Statistica Sinica **24** (2014), 1341-1363
doi:http://dx.doi.org/10.5705/ss.2012.105

SURROGATE DIMENSION REDUCTION IN MEASUREMENT ERROR REGRESSIONS

Jun Zhang, Liping Zhu and Lixing Zhu

*Shenzhen University, Shanghai University of Finance and Economics
Hong Kong Baptist University*

Abstract: We generalize the cumulative slicing estimator to dimension reduction where the predictors are subject to measurement errors. Unlike existing methodologies, our proposal involves neither nonparametric smoothing in estimation nor normality assumption on the predictors or measurement errors. We establish strong consistency and asymptotic normality of the resultant estimators, allowing that the predictor dimension diverges with the sample size. Comprehensive simulations have been carried out to evaluate the performance of our proposal and to compare it with existing methods. A dataset is analyzed to further illustrate the proposed methodology.

Key words and phrases: Central subspace, diverging parameters, inverse regression, measurement error, surrogate dimension reduction.

1. Introduction

When predictors in regressions are observed with measurement errors, statistical analysis that ignores the measurement errors can cause substantial estimation bias. Regression models with semiparametric structures suffer from several additional issues; for instance, Carroll and Li (1992) pointed out that it is computationally demanding to evaluate the likelihood function of semiparametric regressions. Such issues are more serious when the predictor dimension is high. It is thus beneficial to develop some efficient dimension reduction methods for high-dimensional semiparametric regressions, particularly when the predictors are measured with errors. Toward this end, Carroll and Li (1992) investigated the semiparametric model $Y = \ell(B^\tau X, \varepsilon)$, where the predictors are contaminated with measurement errors as

$$W = \gamma + \Gamma X + \delta. \tag{1.1}$$

Here, $X = (X_1, \ldots, X_p)^\tau$ is the p-dimensional predictor vector, B is an unknown $p \times K$ matrix to be estimated from the observed data, τ is the transpose operator, ℓ is an unknown function, and ε is assumed to be independent of X. In (1.1), γ is an r-dimensional nonrandom vector, Γ is an $r \times p$ nonrandom matrix with

$r \geq p$, and δ is an r-dimensional random vector independent of (X, Y). Without loss of generality we take $\gamma = 0$ and $E(X) = E(\delta) = 0$ throughout. From Cook (1998a), we know that the relationship between Y and X is equivalent to the conditional independence statement

$$Y \perp\!\!\!\perp X \mid B^{\tau} X, \tag{1.2}$$

where $\perp\!\!\!\perp$ denotes statistical independence, $B \in \mathbb{R}^{p \times K}$, $K \leq p$. Clearly B is not unique, because any orthogonal transformation of B satisfies this conditional independence as well. Therefore, we are concerned with the column space of B, often referred to as the dimension reduction subspace and denoted by $\mathcal{S}(B)$. Observing that any matrix whose column space contains $\mathcal{S}(B)$ also satisfies (1.2), Cook (1998a) defined the central subspace (CS) as the intersection of all dimensional reduction subspaces satisfying (1.2) if itself is also a dimension reduction subspace. Following the convention in sufficient dimension reduction literature, we denote the CS by $\mathcal{S}_{Y|X}$. With a slight abuse of notation, we take $\mathcal{S}_{Y|X} = \mathcal{S}(B)$. Cook (1998a) provides a detailed account of useful ideas and results toward identifying the CS.

To investigate (1.2) with measurement error structure (1.1), Carroll and Li (1992) proposed to use the surrogate predictors

$$U \overset{\text{def}}{=} LW, \text{ where } L \overset{\text{def}}{=} \text{cov}(X, W)\Sigma_W^{-1} \text{ and } \Sigma_W \overset{\text{def}}{=} \text{cov}(W, W). \tag{1.3}$$

In spirit, the surrogate predictors U form the least squares prediction of X when W is given. The slope vector through regressing the coordinates of X against W constitutes the rows of the linear transformation L. Carroll and Li (1992) pointed out that ordinary least squares (OLS) and sliced inverse regression (SIR) can produce consistent estimators of $\mathcal{S}_{Y|X}$. Lue (2004) developed a modified principal Hessian direction (pHd) method to estimate the CS. Li and Yin (2007) established a general invariance law between the surrogate and the ordinary dimensional reduction subspaces:

$$Y \perp\!\!\!\perp X \mid B^{\tau} X \text{ if and only if } Y \perp\!\!\!\perp U \mid B^{\tau} U \tag{1.4}$$

when X and δ are jointly multivariate normal. Accordingly, all inverse regression based methods using (X, Y) can be readily adapted to the methods using (U, Y). In regression modeling, this equivalence implies that, when X is normally distributed, we can employ the regression calibration method (see, e.g., Carroll et al. (2006)) to deal with semiparametric regressions. If X or δ is not normally distributed, whether or not the invariance law holds true remains an open problem, although Li and Yin (2007) suggested an approximation based on the results of Hall and Li (1993). The adaptation of the OLS and the pHd

(Lue (2004)) is probably easy to implement, yet both the OLS and the pHd are in spirit targeting the central mean subspace (CMS; Cook and Li (2002)) rather than the CS. In addition, pHd requires the constant variance condition, which is more stringent than the linearity condition and may not be true if X deviates from multivariate normality.

Here we first establish connection between the ordinary CS, $\mathcal{S}_{Y|X}$, and the surrogate CS, $\mathcal{S}_{Y|U}$, when X satisfies the linearity condition, then elaborate the notion of cumulative slicing estimation (Zhu, Zhu, and Feng (2010)) to recover $\mathcal{S}_{Y|X}$. Compared with the sliced inverse regression in Carroll and Li (1992), cumulative slicing estimation requires no nonparametric smoothing in estimation. We establish the strong consistency and asymptotic normality of our proposed estimators, allowing the predictor dimension p to diverge to infinity as the sample size n grows. In particular, we show that strong consistency and asymptotic normality hold true when the predictor dimension $p = o(\sqrt{n}/\log n)$ and $p = o(n^{1/3})$, respectively. It is worth mentioning that, although investigation of such results is not the focus of this paper, this paper provides some insights into estimating the CS if some variable selection techniques have been adopted to choose important ones from ultra-high dimensional candidate predictors. Specifically, we can obtain consistent estimators of the CS as long as the number of important predictors is of order $o(\sqrt{n}/\log n)$.

The remainder of this paper is structured as follows. In Section 2, we adapt the notion of cumulative slicing estimation to accommodate measurement error regressions, and justify its theoretical underpinnings for the surrogate dimensional reduction problems at the population level. We then discuss the estimation procedure at the sample level and derive some relevant asymptotics. We report results of several simulation studies in Section 3 and further illustrate our proposed methodology through an application to a dataset in Section 4. In Section 5, we discuss several extensions of our proposal. This paper concludes with a brief discussion in Section 6. All technical details are in the Appendix.

2. Methodology and Asymptotic Properties

2.1. Rationale of the method

We assume here that the predictors satisfy the linearity condition (Li (1991)):

$$E(X \mid B^{\tau}X) = P_B^{\tau}(\Sigma_X)X, \tag{2.1}$$

where $\Sigma_X \overset{\text{def}}{=} \text{cov}(X, X)$ denotes the covariance matrix of X, and $P_B(\Sigma_X) \overset{\text{def}}{=} B(B^{\tau}\Sigma_X B)^{-1}B^{\tau}\Sigma_X$ is the projection operator in the Σ_X inner product of B. Under (1.2), (2.1), and the adjoint property of conditional expectation, Zhu, Zhu, and Feng (2010) obtained that

$$\Lambda(y) \overset{\text{def}}{=} \text{cov}(\mathbf{1}\{Y \le y\}, X)$$

$$
\begin{aligned}
&= E\Big[\mathbf{1}\{Y \le y\} E(X \mid Y, B^\tau X)\Big] \\
&= E\Big[\mathbf{1}\{Y \le y\} E(X \mid B^\tau X)\Big] \\
&= P_B^\tau(\Sigma_X)\mathrm{cov}(\mathbf{1}\{Y \le y\}, X), \qquad (2.2)
\end{aligned}
$$

indicating that $\Sigma_X^{-1}\Lambda(y)$ lies in the CS. It can be proved without much difficulty that $\mathrm{span}\{\Sigma_X^{-1}\Lambda(y), y \in \mathbb{R}\} = \mathrm{span}\{\Lambda\} \subseteq \mathcal{S}_{Y|X}$, where

$$\Lambda \overset{\text{def}}{=} \Sigma_X^{-1} E[\Lambda(\widetilde{Y})\Lambda^\tau(\widetilde{Y})] \qquad (2.3)$$

is a kernel matrix, and $\widetilde{Y}$ is an independent copy of Y. Such an observation implies that the spectral decomposition of Λ helps to infer about $\mathcal{S}_{Y|X}$ through the eigenvectors associated with the nonzero eigenvalues of Λ. In effect, (2.2) and (2.3) make the core of the cumulative mean estimation (CUME; Zhu, Zhu, and Feng (2010)). The advantage of the CUME is that it uses a determining class of unconditional covariances $\mathrm{cov}(\mathbf{1}\{Y \le \cdot\}, X)$, and it is a simple moment method without nonparametric estimation.

In semiparametric regression with measurement errors, the predictors X at (1.2) cannot be observed precisely, but rather via (1.1). Consequently, (2.3) cannot be applied directly to recover $\mathcal{S}_{Y|X}$. We expect to utilize the surrogate prediction of X based on (1.1) to identify $\mathcal{S}_{Y|X}$. Recall the definition of U and L in (1.3). A proposition establishes the connection between the seed vector using the underlying unobservable X and the seed vector using the surrogate predictor vector U.

