Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

On Partial Sufficient Dimension Reduction With Applications to Partially Linear Multi-Index Models

Zhenghui Feng , Xuerong Meggie Wen , Zhou Yu & Lixing Zhu

To cite this article: Zhenghui Feng , Xuerong Meggie Wen , Zhou Yu & Lixing Zhu (2013) On Partial Sufficient Dimension Reduction With Applications to Partially Linear Multi-Index Models, Journal of the American Statistical Association, 108:501, 237-246, DOI: 10.1080/01621459.2012.746065

To link to this article: https://doi.org/10.1080/01621459.2012.746065

Accepted author version posted online: 20 Nov 2012.
Published online: 15 Mar 2013.

Submit your article to this journal

Article views: 837

View related articles

Citing articles: 12 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=uasa20

On Partial Sufficient Dimension Reduction With Applications to Partially Linear Multi-Index Models

Zhenghui FENG, Xuerong Meggie WEN, Zhou YU, and Lixing ZHU

Partial dimension reduction is a general method to seek informative convex combinations of predictors of primary interest, which includes dimension reduction as its special case when the predictors in the remaining part are constants. In this article, we propose a novel method to conduct partial dimension reduction estimation for predictors of primary interest without assuming that the remaining predictors are categorical. To this end, we first take the dichotomization step such that any existing approach for partial dimension reduction estimation can be employed. Then we take the expectation step to integrate over all the dichotomic predictors to identify the partial central subspace. As an example, we use the partially linear multi-index model to illustrate its applications for semiparametric modeling. Simulations and real data examples are given to illustrate our methodology.

KEY WORDS: Partial central subspace; Partial discretization-expectation estimation; Partially linear model.

1. INTRODUCTION

For a regression problem, partial dimension reduction arises when one considers the informational role of all predictors but limits reduction to a subset of them. These predictors are called the predictors of primary interest, and other predictors are called the predictors of secondary interest. Partial dimension reduction is a very general problem; in a certain sense, sufficient dimension reduction (see, e.g., Li 1991; Cook 1998) may be regarded as its special case when the rest of the predictors outside this subset is a constant (vector). This would be of particular interest in applications in which some predictors play a particular role and must therefore be shielded from the reduction process. An example is an alcoholism study (Pfeiffer and Bura 2008), from the publicly available Third National Health and Nutrition Examination Survey (NHANES III), where the goal was to classify men aged 40 years or older into two groups: heavy drinkers and abstainers, using nine serum biomarkers. Age was also included since it is known to influence both the values of the biomarkers and the drinking pattern. Here the dimension reduction should focus on the set of biomarkers while controlling for the age effect.

Let Y be a univariate random response, $\mathbf{X} = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$ be a vector of continuous random predictors of primary interest, and $\mathbf{W} = (W_1, \ldots, W_q) \in \mathbb{R}^q$ be a vector of predictors of secondary interest. When it is desirable to conduct dimension reduction on $\mathbf{X}$ while incorporating the prior information from $\mathbf{W}$, we should not treat all the components of the predictors $(\mathbf{X}, \mathbf{W})$ indiscriminately, as the usual case in sufficient dimension reduction (e.g., Li 1991; Cook 1998). Chiaromonte, Cook, and Li (2002) introduced the partial central subspace $\mathcal{S}^{(W)}_{Y|\mathbf{X}}$ that is defined as the intersection of all subspaces $\mathcal{S}$ satisfying

$$Y \perp\!\!\!\perp \mathbf{X} \mid (P_{\mathcal{S}}\mathbf{X}, \mathbf{W}), \tag{1.1}$$

where $\perp\!\!\!\perp$ indicates independence and $P_{(\cdot)}$ stands for a projection operator with respect to the standard inner product. And $\dim(\mathcal{S}^{(W)}_{Y|\mathbf{X}}) = d$ is called the structural dimension of the partial central subspace.

Chiaromonte, Cook, and Li (2002) proposed an estimation method for the partial central subspace. Although, their method—along with other existing methods in this field (Wen and Cook 2007)—can only be applied to the cases in which the predictors of secondary interest are categorical, and it is difficult to extend these methods to incorporate the continuous $\mathbf{W}$ scenario. However, this scenario is of particular interest in semiparametric modeling as the approach we develop in this article could be applied to many semiparametric models. Details on this perspective will be provided in later sections.

