The Theil-Sen Estimators in a Multiple Linear Regression Model

Xin Dang [a,1], Hanxiang Peng [b,1,*], Xueqin Wang, [c,d,2]
Heping Zhang [c,2]

[a] *Department of Mathematics, the University of Mississippi, University, MS 38677-1848, USA*

[b] *Department of Mathematical Sciences, Indiana University Purdue University Indianapolis, Indiana 46074, USA*

[c] *Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520-8034, USA*

[d] *School of Mathematics & Computational Science, Zhongshan School of Medicine, Sun Yat-Sen University, P.R. China*

Abstract

In this article, we propose the Theil-Sen estimators of parameters in a multiple linear regression based on multivariate medians, generalizing the Theil-Sen estimator in a simple linear regression. We show that the proposed estimators are robust, consistent, asymptotically normal under mild conditions, and super-efficient when the error distribution is discontinuous. The estimators can be chosen to allow for pre-specified robustness and efficiency. Simulations are conducted to compare robustness and efficiency with least squares estimators and to validate super-efficiency. Additionally, we show that a random variable is symmetric if and only if the random vectors whose components are the differences of three i.i.d. copies of the random variable are symmetric.

AMS 2000 subject classification: primary 62G05; 62G20.

Key words: breakdown point; depth function; efficiency; multiple linear regression; robustness; spatial median.

* Corresponding author.
Email address: hpeng@math.iupui.edu (Hanxiang Peng).

[1] This research is supported by the US National Science Foundation under Grant No. DMS-0707074.

[2] This research is supported in part by grants K02DA017713, R01DA016750 and T32MH014235 from the US National Institute of Health.

1 Introduction

In a simple linear regression model, Theil (1950) proposed the median of pairwise slopes as an estimator of the slope parameter. Sen (1968) extended this estimator to handle ties. The Theil-Sen estimator (TSE) is robust with a high breakdown point 29.3%, has a bounded influence function, and possesses high asymptotic efficiency. Thus it is very competitive to other slope estimators (e.g. the least squares estimator), see Sen (1968), Dietz (1989) and Wilcox (1998). The TSE has been acknowledged in several popular textbooks on nonparametric and robust statistics, e.g., Sprent (1993), Hollander and Wolfe (1973, 1999), and Rousseeuw and Leroy (1986). It has important applications, for example, in astronomy by Akritas *et al.* (1995) in censored data, in remote sensing by Fernandes and Leblanc (2005). Sen (1968) obtained unbiasedness and asymptotic normality of the estimator for an absolutely continuous error distribution and a nonidentical covariate. Viewed as a generalized L-statistics, its asymptotics can be obtained from Serfling (1984). Wang (2005) investigated the asymptotic behaviors of the TSE when the covariate is random. Peng, Wang and Wang (2005) obtained the consistency and asymptotic distribution of the TSE when the error distribution is arbitrary and the asymptotic normality obtained by Sen (1968) follows as a special case. They showed further that the TSE is super-efficient when the error distribution is discontinuous.

Despite its many good properties and clear geometric interpretation, the TSE is vastly under-developed and -used because it is only formulated for a simple linear model, although statisticians have made their efforts to extend it, see, e.g., Oja and Niinima (1984), Zhou and Serfling (2006). While the extension of TSE to a multiple linear model is geometrically apparent and appealing, it is technically challenging, delaying the generalization and investigation of the properties. In this article, we propose the use of multivariate medians to generalize the Theil-Sen estimator of the slope parameter in a simple linear model to a multiple linear model in several ways. Multivariate medians (multidimensional medians, as also used by some authors) generalize the univariate median and are a well established notion in the literature, see, e.g., Small (1990). Our approach is essentially a hybrid of two principles, i.e., least-squares estimate and multivariate median. Specifically, for each sub-sample of size k (at least the number of parameters) from a random sample of size n, calculate the least-squares(LS) estimate of the parameter vector in a multiple linear regression, so that we obtain $\binom{n}{k}$ LS estimates. Then a natural robust estimate of the parameter vector is the multivariate median of these LS estimates. The construction itself manifests that it is robust to outliers and efficient in a certain extent.

The proposed estimators contain an integer variable k which controls the amount of robustness and efficiency. The maximal possible robustness (in

terms of breakdown point) is attained when the integer variable is chosen to be equal to the number of the parameters to be estimated, while the maximal efficiency is achieved when the variable is equal to the sample size. Any value of the variable taking values between the number of parameters and the sample size results in an estimator which gives a compromise between robustness and efficiency.

Our construction applies to any multivariate median including, of course, those defined via depth functions. Specifically, a depth-defined multivariate median is a maximizer of the depth function. The theory of depth functions is relatively young and is still under its development. Analogous to *linear order* in one dimension, statistical depth functions provide a center-outward ordering of multidimensional data. Tukey (1975) first introduced *halfspace depth*. Oja (1983) defined *Oja depth*. Liu (1990) proposed *simplicial depth*. Zuo and Serfling (2000a) considered *projection depth*. Other notions include *Zonoid depth* (Koshevoy and Mosler, 1997), *generalized Tukey depth* (Zhang 2002), and *spatial depth* (Chaudhuri 1996) among others. Of the various depths the *spatial depth* is especially appealing because of its computational ease and mathematical tractability. Its complexity is indeed n^2 for sample size n regardless of the dimension. In contrast, for example, the computational complexity for halfspace and simplicial depth is $O(n^{d-1}\log n)$ (Rousseeuw and Ruts, 1996), for projection depth, it is $O([\binom{2(d-1)}{d-1}/d]^2 n^3)$, where d is the dimension. This is an NP-hard problem in high dimensional data (Ghosh and Chaudhuri, 2005).

Thus we shall mainly focus on the spatial-depth-based MTSE's, although analogs for some of other depths-based MTSE can be easily obtained. We shall show that the proposed MTSE's are robust with a relatively high breakdown point and possesses a bounded influence function. We shall establish the strong consistency under mild conditions, super-efficiency for a discontinuous error distribution, and asymptotic normality. We shall conduct simulations to investigate the estimators about its computation, robustness, efficiency, and super-efficiency. Additionally, we shall prove that a random variable is symmetric if and only if all the random vectors whose components are the differences of three independent and identically distributed copies are symmetric about zero, see Theorem 1.

The rest of the article is structured as follows. Section 2 gives the proposed estimators. Section 3 discusses existence and uniqueness. A theorem characterizing the symmetry of a vector is given. Useful facts for the uniqueness are collected. Section 4 deals with asymptotic consistency. Two useful theorems on the convergence of U-statistics are given. Section 5 presents asymptotic normality and super-efficiency. Two useful theorems on the asymptotic normality of U-statistics are given. Section 6 is devoted to robustness considerations. The complexity, breakdown points, and influence function are computed. Section 7 reports simulations. We also discuss the relationships of the estimators among

robustness, efficiency, and computational complexity. Stochastic sampling of subpopulation is described. Some of the technical proofs are collected in the appendix.

2 The Proposed Multivariate Theil-Sen Estimators

In this section, we generalize the TSE in two ways and the third is given in the next section. Consider a multiple linear regression model

$$Y_i = \alpha + X_i^\top \beta + \epsilon_i, \quad i = 1, ..., n, \tag{1}$$

where α is the intercept and β is a p-dimension parameter, and $\epsilon_1, ..., \epsilon_n, \epsilon$ are i.i.d. random errors.

We start with the simple linear regression $p = 1$. Geometrically, in order to estimate the slope β, only two distinct points $(X_i, Y_i), (X_j, Y_j)$ $(X_i \neq X_j$, say) are needed; an estimator of the slope β is $b_{i,j} = (Y_i - Y_j)/(X_i - X_j)$. Alternatively, with every two distinct points, the sum of squares of residuals is $(Y_i - \alpha - \beta X_i)^2 + (Y_j - \alpha - \beta X_j)^2$, which is minimized when α, β satisfy the equations

$$Y_i - \alpha - \beta X_i = 0, \quad Y_j - \alpha - \beta X_j = 0.$$

The solutions $a_{i,j} = Y_i - b_{i,j} X_i$ and $b_{i,j} = (Y_i - Y_j)/(X_i - X_j)$ are the *least squares estimators*. A *robust* estimator $\tilde{\beta}_n$ of the slope β is then the median of these least squares estimates:

$$\tilde{\beta}_n = \text{Med}\left\{b_{i,j} = (Y_i - Y_j)/(X_i - X_j) : \ X_i \neq X_j, \ 1 \leq i < j \leq n\right\},$$

where $\text{Med}\{B_j : j \in J\}$ denotes the median of the numbers $\{B_j : j \in J\}$. This is the well known Theil-Sen estimator which is robust with high breakdown point. If only the estimation of the slope β is concerned, no identifiability assumption on the error is needed. In order to estimate the intercept, however, certain identifiability condition on the error distribution is indispensable. We now assume

Assumption S. The error has a distribution which is symmetric about zero.

This is a sufficient condition and a less restrictive condition is given later. Then, likewise, the intercept may be estimated by the median of the least squares estimates:

$$\tilde{\alpha}_n = \text{Med}\left\{a_{i,j} = (Y_j X_i - Y_i X_j)/(X_i - X_j) : \ X_i \neq X_j, \ 1 \leq i < j \leq n\right\}.$$

These result in a componentwise median estimator $(\tilde{\alpha}_n, \tilde{\beta}_n)$ of the parameter (α, β). It is known that a componentwise median estimator may be a very poor estimator, for example, the componentwise median of the points $(1,0,0)$, $(0,1,0)$, $(0,0,1)$ is $(0,0,0)$ which is not even on the plane passing through the three points. To overcome this flaw, we could use the robust $\tilde{\beta}_n$ to construct a robust estimator of the intercept α, for example, $\text{Med}\{Y_i - \tilde{\beta}_n X_i : 1 \le i, j \le n\}$ in Chatterjee and Olkin (2006) among others. Alternatively, we may estimate (α, β) *simultaneously* by the *multivariate median*:

$$(\tilde{\alpha}_n, \tilde{\beta}_n) = \text{Mmed}\,\{(a_{i,j}, b_{i,j}) : \ X_i \neq X_j, \ 1 \le i < j \le n\},$$

where $\text{Mmed}\,\{B_j : j \in J\}$ stands for the multivariate median of the vectors $\{B_j \in \mathbb{R}^d : j \in J\}$, see Sections 1 and 3 for discussion about multivariate medians. We shall be using the multivariate medians to construct the Theil-Sen estimators of parameters in a multiple linear regression.

