The Asymptotics of the Mahalanobis Depth-Based Theil-Sen Estimators in Multiple Linear Models

Fang Li [a], Hanxiang Peng [a,*,1] and Fei Tan [b]

[a] *Department of Mathematical Sciences, Indiana University – Purdue University At Indianapolis, Indianapolis, IN 46202-3216, USA*

[b] *Institute of Public Health, College of Pharmacy and Pharmaceutical Sciences, Florida A&M University, Tallahassee, FL 32307, USA*

Abstract

In this article, we propose the use of the Mahalanobis depth to construct fully affine equivalent Theil-Sen estimators of parameters in a multiple linear model. Our construction includes the spatial depth-based Theil-Sen estimators as a special case. We exhibit that the efficiency of the proposed estimators relative to the least squares estimators increases with the number of parameters under independent and identically distributed normal errors. We point out that the estimators are robust with a possible maximal breakdown point and possess bounded influences. We show that the estimators are consistent and asymptotic normal for both deterministic and random covariates under suitable conditions.

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

Key words: Breakdown point, efficiency, influence function, Mahalanobis distance, spatial depth function.

July 17, 2009

1 Introduction

In a simple linear model, each pair of distinct observations results in an estimator of the slope. The median of these slope estimators is known as the Theil-Sen estimator (Theil, 1950; Sen, 1968). Having a clear geometric interpretation, the Theil-Sen estimator is robust and relatively efficient, so that it is competitive to other slope estimators such as the least squares estimator. The Theil-Sen estimator is often included in textbooks about robust statistics (Hollander and Wolfe, 1973, 1999; Rousseeuw and Leroy, 1986; Huber 1977; Dietz, 1989; Wilcox, 1998; Sprent 1993; Jurečková and Picek, 2006; and Maronna, *et al.*, 2006). It also has important applications, e.g., in astronomy by Akritas *et al.* (1995) in censored data, in remote sensing by Fernandes and Leblanc (2005). Recently, Wang (2005) investigated the asymptotic behaviors of the Theil-Sen estimator when covariates are random. When covariates are deterministic, Peng, Wang and Wang (2008) obtained the consistency and asymptotic distribution of the Theil-Sen estimator, and showed that it is super-efficient when the error distribution is discontinuous. Chatterjee and Olkin (2006) proposed nonparametric method for fitting a quadratic regression by exploiting the idea of the Theil-Sen estimator.

Using the spatial median, Zhou and Serfling (2008) and Wang, Dang, Peng and Zhang (2009) constructed the multivariate Theil-Sen estimator (MTSE) of parameters in a multiple linear model, generalizing the (univariate) Theil-Sen estimator (UTSE). When covariates are random, the asymptotic normality of

* 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-0940365.

the MTSE can be derived from the spatial quantile theory of Zhou and Serfling. Also for random covariates, Wang, *et al.* demonstrated that the MTSE's are robust with a possible maximal breakdown point, consistent, and asymptotically normal under mild conditions. They also conducted simulations to demonstrate the robustness and efficiency of the MTSE's and compare them with the least squares estimators. Busarova, *et al.* (2006) proposed the multivariate Theil type estimator of the parameter in a multiple linear model based on the Oja median and derived the asymptotic normality for random covariates. When the covariates are deterministic, Shen (2008, 2009) proved the asymptotic normality using the convexity lemma.

In this article, we will construct the Theil-Sen estimators of parameters in a multiple linear model based on the Mahalanobis depth function. Our construction is flexible, for example, we can choose the scale matrix in the Mahalanobis depth function to be the identity matrix, the resulting estimators are then the spatial depth-based Theil-Sen estimators given by Zhou and Serfling (2008) and Wang, *et al.* (2009). Since the Mahalanobis depth function is fully *affine invariant*, the Mahalanobis depth-based Theil-Sen estimators are *affine equivariant*. Whereas the spatial depth-based Theil-Sen estimators are only *orthogonally equivariant*, since the spatial depth is orthogonally but not affine invariant. It is worth to note that both Mahalanobis depth and spatial depth based- Theil-Sen estimators are *regression* and *scale equivarant*.

We show that the efficiency of the MTSE's relative to the least squares estimators increases with the number of parameters and tends to one as the number of parameters goes to infinity under i.i.d. normal errors. This is an advantage of the MTSE's, bearing in mind that the MTSE's have the desired properties such as consistency and asymptotic normality under much weaker assumptions than other estimators such as the least squares estimators. In fact, what is required for the error with the MTSE's is that it is similar to "angularly symmetric", a weaker assumption than the usual symmetry, see Liu (1990). Whereas the error with the least squares estimator must have mean zero and second finite moment. Note that the (univariate) sample median is sometimes described as inefficient, with normal efficiency $2/\pi = 0.637$ compared to the sample mean. The spatial median generalizes the univariate median to high dimension. Brown (1983) showed that the efficiency of the spatial median relative to the sample mean is superior to the efficiency of the univariate median. Higher dimensional results for spherical symmetric distributions show further increase in efficiency, which converges to one as the dimension tends to infinity. Bai, Chen, Miao and Rao (1990) exhibited the asymptotic relative efficiency of their least distance estimator in a location regression model with normal errors intends to one as the dimension of the covariate matrices intends to infinity. Chaudhuri (1992) demonstrated what was observed by Brown (1983) is also true for his multivariate extension of the Hodges-Lehman estimate. Our result confirms that what Brown, Bai, *et al.* and Chaudhuri have found is also true for a multiple linear regression. This shows that the advantage in efficiency of the least squares estimator over the the MTSE's diminishes as the dimension p gets larger, for very large p, the least squares estimator practically loses its merit over the MTSE's. Of course the MTSE's are much more difficult to compute than the least squares estimator. This phenomenon is comparable to the behavior of the well known Stein estimator.

We prove that the MTSE's are asymptotic normal for both deterministic or random covariates under suitable conditions. When covariates are deterministic, the loss function $\Psi_n(b)$ (see (3)) that defines the MTSE's is not a U-statistic. This causes us a technical difficulty, that is, we cannot apply the commonly used theory of U-statistics to derive the asymptotic normality. We will follow Shen (2009) to overcome this difficulty. Specifically, like Shen, we will prove the asymptotic normality by using the convexity lemma (Pollard, 1991), the characterization of minimizers of a convex function, and the central limit theorem for sums of dissociated random variables (Barbour and Eagleson, 1985).

The rest of the article is structured as follows. In Section 2, we construct the Theil-Sen estimators and explore existence, uniqueness and robustness. We also prove the consistency. In Section 3, we introduce the assumptions and give the main results. Section 4 is devoted to the comparison of the efficiency of the MTSE's with the LSE. Proofs can be found in Section 5.

2 The Theil-Sen Estimators and Existence, Uniqueness, Robustness and Consistency

In this section, we construct the MTSE's, discuss existence, uniqueness and robustness. At the end of the section, we give the consistency.

The Multivariate Theil-Sen Estimators. In a multiple linear model, the response y_j and covariate vector x_j satisfy

$$y_j = x_j^\top \beta + \varepsilon_j, \quad j = 1, ..., n, \tag{1}$$

where β is a p-variate (column) vector of regression parameter, x_j's are (either random or deterministic) covariates, ε_j's are random errors, and x_j's and ε_j's are uncorrelated.

Let J be a p-subset of $\{1, ..., n\}$. Let X_J be the $p \times p$ matrix with rows $x_j^\top, j \in J$ and ε_J be the column vector with components $\varepsilon_j, j \in J$. Assuming for now that X_J is invertible. Then an estimate of β can be obtained as the solution to the p equations

$$x_j^\top \beta = y_j, \quad j \in J. \tag{2}$$

This estimate is $b_J = X_J^{-1} Y_J$, which is also the least squares estimate (LSE) of β based on the sub-sample $\{(x_j, y_j) : j \in J\}$. Here the sub-sample size is p. One actually does not have to choose exactly p observations; any m observations can be chosen as long as m is at least the number of parameters. Obviously m is at most the sample size n, so $p \leq m \leq n$. For each m-subset J of $\{1, ..., n\}$ and the sub-sample with indices in J, an estimate of β is the least squares estimate $b_{(J)} = (X_J^\top X_J)^{-1} X_J^\top Y_J$, provided X_J has full rank. Let $\mathcal{J}^{(n)}$ be the collection of the m-subsets of $\{1, ..., n\}$ such that for every $J \in \mathcal{J}^{(n)}$, X_J has full rank. Then an estimator $\hat{\beta}_n$ of β is the *Mahalanobis depth-based median*, $\hat{\beta}_n = \mathrm{MHmed}\{b_{(J)} : J \in \mathcal{J}^{(n)}\}$, which is the minimizer of the loss function

$$\Psi_n(b) \equiv \binom{n}{m}^{-1} \sum_{J \in \mathcal{J}^{(n)}} \left(\|\Sigma_{(J)}^{-1/2} (b - b_{(J)})\| - \|\Sigma_{(J)}^{-1/2} b_{(J)}\| \right), \quad b \in \mathbb{R}^p, \tag{3}$$

where $\|\cdot\|$ is the Euclidean norm and $\Sigma_{(J)}$ is an estimate of the covariance functional Σ (Serfling, 2008) depending on X_J (which is assumed hereafter), so that Σ and $\Sigma_{(J)}$ are symmetric and positive definite. Since the summand in (3) is bounded by $\|\Sigma_{(J)}^{-1/2} b\|$ by the triangle inequality, differentiation of (3) with respect to b can pass the expectation by the dominated convergence theorem, therefore we see that the minimizer $\hat{\beta}_n$ must satisfy

$$\binom{n}{m}^{-1} \sum_{J \in \mathcal{J}^{(n)}} \Sigma_{(J)}^{-\top/2} S\left(\Sigma_{(J)}^{-1/2} (b - b_{(J)})\right) = 0, \tag{4}$$

under certain assumptions, see Remark 3 and (22) below and the discussions therein. Here $S(b) = b/\|b\|, b \in \mathbb{R}^p$ is the *spatial sign function* (Zhou and Serfling, 2008) or the *spatial unit function* (Chaudhuri, 1992), whereas $D_S(b) = 1 - \|\mathbb{E}S(b - \xi)\|$ is termed as the *spatial depth function* of random vector ξ. Note when $\Sigma_{(J)} = I_m$, then $\hat{\beta}_n$ simplifies to the spatial depth-based MTSE's proposed by Wang, *et al.* (2009).

