*The Annals of Statistics*
2020, Vol. 48, No. 3, 1567–1592
https://doi.org/10.1214/19-AOS1859


# ON POST DIMENSION REDUCTION STATISTICAL INFERENCE

BY KYONGWON KIM[1,*], BING LI[1,**], ZHOU YU[2] AND LEXIN LI[3]

[1]*Department of Statistics, Pennsylvania State University,* [*]*kbk5206@psu.edu;* [**]*bxl9@psu.edu*
[2]*KLATASDS-MOE, School of Statistics, East China Normal University, zyu@stat.ecnu.edu.cn*
[3]*Department of Biostatistics and Epidemiology, University of California, Berkeley, lexinli@berkeley.edu*

The methodologies of sufficient dimension reduction have undergone extensive developments in the past three decades. However, there has been a lack of systematic and rigorous development of post dimension reduction inference, which has seriously hindered its applications. The current common practice is to treat the estimated sufficient predictors as the true predictors and use them as the starting point of the downstream statistical inference. However, this naive inference approach would grossly overestimate the confidence level of an interval, or the power of a test, leading to the distorted results. In this paper, we develop a general and comprehensive framework of post dimension reduction inference, which can accommodate any dimension reduction method and model building method, as long as their corresponding influence functions are available. Within this general framework, we derive the influence functions and present the explicit post reduction formulas for the combinations of numerous dimension reduction and model building methods. We then develop post reduction inference methods for both confidence interval and hypothesis testing. We investigate the finite-sample performance of our procedures by simulations and a real data analysis.

**1. Introduction.** Sufficient dimension reduction (SDR) embodies a family of methods that, in a regression setup, seek reduction of dimensionality without loss of regression information. It has proven to be a powerful tool to extract useful information from high-dimensional data, and has found wide applications in high-dimensional data analysis and regression graphics (Cook [4], Li [15] and Li [24]). For a response variable $Y$ and the $p$-dimensional predictor vector $X$, SDR seeks the $q$-dimensional sufficient predictor $\eta^{\mathsf{T}}X$, such that

$$Y \perp\!\!\!\perp X|\eta^{\mathsf{T}}X, \tag{1.1}$$

where $\perp\!\!\!\perp$ denotes statistical independence, and $\eta$ is a $p \times q$ matrix, with $q \leq p$. It is straightforward to see that $\eta$ always exists, as it can trivially take the form of the identify matrix. But it is not unique, as one can rotate or amend $\eta$ so that (1.1) still holds. As such, SDR turns to the subspace spanned by the columns of $\eta$. It is called a dimension reduction subspace, and under very minor conditions (Yin, Li and Cook [41]), the intersection of all such subspaces is itself a dimension reduction subspace. Such an intersection, by definition, is a unique and parsimonious population parameter that captures full regression information of $Y$ given $X$. It is called the central subspace, is denoted as $\mathcal{S}_{Y|X}$, and is the main object of interest in the SDR inquiry. Since the pioneering work of sliced inverse regression (Li [21]), the research in SDR has been flourishing, and numerous SDR methods have been proposed, including sliced average variance estimation (Cook and Weisberg [7]), principal hessian directions (Li [22]), minimum average variance estimation (Xia et al. [40]) and directional regression (Li

Received August 2018; revised April 2019.
*MSC2010 subject classifications.* Primary 62G08; secondary 62H99.
*Key words and phrases.* Central subspace, directional regression, estimating equations, generalized method of moment, influence function, sliced inverse regression, Von Mises expansion.

and Wang [18]), among many others. There have also been developments of SDR based variable selection and screening (Bondell and Li [2], Zhu et al. [42]), semiparametric SDR (Ma and Zhu [26, 27]) and nonlinear SDR (Li, Artemiou and Li [16], Li and Song [17]). For a comprehensive review, see Li [15].

Despite the rapid advances of sufficient dimension reduction methodologies, however, there has been a lack of development on post dimension reduction inference. The outcome of SDR is a vector of sufficient predictors, but this is not the end of a typical data analysis. In most applications, the end product is an estimated statistical model, furnished with confidence intervals and $p$-values for statistical significance. Currently, the common practice is to feed the sufficient predictors obtained from SDR to the subsequent modeling as if they were the true predictors. It then proceeds with the usual model estimation and inference procedures, which completely ignores the estimation error incurred in the dimension reduction step, and thus tends to produce overly optimistic confidence intervals and $p$-values. More specifically, sufficient dimension reduction produces an estimate $\hat{\eta}$ of the $\eta$ in (1.1), which, under mild regularity conditions, converges to $\eta$ at the $n^{-1/2}$ rate. A subsequent modeling step builds a parametric probability model, say $f_\theta(\hat{\eta}^\mathsf{T} X, Y)$, which treats $\hat{\eta}^\mathsf{T} X$ as the new predictor, and from which an estimate $\hat{\theta}$ of $\theta$ is derived. In this process, the error in $\hat{\eta}$ contributes to the error in $\hat{\theta}$, and the contribution is in the same order of magnitude, that is, $O_P(n^{-1/2})$, as the error in $\hat{\theta}$ when $\eta$ is known. If we ignore the error propagated from $\hat{\eta}$, as the current solutions do, then the confidence interval for $\theta$ will be significantly narrower than the true confidence interval, and the $p$-value for testing $\theta$ will be significantly smaller than the true $p$-value. Indeed, our data example in Section 7 shows that in some cases an inference method ignoring the error in $\hat{\eta}$ leads to a statistically significant conclusion, whereas an inference method that takes into account of the error in $\hat{\eta}$ leads to a statistically insignificant one. This lack of formal and rigorous post dimension reduction inference has seriously hindered the applications of sufficient dimension reduction.

In this article, we fill this gap by developing a general and comprehensive framework for post dimension reduction inference. The central issue for post reduction inference is to track how the error induced by dimension reduction propagates into the subsequent model estimation. To do so, we face the challenges that there are a large variety of dimension reduction methods, and as many different methods of estimating a statistical model. A useful post dimension reduction inference framework should be an open system that is capable of adapting to different dimension reduction and model estimation methods. Our idea is to use the influence functions of statistical functionals as a vehicle to achieve this generality. Many SDR methods can be expressed as eigenvectors of matrix-valued statistical functionals. As such, they can be expanded as asymptotic linear forms under mild regularity conditions (Bickel et al. [1]). Likewise, many estimation methods can also be expressed as vector-valued statistical functionals, which again can be expanded as asymptotic linear forms. These two asymptotic linear forms are uniquely determined by the influence functions of the statistical functionals for dimension reduction and estimation, and together would uniquely determine the post dimension reduction asymptotic distribution. Our post reduction framework is designed in such a way that one can input the influence functions of any dimension reduction method and any estimation method to produce the post reduction asymptotic distribution that takes both processes into account.

Within this general framework, we derive explicitly the influence functions for five popular SDR methods and three commonly used model estimation methods. The SDR methods include sliced inverse regression (SIR, Li [21]), sliced average variance estimation (SAVE, Cook and Weisberg [7]), two forms of principal Hessian directions (y-PHD and r-PHD, Li [22], Cook [5]) and directional regression (DR, Li and Wang [18]). The model estimation methods include differentiable estimating equations, nondifferentiable estimating equations

and generalized method of moments (GMM). We note that differentiable estimating equations include generalized linear model (McCullagh and Nelder [29]) as a special case, whereas nondifferentiable estimating equations include median and quantile regression as special cases. Moreover, generalized method of moments (Hansen [9], Hansen, Heaton and Yaron [10]) have been widely used in econometrics. These $5 \times 3$ combinations of SDR and estimation methods cover a wide range of statistical modeling and applications. They also serve as an illustration on how to derive the influence functions and how to plug them into our post dimension reduction inference framework to obtain the desired post reduction asymptotic distribution. As such, more SDR and estimation methods can be incorporated into this framework.

Based on the derived post dimension reduction asymptotic distribution, we proceed further to develop specific methods for conducting statistical inference: constructing confidence intervals and test statistics, and computing the asymptotic null and local alternative distributions of the test statistics. It is our hope that the materials developed in this paper can serve as a first step toward incorporating sufficient dimension reduction and post reduction inference into a systematic and comprehensive statistical method.

The rest of the paper is organized as follows. We develop the general post dimension reduction framework and the post reduction asymptotic distribution under a given pair of influence functions, one from a SDR method and the other from an estimation method, in Section 2. We next derive the explicit influence functions for three estimation methods in Sections 3, and the influence functions for five SDR methods in Section 4. We then develop the post dimension reduction statistical inference, confidence interval and hypothesis testing, in Section 5. We conduct simulations and compare with the naive inference method in Section 6, and illustrate our method with a real data analysis in Section 7. We conclude the paper with a discussion in Section 8. We report some additional simulation results in the online supplementary material [13].

**2. General framework for post reduction inference.** We begin with an introduction of two statistical functionals: one for sufficient dimension reduction, which we call the *reduction functional*, and one for model estimation, which we call the *estimation functional*. We then define the composite functional and derive its influence function, from which we obtain the post dimension reduction asymptotic distribution. Finally, we explicitly compare the asymptotic covariances of the estimated parameter with and without taking into account the error induced by dimension reduction.

2.1. *Reduction, estimation and composite functionals.* Let $(X, Y)$ be random vectors in $\mathbb{R}^p \times \mathbb{R}$ that take values in the measurable space $(\Omega_{XY}, \mathcal{F}_{XY})$. Let $\mathcal{P}$ be the class of all probability distributions of $(X, Y)$. Let $\mathcal{S}$ be a metric space, which in our context is taken as a space of matrices. A statistical functional is a mapping $R$ from $\mathcal{P}$ to $\mathcal{S}$. Let $F_0$ be the true distribution of $(X, Y)$, let $(x, y)$ be a fixed point in $\Omega_{XY}$, and let $\delta_{xy}$ be the Dirac measure at $(x, y)$. The *influence function* of the functional $R$ is defined as

$$R^\star(x, y) = \frac{\partial}{\partial \varepsilon} R\big[(1-\varepsilon)F_0 + \varepsilon \delta_{xy}\big]\big|_{\varepsilon=0}.$$

For more details about influence functions, see Bickel et al. [1]. Throughout this paper, we assume that $R^\star$ satisfies the following conditions.

ASSUMPTION 1. (1) $E[R^\star(X, Y)] = 0$.

(2) $R^\star(X, Y)$ has finite variance; if $R^\star(X, Y)$ is a random vector or a random matrix, then its entries have finite variances.

These assumptions are mild and hold for all the SDR and estimation methods considered in this paper. For a set of sufficient conditions for these assumptions, see Bickel et al. [1], page 19. When there is no ambiguity, we abbreviate $R^\star(X,Y)$ by $R^\star$. In the following, an asterisk on a symbol always indicates the influence function of a statistical functional represented by that symbol. For example, for the statistical functionals $\Phi(F,\eta)$ and $\Lambda(F)$ discussed below, $\Phi^\star$ and $\Lambda^\star$ represent their respective influence functions.

Let $(X_1,Y_1),\dots,(X_n,Y_n)$ be i.i.d. samples of $(X,Y)$. Let $F_n$ be the empirical distribution based on this sample. It is well known that, if $R$ is Hadamard differentiable, then $R(F_n)$ has the following expansion:

$$R(F_n) = R(F_0) + E_n\big(R^\star\big) + o_p\big(n^{-\frac{1}{2}}\big), \tag{2.1}$$

where $E_n(R^\star)$ denotes the sample average $n^{-1}\sum_{i=1}^n R^\star(X_i,Y_i)$. Consequently, by the central limit theorem,

$$\sqrt{n}\big[R(F_n) - R(F_0)\big] \xrightarrow{\mathcal{D}} N\big(0, \operatorname{var}\big(R^\star\big)\big). \tag{2.2}$$

Thus, the influence function $R^\star$ uniquely determines the asymptotic distribution of $R(F_n)$. Conventionally, $R(F_n)$ represents a statistic, and $R(F_0)$ the parameter it estimates. For more information about statistical functionals and influence functions, see, for example, Fernholz [8], Bickel et al. [1] and Li [15].

We first define the reduction functional. Most SDR methods can be written in the form of a generalized eigendecomposition problem. That is, there is a statistical functional $\Lambda : \mathcal{P} \to \mathbb{R}^{p\times p}$ satisfying that

$$\Sigma(F_0)^{-1}\operatorname{span}\big[\Lambda(F_0)\big] \subseteq \mathcal{S}_{Y|X}, \tag{2.3}$$

where $\Sigma(F_0)$ denotes the covariance matrix of $X$. The relation (2.3) implies that the central subspace $\mathcal{S}_{Y|X}$ can be recovered by solving the generalized eigenvalue problem

$$\Lambda(F_0)v = \lambda\Sigma(F_0)v. \tag{2.4}$$

Let $\eta = (\eta_1,\dots,\eta_r)$ denote its first $r$ eigenvectors, where $r$ is the rank of $\Lambda(F_0)$ and $r \le q$. For many SDR methods, the equality in (2.3) holds, and correspondingly, $r = q$. In this case, we say the SDR method is exhaustive. See Li, Zha and Chiaromonte [19] and Li and Wang [18] for sufficient conditions for exhaustiveness. For simplicity, we assume the SDR method is exhaustive in this article; that is, $\mathcal{S}_{Y|X}$ can be fully recovered by $\operatorname{span}(\eta_{01},\dots,\eta_{0r})$. We also note that, the generalized eigenvalue problem in (2.4) can be solved by transforming it into a standard eigenvalue problem. That is, if $\{\beta_{0i}\}_{i=1}^r$ are the first $r$ eigenvectors of $\Sigma(F_0)^{-1/2}\Lambda(F_0)\Sigma(F_0)^{-1/2}$, then $\eta_{0i} = \Sigma(F_0)^{-\frac{1}{2}}\beta_{0i}$, $i = 1,\dots,r$, are the first $r$ eigenvectors of the generalized eigenvalue problem (2.4). Given i.i.d. samples of $(X,Y)$, the corresponding sample version of (2.4) is $\Lambda(F_n)v = \lambda\Sigma(F_n)v$, where $\Sigma(F_n)$ is the sample covariance matrix of $X$. We define $\{\hat\eta_i\}_{i=1}^r$ and $\{\hat\beta_i\}_{i=1}^r$ accordingly.

We call the functional $\Lambda(F)$ the *reduction functional*, and assume it is Hadamard differentiable with the influence function $\Lambda^\star$. Correspondingly, we use $\eta(F)$ to denote the $\mathbb{R}^{p\times q}$-valued statistical functional of the first $q$ eigenvectors of $\Lambda(F)$.

We next define the estimation functional. We start with a set of fixed eigenvectors $(\eta_1,\dots,\eta_q)$ that form an orthonormal set in $\mathbb{R}^p$. Suppose we replace the original $p$-dimensional predictor vector $X$ with the $q$-dimensional sufficient predictor $\eta^T X$, then fit some parametric regression model with the model parameter $\theta$. Assume, for a fixed $\eta$, the estimate of $\theta$ takes the following general form of a statistical functional

$$\Phi : \mathcal{P}\times\mathbb{R}^{p\times q} \to \Theta \subseteq \mathbb{R}^s,$$

where $\Theta$ is the parameter space for the parametric regression model. We call the functional $\Phi$ the *estimation functional*, and assume that, for each fixed $\eta$, the mapping $F \mapsto \Phi(F, \eta)$ is Hadamard differentiable with the influence function $\Phi^\star$. Since we treat $\eta$ as fixed, this functional corresponds to the naive estimator as if $\eta$ is known.

