Statistica Sinica Preprint No: SS-2020-0108				
Title	Bayesian Inference on Multivariate Medians and			
	Quantiles			
Manuscript ID	SS-2020-0108			
URL	http://www.stat.sinica.edu.tw/statistica/			
DOI	10.5705/ss.202020.0108			
Complete List of Authors	Indrabati Bhattacharya and			
	Subhashis Ghosal			
Corresponding Author	Indrabati Bhattacharya			
E-mail	ibhatta@ncsu.edu			
Notice: Accepted version subject to English editing.				

Statistica Sinica

Bayesian Inference on Multivariate Medians and Quantiles

Indrabati Bhattacharya and Subhashis Ghosal

North Carolina State University

Abstract: In this paper, we consider Bayesian inference on a type of multivariate median and the multivariate quantile functionals of a joint distribution using a Dirichlet process prior. Since, unlike univariate quantiles, the exact posterior distribution of multivariate median and multivariate quantiles are not obtainable explicitly, 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 $n^{-1/2}$ -rate and its credible sets have asymptotically correct frequentist coverages. In particular, asymptotic normality results 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 the case 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 analogous results 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 a joint distribution of multivariate quantiles. The accuracy of the asymptotic result is confirmed by a simulation study. We also use the results to obtain Bayesian credible regions for multivariate medians for Fisher's iris data, which consists of four features measured for each of three plant species.

Key words and phrases: Affine equivariance, Bayesian bootstrap, Donsker class, Dirichlet process, Empirical process, Multivariate median.

1. Introduction

It is well known that 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 data 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

$$\hat{\theta}_{n;p} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \|X_i - \theta\|_p.$$
(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 (see Lopuhaa and Rousseeuw (Lopuhaa et al., 1991) for more details). The asymptotic properties of spatial median has also been investigated (see Möttönen et al. (2010) for more details). The ℓ_1 -median functional of a probability distribution Pbased on the ℓ_p -norm is given by

$$\theta_p(P) = \underset{\theta}{\arg\min} P\left(\|X - \theta\|_p - \|X\|_p \right), \tag{1.2}$$

for $Pf = \int f \, dP$ and $1 \leq p \leq \infty$. It can be noted that this definition does not require any moment assumption on X, since $|||X - \theta||_p - ||X||_p| \leq ||\theta||_p$. Henceforth, we fix $1 and drop p from the notations <math>\hat{\theta}_{n;p}$ 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 (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 (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, without any asymptotic approximations being required. Here we take a nonparametric Bayesian approach. We model the random distribution P and treat $\theta(P)$ as a functional of P. The most commonly used prior on a random distribution P is the Dirichlet process prior which we discuss in Section 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 context, the

4

Bernstein-von Mises theorem ensures that the Bayes estimator converges at the parametric rate $n^{-1/2}$ and a Bayesian $(1 - \alpha)$ 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 $\theta_0 \equiv \theta(P_0)$, where P_0 is the true value 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 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 θ_0 , 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 coordinatewise quantile (see Abdous and Theodorescu (1992), Babu and Rao (1989)). As Chaudhuri (1996) pointed out, the coordinatewise 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 $B^{(k)} := \{u : u \in \mathbb{R}^k, \|u\|_2 < 1\}$. For any $u \in B^{(k)}$, 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

7

to the ℓ_p -norm with 1 , we define the multivariate sample quantile process as

$$\hat{Q}_n(u) = \operatorname*{arg\,min}_{\xi \in \mathbb{R}^k} \frac{1}{n} \sum_{i=1}^n \Phi_p(u, X_i - \xi), \tag{1.3}$$

where $\Phi_p(u,t) = ||t||_p + \langle u,t \rangle$ with $u \in B_q^{(k)} \coloneqq \{u : u \in \mathbb{R}^k, ||u||_q < 1\}$ 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

$$Q_P(u) = \arg\min_{\xi \in \mathbb{R}^k} P\{\Phi_p(u, X - \xi) - \Phi_p(u, X)\}.$$
 (1.4)

with $Q_0(u) \equiv Q_{P_0}(u)$ being the multivariate quantile function for the true distribution P_0 .

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 noted that the joint distribution of $\{\sqrt{n}(\hat{Q}_n(u_1) - \hat{Q}_n(u_1)), \ldots, \sqrt{n}(\hat{Q}_n(u_m) - \hat{Q}_0(u_1)), \ldots, \sqrt{n}(\hat{Q}_n(u_m) - Q_0(u_m))\}$ 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 paper is organized as follows. In Section 2, we give the background needed to introduce the main results. In Section 3, we state the Bernstein-von Mises theorem for the multivariate ℓ_1 -median and the theorems we need to prove the same. In Sections 4 and 5, we present Bernstein-von Mises theorems for the affine-equivariant ℓ_1 -median and multivariate quantiles, respectively. In Section 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 7 and all the proofs are given in Section 8.