Difference-Based MTSE's. Taking the pairwise differences of multiple linear model (1), we eliminate the intercept and obtain

$$y_j - y_k = (x_j - x_k)^\top \beta_1 + \varepsilon_j - \varepsilon_k, \quad j, k = 1,\ 2,\ ...,\ n, \tag{5}$$

where β_1 is a $(p-1)$-variate parameter resulting from β with the first component (the intercept) deleted. There are $N = n(n-1)/2$ distinct pairwise differences for n distinct observations. For an integer m between $p-1$ parameters and the sample size n, let $\mathcal{K}^{(n)}$ be the $\binom{N}{m}$ combinations of (j, k) from $\nabla \equiv \{(j, k) : j < k, j, k = 1,\ ...,\ n\}$ and write $\{(k_{1,i},\ k_{2,i}) : i = 1,\ ...,\ m\}$, a generic element in $\mathcal{K}^{(n)}$, $K_j = (k_{j,i} : i = 1,\ ...,\ m)$ for $j = 1, 2$, and write K for either K_1 or K_2. Then (5) can be written in matrix form as

$$Y_{K_1,K_2} = X_{K_1,K_2}\beta_1 + \varepsilon_{K_1,K_2}, \quad (K_1, K_2) \in \mathcal{K}^{(n)}, \tag{6}$$

where $Y_{K_1,K_2} = Y_{K_1} - Y_{K_2}$, $X_{K_1,K_2} = X_{K_1} - X_{K_2}$ and $\varepsilon_{K_1,K_2} = \varepsilon_{K_1} - \varepsilon_{K_2}$ with $\varepsilon_K = (\varepsilon_k : k \in K)^\top$. Let $b_{(K_1,K_2)}$ be the least squares estimator based on the subset of the observations with indices in (K_1, K_2), i.e.,

$$b_{(K_1,K_2)} = ({X_{K_1,K_2}}^\top X_{K_1,K_2})^{-1} {X_{K_1,K_2}}^\top Y_{K_1,K_2}, \qquad (K_1, K_2) \in \mathcal{K}^{(n)}. \tag{7}$$

Like Zhou and Serfling (2008) and Wang, *et al.* (2009), we propose an estimator $\hat{\beta}_{1,n}$ of β_1 as the Mahalanobis-depth median, $\hat{\beta}_{1,n} = \text{MHmed}\{b_{(K_1,K_2)} :$

$(K_1, K_2) \in \mathcal{K}_0^{(n)}\}$, which is the minimizer of

$$\Psi_{n,d}(b) \equiv \binom{N}{m}^{-1} \sum_{(K_1,K_2)\in\mathcal{K}_0^{(n)}} \left(\|\Sigma_{(K_1,K_2)}^{-1/2} (b - b_{(K_1,K_2)})\| - \|\Sigma_{(K_1,K_2)}^{-1/2} b_{(K_1,K_2)}\| \right) \tag{8}$$

for $b \in \mathbb{R}^{p-1}$, where $\Sigma_{(K_1,K_2)}$ is an estimate of the covariance functional Σ depending only on X_{K_1,K_2}, which is assumed hereafter. Here $\mathcal{K}_0$ is a subset of $\mathcal{K}$ in which all the least squares estimates exist. Note when $\Sigma_{(K_1,K_2)} = I_m$, the proposed estimator $\hat{\beta}_{1,n}$ simplifies the MTSE's proposed by Zhou and Serfling (2008) and Wang, *et al.* (2009).

The symmetry of ε_{K_1,K_2} later on is shown to be one of the sufficient conditions that guarantee the uniqueness of the MTSE's. Even though each component of ε_{K_1,K_2} is symmetric, we do not have the symmetry of ε_{K_1,K_2}, i.e., $\varepsilon_{K_1,K_2} \stackrel{d}{=} -\varepsilon_{K_1,K_2}$, without the assumption of central symmetry on the error ε. The reason is that the components of the random vectors ε_{K_1,K_2} may be correlated, for instance, $\varepsilon_2 - \varepsilon_1$ and $\varepsilon_3 - \varepsilon_2$ are correlated. 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 $\varepsilon_{K_1,K_2} = (\varepsilon_1 - \varepsilon_2,\ \varepsilon_3 - \varepsilon_4,\ \ldots,\ \varepsilon_{2m-1} - \varepsilon_{2m})^\top$. Specifically, we choose the pairwise differences which constitute ε_{K_1,K_2} in such a way that each of K_1 and K_2 has distinct elements and K_1, K_2 have no element in common. We denote the set of all possible such (K_1, K_2) as $\mathcal{K}^*$. It is easy to obtain that the total number of (K_1, K_2) in $\mathcal{K}^*$ is $N_{n,m} = \binom{n}{2}\binom{n-2}{2}\cdots\binom{n-2m+2}{2}/(2^m m!) = \binom{n}{2m}(2m)!/(2^m m!)$. Then following the above procedure, we construct the non overlapped difference- based MTSE $\hat{\beta}^*_{1,n} = \text{MHmed}\{b_{(K_1,K_2)} : (K_1, K_2) \in \mathcal{K}^*_0\}$, which is the minimizer of

$$\Psi^*_{n,d}(b) \equiv N_{n,m}^{-1} \sum_{(K_1,K_2)\in\mathcal{K}^*_0} \left(\|\Sigma_{(K_1,K_2)}^{-1/2} (b - b_{(K_1,K_2)})\| - \|\Sigma_{(K_1,K_2)}^{-1/2} b_{(K_1,K_2)}\| \right) \tag{9}$$

for $b \in \mathbb{R}^{p-1}$. Here $\mathcal{K}_0^*$ is a subset of $\mathcal{K}^*$ in which all the least squares estimates exist.

Existence and Uniqueness The proposed estimators are the multivariate medians defined by the Mahalanobis depth function. Kemperman (1987) investigated the median of a finite measure on a Banach space. Here we review some of the facts tailored for our application. Let μ be a probability measure on a Banach space $\mathbb{X}$ with norm $\|\cdot\|$. The median θ of μ is any minimizer in $\mathbb{X}$ of the loss function

$$\Psi(y) = \int (\|x - y\| - \|x\|)\, \mu(dx), \quad y \in \mathbb{X}. \tag{10}$$

Clearly, the integral in (10) is finite as the integrand is bounded by $\|y\|$ for every $y \in \mathbb{X}$. The median θ always exists when $\mathbb{X}$ is *finite dimensional*. Note that Ψ is strictly convex, hence μ has a unique median if μ is *not concentrated on any straight line in* $\mathbb{X}$. A simple calculus shows that θ is a median of μ if and only if

$$\int \phi_{\theta - x}(h)\, \mu(dx) \geq 0,$$

where $\phi_x(h) = \lim_{t \downarrow 0}(\|x + th\| - \|x\|)/t$, x, $h \in \mathbb{X}$ is the right limit at x. This limit exists for $\mathbb{X} = \mathbb{R}^p$, hence θ is a median if and only if

$$\int \phi_{\theta - x}(h) \mu(dx) = 0, \quad x, h \in \mathbb{X}, \tag{11}$$

provided $\mu(\{\theta\}) = 0$. Moreover, *a sufficient condition that θ is a median is that θ satisfies (11).* For the Euclidean norm, a sufficient condition that θ is a spatial median is

$$\int S(\theta - x)\, \mu(dx) = 0. \tag{12}$$

It is worth to note that *if μ is symmetric about θ_0, then θ_0 satisfies (12) ((11)).*

We now apply the above results to acquire the conditions that ensure the existence and uniqueness of the MTSE's. Let β_0 be the true but unknown parameter value and $\beta_{1,0}$ be the true parameter value obtained from β_0 with the first component deleted. There are two cases.

Case i. Random Covariates. We first consider the case that x_1, ..., x_n are i.i.d. random covariates. Let $\Sigma^{1/2}$ be its square root of a covariance functional Σ, so that $\Sigma = \Sigma^{1/2}\Sigma^{\top/2}$. Since $\|\Sigma^{-1/2}b\|$ is convex in $b \in \mathbb{R}^p$, we apply $\|\!|b|\!\| = \|\Sigma^{-1/2}b\|$. Obviously the MTSE's $\hat{\beta}_n$ ($\hat{\beta}_{1,n}$) exists since it is finite dimensional. By an analogous reasoning to Remark 1 of Wang, *et al.* (2009), we conclude that if $p \geq 2$ and the error is not point mass, then the MTSE $\hat{\beta}_n$ ($\hat{\beta}_{1,n}$) is unique almost surely for large n. In order that as n tends to infinity the MTSE $\hat{\beta}_n(\hat{\beta}_{1,n})$ will converge to the true parameter value $\beta_0(\beta_{1,0})$ respectively, the population loss function $b \mapsto \mathbb{E}[\Psi_n(b)]$ ($b \mapsto \mathbb{E}[\Psi_{n,d}(b)]$) must be minimized at the true parameter value $\beta_0(\beta_{1,0})$. We now derive an analog of characterization (12). Since under the true model, $b_{(J)} - \beta_0 = (X_J^\top X_J)^{-1}\varepsilon_J$, it follows that β_0 is the minimizer of $b \mapsto \mathbb{E}[\Psi_n(b)]$ if

$$\mathbb{E}\left\{\Sigma_{(J_0)}^{-\top/2} S\left(\Sigma_{(J_0)}^{-1/2}(X_{J_0}^\top X_{J_0})^{-1}\varepsilon_{J_0}\right)\right\} = 0. \tag{13}$$

Since x_i's and ε_i's are uncorrelated, the above equation is implied by the following sufficient condition

Assumption 1. *The error ε is symmetric about zero.*

The analog of (13) for the difference-based MTSE $\hat{\beta}_{1,n}$ is

$$\mathbb{E}\left\{\Sigma_{(K_1,K_2)}^{-\top/2} S\left(\Sigma_{(K_1,K_2)}^{-1/2}((X_{K_1,K_2}{}^\top X_{K_1,K_2})^{-1} X_{K_1,K_2}{}^\top \varepsilon_{K_1,K_2})\right)\right\} = 0,$$

where $(K_1, K_2) \in \mathcal{K}^0$. By Theorem 1 of Wang, *et al.* (2009), Assumption 1 implies that ε_{K_1,K_2} is also symmetric about zero for any $(K_1, K_2) \in \mathcal{K}^0$. Thus the above analog is satisfied under Assumption 1. If the pairwise differences are not overlapped $K_1 \cap K_2 = \emptyset$, then ε_{K_1,K_2} is symmetric about zero, and the convergence of the non overlapped difference-based MTSE $\hat{\beta}^*_{1,n}$ in Theorem 5 to the true parameter value $\beta_{1,0}$ is automatic. If we choose $\Sigma^{1/2}_{(J)} = X_J^\top X_J$, then (13) simplifies to

$$\mathbb{E}(S(\varepsilon)) = 0. \tag{14}$$

This means that the error ε is *angularly symmetric* about zero. Introduced in Liu (1990), angular symmetry is a weaker symmetry than the usual central symmetry, i.e., central symmetry implies angular symmetry. A systematic discussion about various types of symmetry can be found in Serfling (2006).