Estimating simultaneous intercept and "slope" vector. Consider a multiple linear regression with $p \ge 1$. Following the above procedure, first, an estimator of $\theta = (\alpha, \beta^\top)^\top$ can be found as the solution to the $p+1$ equations

$$Y_i - \alpha - X_i^\top \beta = 0, \quad i \in \mathbf{k}_{p+1} = \{i_1, ..., i_{p+1}\}\,, \tag{2}$$

where $\mathbf{k}_{p+1}$ is a (p+1)-subset of $\{1, ..., n\}$ such that $(p+1) \times (p+1)$ matrix $(X_k : k \in \mathbf{k}_{p+1})$ is invertible. To emphasize the dependence on the $p+1$ observations, we denote this estimator by $\hat{\theta}_{\mathbf{k}_{p+1}}$. Then a natural extension of the Theil-Sen estimator from a simple linear regression to a multiple linear regression is the multivariate median

$$\tilde{\theta}_n = \text{Mmed}\left\{\hat{\theta}_{\mathbf{k}_{p+1}} : \forall \ \ \mathbf{k}_{p+1}\right\}.$$

Note that this $\hat{\theta}_{\mathbf{k}_{p+1}}$ is also the least squares estimator of θ based on $p+1$ observations $\{(X_i, Y_i) : i \in \mathbf{k}_{p+1}\}$. From this point of view and slightly more generally, one may choose an arbitrary combination of m distinct observations $\{(X_i, Y_i) : i \in \mathbf{k}_m\}$, where $p+1 \le m \le n$, and construct a least squares estimator $\hat{\theta}_{\mathbf{k}_m}$. Then a multiple Theil-Sen estimator $\hat{\theta}_n$ of the parameter θ is naturally defined to be the multivariate median of all possible least squares estimators:

$$\hat{\theta}_n = \text{Mmed}\left\{\hat{\theta}_{\mathbf{k}_m} : \forall \ \ \mathbf{k}_m\right\}. \tag{3}$$

Herein a possible least squares estimator is such that

$$\hat{\theta}_{\mathbf{k}} = (X_{\mathbf{k}}^\top X_{\mathbf{k}})^{-1} X_{\mathbf{k}}^\top Y_{\mathbf{k}}, \tag{4}$$

where $X_{\mathbf{k}}^\top X_{\mathbf{k}}$ is assumed invertible with $X_{\mathbf{k}}$ being an $(1+p) \times m$ matrix with rows $(1, X_i^\top) : i \in \mathbf{k}$ and $Y_{\mathbf{k}} = (Y_i : i \in \mathbf{k})^\top$. Here for ease of notation we have written $\mathbf{k} = \mathbf{k}_m$ and hereafter we shall use this notation. We shall point out

here that by choosing the value of m we can compromise between robustness and efficiency. See more discussion in Section 6.

Estimating the "slope" vector. If one is only interested in estimating the "slope" parameter β, then the identifiability condition on the distribution of the error for the intercept α such as the symmetry Assumption S is not required, as in the univariate TSE. Zhou and Serfling (2006) developed a theory of spatial U-quantiles and, as an application of the theory, generalized TSE to MTSE based on pairwise differences of the observations. Here we briefly review their result (slightly more general, in their construction, $m = p+1$). Note that their extension of TSE is based on the spatial depth but can be extended straightforwardly to an arbitrary multivariate median.

Consider the pairwise difference of (1):

$$Y_j - Y_k = (X_j - X_k)^\top \beta + \epsilon_j - \epsilon_k, \quad j, k = 1, 2, ..., n. \tag{5}$$

There are $N = n(n-1)/2$ pairwise differences. For an integer $m \leq N$, let $\mathcal{K}$ be the $\binom{N}{m}$ combinations of (j,k) from $\triangledown \equiv \{(j,k) : j < k, j, k = 1, ..., n\}$ and write by $\{(k_{1,i},\ k_{2,i}) : i = 1, ..., m\} \in \mathcal{K}$ a generic combination, $\mathbf{k}_j = (k_{j,i} : i = 1, ..., m)$ for $j = 1, 2$, and write $\mathbf{k}$ for either $\mathbf{k}_1$ or $\mathbf{k}_2$. Then (5) can be written in matrix form

$$Y_{\mathbf{k}_1,\mathbf{k}_2} = X_{\mathbf{k}_1,\mathbf{k}_2}\beta + \epsilon_{\mathbf{k}_1,\mathbf{k}_2}, \tag{6}$$

where $Y_{\mathbf{k}_1,\mathbf{k}_2} = Y_{\mathbf{k}_1} - Y_{\mathbf{k}_2}$, $X_{\mathbf{k}_1,\mathbf{k}_2} = X_{\mathbf{k}_1} - X_{\mathbf{k}_2}$ and $\epsilon_{\mathbf{k}_1,\mathbf{k}_2} = \epsilon_{\mathbf{k}_1} - \epsilon_{\mathbf{k}_2}$ with $\boldsymbol{\epsilon}_{\mathbf{k}} = (\boldsymbol{\epsilon}_k : k \in \mathbf{k})^\top$. Let $\hat{\beta}_{\mathbf{k}_1,\mathbf{k}_2}$ be the least squares estimator based on the subset of the observations, i.e.,

$$\hat{\beta}_{\mathbf{k}_1,\mathbf{k}_2} = ({X_{\mathbf{k}_1,\mathbf{k}_2}}^\top X_{\mathbf{k}_1,\mathbf{k}_2})^{-1} {X_{\mathbf{k}_1,\mathbf{k}_2}}^\top Y_{\mathbf{k}_1,\mathbf{k}_2}, \tag{7}$$

Accordingly, Serfling and Zhou (2006) extended the TSE to the MTSE as the spatial median,

$$\hat{\beta}_n = \text{Mmed}\left\{\hat{\beta}_{\mathbf{k}_1,\mathbf{k}_2} : (\mathbf{k}_1, \mathbf{k}_2) \in \mathcal{K}_0\right\}. \tag{8}$$

where $\mathcal{K}_0$ is the subset of $\mathcal{K}$ in which all the least squares exist.

In a simple linear regression model, Peng, Wang and Wang (2006) studied the Theil-Sen estimator under no assumption on the distribution of the error (neither symmetry nor continuity on the error distribution is assumed). They showed that the TSE is strongly consistent, has an asymptotic distribution under mild conditions, and is super-efficient if the error distribution is discontinuous. Naturally we might ask whether these results can be extended to the MTSE's and under what conditions. Specifically, we have two questions herein. First, can we remove the assumption of *symmetry* of the error distribution? Second, can we have super-efficiency when the error ϵ is discontinuous? The answers to the two questions are yes as shall be demonstrated below.

3 Existence and Uniqueness

In this section, we first give a theorem which characterizes the symmetry of a vector. We then propose a third construction of the MTSE. The section is ended with the introduction of the spatial depth and its existence and uniqueness.

In order to ensure that $\hat{\beta}_n$ converges to the true parameter β as n tends to infinity, a sufficient condition, as pointed out by Zhou and Serfling (2006) in their spatial-depth-based MTSE, is that $\hat{\beta}_{\mathbf{k}_1,\mathbf{k}_2}$ is *centrally symmetric* about the true unknown parameter β, i.e.,

$$\hat{\beta}_{\mathbf{k}_1,\mathbf{k}_2} - \beta \overset{cd}{=} \beta - \hat{\beta}_{\mathbf{k}_1,\mathbf{k}_2}, \tag{9}$$

where $\overset{cd}{=}$ denotes both sides have an identical distribution. A more general symmetry is *angular symmetry*, see Liu (1992). For more details about various notions of symmetry, see Serfling (2006). They demonstrated that the central symmetry of $\hat{\beta}_{\mathbf{k}_1,\mathbf{k}_2}$ about β follows from the central symmetry of $\epsilon_{\mathbf{k}_1,\mathbf{k}_2}$ about zero,

$$\epsilon_{\mathbf{k}_1,\mathbf{k}_2} \overset{cd}{=} -\epsilon_{\mathbf{k}_1,\mathbf{k}_2}. \tag{10}$$

Surprisingly we found that this is equivalent to Assumption S. The argument is as follows.

Using the method of characteristic function, it is easy to show that Assumption S implies (10). Let $\psi(t) = \mathbb{E}\exp(\mathbf{i}\epsilon)$ be the characteristic function of the error ϵ, where $\mathbf{i}^2 = -1$ is the unit imaginary number. We now calculate the characteristic function $\varphi(\mathbf{t}) = \mathbb{E}\exp(i\mathbf{t}^\top\epsilon_{\mathbf{k}_1,\mathbf{k}_2})$ of $\epsilon_{\mathbf{k}_1,\mathbf{k}_2}$ for $\mathbf{t} = (t_1, ..., t_m)^\top \in \mathbb{R}^m$. To this end we identify ϵ_j from $t_l(\epsilon_{k_{1,l}} - \epsilon_{k_{2,l}})$ for $l = 1, ..., m$ and $j = 1, ..., n$ and let $d_{j,l}$ be the identifier and $\mathbf{d}_j = (d_{j,1}, ..., d_{j,m})^\top$. Then using the independence of $\epsilon_1, ..., \epsilon_n$ one finds

$$\varphi(\mathbf{t}) = \psi(\mathbf{t}^\top\mathbf{d}_1)\cdots\psi(\mathbf{t}^\top\mathbf{d}_n), \tag{11}$$

where the identifier is given by

$$d_{j,l} = \begin{cases} 0, & k_{1,l} \neq j, k_{2,l} \neq j, \\ 1, & k_{1,l} = j, \\ -1, & k_{2,l} = j. \end{cases} \tag{12}$$

Under Assumption S, ϵ is symmetric about zero so that $\psi(t) = \psi(-t)$. Thus the characteristic function of $-\epsilon_{\mathbf{k}_1,\mathbf{k}_2}$ is $\mathbb{E}\exp(-i\mathbf{t}^\top\epsilon_{\mathbf{k}_1,\mathbf{k}_2}) = \varphi(-\mathbf{t}) = \varphi(\mathbf{t})$ by (11). This establishes the symmetry (10) of $\epsilon_{\mathbf{k}_1,\mathbf{k}_2}$.

To show that (10) implies Assumption S, we present the following theorem, which gives a little stronger result stating that it only requires the central symmetry (10) to hold for $m = 3$.

Theorem 1. *Suppose that $\mathscr{E}_1, \mathscr{E}_2, \mathscr{E}_3$ are independent and identically distributed. Then $\mathscr{E}_1$ is symmetric about its median if and only if $\mathscr{E}_1, \mathscr{E}_2, \mathscr{E}_3$ satisfy (10) for*

$$(\mathbf{k}_1, \mathbf{k}_2) = (\{1,1\}, \{2,3\}), (\{1,2\}, \{3,3\}), (\{1,2\}, \{2,3\}). \tag{13}$$

Proof: We only need to show the sufficiency. Let ϕ be the characteristic function of $\mathscr{E}_1$. Since (10) holds for the values of $(\mathbf{k}_1, \mathbf{k}_2)$ in (13), it follows

$$\phi(t+s)\phi(-t)\phi(-s) = \phi(-t-s)\phi(t)\phi(s), \quad s, t \in \mathbb{R}. \tag{14}$$

Let $\Phi(t) = \phi(t)/\phi(-t)$. Then Φ is continuous and, by (14), satisfies the Cauchy functional equation $\Phi(t+s) = \Phi(t)\Phi(s)$. It is well known that the solution of a Cauchy functional equation is exponential, i.e., $\Phi(t) = e^{ct}$ for some complex number c among continuous functions. In addition, it is easy to verify by the definition that the conjugate $\bar{\Phi}(t)$ satisfies $\bar{\Phi}(t)\Phi(t) = 1$, yielding $\bar{c} + c = 0$, so that c is an imaginary number, i.e., $c = \mathbf{i}a$ for some real a. Hence $\phi(t) = e^{\mathbf{i}at}\phi(-t)$. This is equivalent to $\epsilon - a \overset{cd}{=} a - \epsilon$. The proof is complete. □

From the above Theorem 1, we see that Assumption S is necessary and sufficient for the central symmetry of the joint (10) and hence (9), while the latter ensures that the spatial median converges to the true symmetric center, the true parameter value β, as the sample size n tends to infinity. In addition, by Theorem 1, in estimating the "slope" vector β, a slightly more general assumption of symmetry is that the error ϵ is **essentially symmetric** in the sense that it has a distribution symmetric about its median. Such an example is the uniform distribution.