2. Background and Preliminaries

Before giving the background, we introduce some notations that we follow in this paper. Throughout this paper, $N_k(\mu, \Sigma)$ denotes a k-variate multivariate normal distribution with mean vector μ and covariance matrix Σ , and $\text{Gamma}_k(s, r, V)$ 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)). Also, $DP(\alpha)$ denotes a Dirichlet process with centering measure α (See Chapter 4, Ghosal and van

8

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 Pand 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 function

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

$$X_1, X_2, \dots, X_n | P \stackrel{\text{iid}}{\sim} P, \quad P \sim \text{DP}(\alpha).$$
 (2.1)

The posterior distribution of P given X_1, X_2, \ldots, X_n is $DP(\alpha + n\mathbb{P}_n)$, (see Chapter 4, 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 $P \sim$ $DP(\alpha + n\mathbb{P}_n)$ 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 $V_nQ + (1 - V_n)\mathbb{B}_n$, where the processes $Q \sim DP(\alpha)$, $\mathbb{B}_n \sim DP(n\mathbb{P}_n)$ and $V_n \sim \operatorname{Be}(|\alpha|, n)$ are independent and $\operatorname{Be}(a, b)$ 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 $(B_{n1}, B_{n2}, \dots, B_{nn})$ is a random vector following the Dirichlet distribution Dir(n; 1, 1, ..., 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 1), where $\theta(\mathbb{B}_n) = \arg\min_{\theta \in \mathbb{R}^k} ||X - \theta||_p$.

With the approximation in Lemma 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 (1996)), because it satisfies the system of equations $\Psi_n(\theta) = \mathbb{P}_n \psi(\cdot, \theta) = 0$,

where $\psi(\cdot, \theta) = (\psi_1(\cdot, \theta), \dots, \psi_k(\cdot, \theta))^T$ is a $k \times 1$ vector of functions from $\mathbb{R}^k \times \mathbb{R}^k$ to \mathbb{R} with

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

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 $\dot{\Psi}_0 = \int \dot{\psi}_{x,0} dP_0$, where

$$\dot{\psi}_{x,0} = \left[\frac{\partial\psi(x,\theta)}{\partial\theta}\right]_{\theta=\theta_0}$$

The matrix $\psi_{x,0}$ is given by

$$\dot{\psi}_{x,0} = \frac{p-1}{\|x-\theta_0\|_p} \Big[\operatorname{diag}\left(\frac{|x_1-\theta_{01}|^{p-2}}{\|x-\theta_0\|_p^{p-2}}, \dots, \frac{|x_k-\theta_{0k}|^{p-2}}{\|x-\theta_0\|_p^{p-2}}\right) - \frac{yy^T}{\|x-\theta_0\|_p^{2(p-1)}} \Big].$$
(3.1)

Moreover, y is given by

$$y = \left[|x_1 - \theta_{01}|^{p-1} \operatorname{sign}(x_1 - \theta_{01}), \dots, |x_k - \theta_{0k}|^{p-1} \operatorname{sign}(x_k - \theta_{0k}) \right]^T.$$

Also, we denote $\Sigma_0 = \frac{yy^T}{\|x - \theta_0\|_p^{2(p-1)}}.$

Theorem 3.1. Let $p \ge 2$ be a fixed integer. Suppose that the following conditions hold for $k \ge 2$.

- C1. The true probability distribution of $X \in \mathbb{R}^k$, P_0 has a probability density that is bounded on compact subsets of \mathbb{R}^k .
- C2. The ℓ_1 -median of P_0 , given by $\theta_0 = \theta(P_0)$, is unique.

Then

(i)
$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow \mathcal{N}_k(0, \dot{\Psi}_0^{-1} \Sigma_0 \dot{\Psi}_0^{-1}),$$

(ii) given $X_1, \ldots, X_n, \sqrt{n}(\theta(P) - \hat{\theta}_n) \rightsquigarrow N_k(0, \dot{\Psi}_0^{-1} \Sigma_0 \dot{\Psi}_0^{-1})$ in P_0 -probability.

Further if k = 2, (i) and (ii) hold for any 1 .

The uniqueness holds unless P_0 is completely supported on a straight line in \mathbb{R}^k , for $k \ge 2$ (Section 3, Chaudhuri (1996)). As we have pointed out before, finding the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ 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

12

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 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 nonnegative 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 $\sqrt{n}(\hat{\theta}_n - \theta_0)$, 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 1 proves Theorem 3.1. In Section 8, we provide a detailed verification of the conditions of Wellner and Zhan's (1996) theorem in our situation.

3.1 Bootstrapping a Z-estimator

In this subsection, we state Wellner and Zhan's (1996) bootstrap theorem for Z-estimators. Let $W_n = (W_{n1}, W_{n2}, \dots, W_{nn})$ 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 $W = \{W_{ni}, i = 1, 2, ..., n, n = 1, 2, ...\}$ 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},..., W_{nn})^T are exchangeable for every n, i.e., for any permutation π = (π₁,..., π_n) of {1, 2, ..., n}, the joint distribution of π(W_n) = (W_{nπ1}, W_{nπ2},..., W_{nπn})^T is same as that of W_n.
- (ii) The weights $W_{ni} \ge 0$ 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 $0 < K < \infty$

$$\|W_{n1}\|_{2,1} = \int_0^\infty \sqrt{\hat{P}(W_{n1} \ge u)} \,\mathrm{d}u \le K. \tag{3.2}$$