Proposition 1. Suppose $\delta \perp\!\!\!\perp (X, Y)$ at (1.1), and that $\Sigma_X \overset{\text{def}}{=} \mathrm{cov}(X, X)$, $\Sigma_U \overset{\text{def}}{=} \mathrm{cov}(U, U)$, and $\Sigma_W \overset{\text{def}}{=} \mathrm{cov}(W, W)$ are all positive-definite matrices. If $M(y) \overset{\text{def}}{=} \mathrm{cov}(\mathbf{1}\{Y \le y\}, U)$ and $\Lambda(y) \overset{\text{def}}{=} \mathrm{cov}(\mathbf{1}\{Y \le y\}, X)$, then

$$\Sigma_U^{-1} M(y) = \Sigma_X^{-1}\Lambda(y). \qquad (2.4)$$

Relation (2.4) in Proposition 1 holds without the linearity condition. However, this condition is still needed to ensure that the columns of $\Sigma_U^{-1}M(y)$ and $\Sigma_X^{-1}\Lambda(y)$ are vectors in the central subspace. Proposition 1 also shows that $\mathrm{span}\{\Sigma_U^{-1}M(y), y \in \mathbb{R}\} = \mathrm{span}\{\Sigma_X^{-1}\Lambda(y), y \in \mathbb{R}\} = \mathrm{span}\{\Lambda\}$. Thus, to recover $\mathcal{S}_{Y|X}$, it suffices to use the equivalent kernel matrix

$$M \overset{\text{def}}{=} \Sigma_U^{-1} E[M(\widetilde{Y})M^\tau(\widetilde{Y})]. \qquad (2.5)$$

Corollary 1. *Under the conditions in Proposition* 1*,*

$$\mathcal{S}(M) = \mathcal{S}(\Lambda). \qquad (2.6)$$

An immediate consequence of Corollary 1 is that, with the linearity condition (2.1), $\mathcal{S}(\Lambda) = \mathcal{S}(M) \subseteq (\mathcal{S}_{Y|X} \cap \mathcal{S}_{Y|U})$. Thus, (2.6) implies that, under the linearity condition, we can recover at least a subspace of the intersection $\mathcal{S}_{Y|X} \cap \mathcal{S}_{Y|U}$. This allows us to employ a spectral decomposition directly on the matrix M to infer about $\mathcal{S}_{Y|X}$: i.e., $b_1, \dots, b_K$ are the eigenvectors of M corresponding to its K nonzero eigenvalues, then $\mathcal{S}(b_1, \dots, b_K)$ can be used to estimate $\mathcal{S}_{Y|X}$.

2.2. Estimation procedure

We discuss how to estimate the CS of measurement error regressions at the sample level, assuming L is known or unknown. When L is unknown, we present two methods to estimate it.

2.2.1. L is known

Based on (1.3), $M(y)$ in Proposition 1 can be written as $M(y) = LV(y)$, where $V(y) \overset{\text{def}}{=} \text{cov}(\mathbf{1}\{Y \leq y\}, W)$. We take

$$V \overset{\text{def}}{=} E[V(\widetilde{Y})V^{\tau}(\widetilde{Y})]. \tag{2.7}$$

Suppose that n observations $\{(w_i, y_i), i = 1, \dots, n\}$ are available, and our objective is to estimate, using the (w_i, y_i)'s, the kernel matrix V and then its eigenvalues and corresponding eigenvectors. Let $E_n(\cdot)$ be the average over all sample points, so $E_n[f(X,Y)] = n^{-1}\sum_{i=1}^{n} f(x_i, y_i)$. An estimator of $V(y)$ for any given y, denoted by $V_n(y)$, is

$$V_n(y) \overset{\text{def}}{=} n^{-1}\sum_{i=1}^{n}[w_i - E_n(W)]\mathbf{1}\{y_i \leq y\}.$$

Hence, an estimator of V, written as V_n, is

$$V_n \overset{\text{def}}{=} E_n[V_n(Y)V_n^{\tau}(Y)]. \tag{2.8}$$

Accordingly, we can estimate $M(y)$ and M with

$$M_n(y) \overset{\text{def}}{=} LV_n(y), \text{ and } M_n \overset{\text{def}}{=} LE_n[V_n(Y)V_n^{\tau}(Y)]L^{\tau}. \tag{2.9}$$

We can estimate Σ_W using $\Sigma_{W,n} = n^{-1}\sum_{i=1}^{n}[w_i - E_n(W)][w_i - E_n(W)]^{\tau}$ and Σ_U using $\Sigma_{U,n} = L\Sigma_{W,n}L^{\tau}$. Let $\widehat{b}_1, \dots, \widehat{b}_K$ be the K principal eigenvectors of $\Sigma_{U,n}^{-1}M_n$. Then, $\mathcal{S}(\widehat{b}_1, \dots, \widehat{b}_K)$ can be used to estimate $\mathcal{S}_{Y|X}$.

2.2.2. L is unknown

When L is unknown, there are two ways to estimate it depending upon availability: using validation data, or replication data. Consider a validation dataset $(x'_1, w'_1), \ldots, (x'_m, w'_m)$ as an external sample, and assume the size of m is much smaller than n. Using this auxiliary sample to estimate L through least squares, we have

$$L_m^{(1)} \stackrel{\text{def}}{=} \Sigma_{(1)XW,m}\Sigma_{(1)W,m}^{-1}, \tag{2.10}$$

where

$$\Sigma_{(1)XW,m} \stackrel{\text{def}}{=} m^{-1}\sum_{i=1}^{m}[x'_i - E_m(X')][w'_i - E_m(W')]^{\tau} \text{ and}$$
$$\Sigma_{(1)W,m} \stackrel{\text{def}}{=} m^{-1}\sum_{i=1}^{m}[w'_i - E_m(W')][w'_i - E_m(W')]^{\tau}.$$

Next we turn to replication data. Here $\Gamma = I_p$ and W is an unbiased surrogate for X, as considered by Carroll and Li (1992), I_p is the $p \times p$ identity matrix. Suppose $(x'_1, w'_{1j}), \ldots, (x'_m, w'_{mj})$ are generated from

$$w'_{ij} = \gamma + x'_i + \delta_{ij}, j = 1, 2, i = 1, \ldots, m, \tag{2.11}$$

where δ_{ij}'s are independent and identically distributed, and are independent of (w_{ij}, y_i). From (2.11),

$$\text{var}(w'_{i1} - w'_{i2}) = 2\Sigma_\delta, \quad \text{var}(w'_{i1} + w'_{i2}) = 4\Sigma_X + 2\Sigma_\delta.$$

Thus we can use $\{(x'_1, w'_{1j}), \ldots, (x'_m, w'_{mj})\}$ to estimate Σ_δ and Σ_W by

$$\Sigma_{\delta,m} \stackrel{\text{def}}{=} (2m)^{-1}\sum_{i=1}^{m}(w'_{i1} - w'_{i2})(w'_{i1} - w'_{i2})^{\tau}, \text{and}$$
$$\Sigma_{(2)W,m} \stackrel{\text{def}}{=} (4m)^{-1}\sum_{i=1}^{m}\left[(\widetilde{w}'_{i1} + \widetilde{w}'_{i2})(\widetilde{w}'_{i1} + \widetilde{w}'_{i2})^{\tau} + (\widetilde{w}'_{i1} - \widetilde{w}'_{i2})(\widetilde{w}'_{i1} - \widetilde{w}'_{i2})^{\tau}\right],$$

where $\widetilde{w}'_{ij} \stackrel{\text{def}}{=} w'_{ij} - E_m(W'_j), i = 1, \ldots, m, j = 1, 2$. Because $L = \Sigma_{XW}\Sigma_W^{-1} = I_p - \Sigma_\delta\Sigma_W^{-1}$, take

$$L_m^{(2)} \stackrel{\text{def}}{=} I_p - \Sigma_{\delta,m}\Sigma_{(2)W,m}^{-1}. \tag{2.12}$$

With a consistent estimator of L defined in either (2.10) when the validation data are used, or in (2.12) when the replication data are used, the corresponding estimators of $M(y)$ and M are

$$M^*_{mn}(y) \stackrel{\text{def}}{=} L_m^{(j)}V_n(y) \text{ and } M^*_{mn} \stackrel{\text{def}}{=} L_m^{(j)}E_n[V_n(Y)V_n^{\tau}(Y)]L_m^{(j)\tau}, \tag{2.13}$$

where $j = 1$ is for the validation data case and $j = 2$ is for the replication data case. We can now obtain the two estimators of Σ_U, denoted by

$\Sigma_{U,mn} = L_m^{(j)}\Sigma_{W,n}L_m^{(j)\tau}, j = 1, 2$. Let $\widehat{b}_1^*, \ldots, \widehat{b}_K^*$ be the K principal eigenvectors of $\Sigma_{U,mn}^{-1}M_{mn}^*$. Then, $\mathcal{S}(\widehat{b}_1^*, \ldots, \widehat{b}_K^*)$ can be used to estimate $\mathcal{S}_{Y|X}$.

2.3. Asymptotic results

Theorem 1 gives the strong consistency of the kernel matrices of the CUME method in estimating the CS of measurement error regressions.

Theorem 1. *In addition to the linearity condition, we assume the following.*

(A1) $\max_{1\le i\le p} E|X_i^4| < \infty$ *and* $\max_{1\le i\le p} E|\delta_i^4| < \infty$ *hold uniformly for* p.

(A2) *For largest eigenvalue of* LL^τ *and* Σ_W, $\lambda_{\max}(LL^\tau) < +\infty$, $\lambda_{\max}(\Sigma_W) < +\infty$ *uniformly for* p.

(A3) Σ_X *and* Σ_U *are positive definite matrices uniformly for all* p.

Then $\|\Sigma_{U,n}^{-1}M_n - \Sigma_U^{-1}M\| = o(p\log n/\sqrt{n})$ *almost surely when* L *is known, and* $\|\Sigma_{U,mn}^{-1}M_{mn}^* - \Sigma_U^{-1}M\| = o(p\log m/\sqrt{m})$ *almost surely when* L *is unknown and needs to be estimated from the validation or replication data.*

Strong consistency thus holds true even when p goes to infinity at a rate of $o(\sqrt{n}/\log n)$.

Theorem 2. *In addition to the conditions in Theorem* 1, *we assume the following.*

(A4) $\max_{1\le i\le p} E|X_i^8| < \infty$ *and* $\max_{1\le i\le p} E|\delta_i^8| < \infty$ *uniformly for* p.

(A5) $var(e^\tau\Sigma_U^{-1}S_1b_i) = \lambda_i^{-2}e^\tau\Sigma_U^{-1}E(S_1b_ib_i^\tau S_1^\tau)\Sigma_U^{-1}e \to G_i$, *where* G_i *is a positive constant,* $i = 1, \ldots, K$, *and* S_1 *is defined in* (A.8) *in the Appendix.*

For e *a unit length vector orthogonal to* $\mathcal{S}_{Y|X}$, L *is known, and* $p = o(n^{1/3})$,

$$\sqrt{n}e^\tau\widehat{b}_i \to N(0, G_i), \quad i = 1, \ldots, K.$$

If m *is the size of the validation or replicated dataset with* $m < n$, L *is unknown, and* $p = o(m^{1/3})$, *then*

$$\sqrt{m}e^\tau\widehat{b}_i^* \to N(0, G_i), \quad i = 1, \ldots, K.$$

2.4. Dimension determination

Zhu, Zhu, and Feng (2010) stated that, as the predictor dimension p diverges with the sample size n, the usual χ^2-sequential test statistic (Li (1991) and Cook (1998a)) may not be proper because the degree of freedom of the χ^2-sequential test also diverges. Following Zhu, Miao, and Peng (2006), we proposed a BIC-type criterion to determinate K:

$$\widehat{K} \overset{\text{def}}{=} \arg\min_{1\le k\le p}\{G(k)\}, \tag{2.14}$$

where

$$G(k) \overset{\text{def}}{=} \sum_{i=1}^{k}\widehat{\lambda}_i^2 \Big/ \sum_{i=1}^{p}\widehat{\lambda}_i^2 - C_n\frac{k(k+1)}{2}.$$

Here $k(k+1)/2$ is the number of free parameters when the target matrix is of rank k. The $\widehat{\lambda}_i$s are the sample eigenvalues of $\Sigma_{U,n}^{-1}M_n$ or $\Sigma_{U,mn}^{-1}M_{mn}^*$. Zhu, Zhu, and Feng (2010) suggested $C_n = 2n^{3/4}$. How to choose an optimal penalty for C_n in a data-driven manner is a challenging issue, particularly when p is divergent.