Estimating the normal vector using non-overlapping differences. Because $\epsilon_i - \epsilon_j$ and $\epsilon_j - \epsilon_i$ have an identical distribution as long as ϵ_i, ϵ_j are independent and have a common distribution no matter whether or not this distribution is symmetric. Without the assumption of central symmetry on the error ϵ, (10) is no longer true. What happens is that its components are correlated, for instance, $\epsilon_2 - \epsilon_1$ and $\epsilon_3 - \epsilon_2$ are correlated. Therefore one simple remedy to this problem is to choose its components, the pairwise differences, in a way that they are not overlapped, for instance, we may choose $\epsilon_{\mathbf{k}_1,\mathbf{k}_2} = (\epsilon_1 - \epsilon_2,\ \epsilon_3 - \epsilon_4,\ ...,\ \epsilon_{2p-1} - \epsilon_{2p})^\top$. In general we choose the pairwise difference $\epsilon_{\mathbf{k}_1,\mathbf{k}_2}$ in such a way that $\mathbf{k}_1, \mathbf{k}_2$ have no element in common. Then following the procedure of Zhou and Serfling (2006), we construct the multiple Theil-Sen estimator, β_n^* say, of β for a general depth function. For the spatial-depth-based median, the asymptotic normality of the MTSE of Zhou and Serfling, under no assumption of symmetry on the error distribution, follows from their theory of spatial quantiles. In this article we give the strong

consistency and asymptotic normality in Theorem 5 under a set of weaker assumptions as an application of the asymptotic results that we shall present below in this article.

Existence and Uniqueness for the Spatial Median. As an illustration and for later applications, let us recall the spatial depth in the literature. Let Z be a random vector on $\mathbb{R}^d$ with probability distribution Q. The spatial median m of Z is the minimizer of $z \mapsto \int (\|t-z\| - \|t\|)\, dQ(t) = \mathbb{E}_Q(\|Z-z\| - \|Z\|)$ where $\|\cdot\|$ is the Euclidean norm. The existence follows from the tightness of Q. For $z \in \mathbb{R}^d$, let $S(z) = z/\|z\|(S(0)=0)$ be the spatial sign function (or spatial unit function by Chaudhuri). The statistical spatial depth is then defined as

$$D_{\rm sp}(z,Q) = 1 - \|\mathbb{E}_Q S(z-Z)\|, \quad z \in \mathbb{R}^d. \tag{15}$$

For a random sample $Z_1, ..., Z_n$ of Q, the sample version spatial depth is

$$D_{\rm sp}(z,Q_n) = 1 - \left\| \tfrac{1}{n} \textstyle\sum_{i=1}^n S(z-Z_i) \right\|, \quad z \in \mathbb{R}^d, \tag{16}$$

where Q_n is the empirical distribution. Then the *spatial median* m is the multivariate median defined by the spatial depth, which is any maximizer of the spatial depth, i.e.,

$$m = \arg\sup_{x \in \mathbb{R}^d} D_{\rm sp}(x,Q). \tag{17}$$

Note that the above two definitions of the spatial median coincide. The spatial median m can be estimated by the sample spatial median m_n, which maximizes the sample depth, i.e.,

$$m_n = \arg\sup_{x \in \mathbb{R}^d} D_{\rm sp}(x,Q_n). \tag{18}$$

The strong consistency and asymptotic normality of the spatial median are well established in the literature, see Bose (1998), Chaudhuri (1996), Niemiro (1992) among others. Other depth-based multivariate medians are defined analogously, i.e., they are the maximizers of the depths. If a distribution is symmetric in some sense then the depth-based multivariate median is the center of symmetry. There are various notions of symmetry, for example, central symmetry, angular symmetry, halfspace symmetry, etc. For a systematic discussion, see Serfling (2006). In the following we summarize some useful facts about the uniqueness of the spatial medians.

Remark 1. *Z has a unique spatial median if one of the following holds.*
(1) Q is not concentrated on a line (Milasevic and Ducharme (1987)). Hence,
(2) There are two one-dimensional marginal distributions each of which is not point mass for $d \geq 2$. Further,
(3) There are at least two absolute continuous one-dimensional marginal distributions.
(4) Q is angularly symmetric about its median and $\phi'(\mathbf{m}) = \int S(\mathbf{z}-\mathbf{m})\, P(d\mathbf{z})$. Hence,

(5) Q is centrally symmetric about its median.
(6) Q is angularly symmetric about the its median and Q is absolutely continuous.

Both (2) and (3) are clear and for (5) see Milasevic and Ducharme (1987) and we give an argument for (4) from which (6) follows. For $\mathbf{z} \in \mathbb{R}^d$, let $T : \mathbb{R}^d \to [0, \infty) \times S^{d-1}$ be the transformation given by the polar coordinate $\mathbf{u} = \mathbf{z}/\|\mathbf{z}\|, r = \|z\|$ where $S^{d-1} = \left\{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\| = 1\right\}$ is the unit sphere. Let $\nu(\mathbf{u}) = \int_0^\infty P \circ T^{-1}(\mathbf{u}, dr)$ be assumed for $\mathbf{u}$ on S^{d-1}. Then Z is angularly symmetric about zero provided that $\nu(-\mathbf{u}) = \nu(\mathbf{u})$ for every $\mathbf{u}$ on S^{d-1}, so that $\phi'(0) = \int_{S^{d-1}} (\mathbf{u} \int_0^\infty P \circ T^{-1}(\mathbf{u}, dr))\, d\mathbf{u} = \int_{S^{d-1}} \mathbf{u}\nu(\mathbf{u})\, d\mathbf{u} = 0$. Therefore $\min_{\mathbf{m}} \phi(\mathbf{m}) = \phi(0) = 0$ and this is the desired result. □

4 Asymptotic consistency

In this section, we first give two theorems which are useful for proving strong consistency for U-statistics. As an application, the consistency of the spatial depth-based MTSE and pairwise-difference based MTSE are given, followed by the super-efficiency.

Let $(\mathscr{X}, \mathscr{O})$ be a probability space on which F is a probability measure. Let $\{X_i\}_{i=1}^\infty$ be a sequence of independent r.v.'s with common distribution F. Let Θ be an open subset of $\mathbb{R}^d$ and $\vartheta_0 \in \Theta$ is fixed. For a positive integer r, denote the r-tuple product space by $\mathscr{X}^r = \mathscr{X} \otimes \cdots \otimes \mathscr{X}$ and the r-tuple convolution by $F^r = F \otimes \cdots \otimes F$. Let ψ be a kernel which is a symmetric map (invariant under argument permutation) from $\mathscr{X}^r \times \Theta$ into $\mathbb{R}$ satisfying the following conditions C.1–C.5.
(C.1) the map $\mathbf{x} \mapsto \psi(\mathbf{x}, \vartheta)$ is measurable for every $\vartheta \in \Theta$.
(C.2) the map $\vartheta \mapsto \psi(\mathbf{x}, \vartheta)$ is continuous for every $\mathbf{x} \in \mathscr{X}^r$.
For $\vartheta \in \Theta$ set

$$U_n(\vartheta) = \binom{n}{r}^{-1} \sum_{i_1 < \cdots < i_r} \psi(X_{i_1}, \ldots, X_{i_r}, \vartheta).$$

A sequence $\langle \hat{\vartheta}_n \rangle$ is called a U-estimate if $U_n(\hat{\vartheta}_n) = \sup_{\vartheta \in \Theta} U_n(\vartheta)$. It is called a generalized U(V)-estimate if $U_n(\hat{\vartheta}_n) \geq \sup_{\vartheta \in \Theta} U_n(\vartheta) - O_P(n^{-1})$.
(C.3) For every $\vartheta \in \Theta$, there is an F^r-integrable function H_ϑ and positive ϵ_ϑ such that $\psi(\mathbf{x}, t) \leq H_\vartheta(\mathbf{x})$ for all $\mathbf{x} \in \mathscr{X}^r$ and $t \in \Theta$ with $\|t - \vartheta\| \leq \epsilon_\vartheta$.
(C.4) The map $\mu(\vartheta)$ from Θ into $[-\infty, \infty)$ defined by

$$\mu(\vartheta) = \int \psi(\mathbf{x}, \vartheta)\ F^r(d\mathbf{x}), \quad \vartheta \in \Theta$$

is uniquely maximized at ϑ_0.

(C.5) There exists a compact neighborhood $K \subset \Theta$ of ϑ_0 such that

$$\limsup_{n\to\infty} \sup_{\vartheta\in\Theta\setminus K} U_n(\vartheta) < \mu(\vartheta_0) \quad a.s.$$

We have the following theorem and the proof is given later.
Theorem 2. *Suppose that (C.1)-(C.5) hold. If $\langle\hat{\vartheta}_n\rangle$ is a generalized U-estimate, then $\hat{\vartheta}_n \to \vartheta_0$ a.s.*

The above (C.5) can be replaced with the convexity of $\vartheta \mapsto \psi(x, \vartheta)$. This is especially useful because some of the depth functions are concave down, for instance, the spatial depth. This is stated in the following theorem and the proof is relegated in the last section.
Theorem 3. *Suppose that (C.1)-(C.4) hold. If the map $\vartheta \mapsto \psi(\mathbf{x}, \vartheta)$ is concave down for every $\mathbf{x} \in \mathscr{X}^r$ then $\hat{\vartheta}_n \to \vartheta_0$ a.s.*

Now we apply the above theorems to the spatial depth-based MTSE's, see definition and discussion of about the spatial median in Section 1. Denote $\xi_k = (X_k, Y_k), k = 1, ..., n$, $\xi_{\mathbf{k}} = \{\xi_k : k \in \mathbf{k}\}$ and $\mathbf{k}_0 = (1, ..., m)$ and write $\boldsymbol{\xi}_{\mathbf{k}_0} = \boldsymbol{\xi}_{\mathbf{0}}$. For the spatial depth, we apply Theorem 3 with $\psi(\boldsymbol{\xi}_{\mathbf{0}}; \vartheta) = \|h(\boldsymbol{\xi}_{\mathbf{0}})\| - \|\vartheta - h(\boldsymbol{\xi}_{\mathbf{0}})\|$ where $h(\boldsymbol{\xi}_{\mathbf{0}}) = (X_{\mathbf{k}_0}^\top X_{\mathbf{k}_0})^{-1} X_{\mathbf{k}_0}^\top Y_{\mathbf{k}_0}$ clearly satisfies (C.1), (C.2), and (C.3) with integrable $H(\boldsymbol{\xi}_{\mathbf{0}}) = \|\vartheta_0\| + 1$ for $\|\vartheta - \vartheta_0\| \le 1$ by the triangle inequality. By the triangle inequality of the Euclidean norm, the map $\vartheta \mapsto \psi(\boldsymbol{\xi}_{\mathbf{0}}, \vartheta)$ is concave down. Thus by Theorem 3 we have the strong consistency for the spatial-depth based MTSE $\hat{\theta}_n \equiv \hat{\theta}_{n,\mathrm{sp}}$.
Theorem 4. *(Consistency for spatial-depth based MTSE under no symmetry.) Suppose that the distribution of $h(\boldsymbol{\xi}_{\mathbf{0}})$ is not concentrated on a line and the map $\vartheta \mapsto \mathbb{E}\|\vartheta - h(\boldsymbol{\xi}_{\mathbf{0}})\|$ is maximized at the true θ. Then the spatial-depth based MTSE $\hat{\theta}_{n,\mathrm{sp}}$ is strongly consistent, i.e. $\hat{\theta}_{n,\mathrm{sp}} \to \theta$ a.s.*

