Biometrika (2022), **00**, 0, *p.* 1 doi: 10.1093/biomet/asac027

Supplementary material for "Linearized Maximum Rank Correlation Estimation"

BY GUOHAO SHEN

Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

guohao.shen@polyu.edu.hk

KANI CHEN

Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

makchen@ust.hk

JIAN HUANG

Department of Statistics and Actuarial Science, University of Iowa, Schaeffer Hall, Iowa City, Iowa, U.S.A

jian-huang@uiowa.edu

YUANYUAN LIN

Department of Statistics, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong

ylin@sta.cuhk.edu.hk

This supplementary material contains numerical algorithms for computing the proposed general class of estimators studied in Section 2.3 and the penalized estimators studied in Section 4 in the main context, lemmas and technical proofs for the main theorems, as well as some additional simulation results.

1. NUMERICAL ALGORITHMS

We describe two numerical algorithms to compute $\hat{\beta}_n^g$ studied in Section 2.3 in the main context, depending on the differentiability of g. We focus on the case that Σ is unknown.

When g is differentiable, the fixed-point iteration algorithm can be used. Recall that $L_n^g(\beta) \equiv \sum_{i\neq j}^n I(Y_i < Y_j) g\{(X_j - X_i)^{\mathrm{T}}\beta\}/\{n(n-1)\}$. Define $U_n^g(\beta) = \nabla L_n^g(\beta) \equiv \sum_{i\neq j}^n I(Y_i < Y_j) g'\{(X_j - X_i)^{\mathrm{T}}\beta\}(X_j - X_i)/\{n(n-1)\}$ where $g'(\cdot)$ denotes the derivative of $g(\cdot)$. By the definition of the maximizer $\hat{\beta}_n^g$, it is not hard to check that $\hat{\beta}_n^g = \hat{\Sigma}^{-1} U_n^g(\hat{\beta}_n^g)/(U_n^{g\mathrm{T}}\hat{\beta}_n^g)$. Let $f_n^g(\beta) = \hat{\Sigma}^{-1} U_n^g(\beta)/\{\beta^{\mathrm{T}} U_n^g(\beta)\}$ for $\beta \in \mathcal{E}(\hat{\Sigma})$. Then $f_n^g(\cdot)$ is a continuous mapping from the compact set $\mathcal{E}(\hat{\Sigma})$ to itself.

Algorithm S1. Fixed-point iteration

Input data $\{(Y_i, X_i)\}_{i=1}^n$, compute $\hat{\Sigma}^{-1}$ and set $\kappa > 0$
Randomly set initial value $\beta^{(0)} \in \mathcal{E}(\hat{\Sigma})$
For $t \geq 0$, repeat
  $\beta^{(t+1)} \leftarrow f_n^g(\beta^{(t)})$
Until $\|\beta^{(t+1)} - \beta^{(t)}\|_2 \leqslant \kappa$ or $\|\beta^{(t+1)} - \beta^{(t-1)}\|_2 \leqslant \kappa$
$\hat{\beta}_n \leftarrow \arg\min_{\beta \in \{\beta^{(t-1)}, \beta^{(t)}, \beta^{(t+1)}\}} L_n^g(\beta)$
Output $\hat{\beta}_n$

Algorithm S1 is not a direct fixed-point iteration, as f_n^g is defined on the hyper ellipsoid and it is possibly an antipodal map, i.e, $f_n^g(\beta) = -\beta$ and $f_n^g(-\beta) = \beta$ for some $\beta \in \mathcal{E}(\hat{\Sigma})$. To circumvent the problem, the iteration will cease when $\|\beta^{(t+1)} - \beta^{(t)}\|_2 \leqslant \kappa$ or $\|\beta^{(t+1)} - \beta^{(t-1)}\|_2 \leqslant \kappa$, so as to avoid the case that there are potentially two alternating converging sequences. Algorithm S1 is relatively efficient compared with gradient decent methods for differentiable $g(\cdot)$, as no tuning parameter such as the learning rate or the batch size is involved. In our numerical studies, it takes around hundreds of iterations to converge. General convergence analysis for fixed-point iteration can be referred to Huang & Ma (2014) and chapter 10 of Burden et al. (2016). A sufficient condition for the convergence of the algorithm is the contraction mapping condition, i.e., $\|f_n^g(\beta) - f_n^g(\hat{\beta}_n^g)\|_2 \leq C\|\beta - \hat{\beta}_n^g\|_2$ holds for some $0 \leq C < 1$ over a neighborhood of $\hat{\beta}_n^g$, in which case any initial $\beta^{(0)}$ locating in that neighborhood would converge linearly to $\hat{\beta}_n^g$.

For non-differentiable $g(\cdot)$, we provide a simulated annealing algorithm to compute $\hat{\beta}_n^g$. Simulated annealing is an effective optimization method for solving unconstrained or bounded-constrained problem (Kirkpatrick et al., 1983). The detailed steps are given below:

Algorithm S2. Simulated Annealing solver

Input $\{(Y_i, X_i)\}_{i=1}^n$, compute $\hat{\Sigma}$, set $\kappa > 0$ and integer $K > 0$
Randomly set an initial value $\beta^{(0)} \in \mathcal{E}(\hat{\Sigma})$
For $t \geq 0$, repeat
  Generate random vector $e^{(t)} \sim N(0, d_t I_p)$, where d_t is the step size at the t-th iteration
  $\beta^{(t+1/2)} \leftarrow \{\beta^{(t)} + e^{(t)}\}/\|\beta^{(t)} + e^{(t)}\|_2$
  $\beta^{(t+1)} \leftarrow \beta^{(t+1/2)}$ if $L_n^g(\beta^{(t+1/2)}) > L_n^g(\beta^{(t)})$; otherwise, $\beta^{(t+1)} \leftarrow \beta^{(t)}$
Until $\|\beta^{(t+1)} - \beta^{(t-K)}\|_2 \leq \kappa$
Output $\hat{\beta}_n = \beta^{(t+1)}$

In the simulate annealing algorithm, the so-called "temperature" is always zero, which ensures that the objective function is strictly increasing. In our numerical studies, it takes around hundreds of iterations to converge with a satisfactory accuracy. The step size should satisfy $\sum_{t=0}^{\infty} d_t = \infty$ to ensure convergence. Comprehensive theoretical analysis can be found in-Granville et al. (1994). Different from Algorithms S1, there is no need to consider the sign of $\hat{\beta}_n$, and the calculation of the p-dimensional gradient is avoided. Nonetheless, finding the direction of decent by random trials as in Algorithm S2 would be less efficient than direct calculation of the gradient in high-dimensional case.

Lastly, a proximal (stochastic) gradient decent algorithm (Ferreira & Oliveira, 2002; Chen et al., 2020) is introduced to solve the penalized linearized MRC in Section 4 in the main context. When Σ is unknown, a consistent estimator $\hat{\Sigma}$ will be used to estimate Σ.

Algorithm S3. Proximal (stochastic) gradient decent

Input $\{(Y_i, X_i)\}_{i=1}^n$, λ, compute $\hat{\Sigma}$ and set $\kappa > 0$
Set an initial value $\beta^{(0)} \in \mathcal{E}(\hat{\Sigma})$
For $t \geq 0$, repeat
  Set the step size $\alpha_t > 0$
  $\theta^{(t+1/3)} \leftarrow \theta^{(t)} - \alpha_t \nabla_1 L_n\{\beta(\theta^{(t+1/3)}, \hat{\Sigma})\}$ and $\beta^{(t+1/3)} \leftarrow \beta(\theta^{(t+1/3)}, \hat{\Sigma})$
  $\beta^{(t+2/3)} \leftarrow \mathrm{sgn}(\beta^{(t+1/3)})[|\beta^{(t+1/3)}| - \alpha_t\lambda]_+$
  $\beta^{(t+1)} \leftarrow c\beta^{(t+2/3)}$ for some $c > 0$ such that $c^2(\beta^{(t+2/3)})^{\mathrm{T}}\hat{\Sigma}\beta^{(t+2/3)} = 1$
Until $\|\beta^{(t+1)} - \beta^{(t)}\|_2 \leqslant \kappa$
Output $\hat{\beta}_n^g = \beta^{(t+1)}$

In Algorithm S3, a good initial value of $\beta^{(0)}$ can be obtained easily by the LMRC estimation (without penalty), which greatly improves the efficiency of the algorithm. The term $L_n(\beta) = \sum_{i \neq j}^n I(Y_i < Y_j)(X_i - X_j)^{\mathrm{T}}\beta/\{n(n-1)\}$ is the empirical objective function without the penalty term and $\nabla_1 L_n$ denotes the (stochastic) gradient w.r.t θ. Here $\theta^{(t)}$ is updated by the step size α_t along its gradient, and $\beta^{(t)}(\theta^{(t)}, \hat{\Sigma})$ is updated accordingly based on the reparameterization. In Step 3, proximal operation ("soft-threshold" operation) is applied, where $\mathrm{sgn}(\cdot)$ returns the sign of each component. In Step 4, the updated parameter is re-scaled to satisfy the hyper ellipsoid constraint.

2. LEMMAS AND PROOFS OF THE MAIN THEOREMS

In this section, we provide some lemmas and detailed proofs of the theorems in the main context.

2.1. Lemmas

LEMMA S1. *(Hoeffding, 1992) For a U-statistic U_n with symmetric kernel h, let $\mu = E_F\{h(X_{i_1}, \ldots, X_{i_m})\}$. If $E_F|h| < \infty$, then $U_n \to \mu$ almost surely.*

LEMMA S2. *If a random vector X with mean μ and covariance matrix Σ satisfying $\beta_0^{\mathrm{T}}\Sigma\beta_0 \neq 0$. If X is of linearity of expectation in the direction of β_0, i.e., for any direction $b \in \mathbb{R}^p$,*

$$E[X^{\mathrm{T}}b \mid X^{\mathrm{T}}\beta_0] = c_b X^{\mathrm{T}}\beta_0 + a_b,$$

where $a_b, c_b \in \mathbb{R}$ *are some real constants which may depend on* b*, then, for any* $b \in \mathbb{R}^p$, $c_b = b^{\mathrm{T}}\Sigma\beta_0/\beta_0^{\mathrm{T}}\Sigma\beta_0$ *and* $a_b = b^{\mathrm{T}}\mu - c_b\beta_0^{\mathrm{T}}\mu = b^{\mathrm{T}}\mu - \beta_0^{\mathrm{T}}\mu b^{\mathrm{T}}\Sigma\beta_0/\beta_0^{\mathrm{T}}\Sigma\beta_0$.

Proof. Direct calculations give

$$\begin{aligned} b^{\mathrm{T}}\mu = E(b^{\mathrm{T}}X) =& E\{E(b^{\mathrm{T}}X \mid X^{\mathrm{T}}\beta_0)\} \\ =& E(c_b X^{\mathrm{T}}\beta_0 + a_b) \\ =& c_b\mu^{\mathrm{T}}\beta_0 + a_b, \end{aligned}$$

and

$$\begin{aligned} b^{\mathrm{T}}(\Sigma + \mu\mu^{\mathrm{T}})\beta_0 =& E(X^{\mathrm{T}}bX^{\mathrm{T}}\beta_0) \\ =& E\{X^{\mathrm{T}}\beta_0 E(X^{\mathrm{T}}b \mid X^{\mathrm{T}}\beta_0)\} \\ =& E\{X^{\mathrm{T}}\beta_0(c_b X^{\mathrm{T}}\beta_0 + a_b)\} \\ =& E(c_b\beta_0^{\mathrm{T}}XX^{\mathrm{T}}\beta_0 + a_b X^{\mathrm{T}}\beta_0) \\ =& c_b\beta_0^{\mathrm{T}}(\Sigma + \mu\mu^{\mathrm{T}})\beta_0 + a_b\beta_0^{\mathrm{T}}\mu. \end{aligned}$$

Combining these two equations, we have $a_b = b^{\mathrm{T}}\mu - c_b\beta_0^{\mathrm{T}}\mu = b^{\mathrm{T}}\mu - \beta_0^{\mathrm{T}}\mu b^{\mathrm{T}}\Sigma\beta_0/\beta_0^{\mathrm{T}}\Sigma\beta_0$ and $c_b = b^{\mathrm{T}}\Sigma\beta_0/\beta_0^{\mathrm{T}}\Sigma\beta_0$. $\square$

LEMMA S3. *Let* $W \in \mathbb{R}$ *be a random variable and let* $g(\cdot) : \mathbb{R} \to \mathbb{R}$ *be a non-constant increasing function defined on the support of* W*, then*

$$E\{g(W)W\} \geq E\{g(W)\}E(W).$$

Further, if $E\{g(W) - Eg(W)\}^2 > 0$*, i.e.,* $g(W)$ *has non-zero variance, we have*

$$E\{g(W)W\} > E\{g(W)\}E(W).$$

Proof. Considering

$$\begin{aligned} E\{g(W)W\} - E\{g(W)\}E(W) =& E[g(W)\{W - E(W)\}] \\ =& E[g(W)\{W - E(W)\}] - E[g\big(E(W)\big)\{W - E(W)\}] \\ =& E[\{g(W) - g\big(E(W)\big)\}\{W - E(W)\}], \end{aligned}$$

we only need to prove $E[\{g(W) - g(E(W))\}\{W - E(W)\}] \geq 0$. Note that $g(\cdot)$ is non-constant increasing on the support of W, then $W - E(W) \geq 0$ happens if and only if $g(W) - g(E(W)) \geq 0$ holds, which implies $\{g(W) - g\big(E(W)\big)\}\{W - E(W)\} \geq 0$. Thus, the inequality holds. Furthermore, if $g(W)$ has non-zero variance, then W also has non-zero variance. This implies that, there exist a subset $\mathcal{W}$ of the support of W with non-zero measure $\delta_0 > 0$ (i.e., $\mathrm{pr}(W \in \mathcal{W}) = \delta_0$) and constants $\delta_1, \delta_2 > 0$ such that $|W - E(W)| > \delta_1$ and $|g(W) - g(E(W))| > \delta_2$ on $\mathcal{W}$. Then $E\{g(W) - g(E(W))\}\{W - E(W)\} \geq \delta_0\delta_1\delta_2 > 0$. $\square$

2.2. *Proof of Theorem 1*

Define $L_n(\beta) = \sum_{i \neq j}^{n} I(Y_i < Y_j)(X_j - X_i)^{\mathrm{T}}\beta/\{n(n-1)\}$ and $L(\beta) = E\{L_n(\beta)\}$. Both of them are defined on the compact set $\mathcal{E}(\Sigma) = \{\beta \in \mathbb{R}^p : \beta^{\mathrm{T}}\Sigma\beta = 1\}$.