3. Numerical Studies

In this section we report on simulations to evaluate the performance of our proposed method and to compare it with existing competitors. To measure estimation accuracy, we adopt the trace correlation criterion proposed by Férre (1998): $R^2(K) = \text{trace}(P_B P_{\widehat{B}})/K$, where B is a $p\times K$ matrix spanning $\mathcal{S}_{Y|X}$, $\widehat{B}$ is a $p\times K$ matrix whose columns are the eigenvectors associated with the K nonzero eigenvalues of $\Sigma_{U,n}^{-1}M_n$ or $\Sigma_{U,mn}^{-1}M_{mn}^*$, and P_B and $P_{\widehat{B}}$ are the respective projection operators in the standard inner product of B and $\widehat{B}$. The closer $R^2(K)$ is to 1, the better the performance of the K nonzero eigenvalues of $\Sigma_{U,n}^{-1}M_n$ or $\Sigma_{U,mn}^{-1}M_{mn}^*$ in estimating $\mathcal{S}_{Y|X}$.

We compare our method with SIR (Carroll and Li (1992)), pHd (Lue (2004)) and the CR proposed in Li and Yin (2007), all of which utilize the surrogate variable U instead of X. For SIR, different numbers of slices in the estimators were chosen, H at 5 and 10. For CR, the cutting number c was taken to be 0.5. To examine the performance of our method in high-dimensional cases with p associated with the sample size, we considered sample sizes with $(n,p) = (100,19), (225,29), (400,39), (625,49)$. Six models were adopted:

$$Y = X_1 - X_2 + X_3 + 4\varepsilon, \tag{3.1}$$

$$Y = \exp^{X_1 - X_2 + 4\varepsilon}, \tag{3.2}$$

$$Y = \sin(X_1 - X_2 + 1.5\varepsilon), \tag{3.3}$$

$$Y = \log(|X_1 + 1|) + \varepsilon, \tag{3.4}$$

$$Y = X_1(1 + X_2) + 0.5\varepsilon, \tag{3.5}$$

$$Y = X_2/[0.5 + (1.5 + X_1)^2] + \varepsilon. \tag{3.6}$$

In these models, $X = (X_1, X_2, \ldots, X_p)^\tau$ is independent of ε. We want to estimate $\mathcal{S}_{Y|X}$. In (3.1), $K = 1$ and $\mathcal{S}_{Y|X} = \text{span}\{(1,-1,1,0,0,\ldots,0)^\tau\}$;

in (3.2) and (3.3), $K = 1$ and $\mathcal{S}_{Y|X} = \text{span}\{(1, -1, 0, 0, 0, \ldots, 0)^\tau\}$; in (3.4), $K = 1$ and $\mathcal{S}_{Y|X} = \text{span}\{(1, 0, 0, 0, 0, \ldots, 0)^\tau\}$; in (3.5) and (3.6), $K = 2$ and $\mathcal{S}_{Y|X} = \text{span}\{(1, 0, 0, \ldots, 0)^\tau, (0, 1, 0, 0, \ldots, 0)^\tau\}$. The measurement errors in (1.1) are set in each example.

Example 1. We assume that the link L in known. The predictors X_{ij} were generated from the t-distribution $t(6)$, $i = 1, \ldots n$, $j = 1 \ldots p$, and the ε were drawn from uniform distribution over the interval [-0.2, 0.2]. Because X contains measurement errors in the manner of model (1.1), we took Γ as the $p \times p$ matrix with diagonal elements 1, and off-diagonal elements 0.5, and $\delta_i \sim N_p(0, 0.3^2 \times I_p)$, $i = 1, \ldots n$.

By invoking (1.3), we have

$$L = \Gamma^\tau \Big(\Gamma\Gamma^\tau + \frac{0.3^2}{\sigma^2_{t(6)}} \times I_p \Big)^{-1},$$

where $\sigma^2_{t(6)}$ stands for the variance of the t-distribution with six degrees of freedom. With this known L, we estimated $\mathcal{S}_{Y|X}$ following the routines suggested in Section 2.2.1. We conducted 200 simulation replications. The averages and standard deviations of the $R^2(K)$ values are reported in Table 1.

Models (3.1) and (3.2) have been shown to be favorable to SIR, and SIR again showed favorably for (3.1) and (3.2) with a transformation of Y. However, our method is superior to SIR in models (3.3)$-$(3.6), and is slightly better than CR in (3.1)-(3.4). In (3.5), the CR method is the winner over its competitors; further, the pHd method produces much lower $R^2(K)$ values than either the SIR or our method.

Example 2. In this example, we took validation data to be available. We generated X, δ, ε and Γ as in Example 1, but with the sample size m of the validation data much smaller. In our simulations, we chose $m = [n/2]$. We first employed the validation data to estimate $L_m^{(1)}$ by (2.10), then used (2.13) to recover $\mathcal{S}_{Y|X}$. The results for the means and the standard deviations of the $R^2(K)$ values, obtained from 200 simulation replications, are reported in Table 2.

In this example, SIR performs well for (3.1) and (3.2), while our method works well for the other models. The pHd exhibits poor performance here, in line with Cook's (1998b) observations that it is probably not efficient in symmetric models. All three methods in this example exhibit similar performance to those in Example 1, but less well given the need to estimate the unknown L.

Example 3. In this example the link L is estimated using replication data. Following Carroll and Li (1992), we generated the data from (2.11) by taking $\gamma =$

Table 1. With a known link L in Example 1. The means and standard deviations of the $R^2(K)$ values.

	SIR		pHd	CR	our method
$n = 100, p = 19$					
$H \to$	5	10			
model(3.1)	0.8613 ± 0.0436	0.8733 ± 0.0431	0.1547 ± 0.1248	0.8431 ± 0.0467	0.8598 ± 0.0407
model(3.2)	0.8560 ± 0.0508	0.8714 ± 0.0445	0.2936 ± 0.1652	0.8441 ± 0.0494	0.8545 ± 0.0487
model(3.3)	0.7489 ± 0.1459	0.7502 ± 0.1422	0.0356 ± 0.0497	0.7641 ± 0.1201	0.7745 ± 0.1112
model(3.4)	0.6271 ± 0.2122	0.6171 ± 0.2306	0.0872 ± 0.1684	0.6277 ± 0.1610	0.6698 ± 0.1430
model(3.5)	0.4557 ± 0.1191	0.4372 ± 0.1157	0.3666 ± 0.1205	0.5213 ± 0.1007	0.4967 ± 0.1101
model(3.6)	0.5955 ± 0.1028	0.5807 ± 0.1025	0.2253 ± 0.1066	0.4955 ± 0.1201	0.6738 ± 0.0894
$n = 225, p = 29$					
$H \to$	5	10			
model(3.1)	0.9046 ± 0.0266	0.9009 ± 0.0238	0.1809 ± 0.1369	0.8742 ± 0.0301	0.8990 ± 0.0243
model(3.2)	0.9065 ± 0.0310	0.9128 ± 0.0247	0.2634 ± 0.1315	0.8952 ± 0.0288	0.9058 ± 0.0227
model(3.3)	0.8348 ± 0.0682	0.8444 ± 0.0602	0.0305 ± 0.0505	0.8477 ± 0.0620	0.8563 ± 0.0580
model(3.4)	0.7562 ± 0.1040	0.7591 ± 0.1005	0.0918 ± 0.2297	0.7445 ± 0.0987	0.7813 ± 0.0702
model(3.5)	0.5422 ± 0.1118	0.5412 ± 0.1113	0.4329 ± 0.1185	0.6524 ± 0.0743	0.6307 ± 0.0811
model(3.6)	0.7212 ± 0.0642	0.7215 ± 0.0622	0.2303 ± 0.0948	0.5122 ± 0.0899	0.7588 ± 0.0491
$n = 400, p = 39$					
$H \to$	5	10			
model(3.1)	0.9324 ± 0.0166	0.9244 ± 0.0143	0.1646 ± 0.1178	0.9132 ± 0.0201	0.9206 ± 0.0138
model(3.2)	0.9313 ± 0.0165	0.9317 ± 0.0161	0.2826 ± 0.1184	0.9195 ± 0.0181	0.9295 ± 0.0143
model(3.3)	0.8678 ± 0.0427	0.8821 ± 0.0367	0.0223 ± 0.0427	0.8700 ± 0.0421	0.8859 ± 0.0347
model(3.4)	0.8240 ± 0.0610	0.8239 ± 0.0650	0.0893 ± 0.2324	0.7723 ± 0.0711	0.8175 ± 0.0506
model(3.5)	0.6371 ± 0.0715	0.6459 ± 0.0912	0.4983 ± 0.1037	0.7134 ± 0.0561	0.6929 ± 0.0616
model(3.6)	0.7752 ± 0.0398	0.7979 ± 0.0463	0.2661 ± 0.0916	0.5799 ± 0.0654	0.8017 ± 0.0366
$n = 625, p = 49$					
$H \to$	5	10			
model(3.1)	0.9464 ± 0.0096	0.9551 ± 0.0104	0.1745 ± 0.1153	0.9326 ± 0.0132	0.9466 ± 0.0098
model(3.2)	0.9441 ± 0.0120	0.9439 ± 0.0114	0.3029 ± 0.1175	0.9276 ± 0.0141	0.9376 ± 0.0111
model(3.3)	0.9003 ± 0.0282	0.9015 ± 0.0262	0.0183 ± 0.0289	0.8923 ± 0.0291	0.9046 ± 0.0233
model(3.4)	0.8526 ± 0.0502	0.8580 ± 0.0495	0.2064 ± 0.3319	0.8188 ± 0.0497	0.8588 ± 0.0433
model(3.5)	0.6969 ± 0.0578	0.7228 ± 0.0567	0.5558 ± 0.0945	0.7577 ± 0.0366	0.7384 ± 0.0494
model(3.6)	0.8124 ± 0.0369	0.8342 ± 0.0323	0.2917 ± 0.0950	0.6066 ± 0.0539	0.8366 ± 0.0299

0, $\Gamma = I_p$, so $w_{ij} = x_i + \delta_{ij}$, $j = 1, 2, i = 1, \ldots, m$. Here, x_{ij} was generated from the t-distribution $t(6)$, $i = 1, \ldots, m$, $j = 1, \ldots, p$, δ_{ij} was from $N_p(0, 0.3^2 \times I_p)$. $i = 1, \ldots, m, j = 1, 2$, and ε was from the uniform distribution over [-0.2,0.2]. The replication sample size was $m = [n/2]$. We obtained $L_m^{(2)}$ with (2.12), and estimated $\mathcal{S}_{Y|X}$ by (2.13). We investigated the performance of SIR, pHd, and our method. The averages of the 200 replications of the $R^2(K)$ values and the corresponding standard deviations are reported in Table 3.

