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

# 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

## SUMMARY

We propose a linearized maximum rank correlation estimator for the single index model. Unlike the existing maximum rank correlation and other rank-based methods, the proposed estimator has a closed-form expression, making it appealing in theory and computation. The proposed estimator is robust to outliers in the response and its construction does not need the knowledge of the unknown link function or the error distribution. Under mild conditions, it is shown to be consistent and asymptotically normal when the predictors satisfy the linearity of expectation assumption. A more general class of estimators is also studied. Inference procedures based on the plug-in rule or random weighting resampling are employed for variance estimation. The proposed method can be easily modified to accommodate censored data. It can also be extended to deal with high-dimensional data combined with a penalty function. Extensive simulation studies provide strong evidence supporting that the proposed method works well in various practical situations. Its application is illustrated with the Beijing PM 2.5 dataset.

*Some key words*: Closed-form solution; Censored data; Linearized maximum rank correlation; Single index model.

## 1. INTRODUCTION

The single index model with an unknown univariate link function is popular and widely-used in econometrics and statistics. Let $(Y, X)$ be a pair of response and $p$-vector of covariates. A classical single index model assumes

$$Y = g(X^{\mathrm{T}}\beta_0) + \epsilon, \tag{1}$$

where $\epsilon$ is a random error satisfying $E(\epsilon \mid X) = 0$, $\beta_0$ is an unknown $p$-dimensional index coefficient vector, and $g(\cdot) : \mathbb{R} \to \mathbb{R}$ is an unknown link function. There is an extensive literature on statistical inference for model (1), focusing on the estimation of the index parameters $\beta_0$ and the link function $g$ (Xia & Li, 1999; Hristache et al., 2001; Xia & Tong, 2006; Kong & Xia, 2007; Horowitz, 2009; Wang et al., 2010). Han (1987) studied a semiparametric monotonic linear index model

$$Y = D \circ F(X^{\mathrm{T}}\beta_0, \epsilon), \tag{2}$$

where the function $D : \mathbb{R} \to \mathbb{R}$ is non-constant increasing, the function $F : \mathbb{R}^2 \to \mathbb{R}$ is strictly increasing in each of its arguments, and $Y, X, \epsilon, \beta_0$ are the same as in model (1). Given a random sample $(X_i, Y_i)(i = 1, \ldots, n)$ of size $n$, Han (1987) proposed maximum rank correlation estimation based on Kendall's $\tau$ for estimating $\beta_0$ by maximizing $\sum_{i \neq j} I(Y_i < Y_j) I(X_i^{\mathrm{T}}\beta < X_j^{\mathrm{T}}\beta)$ over $\beta$, where $I(\cdot)$ is the indicator function. Maximum rank correlation estimator is a nonparametric and distribution-free estimator. A similar class of monotone rank estimators was considered by Cavanagh & Sherman (1998) by maximizing $\sum_{i=1}^{n} M(Y_i) R_n(X_i^{\mathrm{T}}\beta)$ over $\beta$, where $M(\cdot)$ is a known increasing function on real line and $R_n(X_i^{\mathrm{T}}\beta)$ is the rank of $X_i^{\mathrm{T}}\beta$ among $X_1^{\mathrm{T}}\beta, X_2^{\mathrm{T}}\beta, \ldots, X_n^{\mathrm{T}}\beta$. Khan & Tamer (2007) extended maximum rank correlation to accommodate censored data and introduced the partial rank estimator under a general form of censoring. The objective functions of maximum rank correlation and the monotone rank estimation are neither continuous nor convex. Though the computation of maximum rank correlation can be carried out by the Nelder-Mead simplex search algorithm, it could be numerically unstable or even fail when the number of predictors is relatively large. To alleviate the numerical complications, a smoothed approximation of the partial rank estimator was studied by Song et al. (2007). Rank estimators in high dimensions were studied by Han et al. (2017) and Fan et al. (2020).

In this paper, we consider a general single index model

$$Y = f(X^{\mathrm{T}}\beta_0, \epsilon), \tag{3}$$

where $\epsilon$ is a random error, $\beta_0$ is an unknown $p$-dimensional index coefficient vector, and $f(\cdot, \cdot) : \mathbb{R}^2 \to \mathbb{R}$ is an unknown function satisfying condition (M) in section 2.2. Many popular models, such as the linear model, binary choice model, censored regression, duration model, transformation model, model (1) and model (2) fit into this framework. Recently, Neykov et al. (2016b) proved that when $X$ is Gaussian or satisfies the linearity of expectation in the direction of $\beta_0$, the least squares estimation is consistent for estimating $\beta_0$ up to a multiplicative scalar for model (3), without the monotonicity assumption in the first argument of $f$. This finding was further applied to recover the support of $\beta_0$ in high dimension.

For model (3), we propose a class of new estimators and single out a typical one, linearized maximum rank correlation estimator, which possesses superior properties in both theory and computation. Similar to maximum rank correlation and monotone rank estimation, our proposed estimator directly exploits the monotonic relationship between $Y$ and the linear index, thus it is robust to outliers in $Y$. The objective function of linearized maximum rank correlation is smooth and concave, making it computationally efficient, especially for large $p$. A more general class of estimators is also studied. The proposed methods can be easily extended to handle censored data and high-dimensional data. With or without censoring, the estimation of the unknown link function is avoided. An efficient plug-in method and a random weighting resampling scheme are used for variance estimation.

## 2. Methodologies and Main results

### 2.1. *Linearized maximum rank correlation and its closed-form solution*

For model (3), we restrict $\mathrm{cov}(\beta_0^{\mathrm{T}}X) = \beta_0^{\mathrm{T}}\Sigma\beta_0 = 1$ for identifiability, where $\Sigma$ is the covariance matrix of $X$. The observations $(Y_i, X_i), i = 1, \ldots, n$, are independent and identically distributed copies of $(Y, X)$. Our proposed estimator for $\beta_0$ is defined as

$$\hat{\beta}_n^* = \mathop{\arg\max}_{\beta \in \mathcal{E}(\Sigma)} \left\{ L_n(\beta) = \frac{1}{n(n-1)} \sum_{i \neq j}^{n} I(Y_i < Y_j)(X_j - X_i)^{\mathrm{T}}\beta \right\}, \tag{4}$$

where $\mathcal{E}(\Sigma) = \{\beta \in \mathbb{R}^p : \beta^{\mathrm{T}}\Sigma\beta = 1\}$ is a $p$-dimensional ellipsoid related to $\Sigma$, which coincides with the identifiability condition. When $(X_j - X_i)^{\mathrm{T}}\beta$ in (4) is replaced by $I\{(X_j - X_i)^{\mathrm{T}}\beta > 0\}$, it becomes maximum rank correlation estimator. The objective function in (4) is linear and smooth in $\beta$. We refer to the proposed estimator in (4) as linearized maximum rank correlation. An advantage of using $(X_j - X_i)^{\mathrm{T}}\beta$ in (4) instead of the indicator function is that $\hat{\beta}_n^*$ has a closed form expression derived below.

When $\Sigma$ is known, we solve (4) using the method of Lagrange multipliers:

$$\nu^\star \equiv \mathop{\arg\min}_{\nu \in \mathbb{R}} \max_{\beta \in \mathcal{E}(\Sigma)} L_n(\beta) + \nu(\beta^{\mathrm{T}}\Sigma\beta - 1),$$

where $\nu > 0$ is the Lagrange multiplier. Then, any solution pair $(\nu^\star, \hat{\beta}_n^*)$ necessarily satisfies

$$\nabla L_n(\hat{\beta}_n^*) + 2\nu^\star\Sigma\hat{\beta}_n^* = 0 \quad \text{and} \quad \hat{\beta}_n^{*\mathrm{T}}\Sigma\hat{\beta}_n^* - 1 = 0,$$

where $\nabla L_n(\beta) = \sum_{i \neq j}^{n} I(Y_i < Y_j)(X_j - X_i)/\{n(n-1)\}$ is the gradient of $L_n(\beta)$ w.r.t $\beta$. Note that $\nabla L_n(\beta)$ is independent of $\beta$, thus for notational simplicity, we write the U-statistic $\nabla L_n(\beta)$ as $U_n$. Solving the above two equations yields that $\nu^\star = -\hat{\beta}_n^{*\mathrm{T}}U_n/2$ and

$\hat{\beta}_n^* - \Sigma^{-1}U_n/(\hat{\beta}_n^{*\mathrm{T}}U_n) = 0$. It then follows that

$$\hat{\beta}_n^* = +\frac{\Sigma^{-1}U_n}{(U_n{}^{\mathrm{T}}\Sigma^{-1}U_n)^{1/2}}, \tag{5}$$

where the positive sign is determined by the definition that $\hat{\beta}_n^*$ is the maximizer of the optimization problem in (4). The denominator in (5) is a normalization scalar to ensure that $\hat{\beta}_n^*$ satisfies $\hat{\beta}_n^{*\mathrm{T}}\Sigma\hat{\beta}_n^* = 1$. Intuitively, ignoring the denominator and letting $n \to \infty$, the direction of $\hat{\beta}_n^*$ converges to that of $E(\Sigma^{-1}U_n) = E\{I(Y_i < Y_j)\Sigma^{-1}(X_j - X_i)\}$ for $i \neq j$, which can be shown to have the same direction as $\beta_0$ under the *linearity of expectation* assumption on $X$ (Li & Duan, 1989).

