Biometrika (2013), **100**, 2, *pp.* 371–383
doi: 10.1093/biomet/ass075
Advance Access publication 24 February 2013

Efficiency loss and the linearity condition in dimension reduction

By YANYUAN MA

Department of Statistics, Texas A&M University, College Station Texas 77843, U.S.A.
ma@stat.tamu.edu

and LIPING ZHU

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China
zhu.liping@mail.shufe.edu.cn

Summary

Linearity, sometimes jointly with constant variance, is routinely assumed in the context of sufficient dimension reduction. It is well understood that, when these conditions do not hold, blindly using them may lead to inconsistency in estimating the central subspace and the central mean subspace. Surprisingly, we discover that even if these conditions do hold, using them will bring efficiency loss. This paradoxical phenomenon is illustrated through sliced inverse regression and principal Hessian directions. The efficiency loss also applies to other dimension reduction procedures. We explain this empirical discovery by theoretical investigation.

Some key words: Constant variance condition; Dimension reduction; Estimating equation; Inverse regression; Linearity condition; Semiparametric efficiency.

1. Introduction

In the sufficient dimension reduction literature, two essential conditions are linearity and constant variance. Denote X the p-dimensional random covariate vector, and let the dimension reduction subspace be the column space of a full rank $p \times d$ matrix β. The linearity condition assumes $E(X \mid \beta^{\mathrm{T}} X) = PX$, where $P = \Sigma\beta(\beta^{\mathrm{T}}\Sigma\beta)^{-1}\beta^{\mathrm{T}}$ is a $p \times p$ matrix and $\Sigma = \mathrm{cov}(X)$. The constant variance condition assumes $\mathrm{cov}(X \mid \beta^{\mathrm{T}} X) = Q$, where $Q = \Sigma - P\Sigma P^{\mathrm{T}}$. These two conditions have played a central role throughout the development of the sufficient dimension reduction literature. For example, the linearity condition, sometimes jointly with the constant variance condition, permitted the establishment of sliced inverse regression (Li, 1991), sliced average variance estimation (Cook & Weisberg, 1991), directional regression (Li & Wang, 2007), cumulative slicing estimation (Zhu et al., 2010a), discretization-expectation estimation (Zhu et al., 2010b), ordinary least squares (Li & Duan, 1989), and principal Hessian directions (Li, 1992; Cook & Li, 2002). It is no exaggeration to call linearity and constant variance the fundamental conditions of dimension reduction.

It is a different story regarding the validity of these conditions and their verification in practice. Hall & Li (1993) showed that the linearity condition would hold in an asymptotic sense when p goes to infinity. Yet whether the asymptotically true result suffices for a finite-dimensional problem remains unclear. This has prompted researchers to relax these conditions. For example,

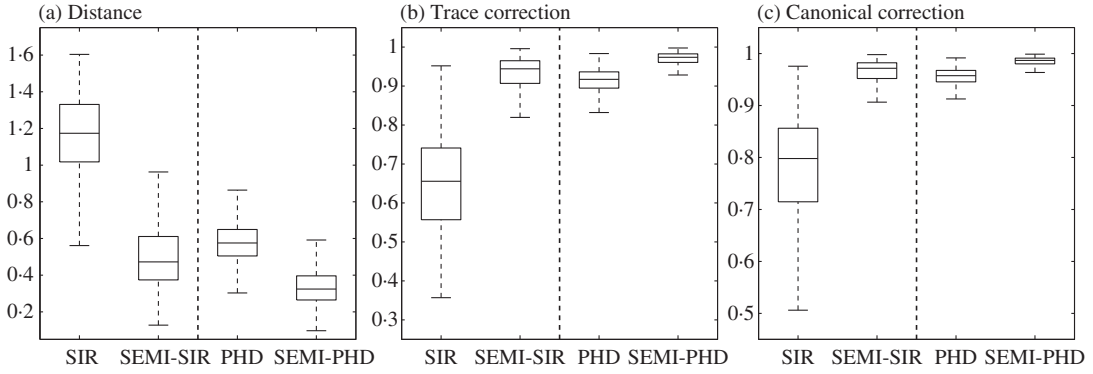

Fig. 1. Comparison of sliced inverse regression, SIR, and principal Hessian directions, PHD, with their semiparametric counterparts, SEMI-SIR, SEMI-PHD, in Model (I) (left half in each panel) and Model (II) (right half in each panel). Results are based on 1000 simulated datasets with sample size $n = 200$.

Li & Dong (2009) and Dong & Li (2010) replaced the linearity condition by a polynomial condition. Ma & Zhu (2012) completely eliminated both conditions.

Following the relaxation of the linearity and constant variance conditions, a natural question arises: what do we lose by ignoring these conditions when they hold? It is natural to conjecture that this will cause estimation variance inflation. However, our discovery is exactly the opposite.

We illustrate this paradoxical phenomenon empirically. Consider

$$\text{Model I}: Y = \beta_1^{\mathrm{T}} X / \{0{\cdot}5 + (\beta_2^{\mathrm{T}} X + 1{\cdot}5)^2\} + \varepsilon,$$
$$\text{Model II}: Y = (\beta_1^{\mathrm{T}} X)^2 + (\beta_2^{\mathrm{T}} X)^2 + \varepsilon,$$

where $\beta_1 = 10^{-1/2}(1, 1, 1, 1, 1, 1, 1, 1, 1, 1)^{\mathrm{T}}$, $\beta_2 = 10^{-1/2}(1, -1, 1, -1, 1, -1, 1, -1, 1, -1)^{\mathrm{T}}$ and ε is a standard normal random variable. Thus, $p = 10$, $d = 2$ and $\beta = (\beta_1, \beta_2)$ in these models. We generate X from a multivariate standard normal distribution. Thus both the linearity and the constant variance conditions hold. For Model I, we implement classical sliced inverse regression and its semiparametric counterpart where the linearity condition is not used. For Model II, we compare classical principal Hessian directions and its semiparametric counterpart where neither condition is used. Here, the sliced inverse regression and principal Hessian directions are identical to their semiparametric counterparts, except that the sliced inverse regression and principal Hessian directions utilize the linearity and the constant variance conditions to obtain $E(X \mid \beta^{\mathrm{T}} X)$ and $\mathrm{cov}(X \mid \beta^{\mathrm{T}} X)$, while their semiparametric counterparts estimate $E(X \mid \beta^{\mathrm{T}} X)$ and $\mathrm{cov}(X \mid \beta^{\mathrm{T}} X)$ nonparametrically. See Ma & Zhu (2012) for details on these semiparametric estimators. We generate 1000 datasets each of size $n = 200$, and summarize the results in Fig. 1. To make a fair comparison, we estimate the kernel matrix of the classical sliced inverse regression by using kernel smoothing rather than the usual slicing estimation. This allows us to avoid selecting the number of slices, which usually adversely affects the performance. Thus, sliced inverse regression is implemented in its improved form.