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

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

Van der Vaart and Wellner (1996) 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

3.1 Bootstrapping a Z-estimator15

scheme with c = 1 because the left hand side in (v) is given by $n^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2 / \bar{Y}_n^2 \xrightarrow{P} \operatorname{Var}(Y) / \{ \operatorname{E}(Y) \}^2 = 1$. To apply the bootstrap theorem, we also need to assume that the function class

$$\mathcal{F}_{R} = \{\psi_{j}(\cdot, \theta) : \|\theta - \theta_{0}\|_{2} \le R, \ j = 1, 2, \dots, k\}$$
(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 $\mathcal{F} \in \mathfrak{m}(P)$ if \mathcal{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 the supplement), or \mathcal{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 3.2 (Wellner and Zhan). Assume that the class of functions $\mathcal{F} \in \mathfrak{m}(P_0)$ and the following conditions hold.

1. There exists a $\theta_0 \equiv \theta(P_0)$ such that

$$\Psi(\theta_0) = P_0 \psi(X, \theta_0) = 0.$$
(3.4)

The function $\Psi(\theta) = P_0 \psi(X, \theta)$ is differentiable at θ_0 with nonsingular

Bootstrapping a Z-estimator16 3.1

derivative matrix $\dot{\Psi}_0$:

$$\dot{\Psi}_0 = \left[\frac{\partial\Psi}{\partial\theta}\right]_{\theta=\theta_0}.$$
(3.5)

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

$$\sup\left\{\frac{\|\mathbb{G}_{n}(\psi(\cdot,\theta) - \psi(\cdot,\theta_{0}))\|_{2}}{1 + \sqrt{n}\|\theta - \theta_{0}\|_{2}} : \|\theta - \theta_{0}\|_{2} \le \delta_{n}\right\} = o_{P_{0}}(1).$$
(3.6)

3. The k-vector of functions ψ is square-integrable at θ_0 with covariance

matrix

$$\Sigma_0 = P_0 \psi(X, \theta_0) \psi^T(X, \theta_0) < \infty.$$
(3.7)

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

$$D_n(x) = \sup\left\{\frac{|\psi_j(x,\theta) - \psi_j(x,\theta_0)|}{1 + \sqrt{n}\|\theta - \theta_0\|_2} : \|\theta - \theta_0\|_2 \le \delta_n, \ j = 1, \ 2, \dots, k\right\}$$
(3.8)

satisfy

$$\lim_{\lambda \to \infty} \limsup_{n \to \infty} \sup_{t \ge \lambda} t^2 P_0(D_n(X_1) > t) = 0.$$
(3.9)

- 4. The estimators $\hat{\hat{\theta}}_n$ and $\hat{\theta}_n$ are consistent for θ_0 , i.e., $\|\hat{\theta}_n \theta_0\|_2 \xrightarrow{P_0} 0$ and $\|\hat{\theta}_n - \hat{\theta}_n\|_2 \xrightarrow{\hat{P}} 0$ in P_0 -probability.
- 5. The bootstrap weights satisfy conditions (i)-(v).

Then

(i) $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow \mathcal{N}_k(0, \dot{\Psi}_0^{-1} \Sigma_0 \dot{\Psi}_0^{-1});$

(ii) $\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n) \rightsquigarrow N_k(0, c^2 \dot{\Psi}_0^{-1} \Sigma_0 \dot{\Psi}_0^{-1})$ in P_0 - 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) satisfies the conditions in Theorem 3.2, then Theorem 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.

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-andretransformation 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 independently and identically distributed 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

$$\hat{\phi}_n^{(\alpha)} = \operatorname*{arg\,min}_{\phi} \sum_{j \notin \alpha} \|Z_j^{(\alpha)} - \phi\|_2. \tag{4.1}$$

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

$$\hat{\theta}_n^{(\alpha)} = X(\alpha)\hat{\phi}_n^{(\alpha)}.$$
(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 4.1 Bernstein-von Mises theorem for the affine-equivariant multivariate median19

Oja (1998) proved that conditional on $X(\alpha)$, the asymptotic distribution of the transformed-and-retransformed spatial median is normal.

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_{Z0} , i.e., the distribution of Z when $X \sim P_0$. Hence the Bayesian model can be described as

$$Z_j^{(\alpha)}|P_Z \stackrel{iid}{\sim} P_Z, \quad P_Z \sim \mathrm{DP}(\beta), \ j \notin \alpha,$$
 (4.3)

which implies that

$$P_Z|\{Z_j^{(\alpha)}: j \notin \alpha\} \sim \mathrm{DP}(\beta + \sum_{j \notin \alpha} \delta_{Z_j}).$$
(4.4)

Following the same arguments used in Section 2, we can approximate the posterior Dirichlet process P_Z by the Bayesian bootstrap process \mathbb{B}_{n-k-1} ,

4.1 Bernstein-von Mises theorem for the affine-equivariant multivariate median20

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

$$\phi^{(\alpha)}(\mathbb{B}_{n}) = \arg\min_{\phi} \mathbb{B}_{n-k-1} \| Z^{(\alpha)} - \phi \|_{p},$$
(4.5)
$$\phi^{(\alpha)}(P_{Z}) = \arg\min_{\phi} \left\{ P_{Z}(\| Z^{(\alpha)} - \phi \|_{p} - \| Z^{(\alpha)} \|_{p}) \right\}.$$
(4.6)