A related approach is the groupwise dimension reduction (GDR; Li, Li, and Zhu 2010), which could be adopted to handle the partial dimension reduction problem. As was mentioned in Li, Li, and Zhu (2010, sect. 4.2), all the predictors of secondary interest are regarded as an extra group with a $q \times q$ identity matrix as a given projection matrix when the secondary predictors are q-dimensional. However, GDR can only be used to infer about the partial conditional mean subspace (Li, Cook, and Chiaromonte 2003), rather than the partial dimension reduction subspace. Directions along the conditional variance cannot be identified by GDR. Further, even for the inference on the partial central mean subspace, GDR is not an efficient approach. The convergence rate of the GDR estimator is highly related to the bandwidth and the number of all predictors in nonparametric smoothing, the estimation efficiency is therefore greatly

Feng and Wen contributed equally to this work. Zhenghui Feng is Assistant Professor, School of Economics & Wang Yanan Institute for Studies in Economics (WISE), Xiamen University, Fujian Province, China (E-mail: *zhfengwise@gmail.com*). Xuerong Meggie Wen is corresponding author and Associate Professor, Missouri University of Science and Technology, Rolla, MO 65409 (E-mail: *wenx@mst.edu*). Her work was supported by the Missouri Research Board. Zhou Yu is Assistant Professor, School of Finance and Statistics, East China Normal University, Shanghai, China (E-mail: *zyu@stat.ecnu.edu.cn*). His research was supported by a grant from the National Natural Science Foundation of China (no. 11201151). Lixing Zhu is Chair Professor, Department of Mathematics, Hong Kong Baptist University, Hong Kong, China (E-mail: *lzhu@hkbu.edu.hk*). His research was supported by a grant from the Research Grants Council of Hong Kong and a Faculty Research Grant (FRG) from Hong Kong Baptist University. The authors thank three anonymous referees, an associate editor, and the coeditor for their many constructive suggestions that helped us improve both the presentation and the substance of this article.

© 2013 American Statistical Association
Journal of the American Statistical Association
March 2013, Vol. 108, No. 501, Theory and Methods
DOI: 10.1080/01621459.2012.746065

deteriorated. Oversmoothing is needed for better convergence rate, which thus increases the difficulty of bandwidth selection (Stute and Zhu 2005; Zhu 2005). It is clear that when treating all those secondary predictors as a group, the estimation efficiency gets worse particularly when q is large. See theorem 5 of Li, Li, and Zhu (2010) for further details.

In this article, we propose a method to deal with the estimation of partial central subspace with a general $\mathbf{W}$. Our method can identify the partial central subspace while enjoying the root-n convergence rate as long as the corresponding estimation for sufficient dimension reduction has such a rate. The basic idea is to transfer the continuous $\mathbf{W}$ to a set of dichotomized $\mathbf{W}$ in terms of a dichotomization transformation. This transformation plays a critical role enabling us to apply existing approaches which could deal with categorical $\mathbf{W}$ successfully. The dichotomization transformation can also be replaced by a discretization transformation with more than two values. Once an existing approach is used to construct kernel matrix (see Yu, Zhu, and Wen 2012 for details) with respect to any dichotomized $\mathbf{W}$, the average over all the transformed $\mathbf{W}$'s in the set can be employed to define a final estimator of the partial central subspace. We call the method partial discretization-expectation estimation (PDEE). Note that although the spirit in discretization and expectation is similar to DEE (Zhu et al. 2010), neither the motivation nor target of our method is the same as DEE. DEE is developed to make slicing estimation more efficient (Li 1991; Zhu and Ng 1995; Li and Zhu 2007). In contrast, PDEE is to make the estimation of $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$ possible when $\mathbf{W}$ is not categorical. This is the first result with continuous $\mathbf{W}$ in the literature. As the discretization is placed on $\mathbf{W}$, rather than on the response Y, we cannot simply use the slicing estimation in DEE to partial dimension reduction. Therefore, the method in DEE cannot be directly applied to PDEE. Further, we also propose a new approximation algorithm to implement the expectation step when $\mathbf{W}$ is high-dimensional.

As an important application, PDEE can be applied to some well-known semiparametric models where the classical methods have difficulties to handle. An example is the well-known partially linear single-index (PLSI) or multi-index model (Carroll et al. 1997; Wang et al. 2010):

$$Y = \boldsymbol{\theta}^T\mathbf{W} + g(\boldsymbol{\beta}^T\mathbf{X}) + \epsilon, \tag{1.2}$$

where $\mathbf{W} \in \mathbb{R}^q$, $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ (a $p \times d$ matrix), $g(.)$ is an unknown link function for the single index (when $d = 1$) or multiple indices (when $d > 1$), and ϵ is the error term with $E(\epsilon) = 0$ and $0 < \text{var}(\epsilon) < \infty$. For this model, there are three main approaches in the literature to the best of our knowledge. The first approach is to estimate $\boldsymbol{\theta}$ by using the conditionally centered Y given $\mathbf{X}$ and then to estimate $\boldsymbol{\beta}$. A relevant reference is Härdle, Liang, and Gao (2000). This type of method involves nonparametrically estimating $E(\mathbf{W}|\mathbf{X})$ with high-dimensional predictor $\mathbf{X}$, which suffers from a typical estimation inefficiency. The second approach is to estimate them simultaneously (see, e.g., Carroll et al. 1997). It is not stable in computation as its estimation procedure is complicated (Carroll et al. 1997; Yu and Ruppert 2002). The third method (Wang et al. 2010) is a computationally more efficient procedure than the second one, assuming that $\mathbf{W}$ is of a dimension reduction structure of $\boldsymbol{\beta}_1^T\mathbf{X}$ for another projection $\boldsymbol{\beta}_1$. This method can only be applied to the single-index model where $d = 1$ with $\mathbf{W}$ being limited to be a function of $\boldsymbol{\beta}_1^T\mathbf{X}$. Thus, it cannot handle the general $\mathbf{W}$ herein. Xia and Härdle (2006) also used a dimension reduction approach to simultaneously estimate both $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. Their method, however, also involves high-dimensional nonparametric smoothing.

