Statistica Sinica Preprint No: SS-2023-0242	
Title	Statistical Inference For Ultrahigh Dimensional Location Parameter Based On Spatial Median
Manuscript ID	SS-2023-0242
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0242
Complete List of Authors	Guanghui Cheng, Liuhua Peng and Changliang Zou
Corresponding Authors	Changliang Zou
E-mails	nk.chlzou@gmail.com
Notice: Accepted author version.	

STATISTICAL INFERENCE FOR ULTRAHIGH DIMENSIONAL LOCATION PARAMETER BASED ON SPATIAL MEDIAN

Guanghui Cheng, Liuhua Peng and Changliang Zou

Guangzhou University, The University of Melbourne, Nankai University

Abstract: Motivated by the widely used geometric median-of-means estimator in machine learning, this paper studies statistical inference for ultrahigh dimensionality location parameter based on the sample spatial median under a general multivariate model, including simultaneous confidence intervals construction, global tests, and multiple testing with false discovery rate control. To achieve these goals, we derive a novel Bahadur representation of the sample spatial median with a maximum-norm bound on the remainder term, and establish Gaussian approximation for the sample spatial median over the class of hyperrectangles. In addition, a multiplier bootstrap algorithm is proposed to approximate the distribution of the sample spatial median. The approximations are valid when the dimension diverges at an exponentially rate of the sample size, which facilitates the application of the spatial median in the ultrahigh dimensional region. The proposed approaches are further illustrated by simulations and analysis of a genomic dataset from a microarray study.

Key words and phrases: Bootstrap approximation, Gaussian approximation, High-dimensional, Spatial median, FDR control

1. Introduction

Geometric median-of-means (GMOM) has been widely used for robust estimation of multivariate means, and it has been broadly adopted in machine learning (Minsker, 2015; Hsu and Sabato, 2016; Prasad et al., 2020). The idea of GMOM is to first divide the data into

disjoint subsamples and calculate the empirical means of each of the subsamples. Then the GMOM estimator is computed as the spatial median (also called geometric median) of the obtained empirical means. The previous studies on the GMOM focused on establishing its non-asymptotic error bounds under certain heavy-tailed assumptions. Its distributional properties, which are essential for statistical inference, remain unknown.

High-dimensional data with the dimension increases to infinity as the number of observations goes to infinity have been encountered in many scientific disciplines. There is a growing evidence of the multivariate normal distribution is problematic to model high-dimensional data due to the presents of heavy-tailedness and inadequate to accommodate tail dependence. For example, the distributions of the microarray expression are observed to be non-normal and have heavy tails even after log transformation in many gene expression data (Purdom and Holmes, 2005; Wang et al., 2015). As another example, elliptical distributions, in particular the multivariate t-distribution and symmetric multivariate normal inverse Gaussian distribution, provided far superior models to the multivariate normal for daily and weekly US stock-return data (McNeil et al., 2005). In such cases, the sample spatial median is favored against the sample mean for estimating the location parameter. The above discussions strongly motivate studying the spatial median under high-dimensionality, especially its distributional properties and the implementation in statistical inference for high-dimensional location parameter.

Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) p-dimensional random vectors from a population X with cumulative distribution function F_X in $\mathbb{R}^p$. In this paper, we work on a general multivariate model where X admits the following stochastic

representation:

$$X = \boldsymbol{\theta} + \nu \Gamma U \,, \tag{1.1}$$

where $\boldsymbol{\theta}$ is the location parameter, ν is a nonnegative univariate random variable and U is a p-dimensional random vector with independent components. Model (1.1) covers many commonly used multivariate models and distribution families, including the independent components model (Yao et al., 2015) and the elliptical distribution family (Fang et al., 1990). We refer to Section 2 for more detailed discussions.

Spatial median, an extension of the univariate median to multivariate distributions, was proposed for robust inference of the location parameter (Haldane, 1948; Weber, 1929). The sample spatial median $\hat{\boldsymbol{\theta}}_n \in \mathbb{R}^p$ minimizes the empirical criteria function $L_n(\boldsymbol{\beta}) = \sum_{i=1}^n (\|X_i - \boldsymbol{\beta}\| - \|X_i\|)$, where $\|\cdot\|$ is the Euclidean norm. Equivalently,

$$\hat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} L_n(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^n (\|X_i - \boldsymbol{\beta}\| - \|X_i\|) \,. \tag{1.2}$$

The function $L_n(\boldsymbol{\beta})$ is convex, and $\hat{\boldsymbol{\theta}}_n$ is unique if $\{X_i\}_{i=1}^n$ are not concentrated on a line in $\mathbb{R}^p$ when $p > 2$ (Milasevic and Ducharme, 1987). When p is fixed, the spatial median has been well studied in the literature. We refer to Chapter 6.2 of Oja (2010) for a nice review.

In the high-dimensional setting, where the dimension p diverges to infinity as $n \to \infty$, there are several existing works that study the asymptotic properties of the sample spatial median. Zou et al. (2014) offered an expansion of $\hat{\boldsymbol{\theta}}_n$ under elliptical distributions with identical shape matrix, and Cheng et al. (2019) extended the result to a general shape matrix. As a recent work, Li and Xu (2022) improved the expansion in Cheng et al. (2019)

with a smaller order remainder term under stronger conditions, and established a central limit theorem for the squared Euclidean distance $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\|^2$. In Zou et al. (2014) and Cheng et al. (2019), they both require that $p = O(n^2)$. In addition, it is required in Li and Xu (2022) that p diverges at the same rate as n. However, in modern areas such as genomics and proteomics, the dimension of the data may grow exponentially with the sample size, which lies in the "ultrahigh dimensional" region (Fan and Lv, 2008). The previous works with restrictions on the polynomial dimensionality limit the usage of the spatial median under ultrahigh-dimensionality. Moreover, the previous results are all under elliptical distributions. Thus, it is of great importance to establish asymptotic properties of the spatial median and investigate its applications under ultrahigh dimensionality and beyond elliptical distributions.

In this paper, we first establish Gaussian and bootstrap approximations hit hyperrectangles for the sample spatial median under the general model (1.1) beyond elliptical distributions, which are valid when the dimension diverges exponentially with the sample size. They serve as the theoretical foundations of statistical inference for the location parameter based on the sample spatial median under ultrahigh dimensionality. Consistent simultaneous confidence intervals (SCIs) and global tests for the location parameters are established. We also study multiple testing for every component of $\boldsymbol{\theta}$ based on $\hat{\boldsymbol{\theta}}_n$. Motivated by simultaneous inference of $\boldsymbol{\theta}$, we define a high-dimensional asymptotic relative efficiency of the sample spatial median relative to the sample mean. Most importantly, our theoretical results guarantee the validity of the proposed inferential methods for exponentially divergent p. The advantages of our proposed approaches have been justified by simulations and a real data analysis.

The main contributions of this paper are summarized as follow. Firstly, we establish SCIs for the location parameter $\boldsymbol{\theta}$ based on the sample spatial median $\hat{\boldsymbol{\theta}}_n$, which is new in the

literature. The consistency of bootstrap approximation guarantees that the probability that the SCIs cover all components of the location parameter approaches the nominal confidence level under ultrahigh dimensionality. We also propose a novel test for ultrahigh dimensional location parameter based on the maximum-norm of the sample spatial median. The proposed test not only maintains nominal significance level asymptotically for exponentially divergent p, but also is more powerful under sparse alternatives compared to those based on L_2-norms (Li and Xu, 2022; Wang et al., 2015). As another major inference, we study multiple testing for every component of the location parameter, and the false discovery rate (FDR) can be well controlled combined with the Benjamini-Hochberg procedure based on the sample spatial median, which extends the existing methods based on the sample mean (Liu and Shao, 2014). In all inferential methods, the procedures based on the sample spatial median advances those based on the sample mean for heavy-tailed distributions.

Secondly, this paper serves as the first work that provides Gaussian and bootstrap approximations for the sample spatial median under ultrahigh dimensionality. Gaussian and bootstrap approximations for high-dimensional sample mean have received extensive attraction in the last decade. Chernozhukov et al. (2013) and Chernozhukov et al. (2017) established Gaussian and bootstrap approximations for the maxima of a sum of centered independent random vectors under Kolmogorov distance and on hyperrectangles, respectively. See also Chen (2018), Chernozhukov et al. (2019) and Chernozhukov et al. (2020) for related works. Compared to the sample mean, which has a simple linear form, the theoretical difficulty for the sample spatial median lies in that it does not enjoy an explicit form. This issue is addressed by deriving a novel Bahadur representation of the sample spatial median with a maximum-norm bound on the remainder term, which extends the results of Zou et al.

(2014), Cheng et al. (2019) and Li and Xu (2022) under elliptical distributions and polynomial dimensionality. Moreover, our results can be applied to the GMOM under reasonable conditions, and thus enhance the practice usage of GMOM.

Thirdly, we propose a novel multiplier bootstrap method for the sample spatial median. Instead of multiplying on the loss function, which is generally the case for M-estimator (Imaizumi and Otsu, 2021), the multiplier is applied on the centralized X_i. Specifically, the bootstrap version of $\hat{\boldsymbol{\theta}}_n$ is defined as $\tilde{\boldsymbol{\theta}}_n = \text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d} \sum_{i=1}^n \|Z_i(X_i - \hat{\boldsymbol{\theta}}_n) - \boldsymbol{\beta}\|$, where $Z_1, \ldots, Z_n$ are the multipliers. The multiplier bootstrap is consistent under ultrahigh dimensionality thanks to this novel formulation. This is, however, different from the multiplier bootstrap method for the sample mean, which again has an explicit form (Chernozhukov et al., 2013, 2017).

The rest of the paper is organized as follows. Section 2 introduces model and assumptions. Section 3 establishes Gaussian and bootstrap approximations to the distribution of the sample spatial median. Statistical inference for the location parameter based on the sample spatial median is presented in Section 4. Section 5 reports numerical results including simulations and a real data analysis. Preliminary lemmas, proofs of main results, and additional simulations are given in the supplementary material.

Notation: Denote $|x|_\infty = \max(|x_1|, \ldots, |x_d|)$ as the maximum-norm of $x = (x_1, \ldots, x_d)^\top$. Denote $a_n \lesssim b_n$ if $a_n \leq Cb_n$ for a positive constant C, and $a_n \asymp b_n$ means $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For $\alpha > 0$, let $\psi_\alpha(x) = \exp(x^\alpha) - 1$ be a function defined on $[0, \infty)$. Then the Orlicz norm $\|\cdot\|_{\psi_\alpha}$ of a random variable X is defined as $\|X\|_{\psi_\alpha} = \inf\{t > 0, \mathbb{E}\{\psi_\alpha(|X|/t)\} \leq 1\}$. We use $\text{tr}(\cdot)$ to denote the trace operator for square matrices. Moreover, we denote I_p as the $p \times p$ identity matrix. For $a, b \in \mathbb{R}$, we write $a \wedge b = \min(a, b)$.