Thus the transformed-and-retransformed medians are given by

$$\hat{\theta}_n^{(\alpha)} = X(\alpha)\hat{\phi}_n^{(\alpha)}, \quad \theta^{(\alpha)}(\mathbb{B}_n) = X(\alpha)\phi^{(\alpha)}(\mathbb{B}_n). \tag{4.7}$$

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 $\Psi_Z(\phi) = P\psi_Z(\cdot, \phi)$. The real-valued elements of the vector $\psi_Z(z, \phi)$ are then given by

$$\psi_{Z;j}(z,\phi) = \frac{|z_j - \phi_j|^{p-1}}{\|z - \phi\|_p^{p-1}} \operatorname{sign}(\phi_j - z_j), \quad j = 1, \dots, k.$$
(4.8)

Let $\phi_0^{(\alpha)} \equiv \phi^{(\alpha)}(P_{Z0})$ satisfy $\Psi_{Z0}(\phi^{(\alpha)}) = P_{Z0}\psi_Z(\cdot,\phi^{(\alpha)}) = 0$. In the following, we denote $\dot{\Psi}_{Z0}^{(\alpha)} = \left[\partial\Psi_{Z0}/\partial\phi\right]_{\phi=\phi_0^{(\alpha)}}$ and $\Sigma_{Z0}^{(\alpha)} = P_{Z0}\psi_Z(\cdot,\phi_0^{(\alpha)})\psi_Z^T(\cdot,\phi_0^{(\alpha)})$.

Theorem 4.1. Let $p \ge 2$ be a fixed integer. For $k \ge 2$ 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.

- 4.1 Bernstein-von Mises theorem for the affine-equivariant multivariate median21
- C1. The true distribution of $Z^{(\alpha)}$, P_{Z0} has a density which is bounded on compact subsets of \mathbb{R}^k .
- C2. The ℓ_1 -median of P_{Z0} , denoted by $\phi_0^{(\alpha)} = \phi^{(\alpha)}(P_{Z0})$, is unique.

Then

- (i) $\sqrt{n}(\hat{\theta}_{n}^{(\alpha)} \theta^{(\alpha)}(P_{0}))|\{X_{i} : i \in \alpha\} \rightsquigarrow$ $N_{k}(0, X(\alpha)\{\dot{\Psi}_{Z0}^{(\alpha)}\}^{-1}\Sigma_{Z0}^{(\alpha)}\{\dot{\Psi}_{Z0}^{(\alpha)}\}^{-1}\{X(\alpha)\}^{T});$
- (ii) given $X_1, \ldots, X_n, \sqrt{n}(\theta^{(\alpha)}(P) \hat{\theta}_n^{(\alpha)}) \rightsquigarrow N_k(0, X(\alpha) \{ \dot{\Psi}_{Z0}^{(\alpha)} \}^{-1} \Sigma_{Z0}^{(\alpha)} \{ \dot{\Psi}_{Z0}^{(\alpha)} \}^{-1} \{ X(\alpha) \}^T)$ in P_0 -probability. Here $\dot{\Psi}_{Z0}^{(\alpha)} = \int \dot{\psi}_{Z,0} \, dP_{Z0}, \text{ where}$

$$\dot{\psi}_{Z,0} = \left[\frac{\partial\psi_Z(z,\phi)}{\partial\phi}\right]_{\phi=\phi_0^{(\alpha)}}.$$
(4.9)

The matrix $\dot{\psi}_{Z,0}$ is given by

$$\dot{\psi}_{Z,0} = \frac{p-1}{\|z-\phi_0^{(\alpha)}\|_p} \Big[\operatorname{diag}\left(\frac{|z_1-\phi_{01}^{(\alpha)}|^{p-2}}{\|z-\phi_0^{(\alpha)}\|_p^{p-2}}, \dots, \frac{|z_k-\phi_{0k}^{(\alpha)}|^{p-2}}{\|z-\phi_0^{(\alpha)}\|_p^{p-2}}\right) - \frac{yy^T}{\|z-\phi_0^{(\alpha)}\|_p^{2(p-1)}}\Big],$$

with y given by

$$y = \left[|z_1 - \phi_{01}^{(\alpha)}|^{p-1} \operatorname{sign}(z_1 - \phi_{01}^{(\alpha)}), \dots, |z_k - \phi_{0k}^{(\alpha)}|^{p-1} \operatorname{sign}(z_k - \phi_{0k}^{(\alpha)}) \right]^T,$$
(4.10)

and
$$\Sigma_{Z0}^{(\alpha)} = \frac{yy^T}{\|z - \phi_{0k}^{(\alpha)}\|_p^{2(p-1)}}.$$

Further if k = 2, (i) and (ii) hold for any 1 .