Figure 1(a) contains the boxplots of the four estimation procedure results, measured by a distance between the estimated and the true dimension reduction subspaces. This distance is defined as the Frobenius norm of $\hat{P} - P$, where P is as defined before, $\hat{P} = \Sigma\hat{\beta}(\hat{\beta}^{\mathrm{T}}\Sigma\hat{\beta})^{-1}\hat{\beta}^{\mathrm{T}}$ and $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)$ is obtained from the aforementioned estimation procedures. This distance criterion is widely used to evaluate the performance of different estimation procedures, with a smaller distance indicating better estimation of the dimension reduction subspace. Figure 1(a) shows clearly that the semiparametric counterparts outperform their classical versions. Thus, not

taking advantage of the linearity condition or the constant variance condition, although both are satisfied, seems to bring a gain in estimating the dimension reduction subspaces.

Figure 1(b) contains the boxplots of the same results measured by the trace correlation, defined as $\text{trace}(P\hat{P})/d$. A larger value of this criterion indicates better performance. Figure 1(b) demonstrates again that the semiparametric counterparts outperform their classical versions, once again indicating that not taking advantage of linearity or constant variance, even though both hold, brings a gain.

Finally, Fig. 1(c) shows the results for yet another popular criterion, the canonical correlation, defined as the average of d canonical correlation coefficients between $\hat{\beta}^{\mathrm{T}}X$ and $\beta^{\mathrm{T}}X$. Under this measure, larger values indicate better estimation results, and the conclusion from Fig. 1(c) is consistent with those from Fig. 1(a) and Fig. 1(b).

Having observed these unexpected results, our goal here is to demonstrate that the improvement is not an accident, but is theoretically verifiable. Because it is already well understood that using linearity and constant variance when they do not hold causes bias, here we consider exclusively the case when the covariate vector satisfies the linearity condition and, if required, the constant variance condition. Thus the aforementioned original methods, such as sliced inverse regression and principal Hessian directions, are valid and will provide consistent estimation. Although this is the classical setting of the dimension reduction literature and seems well understood, we will formally establish that if we ignore the linearity and the constant variance conditions, and instead, estimate the conditional expectation $E(X \mid \beta^{\mathrm{T}}X)$ or more generally $E\{a(X) \mid \beta^{\mathrm{T}}X\}$ nonparametrically, then the performance of sufficient dimension reduction methods will improve. The improvement is in the asymptotic variance of the estimated dimension reduction subspace, in that not using linearity or constant variance will yield a more efficient estimator than using them, even when they are true.

2. Some preliminaries

We first lay out the central subspace and the central mean subspace models and some notation we use throughout. Let Y be a univariate response variable and let X and β be defined as in § 1. Throughout this paper we assume that X satisfies both the linearity and the constant variance conditions. Using an invariance property (Cook, 1998, p. 106), we assume without loss of generality that $E(X) = 0$ and $\text{cov}(X) = I_p$. The essence of the sufficient dimension reduction literature is to assume that Y depends on X only through a few linear combinations $\beta^{\mathrm{T}}X$, and to identify the space spanned by the columns of β. Mainly two types of links between Y and X are commonly studied, the conditional distribution and the conditional mean function. In the first model type (Li, 1991; Cook, 1998), one assumes that

$$F(y \mid X) = F(y \mid \beta^{\mathrm{T}}X), \quad y \in \mathbb{R}, \tag{1}$$

where $F(y \mid X) = \text{pr}(Y \leqslant y \mid X)$ denotes the conditional distribution function of Y given X. The smallest column space of β satisfying (1) is called the central subspace, $\mathcal{S}_{Y|X}$. In the second model type (Cook & Li, 2002), one assumes that the conditional mean function satisfies

$$E(Y \mid X) = E(Y \mid \beta^{\mathrm{T}}X). \tag{2}$$

The corresponding smallest column space of β is called the central mean subspace $\mathcal{S}_{E(Y|X)}$. Estimation of $\mathcal{S}_{Y|X}$ and $\mathcal{S}_{E(Y|X)}$ is the main purpose of sufficient dimension reduction. In the following development, we focus on the classical sliced inverse regression and principal Hessian

directions methods as representative estimators for $\mathcal{S}_{Y|X}$ and $\mathcal{S}_{E(Y|X)}$ respectively, although the conclusion applies to other sufficient dimension reduction methods.

To further ensure that the identifiability of $\mathcal{S}_{Y|X}$ or $\mathcal{S}_{E(Y|X)}$ implies the identifiability of β, we require the upper $d \times d$ submatrix of β to be the identity matrix. Through this parameterization, estimating $\mathcal{S}_{Y|X}$ or $\mathcal{S}_{E(Y|X)}$ is equivalent to estimating the lower $(p-d) \times d$ submatrix in β. This submatrix contains all the unknown parameters involved in the estimation of $\mathcal{S}_{Y|X}$ or $\mathcal{S}_{E(Y|X)}$ and uniquely defines the corresponding space $\mathcal{S}_{Y|X}$ or $\mathcal{S}_{E(Y|X)}$. This particular parameterization is simple and enables us to study the properties of the space estimation by studying those of estimators of the parameters in β. Other parameterizations can also be used.

We further introduce two matrix operators. We use $\mathrm{vecl}(\beta)$ to denote the length $(p-d)d$ vector formed by the concatenation of the columns in the lower $(p-d) \times d$ submatrix of β, and use $\mathrm{vec}(M)$ to denote the concatenation of the columns of an arbitrary matrix M.

3. THEORETICAL EXPLANATION OF THE PARADOXICAL PHENOMENON

Following Ma & Zhu (2012), to identify $\mathcal{S}_{Y|X}$ in model (1), the sliced inverse regression solves for the parameters contained in β from an estimating equation based on the relation

$$E[E(X \mid Y)\{X^{\mathrm{T}} - E(X^{\mathrm{T}} \mid \beta^{\mathrm{T}} X)\}] = 0. \tag{3}$$

When the linearity condition holds, (3) simplifies to

$$QE\{E(X \mid Y)X^{\mathrm{T}}\} = Q\,\mathrm{cov}\{E(X \mid Y)\} = 0,$$

where $Q = I_p - P = I_p - \beta(\beta^{\mathrm{T}}\beta)\beta^{\mathrm{T}}$. Consequently, solving (3) is equivalent to calculating the eigenspace of the matrix $\mathrm{cov}\{E(X \mid Y)\}$ associated with its d nonzero eigenvalues.

Similarly, to identify $\mathcal{S}_{E(Y|X)}$ in model (2), the principal Hessian directions method solves for the parameters contained in β from an estimating equation based on the relation

$$E[\{Y - E(Y)\}\{XX^{\mathrm{T}} - E(XX^{\mathrm{T}} \mid \beta^{\mathrm{T}} X)\}] = 0. \tag{4}$$

When both the linearity and the constant variance conditions hold, (4) simplifies to

$$E[\{Y - E(Y)\}XX^{\mathrm{T}}] = PE[\{Y - E(Y)\}XX^{\mathrm{T}}]P.$$

Thus, solving (4) is equivalent to computing the eigenspace of the matrix $E[\{Y - E(Y)\}XX^{\mathrm{T}}]$ associated with its d nonzero eigenvalues.