Our conclusions are similar to those for Example 2. As n and m grow, the performance of SIR and our method improves, although our method exhibits superior performance for (3.3)-(3.6) and SIR works better for (3.1)$-$(3.2). In (3.5), the CR method performs quite well when both n and m are large.

The three examples convey similar messages: SIR works well for the linear and transformed linear models, while our method outperforms its competitors in some other complicated nonlinear models. Our method is comparable to SIR in (3.1)$-$(3.2) and to CR in (3.5). This suggests that our proposal is indeed worthy

Table 2. We estimated L with validation data in Example 2 and give the means and standard deviations of $R^2(K)$ values.

	SIR		pHd	CR	our method
$n = 100, m = 50, p = 19$					
$H \to$	5	10			
model(3.1)	0.6989 ± 0.1026	0.7253 ± 0.0869	0.1328 ± 0.1176	0.6990 ± 0.0993	0.7190 ± 0.0967
model(3.2)	0.7145 ± 0.0954	0.7062 ± 0.0988	0.2096 ± 0.1533	0.6823 ± 0.1290	0.7077 ± 0.1023
model(3.3)	0.6159 ± 0.1565	0.5882 ± 0.1515	0.0445 ± 0.0621	0.6123 ± 0.1411	0.6313 ± 0.1314
model(3.4)	0.5264 ± 0.1668	0.5145 ± 0.2109	0.0835 ± 0.1702	0.4923 ± 0.1689	0.5715 ± 0.1473
model(3.5)	0.3904 ± 0.1229	0.3726 ± 0.1163	0.3368 ± 0.1189	0.4756 ± 0.0923	0.4406 ± 0.1028
model(3.6)	0.5169 ± 0.1051	0.5322 ± 0.1073	0.2217 ± 0.1028	0.4239 ± 0.1106	0.5811 ± 0.0855
$n = 225, m = 112, p = 29$					
$H \to$	5	10			
model(3.1)	0.8409 ± 0.0389	0.8464 ± 0.0423	0.1513 ± 0.1191	0.8298 ± 0.0513	0.8355 ± 0.0487
model(3.2)	0.8277 ± 0.0457	0.8356 ± 0.0394	0.2422 ± 0.1242	0.8131 ± 0.0617	0.8325 ± 0.0511
model(3.3)	0.7513 ± 0.0824	0.7578 ± 0.0710	0.0278 ± 0.0385	0.7607 ± 0.0734	0.7703 ± 0.0679
model(3.4)	0.7167 ± 0.0937	0.7000 ± 0.0864	0.0797 ± 0.1814	0.6945 ± 0.0911	0.7275 ± 0.0824
model(3.5)	0.5158 ± 0.0896	0.5113 ± 0.1005	0.4061 ± 0.1126	0.5756 ± 0.0744	0.5502 ± 0.0859
model(3.6)	0.6634 ± 0.0723	0.6766 ± 0.0696	0.2383 ± 0.0918	0.5433 ± 0.0772	0.6915 ± 0.0646
$n = 400, m = 200, p = 39$					
$H \to 5$	10				
model(3.1)	0.8771 ± 0.0295	0.8867 ± 0.0247	0.1316 ± 0.1114	0.8600 ± 0.0379	0.8679 ± 0.0279
model(3.2)	0.8725 ± 0.0298	0.8620 ± 0.0301	0.2885 ± 0.1168	0.8270 ± 0.0334	0.8569 ± 0.0301
model(3.3)	0.8275 ± 0.0535	0.8248 ± 0.0524	0.0246 ± 0.0335	0.8055 ± 0.0489	0.8324 ± 0.0467
model(3.4)	0.7717 ± 0.0686	0.7775 ± 0.0664	0.1205 ± 0.2596	0.7461 ± 0.0609	0.7864 ± 0.0574
model(3.5)	0.6063 ± 0.0825	0.6176 ± 0.0905	0.4650 ± 0.1093	0.6734 ± 0.0540	0.6505 ± 0.0628
model(3.6)	0.7411 ± 0.0468	0.7530 ± 0.0455	0.2574 ± 0.0884	0.5876 ± 0.0563	0.7643 ± 0.0487
$n = 625, m = 312, p = 49$					
$H \to$	5	10			
model(3.1)	0.9041 ± 0.0214	0.9049 ± 0.0192	0.1422 ± 0.0961	0.8887 ± 0.0275	0.8997 ± 0.0201
model(3.2)	0.9017 ± 0.0214	0.9239 ± 0.0222	0.2861 ± 0.0995	0.8882 ± 0.0275	0.9178 ± 0.0203
model(3.3)	0.8600 ± 0.0398	0.8616 ± 0.0375	0.0150 ± 0.0191	0.8467 ± 0.0311	0.8632 ± 0.0308
model(3.4)	0.8210 ± 0.0549	0.8281 ± 0.0453	0.1754 ± 0.3037	0.8089 ± 0.0422	0.8301 ± 0.0388
model(3.5)	0.6737 ± 0.0601	0.6854 ± 0.0631	0.5149 ± 0.1081	0.7233 ± 0.0473	0.6999 ± 0.0538
model(3.6)	0.7862 ± 0.0341	0.8024 ± 0.0329	0.2939 ± 0.0905	0.6123 ± 0.0454	0.8045 ± 0.0317

of recommendation.

4. Application to Cardiovascular Disease Factors Two-Township Study

We applied our method to the Cardiovascular Disease Factors Two-Township Study carried out in Taiwan to investigate the risk factors for a high cholesterol level. The dataset was collected by Pan et al. (1997) and was used in Lue (2004). It includes six factors: Age (W_1), waist measurement (W_2), hip measurement (W_3), triglycedrine level (W_4), BMI measurement (W_5) and WHR measurement (W_6). Of these variables, age is observed precisely while the other five are measured with errors. All subjects have three replicates, each with the same size of 1,941. Each subject's cholesterol level (Y) was measured at the third examination. We are interested in investigating whether all six factors have an impact on subject's cholesterol level (Y). Following Lue (2004), we used the first two replicates of the five factors with measurement errors to obtain an estimator of L, denoted by $\widehat{L}_m^{(2)}$. We corrected the five factors in the third replicate by $\widehat{L}_m^{(2)}$ and then combined age (W_1) in the third examination with the transformed variables $\widehat{L}_m^{(2)}(W_2, \ldots, W_6)^\tau$ to obtain the surrogate variable $\widehat{U}$. With the BIC-type cri-

Table 3. We estimated L with replication data in Example 3 and give the means and standard deviations of $R^2(K)$ values.

	SIR		pHd	CR	our method
		$n = 100, m = 50, p = 19$			
$H \rightarrow$	5	10			
model(3.1)	0.8794 ± 0.0437	0.9021 ± 0.0379	0.1660 ± 0.1431	0.8851 ± 0.0458	0.8941 ± 0.0427
model(3.2)	0.8776 ± 0.0491	0.8842 ± 0.0397	0.2778 ± 0.1536	0.8521 ± 0.0556	0.8755 ± 0.0439
model(3.3)	0.7891 ± 0.1259	0.7721 ± 0.1473	0.0496 ± 0.0721	0.7700 ± 0.1078	0.7945 ± 0.1230
model(3.4)	0.6848 ± 0.1749	0.6789 ± 0.2126	0.0806 ± 0.1783	0.6878 ± 0.1092	0.7072 ± 0.1316
model(3.5)	0.4962 ± 0.1363	0.4400 ± 0.1488	0.4303 ± 0.1340	0.5534 ± 0.0899	0.5393 ± 0.1047
model(3.6)	0.6472 ± 0.1008	0.6493 ± 0.1137	0.2632 ± 0.1007	0.5023 ± 0.0877	0.7014 ± 0.0835
		$n = 225, m = 112, p = 29$			
$H \rightarrow$	5	10			
model(3.1)	0.9374 ± 0.0198	0.9212 ± 0.0143	0.1542 ± 0.1201	0.9167 ± 0.0201	0.9267 ± 0.0196
model(3.2)	0.9297 ± 0.0221	0.9172 ± 0.0166	0.3091 ± 0.1412	0.8801 ± 0.0389	0.9107 ± 0.0245
model(3.3)	0.8561 ± 0.0608	0.8521 ± 0.0804	0.0259 ± 0.0548	0.8453 ± 0.0412	0.8704 ± 0.0570
model(3.4)	0.7979 ± 0.0944	0.8066 ± 0.1134	0.1329 ± 0.2719	0.7765 ± 0.0813	0.8101 ± 0.0749
model(3.5)	0.6001 ± 0.1089	0.6176 ± 0.0998	0.4770 ± 0.1291	0.6856 ± 0.0834	0.6555 ± 0.0948
model(3.6)	0.7688 ± 0.0590	0.7718 ± 0.0628	0.2784 ± 0.0991	0.6134 ± 0.0676	0.8006 ± 0.0484
		$n = 400, m = 200, p = 39$			
$H \rightarrow$	5	10			
model(3.1)	0.9525 ± 0.0126	0.9630 ± 0.0100	0.1742 ± 0.1161	0.9328 ± 0.0144	0.9523 ± 0.0128
model(3.2)	0.9507 ± 0.0146	0.9433 ± 0.0158	0.3262 ± 0.1249	0.9217 ± 0.0134	0.9432 ± 0.0119
model(3.3)	0.9040 ± 0.0343	0.8992 ± 0.0416	0.0168 ± 0.0327	0.8832 ± 0.0319	0.9074 ± 0.0324
model(3.4)	0.8592 ± 0.0545	0.8584 ± 0.0531	0.2721 ± 0.3616	0.8133 ± 0.0521	0.8591 ± 0.0430
model(3.5)	0.6898 ± 0.0702	0.7039 ± 0.0772	0.5599 ± 0.1156	0.7445 ± 0.0498	0.7275 ± 0.0618
model(3.6)	0.8179 ± 0.0419	0.8244 ± 0.0427	0.3179 ± 0.0948	0.6455 ± 0.0445	0.8430 ± 0.0387
		$n = 625, m = 312, p = 49$			
$H \rightarrow$	5	10			
model(3.1)	0.9632 ± 0.0079	0.9598 ± 0.0072	0.1928 ± 0.1062	0.9401 ± 0.0085	0.9449 ± 0.0079
model(3.2)	0.9595 ± 0.0105	0.9588 ± 0.0089	0.3401 ± 0.1145	0.9389 ± 0.0121	0.9489 ± 0.0089
model(3.3)	0.9217 ± 0.0246	0.9224 ± 0.0236	0.0138 ± 0.0194	0.9004 ± 0.0187	0.9228 ± 0.0216
model(3.4)	0.8739 ± 0.0414	0.8855 ± 0.0403	0.3370 ± 0.3937	0.8654 ± 0.0423	0.8857 ± 0.0368
model(3.5)	0.7301 ± 0.0574	0.7554 ± 0.0570	0.6136 ± 0.1042	0.7823 ± 0.0396	0.7615 ± 0.0448
model(3.6)	0.8455 ± 0.0309	0.8628 ± 0.0276	0.3476 ± 0.1070	0.6676 ± 0.0348	0.8652 ± 0.0272