Case ii. Deterministic Covariates. We now consider the case that x_1, ..., x_n are deterministic. Analogous to the reasoning of the existence of Kemperman (pages 218-219, 1987) and the uniqueness of Wang, *et al.* (2009), we can show that there exists a minimizer for $\mathbb{E}[\Psi_n(b)]$ in $b \in \mathbb{R}^p$, and this minimizer is unique if $p \geq 2$ and the error ε is not point mass. In order that the MTSE's $\hat{\beta}_n, \hat{\beta}_{1,n}$ will converge to the true parameter values $\beta_0, \beta_{1,0}$ respectively, we need the following assumption and its analogous version for $\Psi_{n,d}(b)$ and $\Psi^*_{n,d}(b)$.

Assumption 2. *For each $b \in \mathbb{R}^p$ in a neighborhood of the true parameter value β_0, there exists an finite constant $\bar{\Psi}(b)$ such that as n tends to infinity, $\mathbb{E}\Psi_n(b) \to \bar{\Psi}(b)$. Further,*

$$\bar{\Psi}(b) = 0 \tag{15}$$

has a unique solution $b = \beta_0$.

The above assumption is mild. In fact, Assumption 1 implies that β_0 satisfies (15), whereas the uniqueness in (15) follows from the positive definiteness of matrix D in Assumption 3.

Remark 1. *Assumption 2 is implied by Assumptions 1, 3 and the existence* *of the finite limit $\bar{\Psi}(b)$ for b in a neighborhood of β_0.*

Consistency. From the above discussion and Corollary II.2 of Andersen and Gill (1982), we can prove the consistency of the MTSE's. The following Theorem 1 is a generalization of convergence from an average of independent random variables in Theorem 2.24 and Corollary 2.26 of Kemperman (1987) to an average of dissociated random variables (Barbour and Eegleson, 1985).

Theorem 1. *Suppose that the error ε is not point mass and $p \geq 2$.*

Case i. *Covariates x_i's are random. (i.1) Suppose $b \mapsto \mathbb{E}[\Psi_n(b)], b \in \mathbb{R}^p$ ($b \mapsto \mathbb{E}[\Psi_{n,d}(b)], b \in \mathbb{R}^{p-1}$) is strictly convex. If Assumption 1 holds, then $\hat{\beta}_n(\hat{\beta}_{1,n})$ will converge in probability to the true parameter value $\beta_0(\beta_{1,0})$, i.e., $\hat{\beta}_n \xrightarrow{P} \beta_0$ ($\hat{\beta}_{1,n} \xrightarrow{P} \beta_{1,0}$). (i.2) If $b \mapsto \mathbb{E}[\Psi^*_{n,d}(b)], b \in \mathbb{R}^{p-1}$ is strictly convex, then $\hat{\beta}^*_{1,n} \xrightarrow{P} \beta_{1,0}$.*

Case ii. *Covariates x_i's are deterministic. (ii.1) If Assumption 2 and its analogs hold, respectively, then $\hat{\beta}_n \xrightarrow{P} \beta_0$, $\hat{\beta}_{1,n} \xrightarrow{P} \beta_{1,0}$ and $\hat{\beta}^*_{1,n} \xrightarrow{P} \beta_{1,0}$, respectively.*

The strict convexity in Case i and the uniqueness of one solution (zero point) in Case ii are guaranteed by the positive definiteness of D in Assumption 3 and the analog of D in Assumption 5. Note also that with non-overlapped differences, symmetry Assumption 1 is not required for the consistency, and this is also true for the asymptotic normality. More discussions can be found in Wang, *et al.* (2009).

Robustness. Kemperman (1987) shows that the median has a maximal breakdown point 50%. Further, a median will not change if transporting mass along open half lines (rays) with the median as endpoint, leaving any mass at the median unchanged.

Remark 2. *Instead of the least squares estimate, we use a robust estimate with maximal breakdown point to construct the MTSE's, then the MTSE's has a maximal breakdown point, see Wang,* et al. *(2009).*

3 Asymptotic Normality

In this section, we first introduce the assumptions. We then give the asymptotic normality of the MTSE's for deterministic and random covariates, and the difference-based MTSE.

The MTSE $\hat{\beta}_n$ is the minimizer of the convex loss function (3). For deterministic covariates, $\Psi_n(b)$ is not a U-statistic, so we cannot apply the elegant theory of U-statistics. Like Shen (2009), we will exploit the convexity (Pollard, 1992), the characterization of minimizers of a convex function, and the central limit theorem for sums of dissociated random variables (Barbour and Eagleson, 1985) to derive the asymptotic normality of the MTSE's. We need the following assumptions.

Assumption 3. $\mathbb{E}\Psi_n(\beta)$ *has continuous second derivative* $\nabla^2\mathbb{E}\Psi_n(\beta)$ *with respect to* β *in a neighborhood of the true parameter value* β_0 *such that*

$$\lim_{n\to\infty} \nabla^2\mathbb{E}(\Psi_n(\beta_0)) = D, \quad \lim_{n\to\infty} n\,\mathbb{V}\mathrm{ar}(\nabla\Psi_n(\beta_0)) = V \tag{16}$$

for some positive definite matrices D *and* V*.*

A simple calculus shows

$$\nabla\Psi_n(b) = \binom{n}{m}^{-1} \sum_{J\in\mathcal{J}^{(n)}} \Sigma_{(J)}^{-\top/2}\, S\left(\Sigma_{(J)}^{-1/2}\,(b - b_{(J)})\right) \tag{17}$$

and

$$\nabla^2 \Psi_n(b) = \binom{n}{m}^{-1} \sum_{J \in \mathcal{J}^{(n)}} \Sigma_{(J)}^{-\top/2} S^{(1)} \left(\Sigma_{(J)}^{-1/2} (b - b_{(J)}) \right) \Sigma_{(J)}^{-1/2}, \qquad (18)$$

where $S^{(1)}(b) = (I_p - S^{\otimes 2}(b))/\|b\|$, $b \neq 0$, $b \in \mathbb{R}^p$ and $S^{(1)}(0) = 0$.

In view of Assumption 3.1 and Lemma 5.3 of Chaudhuri (1992), we have a sufficient condition for Assumption 3.

Remark 3. *A sufficient condition for Assumption 3 to hold is that the random errors* ε_1, ..., ε_n *are i.i.d. with an absolutely continuous distribution (w.r.t. the Lebesgue measure) having a density such that it is bounded on any compact subset of the real line* $\mathbb{R}$*;* $\mathbb{E}(\Sigma_J^{-1/2})$ *is nonsingular for every* $J \in \mathcal{J}^{(n)}$*;* $p \geq 2$*; the second limit in (16) exists for some positive definite matrix* V *when the covariates* x_i*'s are random, and when the covariates are deterministic both limits in (16) exist for some positive definite matrices* D *and* V*.*

Under the assumptions in Remark 3, we have

$$\nabla \mathbb{E}\left[\Psi_n(b)\right] = \binom{n}{m}^{-1} \sum_{J \in \mathcal{J}^{(n)}} \mathbb{E}\left[\Sigma_{(J)}^{-\top/2} S \left(\Sigma_{(J)}^{-1/2} (b - b_{(J)}) \right) \right] \qquad (19)$$

and

$$\nabla^2 \mathbb{E}\left[\Psi_n(b)\right] = \binom{n}{m}^{-1} \sum_{J \in \mathcal{J}^{(n)}} \mathbb{E}\left[\Sigma_{(J)}^{-\top/2} S^{(1)} \left(\Sigma_{(J)}^{-1/2} (b - b_{(J)}) \right) \Sigma_{(J)}^{-1/2} \right]. \qquad (20)$$

If, further, covariates x_1, ..., x_n are i.i.d., then we have

$$\mathbb{E}\left(\Sigma_{(J_0)}^{-\top/2} S^{(1)} (\Sigma_{(J_0)}^{-1/2} (\beta_0 - b_{(J_0)})) \Sigma_{(J_0)}^{-1/2} \right) = D(\beta_0) = D \qquad (21)$$

for some positive definite D. The positive definiteness of D is ensured by the assumptions in Remark 3, more details can be found Lemma 5.3 of Chaudhuri (1992). However, if the covariates x_i's are deterministic, we need assume the limit of (19) exists as n tends to infinity and D is positive definite. Analogous to the discussion in Chaudhuri (page 900, 1992), the MTSE $\hat{\beta}_n$ must satisfy

$$\nabla \mathbb{E}\left[\Psi_n(b)\right] = \binom{n}{m}^{-1} \sum_{J \in \mathcal{J}^{(n)}} \mathbb{E}\left[\Sigma_{(J)}^{-\top/2}\, S\left(\Sigma_{(J)}^{-1/2}\left(b - b_{(J)}\right)\right)\right] = 0. \tag{22}$$

Clearly (4) is the sample equation of (22).

From now on, we assume $m = p$ and $m = p - 1$ for the difference-based MTSE's.

Assumption 4. *The random error* ε *and deterministic covariates* x_1, ..., x_n *satisfy*

$$\binom{n}{m}^{-1} \sum_{J \in \mathcal{J}^{(n)}} \lambda_{\min}^{-1}(\Sigma_{(J)}) \mathbb{E}(\{a_{(J)}^2 \varepsilon^{-2}\} \wedge n) = o(n), \tag{23}$$

and

$$\binom{n}{m}^{-1} \sum_{J \in \mathcal{J}^{(n)}} \lambda_{\min}^{-3/2}(\Sigma_{(J)}) = o(n^{1/2}), \tag{24}$$

where $a_{(J)} = \lambda_{\min}^{-1/2}(\Sigma_{(J)}) \lambda_{\max}^{1/2}(\Sigma_{(J)}(X_J^\top X_J))$, $\lambda_{\min}(M)(\lambda_{\max}(M))$ *denotes the smallest (largest) eigenvalue of matrix* M*, and* $a \wedge b = \min(a, b)$.