The uniqueness holds unless P_{Z0} is completely supported on a straight line in \mathbb{R}^k , for $k \ge 2$, (Section 3, Chaudhuri (1996)). It can be noted that the DP(β) prior on P_Z induces the 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 3.1. Apart from Theorem 3.1, this theorem uses the affine equivariance of the normal family: if a random vector $X \sim N(\mu, \Sigma)$, then $Y = AX + b \sim N(A\mu + b, A\Sigma A^T)$.

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, ..., n, be independently and identically distributed observations from a k-variate distribution Pon \mathbb{R}^k and P is given the DP(α) prior. We study the posterior distributions asymptotically, and for every fixed $u_1, ..., u_m \in B_q^{(k)}$, Theorem 5.1 gives the joint posterior asymptotic distribution of the centered quantiles $\{\sqrt{n}(Q_P(u_1) - \hat{Q}_n(u_1)), ..., \sqrt{n}(Q_P(u_m) - \hat{Q}_n(u_m))\}.$

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 $Q_0(u) \equiv Q_{P_0}(u)$ and it satisfies the system of equations $\Psi_0^{(u)}(\cdot,\xi) = P_0\psi^{(u)}(\cdot,\xi) = 0$. The realvalued components of $\psi^{(u)}(\cdot,\xi)$ are then given by

$$\psi_j^{(u)}(x,\xi) = \frac{|x_j - \xi_j|^{p-1}}{\|x - \xi\|_p^{p-1}} \operatorname{sign}(\xi_j - x_j) + u_j, \quad j = 1, \dots, k.$$
(5.1)

Define $\dot{\Psi}_0^{(u)} = \int \dot{\psi}_{x,0}^{(u)} dP_0$, where

$$\dot{\psi}_{x,0}^{(u)} = \left[\frac{\partial\psi^{(u)}(x,\xi)}{\partial\xi}\right]_{\xi=Q_0(u)}.$$
(5.2)

The matrix $\dot{\psi}_{x,0}^{(u)}$ is given by

$$\dot{\psi}_{x,0}^{(u)} = \frac{p-1}{\|x-Q_0(u)\|_p} \Big[\operatorname{diag} \left(\frac{|x_1-Q_{01}(u)|^{p-2}}{\|x-Q_0(u)\|_p^{p-2}}, \dots, \frac{|x_k-Q_{0k}(u)|^{p-2}}{\|x-Q_0(u)\|_p^{p-2}} \right) - \frac{yy^T}{\|x-Q_0(u)\|_p^{2(p-1)}} \Big], \quad (5.3)$$

with y given by

$$y = \left[|x_j - Q_{0j}(u)|^{p-1} \operatorname{sign}(x_j - Q_{0j}(u)) : j = 1, \dots, k \right]^T.$$
 (5.4)

In the above, $Q_{0j}(u), j = 1, ..., k$ denotes the *j*th component of the vector $Q_0(u)$. We also define $\Sigma_{0;u,v} = P_0 \psi^{(u)}(x, Q_0(u)) \{\psi^{(v)}(x, Q_0(v))\}^T$.

Theorem 5.1. Let $p \ge 2$ be a fixed integer. Suppose that the following conditions hold for $k \ge 2$.

C1. The true distribution of X, P_0 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_0 , denoted by $Q_0(u_1), \ldots, Q_0(u_m)$, are unique.

Then

- (i) the joint distribution of (√n(Q̂_n(u₁) Q₀(u₁)),..., √n(Q̂_n(u_m) Q₀(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 {Ψ̂₀^{(u_j}}⁻¹ Σ_{0;u_j,u_l}{Ψ̂₀^{(u_l}}⁻¹, 1 ≤ j, l ≤ m;
- (ii) given X₁,...,X_n, the posterior joint distribution of {√n(Q_P(u₁) Q̂_n(u₁)),...,√n(Q_P(u_m) Q̂_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 {Ψ̂₀^{(u_j}}⁻¹Σ_{0;u_j,u_l}{Ψ̂₀^{(u_l}}⁻¹, 1 ≤ j, l ≤ m.

Further if k = 2, (i) and (ii) hold for any 1 .

Just like the ℓ_1 -median, the uniqueness of the quantiles holds unless P_0 is completely supported on a straight line on \mathbb{R}^k , (Section 3, Chaudhuri (1996)). We give the proof of the previous theorem in Section 8.

6. Simulation Study and a real data application

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 $P = 0.5 N_k(\mathbb{1}_k, I_k) + 0.5 \text{Gamma}_k(1, 1, V)$, 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 $r_{0.95}$ is the 95% 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:

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

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

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

Table 1 and Table 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.

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 same $DP(\alpha)$ prior with $\alpha = 2 \times N_4(0_4, 10I_4)$ is used. We construct the 95% Bayesian credible ellipsoid of the 4-dimensional spatial median and report its four principal axes in Table 1 of the supplementary document. Also, for the purpose of illustration, we plot 6 pairs of features for each species and the credible ellipsoids for the corresponding two dimensional spatial medians. The figures are given in the supplementary document.

7. Concluding Remarks

• The present paper is the first to study the asymptotic behavior of posterior distributions of multivariate median and quantiles. Multivariate quan-

	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	$0.955\ (5.97)$	0.980(6.19)

Table 1: Estimated coverage probability, mean size of the 95% credible ellipsoids and confidence 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.

tiles 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, skewness and outliers. The Bayesian approach gives us automatic uncertainty quantification through the posterior distributions without requiring any largesample 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

27

	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	$0.961 \ (5.99)$	9.50 (6.11)

Table 2: Estimated coverage probability, mean size of the 95% credible ellipsoids and confidence 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.