terion introduced in Section 2.4, we inferred that $\mathcal{S}_{Y|X}$ is one-dimensional. Our method gives

$$\widehat{\beta} = (0.1977, -0.5201, 0.7358, -0.0203, 0.3844, -0.0283)^{\tau},$$

possibly indicating that the hip measurement has the dominant effect. Using the same dataset, we also tried out the SIR with 10 slices to estimate the direction (Carroll and Li (1992)). The estimated direction is

$$\widehat{\beta}_{sir} = (0.2097, -0.4940, 0.7454, -0.0827, 0.3856, -0.0286)^{\tau},$$

which agrees with our observation that hip measurement plays the dominant role. The angle between the two estimated directions obtained with the two methods is almost 0, as the cosine value of angle $\widehat{\beta}^{\tau}\widehat{\beta}_{sir} = 0.9976$. Thus, for this dataset, our method and SIR agree that a subject's hip measurement has a potential influence on their cholesterol levels.

5. Some Extensions

There are several ways to extend the methodology to handle measurement error regressions. For example, an anonymous referee observed that, to identify the CS, the indicator function $\mathbf{1}\{Y \le y\}$ in (2.2) can be replaced with any measurable function $f_y(Y)$, for $y \in \mathbb{R}$. In parallel to (2.2), we can show that

$$\begin{aligned}\widetilde{\Lambda}(y) &\overset{\text{def}}{=} \mathrm{cov}(f_y(Y), X)\\ &= P_B^{\tau}(\Sigma_X)\mathrm{cov}(f_y(Y), X),\end{aligned} \tag{5.1}$$

indicating that, for an arbitrary measurable function $f_y(Y)$, $\Sigma_X^{-1}\widetilde{\Lambda}(y) \subseteq \mathcal{S}_{Y|X}$, $\widetilde{\Lambda} \overset{\text{def}}{=} \Sigma_X^{-1}E[\widetilde{\Lambda}(\widetilde{Y})\widetilde{\Lambda}^{\tau}(\widetilde{Y})]$ can be used to infer about $\mathcal{S}_{Y|X}$ through using the eigenvectors associated with the nonzero eigenvalues of $\widetilde{\Lambda}$. This modification indeed generalizes the scope of cumulative slicing estimation, which corresponds to the CUME method here if we choose $f_y(Y) = \mathbf{1}\{Y \le y\}$, and the ordinary least squares (OLS) procedure in Carroll and Li (1992) if we choose $f_y(Y) = Y$. Let $\widetilde{M}(y) \overset{\text{def}}{=} \mathrm{cov}(U, f_y(Y))$ for an arbitrary measurable function $f_y(Y)$. Following (5.1) and similar arguments for proving Proposition 1, we can show that

$$\Sigma_U^{-1}\widetilde{M}(y) = \Sigma_X^{-1}\widetilde{\Lambda}(y). \tag{5.2}$$

This enables us to design some other dimension reduction methods to analyze measurement error data.

Another way of generalizing the CUME method is to consider using the second moment of X given Y. Observing that $\mathrm{cov}(\mathbf{1}\{Y \le y\}, X) = \mathrm{cov}(\mathbf{1}\{Y \le y\}, E(X \mid Y))$, Zhu, Zhu, and Feng (2010) pointed out that the CUME method fails if $E(X \mid Y)$ degenerates. They proposed cumulative variance estimation (CUVE) and cumulative directional regression (CUDR). In the measurement error regressions, we can extend the CUME method. Zhu, Zhu, and Feng (2010) noticed that the CUME method replaces $E(X \mid Y)$ in SIR with $E[X\mathbf{1}\{Y \le y\}]$, and then proposed CUVE by replacing $\mathrm{var}(X \mid Y)$ by $\mathrm{var}[X\mathbf{1}\{Y \le y\}]$. Assuming (2.1) and the constant variance condition

$$\mathrm{var}(X \mid B^{\tau}X) = \Sigma_X - \Sigma_X B(B^{\tau}\Sigma_X B)^{-1}B^{\tau}\Sigma_X, \tag{5.3}$$

we now propose the kernel matrix of the CUVE for measurement error problem. Let $V(y) \overset{\text{def}}{=} \mathrm{var}[U1\{Y \le y\}] - F(y)L\Sigma_W L^{\tau}$ and

$$M^* \overset{\text{def}}{=} \Sigma_U^{-1}E[V(\widetilde{Y})V^{\tau}(\widetilde{Y})],$$

where $\widetilde{Y}$ is an independent copy of Y.

Theorem 3. *Assume* (2.1) *and* (5.3). *Then* $\mathcal{S}(M^*) \subseteq \mathcal{S}_{Y|X}$.

Following Zhu, Zhu, and Feng (2010), we can further develop an extension of the cumulative directional regression (CUDR) method under the measurement error setting. The directional regression (DR) method was proposed by Li and Wang (2007) using $E[(X-\widetilde{X})(X-\widetilde{X})^\tau \mid Y, \widetilde{Y}]$ to infer about $\mathcal{S}_{Y|X}$, where $(\widetilde{X}, \widetilde{Y})$ is an independent copy of (X, Y). Zhu, Zhu, and Feng (2010) extended the DR idea and proposed CUDR method by using $E[(X-\widetilde{X})(X-\widetilde{X})^\tau \mathbf{1}\{Y \le y\}\mathbf{1}\{\widetilde{Y} \le \widetilde{y}\}] - 2F(\widetilde{y})F(y)\Sigma_X$. Inspired by this, let

$$R(y, \widetilde{y}) \stackrel{\text{def}}{=} E[(U-\widetilde{U})(U-\widetilde{U})^\tau \mathbf{1}\{Y \le y\}\mathbf{1}\{\widetilde{Y} \le \widetilde{y}\}] - 2F(y)F(\widetilde{y})L\Sigma_W L^\tau,$$

where $(\widetilde{U}, \widetilde{Y})$ is an independent copy of (U, Y). Moreover, let

$$M^{**} \stackrel{\text{def}}{=} \Sigma_U^{-1} E[R(Y, \widetilde{Y})R^\tau(Y, \widetilde{Y})].$$

Theorem 4. *Assume* (2.1) *and* (5.3). *Then* $\mathcal{S}(M^{**}) \subseteq \mathcal{S}_{Y|X}$.

With Theorems 3 and 4, we can estimate $\mathcal{S}_{Y|X}$ as in Section 2. Details are omitted here.

6. Concluding Remarks

In this paper, we have studied sufficient dimension reduction when the predictors in semiparametric regressions are measured with errors. Our method enhances the estimation efficiency in comparison with existing methods. The resultant estimators are consistent when the predictor dimension p diverges with the sample size n at a rate of at most $p = o(n^{1/2}/\log n)$. This inspires us to consider the variable selection issue: When the predictor dimension is very large, how can important variables be selected, and to what extent can the predictor dimension p be brought down to a value much smaller than n while retaining the estimation consistency? There are such variable selection procedures, as the Dantzig selector (Candés and Tao (2007)), the LASSO (Tibshirani, R. (1996)), the SCAD (Fan and Li (2001)) and the SIS (Fan and Lv (2008)), etc. It may be feasible to combine one of these approaches with that proposed herein if an objective function is properly defined. Research along this line is warranted.

Acknowledgement

This work is supported by National Natural Science Foundation of China (11071077, 11201306, 11101157 and 11101063), Innovation Program of Shanghai Municipal Education Commission (13YZ065 and 13ZZ055), Pujiang Project of Science and Technology Commission of Shanghai Municipality (12PJ1403200), Program for New Century Excellent Talents, Ministry of Education of China (NCET-12-0901), the Fundamental Research Project of Shanghai Normal University (SK201207), a GRF grant from the Research Grants Council of Hong Kong and a FRG grant from Hong Kong Baptist University. zhu.liping@mail.shufe.edu.cn. The authors thank the Editor, an associate editor, and two anonymous referees for their valuable comments and suggestions which led to a substantial improvement in the presentation of an early manuscript.

Appendix

Proof of Proposition 1. We first observe that

$$\begin{aligned} L &= \mathrm{cov}(X,W)\Sigma_W^{-1} = \mathrm{cov}(X,\Gamma X+\delta)\Sigma_W^{-1} = \Sigma_X\Gamma^\tau\Sigma_W^{-1} \quad \text{and} \\ \Sigma_U &= \mathrm{cov}(LW) = L\mathrm{cov}(W)L^\tau = \mathrm{cov}(X,W)\Sigma_W^{-1}\mathrm{cov}(X,W)^\tau \\ &= \Sigma_X\Gamma^\tau\Sigma_W^{-1}\Gamma\Sigma_X = L\Gamma\Sigma_X. \end{aligned} \tag{A.1}$$

Consequently,

$$\begin{aligned} \mathrm{cov}(\mathbf{1}\{Y\le y\},U) &= L\Gamma\mathrm{cov}(\mathbf{1}\{Y\le y\},X) + L\mathrm{cov}(\mathbf{1}\{Y\le y\},\delta) \\ &= L\Gamma\Sigma_X\Sigma_X^{-1}\mathrm{cov}(\mathbf{1}\{Y\le y\},X) = \Sigma_U\Sigma_X^{-1}\mathrm{cov}(\mathbf{1}\{Y\le y\},X), \end{aligned}$$

which yields the desired conclusion.

