are fixed. Then the risk is $R(\theta) = P m_\theta$, minimized at $\theta^\star$, and the empirical risk is

$$R_n(\theta) = \mathbb{P}_n m_\theta = \frac{1}{n} \sum_{i=1}^n \{\|X_i - \theta\|_p + \langle u, X_i - \theta \rangle\}, \tag{3.3.1}$$

minimized at $\hat{\theta}_n$. Given a prior distribution Π for θ, the Gibbs posterior distribution Π_n is defined like in (3.2.3). It follows from Hölder's inequality that $R_n(\theta) \geq 0$, so the denominator in (3.2.3) is finite and the Gibbs posterior is well-defined. Therefore, Π_n is just an ordinary probability distribution and it can be summarized via the usual Markov chain Monte Carlo. For the theoretical analysis that follows, we assume that the learning rate ω is a fixed constant, but we will recommend a data-driving choice of ω in Section 3.3.3.

3.3.2 Asymptotic properties

First, we investigate the Gibbs posterior concentration rate, i.e., the radius of the smallest ball around $\theta^\star$ to which the posterior asymptotically assigns all of its mass, as $n \to \infty$. For this, we require a mild condition on the underlying distribution P.

Assumption 1. The distribution P admits a density f that is continuous and bounded away from 0 on a set $\mathbb{X} \subset \mathbb{R}^k$ containing $\theta^\star$.

Assumption 2. The density f is bounded on compact subsets of $\mathbb{R}^k$.

An important consequence (see the proof of Lemma 3.6.1 in Section 3.6) of Assumptions 1–2 is that the function R is twice differentiable at $\theta^\star$, where $\dot{R}(\theta^\star) = 0$ and $V_{\theta^\star} := \ddot{R}(\theta^\star)$

is positive definite; here, dot and double-dot correspond to first and second derivatives with respect to θ, the gradient vector and the Hessian matrix, respectively.

Assumption 3. The prior distribution Π has a density π which is continuous and bounded away from 0 in a neighborhood of $\theta^\star$.

Theorem 3.3.1. *Under Assumptions 1–3, if $\omega > 0$ is sufficiently small, then the Gibbs posterior Π_n for the multivariate quantile satisfies*

$$P^n \Pi_n(\{\theta \in \mathbb{R}^k : \|\theta - \theta^\star\|_2 > a_n n^{-1/2}\}) = o(1), \quad n \to \infty,$$

where $a_n \to \infty$ is any diverging sequence.

Proof. See Section 3.6.2. □

In the proof of Theorem 3.3.1, a bound on what it means for ω to be "sufficiently small" is obtained. The bound we obtain is not sharp, and it depends on unknown features of P, so we opt not to state that bound here. Rather than relying on analytic bounds on ω, we recommend a data-driven choice as described in Section 3.3.3.

Next, we will prove a Bernstein–von Mises theorem for the Gibbs posterior, that is, the Gibbs posterior can be approximated by a Gaussian distribution in a total variation sense as $n \to \infty$. Results of this type were considered in Chernozhukov and Hong (2003), but here we will follow the approach in Kleijn and van der Vaart (2012). Before formally stating this result, we need a bit more notation. The loss $\theta \mapsto m_\theta(x)$ can be differentiated for P-almost all x, and the j^{th} component of the gradient vector, $\psi(x, \theta)$, is given by

$$\psi_j(x, \theta) = \frac{|x_j - \theta_j|^{p-1}}{\|x - \theta\|_p^{p-1}} \operatorname{sign}(\theta_j - x_j) + u_j, \quad j = 1, \ldots, k,$$

where, again, p and u are fixed, and $\operatorname{sign}(\cdot)$ denotes the *signum* function. Now set $\Delta_{n,\theta^\star} = n^{-1/2} \sum_{i=1}^n V_{\theta^\star}^{-1} \psi(X_i, \theta^\star)$.

Theorem 3.3.2. *Under Assumptions 1–3, the sequence of centered and scaled Gibbs posteriors, with learning rate ω as in Theorem 3.3.1, approaches a k-variate normal distributions in total variation, i.e.,*

$$\sup_B \left| \Pi_n(\{\theta : n^{1/2}(\theta - \theta^\star) \in B\}) - \mathrm{N}_k(B \mid \omega \Delta_{n,\theta^\star}, (\omega V_{\theta^\star})^{-1}) \right| = o_P(1), \quad n \to \infty.$$

Proof. See Section 3.6.3. □

Like in van der Vaart (2000), it is possible to center the normal approximation about a different but asymptotically equivalent estimator. In particular, using the location shift invariance of the total variation distance, we can conclude from Theorem 3.3.2 that, when n is large, the Gibbs posterior Π_n is approximately $\mathrm{N}_k(\hat{\theta}_n, (\omega n V_{\theta^\star})^{-1})$. It is also worth mentioning that a similar theorem would hold in examples other than multivariate quantiles. What matters is that the empirical risk R_n satisfies a version of the *local asymptotic normality* condition, i.e., for every compact set $K \subset \mathbb{R}^k$,

$$\sup_{h \in K} \left| n\{R_n(\theta^\star + h n^{-1/2}) - R_n(\theta^\star)\} - h^\top V_{\theta^\star} \Delta_{n,\theta^\star} - \tfrac{1}{2} h^\top V_{\theta^\star} h \right| = o_P(1). \tag{3.3.2}$$

This property is critical to our proof of Theorem 3.3.2, so we can expect similar conclusions in any other problem for which (3.3.2) holds.

In a Bayesian setting, with a regular, well-specified model, a Bernstein–von Mises theorem ensures that inferences derived from the Bayesian posterior distribution are valid in a frequentist sense. For example, a $100(1-\alpha)\%$ posterior credible set will have frequentist coverage probability approximately equal to $1-\alpha$ for large n. The reason for this Bayesian–frequentist connection is that the posterior distribution centers around, in that case, the maximum likelihood estimator, and the covariance matrix in the Bernstein–von Mises theorem is the inverse Fisher information matrix, which agrees with the asymptotic covariance matrix of the maximum likelihood estimator. However, when the model is misspecified, like in Kleijn and van der Vaart (2012), or, like here, where no model is specified at all, then this covariance matching is not guaranteed. Indeed, in our present case, the covariance matrix in the normal approximation to the Gibbs posterior is $(\omega V_{\theta^\star})^{-1}$ whereas the asymptotic covariance matrix of the M-estimator $\hat{\theta}_n$ is

$$\Gamma := V_{\theta^\star}^{-1} U_{\theta^\star} V_{\theta^\star}^{-1}, \tag{3.3.3}$$

where $U_{\theta^\star} = P\psi(\cdot, \theta^\star)\psi(\cdot, \theta^\star)^T$. The above equation comes from the familiar sandwich formula. Since these two matrices are generally different, our Bernstein–von Mises theorem does not guarantee that inference drawn from the Gibbs posterior are valid in a frequentist sense.

If we were in a traditional Bayesian setting and were unfortunate enough that our model was sufficiently misspecified that we get the aforementioned covariance matrix mismatch, then (a) we typically would not be aware of this serious problem and (b) there would be nothing we could do about it anyway, aside from starting over with a different model. However, since we are working within a Gibbs framework, we are aware of and openly acknowledge that our posterior distribution is based on an effectively misspecified model

and, moreover, we have a potential remedy: adjusting the learning rate.

It is easy to see that if $U_{\theta^\star} \propto V_{\theta^\star}$, i.e., if the *generalized information equality* (Chernozhukov and Hong (2003)) holds, then the covariance matrix in the normal approximation to the Gibbs posterior will be proportional to the asymptotic covariance matrix of the M-estimator $\hat{\theta}_n$. In that case, we can exactly correct for the covariance mismatch simply by tuning the learning rate. In general, however, simply tuning the scalar learning rate parameter cannot fully correct for the covariance mismatch, but it is still possible to find a learning rate such that credible sets derived from the Gibbs posterior have approximately the nominal frequentist coverage probability; see Section 3.3.3.

3.3.3 Choice of the learning rate

As we indicated in Section 3.2, the choice of learning rate is critical to the performance of methods derived from a Gibbs posterior distribution. This is especially important in our present situation because, as mentioned in the remarks following Theorem 3.3.2, the Gibbs posterior does not inherit the correct asymptotic shape. This covariance mismatch is a common occurrence when a Bayesian model is misspecified but, unlike the traditional Bayesian setting where nothing can be done to overcome the misspecification bias, the Gibbs posterior has a learning rate that can be suitably chosen to correct for the mismatched asymptotic covariance matrix.

More specifically, following Syring and Martin (2019), if Π_n^ω denotes the Gibbs posterior with learning rate ω, then we aim to choose ω such that the frequentist coverage probabilities of credible sets derived from Π_n^ω are approximately equal to the nominal level. That is, for a desired significance level $\alpha \in (0,1)$, if $C_{\omega,\alpha}(X^n)$ denotes a $100(1-\alpha)\%$ credible set from the Gibbs posterior Π_n^ω, then the coverage probability is

$$c_\alpha(\omega; P) = P\{C_{\omega,\alpha}(X^n) \ni \theta(P)\},$$

i.e., the P-probability that $C_{\omega,\alpha}(X^n)$ contains $\theta(P)$. Of course, if P were known, then it would be possible to approximate the coverage probability using Monte Carlo and solve the equation, $c_\alpha(\omega; P) = 1-\alpha$, using stochastic approximation (Robbins and Monro (1951)). Since P is unknown in practice, Syring and Martin (2019) recommend a bootstrap version that replaces P with the empirical distribution, $\mathbb{P}_n$. We use their *Gibbs posterior calibration* (GPC) algorithm (see Algorithm 1) for choosing the learning rate, which performs well in our experiments below. For a quick visual illustration, consider a bivariate case, $d = 2$. Suppose we have $n = 100$ samples from a bivariate normal distribution as in Example 1 in Section 3.4.1. Using a relatively flat $\mathrm{N}_2(0, 10I_2)$ prior, and with the learning rate chosen

Algorithm 1 — Gibbs Posterior Calibration (Syring and Martin (2019))

Fix a convergence tolerance $\epsilon > 0$ and an initial value $\omega^{(0)}$ of the learning rate ω. Take B bootstrap samples $\tilde{X}_1^n, \ldots, \tilde{X}_B^n$ of size n. Set $t = 0$ and do the following.

1. Construct $100(1-\alpha)\%$ credible set $C_{\omega^{(t)},\alpha}(\tilde{X}_b^n)$ for every $b = 1, \ldots, B$.

2. Evaluate the bootstrap estimate

$$\hat{c}_\alpha(\omega^{(t)}, \mathbb{P}_n) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\big\{ C_{\omega^{(t)},\alpha}(\tilde{X}_b^n) \ni \hat{\theta}_n \big\}$$

   of the empirical coverage probability $c_\alpha(\omega^{(t)}, \mathbb{P}_n)$.

3. If $|\hat{c}_\alpha(\omega^{(t)}, \mathbb{P}_n) - (1-\alpha)| < \epsilon$, then return $\omega^{(t)}$ as the output, else update $\omega^{(t)}$ to $\omega^{(t+1)}$ as

$$\omega^{(t+1)} = \omega^{(t)} + \kappa_t \{ \hat{c}_\alpha(\omega^{(t)}, \mathbb{P}_n) - (1-\alpha) \},$$

   with $\kappa_t = (t+1)^{-0.51}$, set $t \leftarrow t+1$, and go back to Step 1.

according to Algorithm 1, samples from the corresponding Gibbs posterior distribution are shown in Figure 3.1(a). The same is shown in Figure 3.1(b), except where the data are sampled from a bivariate Laplace distribution as in Example 2 of Section 3.4.1. In addition to the Gibbs posterior samples, we also display the 95% credible region based on the normal approximation. That is, we first compute the posterior mean $\bar{\theta}$ and covariance matrix S based on the Monte Carlo samples, and then find the 95th percentile of the marginal posterior distribution for $\vartheta \mapsto (\vartheta - \bar{\theta})^\top S^{-1} (\vartheta - \bar{\theta})$, denoted by $r_{0.95}$. Then the 95% Gibbs posterior credible set is

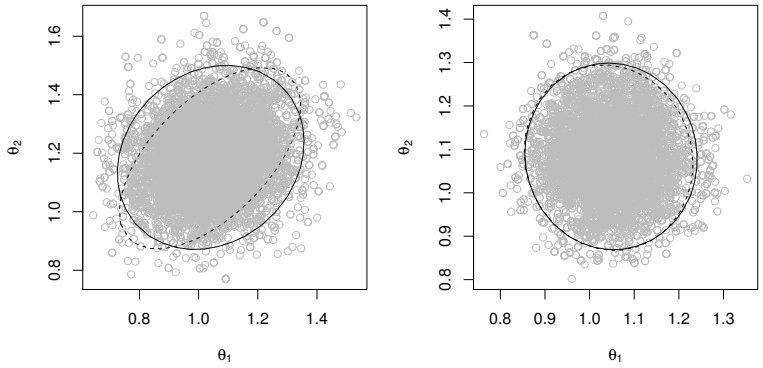
$$\{\vartheta : (\vartheta - \bar{\theta})^\top S^{-1} (\vartheta - \bar{\theta}) \leq r_{0.95}\}.$$

Similarly, we compute the 95% confidence ellipse based on the asymptotic normality of the M-estimator/spatial median, namely,

$$\{\vartheta : (\vartheta - \hat{\theta}_n)^\top \widehat{\Gamma}_n (\vartheta - \hat{\theta}_n) \leq \chi^2_{2;.95}\},$$

where $\widehat{\Gamma}_n$ is a plug-in estimator of Γ in (3.3.3) and $\chi^2_{2;.95}$ is the 95th percentile of the chi-square distribution with 2 degrees of freedom. The boundaries of these two ellipses are overlaid on the plots of the Gibbs posterior samples. Clearly, in both cases, the contours of the Gibbs posterior are not of the same shape as the M-estimator confidence ellipse, a consequence of the covariance mismatch. However, by choosing the learning rate according to Algorithm 1, which is aiming to achieve the nominal 95% frequentist coverage rate, the Gibbs posterior

credible ellipse is stretched to roughly match the confidence ellipse in the direction in which it is widest. And since the confidence ellipse achieves the nominal frequentist coverage probability, at least asymptotically, the Gibbs posterior credible ellipse will too. Of course, there is some loss of efficiency due to the covariance mismatch—which is the price one pays for a model-free posterior distribution—but, as the simulation results in Section 3.4.1 show, this loss of efficiency is not severe. In fact, in some cases, the Gibbs posterior credible regions are more efficient than the credible sets from other Bayesian approaches.



(a) Bivariate normal (b) Bivariate Laplace

Figure 3.1 Gibbs posterior samples (gray), with learning rate chosen according to Algorithm 1, along with the 95% posterior credible region (solid) and the corresponding M-estimator confidence region (dashed).

3.4 Numerical results

3.4.1 Simulation study

In this section, we illustrate the finite sample performance of the Gibbs posterior of bivariate ℓ_1-medians, i.e., for $k = 2$, for ℓ_2 and ℓ_3 norms. We would like to compare the Gibbs posterior's performance to that of both parametric and non-parametric Bayesian methods in situations when the components of the vector are correlated, and when the data has outliers. Aside from the k-variate normal distribution $\mathrm{N}_k(\mu, \Sigma)$ with mean vector μ and covariance matrix Σ, we also consider a k-variate Laplace distribution, denoted by $\mathrm{Lap}_k(\mu, \Sigma)$, with location vector μ and dispersion matrix Σ, with density for the standardized version,

with $\mu = 0$ and $\Sigma = I_k$,

$$f(x) \propto \|x\|_2^{-(k-1)/2} e^{-2^{3/2}\|x\|_2}, \quad x \in \mathbb{R}^k.$$

Also, $\text{Gamma}_k(s, r, V)$ denotes a k-variate gamma distribution with shape s, rate r, and correlation matrix V, constructed using a Gaussian copula (Xue-Kun Song (2000)). The three specific examples we consider are as follows, each with sample size $n = 100$.

Example 1. $P = \text{N}_2(\mu, \Sigma)$, where $\mu = (1,1)^\top$, $\Sigma_{11} = \Sigma_{22} = 1$ and $\Sigma_{12} = 0.7$.

Example 2. $P = \text{lap}_2(\mu, \Sigma)$, where $\mu = (1,1)^\top$ and $\Sigma = I_2$.

Example 3. $P = \text{Gamma}_2(1, 1, V)$, where $V_{11} = V_{22} = 1$ and $V_{12} = 0.5$.

In each case, we consider a bivariate normal prior for θ, namely, $\text{N}_2((0,0)^\top, 10I_2)$. We use a Metropolis–Hastings algorithm with transition kernel $Q(y \mid x) = \text{N}_2(y \mid x, 0.01I_2)$ for drawing samples from the posterior distribution. The relevant summaries would be size and frequentist coverage of the 95% credible ellipses. We compare the performance of the Gibbs posterior to that of both parametric and non-parametric Bayesian methods. We would consider the following parametric Bayesian model:

$$(X_1, \ldots, X_n) | \theta \overset{iid}{\sim} \text{N}_2(\theta, \sigma^2 I_2) \quad \text{and} \quad \theta \sim \text{N}_2((0,0)^\top, 10I_2) \quad \sigma^{-2} \sim \text{Gamma}(1,1).$$

This parametric Bayesian posterior is quite simple, and we can use a Gibbs sampler for posterior inference. While the model and corresponding analysis is simple, the concern is potential model misspecification bias. As an alternative, to avoid these potential biases, one might consider a nonparametric Bayesian formulation. As suggested in Chapter 2, assume

$$(X_1, \ldots, X_n) \mid P \overset{iid}{\sim} P \quad \text{and} \quad P \sim \text{DP}(\alpha),$$

where $\text{DP}(\alpha)$ denotes a Dirichlet process distribution with base or centering measure α (Ghosal and van der Vaart (2017)). Here we choose $\alpha = 2 \times \text{N}_2((0,0)^\top, I_2)$. It is well known that the Dirichlet process prior is conjugate, so the posterior distribution for P, given X^n, is also a Dirichlet process, which is relatively simple to work with. However, the quantity of interest is $\theta = \theta(P)$, a functional of P, so some non-trivial marginalization is required. Specifically, we sample P from the Dirichlet process posterior distribution, and then evaluate θ as the minimizer of $\xi \mapsto P\{\|X - \xi\|_p - \|X\|_p\}$.

To compare the three methods described above, we consider a measure of size and also the frequentist coverage probability of the 95% posterior credible sets. Suppose that we have samples $\theta_1, \ldots, \theta_M$ from any one of the three posterior distributions, where $M = 5000$

Table 3.1 Estimated coverage probability and mean size (in parentheses) of 95% credible ellipses for the Gibbs, parametric, and nonparametric Bayes posterior distributions.

Example	p	Gibbs	Parametric Bayes	Nonparametric Bayes
1	2	0.950 (0.42)	0.900 (0.33)	0.975 (0.37)
	3	0.955 (0.40)	0.900 (0.33)	0.975 (0.38)
2	2	0.950 (0.21)	0.925 (0.34)	0.920 (0.19)
	3	0.960 (0.20)	0.925 (0.34)	0.970 (0.21)
3	2	0.950 (0.31)	0.926 (0.35)	0.950 (0.40)
	3	0.965 (0.21)	0.910 (0.35)	0.949 (0.27)

is the Monte Carlo sample size. The 95% credible set is given by

$$\{\vartheta : (\vartheta - \bar{\theta})^\top S^{-1} (\vartheta - \bar{\theta}) \leq r_{0.95}\},$$

where $\bar{\theta}$ and S are the Monte Carlo sample mean and covariance matrix, respectively, and $r_{0.95}$ is the 95th percentile of the values $(\theta_j - \bar{\theta})^\top S^{-1} (\theta_j - \bar{\theta})$, $j = 1, \ldots, M$. The coverage probability is defined as usual and, as a measure of the credible set's size, we use $|S| r_{0.95}^k$. Table 3.1 summarizes the size and coverage probability over 2000 replications for each of the three examples. It can be seen that the Gibbs posterior performs well compared to the parametric and non-parametric Bayesian approaches in terms of both size and coverage. Indeed, the Gibbs posterior has at least the nominal 95% coverage in every scenario, while the parametric and non-parametric Bayesian credible sets occasionally miss the target coverage, especially the parametric solution. The coverage performance may not be surprising, given that we tuned the learning rate of the Gibbs posterior to achieve the nominal coverage. However, it is interesting to see that the coverage guarantees do not come with a loss of efficiency—in some cases, the Gibbs posterior credible sets are even more efficient.

3.4.2 Real-data analysis

The Egyptian skulls dataset (Hand et al. 1994) consists of $k = 4$ measurements—namely, maximal breadth, basibregmatic height, basialveolar length, and nasal height—taken on $n = 150$ ancient Egyptian skulls from five time epochs; these data are available in the `HSAUR` package in R. The mean effect of time was removed by fitting a linear model and extracting the residuals, and we take these 4-dimensional residual vectors as our data X^n. A first thought would be to assume multivariate normality and carry out a standard analysis. However, formal tests of normality are conflicting: marginal tests of normality reject while

tests of multivariate normality do not reject (Tokdar and Martin (2019)). So, we proceed with the construction of a Gibbs posterior that does not require us to decide about normality in this difficult case.

For simplicity, here we will focus on the median of the 4-dimensional distribution, although other quantiles could be handled similarly. Since these data already have time trends removed, we expect that the distribution's center should be roughly near the origin, so we take a normal prior with zero mean, but with covariance matrix $10I_4$. Since we are only looking at a finite sample situation, we expect some non-elliptical shape in the Gibbs posterior, and the goal of this analysis is to investigate that shape. Figure 3.2 summarizes the marginal and pairwise Gibbs posterior distributions after scaling the learning rate according to Algorithm 1. As expected, even on these limited low-dimensional summaries, the Gibbs posterior does not appear to be exactly normal, but the results are quite reasonable. For comparison, we also show the marginal plug-in densities and 95% pairwise confidence ellipses based on the asymptotically normal sampling distribution of the M-estimator, the sample spatial median. Clearly, these margins of the Gibbs posterior are centered in roughly the correct place but, most importantly, and thanks to the calibration framework in Algorithm 1, the Gibbs posterior spread tends to be wider in some directions than that of the M-estimator sampling distribution. While some might view this wider spread as an indication of some inefficiency, we believe the wider spread is necessary for valid uncertainty quantification in finite samples, not just in the idealistic $n \to \infty$ case.

To compare the performance of our Gibbs posterior with a Bayes solution that assumes normality, we look at the posterior of the empirical risk evaluated on a held-out testing set. That is, we split the data into a training and testing set—the first 100 samples are training and the last 50 are testing—construct a posterior distribution for the median θ using the training data, then evaluate the corresponding posterior distribution for $R_{\text{test}}(\theta)$, where R_{test} is the empirical risk function in (3.3.1) based on the testing data set only. We also compare our method with the non-parametric Bayes method discussed in the previous section. Figure 3.3 plots (kernel density estimates of) the posterior distribution of log empirical risk difference,

$$\log\{R_{\text{test}}(\theta) - \min_{\vartheta} R_{\text{test}}(\vartheta)\}, \tag{3.4.1}$$

based on the testing data. We follow the same Gibbs formulation as above; for the Bayes solution, we assume $P = \text{N}_4(\theta, \Sigma)$ and use the conjugate normal–inverse Wishart prior for (θ, Σ). The figure shows that the Gibbs posterior distribution of the log empirical risk difference is centered to left of that for Bayes, which is an indication that the former is more concentrated around θ values that make the out-of-sample risk small than the latter. Since the training and testing data are not fundamentally different, this suggests that there is

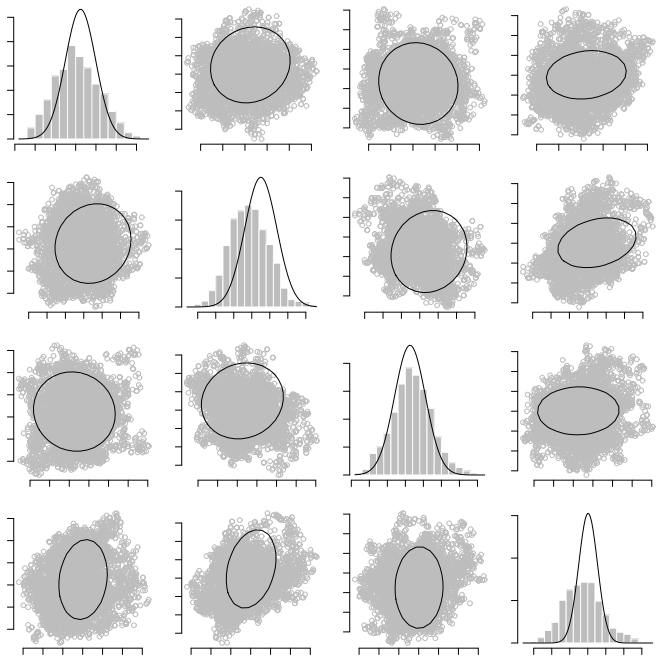


Figure 3.2 Marginal and pairwise Gibbs posterior distributions (gray) for the 4-dimensional spatial median of the Egyptian skull data described in Section 3.4.2. Overlaid (black) are approximate marginal sampling distributions and 95% pairwise confidence ellipses based on asymptotic normality of the M-estimator.

some bias created by the assumption of multivariate normality made by the parametric Bayesian solution. But clearly, the winner is the non-parametric Bayesian method, whose out-of-sample risk is centered to the left most position out of all three.

3.5 Concluding remarks

In this chapter, we have studied multivariate medians and quantiles in a Gibbs posterior framework. Our approach does not need a model and is free from the potential issues that may arise in a model-based parametric Bayesian approach, in particular, model misspecification bias. The Gibbs posterior is simple to use and is also theoretically justified in the sense that the posterior concentrates around the true multivariate quantile at the optimal $n^{-1/2}$ rate, and has a Bernstein–von Mises property, i.e., it can be approximated by a suitable Gaussian distribution centered at the sample spatial median. We also pointed out that this Gaussian approximation holds in other problems, not just multivariate quantiles, provided that the empirical risk satisfies a version of the local asymptotic normality property.

A unique feature of the Gibbs posterior is its dependence on the choice of learning

rate. On one hand, this dependence might seem like a disadvantage, since the learning rate is not determined by the context of the problem, and there is no universally accepted choice. On the other hand, as we argued here, being able to choose the learning rate is an advantage in the sense that it provides the flexibility necessary to at least partially correct for the covariance mismatch in the Gaussian approximation discussed in Section 3.3.2. Here we recommend the data-driven learning rate selection procedure of Syring and Martin (2019), summarized in Algorithm 1, as it aims to set ω so that the Gibbs posterior credible region achieves the nominal frequentist coverage probability.