Since $\mathbb{E}(\{a_{(J)}^2 \varepsilon^{-2}\} \wedge n) = n\, \mathbb{P}(|\varepsilon| \le a_{(J)} n^{-1/2}) + \mathbb{E}(a_{(J)}^2 \varepsilon^{-2} \mathbf{1}[|\varepsilon| > a_{(J)} n^{-1/2}])$, it follows that if the error ε has a density bounded in a neighborhood of the origin, then

$$\mathbb{E}(\{a_{(J)}^2 \varepsilon^{-2}\} \wedge n) \le c\, a_{(J)} n^{1/2},$$

for some constant c. While a mean is highly sensitive to outliers, a median is upset by the mass in the neighborhood of the median. Less mass in a neighborhood will result in more variation of the median estimator. The boundedness of the density in a neighborhood of a median (the origin) is a typical condition that regulates the behavior of the median. Bai, Chen, Miao and Rao (1990) imposed boundedness in a neighborhood of the origin on the error distribution in assumption (a). Therefore, we give the following sufficient condition.

Remark 4. *Suppose that the density of the random error* ε *is bounded in a* *neighborhood of the origin. Then (23) is implied by*

$$\binom{n}{m}^{-1} \sum_{J \in \mathcal{J}^{(n)}} \lambda_{\min}^{-3/2}(\Sigma_{(J)}) \lambda_{\max}^{1/2}(\Sigma_{(J)}(X_J^\top X_J)) = o(n^{1/2}). \tag{25}$$

Conditions (25) and (24) are mild assumptions. In fact, if $\Sigma_{(J)} = I_m$ and the covariates are in the magnitude of $o(n^{1/2})$, then both (25) and (24) are satisfied. An important special case of this situation is that the covariates are *bounded* and $\Sigma_{(J)} = (X_J^\top X_J)^{-1}$, then Assumption 4 is satisfied. We now give our main theorem.

Theorem 2. *Suppose that the error ε is not point mass and $p \geq 2$. Suppose Assumptions 1, 3 and 4 hold. Then $\hat{\beta}_n$ has a bounded influence function $D^{-1}\Sigma_{(J)}^{-\top/2} S$ and satisfies the stochastic expansion:*

$$\hat{\beta}_n = \beta_0 + \binom{n}{m}^{-1} \sum_{J \in \mathcal{J}^{(n)}} D^{-1}\Sigma_{(J)}^{-\top/2} S\left(\Sigma_{(J)}^{-1/2}(b_{(J)} - \beta_0)\right) + o_p(n^{-1/2}). \tag{26}$$

Hence, $n^{1/2}(\hat{\beta}_n - \beta_0) \Longrightarrow \mathcal{N}(0,\ D^{-1}VD^{-\top})$.

I.I.D. Random Covariates. When covariates x_1, ..., x_n are independent and identically distributed random vectors, Wang and *et al.* (2008) have proved the asymptotic normality of the spatial depth-based multivariate Theil-Sen estimators using the theory of U-statistics. With the aid of the convexity lemma (Pollard, 1991) and following the proof of Theorem 2, we can obtain a simple proof for the asymptotic normality of the MTSE. We will introduce the assumptions and give results below with the details of the proof omitted.

Note when x_1, ..., x_n are i.i.d., Assumption 4 is simplified to

$$\mathbb{E}\left(\lambda_{\min}^{-1}(\Sigma_{(J_0)})[\{\lambda_{\min}^{-1}(\Sigma_{(J_0)})\lambda_{\max}(\Sigma_{(J_0)}(X_{J_0}^\top X_{J_0}))\varepsilon^{-2}\} \wedge n]\right) = o(n), \tag{27}$$

and

$$\mathbb{E}\left(\lambda_{\min}^{-3/2}(\Sigma_{(J_0)})\right) = o(n^{1/2}). \tag{28}$$

Remark 4 is reduced to

Remark 5. *Suppose that the density of the random error ε is bounded in a neighborhood of the origin. Then (27) is implied by*

$$\mathbb{E}\left(\lambda_{\min}^{-3/2}(\Sigma_{(J_0)})\lambda_{\max}^{1/2}(\Sigma_{(J_0)}(X_{J_0}^{\top}X_{J_0}))\right) = o(n^{1/2}). \tag{29}$$

We are now ready to give the theorem with the proof omitted.

Theorem 3. *Suppose that the error ε is not point mass and $p \geq 2$. Suppose that the covariates x_1, ..., x_n in multiple linear model (1) are i.i.d. random vectors. Assume Assumptions 1, 3 hold. If (27) and (28) are satisfied, then the multivariate Theil-Sen estimator $\hat{\beta}_n$ has bounded influence function $D_r{}^{-1}\Sigma_{(J)}^{-\top/2}S$ and satisfies the stochastic expansion:*

$$\hat{\beta}_n = \beta_0 + \binom{n}{m}^{-1} \sum_{J \in \mathcal{J}^{(n)}} D_r{}^{-1}\Sigma_{(J)}^{-\top/2} S(\Sigma_{(J)}^{-1/2}(b_{(J)} - \beta_0)) + o_p(n^{-1/2}). \tag{30}$$

Hence, $n^{1/2}(\hat{\beta}_n - \beta_0) \Longrightarrow \mathcal{N}(0,\ D_r{}^{-1}V_r D_r{}^{-\top})$, where $D_r = D(\beta_0), V_r = V(\beta_0)$ computed in (21) and (35).

The Difference-Based MTSE. In a simple linear regression model with deterministic covariates, Peng, Wang and Wang (2008) studied the Theil-Sen estimator under no assumption on the distribution of the error. They showed that the TSE is strongly consistent and has an asymptotic distribution under mild conditions. Wang, *et al.* (2009) extended these results to the spatial depth-based MTSE's and gave the asymptotic normality when covariates are random. Here we will investigate the asymptotic behaviors of the Mahalanobis depth-based MTSE's when covariates are deterministic. Our consideration includes the spatial depth-based MTSE's as a special case.

The proof of the asymptotic normality of $\hat{\beta}_{1,n}$ is analogous to the proof of Theorem 2, and we give the sketches where major differences occur. One of the differences in the proof is that we cannot directly apply Theorem 2.1 of Barbour and Eagleson (1985) to claim the asymptotic normality. What happens is that the sum of random variables defining our estimate $\hat{\beta}_{1,n}$ is not *dissociated*, that is, $b_{(K_1,K_2)}, b_{(K_3,K_4)}$ are not *independent* when the indices $(K_1, K_2), (K_3, K_4)$ are disjoint, while Theorem 2.1 gives the asymptotic normality of a sum of *dissociated random variables*. This technical difficulty can be circumvented by combining the like terms, see details in the proof below. We need the following assumptions.

Assumption 5. *Assumption 3 is satisfied for $\Psi_n(b) = \Psi_{n,d}(b)$ given in (8).*

Under difference model (5), using an analogous reasoning leading (24), (39) and (40) to Assumption 4, we introduce the following assumption.

Assumption 6.

$$\binom{N}{m}^{-1} \sum_{(K_1,K_2)\in\mathcal{K}_0^{(n)}} \lambda_{\min}^{-1}(\Sigma_{(K_1,K_2)}) \times$$
$$\mathbb{E}\left(\{\lambda_{\min}^{-1}(\Sigma_{(K_1,K_2)})\lambda_{\max}(\Sigma_{(K_1,K_2)}(X_{K_1,K_2}^\top X_{K_1,K_2}))(\varepsilon_1 - \varepsilon_2)^{-2}\} \wedge n\right) = o(n)$$

$$\binom{N}{m}^{-1} \sum_{(K_1,K_2)\in\mathcal{K}_0^{(n)}} \lambda_{\min}^{-3/2}(\Sigma_{(K_1,K_2)}) = o(n^{1/2}). \tag{31}$$

Similarly, we give a sufficient condition below.

Remark 6. *Suppose the convolution of the density of the random error ε is bounded in a neighborhood of the origin. Then (31) is implied by*

$$\binom{N}{m}^{-1} \sum_{(K_1,K_2)\in\mathcal{K}_0^{(n)}} \lambda_{\min}^{-3/2}(\Sigma_{(K_1,K_2)})\lambda_{\max}^{1/2}(\Sigma_{(K_1,K_2)}(X_{K_1,K_2}^\top X_{K_1,K_2})) = o(n^{1/2}). \tag{32}$$

With the pairwise differences, the symmetry of the error is still indispensable for the asymptotic normality, see Wang, *et al.* (2009).

Theorem 4. *Suppose $p \geq 3$ and ε is not point mass. Suppose Assumptions 1, 5 and 6 hold. Then $\hat{\beta}_{1,n}$ has a bounded influence function ${D_d}^{-1}\Sigma_{(K_1,K_2)}^{-\top/2}S$ and satisfies the stochastic expansion:*

$$\hat{\beta}_{1,n} = \beta_{1,0} + \binom{N}{m}^{-1} \sum_{(K_1,K_2)\in\mathcal{K}_0^{(n)}} {D_d}^{-1}\Sigma_{(K_1,K_2)}^{-\top/2} S\left(\Sigma_{(K_1,K_2)}^{-1/2}(b_{(K_1,K_2)} - \beta_{1,0})\right) + o_p(n^{-1/2}). \tag{33}$$

Hence, $n^{1/2}(\hat{\beta}_{1,n} - \beta_{1,0}) \Longrightarrow \mathcal{N}(0,\ {D_d}^{-1}V_d{D_d}^{-\top})$, where D_d and V_d are similarly defined as D and V.

If the differences are not overlapped, then the symmetry of the error is no longer required. With a similar proof in Theorem 4, we derive

Theorem 5. *Suppose $p \geq 3$ and ε is not point mass. Suppose analogs of Assumptions 5 and 6 hold. Then $\hat{\beta}^*_{1,n}$ has a bounded influence function $D_*^{-1}\Sigma_{(K_1,K_2)}^{-\top/2}S$ and satisfies the stochastic expansion:*

$$\hat{\beta}^*_{1,n} = \beta_{1,0} + N_{n,m}^{-1} \sum_{(K_1,K_2)\in\mathcal{K}_0^*} D_*^{-1}\Sigma_{(K_1,K_2)}^{-\top/2} S\left(\Sigma_{(K_1,K_2)}^{-1/2}(b_{(K_1,K_2)} - \beta_{1,0})\right) + o_p(n^{-1/2}). \tag{34}$$

Hence, $n^{1/2}(\hat{\beta}^*_{1,n} - \beta_{1,0}) \Longrightarrow \mathcal{N}(0,\ D_*^{-1}V_*D_*^{-\top})$, where D_*, V_* are similarly defined as D, V.