Now we replace the fixed $\eta$ with the estimate $\hat{\eta} = \eta(F_n)$ from a given SDR method, which leads to an estimate of $\theta$, $T(F_n) = \Phi[F_n, \eta(F_n)]$, and the functional

$$T : \mathcal{P} \to \Theta, \qquad F \mapsto \Phi\big[F, \eta(F)\big].$$

We call it the *composite functional*, as it is a composition of the reduction functional $\Lambda(F)$, which is implicitly contained in $\eta(F)$, and the estimation functional $\Phi(F, \eta)$. The functional $T$ accounts for the variations in both dimension reduction and estimation, and its influence function determines the post dimension reduction asymptotic distribution. It corresponds to the inference procedure that does not pretend $\eta$ is known.

2.2. *Influence function and asymptotic distribution.* Next we derive the influence function of $T(F)$ given the influence functions $\Lambda^\star$ and $\Phi^\star$. We derive the influence functions $\Lambda^\star$ and $\Phi^\star$ for a variety of dimension reduction and estimation methods in Sections 3 and 4, respectively. In the following, we use $\otimes$ to denote the Kronecker product. We denote $\Sigma(F_n)$, $\Sigma(F_0)$, $\Sigma(F)$ by $\hat{\Sigma}$, $\Sigma_0$, $\Sigma$, and denote $\Lambda(F_n)$, $\Lambda(F_0)$, $\Lambda(F)$ by $\hat{\Lambda}$, $\Lambda_0$, $\Lambda$, respectively. We first need the following lemma, whose proof can be found in Li [15].

LEMMA 1. *Suppose all moments involved are finite. Then*

(1) $\operatorname{vec}(\Sigma^\star) = X \otimes X - E(X \otimes X) - [X - E(X)] \otimes E(X) - E\{X \otimes [X - E(X)]\}$;
(2) $\operatorname{vec}[(\Sigma^{-\frac{1}{2}})^\star] = -(\Sigma_0^{1/2} \otimes \Sigma_0 + \Sigma_0 \otimes \Sigma_0^{1/2})^{-1} \operatorname{vec}(\Sigma^\star)$;
(3) $(\Sigma^{-1})^\star = -\Sigma^{-1}\Sigma^\star\Sigma^{-1}$

THEOREM 1. *Suppose the following conditions are satisfied*:

(C1) *The statistical functionals $F \mapsto \Lambda(F)$ and $F \mapsto \Phi(F, \eta)$ are Hadamard differentiable with influence functions $\Lambda^\star(X, Y)$ and $\Phi^\star(X, Y, \eta)$. Both $\Lambda^\star$ and $\Phi^\star$ satisfy Assumption 1.*

(C2) *The function $\eta \mapsto \Phi(F_0, \eta)$ is differentiable.*

(C3) *All the nonzero eigenvalues of $\Sigma_0^{-1/2}\Lambda_0\Sigma_0^{-1/2}$ are distinct.*

*Then the influence function of $T(F)$ is*

$$T^\star(X, Y) = \Phi^\star(X, Y, \eta_0) + DC \begin{pmatrix} \operatorname{vec}\big[\Sigma^\star(X, Y)\big] \\ \operatorname{vec}\big[\Lambda^\star(X, Y)\big] \end{pmatrix},$$

*where $D = \partial\Phi(F_0, \eta_0)/\partial \operatorname{vec}(\eta)^\mathsf{T}$ and $C = (A, B)$, in which*

$$\begin{aligned} A &= -\big[\beta_0^\mathsf{T} \otimes I_p + \big(I_q \otimes \Sigma_0^{-1/2}\big)H\big(\Sigma_0^{-1/2}\Lambda_0 \otimes I_p + I_p \otimes \Lambda_0\Sigma_0^{-1/2}\big)\big] \\ &\quad \times \big(\Sigma_0 \otimes \Sigma_0^{\frac{1}{2}} + \Sigma_0^{\frac{1}{2}} \otimes \Sigma_0\big)^{-1}, \\ B &= \big(I_q \otimes \Sigma_0^{-1/2}\big)H\big(\Sigma_0^{-1/2} \otimes \Sigma_0^{-1/2}\big), \\ H &= \big(H_1^\mathsf{T}, \ldots, H_q^\mathsf{T}\big)^\mathsf{T}, \\ H_i &= \beta_{0i}^\mathsf{T} \otimes \left[\sum_{j=1, j\neq i}^{p} (\lambda_{0i} - \lambda_{0j})^{-1}\big(\beta_{0j}\beta_{0j}^\mathsf{T}\big)\right], \quad i = 1, \ldots, q. \end{aligned}$$

PROOF. Recall that the sample estimator of $\eta_{0i}$ is $\hat{\eta}_i = \hat{\Sigma}^{-1/2}\hat{\beta}_i$, where $\hat{\beta}_i$ is $i$th eigenvector of $\hat{\Sigma}^{-1/2}\hat{\Lambda}\hat{\Sigma}^{-1/2}$, $i = 1, \ldots, q$. Thus the influence function of $\hat{\eta}_i$ is

$$\eta_i^\star = (\Sigma^{-1/2})^\star \beta_{0i} + \Sigma_0^{-1/2}\beta_i^\star.$$

Furthermore, by Zhu and Fang [43], the influence function of $\hat{\beta}_i$ is

$$\beta_i^\star = \sum_{j=1, j\neq i}^{p} \frac{\beta_{0j}\beta_{0j}^\mathsf{T}(\Sigma^{-1/2}\Lambda\Sigma^{-1/2})^\star \beta_{0i}}{\lambda_{0i} - \lambda_{0j}} = H_i \operatorname{vec}\big[(\Sigma^{-1/2}\Lambda\Sigma^{-1/2})^\star\big], \tag{2.5}$$

where

$$H_i = \beta_{0i}^\mathsf{T} \otimes \left[\sum_{j=1, j\neq i}^{p} (\lambda_{0i} - \lambda_{0j})^{-1}(\beta_{0j}\beta_{0j}^\mathsf{T})\right].$$

By Lemma 1 and some simple calculation,

$$\begin{aligned}
&\operatorname{vec}\big[(\Sigma^{-1/2}\Lambda\Sigma^{-1/2})^\star\big] \\
&\quad = -(\Sigma_0^{-1/2}\Lambda_0 \otimes I_p + I_p \otimes \Lambda_0\Sigma_0^{-1/2})(\Sigma_0^{1/2} \otimes \Sigma_0 + \Sigma_0 \otimes \Sigma_0^{1/2})^{-1} \operatorname{vec}(\Sigma^\star) \\
&\qquad + (\Sigma_0^{-1/2} \otimes \Sigma_0^{-1/2}) \operatorname{vec}(\Lambda^\star).
\end{aligned}$$

Combination of (2.5) and the above equality yields

$$\begin{aligned}
&\operatorname{vec}(\beta^\star) \\
&\quad = -H(\Sigma_0^{-1/2}\Lambda_0 \otimes I_p + I_p \otimes \Lambda_0\Sigma_0^{-1/2})(\Sigma_0^{1/2} \otimes \Sigma_0 + \Sigma_0 \otimes \Sigma^{1/2})^{-1} \operatorname{vec}(\Sigma^\star) \\
&\qquad + H(\Sigma_0^{-1/2} \otimes \Sigma_0^{-1/2}) \operatorname{vec}(\Lambda^\star),
\end{aligned}$$

where $H = (H_1^\mathsf{T}, \ldots, H_q^\mathsf{T})^\mathsf{T}$. Hence

$$\begin{aligned}
\operatorname{vec}(\eta^\star) &= \operatorname{vec}\big[(\Sigma^{-\frac{1}{2}})^\star \beta_0 + \Sigma_0^{-\frac{1}{2}}\beta^\star\big] \\
&= -(\beta_0^\mathsf{T} \otimes I_p)(\Sigma_0^{\frac{1}{2}} \otimes \Sigma_0 + \Sigma_0 \otimes \Sigma_0^{\frac{1}{2}})^{-1} \operatorname{vec}(\Sigma^\star) + (I_q \otimes \Sigma_0^{-\frac{1}{2}}) \operatorname{vec}(\beta^\star) \\
&= C \begin{pmatrix} \operatorname{vec}(\Sigma^\star) \\ \operatorname{vec}(\Lambda^\star) \end{pmatrix},
\end{aligned}$$

where $C$ is as defined in the theorem. By condition (C2) and the chain rule for differentiation, we have

$$T^\star(X, Y) = \Phi^\star(X, Y, \eta_0) + D \operatorname{vec}\big[\eta^\star(X, Y)\big],$$

which completes the proof. □

Condition (C1) of Theorem 1 is mild as most $\Lambda$ matrices in SDR are functions of sample moments, which are Hadamard differentiable if the moments of $X$ and $Y$ up to a certain order are finite. Condition (C2) is also mild and is easy to verify. As we will see in Section 2.4, Condition (C3) is also satisfied by numerous SDR methods and statistical models. Based on Theorem 1, we next derive the asymptotic distribution of $\hat{\theta} = \Phi(F_n, \hat{\eta})$.

COROLLARY 1. *Suppose the conditions in Theorem 1 are satisfied. Then*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{D}} N(0, \Gamma),$$

*where* $\Gamma = (I_p, DC)B(I_p, DC)^{\mathsf{T}}$ *and*

$$B = \begin{pmatrix} E(\Phi^\star\Phi^{\star\mathsf{T}}) & E[\Phi^\star \operatorname{vec}(\Sigma^\star)^{\mathsf{T}}] & E[\Phi^\star \operatorname{vec}(\Lambda^\star)^{\mathsf{T}}] \\ E[\operatorname{vec}(\Sigma^\star)\Phi^{\star\mathsf{T}}] & E[\operatorname{vec}(\Sigma^\star)\operatorname{vec}(\Sigma^\star)^{\mathsf{T}}] & E[\operatorname{vec}(\Sigma^\star)\operatorname{vec}(\Lambda^\star)^{\mathsf{T}}] \\ E[\operatorname{vec}(\Lambda^\star)\Phi^{\star\mathsf{T}}] & E[\operatorname{vec}(\Lambda^\star)\operatorname{vec}(\Sigma^\star)^{\mathsf{T}}] & E[\operatorname{vec}(\Lambda^\star)\operatorname{vec}(\Lambda^\star)^{\mathsf{T}}] \end{pmatrix}.$$

PROOF. By Theorem 1 and the relation (2.1) between the influence function and its asymptotic linear form, we have

$$\hat\theta = \theta_0 + (I_p, DC)\, E \begin{pmatrix} \Phi^\star(X, Y, \eta_0) \\ \operatorname{vec}[\Sigma^\star(X, Y)] \\ \operatorname{vec}[\Lambda^\star(X, Y)] \end{pmatrix} + o_p(n^{-1/2}).$$

Then applying (2.2) completes the proof. □

At the sample level, $\Sigma_0$, $\lambda_{0i}$, $\beta_0$ in the matrix $C$ are estimated by $\hat\Sigma$, $\hat\lambda_i$ and $\hat\beta$. The matrix $D$ is estimated by $\partial\Phi(F_n, \eta_0)/\partial \operatorname{vec}(\eta)^{\mathsf{T}}$. This is justified by

$$\frac{\partial\Phi(F_n, \eta_0)}{\partial \operatorname{vec}(\eta)^{\mathsf{T}}} \xrightarrow{P} D,$$

which holds under mild regularity conditions.

2.3. *Asymptotic comparison of naive and objective inference.* We compare the asymptotic covariance of the parameter estimate $\hat\theta = T(F_n) = \Phi(F_n, \eta(F_n))$ that takes into account the estimation error induced by dimension reduction, and that of the naive estimate $\tilde\theta(\eta_0) = \Phi(F_n, \eta_0)$ that does not. We denote their corresponding asymptotic covariance matrices by $\Gamma(\eta_0, \theta_0)$ and $\tilde\Gamma(\eta_0, \theta_0)$, respectively. Given the data, $\Gamma(\eta_0, \theta_0)$ and $\tilde\Gamma(\eta_0, \theta_0)$ are estimated by $\Gamma(\hat\eta, \hat\theta)$ and $\tilde\Gamma(\hat\eta, \hat\theta)$. Since $\hat\eta$ and $\hat\theta$ are root-$n$ consistent and $\Gamma$ and $\tilde\Gamma$ are differentiable, the differences, $\Gamma(\hat\eta, \hat\theta) - \Gamma(\eta_0, \theta_0)$ and $\tilde\Gamma(\hat\eta, \hat\theta) - \tilde\Gamma(\eta_0, \theta_0)$, are both of the order $O_P(n^{-1/2})$. Thus it suffices to compare $\Gamma(\eta_0, \theta_0)$ with $\tilde\Gamma(\eta_0, \theta_0)$. The next theorem characterizes the amount of the asymptotic variance increase after taking the dimension reduction error into account.

THEOREM 2. *Suppose the conditions in Theorem 1 are satisfied. Moreover, suppose when* $\eta_0$ *is known,* $\tilde\theta(\eta_0)$ *is an efficient estimator of* $\theta_0$. *Then*

$$\Gamma(\eta_0, \theta_0) - \tilde\Gamma(\eta_0, \theta_0) = DC \begin{pmatrix} E[\operatorname{vec}(\Sigma^\star)\operatorname{vec}(\Sigma^\star)^{\mathsf{T}}] & E[\operatorname{vec}(\Sigma^\star)\operatorname{vec}(\Lambda^\star)^{\mathsf{T}}] \\ E[\operatorname{vec}(\Lambda^\star)\operatorname{vec}(\Sigma^\star)^{\mathsf{T}}] & E[\operatorname{vec}(\Lambda^\star)\operatorname{vec}(\Lambda^\star)^{\mathsf{T}}] \end{pmatrix} C^{\mathsf{T}} D^{\mathsf{T}}.$$

PROOF. The proof echoes the Hajek–LeCam convolution theorem of regular estimators (Bickel et al. [1]). Since, when $\eta_0$ is given, both $\hat\theta$ and $\tilde\theta(\eta_0)$ are regular estimators of $\theta_0$, and $\tilde\theta(\eta_0)$ is efficient, by the LeCam–Hajek convolution theorem, $\sqrt{n}(\hat\theta(\eta_0) - \theta_0)$ can be decomposed into the sum of two asymptotically independent terms

$$\begin{aligned} &\sqrt{n}\big(\tilde\theta(\eta_0) - \theta_0\big) + \big[\sqrt{n}(\hat\theta - \theta_0) - \sqrt{n}\big(\tilde\theta(\eta_0) - \theta_0\big)\big] \\ &\quad = \sqrt{n}E\big(\Phi^\star\big) + \sqrt{n}E\big(T^\star - \Phi^\star\big) + o_P(1), \end{aligned}$$

which implies that $E[\Phi^\star(T^\star - \Phi^\star)^{\mathsf{T}}] = 0$. Hence

$$\operatorname{var}\big[T^\star(X, Y)\big] = \operatorname{var}\big[\Phi^\star(X, Y, \eta_0)\big] + D \operatorname{var}\big\{\operatorname{vec}\big[\eta^\star(X, Y)\big]\big\} D^{\mathsf{T}}.$$

Substituting the form of $\operatorname{vec}(\eta^\star)$ into this equation completes the proof. □

2.4. *Identifiability of reduction parameter.* Here, we briefly discuss the subtle issue of the identifiability for the reduction parameters. In the framework of SDR with the structural dimension $q > 1$, the basis $(\gamma_1, \ldots, \gamma_q)$ of $\mathcal{S}_{Y|X}$ is not identifiable. However, in practice, we always use a specific SDR method, say SIR, to estimate $\mathcal{S}_{Y|X}$. A specific SDR method, when applied to a specific statistical model, almost always yields a fixed set of eigenvectors in $\mathcal{S}_{Y|X}$ up to a sign. Thus, if we agree to take, for example, the first nonzero component of the relevant eigenvectors to be positive, then we have a well-identified set of reduction dimension parameters. As an example, for Model III and Model IV in Section 6, the structural dimension $q = 2$ and the first two eigenvalues of $\Sigma_0^{-1/2}\Lambda_0\Sigma_0^{-1/2}$ for DR are, respectively, 1.30, 1.25 and 1.54, 1.35. These distinct population-level eigenvalues give rise to well-identified reduction parameters $\beta_1$ and $\beta_2$. A parametric statistical model can then be imposed upon the predictors $\beta_1^{\mathsf{T}}X$ and $\beta_2^{\mathsf{T}}X$ without ambiguity.