To simultaneously consider both the sliced inverse regression in (3) and the principal Hessian directions in (4), we consider a unified form

$$E[g^*(Y)\{a^*(X) - E(a^* \mid \beta^{\mathrm{T}} X)\}] = 0, \tag{5}$$

where g^* is a fixed function of Y that satisfies $E(g^* \mid X) = E(g^* \mid \beta^{\mathrm{T}} X)$, and a^* is a fixed function of X. Clearly (5) contains both sliced inverse regression and principal Hessian directions as special cases, by choosing $g^* = E(X \mid Y)$ and $a^* = X^{\mathrm{T}}$ to yield (3) or $g^* = Y - E(Y)$ and $a^* = XX^{\mathrm{T}}$ to yield (4). Denote the observations by (X_i, Y_i) $(i = 1, \ldots, n)$. To facilitate our subsequent inference procedure, we perform the following operations. First we vectorize the sample version of (5), then we use a generalized method of moment argument to reduce the number of

estimating equations to $(p-d)d$, then finally we simplify these estimation equations to obtain

$$\sum_{i=1}^{n}\sum_{j=1}^{q}[g_j(Y_i)\{a_j(X_i) - E(a_j \mid \beta^{\mathrm{T}}X_i)\}] = 0, \tag{6}$$

where $g_j(Y)$ is a scalar or column vector that satisfies $E(g_j \mid X) = E(g_j \mid \beta^{\mathrm{T}}X)$, and a_j is a row vector, for $j = 1, \ldots, q$. A more detailed description of how to obtain (6) from (5) is given in the Appendix. We also assume that (6) has a unique solution when $n \to \infty$. As (6) is equivalent to (5) when $n \to \infty$, we can view (6) as a compact expression of the sample version of (5).

We now study the asymptotic properties of the estimating equation (6), both when the $E(a_j \mid \beta^{\mathrm{T}}X)$s are known and when they are estimated nonparametrically. The analysis of (6) for $q > 1$ can be readily obtained after we study (6) for $q = 1$, so in the sequel, whenever appropriate, we shall focus on the case $q = 1$ in (6) and ignore the subscript j.

When we decide to give up using any properties of the covariate X such as linearity or constant variance, we need to estimate $E(a \mid \beta^{\mathrm{T}}X)$ nonparametrically. For example, we can estimate $E(a \mid \beta^{\mathrm{T}}X)$ through

$$\hat{E}(a \mid \beta^{\mathrm{T}}X) = \sum_{i=1}^{n} K_h(\beta^{\mathrm{T}}X_i - \beta^{\mathrm{T}}X)a(X_i) \Big/ \sum_{i=1}^{n} K_h(\beta^{\mathrm{T}}X_i - \beta^{\mathrm{T}}X)\ .$$

Here $K_h(\cdot) = K(\cdot/h)/h$ is a kernel function and h is a bandwidth which can be estimated by leave-one-out crossvalidation. Replacing $E(a \mid \beta^{\mathrm{T}}X)$ in (6) with $\hat{E}(a \mid \beta^{\mathrm{T}}X)$, we obtain an estimator $\hat{\beta}$ by solving the equation

$$\sum_{i=1}^{n} g(Y_i)\{a(X_i) - \hat{E}(a \mid \beta^{\mathrm{T}}X_i)\} = 0. \tag{7}$$

Crossvalidation is one possible way of selecting h, and its validity in the nonparametric context can be found in Härdle et al. (1988). In the semiparametric context, the final estimate is insensitive to the bandwidth choice. This is reflected in Condition A4, see the Appendix, where a range of bandwidths are allowed, all of which will yield the same first-order asymptotic properties for estimation of β. In terms of finite sample performance, bandwidth has also been observed to have low impact, see for example Maity et al. (2007). Theorem 1 states the asymptotic properties of $\hat{\beta}$.

THEOREM 1. *Under the regularity conditions given in the Appendix, $\hat{\beta}$ satisfies*

$$-n^{1/2}A\,\mathrm{vecl}(\hat{\beta} - \beta) = n^{-1/2}\sum_{i=1}^{n}\mathrm{vec}[\{g(Y_i) - E(g \mid \beta^{\mathrm{T}}X_i)\}\{a(X_i) - E(a \mid \beta^{\mathrm{T}}X_i)\}] + o_p(1),$$

where

$$A = E\left(\frac{\partial\,\mathrm{vec}[\{g(Y) - E(g \mid \beta^{\mathrm{T}}X)\}\{a(X) - E(a \mid \beta^{\mathrm{T}}X)\}]}{\partial\,\mathrm{vecl}(\beta)^{\mathrm{T}}}\right).$$

Hence, when $n \to \infty$, $n^{1/2}\,\mathrm{vecl}(\hat{\beta} - \beta) \to \mathcal{N}\{0, A^{-1}B_1(A^{-1})^{\mathrm{T}}\}$ in distribution, where

$$B_1 = \mathrm{cov}(\mathrm{vec}[\{g(Y) - E(g \mid \beta^{\mathrm{T}}X)\}\{a(X) - E(a \mid \beta^{\mathrm{T}}X)\}]).$$

Remark 1. We consider only the situation where g^* and hence g are fixed functions. In practice, sometimes g^* and hence g are estimated, thus more careful notation is $\hat{g}^*$ and $\hat{g}$. However, as long as the estimated function $\hat{g}$ converges to g at a proper rate (Mack & Silverman, 1982), the first-order result in Theorem 1 stays unchanged if we replace g with $\hat{g}$.

Alternatively, $E(a^* \mid \beta^{\mathrm{T}}X)$ may have a known form, say $E(a^* \mid \beta^{\mathrm{T}}X) = h^*(\beta^{\mathrm{T}}X, \beta)$, where $h^*(\cdot)$ is a known function. This will further yield a known form for $E(a_j \mid \beta^{\mathrm{T}}X)$ in (6), which we denote $h_j(\beta^{\mathrm{T}}X, \beta)$ $(j = 1, \ldots, q)$. For example, under the linearity condition, for $a^*(X) = X^{\mathrm{T}}$, $h^*(\beta^{\mathrm{T}}X, \beta) = PX = \beta^{\mathrm{T}}(\beta^{\mathrm{T}}\beta)^{-1}(\beta^{\mathrm{T}}X)$; under both the linearity and the constant variance conditions, for $a^*(X) = XX^{\mathrm{T}}$, $h^*(\beta^{\mathrm{T}}X, \beta) = Q + PXX^{\mathrm{T}}P = I_p - \beta(\beta^{\mathrm{T}}\beta)^{-1}\beta^{\mathrm{T}} + \beta(\beta^{\mathrm{T}}\beta)^{-1}(\beta^{\mathrm{T}}X)(\beta^{\mathrm{T}}X)^{\mathrm{T}}(\beta^{\mathrm{T}}\beta)^{-1}\beta$. This allows us to solve a simplified version of (6). For $q = 1$ and ignoring the subscript j, we need to solve only

$$\sum_{i=1}^{n} g(Y_i)\{a(X_i) - h(\beta^{\mathrm{T}}X_i, \beta)\} = 0 \tag{8}$$

to obtain an estimator $\tilde{\beta}$. The asymptotic properties of $\tilde{\beta}$ are given in Theorem 2.