Important special cases of the above theorem are given below in view of Remark 1.
Corollary 1. *Theorem 4 holds if one of the following is true.*
(1) Assumption S is met. The derivative can pass the integral $\Delta\mu(\vartheta) = \mathbb{E}S(\vartheta - h(\boldsymbol{\xi}_{\mathbf{0}}))$ for ϑ in a neighborhood of the true θ.
(2) Assumption S is met and the distributions of ϵ and X are absolutely continuous.
(3) There are at least two one-dimensional marginal distributions of $h(\boldsymbol{\xi}_{\mathbf{0}})$ each of which is not point mass for $p \ge 1$ and the true parameter θ satisfies $\mathbb{E}S(\theta - h(\boldsymbol{\xi}_{\mathbf{0}})) = 0$. Hence,
(4) The distributions of ϵ and X are absolutely continuous and the true parameter θ satisfies $\mathbb{E}S(\theta - h(\boldsymbol{\xi}_{\mathbf{0}})) = 0$.

By Theorem 1 and with a similar argument we have the following.
Corollary 2. *(Consistency for spatial-depth based MTSE under symmetry, based on overlapped differences.) Suppose Assumption S is met. Then the*

MTSE $\hat{\beta}_{n,sp}$ based on the spatial depth and the pairwise (overlapped) differences is strongly consistent, i.e. $\hat{\beta}_{n,sp} \to \beta$ a.s.

Peng, Wang and Wang (2006) gave the consistency of the univariate Theil-Sen estimator under no assumption on the distribution of the error. With non-overlapping pairwise differences we have a similar result, i.e., Assumption S is not required for the consistency of $\hat{\beta}^*_{n,sp}$ of the "slope" vector β.

Theorem 5. *(Consistency for spatial-depth under no symmetry.) Suppose the distribution of the error ϵ is not concentrated on a point mass. Then the MTSE $\hat{\beta}^*_{n,sp}$ based on the spatial depth and the non-overlapping pairwise differences is strongly consistent, i.e., $\hat{\beta}^*_{n,sp} \to \beta$ a.s.*

Remark 2. *Using Theorem 2 or Theorem 3, one can establish the consistency of the MTSE's whose defining medians are associated with continuous depth functions. Examples of these include L_p-depth, smoothed Tukey depth, simplicial value depth, etc.*

Super-efficiency. Here we consider the super-efficiency of the spatial-depth based MTSE $\hat{\beta}_{n,sp}$. Let $h_b(\boldsymbol{\xi_0}) = I_p h(\boldsymbol{\xi_0})$ with $I_p = \text{diag}(0, 1, \ldots, 1)$ a diagonal matrix and $\psi_b\ (\boldsymbol{\xi_0}; \vartheta) = \|h_b(\boldsymbol{\xi_0})\| - \|\vartheta - h_b(\boldsymbol{\xi_0})\|$ and the resulting U-statistic $U_{b,n}(\vartheta)$. We have the following theorem and the proof is given in the last section.

Theorem 6. *(Super-efficiency of spatial-depth-based MTSE) Suppose Assumption S holds. Assume $h_b(\boldsymbol{\xi_0})$ is not concentrated on a line. Then if the error distribution is discontinuous,*

$$P(\hat{\beta}_{n,sp} = \beta) \to 1.$$

It follows from the above theorem we have for any $\nu \geq 0$

$$n^{\nu}(\hat{\beta}_{n,sp} - \beta) \to 0.$$

Thus $\hat{\beta}_{n,sp}$ is super-efficient. This result is true for the TSE in a simple linear regression model, see Peng, Wang and Wang (2006). Our simulation validates this fact and exhibits that different samples are required to reach the equality.

5 Asymptotic normality

In this section, we first give a theorem which is useful for proving asymptotic normality of U-statistics. As an application, the asymptotic normality of the MTSE and paired MTSE are obtained under weaker assumptions of Zhou and Serfling (2006). Rates of the remainder are also obtained.

Definition 1. *ψ is* **regular** *at ϑ_0 if there exists a neighborhood Θ_0 of ϑ_0 such that*

(A.1) For every $\mathbf{x} \in \mathscr{X}^r$, the map $\vartheta \mapsto \psi(\mathbf{x}, \vartheta)$ is twice continuously differentiable on Θ_0 with gradient $\nabla\psi(\mathbf{x}, \vartheta)$ and second derivative $\nabla^2\psi(\mathbf{x}, \vartheta)$.
(A.2) There is an F^r-integrable function H such that $\sup_{\vartheta\in\Theta_0} \|\nabla^2\psi(\mathbf{x}, \vartheta)\| \le H(\mathbf{x})$ for all $\mathbf{x} \in \mathscr{X}^r$.

For ψ regular at θ and $\vartheta \in \Theta_0$, let

$$\nabla U_n(\vartheta) = \binom{n}{r}^{-1} \sum_{i_1<\cdots<i_r} \nabla\psi(X_{i_1}, \ldots, X_{i_r}, \vartheta),$$

$$\nabla\tilde{\psi}(x, \vartheta) = \int \nabla\psi(x_1, \ldots, x_{r-1}, x, \vartheta)\ F(dx_1)\ldots F(dx_{r-1}),$$

$$\nabla^2 U_n(\vartheta) = \binom{n}{r}^{-1} \sum_{i_1<\cdots<i_r} \nabla^2\psi(X_{i_1}, \ldots, X_{i_r}, \vartheta), \quad M_\vartheta = \int \nabla^2\psi(\mathbf{x}, \vartheta)\ F^r(d\mathbf{x}).$$

Theorem 7. *Suppose that ψ is regular at θ, M_θ is invertible,*

$$\int \nabla\psi(\mathbf{x}, \vartheta)\ F^r(d\mathbf{x}) = 0 \quad \textit{and} \int \|\nabla\psi(\mathbf{x}, \vartheta)\|^2\ F^r(d\mathbf{x}) < \infty.$$

Let $\langle\hat{\theta}_n\rangle$ be a sequence of Θ-valued random vectors such that $\hat{\vartheta}_n = \vartheta_0 + o_p(1)$ and $\sqrt{n}\nabla U_n(\hat{\vartheta}_n) = o_p(1)$. Then

$$\sqrt{n}(\hat{\vartheta}_n - \vartheta_0) = -\sqrt{n}M_{\vartheta_0}^{-1}\nabla U_n(\vartheta_0) + o_p(1).$$

In particular,

$$\sqrt{n}(\hat{\vartheta}_n - \vartheta_0) \Rightarrow N(0,\ r^2 M_{\vartheta_0}^{-1} V_{\vartheta_0} M_{\vartheta_0}^{-\top}),$$

where $V_{\vartheta_0} = \int \nabla\tilde{\psi}(x, \vartheta_0)\nabla\tilde{\psi}^\top(x, \vartheta_0)\ F(dx)$.

If the above condition (A.2) is not met or difficult to verify (for example, for the spatial depth in a two-dimensional parameter space), the following theorem gives another set of conditions.

Theorem 8. *Suppose the map $\vartheta \mapsto \psi(\mathbf{x}, \vartheta)$ is differentiable at ϑ_0 for almost every $\mathbf{x} \in \mathscr{X}^r$ with gradient $\nabla\psi(\mathbf{x}, \vartheta_0)$ and there exists a neighborhood Θ_0 of ϑ_0 and a measurable function L with $\int \|L(\mathbf{x})\|^2\ F^r(d\mathbf{x}) < \infty$, such that for every ϑ_1, ϑ_1 in Θ_0 and every $\mathbf{x} \in \mathscr{X}^r$,*

$$|\psi(\mathbf{x}, \vartheta_1) - \psi(\mathbf{x}, \vartheta_2)| \le L(\mathbf{x})\,\|\vartheta_1 - \vartheta_2\|\,. \tag{19}$$

If there exists a positive definite symmetric matrix M_{ϑ_0} such that

$$\mu(\vartheta) = \mu(\vartheta_0) + \frac{1}{2}(\vartheta - \vartheta_0)^{\mathsf{T}} M_{\vartheta_0}(\vartheta - \vartheta_0) + o(\|\vartheta - \vartheta_0\|^2). \tag{20}$$

Then for any sequence of Θ-valued random vectors $\langle\hat{\vartheta}_n\rangle$ satisfying $\hat{\vartheta}_n = \vartheta_0 + o_p(1)$, and $U_n(\hat{\vartheta}_n) \ge \sup_{\vartheta\in\Theta} U_n(\vartheta) - o_p(n^{-1})$, one has

$$\sqrt{n}(\hat{\vartheta}_n - \vartheta_0) = -\sqrt{n}M_{\vartheta_0}^{-1}\nabla U_n(\vartheta_0) + o_p(1).$$

In particular,

$$\sqrt{n}(\hat{\vartheta}_n - \vartheta_0) \Rightarrow N(0,\ r^2 M_{\vartheta_0}^{-1} V_{\vartheta_0} M_{\vartheta_0}^{-\top}).$$

Asymptotic normality of the spatial-depth-based MTSE. Zhou and Serfling (2006) gave the Bahadur-Kiefer representation for multivariate spatial U-quantiles. They obtained the faster rate of the remainder than the existing results. The asymptotic normality of the MTSE as a special U quantile can be derived from the representation. Bose(1998) gave the Bahadur presentation of median estimates where the rate of the remainder was also obtained. Here we first give a representation under weak assumptions based on the above theorems. Then with an application to Bose's we obtain the rate of the remainder of the representation. Recall $h(\boldsymbol{\xi_0}) = (X_{\mathbf{k}_0}^\top X_{\mathbf{k}_0})^{-1} X_{\mathbf{k}_0}^\top Y_{\mathbf{k}_0}$, and $\mu(\vartheta) = \mathbb{E}(\|\vartheta - h(\boldsymbol{\xi_0})\| - \|h(\boldsymbol{\xi_0})\|)$. Denote $\mathbf{A}_{\mathbf{k}_0} = (X_{\mathbf{k}_0}^\top X_{\mathbf{k}_0})^{-1} X_{\mathbf{k}_0}^\top$ and

$$D_1(\vartheta) \equiv \mathbb{E}\left\{ \frac{1}{\|\vartheta - h(\boldsymbol{\xi_0})\|} \left(I_m - \frac{(\vartheta - h(\boldsymbol{\xi_0}))^{\otimes 2}}{\|\vartheta - h(\boldsymbol{\xi_0})\|^2} \right) \right\}.$$