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 (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 $B_{2}^{(d)}$. We propose

28

a non-parametric Bayesian analog of the form $\int_{S} Q_{P}(u)\mu(du)$, 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 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 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.

8. Proof

8.1 Proof of Theorem 3.1

We will need some concepts on stochastic and empirical processes theory, which are given in the supplementary document. This includes definitions of covering numbers, bracketing numbers, uniform entropy, bracketing entropy, VC-classes, Glivenko-Cantell and Donsker classes, and a stochastically separable process.

We give the proof in two steps. In the first step, we state and prove Lemma 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 3.2 in our situation and show that the asymptotic conditional distribution of $\sqrt{n}(\theta(\mathbb{B}_n) - \hat{\theta}_n)$ is $N_k(0, \dot{\Psi}_0^{-1} \Sigma_0 \dot{\Psi}_0^{-1})$.

Lemma 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 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 $DP(\alpha + n\mathbb{P}_n)$. From the fact that $||P - \mathbb{B}_n||_{TV} = o_{P^*}(n^{-1/2})$ a.s. $[P_0^{\infty}]$, where $P^* = P^{\infty} \times \mathbb{B}_n$,

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

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

$$\left\|\Psi^{\star}(\theta(P)) - \Psi(\theta(P))\right\|_{2} = \left\|\Psi^{\star}(\theta(P))\right\|_{2} = o_{P^{\star}}(n^{-1/2}).$$
(8.1)

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 3.2 in this paper) holds for any sequence of bootstrap asymptotic Z-estimators $\hat{\theta}_n$ that satisfies

$$\|\Psi^{\star}(\hat{\hat{\theta}}_{n})\| = o_{P^{\star}}(n^{-1/2}).$$
 (8.2)

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) satisfies the conditions in Theorem 3.2. Firstly, we need to show that the function class $\mathcal{F}_R \in \mathfrak{m}(P_0)$ where \mathcal{F}_R is defined in (3.3). To achieve this, we prove that the empirical process $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_0)$ indexed by \mathcal{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 $\|\theta - \theta_0\|_2 \leq R$. Hence there exists a null set N and a countable $\mathcal{G} \subset \mathcal{F}_R$ such that, for every $\omega \notin N$ and $f \in \mathcal{F}_R$, we have a sequence $g_m \in \mathcal{G}$ 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 3.2. By Condition C2 in Theorem 3.1, the ℓ_1 -median of P_0 exists and is unique. Hence there exists a $\theta_0 \equiv \theta(P_0) \in \mathbb{R}^k$ such that (3.4) is satisfied. Also, $\Psi_0(\theta) = P_0 \psi(X, \theta)$ 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 , $P_0(||X - \theta||_2^{-1})$ is finite, which in turn implies that $P_0(||X - \theta||_p^{-1})$ 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 3.2. From Wellner and Zhan (1996), Condition 2 is satisfied if, \mathcal{F}_R in (3.3) is P_0 -Donsker for some R > 0 and

$$\max_{1 \le j \le k} P_0(\psi_j(\cdot, \theta) - \psi_j(\cdot, \theta_0))^2 \to 0,$$
(8.3)

as $\theta \to \theta_0$. In order to prove that \mathcal{F}_R is P_0 -Donsker, we define the following two function classes:

$$\mathcal{F}_{1R} = \Big\{ \frac{|x_j - \theta_j|^{p-1}}{\|x - \theta\|_p^{p-1}} : j = 1, 2, \dots, k, \|\theta - \theta_0\|_2 \le R \Big\},$$
(8.4)

$$\mathcal{F}_{2R} = \left\{ \operatorname{sign}(\theta_j - x_j) : j = 1, 2, \dots, k, \|\theta - \theta_0\|_2 \le R \right\}.$$
(8.5)

From Example 2.10.23 of van der Vaart and Wellner (1996), if \mathcal{F}_{1R} and \mathcal{F}_{2R} satisfy the uniform entropy condition and are suitably measurable, then $\mathcal{F}_{R} = \mathcal{F}_{1R}\mathcal{F}_{2R}$ is P_{0} -Donsker provided their envelopes F_{1R} and F_{2R} satisfy $P_{0}F_{1R}^{2}F_{2R}^{2} < \infty$.

Lemma 2. For k > 2, \mathcal{F}_{1R} and \mathcal{F}_{2R} are P_0 -Donsker classes for some fixed integer p, and hence they satisfy the uniform entropy condition.

The proof is presented in the supplementary document. In view of Lemma 2, next we need to prove (8.3), that is, $\max_{1 \le j \le k} P_0(\psi_j(\cdot, \theta) - \psi_j(\cdot, \theta_0))^2 \to 0$ as $\theta \to \theta_0$. Note that $\psi_j(x, \theta) \to \psi_j(x, \theta_0)$ for every x as $\theta \to \theta_0$ for $j \in \{1, \ldots, k\}$. Also $(\psi_j(x, \theta) - \psi_j(x, \theta_0))^2 \le 4$ for every x and every θ . Hence by the dominated convergence theorem, $P_0(\psi_j(\cdot, \theta) - \psi_j(\cdot, \theta_0))^2 \to 0$ as $\theta \to \theta_0$ for $j \in \{1, \ldots, k\}$. Thus (8.3) is established.