For illustration, we consider a simple example when $X$ follows Gaussian distribution with mean $\mu$ and identity covariance matrix $I_p$, which satisfies linearity of expectation in all directions. Then for $i \neq j$, $E\{I(Y_i < Y_j)I_p^{-1}(X_j - X_i)\}$ is a weighted expectation of $X_j - X_i$, i.e.,

$$\begin{aligned} E\{I(Y_i < Y_j)I_p^{-1}(X_j - X_i)\} =& E\big[E\{I(Y_i < Y_j) \mid X_i, X_j\}(X_j - X_i)\big] \\ =& E\big\{W(X_i^{\mathrm{T}}\beta_0, X_j^{\mathrm{T}}\beta_0)(X_j - X_i)\big\}, \end{aligned}$$

where $W(X_i^{\mathrm{T}}\beta_0, X_j^{\mathrm{T}}\beta_0) \equiv \mathrm{pr}\{f(X_i^{\mathrm{T}}\beta_0, \epsilon_i) < f(X_j^{\mathrm{T}}\beta_0, \epsilon_j) \mid X_i, X_j\}$ is non-negative with mean 1/2, depending on $X_i, X_j$ only through $X_i^{\mathrm{T}}\beta_0$ and $X_j^{\mathrm{T}}\beta_0$. Under the standard multivariate normal assumption, for any vector $b$ satisfying $b^{\mathrm{T}}\beta_0 = 0$, $W(X_i^{\mathrm{T}}\beta_0, X_j^{\mathrm{T}}\beta_0)$ and $(X_j - X_i)^{\mathrm{T}}b$ are independent, as

$$b^{\mathrm{T}}E\big\{W(X_i^{\mathrm{T}}\beta_0, X_j^{\mathrm{T}}\beta_0)(X_j - X_i)\big\} = E\big\{W(X_i^{\mathrm{T}}\beta_0, X_j^{\mathrm{T}}\beta_0)\big\}E\{(X_j - X_i)^{\mathrm{T}}b\} = 0.$$

Next, it follows from the definition of $W(X_i^{\mathrm{T}}\beta_0, X_j^{\mathrm{T}}\beta_0)$ and $f$ is increasing in its first argument that $E\big\{W(X_i^{\mathrm{T}}\beta_0, X_j^{\mathrm{T}}\beta_0)(X_j - X_i)^{\mathrm{T}}\beta_0\big\} > 0$. Hence, $E\big\{W(X_i^{\mathrm{T}}\beta_0, X_j^{\mathrm{T}}\beta_0)(X_j - X_i)\big\}$ is nonzero and has the same direction as $\beta_0$.

The multivariate normal distribution implies independence of $X^{\mathrm{T}}\beta_0$ and $X^{\mathrm{T}}b$ for all $b$ satisfying $b^{\mathrm{T}}\beta_0 = 0$, and thus ensures the consistency of $\hat{\beta}_n^*$. A weaker and sufficient condition to guarantee consistency is that $X$ satisfies linearity of expectation in the direction of $\beta_0$.

DEFINITION 1. *(Linearity of Expectation) A $p$-dimensional random vector $X$ is said to satisfy linearity of expectation in the direction $\beta$ if for any direction $b \in \mathbb{R}^p$,*

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

*where $a_b, c_b \in \mathbb{R}$ are some real constants which may depend on the direction $b$ (Li & Duan, 1989; Li, 1991).*

For a random vector $X$ satisfying linearity of expectation with mean $\mu$ and covariance matrix $\Sigma$, if $\beta^{\mathrm{T}}\Sigma\beta \neq 0$, then $c_b = b^{\mathrm{T}}\Sigma\beta/\beta^{\mathrm{T}}\Sigma\beta$ and $a_b = b^{\mathrm{T}}\mu - c_b\beta_0^{\mathrm{T}}\mu$. The proof is given in Lemma S2 in section 2 of the supplementary material. Elliptical distributions, including the multivariate normal distribution and $t$-distribution, satisfy the linearity of expectation uniformly in all directions. The construction and definition of elliptical distributions can be found in Cambanis et al. (1981).

*Remark 1.* Like maximum rank correlation estimation in Han (1987), the objective function in (4) depends on the responses only through their ranks, making it robust to outliers in $Y$. When $X$ satisfies the linearity of expectation, the proposed estimator is shown to be consistent, more efficient and stable than maximum rank correlation estimator, especially for large $p$. But maximum rank correlation estimator is robust to outliers in the covariates and does not need the linearity of expectation for consistency.

*Remark 2.* For model (3) without the monotonicity assumption on $f$, the least squares estimator by Neykov et al. (2016b) consistently estimates the direction of $\beta_0$ (up to a sign) if the constant $E(YX^{\mathrm{T}}\beta_0) \neq 0$. A main advantage of the proposed method over the least squares estimator by Neykov et al. (2016b) is its robustness to outliers in the response variable.

*Remark 3.* Our linearized maximum rank correlation estimation shares similar assumption on the distribution of $X$ with the well-known sliced inverse regression (SIR) in Li (1991). Specifically, Li (1991) considered a general multiple-index model $Y = f(X^{\mathrm{T}}\beta_1, \ldots, X^{\mathrm{T}}\beta_K, \epsilon)$, where $\beta_k$'s $(k = 1, \ldots, K)$ are unknown parameter vectors, $f : \mathbb{R}^{K+1} \to \mathbb{R}$ is an unknown function, and the error $\epsilon$ is independent of $X$. It reduces to model (3) when $K = 1$ and $f$ is non-constant increasing in its first argument. Linearized maximum rank correlation estimation and sliced inverse regression both focus on the estimation of the direction(s) of or the linear space spanned by the unknown vector(s) $\beta_k$'s, $k = 1, \ldots, K$, without the estimation of $f$. Different from the sliced inverse regression, linearized maximum rank correlation estimation has a closed-form solution and no tuning parameter is involved, making it easy to implement in practice and relatively easy to analyze in theory. Both methods can be extended to handle censored data (Chen et al., 1999) and high-dimensional data with some penalty term (Neykov et al., 2016a; Lin et al., 2018). Numerical comparisons are given in Section 5.

*Remark 4.* Similar to the sliced inverse regression (Li, 1991), our linearized maximum rank correlation estimation can be inefficient and theoretically fails for symmetric single index models. For example, if $Y = f(X^{\mathrm{T}}\beta_0) + \epsilon$ for some symmetric function $f$, and $X^{\mathrm{T}}\beta_0$ is also symmetric about 0, $E(X \mid Y) = 0$ and $E\{I(Y_1 < Y_2)(X_2 - X_1)^{\mathrm{T}}\beta_0\} = 0$, thus the resulting estimator is a poor estimator of $\beta_0$. Cook & Weisberg (1991) proposed a remedy to this problem for sliced inverse regression by exploring higher conditional moments of $X$ given $Y$, called sliced average variance estimate. How to apply the idea of sliced average variance estimate to our method for symmetric single index models is a very interesting research problem. We leave space here for future research.

When $\Sigma$ is unknown, our proposed estimator is defined as

$$\hat{\beta}_n \equiv \operatorname*{arg\,max}_{\beta \in \mathcal{E}(\hat{\Sigma})} \frac{1}{n(n-1)} \sum_{i \neq j}^{n} I(Y_i < Y_j)(X_j - X_i)^{\mathrm{T}} \beta, \tag{6}$$

where $\hat{\Sigma}$ is a consistent estimator of $\Sigma$. A closed-form solution of (6) can be obtained:

$$\hat{\beta}_n = ({U_n}^{\mathrm{T}} \hat{\Sigma}^{-1} U_n)^{-1/2} \hat{\Sigma}^{-1} U_n, \tag{7}$$

which is computationally straightforward and efficient.

## 2.2. *Asymptotic properties*

Let $Z = (Y, X^{\mathrm{T}})^{\mathrm{T}}$ be a random vector with distribution $P$ on the set $S \subset \mathbb{R} \otimes \mathbb{R}^p$, and write the observations as $\mathbb{D} = \{Z_i, i = 1, \ldots, n\}$. Let $\epsilon_i, i = 1, \ldots, n$, be the error terms. Let $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the $\ell_1$ and the $\ell_2$ norms of a vector in Euclidean space. Define the $\ell_2$-norm of a matrix $A$ as $\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2$, and $\|\cdot\|$ denotes the matrix norm: $\|(a_{ij})\| = (\sum_{i,j} a_{ij}^2)^{1/2}$. Define $L(\beta) = E\{L_n(\beta)\}$ and $U = E(U_n)$.

Some basic properties of $L(\beta)$ and $L_n(\beta)$ are first introduced. First, $L(\beta)$ and $L_n(\beta)$ are unbounded in $\mathbb{R}^p$, and achieve their maximum at $\beta = \infty$; namely, without any bounded constraint on $\beta$, maximizing $L_n(\beta)$ results in a trivial solution $\hat{\beta}_n = \infty$. Second, for any constant $c \in \mathbb{R}$ and $n \geq 1$, $L_n(c\beta) = cL_n(\beta)$, i.e., $L_n(\beta)$ is proportional to the norm of those $\beta$ with the same direction in $\mathbb{R}^p$. Thus, the maximum of $L_n(\beta)$ will be achieved on the boundary of the constraint region, suggesting that the optimization can be conducted on a manifold. Moreover, under model (3), $Y$ depends on $X$ only through $X^{\mathrm{T}}\beta_0$. Hence, to search over all directions in $\mathbb{R}^p$ uniformly, $L_n(\beta)$ is maximized subject to the constraint $\operatorname{cov}(X^{\mathrm{T}}\beta) = 1$. Third, by letting $\alpha = \Sigma^{1/2}\beta$ and $\tilde{X}_i = \Sigma^{-1/2} X_i$, $i = 1, 2, \ldots, n$, problem (4) can be cast into

$$\hat{\beta}_n^* = \Sigma^{-1/2} \operatorname*{arg\,max}_{\alpha \in \mathbb{R}^p} \frac{1}{n(n-1)} \sum_{i \neq j}^{n} I(Y_i < Y_j)(\tilde{X}_j - \tilde{X}_i)^{\mathrm{T}} \alpha, \ \text{ subject to } \ \alpha^{\mathrm{T}} \alpha = 1,$$

which may simplify the numerical computation as the feasible region of $\alpha$ is a unit hypersphere.

The following conditions are needed to establish the asymptotic properties.

(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 $E\{f(X^{\mathrm{T}}\beta_0, \epsilon) \mid X^{\mathrm{T}}\beta_0\}$ has non-zero variance. And $X$ and $\epsilon$ are independent.

(C1) The predictor vector $X$ satisfies linearity of expectation in the direction of $\beta_0$.

(C2) The covariance matrix $\Sigma$ exists and is positive definite.

THEOREM 1. *(Consistency, known $\Sigma$) Assume that model (3) is true, and Conditions (M), (C1)-(C2) hold. Then, $\hat{\beta}_n^*$ defined in (4) is consistent for $\beta_0$.*

*Remark 5.* Without the monotonicity assumption on $f$ in Condition (M), our estimator $\hat{\beta}_n$ can still be consistent for $\beta_0$ up to a sign if $(\Sigma^{-1}U)^{\mathrm{T}}\beta_0 \neq 0$, whose proof is given in the proof of Theorem 1 in section 2 of the supplementary material.