**3. Influence functions for estimation functionals.** The asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$ relies on the reduction influence function $\Lambda^\star(X, Y)$, the estimation influence function $\Phi^\star(X, Y, \eta)$, and the form of $D = \partial\Phi(F_0, \eta_0)/\partial\,\mathrm{vec}(\eta)^{\mathsf{T}}$. In this section, we derive the explicit forms of the influence function $\Phi^\star(X, Y, \eta)$ and the derivative $D$ for three estimation methods: the differentiable estimating equations, the non-differentiable estimating equations and the generalized method of moments. They cover a wide variety of regression methods, including generalized linear model, nonlinear mean regression and nonlinear median and quantile regression, among others.

3.1. *Differentiable estimating equations.* Many commonly used parametric models can be formulated as special cases of a general class of estimators of $\theta$, each of which is defined as the solution to the estimating equation

$$E\big[g(\theta, X, Y)\big] = 0, \tag{3.1}$$

where $E_\theta[g(\theta, X, Y)] = 0$, $\mathrm{var}_\theta[g(\theta, X, Y)]$ is a matrix with finite entries, and the dimension of $g$ is the same as the dimension of $\theta$. One example is generalized linear model, which can be expressed as the solution to the estimating equation

$$E\left\{\frac{\partial\mu(\theta^{\mathsf{T}}X)}{\partial\theta^{\mathsf{T}}}V^{-1}(\theta^{\mathsf{T}}X)\big[Y - \mu\big(\theta^{\mathsf{T}}X\big)\big]\right\} = 0,$$

where $\mu(\theta^{\mathsf{T}}X) = E(Y|\theta^{\mathsf{T}}X)$, and $V(\theta^{\mathsf{T}}X) = \mathrm{var}(Y|\theta^{\mathsf{T}}X)$. See, for example, McCullagh and Nelder [29] and Li [14]. Another example is the parametric nonlinear regression, where we minimize the objective function $E[Y - h(\theta^{\mathsf{T}}X)]^2$, and $h$ can take a polynomial form, $h(u_1, \ldots, u_k) = \sum_{i=1}^k \theta_i u_i + \sum_{i,j=1}^k \theta_{ij}u_i u_j$. Correspondingly, the parameter $\theta$ can be expressed as the solution to the estimating equation

$$E\left\{2\frac{\partial h(\theta^{\mathsf{T}}X)}{\partial\theta}\big[Y - h\big(\theta^{\mathsf{T}}X\big)\big]\right\} = 0.$$

In our context of SDR based parametric modeling, the predictor vector $X$ is replaced by the sufficient predictor $\eta^{\mathsf{T}}X$. The statistical functional of the estimator $\theta$ in (3.1) is $\Phi(F, \eta)$, which is implicitly defined by the equation $\int g[\Phi(F, \eta), \eta^{\mathsf{T}}X, Y]\,dF = 0$. We next derive the explicit forms of the corresponding influence function $\Phi^\star$ and the derivative $D$, and summarize the results in the next proposition.

PROPOSITION 1. *For the estimating equations* (3.1), *we have*

$$\Phi^\star(X, Y, \eta) = -\left\{E\left[\frac{\partial g(\theta_0, \eta^{\mathsf{T}}X, Y)}{\partial\theta^{\mathsf{T}}}\right]\right\}^{-1} g\big(\theta_0, \eta^{\mathsf{T}}X, Y\big),$$

$$D = -\left\{E\left[\frac{\partial g(\theta_0, \eta_0^{\mathsf{T}}X, Y)}{\partial\theta^{\mathsf{T}}}\right]\right\}^{-1} E\left[\frac{\partial g(\theta_0, \eta_0^{\mathsf{T}}X, Y)}{\partial u}\big(I_q \otimes X^{\mathsf{T}}\big)\right].$$

PROOF. Let $F_\varepsilon = (1-\varepsilon)F_0 + \varepsilon\delta_{XY}$. Then, for all $\varepsilon \in [0,1]$, we have $\int g[\Phi(F_\varepsilon,\eta), \eta^\mathsf{T}X, Y]\,dF_\varepsilon = 0$. Differentiating (3.1) with respect to $\varepsilon$, and evaluating the derivatives at $\varepsilon = 0$, we have

$$\frac{\partial}{\partial\varepsilon}\int g\big(\Phi(F_\varepsilon,\eta),\eta^\mathsf{T}X,Y\big)\,dF_\varepsilon\bigg|_{\varepsilon=0}$$
$$= \bigg[\int \frac{\partial g(\theta_0,\eta^\mathsf{T}X,Y)}{\partial\theta^\mathsf{T}}\,dF_0\bigg]\Phi^\star + \int g\big(\theta_0,\eta^\mathsf{T}X,Y\big)\,d(\delta_{XY}-F_0) = 0.$$

Since $Eg(\theta_0,\eta^\mathsf{T}X,Y)=0$, the second term on the right-hand side is simply $g(\theta_0,\eta^\mathsf{T}X,Y)$, which leads to the desired form for $\Phi^\star(X,Y,\eta)$.

Next, we note that $\Phi(F_0,\eta)$ satisfies $E[g(\Phi(F_0,\eta),\eta^\mathsf{T}X,Y)]=0$. Differentiating this equation with respect to $\operatorname{vec}(\eta)$, we have

$$\bigg[\int \frac{\partial}{\partial\theta^\mathsf{T}}g\big(\theta_0,\eta^\mathsf{T}X,Y\big)\,dF_0\bigg]\frac{\partial\Phi(F_0,\eta)}{\partial\operatorname{vec}(\eta)^\mathsf{T}} + \int \frac{\partial}{\partial u}g\big(\theta_0,\eta^\mathsf{T}X,Y\big)\frac{\partial(\eta^\mathsf{T}X)}{\partial\operatorname{vec}(\eta)^\mathsf{T}}\,dF_0 = 0,$$

where $\partial g/\partial u$ denotes the partial derivative with respect to the second argument of $g$, which is $\eta^\mathsf{T}X$. Since $\eta^\mathsf{T}X = \operatorname{vec}(X^\mathsf{T}\eta) = \operatorname{vec}(X^\mathsf{T}\eta I_q) = (I_q\otimes X^\mathsf{T})\operatorname{vec}(\eta)$, we have $\partial(\eta^\mathsf{T}X)/\partial\operatorname{vec}(\eta)^\mathsf{T} = I_q\otimes X^\mathsf{T}$. Hence,

$$E\bigg[\frac{\partial g(\theta_0,\eta_0^\mathsf{T}X,Y)}{\partial\theta^\mathsf{T}}\bigg]D + E\bigg[\frac{\partial g(\theta_0,\eta_0^\mathsf{T}X,Y)}{\partial u}(I_q\otimes X^\mathsf{T})\bigg] = 0.$$

Solving this equation yields the desired form for $D$. □

3.2. *Nondifferentiable estimating equations*. Another family of popular models can be formulated as solving a set of nondifferentiable estimating equations. Examples include nonlinear quantile regression (He, Fu and Fung [11], Wang and Wang [37]) and support vector regression (Smola and Scholkopt [35]). In this section, we use nonlinear quantile regression as an illustration. The derivation of the estimation functional for other models follow in a similar fashion.

For a number $\tau\in[0,1]$, define the function $\rho:\mathbb{R}\to\mathbb{R}^+$ as $\rho_\tau(u)=\tau u$ if $u>0$, and $-(1-\tau)u$ if $u<0$. Let $m(\eta^\mathsf{T}X,\theta)$ be a function such that, for the true value $(\eta_0,\theta_0)$ of $(\eta,\theta)$, it is the $\tau$th conditional quantile, $P[Y\le m(\eta_0^\mathsf{T}X,\theta_0)|X]=\tau$. At the population level, nonlinear quantile regression is defined as minimizing the objective function $E\{\rho_\tau[Y-m(\eta^\mathsf{T}X,\theta)]\}$ over $\theta\in\mathbb{R}^d$, which amounts to solving the estimating equations

$$E\bigg\{\dot\rho_\tau\big[Y-m\big(\eta^\mathsf{T}X,\theta\big)\big]\frac{\partial m(\eta^\mathsf{T}X,\theta)}{\partial\theta}\bigg\} = 0, \tag{3.2}$$

where $\dot\rho_\tau(u)=\tau I(u>0)-(1-\tau)I(u\le 0)=\tau - I(u\le 0)$. Rigorously speaking, $\dot\rho_\tau$ is not defined at $u=0$. But since $u=0$ has measure 0, we can assign any value to $\dot\rho(0)$; in our case, we set $\dot\rho(0)$ equal to $-(1-\tau)$.

Next, we write the first argument $\eta^\mathsf{T}X$ of $m(\eta^\mathsf{T}X,\theta)$ as $u$, and use the following notation for partial derivatives, $\dot m_u=\partial m/\partial u$, $\dot m_\theta=\partial m/\partial\theta$, $\ddot m_{uu}=\partial^2 m/\partial u\partial u^\mathsf{T}$, $\ddot m_{u\theta}=\partial^2 m/\partial u\partial\theta^\mathsf{T}$ and $\ddot m_{\theta\theta}=\partial^2 m/\partial\theta\partial\theta^\mathsf{T}$. We derive the influence function $\Phi^\star$ and the derivative $D$ in the next proposition.

PROPOSITION 2. *For the estimating equations* (3.2), *we have*

$$\Phi^\star(X,Y,\eta) = \big(E\big\{f_{Y|X}\big[m\big(\eta_0^\mathsf{T}X,\theta_0\big)|X\big]\dot m_\theta\big(\eta_0^\mathsf{T}X,\theta_0\big)\dot m_\theta^\mathsf{T}\big(\eta_0^\mathsf{T}X,\theta_0\big)\big\}\big)^{-1}$$
$$\times\big\{\tau - I\big[Y\le m\big(\eta_0^\mathsf{T}X,\theta_0\big)\big]\big\}\dot m_\theta\big(\eta_0^\mathsf{T}X,\theta_0\big),$$
$$D = -\big(E\big\{\dot m_\theta\big(\eta_0^\mathsf{T}X,\theta_0\big)\dot m_\theta^\mathsf{T}\big(\eta_0^\mathsf{T}X,\theta_0\big)f_{Y|X}\big[m\big(\eta_0^\mathsf{T},\theta_0\big)|x\big]\big\}\big)^{-1}$$
$$\times E\big\{\dot m_\theta\big(\eta_0^\mathsf{T}X,\theta_0\big)\dot m_u^\mathsf{T}\big(\eta_0^\mathsf{T}X,\theta_0\big)(I_q\otimes X^\mathsf{T})f_{Y|X}\big[m\big(\eta_0^\mathsf{T}X,\theta_0\big)|x\big]\big\}.$$

PROOF. Denote $A(F,\eta_0)=\int \dot\rho_\tau\{Y-m[\eta_0^\top X,\Phi(F,\eta_0)]\}\dot m_\theta[\eta_0^\top X,\Phi(F,\eta_0)]\,dF$. The influence function $\Phi^\star(X,Y,\eta_0)$ can be obtained from the equation

$$\frac{\partial}{\partial\varepsilon}A(F_\varepsilon,\eta_0)\Big|_{\varepsilon=0}=0.$$

In the following, we abbreviate $\partial f(\varepsilon)/\partial\varepsilon|_{\varepsilon=0}$ by $\partial f(\varepsilon)/\partial\varepsilon$. By the chain rule, we decompose the above derivative into three terms:

$$\frac{\partial}{\partial\varepsilon}A(F_\varepsilon,\eta_0)=\frac{\partial}{\partial\varepsilon}A_1(F_\varepsilon,\eta_0)+\frac{\partial}{\partial\varepsilon}A_2(F_\varepsilon,\eta_0)+\frac{\partial}{\partial\varepsilon}A_3(F_\varepsilon,\eta_0), \tag{3.3}$$

where

$$A_1(F_\varepsilon,\eta_0)=\int \dot\rho_\tau\{Y-m[\eta_0^\top X,\Phi(F_\varepsilon,\eta_0)]\}\dot m_\theta[\eta_0^\top X,\Phi(F_0,\eta_0)]\,dF_0,$$

$$A_2(F_\varepsilon,\eta_0)=\int \dot\rho_\tau\{Y-m[\eta_0^\top X,\Phi(F_0,\eta_0)]\}\dot m_\theta[\eta_0^\top X,\Phi(F_\varepsilon,\eta_0)]\,dF_0,$$

$$A_3(F_\varepsilon,\eta_0)=\int \dot\rho_\tau\{Y-m[\eta_0^\top X,\Phi(F_0,\eta_0)]\}\dot m_\theta[\eta_0^\top X,\Phi(F_0,\eta_0)]\,dF_\varepsilon.$$

The term $\partial A_1(F_\varepsilon,\eta_0)/\partial\varepsilon$ can be written as

$$\begin{aligned}
&\frac{\partial}{\partial\varepsilon}A_1(F_\varepsilon,\eta_0)\\
&\quad=\frac{\partial}{\partial\varepsilon}\int\big(\tau-I\{Y\le m[\eta_0^\top X,\Phi(F_\varepsilon,\eta_0)]\}\big)\dot m_\theta[\eta_0^\top X,\Phi(F_0,\eta_0)]\,dF_0\\
&\quad=-\frac{\partial}{\partial\varepsilon}\int I\{Y\le m[\eta_0^\top X,\Phi(F_\varepsilon,\eta_0)]\}\dot m_\theta[\eta_0^\top X,\Phi(F_0,\eta_0)]\,dF_0\\
&\quad=-\int_{\Omega_X}\frac{\partial}{\partial\varepsilon}\int_{-\infty}^{m[\eta_0^\top X,\Phi(F_\varepsilon,\eta_0)]} f_{Y|X}(y|x)\,dy\,\dot m_\theta[\eta_0^\top X,\Phi(F_0,\eta_0)]\\
&\qquad\times f_X(x)\,dx\\
&\quad=-\left\{\int_{\Omega_X} f_{Y|X}[m(\eta_0^\top X,\theta_0)|x]\dot m_\theta(\eta_0^\top X,\theta_0)\dot m_\theta^\top(\eta_0^\top X,\theta_0)f_X(x)\,dx\right\}\Phi^\star\\
&\quad=-E\{f_{Y|X}[m(\eta_0^\top X,\theta_0)|x]\dot m_\theta(\eta_0^\top X,\theta_0)\dot m_\theta^\top(\eta_0^\top X,\theta_0)\}\Phi^\star,
\end{aligned} \tag{3.4}$$

where the first equality is by the definition of $\dot\rho_\tau(u)$, the second equality is because $\tau\int\dot m_\theta(\eta_0^\top X,\Phi(F_0,\eta_0))\,dF_0$ does not depend on $\varepsilon$, and the fourth equality is because $\Phi(F_0,\eta_0)=\theta_0$.