In contrast, our method can be efficient without using the classical nonparametric estimation and without assuming the special structure of $\mathbf{W}$. Once we obtain an estimator of $\boldsymbol{\beta}$ via partial dimension reduction method, we can reduce the high-dimensional $\mathbf{X}$ to a low-dimensional $\boldsymbol{\beta}^T\mathbf{X}$. The least squares method can then be used to estimate $\boldsymbol{\theta}$ in terms of centering Y conditionally on $\boldsymbol{\beta}^T\mathbf{X}$. The approach is different from all existing methods in the literature. Our method is, by nature, one of dimension reduction and an asymptotically normal estimator of $\boldsymbol{\beta}$ can be obtained without an iteration algorithm. More details are given in Section 4. Also, our approach can be applied to multi-index models with $d > 1$. The above discussion is also applicable to a more general model investigated by Li, Zhu, and Zhu (2011), in which the secondary predictor set contains two sets of variables. More discussion is provided in Section 6.

The rest of this article is organized as follows. In Section 2, we present our new estimation method, which we call PDEE, of $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$ with continuous $\mathbf{W}$ and its related asymptotic properties. An approximation algorithm is suggested in Section 2 as well. A modified Bayesian information criterion (BIC)-type criterion will be adopted to estimate the dimension of $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$ in Section 3. Section 4 is dedicated to the inferences of the partially linear multi-index models with the aid of the partial sufficient dimension reduction. We illustrate the performance of our methods via simulation studies in Section 5. Real data analyses will also be discussed. Some further discussion on future research directions are given in Section 6. For the ease of exposition, we defer all proofs to the Appendix.

2. PARTIAL DISCRETIZATION-EXPECTATION ESTIMATION

2.1 Theoretical Development

Let $(\mathbf{X}_w, Y_w)$ denote a generic pair distributed like $(\mathbf{X}, Y)|(\mathbf{W} = w)$ and $\mathcal{S}_{Y_w|\mathbf{X}_w}$ be the central subspace in subpopulation $\mathbf{W} = w$. When $\mathbf{W}$ is discrete and takes value at $\{1, 2, \ldots, K\}$, the following equation connects the marginal central subspace $\mathcal{S}_{Y_w|\mathbf{X}_w}$ with the partial central subspace $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$:

$$\mathcal{S}_{Y|\mathbf{X}}^{(W)} = \sum_1^K \mathcal{S}_{Y_w|\mathbf{X}_w}. \tag{2.1}$$

Chiaromonte, Cook, and Li (2002) and Wen and Cook (2007) proposed estimation methods for the partial central subspace based on (2.1).

Now we introduce a PDEE procedure for continuous $\mathbf{W}$. At the first step, we discretize the continuous $\mathbf{W} = (W_1, \ldots, W_q)^T$ into a set of binary variables. To be precise, for each $\mathbf{t} = (t_1, \ldots, t_q)^T$, we define the new $\mathbf{W}(\mathbf{t}) = (I_{\{W_1\le t_1\}}, \ldots, I_{\{W_q\le t_q\}})^T$, where the indicator function $I_{\{W_i\le t_i\}}$ takes the value 1 if $W_i \le t_i$, and 0 otherwise, for $i = 1, \ldots, q$. In doing so, the multidimension $\mathbf{W}$ is divided into at most 2^q hypercubes because every response coordinate in $\mathbf{W}(\mathbf{t})$ is

binary. This procedure is easy to implement, and we can also use any general discretization procedure. Let $\mathcal{S}_{Y|\mathbf{X}}^{(W(t))}$ be the partial central subspace of $\mathbf{Y}|(\mathbf{X}, \mathbf{W}(\mathbf{t}))$, and $M(\mathbf{t})$ be a $p \times p$ positive semidefinite matrix such that $\text{Span}\{M(\mathbf{t})\} = \mathcal{S}_{Y|\mathbf{X}}^{(W(t))}$. We have the following results.

Proposition 1. $\mathcal{S}_{Y|\mathbf{X}}^{(W(t))} \subseteq \mathcal{S}_{Y|\mathbf{X}}^{(W)}$ for any $\mathbf{t} \in \mathbb{R}^q$ and $\bigcup_{\mathbf{t}} \mathcal{S}_{Y|\mathbf{X}}^{(W(t))} = \mathcal{S}_{Y|\mathbf{X}}^{(W)}$.