2. Model and assumptions

In this paper, we consider a general multivariate model for the distribution F_X such that X_i admits the following stochastic representation:

$$X_i = \boldsymbol{\theta} + \nu_i \Gamma U_i \,, \tag{2.1}$$

where $\boldsymbol{\theta}$ is the location parameter, Γ is a nonrandom and invertible $p \times p$ matrix, U_i is a p-dimensional random vector with independent standardized components, and ν_i is a nonnegative univariate random variable independent with the spatial sign of U_i. The distribution of X_i depends on Γ through the shape matrix $\Omega = \Gamma\Gamma^\top$.

Remark 1. Model (2.1) covers many commonly used multivariate models and distribution families. First, the independent components model (Yao et al., 2015) follows (2.1) with ν_i being a nonnegative constant. Second, model (2.1) also includes elliptical distributions by choosing $U_i \sim N(0, I_p)$ and $\nu_i = \xi_i / \|U_i\|$ for some nonnegative random variable ξ_i independent of U_i. In this case, ν_i is independent of the spatial sign of U_i, but not U_i. The independent components model has received great extension in high-dimensional data analysis as well as signal processing and machine learning (Oja, 2010). In addition, the elliptical distribution family covers many non-Gaussian distributions such as multivariate t-distribution, multivariate logistic distribution, and so on. It is commonly adopted in the literature on studying the sample spatial median (Cheng et al., 2019; Li and Xu, 2022; Zou et al., 2014). In terms of the GMOM, if the data are from the independent components model, the subsample means satisfy model (2.1) clearly. In addition, some subfamilies of elliptical distributions are closed under convolution, and thus the subsample means also follow

model (2.1). Our results can be applied to the GMOM estimator directly in those cases.

For $i = 1, \ldots, n$, denote

$$W_i = S(X_i - \boldsymbol{\theta}) \quad \text{and} \quad R_i = \|X_i - \boldsymbol{\theta}\| \tag{2.2}$$

as the spatial-sign and radius of $X_i - \boldsymbol{\theta}$, where $S(X) = \|X\|^{-1} X \mathbb{I}(X \neq 0)$ is the multivariate sign function with $\mathbb{I}(\cdot)$ being the indicator function. Thus, $\hat{\boldsymbol{\theta}}_n$ satisfies $\sum_{i=1}^n S(X_i - \hat{\boldsymbol{\theta}}_n) = 0$.

Denote $U_i = (U_{i,1}, \ldots, U_{i,p})^\top$, we impose the following three conditions.

Condition C1. $U_{i,1}, \ldots, U_{i,p}$ are i.i.d. symmetric random variables with $\mathbb{E}(U_{i,j}) = 0$, $\mathbb{E}(U_{i,j}^2) = 1$, and $\|U_{i,j}\|_{\psi_\alpha} \leq c_0$ with some constant $c_0 > 0$ and $1 \leq \alpha \leq 2$.

Condition C2. The moments $\zeta_k = \mathbb{E}(R_i^{-k})$ for $k = 1, 2, 3, 4$ exist for large enough p. In addition, there exist two positive constants $\underline{b}$ and $\bar{B}$ such that $\underline{b} \leq \limsup_p \mathbb{E}(R_i/\sqrt{p})^{-k} \leq \bar{B}$ for $k = 1, 2, 3, 4$.

Condition C3. The shape matrix $\Omega = (\omega_{j\ell})_{p\times p}$ satisfies $\mathrm{tr}(\Omega) = p$ and it belongs to the following class:

$$\mathcal{U}(a_0(p), \underline{m}, \bar{M}) = \left\{ \Omega : \underline{m} \leq \omega_{jj} \leq \bar{M}, \sum_{\ell=1}^{p} |\omega_{j\ell}| \leq a_0(p), \text{ for all } j = 1, \ldots, p \right\},$$

where $\underline{m} \leq \bar{M}$ are bounded positive constants.

Remark 2. In Condition C1, the symmetric assumption is to ensure that $\boldsymbol{\theta}$ in model (2.1) coincides with the population spatial median, which minimizes $L(\boldsymbol{\beta}) = \mathbb{E}(\|X - \boldsymbol{\beta}\| - \|X\|)$. It is obvious that Condition C1 is satisfied by elliptical distributions with $U_i \sim N(0, I_p)$. The condition $\|U_{i,j}\|_{\psi_\alpha} \leq c_0$ implies that $U_{i,j}$ has a sub-exponential distribution. It is worth

highlighting that with slight modification of the proofs of main theorems, the i.i.d. condition on $U_{i,1}, \ldots, U_{i,p}$ can be weaken by replacing Condition C1 with the following assumption: $U_{i,1}, \ldots, U_{i,p}$ are independent symmetric random variables with $\mathbb{E}(U_{i,j}) = 0$, $\mathbb{E}(U_{i,j}^2) = 1$ for all $j = 1, \ldots, p$, and $\sup_{1 \le j \le p} \|U_{i,j}\|_{\psi_\alpha} \le c_0$ with some constant $c_0 > 0$ and $1 \le \alpha \le 2$.

Remark 3. The condition $\underline{b} \le \limsup_p \mathbb{E}(R_i/\sqrt{p})^{-k} \le \bar{B}$ indicates that $\zeta_k \asymp p^{-k/2}$ for $k = 1, 2, 3, 4$. It is introduced to avoid X_i from concentrating too much near $\boldsymbol{\theta}$. For elliptical distributions, it is a generalization of Assumption 1 of Zou et al. (2014), which is satisfied by many common distributions. For the independent components model, Condition C2 is equivalent to that $\underline{b} \le \limsup_p \mathbb{E}(\|\Gamma U_i\|/\sqrt{p})^{-k} \le \bar{B}$. According to Lemma A2 in Appendix A, $\mathbb{E}(\|\Gamma U_i\|^k) = p^{k/2}\{1 + o(1)\}$ for $k = 1, 2, 3, 4$. Then the Cauchy-Schwarz inequality implies that $\mathbb{E}(\|\Gamma U_i\|^{-k}) \ge \{\mathbb{E}(\|\Gamma U_i\|^k)\}^{-1} = p^{-k/2}\{1 + o(1)\}$, from which $\mathbb{E}(\|\Gamma U_i\|^{-k}) \gtrsim p^{-k/2}$. Furthermore, denote Γ_j as the jth row of Γ, by the inequality of harmonic and quadratic means,

$$p^2 \|\Gamma U_i\|^{-4} = \left\{ \frac{p}{(\Gamma_1 U_i)^2 + \cdots + (\Gamma_p U_i)^2} \right\} \le \frac{(\Gamma_1 U_i)^{-4} + \cdots + (\Gamma_p U_i)^{-4}}{p}.$$

It follows that $\mathbb{E}(\|\Gamma U_i\|^{-4}) \lesssim p^{-2}$ if $\mathbb{E}\{(\Gamma_1 U_i)^{-4}\}, \ldots, \mathbb{E}\{(\Gamma_p U_i)^{-4}\}$ are uniformly bounded, and from which $\mathbb{E}(\|\Gamma U_i\|^{-k}) \lesssim p^{-k/2}$ by Jensen's inequality. Thus, Condition C2 is satisfied by the independent components models as long as $\Gamma_1 U_i, \ldots, \Gamma_p U_i$ are not concentrating too much near 0. See also discussions in Cardot et al. (2013) on similar conditions.

Remark 4. It is noticed that the shape matrix Ω is only well defined up to a scalar multiple, the condition $\mathrm{tr}(\Omega) = p$ is used to regularize Ω to make model (2.1) identifiable. The class $\mathcal{U}(a_0(p), \underline{m}, \bar{M})$ covers a wide range of symmetric square matrices, and it is commonly

adopted in the literature on high-dimensional analysis. For example, a similar matrix class is introduced in Bickel and Levina (2008). The condition $\underline{m} \leq \omega_{jj} \leq \bar{M}$ requires bounded diagonal elements. The order of $a_0(p)$, which will be specified later, controls the orders of the off-diagonal elements of Ω.

3. Gaussian and bootstrap approximations

3.1 Bahadur representation and Gaussian approximation

In this section, we establish Gaussian approximation for $\hat{\boldsymbol{\theta}}_n$, which is valid when p diverges exponentially over n. The following lemma offers a Bahadur representation of $\hat{\boldsymbol{\theta}}_n$, and it severs as the foundation of the Gaussian approximation result in Theorem 1.

Lemma 1. *(Bahadur representation) Assume Conditions C1, C2 and C3 with* $a_0(p) \asymp p^{1-\delta}$ *for some positive constant* $\delta \leq 1/2$ *hold. If* $\log p = o(n^{1/3})$ *and* $\log n = o(p^{1/3 \wedge \delta})$*, then*

$$n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) = n^{-1/2}\zeta_1^{-1} \sum_{i=1}^{n} W_i + C_n \,,$$

where $|C_n|_\infty = O_p\{n^{-1/4} \log^{1/2}(np) + p^{-(1/6 \wedge \delta/2)} \log^{1/2}(np)\}$.

Remark 5. To the best of our knowledge, Lemma 1 firstly offers the Bahadur representation of the sample spatial median with a maximum-norm bound on the remainder term. In Zou et al. (2014) and Cheng et al. (2019), the same expansion with the remainder term C_n satisfies $\|C_n\| = o_p(\zeta_1^{-1})$ was obtained, and their result was improved to $\|C_n\| = o_p(1)$ in Li and Xu (2022), by replacing ζ_1 with $n^{-1}\sum_{i=1}^{n} R_i^{-1}$ in the linear term, but under a more restricted condition that p and n are of the same order. It is worth noticing that the previous results (Li and Xu, 2022; Zou et al., 2014) are derived under elliptical distributions with a bounded

spectral norm for the shape matrix. In this paper, we first extend the model to (2.1), which covers the elliptical distribution as a special case. Second, we weaken the condition on the shape matrix by controlling the sparsity of its L_1 norm. More specifically, when the shape matrix Ω has a bounded spectral norm, it follows immediately that $\|\Omega\|_1 = O(\sqrt{p})$ with the fact $\|\Omega\|_1 \leq \sqrt{p}\|\Omega\|_2$, and thus Condition C3 is satisfied with $a_0(p) \asymp \sqrt{p}$.

Let $\mathcal{A}^{\mathrm{re}} = \{\prod_{j=1}^{p}[a_j, b_j] : -\infty \leq a_j \leq b_j \leq \infty, j = 1, \ldots, p\}$ be the class of rectangles in $\mathbb{R}^p$. With Lemma 1 on hand, we establish the following Gaussian approximation result for $\hat{\boldsymbol{\theta}}_n$ over hyperrectangles.

Theorem 1. *(Gaussian approximation) Assume Conditions C1, C2 and C3 with* $a_0(p) \asymp p^{1-\delta}$ *for some positive constant* $\delta \leq 1/2$ *hold. If* $\log p = o(n^{1/5})$ *and* $\log n = o(p^{1/3\wedge\delta})$*, then*

$$\rho_n(\mathcal{A}^{\mathrm{re}}) = \sup_{A\in\mathcal{A}^{\mathrm{re}}} \left|\mathbb{P}\{n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \in A\} - \mathbb{P}\left(G \in A\right)\right| \to 0$$

as $n \to \infty$*, where* $G \sim N(0, \zeta_1^{-2}\mathbb{B})$ *with* $\mathbb{B} = \mathbb{E}(W_1 W_1^{\top})$.