While the learning rate selection procedure described in Algorithm 1 works well empirically, there are still some opportunities for improvement and unanswered questions. The main disadvantage of this strategy is having to do multiple Monte Carlo runs on each bootstrap sample; this could be improved by carrying out some of the steps in parallel. In terms of open questions, there so far is no theory to support the claim that choosing the learning rate according to Algorithm 1 will, as advertised, produce credible sets that achieve the nominal frequentist coverage. Beyond that, there is not even any guarantee that the algorithm's choice of ω satisfies conditions required to achieve the $n^{-1/2}$ concentration rate. In Theorem 3.3.1, ω should be smaller than a multiple of the smallest eigenvalue of $V_{\theta\star}$. This is only a sufficient condition—not necessary—but in our numerical experiments, our choice of ω actually does not satisfy the theorem's condition.

There are a couple of possible extensions of the work presented herein:

- It would be interesting to explore cases where the dimension d exceeds the sample size, i.e., a so-called "high-dimensional setting," with $k \gg n$. For such cases, we would need to assume some low-dimensional structure in the high-dimensional $\theta^\star$, and then specify a prior distribution that would encourage this structure. Sparsity-inducing priors [Pas, Szabó & A. v. d. Vaart 2017; Castillo & A. v. d. Vaart 2012; Martin & Ning 2020] have been popular in recent years, and one of the examples in Syring and Martin (2020) shows that this kind of sparsity can be readily handled within the Gibbs framework, but the details for a sparse, high-dimensional multivariate quantile have yet to be worked out.

- The Gibbs posterior approach can also be used in multivariate quantile regression. Consider a linear regression set-up with a k-variate response vector y and a q-dimensional regressor x satisfying the linear model $y = \beta^\top x + e$, with β a $q \times k$ matrix of regression coefficients. For $u \in B_q^{(k)}$, and a sample (x_i, y_i) of response and

regressor pairs, the u^{th} sample geometric quantile of y given x is obtained as

$$Q_{y|x}(u) = \arg\min_{\xi \in \mathbb{R}^k} \frac{1}{n} \sum_{i=1}^{n} \{\|y_i - \beta^\top x_i\|_p + \langle u, y_i - \beta^\top x_i \rangle\}.$$

In a typical Bayesian approach, we would have to choose a model for the errors such that its quantile agrees with the target quantile; of course, there are many such models, so having to make such a choice puts the data analyst at risk of model misspecification bias. On the other hand, it is easy to formulate a Gibbs posterior framework, which is free of such modeling and the associated risks. We expect that the theoretical results for the Gibbs posterior presented herein would carry over to this more general setting, but we have yet to verify this conjecture.

3.6 Proofs

3.6.1 Preliminary results

Recall that $m_\theta(x) = \|x-\theta\|_p + \langle u, x-\theta \rangle$, and $R(\theta) = P m_\theta$, and the second derivative matrix is given by $V_{\theta^\star}$. First, we want to bound

$$\inf_{\|\theta-\theta^\star\|_2 > \delta} R(\theta) - R(\theta^\star). \tag{3.6.1}$$

Lemma 3.6.1. *Under Assumptions 1–2, there exists a constant $c > 0$ such that* (3.6.1) *is lower-bounded by $c\,\delta^2$ for all sufficiently small $\delta > 0$.*

Proof. The function R is twice-differentiable at $\theta^\star$, and the second derivative matrix is given by $V_{\theta^\star} = \ddot{R}(\theta^\star) = \int \dot{\psi}_{x,\theta^\star} \, \mathrm{d}P$, where

$$\dot{\psi}_{x,\theta^\star} = \frac{p-1}{\|x-\theta^\star\|_p} \Big[\operatorname{diag}\left(\frac{|x_1 - \theta_1^\star|^{p-2}}{\|x-\theta^\star\|_p^{p-2}}, \ldots, \frac{|x_k - \theta_k^\star|^{p-2}}{\|x-\theta^\star\|_p^{p-2}} \right) - \psi(x,\theta^\star)\psi(x,\theta^\star)^T \Big].$$

The existence of $V_{\theta^\star}$ can be verified using Assumption 2, i.e., for a fixed $\theta \in \mathbb{R}^k$, if P has a density f that is bounded on compact subsets of $\mathbb{R}^k$, then the expectation of $\|X-\theta\|_p^{-1}$ is finite. Since the first derivative vanishes at $\theta^\star$, the Taylor expansion takes the form

$$R(\theta) - R(\theta^\star) = \tfrac{1}{2}(\theta-\theta^\star)^\top V_{\theta^\star}(\theta-\theta^\star) + o(\|\theta-\theta^\star\|_2^2).$$

Since $V_{\theta^\star}$ is positive definite, the proof follows by taking $c = \frac{1}{2}\lambda_{\min}(V_{\theta^\star})$, where $\lambda_{\min}(M)$ returns the smallest eigenvalue of the matrix M. □

Now, rewrite the Gibbs posterior distribution as

$$\Pi_n(A) = \frac{N_n(A)}{D_n} = \frac{\int_A e^{-\omega n\{R_n(\theta)-R_n(\theta^\star)\}}\,\Pi(d\theta)}{\int e^{-\omega n\{R_n(\theta)-R_n(\theta^\star)\}}\,\Pi(d\theta)}, \quad A \subseteq \mathbb{R}^k.$$

Below we will investigate the limiting behavior of $\Pi_n(A)$ for two kinds of sets A: the first is for a fixed $A = \mathbb{X}^c$, where $\mathbb{X}$ is defined in Assumption 1, and the second is for a sequence of suitably shrinking sets A_n to be defined below.

Lemma 3.6.2. *Under Assumptions 1–2, for any $\omega > 0$, the Gibbs posterior Π_n satisfies $P^n\Pi_n(\mathbb{X}^c) \to 0$ as $n \to \infty$.*

Proof. For the denominator D_n of the Gibbs posterior, we can proceed as we do in Lemma 3.6.3 below and conclude that $D_n > e^{-bn}$ with P^n-probability converging to 1, for any $b > 0$; compare this to Lemma 4.4.1 in Ghosh and Ramamoorthi (2003). For the numerator $N_n(\mathbb{X}^c)$, without loss of generality, assume $\mathbb{X} = \{x : \|x - \theta^\star\|_2 \leq K\}$. Then it is easy to check that $\theta \mapsto R_n(\theta)$ is almost surely convex, from which it follows that, if $\theta \in \mathbb{X}^c$, then

$$R_n(\theta) - R_n(\theta^\star) \geq \inf_u \{R_n(\theta^\star + Ku) - R_n(\theta^\star)\},$$

where the infimum is over all unit vectors u. By the law of large numbers,

$$R_n(\theta^\star + Ku) - R_n(\theta^\star) \to R(\theta^\star + Ku) - R(\theta^\star), \quad \text{for all } u.$$

Moreover, by convexity of R_n and Lemma 1 in Hjort and Pollard (2011), we get

$$\inf_u \{R_n(\theta^\star + Ku) - R_n(\theta^\star)\} \to \Delta \quad \text{in } P^n\text{-probability as } n \to \infty,$$

where $\Delta = \inf_u\{R(\theta^\star + Ku) - R(\theta^\star)\}$ is strictly positive. Then the event $N_n(\mathbb{X}^c) \leq e^{-\omega n(\Delta/2)}$ is implied by

$$\inf_u \{R_n(\theta^\star + Ku) - R_n(\theta^\star)\} > \Delta/2,$$

which has P^n-probability converging to 1. If we take $b > 0$ in the D_n bound described above to be less than $\omega\Delta/2$, then

$$\Pi_n(\mathbb{X}^c) = \frac{N_n(\mathbb{X}^c)}{D_n} \leq e^{-n(\omega\Delta/2 - b)} \to 0 \quad \text{in } P^n\text{-probability.}$$

Since $\Pi_n(\mathbb{X}^c)$ bounded (by 1) and converges to 0 in probability, it follows from the dominated convergence theorem that $P^n\Pi_n(\mathbb{X}^c) \to 0$, which completes the proof. □

Recall that the events of interest in Theorem 3.3.1 are given by $A_n = \{\theta : \|\theta - \theta^\star\|_2 > a_n \epsilon_n\}$, where $\epsilon_n = n^{-1/2}$. To prove that theorem, we need to show that $\Pi_n(A_n) \to 0$ in expectation. Our strategy is to find a lower bound on D_n and an upper bound on $N_n(A_n)$ such that the ratio of these two bounds is vanishing. The next two lemmas accomplish each these two goals in turn.

Lemma 3.6.3. *Under the conditions of Theorem 3.3.1, there exists a constant H such that, for any $\eta > 0$, $P^n(D_n \leq H\epsilon_n^k) \leq \eta$.*

Proof. Define the set $K_n = \{\theta : r(\theta) \vee v(\theta) \leq \epsilon_n^2\}$, where $a \vee b = \max(a, b)$, and

$$r(\theta) = R(\theta) - R(\theta^\star) \quad \text{and} \quad v(\theta) = P\{(m_\theta - m_{\theta^\star}) - r(\theta)\}^2,$$

are the mean and variance of the loss difference, respectively. From the Lipschitz property of the loss and the Taylor approximation in the proof of Lemma 3.6.1, it follows that

$$\|\theta - \theta^\star\|_2 \lesssim \epsilon_n \implies r(\theta) \vee v(\theta) \leq \epsilon_n^2.$$

Therefore, by Assumption 3,

$$\Pi(K_n) \geq \Pi(\{\theta : \|\theta - \theta^\star\|_2 \lesssim \epsilon_n\}) \gtrsim \epsilon_n^k.$$

Next, define a standardized version of the empirical risk difference, i.e.,

$$Z_n(\theta) = \frac{n\{R_n(\theta) - R_n(\theta^\star)\} - n\, r(\theta)}{\{n\, v(\theta)\}^{1/2}}.$$

This is a function of both θ and the data X^n, so define the upper level sets

$$\mathscr{X}_n = \{(\theta, X^n) : |Z_n(\theta)| \geq M\},$$

where $M > 0$ is arbitrary, and also the cross-sections

$$\mathscr{X}_n(\theta) = \{X^n : (\theta, X^n) \in \mathscr{X}_n\} \quad \text{and} \quad \mathscr{X}_n(X^n) = \{\theta : (\theta, X^n) \in \mathscr{X}_n\}.$$

Since we have

$$n\{R_n(\theta) - R_n(\theta^\star)\} = n\, r(\theta) + \{n\, v(\theta)\}^{1/2} Z_n(\theta),$$

and r, v, and Z_n are suitably bounded on $K_n \cap \mathscr{X}_n(X^n)^c$, we immediately get

$$D_n \geq \int_{K_n \cap \mathscr{X}_n(X^n)^c} e^{-\omega n r(\theta) - \omega\{nv(\theta)\}^{1/2} Z_n(\theta)} \Pi(d\theta) \geq e^{-2\omega n \epsilon_n^2} \Pi\{K_n \cap \mathscr{X}_n(X^n)^c\}.$$

From this lower bound, we have

$$\begin{aligned} P^n\{D_n \leq \tfrac{1}{2}\Pi(K_n) e^{-2\omega n\epsilon_n^2}\} &\leq P^n\big[e^{-2\omega n\epsilon_n^2}\Pi\{K_n \cap \mathscr{X}_n(X^n)^c\} \leq \tfrac{1}{2}\Pi(K_n) e^{-2\omega n\epsilon_n^2}\big] \\ &= P^n\big[\Pi\{K_n \cap \mathscr{X}_n(X^n)\} \geq \tfrac{1}{2}\Pi(K_n)\big] \\ &\leq \frac{2P^n\Pi\{K_n \cap \mathscr{X}_n(X^n)\}}{\Pi(K_n)}, \end{aligned}$$

where the last line is by Markov's inequality. Now use Fubini's theorem,

$$\begin{aligned} P^n\Pi\{K_n \cap \mathscr{X}_n(x^n)\} &= \int\int \mathbb{1}\{\theta \in K_n \cap \mathscr{X}_n(x^n)\}\,\Pi(d\theta)\,P^n(dx^n) \\ &= \int\int \mathbb{1}\{\theta \in K_n\}\,\mathbb{1}\{\theta \in \mathscr{X}_n(x^n)\}\,P^n(dx^n)\Pi(d\theta) \\ &= \int_{K_n} P^n\{\mathscr{X}_n(\theta)\}\,\Pi(d\theta). \end{aligned}$$

By Chebyshev's inequality, $P^n\{\mathscr{X}_n(\theta)\} \leq M^{-1}$, and hence

$$P^n\{D_n \leq \tfrac{1}{2}\Pi(K_n) e^{-2\omega n\epsilon_n^2}\} \leq 2M^{-1}.$$

Putting everything together, since $\Pi(K_n) \gtrsim \epsilon_n^k$ and $n\epsilon_n^2 = 1$, we have that

$$P^n(D_n \leq \tfrac{1}{2} e^{-2\omega}\epsilon_n^k) \leq 2M^{-1}.$$

Finally, set $H = \frac{1}{2}e^{-2\omega}$ and choose M so large that $2M^{-1} \leq \eta$. □

Lemma 3.6.4. *Under the conditions of Theorem 3.3.1, $P^n N_n(A_n \cap \mathbb{X}) = o(\epsilon_n^k)$ as $n \to \infty$.*

Proof. Define the empirical process $\mathbb{G}_n f = n^{1/2}(\mathbb{P}_n f - Pf)$ and write

$$R_n(\theta) - R_n(\theta^\star) = R(\theta) - R(\theta^\star) + n^{-1/2}\mathbb{G}_n(m_\theta - m_{\theta^\star}).$$

Then the Gibbs posterior numerator at $A_n \cap \mathbb{X}$ can be decomposed as a sum of integrals

over "shells" as follows:

$$\begin{aligned} N_n(A_n \cap \mathbb{X}) &= \sum_{t=1}^{T_n} \int_{t\,a_n\epsilon_n < \|\theta-\theta^\star\|_2 < (t+1)a_n\epsilon_n} e^{-\omega n\{R_n(\theta)-R_n(\theta^\star)\}}\,\Pi(d\theta) \\ &\le \sum_{t=1}^{T_n} e^{-c\,\omega t^2 a_n^2} \int_{\|\theta-\theta^\star\|_2 < (t+1)a_n\epsilon_n} e^{-\omega n^{1/2}\mathbb{G}_n(m_\theta - m_{\theta^\star})}\,\Pi(d\theta), \end{aligned}$$

where $T_n \to \infty$ is to account for the intersection with $\mathbb{X}$. The exponential term outside the integral results from the bound in Lemma 3.6.1 and the fact that $n\epsilon_n^2 = 1$. Taking expectation of the left-hand side and moving it under the sum and under the integral on the right-hand side, we find that our task is to bound

$$\int_{\|\theta-\theta^\star\|_2 < (t+1)a_n\epsilon_n} P e^{-\omega n^{1/2}\mathbb{G}_n(m_\theta - m_{\theta^\star})}\,\Pi(d\theta), \quad t = 1, \ldots, T_n.$$

By Hoeffding's lemma, if Z is a random variable with mean 0, bounded in $[a, b]$, then its moment generating function satisfies

$$P e^{sZ} \le e^{s^2(b-a)^2/8}, \quad \text{for all } s.$$

In our case, of course, $m_\theta - m_{\theta^\star} - P(m_\theta - m_{\theta^\star})$ has mean 0 and, since $\theta \mapsto m_\theta$ is Lipschitz, $m_\theta - m_{\theta^\star}$ is bounded between $\pm\|\theta - \theta^\star\|_2$. This implies $m_\theta - m_{\theta^\star} - P(m_\theta - m_{\theta^\star})$ is also bounded, and its interval of support is no wider than $2\|\theta - \theta^\star\|_2$. Using independence and the above moment generating function bound, we get

$$P^n e^{-\omega n^{1/2}\mathbb{G}_n(m_\theta - m_{\theta^\star})} \le e^{\omega^2 n\|\theta-\theta^\star\|_2^2/2}.$$

Now plug this upper bound into the above integral:

$$\begin{aligned} P^n N_n(A_n \cap \mathbb{X}) &\le \sum_{t=1}^{T_n} e^{-c\,\omega t^2 a_n^2} e^{\omega^2(t+1)^2 a_n^2/2} \Pi(\{\theta : \|\theta - \theta^\star\|_2 < (t+1)a_n\epsilon_n\}) \\ &\lesssim (a_n\epsilon_n)^k \sum_{t=1}^{\infty} e^{-\omega a_n^2 t^2\{c - \frac{\omega}{2}(1+t^{-1})^2\}}(t+1)^k. \end{aligned}$$

If $\omega < \frac{c}{2}$, then the above sum is finite for all n and $\lesssim e^{-\omega a_n^2}$. Then a_n^k times the sum is vanishing as $n \to \infty$ and, consequently, we find that $P^n N_n(A_n \cap \mathbb{X}) = o(\epsilon_n^k)$. □

3.6.2 Proof of Theorem 3.3.1

To prove Theorem 3.3.1, we need to combine Lemmas 3.6.2, 3.6.3, and 3.6.4. Towards this, write

$$\begin{aligned}\Pi_n(A_n) &\le \Pi_n(A_n \cap \mathbb{X}) + \Pi_n(\mathbb{X}^c) \\ &= \frac{N_n(A_n \cap \mathbb{X})}{D_n} + \Pi_n(\mathbb{X}^c) \\ &= \frac{N_n(A_n \cap \mathbb{X})}{D_n} 1(D_n > H\epsilon_n^k) + \frac{N_n(A_n \cap \mathbb{X})}{D_n} 1(D_n \le H\epsilon_n^k) + \Pi_n(\mathbb{X}^c) \\ &\le \frac{N_n(A_n \cap \mathbb{X})}{H\epsilon_n^k} + 1(D_n \le H\epsilon_n^k) + \Pi_n(\mathbb{X}^c).\end{aligned}$$

Taking expectation, we get

$$P^n \Pi_n(A_n) \le \frac{P^n N_n(A_n \cap \mathbb{X})}{H\epsilon_n^k} + P^n(D_n \le H\epsilon_n^k) + P^n \Pi_n(\mathbb{X}^c).$$

The first and third terms in the upper bound are $o(1)$ as $n \to \infty$ by Lemmas 3.6.4 and 3.6.2, respectively. Lemma 3.6.3 says that the second term is bounded by η, so we have

$$\limsup_{n\to\infty} P^n \Pi_n(A_n) \le \eta,$$

and the claim follows since $\eta > 0$ is arbitrary.

3.6.3 Proof of Theorem 3.3.2

The proof begins by showing that $e^{-\omega n R_n(\theta)}$ satisfies a locally asymptotic normality condition, that is, for every compact set $K \subset \mathbb{R}^k$

$$\sup_{h\in K} \left| \log s_n(h) - \omega h^\top V_{\theta^\star} \Delta_{n,\theta^\star} - \tfrac{\omega}{2} h^\top V_{\theta^\star} h \right| \to 0, \quad \text{in } P^n\text{-probability.} \tag{3.6.2}$$

where $s_n(h) = e^{-\omega n\{R_n(\theta^\star + hn^{-1/2}) - R_n(\theta^\star)\}}$. To show this,

$$\begin{aligned}-\omega n\{R_n(\theta^\star + hn^{-1/2}) - R_n(\theta^\star)\} &= -\omega n\{\mathbb{P}_n(m_{\theta^\star + hn^{-1/2}} - m_{\theta^\star}) \\ &= -\omega n P(m_{\theta^\star + hn^{-1/2}} - m_{\theta^\star}) - \\ &\qquad \omega n^{1/2} \mathbb{G}_n(m_{\theta^\star + hn^{-1/2}} - m_{\theta^\star}).\end{aligned}$$

Since the loss m_θ is Lipschitz, it follows from Lemma 19.31 in van der Vaart (2000), that

$$\mathbb{G}_n\{\omega n^{1/2}(m_{\theta^\star+hn^{-1/2}} - m_{\theta^\star}) - \omega h^\top \dot{m}_{\theta^\star}\} \to 0, \quad \text{in } P^n\text{-probability.} \tag{3.6.3}$$

Since R is twice differentiable at $\theta^\star$ with second derivative matrix $V_{\theta^\star}$, (3.6.2) holds. Given that empirical risk difference has a suitable locally quadratic representation, it is intuitively clear that the Gibbs posterior distribution will take on a Gaussian shape. Confirming this intuition requires some care, but one can follow exactly the arguments used to prove Theorem 2.1 in Kleijn and van der Vaart (2012), so we omit the details here.

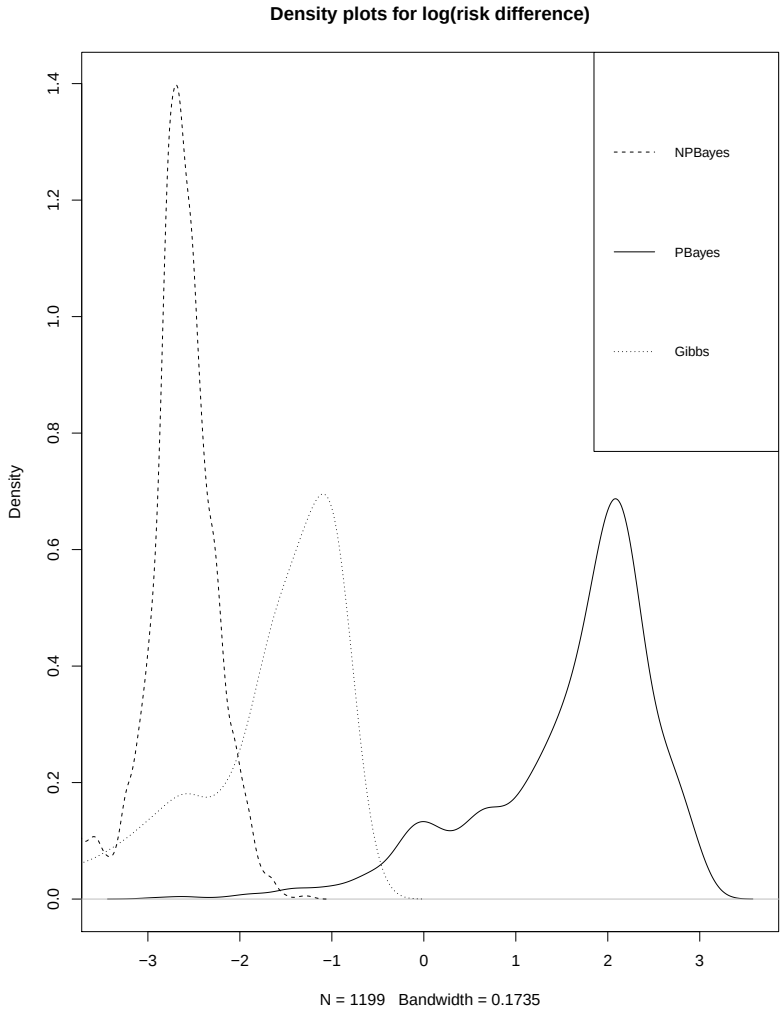


Figure 3.3 Plots of (kernel density estimates of) the Gibbs, parametric Bayes (PBayes) and non-parametric Bayes (NPBayes) posterior distribution for the log risk difference (3.4.1) based on the testing data.

CHAPTER

4

BAYESIAN NON-PARAMETRIC TESTS FOR MULTIVARIATE LOCATIONS

4.1 Introduction

Several frequentist testing procedures for multivariate locations are available in the literature, both parametric and non-parametric. The most well-known parametric procedure is the Hotelling's T^2-test, which is based on the multivariate mean vector and the covariance matrix, and it also relies on the assumption of multivariate normality. This technique performs well if the assumption of multivariate normality is nearly correct, but suffers heavily otherwise, or in the presense of outliers. Non-parametric and robust alternatives based on signs and ranks have been quite popular over the years ([Oja, Randles, et al. 2004]).

The notions of signs and ranks are based on the "ordering" of the data points, but in the multivariate setting, there is no objective basis of ordering. The notions are generalized to higher dimensions using ℓ_1-objective functions (see Section 4.2). The existing one-sample location problem have the following set up. Suppose that, we have n observations $Y = Y_1, \ldots, Y_n \in \mathbb{R}^k$ from a distribution $P(y-\theta)$, where $P(\cdot-\theta)$ is a k-variate continuous

distribution centered at $\theta = (\theta_1, \ldots, \theta_k)^T$. Our objective is to test the hypothesis

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0. \tag{4.1.1}$$

The existing non-parametric test procedures are based on the spatial sign vectors U, muli-variate spatial rank R, and multivariate spatial signed rank Q, which are defined respectively by

$$U(y) = \begin{cases} \|y\|_2^{-1} y, & y \neq 0, \\ 0, & y = 0, \end{cases} \tag{4.1.2}$$

$$R(y, Y) = \frac{1}{n} \sum_{i=1}^{n} U(y - Y_i), \tag{4.1.3}$$

$$Q(y) = \frac{1}{2}[R(y; Y) + R(y, -Y)]. \tag{4.1.4}$$

The estimator of the location associated with spatial signs in (4.1.2) is the spatial median

$$\hat{\theta}_n = \arg\min_{\theta \in \mathbb{R}^k} \mathbb{P}_n \| Y - \theta \|_2, \tag{4.1.5}$$

where $\mathbb{P}_n = n^{-1} \sum_{i=1}^{n} \delta_{Y_i}$ is the empirical measure. The objective functions (4.1.3) and (4.1.4) give rise to multivariate Hodges-Lehmann estimators (Oja and Randles 2004). The p-values of these multivariate sign and rank-based tests rely on a limiting chi-square distribution of the test statistics. Provided the underlying distribution is elliptically symmetric, i.e., its density is of the form

$$f(y - \theta) = |\Sigma|^{-1/2} g((y - \theta)^T \Sigma^{-1} (y - \theta)),$$

with symmetry center θ, and a positive definite scatter matrix Σ, its center of symmetry, location parameter, mean and spatial median are the same. In this chapter, we construct Bayesian non-parametric testing procedures for multivariate locations using spatial median. Such a procedure is attractive because it provides a credible set for spatial median, hence a testing criterion can be formulated without depending on asymptotics. In other words, here we focus on the objective function of type (4.1.2) and propose a non-parametric Bayesian testing procedure. We assume that the observations are drawn from a random distribution P, and we put a Dirichlet process (details given in Section 4.3) prior on it. From P, we can

infer about its median functional

$$\theta(P) = \underset{\theta \in \mathbb{R}^k}{\arg\min}\, P(\|Y-\theta\|_2 - \|Y\|_2), \tag{4.1.6}$$

where $Pf = \int f \mathrm{d}P$. The exact posterior distribution of $\theta(P)$ can be obtained easily by posterior simulation. Thus, we can form a credible region for $\theta(P)$ and our decision will be based on whether the value θ_0 falls into this credible set. For elliptically symmetric distributions, this testing procedure effectively studies the one-sample location problem described above, but our testing procedure can be used to study a wider range of distributions P, where we study the null hypothesis $H_0 : \theta(P) = 0$. We show that our testing procedure is asymptotically non-parametric and further compute the asymptotic power function under Pitman (contiguous) alternatives along possible directions. The two-sample test can be formulated in a similar way.