This motivates us to consider a direct product of the subspaces $\mathcal{S}_{Y|\mathbf{X}}^{(W(t))}$ via a sum of the kernel matrices for all $\mathcal{S}_{Y|\mathbf{X}}^{(W(t))}$ so that we can preserve the integrity of $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$. Specifically, we need to sum up the column spaces of $M(\mathbf{t})$ over all possible values of $\mathbf{t}$. Because $M(\mathbf{t})$ is assumed to be positive semidefinite, it suffices to take the expectation over a random vector $\mathbf{T}$ with support $\mathbb{R}_{\mathbf{T}}^q$ to obtain the target matrix $M = \text{E}\{M(\mathbf{T})\}$, where $\mathbb{R}_{\mathbf{T}}^q$ contains all points in the support of $\mathbf{W}$ ($\mathbb{R}_{\mathbf{W}}^q$). One easy way is to take $\mathbf{T}$ as an independent copy of $\mathbf{W}$. Theorem 1 shows that the above procedure can span $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$ in terms of the matrix M.

Theorem 1. If the support of $\mathbf{W}$ is a subset of the support of $\mathbf{T}$ and $\text{Span}\{M(\mathbf{t})\} = \mathcal{S}_{Y|\mathbf{X}}^{(W(t))}$ for any given $\mathbf{t}$, then $\text{Span}\{M\} = \mathcal{S}_{Y|\mathbf{X}}^{(W)}$, where $M = \text{E}\{M(\mathbf{T})\}$.

In general, we can estimate $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$ using the above two-step procedure by estimating $M = \text{E}\{M(\mathbf{T})\}$. Let $\mathbf{t}_1, \ldots, \mathbf{t}_{l_n}$ be l_n independent copies of $\mathbf{T}$, then

$$\lim_{l_n \to \infty} \frac{1}{l_n} \sum_{i=1}^{l_n} M(\mathbf{t}_i) = \text{E}\{M(\mathbf{T})\}.$$

For any fixed $\mathbf{t}_i \in \mathbb{R}_{\mathbf{T}}^q$, we can obtain $M_n(\mathbf{t}_i)$, a $\sqrt{n}$ consistent estimator of $M(\mathbf{t}_i)$, via available partial dimension reduction methods such as partial sliced inverse regression estimation (partial SIR; Chiaromonte, Cook, and Li 2002), partial sliced average variance estimation (partial SAVE; Shao, Cook, and Weisberg 2009), or partial directional regression (partial DR; Li and Wang 2007).

Let $M_{l_n,n} = \frac{1}{l_n} \sum_{i=1}^{l_n} M_n(\mathbf{t}_i)$, assuming the following conditions:

(a) $M_n(\mathbf{t}) = M(\mathbf{t}) + \text{E}_n\{\phi(\mathbf{X}, Y, \mathbf{W}, \mathbf{t})\} + R_n(\mathbf{t})$, where E_n denotes sample averages, $\text{E}\{\phi(\mathbf{X}, Y, \mathbf{W}, \mathbf{t})\} = 0$ and $\phi(\mathbf{X}, Y, \mathbf{W}, \mathbf{t})$ has a finite second-order moment.

(b) $\sup_{\mathbf{t} \in \mathbb{R}_{\mathbf{T}}^q} \|R_n(\mathbf{t})\|_F = o_p(n^{-\frac{1}{2}})$, where $\|.\|_F$ denotes the Frobenius norm of a matrix.

The following proposition suggests that under some regularity conditions, it suffices to take $O(n)$ random sample points in $\mathbb{R}_{\mathbf{T}}^q$ to obtain a $\sqrt{n}$ consistent estimator of M. Hence we only need to estimate $O(n)$ partial central subspaces $\mathcal{S}_{Y|\mathbf{X}}^{(W(t))}$.

Proposition 2. Assuming conditions (a) and (b), and also assuming that the entries of $M_n(\mathbf{t})$ have finite second moments, for each $\mathbf{t} \in \mathbb{R}_{\mathbf{T}}^q$, we have that if $l_n = O(n)$,

$$M_{l_n,n} = M + O_p\big(n^{-\frac{1}{2}}\big).$$

We now investigate the asymptotic properties of $M_{l_n,n}$. Let $l_n = n$ and $\mathbf{t}_i = \mathbf{W}_i$, which is a natural choice for ease of practical implementation, for $i = 1, \ldots, n$.