Since the parameter space $\mathcal{E}(\Sigma) = \{\beta \in \mathbb{R}^p : \beta^{\mathrm{T}}\Sigma\beta = 1\}$ is a hyper ellipsoid, to study the asymptotic distribution of $\hat{\beta}_n^*$, we focus on a subvector of $\hat{\beta}_n^*$. Without loss of generality, let $A = (0, I_{p-1})$ be a $(p-1) \times p$ matrix with the first column being zeros and $I_{p-1}$ be an identity matrix of order $p-1$. Let $\theta \equiv A\beta = (\beta_2, \ldots, \beta_p)^{\mathrm{T}} \in \mathbb{R}^{p-1}$, a $(p-1)$-subvector of $\beta$ excluding its first component. Actually, $A$ can be other matrices mapping $\beta$ to any of its $(p-1)$-subvector. Due to the constraint $\beta^{\mathrm{T}}\Sigma\beta = 1$, $\theta \in \Theta \equiv \{A\beta : \beta \in \mathcal{E}(\Sigma)\}$, which is a compact set, as it is the projection of $\mathcal{E}(\Sigma)$ onto a $(p-1)$-dimensional hyperplane perpendicular to $\beta_1$ (the first component of $\beta$). Similarly, we write $\theta_0 = A\beta_0$ and $\hat{\theta}_n^* = A\hat{\beta}_n^*$. To establish the asymptotic distribution of $\hat{\theta}_n^*$, an additional regularity condition is needed.

(C3) The true parameter $\theta_0$ is an interior point of $\Theta$, a compact subset of $\mathbb{R}^{p-1}$.

THEOREM 2. *(Asymptotic Normality, known $\Sigma$) Assume that model (3) is true, and Conditions (M), (C1)-(C3) hold. For each $z = (y, x^{\mathrm{T}})^{\mathrm{T}} \in S$, define $\xi(z) = E\{I(y < Y)(X - x) + I(Y < y)(x - X) - 2U\}$. Then, as $n \to \infty$, $n^{1/2}(\hat{\theta}_n^* - \theta_0) \to N(0, AV\Delta V^{\mathrm{T}}A^{\mathrm{T}})$ in distribution, where $V = (\Sigma^{-1} - \beta_0\beta_0^{\mathrm{T}})(U^{\mathrm{T}}\Sigma^{-1}U)^{-1/2}$ and $\Delta = E\{\xi(Z)\xi(Z)^{\mathrm{T}}\}$.*

It is not hard to show that the matrices $V$ and $\Delta$ can be consistently estimated by the plug-in estimators $\hat{V}_n = (\Sigma^{-1} - \hat{\beta}_n^*\hat{\beta}_n^{*\mathrm{T}})(U_n\Sigma^{-1}U_n)^{-1/2}$ and $\hat{\Delta}_n = \sum_{i=1}^n \hat{\xi}_n(Z_i)\hat{\xi}_n(Z_i)^{\mathrm{T}}/n$ respectively, where $\hat{\xi}_n(z) = \sum_{j=1}^n \{I(y < Y_j)(X_j - x) + I(Y_j < y)(x - X_j) - 2U_n\}/n$.

When $\Sigma$ is unknown, the estimation of a covariance matrix and its inverse has been studied by many authors (Muirhead (2009); Yuan (2010); Cai et al. (2010); etc). Theorem 3 establishes the consistency of $\hat{\beta}_n$ defined in (6).

THEOREM 3. *(Consistency, unknown $\Sigma$) Assume that model (3) is true, and Conditions (M), (C1)-(C2) hold. If $\hat{\Sigma}$ is consistent for $\Sigma$, then $\hat{\beta}_n$ is consistent for $\beta_0$.*

The sample covariance matrix $\hat{\Sigma}_S = \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^{\mathrm{T}}/(n-1)$ is a consistent estimator of $\Sigma$ in low-dimensional models, where $\bar{X} = \sum_{i=1}^n X_i/n$. Similar to Theorem 2, we write $\hat{\theta}_n = A\hat{\beta}_n$ and present its asymptotic distribution in the next theorem.

THEOREM 4. *(Asymptotic distribution, unknown $\Sigma$) Assume that model (3) is true, and Conditions (M) and (C1)-(C3) hold.*

*(i) If $\|\hat{\Sigma} - \Sigma\|_2 = o_p(n^{-1/2})$, then as $n \to \infty$, $n^{1/2}(\hat{\theta}_n - \theta_0) \to N(0, AV\Delta V^{\mathrm{T}}A^{\mathrm{T}})$ in distribution, which is the limiting distribution in Theorem 2;*

*(ii) If $\hat{\Sigma} = \hat{\Sigma}_S = \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^{\mathrm{T}}/(n-1)$ and $E\|X\|_2^4 < +\infty$, then as $n \to \infty$, $n^{1/2}(\hat{\theta}_n - \theta_0) \to N(0, ABA^{\mathrm{T}})$ in distribution, where $B = E\{V\xi(Z) + H\psi(Z)\}\{V\xi(Z) +$*

*$H\psi(Z)\}^{\mathrm{T}}$, $\psi(Z) = \{(X-\mu)(X-\mu)^{\mathrm{T}} - \Sigma\}\beta_0$, $H = \beta_0\beta_0^{\top}/2 - \Sigma^{-1}$, $V$ and $\xi(\cdot)$ are defined in Theorem 2.*

Theorem 4 part (i) states that, if $\|\hat{\Sigma} - \Sigma\|_2 = o_p(n^{-1/2})$, then $\hat{\theta}_n = \hat{\theta}_n^* + o_p(n^{-1/2})$; thus, $\hat{\theta}_n$ is asymptotically as efficient as $\hat{\theta}_n^*$ when $\Sigma$ is known. However, if $\hat{\Sigma} = \hat{\Sigma}_S$, the asymptotic distribution in Theorem 4 part (ii) is generally different from that in part (i). Furthermore, if $\hat{\Sigma}$ satisfies $\|\hat{\Sigma} - \Sigma\|_2 = O_p(n^{-1/2})$, it can be shown that $n^{1/2}(\hat{\theta}_n - \theta_0) = O_p(1)$.

*Remark 6.* A surprising and seemingly paradoxical observation in our numerical studies is that, when $\Sigma$ is unknown and a plug-in estimator $\hat{\Sigma}_S$ is used, the resulting estimator is generally more efficient (with smaller variance) than the counterpart when $\Sigma$ is assumed known. Similar observations were reported by Henmi & Eguchi (2004); Henmi et al. (2007) and Tarpey et al. (2014). For more insights, we consider the case when $p = 1$ and focus on the numerators of $\hat{\beta}_n^*$ and $\hat{\beta}_n$. When $\Sigma$ is known, the numerator of $\hat{\beta}_n^*$ in (5), $\Sigma^{-1}U_n$ converges to $c\beta_0$ in probability for some constant $c > 0$, where $\Sigma$ is a positive scalar when $p = 1$ and $U_n = \sum_{i \neq j}^n I(Y_i < Y_j)(X_j - X_i)/\{n(n-1)\}$ can be viewed as a new measure of the correlation between $Y$ and $X$. It is not hard to check that $U_n$ is positively (negatively) correlated with $\hat{\Sigma}_S$ when $\beta_0 > 0$ ($\beta_0 < 0$). When $\Sigma$ is unknown, the numerator of $\hat{\beta}_n$ can be written as $(\Sigma/\hat{\Sigma}_S)(U_n/\Sigma)$, which is also consistent for $c\beta_0$ under the regularity conditions. Without loss of generality, we consider $\beta_0 > 0$. Given the data, if $U_n/\Sigma > c\beta_0$, then $\Sigma/\hat{\Sigma}_S$ is more likely to be less than 1 due to the positive correlation between $U_n$ and $\hat{\Sigma}_S$; when $\Sigma^{-1}U_n$ over-estimates (under-estimates) $c\beta_0$, $\Sigma/\hat{\Sigma}_S$ can help pull down (up) the term $(\Sigma/\hat{\Sigma}_S)(U_n/\Sigma)$ towards the target $c\beta_0$. Thus, it produces estimates with smaller variance. Similarly, the denominator of $\hat{\beta}_n$ can be argued to be more stable than that of $\hat{\beta}_n^*$. The above discussions offer some insights into why $\hat{\beta}_n$ can be more efficient than $\hat{\beta}_n^*$ for the univariate case. Nonetheless, a rigorous proof for a general setting has yet to be found.

### 2.3. *A general class of estimators*

The identity function we adopt in (4) is not necessarily the unique choice. A general class of objective functions can be considered:

$$L_n^g(\beta) \equiv \frac{1}{n(n-1)} \sum_{i \neq j}^n I(Y_i < Y_j) g\{(X_j - X_i)^{\mathrm{T}}\beta\}, \tag{8}$$

where $g(\cdot)$ is a known non-constant increasing function. When $\Sigma$ is known, the estimator is defined as $\hat{\beta}_n^g = \arg\max_{\beta \in \mathcal{E}(\Sigma)} L_n^g(\beta)$. With a slight abuse of notation, when $\Sigma$ is unknown, we plug in a consistent estimator $\hat{\Sigma}$ in $\mathcal{E}(\Sigma)$ and still denote the resulting estimator as $\hat{\beta}_n^g$. When $g(a) = a$ in (8), it is linearized maximum rank correlation estimation in Section 2.1; when $g(a) = I(a > 0)$, it is maximum rank correlation. One may also consider $g(a) = -([-a]_+)^\gamma$, $g(a) = ([a]_+)^\gamma$ for any real number $\gamma > 0$ or $g(a) = \log(1 + [a]_+)$, where $[a]_+ = \max\{0, a\}$ is the rectified linear unit (ReLU). Such choices of $g$ would be valid to produce consistent estimate for $\beta_0$ under suitable conditions, albeit the objective function is no longer concave for some cases such as $g(a) = I(a > 0)$, $g(a) = ([a]_+)^\gamma$ or $g(a) = -([-a]_+)^\gamma$ when $\gamma < 1$.