4 Efficiency Consideration

In this section, we compare the efficiency of the MTSE's with that of the LSE when the random errors are from normal distribution.

To calculate the asymptotic covariance matrix of the MTSE, we need compute V and D. Let us first calculate V. To this end, we must calculate the covariance of the average in (17). Note that this average is not a multivariate U-statistic (vector of U-statistic) since the summand in the average $Z_{(J)} = \Sigma_{(J)}^{-\top/2} S(\Sigma_{(J)}^{-1/2}(b - b_{(J)}))$ is not a symmetric kernel in its arguments. Thus we cannot directly apply the formula of the variance of a U-statistic. Instead, we will exploit the technique used in computing the variance of a U-statistic, see, e.g., van der Vaart (1998). Specifically, since $\mathrm{Cov}(Z_{(I)}, Z_{(J)}) = 0$ if $J_1 \cap J_2 = \varnothing$ and the limiting distribution is non-degenerate, the leading terms contributing to the covariance are those $\mathrm{Cov}(Z_{(I)}, Z_{(J)})$ whose indices $I = \{I^1 < ... < I^m\}$ and $J = \{J^1 < ... < J^m\}$ have only one common index. More specifically, for $i, j = 1, ..., m$, we compute those indices (I, J) such that $I^i = J^j$ and $I \cap J = \{I^i\}$. It then can be seen

$$\begin{aligned}
V &= \lim_{n\to\infty} \frac{n}{\binom{n}{m}^2} \sum_{i,j=1}^{m} \sum_{I,J\in J^{(n)},\, I^i=J^j,\, I\cap J=\{I^i\}} \mathbb{E}(Z_I Z_J^\top) \\
&= \lim_{n\to\infty} \frac{n}{\binom{n}{m}^2} \sum_{i,j=1}^{m} \sum_{I^i=i+j-1}^{n-2p+i+j} \Bigg[\binom{I^i-1}{i-1}\binom{I^i-i}{j-1}\binom{n-I^i}{p-i}\binom{n-I^i-p+i}{p-j} \times \\
&\qquad \mathbb{E}[\mathbb{E}(Z_{J_0}|x_i, \varepsilon_i) E(Z_{J_{i,j}}^\top|x_i, \varepsilon_i)] \Bigg]
\end{aligned}$$

where $J_0 = (1, 2, ..., m)$ and $J_{i,j}$ is the m-subset resulting from augmenting $(m+1, ..., 2m-1)$ with i inserted so that the j-th component equal i for $i, j = 1, ... m$. By Karamta theorem,

$$\begin{aligned}
\sum_{l=1}^{n-2m+2} \frac{(l+i+j-3)!}{(l-1)!} \frac{(n-l-i-j+2)!}{(n-l-2m+2)!} &\sim \sum_{l=1}^{n-2m+2} l^{i+j-2}(n-l)^{2m-i-j} \\
&\sim n^{2m-1} \int_0^1 x^{i+j-2}(1-x)^{2m-i-j}\, dx,
\end{aligned}$$

so that we have

$$
\begin{aligned}
&\sum_{I^i=i+j-1}^{n-2m+i+j} \binom{I^i-1}{i-1}\binom{I^i-i}{j-1}\binom{n-I^i}{m-i}\binom{n-I^i-m+i}{m-j} \\
&\sim \frac{n^{2m-1}}{(i-1)!(j-1)!(m-i)!(m-j)!}\int_0^1 x^{i+j-2}(1-x)^{2m-i-j}\,dx \\
&= \frac{n^{2m-1}}{(2m-1)!}\binom{i+j-2}{i-1}\binom{2m-i-j}{m-i}.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
V &= \lim_{n\to\infty} \frac{n^{2m}}{\binom{n}{m}^2(2m-1)!}\sum_{i,j=1}^{m}\binom{i+j-2}{i-1}\binom{2m-i-j}{m-i}\mathbb{E}[\mathbb{E}(Z_{J_0}|x_i,\varepsilon_i)\mathbb{E}(Z_{J_{i,j}}^\top|x_i,\varepsilon_i)] \\
&= \frac{(m!)^2}{(2m-1)!}\sum_{i,j=1}^{m}\binom{i+j-2}{i-1}\binom{2m-i-j}{m-i}\mathbb{E}[\mathbb{E}(Z_{J_0}|x_i,\varepsilon_i)\mathbb{E}(Z_{J_{i,j}}^\top|x_i,\varepsilon_i)] \qquad (35)
\end{aligned}
$$

It can be verified

$$
\frac{(m!)^2}{(2m-1)!}\sum_{i=1}^{m}\sum_{j=1}^{m}\binom{i+j-2}{i-1}\binom{2m-i-j}{m-i} = m^2. \qquad (36)
$$

If Z_J is symmetric in its arguments, then we obtain a U-statistic vector and V indeed boils down to be the usual variance formula of a U-statistic, namely,

$$
V = m^2\mathbb{E}[\mathbb{E}(Z_{J_0}|x_1,\varepsilon_1)\mathbb{E}(Z_{J_0}^\top|x_1,\varepsilon_1)].
$$

Case 1. In this case, we assume the covariates are deterministic and $X_J \equiv \mathbf{I}_p$. Also we assume the errors follow normal distribution with mean 0 and variance σ^2. To keep the notation simple, we assume $\sigma^2 = 1$. We choose the scale matrix $\Sigma_{(J)} = (X_J^\top X_J)^{-1} = \mathbf{I}_p$. Then the proposed MTSE is the spatial median of the $\binom{n}{p}$ least squares estimators. We now compute D and V in Theorem 2.

It can be seen that $\Sigma_J^{-1/2}(\beta_0 - b_J) = -\varepsilon_J$ has a spherically symmetric p-variable normal distribution around the origin. By a similar argument to Brown (1983),

$$
D = \frac{(p-1)\Gamma(\frac{p-1}{2})}{p\sqrt{2}\Gamma(\frac{p}{2})}\mathbf{I}_p.
$$

Now compute V via (35). Since $Z_J = -\varepsilon_J/\|\varepsilon_J\|$, it follows

$$\mathbb{E}(Z_{J_0}|x_i,\,\varepsilon_i) = -\mathbb{E}(\varepsilon_i\|\varepsilon_{J_0}\|^{-1}\,|\varepsilon_i)\mathbf{1}_i, \quad \mathbb{E}(Z_{J_{i,j}}^\top|x_i,\varepsilon_i) = -\mathbb{E}(\varepsilon_i\|\varepsilon_{J_0}\|^{-1}\,|\varepsilon_i)\mathbf{1}_j^\top.$$

where $\mathbf{1}_i$ is the p-dimensional vector with all zero components except for the i-th unit component. Hence,

$$\begin{aligned}\mathbb{E}[\mathbb{E}(Z_{J_0}|x_i,\varepsilon_i)\mathbb{E}(Z_{J_{i,j}}^\top|x_i,\varepsilon_i)] =& \mathbb{E}[\{\mathbb{E}(\varepsilon_i\|\varepsilon_{J_0}\|^{-1}|\varepsilon_i)\}^2]\mathbf{1}_i\mathbf{1}_j^\top \\ \leq& \mathbb{E}[\{\mathbb{E}(\varepsilon_i^2\|\varepsilon_{J_0}\|^{-2}|\varepsilon_i)\}]\mathbf{1}_i\mathbf{1}_j^\top = p^{-1}\mathbf{1}_i\mathbf{1}_j^\top\end{aligned}$$

and $V = (V_{i,j})_{p\times p}$ with

$$V_{i,j} \leq \frac{(p!)^2\binom{i+j-2}{i-1}\binom{2p-i-j}{p-i}}{p\,(2p-1)!}.$$

Hence, the asymptotic covariance matrix of the proposed MTSE is

$$D^{-1}VD^{-T} = \left(\frac{(p-1)\Gamma(\frac{p-1}{2})}{p\sqrt{2}\Gamma(\frac{p}{2})}\right)^{-2} V.$$

Define $\|M\|_a = \sum_{i,j=1}^m |M_{i,j}|$ as the norm of the covariance matrix M. Notice that this norm is more conservative in a sense that $\|M\|_a \geq GV(M)$, where $GV(M)$ is the Generalized Variance that is defined to be the sum of variances of each random variables. By (36),

$$\begin{aligned}\|D^{-1}VD^{-\top}\|_a =& \left(\frac{(p-1)\Gamma(\frac{p-1}{2})}{p\sqrt{2}\Gamma(\frac{p}{2})}\right)^{-2}\sum_{i,j=1}^p V_{i,j} \\ \leq& \left(\frac{(p-1)\Gamma(\frac{p-1}{2})}{p\sqrt{2}\Gamma(\frac{p}{2})}\right)^{-2}\sum_{i,j=1}^p \frac{(p!)^2\binom{i+j-2}{i-1}\binom{2p-i-j}{p-i}}{p\,(2p-1)!} \\ =& p\left(\frac{(p-1)\Gamma(\frac{p-1}{2})}{p\sqrt{2}\Gamma(\frac{p}{2})}\right)^{-2}\end{aligned} \quad (37)$$

The covariance matrix of the least square estimator is $p\mathbf{I}_p$ with $\|p\mathbf{I}_p\|_a = p^2$. The asymptotic relative efficiency (ARE) of the MTSE relative to the LSE is the ratio of the norm of the covariance matrix of the LSE to the MTSE. Then it can be seen that the ARE satisfies

$$ARE(p) = \frac{\|p\mathbf{I}_p\|_a}{\|D^{-1}VD^{-\top}\|_a} \geq \frac{1}{p}\left(\frac{(p-1)\Gamma(\frac{p-1}{2})}{p\sqrt{2}\Gamma(\frac{p}{2})}\right)^2 \to 1, \quad p \to \infty.$$

As in Brown (1983) and Chaudhuri (1992), the ARE increases with the number of regression parameters. Table 1 reports the above lower bound for several values of p.