The term $\partial A_2(\varepsilon,\eta)/\partial\varepsilon$ can be written as

$$\begin{aligned}
\frac{\partial}{\partial\varepsilon}A_2(\varepsilon,\eta)&=\int\dot\rho_\tau[Y-m(\eta_0^\top X,\theta_0)]\ddot m_{\theta\theta}(\eta_0^\top X,\theta_0)\,dF_0\Phi^\star\\
&=E[E\{\dot\rho_\tau[Y-m(\eta_0^\top X,\theta_0)]|X\}\ddot m_{\theta\theta}(\eta_0^\top X,\theta_0)]\Phi^\star\\
&=0,
\end{aligned} \tag{3.5}$$

where the last equality is due to that, since $m(\eta_0^\top X,\theta_0)$ is the $\tau$th conditional quantile,

$$E\{\dot\rho_\tau[Y-m(\eta_0^\top X,\theta_0)]|X\}=E\{\tau-I[Y\le m(\eta_0^\top X,\theta_0)]|X\}=0.$$

The term $\partial A_3(F_\varepsilon,\eta)/\partial\varepsilon$ can be written as

$$\begin{aligned}
&\frac{\partial}{\partial\varepsilon}A_3(\varepsilon,\eta)\\
&\quad=\dot\rho_\tau[Y-m(\eta_0^\top X,\theta_0)]\dot m_\theta(\eta_0^\top X,\theta_0)-E\{\dot\rho_\tau[Y-m(\eta_0^\top X,\theta_0)]\dot m_\theta(\eta_0^\top X,\theta_0)\}.
\end{aligned}$$

By the fact that $A(F, \eta_0) = 0$, the second term above is 0, leading to

$$\partial A_3(F_\varepsilon, \eta_0)/\partial\varepsilon = \dot\rho_\tau\big[Y - m\big(\eta_0^\mathsf{T} X, \theta_0\big)\big]\dot m_\theta\big(\eta_0^\mathsf{T} X, \theta_0\big). \tag{3.6}$$

Substituting (3.4), (3.5) and (3.6) into (3.3), we obtain

$$\begin{aligned} &-E\big\{f_{Y|X}\big[m\big(\eta_0^\mathsf{T} X, \theta_0\big)|X\big]\dot m_\theta\big(\eta_0^\mathsf{T} X, \theta_0\big)\dot m_\theta^\mathsf{T}\big(\eta_0^\mathsf{T} X, \theta_0\big)\big\}\Phi^\star \\ &\quad + \dot\rho_\tau\big[Y - m\big(\eta_0^\mathsf{T} X, \theta_0\big)\big]\dot m_\theta\big(\eta_0^\mathsf{T} X, \theta_0\big) = 0. \end{aligned}$$

This yields the desired form for $\Phi^\star$.

Next, we note that $\eta \mapsto \Phi(F_0, \eta)$ is defined by the equation

$$\int \dot\rho_\tau\big\{Y - m\big[\eta^\mathsf{T} X, \Phi(F_0, \eta)\big]\big\}\dot m_\theta\big[\eta^\mathsf{T} X, \Phi(F_0, \eta)\big]\,dF_0 = 0.$$

Denote the left-hand side by $B(\eta)$, we have

$$\frac{\partial}{\partial \operatorname{vec}(\eta)^\mathsf{T}} B(\eta_0) = \frac{\partial}{\partial \operatorname{vec}(\eta)^\mathsf{T}} B_1(\eta_0) + \frac{\partial}{\partial \operatorname{vec}(\eta)^\mathsf{T}} B_2(\eta_0),$$

where

$$\begin{aligned} B_1(\eta) &= \int \dot\rho_\tau\big\{Y - m\big[\eta^\mathsf{T} X, \Phi(F_0, \eta)\big]\big\}\dot m\big[\eta_0^\mathsf{T} X, \Phi(F_0, \eta_0)\big]\,dF_0 \\ B_2(\eta) &= \int \dot\rho_\tau\big\{Y - m\big[\eta_0^\mathsf{T} X, \Phi(F_0, \eta_0)\big]\big\}\dot m\big[\eta^\mathsf{T} X, \Phi(F_0, \eta)\big]\,dF_0. \end{aligned}$$

Since $E[\dot\rho_\tau(Y - m(\eta_0^\mathsf{T} X, \theta_0))|X] = 0$, we have

$$\frac{\partial}{\partial \operatorname{vec}(\eta)^\mathsf{T}} B_2(\eta_0) = \int \dot\rho_\tau\big(Y - m\big(\eta_0^\mathsf{T} X, \theta_0\big)\big)\frac{\partial \dot m(\eta_0^\mathsf{T} X, \Phi(F_0, \eta_0))}{\partial \operatorname{vec}(\eta)^\mathsf{T}}\,dF_0 = 0.$$

The term $\partial B_1(\eta)/\partial \operatorname{vec}(\eta)^\mathsf{T}$ can be written as

$$\begin{aligned} &\frac{\partial}{\partial \operatorname{vec}(\eta)^\mathsf{T}} \int_{\Omega_X} \int_{-\infty}^{m[\eta^\mathsf{T} X, \Phi(F_0,\eta)]} f_{Y|X}(y|x)\,dy\,\dot m\big[\eta_0^\mathsf{T} X, \Phi(F_0, \eta_0)\big] f_X(x)\,dx \\ &\quad = \int_{\Omega_X} \dot m_\theta\big(\eta_0^\mathsf{T} X, \theta_0\big) f_{Y|X}\big[m\big(\eta_0^\mathsf{T} X, \theta_0\big)|x\big] \\ &\qquad \times \left[\dot m_u^\mathsf{T}\big(\eta_0^\mathsf{T} X, \theta_0\big)\frac{\partial \eta^\mathsf{T} X}{\partial \operatorname{vec}(\eta)^\mathsf{T}} + \dot m_\theta\big(\eta_0^\mathsf{T} X, \theta_0\big) D\right] f_X(x)\,dx. \end{aligned}$$

Recall that $\partial(\eta^\mathsf{T} X)/\partial \operatorname{vec}(\eta)^\mathsf{T} = I_q \otimes X^\mathsf{T}$. So the above term can be written as

$$E\big\{\dot m_\theta\big(\eta_0^\mathsf{T} X, \theta_0\big) f_{Y|X}\big(m\big(\eta_0^\mathsf{T} X, \theta_0\big)|x\big)\big[\dot m_u^\mathsf{T}\big(\eta_0^\mathsf{T} X, \theta_0\big)(I_q \otimes X^\mathsf{T}) + \dot m_\theta^\mathsf{T}\big(\eta_0^\mathsf{T} X, \theta_0\big) D\big]\big\}.$$

Equating it to 0 and solving for $D$ lead to the desired form for $D$. □

3.3. *Generalized method of moments*. Generalized method of moments [9], GMM, is a popular parametric method in both econometrics and statistics. For instance, it is used to construct optimal estimation and inference procedures based on generalized estimating equations (Qu, Lindsay and Li [33]), or to combine efficient and robust estimators (Park and Lindsay [32]). We next derive the influence function $\Phi^\star(X, Y, \eta_0)$ and $D$ for this approach.

In GMM, we have more estimating equations than parameters. That is, we estimate the $p$-dimensional parameter vector $\theta$ by $k > p$ estimating equations $E_n[g(\theta, \eta^\mathsf{T} X, Y)] = 0$, where

$$g\big(\theta, \eta^\mathsf{T} X, Y\big) = \big[g_1\big(\theta, \eta^\mathsf{T} X, Y\big), \ldots, g_k\big(\theta, \eta^\mathsf{T} X, Y\big)\big]^\mathsf{T},$$

and again we assume $E_{\theta,\eta}[g(\theta,\eta^{\mathsf{T}}X,Y)]=0$ and $\operatorname{var}_{\theta,\eta}[g(\theta,\eta^{\mathsf{T}}X,Y)]<\infty$. For a given $\eta$, $\tilde{\theta}(\eta)=\Phi(F_n,\eta)$ in the optimal version of GMM is defined as the minimizer of the function

$$\begin{aligned}&L(F_n,\theta,\eta)\\&\quad=E_n g(\theta,\eta^{\mathsf{T}}X,Y)^{\mathsf{T}}\big[E_n g(\theta,\eta^{\mathsf{T}}X,Y)g^{\mathsf{T}}(\theta,\eta^{\mathsf{T}}X,Y)\big]^{-1}E_n g(\theta,\eta^{\mathsf{T}}X,Y).\end{aligned}$$

Thus, the functional $\Phi(F,\eta)$ is the minimizer of

$$L(F,\theta,\eta)=V(F,\theta,\eta)^{\mathsf{T}}W(F,\theta,\eta)V(F,\theta,\eta),$$

where $V(F,\theta,\eta)=\int g(\theta,\eta^{\mathsf{T}}X,Y)\,dF$, and

$$W(F,\theta,\eta)=\left(\int g(\theta,\eta^{\mathsf{T}}X,Y)g^{\mathsf{T}}(\theta,\eta^{\mathsf{T}}X,Y)\,dF\right)^{-1}.$$

PROPOSITION 3. *For the generalized method of moments, we have*

$$\begin{aligned}\Phi^{\star}(X,Y,\eta)&=-\left\{E\left(\frac{\partial g^{\mathsf{T}}}{\partial\theta}\right)\big[E(gg^{\mathsf{T}})\big]^{-1}E\left(\frac{\partial g}{\partial\theta^{\mathsf{T}}}\right)\right\}^{-1}E\left(\frac{\partial g^{\mathsf{T}}}{\partial\theta}\right)\big[E(gg^{\mathsf{T}})\big]^{-1}g,\\D&=-\left\{E\left(\frac{\partial g^{\mathsf{T}}}{\partial\theta}\right)\big[E(gg^{\mathsf{T}})\big]^{-1}E\left(\frac{\partial g}{\partial\theta^{\mathsf{T}}}\right)\right\}^{-1}E\left(\frac{\partial g^{\mathsf{T}}}{\partial\theta}\right)\big[E(gg^{\mathsf{T}})^{-1}\big]\\&\quad\times E\left(\frac{\partial g}{\partial u^{\mathsf{T}}}\right)(I_q\otimes X^{\mathsf{T}}),\end{aligned}$$

*where* $g=g(\theta_0,\eta_0^{\mathsf{T}}X,Y)$.

PROOF. Let $H(F,\theta,\eta_0)=\partial L(F,\theta,\eta_0)/\partial\theta$. Then $\Phi(F,\eta_0)$ satisfies

$$H\big(F,\Phi(F,\eta_0),\eta_0\big)=0.$$

Hence the influence function $\Phi^{\star}(X,Y,\eta_0)$ can be solved from the equation

$$\frac{\partial}{\partial\varepsilon}H\big(F_\varepsilon,\Phi(F_\varepsilon,\eta_0),\eta_0\big)=0,$$

which, by the chain rule, yields

$$\Phi^{\star}=-\left[\frac{\partial}{\partial\theta^{\mathsf{T}}}H(F_0,\theta_0,\eta_0)\right]^{-1}\frac{\partial}{\partial\varepsilon}H\big[F_0,\Phi(F_\varepsilon,\eta_0),\eta_0\big]\Big|_{\varepsilon=0}.$$

We now express the above derivatives in terms of $V(F,\theta,\eta)$ and $W(F,\theta,\eta)$. By definition,

$$\begin{aligned}\frac{\partial L(F,\theta,\eta_0)}{\partial\theta}&=\frac{\partial V^{\mathsf{T}}(F,\theta,\eta_0)}{\partial\theta}W(F,\theta,\eta_0)V(F,\theta,\eta_0)\\&\quad+V^{\mathsf{T}}(F,\theta,\eta_0)\frac{\partial W(F,\theta,\eta_0)}{\partial\theta}V(F,\theta,\eta_0)\\&\quad+V^{\mathsf{T}}(F,\theta,\eta_0)W(F,\theta,\eta_0)\frac{\partial V(F,\theta,\eta_0)}{\partial\theta},\end{aligned}\tag{3.7}$$

Differentiating (3.7) with respect to $\theta$, and evaluating the derivative at $\theta_0$, we obtain

$$\begin{aligned}\frac{\partial H}{\partial\theta^{\mathsf{T}}}&=\frac{\partial^2V}{\partial\theta\partial\theta^{\mathsf{T}}}WV+\frac{\partial V^{\mathsf{T}}}{\partial\theta}\frac{\partial W}{\partial\theta^{\mathsf{T}}}V+\frac{\partial V^{\mathsf{T}}}{\partial\theta}W\frac{\partial V}{\partial\theta^{\mathsf{T}}}\\&\quad+\frac{\partial V^{\mathsf{T}}}{\partial\theta^{\mathsf{T}}}\frac{\partial W}{\partial\theta}V+V^{\mathsf{T}}\frac{\partial^2W}{\partial\theta\partial\theta^{\mathsf{T}}}V+V^{\mathsf{T}}\frac{\partial W}{\partial\theta}\frac{\partial V}{\partial\theta^{\mathsf{T}}}\\&\quad+\frac{\partial V^{\mathsf{T}}}{\partial\theta^{\mathsf{T}}}W\frac{\partial V}{\partial\theta}+V^{\mathsf{T}}\frac{\partial W}{\partial\theta^{\mathsf{T}}}\frac{\partial V}{\partial\theta}+V^{\mathsf{T}}W\frac{\partial^2V}{\partial\theta\partial\theta^{\mathsf{T}}}.\end{aligned}\tag{3.8}$$

Since, by construction, $V(F_0, \theta_0, \eta_0) = \int g(\theta_0, \eta_0^\mathsf{T} X, Y)\, dF_0 = 0$, all the terms in (3.8) that involve $V$ vanish, resulting in

$$\frac{\partial H}{\partial \theta^\mathsf{T}} = 2\frac{\partial V^\mathsf{T}}{\partial \theta^\mathsf{T}} W \frac{\partial V}{\partial \theta^\mathsf{T}}. \tag{3.9}$$

Similarly, we have

$$\frac{\partial H}{\partial \varepsilon} = \frac{\partial V^\mathsf{T}}{\partial \theta} W \frac{\partial V}{\partial \varepsilon} + \frac{\partial V^\mathsf{T}}{\partial \varepsilon} W \frac{\partial V}{\partial \theta} = 2\frac{\partial V^\mathsf{T}}{\partial \varepsilon} W \frac{\partial V}{\partial \theta}.$$

Using the fact that

$$\frac{\partial V^\mathsf{T}}{\partial \theta} = \frac{\partial}{\partial \theta} \int g^\mathsf{T}(\theta_0, \eta_0^\mathsf{T} X, Y)\, dF_0 = E\left[\frac{\partial}{\partial \theta} g^\mathsf{T}(\theta_0, \eta_0^\mathsf{T} X, Y)\right],$$

we obtain the desired form for $\Phi^\star(X, Y, \eta_0)$.