For such choices of $g$, asymptotic theories analogous to Theorems 1-4 can be established. For technical convenience, we reparameterize $\beta$ according to the identifiability constraint. Without loss of generality, let $\theta = (\beta_2, \dots, \beta_p)^\mathrm{T} \in \mathbb{R}^{p-1}$ be the $(p-1)$-subvector of $\beta$. Under the constraint $\beta^\mathrm{T}\Sigma\beta = 1$, the first component of $\beta$ can be represented as $\beta_1(\theta, \Sigma) = -\Sigma_{12}\theta/\Sigma_{11} + (s/\Sigma_{11})\left\{(\Sigma_{12}\theta)^2 + \Sigma_{11}(1 - \theta^\mathrm{T}\Sigma_{22}\theta)\right\}^{1/2}$, where $s \in \{1, -1\}$, $\Sigma_{11}$ is a scalar and $\Sigma_{22}$ is a $(p-1) \times (p-1)$ submatrix of $\Sigma$ such that $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. Then, $\beta = \beta(\theta, s, \Sigma)$ is uniquely determined by $(\theta, s, \Sigma)$ for $s \in \{1, -1\}$ and $\theta \in \Theta = \{\theta \in \mathbb{R}^{p-1} : (\Sigma_{12}\theta)^2 + \Sigma_{11}(1 - \theta^\mathrm{T}\Sigma_{22}\theta) \geq 0\}$. Similarly, we write $\beta_0 = \beta_0(\theta_0, s_0, \Sigma)$ and $\hat\beta^g_n = \hat\beta^g_n(\hat\theta^g_n, \hat s^g, \Sigma)$. Without loss of generality, we fix $s_0 = 1$ and write $\beta_0 = \beta_0(\theta_0, \Sigma)$ and $\beta = \beta(\theta, \Sigma)$. With consistent $\hat s^g$, we can write $\hat\beta^g_n(\hat\theta^g_n, \Sigma) = (\hat\beta^g_1, \hat\theta^{g\mathrm{T}}_n)^\mathrm{T}$. Some conditions are assumed to prove the asymptotic distribution of $\hat\theta^g_n$.

(G1) (i) There exists a $\beta \in \mathcal{E}(\Sigma)$ such that $E\|g\{(X_1 - X_2)^\mathrm{T}\beta\}\|_2 < \infty$, and the function $g(t)$ is non-constant increasing in $t$ on the support of $(X_1 - X_2)^\mathrm{T}\beta$. (ii) For any $\beta \in \mathcal{E}(\Sigma)$ and any $t \in \mathbb{R}$, the class of sets $\{(x_1, x_2, t) \in \mathbb{R}^{2p+1} : g\{(x_1 - x_2)^\mathrm{T}\beta\} > t, x_1, x_2 \in \mathbb{R}^p\}$ is a VC class.

(G2) Let $a(t, c) \equiv E\{I(Y_i < Y_j) \mid X_j - X_i = t, t^\mathrm{T}\Sigma^{-1}t = c\}$. For each $c > 0$ and any $t_1, t_2$ satisfying $t_i^\mathrm{T}\Sigma^{-1}t_i = c$, $i = 1, 2$, assume that $a(t_1, c) \geq a(t_2, c)$ if and only if $t_1^\mathrm{T}\beta_0 \geq t_2^\mathrm{T}\beta_0$. Moreover, there exist some $t_1, t_2$ such that $t_1^\mathrm{T}\beta_0 \geq t_2^\mathrm{T}\beta_0$ and $a(t_1, c) > a(t_2, c)$ for each $c > 0$.

(G3) For each $z \in S$ and each $\theta \in \Theta$, define $\tau_g(z, \theta) = E[I(y < Y)g\{(X - x)^\mathrm{T}\beta(\theta, \Sigma)\} + I(Y < y)g\{(x - X)^\mathrm{T}\beta(\theta, \Sigma)\}]$. Let $\nabla_m$ denote the $m$-th partial derivative operator with respect to $\theta$.

(i) For each $z \in S$, the function $\tau_g(z, \cdot)$ is twice differentiable in a neighborhood of $\theta_0$, and $\|\nabla_2\tau_g(z, \theta) - \nabla_2\tau_g(z, \theta_0)\| \leqslant M_g(z)\|\theta\|_2$, where $M_g(z)$ is an integrable function of $z$.

(ii) $E\|\nabla_1\tau_g(Z, \theta_0)\|_2^2 < \infty$ and $E\|\nabla_2\tau_g(Z, \theta_0)\|_2 < \infty$.

(iii) The matrix $E\nabla_2\tau_g(Z, \theta_0)$ is non-singular.

Condition (G1) part (i) is a moment condition for $g(\cdot)$, and part (ii) is a technical condition to prove the uniform convergence. Condition (G3) is a regular condition to ensure a Taylor expansion of the conditional expectation of the objective function.

THEOREM 5. *(Consistency) Assume model (3) is true, and Conditions (M), (C2), (G1) part (i) and (G2) hold. If $X$ is elliptically distributed and (i) $\Sigma$ is known or (ii) $\Sigma$ is unknown and $\hat\Sigma$ is consistent for $\Sigma$, then, $\hat\beta^g_n$ is consistent for $\beta_0$.*

*Remark 7.* Theorem 5 presents the global consistency of $\hat\beta^g_n$ under the elliptical distributed assumption and other suitable conditions. In fact, under Conditions (C1), (C2), (G1) part (i), together with $g$ is concave, we can prove the local consistency of $\hat\beta^g_n$.

THEOREM 6. *(Asymptotic Normality) Assume that model (3) is true, and Conditions (M), (C2)-(C3), (G1)-(G3) hold and $X$ is elliptically distributed. Define* $\mathrm{Diff}(\theta, \Sigma) = g\{(X_j - X_i)^{\mathrm{T}}\beta(\theta, \Sigma)\} - g\{(X_j - X_i)^{\mathrm{T}}\beta(\theta_0, \Sigma)\}$. *If uniformly over $o_p(1)$ neighborhoods of $\theta_0$, $\|\mathrm{Diff}(\theta, \hat{\Sigma}) - \mathrm{Diff}(\theta, \Sigma)\|_2 = o_p(n^{-1/2}\|\theta - \theta_0\|_2)$, then $n^{1/2}(\hat{\theta}_n^g - \theta_0) \to N(0, (V^g)^{-1}\Delta^g(V^g)^{-1})$ in distribution, where $\Delta^g = E[\nabla_1\tau_g(\cdot, \theta_0)\{\nabla_1\tau_g(\cdot, \theta_0)\}^{\mathrm{T}}]$ and* $V^g = (1/2)E\{\nabla_2\tau_g(\cdot, \theta_0)\}$.

Two numerical algorithms are given in Section 1 of the supplementary material to compute $\hat{\beta}_n^g$, depending on the differentiability of $g$.

## 3. EXTENSION TO CENSORED DATA

For model (3) under random censoring, the observations are $(v_i, X_i, d_i)$, $i = 1, \ldots, n$, independent and identically distributed copies of $(v, X, d)$, where $v = \min(Y, C)$, $C$ is the censoring variable and $d = I(Y < C)$ is the censoring indicator. We focus on the case that $\Sigma$ is unknown. Motivated by the partial rank estimator of Khan & Tamer (2007), our proposed linearized partial rank estimator of $\beta_0$ is defined as

$$\hat{\beta}_n^c = \operatorname*{arg\,max}_{\beta \in \mathcal{E}(\hat{\Sigma})} L_n^c(\beta), \quad \text{where} \quad L_n^c(\beta) = \frac{1}{n(n-1)} \sum_{i \neq j}^{n} d_i I(v_i < v_j)(X_j - X_i)^{\mathrm{T}}\beta, \tag{9}$$

and $\mathcal{E}(\hat{\Sigma})$ is given in Section 2.1. Similar to the complete data case, a closed-form solution to (9) is $\hat{\beta}_n^c = (U_n^{c\mathrm{T}}\hat{\Sigma}^{-1}U_n^c)^{-1/2}\hat{\Sigma}^{-1}U_n^c$, where $U_n^c = \sum_{i \neq j}^n d_i I(v_i < v_j)(X_j - X_i)/\{n(n-1)\}$ is the gradient of $L_n^c(\beta)$ w.r.t $\beta$. Additional assumptions are needed.

(A1) The error $\epsilon$ is independent of $(C, X)$, and $C$ is independent of $X$.

(A2) Let $S_X$ denote the support of $X$, and $\mathcal{X}_{uc} = \{x \in S_X : \mathrm{pr}(d = 1 \mid X = x) > 0\}$, then $\mathrm{pr}(X \in \mathcal{X}_{uc}) > 0$.

(A3) The set $\mathcal{X}_{uc}$ is not contained in any proper linear subspace of $\mathbb{R}^p$.

Condition (A2) requires that the probability of censoring is not 1 for all $x \in S_X$. Condition (A3) is slightly stronger than a full rank assumption and it is often needed in the censored case. The independence of $C$ and $X$ in Condition (A1) is imposed for technical convenience.

THEOREM 7. *(Consistency) Assume that model (3) is true, and Conditions (M), (C1)-(C2), (A1)-(A3) hold and $\hat{\Sigma}$ is consistent for $\Sigma$, then $\hat{\beta}_n^c$ is consistent for $\beta_0$.*

For the general transformation model in Khan & Tamer (2007), a special case of model (3), the consistency still holds when condition (A1) is weakened to allow for more general forms of censoring. Similar to section 2.2, we next study the limiting distribution of $\hat{\theta}_n^c = A\hat{\beta}_n^c$, a $(p-1)$-subvector of $\hat{\beta}_n^c$, where $A = (0, I_{p-1})$ is a $(p-1) \times p$ matrix with its first column being zeros and $I_{p-1}$ is an identity matrix of order $p - 1$. Let $Z_i = (d_i, v_i, X_i)$ and $U^c = E(U_n^c)$. For each

$z = (d, v, x)$, define $\xi_c(z) = E\{d_i I(v_i < v)(x - X_i) + dI(v < v_i)(X_i - x)^{\mathrm{T}} - 2U^c\}$. Denote $V^c = (\Sigma^{-1} - \beta_0\beta_0^{\mathrm{T}})(U^{c\mathrm{T}}\Sigma^{-1}U^c)^{-1/2}$ and $\Delta^c = E\{\xi_c(Z)\xi_c(Z)^{\mathrm{T}}\}$.