Table 1

p	2	3	4	5	6	7	8	9	10
ARE $\geq$	0.785	0.849	0.884	0.905	0.920	0.931	0.946	0.951	0.956

Case 2. In this case, we assume that the covariates x_n's are i.i.d random vectors and the errors ε_i's follow the standard normal distribution. This case was studied by Wang, *et al.* (2009) and corresponds to our Theorem 3. We now calculate the D_r, V_r in Theorem 3. It can be seen that under the true model, $\Sigma_J^{-1/2}(\beta_0 - b_J) = -(X_J^\top X_J)^{-1/2}X_J^\top \varepsilon_J$ has a spherically symmetric p-variable normal distribution around the origin. Again as in case 1, with a similar argument in Brown (1983), we have

$$D_r = \frac{(p-1)\Gamma(\frac{p-1}{2})}{p\sqrt{2}\Gamma(\frac{p}{2})}\mathbb{E}(X_{J_0}^\top X_{J_0}) = p\frac{(p-1)\Gamma(\frac{p-1}{2})}{p\sqrt{2}\Gamma(\frac{p}{2})}\mathbb{E}(x_1x_1^\top).$$

Since the covariate X and error ε are uncorrelated. Since $Z_J = X_J^\top \varepsilon_J/\|\varepsilon_J\|$, it follows

$$\begin{aligned}\mathbb{E}(Z_{J_0}|x_i, \varepsilon_i) &= -\mathbb{E}(X_{J_0}^\top|x_i)\mathbb{E}(\varepsilon_{J_0}\|\varepsilon_{J_0}\|^{-1}|\varepsilon_i) \\ &= -\mathbb{E}\left((x_1,\, ...,\, x_p)|x_i\right)\mathbb{E}(\varepsilon_i\|\varepsilon_{J_0}\|^{-1}|\varepsilon_i)\mathbf{1}_i = -\mathbb{E}(\varepsilon_i\|\varepsilon_{J_0}\|^{-1}|\varepsilon_i)x_i.\end{aligned}$$

Hence,

$$\mathbb{E}[\mathbb{E}(Z_{J_0}|x_i, \varepsilon_i)\mathbb{E}(Z_{J_{i,j}}^\top|x_i, \varepsilon_i)] = p^2\mathbb{E}[(\mathbb{E}(\varepsilon_i\|\varepsilon_{J_0}\|^{-1}|\varepsilon_i))^2]\mathbb{E}(x_1x_1^\top)$$

By (35) and (36), we have

$$V_r = p^2\mathbb{E}[(\mathbb{E}(\varepsilon_i\|\varepsilon_{J_0}\|^{-1}|\varepsilon_i))^2]\mathbb{E}(x_1x_1^\top).$$

Hence, the asymptotic covariance matrix of the MTSE is

$$D_r^{-1}V_rD_r^{-\top} = \left(\frac{(p-1)\Gamma(\frac{p-1}{2})}{p\sqrt{2}\Gamma(\frac{p}{2})}\right)^{-2}\mathbb{E}[(\mathbb{E}(\varepsilon_i\|\varepsilon_{J_0}\|^{-1}|\varepsilon_i))^2]\Big(\mathbb{E}(x_1x_1^\top)\Big)^{-1}.$$

so that the norm is

$$\begin{aligned}\|D_r^{-1}V_rD_r^{-\top}\|_a &= \left(\frac{(p-1)\Gamma(\frac{p-1}{2})}{p\sqrt{2}\Gamma(\frac{p}{2})}\right)^{-2}\mathbb{E}[(\mathbb{E}(\varepsilon_i\|\varepsilon_{J_0}\|^{-1}|\varepsilon_i))^2]\,\|(\mathbb{E}(x_1x_1^\top))^{-1}\|_a \\ &\le \frac{1}{p}\left(\frac{(p-1)\Gamma(\frac{p-1}{2})}{p\sqrt{2}\Gamma(\frac{p}{2})}\right)^{-2}\|(E(x_1x_1^\top))^{-1}\|_a. \end{aligned} \tag{38}$$

The covariance matrix of the LSE is $n(\mathbb{E}(X_n^\top X_n))^{-1} = (\mathbb{E}(x_1x_1^\top))^{-1}$ and its norm is $\|(\mathbb{E}(x_1x_1^\top))^{-1}\|_a$. Thus, we have obtained the same conclusion as in Case 1. That is, the asymptotic relative efficiency of the MTSE relative to the LSE satisfies

$$ARE \ge \frac{1}{p}\left(\frac{(p-1)\Gamma(\frac{p-1}{2})}{p\sqrt{2}\Gamma(\frac{p}{2})}\right)^{-2}.$$

5 Proofs

In this section, we provide the proofs.

Proof of Theorem 2. Let

$$R_n(\alpha) = n\{\Psi_n(\beta_0 + n^{-1/2}\alpha) - \Psi_n(\beta_0) - n^{-1/2}\alpha^\top\nabla\Psi_n(\beta_0)\}, \quad \alpha \in \mathbb{R}^p,$$

We show next $\mathbb{V}\mathrm{ar}(R_n(\alpha)) \to 0$ for each $\alpha \in \mathbb{R}^p$. Let $\alpha_n = n^{-1/2}\alpha$ and denote $\gamma_{(J)} = \|\Sigma_{(J)}^{-1/2}(\beta_0 - b_{(J)})\|$ and $\delta_{(J)} = \|\Sigma_{(J)}^{-1/2}(\alpha_n + \beta_0 - b_{(J)})\|$. Using $\|a+b\| - \|b\| = (\|a+b\|^2 - \|b\|^2)/(\|a+b\| + \|b\|)$, we have

$$
\begin{aligned}
R_n(\alpha) = n\binom{n}{m}^{-1} & \sum_{J\in\mathcal{J}^{(n)}} \left(\delta_{(J)} - \gamma_{(J)} - \alpha_n^\top \Sigma_{(J)}^{-\top/2} S(\Sigma_{(J)}^{-1/2}(\beta_0 - b_{(J)}))\right) \\
= n\binom{n}{m}^{-1} & \sum_{J\in\mathcal{J}^{(n)}} \Big(\frac{\|\Sigma_{(J)}^{-1/2}\alpha_n\|^2 + 2\alpha_n^\top \Sigma_{(J)}^{-1}(\beta_0 - b_{(J)})}{\delta_{(J)} + \gamma_{(J)}} \\
& \qquad - \alpha_n^\top \Sigma_{(J)}^{-\top/2} S(\Sigma_{(J)}^{-1/2}(\beta_0 - b_{(J)}))\Big) \\
= n\binom{n}{m}^{-1} & \sum_{J\in\mathcal{J}^{(n)}} \left(\frac{\|\Sigma_{(J)}^{-1/2}\alpha_n\|^2}{\delta_{(J)} + \gamma_{(J)}} + \frac{\gamma_{(J)} - \delta_{(J)}}{\delta_{(J)} + \gamma_{(J)}}\alpha_n^\top \Sigma_{(J)}^{-\top/2} S(\Sigma_{(J)}^{-1/2}(\beta_0 - b_{(J)}))\right) \\
= n\binom{n}{m}^{-1} & \sum_{J\in\mathcal{J}^{(n)}} C_{(J)}, \quad \text{say.}
\end{aligned}
$$

Clearly $\mathrm{Cov}(C_{(J_1)}, C_{(J_2)}) = 0$ if $J_1 \cap J_2 = \emptyset$, and $2\,\mathrm{Cov}(C_{(J_1)}, C_{(J_2)}) \leq \mathrm{Var}(C_{(J_1)}) + \mathrm{Var}(C_{(J_2)})$, $J_1, J_2 \in \mathcal{J}^{(n)}$. Hence, for large n,

$$
\begin{aligned}
\mathbb{V}\mathrm{ar}(R_n(\alpha)) &\leq n^2\binom{n}{m}^{-2}\sum_{i=1}^{m}\binom{m}{i}\binom{n-m}{m-i}\sum_{J\in\mathcal{J}^{(n)}}\mathbb{V}\mathrm{ar}(C_{(J)})) \\
&\leq cn\binom{n}{m}^{-1}\sum_{J\in\mathcal{J}^{(n)}}\mathbb{E}(C_{(J)}^2),
\end{aligned}
$$

where c is a constant independent of n. Since the spatial sign function $S(\cdot)$ is bounded by one and $|\gamma_{(J)} - \delta_{(J)}| \leq \|\Sigma_{(J)}^{-1/2}\alpha_n\|$, it follows

$$
|C_{(J)}| \leq 2\frac{\|\Sigma_{(J)}^{-1/2}\alpha_n\|^2}{\gamma_{(J)} + \delta_{(J)}} \leq 2\frac{\|\Sigma_{(J)}^{-1/2}\alpha_n\|^2}{\gamma_{(J)} \vee \|\Sigma_{(J)}^{-1/2}\alpha_n\|}.
$$

Accordingly, $\mathbb{V}\mathrm{ar}(R_n(\alpha)) \to 0, \alpha \in \mathbb{R}^q$ if we show

$$
\binom{n}{m}^{-1}\sum_{J\in\mathcal{J}^{(n)}}\lambda_{\min}^{-1}(\Sigma_{(J)})\mathbb{E}\left(\{\lambda_{\min}^{-1}(\Sigma_{(J)})\gamma_{(J)}^{-2}\} \wedge n\right) = o(n). \tag{39}
$$

Under the true model, $b_{(J)} = \beta_0 + (X_J^\top X_J)^{-1}X_J^\top \varepsilon_J$. Let Γ be an orthogonal matrix such that $\Gamma X_J(X_J^\top X_J)^{-1}\Sigma_{(J)}^{-1}(X_J^\top X_J)^{-1}X_J^\top\Gamma^\top = \Lambda$, where Λ is a diagonal matrix whose diagonal elements are the eigenvalues of $M_{(J)} = X_J(X_J^\top X_J)^{-1}\Sigma_{(J)}^{-1}(X_J^\top X_J)^{-1}X_J^\top$. Then

$$
\gamma_{(J)}^2 = \|\Sigma_{(J)}^{-1/2}(\beta_0 - b_{(J)})\|^2 = \varepsilon_J^\top M_{(J)}\varepsilon_J = \varepsilon_J^\top\Gamma^\top\Lambda\Gamma\varepsilon_J \geq \lambda_{\min}(M_{(J)})\|\varepsilon_J\|^2.
$$