Proof of Theorem 1. Note that

$$\begin{aligned} \Sigma_{U,n}^{-1}M_n - \Sigma_U^{-1}M = {} & \Sigma_U^{-1}(\Sigma_U - \Sigma_{U,n})\Sigma_{U,n}^{-1}L(V_n - V)L^\tau + \Sigma_U^{-1}(\Sigma_U - \Sigma_{U,n})\Sigma_{U,n}^{-1}LVL^\tau \\ & + \Sigma_U^{-1}L(V_n - V)L^\tau. \end{aligned}$$

To prove Theorem 1, it suffices to investigate the convergence rate of $\|\widehat{V}_n - V\|$ and $\|\Sigma_{W,n} - \Sigma_W\|$, where $\|A\|$ is the Frobenius norm of A. We split the proof into three main steps and several sub-steps.

Step 1. We show that

$$\|V_n - V\| = o\Big(\frac{p\log n}{\sqrt{n}}\Big), \text{ almost surely.}$$

Note that V_n can be recast as a U-statistic:

$$V_n = \frac{6}{n(n-1)(n-2)}\sum_{1\le i<j<k\le n} h(w_i,y_i,w_j,y_j,w_k,y_k),$$

with the kernel

$$\begin{aligned} & h(w_1,y_1,w_2,y_2,w_3,y_3) \\ & \quad = \frac{1}{6}\big[(w_1w_2^\tau + w_2w_1^\tau)\mathbf{1}\{y_1\le y_3\}\mathbf{1}\{y_2\le y_3\} + (w_1w_3^\tau + w_3w_1^\tau)\mathbf{1}\{y_1\le y_2\}\mathbf{1}\{y_3\le y_2\} \\ & \qquad + (w_2w_3^\tau + w_3w_2^\tau)\mathbf{1}\{y_2\le y_1\}\mathbf{1}\{y_3\le y_1\}\big]. \end{aligned}$$

We approximate V_n with its Hájek projection (Serfling (1980)),

$$\widehat{V}_n = \sum_{i=1}^{n} E(V_n \mid w_i, y_i) - (n-1)E(V_n).$$

Step 1.1. We show that

$$\|V_n - \widehat{V}_n\| = o\Big(\frac{p\log n}{n}\Big) \text{ almost surely,} \tag{A.2}$$

because $V_n - \widehat{V}_n$ is itself a U-statistic that has the form

$$V_n - \widehat{V}_n = \frac{6}{n(n-1)(n-2)} \sum_{1\le i<j<k\le n} H(w_i, y_i, w_j, y_j, w_k, y_k),$$

with the symmetric kernel

$$\begin{aligned} H(w_1, y_1, w_2, y_2, w_3, y_3) = {} & h(w_1, y_1, w_2, y_2, w_3, y_3) \\ & -h_1(w_1, y_1) - h_1(w_2, y_2) - h_1(w_3, y_3) - E(V_n), \end{aligned}$$

where $h_1(w_i, y_i) = E[h(W_1, Y_1, W_2, Y_2, W_3, Y_3) \mid W_i = w_i, Y_i = y_i] - E(V_n)$. Using Lemma A in Section 5.2.1 of Serfling (1980, p.183), we have

$$\operatorname{var}(V_n - \widehat{V}_n) = \frac{18\zeta_2}{n(n-1)},$$

where $\zeta_2 = \operatorname{var}[H(W_1, Y_1, W_2, Y_2, W_3, Y_3)]$ because $V_n - \widehat{V}_n$ is a degenerated U-statistic. Next, we calculate the order of $\|\zeta_2\|$. Note that $EV_n = E\widehat{V}_n$,

$$\begin{aligned} \zeta_2 &= \operatorname{var}[H(W_1, Y_1, W_2, Y_2, W_3, Y_3)] \\ &= E[\operatorname{vec}(H(W_1, Y_1, W_2, Y_2, W_3, Y_3))\operatorname{vec}^\tau(H(W_1, Y_1, W_2, Y_2, W_3, Y_3))]. \end{aligned}$$

Therefore,

$$\begin{aligned} \|\zeta_2\| &= \big\|E[\operatorname{vec}(H(W_1, Y_1, W_2, Y_2, W_3, Y_3))\operatorname{vec}^\tau(H(W_1, Y_1, W_2, Y_2, W_3, Y_3))]\big\| \\ &\le \big\|E[\operatorname{vec}^\tau(h(W_1, Y_1, W_2, Y_2, W_3, Y_3) - E(V_n)) \\ &\qquad \operatorname{vec}(h(W_1, Y_1, W_2, Y_2, W_3, Y_3) - E(V_n))]\big\| \\ &\quad +9\big\|E[\operatorname{vec}^\tau(h_1(W_1, Y_1)\operatorname{vec}(h_1(W_1, Y_1)]\big\|. \end{aligned}$$

Using the assumption that $\max\limits_{1\le i\le p} E(W_i^4) < \infty$ uniformly in p, we can show without difficulty that $\|\zeta_2\| \le Cp\sum_{i=1}^{p} E(W_i^4) = O(p^2)$, which ensures that

$$\big\|\operatorname{var}(V_n - \widehat{V}_n)\big\| = O\Big(\frac{p^2}{n^2}\Big).$$

Take $\lambda_n = (p\log n/n)^{-1}$. It suffices to prove that, for any $\varepsilon > 0$,

$$P(\limsup \lambda_n \|V_n - \widehat{V}_n\| > \varepsilon) = 0.$$

Let $\varepsilon > 0$ be given by the Borel-Cantelli lemma. As λ_n is nondecreasing for a large n, it suffices to show that

$$\sum_{k=0}^{\infty} P\Big(\lambda_{2^{k+1}} \max_{2^k \le n \le 2^{k+1}} \|V_n - \widehat{V}_n\| \ge \varepsilon\Big) < \infty. \tag{A.3}$$

Observe that $V_n - \widehat{V}_n$ is a reverse martingale. By applying the standard result in Loeve (1978, Sec. 32), we have

$$P\Big(\sup_{j\ge n} \|V_j - \widehat{V}_j\| > \varepsilon\Big) \le \varepsilon^{-2} E\|V_n - \widehat{V}_n\|^2.$$

The right-hand side of the k-th term of $(A.3)$ can thus be bounded by

$$\begin{aligned}\varepsilon^{-2}\lambda_{2^{k+1}}^2 E\|V_{2^k} - \widehat{V}_{2^k}\|^2 &= \varepsilon^{-2} p_{2^{k+1}}^{-2} 2^{2(k+1)} (\log 2^{2(k+1)})^{-2} O\Big(\frac{p_{2^k}^2}{2^{2k}}\Big)\\ &= O((k+1)^{-2}).\end{aligned}$$

The series is convergent. Thus the Borel-Cantelli lemma yields

$$\|V_n - \widehat{V}_n\| = o\Big(\frac{p\log n}{n}\Big), \text{ almost surely.}$$

Step 1.2. We show that

$$\|\widehat{V}_n - V\| = o\Big(\frac{p\log n}{\sqrt{n}}\Big) \text{ almost surely.} \tag{A.4}$$

Note that $\widehat{V}_n - V = 3n^{-1}\sum_{i=1}^n h_1(w_i, y_i)$, where

$$\begin{aligned}h_1(w_1,y_1) &= E(h(w_1,y_1,w_2,y_2,w_3,y_3) \mid w_1,y_1) - E(V_n)\\ &= 6^{-1}\big\{E[(w_1w_2^\tau + w_2w_1^\tau)\mathbf{1}\{y_1 \le y_3\}\mathbf{1}\{y_2 \le y_3\} \mid w_1,y_1] - 2E(V_n)\big\}\\ &\quad +6^{-1}\big\{E[(w_1w_3^\tau + w_3w_1^\tau)\mathbf{1}\{y_1 \le y_2\}\mathbf{1}\{Y_3 \le y_2\} \mid w_1,y_1] - 2E(V_n)\big\}\\ &\quad +6^{-1}\big\{E[(w_2w_3^\tau + w_3w_2^\tau)\mathbf{1}\{y_2 \le y_1\}\mathbf{1}\{y_3 \le y_1\} \mid w_1,y_1] - 2E(V_n)\big\}\\ &\overset{\text{def}}{=} I_1(w_1,y_1) + I_2(w_1,y_1) + I_3(w_1,y_1).\end{aligned}$$

After some algebraic calculation, we derive

$$\begin{aligned}I_1(w_1,y_1) &= I_2(w_1,y_1) = 6^{-1}\big\{w_1E[\mathbf{1}\{y_1 \le y_3\}V^\tau(y_3) \mid (w_1,y_1)]\\ &\qquad +E[\mathbf{1}\{y_1 \le y_3\}V(y_3) \mid (w_1,y_1)]w_1^\tau - 2E(V_n)\big\} \text{ and}\\ I_3(w_1,y_1) &= 3^{-1}\big\{V(y_1)V^\tau(y_1) - E(V_n)\big\}.\end{aligned}$$

Recall that $EX = E\delta = 0$, and thus $E(W) = 0$. Note that $\widehat{V}_n - V$ is approximated with an average of independent and identically distributed random variables. By Theorem 2.3.2 in Section 2.3 of Stout (1974, p.20), we have

$$\Big\|3n^{-1}\sum_{i=1}^{n} I_j(w_i, y_i)\Big\| = o\Big(\frac{p\log n}{\sqrt{n}}\Big), \quad j = 1, 2, 3.$$

Now the result of $(A.4)$ is verified, and $(A.2)$ follows from the triangle inequality.