Theorem 9. *Suppose that (i) the distributions of ϵ and $\mathbf{A}_{\mathbf{k}_0}$ are absolutely continuous w.r.t. the Lebesgue measure; (ii) $\Delta\mu(\vartheta)$ is continuously differentiable with derivative $\Delta^2\mu(\vartheta) = D_1(\vartheta)$ in a neighborhood $\mathbf{N}$ of θ; (iii) the map $\vartheta \mapsto \mathbb{E}\|\vartheta - h(\boldsymbol{\xi_0})\|$ is maximized at true θ. Then the MTSE $\hat{\theta}_{n,\mathrm{sp}}$ satisfies the stochastic approximation*

$$\hat{\theta}_{n,\mathrm{sp}} = \theta + D^{-1}\bar{S}_n + R_n, \tag{21}$$

where $D = D_1(\theta)$, $\bar{S}_n = \sum_{\mathbf{k}} S(\theta - h(\boldsymbol{\xi}_{\mathbf{k}}))/\binom{n}{m}$, and $R_n = o_p(n^{-1/2})$, assuming that D is invertible. Hence $\hat{\theta}_{n,\mathrm{sp}}$ is asymptotic normal with mean zero and covariance Σ, i.e.,

$$\sqrt{n}(\hat{\theta}_{n,\mathrm{sp}} - \theta) \stackrel{D}{\Longrightarrow} \mathcal{N}(0, \Sigma), \tag{22}$$

where $\Sigma = m^2 D_1^{-1}(\theta)\mathbb{E}[\tilde{h}(\xi_1)^{\otimes 2}]D_1^{-1}(\theta)$ with $\tilde{h}(\xi_1) = \mathbb{E}(S(\theta - h(\xi_1, ..., \xi_m))|\xi_1)$.

Proof: The absolute continuity of ϵ and $\mathbf{A}_{\mathbf{k}_0}$ implies that the distribution of $h(\boldsymbol{\xi_0})$ is also absolutely continuous. The support of the density function of ϵ is not congregated at only one point, hence the distribution of h is not concentrated on a line, see Remark 1, so that the spatial median uniquely exists. The absolute continuity of h also implies that $\vartheta \mapsto \psi(\mathbf{x}, \vartheta) = \|\mathbf{x}\| - \|\vartheta - \mathbf{x}\|$ is differentiable for almost every $\mathbf{x} \in \mathbb{R}^m$ with gradient $\Delta\psi(\mathbf{x}, \vartheta) = S(\vartheta - \mathbf{x})$ and $D(\vartheta) = \Delta\mu(\vartheta) = \mathbb{E}S(\vartheta - \mathbf{X})$ for $\vartheta \in \mathbf{N}$ by the dominated convergence theorem ($S(\vartheta - \mathbf{x})$ is dominated by 1) and $\Delta\mu(\theta) = \mathbb{E}S(\theta - \mathbf{X}) = 0$ by the uniqueness. Clearly (19) is satisfied because

$$|\psi(\mathbf{x}; \vartheta_1) - \psi(\mathbf{x}; \vartheta_2)| \le \|\vartheta_1 - \vartheta_2\|, \quad \mathbf{x} \in \mathbb{R}^m, \vartheta_1, \vartheta_2 \in \mathbf{N}.$$

The differentiability of $D(\vartheta)$ with continuous gradient $D_1(\vartheta)$ in $\mathbf{N}$ with Taylor expansion of $\mu(\vartheta)$ at θ yield (20). Thus an application of Theorem 8 completes the proof. □

Remark 3. *Theorem 9 holds without assuming the boundedness of the densities* $\mathbf{A}_{\mathbf{k}_0}$ *and* ϵ*, while the boundedness is assumed in Chaudhuri and Zhou and Serfling.*
Remark 4. *The absolute continuity of* $\mathbf{A}_{\mathbf{k}_0}$ *and* ϵ *in Theorem 9 is necessary for the asymptotic normality, noticing that Peng, Wang and Wang (2005) demonstrate that the asymptotic distribution is not normal when the absolute continuity is not assumed for the Theil-Sen estimator in a simple linear regression. We believe the latter shall also hold for MTSE.*
Remark 5. *Theorem 9 holds if one of the following is true instead of (iii). (1) Assumption S is met. (2)* $\mathbb{E}S(\theta - h(\boldsymbol{\xi}_{\mathbf{0}})) = 0$.

Denote $\boldsymbol{\xi}_{\boldsymbol{o},\boldsymbol{e}} = \boldsymbol{\xi}_{(1,3,...,(2m-1))} - \boldsymbol{\xi}_{(2,4,...,2m)}$. With the non-overlapping pairwise differences symmetry is automatic so that the MTSE $\hat{\beta}^*_{n,sp}$ uniquely exists. Let $\mu^*(b) = \mathbb{E}(\|b - h(\boldsymbol{\xi}_{\boldsymbol{o},\boldsymbol{e}})\| - \|h(\boldsymbol{\xi}_{\boldsymbol{o},\boldsymbol{e}})\|)$ and

$$D^*_1(b) = \mathbb{E}\left\{\frac{1}{\|b - h(\boldsymbol{\xi}_{\boldsymbol{o},\boldsymbol{e}})\|}\left(I_m - \frac{(b - h(\boldsymbol{\xi}_{\boldsymbol{o},\boldsymbol{e}}))^{\otimes 2}}{\|b - h(\boldsymbol{\xi}_{\boldsymbol{o},\boldsymbol{e}})\|^2}\right)\right\}.$$

Theorem 10. *Suppose the conditions of Theorem 9 are fulfilled with* $\Delta\mu^*(b)$ *and its derivative* $\Delta^2\mu^*(b) = D^*_1(b)$ *for* b *in a neighborhood* $\mathbf{N}$ *of* β*. Then the MTSE* $\hat{\beta}^*_{n,sp}$ *satisfies (21) with* $D = D^*_1(\beta)$, $\bar{S}_n = \sum_{\mathbf{k}_1,\mathbf{k}_2} S(\beta - h(\boldsymbol{\xi}_{\mathbf{k}_1,\mathbf{k}_2}))/\binom{N}{m}$*, assuming that* $D^*_1(\beta)$ *is invertible. Hence* $\hat{\beta}^*_{n,sp}$ *is asymptotic normal with mean zero and covariance matrix* $\Sigma^* = m^2(D^*_1)^{-1}(\beta)\mathbb{E}\tilde{h}^*(\xi_1 - \xi_2)^{\otimes 2}(D^*_1)^{-1}(\beta)$ *with* $\tilde{h}^*(\xi_1 - \xi_2) = \mathbb{E}(S(\beta - h(\boldsymbol{\xi}_{\mathbf{k}_1,\mathbf{k}_2}))|\xi_1 - \xi_2)$.
Remark 6. *For the pairwise overlapped differences we may also derive the asymptotic distribution of the estimator* $\hat{\beta}_{n,sp}$*. Nevertheless it may be different from the above because of the following fact. There are at least two types of errors. One is the overlapped, for example,* $(\epsilon_1 - \epsilon_2, \epsilon_1 - \epsilon_3, ..., \epsilon_1 - \epsilon_{m+1})$*, and the other is the non-overlapped, for example,* $(\epsilon_1 - \epsilon_2, \epsilon_3 - \epsilon_4, ..., \epsilon_{2m-1} - \epsilon_{2m})$*. Thus the the kernels have at least two types, so that the above results do not apply here.*

Using Bose's proposition 1 we find the rate of the remainder in the stochastic approximation, which slightly improves the rate given by Zhou and Serfling (2006) under a slightly weaker assumptions.
Theorem 11. *Suppose* ϵ *fulfills Assumption S . Assume* $\mathbb{E}\|h(\boldsymbol{\xi}_{\mathbf{0}}) - \theta\|^{(3+\nu)/2} < \infty$ *for some* $0 \leq \nu \leq 1$*. Then the MTSE* $\hat{\theta}_{n,\mathrm{sp}}$ *satisfies the stochastic approximation (21) with the remainder*

$$R_n = O(n^{-(3+\nu)/4}(\log n)^{1/2}(\log\log)^{(1+\nu)/4}). \tag{23}$$

Theorem 12. *Suppose that* ϵ *fulfills Assumption S . Assume* $\mathbb{E}\|h(\boldsymbol{\xi}_{\mathbf{k}_1,\mathbf{k}_2}) - \beta\|^{(3+\nu)/2} < \infty$ *for some* $0 \leq \nu \leq 1$*. Then the MTSE* $\hat{\beta}^*_{n,sp}$ *satisfies the stochastic approximation (21) with the remainder* R_n *in (23).*

6 Robustness considerations

In this section, we study the robustness of our estimators in terms of the two prevailing notions: breakdown point (BP) and influence function (IF).

Breakdown Point. The breakdown point measures the ability of an estimator or a statistic to resist contamination of the data. Roughly speaking, the finite sample breakdown point of an estimator is the minimum fraction of 'bad' samples in a data set that can render the estimator useless. For our proposed estimator, the finite sample breakdown point is at least $(1-(1/2)^{1/m})(n-m+1)/n$. Since the breakdown point of the spatial median is $1/2$, we need at least $1/2$ LSE's to be "good". Suppose that there is a fraction ε of 'bad' observations in the data set with size n, then there are $\binom{(1-\varepsilon)n}{m}$ "good" LSE's out of $\binom{n}{m}$. So we need $\binom{(1-\varepsilon)n}{m}/\binom{n}{m} > 1/2$. Since

$$\binom{(1-\varepsilon)n}{m}/\binom{n}{m} > \left(\frac{(1-\varepsilon)n-m+1}{n-m+1}\right)^m,$$

it follows that if $\varepsilon \leq (1/2)^{1/m}(n-m+1)/n$, our proposed estimator will never break down. Accordingly, the asymptotic BP is $1-(1/2)^{1/m}$.

As one can see, the BP depends on the choice of m. On one hand, a smaller m results in a higher BP; on the other hand, a smaller m means lower efficiency. The highest BP is reached when m takes its minimal value $m = p+1$ and this also leads to the lowest efficiency, while the maximal efficiency is attained when m assumes its maximal value n, where the LSE is recovered, and this leads to the lowest BP, assuming the error is Gaussian. Any m taking values in between $p+1$ and n results in an estimator which gives the compromise between robustness and efficiency. Hence one can choose the value of m to gain the desired robustness and efficiency. It should be noted that the highest computational intensity is reached at $m = \lfloor n/2 \rfloor$ because the computational intensity is an order of magnitude $\binom{n}{m}$.