THEOREM 8. *Assume that model (3) is true, and Conditions (M), (C1)-(C2) and (A1)-(A3) hold.*
*(i) If $\|\hat{\Sigma} - \Sigma\|_2 = o_p(n^{-1/2})$, then $n^{1/2}(\hat{\theta}_n^c - \theta_0) \rightarrow N(0, AV^c\Delta^c(V^c)^{\mathrm{T}}A^{\mathrm{T}})$ in distribution as $n \to \infty$.*
*(ii) If $\hat{\Sigma} = \hat{\Sigma}_S = \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^{\mathrm{T}}/(n-1)$ and $E\|X\|_2^4 < +\infty$, then $n^{1/2}(\hat{\theta}_n^c - \theta_0) \rightarrow N(0, AB^cA^{\mathrm{T}})$ in distribution as $n \to \infty$, where $B^c = E\{V^c\xi_c(Z) + H\psi(Z)\}\{V^c\xi_c(Z) + H\psi(Z)\}^{\mathrm{T}}$, $\psi(Z) = \{(X - \mu)(X - \mu)^{\mathrm{T}} - \Sigma\}\beta_0$ and $H = (\beta_0\beta_0^{\mathrm{T}}/2 - \Sigma^{-1})$ .*

For a general estimator $\hat{\Sigma}$ satisfying $\|\hat{\Sigma} - \Sigma\|_2 = O_p(n^{-1/2})$, we can similarly show that $n^{1/2}(\hat{\theta}_n^c - \theta_0) = O_p(1)$. Although linearized partial rank estimator in (9) is designed for right censoring, it can be easily modified to accommodate left-censored data or doubly-censored data.

## 4. EXTENSION TO HIGH DIMENSIONS

In this section, we extend our methods to handle variable selection and parameter estimation for model (3) with high-dimensional predictors. For simplicity, we write $X_i$ as the $i$-th observation of $X$, the $p_n$-vector of predictors. Let $\mathcal{A} = \{j : \beta_{0j} \neq 0\}$ be the true active set, $\mathcal{A}^c = \{j : \beta_{0j} = 0\}$, $s_n$ be the cardinality of $\mathcal{A}_T$, $\beta_{\mathcal{A}}$ be the subvector of $\beta$ indexed by $\mathcal{A}$ and let $d_n = \inf_{j\in\mathcal{A}} |\beta_{0j}|$. Both $p_n$ and $s_n$ may vary with $n$. We consider the following objective function with lasso penalty:

$$PL_n(\beta) \equiv \frac{1}{n(n-1)} \sum_{i \neq j}^{n} I(Y_i < Y_j)(X_i - X_j)^{\mathrm{T}}\beta + \lambda_n\|\beta\|_1, \tag{10}$$

where $\lambda_n$ is the regularization parameter. With a slight abuse of notation, we define the penalized linearized maximum rank correlation estimator as $\hat{\beta}_n \equiv \arg\min_{\beta\in\mathcal{E}(\Sigma)} PL_n(\beta)$.

THEOREM 9. *(Oracle inequality) Assume that model (3) is true, Conditions (C1)-(C2) and Assumptions (M*), (D1)-(D2) given in the supplementary material hold and $\Sigma$ is known. Let $a_n$ be a sequence of positive numbers diverging to $\infty$ at an arbitrarily slow rate. If $\lambda_n = \{a_n \log(n)\log(p_n)/n\}^{1/2}$ and $a_n s_n^2 \log(n)\log(p_n)/n \to 0$ as $n \to \infty$, then with probability at least $1 - 4\exp(-a_n) - (2 + c_1)/p_n$,*

$$\|\hat{\beta}_n - \beta_0\|_2 \leq c_2 a_n^{1/2}\left\{\frac{s_n \log(n)\log(p_n)}{n}\right\}^{1/2}, \quad \|\hat{\beta}_n - \beta_0\|_1 \leq c_2 a_n^{1/2}\left\{\frac{s_n^2 \log(n)\log(p_n)}{n}\right\}^{1/2},$$

*where $c_1, c_2 > 0$ are some universal constants, $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the $\ell_1$ and the $\ell_2$ norm of a vector respectively.*

Theorem 9 implies that $a_n s_n^2 \log(n)\log(p_n) = o(n)$, i.e., our method can handle high-dimensional case when $\Sigma$ is known. Nevertheless, it is nontrivial to prove the oracle inequality when $\Sigma$ is unknown. We leave space here for future research. A proximal (stochastic) gradient decent algorithm is given to solve the penalized linearized maximum rank correlation estimation in Algorithm S3 in Section 1 of the supplementary material.

## 5. NUMERICAL STUDIES

### 5.1. *Complete data*

The covariate $X$ are generated from a multivariate normal distribution with mean 0 and covariance matrix $\Sigma = (\rho_{ij})$ with $\rho_{ij} = \rho_0^{|i-j|}$, $\rho_0 = 0.3$ or $0.8$. Set $p = 5$ and $\beta_0 = (1, 1, 0, 0, -1)^{\mathrm{T}}$, $p = 15$ and $\beta_0 = (1, 1, 0, 0, -1, 1, 1, 0, 0, -1, 1, 1, 0, 0, -1)^{\mathrm{T}}$, or $p = 30$ and $\beta_0 = (1, 1, 0, 0, -1, 1, 1, 0, 0, -1, 1, 1, 0, 0, -1, 1, 1, 0, 0, -1, 1, 1, 0, 0, -1, 1, 1, 0, 0, -1)^{\mathrm{T}}$. Set $n = 100$, 200 or 300. Four models are considered: M1: $Y = X^{\mathrm{T}}\beta_0 + \epsilon$; M2: $Y = \exp\{X^{\mathrm{T}}\beta_0 + \sin(X^{\mathrm{T}}\beta_0) - \epsilon\} - \epsilon$; M3: $Y = \exp(X^{\mathrm{T}}\beta_0) + 0.5X^{\mathrm{T}}\beta_0 \times |\epsilon| + 2\cos(\epsilon)$; M4: $Y = I[\{1 + \exp(-X^{\mathrm{T}}\beta_0 - \epsilon)\}^{-1} > 0.5]$. In M2 and M3, the model assumptions of maximum rank correlation estimation and monotone rank estimation are not satisfied, as the link function is not increasing in $\epsilon$; and $Y$ is binary in M4. Two error distributions are tried: Chi-square distribution with 1 degree of freedom (denoted by $\chi^2(1)$) and Poisson distribution with parameter 1 (denoted by Pois$(1)$). For each setting, $a = \beta_0^{\mathrm{T}}\Sigma\beta_0$ can be computed. We compute the proposed linearized maximum rank correlation estimator (LMRC), the MRC estimator, the monotone rank estimator (MRE) by Cavanagh & Sherman (1998) with $M(\cdot) = R_n(\cdot)$, and the least squares estimator (LSE) in Neykov et al. (2016b) and the sliced inverse regression by Li (1991) for comparison. For all methods, we scale the resulting estimators to satisfy $\beta^{\mathrm{T}}\Sigma\beta = a$ in each setting. With a known $\Sigma$, we refer our method as LMRC*; otherwise, we refer it as LMRC when $\Sigma$ is unknown and the sample covariance matrix $\hat{\Sigma}_S$ is used to estimate $\Sigma$.

The results in the tables are based on 500 replications. We report the averaged absolute bias (BIAS), the averaged standard error (SE), the averaged empirical coverage probability (CP) of 95% confidence intervals, and the mean $\ell_1$ and $\ell_2$ distances. To be precise, let $\{\hat{\beta}^{(r)}\}_{r=1}^{500}$ denote the estimates. The averaged absolute bias is the averaged absolute value of $\{(\sum_{r=1}^{R}\hat{\beta}^{(r)}/R) - \beta_0\}$ over its components; the standard error is the averaged standard error of $\{\hat{\beta}^{(r)}\}_{r=1}^{500}$ over components; the coverage probability of 95% confidence intervals is the averaged empirical coverage probability over components; the mean $\ell_1$ distance is the mean of $\{\|\hat{\beta}^{(r)} - \beta_0\|_1\}_{r=1}^{500}$ and the mean $\ell_2$ distance is the mean of $\{\|\hat{\beta}^{(r)} - \beta_0\|_2\}_{r=1}^{500}$. The estimated standard errors are computed by the bootstrap method with resampling size 200 for maximum rank correlation estimator in Han (1987) and monotone rank estimator in Cavanagh & Sherman (1998).

Table 1: Summary statistics with dimension $p = 5$ and correlation $\rho_0 = 0.3$.

| Model | Error | Method | $n = 100$ | | | | | $n = 200$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | BIAS | SE | CP | $\ell_1$ | $\ell_2$ | BIAS | SE | CP | $\ell_1$ | $\ell_2$ |
| M1 | $\chi^2(1)$ | LMRC* | 0.021 | 0.227 | 0.939 | 0.910 | 0.264 | 0.012 | 0.159 | 0.953 | 0.638 | 0.128 |
| | | LMRC | 0.007 | 0.140 | 0.952 | 0.547 | 0.098 | 0.005 | 0.099 | 0.950 | 0.389 | 0.049 |
| | | SIR | 0.018 | 0.199 | - | 0.775 | 0.201 | 0.007 | 0.139 | - | 0.566 | 0.105 |
| | | LSE | 0.014 | 0.181 | - | 0.718 | 0.167 | 0.008 | 0.126 | - | 0.497 | 0.081 |
| | | MRC | 0.083 | 0.312 | 0.806 | 1.132 | 0.555 | 0.060 | 0.267 | 0.758 | 0.899 | 0.416 |
| | | MRE | 0.081 | 0.324 | 0.891 | 1.197 | 0.588 | 0.067 | 0.283 | 0.884 | 0.979 | 0.458 |
| | Pois(1) | LMRC* | 0.018 | 0.209 | 0.944 | 0.842 | 0.221 | 0.011 | 0.154 | 0.947 | 0.620 | 0.120 |
| | | LMRC | 0.008 | 0.128 | 0.940 | 0.513 | 0.082 | 0.004 | 0.086 | 0.939 | 0.340 | 0.037 |
| | | SIR | 0.013 | 0.156 | - | 0.625 | 0.124 | 0.009 | 0.110 | - | 0.440 | 0.062 |
| | | LSE | 0.010 | 0.142 | - | 0.568 | 0.102 | 0.006 | 0.101 | - | 0.401 | 0.051 |
| | | MRC | 0.081 | 0.313 | 0.808 | 1.123 | 0.554 | 0.061 | 0.261 | 0.783 | 0.851 | 0.402 |
| | | MRE | 0.078 | 0.302 | 0.895 | 1.105 | 0.513 | 0.059 | 0.263 | 0.892 | 0.887 | 0.395 |
| M2 | $\chi^2(1)$ | LMRC* | 0.021 | 0.228 | 0.955 | 0.895 | 0.264 | 0.014 | 0.170 | 0.953 | 0.682 | 0.147 |
| | | LMRC | 0.004 | 0.143 | 0.942 | 0.562 | 0.103 | 0.003 | 0.098 | 0.950 | 0.386 | 0.048 |
| | | SIR | 0.012 | 0.173 | - | 0.679 | 0.151 | 0.005 | 0.115 | - | 0.451 | 0.067 |
| | | LSE | 0.239 | 0.704 | - | 2.969 | 2.887 | 0.245 | 0.715 | - | 3.002 | 2.984 |
| | | MRC | 0.091 | 0.334 | 0.807 | 1.236 | 0.625 | 0.054 | 0.263 | 0.800 | 0.879 | 0.399 |
| | | MRE | 0.089 | 0.332 | 0.886 | 1.243 | 0.626 | 0.055 | 0.264 | 0.890 | 0.917 | 0.398 |
| | Pois(1) | LMRC* | 0.024 | 0.215 | 0.958 | 0.859 | 0.236 | 0.011 | 0.161 | 0.947 | 0.643 | 0.131 |
| | | LMRC | 0.006 | 0.125 | 0.954 | 0.498 | 0.078 | 0.007 | 0.088 | 0.941 | 0.352 | 0.039 |
| | | SIR | 0.007 | 0.122 | - | 0.483 | 0.075 | 0.006 | 0.086 | - | 0.344 | 0.037 |
| | | LSE | 0.081 | 0.431 | - | 1.718 | 0.976 | 0.055 | 0.346 | - | 1.352 | 0.618 |
| | | MRC | 0.086 | 0.324 | 0.810 | 1.164 | 0.599 | 0.059 | 0.260 | 0.787 | 0.852 | 0.398 |
| | | MRE | 0.075 | 0.308 | 0.894 | 1.118 | 0.526 | 0.061 | 0.265 | 0.892 | 0.889 | 0.397 |