THEOREM 2. *If h is differentiable with respect to β, then $\tilde{\beta}$ satisfies*

$$-n^{1/2}A\,\mathrm{vecl}(\tilde{\beta} - \beta) = n^{-1/2}\sum_{i=1}^{n}\mathrm{vec}[g(Y_i)\{a(X_i) - E(a \mid \beta^{\mathrm{T}}X_i)\}] + o_p(1),$$

where A is given in Theorem 1. Hence, when $n \to \infty$, $n^{1/2}\,\mathrm{vecl}(\tilde{\beta} - \beta) \to \mathcal{N}\{0, A^{-1}B_2(A^{-1})^{\mathrm{T}}\}$ in distribution, where

$$B_2 = \mathrm{cov}(\mathrm{vec}[g(Y)\{a(X) - E(a \mid \beta^{\mathrm{T}}X)\}]).$$

Comparison of B_1 and B_2 reveals the difference between $\hat{\beta}$ from solving (7) and $\tilde{\beta}$ from solving (8), as stated in Proposition 1.

PROPOSITION 1. *Under the conditions in Theorems 1 and 2, $n[\mathrm{cov}\{\mathrm{vecl}(\tilde{\beta})\} - \mathrm{cov}\{\mathrm{vecl}(\hat{\beta})\}]$ is positive definite when $n \to \infty$.*

Combining the results in Theorems 1, 2 and Proposition 1, we are now ready to state our main results, in Theorem 3. Its proof combines that of Theorems 1, 2 and Proposition 1, and is omitted to avoid redundancy.

THEOREM 3. *Let $\tilde{\beta}$ and $\hat{\beta}$ solve (6) with $E(a_j \mid \beta^{\mathrm{T}}X)$ replaced by $h_j(\beta^{\mathrm{T}}X, \beta)$ and by its nonparametric kernel estimator $\hat{E}(a_j \mid \beta^{\mathrm{T}}X)$ respectively. Under the regularity conditions given in the Appendix, $n[\mathrm{cov}\{\mathrm{vecl}(\tilde{\beta})\} - \mathrm{cov}\{\mathrm{vecl}(\hat{\beta})\}]$ is positive definite when $n \to \infty$.*

We emphasize that $\tilde{\beta}$ and $\hat{\beta}$ solve the same estimating equation (6), except that $\tilde{\beta}$ takes advantage of the known forms of $E(a_j \mid \beta^{\mathrm{T}}X)$ while $\hat{\beta}$ does not. They are estimators of the same parameter β. Therefore, Theorem 3 states the interesting result that by giving up using the linearity and constant variance conditions, we enjoy a decreased estimation variance.

In estimating the central subspace $\mathcal{S}_{Y|X}$, any function of Y is a qualified g^* function, because $E(g^* \mid X) = E(g^* \mid \beta^{\mathrm{T}}X)$ by the model assumption (1). Specifically, choose $g^*(Y) = E(X \mid Y)$, $a^*(X) = X^{\mathrm{T}}$, and $h^*(\beta^{\mathrm{T}}X, \beta) = \beta^{\mathrm{T}}(\beta^{\mathrm{T}}\beta)^{-1}(\beta^{\mathrm{T}}X)$. After vectorizing and using a generalized method of moments to reduce the number of estimating equations, the over-identified

estimating equation $\sum_{i=1}^{n} g^*(Y_i)\{a^*(X_i) - E(a^* \mid \beta^{\mathrm{T}} X_i)\} = 0$ reduces to (6). Theorem 3 then directly shows that giving up using the h_j functions, hence giving up the linearity condition, will reduce the variance of sliced inverse regression. In estimating the central mean subspace $\mathcal{S}_{E(Y|X)}$ defined in (2), $Y - c$ is the only qualified g^* function, where c is an arbitrary constant. Specifically, choose $g^*(Y) = Y - E(Y)$, $a^*(X) = XX^{\mathrm{T}}$, $h^*(\beta^{\mathrm{T}} X, \beta) = I_p - \beta(\beta^{\mathrm{T}}\beta)^{-1}\beta^{\mathrm{T}} + \beta(\beta^{\mathrm{T}}\beta)^{-1}(\beta^{\mathrm{T}}X)(\beta^{\mathrm{T}}X)^{\mathrm{T}}(\beta^{\mathrm{T}}\beta)^{-1}\beta$. Vectorizing the estimating equations and using the generalized method of moments to reduce the number of estimating equations will again yield an estimating equation of the form (6). Theorem 3 shows that giving up the linearity and constant variance conditions will reduce the variance of principal Hessian directions.

Ma & Zhu (2012) studied many other forms of the sufficient dimension reduction estimators that use the linearity and constant variance conditions. Those estimators can all be written in the form (6). Following Theorem 3, these estimators suffer the same efficiency loss as sliced inverse regression and principal Hessian directions, in that their estimation variance can be decreased by nonparametric estimation of $E(a_j \mid \beta^{\mathrm{T}} X)$. Since we work in the semiparametric framework, our analysis does not apply when Y is categorical and X is multivariate normal given Y. Under these two conditions, the model is parametric and the sliced inverse regression is the maximum likelihood estimator and cannot be further improved.

To keep the vital information simple and clear, we have presented the inverse regression methods in their ideal forms, where the knowledge $E(X) = 0$ and $\mathrm{cov}(X) = I$ is directly incorporated into estimation. In practice, one might need to replace $E(X)$ with $\bar{X}$ and $\mathrm{cov}(X)$ with $\hat{\mathrm{cov}}(X)$, and proceed with the estimation. Denote the resulting estimator by $\check{\beta}$. However, the estimation of $E(X)$ and $\mathrm{cov}(X)$ does not recover the efficiency loss caused by using the linearity and constant variance conditions. In other words, $n[\mathrm{cov}\{\mathrm{vecl}(\check{\beta})\} - \mathrm{cov}\{\mathrm{vecl}(\hat{\beta})\}]$ is still a positive definite matrix. We omit the proof because it is very similar to the proofs of Theorems 1–3 and Proposition 1.

4. Discussion

The surprising discovery that the linearity and constant variance conditions cause efficiency loss reminds us of the situation widely experienced in using the inverse probability weighting idea to handle missing covariates. There it is well known that using the true weights yields a less efficient estimator than using the estimated weights. Such a phenomenon has been well studied in Henmi & Eguchi (2004), and a nice geometrical explanation was provided there. However, our problem shows several important differences. First, the efficiency improvement in the inverse probability weighting scheme can be obtained through any valid parametric estimation of the weights, while the efficiency improvement is not guaranteed when we view the linearity condition as a truth, and replace it with an arbitrary valid parametric estimation. This makes the geometric explanation in Henmi & Eguchi (2004) invalid in our context. Second, our efficiency gain is achieved through replacing the linearity condition with nonparametric estimation, while only parametric estimation is considered in Henmi & Eguchi (2004).