Next, we note that $H[F_0, \Phi(F_0, \eta), \eta] = 0$ for all $\eta$. Hence,

$$\frac{\partial}{\partial \theta^\mathsf{T}} H(F_0, \theta_0, \eta) \frac{\partial \Phi(F_0, \eta)}{\partial \operatorname{vec}(\eta)^\mathsf{T}} + \frac{\partial H(F_0, \theta_0, \eta)}{\partial \operatorname{vec}(\eta)^\mathsf{T}} = 0.$$

Solving this equation, we have

$$D = -\left(\frac{\partial H}{\partial \theta^\mathsf{T}}\right)^{-1} \frac{\partial H}{\partial \operatorname{vec}(\eta)^\mathsf{T}}. \tag{3.10}$$

The computation of $\partial H / \partial \operatorname{vec}(\eta)^\mathsf{T}$ is similar to that of $\partial H / \partial \theta^\mathsf{T}$: there are 9 terms in total, and all the terms that involve $V$ vanish, resulting in

$$\begin{aligned} \frac{\partial H(F_0, \theta_0, \eta_0)}{\partial \operatorname{vec}(\eta)^\mathsf{T}} &= \frac{\partial V^\mathsf{T}}{\partial \theta} W \frac{\partial V}{\partial \operatorname{vec}(\eta)^\mathsf{T}} + \frac{\partial V^\mathsf{T}}{\partial \operatorname{vec}(\eta)^\mathsf{T}} W \frac{\partial V}{\partial \theta} \\ &= 2\frac{\partial V^\mathsf{T}}{\partial \theta} W \frac{\partial V}{\partial \operatorname{vec}(\eta)^\mathsf{T}}. \end{aligned} \tag{3.11}$$

Furthermore,

$$\begin{aligned} \frac{\partial V}{\operatorname{vec}(\eta)^\mathsf{T}} &= E\left[\frac{\partial}{\partial u^\mathsf{T}} g(\theta_0, \eta_0^\mathsf{T} X, Y)\right] \frac{\partial \operatorname{vec}(\eta^\mathsf{T} X)}{\partial \operatorname{vec}(\eta)^\mathsf{T}} \\ &= E\left[\frac{\partial}{\partial u^\mathsf{T}} g(\theta_0, \eta_0^\mathsf{T} X, Y)\right] (I_q \otimes X^\mathsf{T}). \end{aligned} \tag{3.12}$$

Substituting (3.9), (3.11), (3.12) into (3.10), we obtain the desired form of $D$. □

**4. Influence functions for reduction functionals.** In this section, we derive the influence function $\Lambda^\star(X, Y)$ for some popular SDR methods, including SIR, SAVE, DR and two forms of PHD. Although some forms of asymptotic expansions exist in the SDR literature (Li [21, 22], Li and Wang [18], Shao, Cook and Weisberg [34], Li [15]), they have all been developed for sequential tests, and none was in the form suitable for post reduction inference. Also, the development here can be extended to other regression-based SDR methods, for example, the minimal discrepancy method (Cook and Ni [6]), in a similar fashion.

Many SDR methods begin with slicing the range of the response to a fixed number of nonoverlapping intervals; let $\{J_k : k = 1, \ldots, H\}$ be a set of intervals that partition $\Omega_Y$. Let $D_k = I(Y \in J_k)$, $p_k = E(D_k)$, $\mu_k = E(X|Y \in J_k)$, and $\Sigma_k = \operatorname{var}(X|Y \in J_k)$. Let $\mu = E(X)$, $\nu = E(Y)$. The specific form of $\Lambda$ for the above SDR methods are as follows:

(1) For SIR (Li [21]), $\Lambda_{\mathrm{SIR}}(F) = \sum_{k=1}^H p_k(\mu_k - \mu)(\mu_k - \mu)^\mathsf{T}$.

(2) For SAVE (Cook and Weisberg [7]),

$$\Lambda_{\mathrm{SAVE}}(F) = \sum_{k=1}^{H} p_k(\Sigma - \Sigma_k)\Sigma^{-1}(\Sigma - \Sigma_k)^{\mathsf{T}}.$$

(3) For DR (Li and Wang [18]), $\Lambda_{\mathrm{DR}}(F) = 2\Lambda_{\mathrm{DR},1}(F) + 2\Lambda_{\mathrm{DR},2}(F) + 2\Lambda_{\mathrm{DR},3}(F)$, where

$$\begin{aligned}
&\Lambda_{\mathrm{DR},1}(F)\\
&\quad = E\{E[(X-\mu)(X-\mu)^{\mathsf{T}} - \Sigma|\tilde{Y}]\Sigma^{-1}E[(X-\mu)(X-\mu)^{\mathsf{T}} - \Sigma|\tilde{Y}]\},\\
&\Lambda_{\mathrm{DR},2}(F)\\
&\quad = E[E(X-\mu|\tilde{Y})E((X-\mu)^{\mathsf{T}}|\tilde{Y})]\Sigma^{-1}E[E(X-\mu|\tilde{Y})E((X-\mu)^{\mathsf{T}}|\tilde{Y})],\\
&\Lambda_{\mathrm{DR},3}(F)\\
&\quad = E[E((X-\mu)^{\mathsf{T}}|\tilde{Y})\Sigma^{-1}E(X-\mu|\tilde{Y})]E[E(X-\mu|\tilde{Y})E((X-\mu)^{\mathsf{T}}|\tilde{Y})],
\end{aligned}$$

with $\tilde{Y}$ being the discretized $Y$ according to the partition $(J_1, \ldots, J_h)$; that is, $\tilde{Y} = \sum_{k=1}^{h} kI(Y \in J_k)$.

(4) For y-based PHD (Li [22]), $\Lambda_{y\text{-PHD}}(F) = \Sigma_{YXX}\Sigma^{-1}\Sigma_{YXX}$, where

$$\Sigma_{YXX} = E((Y-\nu)(X-\mu)(X-\mu)^{\mathsf{T}}).$$

(5) For r-based PHD (Li [22], Cook [5]), $\Lambda_{r\text{-PHD}}(F) = \Sigma_{RXX}\Sigma^{-1}\Sigma_{RXX}$, where

$$\Sigma_{RXX} = E\{[(Y-\nu) - \beta^{\mathsf{T}}(X-\mu)](X-\mu)(X-\mu)^{\mathsf{T}}\},$$

and $\beta$ is the regression coefficient vector $\Sigma^{-1}\Sigma_{XY}$, with $\Sigma_{XY} = \mathrm{cov}(X, Y)$.

The next proposition gives the explicit forms of $\mathrm{vec}(\Lambda^\star)$ for these SDR methods. The derivations are tedious but straightforward; the details are omitted here. We first write down some simple influence functions:

$$\begin{aligned}
&p_k^\star = D_k - p_k, \qquad \mu^\star = X - \mu, \quad \text{and} \quad \nu^\star = Y - E(Y),\\
&\mu_k^\star = -p_k^{-2}p_k^\star E(XD_k) + p_k^{-1}[XD_k - E(XD_k)],\\
&\Sigma_k^\star = -p_k^{-2}p_k^\star E(XX^{\mathsf{T}}D_k) - p_k^{-1}[XX^{\mathsf{T}}D_k - E(XX^{\mathsf{T}}D_k)] - \mu_k^\star\mu_k^{\mathsf{T}} - \mu_k(\mu_k^\star)^{\mathsf{T}}.
\end{aligned}$$

The influence function of $\beta$ is

$$\beta^\star = (\Sigma^{-1})^\star\Sigma_{XY} + (\Sigma^{-1})\Sigma_{XY}^\star,$$

where $\Sigma_{XY}^\star = XY - E(XY) - (X-\mu)\nu - \mu(Y-\nu)$.

PROPOSITION 4. *The influence functions for the above five reduction functionals are given by the following formulas*:

(1) *For SIR*,

$$\begin{aligned}
\mathrm{vec}(\Lambda_{\mathrm{SIR}}^\star) = &\sum_{k=1}^{H}(\mu_k - \mu)\otimes(\mu_k - \mu)p_k^\star\\
&+ [p_k(\mu_k - \mu)\otimes I_p + I_p \otimes p_k(\mu_k - \mu)](\mu_k^\star - \mu^\star).
\end{aligned}$$

(2) *For SAVE*,

$$\begin{aligned}\operatorname{vec}\bigl(\Lambda^{\star}_{\mathrm{SAVE}}\bigr) &= \sum_{k=1}^{H}\bigl[(\Sigma-\Sigma_k)\otimes(\Sigma-\Sigma_k)\bigr]\operatorname{vec}\bigl(\Sigma^{-1}\bigr)p_k^{\star}\\ &\quad + p_k\bigl[(\Sigma-\Sigma_k)\otimes(\Sigma-\Sigma_k)\bigr]\operatorname{vec}\bigl[\bigl(\Sigma^{-1}\bigr)^{\star}\bigr]\\ &\quad + p_k\bigl[(\Sigma-\Sigma_k)\Sigma^{-1}\otimes I_p + I_p\otimes(\Sigma-\Sigma_k)\Sigma^{-1}\bigr]\operatorname{vec}\bigl(\Sigma^{\star}-\Sigma_k^{\star}\bigr),\end{aligned}$$

*where* $\Sigma^{\star}$ *and* $(\Sigma^{-1})^{\star}$ *are as given in Lemma* 1.

(3) *For DR*,

$$\operatorname{vec}\bigl(\Lambda^{\star}_{\mathrm{DR}}\bigr) = 2\operatorname{vec}\bigl(\Lambda^{\star}_{\mathrm{DR},1}\bigr) + 2\operatorname{vec}\bigl(\Lambda^{\star}_{\mathrm{DR},2}\bigr) + 2\operatorname{vec}\bigl(\Lambda^{\star}_{\mathrm{DR},3}\bigr),$$

*where*

$$\begin{aligned}\operatorname{vec}\bigl(\Lambda^{\star}_{\mathrm{DR},1}\bigr) &= \sum_{k=1}^{H}(A_k\otimes A_k)\operatorname{vec}\bigl(\Sigma^{-1}\bigr)p_k^{\star} + p_k(A_k\otimes A_k)\operatorname{vec}\bigl[\bigl(\Sigma^{-1}\bigr)^{\star}\bigr]\\ &\quad + p_k\bigl(A_k\Sigma^{-1}\otimes I_p + I_p\otimes A_k\Sigma^{-1}\bigr)\operatorname{vec}\bigl(A_k^{\star}\bigr),\\ \operatorname{vec}\bigl(\Lambda^{\star}_{\mathrm{DR},2}\bigr) &= \bigl(B\Sigma^{-1}\otimes I_p + I_p\otimes B\Sigma^{-1}\bigr)\operatorname{vec}\bigl(B^{\star}\bigr) + (B\otimes B)\operatorname{vec}\bigl(\bigl(\Sigma^{-1}\bigr)^{\star}\bigr),\\ \operatorname{vec}\bigl(\Lambda^{\star}_{\mathrm{DR},3}\bigr) &= C^{\star}\operatorname{vec}(B) + C\operatorname{vec}\bigl(B^{\star}\bigr),\end{aligned}$$

*in which* $A_k = E[(X-\mu)(X-\mu)^{\mathsf{T}} - \Sigma | Y\in J_k]$, $B = \sum_{k=1}^{H} p_k(\mu_k-\mu)$, *and* $C = \sum_{k=1}^{H} p_k(\mu_k-\mu)^{\mathsf{T}}\Sigma^{-1}(\mu_k-\mu)$, *with the influence functions*

$$\begin{aligned}A_k^{\star} &= -p_k^{-2}p_k^{\star}E\bigl(XX^{\mathsf{T}}D_k\bigr) + p_k^{-1}\bigl[XX^{\mathsf{T}}D_k - E\bigl(XX^{\mathsf{T}}D_k\bigr)\bigr]\\ &\quad - \mu_k^{\star}\mu^{\mathsf{T}} - \mu_k\mu^{*\mathsf{T}} + \mu^{\star}\mu^{\mathsf{T}} + \mu\mu^{*\mathsf{T}} - \Sigma^{\star},\\ B^{\star} &= \sum_{k=1}^{H} p_k^{\star}(\mu_k-\mu)(\mu_k-\mu)^{\mathsf{T}} + p_k\bigl(\mu_k^{\star}-\mu^{\star}\bigr)(\mu_k-\mu)^{\mathsf{T}}\\ &\quad + p_k(\mu_k-\mu)\bigl(\mu_k^{\star}-\mu^{\star}\bigr)^{\mathsf{T}},\\ C^{\star} &= \sum_{k=1}^{H} p_k^{\star}(\mu_k-\mu)^{\mathsf{T}}\Sigma^{-1}(\mu_k-\mu) + p_k\bigl(\mu_k^{\star}-\mu^{\star}\bigr)^{\mathsf{T}}\Sigma^{-1}(\mu_k-\mu)\\ &\quad + p_k(\mu_k-\mu)^{\mathsf{T}}\Sigma^{-1*}(\mu_k-\mu) + p_k(\mu_k-\mu)^{\mathsf{T}}\Sigma^{-1}\bigl(\mu_k^{\star}-\mu^{\star}\bigr).\end{aligned}$$

(4) *For y-based PHD*,

$$\begin{aligned}\operatorname{vec}\bigl(\Lambda^{\star}_{\text{y-PHD}}\bigr) &= \bigl(\Sigma_{YXX}\Sigma^{-1}\otimes I_p + I_P\otimes\Sigma_{YXX}\Sigma^{-1}\bigr)\operatorname{vec}\bigl(\Sigma^{\star}_{YXX}\bigr)\\ &\quad + (\Sigma_{YXX}\otimes\Sigma_{YXX})\operatorname{vec}\bigl(\Sigma^{-1*}\bigr),\end{aligned}$$

*where*

$$\begin{aligned}\Sigma_{YXX} &= E\bigl[(Y-\nu)(X-\mu)(X-\mu)^{\mathsf{T}}\bigr],\\ \Sigma^{\star}_{YXX} &= YXX^{\mathsf{T}} - E\bigl(YXX^{\mathsf{T}}\bigr) - \nu^{\star}E\bigl(XX^{\mathsf{T}}\bigr) - \nu\bigl[XX^{\mathsf{T}} - E\bigl(XX^{\mathsf{T}}\bigr)\bigr]\\ &\quad - \mu^{\star}E\bigl(YX^{\mathsf{T}}\bigr) - \mu\bigl[YX^{\mathsf{T}} - E\bigl(YX^{\mathsf{T}}\bigr)\bigr] - \bigl[YX - E(YX)\bigr]\mu^{\mathsf{T}}\\ &\quad - E(YX)\mu^{*\mathsf{T}} + \nu^{\star}\mu\mu^{\mathsf{T}} + \nu\mu^{\star}\mu^{\mathsf{T}} + \nu\mu\mu^{*\mathsf{T}}.\end{aligned}$$

(5) *For r-based PHD*,

$$\operatorname{vec}(\Lambda^{\star}_{\text{r-PHD}}) = (\Sigma_{RXX}\Sigma^{-1} \otimes I_p + I_p \otimes \Sigma_{RXX}\Sigma^{-1})\operatorname{vec}(\Sigma^{\star}_{RXX}) + (\Sigma_{RXX} \otimes \Sigma_{RXX})\operatorname{vec}(\Sigma^{-1*}),$$

*where the matrix* $\Sigma_{RXX}$ *is defined as* $\Sigma_{YXX} - R$, *and*

$$R = E(XX^{\mathsf{T}}\beta X^{\mathsf{T}}) - E(XX^{\mathsf{T}})\beta\mu^{\mathsf{T}} - \mu\beta^{\mathsf{T}}E(XX^{\mathsf{T}}) - E(X\mu^{\mathsf{T}}\beta X^{\mathsf{T}}) + 2E(X\mu^{\mathsf{T}}\beta\mu^{\mathsf{T}}).$$