Theorem 2. Let $\widetilde{\mathbf{W}}$ be an independent copy of $\mathbf{W}$. Assume that all conditions in Proposition 2 hold, and $\text{E}\{M^2(\mathbf{T})\} < \infty$ componentwise. Then,

$$\sqrt{n}(\text{vec}(M_{n,n}) - \text{vec}(M)) \xrightarrow{D} \text{Normal}(\mathbf{0}, \text{var}\{\text{vec}(\mathbf{C})\}),$$

where vec(.) is the operator stacking the columns of a matrix to vectorize it and $\mathbf{C} = M(\widetilde{\mathbf{W}}) - \text{E}(M(\widetilde{\mathbf{W}})) + \text{E}\{\phi(\mathbf{X}, Y, \mathbf{W}, \widetilde{\mathbf{W}})|(\mathbf{X}, Y, \mathbf{W})\} + \text{E}\{\phi(\mathbf{X}, Y, \mathbf{W}, \widetilde{\mathbf{W}})|\widetilde{\mathbf{W}}\}$.

As Zhu and Ng (1995) and Zhu and Fang (1996) showed, under certain regularity conditions, the above root-n consistency leads to the root-n consistency of the eigenvectors of $M_{n,n}$. A subset of those eigenvectors can be used to estimate the base vectors of $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$.

2.2 A Discussion on Implementation

We now discuss the implementation of the estimation procedure. From the above theorem, we can see that the law of large numbers ensures that $M_n = \frac{1}{n} \sum_{i=1}^n M_n(\mathbf{W}_i) \to M$ as $n \to \infty$. Thus, theoretically, when we use n points $\mathbf{W}_i$'s, the estimator is consistent. More generally, at the sample level, we may only need to choose l_n points $\mathbf{t}_i$'s of which l_n is of the order $O(n)$ to construct an estimate of M. However, when q is large, for many $\mathbf{W}_i$, the set of contaminated points $\{(\mathbf{X}_i, Y_i)\}$ associated with the super-cube $\{\mathbf{W}_j : I(\mathbf{W}_j \le \mathbf{W}_i)\}$ are very few and then the corresponding partial central subspace $\text{Span}\{M(\mathbf{W}_i)\}$ associated with $M_n(\mathbf{W}_i)$ cannot be estimated well. Hence, $\frac{1}{n} \sum_{i=1}^n M_n(\mathbf{W}_i)$ cannot provide a good estimator of the partial central subspace $\text{Span}\{M\}$. Another immediate way is to use all grid points in the set $\mathbf{A} = \{\mathbf{t}_{i_1,\ldots,i_q} = (W_{1i_1}, \ldots, W_{qi_q})^T : 1 \le i_1, \ldots, i_q \le n\}$ to exhaustively compute the corresponding $M_n(\mathbf{t}_{i_1,\ldots,i_q})$, where $\mathbf{W}_i = (W_{1i}, \ldots, W_{qi})^T$. Then we can have an estimator of M

$$\frac{1}{(n-1)^q} \sum_{1 \le i_1, \ldots, i_q \le n} M_n(\mathbf{t}_{i_1,\ldots,i_q}) := M_n.$$

However, the above strategy is clearly impractical when q is large as it needs to compute n^q matrices in total and thus the computational burden is very heavy. More importantly, such an exhaustive average may not provide a good estimator or even deteriorate the estimation accuracy. This is because, as commented above, there will be many $M_n(\mathbf{t}_{i_1,\ldots,i_q})$ based only on a few points and thus do not estimate the corresponding partial central subspaces efficiently. Actually, in a small-scale simulation, we did observe this phenomenon. Furthermore, the computational burden makes the algorithm infeasible and the resulting estimator does not perform well. As such, we do not use this algorithm. In the simulation section, we adopt the following approximation algorithm. Let $\mathbf{W}_{ik}^{\infty}$ be the column vector of which only the kth component is the same as that of $\mathbf{W}_i$ and the other components are the maximum values of the corresponding components of all $\mathbf{W}_i$'s. We then use

$$\frac{1}{qn} \sum_{k=1}^{q} \sum_{i=1}^{n} M_n\left(\mathbf{W}_{ik}^{\infty}\right) := \tilde{M}_n \tag{2.2}$$

as an estimator. In effect, this is an estimator of

$$\tilde{M} = \int M(\mathbf{T}) dF_1(t_1) \ldots dF_q(t_q), \tag{2.3}$$

where $F_k(\cdot)$'s are the marginal distribution of t_k. $\tilde{M}$ may not be equal to $M = \int M(\mathbf{T}) dF(\mathbf{T})$. Thus, the partial central subspace

that is based on $\tilde{M}$ might not be equal to that of M. Proposition 1 shows that the space identified by $\tilde{M}$ is contained in Span(M). In theory, it is hard to know in which cases the space of $\tilde{M}$ is identical to that of M. However, from the simulation results reported in Section 5, we can see that this approximation algorithm never fails to identify Span(M). Our experiences also show that it always yields satisfactory performances. Therefore, we leave the theoretical development to further studies.