First, for any non-empty compact set $\mathcal{E}(\Sigma)$, the maximizer of $L_n(\cdot)$ defined in (4) in Section 2.1 in the main context always exists. The strong duality of the primal and dual problem by the Lagrange method has been shown in Section 2.2 in the main context. Then by the KKT conditions and the definition of the maximizer $\hat{\beta}_n^*$, we have

$$\hat{\beta}_n^* = \frac{\Sigma^{-1}U_n}{({U_n}^{\mathrm{T}}\Sigma^{-1}U_n)^{1/2}},$$

where $U_n = \nabla L_n(\beta) = \sum_{i \neq j}^{n} I(Y_i < Y_j)(X_j - X_i)/\{n(n-1)\}$ is irrelevant to β. Under Condition (C2), the first moment of X exists and U_n converges to $U = E\{I(Y_i < Y_j)(X_j - X_i)\} \in \mathbb{R}^p$ $(i \neq j)$almost surely by Lemma S1. By Slutsky's theorem, $\hat{\beta}_n^*$ converges to $\beta_\infty \equiv \Sigma^{-1}U/(U^{\mathrm{T}}\Sigma^{-1}U)^{1/2}$ in probability.

Next, we show that $\beta_\infty = \beta_0$. Note that the denominator of β_∞ is a normalizing scalar to make β_∞ satisfy the constraint $\beta_\infty^{\mathrm{T}}\Sigma\beta_\infty = 1$, which does not affect the direction of β_∞. In this regard, we concentrate our effort to show that the numerator $\Sigma^{-1}U$ has the same direction as β_0. To this end, we first show that any direction b perpendicular to β_0 is also perpendicular to $\Sigma^{-1}U$. For any b satisfying $b^{\mathrm{T}}\beta_0 = 0$ and for any $i \neq j$, the inner product of $\Sigma^{-1}U$ and b is

$$\begin{aligned}
(\Sigma^{-1}U)^{\mathrm{T}}b =& E\{I(Y_i < Y_j)(X_j - X_i)^{\mathrm{T}}\Sigma^{-1}b\} \\
=& E[E\{I(Y_i < Y_j)(X_j - X_i)^{\mathrm{T}}\Sigma^{-1}b \mid X_i^{\mathrm{T}}\beta_0, X_j^{\mathrm{T}}\beta_0, \epsilon_i, \epsilon_j\}] \\
=& E[I(Y_i < Y_j)E\{(X_j - X_i)^{\mathrm{T}}\Sigma^{-1}b \mid X_i^{\mathrm{T}}\beta_0, X_j^{\mathrm{T}}\beta_0\}] \\
=& E[I(Y_i < Y_j)\{E(X_j^{\mathrm{T}}\Sigma^{-1}b \mid X_j^{\mathrm{T}}\beta_0) - E(X_i^{\mathrm{T}}\Sigma^{-1}b \mid X_i^{\mathrm{T}}\beta_0)\}] \\
=& E\{I(Y_i < Y_j)(b^{\mathrm{T}}\Sigma^{-1}\mu - b^{\mathrm{T}}\Sigma^{-1}\mu)\} \\
=& 0,
\end{aligned}$$

where μ and Σ are the mean and covariance matrix of X. Under the linearity of expectation assumption, the second last equality holds by applying Lemma S2 on $E(X_j^{\mathrm{T}}\Sigma^{-1}b \mid X_j^{\mathrm{T}}\beta_0)$ and $E(X_i^{\mathrm{T}}\Sigma^{-1}b \mid X_i^{\mathrm{T}}\beta_0)$. Now, it remains to show that $\Sigma^{-1}U$ is a non-zero vector by verifying

$(\Sigma^{-1}U)^{\mathrm{T}}\beta_0 > 0$. To this end, we write

$$
\begin{aligned}
(\Sigma^{-1}U)^{\mathrm{T}}\beta_0 =& E\{I(Y_i < Y_j)(X_j - X_i)^{\mathrm{T}}\Sigma^{-1}\beta_0\} \\
=& E[E\{I(Y_i < Y_j)(X_j - X_i)^{\mathrm{T}}\Sigma^{-1}\beta_0 \mid X_i^{\mathrm{T}}\beta_0, X_j^{\mathrm{T}}\beta_0, \epsilon_i, \epsilon_j\}] \\
=& E[I(Y_i < Y_j)E\{(X_j - X_i)^{\mathrm{T}}\Sigma^{-1}\beta_0 \mid X_i^{\mathrm{T}}\beta_0, X_j^{\mathrm{T}}\beta_0\}] \\
=& E\{I(Y_i < Y_j)(\beta_0^{\mathrm{T}}\beta_0)(X_j^{\mathrm{T}}\beta_0 - X_i^{\mathrm{T}}\beta_0)\} \\
=& \beta_0^{\mathrm{T}}\beta_0 \times (E[E\{I(Y_i < Y_j \mid X_j^{\mathrm{T}}\beta_0)\}X_j^{\mathrm{T}}\beta_0] - E[E\{I(Y_i < Y_j \mid X_i^{\mathrm{T}}\beta_0)\}X_i^{\mathrm{T}}\beta_0]) \\
=& \beta_0^{\mathrm{T}}\beta_0 \times [E\{\mathrm{pr}(Y_i < Y_j \mid X_j^{\mathrm{T}}\beta_0)X_j^{\mathrm{T}}\beta_0\} - E\{\mathrm{pr}(Y_i < Y_j \mid X_i^{\mathrm{T}}\beta_0)X_i^{\mathrm{T}}\beta_0\}] \\
>& \beta_0^{\mathrm{T}}\beta_0 \times [E\{\mathrm{pr}(Y_i < Y_j \mid X_j^{\mathrm{T}}\beta_0)E(X_j^{\mathrm{T}}\beta_0)\} - E\{\mathrm{pr}(Y_i < Y_j \mid X_i^{\mathrm{T}}\beta_0)E(X_i^{\mathrm{T}}\beta_0)\}] \\
=& \beta_0^{\mathrm{T}}\beta_0 \times \{E(X_j^{\mathrm{T}}\beta_0)/2 - E(X_i^{\mathrm{T}}\beta_0)/2\} \\
=& 0.
\end{aligned}
$$

The fourth equality holds by Lemma S2. In the last third line, $\mathrm{pr}(Y_i < Y_j \mid X_j^{\mathrm{T}}\beta_0)$ is non-constant increasing in $X_j^{\mathrm{T}}\beta_0$ as Y_j is non-constant increasing in $X_j^{\mathrm{T}}\beta_0$, implying that $E\{\mathrm{pr}(Y_i < Y_j \mid X_j^{\mathrm{T}}\beta_0)X_j^{\mathrm{T}}\beta_0\} > E\{\mathrm{pr}(Y_i < Y_j \mid X_j^{\mathrm{T}}\beta_0)\}E(X_j^{\mathrm{T}}\beta_0) = E(X_j^{\mathrm{T}}\beta_0)/2$ by Assumption (M) and Lemma S3. Similar arguments can be applied to the other term and thus the inequality holds. As a result, $(\Sigma^{-1}U)^{\mathrm{T}}\beta_0 > 0$ and $\beta_\infty = \beta_0$. The proof of Theorem 1 is complete.

We wish to note that, without the monotonicity assumption on the first argument of $f(\cdot,\cdot)$, the closed form solution $\hat\beta_n$ can still be consistent for β_0 up to a sign as long as $(\Sigma^{-1}U)^{\mathrm{T}}\beta_0 \neq 0$. Actually, the condition $(\Sigma^{-1}U)^{\mathrm{T}}\beta_0 \neq 0$ ensures that $\Sigma^{-1}U$ is in the linear space spanned by β_0, since $(\Sigma^{-1}U)^{\mathrm{T}}b = 0$ still holds for any $b^{\mathrm{T}}\beta_0 = 0$ according to the above proofs.

2.3. *Proof of Theorem 2*

In view of the closed-form expression $\hat\beta_n^* = \Sigma^{-1}U_n/({U_n}^{\mathrm{T}}\Sigma^{-1}U_n)^{1/2}$, a standard Hoeffding's decomposition of U_n would be applied to obtain an asymptotic expression of $\hat\beta_n^*$, so as to prove the asymptotic normality.

Proof. Recall that $U_n = \sum_{i\neq j} I(Y_i < Y_j)(X_j - X_i)/\{n(n-1)\}$ is a U-statistic of order 2. By Hoeffding's decomposition,

$$U_n = U + \frac{1}{n}\sum_{i=1}^{n}\xi(Z_i) + \frac{1}{n(n-1)}\sum_{i\neq j}\phi(Z_i, Z_j),$$

where $U = EU_n$ and for each z, z_1, z_2 in S,

$$
\begin{aligned}
\xi(z) &= E\{I(y < Y)(X - x) + I(Y < y)(x - X) - 2U\}, \\
\phi(z_1, z_2) &= I(y_1 < y_2)(x_2 - x_1) - E\{I(y_1 < Y)(X - x_1)\} - E\{I(Y < y_2)(x_2 - X)\} + U.
\end{aligned}
$$

Since X has finite second moment, by the main corollary in section 6 in Sherman (1994), we have $\sum_{i\neq j}\phi(Z_i,Z_j)/\{n(n-1)\}=o_p(n^{-1/2})$, and

$$U_n = U + n^{-1/2}W_n + o_p(n^{-1/2}), \tag{S1}$$

where $W_n = n^{-1/2}\sum_{i=1}^n \xi(Z_i)$. By the central limit theorem, W_n converges in distribution to a normal random vector $N(0,\Delta)$ with $\Delta = E\{\xi(Z)\xi(Z)^{\mathrm{T}}\}$. By Theorem 1, $U = EU_n = c\Sigma\beta_0$ with $c = (U^{\mathrm{T}}\Sigma^{-1}U)^{1/2}$. Observe that

$$\begin{aligned}
\frac{1}{({U_n}^{\mathrm{T}}\Sigma^{-1}U_n)^{1/2}} &= \frac{1}{(U^{\mathrm{T}}\Sigma^{-1}U)^{1/2}} - \frac{1}{2(U^{\mathrm{T}}\Sigma^{-1}U)^{3/2}}({U_n}^{\mathrm{T}}\Sigma^{-1}U_n - U^{\mathrm{T}}\Sigma^{-1}U) + o_p(n^{-1/2})\\
&= \frac{1}{(c^2\beta_0^{\mathrm{T}}\Sigma\beta_0)^{1/2}} - \frac{1}{2(c^2\beta_0^{\mathrm{T}}\Sigma\beta_0)^{3/2}}({U_n}^{\mathrm{T}}\Sigma^{-1}U_n - c^2{\beta_0}^{\mathrm{T}}\Sigma\beta_0) + o_p(n^{-1/2})\\
&= \frac{1}{c} - \frac{1}{2c^3}({U_n}^{\mathrm{T}}\Sigma^{-1}U_n - c^2) + o_p(n^{-1/2})\\
&= \frac{1}{c} - \frac{1}{2c^3}\left(c^2 + \frac{2c\beta_0 W_n}{n^{1/2}} - c^2\right) + o_p(n^{-1/2})\\
&= \frac{1}{c} - \frac{1}{c^2}\frac{\beta_0 W_n}{n^{1/2}} + o_p(n^{-1/2}).
\end{aligned} \tag{S2}$$

Plugging (S2) into the closed form expression of $\hat\beta_n^*$, we have

$$\hat\beta_n^* = \beta_0 + \frac{n^{-1/2}}{(U^{\mathrm{T}}\Sigma^{-1}U)^{1/2}}(\Sigma^{-1} - \beta_0\beta_0^{\mathrm{T}})W_n + o_p(n^{-1/2}).$$

Let $V = (\Sigma^{-1} - \beta_0\beta_0^{\mathrm{T}})/(U^{\mathrm{T}}\Sigma^{-1}U)^{1/2}$ be a $p\times p$ matrix, $A = (0, I_{p-1})$ be a $(p-1)\times p$ matrix with its first column being zeros and I_{p-1} be an identity matrix of order $(p-1)$. Then, $\hat\theta_n^* = A\hat\beta_n^*$, $\theta_0 = A\beta_0$ and

$$n^{1/2}(\hat\theta_n^* - \theta_0) = \frac{1}{(U^{\mathrm{T}}\Sigma^{-1}U)^{1/2}}A(\Sigma^{-1} - \beta_0\beta_0^{\mathrm{T}})W_n + o_p(1).$$

Then, by the central limit theorem, $n^{1/2}(\hat\theta_n^* - \theta_0) \to N(0, AV\Delta V^{\mathrm{T}}A^{\mathrm{T}})$ in distribution as $n\to\infty$. We complete the proof of Theorem 2.

2.4. *Proof of Theorem 3*

In view of the closed form expression of $\hat\beta_n$ and the consistency of $\hat\Sigma$, Theorem 3 can be shown along similar lines of the proofs of Theorem 1. The details are omitted.