The rest of this chapter is organized as follows. In Section 4.2, we give an overview of the existing multivariate testing procedures. In Section 4.3, we describe our Bayesian non-parametric test procedures. Section 4.4 gives the local asymptotic power under contiguous alternatives and Section 4.5 presents a simulation study. All the proofs are given in Section 4.6.

4.2 Overview of existing tests

We begin this section by describing existing non-parametric testing procedures for one-sample location problems, and later move on to two-sample and several samples problems. Let $Y_1, \ldots, Y_n \in \mathbb{R}^k$ be n observations from a k-variate probability distribution P. According to Sirkiä et al. (2007), the non-parametric testing methods can be classified as based on multivariate spatial sign function U, multivariate spatial rank R, and multivariate spatial signed rank Q, which are defined as follows. The test statistic based on the score statistic $T(Y)$, which is a general notation for the score functions described above, is given by $n^{-1}\sum_{i=1}^n T(Y_i)$. Under H_0, $n^{-1/2}\sum_{i=1}^n T(Y_i) \rightsquigarrow \mathrm{N}_k(0,\Sigma)$, where $\Sigma = P\{T(Y)T(Y)^T\}$. The usual estimator for Σ is $\hat{\Sigma} = n^{-1}\sum_{i=1}^n T(Y_i)T(Y_i)^T$.

The assumption of elliptical symmetry is needed to decide the appropriate cut-off for constructing the test procedure. Under H_0,

$$Q^2 = n\left\|\hat{\Sigma}^{-1/2}\frac{1}{n}\sum_{i=1}^n T(Y_i)\right\|^2 \rightsquigarrow \chi^2_k,$$

where $\rightsquigarrow$ denotes convergence in distribution, and χ^2_k denotes a chi-square distribution

with degrees of freedom k (Sirkiä et al. (2007)). For elliptically symmetric distributions, Q^2 is strictly distribution free (Oja and Randles (2004)). An approximate p-value can be obtained from the above chi-square distribution. For small sample sizes, a conditional distribution-free p-value can be obtained under the assumption of directional symmetry (under which $(Y-\theta)/\|Y-\theta\|_2$ has the same distribution as $(\theta-Y)/\|\theta-Y\|_2$). This p-value can be obtained as $\mathrm{E}_\delta[\mathbb{1}\{Q_\delta^2 \geq Q^2\}]$, where E_δ is the expectation for the uniform distribution δ over the 2^n k-dimensional with each component being $+1$ or -1, and Q_δ^2 is the value of the test statistic for the data points $\delta_1 Y_1, \ldots, \delta_n Y_n$ (Oja and Randles (2004)).

The one sample testing procedures have been naturally extended to two samples. Suppose that, we have 2 independent random samples $Y_1^{(j)}, \ldots, Y_{n_j}^{(j)}$, from k-variate distributions $P(\cdot-\theta^{(j)})$, $j=1,2$. We test the hypothesis

$$H_0 : \theta^{(1)} = \theta^{(2)}, \quad \text{against } H_1 : \theta^{(1)} \neq \theta^{(2)}.$$

Sirkiä et al. (2007) developed a testing procedure using the general score function $T(Y)$ based on the following *inner standardization* approach. First, a $k \times k$ matrix H and a k-vector have to be found such that, for $Z_i^{(j)} = H(Y_i^{(j)} - h)$, $i=1,\ldots,n_j$, $j=1,2$,

$$\frac{1}{n}\sum_{j=1}^{2}\sum_{i=1}^{n_j} T(Z_i^{(j)}) = 0,$$

$$\frac{k}{n}\sum_{j=1}^{2}\sum_{i=1}^{n_j} T(Z_i^{(j)})T(Z_i^{(j)})^T = \left\{\frac{1}{n}\sum_{j=1}^{2}\sum_{i=1}^{n_j} \|T(Z_i^{(j)})\|_2^2\right\} I_k,$$

where $n = n_1 + n_2$, and I_k denotes the identity matrix of order $k \times k$. The test statistic has the form

$$Q^2 = k\frac{\sum_{j=1}^{2} n_j \|\frac{1}{n_j}\sum_{i=1}^{n_j} T(Z_i^{(j)})\|_2^2}{\frac{1}{n}\sum_{j=1}^{2}\sum_{i=1}^{n_j} \|T(Z_i^{(j)})\|_2^2}.$$

It has been shown that Q^2 has a limiting chi-square distribution with k degrees of freedom. Thus, for large samples, a p-value can be constructed using the quantiles of the chi-square distribution. For smaller samples, an approximate p-value can be obtained using a conditionally distribution-free *permutation test* verion (Sirkiä et al. (2007)). This approach has been extended to a general c number of samples as well.

4.3 Bayesian Non-parametric Tests

4.3.1 One-sample Problem

Suppose that, we have n observations $Y_1, \ldots, Y_n \in \mathbb{R}^k$ from a k-dimensional distribution P. We choose a non-parametric Bayesian approach, i.e., we impose a prior on the underlying random distribution P, and we form a credible set based on the posterior distribution of the spatial-median functional

$$\theta(P) = \underset{\theta \in \mathbb{R}^k}{\arg\min}\, P\{\|Y-\theta\|_2 - \|Y\|_2\}. \tag{4.3.1}$$

The most commonly used prior on P is a Dirichlet process prior with centering measure α (DP(α)) (see Chapter 4, Ghosal and van der Vaart (2017)). A Dirichlet process prior can be alternatively denoted as DP(MG), where $M = |\alpha|$, and $\bar{\alpha} = \alpha/M$ has cumulative distribution function G. The notations DP(α) and DP(MG) will be used interchangeably in this chapter.

The process DP(α) is a conjugate prior for i.i.d. observations from P, and the posterior distribution of P given $Y_1, \ldots, Y_n$ is DP($\alpha + n\mathbb{P}_n$). The exact posterior distribution of $\theta(P)$ cannot be obtained analytically, but posterior samples can be drawn via the stick-breaking construction of a Dirichlet process (chapter 4, Ghosal and van der Vaart (2017)). If $\xi_1, \xi_2, \ldots \overset{iid}{\sim} \bar{\alpha}$, and $V_1, V_2, \ldots, \overset{iid}{\sim}$ Be$(1, M)$ are independent random variables and $W_j = V_j \prod_{l=1}^{j-1}(1-V_l)$, then $P = \sum_{j=1}^{\infty} W_j \delta_{\xi_j} \sim$ DP($M\bar{\alpha}$). Thus the posterior distribution of P given $Y_1, \ldots, Y_n$ takes the same form $\sum_{j=1}^{\infty} W_j \delta_{\xi_j}$ with $V_1, V_2, \ldots, \overset{iid}{\sim}$ Be$(1, M+n)$. For computation, we use a truncated approximation $P_N = \sum_{j=1}^{N} W_j \delta_{\xi_j}$ to the stick-breaking representation. Thus, a posterior $100(1-\alpha)\%$ credible region can be formed using the following steps.

- Draw $V_j,\ j = 1, \ldots, N-1 \overset{iid}{\sim}$ Be$(1,\ M+n)$ and $V_N = 1$. Then, we calculate the stick-breaking weights as $W_1 = V_1,\ W_j = V_j \prod_{l=1}^{j-1}(1-V_l),\ j = 2, \ldots, N$.
- With probability $M/(M+n)$, draw ξ_j from G, and with probability $n/(M+n)$, ξ_j is drawn from $\mathbb{P}_n$.
- Draw posterior samples $\hat{\theta}_b,\ b = 1, \ldots, B$

$$\hat{\theta}_b = \underset{\theta}{\arg\min} \sum_{j=1}^{N} W_{jb} \|Y_{jb} - \theta\|_2.$$

- Compute the posterior mean $\bar{\theta} = B^{-1} \sum_{b=1}^{B} \hat{\theta}_b$ and the posterior covariance matrix $S = B^{-1} \sum_{b=1}^{B} (\hat{\theta}_b - \bar{\theta})(\hat{\theta}_b - \bar{\theta})'$.

- A $100(1-\alpha)\%$ credible set for $\theta(P)$ is then given by

$$C(Y_1,\ldots,Y_n;\alpha)=\{\theta:(\theta-\bar{\theta})'S^{-1}(\theta-\bar{\theta})\leq r_{1-\alpha}\},$$

  where $r_{1-\alpha}$ is the $100(1-\alpha)$th percentile of $(\hat{\theta}_b-\bar{\theta})'S^{-1}(\hat{\theta}_b-\bar{\theta}),\ b=1,\ldots,B$.

- We reject H_0 if $\theta_0\notin C(Y_1,\ldots,Y_n;\alpha)$.

The non-informative limit as $M\to 0$ of the posterior of P is $\mathrm{DP}(n\mathbb{P}_n)$ is called the Bayesian bootstrap distribution. Its centering measure is $\mathbb{P}_n$, and a random distribution generated from it is supported on the observation points. If we choose the non-informative limit of the posterior Dirichlet process, we do not need to generate posterior samples from the full Dirichlet process, and we only need to sample n independently and identically distributed (i.i.d.) observations from an exponential distribution with parameter 1, which saves some computational cost.

Theorem 4.3.1. *The one-sample Bayesian non-parametric test for $H_0:\theta(P)=\theta_0$ is asymptotically non-parametric, i.e.,*

$$P_{\theta_0}(\theta_0\in C(Y_1,\ldots,Y_n;\alpha))\to 1-\alpha,$$

for any P_{θ_0} such that $\theta(P_{\theta_0})=\theta_0$.

As we have already mentioned, the testing procedure has been constructed only using the posterior samples, without relying on any asymptotic properties.

4.3.2 Two Sample Problem

The Bayesian non-parametric testing procedure for two-sample location problem can be easily constructed generalizing the one-sample procedure. Suppose that, we have n_1 observations $Y_1^{(1)},\ldots,Y_{n_1}^{(1)}\in\mathbb{R}^k$ from a distribution $P^{(1)}$, and n_2 observations $Y_1^{(2)},\ldots,Y_{n_2}^{(2)}\in\mathbb{R}^k$ from $P^{(2)}$. We want to test the hypothesis

$$H_0:\theta(P^{(1)})-\theta(P^{(2)})=0\quad\text{against}\quad H_1:\theta(P^{(1)})-\theta(P^{(2)})\neq 0.$$

As we have previously mentioned, if $P^{(1)}=P(\cdot-\theta^{(1)})$ and $P^{(2)}=P(\cdot-\theta^{(2)})$ are elliptically symmetric distributions, then this problem boils down to studying the two-sample location problem $H_0:\theta^{(1)}-\theta^{(2)}=0$ against $H_1:\theta^{(1)}-\theta^{(2)}\neq 0$. We put a $\mathrm{DP}(MG)$ prior on both $P^{(1)}$ and $P^{(2)}$, for some $M>0$ and G. Thus $P^{(1)}$ and $P^{(2)}$ have the stick-breaking representations $P^{(1)}=\sum_{j=1}^{\infty}W_j^{(1)}\delta_{\xi_j}^{(1)}$ and $P^{(2)}=\sum_{j=1}^{\infty}W_j^{(2)}\delta_{\xi_j}^{(2)}$ respectively. where both $W_j^{(1)},\ j=1,2,\ldots$ and

$W_j^{(2)},\ j=1,2,\ldots$ are drawn from Be$(1,M)$, and $\xi_j^{(1)},\ j=1,2,\ldots$ and $\xi_j^{(2)},\ j=1,2,\ldots$ are i.i.d. samples from G. We truncate both sets of weights at N and construct a $100(1-\alpha)\%$ posterior credible set through the following steps.

- Draw $V_j^{(1)},\ j=1,\ldots,N-1 \overset{iid}{\sim} \text{Be}(1,\ M+n_1)$ and $V_N^{(1)}=1$. Then the stick-breaking weights are $W_1^{(1)}=V_1^{(1)},\ W_j^{(1)}=V_j^{(1)}\prod_{l=1}^{j-1}(1-V_l^{(1)}),\ j=2,\ldots,N$. Similarly draw $V_j^{(2)}$ from Be$(1,M+n_2)$ and construct $W_j^{(2)}$ accordingly.

- Next, with probability $M/(M+n_l)$, draw $Y_{1b}^{(l)},\ldots,Y_{Nb}^{(l)}$ from G, and with probability $n_l/(M+n_l)$, draw $Y_{1b}^{(l)},\ldots,Y_{Nb}^{(l)}$ from $\mathbb{P}_{n_l}$, $l=1,2$.

- Draw posterior samples $\hat{\theta}_b^{(1)}$ and $\hat{\theta}_b^{(2)},\ b=1,\ldots,B$,

$$\hat{\theta}_b^{(1)}=\arg\min_{\theta}\sum_{j=1}^{N}W_{jb}^{(1)}\|Y_{jb}^{(1)}-\theta\|_2,$$
$$\hat{\theta}_b^{(2)}=\arg\min_{\theta}\sum_{j=1}^{N}W_{jb}^{(2)}\|Y_{jb}^{(2)}-\theta\|_2.$$

- Compute $\bar{\theta}^{(l)}=B^{-1}\sum_{b=1}^{B}\hat{\theta}_b^{(l)}$ and $S^{(l)}=B^{-1}\sum_{b=1}^{B}(\hat{\theta}_b^{(l)}-\bar{\theta}^{(l)})(\hat{\theta}_b^{(l)}-\bar{\theta}^{(l)})'$, for $l=1,2$.

- A $100(1-\alpha)\%$ credible set for $\theta(P_1)-\theta(P_2)$ is then given by

$$\begin{aligned}C(Y_1^{(1)},\ldots,Y_{n_1}^{(1)},Y_1^{(2)},\ldots,Y_{n_2}^{(2)};\alpha)=\{\theta_1-\theta_2:(\theta_1-\theta_2-\bar{\theta}^{(1)}+\bar{\theta}^{(2)})'\\ (S^{(1)}+S^{(2)})^{-1}(\theta_1-\theta_2-\bar{\theta}^{(1)}+\bar{\theta}^{(2)})\le r_{1-\alpha}\},\end{aligned}\tag{4.3.2}$$

  where $r_{1-\alpha}$ is the $100(1-\alpha)$th percentile of $(\hat{\theta}_b^{(1)}-\hat{\theta}_b^{(2)}-\bar{\theta}^{(1)}+\bar{\theta}^{(2)})'(S^{(1)}+S^{(2)})^{-1}(\hat{\theta}_b^{(1)}-\hat{\theta}_b^{(2)}-\bar{\theta}^{(1)}+\bar{\theta}^{(2)}),\ b=1,\ldots,B$.

- We reject H_0 if $0\notin C(Y_1^{(1)},\ldots,Y_{n_1}^{(1)},Y_2^{(1)},\ldots,Y_{n_2}^{(2)};\alpha)$.

Again, we can consider the non-informative limit of the posterior Dirichlet processes, and use the Bayesian bootstrap to form the credible region.

Theorem 4.3.2. *The two-sample Bayesian non-parametric test is asymptotically non-parametric, i.e., for* $0<\alpha<1$, *and* $\theta_0\in\mathbb{R}^k$

$$P_{\theta_0}^{(1)}P_{\theta_0}^{(2)}(0\in C(Y_1^{(1)},\ldots,Y_{n_1}^{(1)},Y_1^{(2)},\ldots,Y_{n_2}^{(2)};\alpha))\to 1-\alpha,$$

for any $P_{\theta_0}^{(1)},\ P_{\theta_0}^{(2)}$ *such that* $\theta(P_{\theta_0}^{(1)})=\theta(P_{\theta_0}^{(2)})=\theta_0$.

4.4 Asymptotic power Under Contiguous Alternatives

In this section, we analyze the local asymptotic power of the proposed Bayesian non-parametric tests, i.e., the limiting power under a sequence of alternatives converging to the null value. For the one-sample problem, we consider differentiable in quadratic mean (DQM) densities $\mathscr{P} = \{p_\theta = \mathrm{d}P_\theta/\mathrm{d}\mu : \theta \in \mathbb{R}^k\}$, i.e., there exists a vector valued measurable function $\dot{\ell}_\theta : \mathbb{R}^k \to \mathbb{R}^k$ such that, for $h \to 0$,

$$\int \left(\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2} h^T \dot{\ell}_\theta \sqrt{p_\theta}\right)^2 \mathrm{d}\mu = o(\|h\|_2^2).$$

We consider shrinking alternatives of the form

$$H_{1n} : \theta = \theta_0 + \frac{h}{\sqrt{n}}, \tag{4.4.1}$$

for models $P_\theta \in \mathscr{P}$. Then we derive the limiting power for the sequence of distributions $P_{\theta_0+h/\sqrt{n}} \in \mathscr{P}$. As a consequence of the DQM condition, the models $P^n_{\theta_0+h/\sqrt{n}}$ satisfy the local asymptotically normal (LAN) condition, i.e., there exist a matrix I_θ and a random vector $\Delta_{n,\theta} \sim \mathrm{N}_k(0, I_\theta)$ such that, for every converging sequence $h_n \to h$,

$$\log \frac{\mathrm{d}P^n_{\theta+h_n/\sqrt{n}}}{\mathrm{d}P^n_\theta} = h^T \Delta_{n,\theta} - \frac{1}{2} h^T I_\theta h + o_{P^n_\theta}(1). \tag{4.4.2}$$

In this context, specifically $\Delta_{n,\theta} = n^{-1/2} \sum_{i=1}^n h^T \dot{\ell}_\theta(Y_i)$, and $I_\theta = P_\theta \dot{\ell}_\theta \dot{\ell}_\theta^T$. The next theorem gives the limiting power for the one-sample test under a sequence of alternatives of the form H_{1n} for the DQM models. Before stating the theorem, we introduce some notations. Define

$$U_{\theta,P} = P\left(\frac{(Y-\theta)(Y-\theta)^T}{\|Y-\theta\|_2^2}\right) \tag{4.4.3}$$

$$V_{\theta,P} = P\left\{\frac{1}{\|Y-\theta\|_2}\left(I_k - \frac{(Y-\theta)(Y-\theta)^T}{\|Y-\theta\|_2^2}\right)\right\}, \tag{4.4.4}$$

Let $P^\star$ be the true distribution of Y, i.e., the truth of P, and $\theta^\star$ be the spatial median for the true distribution $P^\star$, i.e., $\theta^\star = \theta(P^\star)$.

Theorem 4.4.1. *For a sequence of shrinking alternatives of the form* (4.4.1)*, i.e., under a sequence of differentiable in quadratic mean (DQM) models $P_{\theta_0+h/\sqrt{n}} \in \mathscr{P}$, the asymptotic power of the one-sample Bayesian non-parametric test for $H_0 : \theta(P) = \theta_0$ is given by*

$F_{\chi^2}(\chi^2_{k;\alpha}; k, \delta_1'(V^{-1}_{\theta_0,P_{\theta_0}} U_{\theta_0,P_{\theta_0}} V^{-1}_{\theta_0,P_{\theta_0}})^{-1}\delta_1)$, *where*

$$\delta_1 = P_{\theta_0}\bigg(- V^{-1}_{\theta_0,P_{\theta_0}} \frac{Y-\theta_0}{\|Y-\theta_0\|_2} \dot{\ell}^T_{\theta_0} I^{-1}_{\theta_0} h \bigg), \tag{4.4.5}$$

and $F_{\chi^2}(x; k, \delta)$ is the CDF of a non-central chi-square distribution with degrees of freedom k and non-centrality parameter δ, with $\chi^2_{k;\alpha}$ being the $100(1-\alpha)$th percentile of a χ^2_k distribution.

For the two sample problem, we again consider DQM models $P^{(1)}_{\theta_0+h_1/\sqrt{n_1}}$, $P^{(2)}_{\theta_0+h_2/\sqrt{n_2}} \in \mathscr{P}$, i.e., the contiguous alternatives are of the form

$$H_{1n} : \theta^{(j)}_{n_j} = \theta_0 + \frac{h_j}{\sqrt{n_j}},\ j = 1,2, \tag{4.4.6}$$

with $n = n_1 + n_2$, such that $n_1/n \to \lambda$, and $n_2/n \to 1-\lambda$. The following theorem gives the limiting power of the two-sample test under contiguous alternatives of the form (4.4.6), and the notations from Theorem 4.4.1 directly translate to the next theorem.

Theorem 4.4.2. *For a sequence of shrinking alternatives of the form* (4.4.6)*, i.e., for a sequence of DQM models $P^{(1)}_{\theta_0+h_1/\sqrt{n_1}}$, $P^{(2)}_{\theta_0+h_2/\sqrt{n_2}} \in \mathscr{P}$, the asymptotic power of the two-sample Bayesian non-parametric test for testing $H_0 : \theta(P^{(1)}) = \theta(P^{(2)}) = \theta_0$ for $\theta_0 \in \mathbb{R}^k$, is given by*

$$F_{\chi^2}(\chi^2_{k;\alpha}; \delta_2'(V^{-1}_{\theta_0;P^{(1)}_{\theta_0}} U_{\theta_0;P^{(1)}_{\theta_0}} V^{-1}_{\theta_0,P^{(1)}_{\theta_0}} + V^{-1}_{\theta_0;P^{(2)}_{\theta_0}} U_{\theta_0;P^{(2)}_{\theta_0}} V^{-1}_{\theta_0,P^{(2)}_{\theta_0}})^{-1}\delta_2),$$

where

$$\begin{aligned} \delta_2 = \frac{1}{\sqrt{\lambda}} P^{(1)}_{\theta_0}\{ - V^{-1}_{\theta_0,P^{(1)}_{\theta_0}} \frac{Y^{(1)}-\theta_0}{\|Y^{(1)}-\theta_0\|_2} \dot{\ell}^{(1)^T}_{\theta_0} I^{(1)^{-1}}_{\theta_0} h_1 \} \\ + \frac{1}{\sqrt{1-\lambda}} P^{(2)}_{\theta_0}\{ - V^{-1}_{\theta_0,P^{(2)}_{\theta_0}} \frac{Y^{(2)}-\theta_0}{\|Y^{(2)}-\theta_0\|_2} \dot{\ell}^{(2)^T}_{\theta_0} I^{(2)^{-1}}_{\theta_0} h_2 \}, \end{aligned} \tag{4.4.7}$$

for any $\theta_0 \in \mathbb{R}^k$.

4.5 Simulation Study

We perform the simulation study to demonstrate the finite sample performance of the proposed one-sample and two-sample Bayesian non-parametric tests. We compare our tests with the Hotelling T^2-test, and the spatial sign and rank tests. The underlying distributions are bivariate Gaussian, bivariate t (both elliptically symmetric), and bivariate

gamma (asymmetric). The bivariate gamma distribution is constructed using Gaussian copula ([Xue-Kun Song 2000]). To describe the construction briefly, let $Y_1, \ldots, Y_k$ be k many univariate gamma distributions $\mathrm{Ga}(s, r)$ with cumulative distribution functions (CDF) and probability density functions (PDF) being denoted by F_j and f_j, $j = 1, \ldots, k$. Then the joint density of $Y = (Y_1, \ldots, Y_k)$ is then given by

$$g(y, s, r, V) = c_\phi\{F_1, \ldots, F_k | V\} \prod_{j=1}^{k} f_j(y_j, s, r),$$

where $c_\phi(\cdot|V)$ denotes the density of the k-dimensional Gaussian copula. For comparison, we choose a general version of Hotelling's T^2-test, where the Gaussian assumption can be relaxed to existence of second moments. Here the p-value is based on a chi-square approximation instead of the usual F-distribution. For one-sample test, we consider a sample size of $n = 100$ and for the two-sample test, we choose $n_1 = 100$, $n_2 = 90$. We calculate the power, i.e., the proportion of times the null hypotheses are rejected off 2000 replications. The location parameters are chosen suitably to show a good range of powers, and the scatter matrices are chosen to be I_2. For our testing method, we choose a $\mathrm{DP}(\alpha)$ prior with $\alpha = 2 \times \mathrm{N}_2(0, 10I_2)$.

Tables **??** and 4.2 show the power values, and it should be noted that our test procedures attain the nominal level 0.05, and outperforms all other procedures in most cases. When the underlying distributions are not Gaussian, our method performs better than other methods, especially compared with the Hotelling's T^2 -test.

Remark 1. Here we have considered tests for multivariate locations based on spatial medians, but these tests can be constructed using multivariate ℓ_1-medians (with ℓ_p-norms) as well. For some fixed $p > 1$, the ℓ_1-median for a k-variate distribution P can be defined as

$$\theta_p(P) = \underset{\theta \in \mathbb{R}^k}{\arg\min}\{\|Y - \theta\|_p - \|\theta\|_p\}.$$

Bernstein-von Mises theorems of ℓ_1-medians are available in the literature ([Bhattacharya & Ghosal 2019]). Hence the expressions for local asymptotic powers under shrinking alternatives can be obtained using those theorems.

4.6 Proofs

Before giving the proofs of the main theorems, we need a couple of auxiliary results. The first one gives convergence results for the posterior mean $\bar{\theta}$ and covariance matrix S.

Table 4.1 Power for testing $H_0 : \theta(P) = 0$ for bivariate Gaussian, bivariate t (with 1 degree of freedom), and bivariate gamma distributions with different location parameters (θ).

θ	NPBayes	Sign Test	Rank Test	Hotelling's T^2
	Bivariate Gaussian Distribution			
(0, 0)	0.050	0.046	0.051	0.055
(0.05, 0.05)	0.139	0.086	0.084	0.099
(0.1, 0.05)	0.169	0.125	0.141	0.156
(0.1, -0.1)	0.221	0.188	0.213	0.234
	Bivariate t_1 Distribution			
(0, 0)	0.050	0.053	0.041	0.02
(0.05, 0.05)	0.174	0.058	0.053	0.025
(0.1, 0.05)	0.179	0.094	0.082	0.018
(0.1, -0.1)	0.201	0.171	0.197	0.026
	Bivariate Gamma Distribution			
(0, 0)	0.050	0.016	0.025	0.294
(0.05, 0.05)	0.027	0.021	0.039	0.528
(0.1, 0.05)	0.029	0.013	0.058	0.607
(0.1, -0.1)	0.034	0.009	0.018	0.255

Lemma 4.6.1. *Suppose the true distribution of $Y_1, \ldots, Y_n \in \mathbb{R}^k$ is $P^\star$, and the following conditions hold for $P^\star$.*

1. *The distribution $P^\star$ has a density that is bounded on bounded subsets of $\mathbb{R}^k$.*
2. *The spatial median of $P^\star$, $\theta^\star = \theta(P^\star)$ is unique.*

Then, $\bar{\theta} = \hat{\theta}_n + o_{P^\star}(n^{-1/2})$ and $nS = V^{-1}_{\theta^\star,P^\star} U_{\theta^\star,P^\star} V^{-1}_{\theta^\star,P^\star} + o_{P^\star}(1)$, where $\hat{\theta}_n$ is the sample spatial median of $Y_1, \ldots, Y_n$.