Averaged absolute bias (BIAS), averaged standard errors (SE) and averaged coverage probability (CP) of 95% confidence interval over components of the index parameter. Mean $\ell_1$ and $\ell_2$ distances between the estimate and the true parameter; LMRC*, the Linearized maximum rank estimator with a known $\sigma$; LMRC, the Linearized maximum rank estimator when $\sigma$ is unknown and the sample covariance matrix $\hat{\Sigma}_S$ is used to estimate $\Sigma$; SIR, sliced inverse regression by Li (1991); LSE, the MRC estimator the least squares estimator of Neykov et al. (2016b); MRC, the maximum correlation estimator; MRE, the monotone rank estimator by Cavanagh & Sherman (1998).

Tables 1-2 show that our proposed estimators LMRC* and LMRC are generally unbiased for all cases and robust to outliers under M1-M2. And LMRC* and LMRC are more efficient than MRE and MRC for all configurations. Table 2 contains the results for $p = 30$ and $\rho_0 = 0.8$ (high dependence among predictors), indicating that LMRC* and LMRC perform reasonably well. The simulation results under M3-M4 with $p = 5$ or $p = 30$, and those with $p = 15$ are given in Table S1-S3 in Section 3 of the supplementary material.

A sensitivity analysis is conducted to check the robustness of the proposed method when the linearity in expectation assumption of $X$ is violated. Set $p = 5$, $\beta_0 = (1, 1, 0, 0, -1)^{\mathrm{T}}$ and $\epsilon \sim N(0, 0.5^2)$. For models M1-M4, three distributions of $X$ are tried: (i) multivariate Student's $t$ distribution with 2 degrees of freedom and correlation matrix $\Sigma^{corr} = (\rho_{ij}^{corr})$, where $\rho_{ij}^{corr} = 0.3^{|i-j|}$, denoted by $t(2)$; (ii) each component of $X$ are independent Unif$[-2, 10]$, denoted by

$U[-2, 10]$; (iii) a hybrid of Uniform and Bernoulli distributions with the first $4$ components of $X$ being i.i.d from $U[-2, 10]$ and the last element follows Bernoulli distribution with parameter $0.5$. The results are summarized in Table 3. It can be observed that the proposed estimator is fairly robust to the linearity of expectation assumption. We also conduct additional simulations with three models in which the monotonicity condition of $f$ is not satisfied. The results can be found in Table S4 in Section 3 of the supplementary material, which contain supportive evidence that our estimators are still consistent for $\beta_0$ up to a sign when the monotonic assumption of $f$ is violated.

Table 2: Summary statistics with dimension $p = 30$ and $\rho = 0.8$.

| Model | Error | Method | $n = 200$ | | | | | $n = 300$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | BIAS | SE | CP | $\ell_1$ | $\ell_2$ | BIAS | SE | CP | $\ell_1$ | $\ell_2$ |
| M1 | $\chi^2(1)$ | LMRC* | 0.054 | 0.457 | 0.940 | 11.029 | 2.498 | 0.039 | 0.376 | 0.942 | 9.053 | 2.047 |
| | | LMRC | 0.013 | 0.258 | 0.956 | 6.190 | 2.008 | 0.010 | 0.202 | 0.958 | 4.831 | 1.230 |
| | | SIR | 0.019 | 0.375 | - | 8.857 | 4.217 | 0.026 | 0.296 | - | 7.112 | 2.651 |
| | | LSE | 0.021 | 0.269 | - | 6.335 | 2.187 | 0.018 | 0.228 | - | 5.446 | 1.575 |
| | | MRC | 0.303 | 0.605 | 0.445 | 18.127 | 16.304 | 0.244 | 0.565 | 0.461 | 16.331 | 12.972 |
| | | MRE | 0.283 | 0.588 | 0.600 | 17.342 | 14.935 | 0.263 | 0.572 | 0.603 | 16.713 | 13.719 |
| | Pois(1) | LMRC* | 0.058 | 0.442 | 0.942 | 10.657 | 2.420 | 0.040 | 0.365 | 0.945 | 8.806 | 1.994 |
| | | LMRC | 0.014 | 0.223 | 0.956 | 5.309 | 1.496 | 0.007 | 0.176 | 0.951 | 4.211 | 0.937 |
| | | SIR | 0.020 | 0.286 | - | 6.754 | 2.467 | 0.018 | 0.205 | - | 4.917 | 1.280 |
| | | LSE | 0.018 | 0.218 | - | 5.239 | 1.437 | 0.012 | 0.176 | - | 4.197 | 0.939 |
| | | MRC | 0.286 | 0.595 | 0.451 | 17.520 | 15.155 | 0.235 | 0.551 | 0.481 | 15.675 | 12.361 |
| | | MRE | 0.280 | 0.589 | 0.610 | 17.142 | 14.446 | 0.229 | 0.554 | 0.612 | 15.746 | 12.385 |
| M2 | $\chi^2(1)$ | LMRC* | 0.058 | 0.454 | 0.945 | 10.963 | 2.492 | 0.047 | 0.379 | 0.943 | 9.140 | 2.069 |
| | | LMRC | 0.014 | 0.260 | 0.958 | 6.154 | 2.032 | 0.011 | 0.207 | 0.954 | 4.937 | 1.289 |
| | | SIR | 0.033 | 0.385 | - | 9.099 | 4.505 | 0.019 | 0.283 | - | 6.714 | 2.422 |
| | | LSE | 0.417 | 2.204 | - | 54.038 | 151.650 | 0.372 | 2.148 | - | 52.412 | 143.839 |
| | | MRC | 0.290 | 0.602 | 0.451 | 17.765 | 15.566 | 0.244 | 0.554 | 0.479 | 15.785 | 12.601 |
| | | MRE | 0.296 | 0.600 | 0.595 | 17.675 | 15.600 | 0.231 | 0.556 | 0.620 | 15.823 | 12.265 |
| | Pois(1) | LMRC* | 0.056 | 0.445 | 0.939 | 10.736 | 2.4319 | 0.039 | 0.372 | 0.944 | 8.937 | 2.032 |
| | | LMRC | 0.011 | 0.221 | 0.953 | 5.259 | 1.471 | 0.007 | 0.176 | 0.951 | 4.211 | 0.937 |
| | | SIR | 0.016 | 0.287 | - | 6.837 | 2.474 | 0.018 | 0.216 | - | 5.128 | 1.405 |
| | | LSE | 0.361 | 2.162 | - | 53.064 | 145.540 | 0.383 | 2.065 | - | 50.118 | 133.249 |
| | | MRC | 0.279 | 0.586 | 0.457 | 17.272 | 14.600 | 0.246 | 0.559 | 0.465 | 16.016 | 12.748 |
| | | MRE | 0.280 | 0.583 | 0.595 | 16.882 | 14.195 | 0.242 | 0.548 | 0.619 | 15.661 | 12.250 |

### 5.2. *Censored data*

We generate $X$ in the same way as in Section 5.1 and the error $\epsilon$ follows normal distribution with mean $0$ and standard deviation $0.5$. We generate $Y$ from M2 and M4. Set $p = 5$ and $\beta_0 = (1, 1, 0, 0, -1)^{\mathrm{T}}$. Three censoring mechanisms are considered: CV1: $C \sim 2\chi^2(2)$; CV2: $C = 5 \times (X^{\mathrm{T}}\beta_0)^2 - 3 \times \sin(|X^{\mathrm{T}}\beta_0|)$; CV3: $C = 10 \times (X_1^2 \times X_2 + X_3)$, where $X_j$ is the $j$-th component of $X$. In (CV1), $C$ is independent of $X$; $C$ is correlated with $X^{\mathrm{T}}\beta_0$ in (CV2) and intricately depends on $X$ in (CV3). For comparison, we also compute the partial rank estimator (PRE) by Khan & Tamer (2007). Model M4 violates the transformation model assumption of

PRE. Other than the BIAS, SE, CP, $\ell_1$ and $\ell_2$, we report the censoring rate (CR) in Table 4. One can see that LMRC* and LMRC give smaller $\ell_1$ and $\ell_2$ distances compared with the PRE in all scenarios, especially under M4. Our methods are generally more efficient in terms of smaller standard errors and robust to different censoring mechanisms.