2.5. *Proof of Theorem 4*

The notations $c, U, U_n, \xi(\cdot), W_n, A, \Delta$ and V are defined in the proof of Theorem 2.

Proof of Theorem 4 part (i). Since $\hat\Sigma^{-1} = \Sigma^{-1} - \Sigma^{-1}(\hat\Sigma - \Sigma)\Sigma^{-1} + O(\|\hat\Sigma - \Sigma\|_2)$ almost surely, we have

$$\hat\Sigma^{-1} = \Sigma^{-1} + o_p(n^{-1/2}), \tag{S3}$$

under the assumption that $\|\hat{\Sigma} - \Sigma\|_2 = o_p(n^{-1/2})$. Plugging (S3) into the expression of $\hat{\beta}_n$, similar to the proof of Theorem 2, we obtain that $\hat{\beta}_n = \hat{\beta}_n^* + o_p(n^{-1/2})$. Hence, the conclusion of Theorem 4 part (i) holds.

Proof of Theorem 4 part (ii). When Σ is estimated by the sample covariance matrix $\hat{\Sigma}_S = \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^{\mathrm{T}}/(n-1)$, it is not hard to get that

$$\hat{\Sigma}_S = \Sigma + n^{-1/2}\Xi_n + O_p(\frac{1}{n}), \tag{S4}$$

where $\Xi_n = \sum_{i=1}^n \{(X_i - \mu)(X_i - \mu)^{\mathrm{T}} - \Sigma\}/n^{1/2}$. Then, plugging (S1) and (S4) into the closed-form expression of $\hat{\beta}_n$, some simple algebra yields that

$$\hat{\beta}_n = \beta_0 + \frac{n^{-1/2}}{(U^{\mathrm{T}}\Sigma^{-1}U)^{1/2}}(\Sigma^{-1} - \beta_0\beta_0^{\mathrm{T}})W_n + n^{-1/2}(\frac{\beta_0\beta_0^{\mathrm{T}}}{2} - \Sigma^{-1})\Xi_n\beta_0 + o_p(n^{-1/2}).$$

Then, since $\hat{\theta}_n = A\hat{\beta}_n$ and $\theta_0 = A\beta_0$ with $A = (0, I_{p-1})$ being a $(p-1)\times p$ matrix, we have the following asymptotic expression

$$n^{1/2}(\hat{\theta}_n - \theta_0) = \frac{1}{(U^{\mathrm{T}}\Sigma^{-1}U)^{1/2}}A(\Sigma^{-1} - \beta_0\beta_0^{\mathrm{T}})W_n + A(\beta_0\beta_0^{\mathrm{T}}/2 - \Sigma^{-1})\Xi_n\beta_0 + o_p(1).$$

Since W_n and $\Xi_n\beta_0$ are both sum of independent and identically distributed random vectors, under the moment condition of X and by the central limit theorem, $n^{1/2}(\hat{\theta}_n - \theta_0) \to N(0, ABA^{\mathrm{T}})$ in distribution, where $B = E\{V\xi(Z) + H\psi(Z)\}\{V\xi(Z) + H\psi(Z)\}^{\mathrm{T}}$, $\psi(Z) = \{(X-\mu)(X-\mu)^{\mathrm{T}} - \Sigma\}\beta_0$ and $H = \beta_0\beta_0^{\mathrm{T}}/2 - \Sigma^{-1}$. The proof of Theorem 4 is complete. □

2.6. *Proof of Theorem 5*

Define $L_n^g(\beta) = \sum_{i\neq j}^n I(Y_i < Y_j)g\{(X_j - X_i)^{\mathrm{T}}\beta\}/\{n(n-1)\}$ and $L^g(\beta) = E\{L_n^g(\beta)\}$. For reader's convenience, we first give the definition of elliptical distribution below (Theorem 1, Cambanis et al. (1981)).

DEFINITION S1. *(Elliptical Distributions) A p-dimensional random variable X is said to be elliptical distributed if and only if there exist a vector $\mu \in \mathbb{R}^p$ and a positive semidefinite matrix $\Sigma \in \mathbb{R}^{p\times p}$ with rank k, such that $X = \mu + \mathcal{R}\Lambda\mathcal{U}^{(k)}$, where $\mathcal{U}^{(k)}$ is a k-dimensional random vector uniformly distributed on a unit $(k-1)$-sphere $\mathcal{S}^{k-1}$, $\mathcal{R}$ is a non-negative random variable stochastically independent of $\mathcal{U}^{(k)}$ and $\Lambda\Lambda^{\mathrm{T}} = \Sigma$.*

Proof of Theorem 5. We intend to prove the consistency in 3 steps.

Step 1. To prove the maximizer of $L_n^g(\beta)$, $\hat{\beta}_n^g$ converges to the maximizer of $L^g(\beta)$ in probability. By the properties of elliptical distributions as shown in chapter 1 of 2004 University of Cologne Faculty of Management PhD thesis by Frahm. G, there are two facts: first, under elliptical distribution assumption, the difference of each pair of observations, i.e, $X_j - X_i$, $i \neq j$, is also elliptical distributed; second, for any $\beta \in \mathcal{E}(\Sigma)$ and any $i \neq j$, $g\{(X_j - X_i)^{\mathrm{T}}\beta\}$ follows the same distribution, as $(X_j - X_i)^{\mathrm{T}}\beta$ have the same distribution since

$$X_j - X_i \overset{d}{=} \mathcal{R}\Lambda U, \tag{S5}$$

where U is a p-dimensional random vector uniformly distributed on a unit $(p-1)$-sphere $\mathcal{S}^{p-1}$, $\mathcal{R}$ is a non-negative random variable stochastically independent of U and $\Lambda\Lambda^{\mathrm{T}} = \Sigma$. Then $(X_j - X_i)^{\mathrm{T}}\beta = \mathcal{R}U^{\mathrm{T}}\Lambda^{\mathrm{T}}\beta = \mathcal{R}U^{\mathrm{T}}\alpha$, where $\alpha = \Lambda^{\mathrm{T}}\beta = \Sigma^{1/2}\beta \in \mathcal{S}^{p-1} = \{\alpha \in \mathbb{R}^p : \alpha^{\mathrm{T}}\alpha = 1\}$. As a result, $(X_j - X_i)^{\mathrm{T}}\beta$ has the same distribution for any $\beta \in \mathcal{E}(\Sigma)$ since $U^{\mathrm{T}}\alpha$ has the same distribution for any $\alpha \in \mathcal{S}^{p-1}$. Therefore, by Condition (G1) part (i), $E\|g\{(X_1 - X_2)^{\mathrm{T}}\beta\}\|_2 < \infty$ for all $\beta \in \mathcal{E}(\Sigma)$ and L^g is well-defined on $\mathcal{E}(\Sigma)$. In addition, since $|I(Y_i < Y_j)| \leq 1$, $E\|L_n^g\|_\infty := E\sup_{\beta\in\mathcal{E}(\Sigma)} |L_n^g(\beta)| < \infty$. Next we show that $\|L_n^g - L^g\|_\infty \to 0$ in probability uniformly on $\mathcal{E}(\Sigma)$ as $n \to \infty$, i.e.,

$$\sup_{\beta\in\mathcal{E}(\Sigma)} |L_n^g(\beta) - L^g(\beta)| \to 0$$

in probability. Firstly, since $E\|L_n^g\|_\infty < \infty$, we have

$$\sup_{\beta\in\mathcal{E}(\Sigma)} E \sup_{\alpha:\|\alpha-\beta\|_2<\epsilon} |L_n^g(\alpha) - L_n^g(\beta)| \to 0 \qquad \text{as } \epsilon \downarrow 0, \tag{S6}$$

and $L^g(\beta) = E\{L_n^g(\beta)\}$ is continuous on β (Lemma 9.1, Keener, 2010). By Lemma S1, for any $\beta \in \mathcal{E}(\Sigma)$, we have $L_n^g(\beta) \to L^g(\beta)$ in probability. For $\delta > 0$, let

$$M_{\delta,ij}(\beta) = \sup_{\alpha:\|\alpha-\beta\|_2<\delta} |I(Y_i < Y_j)g\{(X_j - X_i)^{\mathrm{T}}\alpha\} - I(Y_i < Y_j)g\{(X_j - X_i)^{\mathrm{T}}\beta\}|$$

and $L_\delta^g(\beta) = EM_{\delta,ij}(\beta)$. Given any $\epsilon > 0$, by (S6), we can choose δ such that

$$E \sup_{\alpha:\|\alpha-\beta\|_2<\epsilon} |L_n^g(\alpha) - L_n^g(\beta)| \leq L_\delta^g(\beta) < \epsilon \qquad \forall \beta \in \mathcal{E}(\Sigma),$$

and with such choice of δ, if $\|\alpha - \beta\|_2 < \delta$, then

$$|L^g(\alpha) - L^g(\beta)| = |E\{L_n^g(\alpha) - L_n^g(\beta)\}| \leq E|L_n^g(\alpha) - L_n^g(\beta)| \leq \epsilon.$$

Let $B_\delta(\beta) = \{\alpha : \|\alpha - \beta\|_2 < \delta\}$ be the open ball with radius δ and center β. Since $\mathcal{E}(\Sigma)$ is compact, the open sets $\{B_\delta(\beta) : \beta \in \mathcal{E}(\Sigma)\}$ covering $\mathcal{E}(\Sigma)$ have a finite subcover $\{O_t = B_\delta(\alpha_t) : t = 1, \ldots, m\}$. Then,

$$\begin{aligned}
\|L_n^g - L^g\|_\infty &= \max_{t=1,\ldots,m} \sup_{\alpha\in O_t} |L_n^g(\alpha) - L^g(\alpha)| \\
&\leq \max_{t=1,\ldots,m} \sup_{\alpha\in O_t} \left\{|L_n^g(\alpha) - L_n^g(\alpha_t)| + |L_n^g(\alpha_t) - L^g(\alpha_t)| + |L^g(\alpha_t) - L^g(\alpha)|\right\} \\
&\leq \max_{t=1,\ldots,m} \sup_{\alpha\in O_t} |L_n^g(\alpha) - L_n^g(\alpha_t)| + \max_{t=1,\ldots,m} |L_n^g(\alpha_t) - L^g(\alpha_t)| + \epsilon.
\end{aligned}$$

Now,

$$\begin{aligned}\sup_{\alpha\in O_t}|L_n^g(\alpha)-L_n^g(\alpha_t)| =&\frac{1}{n(n-1)}\sup_{\alpha\in O_t}|\sum_{i\neq j}I(Y_i<Y_j)g\{(X_j-X_i)^{\mathrm{T}}\alpha\}\\ &\qquad\qquad\qquad -I(Y_i<Y_j)g\{(X_j-X_i)^{\mathrm{T}}\alpha_t\}|\\ \leq&\frac{1}{n(n-1)}\sum_{i\neq j}M_{\delta,ij}(\alpha_t){:=}\bar{M}_{\delta,n}(\alpha_t).\end{aligned}$$

By the law of large numbers,

$$\bar{M}_{\delta,n}(\alpha_t)\to L_\delta^g(\alpha_t)<\epsilon$$

in probability. Thus, we have

$$\|L_n^g-L^g\|_\infty<2\epsilon+\max_{t=1,\ldots,m}\left\{\bar{M}_{\delta,n}(\alpha_t)-L_\delta^g(\alpha_t)\right\}+\max_{t=1,\ldots,m}|L_n^g(\alpha_t)-L^g(\alpha_t)|.$$

The two maximums on the right hand side of the above inequality both tend to zero in probability, with which it is easy to show that $\mathrm{pr}(\|L_n^g-L^g\|_\infty>3\epsilon)\to 0$ as $n\to\infty$. This proves that $\|L_n^g-L^g\|_\infty\to 0$ in probability uniformly on $\mathcal{E}(\Sigma)$. With the same lines of proof coupled with the strong law of large numbers (Lemma S1), we can prove that $\|L_n^g-L^g\|_\infty\to 0$ almost surely.