Proof. The posterior distribution of $\theta(P)$ can be approximated by a Gaussian distribution in the Bernstein-von Mises sense ([Bhattacharya & Ghosal 2019]), i.e., given $Y_1, \ldots, Y_n$

$$\sqrt{n}(\theta(P) - \hat{\theta}_n) \rightsquigarrow \mathrm{N}_k(0, V^{-1}_{\theta^\star,P^\star} U_{\theta^\star,P^\star} V^{-1}_{\theta^\star,P^\star}).$$

Let $\mathbb{B}_n$ is the Bayesian bootstrap process defined by $\mathbb{B}_n f = \sum_{i=1}^n W_{ni} \delta_{Y_i}$, where $(W_{n1}, \ldots, W_{nn})$ follows a Dirichlet distribution $\mathrm{Dir}(n; 1, \ldots, 1)$. Define $\theta(\mathbb{B}_n) = \arg\min_\theta \mathbb{B}_n \| Y - \theta \|_2$. It has been shown in Theorem 2.3.1 that, asymptotically, $\theta(P)$ is a Bayesian bootstrapped analog of a Z-estimator. Thus, our problem boils down to showing the first and second moment consistency of the bootstrap Z-estimator $\theta(\mathbb{B}_n)$.

Cheng (2015) proved the consistency of the bootstrap moment estimators for the class of exchangeably weighted bootstrap (see Section 2.2, Cheng (2015)). The Bayesian bootstrap

Table 4.2 Power for testing $H_0 : \theta^{(1)} = \theta^{(2)}$ for bivariate Gaussian, bivariate t (with 1 degree of freedo), and bivariate gamma distributions with different location parameters ($\theta^{(1)}$ and $\theta^{(2)}$).

$\theta^{(1)}$	$\theta^{(2)}$	NPBayes	Sign Test	Rank Test	Hotelling's T^2
		Bivariate Gaussian Distribution			
(0, 0)	(0, 0)	0.050	0.057	0.051	0.037
(0, 0)	(0.1, 0)	0.135	0.091	0.085	0.083
(0, 0)	(0.1, 0.1)	0.225	0.098	0.122	0.136
(0, 0)	(0, 0.3)	0.402	0.337	0.346	0.146
		Bivariate t_1 Distribution			
(0, 0)	(0, 0)	0.059	0.041	0.052	0.011
(0, 0)	(0,1. 0)	0.141	0.060	0.074	0.026
(0, 0)	(0.1, 0.1)	0.158	0.087	0.099	0.022
(0, 0)	(0, 0.3)	0.307	0.248	0.213	0.023
		Bivariate Gamma Distribution			
(0, 0)	(0, 0)	0.024	0.020	0.019	0.017
(0, 0)	(0.1, 0)	0.020	0.019	0.018	0.025
(0, 0)	(0.1, 0.1)	0.030	0.015	0.014	0.030
(0, 0)	(0, 0.3)	0.033	0.018	0.023	0.028

weights fall into the class of the exchangeable bootstrap weights, and we have to show that the ℓ_1 criterion function $m_\theta(y) = -\|y-\theta\|_2 + \|y\|_2$ satisfies the following two sufficient conditions. Let $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P^\star)$ denotes the empirical process and $\mathbb{G}_n^\star = \sqrt{n}(\mathbb{B}_n - \mathbb{P}_n)$ denotes the bootstrap empirical process. Suppose the following conditions hold.

1. Let Θ be the compact parameter space. For any $\theta \in \Theta$,

$$P^\star(m_\theta - m_{\theta^\star}) \lesssim -\|\theta - \theta^\star\|_2^2.$$

2. Define $N_\delta = \{m_\theta - m_{\theta_0} : \|\theta - \theta\|_2 \le \delta\}$. We have to show

$$\left(\mathrm{E}_X \|\mathbb{G}_n\|_{N_\delta}^{p'}\right)^{1/p'} \lesssim \delta \tag{4.6.1}$$

$$\left(\mathrm{E}_{XW} \|\mathbb{G}_n^\star\|_{N_\delta}^{p'}\right)^{1/p'} \lesssim \delta, \tag{4.6.2}$$

for some $p' > 2$.

Then the assertion in Lemma 4.6.1 holds. First, we need to show that the parameter space can be restricted to a compact subset of $\mathbb{R}^k$ with high probability. In Theorem 2.3.1, it has been shown that for some $0 < \epsilon < 1/4$ and $K > 0$ such that $P(\|Y\|_2 \le K) > 1 - \epsilon$, given

$Y_1, \ldots, Y_n$, $\|\theta(\mathbb{B}_n)\|_2 \leq 3K$ with high $P^{\star n}$-probability, which implies that asymptotically, given $Y_1, \ldots, Y_n$, $\|\theta(P)\|_2 \leq 3K$ with high $P^{\star n}$-probability.

After fixing $K > 0$, we choose $\Theta = \{\theta : \|\theta\|_2 \leq 3K\}$. Since Θ is compact, Condition 1 can be shown from a Taylor series expansion around $\theta^\star$

$$P^\star m_\theta - P^\star m_{\theta^\star} = (\theta - \theta^\star)' P^\star \dot{m}_{\theta^\star} + \frac{(\theta - \theta^\star)' V_{\theta^\star, P^\star}(\theta - \theta^\star)}{2} + o(\|\theta - \theta^\star\|^2). \tag{4.6.3}$$

Since $\theta^\star$ is the maximizer of $P^\star m_\theta$, $P^\star \dot{m}_{\theta^\star}$ vanishes. The matrix $V_{\theta^\star, P^\star}$ is negative definite, and hence, the second term in the right hand side of (4.6.3) is bounded above by $-c\|\theta - \theta^\star\|_2^2$, for a positive constant c.

Before proving Condition 2, we introduce some notations. For any class of functions $\mathscr{A}$, and metric ℓ, its ϵ.-bracketing number is denoted as $N_{[]}(\epsilon, \mathscr{A}, \ell)$. The corresponding bracketing entropy integral is defined as

$$J_{[]}(\epsilon, \mathscr{A}, \ell) = \int_0^\delta \sqrt{1 + \log N_{[]}(\epsilon, \mathscr{A}, \ell)} \mathrm{d}\epsilon.$$

Following [Cheng 2015], a simple sufficient condition for (4.6.1) is the following global Lipschitz condition

$$|m_\theta(x) - m_{\theta^\star}(x)| \leq \|\theta - \theta^\star\|_2 \tag{4.6.4}$$

for any $\theta \in \Theta$, and

$$J_{[]}(1, N_\delta, L_2(P^\star)) + \|M\|_{\psi_{p'}} < \infty, \tag{4.6.5}$$

for some $p' > 2$, where $\|\cdot\|_{\psi_p}$ is the Orlicz norm with respect to the convex function $\psi_p(t) = e^{(t^p - 1)}$. In our case, (4.6.4) holds from triangle inequality $|m_\theta(y) - m_{\theta^\star}(y)| \leq \|\theta - \theta^\star\|_2$. Since $M(y) = 1$ for every y, we just have to show $J_{[]}(1, N_\delta, L_2(P^\star)) < \infty$.

By Example 19.7 in van der Vaart 2000, since $|m_\theta(y) - m_{\theta'}(y)| \leq \|\theta - \theta'\|_2$, for every θ, $\theta' \in \Theta$, there exists a constant K such that

$$N_{[]}(1, N_\delta, L_2(P^\star)) \leq \left(\frac{\operatorname{diam} \Theta}{\epsilon}\right)^k, \text{ for every } 0 < \epsilon < \operatorname{diam} \Theta.$$

Then, the entropy is of the order $\log(1/\epsilon)$. By a change of variable, it can be shown that $J_{[]}(1, N_\delta, L_2(P^\star)) < \infty$. □

The next lemma gives a Bernstein-von Mises theorem for the difference of spatial

medians for two independent samples $Y_1^{(1)},\ldots,Y_{n_1}^{(1)} \sim P^{(1)}$, and $Y_1^{(2)},\ldots,Y_{n_2}^{(2)} \sim P^{(2)}$. The sample spatial medians are denoted by $\hat{\theta}_{n_1}$ and $\hat{\theta}_{n_2}$ respectively. We put a DP(α) prior on both $P^{(1)}$ and $P^{(2)}$, and construct a posterior for $\theta(P^{(1)})-\theta(P^{(2)})$. The asymptotic result follows almost immediately from Theorem 3.1 in [Bhattacharya & Ghosal 2019].

Lemma 4.6.2. *Suppose the following conditions hold.*

1. *The true distributions $P^{\star(1)}$ and $P^{\star(2)}$ have probability densities that are bounded on compact subsets of $\mathbb{R}^k$.*

2. *The true spatial medians, $\theta^{\star(1)}=\theta(P^{\star(1)})$ and $\theta^{\star(2)}=\theta(P^{\star(2)})$ are unique.*

Then, denoting $n=n_1+n_2$, such that $n_1/n\to\lambda$, and $n_2/n\to 1-\lambda$,

(i) $\sqrt{n}(\hat{\theta}^{(1)}_{n_1}-\theta^{\star(1)}-\hat{\theta}^{(2)}_{n_2}+\theta^{\star(2)}) \rightsquigarrow \mathrm{N}_k(0,\lambda^{-1}V^{-1}_{\theta^{\star(1)},P^{\star(1)}}U_{\theta^{\star(1)},P^{\star(1)}}V^{-1}_{\theta^{\star(1)},P^{\star(1)}}+$
$(1-\lambda)^{-1}V^{-1}_{\theta^{\star(2)},P^{\star(2)}}U_{\theta^{\star(2)},P^{\star(1)}}V^{-1}_{\theta^{\star(2)},P^{\star(1)}})$

(ii) *Given* $Y_1^{(1)},\ldots,Y_{n_1}^{(1)}$, $Y_2^{(1)},\ldots,Y_{n_2}^{(2)}$,

$$\sqrt{n}(\theta(P^{(1)})-\hat{\theta}^{(1)}_{n_1}-\theta(P^{(2)})+\hat{\theta}^{(2)}_{n_2}) \rightsquigarrow \mathrm{N}_k(0,\lambda^{-1}V^{-1}_{\theta^{\star(1)},P^{\star(1)}}U_{\theta^{\star(1)},P^{\star(1)}}V^{-1}_{\theta^{\star(1)},P^{\star(1)}}+ (1-\lambda)^{-1}V^{-1}_{\theta^{\star(2)},P^{\star(2)}}U_{\theta^{\star(2)},P^{\star(2)}}V^{-1}_{\theta^{\star(2)},P^{\star(2)}}).$$

Proof. From Theorem 2.3.1, for $j=1,2$,

(i) $\sqrt{n_j}(\hat{\theta}^{(j)}_{n_j}-\theta^{\star(j)}) \rightsquigarrow \mathrm{N}_k(0,V^{-1}_{\theta^{\star(j)},P^{\star(j)}}U_{\theta^{\star(j)},P^{\star(j)}}V^{-1}_{\theta^{\star(j)},P^{\star(j)}})$,

(ii) Given $Y_1^{(j)},\ldots,Y_{n_j}^{(j)}$,

$$\sqrt{n_j}(\theta(P^{(j)})-\hat{\theta}^{(j)}_{n_j}) \rightsquigarrow \mathrm{N}_k(0,V^{-1}_{\theta^{\star(j)},P^{\star(j)}}U_{\theta^{\star(j)},P^{\star(j)}}V^{-1}_{\theta^{\star(j)},P^{\star(j)}}).$$

From independence of the two samples, the conclusion follows. □

Proof of Theorem 4.3.1. The probability of accepting the null hypothesis under the null is

$$P_{\theta_0}(\theta_0\in C(Y_1,\ldots,Y_n))=P_{\theta_0}((\bar{\theta}-\theta_0)'S^{-1}(\bar{\theta}-\theta_0)\le r_{1-\alpha})$$

Using Lemma 4.6.1 and Theorem 2.3.1,

$$n(\theta-\bar{\theta})^T S^{-1}(\theta-\bar{\theta}) \rightsquigarrow \chi^2_k, \tag{4.6.6}$$

which implies that $r_{1-\alpha} = \chi^2_{k;\alpha} + o_{P_{\theta_0}}(1)$. The weak convergence in (4.6.6) uses the fact that if $X \sim \mathrm{N}_k(0, I_k)$, then $X^T X \sim \chi^2_k$. Next, again using Lemma 4.6.1, Theorem 2.3.1 and Slutsky's theorem,

$$\begin{aligned} P_{\theta_0}(\theta_0 \in C(Y_1, \ldots, Y_n)) &= P_{\theta_0}((\bar{\theta} - \theta_0)' S^{-1} (\bar{\theta} - \theta_0) \leq r_{1-\alpha}) \\ &= P_{\theta_0}((\hat{\theta}_n - \theta_0)' (V^{-1}_{\theta_0, P_{\theta_0}} U_{\theta_0, P_{\theta_0}} V^{-1}_{\theta_0, P_{\theta_0}})^{-1} (\hat{\theta}_n - \theta_0) + \\ &\quad o_{P_{\theta_0}}(1) \leq \chi^2_{k;\alpha} + o_{P_{\theta_0}}(1)) \to 1 - \alpha. \end{aligned}$$

. □

Proof of Theorem 4.3.2. The proof is similar to that of Theorem 4.3.1. Using Lemma 4.6.2 under H_0, $\sqrt{n}(\hat{\theta}^{(1)}_{n_1} - \hat{\theta}^{(2)}_{n_2})$ converges to a Gaussian distribution with mean 0 and covariance matrix $\lambda^{-1} V^{-1}_{\theta_0, P^{(1)}_{\theta_0}} U_{\theta_0, P^{(1)}_{\theta_0}} V^{-1}_{\theta_0, P^{(1)}_{\theta_0}} + (1-\lambda)^{-1} V^{-1}_{\theta_0, P^{(2)}_{\theta_0}} U_{\theta_0, P^{(2)}_{\theta_0}} V_{\theta_0, P^{(2)}_{\theta_0}}$. Using Lemma 4.6.1, Lemma 4.6.2 and Slutsky's theorem,

$$n(\bar{\theta}^{(1)} - \bar{\theta}^{(2)})' (S^{(1)} + S^{(2)})^{-1} (\bar{\theta}^{(1)} - \bar{\theta}^{(2)}) \rightsquigarrow \chi^2_k,$$

which implies that $r_{1-\alpha} = \chi^2_{k;\alpha} + o_{P^{(1)}_{\theta_0}}(1) + o_{P^{(2)}_{\theta_0}}(1)$. Next, using Lemma 4.6.2, under H_0,

$$\begin{aligned} &(P^{(1)}_{\theta_0} \times P^{(2)}_{\theta_0})[(\bar{\theta}^{(1)} - \bar{\theta}^{(2)})' (S^{(1)} + S^{(2)})^{-1} (\bar{\theta}^{(1)} - \bar{\theta}^{(2)}) \leq r_{1-\alpha}] \\ &\quad = (P^{(1)}_{\theta_0} \times P^{(2)}_{\theta_0})(n(\hat{\theta}^{(1)}_{n_1} - \hat{\theta}^{(2)}_{n_2})' (\frac{1}{\lambda} V^{-1}_{\theta_0, P^{(1)}_{\theta_0}} U_{\theta_0, P^{(1)}_{\theta_0}} V^{-1}_{\theta_0, P^{(1)}_{\theta_0}} \\ &\quad + \frac{1}{1-\lambda} V^{-1}_{\theta_0, P^{(2)}_{\theta_0}} U_{\theta_0, P^{(2)}_{\theta_0}} V^{-1}_{\theta_0, P^{(2)}_{\theta_0}})^{-1} (\hat{\theta}^{(1)}_{n_1} - \hat{\theta}^{(2)}_{n_2}) + o_{P^{(1)}_{\theta_0}}(1) + o_{P^{(2)}_{\theta_0}}(1) \\ &\quad \leq \chi^2_{k;\alpha} + o_{P^{(1)}_{\theta_0}}(1) + o_{P^{(2)}_{\theta_0}}(1)) \to 1 - \alpha. \end{aligned}$$

□

Proof of Theorem 4.4.1. It is well known that the models $P^n_{\theta_0}$ and $P^n_{\theta_0 + h/\sqrt{n}}$ are mutually contiguous. Under H_0, using Theorem 5.23 in [A. W. v. d. Vaart 2000], $\hat{\theta}_n$ can be written as

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} V^{-1}_{\theta_0, P_{\theta_0}} \frac{Y_i - \theta_0}{\|Y_i - \theta_0\|_2} + o_{P_{\theta_0}}(1).$$

Let L_n denote the log-likelihood ratio $L_n = \log(\mathrm{d}P^n_{\theta_0 + h/\sqrt{n}} / \mathrm{d}P^n_{\theta_0})$. By central limit theorem, $(\sqrt{n}(\hat{\theta}_n - \theta_0), L_n)$ tends to a $(k+1)$-dimensional Gaussian distribution with mean zero and covariance

$$\delta_1 = P_{\theta_0}\big(- V^{-1}_{\theta_0, P_{\theta_0}} \frac{Y - \theta_0}{\|Y - \theta_0\|_2} \dot{\ell}^T_{\theta_0} I^{-1}_{\theta_0} h \big).$$

Then by Le Cam's third lemma (Example 6.7, van der Vaart (2000)), under $P_{\theta_0+h/\sqrt{n}}$, $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges weakly to a Gaussian distribution with mean δ_1. Following the arguments used in Theorem 4.3.1, the local asymptotic power of the test is given by

$$\begin{aligned}
&P_{\theta_0+h/\sqrt{n}}(\bar{\theta}-\theta_0)'S^{-1}(\bar{\theta}-\theta_0) \le r_{1-\alpha}) = \\
&\qquad P_{\theta_0+h/\sqrt{n}}(n(\hat{\theta}_n-\theta_0)'(V^{-1}_{\theta_0,P_{\theta_0}}U_{\theta_0,P_{\theta_0}}V^{-1}_{\theta_0,P_{\theta_0}})^{-1}(\hat{\theta}_n-\theta_0)+o_{P_{\theta_0}}(1) \\
&\qquad \le \chi^2_{k;\alpha}+o_{P_{\theta_0}}(1)).
\end{aligned}$$

Under $P_{\theta_0+h/\sqrt{n}}$, $n(\hat{\theta}_n-\theta_0)'(V^{-1}_{\theta_0,P_{\theta_0}}U_{\theta_0,P_{\theta_0}}V^{-1}_{\theta_0,P_{\theta_0}})^{-1}(\hat{\theta}_n-\theta_0)$ tends to a chi-square distribution with non-centrality parameter $\delta_1'(V^{-1}_{\theta_0,P_{\theta_0}}U_{\theta_0,P_{\theta_0}}V^{-1}_{\theta_0,P_{\theta_0}})^{-1}\delta_1$, which gives us the asymptotic power given in the statement of Theorem 4.4.1. □

Proof of Theorem 4.4.2. The proof proceeds along the lines of Theorem 4.4.1. The models $\{P^{(1)}_{\theta_0}\times P^{(2)}_{\theta_0}\}$ and $\{P^{(1)}_{\theta_0+h_1/\sqrt{n_1}}\times P^{(2)}_{\theta_0+h_2/\sqrt{n_2}}\}$ are mutually contiguous. From Theorem 5.23 in van der Vaart (2000), sample spatial medians have the following linearization,

$$\sqrt{n_1}(\hat{\theta}^{(1)}_{n_1}-\theta_0) = -\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1} V^{-1}_{\theta_0,P^{(1)}_{\theta_0}}\frac{Y^{(1)}_i-\theta_0}{\|Y^{(1)}_i-\theta_0\|_2}+o_{P^{(1)}_{\theta_0}}(1), \tag{4.6.7}$$

$$\sqrt{n_2}(\hat{\theta}^{(2)}_{n_2}-\theta_0) = -\frac{1}{\sqrt{n_2}}\sum_{i=1}^{n_2} V^{-1}_{\theta_0,P^{(2)}_{\theta_0}}\frac{Y^{(2)}_i-\theta_0}{\|Y^{(2)}_i-\theta_0\|_2}+o_{P^{(2)}_{\theta_0}}(1). \tag{4.6.8}$$

Subtracting (4.6.8) from (4.6.7), under H_0,

$$\begin{aligned}
\sqrt{n}(\hat{\theta}^{(1)}_{n_1}-\hat{\theta}^{(2)}_{n_2}) = -\Bigg\{&\frac{1}{\sqrt{n_1\lambda}}\sum_{i=1}^{n_1} V^{-1}_{\theta_0,P^{(1)}_{\theta_0}}\frac{Y^{(1)}_i-\theta_0}{\|Y^{(1)}_i-\theta_0\|_2} \\
&-\frac{1}{\sqrt{n_2(1-\lambda)}}\sum_{i=1}^{n_2} V^{-1}_{\theta_0,P^{(2)}_{\theta_0}}\frac{Y^{(2)}_i-\theta_0}{\|Y^{(2)}_i-\theta_0\|_2}\Bigg\}+o_{P^{(1)}_{\theta_0}}(1)+o_{P^{(2)}_{\theta_0}}(1).
\end{aligned}$$

Define the log-likelihood ratio $L'_N = \log(\mathrm{d}P^{(1)}_{\theta_0+h_1/\sqrt{n_1}}\mathrm{d}P^{(2)}_{\theta_0+h_2/\sqrt{n_2}}/\mathrm{d}P^{(1)}_{\theta_0}\mathrm{d}P^{(2)}_{\theta_0})$, which looks like

$$\begin{aligned}
L'_N = \frac{1}{\sqrt{n_1}}h_1^T\sum_{i=1}^{n_1}\dot{\ell}^{(1)}_{\theta_0}(Y^{(1)}_i)-\frac{1}{2}h_1^T I^{(1)}_{\theta_0}h_1+\frac{1}{\sqrt{n_2}}h_2^T\sum_{i=1}^{n_2}\dot{\ell}^{(2)}_{\theta_0}(Y^{(2)}_i)-\frac{1}{2}h_2^T I^{(2)}_{\theta_0}h_2 \\
+o_{P^{(1)}_{\theta_0}}(1)+o_{P^{(2)}_{\theta_0}}(1).
\end{aligned}$$

By central limit theorem, the joint distribution of $\sqrt{n}(\hat{\theta}^{(1)}_{n_1}-\hat{\theta}^{(2)}_{n_2})$ and $\log L'_n$ tends to a $(k+1)$-

dimensional Gaussian distribution with mean zero and covariance

$$\delta_2 = \frac{1}{\sqrt{\lambda}} P^{(1)}_{\theta_0}\{-V^{-1}_{\theta_0,P^{(1)}_{\theta_0}} \frac{Y^{(1)}-\theta_0}{\|Y^{(1)}-\theta_0\|_2} \dot{\ell}^{(1)T}_{\theta_0} I^{(1)^{-1}}_{\theta_0} h_1\} + \frac{1}{\sqrt{1-\lambda}} P^{(2)}_{\theta_0}\{-V^{-1}_{\theta_0,P^{(2)}_{\theta_0}} \frac{Y^{(2)}-\theta_0}{\|Y^{(2)}-\theta_0\|_2} \dot{\ell}^{(2)T}_{\theta_0} I^{(2)^{-1}}_{\theta_0} h_2\}.$$

Again, by Le Cam's third lemma, under $P^{(1)}_{\theta_0+h_1/\sqrt{n_1}} \times P^{(2)}_{\theta_0+h_2/\sqrt{n_2}}$, $\sqrt{n}(\hat{\theta}^{(1)}_{n_1} - \hat{\theta}^{(2)}_{n_2})$ converges weakly to a Gaussian distribution with mean δ_2. Thus following the arguments used in Theorem 4.3.2, the asymptotic power is given by

$$\begin{aligned} P^{(1)}_{\theta_0+h_1/\sqrt{n_1}} \times P^{(2)}_{\theta_0+h_2/\sqrt{n_2}}\{(\bar{\theta}^{(1)}_{n_1} - \bar{\theta}^{(2)}_{n_2})'(S^{(1)}+S^{(2)})^{-1}(\bar{\theta}^{(1)}_{n_1} - \bar{\theta}^{(2)}_{n_2}) \le r_{1-\alpha}\} = \\ P^{(1)}_{\theta_0+h_1/\sqrt{n_1}} \times P^{(2)}_{\theta_0+h_2/\sqrt{n_2}}\{n(\hat{\theta}^{(1)}_{n_1} - \hat{\theta}^{(2)}_{n_2})'(\lambda^{-1} V^{-1}_{\theta_0,P^{(1)}_{\theta_0}} U_{\theta_0,P^{(1)}_{\theta_0}} V^{-1}_{\theta_0,P^{(1)}_{\theta_0}} + \\ (1-\lambda)^{-1} V^{-1}_{\theta_0,P^{(2)}_{\theta_0}} U_{\theta_0,P^{(2)}_{\theta_0}} V^{(1)}_{\theta_0,P^{(2)}_{\theta_0}})^{-1}(\hat{\theta}^{(1)}_{n_1} - \hat{\theta}^{(2)}_{n_2}) + \\ o_{P^{(1)}_{\theta_0}}(1) + o_{P^{(2)}_{\theta_0}}(1) \le \chi^2_{k;\alpha} + o_{P^{(1)}_{\theta_0}}(1) + o_{P^{(2)}_{\theta_0}}(1)\}. \end{aligned}$$

Therefore under $P^{(1)}_{\theta_0+h_1/\sqrt{n_1}} \times P^{(2)}_{\theta_0+h_2/\sqrt{n_2}}$,

$$n(\hat{\theta}^{(1)}_{n_1} - \hat{\theta}^{(2)}_{n_2})'(\frac{1}{\lambda} V^{-1}_{\theta_0,P^{(1)}_{\theta_0}} U_{\theta_0,P^{(1)}_{\theta_0}} V^{-1}_{\theta_0,P^{(1)}_{\theta_0}} + \frac{1}{1-\lambda} V^{-1}_{\theta_0,P^{(2)}_{\theta_0}} U_{\theta_0,P^{(2)}_{\theta_0}} V^{-1}_{\theta_0,P^{(2)}_{\theta_0}})^{-1} (\hat{\theta}^{(1)}_{n_1} - \hat{\theta}^{(2)}_{n_2})$$

tends to a non-central chi-square distribution with non-centrality parameter $\delta_2'(\lambda^{-1} V^{-1}_{\theta_0,P^{(1)}_{\theta_0}} U_{\theta_0,P^{(1)}_{\theta_0}} V^{-1}_{\theta_0,P^{(1)}_{\theta_0}} + (1-\lambda)^{-1} V^{-1}_{\theta_0,P^{(2)}_{\theta_0}} U_{\theta_0,P^{(2)}_{\theta_0}} V^{-1}_{\theta_0,P^{(2)}_{\theta_0}})^{-1}\delta_2$, which gives the asymptotic power. □

CHAPTER

5

BAYESIAN MULTIVARIATE QUANTILE REGRESSION USING DEPENDENT DIRICHLET PROCESS PRIOR

5.1 Introduction

Quantile regression is a popular alternative to the usual mean regression which models the relationship between the predictor and a specific quantile of the response. Univariate linear quantile regression was first proposed by Koenker and Bassett Jr. (1978) and was extensively studied in the literature since then. Given a covariate $X \in \mathbb{R}^m$, the αth linear quantile regression model for the response $Y \in \mathbb{R}$ can be written as $Q_{Y|x}(\alpha) = x^T\beta$, for $\alpha \in (0,1)$. Based on a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, Koenker and Bassett Jr. (1978) estimated the regression coefficient β by the estimator

$$\hat{\beta} = \underset{b \in \mathbb{R}^m}{\arg\min} \sum_{i=1}^{n} \rho_\alpha(Y_i - X_i^T b), \tag{5.1.1}$$

with $\rho_\alpha(u) = u(\alpha - \mathbb{1}(u < 0))$, where $\mathbb{1}$ is the indicator function. Fast algorithms to compute $\hat{\beta}$ were obtained in the literature; there is an R package called *quantreg*.