3. DIMENSION DETERMINATION OF THE PARTIAL CENTRAL SUBSPACE

There are several approaches for determining the structural dimension d. Sequential test method and weighted sequential test method (Li 1991; Bura and Cook 2001) are frequently used. However, they are not consistent and generally require a relatively large sample size for good performances. Zhu, Miao, and Peng (2006) first proposed the BIC-type criterion to obtain consistent estimation of the structural dimension. Here, we use a modified BIC-type criterion:

$$\hat{d} = \arg\max_{k\in\{1,2,\ldots,p\}} \left(\frac{n\sum_{m=1}^{k}(\log(\hat{\lambda}_m+1)-\hat{\lambda}_m)}{2\sum_{m=1}^{p}(\log(\hat{\lambda}_m+1)-\hat{\lambda}_m)} - 2C_n \times \frac{k(k+1)/2}{p} \right), \tag{3.1}$$

where $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$ denote the eigenvalues of the matrix of $M_{n,n}$; C_n is a penalty constant; and $k(k+1)/2$ equals to the number of free parameters. The following theorem provides the consistency of $\hat{d}$.

Theorem 3. Assuming that $\frac{C_n}{n} \to 0$ and $C_n \to \infty$ as $n \to \infty$, also assuming the conditions of Theorem 2, the estimated structural dimension $\hat{d}$ obtained via (3.1) converges to the true structural dimension d with probability tending to one.

The proof is similar to that of theorem 4 of Zhu et al. (2010) and is omitted. In Zhu et al. (2010), $C_n = \sqrt{n}$ was recommended. In the original BIC proposed by Zhu, Miao, and Peng (2006), some values with leading order $C_n = n^{1/3}$ were considered. In our simulations, we also tried similar values and found that $C_n = n^{1/3}$ was a good choice, but when p is large, BIC tended to overestimate the structural dimension. In contrast, when C_n contains a factor of p, BIC works better. Thus, we recommend a value of $C_n = n^{1/3}p^{2/3}$.

4. PARTIALLY LINEAR MULTI-INDEX MODEL

For model (1.2), we recommend a new estimation approach in which we first estimate $\boldsymbol{\beta}$ to reduce the dimension of $\mathbf{X}$ without dealing with the unknown link function $g(\cdot)$. Its estimator is of the asymptotic normality in terms of Theorem 2. Specifically, our estimation method relies on the following equation:

$$\text{Span}\{\boldsymbol{\beta}\} = \mathcal{S}^{(W)}_{Y|\mathbf{X}}. \tag{4.1}$$

The estimation procedure is as follows:

Step 1. Use partial dimension reduction to construct an estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$.

Step 2. Center Y as $Y - \hat{E}(Y|\hat{\boldsymbol{\beta}}^T\mathbf{X})$ where $\hat{E}$ stands for a nonparametric estimator of $E(Y|\hat{\boldsymbol{\beta}}^T\mathbf{X})$. For example, we can use kernel estimation procedure to produce an estimator of $E(Y|\hat{\boldsymbol{\beta}}^T\mathbf{X})$.

Step 3. Define a least squares estimator $\hat{\boldsymbol{\theta}}$ by $Y - \hat{E}(Y|\hat{\boldsymbol{\beta}}^T\mathbf{X})$ versus $\mathbf{W} - \hat{E}(\mathbf{W}|\hat{\boldsymbol{\beta}}^T\mathbf{X})$.

Following the arguments parallel to that in theorem 1 of Wang et al. (2010), the asymptotic normality of $\hat{\boldsymbol{\theta}}$ can be achieved under some regularity conditions. For this, we can see that

$$\hat{\boldsymbol{\theta}} = (\widehat{\text{cov}}(\mathbf{W} - \hat{E}(\mathbf{W}|\hat{\boldsymbol{\beta}}^T\mathbf{X})))^{-1}\hat{E}((\mathbf{W} - \hat{E}(\mathbf{W}|\hat{\boldsymbol{\beta}}^T\mathbf{X})) \times (Y - \hat{E}(Y|\hat{\boldsymbol{\beta}}^T\mathbf{X}))), \tag{4.2}$$

where $\widehat{\text{cov}}(\mathbf{W} - \hat{E}(\mathbf{W}|\hat{\boldsymbol{\beta}}^T\mathbf{X}))$ is the sample version of $\text{cov}(\mathbf{W} - E(\mathbf{W}|\boldsymbol{\beta}^T\mathbf{X}))$ and $\hat{E}((\mathbf{W} - \hat{E}(\mathbf{W}|\hat{\boldsymbol{\beta}}^T\mathbf{X}))(Y - \hat{E}(Y|\hat{\boldsymbol{\beta}}^T\mathbf{X})))$ is the sample version of $E((\mathbf{W} - E(\mathbf{W}|\boldsymbol{\beta}^T\mathbf{X}))(Y - E(Y|\boldsymbol{\beta}^T\mathbf{X})))$. Under certain regularity conditions (see, e.g., Wang et al. 2010), $\widehat{\text{cov}}(\mathbf{W} - \hat{E}(\mathbf{W}|\hat{\boldsymbol{\beta}}^T\mathbf{X}))$ converges in probability to $\text{cov}(\mathbf{W} - E(\mathbf{W}|\boldsymbol{\beta}^T\mathbf{X}))$ and $\hat{E}((\mathbf{W} - \hat{E}(\mathbf{W}|\hat{\boldsymbol{\beta}}^T\mathbf{X}))(Y - \hat{E}(Y|\hat{\boldsymbol{\beta}}^T\mathbf{X})))$ admits an asymptotic linear presentation as $\hat{E}((\mathbf{W} - E(\mathbf{W}|\boldsymbol{\beta}^T\mathbf{X}))(Y - E(Y|\boldsymbol{\beta}^T\mathbf{X}))) + o_p(1/\sqrt{n})$, which is a sum of independent identically distributed variables plus a negligible remainder. This leads to the asymptotic normality by the central limit theorem. We will not present the details of conditions and proof while only present a general result parallel to that in theorem 1 of Wang et al. (2010), although our model is more general than theirs without imposing a special structure on the predictors related to the parameter $\boldsymbol{\theta}$.