Influence function. While the breakdown point captures the global robustness properties, the local robustness information is provided by the influence function. By (21), the influence function of the MTSE $\hat{\beta}_n$ is

$$IF((y,\mathbf{x});\hat{\beta}_n) = D^{-1}\mathbb{E}\left\{\frac{\beta-(X_{\mathbf{x}}^\top X_{\mathbf{x}})^{-1}X_{\mathbf{x}}^\top Y_y}{\|\beta-(X_{\mathbf{x}}^\top X_{\mathbf{x}})^{-1}X_{\mathbf{x}}^\top Y_y\|}\right\}, \quad \mathbf{x}\in\mathbb{R}^p, y\in\mathbb{R},$$

where D is the previous D_1 or D_1^*, $X_{\mathbf{x}} = [\mathbf{1}_m, X(\mathbf{x})]$ with column $\mathbf{1}_m \in \mathbb{R}^m$ of all entries 1 and $X(\mathbf{x}) = [\mathbf{x}, X_1, ..., X_{m-1}]^\top$ and $Y_y = (y, Y_1, ..., Y_{m-1})^\top$. The above expression shows that the estimator is only influenced by the direction and is irrelevant to the magnitudes of y and $\mathbf{x}$. Consequently our MTSE is

robust against both $\mathbf{x}$ and y outlying. The gross error sensitivity is

$$\gamma^* = \sup_{\mathbf{y},\mathbf{x}} \|IF((y,\mathbf{x}),\hat{\beta}_n)\| = \sup_{\|\mathbf{s}\|\leq 1} \|D^{-1}\mathbf{s}\| \leq \max \lambda^{1/2}((D^{-1})^\top D^{-1}),$$

where $\max \lambda^{1/2}(M)$ denotes the square root of the largest eigenvalue of the matrix M. Since D is invertible, the influence function is bounded.

7 Computation and Simulation Study

In this section, we describe the stochastic sampling of subpopulation to calculate the estimator for a large sample size. A simulation is also conducted.

To investigate the behavior of the proposed MTSE, three simulations are carried out for robustness, efficiency and super-efficiency. Samples are generated from the multiple regression model $Y_i = 1 + 5X_{1i} + 10X_{2i} + \varepsilon_i$, where $X_{1i} \sim \mathcal{N}(0,1)$, $X_{2i} \sim U(0,1)$, and the error ε_i's are from different distributions for different purposes.

Table 1
ROBUSTNESS. (a) The upper part of the table lists the estimators based on a sample of size n without outliers. (b) The lower part lists the estimators based on a sample of size $n_1 + n_2$ with n_2 outliers added to the "good" sample of size n_1.

	True Parameter $\beta = (5, 10)$		
	MTSE	Diff-based MTSE	LSE
$n = 20$	(4.31, 10.43)	(4.38, 10.93)	(4.38, 10.59)
$n = 30$	(4.88, 10.38)	(4.61, 10.39)	(4.91, 10.25)
$n = 40$	(4.97, 9.88)	(4.98, 9.66)	(5.01, 9.87)
$n_1 = 16, n_2 = 4$	(5.01, 9.95)	(5.06, 9.71)	(4.18, 7.76)
$n_1 = 15, n_2 = 5$	(5.30, 9.46)	(5.25, 9.33)	(5.65, 2.27)
$n_1 = 14, n_2 = 6$	(4.37, 9.68)	(4.22, 9.41)	(-2.65, 7.72)
$n_1 = 13, n_2 = 7$	(4.14, 9.17)	(4.88, 9.59)	(-2.37, 3.34)
$n_1 = 12, n_2 = 8$	(3.98, 9.12)	(0.72, 5.65)	(-3.37, 5.18)
$n_1 = 11, n_2 = 9$	(-2.06, -6.67)	(-3.38, 5.85)	(-0.33, -2.12)

Computation and Sampling of Stochastic Subpopulation. The algorithm for computing the spatial-depth-based MTSE is very straightforward and the codes are available upon request. We used these codes to carry out the simulations. For a sample size less than 50, it takes less than half a minute

to compute one MTSE. For a large sample, we suggest to use the sampling of stochastic subpopulation. The details that we used it to conduct our simulations are as follows. Instead of computing the MTSE based on all possible $\binom{n}{m}$ LSE's, we calculate it based on a subpopulation of $\binom{n}{m}$ LSE's. Specifically, we take a random sample of size m from the whole sample and compute the LSE based this random sample and this process is repeated K times, then the MTSE is calculated based on those K LSE's. Here K is a pre-specified number not exceeding $\binom{n}{m}$. In our example below, we take K to be one percent of $\binom{n}{m}$ and the result seems satisfactory, although it warrants further investigation. For example, how large should K be so that the probability, conditional on the sample, of the error that the MTSE is not caught is less than a pre-specified level. For some discussions, see Rousseeuw and Leroy (1987).

Simulation on Robustness. (1) Samples of size $n = 20, 30, 40$ are generated from the multiple linear model with $\epsilon_i \sim \mathcal{N}(0, 0.5)$. The MTSE, the pairwise differencing MTSE, and the LSE are calculated and reported at the upper part of Table 1. (2) Contaminate the data with outliers (X_i, Y_i) from the multiple linear model $Y_i = 1 - 6X_{1i} - 7X_{2i} + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, 0.5)$. Here n_1, n_2 represent the number of "good", "bad" (outliers) observations, respectively. See Table 1. Observe that without contamination, all the MTSE, the difference-based MTSE and LSE work well. However, with the presence of outliers, the LSE's completely break down and are useless, while the difference-based MTSE's work well until the fraction of outliers reaches 35%, and the MTSE's preform well up to 40%.

Simulation on Super-Efficiency. A simulation is run to exhibit the super-efficiency. Specifically, we want to investigate how large a sample size n shall be in order to reach $\hat{\beta}_n = \beta$. For the sample size n, generate errors from discrete distributions: uniform on $\{-1, 1\}$ (each with probability 1/2), binomial with parameters $(4, 0.5)$, hypergeometric with parameters $(6, 3, 2)$. Based on the simulated data, the MTSE $\hat{\beta}_n$ is calculated. Repeat this procedure $N = 200$ times, the ratio of the frequency for which $\hat{\beta}_n = \beta$ is computed. See Table 2. For the sample size 80 and 100, a stochastic procedure is used for the calculation of the spatial median.

Table 2
SUPER-EFFICIENCY. Proportions of $\hat{\beta}_n = \beta$ with repetitions $N = 200$ for different sample sizes n and different error distributions: uniform on $\{-1, 1\}$, binomial $(4, 0.5)$ and hypergeometric $(6, 3, 2)$.

n	10	20	30	50	80	100
Unif on$\{-1, 1\}$	0.405	0.680	0.840	0.930	0.995	1.000
Bin$(4, 0.5)$	0.210	0.335	0.385	0.500	0.630	0.765
HyperGeo$(6, 3, 2)$	0.360	0.630	0.670	0.880	0.960	1.000

Simulation on Relative Efficiency. To investigate the efficiency, a simulation is conducted as follows. For sample sizes $n = 10, 20, 30$, generate a random sample of the error $\epsilon_i, i = 1, ..., n$ from $\mathcal{N}(0, 1)$, heavy tailed distributions t with $df = 3$ and $df = 1$(Cauchy), compute the response Y_i based on the multiple linear model $Y_i = \beta^\top X_i + \epsilon_i$ with $\beta = (1, 5, 10)^\top$, and repeat this process $N = 1000$ times to obtain the MTSE's $\hat{\beta}_{n,k} : k = 1, ..., N$. Now compute the empirical mean squared error EMSE $= (1/N) \sum_{k=1}^{N} \|\hat{\beta}_{n,k} - \beta\|$. The relative efficiency (RE) of $\hat{\beta}_n$ is the ratio of the EMSE of the LSE to the EMSE of $\hat{\beta}_n$. The results are reported in the Table 3. We observe that under the Gaussian model, the finite sample RE of MTSE is about 70-80%, which is acceptable. However, when the error comes from the heavy tailed distributions, the MTSE competes LSE, especially for Cauchy. Note that the EMSE's of LSE under Cauchy are very large (over 2000) and divergent, for the variance of Cauchy does not exist. The MTSE is much stable and achieves a good balance between robustness and efficiency.

Table 3
RELATIVE EFFIENCEY. The empirical mean squared error (EMSE) and the relative efficiency (RE) of MTSE when the errors are from $\mathcal{N}(0, 1)$, t distribution T_3 with df=3, and Cauchy (i.e. t distribution T_1 with df=1). Repetitions $N = 1000$.

		$\mathcal{N}(0,1)$		t distribution T_3		Cauchy	
		MTSE	LSE	MTSE	LSE	MTSE	LSE
n=10	EMSE	3.716	2.643	7.058	7.628	45.97	2613
	RE	0.711	1.000	1.081	1.000	56.84	1.000
n=20	EMSE	1.339	1.075	2.111	2.627	5.667	816.2
	RE	0.803	1.000	1.245	1.000	144.0	1.000
n=30	EMSE	0.739	0.596	1.161	1.569	3.032	2207
	RE	0.806	1.000	1.352	1.000	728.0	1.000

8 Appendix

In this section, we collect some of the proofs. We first give two lemmas.
Lemma 1. *Suppose (C.1)-(C.4) hold. Then*

$$\sup_{\vartheta \in K} \mu(\vartheta) < \mu(\theta)$$

for every compact subset $K \subset \Theta$ that does not contain ϑ_0

Proof. It follows from (C.1)-(C.3) and Fatou's Lemma that

$$\limsup_{\alpha\to\vartheta}\mu(\xi) = \int H_\vartheta(\mathbf{x})\ F^r(d\mathbf{x}) - \liminf_{\alpha\to\vartheta}\int H_\vartheta(\mathbf{x}) - \psi(\mathbf{x},\xi)\ F^r(d\mathbf{x})$$
$$\leq \int H_\vartheta(\mathbf{x}) - \int H_\vartheta(\mathbf{x}) - \psi(\mathbf{x},\vartheta)\ F^r(d\mathbf{x}) = \mu(\vartheta)$$

for each $\vartheta \in \Theta$. Thus μ is upper semi-continuous and achieves a maximum over each compact subset of Θ. The desired result follows this and (C.4). □

Lemma 2. *Suppose (C.1)-(C.3) hold. Then*

$$\limsup_{n\to\infty}\sup_{\vartheta\in K} U_n(\vartheta) \leq \sup_{\vartheta\in K}\mu(\vartheta) \quad a.s.$$

for every compact subset $K \subset \Theta$.