Step 2. We show that

$$\|\Sigma_{W,n} - \Sigma_W\| = o\Big(\frac{p\log n}{\sqrt{n}}\Big) \text{ almost surely.}$$

Using similar arguments to those in Step 1.2, and invoking the assumption that $\lambda_{\max}(LL^\tau) < +\infty$ uniformly in p, we can show that

$$\begin{aligned} \|\Sigma_{W,n} - \Sigma_W\| &= o\Big(\frac{p\log n}{\sqrt{n}}\Big) \text{ almost surely and} \\ \|\Sigma_{U,n} - \Sigma_U\| &= \|L(\Sigma_{W,n} - \Sigma_W)L^\tau\| = o\Big(\frac{p\log n}{\sqrt{n}}\Big) \text{ almost surely.} \end{aligned}$$

Step 3. We prove that

$$\|L_m^{(i)} - L\| = o\Big(\frac{p\log m}{\sqrt{m}}\Big),\ i = 1, 2, \text{ almost surely.}$$

Using similar arguments to those in Step 1.2, we obtain

$$\begin{aligned} \|\Sigma_{\delta,m} - \Sigma_\delta\| &= o\Big(\frac{p\log m}{\sqrt{m}}\Big), \text{ almost surely;} \\ \|\Sigma_{(i)XW,n} - \Sigma_{XW}\| &= o\Big(\frac{p\log m}{\sqrt{m}}\Big),\ i = 1, 2, \text{ almost surely; and} \\ \|\Sigma_{(i)W,n} - \Sigma_W\| &= o\Big(\frac{p\log m}{\sqrt{m}}\Big),\ i = 1, 2, \text{ almost surely.} \end{aligned} \tag{A.5}$$

We can now turn to the asymptotic properties of $L_m^{(i)}, i = 1, 2$, respectively:

$$\begin{aligned} \|L_m^{(1)} - L\| &= \|\Sigma_{(1)XW,m}\Sigma_{(1)W,m}^{-1} - \Sigma_{XW}\Sigma_W^{-1}\| \\ &\le \|(\Sigma_{(1)XW,m} - \Sigma_{XW})\Sigma_{(1)W,m}^{-1}(\Sigma_W - \Sigma_{(1)W,m})\Sigma_W^{-1}\| \\ &\quad + \|(\Sigma_{(1)XW,m} - \Sigma_{XW})\Sigma_W^{-1}\| + \|\Sigma_{XW}\Sigma_{(1)W,m}^{-1}(\Sigma_W - \Sigma_{(1)W,m})\Sigma_W^{-1}\|. \\ \|L_m^{(2)} - L\| &= \|\Sigma_{\delta,m}\Sigma_{(2)W,m}^{-1} - \Sigma_\delta\Sigma_W^{-1}\| \\ &\le \|(\Sigma_{\delta,m} - \Sigma_\delta)\Sigma_{(2)W,m}^{-1}(\Sigma_W - \Sigma_{(2)W,m})\Sigma_W^{-1}\| + \|(\Sigma_{\delta,m} - \Sigma_\delta)\Sigma_W^{-1}\| \\ &\quad + \|\Sigma_\delta\Sigma_{(2)W,m}^{-1}(\Sigma_W - \Sigma_{(2)W,m})\Sigma_W^{-1}\|. \end{aligned}$$

With the assumptions that $\lambda_{\max}(LL^\tau) < +\infty$ and $\lambda_{\max}(\Sigma_\delta) < +\infty$, $\lambda_{\max}(\Sigma_W^{-1}) < +\infty$ uniformly in p, together with $(A.5)$, we find that both $\|L_m^{(1)} - L\|$ and $\|L_m^{(2)} - L\|$ have the rate $o(p\log m/\sqrt{m})$ almost surely.

The almost sure convergence of $\|\Sigma_{U,mn}^{-1}M_{mn}^* - \Sigma_U^{-1}M\| = o(p\log m/\sqrt{m})$ follows from the three foregoing steps.

Proof of Theorem 2. Without loss of generality, we assume that $b_i^\tau \Sigma_U b_i = 1$ for all i. Because we assume that e is orthogonal to $\mathcal{S}_{Y|X}$, by Proposition 1 we have $e^\tau \Sigma_U^{-1} M = 0$, and thus $e^\tau b_i = 0$.

Step 1. When L is known, we can first calculate the eigenvectors for the eigenvalue decomposition of M_n with respect to $\Sigma_{U,n}$: $M_n\widehat{b}_i = \widehat{\lambda}_i \Sigma_{U,n}\widehat{b}_i$ with the constraint $\widehat{b}_i^\tau \Sigma_{U,n}\widehat{b}_i = 1$.

$$
\begin{aligned}
e^\tau \widehat{b}_i &= \widehat{\lambda}_i^{-1} e^\tau \Sigma_{U,n}^{-1} M_n \widehat{b}_i = \widehat{\lambda}_i^{-1} e^\tau \left(\Sigma_{U,n}^{-1}M_n - \Sigma_U^{-1}M\right)\widehat{b}_i \\
&= \widehat{\lambda}_i^{-1} e^\tau \left(\Sigma_{U,n}^{-1}M_n - \Sigma_U^{-1}M\right) b_i + \widehat{\lambda}_i^{-1} e^\tau \left(\Sigma_{U,n}^{-1}M_n - \Sigma_U^{-1}M\right)\left(\widehat{b}_i - b_i\right) \\
&= \frac{\lambda_i - \widehat{\lambda}_i}{\widehat{\lambda}_i \lambda_i} e^\tau \left(\Sigma_{U,n}^{-1}M_n - \Sigma_U^{-1}M\right) b_i + \frac{1}{\widehat{\lambda}_i} e^\tau \left(\Sigma_{U,n}^{-1}M_n - \Sigma_U^{-1}M\right)\left(\widehat{b}_i - b_i\right) \\
&\quad + \widehat{\lambda}_i^{-1} e^\tau \left(\Sigma_{U,n}^{-1}M_n - \Sigma_U^{-1}M\right) b_i \\
&\overset{\text{def}}{=} J_1 + J_2 + J_3.
\end{aligned}
$$

We then show that for $J_1 = o_P(1/\sqrt{n})$, $J_2 = o_P(1/\sqrt{n})$, and J_3 is asymptotically normal. We split this step into three sub-steps, as follows.

Step 1.1. We show that $J_1 = o_P(1/\sqrt{n})$. Towards this end, we prove that $\|\Sigma_{U,n}^{-1}M_n - \Sigma_U^{-1}M\|$ is of order $O_P(p/n)$. Because $E(U) = 0$, the following weak convergence can be derived from Step 1 in the proof of Theorem 1 as

$$
\begin{aligned}
\|\Sigma_{U,n} - \Sigma_U - T_1\| &= \left\| n^{-2}\sum_{i=1}^n U_i \sum_{i=1}^n U_i^\tau \right\| = O_P(\frac{p}{n}), \\
\|M_n - M - T_2\| &= \|L(V_n - \widehat{V}_n)L^\tau\| = O_P(\frac{p}{n}),
\end{aligned}
$$

where $T_1 \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n U_iU_i^\tau - \Sigma_U$ and $T_2 \overset{\text{def}}{=} L(\widehat{V}_n - V)L^\tau$. Because both T_1 and T_2 are sums of the i.i.d. random variables, we have

$$
\|(\Sigma_{U,n}^{-1}M_n - \Sigma_U^{-1}M) - \Sigma_U^{-1}T_1\Sigma_U^{-1}M + \Sigma_U^{-1}T_2\| = O_P(\frac{p}{n}). \tag{A.6}
$$

From (A.6), we observe that $\Sigma_{U,n}^{-1}M_n - \Sigma_U^{-1}M$ can be replaced by $\Sigma_U^{-1}T_1\Sigma_U^{-1}M - \Sigma_U^{-1}T_2 + O_P(p/n)$. Using similar arguments as those used to prove Corollary 1 in

Zhu, Miao, and Peng (2006), we obtain

$$\sum_{i=1}^{p} |\lambda_i - \widehat{\lambda}_i| \leq \left\|\Sigma_{U,n}^{-1} M_n - \Sigma_U^{-1} M\right\|. \tag{A.7}$$

Theorem 1 and (A.7) entail that $|\lambda_i - \widehat{\lambda}_i| = o_P(1)$. Furthermore, invoking $e^T b_i = 0$, we have

$$\begin{aligned}\sqrt{n} J_1 &= \frac{\lambda_i - \widehat{\lambda}_i}{\widehat{\lambda}_i \lambda_i} \sqrt{n} e^\tau \big(\Sigma_{U,n}^{-1} M_n - \Sigma_U^{-1} M\big) b_i \\ &= \frac{\lambda_i - \widehat{\lambda}_i}{\widehat{\lambda}_i \lambda_i} \sqrt{n} e^\tau \big(\Sigma_U^{-1} T_1 \Sigma_U^{-1} M - \Sigma_U^{-1} T_2\big) + O_P(\frac{p}{n}) e^\tau b_i \\ &= \sqrt{n} e^\tau \big(\Sigma_U^{-1} T_1 \Sigma_U^{-1} M - \Sigma_U^{-1} T_2\big) o_P(1) + o_P(1).\end{aligned}$$

From the Linderberg-Feller Central Limit Theorem, we obtain that $\sqrt{n}\big(\Sigma_U^{-1} T_1 \Sigma_U^{-1} M - \Sigma_U^{-1} T_2\big) = O_P(1)$ and $\sqrt{n} J_1 = o_P(1)$.

Step 1.2. We prove $\sqrt{n} J_2 = o_P(1)$. Corollary 1 of Zhu, Miao, and Peng (2006) proved that $\|\widehat{b}_i - b_i\| = o_P(1)$. With $e^\tau e = 1$, $\|e^\tau(\widehat{b}_i - b_i)\| = o_P(1)$. Consequently,

$$\begin{aligned}\sqrt{n} J_2 &= \widehat{\lambda}_i^{-1} \sqrt{n} e^\tau \left(\Sigma_{U,n}^{-1} M_n - \Sigma_U^{-1} M\right) \left(\widehat{b}_i - b_i\right) \\ &= \frac{1}{\lambda_i + o_P(1)} \sqrt{n} e^\tau \big(\Sigma_U^{-1} T_1 \Sigma_U^{-1} M - \Sigma_U^{-1} T_2\big) o_P(1) + O_P(\frac{p}{\sqrt{n}}) o_P(1).\end{aligned}$$

Using similar arguments, we have $\sqrt{n} J_2 = o_P(1)$ for $p = o(n^{1/2})$.

Step 1.3. For the asymptotic distribution of $\sqrt{n} J_3$.

$$\begin{aligned}\sqrt{n} J_3 &= \lambda_i^{-1} \sqrt{n} e^\tau \big(\Sigma_{U,n}^{-1} M_n - \Sigma_U^{-1} M\big) b_i \\ &= \lambda_i^{-1} \sqrt{n} e^\tau \big(\Sigma_U^{-1} T_1 \Sigma_U^{-1} M - \Sigma_U^{-1} T_2\big) b_i + O_P(p/n) e^\tau b_i \\ &= \lambda_i^{-1} \sqrt{n} e^\tau \big(\Sigma_U^{-1} T_1 \Sigma_U^{-1} M - \Sigma_U^{-1} T_2\big) b_i + o_P(1) \\ &= \frac{e^\tau \Sigma_U^{-1}}{\lambda_i \sqrt{n}} \sum_{j=1}^{n} \big\{(U_j U_j^\tau - \Sigma_U) \Sigma_U^{-1} M - 6 L I_1(w_j, y_j) L^\tau - 3 L I_3(w_j, y_j) L^\tau\big\} b_i \\ &\quad + o_P(1) \\ &\overset{\text{def}}{=} \frac{1}{\sqrt{n}} \sum_{j=1}^{n} e^\tau \Sigma_U^{-1} S_j b_i + o_P(1),\end{aligned}$$

where

$$S_j = \lambda_i^{-1} \big\{(U_j U_j^\tau - \Sigma_U) \Sigma_U^{-1} M - 6 L I_1(w_j, y_j) L^\tau - 3 L I_3(w_j, y_j) L^\tau\big\} \tag{A.8}$$

and $I_1(w_j, y_j)$ and $I_3(w_j, y_j)$ are those in (A.5) of Step 1.2.