*The influence function of* $\Sigma_{RXX}$ *is*

$$\operatorname{vec}(\Sigma^{\star}_{RXX}) = \operatorname{vec}(\Sigma^{\star}_{YXX}) - \operatorname{vec}(R^{\star}),$$
$$\operatorname{vec}(R^{\star}) = R^{\star}_1 - R^{\star}_2 - R^{\star}_3 - R^{\star}_4 + R^{\star}_5,$$

*where*

$$\begin{aligned}
R^{\star}_1 &= \{X \otimes (XX^{\mathsf{T}}) - E[X \otimes (XX^{\mathsf{T}})]\}\beta + E(X \otimes XX^{\mathsf{T}})\beta^{\star},\\
R^{\star}_2 &= \{I_p \otimes [XX^{\mathsf{T}} - E(XX^{\mathsf{T}})]\}(\mu \otimes \beta) + [I_p \otimes E(XX^{\mathsf{T}})]\mu^{\star} \otimes \beta\\
&\quad + [I_p \otimes E(XX^{\mathsf{T}})]\mu \otimes \beta^{\star},\\
R^{\star}_3 &= \{[XX^{\mathsf{T}} - E(XX^{\mathsf{T}})] \otimes I_p\}(\beta \otimes \mu) + \{[E(XX^{\mathsf{T}})] \otimes I_p\}(\beta^{\star} \otimes \mu)\\
&\quad + \{[E(XX^{\mathsf{T}})] \otimes I_p\}(\beta \otimes \mu^{\star}),\\
R^{\star}_4 &= [X \otimes X - E(X \otimes X)]\mu^{\mathsf{T}}\beta + E(X \otimes X)\mu^{*\mathsf{T}}\beta + E(X \otimes X)\mu^{\mathsf{T}}\beta^{\star},\\
R^{\star}_5 &= 2[(\mu\beta^{\mathsf{T}} \otimes I_p)\operatorname{vec}(\mu^{\star}\mu^{\mathsf{T}}) + (\mu\beta^{\mathsf{T}} \otimes I_p)\operatorname{vec}(\mu\mu^{*\mathsf{T}})\\
&\quad + (I_p \otimes \mu\mu^{\mathsf{T}})\operatorname{vec}(\beta^{\star}\mu^{\mathsf{T}}) + (I_p \otimes \mu\mu^{\mathsf{T}})\operatorname{vec}(\beta\mu^{*\mathsf{T}})].
\end{aligned}$$

The five influence functions in Proposition 4 can be easily estimated by replacing, whenever applicable, the expectation $E(\cdot)$ with the sample average $E_n(\cdot)$. We can then substitute into the formulas for $B$ and $\Gamma$ in Corollary 1 to obtain the estimated asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$.

**5. Post dimension reduction inference.** In this section, we develop the formal statistical inference procedures for $\theta$ based on the asymptotic distribution of $\hat{\theta} = \Phi(F_n, \hat{\eta})$ derived in Sections 2 through 4. First, we consider the confidence interval for an arbitrary linear combination of $\theta$. Let $c \in \mathbb{R}^s$ be a vector and let $z_\alpha$ be the $(1-\alpha)$th percentile of the standard normal distribution. Because $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{D}} N(0, \Gamma)$, the interval $(c^{\mathsf{T}}\hat{\theta} - z_{\alpha/2}\sqrt{c^{\mathsf{T}}\Gamma c},\ c^{\mathsf{T}}\hat{\theta} + z_{\alpha/2}\sqrt{c^{\mathsf{T}}\Gamma c})$ covers the true parameter $\theta_0$ with probability tending to $1-\alpha$. Therefore, by Slutsky's theorem, the asymptotic $(1-\alpha)$-level confidence interval for $\theta$ is

$$\big(c^{\mathsf{T}}\hat{\theta} - z_{\frac{\alpha}{2}}\sqrt{c^{\mathsf{T}}\widehat{\Gamma}c},\ c^{\mathsf{T}}\hat{\theta} + z_{\frac{\alpha}{2}}\sqrt{c^{\mathsf{T}}\widehat{\Gamma}c}\big),$$

where $\widehat{\Gamma} = \Gamma(\hat{\eta}, \hat{\theta})$ is an estimate of $\Gamma$ as defined in Corollary 1.

Next, we consider testing the null hypothesis

$$H_0 : h(\theta) = h(\theta_0),$$

where $h : \mathbb{R}^s \to \mathbb{R}^k$ is a differentiable function. We use the function $h$ to accommodate the situation where only part of the parameter $\theta$, for example, the first component of $\theta$, is of interest. For power assessment, we consider the local alternative hypothesis

$$H_{1,n}(\lambda) : h(\theta) = h\left(\theta_0 + \frac{\lambda}{\sqrt{n}}\right),$$

where $\lambda$ is a fixed vector in $\mathbb{R}^s$. Let $H(\theta) = \partial h^{\mathsf{T}}(\theta)/\partial\theta \in \mathbb{R}^{s\times k}$ be the gradient matrix of $h$ at $\theta$, $\theta_n = \theta_0 + \lambda/\sqrt{n}$, $\widehat{H} = H(\hat{\theta})$, and $H = H(\theta_0)$. We propose the following Wald-type test statistic

$$T = \sqrt{n}\big[h(\hat{\theta}) - h(\theta_0)\big]\big(\widehat{H}^{\mathsf{T}}\widehat{\Gamma}\,\widehat{H}\big)^{-1}\sqrt{n}\big[h(\hat{\theta}) - h(\theta_0)\big].$$

The next theorem gives the asymptotic distributions of $T$ under the null and the local alternative distribution. In the following, convergence in distribution under the null hypothesis is written as $\xrightarrow[\theta_0]{\mathcal{D}}$, while convergence in distribution under the local alternative hypothesis is written as $\xrightarrow[\theta_n]{\mathcal{D}}$.

THEOREM 3. *Suppose the conditions in Theorem 1 are satisfied and the matrices* $\Gamma$ *and* $H$ *are nonsingular, then*

$$T \xrightarrow[\theta_0]{\mathcal{D}} \chi^2_k. \tag{5.1}$$

*Suppose, moreover, that* $\hat{\theta}$ *is a regular estimator, then*

$$T \xrightarrow[\theta_n]{\mathcal{D}} \chi^2_k\big(\lambda^{\mathsf{T}}H\Gamma H^{\mathsf{T}}\lambda\big). \tag{5.2}$$

PROOF. By Corollary 1 and the delta method, we have

$$\sqrt{n}\big[h(\hat{\theta}) - h(\theta_0)\big] \xrightarrow[\theta_0]{\mathcal{D}} N\big(0, H^{\mathsf{T}}\Gamma H\big),$$

which implies (5.1).

Since $\hat{\theta}$ is a regular estimator and $h$ is differentiable, the asymptotic distribution of $\sqrt{n}[h(\hat{\theta}) - h(\theta_n)]$ under $H_{1,n}(\lambda)$ is the same as the asymptotic distribution of $\sqrt{n}[h(\hat{\theta}) - h(\theta_0)]$ under $H_0$. Next, we decompose $\sqrt{n}[h(\hat{\theta}) - h(\theta_0)]$ as

$$\begin{aligned}\sqrt{n}\big[h(\hat{\theta}) - h(\theta_0)\big] &= \sqrt{n}\big[h(\hat{\theta}) - h(\theta_n)\big] + \sqrt{n}\big[h(\theta_n) - h(\theta_0)\big]\\ &= \sqrt{n}\big[h(\hat{\theta}) - h(\theta_n)\big] + H^{\mathsf{T}}\lambda + o\big(n^{-1/2}\big).\end{aligned}$$

By Slutsky's theorem,

$$\sqrt{n}\big[h(\hat{\theta}) - h(\theta_0)\big] \xrightarrow[\theta_n]{\mathcal{D}} N\big(H^{\mathsf{T}}\lambda, H^{\mathsf{T}}\Gamma H\big),$$

which implies

$$\sqrt{n}\big[H^{\mathsf{T}}\Gamma H\big]^{-\frac{1}{2}}\big[h(\hat{\theta}) - h(\theta_0)\big] \xrightarrow[\theta_n]{\mathcal{D}} N\big(\big(H^{\mathsf{T}}\Gamma H\big)^{-\frac{1}{2}}H^{\mathsf{T}}\lambda, I_k\big).$$

Together we have

$$\sqrt{n}\big[h(\hat{\theta}) - h(\theta_0)\big]\big[H^{\mathsf{T}}\Gamma H\big]^{-1}\big[h(\hat{\theta}) - h(\theta_0)\big] \xrightarrow[\theta_n]{\mathcal{D}} \chi^2_k\big[\lambda^{\mathsf{T}}H\big(H^{\mathsf{T}}\Gamma H\big)^{-1}H^{\mathsf{T}}\lambda\big].$$

Applying Slutsky's theorem again, we obtain (5.2). □

We briefly comment that the requirement $\hat{\theta}$ is a regular estimator is rather mild, and is satisfied by most estimators. See Bickel et al. [1] and Van der Vaart [36].

**6. Simulations.** We next investigate the finite-sample performance of our post dimension reduction inference method, and compare it with the naive inference method that pretends $\hat{\eta}^{\mathsf{T}}X$ were the true predictor. As discussed in Section 2.3, the asymptotic covariances of the two methods are $\Gamma(\eta_0, \theta_0)$ and $\tilde{\Gamma}(\eta_0, \theta_0)$, respectively. Given the data, $\Gamma(\eta_0, \theta_0)$ and $\tilde{\Gamma}(\eta_0, \theta_0)$ are estimated by $\Gamma(\hat{\eta}, \hat{\theta})$ and $\tilde{\Gamma}(\hat{\eta}, \hat{\theta})$. We consider five dimension reduction methods, SIR, SAVE, DR, y-PHD and r-PHD, and one estimation method, GMM. For GMM, let $m(\eta^{\mathsf{T}}X, \theta)$ denote the mean function, which is the same as the median function in our simulations as a symmetric error distribution is employed, and we set

$$g_1(\theta, \eta^{\mathsf{T}}X, Y) = Y - m(\eta^{\mathsf{T}}X, \theta), \qquad g_2(\theta, \eta^{\mathsf{T}}X, Y) = \mathrm{I}(Y \le m(\eta^{\mathsf{T}}X, \theta)) - 1/2.$$

That is, the GMM combines mean regression and median regression, which strikes a balance between efficiency and robustness. We compare the performance in terms of the coverage probability of confidence interval and the local power in hypothesis testing.

6.1. *Comparison of confidence interval.* For confidence interval comparison, we consider two models. The first model is

$$\text{Model I:} \quad Y = \theta_1(\eta^{\mathsf{T}}X) + \theta_2(\eta^{\mathsf{T}}X)^2 + \sigma\varepsilon,$$

where $X \sim N(0, I_5)$, $\varepsilon \sim N(0, 1)$, $X \perp\!\!\!\perp \varepsilon$, $\theta_1 = \theta_2 = 1$, $\sigma = 0.5, 1$, the predictor dimension $p = 5$, and the sample size $n = 300, 400, 800, 1200$. In this example, $\mathcal{S}_{Y|X} = \mathrm{span}(\eta)$ with $\eta = (1, 0, 0, 0, 0)^{\mathsf{T}}$. For the number of slices for SIR, SAVE and DR, the general rule of thumb is to choose a larger value for SIR, and a smaller value for SAVE and DR (Li [15]). In our simulations, we have chosen $H = 20$ for SIR, $H = 2$ for SAVE and $H = 8$ for DR. After obtaining $\hat{\eta}$ and $\hat{\theta}(\hat{\eta})$, we calculate the 95% confidence intervals for $\theta_1$ and $\theta_2$. We report the coverage probabilities of the two methods based on 200 data replications in Table 1. We see that the coverage probability from the naive method is considerably smaller than the nominal value, whereas the coverage probability from our proposed method is much closer. Table 1

TABLE 1
*Coverage probability of confidence interval for $\theta_1$ and $\theta_2$ in model* I

| | | | SIR | | SAVE | | DR | | y-PHD | | r-PHD | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | $\Theta$ | $\sigma^2$ | $\Gamma$ | $\tilde{\Gamma}$ | $\Gamma$ | $\tilde{\Gamma}$ | $\Gamma$ | $\tilde{\Gamma}$ | $\Gamma$ | $\tilde{\Gamma}$ | $\Gamma$ | $\tilde{\Gamma}$ |
| 300 | $\theta_1$ | 0.5 | 0.96 | 0.82 | 0.96 | 0.81 | 0.96 | 0.83 | 0.95 | 0.81 | 0.96 | 0.82 |
| | | 1 | 0.95 | 0.79 | 0.95 | 0.78 | 0.94 | 0.81 | 0.96 | 0.80 | 0.93 | 0.79 |
| | $\theta_2$ | 0.5 | 0.93 | 0.80 | 0.94 | 0.80 | 0.94 | 0.79 | 0.96 | 0.81 | 0.96 | 0.81 |
| | | 1 | 0.94 | 0.78 | 0.93 | 0.80 | 0.94 | 0.80 | 0.94 | 0.79 | 0.96 | 0.79 |
| 400 | $\theta_1$ | 0.5 | 0.95 | 0.85 | 0.96 | 0.85 | 0.95 | 0.85 | 0.96 | 0.84 | 0.95 | 0.85 |
| | | 1 | 0.96 | 0.81 | 0.94 | 0.83 | 0.93 | 0.82 | 0.93 | 0.81 | 0.93 | 0.83 |
| | $\theta_2$ | 0.5 | 0.95 | 0.83 | 0.94 | 0.85 | 0.94 | 0.84 | 0.96 | 0.84 | 0.96 | 0.84 |
| | | 1 | 0.95 | 0.82 | 0.94 | 0.81 | 0.94 | 0.81 | 0.94 | 0.82 | 0.94 | 0.81 |
| 800 | $\theta_1$ | 0.5 | 0.96 | 0.88 | 0.96 | 0.89 | 0.94 | 0.89 | 0.96 | 0.88 | 0.95 | 0.88 |
| | | 1 | 0.96 | 0.88 | 0.95 | 0.87 | 0.93 | 0.87 | 0.95 | 0.86 | 0.93 | 0.86 |
| | $\theta_2$ | 0.5 | 0.96 | 0.87 | 0.94 | 0.88 | 0.96 | 0.87 | 0.95 | 0.86 | 0.95 | 0.87 |
| | | 1 | 0.93 | 0.86 | 0.96 | 0.85 | 0.94 | 0.86 | 0.93 | 0.85 | 0.94 | 0.85 |
| 1200 | $\theta_1$ | 0.5 | 0.95 | 0.92 | 0.96 | 0.91 | 0.96 | 0.91 | 0.96 | 0.92 | 0.96 | 0.91 |
| | | 1 | 0.94 | 0.90 | 0.94 | 0.90 | 0.94 | 0.88 | 0.95 | 0.89 | 0.96 | 0.90 |
| | $\theta_2$ | 0.5 | 0.94 | 0.92 | 0.95 | 0.91 | 0.95 | 0.91 | 0.96 | 0.91 | 0.95 | 0.91 |
| | | 1 | 0.94 | 0.91 | 0.93 | 0.88 | 0.94 | 0.90 | 0.95 | 0.90 | 0.96 | 0.89 |

TABLE 2
*Coverage probability of confidence interval for $\theta_1$ in model* II

| | | | SIR | | SAVE | | DR | | y-PHD | | r-PHD | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | $\Theta$ | $\sigma^2$ | $\Gamma$ | $\tilde{\Gamma}$ | $\Gamma$ | $\tilde{\Gamma}$ | $\Gamma$ | $\tilde{\Gamma}$ | $\Gamma$ | $\tilde{\Gamma}$ | $\Gamma$ | $\tilde{\Gamma}$ |
| 300 | $\theta_1$ | 0.5 | 0.94 | 0.85 | 0.93 | 0.83 | 0.95 | 0.82 | 0.93 | 0.83 | 0.96 | 0.83 |
| | | 1 | 0.93 | 0.84 | 0.94 | 0.80 | 0.94 | 0.81 | 0.95 | 0.80 | 0.93 | 0.81 |
| 400 | $\theta_1$ | 0.5 | 0.95 | 0.86 | 0.95 | 0.87 | 0.95 | 0.86 | 0.93 | 0.86 | 0.95 | 0.85 |
| | | 1 | 0.94 | 0.84 | 0.95 | 0.84 | 0.94 | 0.87 | 0.94 | 0.84 | 0.94 | 0.85 |
| 800 | $\theta_1$ | 0.5 | 0.95 | 0.90 | 0.94 | 0.88 | 0.96 | 0.89 | 0.96 | 0.90 | 0.95 | 0.90 |
| | | 1 | 0.94 | 0.87 | 0.93 | 0.86 | 0.93 | 0.90 | 0.95 | 0.88 | 0.94 | 0.89 |
| 1200 | $\theta_1$ | 0.5 | 0.96 | 0.91 | 0.95 | 0.91 | 0.94 | 0.92 | 0.94 | 0.91 | 0.96 | 0.92 |
| | | 1 | 0.95 | 0.92 | 0.94 | 0.90 | 0.95 | 0.90 | 0.95 | 0.89 | 0.95 | 0.90 |

also shows that the coverage probability for the naive method becomes closer to the nominal value as the sample size increases, but it does not converge to the nominal value.