Step 2. To show that for any $\beta\in\mathcal{E}(\Sigma)$, we can write $L^g(\beta)=\int F(x,\beta)G(x,\beta)dx$, where F and G are integrable functions. Recall that X_j-X_i follows the same symmetric elliptical distribution for any $i\neq j$ by (S5). Without loss of generality, we assume $\mathrm{cov}(X)=I_p$ and $\beta\in\mathcal{E}(I_p)=\mathcal{S}^{p-1}=\{\beta\in\mathbb{R}^p:\beta^{\mathrm{T}}\beta=1\}$, and decompose X_j-X_i into two independent random variables $\mathcal{R}$ and $\mathcal{U}$, where $\mathcal{R}\equiv\|X_j-X_i\|_2$ is a nonnegative random variable and $\mathcal{U}\equiv(X_j-X_i)/\|X_j-X_i\|_2$ is the direction of X_j-X_i uniformly distributed on a unit $(p-1)$-sphere $\mathcal{S}^{p-1}$. Then,

$$\begin{aligned}L^g(\beta)=&E[I(Y_i<Y_j)g\{(X_j-X_i)^{\mathrm{T}}\beta\}]\\ =&E(E[I(Y_i<Y_j)g\{(X_j-X_i)^{\mathrm{T}}\beta\}\mid X_j-X_i])\\ =&E[g\{(X_j-X_i)^{\mathrm{T}}\beta\}E\{I(Y_i<Y_j)\mid X_j-X_i\}]\\ =&E[g(\mathcal{R}\mathcal{U}^{\mathrm{T}}\beta)E\{I(Y_i<Y_j)\mid\mathcal{R},\mathcal{U}\}].\end{aligned}$$

Define $F(\mathcal{R},\mathcal{U};\beta)=g(\mathcal{R}\mathcal{U}^{\mathrm{T}}\beta)$ and $G(\mathcal{R},\mathcal{U})=E\{I(Y_i<Y_j)\mid\mathcal{R},\mathcal{U}\}$. Let $\sigma(\mathcal{S}^{p-1})$ denote the area of unit sphere and let $f_{\mathcal{R}}(\cdot)$ denote the density function of $\mathcal{R}$. Then, it follows from the independence of $\mathcal{R}$ and $\mathcal{U}$ that

$$L^g(\beta)=\int_0^\infty\int_{\mathcal{S}^{p-1}}F(\mathcal{R},\mathcal{U};\beta)G(\mathcal{R},\mathcal{U})f_{\mathcal{R}}(\mathcal{R})/\sigma(\mathcal{S}^{p-1})d\mathcal{U}d\mathcal{R}. \tag{S7}$$

Step 3. To apply Hardy-Littlewood inequality (Burchard, 2009) on (S7), so as to prove β_0 is the unique maximizer of $L^g(\beta)$. For each $\mathcal{R}\in[0,+\infty)$, once $\int_{\mathbb{S}^{p-1}}F(\mathcal{R},\mathcal{U};\beta)G(\mathcal{R},\mathcal{U})/\sigma(\mathbb{S}^{p-1})d\mathcal{U}$ is maximized over β, then L^g is maximized. Next, we focus on $G(\mathcal{R},\mathcal{U})$ and $F(\mathcal{R},\mathcal{U};\beta)$. By definition, $|G(\mathcal{R},\mathcal{U})| \le 1$, and all its moments exist. When $\mathcal{R}=0$, it is easy to see that $G(0,\mathcal{U}) \equiv 1/2$. For each $\mathcal{R}>0$, $G(\mathcal{R},\mathcal{U})$ is symmetric about β_0 on the unit sphere and increasing in $\mathcal{U}^{\mathrm{T}}\beta_0$ by condition (G2). To be exact, $G(\mathcal{R},\mathcal{U}_1)=G(\mathcal{R},\mathcal{U}_2)$ if $\mathcal{U}_1^{\mathrm{T}}\beta_0=\mathcal{U}_2^{\mathrm{T}}\beta_0$ and $G(\mathcal{R},\mathcal{U}_1)\ge G(\mathcal{R},\mathcal{U}_2)$ if $\mathcal{U}_1^{\mathrm{T}}\beta_0>\mathcal{U}_2^{\mathrm{T}}\beta_0$. Meanwhile, for fixed $\mathcal{R}>0$, $1-G(\mathcal{R},\mathcal{U})$ is symmetric about β_0 and decreasing in $\mathcal{U}^{\mathrm{T}}\beta_0$. By condition (G1) part (i) and definition of $F(\mathcal{R},\mathcal{U};\beta)$, its first moment exists. For $\mathcal{R}=0$, $F(0,\mathcal{U};\beta)\equiv 0$. For each fixed $\mathcal{R}>0$, $F(\mathcal{R},\mathcal{U};\beta)$ has the same distribution for all $\beta\in\mathcal{S}^{p-1}$, and β is actually a parameter rotating the function graph of $g(\mathcal{R}\mathcal{U}^{\mathrm{T}}\beta)$ over the support of $\mathcal{U}$. When $\beta=\beta_0$, $F(\mathcal{R},\mathcal{U};\beta_0)=g(\mathcal{R}\mathcal{U}^{\mathrm{T}}\beta_0)$ is symmetric about β_0 on the unit sphere (the support of $\mathcal{U}$), and it is non-constant increasing in $\mathcal{U}^{\mathrm{T}}\beta_0$.

Hence, for each $\mathcal{R}>0$, nonnegative measurable functions $F(\mathcal{R},\mathcal{U};\beta_0)=g(\mathcal{R}\mathcal{U}^{\mathrm{T}}\beta_0)$ and $G(\mathcal{R},\mathcal{U})$ are concordant with each other, i.e. they have the same monotonicity over the support of $\mathcal{U}$. Applying Hardy-Littlewood inequality (Burchard, 2009), we have

$$\begin{aligned}&\int_{\mathcal{S}^{p-1}} -F(\mathcal{R},\mathcal{U};\beta)(1-G(\mathcal{R},\mathcal{U}))/\sigma(\mathcal{S}^{p-1})d\mathcal{U}\\ &\qquad\qquad \le \int_{\mathcal{S}^{p-1}} -F(\mathcal{R},\mathcal{U};\beta_0)(1-G(\mathcal{R},\mathcal{U}))/\sigma(\mathcal{S}^{p-1})d\mathcal{U}\end{aligned}$$

for any $\beta\in\mathcal{S}^{p-1}$. Furthermore,

$$\begin{aligned}L^g(\beta_0)=&\int_0^\infty\int_{\mathcal{S}^{p-1}} F(\mathcal{R},\mathcal{U};\beta_0)G(\mathcal{R},\mathcal{U})/\sigma(\mathcal{S}^{p-1})f_{\mathcal{R}}(\mathcal{R})d\mathcal{U}d\mathcal{R}\\ \ge&\int_0^\infty\int_{\mathcal{S}^{p-1}} F(\mathcal{R},\mathcal{U};\beta)G(\mathcal{R},\mathcal{U})/\sigma(\mathcal{S}^{p-1})f_{\mathcal{R}}(\mathcal{R})d\mathcal{U}d\mathcal{R}\\ =&L^g(\beta).\end{aligned}$$

By condition (G2), both $F(\mathcal{R},\mathcal{U};\beta)$ and $G(\mathcal{R},\mathcal{U})$ are non-constant increasing in their arguments, thus $L^g(\beta_0)>L^g(\beta)$ for any $\beta\ne\beta_0$. This completes the proof of Theorem 5.

2.7. *Proof of Theorem 6*

Proof. With the parameterization $\beta=\beta(\theta,\Sigma)$ in Section 2.4 in the main context, define $\Gamma_n(\theta,\Sigma)\equiv L_n^g(\beta(\theta,\Sigma))-L_n^g(\beta_0(\theta_0,\Sigma))$, and $\Gamma(\theta,\Sigma)\equiv E\Gamma_n(\theta,\Sigma)$ for each $\theta\in\Theta$. Note that $\Gamma_n(\theta_0,\cdot)=0$ and $\Gamma(\theta_0,\cdot)=0$. Firstly, under the assumption that $\|\mathrm{Diff}(\theta,\hat{\Sigma})-\mathrm{Diff}(\theta,\Sigma)\|_2 = o_p(n^{-1/2}\|\theta-\theta_0\|_2)$ uniformly over $o_p(1)$ neighborhoods of θ_0, it is not hard to obtain that

$$\Gamma_n(\theta,\hat{\Sigma})=\Gamma_n(\theta,\Sigma)+o_p(n^{-1/2}\|\theta-\theta_0\|_2)$$

uniformly over $o_p(1)$ neighborhoods of θ_0. Thereafter, we focus on handling $\Gamma_n(\theta,\Sigma)$ and write it as $\Gamma_n(\theta)$ for simplicity. It follows from the standard Hoeffding's decomposition of U-process that

$$\Gamma_n(\theta) = \Gamma(\theta) + \frac{1}{n}\sum_{i=1}^{n}\eta(Z_i,\theta) + \frac{1}{n(n-1)}\sum_{i\neq j}\omega(Z_i,Z_j,\theta),$$

where for each z in S and each $\theta \in \Theta$,

$$\begin{aligned}
&\eta(z,\theta) = \tau(z,\theta) - \tau(z,\theta_0) - 2\Gamma(\theta),\\
&\tau(z,\theta) = E[I(y<Y)g\{(X-x)^{\mathrm{T}}\beta(\theta,\Sigma)\} + I(Y<y)g\{(x-X)^{\mathrm{T}}\beta(\theta,\Sigma)\}],
\end{aligned}$$

and

$$\begin{aligned}
&\omega(z_i,z_j,\theta) = \phi_g(z_1,z_2,\theta) - \phi_g(z_1,z_2,\theta_0),\\
&\phi_g(z_1,z_2,\theta) = I(y_1<y_2)g\{(x_2-x_1)^{\mathrm{T}}\beta(\theta,\Sigma)\} + \Gamma(\theta)\\
&\qquad - E[I(y_1<Y)g\{(Y-x_1)^{\mathrm{T}}\beta(\theta,\Sigma)\} + I(Y<y_2)g\{(x_2-Y)^{\mathrm{T}}\beta(\theta,\Sigma)\}].
\end{aligned}$$

By referring to the main theorems in Sherman (1993), we shall first prove the following three statements, which are key steps to establish the $n^{1/2}$-consistency and asymptotic distribution of $\hat\theta_n^g$:

(i) There exist a neighborhood $\mathcal{N} \subset \Theta$ of θ_0 and a constant $\kappa > 0$ such that, for all θ in $\mathcal{N}$,

$$\Gamma(\theta) = \frac{1}{2}(\theta-\theta_0)^{\mathrm{T}}V^g(\theta-\theta_0) + o(\|\theta-\theta_0\|_2^2) \leqslant -\kappa\|\theta-\theta_0\|_2^2,$$

where $V = E\{\nabla_2\tau_g(Z,\theta_0)\}/2$.

(ii) Uniformly over $o_p(1)$ neighborhoods of $\theta_0 \in \Theta$,

$$\frac{1}{n}\sum_{i=1}^{n}\eta(Z_i,\theta) = n^{-1/2}(\theta-\theta_0)^{\mathrm{T}}W_n^g + o(\|\theta-\theta_0\|_2),$$

where W_n^g is a random vector converging to $N(0,\Delta^g)$ in distribution with $\Delta^g = E[\nabla_1\tau_g(Z,\theta_0)\{\nabla_1\tau_g(Z,\theta_0)\}^{\mathrm{T}}]$.

(iii) Uniformly over $o_p(1)$ neighborhoods of θ_0,

$$\frac{1}{n(n-1)}\sum_{i\neq j}\omega(Z_i,Z_j,\theta) = o_p(\frac{1}{n}).$$

To prove (i), we fix $z \in S$ and $\theta \in \mathcal{N}$. By condition (G3) and Taylor expansion of $\tau_g(z,\theta)$ around θ_0,

$$\tau_g(z,\theta) - \tau_g(z,\theta_0) = (\theta-\theta_0)^{\mathrm{T}}\nabla_1\tau_g(z,\theta_0) + \frac{1}{2}(\theta-\theta_0)^{\mathrm{T}}\nabla_2\tau_g(z,\theta^\star)(\theta-\theta_0), \qquad \text{(S8)}$$

where $\theta^\star$ is between θ_0 and θ. Besides, under condition (G3), for each $z \in S$ and each $\theta \in \mathcal{N}$,

$$\|(\theta-\theta_0)^{\mathrm{T}}\{\nabla_2\tau_g(z,\theta) - \nabla_2\tau_g(z,\theta_0)\}(\theta-\theta_0)\| \leqslant M_g(z)\|\theta-\theta_0\|_2^3$$

with integrable M_g. Notice that $E\{\tau_g(Z,\theta)-\tau_g(Z,\theta_0)\}=2\Gamma(\theta)$. Then,

$$2\Gamma(\theta)=(\theta-\theta_0)^{\mathrm{T}}E\nabla_1\tau_g(Z,\theta_0)+(\theta-\theta_0)^{\mathrm{T}}V^g(\theta-\theta_0)+o(\|\theta-\theta_0\|_2^2). \quad \text{(S9)}$$

As shown in the proofs of Theorem 5, β_0 is the global (local) maximizer of L^g on $\mathcal{E}(\Sigma)$; thus $E\{\nabla_1\tau(Z,\theta_0)\}=0$ and V^g is negative definite. Hence,

$$\Gamma(\theta)=\frac{1}{2}(\theta-\theta_0)^{\mathrm{T}}V^g(\theta-\theta_0)+o(\|\theta-\theta_0\|_2^2)\leqslant-\kappa\|\theta-\theta_0\|_2^2.$$

To show (ii), in view of (S8) and (S9), it follows from the definition of $\eta(\cdot,\theta)=\tau(z,\theta)-\tau(z,\theta_0)-2\Gamma(\theta)$ that

$$\frac{1}{n}\sum_{i=1}^{n}\eta(\cdot,\theta)=n^{-1/2}(\theta-\theta_0)^{\mathrm{T}}W_n^g+\frac{1}{2}(\theta-\theta_0)^{\mathrm{T}}D_n^g(\theta-\theta_0)+o(\|\theta-\theta_0\|_2^2)+R_n^g(\theta),$$

where $W_n^g=n^{-1/2}\sum_{i=1}^n\nabla_1\tau_g(Z_i,\theta_0)$, $D_n^g=\sum_{i=1}^n\nabla_2\tau_g(Z_i,\theta_0)/n-2V^g$ and $\|R_n^g(\theta)\|_2\leqslant\|\theta-\theta_0\|_2^3\sum_{i=1}^n M_g(Z_i)/n$. By the central limit theorem, $W_n^g\to N(0,\Delta)$ in distribution. And according to the weak law of large numbers, $D_n^g\to0$ in probability as $n\to\infty$. Next, by the integrability of M_g and the weak law of large numbers, it can be shown that $R_n^g(\theta)=o_p(\|\theta-\theta_0\|_2^2)$ uniformly over $o_p(1)$ neighborhoods of θ_0.