Proof. We have shown in the proof of Lemma 1 that μ is upper semi-continuous under (C.1)-(C.3). Thus μ achieves a maximum on the compact set K and $M_K = \sup_{\vartheta\in K}\mu(\vartheta) < \infty$. Now select $N > M_K$. For each $\eta > 0$ and $\vartheta \in K$, define a map $\psi_{\vartheta,\eta}$ on $\mathscr{X}^r$ by

$$\psi_{\vartheta,\eta}(\mathbf{x}) = \sup_{\alpha\in K:\|\alpha-\vartheta\|\leq\eta}\psi(\mathbf{x},\alpha), \quad \mathbf{x}\in\mathscr{X}^r.$$

These maps are measurable(since the supremum can be taken over a countable set) and $\psi_{\vartheta,\eta}(\mathbf{x}) \leq H_\vartheta(\mathbf{x})$ if $\eta < \epsilon_\vartheta$, where H_ϑ and ϵ_ϑ are as in (C.3). Moreover, $\psi_{\vartheta,\eta}(\mathbf{x}) \downarrow \psi(\mathbf{x},\vartheta)$ as $\eta \downarrow 0$ for each $\mathbf{x} \in \mathscr{X}^r$ and $\vartheta \in K$. Thus, it follows from the monotone convergence Theorem that

$$\int H_\vartheta(\mathbf{x}) - \psi_{\vartheta,\eta}(\mathbf{x})\ F^r(d\mathbf{x}) \uparrow \int H_\vartheta(\mathbf{x}) - \psi(\mathbf{x},\vartheta)\ F^r(d\mathbf{x}),$$

moreover,

$$\int \psi_{\vartheta,\eta}(\mathbf{x})\ F^r(d\mathbf{x}) \downarrow \int \psi(\mathbf{x},\vartheta)\ F^r(d\mathbf{x})$$

for every $\vartheta \in K$. Consequently, for each $\vartheta \in K$, there exists an $\eta_\vartheta > 0$ such that $\int \psi_{\vartheta,\eta}(\mathbf{x})\ F^r(d\mathbf{x}) < N$. Let $S(\vartheta) = \{\alpha \in K : \|\alpha - \vartheta\| \leq \eta_\vartheta\}$, $\vartheta \in K$. Then it forms an open cover of K, so that there is a finite subcover. Namely, there are $\vartheta_1, \ldots, \vartheta_m$ in K such that $K = \bigcup_{i=1}^m S(\vartheta_i)$. From this we can conclude

$$\sup_{\vartheta\in K} U_n(\vartheta) \leq \max_{1\leq i\leq m}\binom{n}{r}^{-1}\sum_{i_1<\cdots<i_r} h_i(X_{i_1},\ldots,X_{i_r},\vartheta),$$

where $h_i = \psi_{\vartheta_i,\eta_{\vartheta_i}},\ i = 1,\ldots,m$. By the SLLN of U-statistic,

$$\limsup_{n\to\infty}\sup_{\vartheta\in K} U_n(\vartheta) \leq \max_{1\leq i\leq m}\int h_i(\mathbf{x})F^r(d\mathbf{x}) < N \quad a.s..$$

This yields the desired result by letting $N \downarrow M_K$. □

Proof of Theorem 2. Let

$$A = \{\limsup_{n\to\infty}\|\hat{\vartheta}_n - \vartheta_0\| > 0\}\bigcap\{\lim_{n\to\infty} U_n(\vartheta_0) = \mu(\vartheta_0)\}.$$

Since $U_n(\vartheta_0) \to \mu(\vartheta_0)$ a.s. by the SLLN of U-statistic, it is enough to show that $P(A) = 0$. Fix $\omega \in A$. Then there exists an $\epsilon > 0$ and an increasing sequence $< m_n >$ of positive integers such that

$$\|\hat{\vartheta}_{m_n} - \vartheta_0\| \geq \epsilon, \quad \text{for all } n.$$

This yields

$$\sup_{\|\vartheta - \vartheta_0\| \geq \epsilon} U_{m_n}(\omega, \vartheta) \geq U_{m_n}(\omega, \hat{\vartheta}_{m_n}) \geq U_{m_n}(\omega, \vartheta_0) - O(\frac{1}{m_n})$$

for all n. Thus

$$T(\epsilon) \equiv \limsup_{n \to \infty} \sup_{\|\vartheta - \vartheta_0\| \geq \epsilon} U_n(\omega, \vartheta) \geq \mu(\vartheta_0).$$

Consequently, $\omega \in B_\epsilon = \{T(\epsilon) \geq \mu(\vartheta_0)\}$. This shows that $A \subset \bigcup_{\epsilon > 0} B_\epsilon$. We shall now show that $P(B_\epsilon) = 0$ for every $\epsilon > 0$. This will imply that desired $\mathbb{P}(A) = 0$.

Let K be as in (C.5). Fix a small $\epsilon > 0$ so that $C_\epsilon = \{\vartheta \in K : \|\vartheta - \vartheta_0\| \geq \epsilon\}$ is not empty. Then C_ϵ is compact, and it follows from Lemma 1 and Lemma 2 that

$$T_1(\epsilon) \equiv \limsup_{n \to \infty} \sup_{\vartheta \in C_\epsilon} U_n(\vartheta) \leq \sup_{\vartheta \in C_\epsilon} \mu(\vartheta) < \mu(\vartheta_0) \quad a.s.$$

and from (C.5) that

$$T_2(\epsilon) \equiv \limsup_{n \to \infty} \sup_{\vartheta \in \Theta \setminus C_\epsilon : \|\vartheta - \vartheta_0\| \geq \epsilon} U_n(\vartheta) < \mu(\vartheta_0) \quad a.s.$$

Combining the above shows that $T(\epsilon) \leq T_1(\epsilon) \vee T_2(\epsilon) < \mu(\vartheta_0)$ a.s. This is the desired $\mathbb{P}(B_\epsilon) = 0$. □

Proof of Theorem 3. Let $\eta > 0$ be small enough so that the closed ball $B_\eta = \left\{\vartheta \in \mathbb{R}^k : \|\vartheta - \vartheta_0\| \leq \eta\right\} \subset \Theta$. We shall verify (C.5) with $K = B_\eta$. Let $\vartheta \in \Theta$ with $\|\vartheta - \vartheta_0\| > \eta$. Then there exist a $\upsilon \in \mathbb{R}^k$ of length $\|\upsilon\| = \eta$ and an $a > 1$ such that $\vartheta = \vartheta_0 + a\upsilon$. It follows from the assumed concavity that $\vartheta \mapsto U_n(\vartheta)$ is concave down. Thus

$$U_n(\vartheta_0 + \upsilon) \geq \frac{1}{a} U_n(\vartheta_0 + a\upsilon) + \frac{a-1}{a} U_n(\vartheta_0).$$

This yields

$$U_n(\vartheta_0 + a\upsilon) \leq U_n(\vartheta_0) - a\left(U_n(\vartheta_0) - \sup_{\|\upsilon\| = \eta} U_n(\vartheta_0 + \upsilon)\right)$$

and shows that

$$\sup_{\|\vartheta - \vartheta_0\| > \eta} U_n(\vartheta) \leq U_n(\vartheta_0) - \inf_{a > 1} a\left(U_n(\vartheta_0) - \sup_{\|\vartheta - \vartheta_0\| = \eta} U_n(\vartheta)\right).$$

In view of Lemma 2,

$$\liminf_{n\to\infty}\left(U_n(\vartheta_0)-\sup_{\|\vartheta-\vartheta_0\|=\eta}U_n(\vartheta)\right)\geq\mu(\vartheta_0)-\sup_{\|\vartheta-\vartheta_0\|=\eta}\mu(\vartheta)\quad a.s.$$

Since $\Delta_\eta=\mu(\vartheta_0)-\sup_{\|\vartheta-\vartheta_0\|=\eta}\mu(\vartheta)$ is positive by Lemma 2, we obtain

$$\limsup_{n\to\infty}\left(\sup_{\|\vartheta-\vartheta_0\|>\eta}U_n(\vartheta)\right)\geq\mu(\vartheta_0)-\Delta_\eta\quad a.s.$$

This shows that (C.5) holds with $K=B_\eta$. Thus, the desired result follows from Theorem 7.

Lemma 3. *Suppose ψ is regular at ϑ_0. Then the map $\vartheta\mapsto M_\vartheta$ is continuous at ϑ_0. Moreover, if $\{a_n\}$ is a sequence of positive numbers converging to 0, then*

$$\sup_{\|\vartheta-\vartheta_0\|\leq a_n}\left\|\nabla^2U_n(\vartheta)-M_{\vartheta_0}\right\|\to 0\quad a.s.,$$

$$\sup_{\|\vartheta-\vartheta_0\|\leq a_n}\frac{\|\nabla U_n(\vartheta)-\nabla U_n(\vartheta_0)-M_{\vartheta_0}(\vartheta-\vartheta_0)\|}{\|\vartheta-\vartheta_0\|}\to 0\quad a.s.$$

and almost surely,

$$\sup_{\|\vartheta-\vartheta_0\|\leq a_n}\frac{\left\|U_n(\vartheta)-U_n(\vartheta_0)-\nabla U_n(\vartheta_0)(\vartheta-\vartheta_0)-\frac{1}{2}(\vartheta-\vartheta_0)^\intercal M_{\vartheta_0}(\vartheta-\vartheta_0)\right\|}{\|\vartheta-\vartheta_0\|^2}\to 0.$$

Proof. For $a>0$, let h_a denote the map defined by

$$h_a(\mathbf{x})=\sup_{\|\vartheta-\vartheta_0\|\leq a}\left\|\nabla^2\psi(\mathbf{x},\vartheta)-\nabla^2\psi(\mathbf{x},\vartheta_0)\right\|,\ \mathbf{x}\in\mathscr{X}^r$$

This map is measurable as the supremum can be achieved over a countable subset. Moreover, for each $\mathbf{x}\in\mathscr{X}^r$, $h_a(\mathbf{x})\downarrow 0$ as $a\downarrow 0$. Also, for small enough a, $0<h_a(\mathbf{x})\leq H(\mathbf{x})$ for all $\mathbf{x}\in\mathscr{X}^r$. Thus by the Lebesgue Dominated Convergence Theorem,

$$\lim_{a\to 0}\int h_a(\mathbf{x})\ F^r(d\mathbf{x})=0.$$

This shows that the map $\vartheta\mapsto M_\vartheta$ is continuous at ϑ_0. Since

$$\begin{aligned}&\limsup_{n\to\infty}\binom{n}{r}^{-1}\sum_{i_1<\cdots<i_r}h_{a_n}(X_{i_1},\ldots,X_{i_r})\\ \leq&\limsup_{n\to\infty}\binom{n}{r}^{-1}\sum_{i_1<\cdots<i_r}h_a(X_{i_1},\ldots,X_{i_r})=\int h_a(\mathbf{x})\ F^r(d\mathbf{x})\end{aligned}$$

for every $a > 0$. We also find that

$$\binom{n}{r}^{-1} \sum_{i_1<\cdots<i_r} h_{a_n}(X_{i_1}, \ldots, X_{i_r}) \to 0 \quad a.s..$$

The above and $\|\nabla^2 U_n(\vartheta_0) - M_{\vartheta_0}\| \to 0$ by the SLLN yield

$$\sup_{\|\vartheta-\vartheta_0\|\le a_n} \left\|\nabla^2 U_n(\vartheta) - M_{\vartheta_0}\right\| \to 0 \quad a.s..$$

By the Taylor Theorem, each coordinate of

$$\|\vartheta - \vartheta_0\|^{-1} \|\nabla U_n(\vartheta) - \nabla U_n(\vartheta_0) - M_{\vartheta_0}(\vartheta - \vartheta_0)\|$$

is bounded by $\sup_{\|\vartheta-\vartheta_0\|\le a_n} \|\nabla U_n(\vartheta) - M_{\vartheta_0}\|$ provided $\|\vartheta - \vartheta_0\| \le a_n$. Thus,

$$\sup_{\|\vartheta-\vartheta_0\|\le a_n} \frac{\|\nabla U_n(\vartheta) - \nabla U_n(\vartheta_0) - M_{\vartheta_0}(\vartheta - \vartheta_0)\|}{\|\vartheta - \vartheta_0\|} \to 0 \quad a.s..$$