The second model is

$$\text{Model II:} \quad Y = \theta_1 \frac{\eta^{\mathsf{T}} X}{(\eta^{\mathsf{T}} X + 2)^2 + 0.1} + \sigma\varepsilon,$$

where $X \sim N(0, I_{10})$, $X \perp\!\!\!\perp \varepsilon$, $\theta_1 = 1$, and $\eta = (1, 0, \ldots, 0)^{\mathsf{T}}$. The rest of the setup is the same as model I. We report the coverage probabilities in Table 2. Again, the coverage probability of our method is much closer to 95% than the naive method.

Since the true model is known in the simulation experiments, we can also estimate $(\eta, \theta)$ and make inference about them directly using the maximum likelihood method without going through dimension reduction. It would be informative to compare this "oracle" inference method with the naive and objective inference methods. We have carried out this comparison using Model I, with $n = 400$. We read off the standard errors for $\hat{\theta}_1^{\mathrm{MLE}}, \hat{\theta}_2^{\mathrm{MLE}}$ from the asymptotic variance matrix of $(\hat{\eta}^{\mathrm{MLE}}, \hat{\theta}^{\mathrm{MLE}})$, which is the inverted Fisher information evaluated at the MLE. We also compute the standard errors for the $(\hat{\theta}_1, \hat{\theta}_2)$ obtained by SIR + GMM as described above, using the naive and objective inference methods. We repeat the process 200 times to compute the average standard errors. The results are reported in Table 3.

In theory, we would expect the standard errors for $\hat{\theta}_1^{\mathrm{MLE}}$ and $\hat{\theta}_2^{\mathrm{MLE}}$ using the oracle inference method to be smaller than their counterparts for $\hat{\theta}_1$ and $\hat{\theta}_2$ using the objective method, because MLE is asymptotically efficient. But this is not necessarily true in finite-sample, as indicated by our results. Also, Table 3 shows that both the objective and oracle mean standard errors are substantially larger than their counterparts by the naive method, which is not surprising because the naive method claims more information than it actually possesses.

TABLE 3
*Standard errors for $\theta_1$ and $\theta_2$ in model I and comparison to the oracle method*

| Parameter | Naive | Objective | Oracle |
|---|---|---|---|
| $\theta_1$ | 0.03 | 0.06 | 0.09 |
| $\theta_2$ | 0.02 | 0.04 | 0.03 |

6.2. *Comparison of local power.* For power comparison, we again consider two models. The first model is

$$\text{Model III:} \quad Y = \theta_1 (\eta_1^\top X)^2 + \theta_2 \exp(\eta_2^\top X) + \sigma\varepsilon,$$

where $X \sim N(0, I_{10})$, $\varepsilon \sim N(0,1)$, $X \perp\!\!\!\perp \varepsilon$, $\theta_1 = \theta_2 = 1$, $\sigma = 0.5$, $p = 10$, and $n = 300, 500, 800, 1200$ with 50 replications. In this example, $\mathcal{S}_{Y|X} = \text{span}(\eta_1, \eta_2)$ with $\eta_1 = (1, 0, \ldots, 0)^\top$ and $\eta_2 = (0, 1, 0, \ldots, 0)^\top$. We consider the pair of hypotheses

$$H_0 : \theta_1 = 0 \quad \text{vs} \quad H_1 : \theta_1 \neq 0$$

which amounts to taking $h(\theta_1, \theta_2) = \theta_1$ in Section 5. The asymptotic power is computed as in (5.2). Figure 1 reports this asymptotic power as a function of the local parameter $\lambda$ when the sample size is 500, with one panel corresponding to one of the five SDR methods. Figures A.1, A.2, A.3 in the online supplementary material present the results for $n = 300, 800, 1200$, respectively. It is seen that the powers of the naive method, as shown by the red curves, are higher than those by our proposed method, as shown by the blue curves. This reflects that the naive method yields an overly optimistic power, as it does not take into account the estimation error induced by the dimension reduction step. Furthermore, by comparing Figures 1, A.1, A.2 and A.3, we observe that the difference between the local powers of the naive and the objective methods tends to be smaller as the sample size increases, which echoes the pattern in the comparison of confidence intervals.

Our second model for the local power comparison is

$$\text{Model IV:} \quad Y = \theta_1 \frac{\eta_1^\top X}{(\eta_2^\top X + 1)^2 + 0.5} + \sigma\varepsilon,$$

where $\theta_1 = 1$, and the rest of the setup is the same as model III. Figure 2 reports the results for $n = 500$. The same pattern is observed as in model III. Figures A.4, A.5 and A.6 in the online supplementary material present the results for $n = 300, 800$ and 1200, respectively.

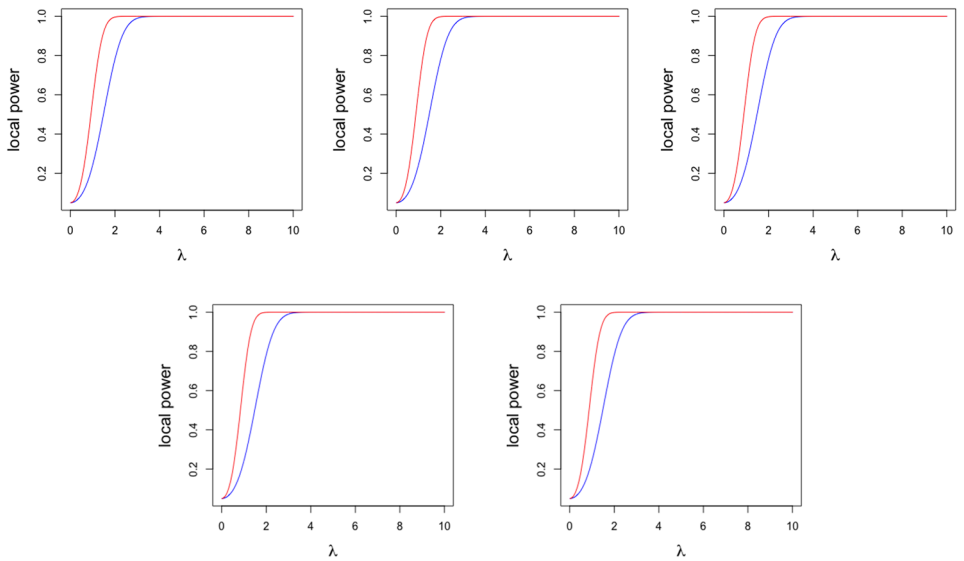


FIG. 1. *Local power of hypothesis testing in model* III *with sample size* $n = 500$. *The five panels, left to right, top to bottom, correspond to the results from five SDR methods, SIR, SAVE, DR, y-PHD and r-PHD. The red curve denotes the naive inference method, and the blue curve denotes our proposed inference method.*

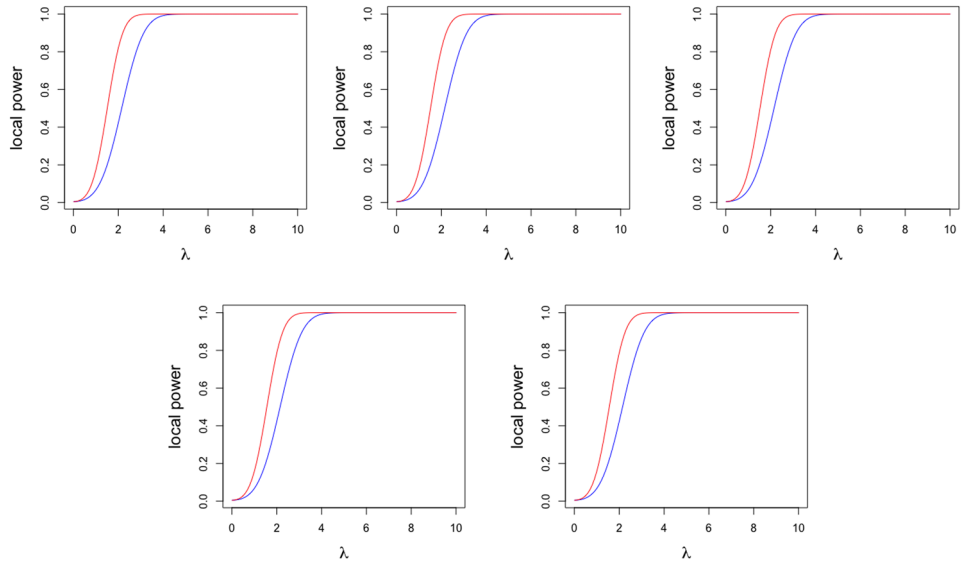


FIG. 2. *Local power of hypothesis testing in model* IV *with sample size* $n = 500$. *The five panels, left to right, top to bottom, correspond to the results from five SDR methods, SIR, SAVE, DR, y-PHD and r-PHD. The red curve denotes the naive inference method, and the blue curve denotes our proposed inference method.*

**7. Application.** We use the BigMac dataset to illustrate our post dimension reduction inference. The data concerns the relation between the minimum labor to buy a McDonald BigMac and fries, which serves as the response variable, and $p = 9$ economic predictors: minimum labor to buy one kilogram bread, lowest cost of 10k public transit, electrical engineer annual salary, tax rate paid by engineer, annual cost of 19 services, primary teacher salary, tax rate paid by primary teacher, average days of vacation per year and average hours of work per year. The data is at http://www.stat.umn.edu/arc/software.html. Before the dimension reduction analysis, we applied the box-cox transformation to each individual predictor.

The sequential tests based on SIR yielded the p-values, 0.02, 0.20, 0.77, for the hypotheses $q = 0$ versus $q > 0$, $q = 1$ versus $q > 1$ and $q = 2$ versus $q > 2$, respectively, suggesting that the dimension of the central subspace is one and a single linear combination is sufficient to fully capture the relationship between the response and the nine predictors. Figure 3 shows the scatter plot of the response versus the estimated sufficient predictor based on SIR.

The scatter plot shows a clear nonlinear trend and a possible heteroscedastic pattern. For this reason, we consider the following model:

$$Y = \theta_0 + \theta_1 \eta^{\mathsf{T}} X + \theta_2 (\eta^{\mathsf{T}} X)^2 + (\theta_3 + \theta_4 \eta^{\mathsf{T}} X)\varepsilon,$$

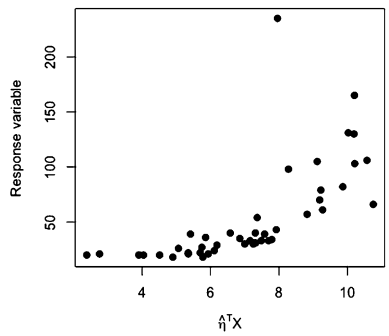


FIG. 3. *Response versus the first SIR predictor in BigMac data.*

where $\varepsilon \sim N(0, 1)$. Based on this model, we aim to address two questions: First, is the nonlinear trend in Figure 3 significant? Second, is the heteroscedasticity in Figure 3 significant? These lead to the following two pairs of hypotheses:

$$H_0^{(1)} : \theta_2 = 0 \quad \text{vs} \quad H_1^{(1)} : \theta_2 \neq 0,$$
$$H_0^{(2)} : \theta_4 = 0 \quad \text{vs} \quad H_1^{(2)} : \theta_4 \neq 0.$$

To test these hypotheses, we applied the naive method and the post dimension reduction method to the five SDR methods combined with the differential estimation equations. We use each method to construct confidence intervals for $\theta_2$ and $\theta_4$. The estimating equations are 5-dimensional $g(\theta, \eta^{\mathsf{T}} X, Y)$ obtained by differentiating with respect to $\theta$ the objective function

$$\left[\frac{Y - \theta_0 - \theta_1(\eta^{\mathsf{T}} X) - \theta_2(\eta^{\mathsf{T}} X)^2}{\theta_3 + \theta_4(\eta^{\mathsf{T}} X)}\right]^2.$$

Figure 4 shows the confidence intervals for $\theta_2$ (the upper panel) and $\theta_4$ (the bottom panel) obtained by different methods. In each plot, the left bar corresponds to the naive inference method, and the right one our proposed inference method. It is seen that, for $\theta_2$, the confidence intervals produced by both inference methods do not cover 0, a clear evidence for the nonlinearity. For $\theta_4$, none of the confidence intervals covers 0, a strong evidence for the heteroscedasticity. Moreover, the confidence intervals produced by the naive method are consistently narrower than those by our objective inference method.