To prove (iii), by Corollary 17, Corollary 21 in Nolan & Pollard (1987) and Theorem 3 in Sherman (1993), it suffices to prove that $\mathcal{H}=\{h_g(\cdot,\cdot,\beta(\theta,\Sigma)):\theta\in\Theta\}$ is Euclidean with a constant envelope, where $h_g(z_1,z_2;\beta(\theta))=I(y_1<y_2)g\{(x_2-x_1)^{\mathrm{T}}\beta(\theta,\Sigma)\}$ for each $(z_1,z_2)\in S\otimes S$ and each $\theta\in\ \Theta$. Then, according to Lemma 2.12 in Pakes & Pollard (1989), if {subgraph(h_g):$h\in\mathcal{H}$} is a VC class of sets, then $\mathcal{H}$ is Euclidean for every envelope. Next, we intend to show that {subgraph(h):$h\in\mathcal{H}$} is a VC class of sets. For each $\theta\in\Theta$,

$$\begin{aligned}\text{subgraph}(h_g(\cdot,\cdot,\beta(\theta)))&=\{(z_1,z_2,t)\in\mathcal{X}\otimes\mathbb{R}:0<t<h_g(z_1,z_2,\beta(\theta,\Sigma))\}\\&=\{t>0\}\{y_2-y_1>0\}\{g\{(x_2-x_1)^{\mathrm{T}}\beta(\theta)\}-t>0\}\\&=\{s_1>0\}\{s_2>0\}\{s_3>0\}\end{aligned}$$

For any $(z_1,z_2,t)\in\mathcal{X}\otimes\mathbb{R}$, the class of sets $\{s_1>0\}$ and $\{s_2>0\}$ are both VC class according to Lemma 2.4 in Pakes & Pollard (1989). And, by condition (G1) part (ii), $\{s_3>0\}$ also belongs to VC class. Since the intersection of sets in VC classes are still a VC class, as a result, {subgraph(h):$h\in\mathcal{H}$} is a VC class of sets.

Combining statements (i)-(iii), according to Theorem 1 of Sherman (1993), we have shown the $n^{1/2}$-consistency and asymptotic normality of $\hat\theta_n^g$, that is, $\|\hat\theta_n^g-\theta_0\|_2=O_p(n^{-1/2})$ and $n^{1/2}(\hat\theta_n^g-\theta_0)\to N(0,(V^g)^{-1}\Delta^g(V^g)^{-1})$ in distribution. The proof of Theorem 6 is complete.□

2.8. *Proof of Theorem 7*

Proof. In the presence of censoring, recall that $L_n^c(\beta) = \sum_{i\neq j} d_i I(v_i < v_j)(X_j - X_i)^{\mathrm{T}}\beta/\{n(n-1)\}$ and $U_n^c = \sum_{i\neq j} d_i I(v_i < v_j)(X_j - X_i)/\{n(n-1)\}$. Define $L^c(\beta) = EL_n^c(\beta)$ and $U^c = E(U_n^c)$. Invoke the closed-form expression $\hat{\beta}_n^c = \hat{\Sigma}^{-1}U_n^c/(U_n^{c\mathrm{T}}\hat{\Sigma}^{-1}U_n^c)^{1/2}$.

With consistent estimate $\hat{\Sigma}^{-1}$, to establish the consistency of $\hat{\beta}_n^c$, it suffices to show that $\Sigma^{-1}U^c$ lies in the linear space of β_0. For any b satisfying $b^{\mathrm{T}}\beta_0 = 0$, the inner product of $\Sigma^{-1}U^c$ and b is

$$
\begin{aligned}
(\Sigma^{-1}U^c)^{\mathrm{T}}b =& E\{d_i I(v_i < v_j)(X_j - X_i)^{\mathrm{T}}\Sigma^{-1}b\}\\
=& E[E\{d_i I(v_i < v_j)(X_j - X_i)^{\mathrm{T}}\Sigma^{-1}b \mid X_i^{\mathrm{T}}\beta_0, X_j^{\mathrm{T}}\beta_0, \epsilon_i, \epsilon_j, C_i, C_j\}]\\
=& E[d_i I(v_i < v_j)E\{(X_j - X_i)^{\mathrm{T}}\Sigma^{-1}b \mid X_i^{\mathrm{T}}\beta_0, X_j^{\mathrm{T}}\beta_0, \epsilon_i, \epsilon_j, C_i, C_j\}]\\
=& E[d_i I(v_i < v_j)E\{(X_j - X_i)^{\mathrm{T}}\Sigma^{-1}b \mid X_i^{\mathrm{T}}\beta_0, X_j^{\mathrm{T}}\beta_0\}]\\
=& E[d_i I(v_i < v_j)\{E(X_j^{\mathrm{T}}\Sigma^{-1}b \mid X_j^{\mathrm{T}}\beta_0) - E(X_i^{\mathrm{T}}\Sigma^{-1}b \mid X_i^{\mathrm{T}}\beta_0)\}]\\
=& E\{d_i I(v_i < v_j)(b^{\mathrm{T}}\Sigma^{-1}\mu - b^{\mathrm{T}}\Sigma^{-1}\mu)\}\\
=& 0,
\end{aligned}
$$

where the second last equation holds by Lemma S2 and the fourth equality is due to the independence assumption in condition (A1). Then, the inner product of $\Sigma^{-1}U^c$ and β_0 is

$$
\begin{aligned}
(\Sigma^{-1}U^c)^{\mathrm{T}}\beta_0 =& E\{d_i I(v_i < v_j)(X_j - X_i)^{\mathrm{T}}\Sigma^{-1}\beta_0\}\\
=& E[E\{d_i I(v_i < v_j)(X_j - X_i)^{\mathrm{T}}\Sigma^{-1}\beta_0 \mid X_i^{\mathrm{T}}\beta_0, X_j^{\mathrm{T}}\beta_0, \epsilon_i, \epsilon_j, C_i, C_j\}]\\
=& E[d_i I(v_i < v_j)E\{(X_j - X_i)^{\mathrm{T}}\Sigma^{-1}\beta_0 \mid X_i^{\mathrm{T}}\beta_0, X_j^{\mathrm{T}}\beta_0, \epsilon_i, \epsilon_j, C_i, C_j\}]\\
=& E[d_i I(v_i < v_j)E\{(X_j - X_i)^{\mathrm{T}}\Sigma^{-1}\beta_0 \mid X_i^{\mathrm{T}}\beta_0, X_j^{\mathrm{T}}\beta_0\}]\\
=& E[d_i I(v_i < v_j)\{E(X_j^{\mathrm{T}}\Sigma^{-1}\beta_0 \mid X_j^{\mathrm{T}}\beta_0) - E(X_i^{\mathrm{T}}\Sigma^{-1}\beta_0 \mid X_i^{\mathrm{T}}\beta_0)\}]\\
=& E[d_i I(v_i < v_j)\{\beta_0^{\mathrm{T}}\beta_0(X_j^{\mathrm{T}}\beta_0) - \beta_0^{\mathrm{T}}\beta_0(X_i^{\mathrm{T}}\beta_0)\}]\\
=& \beta_0^{\mathrm{T}}\beta_0 E\{d_i I(v_i < v_j)(X_j^{\mathrm{T}}\beta_0 - X_i^{\mathrm{T}}\beta_0)\}\\
>& \beta_0^{\mathrm{T}}\beta_0 E\{d_i I(v_i < v_j)\}E(X_j^{\mathrm{T}}\beta_0 - X_i^{\mathrm{T}}\beta_0)\\
=& 0,
\end{aligned}
$$

where the sixth equation follows from Lemma S2. And the last second inequality follows from Lemma (S3), Assumption (M), and the independence assumption in condition (A1), which implies that $E\{\mathrm{pr}(Y_i < C_i, v_i < v_j \mid X_j^{\mathrm{T}}\beta_0)\}$ is non-constant increasing in $X_j^{\mathrm{T}}\beta_0$ while non-constant decreasing in $X_i^{\mathrm{T}}\beta_0$. This completes the proof of consistency. □

2.9. *Proof of Theorem 8*

The proof of the asymptotic normality for the proposed linearized partial rank estimation is structurally the same as the uncensored case. We omit the details here.

2.10. *Proof of Theorem 9*

More notations are introduced. Recall that $PL_n(\beta) = \sum_{i \neq j}^{n} I(Y_i < Y_j)(X_i - X_j)^{\mathrm{T}}\beta/\{n(n-1)\} + \lambda_n\|\beta\|_1 = -L_n(\beta) + \lambda_n\|\beta\|_1$ and $L(\beta) = E\{L_n(\beta)\}$, where $L_n(\cdot)$, $L(\cdot)$ are defined in the main context. Since the optimization is implemented on the manifold $\mathcal{E}(\Sigma)$, other than Conditions (C1)-(C2), to establish the oracle inequalities for high dimensional case, additional assumptions are needed.

(M*) The unknown function $f(\cdot,\cdot) : \mathbb{R}^2 \to \mathbb{R}$ is non-constant increasing in its first argument on the support of $(X^{\mathrm{T}}\beta_0, \epsilon)$ and X is independent of ϵ. And for all n, the variance of the random variable pr$\{Y_1 < Y_2 \mid X_2^{\mathrm{T}}\beta_0\}$ is bounded below by some universal positive constant.

(D1) (i) There exists a positive constant A_0 such that for any n and $\beta \in \mathbb{R}^{p_n}$ and any $t > 0$ such that pr$\{|\beta^{\mathrm{T}}X| \geq tA_0\|\beta\|_2\} \leq 2\exp(-t^2)$. (ii) There exist universal constants $\delta_0, \varepsilon_0 > 0$ such that pr$(|X^{\mathrm{T}}\beta_0 - E(X^{\mathrm{T}}\beta_0)| > \delta_0) \geq \varepsilon_0$ for any n.

(D2) There exists a universal positive constant c_0 such that all the eigenvalues of Σ, the covariance matrix of X, are bounded below by c_0.

Under fixed-dimensional case, we need the random variable $E\{f(X^{\mathrm{T}}\beta_0, \epsilon) \mid X^{\mathrm{T}}\beta_0\} = E\{Y \mid X^{\mathrm{T}}\beta_0\}$ has non-zero variance as in Assumption (M) to prove the consistency of the proposed estimator. This condition is actually a minimal model assumption to ensure a non-zero signal such that the parameter β_0 can be estimated under fixed dimensional settings. Similarly, under high-dimensional settings, Assumption (M*) is also a minimal model assumption to ensure non-zero signals for all n, which avoids the signals decay to 0 as $n \to \infty$. Note that Assumption (M*) is imposed for high-dimensional case, and Assumption (M) is sufficient for fixed dimensional case. For a high-dimensional linear model $Y = X^{\mathrm{T}}\beta_0 + \epsilon$, Assumption (M*) basically requires the variance of $X^{\mathrm{T}}\beta_0$ is uniformly greater than some universal positive constant for all n, which avoids the case that the linear model reduces to a degenerate and trivial model $Y = \epsilon$ as $n \to \infty$. Fan et al. (2020) studied rank estimators in increasing dimensions and imposed a similar identification condition by positing non-constant requirement on the objective function (at the population level) around the true parameter, whose first component is restricted to be 1 for identifiability.

Note that Assumption (D1) part (i) can lead to an upper bound on the spectrum of the covariance matrix Σ, i.e., there exists a universal positive constant C_0 such that all the eigenvalues of Σ, the covariance matrix of X, are bounded above by C_0. To show this, for any $\beta \in \mathbb{R}^{p_n}$, we have

$$\beta^{\mathrm{T}}\Sigma\beta = \mathrm{cov}(X^{\mathrm{T}}\beta) \leq E(|X^{\mathrm{T}}\beta|^2) = \int_0^\infty \mathrm{pr}\{|X^{\mathrm{T}}\beta|^2 > t\}dt.$$

Let $u = t^{1/2}/(A_0\|\beta\|_2)$. Then,

$$
\begin{aligned}
\beta^{\mathrm{T}}\Sigma\beta \leq & \int_0^\infty \mathrm{pr}\{|X^{\mathrm{T}}\beta|^2 > t\}dt \\
= & \int_0^\infty 2uA_0^2\|\beta\|_2^2 \mathrm{pr}\{|X^{\mathrm{T}}\beta| > uA_0\|\beta\|_2\}du \\
\leq & \int_0^\infty 4u\exp(-u^2)A_0^2\|\beta\|_2^2 du \\
= & \{-\exp(-u^2)\}\big|_0^\infty 2A_0^2\|\beta\|_2^2 \\
= & 2A_0^2\|\beta\|_2^2,
\end{aligned}
$$

where the second inequality follows from Assumption (D1) part (i). This implies that there exists a universal positive constant $C_0 \leq 2A_0^2$ such that all the eigenvalues of Σ, the covariance matrix of X, are bounded above by C_0.

Besides, we impose an additional part (ii) in Assumption (D1), which requires that the probability mass of $X^{\mathrm{T}}\beta_0$ does not concentrate around its mean, which is generally satisfied for many common continuous distributions. For example, if X follows normal distribution with mean μ and covariance matrix Σ, then $X^{\mathrm{T}}\beta_0$ follows $\mathcal{N}(\mu^{\mathrm{T}}\beta_0, 1)$ by the identifiability condition $\beta_0^{\mathrm{T}}\Sigma\beta_0 = 1$, which satisfies Assumption (D2) part (ii).