Yu and Moyeed (2001) proposed a parametric Bayesian approach to quantile regression by assuming an asymmetric Laplace likelihood. As Chang (2015) mentioned, modeling the error distribution directly by an asymmetric Laplace distribution is too restrictive. To avoid this restrictive parametric assumption, a number of non-parametric Bayesian approaches have been developed in the literature. A lot of these non-parametric methods are based on Dirichlet process mixture (DPM) models; see Chang (2015) for a comprehensive review and references.

Univariate quantiles can be extended to the multivariate setting in a number of ways. There is no unique definition of quantiles in higher dimension because of the lack of a natural ordering of Euclidean space in higher dimension. Chaudhuri (1996) introduced the notion of geometric quantiles which arises as a natural generalization of the spatial median (see Small (1990) for a review on spatial median). Chakraborty (2003) extended this idea to a regression framework. There are various other ways to define a multivariate quantile (Serfling (2002)). Hallin et al. (2010) introduced the notion of directional quantiles for multivariate location and multiple-output regression problems. The Bayesian literature on multivariate quantiles is very limited; only a few papers exist to the authors' knowledge. Waldmann and Kneib (2015) considered bivariate quantile regression using a multivariate asymmetric Laplace likelihood; while Drovandi and Pettitt (2011) used a copula approach. Recently, Guggisberg (2019) proposed a Bayesian approach to the directional quantile framework developed in Hallin et al. (2010).

Our approach here is non-parametric, i.e., we model the collection of conditional distributions of the response $Y \in \mathbb{R}^k$ given a predictor $X \in \mathbb{R}$, without using any parametric family of distributions and then estimate the desired geometric quantile of the conditional distribution. The most commonly used prior for a probability distribution is the Dirichlet process prior. The collection of conditional distributions are viewed as related quantities and hence are modeled by a Dependent Dirichlet Process (DDP) (defined in Section 5.2). One major drawback of using a Dirichlet process prior is that almost all realizations from a Dirichlet process are discrete. This issue can be handled by convolving it with a kernel. To model the error distribution flexibly without any particular parametric form, we use a countable mixture of k-dimensional normal distributions as our kernel. We use a Block Gibbs sampler (see Ishwaran and James (2001)) for our posterior computations, which is considerably fast. The illustrations discussed here are focused on cases with bivariate response. It should be noted that our proposed method is a new contribution even in the context of univariate quantile regression.

We use geometric quantiles for our treatment of multivariate quantile regression here and hence we define it formally below. Consider the situation when the variable Y is observed along with a univariate predictor X lying in a compact interval $\mathfrak{X}$. For a given value x of $X \in \mathfrak{X}$, let $P_{Y|x}$ stand for the conditional distribution of Y given $X = x$, and $F_{Y|x}$ denote the CDF. Then the non-parametric multivariate quantile regression function of Y is given by

$$Q_{Y|x}(u) = \arg\min_{q \in \mathbb{R}^k} P_{Y|x}\{\Phi_2(u, Y-q) - \Phi_2(u, Y)\}. \tag{5.1.2}$$

The true conditional distribution of Y given x is denoted by $P^\star_{Y|x}$, with the CDF being $F^\star_{Y|x}$. The uth geometric quantile of the distribution $P^\star_{Y|x}$ is denoted by $Q^\star_{Y|x}(u)$. To estimate $Q_{Y|x}(u)$, it is sensible to assume that it changes gradually in x. Hence for sensible inference, we should pull information across neighboring values of x by a smoothing technique. In a Bayesian setting that we follow, we achieve the objective by putting a suitable prior on the family of distributions $\{P_{Y|x} : x \in \mathfrak{X}\}$. The borrowing of information across neighboring values of x may be introduced in the prior by using a Dependent Dirichlet Process (DDP) (see Section 5.2). However, as marginal distributions of a DDP are Dirichlet processes and hence $P_{Y|x}$ are discrete almost surely, a smoother version of the DDP by convolving with a kernel will be used as a prior. Then the resulting posterior distribution can be computed and the induced posterior distribution on the multivariate quantile regression can be used to obtain Bayes estimates and credible sets.

The rest of this chapter is organized as follows. In Section 5.2, we give a brief background on Dependent Dirichlet Processes. In Section 5.3, we describe our non-parametric Bayesian modeling approach for multivariate quantile regression. Section 5.4 gives the details of the posterior computations, and in Section 5.5, we give the posterior consistency theorems. In Sections 5.6 and 5.7, we demonstrate the performance of our method on simulated and real data. We close the paper with the proofs in Section 5.8.

5.2 Overview on Dependent Dirichlet Process

We use the notation $P \sim \mathrm{DP}(\alpha)$ to state that the random measure P has a Dirichlet process distribution with base measure α. We also use $P \sim \mathrm{DP}(MG)$ where $M = |\alpha|$ and $\bar{\alpha} = \alpha/M$ has a distribution function G.

Several papers have considered extending Dirichlet process models over related random distributions, for example, Cifarelli and Regazzini (1978), Tomlinson and Escobar (1999) and Kottas and Gelfand (2001) (see De Iorio et al. (2004) for a comprehensive review), but these models were not naturally extended to include regression on covariates. Dependent

Dirichlet Processes were introduced by MacEachern (1999), to address regression on a predictor variable. Following the notation in Ghosal and van der Vaart (2017), suppose that, we have a collection of distributions P_z on a sample space Ω, indexed by a parameter z belonging to some covariate space $\mathfrak{Z}$. A useful prior distribution on P_z should treat them as related quantities.

It is reasonable to equip each marginal measure P_z with a Dirichlet process prior. By the stick-breaking construction of Dirichlet process, we can write P_z as

$$P_z = \sum_{i=1}^{\infty} W_j(z)\delta_{\theta_j(z)}, \tag{5.2.1}$$

where $\{W_j(z) : z \in \mathfrak{Z}\}$ are called the "stick-breaking weights" and $\{\theta_j(z) : z \in \mathfrak{Z}\}$ are called "locations". The stick breaking weights are constructed as

$$W_j(z) = V_j(z)\prod_{l=1}^{j-1}(1 - V_l(z)), \tag{5.2.2}$$

where $V_j(z) \overset{iid}{\sim} \text{Be}(1, M_z)$ for $M_z > 0$ with $\text{Be}(a, b)$ being a beta distribution with parameters a and b. The locations $\theta_j(z)$ are i.i.d. draws from the base measure G_z. This representation ensures that $P_z \sim \text{DP}(M_z G_z)$ for every $z \in \mathfrak{Z}$. A process satisfying this requirement is called a Dependent Dirichlet Process (DDP). For more details, see Section 14.9, Ghosal and van der Vaart (2017).

Several variations of DDP models have been proposed over the years. De Iorio et al. (2004) described dependence across the random distributions in an ANOVA-type fashion; which is particularly useful for multivariate categorical covariates. Nieto-Bajaras et al. (2012) proposed a version of DDP that is suitable as a prior for a time series of random probability measures. Gelfand et al. (2005) proposed a DDP for point-referenced spatial data, which was later extended by Duan et al. (2007). Sun et al. (2017) proposed location dependent Dirichlet process (LDDP) that incorporates non-parametric Gaussian processes in the DP modeling framework to model dependency information among data arising from space and time.

5.3 Bayesian Multivariate Quantile Regression with DDP

Suppose that, we have a set of independent observations $(Y^n, X^n) = (Y_1, X_1), \ldots, (Y_n, X_n)$ on a univariate predictor X and a k-variate response Y. Let $\mathscr{F} = \{f : \mathbb{R}^k \times \mathfrak{X} \to [0, \infty],$ $\int_{\mathbb{R}^k} f(y|x)\mathrm{d}y = 1\}$ be the space of conditional densities of Y given X, where $\mathfrak{X}$ is compact. Our goal is to infer about the collection $\{Q_{Y|x}(u) : x \in \mathfrak{X}\}$, for some fixed $u \in B_2^{(k)}$. We adopt a non-parametric modeling framework in the following way,

$$Y_i|x_i \sim f(\cdot|x_i),\ \{f(\cdot|x), x \in \mathfrak{X}\} \sim \Pi, \tag{5.3.1}$$

where Π denotes a prior for the class of conditional densities $\{f(\cdot|x) : x \in \mathfrak{X}\}$. We choose a DDP prior here, but as we have mentioned, distributions drawn from a Dirichlet process prior are discrete. Hence, we have to use a kernel for convolving it with. This leads us to the model

$$Y_i = \xi(X_i) + \varepsilon_i,\ i = 1, \ldots, n, \tag{5.3.2}$$

with $\xi \sim \{G_x : x \in \mathfrak{X}\}$ independently, and ε_i are the random errors. Since we are interested in the uth geometric quantile of $\{P_{Y|x} : x \in \mathfrak{X}\}$, we may choose the kernel in such a way that the uth geometric quantile of the error distribution is zero, that is, $Q_\varepsilon(u) = 0$.

The collection of distributions $\{G_x : x \in \mathfrak{X}\}$ follows a DDP prior. We use a stick-breaking representation for G_x to introduce dependence across x as

$$G_x = \sum_{l=1}^{\infty} W_l \delta_{\xi_l(x)}, \tag{5.3.3}$$

for any $x \in \mathfrak{X}$. We use a common set of stick-breaking weights W_l which are constructed from $V_l \overset{iid}{\sim} \mathrm{Be}(1, M_1)$, for some $M_1 > 0$, where the locations are drawn from a k-dimensional Gaussian distribution with mean vector and covariance matrix varying with x. We represent the locations $\xi_l(x)$ as

$$\xi_l(x) = \alpha_l + \beta_l(x), \tag{5.3.4}$$

where $\alpha_l \overset{iid}{\sim} \mathrm{N}_k(c_0, \Sigma_0)$ and β_l follows a Gaussian Process (GP) with mean function $c_1 x$ and covariance kernel $\Sigma(x - x')$ where $\Sigma(x) = \gamma \mathrm{diag}(e^{-\lambda|x|}, \ldots, e^{-\lambda|x|})$. Here c_0 and c_1 are constant k-vectors and Σ_0 is a positive definite matrix of order $k \times k$. Also, $\gamma > 0$ and $\lambda > 0$ are constants. The justification behind choosing such a prior structure is that under this prior specification, $\mathrm{E}(\xi_l(x_i)) = c_0 + c_1 x_i$, that is, under the prior, the expected value of the locations $\xi_l(x)$ varies linearly with x, and the sample paths vary smoothly.

We model the distributions of the observations by a countable location-scale mixture of k-dimensional Gaussian distributions through the relations

$$Y|\{\xi,\ X\} \sim \sum_{j=1}^{\infty} p_j \mathrm{N}_k(\xi(X)+\eta_j, \sigma_j^2 I_k), \tag{5.3.5}$$

where I_k denotes the identity matrix of order k. This model allows for asymmetry in the error distribution. The weights $p_j,\ j=1,2,\ldots$ are again constructed using stick-breaking, using $q_j \overset{iid}{\sim} \mathrm{Be}(1, M_2)$. We put independent k-dimensional Gaussian and inverse gamma hyper-priors on η_j and σ_j^2 respectively. Note that the error distribution $\sum_{j=1}^{\infty} p_j \mathrm{N}_k(\eta_j, \sigma_j^2 I_k)$ does not satisfy $Q_\varepsilon(u)=0$ for any $u \in B_2^{(k)}$. So for any $u \in B_2^{(k)}$, our quantile regression model is formulated as

$$Q_{Y|x}(u) = \xi(x) + Q_\varepsilon(u). \tag{5.3.6}$$

To reduce the computational burden, we use a truncation approximation on both sets of stick-breaking weights. Thus the full hierarchical Bayesian model is given by

$$\begin{aligned}
&Y_i = \xi(X_i) + \varepsilon_i,\ i=1,\ldots,n,\\
&\varepsilon_i \overset{iid}{\sim} \sum_{j=1}^{J} p_j \mathrm{N}_k(\eta_j,\ \sigma_j^2 I_k),\\
&p_1 = q_1,\ p_l = q_l \prod_{r=1}^{l-1}(1-q_r),\ l=2,\ldots,J-1,\ p_J = 1-\sum_{l=1}^{J-1} q_l,\\
&\eta_j \overset{iid}{\sim} \mathrm{N}_k(c_\eta,\ s_\eta^2 I_k),\ \sigma_j^2 \overset{iid}{\sim} \mathrm{IG}(a,b)\ \ j=1,2,\ldots,J,\\
&q_l \overset{iid}{\sim} \mathrm{Be}(1, M_2),\ l=1,2,\ldots,J-1,\ M_2 \sim \mathrm{Ga}(a_{M_2}, b_{M_2})\\
&\xi(X_i) \overset{ind}{\sim} G_{X_i},\ G_{X_i} = \sum_{l=1}^{N} W_l \delta_{\xi_l(X_i)}\\
&\xi_l(X_i) = \alpha_l + \beta_l(X_i),\ l=1,\ldots,N,\\
&\alpha_1,\ldots,\alpha_N \overset{iid}{\sim} \mathrm{N}_k(c_0,\ \Sigma_0),\\
&(\beta_l(X_1),\ldots,\beta_l(X_n)) \overset{iid}{\sim} \mathrm{N}_{kn}((c_1 X_1,\ldots,c_1 X_n)),\ (\!(\Sigma(X_i - X_j))\!)),\ j=1,\ldots,N,\\
&W_1 = V_1,\ W_l = V_l \prod_{r=1}^{l-1}(1-V_r),\ l=2,\ldots,N-1,\ W_N = 1-\sum_{l=1}^{N-1} W_l,\\
&V_l \overset{iid}{\sim} \mathrm{Be}(1, M_1),\ l=1,\ldots,N-1,\ M_1 \sim \mathrm{Ga}(a_{M_1}, b_{M_1}),
\end{aligned}$$

where $\mathrm{Ga}(a,b)$ denotes a Gamma distribution with parameters a and b, and $\mathrm{IG}(a,b)$ denotes an inverse-gamma distribution. The truncated stick-breaking construction reduces

the computations to finitely many terms. We use a Markov Chain Monte Carlo (MCMC) method to compute the posterior estimates, which we describe in detail in the next section.

5.4 Posterior Computation

We use a block Gibbs sampler (Ishwaran and James 2001) for estimating the parameters. Suppose that, we have d distinct observations $X_1,\ldots,X_d$ on the covariate X, and for each X_i, we have n_i observations $Y_{i1},\ldots,Y_{in_i}$ on the response Y. We introduce two sets of latent variables as below for the ease of computation.

- Define $L=(L_1,\ldots,L_d)$ such that $L_i=l$ if and only if $\xi(X_i)=\xi_l(X_i)$, $i=1,\ldots,d$.
- Also define $Z=(Z_{11},\ldots,Z_{dn_d})$ such that $\varepsilon_{im}|\{Z_{im}=j\}\sim \mathrm{N}_k(\eta_j,\sigma_j^2 I_k)$.

We describe the posterior full conditional distributions for each parameter below.

1. To update $\alpha_1,\ldots,\alpha_N$:
   - Let $d^\star$ be the number of distinct values $\{L_j^\star : j=1,\ldots,d^\star\}$ of the vector L. If $l\notin\{L_j^\star,: j=1,\ldots,d^\star\}$, α_l is drawn from $\mathrm{N}_k(c_0,\Sigma_0)$.
   - If $l=L_j^\star,\ j=1,\ldots,d^\star$,

$$p(\alpha_{L_j^\star}|-)\propto \exp\{-\frac{1}{2}(\alpha_{L_j^\star}-c_0)'\Sigma_0^{-1}(\alpha_{L_j^\star}-c_0)\}$$
$$\times\prod_{\{i:L_i=L_j^\star\}}\prod_{r=1}^{n_i}\exp\{-\frac{1}{2\sigma^2_{Z_{ir}}}(Y_{ir}-\alpha_{L_j^\star}-\beta_{L_j^\star}(X_i)-\eta_{Z_{ir}})'$$
$$(Y_{ir}-\alpha_{L_j^\star}-\beta_{L_j^\star}(X_i)-\eta_{Z_{ir}})\},$$

     which is a k-dimensional Gaussian distribution with mean

$$\left\{\Sigma_0^{-1}+\sum_{i:L_i=L_j^\star}\sum_{r=1}^{n_i}\frac{1}{\sigma^2_{Z_{ir}}}I_k\right\}^{-1}\left\{\Sigma_0^{-1}c_0+\sum_{i:L_i=L_j^\star}\sum_{r=1}^{n_i}\frac{(Y_{ir}-\alpha_{L_j^\star}-\beta_{L_j^\star}(x_i)-\eta_{Z_{ir}})}{\sigma^2_{Z_{ir}}}\right\}$$

     and covariance matrix $\left\{\Sigma_0^{-1}+\sum_{i:L_i=L_j^\star}\sum_{r=1}^{n_i}\frac{1}{\sigma^2_{Z_{ir}}}I_k\right\}^{-1}$.

2. To update $\beta_l(X_1),\ldots,\beta_l(X_d),\ l=1,\ldots,N$:
   - If $l\notin\{L_j^\star,: j=1,\ldots,d^\star\}$,

$$(\beta_l(X_1),\ldots,\beta_l(X_d))\sim \mathrm{N}_{kd}(c_1X_1,\ldots,c_1X_d),((\Sigma(X_i-X_j)))).$$

- Denote $X^j = (X_i,\ i : L_i = L_j^\star)$ and $X^{-j} = (X_i,\ i : L_i \neq L_j^\star)$. Also, let $\beta_{L_j^\star}(X^j) = (\beta_{L_j^\star}(X_i),\ i : L_i = L_j^\star)$ and $\beta_{L_j^\star}(X^{-j}) = (\beta_{L_j^\star}(X_i),\ i : L_i \neq L_j^\star)$. Let $m_{j,-j}$ denote the conditional prior mean for $\beta_{L_j^\star}(X^j)$ given $\beta_{L_j^\star}(X^{-j})$. Also, let $V_{j,-j}$ denote the conditional prior covariance matrix for $\beta_{L_j^\star}(X^j)$ given $\beta_{L_j^\star}(X^{-j})$. Similarly, $m_{-j,j}$ and $V_{-j,j}$ denote the conditional prior mean and covariance matrix of $\beta_{L_j^\star}(X^{-j})$ given $\beta_{L_j^\star}(X^j)$ respectively. If $l = L_j^\star,\ j = 1,\ldots,d^\star$, then for i such that $L_i = L_j^\star$

$$p(\beta_{L_j^\star}(X^j)|-) \propto \exp\{-\frac{1}{2}(\beta_{L_j^\star}(X^j) - m_{j,-j})' V_{j,-j}^{-1}(\beta_{L_j^\star}(X^{-j}) - m_{j,-j})\} \times \prod_{\{i:L_i=L_j^\star\}} \prod_{r=1}^{n_i} \exp\{-\frac{1}{2\sigma^2_{Z_{ir}}}(Y_{ir} - \alpha_{L_j^\star} - \beta_{L_j^\star}(X_i) - \eta_{Z_{ir}})' (Y_{ir} - \alpha_{L_j^\star} - \beta_{L_j^\star}(X_i) - \eta_{Z_{ir}})\},$$

  which is a $kU_{L_j^\star}$-dimensional Gaussian distribution, where $U_l = \#\{i : L_i = l\}$.

- Also $p(\beta_{L_j^\star}(X^{-j})|-)$ is proportional to

$$\exp\{-\frac{1}{2}(\beta_{L_j^\star}(X^{-j}) - m_{-j,j})' V_{-j,j}^{-1}(\beta_{L_j^\star}(X^{-j}) - m_{-j,j})\},$$

  which is a $k(d - U_{L_j^\star})$-dimensional Gaussian distribution with mean vector $m_{-j,j}$ and covariance matrix $V_{-j,j}$.

- If our goal to estimate $Q_{Y|x}(u)$ for an arbitrary $x \in \mathfrak{X}$, for a fixed $u \in B_2^{(k)}$, $\beta_l(x)$ has to be sampled conditional on $\beta_l(X_1),\ldots,\beta_l(X_d)$, for $l = 1,\ldots,N$, which is a k-dimensional Gaussian distribution.

3. To update $W_1,\ldots,W_N$:

   The posterior full conditional for W is given by

$$p(w|-) \propto f_W(w|M_1) \prod_{l=1}^{N} w_l^{U_l},$$

   where $f_W(\cdot|M_1)$ is the generalized Dirichlet distribution (Wong (1998))

$$f_W(w|M_1) = M_1^{N-1} w_N^{M_1-1}(1-w_1)^{-1}(1-(w_1+w_2))^{-1} \times \cdots \times (1 - \sum_{l=2}^{N-2} w_l)^{-1}.$$

   The posterior for W is a generalized Dirichlet distribution as well and can be sampled as follows using latent Beta variables as follows.

- Generate $V_l \overset{ind}{\sim} \text{Be}(1+U_l, M_1+\sum_{r=l+1}^{N} U_r)$ for $l=2,\ldots,N-1$.
- Set $W_1=V_1$, $W_l=V_l\prod_{r=1}^{l-1}(1-V_r)$, for $l=2,\ldots,N-1$, and $W_N=1-\sum_{l=1}^{N-1}W_l$.

4. To update $L_1,\ldots,L_d$:

   Each L_i is drawn from $\{1,\ldots,N\}$ with probabilities proportional to

$$W_l\prod_{r=1}^{n_i}\exp\{-\frac{1}{2\sigma_j^2}(Y_{ir}-\alpha_l-\beta_l(X_i)-\eta_{Z_{ir}})'(Y_{ir}-\alpha_l-\beta_l(X_i)-\eta_{Z_{ir}})\},$$

   for $l=1,\ldots,N$.

5. To update $\eta_1,\ldots,\eta_J$:

   Suppose there are s distinct values of $Z_{ij}, i=1,\ldots,d, j=1,\ldots,n_i$, and we denote them by $Z_1^\star,\ldots,Z_s^\star$.

   - If $j\notin\{Z_r^\star : r=1,\ldots,s\}$, draw η_j from $\text{N}_k(c_\eta, s_\eta^2 I_k)$.
   - If $j\in\{Z_r^\star : r=1,\ldots,s\}$, draw η_j from

$$\begin{aligned}p(\eta_{Z_r^\star}|-)\propto \exp\{-\frac{1}{2s_\eta^2}(\eta_{Z_r^\star}-c_\eta)'(\eta_{Z_r^\star}-c_\eta)\}\\ \prod_{\{(i,l):Z_{il}=Z_r^\star\}}\exp\{-\frac{1}{2\sigma_{Z_r^\star}^2}(Y_{il}-\alpha_{L_i}-\beta_{L_i}(X_i)-\eta_{Z_r^\star})'\\ (Y_{il}-\alpha_{L_i}-\beta_{L_i}(X_i)-\eta_{Z_r^\star})\},\end{aligned}$$

     which is a k-dimensional Gaussian distribution, with mean vector

$$\left(\frac{1}{s_\eta^2}+\frac{1}{\sigma_{Z_r^\star}^2}\right)^{-1}\left(\frac{1}{s_\eta^2}c_\eta+\sum_{i,l:Z_{il}=Z_r^\star}\frac{Y_{il}-\alpha_{L_i}-\beta_{L_i}(X_i)-\eta_{Z_r^\star}}{\sigma_{Z_r^\star}^2}\right)$$

     and covariance matrix $\left(\frac{1}{s_\eta^2}+\frac{1}{\sigma_{Z_r^\star}^2}\right)^{-1}I_k$.

6. To update $\sigma_1^2,\ldots,\sigma_J^2$:

   - If $j\notin\{Z_r^\star : r=1,\ldots,s\}$, draw σ_j^2 from $\text{IG}(a,b)$.
   - If $j\in\{Z_r^\star : r=1,\ldots,s\}$, $p(\sigma_{Z_r^\star}^2|-)$ is proportional to

$$\begin{aligned}(\sigma_{Z_r^\star}^2)^{-a-1}e^{-b/\sigma_{Z_r^\star}^2}\prod_{\{(i,l):Z_{il}=Z_r^\star\}}\exp\{-\frac{1}{2\sigma_{Z_r^\star}^2}(Y_{il}-\alpha_{L_i}-\beta_{L_i}(X_i)-\eta_{Z_r^\star})'\\ (Y_{il}-\alpha_{L_i}-\beta_{L_i}(X_i)-\eta_{Z_r^\star})\}.\end{aligned}$$

7. To update $Z_{il}, i = 1,\ldots,d, l = 1,\ldots,n_i$:

   - Draw Z_{il} from $\{1,\ldots,J\}$ with probability proportional to

$$q_j \prod_{\{(i,l):Z_{il}=j\}} \exp\{-\frac{1}{2\sigma_j^2}(Y_{il}-\alpha_{L_i}-\beta_{L_i}(X_i)-\eta_j)'(Y_{il}-\alpha_{L_i}-\beta_{L_i}(X_i)-\eta_j)\}.$$

8. To update $p_1,\ldots,p_J$:

   The posterior full conditional for $p=(p_1,\ldots,p_J)$ is given by

$$p(p_1,\ldots,p_J|-) \propto f_p(p_1,\ldots,p_J|M_2)\prod_{j=1}^{J} p_j^{U_j^\star},$$

   with $U_j^\star = \#\{(i,l): Z_{il}=j\}$, and $f_p(p_1,\ldots,p_J|M_2)$ is the generalized Dirichlet distribution

$$f_p(p_1,\ldots,p_J|M_2) = M_2^{J-1} p_J^{M_2-1}(1-p_1)^{-1}(1-(p_1+p_2))^{-1}\times\cdots\times(1-\sum_{l=2}^{J-2} p_l)^{-1}.$$

   The posterior for $(p_1,\ldots,p_J)$ is a generalized Dirichlet distribution as well and can be sampled as follows using latent Beta variables as follows.