Table 3: Summary statistics for differently distributed covariates when dimension $p = 5$ and correlation $\rho_0 = 0.3$.

| Model | Covariate | Method | $n = 100$ | | | | | $n = 200$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | BIAS | SE | CP | $\ell_1$ | $\ell_2$ | BIAS | SE | CP | $\ell_1$ | $\ell_2$ |
| M1 | $t(2)$ | LMRC | 0.063 | 0.328 | 0.885 | 1.329 | 0.702 | 0.045 | 0.275 | 0.903 | 1.089 | 0.574 |
| | | LMRC* | 0.070 | 0.348 | 0.911 | 1.361 | 0.724 | 0.054 | 0.315 | 0.874 | 1.198 | 0.633 |
| | $U[-2, 10]$ | LMRC | 0.005 | 0.059 | 0.950 | 0.234 | 0.019 | 0.001 | 0.040 | 0.953 | 0.159 | 0.009 |
| | | LMRC* | 0.021 | 0.186 | 0.959 | 0.745 | 0.399 | 0.006 | 0.089 | 0.928 | 0.355 | 0.190 |
| | Hybrid | LMRC | 0.015 | 0.076 | 0.972 | 0.309 | 0.050 | 0.011 | 0.053 | 0.957 | 0.215 | 0.025 |
| | | LMRC* | 0.041 | 0.274 | 0.946 | 1.118 | 0.824 | 0.030 | 0.199 | 0.899 | 0.791 | 0.594 |
| M2 | $t(2)$ | LMRC | 0.066 | 0.345 | 0.886 | 1.382 | 0.740 | 0.034 | 0.209 | 0.818 | 0.824 | 0.476 |
| | | LMRC* | 0.065 | 0.337 | 0.916 | 1.320 | 0.702 | 0.025 | 0.135 | 0.925 | 0.534 | 0.303 |
| | $U[-2, 10]$ | LMRC | 0.006 | 0.066 | 0.972 | 0.262 | 0.022 | 0.003 | 0.048 | 0.947 | 0.191 | 0.012 |
| | | LMRC* | 0.016 | 0.143 | 0.943 | 0.571 | 0.302 | 0.006 | 0.008 | 0.956 | 0.384 | 0.203 |
| | Hybrid | LMRC | 0.011 | 0.082 | 0.974 | 0.337 | 0.055 | 0.016 | 0.063 | 0.944 | 0.266 | 0.040 |
| | | LMRC* | 0.027 | 0.258 | 0.963 | 0.788 | 0.467 | 0.013 | 0.130 | 0.934 | 0.517 | 0.288 |
| M3 | $t(2)$ | LMRC | 0.069 | 0.350 | 0.874 | 1.399 | 0.739 | 0.049 | 0.281 | 0.893 | 1.099 | 0.581 |
| | | LMRC* | 0.072 | 0.327 | 0.920 | 1.257 | 0.670 | 0.059 | 0.309 | 0.880 | 1.192 | 0.631 |
| | $U[-2, 10]$ | LMRC | 0.004 | 0.060 | 0.957 | 0.237 | 0.019 | 0.001 | 0.041 | 0.961 | 0.164 | 0.009 |
| | | LMRC* | 0.013 | 0.146 | 0.948 | 0.582 | 0.313 | 0.009 | 0.104 | 0.951 | 0.423 | 0.225 |
| | Hybrid | LMRC | 0.011 | 0.070 | 0.975 | 0.273 | 0.042 | 0.011 | 0.048 | 0.968 | 0.204 | 0.022 |
| | | LMRC* | 0.022 | 0.170 | 0.940 | 0.670 | 0.379 | 0.014 | 0.124 | 0.949 | 0.490 | 0.274 |
| M4 | $t(2)$ | LMRC | 0.064 | 0.332 | 0.904 | 1.336 | 0.708 | 0.039 | 0.254 | 0.923 | 1.006 | 0.532 |
| | | LMRC* | 0.070 | 0.348 | 0.901 | 1.368 | 0.724 | 0.069 | 0.337 | 0.834 | 1.313 | 0.696 |
| | $U[-2, 10]$ | LMRC | 0.012 | 0.148 | 0.948 | 0.599 | 0.110 | 0.008 | 0.108 | 0.941 | 0.434 | 0.058 |
| | | LMRC* | 0.026 | 0.198 | 0.935 | 0.797 | 0.425 | 0.013 | 0.138 | 0.941 | 0.543 | 0.292 |
| | Hybrid | LMRC | 0.181 | 0.426 | 0.729 | 1.853 | 1.141 | 0.105 | 0.341 | 0.784 | 1.388 | 0.839 |
| | | LMRC* | 0.163 | 0.420 | 0.848 | 1.779 | 1.074 | 0.103 | 0.344 | 0.911 | 1.395 | 0.835 |

### 5.3. *Variable selection*

The covariates $X$ are generated from a multivariate normal distribution with mean 0 and $\Sigma = (\rho_{ij})$ with $\rho_{ij} = 0.1^{|i-j|}$. We consider models M1-M4 with $n = 100$ and $p = 40$ or $p = 80$. The first 5 components of $\beta_0$ are 2 and the rest are 0. Two error distributions are tried: Chi-square distribution with 1 degree of freedom (denoted by $\chi^2(1)$) and Poisson distribution with mean 1 (denoted by $\text{Pois}(1)$). The sample covariance matrix $X$ is computed with $n$ samples. The tuning parameter $\lambda_n$ is selected by the generalized information criterion (Nishii, 1984): $\text{GIC}_{a_n}(\lambda) = L_n\{\hat{\beta}_n(\lambda)\} - (a_n/n)\|\hat{\beta}_n(\lambda)\|_0$, where $\hat{\beta}_n(\lambda)$ is defined in Section 4, $\|\beta\|_0$ is the number of nonzero components of $\beta$ and $a_n$ is a positive sequence depending on $n$. In practice, we let $a_n = \log(n)$, and select $\lambda$ by maximizing the $\text{GIC}_{a_n}$ over $[0, 1]$ with 100 lattice points (Fan & Tang, 2013). For comparison, we also compute the sliced inverse regression with lasso (Lin et al., 2018), denoted by Lasso sliced inverse regression. Based on $500$ replications, the simulation results with $p = 80$ are presented in Table 5, and more results with $p = 40$ are provided in Table S5 in Section 3 of the supplementary material. We report the averaged false positive rate (FP), false negative rate (FN), the empirical probability of choosing the correct model (CM), absolute bias (BIAS), standard errors (SE), and the mean $\ell_1$, $\ell_2$ distances between $\beta_0$ and its estimates. It can be seen that our method performs better than Lasso sliced inverse regression in terms of smaller BIAS, SE and FP.

Table 4: Summary statistics for censored data with dimension $p = 5$. We consider different censoring variable mechanisms and censoring rates (in parentheses)

| Model | Censoring (CR) | Method | $n = 100$ | | | | | $n = 200$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Bias | SE | CP | $\ell_1$ | $\ell_2$ | Bias | SE | CP | $\ell_1$ | $\ell_2$ |
| M2 | CV1 (0.521) | LMRC* | 0.032 | 0.218 | 0.937 | 0.881 | 0.463 | 0.016 | 0.151 | 0.943 | 0.603 | 0.321 |
| | | LMRC | 0.009 | 0.144 | 0.942 | 0.572 | 0.104 | 0.006 | 0.103 | 0.934 | 0.408 | 0.054 |
| | | PRE | 0.161 | 0.469 | 0.898 | 1.826 | 0.990 | 0.105 | 0.385 | 0.895 | 1.310 | 0.722 |
| | CV2 (0.391) | LMRC* | 0.026 | 0.212 | 0.938 | 0.851 | 0.450 | 0.014 | 0.154 | 0.947 | 0.615 | 0.326 |
| | | LMRC | 0.004 | 0.138 | 0.945 | 0.545 | 0.095 | 0.005 | 0.092 | 0.953 | 0.369 | 0.043 |
| | | PRE | 0.171 | 0.488 | 0.884 | 1.967 | 1.069 | 0.105 | 0.382 | 0.888 | 1.321 | 0.732 |
| | CV3 (0.196) | LMRC* | 0.078 | 0.192 | 0.889 | 0.901 | 0.482 | 0.077 | 0.141 | 0.837 | 0.740 | 0.404 |
| | | LMRC | 0.092 | 0.129 | 0.837 | 0.711 | 0.171 | 0.095 | 0.089 | 0.854 | 0.603 | 0.124 |
| | | PRE | 0.129 | 0.430 | 0.896 | 1.589 | 0.867 | 0.083 | 0.358 | 0.876 | 1.186 | 0.666 |
| M4 | CV1 (0.685) | LMRC* | 0.041 | 0.256 | 0.938 | 1.049 | 0.552 | 0.023 | 0.186 | 0.937 | 0.743 | 0.394 |
| | | LMRC | 0.014 | 0.210 | 0.938 | 0.843 | 0.222 | 0.015 | 0.148 | 0.950 | 0.596 | 0.111 |
| | | PRE | 0.225 | 0.561 | 0.837 | 2.331 | 1.243 | 0.109 | 0.376 | 0.886 | 1.347 | 0.743 |
| | CV2 (0.441) | LMRC* | 0.033 | 0.220 | 0.926 | 0.885 | 0.469 | 0.015 | 0.150 | 0.947 | 0.597 | 0.314 |
| | | LMRC | 0.005 | 0.140 | 0.947 | 0.547 | 0.097 | 0.004 | 0.095 | 0.958 | 0.371 | 0.045 |
| | | PRE | 0.208 | 0.544 | 0.836 | 2.159 | 1.163 | 0.109 | 0.387 | 0.868 | 1.291 | 0.714 |
| | CV3 (0.214) | LMRC* | 0.087 | 0.187 | 0.871 | 0.911 | 0.492 | 0.088 | 0.138 | 0.895 | 0.787 | 0.432 |
| | | LMRC | 0.098 | 0.131 | 0.829 | 0.741 | 0.188 | 0.105 | 0.088 | 0.087 | 0.643 | 0.143 |
| | | PRE | 0.179 | 0.513 | 0.850 | 1.954 | 1.067 | 0.126 | 0.451 | 0.874 | 1.554 | 0.853 |

CV, censoring variable; CR, censoring rate.

### 5.4. *Analysis of Beijing PM 2.5 dataset*

We apply our method to analyze a Beijing PM 2.5 dataset originally studied by Liang et al. (2015). The dataset consists of hourly PM 2.5 readings taken at the US Embassy in Beijing located at (116.47 E, 39.95 N) and hourly meteorological measurements at Beijing Capital International Airport (BCIA) obtained from weather.nocrew.org. Both data series run from 1 January 2010 to 31 December 2014 with 43824 observations. There are 8 attributes in total, i.e, PM 2.5 concentration (PM, $\mu g/m^3$), Dew Point (DEWP, in Celsius Degree), Temperature (TEMP, in Celsius Degree), Pressure (PRES, in hPa), Combined wind direction (CBWD), Cumulated wind speed (CWP, $m/s$), Cumulated hours of snow (CS), Cumulated hours of rain (CR). The weather data have 16 wind directions. Based on the locations of major polluting industries around Beijing, the directions can be grouped into four broad categories: northwest (NW), northeast (NE), south (S), and calm and variable (CV). CWP is the cumulated wind speed from the start of the wind direction to the time of interest. CS and CR can be similarly defined.