To compare the local powers of the naive method and the post dimension reduction method, we applied them to the five SDR methods combined with the GMM estimation method. The GMM is based on two 5-dimensional estimating equations, with the first one,

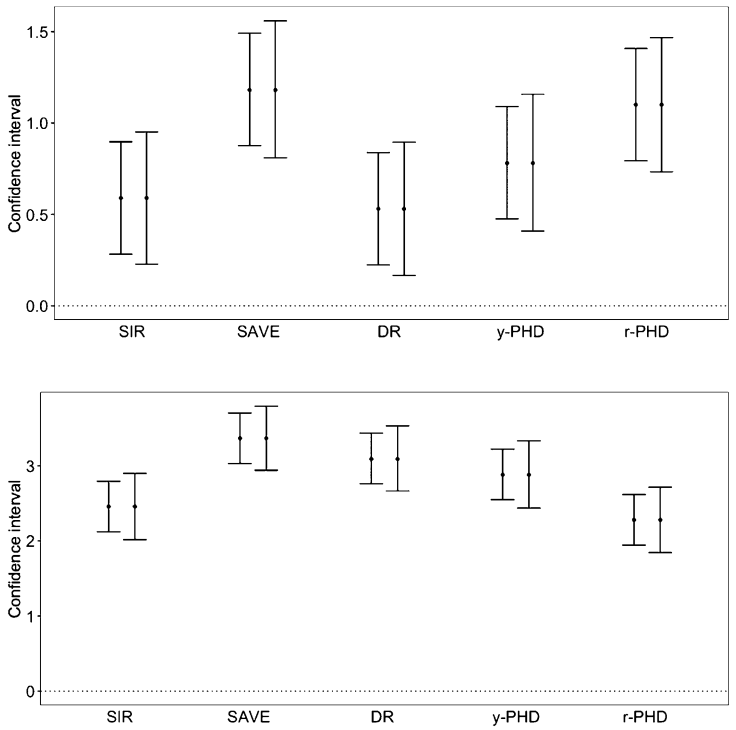


FIG. 4. *Confidence intervals for $\theta_2$ (upper panel) and $\theta_4$ (lower panel) in the BigMac data analysis.*

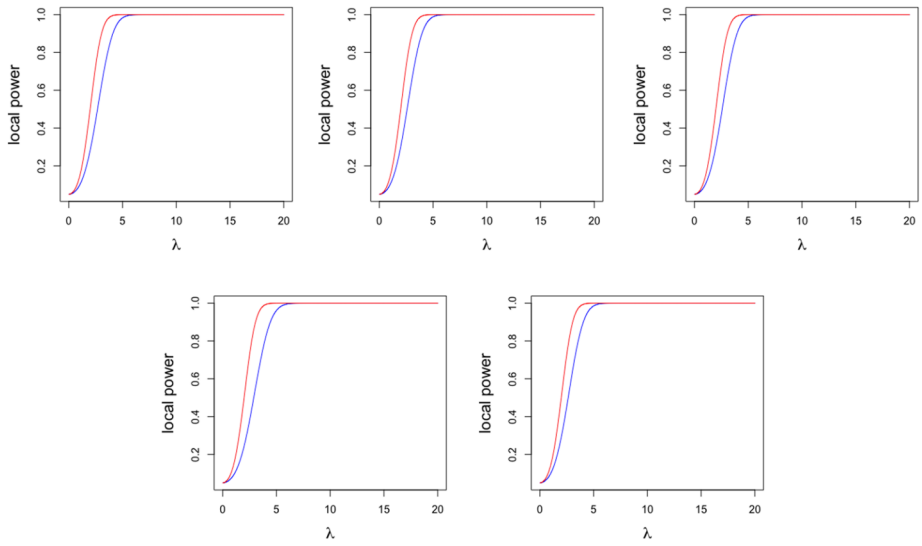


FIG. 5. *Local power for $\theta_2$ in the BigMac data analysis. The five panels, left to right, top to bottom, correspond to the results from five SDR methods, SIR, SAVE, DR, y-PHD and r-PHD. The red curve denotes the naive inference method, and the blue curve denotes our proposed inference method.*

$g_1(\theta, \eta^\top X, Y)$, being obtained by differentiating the objective function $[Y - \theta_0 - \theta_1(\eta^\top X) - \theta_2(\eta^\top X)^2]^2$ with respect to $\theta$, and the second one, $g_2(\theta, \eta^\top X, Y)$, being the the function

$$[Y - \theta_0 - \theta_2(\eta^\top X) - \theta_2(\eta^\top X)^2]^2 - (\theta_3 + \theta_4 \eta^\top X)^2,$$

which is derived from the second moment assumption. Figure 5 shows the local powers of the five SDR methods based on the GMM. To save space, we only report the results for $\theta_2$; the results for $\theta_4$ exhibit a similar pattern. Again, the naive method yields an overly optimistic power compared with the objective method, which agrees with what we have observed in the simulations.

While the above analysis shows substantial differences in the confidence intervals by the naive and the objective inference methods, none of them is large enough to make the parameter statistically significant by one method and insignificant by the other. This turns out to be the case for the intercept parameter $\theta_1$ when DR is used for dimension reduction. Figure 6 shows the confidence interval for $\theta_1$ by the five SDR methods and the two inference methods. For DR, the naive inference method produces a confidence interval that does not contain 0, whereas the objective inference method produces a confidence interval that does. Thus $\theta_1$ is statistically significant by the naive method but insignificant by the objective method.

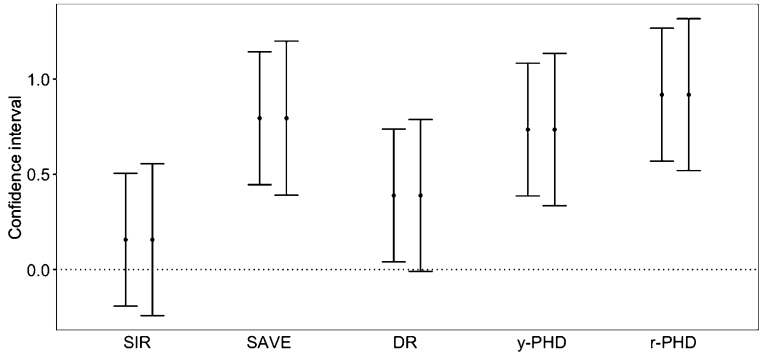


FIG. 6. *Confidence interval for $\theta_1$ in the BigMac data analysis.*

**8. Conclusions.** Despite the extensive development of sufficient dimension reduction in the past three decades, the critical step of post dimension reduction inference has never been taken—at least not in a systematic and rigorous manner. SDR is not complete without a proper post reduction inference procedure that takes the estimation error induced in the dimension reduction step into the subsequent model estimation step. We fill this gap by developing a general post dimension reduction inference framework that is adaptive to a multitude of dimension reduction and model estimation methods. We derive the inference procedures for confidence interval and hypothesis testing based on a combination of commonly used SDR and model building methods.

The framework laid out in this paper also opens the door for developing objective inference procedures for a much broader class of dimension reduction problems than considered here. Potential extensions include unsupervised dimension reduction methods such as principal components analysis and independent components analysis (Hyvärinen, Karhunen and Oja [12]), sparse sufficient dimension reduction methods (Li [23], Bondell and Li [2], Chen, Zou and Cook [3], Wang and Yin [38]) and nonparametric sufficient dimension reduction methods (Xia et al. [40], Xia [39]). A particularly promising direction of extension is to the semiparametrically efficient SDR methods developed in Ma and Zhu [26–28] and Luo, Li and Yin [25]. For these methods, the influence function can be readily developed from the efficient score, and it is plausible that semiparametric efficiency for sufficient dimension reduction can be inherited, to some degree at least, by the post dimension reduction inference procedure.

Beyond the asymptotic normality-based procedures considered in this paper, it is also useful to develop nonparametric inference procedures for post dimension reduction inference. For example, it is possible to employ the empirical likelihood approach (Owen [30, 31]) to conduct post dimension inference. In this direction, Li, Zhu and Zhu [20] proposed an empirical likelihood inference procedure for the single-index model, and the ideas and techniques there might be adaptable to the current setting. The full potential and scope of the general framework of post dimension reduction inference will be explored in future research.

**Acknowledgments.** We thank the Editor, the Associate Editor and two referees for their prompt and thoughtful reviews, which contain many useful comments and suggestions that have helped us to improve this paper.

The second author was supported in part by the NSF Grant DMS-1713078.

The third author was supported in part by the National Natural Science Foundation of China 11831008, 11571111.

The fourth author was supported in part by the NSF Grant DMS-1613137 and NIH Grant AG034570.

## SUPPLEMENTARY MATERIAL

**Supplement to "On post dimension reduction statistical inference"** (DOI: 10.1214/19-AOS1859SUPP; .pdf). The supplement provides additional comparisons of local powers, which are cited in the paper.

## REFERENCES

[1] BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models* **2**. Springer, New York. MR1623559

[2] BONDELL, H. D. and LI, L. (2009). Shrinkage inverse regression estimation for model-free variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 287–299. MR2655534 https://doi.org/10.1111/j.1467-9868.2008.00686.x

[3] CHEN, X., ZOU, C. and COOK, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *Ann. Statist.* **38** 3696–3723. MR2766865 https://doi.org/10.1214/10-AOS826

[4] COOK, R. D. (1998). *Regression Graphics*: *Ideas for Studying Regressions Through Graphics*. *Wiley Series in Probability and Statistics*: *Probability and Statistics*. Wiley, New York. MR1645673 https://doi.org/10.1002/9780470316931

[5] COOK, R. D. (1998). Principal Hessian directions revisited. *J. Amer. Statist. Assoc.* **93** 84–100. MR1614584 https://doi.org/10.2307/2669605

[6] COOK, R. D. and NI, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *J. Amer. Statist. Assoc.* **100** 410–428. MR2160547 https://doi.org/10.1198/016214504000001501

[7] COOK, R. D. and WEISBERG, S. (1991). Comment. *J. Amer. Statist. Assoc.* **86** 328–332.

[8] FERNHOLZ, L. T. (2012). *Von Mises Calculus for Statistical Functionals*. *Lecture Notes in Statistics* **19**. Springer, New York. MR0713611 https://doi.org/10.1007/978-1-4612-5604-5

[9] HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50** 1029–1054. MR0666123 https://doi.org/10.2307/1912775

[10] HANSEN, L. P., HEATON, J. and YARON, A. (1996). Finite-sample properties of some alternative GMM estimators. *J. Bus. Econom. Statist.* **14** 262–280.

[11] HE, X., FU, B. and FUNG, W. (2003). Median regression for longitudinal data. *Stat. Med.* **22** 3655–3669.

[12] HYVÄRINEN, A., KARHUNEN, J. and OJA, E. (2004). *Independent Component Analysis* **46**. Wiley, New York.

[13] KIM, K., LI, B., YU, Z. and LI, L. (2020). Supplement to "On post dimension reduction statistical inference." https://doi.org/10.1214/19-AOS1859SUPP.

[14] LI, B. (1993). A deviance function for the quasi-likelihood method. *Biometrika* **80** 741–753. MR1282783 https://doi.org/10.1093/biomet/80.4.741

[15] LI, B. (2018). *Sufficient Dimension Reduction*: *Methods and Applications with R*. *Monographs on Statistics and Applied Probability* **161**. CRC Press, Boca Raton, FL. MR3838449 https://doi.org/10.1201/9781315119427

[16] LI, B., ARTEMIOU, A. and LI, L. (2011). Principal support vector machines for linear and nonlinear sufficient dimension reduction. *Ann. Statist.* **39** 3182–3210. MR3012405 https://doi.org/10.1214/11-AOS932

[17] LI, B. and SONG, J. (2017). Nonlinear sufficient dimension reduction for functional data. *Ann. Statist.* **45** 1059–1095. MR3662448 https://doi.org/10.1214/16-AOS1475

[18] LI, B. and WANG, S. (2007). On directional regression for dimension reduction. *J. Amer. Statist. Assoc.* **102** 997–1008. MR2354409 https://doi.org/10.1198/016214507000000536

[19] LI, B., ZHA, H. and CHIAROMONTE, F. (2005). Contour regression: A general approach to dimension reduction. *Ann. Statist.* **33** 1580–1616. MR2166556 https://doi.org/10.1214/009053605000000192

[20] LI, G.-R., ZHU, L.-P. and ZHU, L.-X. (2010). Adaptive confidence region for the direction in semiparametric regressions. *J. Multivariate Anal.* **101** 1364–1377. MR2609498 https://doi.org/10.1016/j.jmva.2010.02.002

[21] LI, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86** 316–342. MR1137117

[22] LI, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *J. Amer. Statist. Assoc.* **87** 1025–1039. MR1209564

[23] LI, L. (2007). Sparse sufficient dimension reduction. *Biometrika* **94** 603–613. MR2410011 https://doi.org/10.1093/biomet/asm044

[24] LI, L. (2018). Sufficient dimension reduction. In *Wiley StatsRef*: *Statistics Reference Online* 1–8. American Cancer Society, New York. https://doi.org/10.1002/9781118445112.stat08042.

[25] LUO, W., LI, B. and YIN, X. (2014). On efficient dimension reduction with respect to a statistical functional of interest. *Ann. Statist.* **42** 382–412. MR3189490 https://doi.org/10.1214/13-AOS1195

[26] MA, Y. and ZHU, L. (2012). A semiparametric approach to dimension reduction. *J. Amer. Statist. Assoc.* **107** 168–179. MR2949349 https://doi.org/10.1080/01621459.2011.646925

[27] MA, Y. and ZHU, L. (2013). Efficient estimation in sufficient dimension reduction. *Ann. Statist.* **41** 250–268. MR3059417 https://doi.org/10.1214/12-AOS1072

[28] MA, Y. and ZHU, L. (2014). On estimation efficiency of the central mean subspace. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 885–901. MR3271171 https://doi.org/10.1111/rssb.12044

[29] MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*. *Monographs on Statistics and Applied Probability*. CRC Press, London. MR3223057 https://doi.org/10.1007/978-1-4899-3242-6

[30] OWEN, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18** 90–120. MR1041387 https://doi.org/10.1214/aos/1176347494

[31] OWEN, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237–249. MR0946049 https://doi.org/10.1093/biomet/75.2.237

[32] PARK, C. and LINDSAY, B. G. (1999). Robust estimation and tests based on quadratic inference function. Technical report.

[33] QU, A., LINDSAY, B. G. and LI, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika* **87** 823–836. MR1813977 https://doi.org/10.1093/biomet/87.4.823

[34] SHAO, Y., COOK, R. D. and WEISBERG, S. (2007). Marginal tests with sliced average variance estimation. *Biometrika* **94** 285–296. MR2331487 https://doi.org/10.1093/biomet/asm021

[35] SMOLA, A. J. and SCHÖLKOPF, B. (2004). A tutorial on support vector regression. *Stat. Comput.* **14** 199–222. MR2086398 https://doi.org/10.1023/B:STCO.0000035301.49549.88

[36] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

[37] WANG, H. J. and WANG, L. (2009). Locally weighted censored quantile regression. *J. Amer. Statist. Assoc.* **104** 1117–1128. MR2562007 https://doi.org/10.1198/jasa.2009.tm08230

[38] WANG, Q. and YIN, X. (2008). A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse MAVE. *Comput. Statist. Data Anal.* **52** 4512–4520. MR2432477 https://doi.org/10.1016/j.csda.2008.03.003

[39] XIA, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *Ann. Statist.* **35** 2654–2690. MR2382662 https://doi.org/10.1214/009053607000000352

[40] XIA, Y., TONG, H., LI, W. K. and ZHU, L.-X. (2002). An adaptive estimation of dimension reduction space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 363–410. MR1924297 https://doi.org/10.1111/1467-9868.03411

[41] YIN, X., LI, B. and COOK, R. D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *J. Multivariate Anal.* **99** 1733–1757. MR2444817 https://doi.org/10.1016/j.jmva.2008.01.006

[42] ZHU, L.-P., LI, L., LI, R. and ZHU, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.* **106** 1464–1475. MR2896849 https://doi.org/10.1198/jasa.2011.tm10563

[43] ZHU, L.-X. and FANG, K.-T. (1996). Asymptotics for kernel estimate of sliced inverse regression. *Ann. Statist.* **24** 1053–1068. MR1401836 https://doi.org/10.1214/aos/1032526955