   - Generate $q_l \overset{ind}{\sim} \text{Be}(1+U_l^\star, M_2+\sum_{r=l+1}^{J} U_r^\star)$ for $l=2,\ldots,J-1$.
   - Set $p_1 = q_l$, $p_l = q_l\prod_{r=1}^{l-1}(1-q_l)$, for $l=2,\ldots,J-1$, and $W_J = 1-\sum_{l=1}^{J-1} W_l$.

9. To update $W_1,\ldots,W_N$:

   The posterior for $(W_1,\ldots,W_N)$ is also generalized Dirichlet distribution and can be sampled as follows using latent Beta variables as follows.

   - Generate $V_l \overset{ind}{\sim} \text{Be}(1+U_l, M_2+\sum_{r=l+1}^{N} U_r)$ for $l=2,\ldots,N-1$.
   - Set $W_1 = V_1$, $W_l = V_l\prod_{r=1}^{l-1}(1-V_r)$, for $l=2,\ldots,N-1$, and $W_N = 1-\sum_{l=1}^{N-1} W_l$.

10. To update M_1:

    - The posterior full conditional for M_1 is given by

$$p(M_1|-) \propto e^{-(b_{M_1}-\log W_N)M_1} M_1^{a_{M_1}+N-1},$$

      which is a $\text{Ga}(a_{M_1}+N, b_{M_1}-\log W_N)$ distribution.

11. To update M_2:

- The posterior full conditional for M_2 is given by

$$p(M_2|-) \propto e^{-(b_{M_2}-\log p_J)M_2} M_2^{a_{M_2}+J-1},$$

which is a $\text{Ga}(a_{M_2}+J, b_{M_2}-\log p_J)$ distribution.

The MCMC sampling scheme is implemented using the *nimble* package in R, which is a system for writing hierarchical Bayesian models highly compatible with *BUGS* and *JAGS*. The advantage of using *nimble* is that it compiles the models by generating C++ code, which makes the computation faster.

After generating the posterior samples, we need to take care of the violation of the condition $Q_\varepsilon(u) = 0$. For each of the B many MCMC iterations of $p = (p_1, \ldots, p_J)$, $\eta = (\eta_1, \ldots, \eta_J)$ and $\sigma^2 = (\sigma_1^2, \ldots, \sigma_J^2)$, we have to compute

$$Q_\varepsilon^b(u) = \underset{\theta \in \mathbb{R}^k}{\arg\min} \sum_{j=1}^{J} p_j^b \int \Phi_2(u, x-\theta) e^{-\|x-\eta_j^b\|^2/2(\sigma_j^2)^b} \, \mathrm{d}x, \; b = 1, \ldots, B. \tag{5.4.1}$$

When the response variable is two-dimensional, the integral inside (5.4.1) can be reduced to a one-dimensional integral by the following trick. For $k = 2$, we can do a polar transform and reduce the integral for each j in (5.4.1) to

$$\begin{aligned} &\left(\frac{1}{2\pi}\right)^{k/2} e^{-\|\theta-\eta_j\|^2/2\sigma_j^2} \sigma_j^2 \int_0^\infty r^2 e^{-r^2/2} \left(\int_0^{2\pi} e^{-r(\theta_1-\eta_{j1})\cos u - r(\theta_2-\eta_{j2})\sin u} \mathrm{d}u \right) \mathrm{d}r \\ &+ u_1(\eta_{j1}-\theta_1) + u_2(\eta_{j2}-\theta_2). \end{aligned} \tag{5.4.2}$$

Then, (5.4.2) reduces to

$$\begin{aligned} &\left(\frac{1}{2\pi}\right)^{k/2} e^{-\|\theta-\eta\|^2/2\sigma_j^2} \sigma_j^2 \int_0^\infty r^2 e^{-r^2/2} 2\pi I_0\left(r\sqrt{(\theta_1-\eta_1)^2 + (\theta_2-\eta_2)^2} \right) \mathrm{d}r \\ &u_1(\eta_1-\theta_1) + u_2(\eta_2-\theta_2), \end{aligned} \tag{5.4.3}$$

where I_0 is the modified Bessel function of first kind (Ifantis and Siafarikas (1990)). We can take a rectangular grid for θ and for each θ in that grid, we compute the integral in (5.4.3) and the minimizer gives us an approximation for $Q_\varepsilon^b(u)$. We then calculate $B^{-1}\sum_{b=1}^B Q_\varepsilon^b(u)$ and add this to the posterior mean of $\xi(X_i)$ for each $i = 1, \ldots, d$, which gives us the uth geometric quantile for Y given X_i, $i = 1, \ldots, d$.

There is another way of numerically computing $Q_\epsilon^b(u)$, which is a Monte Carlo integration. This method extends to dimensions higher than 2. For each $b \in \{1, \ldots, B\}$, we generate a large number of samples $t_1^b, \ldots, t_R^b$ from the mixture of k-dimensional Gaussian

distributions $\sum_{j=1}^{J} p_j \mathrm{N}_k(\eta_j, \sigma_j^2 I_k)$. We again take a grid for θ and for each θ in the grid and for each $b \in \{1,\ldots,B\}$. We compute the Monte Carlo average $R^{-1}\sum_{r=1}^{R}\Phi_2(u, t_r^b - \theta)$. We look for the minimizer of this Monte Carlo average which approximates $Q_\varepsilon^b(u)$.

5.5 Posterior Consistency

In this section, we prove the weak posterior consistency for any fixed uth geometric quantile of the collection of true conditional densities $\{f^\star(\cdot|x) : x \in \mathfrak{X} = [0,1]\}$. The model described in Section 5.3 can be alternatively written as

$$f(y|x) = \int \frac{1}{\sigma^k}\phi_k\left(\frac{y-\xi(x)-\eta}{\sigma}\right)\mathrm{d}G(\xi)\mathrm{d}Q(\eta,\sigma), \tag{5.5.1}$$

where $\phi_k(\cdot)$ denotes a k-dimesnional standard normal kernel, and $G \times Q$ is the mixing distribution. The measures G and Q are of the form

$$G = \sum_{h=1}^{\infty} W_h \delta_{\xi_h},$$
$$Q = \sum_{j=1}^{\infty} p_j \delta_{(\eta_j,\sigma_j)}.$$

For every $x \in \mathfrak{X} = [0,1]$, G_x is the induced measure of G through the evaluation map $\xi \mapsto \xi(x)$, which can be written as

$$G_x = \sum_{h=1}^{\infty} W_h \delta_{\xi_h(x)}.$$

The mixing distribution $G \times Q$ is given the prior $\mathscr{P}$ on $\mathfrak{M}(\mathscr{C}(\mathfrak{X}) \times \mathbb{R}^k \times \mathbb{R}^+)$, where $\mathfrak{M}(\Theta)$ is the space of probability distributions on Θ, and $\mathscr{C}(\mathfrak{X})$ is the space of continuous functions on $\mathfrak{X}$. The prior $\mathscr{P}$ on $\mathfrak{M}(\mathscr{C}(\mathfrak{X}) \times \mathbb{R}^k \times \mathbb{R}^+)$ induces a prior Π on the space of conditional densities $\mathscr{F}$ through the map $G \times Q \mapsto \int \phi_k\left(\frac{y-\xi(x)-\eta}{\sigma}\right)\mathrm{d}G(\xi)Q(\eta,\sigma)$. We put priors on G and Q through stick-breaking weights, as described in Section 5.3. The true density $f^\star \in \mathscr{F}$ is assumed to be of the form

$$f^\star(y|x) = \int \frac{1}{\sigma^k}\phi_k\left(\frac{y-\xi(x)-\eta}{\sigma}\right)\mathrm{d}G^\star(\xi)\mathrm{d}Q^\star(\eta,\sigma), \tag{5.5.2}$$

where $G^\star$ and $Q^\star$ are compactly supported probability measures on $\mathscr{C}(\mathfrak{X})$ and $\mathbb{R}^k \times \mathbb{R}^+$ respectively. Just like before, $G_x^\star$ is the induced measure from $G^\star$ through the evaluation

map $\xi \mapsto \xi(x)$, with $G^\star$ being compactly supported on $\mathbb{R}^k$. To prove the posterior consistency of the conditional geometric quantiles $Q_{Y|x}(u)$, we need to show the posterior consistency of the conditional distribution $F_{Y|x}$ around a neighborhood of the true distribution $F^\star_{Y|x}$. This cannot be derived from the weak posterior consistency of the joint distribution of X and Y. However, we can derive the posterior consistency of the δ-smoothed conditional distribution of Y given x, defined below. Let $P_{X,Y}$ denote the joint distribution of X and Y, and let $F_{X,Y}$ denote the CDF. for a chosen $\delta > 0$, the δ-smoothed posterior distribution function of Y given X is defined as

$$F_{\delta;Y|x}(y) = \frac{P_{X,Y}(|X-x| \le \delta, Y \le y)}{P_X(|X-x| \le \delta)}. \tag{5.5.3}$$

For $u \in B_2^{(k)}$, the uth geometric quantile of the δ-smoothed conditional distribution is defined as

$$Q_{\delta;Y|x}(u) = \underset{\theta \in \mathbb{R}^k}{\arg\min} \int \{\Phi_2(u, Y-\theta) - \Phi_2(u, Y)\} \mathrm{d}F_{\delta;Y|x}(y). \tag{5.5.4}$$

Our main theorem, which gives the posterior consistency for $\{Q_{\delta;Y|x}(u) : x \in \mathfrak{X}\}$ for a fixed $u \in B_2^{(k)}$ is stated below.

Theorem 5.5.1. *Assume that, for every $x \in \mathfrak{X}$ and $u \in B_2^{(k)}$,*

$$\inf_{\theta : \|\theta - \theta^\star\|_2 \ge \epsilon} \int \{\Phi_2(u, y-\theta) - \Phi_2(u, y)\} \mathrm{d}F^\star_{Y|x}(y) > \int \{\Phi_2(u, y-\theta^\star) - \Phi_2(u, y)\} \mathrm{d}F^\star_{Y|x}(y).$$

Then for $\delta_n \to 0$ sufficiently slowly and for every $x \in \mathfrak{X}$, $\Pi\{|Q_{\delta_n;Y|x}(u) - Q^\star_{Y|x}(u)| < \epsilon | (Y^n, X^n)\} \to 1$ a.s., for every $\epsilon > 0$.

Next, we state some auxiliary results. The first result Lemma guarantees that our chosen prior is sensible in the non-parametric setting, i.e., it has a large topological support. We consider support properties for weak neighborhoods of the type

$$\left\{ f : \sup_x \left| \int \{g(y) f(y|x) \mathrm{d}y - g(y) f^\star(y|x)\} \mathrm{d}y \right| < \epsilon \right\}, \tag{5.5.5}$$

for every bounded and continuous function $g : \mathbb{R}^k \to \mathbb{R}$.

Lemma 5.5.1. *For every bounded and continuous $g : \mathbb{R}^k \to [0,1]$,*

$$\Pi\left\{ f : \sup_x \left| \int \{g(y) f(y|x) \mathrm{d}y - g(y) f^\star(y|x)\} \mathrm{d}y \right| < \epsilon \right\} > 0. \tag{5.5.6}$$

The proof of Lemma in Section 5.8. We will need one more auxiliary result for proving

the weak consistency, and for that we will need a bit more notation here. We define the joint density of X and Y by $h(x,y)=f(y|x)q(x)$, where $q(x)$ is the fixed marginal density of X, assumed to be bounded away from 0 on $\mathfrak{X}$. Similarly, the true joint density of X and Y is given by $h^\star(x,y)=f^\star(y|x)q(x)$. Next, we define the notion of a weak neighborhood for the true conditional density $f^\star \in \mathscr{F}$.

Definition 5.5.1. A sub-base of a weak-neighborhood for a collection of conditional densities $\{f^\star(\cdot|x) : x \in \mathfrak{X}\}$ is defined as

$$W_{\epsilon,g}(f^\star)=\left\{f : \Big| \int_{\mathfrak{X}\times\mathbb{R}^k} g\,h - \int_{\mathfrak{X}\times\mathbb{R}^k} g\,h^\star \Big| < \epsilon \right\}, \tag{5.5.7}$$

for a bounded and continuous function $g : \mathfrak{X}\times\mathbb{R}^k$. A weak neighborhood base is formed by finite intersection of neighborhoods of the type (5.5.7).

Definition 5.5.2. The posterior $\Pi\{\cdot|(Y^n,X^n)\}$ is weakly consistent at $f^\star \in \mathscr{F}$ if for any bounded and continuous function g, $\Pi\{W_{\epsilon,g}(f^\star)|(Y^n,X^n)\} \to 1$ a.s.

However, the large topological support of the prior is not sufficient for proving the weak posterior consistency, The weak posterior consistency at $h^\star$ holds if the prior Π puts positive probability on Kullback-Leibler neighborhoods of $f^\star$, which is defined below.

Definition 5.5.3. For any $\epsilon > 0$, an ϵ-sized Kullback-Leibler (KL) neighborhood around $f^\star$ is defined as

$$K_\epsilon(f^\star)=\{f : \mathrm{KL}(h^\star,h) < \epsilon,\ h(x,y)=f(y|x)q(x), x \in \mathfrak{X}, y \in \mathbb{R}^k\},$$

where $\mathrm{KL}(h^\star,h)=\int h^\star \log(h^\star/h)$. Then if $\Pi\{K_\epsilon(f^\star)\}>0$ for every $\epsilon>0$, we say $f^\star \in \mathrm{KL}(\Pi)$.

Lemma 5.5.2. *For $f^\star \in \mathscr{F}$ of the form in* (5.5.2)*, $f^\star \in \mathrm{KL}(\Pi)$.*

The proof of Lemma 5.5.2 is presented in Section 5.8, which ensures the weak consistency of the posterior at $h^\star$.

5.6 Simulation Study

In this section, we demonstrate Bayesian median regression with a bivariate response Y and a univariate predictor X on simulated data, which is a special case of the method illustrated above. We compare our method with a couple of frequentist methods.

It would be interesting to investigate the cases where the error distribution is something other than Gaussian, so here we choose a bivariate t-distribution with degree of freedom 1,

non-centrality parameter $(0,0)$ and scale matrix I_2 (a symmetric heavy-tailed distribution) and a bivariate gamma distribution with shape and rate parameter 1 and correlation matrix

$$V = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix},$$

which is a skewed distribution. We draw 100 samples $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_{100})$ from the above mentioned error distributions, and the predictors $X = (X_1, \ldots, X_{100})$ are drawn from a $N(0,1)$ distribution. We form the response vector $Y = (Y_1, Y_2)$ as follows.

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ X^2 \end{pmatrix} + \epsilon. \tag{5.6.1}$$

We are interested in the $(0,0)$-th geometric quantile, that is, the spatial median of y given x. Next, we describe our chosen prior specifications. We have two sets of stick-breaking weights, $(p_1, p_2, \ldots,)$ and $(W_1, W_2, \ldots,)$. Both sets of stick-breaking weights are truncated at 20, i.e., we have chosen both N and J to be 20. Both set of stick-breaking weights are generated from the variables $V_l \overset{iid}{\sim} \text{Be}(1, M_1)$ and $q_j \overset{iid}{\sim} \text{Be}(1, M_2)$, $l = 1, \ldots, 20,\ j = 1, \ldots, 20$, where M_1 and M_2 are drawn from Gamma$(1,1)$ prior. Next, $(\eta_1, \ldots, \eta_{20})$ is drawn from a $N_2((0,0)^T, 10I_2)$ prior, and $(\sigma_1^2, \ldots, \sigma_{20}^2)$ is drawn from a IG$(1,1)$ prior. We choose c_0 to be $(1,1)^T$ and c_1 to be $(2, 1/2)$. Also, S_0 is chosen to be $10I_2$. For the matrix $((\Sigma(X_i - X_j)))$, γ is chosen to be 10 and λ is chosen to be $1/2$.

Since we could only prove a weak consistency theorem for quantiles of δ-smoothed posterior distributions, we here demonstrate quantile regression for δ-smoothed posterior distribution of Y given X, for some chosen δ. The weak posterior consistency ensures that, for some $\epsilon > 0$

$$\Pi\{\sup_{x,y} |F_{X,Y}(x,y) - F^{\star}_{X,Y}(x,y)| < \epsilon | (Y^n, X^n)\} \to 1, \tag{5.6.2}$$

It will be shown in Section 5.8 that δ should be bigger than ϵ for Theorem 5.5.1 to hold. We do not prove a convergence rate theorem here, but intuitively the rate of convergence should be $n^{-1/2}$. Hence we choose δ_n to be bigger then $n^{-1/2}$, say $n^{-1/3}$.

To compute the spatial median for the δ-smoothed posterior for the conditional distribution of Y given X, for each value x of X, we draw a sample $\tilde{x}$ from $N(0,1)$ truncated in $[x - \delta, x + \delta]$, and we draw samples from the posterior distribution of Y given $\tilde{x}$ following the steps in Section 5.4. Note that the spatial median of the true error distribution is equal to its center of symmetry, 0. We generate 5000 samples from the posterior distribution with a burn-in of 500. We report a mean square error (MSE) for the conditional spatial median ξ,

which is given by

$$\text{MSE} = \frac{1}{100}\sum_{i=1}^{100} \|\xi^\star(X_i) - \bar{\xi}(X_i)\|_2^2,$$

where $\bar{\xi}(x)$ is the posterior mean of the δ-smoothed spatial median of Y given x and $\xi^\star(x)$ is the true conditional spatial median of y given x. One well-known competing method is the frequentist linear multivariate median regression proposed by Bai et al. (1990), where the regression estimates are obtained by minimizing $\sum_{i=1}^{n} \|Y_i - \alpha - \beta X_i\|_2$ with respect to α and β. Another method is a non-parametric version of bivariate median regression method. For every x, we estimate the conditional median $\xi(x)$ by minimizing $\arg\min_{\theta\in\mathbb{R}^k} \sum_{i=1}^{100} \|y_i - \theta\|_2 p_i(x)$, where

$$p_j(x) = \frac{K((x - X_j)/h)}{\sum_{i=1}^{100} K((x - X_i)/h)}.$$

The bandwith h has been chosen using cross-validation, and K here is a standard normal kernel. The mean square errors for all three methods are shown in Table 5.1. Our method gives a lesser MSE than the other two methods, thus achieving a gain.

Table 5.1 MSE's for conditional spatial medians for our method (NP-Bayes), the frequentist linear median regression and the non-parametric median regression (NP-frequentist).

Error distribution	NP-Bayes	Frequentist linear	NP-frequentist
Bivariate t	9.40	17.43	14.52
Bivariate gamma	7.29	18.02	14.98

5.7 Application to Blood Pressure data

We do Bayesian bivariate median regression using our method on a data that appeared in Chakraborty (1999). This data (shown in Table 5.2) was collected by the Biological Sciences Division of the Indian Statistical Institute, Kolkata. This data contains systolic and diastolic blood pressures of 40 Marwari (an Indian ethnic group) females living in the Burrabazar area of Kolkata. Our objective is to model the relationship between the systolic and diastolic blood pressures and age for a normal Marwari female living in Kolkata. As demonstrated by Chakraborty (1999) (Also in Figure 5.1, which shows the scatterplot of the blood pressure values against age), the data has very high spread and a few outliers. Hence the mean

Table 5.2 Systolic and Diastolic Blood Pressure of Marwari females in Kolkata

Serial No.	Age	Systolic Pressure	Diastolic Pressure	Serial No.	Age	Systolic Pressure	Diastolic Pressure
1	52	130	80	21	26	130	84
2	21	120	88	22	76	160	90
3	60	180	100	23	37	110	80
4	38	110	90	24	48	130	90
5	19	100	70	25	40	160	112
6	50	170	100	26	36	150	90
7	32	130	84	27	39	140	100
8	41	120	80	28	38	110	74
9	36	140	84	29	16	110	70
10	57	170	106	30	48	130	100
11	52	110	80	31	22	120	80
12	19	120	80	32	30	110	70
13	17	110	70	33	19	120	80
14	16	120	80	34	39	124	84
15	67	160	90	35	38	130	94
16	42	130	90	36	45	120	84
17	44	140	90	37	22	130	80
18	56	170	100	38	20	120	86
19	32	150	94	39	18	120	80
20	21	140	94	40	31	112	80

regression is not very efficient, because the mean is sensitive to outliers. Thus a median regression would be appropriate in this situation.

We again model the conditional spatial median of Y=(Systolic Pressure (Y_1), Diastolic Pressure (Y_2) against X=age. For computing the δ-smoothed posterior for the conditional distribution of Y given X, we need to sample from the distribution of X truncated in the interval $[x-\delta, x+\delta]$ for every value x of X. For choosing c_0 and c_1, we run linear regressions Y_1 on X and Y_2 on X separately. Of course, the distribution of X is unknown, and we estimate the density of X using a Gaussian kernel. Based on the regression coefficients, we choose $c_0 = (100, 73)$ and $c_1 = (0.8, 0.35)$. We draw 20000 samples from the posterior distribution with a burn-in of 1000, and in Table 5.3, we show the estimated spatial medians for selected age values along with their coordinate-wise 95% credible intervals, constructed from the posterior samples. In Figure 5.2, we plot the conditional quantiles as a function of X, and it shows that both functions increase with X, which is expected. In Figure 5.3, we plot the two components of the spatial median against each other, with X=age as labels. This

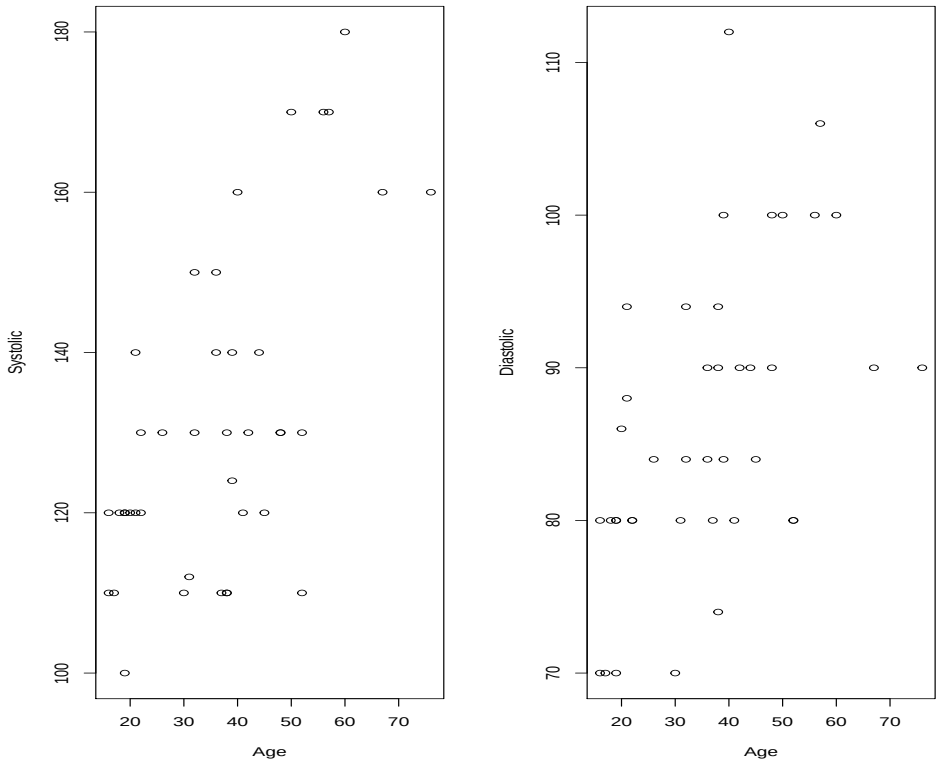


Figure 5.1 The Systolic and Diastolic Blood Pressure Against Age of 40 Marwari Females in Kolkata, India.

plot also shows an upward trend as well, i.e., diastolic pressure increases as the systolic pressure increases. It also shows that both pressures increase with age, which supports our conclusion from Figure 5.2.

Remark 2. Here we have considered Bayesian non-parametric quantile regression of a k-dimensional response on a univariate predictor, but the method can be extended to a general m-dimensional covariate as well. A DDP prior can be constructed in the same way, and a block Gibbs sampler algorithm can be used, but it would be a lot more computationally extensive.

Remark 3. Here we have considered multivariate quantile regression for geometric quantiles which are obtained by minimizing $P_{Y|x}\{\Phi_2(u, Y-\theta)-\Phi_2(u, Y)\}$ with $u \in B_2^{(k)}$. The method can be extended to a more general version of geometric quantiles with general ℓ_p-norm for $p > 1$, which are obtained by minimizing $P_{Y|x}\{\Phi_p(u, Y-\theta)-\Phi_p(u, Y)\}$ (See Section 2.1) for $u \in B_q^{(k)}$, where $p^{-1}+q^{-1}=1$.

Table 5.3 Spatial medians for Systolic (Y_1) and Diastolic (Y_2) pressure along with their coordinate-wise 95% credible intervals (in parenthesis) against selected x age values of Marwari females in Kolkata.