Remark 6. The Gaussian approximation for $\hat{\boldsymbol{\theta}}_n$ indicates that the probabilities $\mathbb{P}\{n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \in A\}$ can be approximated by that of a centered Gaussian random vector with covariance matrix $\zeta_1^{-2}\mathbb{B}$ for hyperrectangles $A \in \mathcal{A}^{\mathrm{re}}$. The condition of $\log p = o(n^{1/5})$, commonly adopted in Gaussian approximation for independent partial sums (Chernozhukov et al., 2013, 2017), allows for the dimension p divergence exponentially with the sample size n and thus is in line with the ultrahigh dimensional case. The condition of $\log(n) = o(p^{1/3\wedge\delta})$ restricts n to be diverging too fast compared to p, which is still in line with the high-dimensional scenario. It is worth mentioning that this condition can be satisfied for a broad range of n and p, especially under the high-dimensional case. For example, $\log(n) = o(p^{1/3\wedge\delta})$ is true

when $p/n \to \gamma$ for some constant $\gamma > 0$ (assumed in Li and Xu (2022)) or $p/n \to \infty$. Thus, the Gaussian approximation result in Theorem 1 requires much weaker restrictions on the rates of n and p compared to the asymptotic normality of $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\|^2$.

Let $\mathbb{B}_{j\ell}$ be the (j,ℓ)th element of $\mathbb{B}$. According to Lemma A4 (iii) in Appendix A, $\zeta_1^{-2}\mathbb{B}_{j\ell} = \zeta_1^{-2}p^{-1}\omega_{j,\ell} + O(p^{-\delta/2})$ for all $1 \leq j, \ell \leq p$. Thus, the covariance matrix of G in Theorem 1 is asymptotically proportional to the shape matrix Ω.

Remark 7. As the sample spatial median is a special M-estimator, Gaussian approximation for M-estimator in Imaizumi and Otsu (2021) is potentially applicable to the spatial median under high-dimensionality. However, it is worth highlighting that the results in Imaizumi and Otsu (2021) cannot be applied to our framework. To be precise, Assumption 1 (A3) in Imaizumi and Otsu (2021) assumes that there exist constants $C > 0$ and $\alpha \in (0, 2)$ such that $\log \mathcal{N}(\varepsilon, \Theta, \|\cdot\|) \leq C\varepsilon^{-\alpha}$ holds for all $\varepsilon \in (0, 1)$, where Θ is the parameter space, and $\mathcal{N}(\varepsilon, \Theta, \|\cdot\|)$ is the ε-covering number of Θ under the Euclidean norm $\|\cdot\|$ (van der Vaart and Wellner, 1996). When Θ is a compact subset of $\mathbb{R}^p$, $\mathcal{N}(\varepsilon, \Theta, \|\cdot\|)$ is of order $O(\varepsilon^{-p})$. In this case, $\log \mathcal{N}(\varepsilon, \Theta, \|\cdot\|) \leq C\varepsilon^{-\alpha}$ cannot be satisfied when $p \to \infty$. Thus, our theoretical findings are independent of those in Imaizumi and Otsu (2021).

Theorem 1 immediately implies the following corollary since the Kolmogorov distance of sup-norm is a subset of $\mathcal{A}^{\mathrm{re}}$ corresponding to max-hyperrectangles in $\mathbb{R}^p$.

Corollary 1. *Under the conditions assumed in Theorem 1, as* $n \to \infty$,

$$\rho_n = \sup_{t \in \mathbb{R}} \left| \mathbb{P}(n^{1/2}|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}|_\infty \leq t) - \mathbb{P}(|G|_\infty \leq t) \right| \to 0.$$

3.2 Multiplier bootstrap approximation

Theorem 1 allows us to approximate the distribution of $n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ by that of G hit hyperrectangles, where $G \sim N(0, \zeta_1^{-2}\mathbb{B})$. However, it cannot be used directly in statistical inference for $\boldsymbol{\theta}$ as the quantity ζ_1 and the matrix $\mathbb{B}$ depend on the underlying distribution F_X and are thus unknown. Estimating the matrix $\mathbb{B}$ is challenging under high-dimensionality and the resampling with replacement approach often exhibits poor finite-sample performance and lacks theoretical justification for the spatial median. Furthermore, the multiplier bootstrap method in Chernozhukov et al. (2019, 2020), developed for the sample mean, cannot be easily adapted to the spatial median due to the latter does not enjoy a simple average form. To solve this issue, we propose an easy-to-implement bootstrap method to approximate the distribution of $n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$.

Recall that the sample spatial median of $X_1, \ldots, X_n$ is $\hat{\boldsymbol{\theta}}_n = \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} L_n(\boldsymbol{\beta}) = \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^n (\|X_i - \boldsymbol{\beta}\| - \|X_i\|)$. Now, consider the sequence $Z_1(X_1 - \hat{\boldsymbol{\theta}}_n), \ldots, Z_n(X_n - \hat{\boldsymbol{\theta}}_n)$, where $Z_1, \ldots, Z_n$ are i.i.d. random variables independent of $X_1, \ldots, X_n$. As $\mathbb{E}\left\{\frac{Z_1(X_1-\hat{\boldsymbol{\theta}}_n)}{\|Z_1(X_1-\hat{\boldsymbol{\theta}}_n)\|}\right\} = \mathbb{E}\left(\frac{Z_1}{|Z_1|}\right)\mathbb{E}\left(\frac{X_1-\hat{\boldsymbol{\theta}}_n}{\|X_1-\hat{\boldsymbol{\theta}}_n\|}\right)$, the population spatial median of $Z_1(X_1 - \hat{\boldsymbol{\theta}}_n), \ldots, Z_n(X_n - \hat{\boldsymbol{\theta}}_n)$ is 0 if $\mathbb{E}(Z_1/|Z_1|) = 0$, which holds for any symmetric random variable. Thus, we propose the bootstrap version of the sample spatial median as

$$\tilde{\boldsymbol{\theta}}_n = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^n \|Z_i(X_i - \hat{\boldsymbol{\theta}}_n) - \boldsymbol{\beta}\|. \tag{3.1}$$

Then, the distribution of $n^{1/2}\tilde{\boldsymbol{\theta}}_n$ conditional on $X_1, \ldots, X_n$ is used to approximate that of $n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$. This algorithm is called the multiplier bootstrap, and $Z_1, \ldots, Z_n$ are the multiplier weights.

To ensure that the asymptotic distribution of $n^{1/2}\tilde{\boldsymbol{\theta}}_n$ mimics that of $n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$, the multiplier weights $Z_1, \ldots, Z_n$ should be chosen such that the gradient and Hessian matrix of the loss function $L_n^{(B)}(\boldsymbol{\beta}) = \sum_{i=1}^n \|Z_i(X_i - \hat{\boldsymbol{\theta}}_n) - \boldsymbol{\beta}\|$ at $\boldsymbol{\beta} = 0$, given by $\sum_{i=1}^n \frac{Z_i(X_i - \hat{\boldsymbol{\theta}}_n)}{\|Z_i(X_i - \hat{\boldsymbol{\theta}}_n)\|}$ and $\sum_{i=1}^n \frac{1}{\|Z_i(X_i - \hat{\boldsymbol{\theta}}_n)\|}\left[I_p - \frac{\{Z_i(X_i - \hat{\boldsymbol{\theta}}_n)\}\{Z_i(X_i - \hat{\boldsymbol{\theta}}_n)\}^\top}{\|Z_i(X_i - \hat{\boldsymbol{\theta}}_n)\|^2}\right]$, approximate those of $L_n(\boldsymbol{\beta}) = \sum_{i=1}^n(\|X_i - \boldsymbol{\beta}\| - \|X_i\|)$ at $\boldsymbol{\beta} = \boldsymbol{\theta}$, which are $\sum_{i=1}^n \frac{X_i - \boldsymbol{\theta}}{\|X_i - \boldsymbol{\theta}\|}$ and $\sum_{i=1}^n \frac{1}{\|X_i - \boldsymbol{\theta}\|}\left\{I_p - \frac{(X_i - \boldsymbol{\theta})(X_i - \boldsymbol{\theta})^\top}{\|X_i - \boldsymbol{\theta}\|^2}\right\}$. In view of this, Rademacher variables (Chernozhukov et al., 2019), with $\mathbb{P}(Z_i = 1) = \mathbb{P}(Z_i = -1) = 1/2$ for $i = 1, \ldots, n$, offers a natural choice for the multiplier bootstrap method, ensuring that the bootstrap version of the sample spatial median $\tilde{\boldsymbol{\theta}}_n$ has the same asymptotic distribution as $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}$. Additionally, it is fast and easy to implement. Although other multiplier weights might be applicable, investigating alternatives is beyond the scope of this paper.

The next theorem shows the validity of the multiplier bootstrap.

Theorem 2. *(Bootstrap approximation) Under the conditions assumed in Theorem 1,*

$$\rho_n^{\mathrm{MB}}(\mathcal{A}^{\mathrm{re}}) = \sup_{A \in \mathcal{A}^{\mathrm{re}}} \left|\mathbb{P}\{n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \in A\} - \mathbb{P}^*(n^{1/2}\tilde{\boldsymbol{\theta}}_n \in A)\right| \to 0$$

in probability as $n \to \infty$, where $\mathbb{P}^*$ denotes the conditional probability given $X_1, \ldots, X_n$.

Under the same conditions on the divergence rates of n and p as in Theorem 1, Theorem 2 validates that conditional on $X_1, \ldots, X_n$, the distribution of the bootstrap sample spatial median $\tilde{\boldsymbol{\theta}}_n$ approximates that of $\hat{\boldsymbol{\theta}}_n$ consistently over hyperrectangles.

Remark 8. The proof of Theorem 2 is nontrivial and does not follow directly from existing results since $\tilde{\boldsymbol{\theta}}_n$ has no explicit form, which is different from the multiplier bootstrap methods for high-dimensional sample mean that have been analysed in the literature. The key step in the proof is to obtain a Bahadur representation of $\tilde{\boldsymbol{\theta}}_n$ similar

as $\hat{\boldsymbol{\theta}}_n$ in Lemma 1. Specifically, we show that $n^{1/2}\tilde{\boldsymbol{\theta}}_n = n^{-1/2}\zeta_1^{-1}\sum_{i=1}^n Z_i W_i + \tilde{C}_n$ with $|\tilde{C}_n|_\infty = O_p\{n^{-1/4}\log^{1/2}(np) + p^{-(1/6\wedge\delta/2)}\log^{1/2}(np)\}$ in Lemma A5 in Appendix A.

The next corollary is an immediate consequence of Theorem 2.

Corollary 2. *Under the conditions assumed in Theorem 2, as* $n \to \infty$,

$$\rho_n^{\mathrm{MB}} = \sup_{t\in\mathbb{R}} \left| \mathbb{P}\{n^{1/2}|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}|_\infty \le t\} - \mathbb{P}^*(n^{1/2}|\tilde{\boldsymbol{\theta}}_n|_\infty \le t) \right| \to 0 \quad \textit{in probability.}$$

4. Statistical inference

The Gaussian and multiplier bootstrap approximations for the sample spatial median enable many statistical inferential methods for ultrahigh dimensional population location parameter. In this section, we present the following statistical inferences: simultaneous confidence intervals (SCIs) and global tests for the population location parameter, multiple testing for every component of $\boldsymbol{\theta}$, and high-dimensional relative efficiency of the sample spatial median compared to the sample mean.