Finally, having focused on the drawback of using the linearity and the constant variance conditions when they do hold, we acknowledge that using these conditions does ease the computation. When these conditions hold, a nonparametric estimation procedure can be avoided, leading to less programming effort and fast computation. However, nonparametric estimation is more or less routine in modern statistics, while the linearity and constant variance conditions remain uncheckable. As estimation is the final goal and this is better achieved without using the linearity or constant variance conditions, whether or not they hold, one would expect that giving up computational convenience for better statistical results can be sensible.

ACKNOWLEDGEMENT

Ma was supported by the National Science Foundation and National Institute of Neurological Disorders and Stroke, U.S.A. Zhu, the corresponding author, was supported by the Natural Science Foundation of China, Program for New Century Excellent Talents in University, the Innovation Program of Shanghai Municipal Education Commission and the Pujiang Project of the Science and Technology Commission of Shanghai Municipality. Zhu is also affiliated with the Key Laboratory of Mathematical Economics at Shanghai University of Finance and Economics.

APPENDIX

Obtaining (6) *from* (5)

We first vectorize the estimating function in (6) to obtain

$$f(Y, X, \beta^{\mathrm{T}}X) = \begin{pmatrix} g^*(Y)[a_1^*(X) - E\{a_1^*(X) \mid \beta^{\mathrm{T}}X\}] \\ \vdots \\ g^*(Y)[a_l^*(X) - E\{a_l^*(X) \mid \beta^{\mathrm{T}}X\}] \end{pmatrix}. \tag{A1}$$

Here we assume a^* contains l columns, denoted $a_1^*, \ldots, a_l^*$. Assume $g^*(Y)a_1^*(X)$ contains l' rows. We then perform the generalized method of moments step to obtain

$$Df(Y, X, \beta^{\mathrm{T}}X) = \sum_{j=1}^{l} D_j g^*(Y)[a_j^*(X) - E\{a_j^*(X) \mid \beta^{\mathrm{T}}X\}] = \sum_{j=1}^{l} g_j(Y)[a_j^*(X) - E\{a_j^*(X) \mid \beta^{\mathrm{T}}X\}],$$

where

$$D = (D_1, \ldots, D_l) = E\left\{ \frac{\partial f^{\mathrm{T}}(Y, X, \beta^{\mathrm{T}}X)}{\partial\, \mathrm{vecl}(\beta)} D^* \right\}$$

has $(p-d)d$ rows and D_j is a $(p-d)d \times l'$ matrix, for $j = 1, \ldots, l$. Here D^* is an arbitrary positive-definite $ll' \times ll'$ matrix, with the optimal choice being $D^* = E(ff^{\mathrm{T}})^{-1}$, and $g_j(Y) = D_j g^*(Y)$. In (A1) $a_j^*(X)$ is either a scalar, such as in sliced inverse regression, or a column vector, such as in principal Hessian directions. When the $a_j^*(X)$s are column vectors, we further expand the matrix multiplication $g_j(Y)[a_j^*(X) - E\{a_j^*(X) \mid \beta^{\mathrm{T}}X\}]$ $(j = 1, \ldots, l)$ so that eventually, after simplification through combining terms that are redundant, each summand contains a scalar function a_j^* and a scalar or column vector $g_j(Y)$. We use q to denote the total number of summands. By now the form of $Df(Y, X, \beta^{\mathrm{T}}X)$ is almost the desired form in (6), except that if several different $a_j^*(X) - E\{a_j^*(X) \mid \beta^{\mathrm{T}}X\}$ are multiplied by the same $g_j(Y)$ functions, we can write them more concisely by forming a row vector of the corresponding vectors a_j^* and this is what we name a_j in (6). For example, writing $X = (X_1, \ldots, X_p)^{\mathrm{T}}$, for sliced inverse regression, we have

$$Df(Y, X, \beta^{\mathrm{T}}X) = \sum_{j=1}^{p} D_j E(X \mid Y)\{X_j - E(X_j \mid \beta^{\mathrm{T}}X)\},$$

hence $q = p$, $g_j(Y) = D_j E(X \mid Y)$, which is a column vector, and $a_j(X) = X_j$, which is a scalar, in (6). For principal Hessian directions, we have

$$\begin{aligned} Df(Y, X, \beta^{\mathrm{T}}X) &= \sum_{j=1}^{p} \{Y - E(Y)\} D_j \{XX_j - E(XX_j \mid \beta^{\mathrm{T}}X)\} \\ &= \sum_{k=1}^{p} \sum_{j=1}^{p} \{Y - E(Y)\} D_{jk} \{X_k X_j - E(X_k X_j \mid \beta^{\mathrm{T}}X)\}, \end{aligned} \tag{A2}$$

where D_j has dimension $(p-d)d \times p$, and D_{jk} stands for the kth column of D_j. We rewrite (A2) into a matrix form as

$$\sum_{k=1}^{p}\{Y - E(Y)\}\{X^{\mathrm{T}} D_{\cdot,k} X_k - E(X^{\mathrm{T}} D_{\cdot,k} X_k \mid \beta^{\mathrm{T}} X)\}, \tag{A3}$$

where $D_{\cdot,k}$ is a $p \times (p-d)d$ matrix with jth row D_{jk}^{T} $(k = 1, \ldots, p)$. Using (A3) to form (6), we hence have $q = p$, $g_j(Y) = Y - E(Y)$, a scalar, and $a_j(X) = X^{\mathrm{T}} D_{\cdot,j} X_j$, a row vector, in (6).

List of regularity conditions

Condition A1. The univariate kernel function $K(\cdot)$ is symmetric, has compact support and is Lipschitz continuous on its support. It satisfies

$$\int K(u)\,\mathrm{d}u = 1, \quad \int u^i K(u)\,\mathrm{d}u = 0 \ (i = 1, \ldots, m-1), \quad 0 \neq \int |u|^m K(u)\,\mathrm{d}u < \infty.$$

Thus K is a mth order kernel. The d-dimensional kernel function is a product of d univariate kernel functions, that is, $K_h(u) = K(u/h)/h^d = \Pi_{j=1}^d K_h(u_j) = \Pi_{j=1}^d K(u_j/h)/h^d$ for $u = (u_1, \ldots, u_d)^{\mathrm{T}}$. Here we abuse notation and use the same K regardless of the dimension of its argument.

Condition A2. The probability density function of $\beta^{\mathrm{T}} X$, denoted by $f(\beta^{\mathrm{T}} X)$, is bounded away from zero and infinity.