Verification of Condition 3 in Theorem 3.2. For every $j \in \{1, 2, ..., k\}$ 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 $\Sigma_0 = P_0 \psi(x, \theta_0) \psi^T(x, \theta_0)$ is given by

$$\sigma_{ij} = \int \frac{|x_i - \theta_{0i}|^{p-1} |x_j - \theta_{0j}|^{p-1}}{\|x - \theta_0\|_p^{2(p-1)}} \operatorname{sign}(\theta_{0i} - x_i) \operatorname{sign}(\theta_{0j} - x_j) dP_0 \quad (8.6)$$

$$\leq \int 1 dP_0 < \infty.$$

The class of functions $\{\psi_j(x,\theta) : j = 1, 2, ..., k, \|\theta - \theta_0\|_2 \leq R\}$ has a constant envelope 1. Hence $D_n(x)$ defined in (3.8) is equal to 2 and it satisfies (3.9).

Verification of Condition 4. First we prove $\|\hat{\theta}_n - \theta_0\|_2 \xrightarrow{P_0} 0$. Note that $\hat{\theta}_n$ can be written as

$$\hat{\theta}_n = \arg\max_{\theta} \mathbb{P}_n m_{\theta}, \tag{8.7}$$

where $m_{\theta}(x) = -\|x - \theta\|_p + \|x\|_p$. Naturally the population analog of $\hat{\theta}_n$ is

given by

$$\theta(P) = \arg\max_{\theta} \mathbb{P}m_{\theta}.$$
(8.8)

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

- (a) $\sup_{\theta} |\mathbb{P}_n m_{\theta} P_0 m_{\theta}| \to 0$ in probability;
- (b) there exists a θ_0 such that $P_0 m_{\theta_0} > \sup_{\theta \notin G} P_0 m_{\theta}$ for every open set G containing θ_0 .

The first condition can be proved by showing that the class of functions $\{m_{\theta} : \theta \in \mathbb{R}^k\}$ forms a P_0 -Glivenko-Cantelli class. From Theorem 19.4 of van der Vaart (2000), the class $\mathcal{M} = \{m_{\theta} : \theta \in \Theta \subset \mathbb{R}^k\}$ will be P_0 -Glivenko-Cantelli if $N_{[]}(\epsilon, \mathcal{M}, L_1(P_0)) < \infty$ for every $\epsilon > 0$.

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

$$|f_1(x) - f_2(x)| \le m(x) \|\theta_1 - \theta_2\|_2, \tag{8.9}$$

for every θ_1, θ_2 and $P_0|m|^r < \infty$, then there exists a constant K, depending on Θ and k only, such that the bracketing numbers satisfy

$$N_{[]}(\epsilon \|m\|_{P_0,r}, \mathcal{F}, \mathcal{L}_r(P_0)) \le K \left(\frac{\operatorname{diam} \Theta}{\epsilon}\right)^k, \qquad (8.10)$$

for every $0 < \epsilon < \text{diam } \Theta$. 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 3. If for some $0 < \epsilon < 1/4$ and K > 0, P_0 on $(\mathbb{R}^k, \mathscr{R}^k)$ satisfying $P_0(||X||_p \leq K) > 1 - \epsilon$ for any probability measure P_0 , then $||\theta(P_0)||_p \leq 3K$; where \mathscr{R}^k denotes the Borel sigma field on \mathbb{R}^k .

The proof of Lemma 3 is given in the supplementary document. Because of Lemma 3, it suffices to establish (8.9). 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 $1 , the expression is bounded by <math>2^{(1/p)-(1/2)}\|\theta - \theta'\|_2$. Hence we choose m(x) = 1 for every x and therefore $P_0|m| = 1$. This ensures that $N_{[]}(\epsilon, \mathcal{M}, \mathcal{L}_1(P_0)) < \infty$ and hence Condition (a) is satisfied. From Condition (C2) in Theorem 3.1, Condition (b) holds. Therefore $\hat{\theta}_n \to \theta_0$ in P_0 -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 $\sup_{\theta} |\mathbb{B}_n m_{\theta} -$ $P_0m_{\theta}| \stackrel{P_0 \times \mathbb{B}_n}{\to} 0$. 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 $(B_{n1}, \ldots, B_{nn}) \sim \operatorname{Dir}(n; 1, \ldots, 1)$, it follows that $B_{ni} \geq 0$, $\sum_{i=1}^n B_{ni} = 1$ and $B_{ni} \sim \operatorname{Be}(1, n - 1)$. Therefore, for every $\epsilon > 0$, as $n \to \infty$

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

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 $\hat{\theta}_n \xrightarrow{P_0} \theta_0$ and $\theta(\mathbb{B}_n) \xrightarrow{P_0 \times \mathbb{B}_n} \theta_0$. Hence by an application of the triangle inequality, $\theta(\mathbb{B}_n) \xrightarrow{\mathbb{B}_n} \hat{\theta}_n$ in P_0 -probability. \Box

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

Proof for arbitrary p > 1 when k = 2. When k = 2, we do not need p to be an integer because we can show that F_{1R} is a P_0 -Donsker class for any fixed p > 1, which we formally state in the following lemma.