Serial No.	Age	Spatial Median (Y_1)	Spatial Median (Y_2)
1	16	119.66 (118.69, 120.92)	89.98 (88.47, 91.45)
2	21	120.18 (118.95, 121.39)	90.50 (89.41, 91.66)
3	26	130.42 (128.91, 131.80)	100.79 (99.44, 102.01)
4	31	116.92 (115.86, 118.32)	89.48 (86.07, 88.65)
5	36	116.85 (115.91, 117.74)	87.25 (85.84, 88.73)
6	41	115.66 (114.55, 116.84)	85.95 (84.76, 87.04)
7	46	132.10 (131.08, 133.39)	85.24 (84.65, 85.92)
8	51	140.62 (139.60, 141.91)	90.05 (89.46, 90.74)
9	56	117.79 (116.56, 118.97)	87.22 (86.02, 88.31)
10	61	152.30 (151.28, 153.59)	96.19 (95.59, 96.87)
11	66	153.89 (152.87, 155.18)	93.76 (93.17, 94.44)
12	71	160.91 (159.89, 162.20)	96.45 (95.86, 97.13)
13	76	163.19 (162.18, 164.49)	95.57 (94.97, 96.25)

5.8 Proofs

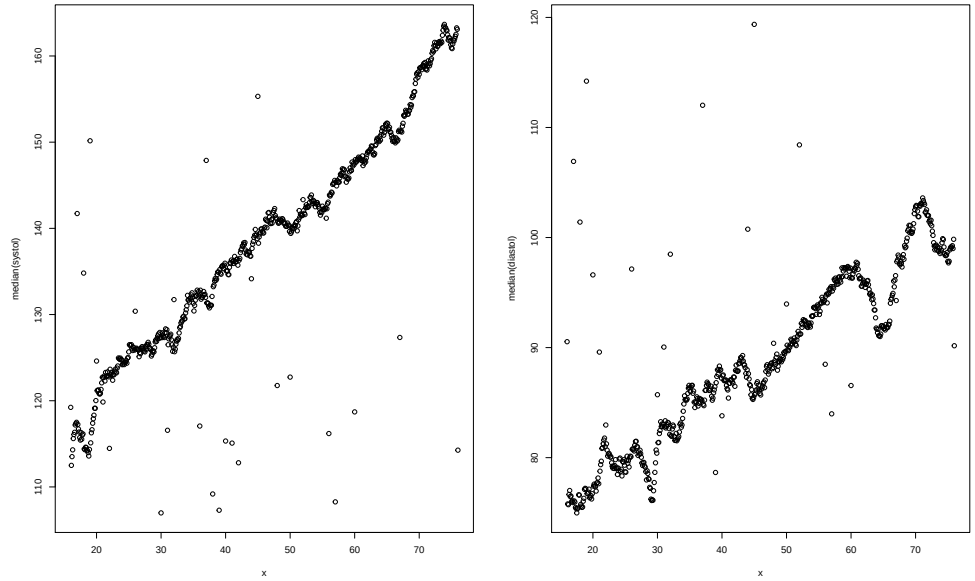
Proof of Lemma 5.5.1. The neighborhood in (5.5.5) can be written as

$$\left\{G\times Q : \sup_x \left| \int \gamma(\xi,\eta,\sigma)(x)\mathrm{d}G(\xi)\mathrm{d}Q(\eta,\sigma) - \int \gamma(\xi,\eta,\sigma)(x)\mathrm{d}G^\star(\xi)\mathrm{d}Q^\star(\eta,\sigma)\right| < \epsilon\right\}, \quad (5.8.1)$$

where $\gamma(\xi,\eta,\sigma)(x) = \int g(y)\sigma^{-k}\phi_k\big(\frac{y-\xi(x)-\eta}{\sigma}\big)\mathrm{d}y$. Note that

$$\begin{aligned}\int \gamma(\xi,\eta,\sigma)(x)\mathrm{d}G(\xi)\mathrm{d}Q(\eta,\sigma) = G(D_1)Q(D_2)\int \gamma(\xi,\eta,\sigma)(x)\mathrm{d}G_{D_1}(\xi)\mathrm{d}Q_{D_2}(\eta,\sigma)\\ + \int_{(D_1\times D_2)^c} \gamma(\xi,\eta,\sigma)(x)\mathrm{d}G(\xi)\mathrm{d}Q(\eta,\sigma),\end{aligned} \quad (5.8.2)$$

where G_{D_1} denotes the measure G restricted and normalized to a compact set $D_1 \subset \mathscr{C}(\mathfrak{X})$, and Q_{D_2} denotes the measure Q restricted and normalized to the compact set $D_2 \subset \mathbb{R}^k \times \mathbb{R}^+$. For every $x \in \mathfrak{X}$, the 2nd term in the right hand side of (5.8.2) can be bounded above by



(a) Systolic Pressure (Y_1) (b) Diastolic Pressure (Y_2)

Figure 5.2 Conditional spatial median for Systolic pressure (Y_1) and Diastolic pressure (Y_2) as a function of age (X).

$(G \times Q)(D_1 \times D_2)^c$. Then

$$\begin{aligned}
&\sup_x \left| \int \gamma(\xi,\eta,\sigma)(x) \mathrm{d}G(\xi)\mathrm{d}Q(\eta,\sigma) - \int \gamma(\xi,\eta,\sigma)\mathrm{d}G^\star(\xi)\mathrm{d}Q^\star(\eta,\sigma) \right| \\
&\quad \le \sup_x \left| \int \gamma(\xi,\eta,\sigma)(x)\mathrm{d}G_{D_1}(\xi)\mathrm{d}Q_{D_2}(\eta,\sigma) - \int \gamma(\xi,\eta,\sigma)(x)\mathrm{d}G^\star(\xi)Q^\star(\eta,\sigma) \right| \\
&\quad + \left| \frac{1}{G(D_1)Q(D_2)} - 1 \right| \left| \sup_x \left| \int \gamma(\xi,\eta,\sigma)(x)\mathrm{d}G(\xi)\mathrm{d}Q(\eta,\sigma) \right| + (G\times Q)(D_1\times D_2)^c \right. \qquad (5.8.3) \\
&\quad \le \sup_x \left| \int \gamma(\xi,\eta,\sigma)(x)\mathrm{d}G_{D_1}(\xi)\mathrm{d}Q_{D_2}(\eta,\sigma) - \int \gamma(\xi,\eta,\sigma)(x)\mathrm{d}G^\star(\xi)Q^\star(\eta,\sigma) \right| \\
&\quad + 2\frac{(G\times Q)(D_1\times D_2)^c}{(G\times Q)(D_1\times D_2)}.
\end{aligned}$$

The following lemma (Lemma 5.8.1) says that the family of functions $(\xi,\eta,\sigma) \mapsto \{\gamma(\xi,\eta,\sigma)(x) : x \in \mathfrak{X}\}$ is uniformly bounded and equicontinuous. Hence by the Arzela-Ascoli theorem, the family is pre-compact, and hence totally bounded. Hence for any $\epsilon > 0$, there exist $x_1,\ldots,x_s \in \mathfrak{X}$, such that for every $x \in \mathfrak{X}$, there exist $i = 1,\ldots,s$ such that

$$|\gamma(\xi,\eta,\sigma)(x) - \gamma(\xi,\eta,\sigma)(x_i)| < \epsilon, \qquad (5.8.4)$$

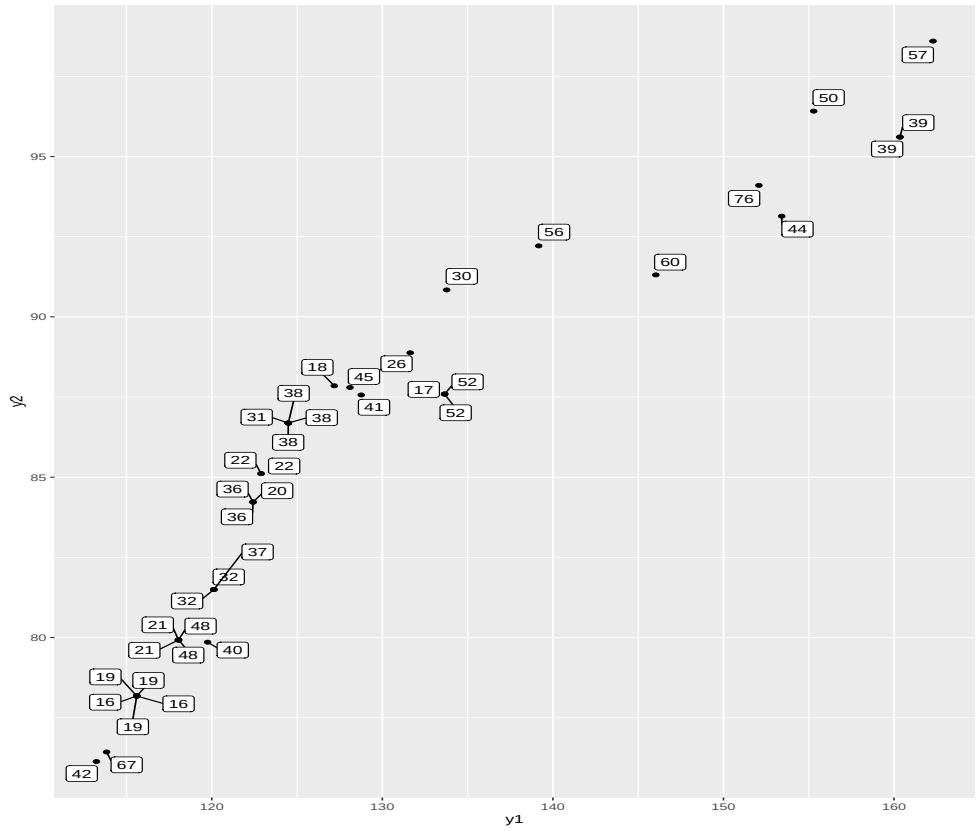


Figure 5.3 Conditional spatial medians (y_1, y_2) for Systolic and Diastolic Blood Pressure of 40 Marwari Females in Kolkata, India, with their age as labels.

for every $(\xi,\eta,\sigma)\in D$. Hence for every $x\in\mathfrak{X}$

$$\begin{aligned}
&\left|\int \gamma(\xi,\eta,\sigma)(x)\mathrm{d}G_{D_1}(\xi)\mathrm{d}Q_{D_2}(\eta,\sigma)-\int \gamma(\xi,\eta,\sigma)(x)\mathrm{d}G^\star(\xi)\mathrm{d}Q^\star(\eta,\sigma)\right| \\
&\leq\left|\int \gamma(\xi,\eta,\sigma)(x)\mathrm{d}G_{D_1}(\xi)\mathrm{d}Q_{D_2}(\eta,\sigma)-\int \gamma(\xi,\eta,\sigma)(x_i)\mathrm{d}G_{D_1}(\xi)\mathrm{d}Q_{D_2}(\eta,\sigma)\right| \\
&+\left|\int \gamma(\xi,\eta,\sigma)(x_i)\mathrm{d}G_{D_1}(\xi)\mathrm{d}Q_{D_2}(\eta,\sigma)-\int \gamma(\xi,\eta,\sigma)(x_i)\mathrm{d}G^\star(\xi)Q^\star(\eta,\sigma)\right| \\
&+\left|\int \gamma(\xi,\eta,\sigma)(x_i)\mathrm{d}G^\star(\xi)\mathrm{d}Q^\star(\eta,\sigma)-\int \gamma(\xi,\eta,\sigma)(x)\mathrm{d}G^\star(\xi)Q^\star(\eta,\sigma)\right|.
\end{aligned} \tag{5.8.5}$$

The first and third terms in the right hand side of (5.8.4) are bounded above by ϵ, using (5.8.3). For the second term, note that, since $G^\star(D_1)=Q^\star(D_2)=1$, for every $\epsilon>0$, there exists a weak neighborhood $W_1^\star$ of $G^\star$ and $W_2^\star$ of $Q^\star$ respectively in $\mathfrak{M}(\mathscr{C}(\mathfrak{X}))$ and $\mathfrak{M}(\mathbb{R}^k\times\mathbb{R}^+)$ such

that for every $G \in W_1^\star$ and $Q \in W_2^\star$, $G(D_1) > 1-\epsilon$, and $Q(D_2) > 1-\epsilon$, and for all $i = 1, \ldots, s$,

$$\left| \int \gamma(\xi, \eta, \sigma)(x_i) \mathrm{d}G(\xi) \mathrm{d}Q(\eta, \sigma) - \int \gamma(\xi, \eta, \sigma)(x_i) \mathrm{d}G^\star(\xi) \mathrm{d}Q^\star(\eta, \sigma) \right| < \epsilon. \tag{5.8.6}$$

Then for $G \in W_1^\star$, and $Q \in W_2^\star$,

$$\begin{aligned} &\left| \int \gamma(\xi, \eta, \sigma)(x_i) \mathrm{d}G_{D_1}(\xi) \mathrm{d}Q_{D_2}(\eta, \sigma) - \int \gamma(\xi, \eta, \sigma)(x_i) \mathrm{d}G^\star(\xi) \mathrm{d}Q^\star(\eta, \sigma) \right| \\ &\quad \leq \left| \frac{1}{G(D_1)Q(D_2)} - 1 \right| + \epsilon \\ &\quad \leq \epsilon + \frac{1-(1-\epsilon)^2}{(1-\epsilon)^2} \leq 4\epsilon \end{aligned}$$

if $\epsilon < 1 - \sqrt{3}/2$. Therefore, the right hand side in (5.8.3) is less than 6ϵ. Plugging everything in (5.8.3), the right hand side of it can be bounded above by 10ϵ. Thus, to show that the left hand side of (5.8.3) has positive prior probability, all we need to show is any weak neighborhood $W_1^\star$ of $G^\star$ and $W_2^\star$ of $Q^\star$ have positive prior probability. The measure G has a $\mathrm{DP}(M_1 G_0)$ prior with G_0 being a Gaussian process, having full support on $\mathscr{C}(\mathfrak{X})$. Similarly, Q has a $\mathrm{DP}(M_2 Q_0)$ prior, where Q_0 is the product measure of a k-dimensional Gaussian and an inverse gamma distribution, which also has a full support on $\mathbb{R}^k \times \mathbb{R}^+$. Thus, by Lemma 3.6 in Ghosal and van der Vaart (2017), the weak neighborhoods have positive prior probability. □

Lemma 5.8.1. *Define* $\gamma(\xi, \eta, \sigma)(x) = \int g(y) \sigma^{-k} \phi_k\big(\frac{y - \xi(x) - \eta}{\sigma}\big) \mathrm{d}y$, *where* $g : \mathbb{R}^k \to [0,1]$ *is bounded and continuous. The family of maps* $(\xi, \eta, \sigma) \mapsto \{\gamma(\xi, \eta, \sigma)(x) : x \in \mathfrak{X}\}$ *is uniformly equicontinuous as a family of functions of* (ξ, η, σ) *on the compact metric space* $D = D_1 \times D_2$, *i.e., for all* $x \in \mathfrak{X}$, *and all* $\|(\xi, \eta, \sigma) - (\xi', \eta', \sigma')\| < \delta$, *we have*

$$|\gamma(\xi, \eta, \sigma)(x) - \gamma(\xi', \eta', \sigma')(x)| < \epsilon. \tag{5.8.7}$$

Proof. For this proof, we borrow some ideas from the proof of Theorem 3 in Ghosal, Ghosh and Ramamoorthi (1999). Using the fact that $0 \leq g(\cdot) \leq 1$, for each $x \in \mathfrak{X}$, the left hand side of (5.8.7) can be bounded as

$$\begin{aligned} &\left| \int g(y) \sigma^{-k} \phi_k\left(\frac{y - \xi(x) - \eta}{\sigma}\right) \mathrm{d}y - \int g(y) \sigma'^{-k} \phi_k\left(\frac{y - \xi'(x) - \eta'}{\sigma'}\right) \mathrm{d}y \right| \\ &\qquad \leq \|\xi(x) + \eta - \xi'(x) - \eta'\| + |\sigma - \sigma'| \\ &\qquad \leq \|\xi - \xi'\|_\infty + \|\eta - \eta'\|_2 + |\sigma - \sigma'|. \end{aligned} \tag{5.8.8}$$

The last inequality follows from the Lipschitz continuity of $\phi_k(\cdot)$ as a function of (ξ,η,σ), which gives us the conclusion. □

Proof of Lemma 5.5.2. Note that $\mathrm{KL}(h^\star,h)$ can be decomposed as

$$\begin{aligned}\mathrm{KL}(h^\star,h)=\int_{\mathfrak{X}}\int_{\mathscr{K}} f^\star(y|x)\log\frac{f^\star(y|x)}{f(y|x)}\mathrm{d}y\,q(x)\mathrm{d}x+\\ \int_{\mathfrak{X}}\int_{\mathscr{K}^c} f^\star(y|x)\log\frac{f^\star(y|x)}{f(y|x)}\mathrm{d}y\,q(x)\mathrm{d}x,\end{aligned}\tag{5.8.9}$$

where $\mathscr{K}=\{y:\|y\|_2\le K\}$, for some $K>0$. First, we show that the 2nd term in the RHS of (5.8.9) is sufficiently small. Note that

$$\begin{aligned}&\int_{\mathfrak{X}}\int_{\mathscr{K}^c} f^\star(y|x)\log\frac{f^\star(y|x)}{f(y|x)}\mathrm{d}y\,q(x)\mathrm{d}x\\ &\qquad\le\int_{\mathfrak{X}}\int_{\mathscr{K}^c} f^\star(y|x)\log\frac{\sup_{(\xi,\eta,\sigma)\in D}\frac{1}{\sigma^k}\phi_k\left(\frac{y-\xi(x)-\eta}{\sigma}\right)}{\inf_{(\xi,\eta,\sigma)\in D}\frac{1}{\sigma^k}\phi_k(\frac{y-\xi(x)-\eta}{\sigma})G^\star(D_1)Q^\star(D_2)}\mathrm{d}y\,q(x)\mathrm{d}x,\end{aligned}$$

where D_1 and D_2 are compact metric spaces. For $\xi\in D_1$, $\sup_x\|\xi(x)\|<b^\star$ for some $b^\star>0$. Also, for $(\eta,\sigma)\in D_2$, $\|\eta\|_2<a^\star$, and $\underline{\sigma}<\sigma<\bar{\sigma}$, for $a^\star,\underline{\sigma},\bar{\sigma}>0$. For $\|y\|_2>K>a^\star+b^\star$,

$$\log\inf_{(\xi,\eta,\sigma)\in D}\frac{1}{\sigma^k}\phi_k\left(\frac{y-\xi(x)-\eta}{\sigma}\right)=\log\left\{\frac{1}{\underline{\sigma}^k}\phi_k\left(\frac{y+(a^\star+b^\star)\frac{y}{\|y\|}}{\underline{\sigma}^k}\right)\right\}.$$

Let $\mathscr{V}=\{G\times Q:(G\times Q)(D)>\underline{\sigma}^k/\bar{\sigma}^k\}$. Since $(G^\star\times Q^\star)(D)=1$, and D is an open set, $\mathscr{V}$ contains a neighborhood of $G^\star\times Q^\star$ of the form (5.5.5). Thus for every $G\times Q\in\mathscr{V}$,

$$
\begin{aligned}
&\int_{\mathfrak{X}}\int_{\mathscr{K}^C} f^\star(y|x)\log\frac{f^\star(y|x)}{f(y|x)}\mathrm{d}y\,q(x)\mathrm{d}x\\
&\quad\le \int_{\mathfrak{X}}\int_{\mathscr{K}^c} f^\star(y|x)\log\frac{\sup_{(\xi,\eta,\sigma)\in D}\phi_k\big(\frac{y-\xi(x)-\eta}{\sigma}\big)}{\inf_{(\xi,\eta,\sigma)\in D}\phi_k\big(\frac{y-\xi(x)-\eta}{\sigma}\big)}\mathrm{d}y\,q(x)\mathrm{d}x\\
&\quad= \int_{\mathfrak{X}}\int_{\mathscr{K}^c} f^\star(y|x)\log\frac{\phi_k\big(\frac{y+(a^\star+b^\star)\frac{y}{\|y\|_2}}{\bar\sigma}\big)}{\phi_k\big(\frac{y-(a^\star+b^\star)\frac{y}{\|y\|_2}}{\underline\sigma}\big)}\mathrm{d}y\,q(x)\mathrm{d}x\\
&=\int_{\mathfrak{X}}\int_{\mathscr{K}^c}\Big\{-\frac{1}{2\bar\sigma^k}\Big\|y+(a^\star+b^\star)\frac{y}{\|y\|_2}\Big\|_2^2+\frac{1}{2\bar\sigma^k}\Big\|y-(a^\star+b^\star)\frac{y}{\|y\|_2}\Big\|_2^2\Big\}f^\star(y|x)\mathrm{d}y\,q(x)\mathrm{d}x\\
&=\int_{\mathfrak{X}}\int_{\mathscr{K}^c}\{-\frac{1}{2\bar\sigma^k}\Big\|\|y\|_2+(a^\star+b^\star)\Big\|_2^2+\frac{1}{2\bar\sigma^k}\Big\|\|y\|_2-(a^\star+b^\star)\Big\|_2^2\}f^\star(y|x)\mathrm{d}y\,q(x)\mathrm{d}x.
\end{aligned}
$$

Since $f^\star$ is of the form (5.5.2), with ξ being uniformly bounded on D_1, and (η,σ) being bounded on D_2, for every $\epsilon>0$, we can find a compact set $\mathscr{K}$ such that

$$
\int_{\mathfrak{X}}\int_{\mathscr{K}^C} f^\star(y|x)\log\frac{f^\star(y|x)}{f(y|x)}\mathrm{d}y\,q(x)\mathrm{d}x<\frac{\epsilon}{2}. \tag{5.8.10}
$$

Next, we show that,

$$
\int_{\mathfrak{X}}\int_{\mathscr{K}} f^\star(y|x)\log\frac{f^\star(y|x)}{f(y|x)}\mathrm{d}y\,q(x)\mathrm{d}x<\frac{\epsilon}{2}. \tag{5.8.11}
$$

Following the arguments in Lemma 5.8.1, it can be shown that the family of maps $(\xi,\eta,\sigma)\mapsto\big\{\frac{1}{\sigma^k}\phi_k\big(\frac{y-\xi(x)-\eta}{\sigma}\big):y\in\mathscr{K},x\in\mathfrak{X}\big)\big\}$ is uniformly equicontinuous on D. Thus, the family is uniformly bounded on D, and pre-compact by Arzela-Ascoli theorem. Hence there exist $x_i,y_i,\ i=1,\ldots,m$ such that, for any $y\in\mathscr{K}$, $x\in\mathfrak{X}$,

$$
\sup_{(\xi,\eta,\sigma)\in D}\Big|\sigma^{-k}\phi_k(\frac{y-\xi(x)-\eta}{\sigma})-\sigma^{-k}\phi_k(\frac{y_i-\xi(x_i)-\eta}{\sigma})\Big|<c^\star\delta, \tag{5.8.12}
$$

where $c^\star=\sup_{x\in\mathfrak{X},y\in\mathscr{K}}\sup_{(\xi,\eta,\sigma)\in D}\Big|\sigma^{-k}\phi_k\big(\frac{y_i-\xi(x_i)-\eta}{\sigma}\big)\Big|$. Define

$$
\begin{aligned}
\mathscr{U}=\{G\times Q:|&\int_D\sigma^{-k}\phi_k\big(\frac{y-\xi(x)-\eta}{\sigma}\big)\mathrm{d}G^\star(\xi)\mathrm{d}Q^\star(\eta,\sigma)-\\
&\int_D\sigma^{-k}\phi_k\big(\frac{y_i-\xi(x_i)-\eta}{\sigma}\big)\mathrm{d}G(\xi)\mathrm{d}Q(\eta,\sigma)|<c^\star\delta,\ i=1,\ldots,m\}.
\end{aligned} \tag{5.8.13}
$$

Then $\mathscr{U}$ is a finite intersection of neighborhoods of $G^\star \times Q^\star$ of the form (5.5.5). Since $\mathrm{supp}(G \times Q) \subset D$,

$$\int_{\mathfrak{X}} \int_{\mathscr{K}} f^\star(y|x) \log \frac{f^\star(y|x)}{f(y|x)} \mathrm{d}y\, q(x) \mathrm{d}x < \int_{\mathfrak{X}} \int_{\mathscr{K}} f^\star(y|x) \log \frac{\int_D \phi_k(\frac{y-\xi(x)-\eta}{\sigma}) \mathrm{d}G^\star(\xi) \mathrm{d}Q^\star(\mu,\sigma)}{\int_D \phi_k(\frac{y-\xi(x)-\eta}{\sigma}) \mathrm{d}G(\xi) \mathrm{d}Q(\mu,\sigma)} \mathrm{d}y\, q(x) \mathrm{d}x.$$

Without loss of generality, we assume $(G^\star \times Q^\star)(\partial D) = 0$, where ∂X denotes the boundary of the set X. For any $(G \times Q) \in \mathscr{U}$, $y \in \mathscr{K}$ and $x \in \mathfrak{X}$,

$$\begin{aligned}
&\left| \int_D \frac{1}{\sigma^k} \phi_k\Big(\frac{y-\xi(x)-\eta}{\sigma}\Big) \mathrm{d}G^\star(\xi) \mathrm{d}Q^\star(\eta,\sigma) - \int_D \frac{1}{\sigma^k} \phi_k\Big(\frac{y-\xi(x)-\eta}{\sigma}\Big) \mathrm{d}G(\xi) \mathrm{d}Q(\eta,\sigma) \right| \\
&\leq \left| \int_D \frac{1}{\sigma^k} \phi_k\Big(\frac{y-\xi(x)-\eta}{\sigma}\Big) \mathrm{d}G^\star(\xi) \mathrm{d}Q^\star(\eta,\sigma) - \int_D \frac{1}{\sigma^k} \phi_k\Big(\frac{y_i-\xi(x_i)-\eta}{\sigma}\Big) \mathrm{d}G^\star(\xi) \mathrm{d}Q^\star(\eta,\sigma) \right| \\
&+ \left| \int_D \frac{1}{\sigma^k} \phi_k\Big(\frac{y_i-\xi(x_i)-\eta}{\sigma}\Big) \mathrm{d}G^\star(\xi) \mathrm{d}Q^\star(\eta,\sigma) - \int_D \frac{1}{\sigma^k} \phi_k\Big(\frac{y_i-\xi(x_i)-\eta}{\sigma}\Big) \mathrm{d}G(\xi) \mathrm{d}Q(\eta,\sigma) \right| \\
&+ \left| \int_D \frac{1}{\sigma^k} \phi_k\Big(\frac{y_i-\xi(x_i)-\eta}{\sigma}\Big) \mathrm{d}G(\xi) \mathrm{d}Q(\eta,\sigma) - \int_D \frac{1}{\sigma^k} \phi_k\Big(\frac{y-\xi(x)-\eta}{\sigma}\Big) \mathrm{d}G(\xi) \mathrm{d}Q(\eta,\sigma) \right|
\end{aligned} \tag{5.8.14}$$