Table 5: Summary statistics with dimension $p = 80$.

| Dimension | Model | Error | Method | $n = 100$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | BIAS | SE | $\ell_1$ | $\ell_2$ | FP | FN | CM |
| $p = 80$ | M1 | $\chi^2(1)$ | Lasso LMRC | 0.001 | 0.022 | 1.279 | 0.688 | 0.000 | 0.001 | 0.996 |
| | | | Lasso SIR | 0.009 | 0.131 | 4.634 | 1.152 | 0.262 | 0.001 | 0.997 |
| | | Pois(1) | Lasso LMRC | 0.002 | 0.017 | 1.013 | 0.538 | 0.001 | 0.000 | 1.000 |
| | | | Lasso SIR | 0.008 | 0.124 | 4.306 | 1.059 | 0.267 | 0.000 | 1.000 |
| | M2 | $\chi^2(1)$ | Lasso LMRC | 0.001 | 0.023 | 1.202 | 0.635 | 0.002 | 0.002 | 0.992 |
| | | | Lasso SIR | 0.009 | 0.134 | 4.769 | 1.181 | 0.271 | 0.001 | 0.997 |
| | | Pois(1) | Lasso LMRC | 0.001 | 0.017 | 1.041 | 0.562 | 0.000 | 0.000 | 1.000 |
| | | | Lasso SIR | 0.008 | 0.111 | 3.920 | 0.982 | 0.255 | 0.000 | 1.000 |
| | M3 | $\chi^2(1)$ | Lasso LMRC | 0.002 | 0.026 | 1.444 | 0.776 | 0.001 | 0.002 | 0.992 |
| | | | Lasso SIR | 0.022 | 0.227 | 8.343 | 2.069 | 0.273 | 0.005 | 0.983 |
| | | Pois(1) | Lasso LMRC | 0.000 | 0.023 | 1.293 | 0.696 | 0.001 | 0.001 | 0.996 |
| | | | Lasso SIR | 0.019 | 0.203 | 7.253 | 1.831 | 0.260 | 0.003 | 0.987 |
| | M4 | $\chi^2(1)$ | Lasso LMRC | 0.001 | 0.014 | 0.881 | 0.471 | 0.000 | 0.000 | 1.000 |
| | | | Lasso SIR | 0.012 | 0.153 | 5.438 | 1.395 | 0.249 | 0.001 | 0.997 |
| | | Pois(1) | Lasso LMRC | 0.001 | 0.016 | 0.955 | 0.499 | 0.000 | 0.001 | 0.996 |
| | | | Lasso SIR | 0.012 | 0.148 | 5.112 | 1.313 | 0.243 | 0.000 | 1.000 |

FP, averaged false positive rate; FN, averaged false negative rate; CM, the empirical probability of choosing the correct model.

Table 6: Correlation matrix of six covariates in Beijing PM 2.5 data.

| | DEWP | TEMP | PRES | CWP | CS | CR |
|---|---|---|---|---|---|---|
| DEWP | 1.000 | | | | | |
| TEMP | 0.824 | 1.000 | | | | |
| PRES | -0.778 | -0.827 | 1.000 | | | |
| CWP | -0.293 | -0.150 | 0.179 | 1.000 | | |
| CS | -0.035 | -0.095 | 0.071 | 0.023 | 1.000 | |
| CR | 0.125 | 0.050 | -0.081 | -0.009 | -0.010 | 1.000 |

DEWP, dew point in degrees Celsius; TEMP, temperature in degrees Celsius; PRES, pressure in hPa; CWP, cumulated wind speed in m/s; CS, cumulated hours of snow; CR, cumulated hours of rain.

The response variable is PM 2.5. The sample covariance matrix of 6 continuously-distributed predictors is calculated and their sample correlation are given in Table 6. Under different wind directions (NW,NE,S and CV), we fit model (3) to this dataset with the proposed linearized MRC, and present the results in Table 7, suggesting that the main source of pollution in Beijing is the heating in winter and the dense heavy industries nearby. First, PM 2.5 significantly increases with a lower temperature, indicating that the main reason of pollution in Beijing is the use of coal for heating in winter. Second, there is more pollution under a southern wind, and the southern wind reduces the pollution less significantly than wind from other directions, as Beijing is hemmed in by mountains to the west and north, and the south and the east of Beijing are dense with heavy industries. Third, the meteorological variables are mutually correlated. Thus, a decrease in the dew point and an increase in the pressure are usually accompanied by the arrival of the northerly wind, which brings in drier and fresher air, and reduces the pollution.

Table 7: The estimation results (estimated standard errors in parenthesis) of the Beijing PM 2.5 Data. The first row shows the mean (standard errors in parenthesis) PM 2.5 reading under different wind directions.

| | NW | NE | S | CV |
|---|---|---|---|---|
| PM2.5 | 70.128(89.064) | 90.178(95.197) | 110.822(80.050) | 126.152(100.403) |
| DEWP | 7.237(0.068) | 9.336(0.131) | 12.665(0.061) | 12.564(0.098) |
| TEMP | -11.190(0.085) | -12.65(0.158) | -15.881(0.098) | -16.849(0.071) |
| PRES | -4.725(0.113) | -3.328(0.204) | -2.830(0.193) | -1.809(0.196) |
| CWP | -0.505(0.009) | -2.008(0.093) | -0.379(0.040) | -1.106(0.184) |
| CS | 6.342(1.355) | -3.075(1.739) | -11.630(0.647) | -9.564(1.287) |
| CR | -9.942(0.351) | -11.456(0.561) | -18.523(0.870) | -20.568(1.158) |

NW, northwest; NE, northeast; S, south; CV, calm and variable.

## ACKNOWLEDGEMENT

The authors wish to thank the editor, the associate editor and two reviewers for their insightful comments and constructive suggestions that significantly helped improve the paper. The work of Chen was supported by the Hong Kong Research Grants Council. Lin was supported by the Hong Kong Research Grants Council and The Chinese University of Hong Kong.

## SUPPLEMENTARY MATERIAL

The supplementary material contains three numerical algorithms, lemmas and technical proofs for the main theorems and additional simulation results.

## REFERENCES

CAI, T. T., ZHANG, C.-H. & ZHOU, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics* **38**, 2118–2144.

CAMBANIS, S., HUANG, S. & SIMONS, G. (1981). On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis* **11**, 368–385.

CAVANAGH, C. & SHERMAN, R. P. (1998). Rank estimators for monotonic index models. *Journal of Econometrics* **84**, 351–381.

CHEN, C.-H., LI, K.-C. & WANG, J.-L. (1999). Dimension reduction for censored regression data. *The Annals of Statistics* **27**, 1–23.

COOK, R. D. & WEISBERG, S. (1991). Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association* **86**, 328–332.

FAN, Y., HAN, F., LI, W. & ZHOU, X.-H. (2020). On rank estimators in increasing dimensions. *Journal of Econometrics* **214**, 379–412.

FAN, Y. & TANG, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: SERIES B: Statistical Methodology* , 531–552.

HAN, A. K. (1987). Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics* **35**, 303–316.

HAN, F., JI, H., JI, Z. & WANG, H. (2017). A provable smoothing approach for high dimensional generalized regression with applications in genomics. *Electronic Journal of Statistics* **11**, 4347–4403.

HENMI, M. & EGUCHI, S. (2004). A paradox concerning nuisance parameters and projected estimating functions. *Biometrika* **91**, 929–941.

HENMI, M., YOSHIDA, R. & EGUCHI, S. (2007). Importance sampling via the estimated sampler. *Biometrika* **94**, 985–991.

HOROWITZ, J. L. (2009). *Semiparametric and nonparametric methods in econometrics*, vol. 12. Springer.

HRISTACHE, M., JUDITSKY, A., POLZEHL, J. & SPOKOINY, V. (2001). Structure adaptive approach for dimension reduction. *Annals of Statistics* , 1537–1566.

KHAN, S. & TAMER, E. (2007). Partial rank estimation of duration models with general forms of censoring. *Journal of Econometrics* **136**, 251–280.

KONG, E. & XIA, Y. (2007). Variable selection for the single-index model. *Biometrika* **94**, 217–229.

LI, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86**, 316–327.

LI, K.-C. & DUAN, N. (1989). Regression analysis under link violation. *The Annals of Statistics* , 1009–1052.

LIANG, X., ZOU, T., GUO, B., LI, S., ZHANG, H., ZHANG, S., HUANG, H. & CHEN, S. X. (2015). Assessing beijing's pm2. 5 pollution: severity, weather impact, apec and winter heating. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **471**, 20150257.

LIN, Q., ZHAO, Z. & LIU, J. S. (2018). On consistency and sparsity for sliced inverse regression in high dimensions. *The Annals of Statistics* **46**, 580–610.

MUIRHEAD, R. J. (2009). *Aspects of multivariate statistical theory*, vol. 197. John Wiley & Sons.

NEYKOV, M., LIN, Q. & LIU, J. S. (2016a). Signed support recovery for single index models in high-dimensions. *Annals of Mathematical Sciences and Applications* **1**, 379–426.

NEYKOV, M., LIU, J. S. & CAI, T. (2016b). L1-regularized least squares for support recovery of high dimensional single index models with gaussian designs. *The Journal of Machine Learning Research* **17**, 2976–3012.

NISHII, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics* , 758–765.

SONG, X., MA, S., HUANG, J. & ZHOU, X.-H. (2007). A semiparametric approach for the nonparametric transformation survival model with multiple covariates. *Biostatistics* **8**, 197–211.

TARPEY, T., OGDEN, R. T., PETKOVA, E. & CHRISTENSEN, R. (2014). A paradoxical result in estimating regression coefficients. *The American Statistician* **68**, 271–276.

WANG, J.-L., XUE, L., ZHU, L. & CHONG, Y. S. (2010). Estimation for a partial-linear single-index model. *The Annals of statistics* **38**, 246–274.

XIA, Y. & LI, W. K. (1999). On single-index coefficient regression models. *Journal of the American Statistical Association* **94**, 1275–1285.

XIA, Y. & TONG, H. (2006). On the efficiency of estimation for a single-index model. In *Frontiers in statistics*. World Scientific, pp. 63–85.

YUAN, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research* **11**, 2261–2286.