Proposition 3. Assume that an estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is $\sqrt{n}$-consistent. Under the regularity conditions specified in Wang et al. (2010), $\hat{\boldsymbol{\theta}}$ is also asymptotically normal with mean zero and variance matrix $\Sigma = (\text{cov}(\mathbf{W} - E(\mathbf{W}|\boldsymbol{\beta}^T\mathbf{X})))^{-1}\text{cov}((\mathbf{W} - E(\mathbf{W}|\boldsymbol{\beta}^T\mathbf{X}))(Y - E(Y|\boldsymbol{\beta}^T\mathbf{X})))(\text{cov}(\mathbf{W} - E(\mathbf{W}|\boldsymbol{\beta}^T\mathbf{X})))^{-1}$ provided that it is a positive definite matrix.

Another issue is about estimation efficiency for $\boldsymbol{\beta}$. Since there are no results available in the literature for multiple indices, we limit the comparison of our method with existing ones to single-index model ($d = 1$). Comparing the limiting variance matrix with those in the literature (see, e.g., Carroll et al. 1997; Wang et al. 2010; its correction in Li, Zhu, and Zhu 2011), our estimator $\hat{\boldsymbol{\beta}}$ may not be as efficient as existing ones which are asymptotically efficient in a semiparametric sense. However, this can be easily fixed because the following algorithm with one more iteration can be applied to achieve the asymptotic efficiency. Regard $\hat{\boldsymbol{\beta}}$ as the initial estimator of $\boldsymbol{\beta}$ to obtain $\hat{\boldsymbol{\theta}}$. We then use $\hat{Y}_j = Y_j - \hat{\boldsymbol{\theta}}^T\mathbf{W}_j$ and $\mathbf{X}_j$ to update the estimator of the index $\boldsymbol{\beta}$. This is because we actually rewrite the model as $Y - \boldsymbol{\theta}^T\mathbf{W} = g(\boldsymbol{\beta}^T\mathbf{X}) + \epsilon$ and regard the model as the single-index model. The techniques in Wang et al. (2010) or Cui, Härdle, and Zhu (2011) may be useful for proving the asymptotic efficiency. Research along this line is ongoing. On the other hand, our method also has its limitation on handling very high-dimensional $\mathbf{W}$ because the PDEE algorithm with the average over all $\mathbf{W}$ neglects the special linear structure about $\mathbf{W}$. Thus, it deserves a further investigation on more efficient algorithms.

5. NUMERICAL STUDIES

In this section, we first conduct extensive simulation studies to evaluate the performance of the three partial discretization-expectation estimators: PDEE-SIR, PDEE-SAVE, and PDEE-DR. To assess the accuracy of our proposed method, we use the squared trace correlation coefficient (Li and Dong 2009). For a pair of generic random vectors U and V, the squared trace correlation coefficient is defined as $r^2 = \mathrm{tr}(A)/\dim(A)$, where $A = \Sigma_V^{-1/2}\Sigma_{VU}\Sigma_U^{-1}\Sigma_{UV}\Sigma_V^{-1/2}$, Σ_V, Σ_U are the variance matrices of U and V, respectively, and Σ_{VU} is the covariance matrix between U and V. For a sample estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, we can then compute a sample estimate of $r^2(U, V)$ with $U = \boldsymbol{\beta}^T X$ and $V = \hat{\boldsymbol{\beta}}^T \mathbf{X}$. A squared trace correlation coefficient closer to unity indicates higher estimation efficiency. See Li and Dong (2009) and references therein for further details. We then apply PDEE-based dimension reduction methods to analyze the NHAMES III data (Pfeiffer and Bura 2008) and the Boston Housing data (Wang et al. 2010).

5.1 Simulation Studies

In this section, we use simulations to evaluate the performance of the three partial discretization-expectation estimators: PDEE-SIR, PDEE-SAVE, and PDEE-DR. Our comparison is twofold: we compare the performances among the three partial discretization-expectation estimators themselves and compare them with other well-developed methods for PLSI model.