The first and third terms in the RHS of (5.8.14) are each less than $c^\star \delta$ by (5.8.12). The second term is also less than $c^\star \delta$, since $(G \times Q) \in \mathscr{U}$. Thus,

$$\left| \int_D \frac{1}{\sigma^k} \phi_k\Big(\frac{y-\xi(x)-\eta}{\sigma}\Big) \mathrm{d}G^\star(\xi) \mathrm{d}Q^\star(\eta,\sigma) - \int_D \frac{1}{\sigma^k} \phi_k\Big(\frac{y-\xi(x)-\eta}{\sigma}\Big) \mathrm{d}G(\xi) \mathrm{d}Q(\eta,\sigma) \right| < 3c^\star \delta. \tag{5.8.15}$$

Therefore, for $(G \times Q) \in \mathscr{U}$,

$$\left| \frac{\int_D \frac{1}{\sigma^k} \phi_k\big(\frac{y-\xi(x)-\eta}{\sigma}\big) \mathrm{d}G^\star(\xi) \mathrm{d}Q^\star(\eta,\sigma)}{\int_D \frac{1}{\sigma^k} \phi_k\big(\frac{y-\xi(x)-\eta}{\sigma}\big) \mathrm{d}G(\xi) \mathrm{d}Q(\eta,\sigma)} - 1 \right| < \frac{3\delta}{1-3\delta}, \tag{5.8.16}$$

for $\delta < \frac{1}{3}$. Thus, by choosing δ small enough

$$\int_{\mathfrak{X}} \int_{\mathscr{K}} f^\star(y|x) \log \frac{f^\star(y|x)}{f(y|x)} \mathrm{d}y\, q(x) \mathrm{d}x \leq \sup_{x \in \mathfrak{X}, y \in \mathscr{K}} \left| \frac{\int_D \frac{1}{\sigma^k} \phi_k\big(\frac{y-\xi(x)-\eta}{\sigma}\big) \mathrm{d}G^\star(\xi) \mathrm{d}Q^\star(\eta,\sigma)}{\int_D \frac{1}{\sigma^k} \phi_k\big(\frac{y-\xi(x)-\eta}{\sigma}\big) \mathrm{d}G(\xi) \mathrm{d}Q(\eta,\sigma)} - 1 \right| < \frac{\epsilon}{2},$$

for $G \times Q \in \mathscr{U}$. Thus for any $\epsilon > 0$ and $G \times Q \in \mathscr{V} \cap \mathscr{U}$,

$$\int_{\mathfrak{X}} \int_{\mathbb{R}^k} f^\star(y|x) \log \frac{f^\star(y|x)}{f(y|x)} \mathrm{d}y\, q(x) \mathrm{d}x < \epsilon. \tag{5.8.17}$$

Thus Lemma 5.5.2 is proved. □

5.8.1 Proof of Theorem 5.5.1

Proof. Using Example 6.20 in Ghosal and van der Vaart (2017), Theorem 5.5.1 implies that the posterior is weakly consistent at $f^\star$, i.e., for any $W_{\epsilon,g}(f^\star)$

$$\Pi\{W_{\epsilon,g}(f^\star)|(Y^n, X^n)\} \to 1. \tag{5.8.18}$$

The above fact further implies that, for any $\epsilon > 0$

$$\Pi\{\sup_{x,y} |F_{X,Y}(x,y) - F^\star_{X,Y}(x,y)| < \epsilon |(Y^n, X^n)\} \to 1, \tag{5.8.19}$$

Thus there exists $\epsilon_n \downarrow 0$ such that (5.8.19) holds with ϵ_n replacing ϵ. Note that for any $\epsilon > 0$, $\delta > 0$ and F such that $\sup_{x,y} |F_{X,Y}(x,y) - F^\star_{X,Y}(x,y)| < \epsilon$, we have

$$\sup_{x,y} \left| \frac{P_{X,Y}(|X-x| \le \delta, Y \le y)}{P_X(|X-x| \le \delta)} - \frac{P^\star_{X,Y}(|X-x| \le \delta, Y \le y)}{P_X(|X-x| \le \delta)} \right| < \frac{\epsilon}{P_X(|X-x| \le \delta)},$$

Note that $P_X(|X-x| \le \delta) = \int_{x-\delta}^{x+\delta} q(x)\mathrm{d}x \ge 2\delta a$, with $a = \min_x q(x)$. Choosing a fixed sequence $\delta_n \downarrow 0$ at a rate slower than ϵ_n,

$$\sup_{x,y} \left| \frac{P_{X,Y}(|X-x| \le \delta_n, Y \le y)}{P_X(|X-x| \le \delta_n)} - \frac{P^\star_{X,Y}(|X-x| \le \delta_n, Y \le y)}{P_X(|X-x| \le \delta_n)} \right| < \epsilon_n,$$

for every n. Notice that $\lim_{\delta_n \to 0} \dfrac{P^\star_{X,Y}(|X-x| \le \delta_n, Y \le y)}{P_X(|X-x| \le \delta_n)} \to F^\star_{Y|x}(y)$. For a fixed $u \in B_2^{(k)}$, note that $Q_{\delta;Y|x}(u)$ can be written as $Q_{\delta;Y|x}(u) = \arg\max_\theta \int g(u;y,\theta) \mathrm{d}F_{\delta;Y|x}(y)$, where $g(u;y,\theta)$ is defined as

$$g(u;y,\theta) = -\{\|y-\theta\|_2 + \langle u, y-\theta\rangle - \|y\|_2 + \langle u, y\rangle\}. \tag{5.8.20}$$

Since $g(u;y,\theta)$ is a bounded and continuous function in y for every fixed $\theta \in \mathbb{R}^k$, and $x \in \mathfrak{X}$,

$$\Pi\left\{ \left| \int g(u;y,\theta) \mathrm{d}F_{\delta_n;Y|x}(y) - \int g(u;y,\theta) \mathrm{d}F^{\star}_{Y|x}(y) \right| < \epsilon | (Y^n, X^n) \right\} \to 1 \ a.s., \tag{5.8.21}$$

for every $\epsilon > 0$. We use the argmax theorem (Theorem 5.7 in van der Vaart (2000)) to achieve the assertion in Theorem 5.5.1. We need the following two conditions:

1. For every $\epsilon > 0$ and fixed $u \in B_2^{(k)}$, and for all $x \in \mathfrak{X}$,

$$\Pi\left\{ \sup_{\theta} \left| \int g(u;y,\theta) \mathrm{d}F_{\delta_n;Y|x}(y) - \int g(u;y,\theta) \mathrm{d}F^{\star}_{Y|x}(y) \right| < \epsilon | (Y^n, X^n) \right\} \to 1. \tag{5.8.22}$$

2. $\sup_{\theta : \|\theta - \theta^{\star}\|_2 \geq \epsilon} \int g(u;y,\theta) \mathrm{d}F^{\star}_{Y|x}(y) < \int g(u;y,Q^{\star}_{Y|X}(u|x)) \mathrm{d}F^{\star}_{Y|x}(y)$, which is also known as the "well-separatedness" condition.

To prove the above conditions, we need to restrict the parameter space to a compact subset of $\mathbb{R}^k$, which leads us to the following lemma, which says that the parameter space can be taken to be a compact set with high probability.

Lemma 5.8.2. *For every $x \in \mathfrak{X}$, and for every fixed $u \in B_2^{(k)}$, for every $0 < \epsilon < c^{-1}/(c^{-1} + \|u\|_2 + 1)$ and $K_x > 0$ such that $P^{\star}_{Y|x}(\|Y\|_2 \leq K_x) > 1 - \epsilon$, the posterior probability of $Q_{Y|x}(u) \leq c K_x$ given (Y^n, X^n) tends to 1, a.s. $n \to \infty$, where $c = 3/(1 - \|u\|_2)$.*

Proof of Lemma 5.8.2 is given at the end of this proof. The proof makes use of the arguments given in Lemma 2.8.3. Using Lemma 5.8.2, the parameter space can be taken to be Θ, which is a compact subset of $\mathbb{R}^k$. Condition 1 is proved using Bickel and Millar (1992). We have to show that, for every $\theta \in \Theta$ with Θ compact, and $u \in B_2^{(k)}$,

- $\sup_y |g(u;y,\theta)| \leq k_0$
- $\sup_y \{|g(u;y,\theta) - g(u;y',\theta)| / \|y - y'\|_2\} \leq k_0$.

The first condition follows from

$$|g(u;y,\theta)| \leq \|\theta\|_2 + \langle u, \theta \rangle \leq 2\|\theta\|_2 \leq 2c K_x,$$

The second condition follows from the Lipschitz continuity of the functions $g(u;y,\theta)$,

$$\begin{aligned}|g(u;y,\theta)-g(u;y',\theta)| =& |\|y-\theta\|_2-\|y'-\theta\|_2-\|y\|_2+\|y'\|_2| \\ \leq& 2\|y-y'\|_2.\end{aligned}$$

Then

$$\sup_y\{|g(u;y,\theta)-g(u;y',\theta)|/\|y-y'\|_2\}\leq 2.$$

Condition 2 follows from our assumption, which proves Lemma 5.5.2. □

Proof of Lemma 5.8.2. Define $M(F^\star_{Y|x},\theta)=F^\star_{Y|x}\{\Phi_2(u,Y-\theta)-\Phi_2(u,Y)\}=F^\star_{Y|x}(\|Y-\theta\|_2-\|Y\|_2-\langle u,\theta\rangle)$. We show that for $0<\epsilon<c^{-1}/(c^{-1}+\|u\|_2+1)$, there exists $K_x>0$ such that $\|\theta\|_2\geq cK_x$ implies $M(F^\star_{Y|x},\theta)>0$. If $\|Y\|_2\leq K_x$ and $\|\theta\|_2\geq cK_x$, then

$$\|Y-\theta\|_2\geq\|\theta\|_2-\|Y\|_2\geq\frac{(c-1)\|\theta\|_2}{c}+K-\|Y\|_2\geq\frac{\|\theta\|_2}{c},$$

Hence as $\|Y\|_2\leq K_x\leq\|\theta\|_2/c$,

$$\|Y-\theta\|_2-\|Y\|_2-\langle u,\theta\rangle\geq\frac{(c-1)\|\theta\|_2}{c}-\frac{\|\theta\|_2}{c}-\|u\|_2\|\theta\|_2.$$

Using the relation $c=3/(1-\|u\|_2)$

$$\|Y-\theta\|_2-\|Y\|_2-\langle u,\theta\rangle\geq\frac{\|\theta\|_2}{c}.$$

Now since always $\big|\|Y-\theta\|_2-\|Y\|_2-\langle u,\theta\rangle\big|\leq(1+\|u\|_2)\|\theta\|_2$, we can write

$$\begin{aligned}M(F^\star_{Y|x},\theta)=&\int_{\|Y\|_2\leq K_x}(\|Y-\theta\|_2-\|Y\|_2-\langle u,\theta\rangle)\mathrm{d}F^\star_{Y|x}+\\ &\qquad\int_{\|Y\|_2>K_x}(\|Y-\theta\|_2-\|Y\|_2-\langle u,\theta\rangle)\mathrm{d}F^\star_{Y|x}\\ &\geq\|\theta\|_2\big(\frac{1}{c}F^\star_{Y|x}(\|Y\|_2\leq K_x)-(1+\|u\|_2)F^\star_{Y|x}(\|Y\|_2>K_x)\big)\\ &=\|\theta\|_2\big(\frac{1}{c}-(1+\|u\|_2+\frac{1}{c})F^\star_{Y|x}(\|Y\|_2>K_x)\big)\\ &\geq\|\theta\|_2\Big\{\frac{1}{c}-\big(1+\|u\|_2+\frac{1}{c}\big)\epsilon\Big\}>0.\end{aligned}$$

Thus, for $u\in B_2^{(k)}$, and every $x\in\mathfrak{X}$, $Q^\star_{Y|x}(u)\leq cK_x$, where K_x is chosen such that $P^\star_{Y|x}(\|Y\|_2\leq K_x)>1-\epsilon$, where $0<\epsilon<c^{-1}/(c^{-1}+\|u\|_2+1)$. Since the δ_n-smoothed conditional distribu-

tion $F_{\delta_n;Y|x}$ is weakly consistent at $F^{\star}_{Y|x}$, the posterior probability $Q_{Y|x}(u) \leq c\,K_x$ tends to 1 almost surely. □

BIBLIOGRAPHY

[1] Abdous, B. & Theodorescu, R. "Note on the spatial quantile of a random vector". *Statistics and Probability Letters* **13**.4 (1992), pp. 333–336.

[2] Alquier, P. "PAC-Bayesian bounds for randomized empirical risk minimizers". *Mathematical Methods of Statistics* **17**.4 (2008), pp. 279–304. ISSN: 1066-5307.

[3] Babu, G. J. & Rao, C. R. "Joint asymptotic distribution of marginal quantiles and quantile functions in samples from a multivariate population". *Journal of Multivariate Analysis* (1989), pp. 15–23.

[4] Bai, Z., Chen, X., Miao, B. & Radhakrishna Rao, C. "Asymptotic theory of least distances estimate in multivariate linear models". *Statistics* **21**.4 (1990), pp. 503–519.

[5] Bhattacharya, I. & Ghosal, S. "Bayesian Inference on Multivariate Medians and Quantiles". *arXiv preprint arXiv:1909.10110* (2019).

[6] Bhattacharya, I. & Martin, R. "Gibbs posterior inference on multivariate quantiles". *arXiv preprint arXiv:2002.01052* (2020).

[7] Bickel, P. & Millar, P. "Uniform convergence of probability measures on classes of functions". *Statistica Sinica* (1992), pp. 1–15.

[8] Bissiri, P. G., Holmes, C. C. & Walker, S. G. "A general framework for updating belief distributions". *Journal of Royal Statistical Society B* **78**.5 (2016), pp. 1103–1130.

[9] Castillo, I. & Vaart, A. van der. "Needles and straw in a haystack: posterior concentration for possibly sparse sequences". *Annals of Statistics* **40**.4 (2012), pp. 2069–2101. ISSN: 0090-5364.

[10] Chakraborty, B. "On multivariate median regression". *Bernoulli* **5**.4 (1999), pp. 683–703.

[11] Chakraborty, B. "On multivariate quantile regression". *Journal of Statistical Planning and Inference* **110**.1-2 (2003), pp. 109–132.

[12] Chakraborty, B. & Chaudhuri, P. "On a transformation and re-transformation technique for constructing an affine equivariant multivariate median". *Proceedings of the American Mathematical Society* **124**.8 (1996), pp. 2539–2547.

[13] Chakraborty, B., Chaudhuri, P. & Oja, H. "Operating transformation retransformation on spatial median and angle test". *Statistica Sinica* (1998), pp. 767–784.

[14] Chakraborty, B. & Chaudhuri, P. "On an adaptive transformation–retransformation estimate of multivariate location". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**.1 (1998), pp. 145–157.

[15] Chang, C. “Nonparametric Bayesian Quantile Regression via Dirichlet Process Mixture Models” (2015).

[16] Chaudhuri, P. “On a geometric notion of quantiles for multivariate data”. *Journal of American Statistical Association* **91**.434 (1996), pp. 862–872.

[17] Cheng, G. “Moment consistency of the exchangeably weighted bootstrap for semiparametric M-estimation”. *Scandinavian Journal of Statistics* **42**.3 (2015), pp. 665–684.

[18] Cheng, G., Huang, J. Z., et al. “Bootstrap consistency for general semiparametric M-estimation”. *The Annals of Statistics* **38**.5 (2010), pp. 2884–2915.

[19] Chernozhukov, V. & Hong, H. “An MCMC approach to classical estimation”. *J.Econom* **115**.2 (2003), pp. 293–346.

[20] Cifarelli, D. & Regazzini, E. *Nonparametric statistical problems under partial exchangeability: The role of associative means*. Tech. rep. Quaderni Istituto Matematica Finanziaria of the University of Turin, 1978.

[21] De Iorio, M., Müller, P., Rosner, G. L. & MacEachern, S. N. “An ANOVA model for dependent random measures”. *Journal of the American Statistical Association* **99**.465 (2004), pp. 205–215.

[22] Drovandi, C. C. & Pettitt, A. N. “Likelihood-free Bayesian estimation of multivariate quantile distributions”. *Computational Statistics & Data Analysis* **55**.9 (2011), pp. 2541–2556.

[23] Duan, J. A., Guindani, M. & Gelfand, A. E. “Generalized spatial Dirichlet process models”. *Biometrika* **94**.4 (2007), pp. 809–825.

[24] Dudley, R. M. *Uniform Central Limit Theorems*. Cambridge University Press, 2014.

[25] Fraser, D. A. S. “Is Bayes posterior just quick and dirty confidence?” *Statistical Science* **26**.3 (2011), pp. 299–316. ISSN: 0883-4237.

[26] Gelfand, A. E., Kottas, A. & MacEachern, S. N. “Bayesian nonparametric spatial modeling with Dirichlet process mixing”. *Journal of the American Statistical Association* **100**.471 (2005), pp. 1021–1035.

[27] Ghosal, S., Ghosh, J. K., Ramamoorthi, R., et al. “Posterior consistency of Dirichlet mixtures in density estimation”. *Ann. Statist* **27**.1 (1999), pp. 143–158.

[28] Ghosal, S. & Vaart, A. van der. *Fundamentals of Nonparametric Bayesian Inference*. Vol. 44. Cambridge University Press, 2017.

[29] Ghosh, J. K. & Ramamoorthi, R. V. *Bayesian Nonparametrics*. New York: Springer-Verlag, 2003, pp. xii+305. ISBN: 0-387-95537-2.

[30] Grünwald, P. "The safe Bayesian: learning the learning rate via the mixability gap". *Algorithmic Learning Theory*. Vol. 7568. Lecture Notes in Computer Science. Springer, Heidelberg, 2012, pp. 169–183.

[31] Grünwald, P. & Ommen, T. van. "Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it". *Bayesian Analysis* **12**.4 (2017), pp. 1069–1103. ISSN: 1936-0975.

[32] Guggisberg, M. "A Bayesian approach to multiple-output quantile regression". *arXiv preprint arXiv:1909.02623* (2019).

[33] Hallin, M., Paindaveine, D., Šiman, M., Wei, Y., Serfling, R., Zuo, Y., Kong, L. & Mizera, I. "Multivariate quantiles and multiple-output regression quantiles: From L1 Optimization to halfspace depth [with Discussion and Rejoinder]". *Annals of Statistics* (2010), pp. 635–703.

[34] Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J. & Ostrowski, E. *A Handbook of Small Datasets*. London: Chapman and Hall/CRC, 1994.

[35] Hayford, J. F. "What is the center of an area, or the center of a population ?" *Journal of American Statistical Association* **8** (1992), pp. 47–58.

[36] Hjort, N. L. & Pollard, D. "Asymptotics for minimisers of convex processes". *arXiv preprint arXiv:1107.3806* (2011).

[37] Holmes, C. C. & Walker, S. G. "Assigning a value to a power likelihood in a general Bayesian model". *Biometrika* **104**.2 (2017), pp. 497–503. ISSN: 0006-3444.

[38] Huber, P. J. "The behavior of maximum likelihood estimates under nonstandard conditions". *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1. 1. University of California Press. 1967, pp. 221–233.

[39] Huber, P. J. *Robust Statistics*. New York: John Wiley & Sons Inc., 1981, pp. ix+308. ISBN: 0-471-41805-6.

[40] Ifantis, E. & Siafarikas, P. "Inequalities involving Bessel and modified Bessel functions". *Journal of mathematical analysis and applications* **147**.1 (1990), pp. 214–227.

[41] Ishwaran, H. & James, L. F. "Gibbs sampling methods for stick-breaking priors". *Journal of the American Statistical Association* **96**.453 (2001), pp. 161–173.

[42] Jiang, W. & Tanner, M. A. . "Gibbs posterior for variable selection in high-dimensional classification and data mining". *Annals of Statistics* **36**.5 (2008), pp. 2207–2231.

[43] Kleijn, B. J. K. & Vaart, A. W. van der. "The Bernstein-Von-Mises theorem under misspecification". *Electronic Journal of Statistics* **6** (2012), pp. 354–381.

[44] Koenker, R. & Bassett Jr, G. “Regression quantiles”. *Econometrica: Journal of the Econometric Society* (1978), pp. 33–50.

[45] Kottas, A. & Gelfand, A. E. “Bayesian semiparametric median regression modeling”. *Journal of the American Statistical Association* **96**.456 (2001), pp. 1458–1468.

[46] Liu, R. Y. “On a notion of simplicial depth”. *Proceedings of the National Academy of Sciences* **85**.6 (1988), pp. 1732–1734.

[47] Lo, A. Y. “A remark on the limiting posterior distribution of the multiparameter Dirichlet process”. *Sankhyā: The Indian Journal of Statistics, Series A* (1986), pp. 247–249.

[48] Lo, A. Y. “Weak convergence for Dirichlet processes”. *Sankhyā: The Indian Journal of Statistics, Series A* **45** (1983), pp. 105–111.

[49] Lopuhaa, H. P., Rousseeuw, P. J., et al. “Breakdown points of affine equivariant estimators of multivariate location and covariance matrices”. *The Annals of Statistics* **19**.1 (1991), pp. 229–248.

[50] Lyddon, S. P., Holmes, C. C. & Walker, S. G. “General Bayesian updating and the loss-likelihood bootstrap”. *Biometrika* **106**.2 (2019), pp. 465–478. ISSN: 0006-3444.

[51] MacEachern, S. N. “Dependent nonparametric processes”. *ASA proceedings of the section on Bayesian statistical Science*. Vol. 1. Alexandria, Virginia. Virginia: American Statistical Association. 1999, pp. 50–55.

[52] Martin, R. “False confidence, non-additive beliefs, and valid statistical inference”. *International Journal of Approximate Reasoning* **113** (2019), pp. 39–73. ISSN: 0888-613X.

[53] Martin, R. & Ning, B. “Empirical priors and coverage of posterior credible sets in a sparse normal mean model”. *Sankhya A*, to appear; `arXiv:1812.02150`. 2020.

[54] McAllester, D. A. “PAC-Bayesian model averaging”. *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (Santa Cruz, CA, 1999)*. ACM, New York, 1999, pp. 164–170.

[55] Möttönen, J., Nordhausen, K. & Oja, H. “Asymptotic theory of the spatial median”. *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková*. Institute of Mathematical Statistics, 2010, pp. 182–193.

[56] Nieto-Barajas, L. E., Müller, P., Ji, Y., Lu, Y. & Mills, G. B. “A Time-Series DDP for Functional Proteomics Profiles”. *Biometrics* **68**.3 (2012), pp. 859–868.

[57] Oja, H. “Descriptive statistics for multivariate distributions”. *Statistics & Probability Letters* **1**.6 (1983), pp. 327–332.

[58] Oja, H., Randles, R. H., et al. “Multivariate nonparametric tests”. *Statistical Science* **19**.4 (2004), pp. 598–605.

[59] Pas, S. van der, Szabó, B. & Vaart, A. van der. “Uncertainty quantification for the horseshoe (with discussion and rejoinder)”. *Bayesian Analysis* **12**.4 (2017), pp. 1221–1274. ISSN: 1936-0975.

[60] Robbins, H. & Monro, S. “A stochastic approximation method”. *Annals of Mathematical Statistics* **22** (1951), pp. 400–407. ISSN: 0003-4851.

[61] Scates, D. “Locating the median of the population in the United States.” *Metron* **11** (1933), pp. 49–66.

[62] Serfling, R. “Quantile functions for multivariate analysis: approaches and applications”. *Statistica Neerlandica* **56**.2 (2002), pp. 214–232.

[63] Sirkiä, S., Taskinen, S., Nevalainen, J. & Oja, H. “Multivariate nonparametrical methods based on spatial signs and ranks: The R package SpatialNP” (2007).

[64] Small, C. G. “A survey of multidimensional medians”. *International Statistical Review* (1990), pp. 263–277.

[65] Sun, S., Paisley, J. & Liu, Q. “Location Dependent Dirichlet Processes”. *International Conference on Intelligent Science and Big Data Engineering*. Springer. 2017, pp. 64–76.

[66] Syring, N., Hong, L. & Martin, R. “Gibbs posterior inference on value-at-risk”. *Scandinavian Actuarial Journal* (2019), pp. 1–10.

[67] Syring, N. & Martin, R. “Calibrating general posterior credible regions”. *Biometrika* **106**.2 (2019), pp. 479–486.

[68] Syring, N. & Martin, R. “Gibbs posterior concentration rates”. In preparation. 2020.

[69] Syring, N. & Martin, R. “Gibbs posterior inference on the minimum clinically important difference”. *Journal of Statistical Planning and Inference* **187** (2017), pp. 67–77.

[70] Syring, N. & Martin, R. “Robust and rate-optimal Gibbs posterior inference on the boundary of a noisy image”. *arXiv preprint arXiv:1606.08400* (2016).

[71] Tokdar, S. T. & Martin, R. “Bayesian test of normality versus a Dirichlet process mixture alternative”. *Sankhya B* (2019), pp. 1–31.

[72] Tomlinson, G. & Escobar, M. *Analysis of densities.* University of Toronto Technical report, 1999.

[73] Tukey, J. W. “Mathematics and the picturing of data”. *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*. Vol. 2. 1975, pp. 523–531.

[74] Vaart, A. W. van der. *Asymptotic Statistics*. Cambridge University Press, 2000.

[75] Van Der Vaart, A. W. & Wellner, J. A. “Weak Convergence and Empirical Processes”. Springer, 1996.

[76] Van der Vaart, A. “Efficiency of infinite dimensional M-estimators”. *Statistica Neerlandica* **49**.1 (1995), pp. 9–30.

[77] Waldmann, E. & Kneib, T. “Bayesian bivariate quantile regression”. *Statistical Modelling* **15**.4 (2015), pp. 326–344.

[78] Wang, Z. & Martin, R. “Model-free posterior inference on the area under the receiver operating characteristic curve”. *Journal of Statistical Planning and Inference* (2020).

[79] Wellner, J. A. & Zhan, Y. “Bootstrapping Z-estimators”. *University of Washington Department of Statistics Technical Report* **308** (1996).

[80] Wong, T.-T. “Generalized Dirichlet distribution in Bayesian analysis”. *Applied Mathematics and Computation* **97**.2-3 (1998), pp. 165–181.

[81] Xue-Kun Song, P. “Multivariate dispersion models generated from Gaussian copula”. *Scandinavian Journal of Statistics* **27**.2 (2000), pp. 305–320.

[82] Yu, K. & Moyeed, R. A. “Bayesian quantile regression”. *Statistics & Probability Letters* **54**.4 (2001), pp. 437–447.

[83] Zhang, T. “From ϵ-entropy to KL-entropy: analysis of minimum information complexity density estimation”. *Annals of Statistics* **34**.5 (2006), pp. 2180–2210. ISSN: 0090-5364.

[84] Zhang, T. “Information theoretical upper and lower bounds for statistical estimation”. *IEEE Transactions on Information Theory* **52**.4 (2006), pp. 1307–1321.