4.1 Simultaneous confidence intervals

We are interested in building SCIs for all components of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$. Corollary 2 motivates the following way of constructing SCIs for $\boldsymbol{\theta}$. Given a nominal confidence level $1-\tau$, define the set $\mathcal{C}_\tau$ as

$$\mathcal{C}_\tau = \left\{ \boldsymbol{\theta} \in \mathbb{R}^p, n^{1/2}|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}|_\infty < q_{1-\tau}^B \right\},$$

where $q^B_{1-\tau}$ is the $(1-\tau)$th quantile of $n^{1/2}|\tilde{\boldsymbol{\theta}}_n|_\infty$ given $X_1, \ldots, X_n$. Denote $\hat{\boldsymbol{\theta}}_n = (\hat{\theta}_{n,1}, \ldots, \hat{\theta}_{n,p})^\top$, the confidence intervals are $[\theta^-_{n,j}, \theta^+_{n,j}]$ for $j = 1, \ldots, p$, where

$$\theta^-_{n,j} = \hat{\theta}_{n,j} - n^{-1/2} q^B_{1-\tau} \quad \text{and} \quad \theta^+_{n,j} = \hat{\theta}_{n,j} + n^{-1/2} q^B_{1-\tau}.$$

The next theorem shows that $\mathcal{C}_\tau$ preserves the nominal simultaneous confidence level $1-\tau$ asymptotically under ultrahigh dimensionality.

Theorem 3. *Suppose the conditions of Theorem 2 hold, then* $\mathbb{P}(\boldsymbol{\theta} \in \mathcal{C}_\tau) \to 1-\tau$ *as* $n \to \infty$. *Equivalently,* $\mathbb{P}(\theta_j \in [\theta^-_{n,j}, \theta^+_{n,j}] \text{ for all } 1 \le j \le p) \to 1-\tau$ *as* $n \to \infty$.

Remark 9. Unlike the fixed dimensional setting, $n^{1/2}|\tilde{\boldsymbol{\theta}}_n|_\infty$ and $n^{1/2}|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}|_\infty$ are maxima of divergent numbers of variables, and their quantiles are generally divergent as $p \to \infty$. Thus, Theorem 3 is not a direct consequence of Corollary 2. To ascertain the consistency of $\mathcal{C}_\tau$ theoretically, we show that, with probability approaching one, $q^B_{1-\tau}$ is bounded by two quantiles of $n^{1/2}|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}|_\infty$ with quantile levels close enough to $1-\tau$ using an anti-concentration inequality for divergent random sequences.

Remark 10. The Gaussian approximation for the sample mean $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ (Chernozhukov et al., 2013, 2017, 2019) indicate that if $\log p = o(n^{1/5})$,

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}(n^{1/2}|\bar{X}_n - \boldsymbol{\theta}|_\infty \le t) - \mathbb{P}(|G_0|_\infty \le t) \right| \to 0 \tag{4.1}$$

as $n \to \infty$ under some moderate conditions, where $G_0 \sim N(0, \Sigma)$ with $\Sigma = \mathbb{E}(XX^\top)$. Define $X^*_i = Z_i(X_i - \bar{X}_n)$ for $i = 1, \ldots, n$, where $Z_1, \ldots, Z_n$ are the Rademacher weights. Denote

$\bar{X}_n^* = n^{-1}\sum_{i=1}^n X_i^*$, it has been shown in Chernozhukov et al. (2019) that

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}(n^{1/2}|\bar{X}_n - \boldsymbol{\theta}|_\infty \le t) - \mathbb{P}^*(n^{1/2}|\bar{X}_n^*|_\infty \le t)\right| \to 0 \tag{4.2}$$

in probability as $n \to \infty$ when $\log p = o(n^{1/5})$. Based on (4.2), define

$$\mathcal{C}_\tau' = \left\{\boldsymbol{\theta} \in \mathbb{R}^p, n^{1/2}|\bar{X}_n - \boldsymbol{\theta}|_\infty < q_{1-\tau}^{B\prime}\right\},$$

where $q_{1-\tau}^{B\prime}$ is the $(1-\tau)$th quantile of $n^{1/2}|\bar{X}_n^*|_\infty$ conditional on $X_1, \ldots, X_n$. Then $\mathcal{C}_\tau'$ is also an asymptotic $1-\tau$ SCIs for $\boldsymbol{\theta}$. Based on the discussion in Section 4.4, $\mathcal{C}_\tau$ has advantage (relative shorter intervals) over $\mathcal{C}_\tau'$ under heavy-tailed distributions. We refer to Section 5.1 for finite-sample justifications on this.

4.2 Global tests for high-dimensional location parameters

In this section, we propose a novel approach for global tests on high-dimensional location parameters. Let $\boldsymbol{\theta}_0$ be a known p-dimensional vector, we are interested in testing

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0. \tag{4.3}$$

Theorems 1 and 2 motivate us proposing a maximum-norm type test statistic. Define

$$T_n = n^{1/2}|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0|_\infty \tag{4.4}$$

as the test statistic, and H_0 is rejected when T_n is larger than a critical value. We can use the multiplier bootstrap to approximate the distribution of T_n under H_0. Specifically,

with a nominal significance level τ, the null hypothesis is rejected if $T_n > q^B_{1-\tau}$. Theorem 3 guarantees that the test based on T_n maintains nominal significance level asymptotically under ultrahigh dimensionality, that is, $\mathbb{P}(T_n > q^B_{1-\tau} \mid H_0) \to \tau$ as $n \to \infty$ when $\log p = o(n^{1/5})$.

Remark 11. An alternative test for (4.3) can be constructed based on $\bar{X}_n$ by defining the test statistic as $T_{\text{Mean}} = n^{1/2}|\bar{X}_n - \boldsymbol{\theta}_0|_\infty$. Then, the null hypothesis is rejected if $T_{\text{Mean}} > q^{B\prime}_{1-\tau}$. The test based on T_n can be deemed as a nonparametric extension of the test based on T_{Mean} . As $\hat{\boldsymbol{\theta}}_n$ is more efficient than $\bar{X}_n$ for simultaneous inference of $\boldsymbol{\theta}$ under heavy-tailed distributions as discussed in Section 4.4, we expect that the proposed test based on T_n is more powerful than that based on T_{Mean} in those cases. This has been reflected by the simulation results in Appendix D of the supplementary material.

The next theorem summarises the asymptotic power of the proposed test based on T_n.

Theorem 4. *Suppose the conditions of Theorem 2 hold, where* $\log p = o(n^{1/5})$ *and* $\log n = o(p^{1/3\wedge\delta})$. *For any given* $0 < \tau < 1$, *if* $|\boldsymbol{\theta} - \boldsymbol{\theta}_0|_\infty \geq Cn^{-1/2}\log^{1/2}(np)$ *for some sufficient large constant* $C > 0$, *then* $\mathbb{P}(T_n > q^B_{1-\tau} \mid H_1) \to 1$ *as* $n, p \to \infty$.

Theorem 4 indicates that the test based on T_n possesses non-trivial power when the order of $|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0|_\infty$ is at least $n^{-1/2}\log^{1/2}(np)$ for a fixed significant level τ. The constant C is introduced to ensure that the signal term $|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0|_\infty$ is large enough to achieve the non-trivial power. According to the proof of Theorem 4 in the supplementary materials, the test is consistent as long as the universal constant C is larger than $4\bar{B}\bar{M}$, where $\bar{B}$ and $\bar{M}$ are defined in Conditions C2 and C3, respectively.

Remark 12. Wang et al. (2015) proposed a L_2-norm type test (WPL test) for (4.3) with $\boldsymbol{\theta}_0 = 0$ based on $T_{\text{WPL}} = \sum_{i=1}^n \sum_{j=1}^{i-1} W_i^\top W_i$. It has been argued in Wang et al. (2015) and

Li and Xu (2022) that the signal of the WPL test is determined by the magnitude of $\|\boldsymbol{\theta}\|$, the L_2-norm of $\boldsymbol{\theta}$. As a contrast, the power of the test based on T_n depends on $|\boldsymbol{\theta}|_\infty$. Thus, the proposed test based on T_n is expected to be more powerful under sparse alternatives, when $\boldsymbol{\theta}$ contains only a limited number of non-zero components and its maximum element has certain order of magnitude. In such cases, $\|\boldsymbol{\theta}\|$ is not big enough for the rejection of the WPL test. See Appendix D in the supplementary material and Section 5.3 for numerical justifications.

4.3 Multiple testing with FDR control in large-scale tests

Multiple testing with false discovery rate (FDR) control has been applied to many real problems, such as detecting differentially expressed genes in genomic study. In this section, we study multiple testing for every component of $\boldsymbol{\theta}$ based on the spatial median with the Benjamini and Hochberg (B-H) method for FDR control. For $j = 1, \ldots, p$, we are interested in testing

$$H_{0j} : \theta_j = \theta_{0,j} \quad \text{versus} \quad H_{1j} : \theta_j \neq \theta_{0,j}$$

simultaneously, where $\theta_{0,1}, \ldots, \theta_{0,p}$ are given values. Define the test statistics as

$$T_{n,j} = n^{1/2}(\hat{\theta}_{n,j} - \theta_{0,j})/s_{n,j}$$

for $j = 1, \ldots, p$, where $s_{n,j}^2 = \hat{\zeta}_1^{-2}\hat{\mathbb{B}}_{jj}$ with $\hat{\zeta}_1 = n^{-1}\sum_{i=1}^n \|X_i - \hat{\boldsymbol{\theta}}_n\|^{-1}$, and $\hat{\mathbb{B}}_{jj}$ is the jth diagonal element of $\hat{\mathbb{B}} = n^{-1}\sum_{i=1}^n \|X_i - \hat{\boldsymbol{\theta}}_n\|^{-2}(X_i - \hat{\boldsymbol{\theta}}_n)(X_i - \hat{\boldsymbol{\theta}}_n)^\top$.

According to the proof of Theorem 5 in Appendix B of the supplementary materials, $T_{n,j}$

converges in distribution to a standard normal as $n, p \to \infty$, provided that $\log p = o(n^{1/5})$ and $\log n = o(p^{1/3 \wedge \delta})$ under H_{0j} for all $j = 1, \ldots, p$ simultaneously. Specifically,

$$\max_{1 \le j \le p} \sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(n^{1/2}(\hat{\theta}_{n,j} - \theta_{0,j})/s_{n,j} \le t \right) - \Phi(t) \right| \to 0$$

as $n, p \to \infty$ with $\log p = o(n^{1/5})$ and $\log n = o(p^{1/3 \wedge \delta})$ under H_{0j}, where $\Phi(\cdot)$ denotes the cumulative distribution function (cdf) of the standard normal distribution. Thus, we utilize the standard normal distribution to estimate the marginal p-values. For $j = 1, \ldots, p$, define the p-value for H_{0j} as $P_j = 2 - 2\Phi(|T_{n,j}|)$. Denote $P_{(1)} \le \cdots \le P_{(p)}$ be the ordered p-values, and define

$$\hat{k} = \max\left\{ j = 1, \ldots, p : P_{(j)} \le \tau j / p \right\}$$

for a pre-specific significance level τ. Then, the B-H procedure rejects the null hypotheses for which $P_j \le P_{(\hat{k})}$. Denote $\mathcal{H}_R = \{j : P_j \le P_{(\hat{k})}\}$ as the set of indices j such that H_{0j} is rejected by the B-H method, and let $|\mathcal{H}_R|$ be the cardinality of $\mathcal{H}_R$ that equals the total number of rejected null hypotheses.