The asymptotic distribution can be obtained by checking the Linderberg-Feller conditions. Specifically, for any given $\varepsilon > 0$,

$$\begin{aligned}
n^{-1}\sum_{j=1}^{n} & E|e^\tau\Sigma_U^{-1}S_jb_i|^2 I(|e^\tau\Sigma_U^{-1}S_jb_i| > \sqrt{n}\varepsilon) \\
&= E|e^\tau\Sigma_U^{-1}S_jb_i|^2 I(|e^\tau\Sigma_U^{-1}S_jb_i| > \sqrt{n}\varepsilon) \\
&\le \left\{E(e^\tau\Sigma_U^{-1}S_jb_i)^4\right\}^{1/2} P(|e^\tau\Sigma_U^{-1}S_jb_i| > \sqrt{n}\varepsilon) \\
&= \left\{E(b_i^\tau S_j^\tau\Sigma_U^{-1}ee^\tau\Sigma_U^{-1}S_jb_i)^2\right\}^{1/2} P(|e^\tau\Sigma_U^{-1}S_jb_i| > \sqrt{n}\varepsilon) \\
&\le \left\{\lambda_{\max}^2(\Sigma_U^{-1}ee^\tau\Sigma_U^{-1})E(b_i^\tau S_j^\tau S_jb_i)^2\right\}^{1/2} P(|e^\tau\Sigma_U^{-1}S_jb_i| > \sqrt{n}\varepsilon) \\
&\le O(p^{3/2})\lambda_{\min}^{-1}(\Sigma_U)\lambda_{\max}(LL^\tau)\max_{1\le i\le p} E|W_i|^8(b_i^\tau b_i)^2 P(|e^\tau\Sigma_U^{-1}S_jb_i| > \sqrt{n}\varepsilon).
\end{aligned}$$

Note that $1 = b_i^\tau\Sigma_U b_i \ge \lambda_{\min}(\Sigma_U)b_i^\tau b_i$, and the assumption is that Σ_U are positive-definite matrices uniformly in p. We also know that $b_i^\tau b_i$ are uniformly $O(1)$ in p. Furthermore, $\max_{1\le i\le p} E(|W_i|^8) < \infty$ and $\lambda_{\max}(LL^\tau) < \infty$ uniformly for p. Appealing to the Markov inequality that entails $P(|e^\tau\Sigma_U^{-1}S_jb_i| > \sqrt{n}\varepsilon) \le E|e^\tau\Sigma_U^{-1}S_jb_i|/\sqrt{n}\varepsilon$, we derive

$$n^{-1}\sum_{j=1}^{n} E|e^\tau\Sigma_U^{-1}S_jb_i|^2 I(|e^\tau\Sigma_U^{-1}S_jb_i| > \sqrt{n}\varepsilon) = O(p^{3/2}/\sqrt{n}). \tag{A.9}$$

When $p = o(n^{1/3})$, (A.9) is $o_P(1)$. Together with (A5) that $\mathrm{var}(e^\tau\Sigma_U^{-1}S_1b_i) \to G_i$, we prove that $\{e^\tau\Sigma_U^{-1}S_jb_i, j = 1,\ldots,n\}$ satisfies the conditions of the Linderberg-Feller Central Limit Theorem.

Step 2. When L is unknown and estimated by either validation or replicated data with a sample size of $m \le n$, we can show with little difficulty that

$$\begin{aligned}
\|\Sigma_{U,mn} - \Sigma_{U,n}\| \le{}& \|(L_m^{(i)} - L)\Sigma_{W,n}(L_m^{(i)} - L)^\tau\| + 2\|(L_m^{(i)} - L)(\Sigma_{W,n} - \Sigma_W)L^\tau\| \\
&+2\|(L_m^{(i)} - L)\Sigma_W L^\tau\| = O_P(\frac{p}{m}),
\end{aligned}$$

$$\begin{aligned}
\|M_{mn}^* - M_n\| \le{}& \|(L_m^{(i)} - L)V_n(L_m^{(i)} - L)^\tau\| + 2\|(L_m^{(i)} - L)(V_n - V)L^\tau\| \\
&+2\|(L_m^{(i)} - L)VL^\tau\| = O_P(\frac{p}{m}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\Sigma_{U,mn} - \Sigma_U - T_1\| &\le \|\Sigma_{U,mn} - \Sigma_{U,n}\| + \|\Sigma_{U,n} - \Sigma_U - T_1\| = O_P(\frac{p}{m}). \\
\|M_{mn}^* - M - T_2\| &\le \|M_{mn}^* - M_n\| + \|M_n - M - T_2\| = O_P(\frac{p}{m}).
\end{aligned}$$

$$\|(\Sigma_{U,mn}^{-1} M_n^* - \Sigma_U^{-1} M) - \Sigma_U^{-1} T_1 \Sigma_U^{-1} M + \Sigma_U^{-1} T_2\| = O_P(\frac{p}{m}).$$

The rest of the proof of the second part of Theorem 2 is similar to Step 1. We omit the details here and emphasize that the optimal rate we can achieve is $p = o(m^{1/3})$.

Proof of Theorems 3 and 4. Note that $U = LW = L\Gamma X + L\delta$ and $\delta \perp\!\!\!\perp (X, Y)$. We have

$$\begin{aligned}
\text{var}[UI\{Y \le y\}] &= \text{var}[L\Gamma XI\{Y \le y\} + L\delta I\{Y \le y\}] \\
&= L\Gamma \text{var}[XI\{Y \le y\}]\Gamma^\tau L^\tau + L\Sigma_\delta L^\tau F(y) \\
&= L\Gamma \left\{\text{var}[XI\{Y \le y\}] - \Sigma_X F(y)\right\} \Gamma^\tau L^\tau + L\left(\Gamma\Sigma_X\Gamma^\tau + \Sigma_U\right) L^\tau F(y) \\
&= L\Gamma \left\{\text{var}[XI\{Y \le y\}] - \Sigma_X F(y)\right\} \Gamma^\tau L^\tau + L\Sigma_W L^\tau F(y).
\end{aligned}$$

The last equality follows from the fact that $\Sigma_W = \Gamma\Sigma_X\Gamma^\tau + \Sigma_\delta$. Following the arguments for Theorem 4 of Zhu, Zhu, and Feng (2010) using the linearity and constant variance conditions, we can prove without much difficulty that $\Sigma_X^{-1}\left\{\text{var}[XI\{Y \le Y\}] - \Sigma_X F(y)\right\} \subseteq \mathcal{S}_{Y|X}$. With (A.1), we have $\Sigma_U^{-1}\Lambda^* = \Sigma_X^{-1}\left\{\text{var}[XI\{Y \le Y\}] - \Sigma_X F(y)\right\}\Gamma^\tau L^\tau \subseteq \mathcal{S}_{Y|X}$.

The proof of Theorem 4 can be completed similarly, we omit the details.

References

Candés, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n (with discussion). *Ann. Statist.* **35**, 2313-2351.

Carroll, R. J. and Li, K. C. (1992). Measurement error regression with unknown link: Dimension reduction and data visualization. *J. Amer. Statist. Assoc.* **87**, 1040-1050.

Carroll, R. J., Ruppert, D., Stefanski L. A. and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective.* Second Edition, Chapman and Hall, London.

Cook, R. D. (1998a). *Regression Graphics: Ideas for Studying Regressions Through Graphics.* Wiley & Sons, New York.

Cook, R. D. (1998b). Principal Hessian directions revisited. *J. Amer. Statist. Assoc.* **93**, 84-94.

Cook, R. D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *Ann. Statist.* **30**, 455-474.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *J. Roy. Statist. Soc. Ser. B* **70**, 849-911.

Férre, L. (1998). Determining the dimension in sliced inverse regression and related methods, *J. Amer. Statist. Assoc.* **93**, 132-140.

Hall, P. and Li, K. C. (1993). On almost linearity of low dimensional projection from high dimensional data. *Ann. Statist.* **21**, 867-889.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *J. Amer. Statist. Assoc.* **102**, 997-1008.

Li, B. and Yin, X. (2007). On surrogate dimension reduction for measurement error regression: An invariance law. *Ann. Statist.* **35**, 2143-2172.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86** 316-342.

Loeve, M. (1978). *Probability Theory II.* The Fourth Edition, Springer-Verlag, New York.

Lue, H. H. (2004). Principal Hessian directions for regresson with measurement error. *Biometrika* **91**, 409-423.

Pan, W. H., Bai, C. H., Chen, J. R. and Chin, H. C. (1997). Associations between carotid atherosclerosis and high factor VIII activity, dyslipidemia, and hypertension. *Stroke.* **28**, 88-94.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* Wiley, New York.

Stout, W. F. (1974). *Almost Sure Convergence.* Academic Press, New York, San Francisco, London.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Zhu, L. P. and Zhu, L. X. (2009). On distribution weighted partial least squares with diverging number of highly correlated predictors. *J. Roy. Statist. Soc. Ser. B* **71**, 525-548.

Zhu, L. P., Zhu, L. X. and Feng, Z. H. (2010). Dimension reduction in regressions through cumulative slicing estimation. *J. Amer. Statist. Assoc.* **105**, 1455-1466.

Zhu, L. X. and Fang, K. T. (1996). Asymptotics for the kernel estimates of sliced inverse regression. *Ann. Statist.* **24**, 1053-1067.

Zhu, L. X., Miao, B. Q. and Peng, H. (2006). Sliced inverse regression with large dimensional covariates. *J. Amer. Statist. Assoc.* **101**, 630-643.

Shen Zhen-Hong Kong Joint Research Center for Applied Statistical Sciences; College of Mathematics and Computational Science; Institute of Statistical Sciences at Shenzhen University; Shenzhen University, Shenzhen, P.R. China.

E-mail: zhangjunstat@gmail.com

School of Statistics and Management, Shanghai University of Finance and Economics, and the Key Laboratory of Mathematical Economics (SUFE), Ministry of Education, Shanghai, P. R. China.

E-mail: zhu.liping@mail.shufe.edu.cn

Department of Mathematics, Hong Kong Baptist University, Hong Kong, P. R. China.

E-mail: lzhu@hkbu.edu.hk

(Received March 2012; accepted September 2013)