Condition A3. Let $r(\beta^{\mathrm{T}} X) = E\{a(X) \mid \beta^{\mathrm{T}} X\} f(\beta^{\mathrm{T}} X)$. The $(m-1)$th derivatives of $r(\beta^{\mathrm{T}} X)$ and $f(\beta^{\mathrm{T}} X)$ are locally Lipschitz-continuous as functions of $\beta^{\mathrm{T}} X$.

Condition A4. The bandwidth $h = O(n^{-\kappa})$ for $(2m)^{-1} < \kappa < (2d)^{-1}$. This implies $m > d$.

Technical details

Let $\hat{f}(\beta^{\mathrm{T}} X_i) = (n-1)^{-1} \sum_{j \neq i} K_h(X_j^{\mathrm{T}}\beta - X_i^{\mathrm{T}}\beta)$ and $\hat{r}(\beta^{\mathrm{T}} X_i) = (n-1)^{-1} \sum_{j \neq i} K_h(X_j^{\mathrm{T}}\beta - X_i^{\mathrm{T}}\beta) a(X_j)$. Write $\eta_i = \beta^{\mathrm{T}} X_i$, $\eta = \beta^{\mathrm{T}} X$, $\hat{\eta}_i = \hat{\beta}^{\mathrm{T}} X_i$.

LEMMA A1. *Assume* $E(\varepsilon \mid \eta) = 0$. *Under Conditions* A1–A3, *we have*

$$\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \{\hat{E}(a \mid \eta_i) - E(a \mid \eta_i)\} = O_p\{h^m/n^{1/2} + h^{2m} + \log^2 n/(nh^d)\}.$$

Proof of Lemma A1. Recall that $r(\eta) = E(a \mid \eta) f(\eta)$ from Condition A3, and $\hat{r}(\eta) = \hat{E}(a \mid \eta)\hat{f}(\eta)$. We write

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \{\hat{E}(a \mid \eta_i) - E(a \mid \eta_i)\} &= \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \left\{\frac{\hat{r}(\eta_i)}{\hat{f}(\eta_i)} - \frac{r(\eta_i)}{f(\eta_i)}\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \left\{\frac{\hat{r}(\eta_i) - r(\eta_i)}{f(\eta_i)}\right\} - \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \left[\frac{r(\eta_i)\{\hat{f}(\eta_i) - f(\eta_i)\}}{f^2(\eta_i)}\right] \\
&\quad - \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \left[\frac{\{\hat{r}(\eta_i) - r(\eta_i)\}\{\hat{f}(\eta_i) - f(\eta_i)\}}{f(\eta_i)\hat{f}(\eta_i)}\right] \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \left[\frac{r(\eta_i)\{\hat{f}(\eta_i) - f(\eta_i)\}^2}{f^2(\eta_i)\hat{f}(\eta_i)}\right].
\end{aligned} \tag{A4}$$

By the uniform convergence of nonparametric regression (Mack & Silverman, 1982), the third and the fourth summations are order $O_p\{h^{2m} + \log^2 n/(nh^d)\}$. The first two summations in the right-hand side of

(A4) have similar structures, thus we explain in detail only the first one. We write $n^{-1}\sum_{i=1}^{n}\hat{r}(\eta_i)\varepsilon_i/f(\eta_i)$ as a second-order U-statistic:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{r}(\eta_i)}{f(\eta_i)}\varepsilon_i=\frac{1}{2n(n-1)}\sum_{i\neq j}K_h(\eta_i-\eta_j)\{\varepsilon_i a(X_j)/f(\eta_i)+\varepsilon_j a(X_i)/f(\eta_j)\}.$$

By using Lemma 5.2.1.A of Serfling (1980, p. 183), it follows that

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\frac{\hat{r}(\eta_i)}{f(\eta_i)}-\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\frac{E\{K_h(\eta_i-\eta_j)E(a\mid\eta_j)\mid\eta_i\}}{f(\eta_i)}=O_p\{1/(nh^{d/2})\},\tag{A5}$$

because the difference on the left-hand side is a degenerate U-statistic. Next we show that

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left[\frac{E\{K_h(\eta_i-\eta_j)E(a\mid\eta_j)\mid\eta_i\}-r(\eta_i)}{f(\eta_i)}\right]=O_p(n^{-1/2}h^m).\tag{A6}$$

Following similar arguments in Lemma 3.3 of Zhu & Fang (1996) for calculating the bias term, we easily have $\sup_{X_i}|E\{K_h(\eta_i-\eta_j)E(a\mid\eta_j)\mid\eta_i\}-r(\eta_i)|=O(h^m)$ by assuming that the $(m-1)$th derivative of $r(\eta)$ is locally Lipschitz-continuous. This proves (A6). Combining (A5) and (A6), we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left\{\frac{\hat{r}(\eta_i)-r(\eta_i)}{f(\eta_i)}\right\}=O_p\{1/(nh^{d/2})+h^m/n^{1/2}\}.$$

This result together with (A4) entails the desired result, which completes the proof. □

LEMMA A2. *Under Conditions* A1–A4*, we have*

$$\sum_{i=1}^{n}E(g\mid\eta_i)\{E(a\mid\eta_i)-\hat{E}(a\mid\eta_i)\}=\sum_{i=1}^{n}E(g\mid\eta_i)\{E(a\mid\eta_i)-a(X_i)\}+o_p(n^{1/2}).$$

Proof of Lemma A2. Using the definition of the function $r(\cdot)$ in Condition A3, and the bandwidth conditions $nh^{2d}\to\infty$ and $nh^{4m}\to 0$ in Condition A4, we obtain

$$\begin{aligned}&\sum_{i=1}^{n}E(g\mid\eta_i)\{E(a\mid\eta_i)-\hat{E}(a\mid\eta_i)\}\\&\quad=\sum_{i=1}^{n}E(g\mid\eta_i)\left[\frac{r(\eta_i)\{\hat{f}(\eta_i)-f(\eta_i)\}}{f^2(\eta_i)}-\frac{\{\hat{r}(\eta_i)-r(\eta_i)\}}{f(\eta_i)}\right]+o_p(n^{1/2}).\end{aligned}$$

Furthermore, since the bandwidth h satisfies $nh^{2m}\to 0$ under Condition A4, Lemma A2 of Zhu & Zhu (2007) yields

$$\sum_{i=1}^{n}E(g\mid\eta_i)\frac{r(\eta_i)\{\hat{f}(\eta_i)-f(\eta_i)\}}{f^2(\eta_i)}=\sum_{i=1}^{n}[E(g\mid\eta_i)E(a\mid\eta_i)-E\{E(g\mid\eta_i)E(a\mid\eta)\}]+o_p(n^{1/2}).\tag{A7}$$

Similarly, invoking Lemma A3 of Zhu & Zhu (2007), we obtain

$$\sum_{i=1}^{n}E(g\mid\eta_i)\frac{\{\hat{r}(\eta_i)-r(\eta_i)\}}{f(\eta_i)}=\sum_{i=1}^{n}[E(g\mid\eta_i)a(X_i)-E\{E(g\mid\eta)E(a\mid\eta)\}]+o_p(n^{1/2}).\tag{A8}$$