Let $\mathcal{H}_0 \subset \{1, \ldots, p\}$ be the set of indices j corresponding to the true null hypotheses H_{0j}. The false discovery proportion (FDP) and false discovery rate (FDR) of the B-H method are defined as

$$\mathrm{FDP}_M = \frac{|\mathcal{H}_0 \cap \mathcal{H}_R|}{|\mathcal{H}_R| \vee 1} \quad \text{and} \quad \mathrm{FDR}_M = \mathbb{E}(\mathrm{FDP}_M).$$

Regarding that $T_{n,1}, \ldots, T_{n,p}$ are dependent, we impose the following condition on the

weak dependence between any two components of W_i. Define the correlation matrix as $(r_{j\ell})_{p\times p} = \{\text{diag}(\mathbb{B})\}^{-1/2}\mathbb{B}\{\text{diag}(\mathbb{B})\}^{-1/2}$, where diag($\mathbb{B}$) is the diagonal matrix of $\mathbb{B}$.

Condition C4. Suppose $\max_{1\le j<\ell\le p}|r_{j\ell}| \le r$ for some constant $0 < r < 1$. In addition, for all $1 \le j \le p$, the cardinality of $\{\ell : |r_{j\ell}| \ge (\log p)^{-3}\}$ is at most p^η for some constant $0 < \eta < (1-r)/(1+r)$.

Condition C4, which imposes weak dependence among $T_{n,1}, \ldots, T_{n,j}$, is similarly assumed in Liu and Shao (2014) and Belloni et al. (2018). First, as the correlation between $T_{n,j}$ and $T_{n,\ell}$ tends to $r_{j\ell}$, the condition that $\max_{1\le j<\ell\le p}|r_{j\ell}| \le r$ ensures that the correlations between $T_{n,1}, \ldots, T_{n,p}$ are uniformly bounded by r. In addition, the condition that the cardinality of $\{\ell : |r_{j\ell}| \ge (\log p)^{-3}\}$ is at most p^η for all $1 \le j \le p$ requires that each statistic $T_{n,j}$ is weakly correlated or uncorrelated with at least $p - p^\eta$ other test statistics. Thus, the parameter η controls the number of highly correlated test statistics. Under the condition $\eta < (1-r)/(1+r)$, a larger r results in a smaller allowable η, meaning fewer test statistics can be highly correlated. Finally, the constrain between η and r is reflected by (S.11) in the proof of Theorem 5 in the supplementary materials, where it requires $-1 + \eta + \frac{2(r+\epsilon)}{1+r} < 0$, or equivalently, $\eta < (1-r-2\epsilon)/(1+r)$ for any $\epsilon > 0$ for the validity of FDR control.

Theorem 5. *Suppose Condition C4 and the conditions of Theorem 1 hold. In addition, there exists* $\mathcal{H} \subset \{1, \ldots, p\}$ *such that* $\mathcal{H} = \{j : \zeta_1 \mathbb{B}_{jj}^{-1/2} n^{1/2}|\theta_j - \theta_{0,j}| \ge 2\log^{1/2}(p)\}$ *and* $|\mathcal{H}| \ge \log\log p \to \infty$ *as* $p \to \infty$. *Assume that the number of false null hypotheses* $p_1 \le p^\varpi$ *for some* $0 < \varpi < 1$. *Then,* $\text{FDR}_M/(\tau p_0/p) \to 1$ *as* $n \to \infty$.

Theorem 5 shows the B-H procedure based on $P_1, \ldots, P_p$ controls the FDR asymptotically, and it extends Theorem 4.1 in Liu and Shao (2014) to spatial median-based test statistic.

4.4 High-dimensional asymptotic relative efficiency

As two candidate estimators of the location parameter $\boldsymbol{\theta}$, it is of interest to study the asymptotic relative efficiency (ARE) of the sample spatial median $\hat{\boldsymbol{\theta}}_n$ relative to the sample mean $\bar{X}_n$. When p is fixed, for spherical multivariate normal distribution, Brown (1983) showed that the asymptotic efficiency of $\hat{\boldsymbol{\theta}}_n$ relative $\bar{X}_n$, denoted as $\mathrm{ARE}(\hat{\boldsymbol{\theta}}_n, \bar{X}_n)$, exceeds the usual univariate case $2/\pi$. In addition, $\mathrm{ARE}(\hat{\boldsymbol{\theta}}_n, \bar{X}_n)$ increases as the dimension increases, and it approaches to 1 as p tends to be sufficient large (Magyar and Tyler, 2011). However, when $p \to \infty$, the ARE is not straightforward to quantify as there are no obvious "final" limit distributions for $\hat{\boldsymbol{\theta}}_n$ and $\bar{X}_n$. Motivated by the discussions in Sections 4.1 and 4.2, we compare $\hat{\boldsymbol{\theta}}_n$ and $\bar{X}_n$ in terms of their efficiencies in simultaneous inference for $\boldsymbol{\theta}$, which are determined by the variations of $|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}|_\infty$ and $|\bar{X}_n - \boldsymbol{\theta}|_\infty$. According to Corollary 1 and (4.1), we define the high-dimensional ARE of $\hat{\boldsymbol{\theta}}_n$ compared to $\bar{X}_n$ in simultaneous inference for $\boldsymbol{\theta}$ as

$$\mathrm{ARE}(\hat{\boldsymbol{\theta}}_n, \bar{X}_n) = \mathrm{Var}(|G_0|_\infty)/\mathrm{Var}(|G|_\infty)\,, \tag{4.5}$$

which approximates $\mathrm{Var}(|\bar{X}_n - \boldsymbol{\theta}|_\infty)/\mathrm{Var}(|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}|_\infty)$. If $\lim_{p\to\infty} \mathrm{ARE}(\hat{\boldsymbol{\theta}}_n, \bar{X}_n) > 1$, we say that $\hat{\boldsymbol{\theta}}_n$ is more efficient than $\bar{X}_n$ in simultaneous inference for $\boldsymbol{\theta}$ under high-dimensionality.

As discussed in Remark 6, $G \sim N(0, \zeta_1^{-2}\mathbb{B})$ with $\zeta_1^{-2}\mathbb{B}_{j\ell} = \zeta_1^{-2}p^{-1}\omega_{j\ell}$ for all $1 \le j, \ell \le p$. Moreover, we can show that $\Sigma_{j\ell} = \mathbb{E}(\nu_i^2)\omega_{j\ell} + O(p^{-1/2})$ similar to the proof of Lemma A3 in Appendix C of the supplementary material, where $\Sigma_{j\ell}$ is the (j,ℓ)th element of Σ. Thus, both the covariance matrix Σ and $\zeta_1^{-2}\mathbb{B}$ are proportional to Ω asymptotically, and $\mathrm{ARE}(\hat{\boldsymbol{\theta}}_n, \bar{X}_n)$ is approximately $\mathbb{E}(\nu_i^2)\zeta_1^2 p$.

As Σ and $\zeta_1^{-2}\mathbb{B}$ are rarely known in practice, we use bootstrap approximation to estimate the value of $\mathrm{Var}(|G_0|_\infty)/\mathrm{Var}(|G|_\infty)$. Combining Corollary 2 and (4.2), we propose using

$$\mathrm{Var}^*(|\bar{X}_n^*|_\infty)/\mathrm{Var}^*(|\tilde{\boldsymbol{\theta}}_n|_\infty),$$

to estimate $\mathrm{ARE}(\hat{\boldsymbol{\theta}}_n, \bar{X}_n)$.

Example 1. Suppose $X_1, \ldots X_n$ are i.i.d. from $N(\boldsymbol{\theta}, I_p)$, then $X_i = \boldsymbol{\theta} + \nu_i U_i$ with $\nu_i = 1$ and $U_i \sim N(0, I_p)$. Thus, $\|U_i\|^2$ follows a chi-squared distribution with p degrees of freedom. It follows that $\mathbb{E}(\nu_i^2) = 1$ and $\zeta_1 = \mathbb{E}(\|U_i\|^{-1}) = \Gamma(p/2 - 1/2)/\{2^{1/2}\Gamma(p/2)\}$, where $\Gamma(\cdot)$ is the gamma function. Then, $\mathrm{ARE}(\hat{\boldsymbol{\theta}}_n, \bar{X}_n) = p\{\Gamma(p/2 - 1/2)\}^2/\{2^{1/2}\Gamma(p/2)\}^2$. Using Stirling's formula, $\lim_{p\to\infty} \mathrm{ARE}(\hat{\boldsymbol{\theta}}_n, \bar{X}_n) = 1$. Thus, for high-dimensional Gaussian data, the sample spatial median has the same asymptotically efficiency as the sample mean in simultaneous inference for $\boldsymbol{\theta}$.

Example 2. When the data are from the multivariate t-distribution with degrees of freedom $v > 2$ and shape matrix $\Omega = I_p$, we have $X_i = \boldsymbol{\theta} + \nu_i U_i$ with $\nu_i = \sqrt{v/\xi_i}$ and $U_i \sim N(0, I_p)$, where ξ_i is a chi-squared random variable with degrees of freedom v independent of U_i. Then, $\mathbb{E}(\nu_i^2) = \mathbb{E}(v/\xi_i) = v/(v-2)$ and $\zeta_1 = \mathbb{E}(v^{-1/2}\xi_i^{1/2}\|U_i\|^{-1}) = \Gamma(v/2 + 1/2)\Gamma(p/2 - 1/2)/\{v^{1/2}\Gamma(v/2)\Gamma(p/2)\}$. Thus, the ARE is $\mathrm{ARE}(\hat{\boldsymbol{\theta}}_n, \bar{X}_n) = (v-2)^{-1}p\{\Gamma(v/2+1/2)\Gamma(p/2-1/2)\}^2/\{\Gamma(v/2)\Gamma(p/2)\}^2$. It is clear that $\mathrm{ARE}(\hat{\boldsymbol{\theta}}_n,\ \bar{X}) > 1$ for large enough p. In addition,

$$\lim_{p\to\infty} \mathrm{ARE}(\hat{\boldsymbol{\theta}}_n, \bar{X}_n) = 2(v-2)^{-1}\{\Gamma(v/2+1/2)\}^2/\{\Gamma(v/2)\}^2 > 1\,.$$

Thus, for high-dimensional t-distribution, the sample spatial median is asymptotically more efficient than the sample mean in simultaneous inference for $\boldsymbol{\theta}$. Table 1 and Figure 1 presents

the values of the limiting ARE $\lim_{p\to\infty} \mathrm{ARE}(\hat{\boldsymbol{\theta}}_n, \bar{X}_n)$ for different degrees of freedom v. The limiting ARE starts at 2.546 when $v = 3$, and decreases as v increases. As v approaches infinity, the value of the limiting ARE $\lim_{p\to\infty} \mathrm{ARE}(\hat{\boldsymbol{\theta}}_n, \bar{X}_n)$ converges to 1, which corresponding to the scenario of a multivariate normal distribution, where the ARE limit is 1.