Therefore, in view of $\|\varepsilon_J\| \geq |\varepsilon_j|$ for $j \in J$, (39) follows from

$$\binom{n}{m}^{-1} \sum_{J \in \mathcal{J}^{(n)}} \lambda_{\min}^{-1}(\Sigma_{(J)}) \mathbb{E}\left(\{\lambda_{\min}^{-1}(\Sigma_{(J)}) \lambda_{\min}^{-1}(M_{(J)}) \varepsilon_1^{-2}\} \wedge n\right) = o(n), \qquad (40)$$

provided that covariates X_j's and errors ε_j's are uncorrelated, which holds if the covariates are deterministic. Since $\lambda_{\min}^{-1}(M_{(J)}) = \lambda_{\min}^{-1}((X_J^\top X_J)^{-1}\Sigma_{(J)}^{-1}) = \lambda_{\max}(\Sigma_{(J)}(X_J^\top X_J))$, it follows that (40) is implied by (23). Hence $R_n(\alpha) - \mathbb{E}(R_n(\alpha)) \to 0$ in probability for $\alpha \in \mathbb{R}^q$. By Assumption 3,

$$\mathbb{E}(R_n(\alpha)) = \frac{1}{2}\alpha^\top D\alpha + o(1), \qquad (41)$$

where D is given in Assumption 3. Therefore,

$$n\{\Psi_n(\beta_0 + n^{-1/2}\alpha) - \Psi_n(\beta_0) - n^{-1/2}\alpha^\top \nabla\Psi_n(\beta_0)\} = \frac{1}{2}\alpha^\top D\alpha + o_p(1), \ \alpha \in \mathbb{R}^p.$$

Since the summand in $\Psi_n(\beta)$ is convex in β, it follows from the convexity lemma (Pollard, 1991) that for any $M > 0$,

$$\sup_{\|\alpha\| \leq M} \left| n\{\Psi_n(\beta_0 + n^{-1/2}\alpha) - \Psi_n(\beta_0) - n^{-1/2}\alpha^\top \nabla\Psi_n(\beta_0)\} - \frac{1}{2}\alpha^\top D\alpha \right| = o_p(1), \qquad (42)$$

Let $\Delta_n(\alpha) = n[\Psi_n(\beta_0 + n^{-1/2}\alpha) - \Psi_n(\beta_0)]$ and $\hat{\alpha}_n = \arg\min_{\alpha \in \mathbb{R}^q} \Delta_n(\alpha)$. Then $\hat{\alpha}_n = n^{1/2}(\hat{\beta}_n - \beta_0)$. Further, for any random variable γ_n bounded in probability, it follows from (42) that

$$\Delta_n(\gamma_n) = \gamma_n^\top n^{1/2}\nabla\Psi_n(\beta_0) + \frac{1}{2}\gamma_n^\top D\gamma_n + o_p(1). \qquad (43)$$

This shows that $\Delta_n(\gamma_n)$ can be approximated by a quadratic function in γ_n, which is uniquely minimized at $\hat{\gamma}_n = -D^{-1}n^{1/2}\nabla\Psi_n(\beta_0)$. Like Shen (2009), we shall use the convexity of the criterion function and characterization of a minimizer to show that $\hat{\gamma}_n$ and $\hat{\alpha}_n$ are equivalent. Arbitrarily fix $\epsilon > 0$. If $\|\hat{\alpha}_n - \hat{\gamma}_n\| > \epsilon$, then there exists $\hat{\gamma}_n^*$ on the line segment joining $\hat{\alpha}_n$ and $\hat{\gamma}_n$ such that $\|\hat{\gamma}_n^* - \hat{\gamma}_n\| = \epsilon$. Clearly $\hat{\gamma}_n^* = O_p(1)$. Substituting it in (43), we obtain

$$\Delta_n(\hat{\gamma}_n^*) = \Delta_n(\hat{\gamma}_n) + \frac{1}{2}(\hat{\gamma}_n^* - \hat{\gamma}_n)^\top D(\hat{\gamma}_n^* - \hat{\gamma}_n) + o_p(1).$$

Since $\Delta_n(\cdot)$ is convex and $\Delta_n(\hat{\alpha}_n) \leq \Delta_n(\hat{\gamma}_n)$, it follows $\Delta_n(\hat{\gamma}_n^*) \leq \Delta_n(\hat{\gamma}_n)$, so that

$$\frac{1}{2}\epsilon^2\lambda_{\min}(D) + o_p(1) \leq \frac{1}{2}(\hat{\gamma}_n^* - \hat{\gamma}_n)^\top D(\hat{\gamma}_n^* - \hat{\gamma}_n) + o_p(1) \leq 0.$$

This shows

$$\mathbb{P}(\|\hat{\alpha}_n - \hat{\gamma}_n\| > \epsilon) \leq \mathbb{P}\left(\Delta_n(\hat{\gamma}_n^*) \leq \Delta_n(\hat{\gamma}_n)\right) \leq \mathbb{P}\Big(\frac{1}{2}\epsilon^2\lambda_{\min}(D) + o_p(1) \leq 0\Big).$$

The last probability converges to zero as n tends to infinity by Assumption 3, and this gives the desired equivalence $\hat{\alpha}_n - \hat{\gamma}_n \to 0$ in probability. We now apply the central limit theorem of U-statistics to determine the asymptotic distribution of $\hat{\gamma}_n$. By the Cramér-Wold device and Assumption 3, it suffices to show

$$c^\top n^{1/2}\nabla\Psi_n(\beta_0) \Longrightarrow \mathcal{N}(0, c^\top V c) \tag{44}$$

for every $c \in \mathbb{R}^p$. We shall apply Theorem 2.1 of Barbour and Eagleson (1985) to prove (44). To this end, we verify their condition (2.7) for

$$X_J^{(n)} := n^{1/2}\binom{n}{m}^{-1} c^\top \Sigma_{(J)}^{-\top/2} S(\Sigma_{(J)}^{-1/2}(\beta - b_{(J)})).$$

Let $(s^{(n)})^2 := \mathbb{V}\mathrm{ar}[c^\top n^{1/2}\nabla\Psi_n(\beta_0)] = \mathbb{V}\mathrm{ar}\left[\sum_{J\in J^{(n)}} X_J^{(n)}\right]$. Then by Assumption 3, $(s^{(n)})^2 \to c^\top V c$ as n tends to infinity. This together with (24) gives

$$\begin{aligned} n^{2m-2}(s^{(n)})^{-3}\sum_{J\in J^{(n)}} E|X_J^{(n)}|^3 &\leq n^{2m-2}(s^{(n)})^{-3} n^{3/2}\binom{n}{m}^{-3}\sum_{J\in J^{(n)}}\|\Sigma_{(J)}^{-1/2}c\|^3 \\ &= O\Big(n^{-1/2}\binom{n}{m}^{-1}\sum_{J\in J^{(n)}}\lambda_{\min}^{-3/2}(\Sigma_{(J)})\Big) \to 0. \end{aligned}$$

This verifies their (27) and hence shows the desired (44). □

Proof of Theorem 4. First, as in the proof of Theorem 2, we have for $\alpha \in \mathbb{R}^q$ that

$$\begin{aligned} R_{n,d}(\alpha) &:= n\{\Psi_{n,d}(\beta_{1,0} + n^{-1/2}\alpha) - \Psi_{n,d}(\beta_0) - n^{-1/2}\alpha^\top \nabla \Psi_{n,d}(\beta_{1,0})\}, \\ &= \frac{n}{\binom{N}{m}} \sum_{(K_1,K_2)\in\mathcal{K}_0^{(n)}} \Bigg(\frac{\|\Sigma_{(K_1,K_2)}^{-1/2}\alpha_n\|^2}{\gamma_{(K_1,K_2)} + \delta_{(K_1,K_2)}} \\ &\qquad + \frac{\alpha_n^\top \Sigma_{(K_1,K_2)}^{-\top/2}\Sigma_{(K_1,K_2)}^{-1/2}(\beta_0 - b_{(K_1,K_2)})(\gamma_{K_1,K_2} - \delta_{(K_1,K_2)})}{(\gamma_{(K_1,K_2)} + \delta_{(K_1,K_2)})\gamma_{(K_1,K_2)}} \Bigg), \end{aligned}$$

where $\gamma_{(K_1,K_2)} = \|\Sigma_{(K_1,K_2)}^{-1/2}(\beta_{1,0} - b_{(K_1,K_2))}\|$ and $\delta_{(K_1,K_2)} = \|\Sigma_{(K_1,K_2)}^{-1/2}(\alpha_n + \beta_{1,0} - b_{(K_1,K_2)})\|$. Denote the summand by $C_{(K_1,K_2)}$. Then $\mathrm{Cov}(C_{(K_1,K_2)}, C_{(K_1',K_2')}) = 0$ if $\{K_1, K_2\} \cap \{K_1', K_2'\} = \emptyset$, and $2\,\mathrm{Cov}(C_{(K_1,K_2)}, C_{(K_1',K_2')}) \le \mathrm{Var}(C_{(K_1,K_2)}) + \mathrm{Var}(C_{(K_1',K_2')})$ for $(K_1,K_2), (K_1',K_2') \in \mathcal{K}_0^{(n)}$. Since for each $(K_1,K_2) \in \mathcal{K}_0^{(n)}$, there are at most $(2n - 2m - 1)m$ possible (j,k)'s from $\triangledown$ that $\{j,k\} \cap \{K_1,K_2\} \neq \emptyset$, it follows that $\mathbb{V}\mathrm{ar}(R_n(\alpha))$ is bounded by

$$\begin{aligned} &\frac{n^2}{\binom{N}{m}^2} \sum_{i=1}^{m} \binom{(2n-2m-1)m}{i} \binom{N - (2n-2m-1)m}{m-i} \sum_{(K_1,K_2)\in\mathcal{K}_0^{(n)}} \mathbb{V}\mathrm{ar}(C_{(K_1,K_2)}) \\ \le\, & cn \binom{N}{m}^{-1} \sum_{(K_1,K_2)\in\mathcal{K}_0^{(n)}} \mathbb{E}(C_{(K_1,K_2)}^2), \end{aligned}$$

for large n, where c is a constant independent of n. Arguing as in (39), we derive $\mathbb{V}\mathrm{ar}(R_{n,d}(\alpha)) \to 0$ for $\alpha \in \mathbb{R}^q$ if we can show

$$\begin{aligned} &\binom{N}{m}^{-1} \sum_{(K_1,K_2)\in\mathcal{K}_0^{(n)}} \lambda_{\min}^{-1}\left(\Sigma_{(K_1,K_2)}^{1/2}\right) \times \\ &\qquad \mathbb{E}\left(\{\lambda_{\min}^{-1}\left(\Sigma_{(K_1,K_2)}^{1/2}\right) \|\Sigma_{(K_1,K_2)}^{-1/2}(\beta_{1,0} - b_{(K_1,K_2)})\|^{-2}\} \wedge n\right) = o(n). \qquad (45) \end{aligned}$$