Combining (A7) and (A8), we obtain the results of Lemma A2. □

Proof of Theorem 1. We rewrite the estimating equation $\sum_{i=1}^{n} g(Y_i)\{a(X_i) - \hat{E}(a \mid \hat{\eta}_i)\} = 0$ to obtain

$$-\sum_{i=1}^{n} E(g \mid \eta_i)\{a(X_i) - \hat{E}(a \mid \hat{\eta}_i)\} = \sum_{i=1}^{n} \{g(Y_i) - E(g \mid \eta_i)\}\{a(X_i) - \hat{E}(a \mid \hat{\eta}_i)\}. \tag{A9}$$

We first study the left-hand side of (A9):

$$\begin{aligned}\sum_{i=1}^{n} E(g \mid \eta_i)\{a(X_i) - \hat{E}(a \mid \hat{\eta}_i)\} &= \sum_{i=1}^{n} E(g \mid \eta_i)\{a(X_i) - E(a \mid \eta_i)\} + \sum_{i=1}^{n} E(g \mid \eta_i)\{E(a \mid \eta_i) \\ &\quad - \hat{E}(a \mid \eta_i)\} + \sum_{i=1}^{n} E(g \mid \eta_i)\{\hat{E}(a \mid \eta_i) - \hat{E}(a \mid \hat{\eta}_i)\}. \end{aligned} \tag{A10}$$

From Lemma A2, the summation of the first two terms on the right-hand side of (A10) is $o_p(n^{1/2})$. Using Taylor's expansion and the weak law of large numbers, denoting the Kronecker product as $\otimes$, i.e., $M \otimes B = (m_{ij}B)$ for any matrices M and B, we rewrite the vectorized form of the third summation on the right-hand side of (A10) as

$$\begin{aligned}\operatorname{vec}\left[\sum_{i=1}^{n} E(g \mid \eta_i)\{\hat{E}(a \mid \eta_i) - \hat{E}(a \mid \hat{\eta}_i)\}\right] &= -\sum_{i=1}^{n} \left\{\frac{\partial \hat{E}(a^{\mathrm{T}} \mid \eta_i)}{\partial \operatorname{vecl}(\beta)^{\mathrm{T}}} \otimes E(g \mid \eta_i)\right\} \operatorname{vecl}(\hat{\beta} - \beta) + o_p(n^{1/2}) \\ &= -nE\left\{\frac{\partial E(a^{\mathrm{T}} \mid \eta)}{\partial \operatorname{vecl}(\beta)^{\mathrm{T}}} \otimes E(g \mid \eta)\right\} \operatorname{vecl}(\hat{\beta} - \beta) + o_p(n^{1/2}).\end{aligned}$$

Using $\partial \operatorname{vec}(fg^{\mathrm{T}})/\partial X^{\mathrm{T}} = g \otimes \partial f/\partial X^{\mathrm{T}} + \partial g/\partial X^{\mathrm{T}} \otimes f$, we obtain that

$$\begin{aligned}A &= E\left(\frac{\partial \operatorname{vec}[\{g(Y) - E(g \mid \eta)\}\{a(X) - E(a \mid \eta)\}]}{\partial \operatorname{vecl}(\beta)^{\mathrm{T}}}\right) \\ &= E\left[\{a(X) - E(a \mid \eta)\}^{\mathrm{T}} \otimes \frac{\partial\{g(Y) - E(g \mid \eta)\}}{\partial \operatorname{vecl}(\beta)^{\mathrm{T}}}\right] + E\left[\frac{\partial\{a(X) - E(a \mid \eta)\}^{\mathrm{T}}}{\partial \operatorname{vecl}(\beta)^{\mathrm{T}}} \otimes \{g(Y) - E(g \mid \eta)\}\right] \\ &= -E\left[\{a(X) - E(a \mid \eta)\}^{\mathrm{T}} \otimes \frac{\partial E(g \mid \eta)}{\partial \operatorname{vecl}(\beta)^{\mathrm{T}}}\right] - E\left[\frac{\partial E(a \mid \eta)^{\mathrm{T}}}{\partial \operatorname{vecl}(\beta)^{\mathrm{T}}} \otimes \{g(Y) - E(g \mid \eta)\}\right] \\ &= -E\left[\{a(X) - E(a \mid \eta)\}^{\mathrm{T}} \otimes \frac{\partial E(g \mid \eta)}{\partial \operatorname{vecl}(\beta)^{\mathrm{T}}}\right],\end{aligned}$$

where the last equality is because $E(g \mid X) = E(g \mid \eta)$. Hence

$$E\left\{\frac{\partial E(a^{\mathrm{T}} \mid \eta)}{\partial \operatorname{vecl}(\beta)^{\mathrm{T}}} \otimes E(g \mid \eta)\right\} + A = -E\left(\frac{\partial \operatorname{vec}[E(g \mid \eta)\{a(X) - E(a \mid \eta)\}]}{\partial \operatorname{vecl}(\beta)^{\mathrm{T}}}\right) = 0,$$

since $E[E(g \mid \eta)\{a(X) - E(a \mid \eta)\}] = 0$ for all β. Thus, the vectorized form of the left-hand side of (A9) is $-nA \operatorname{vecl}(\hat{\beta} - \beta) + o_p(n^{1/2})$. Next we study the right-hand side of (A9). We write

$$\begin{aligned}&\sum_{i=1}^{n} \{g(Y_i) - E(g \mid \eta_i)\}\{a(X_i) - \hat{E}(a \mid \hat{\eta}_i)\} \\ &\quad = \sum_{i=1}^{n} \{g(Y_i) - E(g \mid \eta_i)\}\{a(X_i) - E(a \mid \eta_i)\} + \sum_{i=1}^{n} \{g(Y_i) - E(g \mid \eta_i)\}\{E(a \mid \eta_i) - \hat{E}(a \mid \eta_i)\} \\ &\qquad + \sum_{i=1}^{n} \{g(Y_i) - E(g \mid \eta_i)\}\{\hat{E}(a \mid \eta_i) - \hat{E}(a \mid \hat{\eta}_i)\}.\end{aligned}$$

Because $E\{g(Y_i) - E(g \mid \eta_i) \mid \eta_i\} = 0$, a direct application of Lemma A1 entails that the second term is of order $O_P\{n^{1/2}h^m + nh^{2m} + (\log^2 n)h^{-d}\}$, which is $o_p(n^{1/2})$ when $nh^{2d} \to \infty$ and $nh^{4m} \to 0$. By Taylor expansion, the vectorized form of the third term is

$$\begin{aligned}
&\operatorname{vec}\left[\sum_{i=1}^{n}\{g(Y_i) - E(g \mid \eta_i)\}\{\hat{E}(a \mid \eta_i) - \hat{E}(a \mid \hat{\eta}_i)\}\right] \\
&\quad = -\sum_{i=1}^{n}\left[\frac{\partial \hat{E}(a^{\mathrm{T}} \mid \eta_i)}{\partial \operatorname{vecl}(\beta)^{\mathrm{T}}} \otimes \{g(Y_i) - E(g \mid \eta_i)\}\right] \operatorname{vecl}(\hat{\beta} - \beta) + o_p(n^{1/2}) \\
&\quad = -nE\left[\frac{\partial E(a^{\mathrm{T}} \mid \eta)}{\partial \operatorname{vecl}(\beta)^{\mathrm{T}}} \otimes \{g(Y) - E(g \mid \eta)\}\right] \operatorname{vecl}(\hat{\beta} - \beta) + o_p(n^{1/2}) \\
&\quad = o_p(n^{1/2}),
\end{aligned}$$

where the last equality is again because $E(g \mid X) = E(g \mid \eta)$. The proof of Theorem 1 is completed by combining the results concerning the left- and right-hand sides of (A9). □