In a similar way, the two term Taylor expansion, we also obtain almost surely

$$\sup_{\|\vartheta-\vartheta_0\|\le a_n} \frac{\left\|U_n(\vartheta) - U_n(\vartheta_0) - \nabla U_n(\vartheta_0)(\vartheta - \vartheta_0) - \frac{1}{2}(\vartheta - \vartheta_0)^{\mathsf{T}} M_{\vartheta_0}(\vartheta - \vartheta_0)\right\|}{\|\vartheta - \vartheta_0\|^2} \to 0.$$

□

Proof of Theorem 7. Set $R_n = \nabla U_n(\hat{\vartheta}_n) - \nabla U_n(\vartheta_0) - M_{\vartheta_0}(\hat{\vartheta}_n - \vartheta_0)$, since $\hat{\vartheta}_n = \vartheta_0 + o_p(1)$, We obtain from the second part of Lemma 3 that $R_n = o_p(\left\|\hat{\vartheta}_n - \vartheta_0\right\|)$. This and $\sqrt{n}U_n(\hat{\vartheta}_n) = o_p(1)$ yields

$$\sqrt{n}(\hat{\vartheta}_n - \vartheta_0) = -\sqrt{n}M_{\vartheta_0}^{-1}\nabla U_n(\vartheta_0) + o_p(1) + \sqrt{n}o_p(\left\|\hat{\vartheta}_n - \vartheta_0\right\|).$$

Since $\int \|\nabla\psi(\mathbf{x}, \vartheta)\|^2 \ F^r(d\mathbf{x}) < \infty$, $-\sqrt{n}M_{\vartheta_0}^{-1}\nabla U_n(\vartheta_0) = O_p(1)$. By the invertibility of M_{ϑ_0},

$$\sqrt{n}\left\|\hat{\vartheta}_n - \vartheta_0\right\| \le \left\|M_{\vartheta_0}^{-1}\right\| \left\|\sqrt{n}M_{\vartheta_0}(\hat{\vartheta}_n - \vartheta_0)\right\| = O_p(1) + \sqrt{n}o_p(\left\|\hat{\vartheta}_n - \vartheta_0\right\|).$$

This implies that $\sqrt{n}\|\hat{\vartheta}_n - \vartheta_0\| = O_p(1)$. Consequently, we obtain

$$\sqrt{n}(\hat{\vartheta}_n - \vartheta_0) = -\sqrt{n}M_{\vartheta_0}^{-1}\nabla U_n(\vartheta_0) + o_p(1).$$

Hence,

$$\sqrt{n}(\hat{\vartheta}_n - \vartheta_0) \Rightarrow N(0, r^2 M_{\vartheta_0}^{-1} V_{\vartheta_0} (M_{\vartheta_0}^{-1}))^{\mathsf{T}})$$

follows from the CLT of U-statistic directly. □

Proof of Theorem 6. Denote the cdf of $h_b(\boldsymbol{\xi_0})$ by H and $\mu_b(\vartheta) = E[\|h_b(\boldsymbol{\xi_0})\| - \|\vartheta - h_b(\boldsymbol{\xi_0})\|]$. Then

$$\begin{aligned}
&\mu_b(\vartheta) - \mu_b(\vartheta_0) = E\left(\|\vartheta_0 - h_b(\boldsymbol{\xi_0})\| - \|\vartheta - h_b(\boldsymbol{\xi_0})\|\right) \\
&= -\|\vartheta - \vartheta_0\|B + E\left(\|\vartheta_0 - h_b(\boldsymbol{\xi_0})\| - \|\vartheta - h_b(\boldsymbol{\xi_0})\|\mathbf{1}[h_b(\boldsymbol{\xi_0}) \neq \beta(\vartheta_0)]\right) \\
&\leq -\|\vartheta - \vartheta_0\|B
\end{aligned}$$

the last inequality yields from the symmetry of the $h_b(\boldsymbol{\xi_0})$ at ϑ_0 and the convexity of the norm. By the Lemma 5, we obtain that

$$\mathrm{Sup}_\vartheta \left| \frac{U_{b,n}(\vartheta) - U_{b,n}(\vartheta_0) - (\mu_b(\vartheta) - \mu_b(\vartheta_0))}{\|\vartheta - \vartheta_0\|} \right| \to 0. \quad a.s.$$

Thus, for almost every ω and ϵ, there exists $N_{\omega,\epsilon} > 0$, such that for $n > N_{\omega,\epsilon}$ and $\vartheta \in B_\epsilon = \{\vartheta : 0 < \|\vartheta - \vartheta_0\| \leq B/2\epsilon\}$,

$$U_{b,n}(\vartheta) - U_{b,n}(\vartheta_0) \leq \epsilon \|\vartheta - \vartheta_0\| - B \leq -B/2.$$

This combine with Condition (C.5) yields the super-efficiency of the spatial-depth based MTSE $\hat{\beta}_{n,\mathrm{sp}}$. □

Lemma 4. *If the error distribution is discontinuous then* $P(h_\beta(\boldsymbol{\xi_0}) = \beta) > 0$.

Let

$$\varphi_\vartheta(\boldsymbol{\xi_0}) = \frac{\psi(\boldsymbol{\xi_0}; \vartheta)}{\|\vartheta\|} = \frac{\|h(\boldsymbol{\xi_0})\| - \|\vartheta - h(\boldsymbol{\xi_0})\|}{\|\vartheta\|}$$

and $\Phi = \{\varphi_\vartheta : \vartheta\}$

Lemma 5. *For all* $\epsilon > 0$, $N_{[\ ]}(\epsilon, \Phi, P) < \infty$. *Furthermore,* $\left\| \frac{U_n(\vartheta) - \mu(\vartheta)}{\|\vartheta\|} \right\|_\Phi \to 0, \quad a.s.$

Proof. $\varphi_\vartheta(x) = \frac{\|x\| - \|\vartheta - x\|}{\|\vartheta\|}$ can be bracketed as the indicator functions of cells considered in Example 3.7.4C in Van de Geer (2000). Thus, $N_{[\ ]}(\epsilon, \Phi, P) < \infty$. By the Corollary 3.5 in Arcones, Chen and Giné (1994), we have

$$\left\| \frac{U_n(\vartheta) - \mu(\vartheta)}{\|\vartheta\|} \right\|_\Phi \to 0, \quad a.s.$$

References

[1] Arcones, M., Chen, Z., and Giné, E. (1994). Estimators Related to U-Processes with Applications to Multivariate Medians: Asymptotic Normality. *Ann. Stat* **22**, 1460 - 1477.

[2] Akritas, M., Murphy, S., and LaValley, M. (1995). The Theil-Sen Estimator with Doubly Censored Data and Applications to Astronomy. *J. Amer. Statist. Assoc.* **90**, 170 -177.

[3] Bose, A.(1998). Bahadur representation of M_m estimates. *Ann. Stat* **26** 771 - 777.

[4] Chaudhuri, P.(1996). Multivariate Location Estimation Using Extension of R-Estimates Through U-Statistics Type Approach. *Ann. Statist.* **20**, 897-916.

[5] Chatterjee, S. and Olkin, I.(2006). Nonparametric estimation for quadratic regression. *Statist. & Probil. Lett.* **76**, 1156-1163.

[6] Dietz, E. J. (1989). Teaching Regression in a Nonparametric Statistics Course. *The American Statistician.* **43**, 35-40.

[7] Fernandes, R. and Leblanc, S. (2005). Parametric (modified least squares) and non-parametric (Theil-Sen) linear regressions for predicting biophysical parameters in the presence of measurement errors. *Remote Sensing of Environment.* **95**, 303-316.

[8] Ghosh, A and Chaudhuri, P. (2005). On data depth and distribution-free discriminant analysis using separating surfaces. *Bernoulli* **11**, 127.

[9] Hollander, M. and Wolfe, D. A. (1999). *Nonparametric statistical methods.* 2nd ed. John Wiley & Sons, NY. 421-423.

[10] Koshevoy, G and Mosler, K. (1997). Zonoid trimming for multivariate distributions. *Ann. of statist.* **95**, 1998-2017.

[11] Liu. R. (1990). On a notion of data depth based on random simplices. *Ann. Stat* **18** 405–414.

[12] Liu, R, Parelius, J.M., and Singh, K. (1999). Multivariate analysis by data depth descriptive statistics, graph and inference. *Ann. Stat* **27**, 783-858.

[13] Niemiro, W.(1992). Asympotics for M-Estimators Defined by Convex Minimization. *Ann. Stat* **20**, 1514 – 1533.

[14] Oja, H. (1983). Descriptive statistics for multivariate distributions. *Stat. & Prob. Letters* **6**, 327-332.

[15] Oja, H. and Niinimaa, A. (1984). On Robust Estimation of Regression Coefficients, *Research Report*, Department of Applied Mathematics and Statistics, University of Oulu, Finland.

[16] Peng, H., Wang, S., and Wang, X. (2007). Consistency and asymptotic distribution of the Theil-Sen estimator. *J. Statist. Plann. Infer.*, In press (www.sciencedirect.com).

[17] Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection.* John Wiley & Sons, Inc.

[18] Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *J. Amer. Statist. Assoc.*, **63**, 1379-1389.

[19] Serfling, R. (1984). Generalized $L-$, $M-$, and $R-$ statistics. *Ann. Statist.* **1**, 76-86.

[20] Serfling, R. (2006). Multivariate symmetry and asymmetry. *In Encyclopedia of Statistical Sciences*, 2nd Ed. (Kotz, Balakrishnan, Read and Vidakovic, eds.), Wiley, 5338-5345.

[21] Small, C. (1990). A survey of multidimensional medians. *International Statistical Review / Revue Internationale de Statistique* **58**, 263-277.

[22] Sprent, P. (1993). *Applied nonparametric statistical methods.* 2nd Ed. CRC Press, NY.

[23] Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis, *I. Proc. Kon. Ned. Akad. v. Wetensch.***A53**, 386-392.

[24] Tukey, J.W. (1975). Mathematics and picturing data. In *Proceedings of the 1974 International Congress of Mathematicians* (R. James, ed.) 523-531.

[25] Van de Geer, S (2000). Empirical processes in M-estimation, *Cambridge university press*

[26] Wang, X. (2005). Asymptotics of the Theil-Sen estimator in a simple linear regression model with a random covariate. *J. Nonparam. Statist.* **17**, 107-120.

[27] Wilcox, R. (1998). Simulations on the Theil-Sen regression estimator with right-censored data. *Stat. & Prob. Letters* **39**, 43-47.

[28] Wilcox, R. (2004). Some results on extensions and modifications of the Theil-Sen regression estimator. *British J. Math. Statist. Psych.*, **57**, 265-280.

[29] Zhang, J. (2002). Some extensions of Tukey depth function. *J. of Multi. Analysis* **82**, 134–165.

[30] Zhou, W. and Serfling, R. (2006). Multivariate spatial U-quantiles: a Bahadur-Kiefer representation, a Theil-Sen estimator for multiple regression, and a robust dispersion estimator. *Manuscript.*

[31] Zuo, Y. and Serfling, R. (2000a). General notions of statistical depth function. *Ann. Statist.* **28**, 461-482.