Table 1: Asymptotic relative efficiency $\lim_{p\to\infty} \mathrm{ARE}(\hat{\boldsymbol{\theta}}_n, \bar{X}_n) = 2(v-2)^{-1}\{\Gamma(v/2+1/2)\}^2/\{\Gamma(v/2)\}^2$ for multivariate t-distributions with different degrees of freedom.

v	3	4	5	10	20	50	100
$\lim_{p\to\infty} \mathrm{ARE}(\hat{\boldsymbol{\theta}}_n, \bar{X}_n)$	2.546	1.767	1.509	1.189	1.084	1.031	1.015

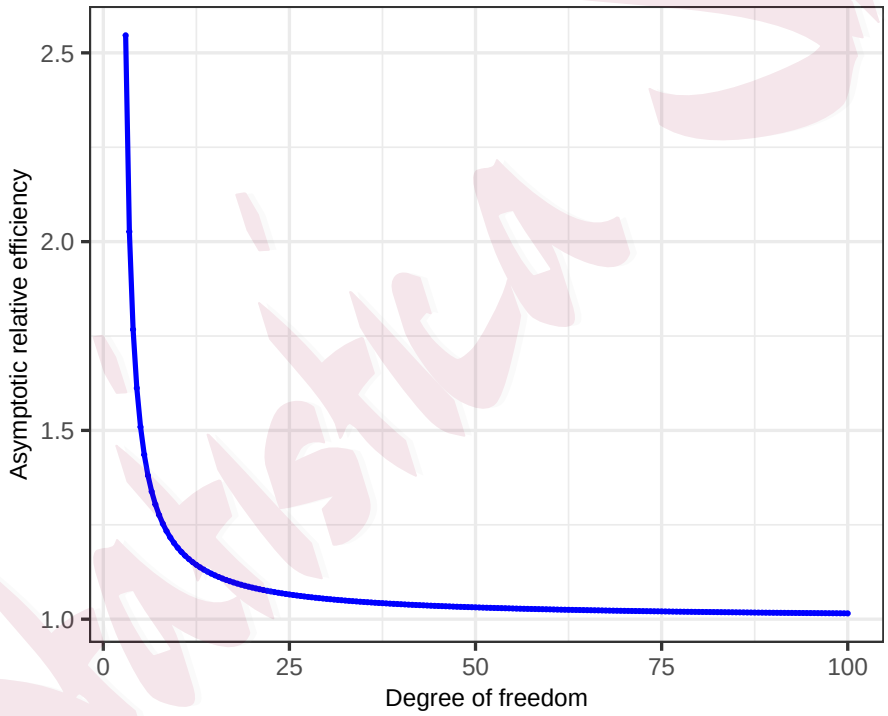

Figure 1: Asymptotic relative efficiency $\lim_{p\to\infty} \mathrm{ARE}(\hat{\boldsymbol{\theta}}_n, \bar{X}_n) = 2(v-2)^{-1}\{\Gamma(v/2+1/2)\}^2/\{\Gamma(v/2)\}^2$ for multivariate t-distributions with different degrees of freedom.

Figure 2 plots the simulated values of $\mathrm{Var}(|\bar{X}_n|_\infty)/\mathrm{Var}(|\hat{\boldsymbol{\theta}}_n|_\infty)$ with a range of dimensions and sample sizes under different models. For Gaussian data, the relative efficiency kept increasing in p, and it approached 1 as p getting larger. For the data simulated from multivariate t-distribution, the relative efficiency was greater than 1 for all combinations of n and

p. This indicates that the sample spatial median is more efficient than the sample mean for t-distribution. The results were consistent under different covariance structure considered in the simulation.

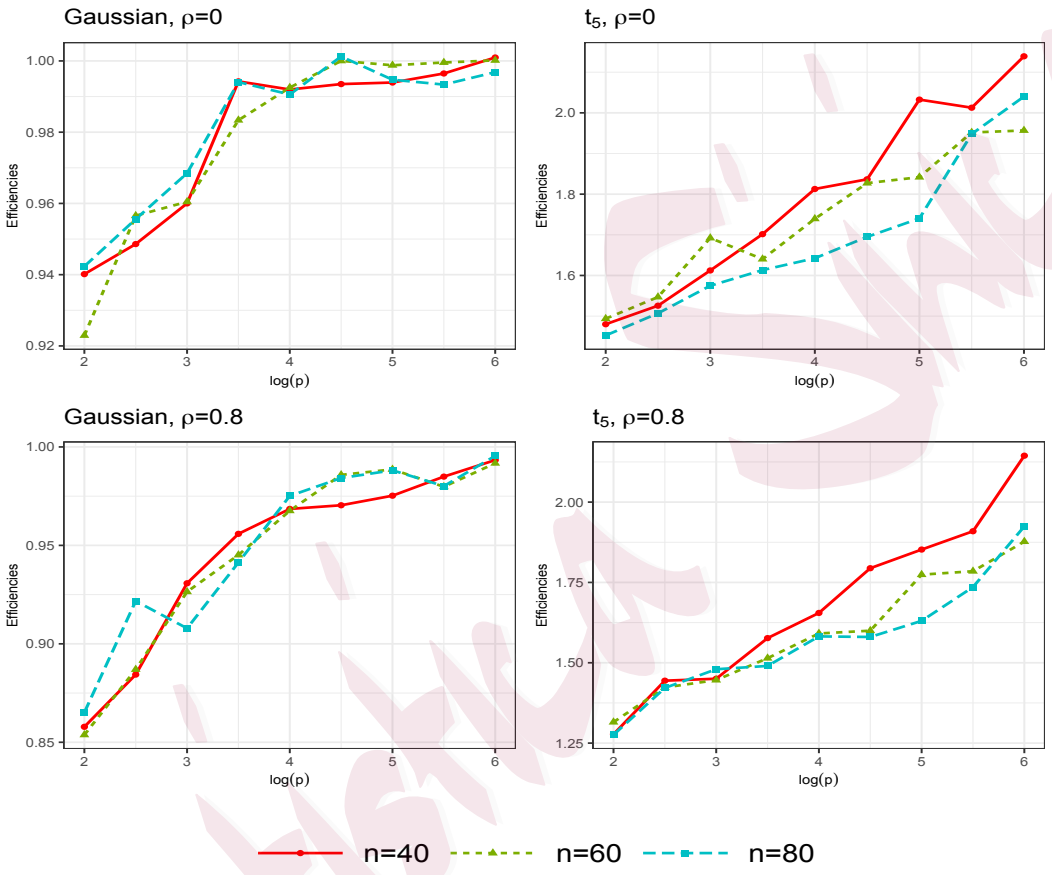

Figure 2: Finite sample relative efficiency of $|\hat{\boldsymbol{\theta}}_n|_\infty$ compared to $|\bar{X}_n|_\infty$ based on 5000 replications, the data are generated from multivariate normal distribution (Gaussian) and t-distribution with 5 degrees of freedom (t_5). The shape matrix $\Omega = (\rho^{|j-\ell|})_{p\times p}$ with $\rho = 0$ and 0.8.

5. Numerical studies

In this section, we report Monte Carlo simulations on simultaneous confidence intervals and multiple testing with FDR control, along with a real data analysis, to demonstrate the performance of the proposed approaches. Additional simulations on global tests can be found

in Appendix D of the supplementary material. In the simulations, all results were based on 2500 replications. In the bootstrap implementation, the number of bootstrap iterations was set to $B = 400$. To calculate the sample spatial median, we utilized the algorithm proposed by Vardi and Zhang (2000), which only requires $O(np)$ operations at each iteration. This efficient algorithm is implemented as the `Weiszfeld` function in the R package "Gmedian". A comparison of the computation times for the sample spatial median and sample mean, as well as the implementation time for both the spatial median-based and sample mean-based multiplier bootstrap methods can be found in Tables A1 and A2 of the supplementary materials.

5.1 Simulations on simultaneous confidence intervals

We first examine the performance of the SCIs based on $\hat{\boldsymbol{\theta}}_n$, and compare it with the SCIs based on $\bar{X}_n$. The sample size n is taken to be 100 or 200, and the dimensions $p = 100$ and 1000 are considered for each sample size. Two types of commonly used elliptical distributions are considered: (I) the multivariate normal distribution $N(\boldsymbol{\theta}, \Sigma)$; (II) the multivariate t-distribution with 3 degrees of freedom, mean vector $\boldsymbol{\theta}$, and covariance matrix Σ. In addition, we include the following independent components model: (III) $X_i = \boldsymbol{\theta} + \Sigma^{1/2} Z_i$, where each component of Z_i are i.i.d. from the standard Laplace distribution. We set $\Sigma = (\rho^{|j-\ell|})$ with $\rho = 0, 0.2, 0.5$ and 0.8. To save space, we present the results for $\rho = 0$ and 0.8 here. The results for $\rho \in \{0.2, 0.5\}$ are similar and are reported in the supplementary material. We consider both sparse and dense case scenarios for $\boldsymbol{\theta}$: (i) $\boldsymbol{\theta}_1 = (2, -2, 3, 0, \ldots, 0)$; (ii) $\boldsymbol{\theta}_2 = (0.2, \ldots, 0.2_{\lfloor p/4 \rfloor}, 0, \ldots, 0)$. Here $\lfloor \cdot \rfloor$ is the floor function.

Table 2 reports the coverage probability and median length of the SCIs based on $\hat{\boldsymbol{\theta}}_n$, the

results of the SCIs based on $\bar{X}_n$ are presented in parentheses. Recall that the SCIs base on $\hat{\boldsymbol{\theta}}_n$ are $[\theta^-_{n,j}, \theta^+_{n,j}]$ for $j = 1, \ldots, p$, where $\theta^-_{n,j} = \hat{\theta}_{n,j} - n^{-1/2} q^B_{1-\tau}$ and $\theta^+_{n,j} = \hat{\theta}_{n,j} + n^{-1/2} q^B_{1-\tau}$. The length of each confidence interval based on $\hat{\boldsymbol{\theta}}_n$ is the same for all θ_j, given by $2n^{-1/2} q^B_{1-\tau}$ for all $j = 1, \ldots, p$. It is important to note that $q^B_{1-\tau}$ depends on the dimension p. A similar argument applies to the SCIs based on $\bar{X}_n$. Thus, The "Median length" reported in Table 2 refers to the median of the lengths of the confidence intervals for all θ_j's across 2500 replications. For Models I and II from elliptical distributions, we observe that the SCIs based on $\hat{\boldsymbol{\theta}}_n$ and $\bar{X}_n$ both achieve satisfying coverage probability for different choices for ρ, $\boldsymbol{\theta}$, n and p. For the data simulated from the multivariate normal distribution, the median length of the SCIs based on $\hat{\boldsymbol{\theta}}_n$ is very close to that of the the SCIs based on $\bar{X}_n$. These results indicate that $\hat{\boldsymbol{\theta}}_n$ has similar asymptotic efficiency as $\bar{X}_n$ in simultaneous inference for $\boldsymbol{\theta}$ under high-dimensional Gaussian model as discussed in Section 4.4. For the multivariate t-distribution, the SCIs based on $\hat{\boldsymbol{\theta}}_n$ is much narrower than the SCIs based on $\bar{X}_n$. These results suggest that the SCIs based on $\hat{\boldsymbol{\theta}}_n$ is more efficient than the SCIs based on $\bar{X}_n$ for multivariate t-distribution, which is heavy-tailed. This is consistent with the asymptotic analysis in Section 4.4. Moreover, the results for Model III, which does not belong to the elliptical distribution family, shows the robustness of the SCIs based on the spatial median, and it performs similar to the SCIs based on the sample mean. It is shown that the median length of the SCIs decreases when n increases or p decreases for each model.