Lemma 4. For k = 2, \mathcal{F}_{1R} is a P_0 -Donsker class for any p > 1, and hence it satisfies the uniform entropy condition.

Proof. Presented in the supplementary document. \Box

8.2 Proof of Theorem 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

$$\psi_{lj}(x,\xi_{lj}) = \frac{|x_j - \xi_{lj}|^{p-1}}{\|x - \xi_l\|_p^{p-1}} \operatorname{sign}(\xi_{lj} - x_j) + u_{lj}.$$
(8.11)

We define $Q_{\mathbb{B}_n}(u_1), \ldots, Q_{\mathbb{B}_n}(u_m)$ as the corresponding "Bayesian bootstrapped" 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 1 to the quantile case.

Lemma 5. 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 5 is same as that of Lemma 1, hence is omitted. The rest of the proof of Theorem 5.1 is along the lines of that of Theorem 3.1 as well, hence that is skipped too.

Bayesian Inference on Multivariate MEdians

Supplementary Document

In the supplementary document, we provide some background of the empirical process theory. We give definitions of covering numbers and uniform entropy, bracketing numbers, VC class of sets, Glivenko-Cantelli class of functions and Donsker class of functions. Additionally, we provide proofs of Lemmas 2, 3 and 4 and some details of the application on iris data. Finally, we provide some R codes used for the computation of our method.

References

- Abdous, B. and Theodorescu, R. (1992), 'Note on the spatial quantile of a random vector', Statistics & Probability Letters 13(4), 333–336.
- Babu, G. J. and Rao, C. R. (1989), Joint asymptotic distribution of marginal quantiles and quantile functions in samples from a multivariate population, in 'Multivariate Statistics and Probability', Elsevier, pp. 15–23.
- Chakraborty, B. and Chaudhuri, P. (1996), 'On a transformation and re-transformation technique for constructing an affine equivariant multivariate median', *Proceedings of the American Mathematical Society* **124**(8), 2539–2547.
- Chakraborty, B. and Chaudhuri, P. (1998), 'On an adaptive transformation-retransformation estimate of multivariate location', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**(1), 145–157.

REFERENCES

- Chakraborty, B., Chaudhuri, P. and Oja, H. (1998), 'Operating transformation retransformation on spatial median and angle test', *Statistica Sinica* 8, 767–784.
- Chaudhuri, P. (1996), 'On a geometric notion of quantiles for multivariate data', *Journal of the* American Statistical Association **91**(434), 862–872.
- Cheng, G., Huang, J. Z. et al. (2010), 'Bootstrap consistency for general semiparametric mestimation', *The Annals of Statistics* **38**(5), 2884–2915.
- Dudley, R. M. (2014), Uniform Central Limit Theorems, Cambridge University Press.
- Ghosal, S. and Van der Vaart, A. (2017), Fundamentals of Nonparametric Bayesian Inference, Cambridge University Press.
- Huber, P. J. (1967), The behavior of maximum likelihood estimates under nonstandard conditions, in 'Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability', Vol. 1, University of California Press, pp. 221–233.
- Lo, A. Y. (1983), 'Weak convergence for dirichlet processes', Sankhyā: The Indian Journal of Statistics, Series A 45, 105–111.
- Lo, A. Y. (1986), 'A remark on the limiting posterior distribution of the multiparameter dirichlet process', Sankhyā: The Indian Journal of Statistics, Series A pp. 247–249.
- Lopuhaa, H. P., Rousseeuw, P. J. et al. (1991), 'Breakdown points of affine equivariant estimators of multivariate location and covariance matrices', *The Annals of Statistics* **19**(1), 229– 248.

REFERENCES

- Möttönen, J., Nordhausen, K., Oja, H. et al. (2010), Asymptotic theory of the spatial median, *in* 'Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková', Institute of Mathematical Statistics, pp. 182–193.
- Small, C. G. (1990), 'A survey of multidimensional medians', International Statistical Review/Revue Internationale de Statistique 58, 263–277.
- Van der Vaart, A. (1995), 'Efficiency of infinite dimensional m-estimators', *Statistica Neerlandica* **49**(1), 9–30.

Van der Vaart, A. W. (2000), Asymptotic Statistics, Cambridge University Press.

- Van Der Vaart, A. W. and Wellner, J. A. (1996), Weak convergence and empirical processes, Springer.
- Wellner, J. A. and Zhan, Y. (1996), 'Bootstrapping z-estimators', University of Washington Department of Statistics Technical Report 308.

Xue-Kun Song, P. (2000), 'Multivariate dispersion models generated from gaussian copula', Scandinavian Journal of Statistics **27**(2), 305–320.

Department of Statistics, North Carolina State University

E-mail: ibhatta@ncsu.edu

Department of Statistics, North Carolina State University

E-mail: sghosal@ncsu.edu