Proof. We first show that, under Condition (M*), (D1) and (D2), there is a local quadratic curvature of $L(\cdot)$ on the manifold $\mathcal{E}(\Sigma)$, i.e., there exists some universal $\kappa_L > 0$ such that for all n and any $\beta \in \mathcal{E}(\Sigma)$

$$
L(\beta) - L(\beta_0) \leq -\kappa_L\|\beta - \beta_0\|_2^2. \tag{S10}
$$

To this end, recall that $U = E\{I(Y_1 < Y_2)(X_2 - X_1)\}$ and $L(\beta) = U^{\mathrm{T}}\beta$. By the proof of Theorem 1, we have $\Sigma^{-1}U = c\beta_0$ where $c = (U^{\mathrm{T}}\Sigma^{-1}U)^{1/2} > 0$. Define $\Delta(\beta) := \beta - \beta_0$. In view of the identifiability condition and $\beta \in \mathcal{E}(\Sigma)$, we have $\beta_0^{\mathrm{T}}\Sigma\beta_0 = \beta^{\mathrm{T}}\Sigma\beta = \{\beta_0 + \Delta(\beta)\}^{\mathrm{T}}\Sigma\{\beta_0 + \Delta(\beta)\}$, which implies

$$
\beta_0^{\mathrm{T}}\Sigma\Delta(\beta) = -\frac{1}{2}\Delta(\beta)^{\mathrm{T}}\Sigma\Delta(\beta). \tag{S11}
$$

Then,

$$
\begin{aligned}
L(\beta) - L(\beta_0) &= U^{\mathrm{T}}(\beta - \beta_0) \\
&= U^{\mathrm{T}}\Sigma^{-1}\Sigma\Delta(\beta) \\
&= (\Sigma^{-1}U)^{\mathrm{T}}\Sigma\Delta(\beta) \\
&= c\beta_0^{\mathrm{T}}\Sigma\Delta(\beta) \\
&= -\frac{c}{2}\Delta(\beta)^{\mathrm{T}}\Sigma\Delta(\beta) \\
&= -\frac{1}{2}(U^{\mathrm{T}}\Sigma^{-1}U)^{1/2}\Delta(\beta)^{\mathrm{T}}\Sigma\Delta(\beta).
\end{aligned}
$$

Under Condition (D1) and (D2), all eigenvalues of Σ lie in $[c_0, C_0]$. Then,

$$
\begin{aligned}
L(\beta) - L(\beta_0) &= -\frac{1}{2}(U^{\mathrm{T}}\Sigma^{-1}U)^{1/2}\Delta(\beta)^{\mathrm{T}}\Sigma\Delta(\beta)\\
&\le -\frac{1}{2}C_0^{-1/2}\|U\|_2 \times c_0\|\Delta(\beta)\|_2^2\\
&= -\Big(\frac{1}{2}c_0C_0^{-1/2}\|U\|_2\Big) \times \|\beta-\beta_0\|_2^2.
\end{aligned}
$$

Let $\kappa_L := c_0C_0^{-1/2}\|U\|_2/2$, we intend to prove that $\kappa_L \ge c > 0$ for some universal positive constant c, as $U \in \mathbb{R}^{p_n}$ can change with the sample size n under the triangular array setting. To this end, we only need to show that there exists a universal constant c' such that $\|U\|_2 \ge c' > 0$ for all n.

For notational simplicity, we still use the notations $U, \beta_0, \mu, \Sigma, X$ and suppress their dependence on n. First, by Assumptions (D1)-(D2) and the identifiability condition $\beta_0^{\mathrm{T}}\Sigma\beta_0 = 1$, we have for all n,

$$
c_0\|\beta_0\|_2^2 \le \beta_0^{\mathrm{T}}\Sigma\beta_0 \le C_0\|\beta_0\|_2^2, \qquad C_0^{-1/2} \le \|\beta_0\|_2 \le c_0^{-1/2}.
$$

Note that

$$
\|U\|_2^2 = U^{\mathrm{T}}U \ge U^{\mathrm{T}}\beta_0\frac{\|U\|_2}{\|\beta_0\|_2} \ge c_0^{1/2}\|U\|_2U^{\mathrm{T}}\beta_0.
$$

If we can show that there exists some universal constant c'' such that $U^{\mathrm{T}}\beta_0 \ge c'' > 0$ for all n, then $\|U\|_2 > 0$ and $\|U\|_2 \ge c_0^{1/2}U^{\mathrm{T}}\beta_0 = c_0^{1/2}c'' > 0$. By the definition of U and the proof of Theorem 1, we have

$$
\begin{aligned}
U^{\mathrm{T}}\beta_0 =& E\{I(Y_1 < Y_2)(X_2 - X_1)^{\mathrm{T}}\beta_0\}\\
=& E\{I(Y_1 < Y_2)(X_2 - \mu)^{\mathrm{T}}\beta_0\} - E\{I(Y_1 < Y_2)(X_1 - \mu)^{\mathrm{T}}\beta_0\}\\
=& E[\{\mathrm{pr}(Y_1 < Y_2 \mid X_2^{\mathrm{T}}\beta_0) - 1/2\}\{(X_2 - \mu)^{\mathrm{T}}\beta_0\}]\\
&\qquad - E[\{\mathrm{pr}(Y_1 < Y_2 \mid X_1^{T}\beta_0) - 1/2\}\{(X_1 - \mu)^{\mathrm{T}}\beta_0\}].
\end{aligned}
$$

Now we prove that there exists a positive constant $c_1 > 0$ such that $E[\{\mathrm{pr}(Y_1 < Y_2 \mid X_2^{\mathrm{T}}\beta_0) - 1/2\}\{(X_2 - \mu)^{\mathrm{T}}\beta_0\}] \ge c_1 > 0$ for all n. Let $v_0 > 0$ denote the uniform lower bound of the variance of $\mathrm{pr}\{Y_1 < Y_2 \mid X_2^{\mathrm{T}}\beta_0\}$ under Assumption (M*). Note that $\mathrm{pr}(Y_1 < Y_2 \mid X_2^{\mathrm{T}}\beta_0) - 1/2 \in [-1/2, 1/2]$ is bounded for any n, then $v_0 \le \mathrm{var}\{\mathrm{pr}(Y_1 < Y_2 \mid X_2^{\mathrm{T}}\beta_0)\} \le 1/4$. For brevity, we use $g(X_2^{\mathrm{T}}\beta_0)$ to denote $\mathrm{pr}(Y_1 < Y_2 \mid X_2^{\mathrm{T}}\beta_0) - 1/2$. Then,

$$
\begin{aligned}
v_0 \le& \mathrm{var}\{g(X_2^{\mathrm{T}}\beta_0)\}\\
=& E[\{g(X_2^{\mathrm{T}}\beta_0)\}^2I(|g(X_2^{\mathrm{T}}\beta_0)| \le v_0^{1/2}/2)] + E[\{g(X_2^{\mathrm{T}}\beta_0)\}^2I(|g(X_2^{\mathrm{T}}\beta_0)| > v_0^{1/2}/2)]\\
\le& \mathrm{pr}\{|g(X_2^{\mathrm{T}}\beta_0)| \le v_0^{1/2}/2\}v_0/4 + \mathrm{pr}\{|g(X_2^{\mathrm{T}}\beta_0)| > v_0^{1/2}/2\}/4,
\end{aligned}
$$

which leads to

$$\mathrm{pr}\{|g(X_2^{\mathrm{T}}\beta_0)| > v_0^{1/2}/2\} \geq \frac{3v_0}{1-v_0}.$$

On the other hand, under Condition (D1) part (ii), there exist constants $\delta_0, \varepsilon_0 > 0$ such that $\mathrm{pr}(|X_2^{\mathrm{T}}\beta_0 - E(X_2^{\mathrm{T}}\beta_0)| > \delta_0) \geq \varepsilon_0$ for all n. Then, for any n,

$$\begin{aligned}
U^{\mathrm{T}}\beta_0 =& E[\{\mathrm{pr}(Y_1 < Y_2 \mid X_2^{\mathrm{T}}\beta_0) - 1/2\}\{(X_2-\mu)^{\mathrm{T}}\beta_0\}] \\
& \quad - E[\{\mathrm{pr}(Y_1 < Y_2 \mid X_1^{\mathrm{T}}\beta_0) - 1/2\}\{(X_1-\mu)^{\mathrm{T}}\beta_0\}] \\
\geq& E[\{\mathrm{pr}(Y_1 < Y_2 \mid X_2^{\mathrm{T}}\beta_0) - 1/2\}\{(X_2-\mu)^{\mathrm{T}}\beta_0\}] \\
=& E\{g(X_2^{\mathrm{T}}\beta_0)(X_2-\mu)^{\mathrm{T}}\beta_0\} \\
\geq& E[g(X_2^{\mathrm{T}}\beta_0)(X_2-\mu)^{\mathrm{T}}\beta_0 I\{|g(X_2^{\mathrm{T}}\beta_0)| > v_0^{1/2}/2\} I\{|X_2^{\mathrm{T}}\beta_0 - E(X_2^{\mathrm{T}}\beta_0)| > \delta_0\}] \\
\geq& \frac{v_0^{1/2}}{2}\delta_0 \min\left\{\frac{3v_0}{1-v_0}, \varepsilon_0\right\} > 0.
\end{aligned}$$

Hence, it has been shown that there exists a universal constant c' such that $\|U\|_2 \geq c' > 0$ for all n. As a result, we have shown that there exists some universal constant $\kappa_L > 0$ such that for all n, any $\beta \in \mathcal{E}(\Sigma)$

$$L(\beta) - L(\beta_0) \leq -\kappa_L \|\beta - \beta_0\|_2^2.$$

Next we carry out the proof of Theorem 9 in three steps.

Step 1. If $\lambda_n \geq 2\|\nabla L_n(\beta_0)\|_\infty$, then $\hat{\beta}_n - \beta_0 \in C(\mathcal{A})$, where $C(\mathcal{A}) = \{\alpha \in \mathbb{R}^{p_n} : \|\alpha_{\mathcal{A}^c}\|_1 \leq 3\|\alpha_{\mathcal{A}}\|_1\}$, i.e. $\|(\hat{\beta}_n - \beta_0)_{\mathcal{A}^c}\|_1 \leq 3\|(\hat{\beta}_n - \beta_0)_{\mathcal{A}}\|_1$. By the definition of $\hat{\beta}_n$, we have

$$\begin{aligned}
0 \geq& PL_n(\hat{\beta}_n) - PL_n(\beta_0) \\
=& \{-L_n(\hat{\beta}_n)\} - \{-L_n(\beta_0)\} + \lambda_n(\|\hat{\beta}_n\|_1 - \|\beta_0\|_1) \\
=& (\hat{\beta}_n - \beta_0)^{\mathrm{T}}\nabla(-L_n)(\beta_0) + \lambda_n(\|\hat{\beta}_n\|_1 - \|\beta_0\|_1) \\
\geq& -\|\hat{\beta}_n - \beta_0\|_1\|\nabla L_n(\beta_0)\|_\infty + \lambda_n(\|\hat{\beta}_n\|_1 - \|\beta_0\|_1) \\
\geq& -\frac{\lambda_n}{2}\|\hat{\beta}_n - \beta_0\|_1 + \lambda_n(\|\hat{\beta}_n\|_1 - \|\beta_0\|_1) \\
=& -\frac{\lambda_n}{2}\|\hat{\beta}_n - \beta_0\|_1 + \lambda_n(\|(\hat{\beta}_n - \beta_0 + \beta_0)_{\mathcal{A}^c}\|_1 + \|(\hat{\beta}_n - \beta_0 + \beta_0)_{\mathcal{A}}\|_1 - \|\beta_0\|_1) \\
=& -\frac{\lambda_n}{2}\|\hat{\beta}_n - \beta_0\|_1 + \lambda_n(\|(\hat{\beta}_n - \beta_0)_{\mathcal{A}^c}\|_1 + \|(\hat{\beta}_n - \beta_0 + \beta_0)_{\mathcal{A}}\|_1 - \|\beta_0\|_1) \\
\geq& -\frac{\lambda_n}{2}(\|(\hat{\beta}_n - \beta_0)_{\mathcal{A}}\|_1 + \|(\hat{\beta}_n - \beta_0)_{\mathcal{A}^c}\|_1) + \lambda_n(\|(\hat{\beta}_n - \beta_0)_{\mathcal{A}^c}\|_1 - \|(\hat{\beta}_n - \beta_0)_{\mathcal{A}}\|_1) \\
=& -\frac{\lambda_n}{2}(3\|(\hat{\beta}_n - \beta_0)_{\mathcal{A}}\|_1 - \|(\hat{\beta}_n - \beta_0)_{\mathcal{A}^c}\|_1).
\end{aligned}$$

The inequality follows from $\|\hat{\beta}_n - \beta_0\|_1 = \|(\hat{\beta}_n - \beta_0)_{\mathcal{A}}\|_1 + \|(\hat{\beta}_n - \beta_0)_{\mathcal{A}^c}\|_1$ and $\beta_{0,\mathcal{A}^c} = 0$.

Step 2. For $\lambda_n = a_n\{\log(n)\log(p_n)/n\}^{1/2}$, the probability of $\lambda_n \geq 2\|\nabla L_n(\beta_0)\|_\infty$ is greater than $1 - 2\exp(-a_n) - 2/p_n$. Let e_j be the unit vector with its j-th component being 1 and others 0. Taking β in condition (D1) as e_j, for each $i = 1, \ldots, n$, we have

$$\mathrm{pr}\{|e_j^{\mathrm{T}} X_i| \geq tA_0\} \leq 2\exp(-t^2),$$

and

$$\mathrm{pr}\{|e_j^{\mathrm{T}} X_1| \leq BA_0, \ldots, |e_j^{\mathrm{T}} X_n| \leq BA_0\} \geq 1 - 2n\exp(-B^2).$$

For the j-th component of $\nabla L_n(\beta_0)$, given $\{|e_j^{\mathrm{T}} X_1| \leq BA_0, \ldots, |e_j^{\mathrm{T}} X_n| \leq BA_0\}$, then $e_j^{\mathrm{T}} \nabla L_n(\beta_0)$ is a U-statistic with kernel bounded by $2BA_0$. By the concentration inequality with bounded kernel in Hoeffding (1994), there exists some constant $c_1 > 0$ depending only on A_0, such that with probability at least $1 - 2n\exp(-B^2)$,

$$\mathrm{pr}\{2|e_j^{\mathrm{T}} \nabla L_n(\beta_0)| \geq \lambda_n\} \leq c_1 \exp\{-n\lambda_n^2/(16B^2A_0^2)\}.$$

By Condition (D1), the above inequality holds for any e_j. Thus,

$$\mathrm{pr}\{2\|\nabla L_n(\beta_0)\|_\infty \leq \lambda_n\} \geq 1 - c_1\exp\{-n\lambda_n^2/(16B^2A_0^2)\} - 2n\exp(-B^2).$$

Taking $B = \{a_n \log(n)\}^{1/2}$ and $\lambda_n = 4A_0 B\{\log(p_n)/n\}^{1/2} = 4A_0\{a_n \log(n)\log(p_n)/n\}^{1/2}$, we obtain that with probability at least $1 - 2\exp(-a_n) - c_1/p_n$, $\lambda_n \geq 2\|\nabla L_n(\beta_0)\|_\infty$. Here a_n is a sequence of positive numbers diverging to ∞ as $n \to \infty$, and the rate of a_n diverging to ∞ can be arbitrarily slow.