5.2 Simulations on multiple testing with FDR control

In this section, we examine the performance of the sample spatial median-based B-H method introduced in Section 4.3, and compare it to the B-H procedure based on the sample mean

Table 2: Coverage probability (in %) and median length of the SCIs based on $\hat{\boldsymbol{\theta}}_n$, the results of the SCIs based on $\bar{X}_n$ are in parentheses.

				$\boldsymbol{\theta}=\boldsymbol{\theta}_1$				$\boldsymbol{\theta}=\boldsymbol{\theta}_2$			
				Coverage probability		Median length		Coverage probability		Median length	
Model	ρ	n	p	90%	95%	90%	95%	90%	95%	90%	95%
I	0	100	100	89.6 (89.9)	94.4 (94.4)	0.65 (0.65)	0.69 (0.69)	88.9 (88.8)	94.1 (93.9)	0.65 (0.65)	0.69 (0.69)
			1000	89.5 (89.6)	94.7 (94.4)	0.77 (0.77)	0.80 (0.80)	89.5 (89.5)	94.0 (94.0)	0.77 (0.77)	0.81 (0.80)
		200	100	89.8 (89.8)	95.1 (95.1)	0.46 (0.46)	0.49 (0.49)	88.6 (88.8)	94.4 (94.7)	0.46 (0.46)	0.49 (0.49)
			1000	89.7 (89.7)	94.4 (94.6)	0.55 (0.55)	0.57 (0.57)	89.1 (89.2)	94.7 (94.6)	0.55 (0.55)	0.57 (0.57)
	0.8	100	100	89.1 (88.7)	94.6 (94.6)	0.64 (0.63)	0.68 (0.67)	88.4 (88.6)	93.7 (94.1)	0.64 (0.63)	0.68 (0.67)
			1000	88.4 (88.4)	93.8 (93.7)	0.76 (0.76)	0.80 (0.79)	89.0 (89.2)	94.6 (94.6)	0.76 (0.76)	0.80 (0.79)
		200	100	90.5 (90.1)	95.2 (94.9)	0.45 (0.45)	0.48 (0.48)	89.6 (89.6)	94.0 (94.1)	0.45 (0.45)	0.48 (0.48)
			1000	90.4 (90.4)	94.5 (94.4)	0.54 (0.54)	0.56 (0.56)	88.4 (88.5)	93.6 (93.8)	0.54 (0.54)	0.56 (0.56)
II	0	100	100	89.7 (88.6)	94.7 (93.7)	0.71 (1.05)	0.75 (1.11)	88.8 (88.8)	94.5 (94.2)	0.71 (1.05)	0.75 (1.11)
			1000	89.4 (91.0)	95.8 (95.0)	0.84 (1.25)	0.88 (1.30)	89.1 (89.0)	94.4 (94.5)	0.84 (1.25)	0.88 (1.31)
		200	100	88.6 (89.1)	94.2 (95.1)	0.50 (0.76)	0.53 (0.81)	89.5 (89.7)	94.4 (94.8)	0.50 (0.76)	0.53 (0.80)
			1000	89.6 (88.7)	94.8 (94.6)	0.59 (0.90)	0.62 (0.94)	90.1 (89.5)	94.8 (93.9)	0.59 (0.90)	0.62 (0.94)
	0.8	100	100	89.1 (90.7)	94.4 (94.9)	0.69 (1.02)	0.74 (1.09)	89.4 (89.7)	94.2 (94.4)	0.69 (1.02)	0.74 (1.09)
			1000	89.3 (89.1)	94.6 (94.4)	0.83 (1.23)	0.87 (1.29)	89.8 (88.8)	94.7 (94.4)	0.83 (1.23)	0.87 (1.29)
		200	100	87.6 (87.7)	93.4 (93.6)	0.49 (0.73)	0.52 (0.78)	90.3 (90.1)	94.9 (95.2)	0.49 (0.73)	0.52 (0.78)
			1000	88.7 (89.7)	94.7 (94.6)	0.59 (0.88)	0.61 (0.92)	90.2 (90.8)	94.7 (95.7)	0.59 (0.89)	0.61 (0.93)
III	0	100	100	89.8 (89.4)	94.6 (94.5)	0.65 (0.66)	0.69 (0.70)	89.1 (89.0)	94.4 (94.4)	0.65 (0.66)	0.69 (0.70)
			1000	88.3 (88.2)	93.6 (93.7)	0.78 (0.78)	0.82 (0.82)	89.1 (89.0)	94.2 (93.8)	0.78 (0.78)	0.82 (0.82)
		200	100	90.6 (91.1)	95.0 (95.0)	0.46 (0.46)	0.49 (0.49)	90.6 (90.1)	95.2 (95.2)	0.46 (0.46)	0.49 (0.49)
			1000	90.1 (90.4)	95.0 (94.6)	0.55 (0.55)	0.57 (0.58)	88.7 (89)	93.6 (93.8)	0.55 (0.55)	0.57 (0.58)
	0.8	100	100	90.4 (89.7)	95.0 (94.8)	0.63 (0.63)	0.68 (0.68)	89.0 (88.9)	95.0 (94.9)	0.63 (0.63)	0.67 (0.68)
			1000	88.7 (88.9)	93.8 (94.0)	0.77 (0.77)	0.80 (0.80)	89.0 (89.0)	94.6 (94.3)	0.76 (0.76)	0.80 (0.80)
		200	100	88.8 (89.1)	94.2 (94.0)	0.45 (0.45)	0.48 (0.48)	90.2 (89.7)	94.8 (95.0)	0.45 (0.45)	0.48 (0.48)
			1000	90.0 (90.3)	95.0 (95.0)	0.54 (0.54)	0.57 (0.57)	88.8 (89.1)	94.2 (94.1)	0.54 (0.54)	0.57 (0.57)

with p-values calculated from $N(0,1)$ in Liu and Shao (2014). We set $\theta_{0,j}=0$ for all $j=1,\ldots,p$. The data are generated from Models I and II with $p=1000$. For $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_p)^\top$, let $\theta_j=2(\log p/n)^{1/2}$ for $1\le j\le p_1$ and $\theta_j=0$ for $(p_1+1)\le j\le p$, where $p_1=0.1p$.

Table 3 reports the empirical FDR and power for the sample spatial median-based (FDR_M and power_M) and the sample mean-based (FDR_A and power_A) B-H procedures (Liu and Shao, 2014) with nominal level $\alpha=0.1$ and 0.2. The results indicate that the FDR are well controlled by both methods. For the multivariate normal distribution, the B-H procedures based on the spatial median and the sample mean have similar performance. However, the spatial median-based B-H method outperforms the sample mean-based B-H procedure in terms of empirical power under multivariate t-distribution, which is heavy-tailed.

Table 3: Empirical FDR and power for the spatial median-based (FDR_M and power_M) and the sample mean-based (FDR_A and power_A) in Liu and Shao (2014) via B-H procedures.

			$\alpha = 0.1$				$\alpha = 0.2$			
Model	ρ	n	FDR_M	FDR_A	power_M	power_A	FDR_M	FDR_A	power_M	power_A
I	0	50	0.124	0.124	0.996	0.996	0.224	0.222	0.999	0.999
		100	0.107	0.106	0.997	0.997	0.202	0.201	0.999	0.999
	0.2	50	0.125	0.124	0.996	0.996	0.224	0.223	0.999	0.999
		100	0.107	0.106	0.997	0.997	0.202	0.201	0.999	0.999
	0.5	50	0.125	0.124	0.996	0.996	0.225	0.223	0.999	0.999
		100	0.107	0.105	0.997	0.997	0.202	0.201	0.999	0.999
	0.8	50	0.127	0.124	0.996	0.996	0.227	0.223	0.999	0.999
		100	0.108	0.105	0.997	0.997	0.204	0.199	0.999	0.999
II	0	50	0.117	0.099	0.984	0.728	0.215	0.193	0.992	0.805
		100	0.103	0.088	0.987	0.710	0.197	0.179	0.994	0.795
	0.2	50	0.117	0.098	0.984	0.727	0.215	0.194	0.992	0.805
		100	0.103	0.087	0.987	0.709	0.198	0.179	0.994	0.795
	0.5	50	0.118	0.099	0.984	0.727	0.216	0.194	0.992	0.803
		100	0.103	0.087	0.987	0.708	0.198	0.178	0.994	0.794
	0.8	50	0.120	0.098	0.984	0.724	0.218	0.192	0.992	0.800
		100	0.104	0.087	0.987	0.705	0.199	0.177	0.994	0.791

5.3 Real data analysis

Type 2 diabetesis a disease in which the body becomes resistant to normal effects of insulin and gradually loses the capacity to produce enough insulin. Because skeletal muscle is the main tissue for insulin-stimulated glucose disposal, skeletal muscle insulin resistance is commonly viewed as the critical component of whole-body insulin resistance, and thus is critical to the pathogenesis of Type 2 diabetes. To investigate the effects of insulin on gene expression in skeletal muscle, a microarray study was performed in 15 diabetic patients using the Affymetrix Hu95A chip of muscle biopsies both before and after insulin treatment (Wu et al., 2007). In this paper, we are interested in the gene expression alteration, that is, the change of the gene expression level, due to the treatment. The data are available at `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE22309`. The data were

normalized by the quantile normalization method implemented by the `normalizeQuantiles` function in the `limma` R package. Follow Wang et al. (2015), we focused on 2547 curated gene sets with at least 15 genes, which are from the C2 collection of the GSEA online pathway databases. The gene expression values are consolidated by taking the average when multiple probes are associated with the same gene.

We implemented the Median test based on T_n on the 2519 gene sets. This is equivalent to testing whether the median change vector of gene expression levels is equal to 0. The number of bootstrap iterations is $B = 10^5$. With the Bonferroni correction, there are 1242 gene sets identified as significant at 5% level. For comparison, we applied the WPL test (Wang et al., 2015) and the CQ test (Chen and Qin, 2010) on the same gene sets. For the WPL test, 1060 gene sets are selected as significant; and for the CQ test, 630 gene sets are identified as significant. Out of the 630 gene sets selected by the CQ test, 605 of them are also identified by our proposed method, and 629 of them are identified by the WPL test. It has been argued in Wang et al. (2015) that some gene expression levels have heavy tails as their kurtosises are much larger than the kurtosis of a normal distribution, 3. Thus, the methods based on the spatial median (Median test and the WPL test) are expected to be more robust and efficient than those based on moments (CQ test). In addition, out of the 1060 gene sets identified by the WPL test, 958 of them are significant based on our proposed approach.