Under the true model (5), $b_{(K_1,K_2)} = \beta_{1,0} + ({X_{K_1,K_2}}^\top X_{K_1,K_2})^{-1} {X_{K_1,K_2}}^\top \varepsilon_{K_1,K_2}$, so that

$$\|\Sigma_{(K_1,K_2)}^{-1/2}(\beta_{1,0} - b_{(K_1,K_2)})\|^2 \ge \lambda_{\min}(M_{(K_1,K_2)})\|\varepsilon_{K_1,K_2}\|^2,$$

where $M_{(K_1,K_2)} = (X_{K_1,K_2}{}^\top X_{K_1,K_2})^{-1}\Sigma^{-1}_{(K_1,K_2)}$. Further,

$$\lambda^{-1}_{\min}(M_{(K_1,K_2)}) = \lambda_{\max}(\Sigma_{(K_1,K_2)}(X_{K_1,K_2}{}^\top X_{K_1,K_2})).$$

In view of $\|\varepsilon_{K_1,K_2}\| > |\varepsilon_i - \varepsilon_j|$ for some $i \in k_1$, $j \in k_2$ with $i \neq j$, we see that (45) is implied by (31). Analogously, we derive $\hat{\alpha}_n - \hat{\gamma}_n \to 0$ in probability, where $\hat{\alpha}_n = n^{1/2}(\hat{\beta}_{1,n} - \beta_{1,0})$ for $\hat{\beta}_{1,n}$ given in Section 2, and

$$\begin{aligned}\hat{\gamma}_n &= -D^{-1}n^{1/2}\nabla\Psi_{n,d}(\beta_{1,0})\\ &= -D^{-1}n^{1/2}\binom{N}{m}^{-1}\sum_{(K_1,K_2)\in\mathcal{K}_0^{(n)}}\Sigma^{-\top/2}_{(K_1,K_2)}S\left(\Sigma^{-1/2}_{(K_1,K_2)}(\beta_{1,0} - b_{(K_1,K_2)})\right).\end{aligned}$$

Next we will prove the asymptotic normality of $\hat{\gamma}_n$, so that we obtain the asymptotic normality of $\hat{\beta}_{1,n}$. First, let $\mathcal{I}^{(n)}$ be the collection of the $2m$-subsets of $\{1, ..., n\}$. There are in total $\binom{n}{2m}$ such subsets. Then, we do a matching, for each $(K_1, K_2) \in \mathcal{K}_0^{(n)}$, we match it with one of the $2m$-subset $I \in \mathcal{I}^{(n)}$ such that $K_1 \cup K_2 \subseteq I$, and we denote the match by $(K_1, K_2) \hookrightarrow I$. Note, for each I, the number of (K_1, K_2) that are matched is bounded by a constant, $c(m)$, depending only upon m. Now for any $\mathbf{c} \in \mathbb{R}^{p-1}$, we have

$$\begin{aligned}&\mathbf{c}^\top n^{1/2}\nabla\Psi_{n,d}(\beta_0)\\ &= n^{1/2}\binom{N}{m}^{-1}\sum_{(K_1,K_2)\in\mathcal{K}_0^{(n)}}\mathbf{c}^\top\Sigma^{-\top/2}_{(K_1,K_2)}S\left(\Sigma^{-1/2}_{(K_1,K_2)}(b - b_{(K_1,K_2)})\right)\\ &= \sum_{I\in\mathcal{I}^{(n)}} n^{1/2}\binom{N}{m}^{-1}\sum_{\substack{(K_1,\,K_2)\,\in\,\mathcal{K}_0^{(n)}\\ (K_1,\,K_2)\,\hookrightarrow\, I}}\mathbf{c}^\top\Sigma^{-\top/2}_{(K_1,K_2)}S\left(\Sigma^{-1/2}_{(K_1,K_2)}(b - b_{(K_1,K_2)})\right)\\ &\equiv \sum_{I\in\mathcal{I}^{(n)}} W_I, \quad \text{say.}\end{aligned}$$

The last sum is a sum of *dissociated random variables*, that is, W_I and W_J are independent if I and J are disjoint. Thus, we can apply Theorem 2.1 of Barbour and Eagleson (1985) to prove the asymptotic normality of $\hat{\gamma}_n$. To do so, it suffices to verify their condition (2.7) for $X_I^{(n)} = W_I$ and $k = 2m$. By Assumption 5, $(s^{(n)})^2 \equiv \mathbb{V}\mathrm{ar}[\mathbf{c}^\top n^{1/2}\nabla\Psi_{n,d}(\beta_0)] \to \mathbf{c}^\top V\mathbf{c}$. Thus,

$$\begin{aligned}
& n^{2(2m)-2}\{s^{(n)}\}^{-3}\sum_{I\in\mathcal{I}^{(n)}}|W_I|^3 \\
&\quad = n^{4m-2}\{s^{(n)}\}^{-3}n^{3/2}\frac{1}{\binom{N}{m}^3}\sum_{I\in\mathcal{I}^{(n)}}\Bigg|\sum_{\substack{(K_1,K_2)\in\mathcal{K}_0^{(n)}\\(K_1,K_2)\hookrightarrow I}}\mathbf{c}^\top\Sigma_{(K_1,K_2)}^{-\top/2}\times \\
&\qquad\qquad S\left(\Sigma_{(K_1,K_2)}^{-1/2}(\beta_{1,0}-b_{(K_1,K_2)})\right)\Bigg|^3 \\
&\quad \le cn^{4m-2}\{s^{(n)}\}^{-3}n^{3/2}\frac{1}{\binom{N}{m}^3}\sum_{I\in\mathcal{I}^{(n)}}\sum_{\substack{(K_1,K_2)\in\mathcal{K}_0^{(n)}\\(K_1,K_2)\hookrightarrow I}}\Bigg|\mathbf{c}^\top\Sigma_{(K_1,K_2)}^{-\top/2}\times \\
&\qquad\qquad S\left(\Sigma_{(K_1,K_2)}^{-1/2}(\beta_{1,0}-b_{(K_1,K_2)})\right)\Bigg|^3 \\
&\quad \le cn^{4m-2}\{s^{(n)}\}^{-3}n^{3/2}\frac{1}{\binom{N}{m}^3}\sum_{(K_1,K_2)\in\mathcal{K}_0^{(n)}}\lambda_{\min}^{-3/2}\left(\Sigma_{(K_1,K_2)}\right) \\
&= O\Bigg(n^{-1/2}\frac{1}{\binom{N}{m}}\sum_{(K_1,K_2)\in\mathcal{K}_0^{(n)}}\lambda_{\min}^{-3/2}\left(\Sigma_{(K_1,K_2)}\right)\Bigg) \to 0,
\end{aligned}$$

by Assumption 6. This verifies their condition (2.7), so that we obtain the asymptotic normality of $\mathbf{c}^\top n^{1/2}\nabla\Psi_{n,d}(\beta_{1,0})$ for any $\mathbf{c} \in \mathbb{R}^{p-1}$ and hence $\hat{\gamma}_n$. Combining with the established equivalence $\hat{\alpha}_n - \hat{\gamma}_n = n^{1/2}(\hat{\beta}_{1,n} - \beta_{1,0}) - \hat{\gamma}_n \to 0$ in probability, we obtain the desired asymptotic normality of the difference-based MTSE $\hat{\beta}_{1,n}$. □

References

[1] Andersen, P.K. and Gill, R.D. (1982). Cox's Regression Model for Counting Processes: A Large Sample Study. *Ann. Statist.* **10**, 1100–1120.

[2] Bai, Z.D., Chen, X.R., Miao, B.Q. and Rao, C.R. (1990). Asymptotic Theory of Least Distance Estimate in Multivariate Linear Models. *J. Amer. Statist. Assoc.* **90**, 170 – 177.

[3] 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.

[4] Barbour, A.D. and Eagleson, G.K. (1985). Multiple Comparisons and Sums of Dissociated Random Variables. *Adv. Appl. Prob.* **17**, 147 – 185.

[5] Brown, B.M. (1983). Statistical use of the spatial median. *J. Roy. Statist. Soc. Ser. B* **45**, 25 – 30.

[6] Busarova, D., Tyurin, Y., Mottonen, j. and Oja, H. (2006). Multivariate Theil estimator with the corresponding test. *Math. Meth. Statist.* **15**, 1–19.

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

[8] Chaudhuri, P. (1992). Multivariate Location Estimation Using Extension of R-Estimates Through U-Statistics Type Approach. *Ann. Statist.* **20**, 897 – 916.

[9] Dietz, E. J. (1989). Teaching Regression in a Nonparametric Statistics Course. *Amer. Statist.* **43**, 35 – 40.

[10] 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.

[11] Hollander, M. and Wolfe, D. A. (1999). *Nonparametric Statistic Methods*, 2nd Ed. John Wiley & Sons, NY. 421 – 423.

[12] Kemperman (1987). The Median of a Finite Measure on a Banach Space. *Probability and Related Methods.* Elsevier Science Publishers B.V. (North-Holland). 217 – 230.

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

[14] Peng, H., Wang, S. and Wang, X. (2008). Consistency and asymptotic distri-

bution of the Theil-Sen estimator. *J. Statist. Plann. Inferr.* **138**, 1836 – 1850.

[15] Pollard, D. (1991). Asymptotics for Least Absolute Deviation Regression Estimators. *Econometric Theory* **7**, 186 – 199.

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

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

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

[19] Shen, G. (2008). Asymptotics of Oja Median Estimate. *Statist. & Probabil. Lett.* **78**, 2137-2141.

[20] Shen, G. (2009). Asymptotics of A Theil-type Estimate in Linear Regression. *Stat & Prob Lett.* **79**, 1053-1064.

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

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

[24] 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.

[24] Wang, X., Dang, X., Peng, H. and Zhang, H. (2009). The Theil-Sen estimators in multiple linear regression models. *Manuscript*, Available at URL: *hpeng@math.iupui.edu*.

[25] Wilcox, R. (1998). Simulations on the Theil-Sen regression estimator with right-censored data. *Statist. & Probabil. Lett.* **39**, 43 – 47.

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

[27] Zhou, W. and Serfling, R. (2008). Multivariate spatial U-quantiles: a Bahadur-Kiefer representation, a Theil-Sen estimator for multiple regression, and a robust dispersion estimator. *J. Statist. Plann. Inferr.* **138**, 1660 – 1678.