Step 3. We will show that with probability at least $1 - 2\exp(-a_n) - 2/p_n$,

$$\{-L_n(\hat{\beta}_n)\} - \{-L_n(\beta)\} \geq \kappa_L\|\hat{\beta}_n - \beta_0\|_2^2 - 2^{3/2}A_0\left\{\frac{a_n\log(n)\log(p_n)}{n}\right\}^{1/2}\|\hat{\beta}_n - \beta_0\|_2. \tag{S12}$$

For any $\alpha \in \mathbb{R}^{p_n}$ and given $B > 0$, $\mathrm{pr}\{|\alpha^{\mathrm{T}} X_i| \geq BA_0\|\alpha\|_2\} \leq 2\exp(-B^2)$. Then,

$$\mathrm{pr}\{|\alpha^{\mathrm{T}} X_1| \leq BA_0\|\alpha\|_2, \ldots, |\alpha^{\mathrm{T}} X_n| \leq BA_0\|\alpha\|_2\} \geq 1 - 2n\exp(-B^2).$$

For any $\delta > 0$, define $C(\mathcal{A}, \delta) = C(\mathcal{A}) \cap \{\alpha \in \mathcal{R}^{p_n} : \|\alpha\|_2 = \delta\}$. Note that if $\hat{\beta}_n - \beta_0 \in C(\mathcal{A}, \delta)$, the probability for the occurrence of the event $\{|I(Y_i < Y_j)(X_j - X_i)^{\mathrm{T}}(\hat{\beta}_n - \beta_0)| \leq BA_0\delta, \text{for all } i, j = 1, \ldots, n\}$ is at least $1 - 2n\exp(-B^2)$. Then, by the bounded difference inequality in Hoeffding (1994), with probability at least $1 - 2n\exp(-B^2)$, we have

$$\mathrm{pr}\{|L_n(\hat{\beta}_n) - L_n(\beta) - \{L(\hat{\beta}_n) - L(\beta)\}| \geq t\} \leq 2\exp\Big(-\frac{nt^2}{8B^2A_0^2\|\hat{\beta}_n - \beta_0\|_2^2}\Big).$$

Recall that $B = \{a_n \log(n)\}^{1/2}$. Taking $t = 2^{3/2}A_0\|\hat{\beta}_n - \beta_0\|_2 B\{\log(p_n)/n\}^{1/2}$, together with inequality (S10), we have proved (S12).

Step 4. For any $\delta > 0$ and all $\hat{\beta}_n - \beta_0 \in C(\mathcal{A}, \delta)$, with probability at least $1 - 4\exp(-a_n) - (2 + c_1)/p_n$, we have

$$
\begin{aligned}
0 &\geq PL_n(\hat{\beta}_n) - PL_n(\beta_0) \\
&\geq \kappa_L\|\hat{\beta}_n - \beta_0\|_2^2 - 2^{3/2}A_0 a_n \left\{\frac{\log(n)\log(p_n)}{n}\right\}^{1/2} \|\hat{\beta}_n - \beta_0\|_2 - \lambda_n\|(\hat{\beta}_n - \beta_0)_{\mathcal{A}}\|_1 \\
&\geq \kappa_L\|\hat{\beta}_n - \beta_0\|_2^2 - 2^{3/2}A_0 a_n \left\{\frac{\log(n)\log(p_n)}{n}\right\}^{1/2} \|\hat{\beta}_n - \beta_0\|_2 - (s_n)^{1/2}\lambda_n\|(\hat{\beta}_n - \beta_0)_{\mathcal{A}}\|_2 \\
&\geq \kappa_L\|\hat{\beta}_n - \beta_0\|_2^2 - \left(2^{3/2}A_0 a_n \left\{\frac{\log(n)\log(p_n)}{n}\right\}^{1/2} + (s_n)^{1/2}\lambda_n\right) \|\hat{\beta}_n - \beta_0\|_2 \\
&\geq \kappa_L\|\hat{\beta}_n - \beta_0\|_2^2 - \left(2^{-1/2} + (s_n)^{1/2}\right)\lambda_n\|\hat{\beta}_n - \beta_0\|_2,
\end{aligned}
$$

which suggests that $\delta \leq \{2^{-1/2} + (s_n)^{1/2}\}\lambda_n/\kappa_L$. Since $\beta_0 \in \mathcal{E}(\Sigma)$, we have $s_n \geq 1$. Consequently, when $\lambda_n = 2^{3/2}A_0\{a_n \log(n)\log(p_n)/n\}^{1/2}$, with probability at least $1 - 4\exp(-a_n) - (c_1 + 2)/p_n$,

$$
\begin{aligned}
\|\hat{\beta}_n - \beta_0\|_2 &\leq 2(s_n)^{1/2}\lambda_n/\kappa_L, \\
\|\hat{\beta}_n - \beta_0\|_1 &\leq 2s_n\lambda_n/\kappa_L,
\end{aligned}
$$

where $c_1 > 0$ is a constant depending only on A_0. We have completed the proof of Theorem 9.□

3. ADDITIONAL SIMULATION RESULTS

In this section, we present some additional simulation results in Tables S1-S3 and Table S5. For checking the robustness of our methods without the monotonicity assumption of the link function f, additional simulation results under three models: M5: $Y = (X^{\mathrm{T}}\beta_0)^2 + \epsilon$; M6: $Y = (X^{\mathrm{T}}\beta_0)^3 + 5(X^{\mathrm{T}}\beta_0)^2 - 3(X^{\mathrm{T}}\beta_0) + \epsilon$; M7: $Y = 5\sin(X^{\mathrm{T}}\beta_0) + \epsilon$, are presented in Table S4. We set $p = 5$ and $\beta_0 = (1, 1, 0, 0, -1)^{\mathrm{T}}$. And the covariate X are generated from a multivariate normal distribution with mean 0 and covariance matrix $\Sigma = (\rho_{ij})$ with $\rho_{ij} = \rho_0^{|i-j|}$ and $\rho_0 = 0.3$.

Table S1. *Summary statistics with dimension $p = 5$ and correlation $\rho_0 = 0.3$. Averaged absolute bias (BIAS), standard errors (SE) and coverage probability (CP) of 95% confidence interval over components of the index parameter. Mean ℓ_1 and ℓ_2 distances between the estimate and the true parameter.*

Model	Error	Method	$n = 100$					$n = 200$				
			BIAS	SE	CP	ℓ_1	ℓ_2	BIAS	SE	CP	ℓ_1	ℓ_2
M3	$\chi^2(1)$	LMRC*	0.021	0.193	0.947	0.783	0.411	0.013	0.138	0.953	0.553	0.293
		LMRC	0.005	0.097	0.947	0.382	0.047	0.003	0.066	0.942	0.266	0.022
		SIR	0.008	0.139	-	0.539	0.097	0.006	0.094	-	0.375	0.045
		LSE	0.034	0.236	-	0.942	0.496	0.029	0.189	-	0.756	0.397
		MRC	0.200	0.538	0.864	2.096	1.142	0.082	0.337	0.923	0.993	0.557
		MRE	0.444	0.739	0.990	3.597	1.930	0.392	0.725	0.998	3.373	1.870
	Pois(1)	LMRC*	0.024	0.192	0.956	0.771	0.409	0.010	0.138	0.953	0.554	0.291
		LMRC	0.004	0.096	0.957	0.378	0.046	0.001	0.071	0.944	0.286	0.026
		SIR	0.008	0.124	-	0.493	0.078	0.003	0.082	-	0.320	0.034
		LSE	0.037	0.234	-	0.924	0.488	0.023	0.191	-	0.752	0.391
		MRC	0.183	0.501	0.862	1.923	1.047	0.081	0.349	0.924	1.018	0.571
		MRE	0.399	0.722	0.994	3.391	1.835	0.395	0.728	0.994	3.432	1.859
M4	$\chi^2(1)$	LMRC*	0.040	0.275	0.948	1.113	0.590	0.026	0.199	0.940	0.800	0.422
		LMRC	0.017	0.241	0.925	0.957	0.293	0.009	0.161	0.942	0.650	0.130
		SIR	0.018	0.018	-	0.804	0.206	0.008	0.137	-	0.554	0.095
		LSE	0.378	0.190	-	2.051	1.184	0.379	0.136	-	1.982	1.156
		MRC	0.185	0.501	0.828	2.037	1.093	0.105	0.377	0.911	1.392	0.771
		MRE	0.378	0.694	0.990	3.095	1.709	0.378	0.707	1.000	3.205	1.755
	Pois(1)	LMRC*	0.044	0.282	0.940	1.152	0.610	0.024	0.208	0.949	0.842	0.444
		LMRC	0.036	0.234	0.941	0.937	0.281	0.007	0.172	0.951	0.683	0.147
		SIR	0.022	0.199	-	0.796	0.201	0.009	0.141	-	0.560	0.100
		LSE	0.403	0.199	-	2.163	1.248	0.396	0.135	-	2.056	1.209
		MRC	0.190	0.507	0.834	2.097	1.133	0.125	0.415	0.907	1.564	0.857
		MRE	0.405	0.718	0.994	3.317	1.819	0.466	0.744	0.998	3.614	2.002

Table S2. *Summary statistics with dimension $p = 15$ and correlation $\rho_0 = 0.3$. Averaged absolute bias (BIAS), standard errors (SE) and coverage probability (CP) of 95% confidence interval over components of the index parameter. Mean ℓ_1 and ℓ_2 distances between the estimate and the true parameter.*

Model	Error	Method	$n = 200$					$n = 400$				
			BIAS	SE	CP	ℓ_1	ℓ_2	BIAS	SE	CP	ℓ_1	ℓ_2
M1	$\chi^2(1)$	LMRC*	0.029	0.240	0.951	2.900	0.922	0.018	0.175	0.944	2.118	0.670
		LMRC	0.005	0.123	0.945	1.467	0.228	0.006	0.085	0.938	1.023	0.109
		SIR	0.013	0.166	-	1.980	0.418	0.007	0.109	-	1.302	0.181
		LSE	0.009	0.135	-	1.614	0.275	0.049	0.073	-	1.133	0.360
		MRC	0.202	0.508	0.757	6.636	2.136	0.144	0.442	0.904	5.360	1.775
		MRE	0.517	0.764	0.827	11.750	3.849	0.563	0.772	0.999	12.344	4.033
	Pois(1)	LMRC*	0.033	0.231	0.945	2.806	0.891	0.016	0.167	0.947	2.007	0.637
		LMRC	0.006	0.100	0.958	1.192	0.149	0.004	0.073	0.942	0.862	0.080
		SIR	0.008	0.121	-	1.446	0.220	0.004	0.085	-	1.013	0.108
		LSE	0.006	0.110	-	1.312	0.181	0.050	0.053	-	0.967	0.306
		MRC	0.191	0.509	0.750	6.603	2.124	0.142	0.442	0.890	5.276	1.769
		MRE	0.095	0.332	0.841	3.879	1.980	0.595	0.767	0.998	12.277	4.077
M2	$\chi^2(1)$	LMRC*	0.036	0.247	0.943	2.977	0.949	0.019	0.181	0.946	2.193	0.694
		LMRC	0.011	0.138	0.956	1.643	0.286	0.006	0.098	0.938	1.176	0.145
		SIR	0.021	0.204	-	2.433	0.634	0.008	0.122	-	1.451	0.226
		LSE	0.244	0.766	-	9.703	10.152	0.254	0.696	-	8.939	2.852
		MRC	0.203	0.513	0.769	6.681	2.150	0.167	0.470	0.879	5.860	1.908
		MRE	0.515	0.762	0.834	11.829	3.890	0.570	0.767	0.999	12.119	4.020
	Pois(1)	LMRC*	0.034	0.242	0.951	2.948	0.934	0.018	0.174	0.951	2.099	0.667
		LMRC	0.009	0.123	0.950	1.474	0.227	0.007	0.089	0.937	1.067	0.119
		SIR	0.008	0.122	-	1.451	0.226	0.004	0.075	-	0.895	0.085
		LSE	0.158	0.533	-	6.608	2.108	0.131	0.478	-	5.887	1.866
		MRC	0.202	0.515	0.863	6.724	2.156	0.138	0.439	0.882	5.466	1.804
		MRE	0.521	0.765	0.850	11.869	3.857	0.562	0.765	0.998	12.118	3.980
M3	$\chi^2(1)$	LMRC*	0.031	0.224	0.953	2.721	0.864	0.016	0.163	0.952	1.969	0.624
		LMRC	0.007	0.076	0.952	0.912	0.087	0.003	0.052	0.937	0.621	0.041
		SIR	0.013	0.156	-	1.854	0.367	0.006	0.102	-	1.217	0.157
		LSE	0.133	0.484	-	6.019	1.915	0.115	0.429	-	5.266	1.680
		MRC	0.162	0.462	0.866	5.742	1.879	0.122	0.401	0.883	4.685	1.583
		MRE	0.465	0.740	0.995	10.524	3.589	0.504	0.749	0.997	11.206	3.728
	Pois(1)	LMRC*	0.029	0.222	0.951	2.690	0.851	0.016	0.161	0.947	1.939	0.614
		LMRC	0.004	0.076	0.952	0.905	0.087	0.002	0.051	0.950	0.602	0.038
		SIR	0.008	0.120	-	1.419	0.216	0.004	0.077	-	0.919	0.090
		LSE	0.136	0.492	-	6.126	1.946	0.112	0.422	-	5.213	1.657
		MRC	0.160	0.458	0.858	5.700	1.882	0.123	0.404	0.885	4.677	1.583
		MRE	0.457	0.735	0.997	10.636	3.533	0.528	0.753	0.999	11.405	3.824
M4	$\chi^2(1)$	LMRC*	0.068	0.350	0.938	4.282	1.368	0.037	0.259	0.944	3.168	1.001
		LMRC	0.041	0.319	0.925	3.835	1.561	0.021	0.230	0.933	2.776	0.803
		SIR	0.021	0.209	-	2.505	0.665	0.013	0.147	-	1.767	0.327
		LSE	0.361	0.241	-	6.134	2.006	0.352	0.169	-	5.682	1.878
		MRC	0.225	0.554	0.817	7.440	2.373	0.170	0.480	0.850	6.105	1.970
		MRE	0.472	0.734	0.999	10.533	3.575	0.508	0.756	0.999	11.178	3.782
	Pois(1)	LMRC*	0.073	0.355	0.920	4.352	1.389	0.041	0.267	0.943	3.250	1.029
		LMRC	0.045	0.334	0.922	4.026	1.717	0.026	0.234	0.926	2.794	0.832
		SIR	0.020	0.202	-	2.438	0.618	0.009	0.144	-	1.734	0.313
		LSE	0.371	0.242	-	6.249	2.051	0.369	0.167	-	5.889	1.945
		MRC	0.229	0.565	0.811	7.568	2.398	0.164	0.478	0.850	6.049	1.961
		MRE	0.485	0.748	0.999	10.850	3.669	0.541	0.757	0.998	11.415	3.866