As argued in Remark 12, the Median test based on T_n is more powerful in detecting strong sparse signal compared to the WPL test. To see this, we look into the following three gene sets:

(1) ZHAN_MULTIPLE_MYELOMA_UP;

(2) MIKKELSEN_MEF_HCP_WITH_H3K27ME3;

(3) JAZAG_TGFB1_SIGNALING_VIA_SMAD4_UP.

The p-values of the WPL test for these three gene sets are 0.41, 0.31, 0.27, respectively. However, the p-values of the Median test are all less than 1.0×10^{-5} with $B = 10^5$ bootstrap iterations for these three gene sets. Figure 3 plots the SCIs for the spatial median vectors of the change of gene expression levels for these three gene sets. The confidence intervals that do not cover 0 are colored in red. It is very clear that the only one or two big values in the spatial median results in a rejection of the Median test, while the signals from other dimensions are not strong enough to land a rejection by the WPL test.

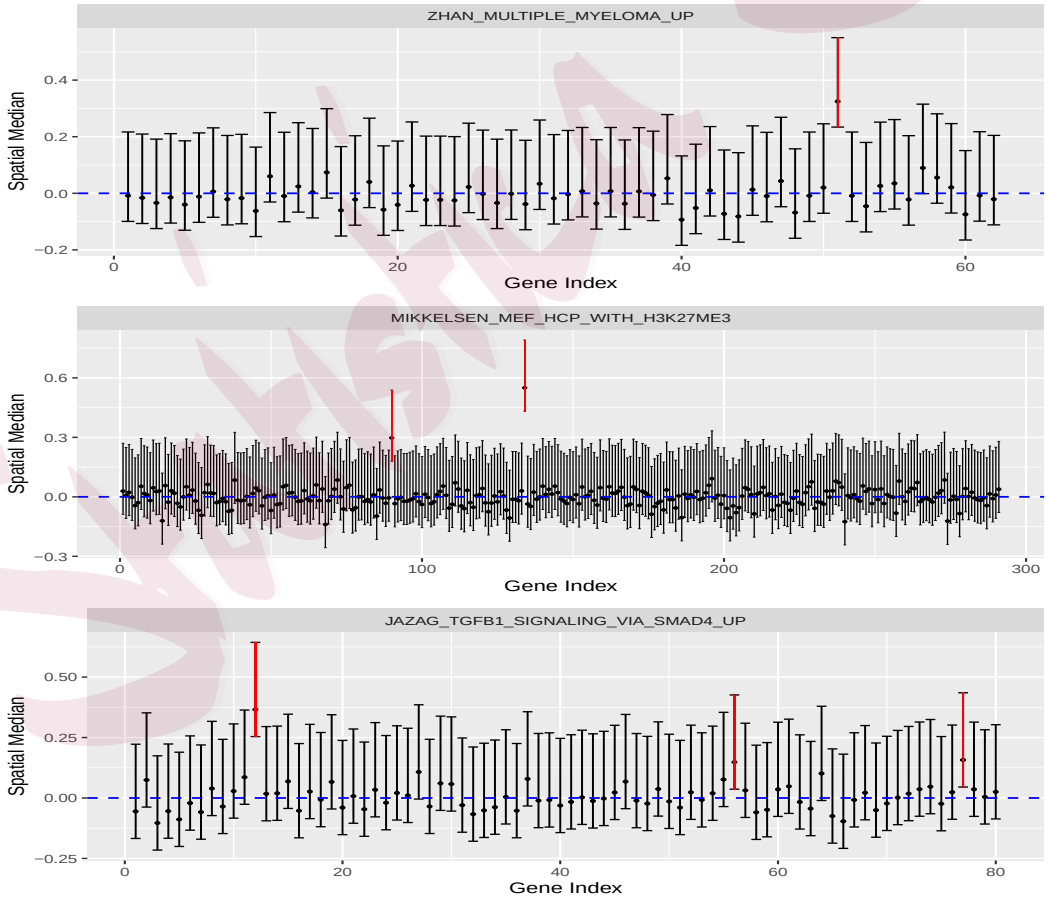

Figure 3: Simultaneous Confidence intervals (SCIs) for spatial medians of three gene sets.

Finally, we use the spatial median-based B-H procedure to perform multiple testing with FDR control on the three gene sets to detect differentially expressed genes (DEG), which is one of the most important targets in genomic analysis. Table 4 reports the detected differentially expressed genes (DEG) in each gene set with nominal level $\alpha = 0.1$, along with the corresponding marginal p-value $P_j = 2 - 2\Phi(|T_{n,j}|)$ and the confidence interval in the SCIs for the selected genes. It can be seen that for all the selected genes, the marginal p-values are very small, and the corresponding confidence intervals do not cover 0.

Table 4: Detected differentially expressed genes (DEG) by the spatial median-based B-H procedure for three gene sets with $\alpha = 0.1$; "p-value" refers to the marginal p-value $P_j = 2 - 2\Phi(|T_{n,j}|)$, and "CI" refers to the confidence interval in the SCIs for the selected genes.

Gene set	DEG	p-value	CI
ZHAN_MULTIPLE_MYELOMA_UP	CDKN1A	0.00082	(0.234, 0.550)
MIKKELSEN_MEF_HCP_WITH_H3K27ME3	MYOD1	< 0.00001	(0.433, 0.791)
JAZAG_TGFB1_SIGNALING_VIA_SMAD4_UP	HDAC4	0.00058	(0.254, 0.644)

6. Discussion

In this paper, we established one-sample Gaussian and bootstrap approximations for ultra-high dimensional sample spatial median under a general model beyond elliptical distributions. It is of interest to study whether our results are potentially extendable to some other distribution families and two-sample or multi-sample Gaussian and bootstrap approximations. We leave this to a future work. In addition, the proposed test based on the maxima of the sample spatial median is more powerful under sparse alternatives compared to those based on L_2-norms. It is well known that the L_2-norm type tests are more powerful under dense alternatives. Thus, it is of interest to consider combining the test based on the maximum-norm and L_2-norm, which could be potentially powerful under both sparse and dense alternatives.

We also leave this to a future study.

Supplementary Materials

The supplementary materials consist of the proofs of main results in the paper, preliminary lemmas, and additional simulation results.

References

Belloni, A., V. Chernozhukov, D. Chetverikov, C. Hansen, and K. Kato (2018). High-dimensional econometrics and generalized gmm. *arXiv*, 1806.01888.

Bickel, P. J. and E. Levina (2008). Covariance regularization by thresholding. *Ann. Statist. 36*, 2577–2604.

Brown, B. (1983). Statistical uses of the spatial median. *J. R. Statist. Soc.* B *45*, 25–30.

Cardot, H., P. Cénac, and P.-A. Zitt (2013). Efficient and fast estimation of the geometric median in hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli 19*, 18–43.

Chen, S. X. and Y. Qin (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist. 38*, 808–835.

Chen, X. (2018). Gaussian and bootstrap approximations for high-dimensional U-statistics and their applications. *Ann. Statist. 46*, 642–678.

Cheng, G., B. Liu, L. Peng, B. Zhang, and S. Zheng (2019). Testing the equality of two high-dimensional spatial sign covariance matrices. *Scand. J. Statist. 46*, 257–271.

Chernozhukov, V., D. Chetverikov, and K. Kato (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist. 41*, 2786–2819.

Chernozhukov, V., D. Chetverikov, and K. Kato (2017). Central limit theorems and bootstrap in high dimensions. *The Annals of Probability 45*, 2309–2352.

Chernozhukov, V., D. Chetverikov, and K. Kato (2019). Improved central limit theorem and bootstrap approximation in high dimensions. *arXiv*, 1912.10529.

Chernozhukov, V., D. Chetverikov, K. Kato, and Y. Koike (2020). Nearly optimal central limit theorem and bootstrap approximations in high dimensions. *arXiv*, 2012.09513.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Statist. Soc.* B *70*, 849–911.

Fang, K. W., S. Kotz, and K. W. Ng (1990). *Symmetric multivariate and related distributions.* Boca Raton, FL: CRC Press.

Haldane, J. B. S. (1948). Note on the median of a multivariate distribution. *Biometrika 35*, 414–417.

Hsu, D. and S. Sabato (2016). Loss minimization and parameter estimation with heavy tails. *J. Mach. Learn. Res. 17*, 1–40.

Imaizumi, M. and T. Otsu (2021). On gaussian approximation for m-estimator. *arXiv*, 2012.15678v2.

Li, W. and Y. Xu (2022). Asymptotic properties of high-dimensional spatial median in elliptical distributions with application. *Journal of Multivariate Analysis 190*, 104975.

Liu, W. and Q.-M. Shao (2014). Phase transition and regularized bootstrap in large scale t-tests with false discovery rate control. *Ann. Statist. 42*, 2003–2025.

Magyar, A. and D. E. Tyler (2011). The asymptotic efficiency of the spatial median for elliptically symmetric distributions. *Sankhya B 73*, 165–192.

McNeil, A. J., R. Frey, and P. Embrechts (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton, NJ: Princeton University Press.

Milasevic, P. and G. R. Ducharme (1987). Uniqueness of the spatial median. *Ann. Statist. 15*, 1332–1333.

Minsker, S. (2015). Geometric median and robust estimation in banach spaces. *Bernoulli 21*, 2308–2335.

Oja, H. (2010). *Multivariate nonparametric methods with R: An approach based on spatial signs and ranks*. Lecture Notes in Statistics, Springer, New York.

Prasad, A., A. S. Suggala, S. Balakrishnan, and P. Ravikumar (2020). Robust estimation via robust gradient estimation. *J. R. Statist. Soc. B 82*, 601–627.

Purdom, E. and S. P. Holmes (2005). Error distribution for gene expression data. *Statistical Applications in Genetics and Molecular Biology 4*, 1–35.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.

Vardi, Y. and C. H. Zhang (2000). The multivariate l 1-median and associated data depth. *Proceedings of the National Academy of Sciences 97*, 1423–1426.

Wang, L., B. Peng, and R. Li (2015). A high-dimensional nonparametric multivariate test for mean vector. *J. Am. Statist. Assoc. 110*, 1658–1669.

Weber, A. (1929). *Uber Den Standort der Industrien (Alfred Weber?s Theory of the Location of Industries)*. Chicago, IL: Univ. Chicago Press.

Wu, X., J. Wang, X. Cui, L. Maianu, B. Rhees, J. Rosinski, W. V. So, S. M. Willi, M. V. Osier, H. S. Hill, G. P. Page, D. B. Allison, M. Maritin, and W. T. Garvey (2007). The effect of insulin on expression of genes and biochemical pathways in human skeletal muscle. *Endocrine 31*, 5–17.

Yao, J., S. Zheng, and Z. Bai (2015). *Sample covariance matrices and high-dimensional data analysis*. Cambridge University Press, Cambridge.

Zou, C., L. Peng, L. Feng, and Z. Wang (2014). Multivariate sign-based high-dimensional tests for sphericity. *Biometrika 101*, 229–236.

Guangzhou Institute of International Finance, Guangzhou University, Guangzhou 510006, China

E-mail: chenggh845@nenu.edu.cn

School of Mathematics and Statistics, the University of Melbourne, Victoria 3010, Australia

E-mail: liuhua.peng@unimelb.edu.au

School of Statistics and Data Science, Nankai University, Tianjin 300071, China

E-mail: nk.chlzou@gmail.com