Proof of Theorem 2. A standard Taylor expansion around β yields

$$\begin{aligned}
0 &= n^{-1/2}\sum_{i=1}^{n}\operatorname{vec}[g(Y_i)\{a(X_i) - h(\tilde{\eta}_i, \tilde{\beta})\}] \\
&= n^{-1/2}\sum_{i=1}^{n}\operatorname{vec}[g(Y_i)\{a(X_i) - h(\eta_i, \beta)\}] \\
&\quad + n^{-1}\sum_{i=1}^{n}\frac{-\partial[\operatorname{vec}\{g(Y_i)h(\eta_i, \beta)\}]}{\partial \operatorname{vecl}(\beta)^{\mathrm{T}}}\bigg|_{\beta=\beta^*} n^{1/2}\{\operatorname{vecl}(\tilde{\beta} - \beta)\} + o_p(1) \\
&= n^{-1/2}\sum_{i=1}^{n}\operatorname{vec}[g(Y_i)\{a(X_i) - E(a \mid \eta_i)\}] - E\left\{\frac{E(a^{\mathrm{T}} \mid \eta)}{\partial \operatorname{vecl}(\beta)^{\mathrm{T}}} \otimes g(Y)\right\} n^{1/2}\{\operatorname{vecl}(\tilde{\beta} - \beta)\} + o_p(1),
\end{aligned}$$

where β^* is on the line connecting $\tilde{\beta}$ and β. From the proof of Theorem 1, we have

$$E\left\{\frac{\partial E(a^{\mathrm{T}} \mid \eta)}{\partial \operatorname{vecl}(\beta)^{\mathrm{T}}} \otimes g(Y)\right\} = E\left\{\frac{\partial E(a^{\mathrm{T}} \mid \eta)}{\partial \operatorname{vecl}(\beta)^{\mathrm{T}}} \otimes E(g \mid \eta)\right\} = -A.$$

Thus, we have

$$0 = n^{-1/2}\sum_{i=1}^{n}\operatorname{vec}[g(Y_i)\{a(X_i) - E(a \mid \eta_i)\}] + An^{1/2}\{\operatorname{vecl}(\tilde{\beta} - \beta)\} + o_p(1),$$

hence the theorem is proven. □

Proof of Proposition 1. From Theorems 1 and 2, we can easily obtain that

$$\begin{aligned}
B_2 - B_1 &= \operatorname{cov}(\operatorname{vec}[E(g \mid \eta)\{a(X) - E(a \mid \eta)\}]) \\
&\quad + \operatorname{cov}(\operatorname{vec}[\{g(Y) - E(g \mid \eta)\}\{a(X) - E(a \mid \eta)\}], \operatorname{vec}[E(g \mid \eta)\{a(X) - E(a \mid \eta)\}]) \\
&\quad + \operatorname{cov}(\operatorname{vec}[E(g \mid \eta)\{a(X) - E(a \mid \eta)\}], \operatorname{vec}[\{g(Y) - E(g \mid \eta)\}\{a(X) - E(a \mid \eta)\}]) \\
&= \operatorname{cov}(\operatorname{vec}[E(g \mid \eta)\{a(X) - E(a \mid \eta)\}]),
\end{aligned}$$

which is clearly positive definite. The last equality holds because $E(g \mid X) = E(g \mid \eta)$. Hence $\operatorname{cov}\{\operatorname{vecl}(\tilde{\beta})\} - \operatorname{cov}\{\operatorname{vecl}(\hat{\beta})\} = n^{-1}A^{-1}(B_2 - B_1)(A^{-1})^{\mathrm{T}}$ is positive definite. □

References

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. New York: Wiley.

Cook, R. D. & Li, B. (2002). Dimension reduction for conditional mean in regression. *Ann. Statist.* **30**, 455–74.

Cook, R. D. & Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction" by Li, K. C. *J. Am. Statist. Assoc.* **86**, 28–33.

Dong, Y. & Li, B. (2010). Dimension reduction for non-elliptically distributed predictors: second-order moments. *Biometrika* **97**, 279–94.

Hall, P. & Li, K. C. (1993). On almost linearity of low dimensional projections from high dimensional data. *Ann. Statist.* **21**, 867–89.

Härdle, W., Hall, P. & Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? (with discussion). *J. Am. Statist. Assoc.* **83**, 86–99.

Henmi, M. & Eguchi, S. (2004). A paradox concerning nuisance parameters and projected estimating functions. *Biometrika* **91**, 929–41.

Li, B. & Dong, Y. (2009). Dimension reduction for non-elliptically distributed predictors. *Ann. Statist.* **37**, 1272–98.

Li, B. & Wang, S. (2007). On directional regression for dimension reduction. *J. Am. Statist. Assoc.* **102**, 997–1008.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Am. Statist. Assoc.* **86**, 316–42.

Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *J. Am. Statist. Assoc.* **87**, 1025–39.

Li, K. C. & Duan, N. (1989). Regression analysis under link violation. *Ann. Statist.* **17**, 1009–52.

Ma, Y. & Zhu, L. P. (2012). A semiparametric approach to dimension reduction. *J. Am. Statist. Assoc.* **107**, 168–79.

Mack, Y. P. & Silverman, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Prob. Theory Rel. Fields* **61**, 405–15.

Maity, A., Ma, Y. & Carroll, R. J. (2007). Efficient estimation of population-level summaries in general semiparametric regression models. *J. Am. Statist. Assoc.* **102**, 123–39.

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

Zhu, L. P. & Zhu, L. X. (2007). On kernel method for sliced average variance estimation. *J. Mult. Anal.* **98**, 970–91.

Zhu, L. P., Zhu, L. X. & Feng, Z. H. (2010a). Dimension reduction in regressions through cumulative slicing estimation. *J. Am. Statist. Assoc.* **105**, 1455–66.

Zhu, L. P., Wang, T., Zhu, L. X. & Ferré, L. (2010b). Sufficient dimension reduction through discretization-expectation estimation. *Biometrika* **97**, 295–304.

Zhu, L. X. & Fang, K. T. (1996). Asymptotics for kernel estimate of sliced inverse regression. *Ann. Statist.* **24**, 1053–68.

[*Received April* 2012. *Revised November* 2012]