Table S3. *Summary statistics with dimension $p = 30$ and $\rho = 0.8$. Averaged absolute bias (BIAS), standard errors (SE) and coverage probability (CP) of 95% confidence interval over components of the index parameter. Mean ℓ_1 and ℓ_2 distances between the estimate and the true parameter.*

Model	Error	Method	$n = 200$					$n = 300$				
			BIAS	SE	CP	ℓ_1	ℓ_2	BIAS	SE	CP	ℓ_1	ℓ_2
M3	$\chi^2(1)$	LMRC*	0.043	0.559	0.944	13.529	3.042	0.044	0.468	0.943	11.290	2.550
		LMRC	0.009	0.149	0.972	3.538	0.672	0.009	0.116	0.956	2.776	0.408
		SIR	0.095	0.786	-	18.200	19.008	0.084	0.634	-	14.173	12.435
		LSE	0.411	2.274	-	55.602	160.928	0.369	2.092	-	51.313	136.856
		MRC	0.332	0.614	0.449	18.458	16.963	0.280	0.583	0.460	17.005	14.365
		MRE	0.326	0.618	0.590	18.830	17.451	0.280	0.595	0.626	17.338	14.835
	Pois(1)	LMRC*	0.051	0.551	0.950	13.342	3.002	0.039	0.465	0.942	11.189	2.523
		LMRC	0.010	0.149	0.965	3.561	0.672	0.006	0.122	0.954	2.921	0.452
		SIR	0.045	0.519	-	11.835	8.176	0.032	0.357	-	8.465	3.854
		LSE	0.349	2.307	-	56.571	164.265	0.394	2.114	-	51.931	140.305
		MRC	0.314	0.622	0.466	18.264	16.363	0.274	0.580	0.459	16.844	13.937
		MRE	0.321	0.625	0.607	18.424	16.725	0.270	0.584	0.622	17.041	14.221
M4	$\chi^2(1)$	LMRC*	0.175	1.028	0.874	25.198	5.687	0.128	0.880	0.905	21.413	4.831
		LMRC	0.155	1.023	0.840	24.802	32.333	0.141	0.873	0.891	21.152	23.608
		SIR	0.171	0.910	-	21.339	26.060	0.099	0.730	-	17.032	16.456
		LSE	0.156	1.534	-	36.985	71.364	0.153	1.267	-	30.695	49.170
		MRC	0.361	0.658	0.477	19.620	19.047	0.321	0.631	0.473	18.638	16.934
		MRE	0.362	0.641	0.614	19.486	18.621	0.319	0.619	0.635	18.258	16.323
	Pois(1)	LMRC*	0.182	1.002	0.880	24.524	5.544	0.144	0.899	0.903	21.825	4.940
		LMRC	0.161	1.057	0.790	25.563	34.538	0.115	0.887	0.885	21.341	24.105
		SIR	0.149	0.923	-	21.490	26.604	0.078	0.694	-	16.230	14.645
		LSE	0.128	1.494	-	35.762	67.525	0.141	1.237	-	30.202	46.684
		MRC	0.370	0.659	0.463	19.860	19.577	0.297	0.604	0.487	17.810	15.466
		MRE	0.357	0.652	0.622	19.661	18.980	0.301	0.616	0.642	17.918	15.731

Table S4. *Summary statistics with dimension $p = 5$ and correlation $\rho_0 = 0.3$. Averaged absolute bias (BIAS), standard errors (SE) and coverage probability (CP) of 95% confidence interval over components of the index parameter. Mean ℓ_1 and ℓ_2 distances between the estimate and the true parameter.*

Model	Error	Method	$n = 100$					$n = 200$				
			BIAS	SE	CP	ℓ_1	ℓ_2	BIAS	SE	CP	ℓ_1	ℓ_2
M5	$\chi^2(1)$	LMRC*	0.085	0.401	0.936	1.526	0.809	0.032	0.249	0.945	0.996	0.528
		LMRC	0.057	0.348	0.959	1.360	0.629	0.020	0.231	0.949	0.916	0.276
		SIR	0.548	0.818	-	4.434	2.525	0.564	0.826	-	4.524	2.530
		LSE	0.361	0.780	-	3.551	2.084	0.372	0.772	-	3.524	2.069
		MRC	0.464	0.704	0.821	3.719	2.116	0.477	0.713	0.824	3.775	2.143
		MRE	0.473	0.704	0.861	3.729	2.110	0.471	0.714	0.848	3.736	2.099
	Pois(1)	LMRC*	0.058	0.375	0.911	1.457	0.876	0.030	0.253	0.931	1.006	0.728
		LMRC	0.022	0.300	0.955	1.166	0.451	0.017	0.226	0.954	0.877	0.258
		SIR	0.538	0.823	-	4.512	2.509	0.494	0.809	-	4.294	2.393
		LSE	0.366	0.772	-	3.565	2.084	0.385	0.772	-	3.574	2.099
		MRC	0.443	0.708	0.793	3.607	2.050	0.479	0.717	0.819	3.766	2.129
		MRE	0.448	0.711	0.869	3.726	2.111	0.504	0.715	0.869	3.837	2.157
M6	$\chi^2(1)$	LMRC*	0.047	0.336	0.926	1.351	0.843	0.024	0.237	0.945	0.960	0.708
		LMRC	0.019	0.282	0.953	1.110	0.401	0.025	0.214	0.946	0.858	0.234
		SIR	0.100	0.496	-	1.764	1.161	0.027	0.252	-	0.947	0.569
		LSE	0.148	0.536	-	2.064	1.266	0.056	0.359	-	1.339	0.820
		MRC	0.458	0.710	0.813	3.671	2.090	0.457	0.704	0.818	3.619	2.071
		MRE	0.456	0.708	0.863	3.632	2.072	0.468	0.708	0.872	3.725	2.104
	Pois(1)	LMRC*	0.041	0.313	0.935	1.267	0.666	0.027	0.238	0.933	0.951	0.509
		LMRC	0.021	0.285	0.942	1.130	0.409	0.016	0.202	0.956	0.798	0.207
		SIR	0.081	0.425	-	1.526	0.983	0.024	0.251	-	0.966	0.564
		LSE	0.111	0.486	-	1.824	1.135	0.060	0.392	-	1.428	0.894
		MRC	0.454	0.710	0.790	3.715	2.101	0.454	0.715	0.816	3.702	2.108
		MRE	0.466	0.714	0.856	3.741	2.119	0.427	0.699	0.876	3.664	2.059
M7	$\chi^2(1)$	LMRC*	0.280	0.711	0.878	3.069	1.627	0.130	0.539	0.910	2.134	1.121
		LMRC	0.213	0.666	0.918	2.751	2.649	0.163	0.560	0.926	2.181	1.785
		SIR	0.353	0.794	-	3.553	2.118	0.264	0.685	-	2.829	1.735
		LSE	0.084	0.462	-	1.728	1.063	0.037	0.316	-	1.198	0.713
		MRC	0.114	0.394	0.855	1.373	0.929	0.082	0.311	0.819	1.010	0.737
		MRE	0.125	0.410	0.878	1.489	0.976	0.080	0.322	0.894	1.003	0.777
	Pois(1)	LMRC*	0.260	0.674	0.886	2.778	1.496	0.148	0.551	0.911	2.180	1.146
		LMRC	0.234	0.699	0.896	2.911	3.004	0.147	0.566	0.915	2.222	1.805
		SIR	0.082	0.435	-	1.528	1.006	0.264	0.695	-	2.906	3.088
		LSE	0.136	0.536	-	2.046	1.607	0.036	0.284	-	1.102	0.413
		MRC	0.441	0.698	0.802	3.584	4.164	0.071	0.286	0.802	0.932	0.480
		MRE	0.459	0.709	0.838	3.706	4.318	0.079	0.300	0.905	0.974	0.521

Table S5. *Summary statistics with dimension $p = 40$. Averaged absolute bias (BIAS), standard errors (SE) over components of the index parameter. Mean ℓ_1 and ℓ_2 distances between the estimate and the true parameter. Averaged false positive rate (FP), false negative rate (FN), the empirical probability of choosing the correct model (CM).*

Dimension	Model	Error	Method	$n = 100$						
				BIAS	SE	ℓ_1	ℓ_2	FP	FN	CM
$p = 40$	M1	$\chi^2(1)$	Lasso LMRC	0.003	0.036	1.130	0.602	0.000	0.000	1.000
			Lasso SIR	0.010	0.135	2.823	0.850	0.337	0.000	1.000
		Pois(1)	Lasso LMRC	0.003	0.032	0.992	0.526	0.000	0.000	1.000
			Lasso SIR	0.008	0.110	2.291	0.698	0.326	0.000	1.000
	M2	$\chi^2(1)$	Lasso LMRC	0.002	0.034	1.078	0.570	0.000	0.000	1.000
			Lasso SIR	0.010	0.134	2.794	0.837	0.347	0.000	1.000
		Pois(1)	Lasso LMRC	0.002	0.032	0.994	0.528	0.000	0.000	1.000
			Lasso SIR	0.008	0.115	2.423	0.731	0.331	0.000	1.000
	M3	$\chi^2(1)$	Lasso LMRC	0.001	0.036	1.130	0.609	0.000	0.000	1.000
			Lasso SIR	0.022	0.236	5.065	1.518	0.344	0.001	0.997
		Pois(1)	Lasso LMRC	0.002	0.034	1.054	0.563	0.007	0.001	0.996
			Lasso SIR	0.022	0.221	4.725	1.407	0.340	0.000	1.000
	M4	$\chi^2(1)$	Lasso LMRC	0.001	0.025	0.792	0.413	0.000	0.000	1.000
			Lasso SIR	0.012	0.164	3.370	1.045	0.318	0.000	1.000
		Pois(1)	Lasso LMRC	0.002	0.026	0.826	0.431	0.000	0.000	1.000
			Lasso SIR	0.011	0.149	3.079	0.939	0.325	0.000	1.000

References

Burchard, A. (2009). A short course on rearrangement inequalities. *Lecture notes, IMDEA Winter School, Madrid* .

Burden, A. M., Burden, R. L. & Faires, J. D. (2016). *Numerical Analysis, 10th ed.* Cengage.

Cambanis, S., Huang, S. & Simons, G. (1981). On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis* **11**, 368–385.

Chen, S., Ma, S., Man-Cho So, A. & Zhang, T. (2020). Proximal gradient method for nonsmooth optimization over the stiefel manifold. *SIAM Journal on Optimization* **30**, 210–239.

Fan, Y., Han, F., Li, W. & Zhou, X.-H. (2020). On rank estimators in increasing dimensions. *Journal of Econometrics* **214**, 379–412.

Ferreira, O. & Oliveira, P. (2002). Proximal point algorithm on riemannian manifolds. *Optimization* **51**, 257–270.

Granville, V., Křivánek, M. & Rasson, J.-P. (1994). Simulated annealing: A proof of convergence. *IEEE transactions on pattern analysis and machine intelligence* **16**, 652–656.

Hoeffding, W. (1992). A class of statistics with asymptotically normal distribution. In *Breakthroughs in statistics*. Springer, pp. 308–334.

Hoeffding, W. (1994). Probability inequalities for sums of bounded random variables. In *The collected works of Wassily Hoeffding*. Springer, pp. 409–426.

Huang, N. & Ma, C. (2014). Convergence analysis and numerical study of a fixed-point iterative method for solving systems of nonlinear equations. *The Scientific World Journal* **2014**.

Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. (1983). Optimization by simulated annealing. *science* **220**, 671–680.

Nolan, D. & Pollard, D. (1987). U-processes: rates of convergence. *The Annals of Statistics* , 780–799.

Pakes, A. & Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica: Journal of the Econometric Society* , 1027–1057.

Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica: Journal of the Econometric Society* , 123–137.