5.1.1 Study I: Comparisons Among the Three PDEEs. Consider the following six models:

$$
\begin{aligned}
&\text{(I) } Y = (5 + \mathbf{X}_1 + \mathbf{X}_2 - \mathbf{X}_3 - \mathbf{X}_4 + W + 0.5\epsilon)^2,\\
&\text{(II) } Y = 3W\mathbf{X}_1/(0.5 + (1.5 + \mathbf{X}_2)^2) + 0.2\epsilon,\\
&\text{(III) } Y = \theta^T\mathbf{W} + (\mathbf{X}_1 + \mathbf{X}_2)^4 + 0.2\epsilon,\\
&\text{(IV) } Y = 0.8W_1 - 0.3W_2 + \exp\left(\boldsymbol{\beta}_1^T\mathbf{X}\right)\mathrm{sgn}\left(\boldsymbol{\beta}_2^T\mathbf{X}\right) + 0.2\epsilon,\\
&\text{(V) } Y = 0.3W_1 + 3\sin\left(\boldsymbol{\beta}_1^T\mathbf{X}/4\right) + 0.5(1+W_2)\left(\boldsymbol{\beta}_2^T\mathbf{X}\right)^2 + 0.2\epsilon,\\
&\text{(VI) } Y = 0.3W_1 + 3\sin\left(1 + \left(W_2 + \boldsymbol{\beta}_2^T\mathbf{X}\right)/4\right)\\
&\qquad + 0.4\left(2 + \boldsymbol{\beta}_1^T\mathbf{X} + 0.5W_3\right)^2\\
&\qquad + 0.1(W_4 + \cdots + W_q) + 0.5\epsilon.
\end{aligned}
$$

For the aforementioned six models, the error term ϵ is standard normal $N(0, 1)$ and is independent of $\mathbf{X}$ and $\mathbf{W}$. In Models I–II, $\mathbf{X}$ and W are generated independently from $N(0, I_p)$ and $N(0, 1)$. Model II has two directions with $\beta_1 = (1, 0, \ldots, 0)$ and $\beta_2 = (0, 1, 0, \ldots, 0)$. Different versions of Models I–II had been considered by Yin (2005) with discrete W. In Model III, $\mathbf{X}$ follows $N(0, I_p)$, $\theta = (0.8, -0.6, 0.5)^T$, and $\mathbf{W} = (W_1, W_2, W_3)^T$ is a three-dimensional random vector with W_i being independently generated from $U[0, 1]$. For Models IV, V, and VI, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are p-dimensional vectors with their first six components being $(1, 1, 1, 0, 0, 0)$ and $(0, 0, 0, 0, -1, 1)$ and other elements being 0 if $p > 6$. In Model IV, $\mathbf{V} = (\mathbf{W}, \mathbf{X})$ follows a multivariate normal distribution with mean 0, and the correlation between $\mathbf{V}_i$ and $\mathbf{V}_j$ is $0.5^{|i-j|}$, $1 \le i, j \le p + 2$. In Model V, $\mathbf{X}$ is generated independently from $N(0, I_p)$, while $\mathbf{W}$ is generated independently from $N(0, I_2)$. In Model VI, $q = 15$, $\mathbf{X} \sim N(0, I_p)$, and $\mathbf{W} \sim N(0, I_q)$.

Models I and III are of one-dimensional structure, while Models II, IV, V, and VI are two-dimensional structure. Model III is standard PLSI models. Models IV–V are different variations of partially linear multi-index models with relatively complicated structures. Model VI is a partially linear multi-index model with high-dimensional $\mathbf{W}$.

We compare the performances among the three PDEE-based sufficient dimension reduction methods with different choices of n and p. The number of slices is taken as 5. Table 1 gives the median values and the interquartile ranges (IQRs) of the estimated squared trace correlation coefficients across 200

Table 1. Medians (IQR) of $\hat{r}^2$ for Models I–VI

		PDEE-SIR		PDEE-SAVE		PDEE-DR	
Model	n	$p = 6$	$p = 12$	$p = 6$	$p = 12$	$p = 6$	$p = 12$
	100	0.9759(0.0228)	0.9423(0.0288)	0.1291(0.6197)	0.0077(0.0207)	0.9731(0.0270)	0.9354(0.0416)
I	200	0.9870(0.0109)	0.9733(0.0179)	0.6461(0.8269)	0.0069(0.0211)	0.9842(0.0125)	0.9698(0.0189)
	400	0.9935(0.0059)	0.9844(0.0103)	0.9478(0.0786)	0.0205(0.0987)	0.9926(0.0062)	0.9836(0.0101)
	100	0.9193(0.0714)	0.8265(0.0758)	0.4934(0.1384)	0.0527(0.0782)	0.8789(0.0943)	0.7789(0.1205)
II	200	0.9590(0.0341)	0.9309(0.0505)	0.5213(0.1675)	0.1633(0.3049)	0.9410(0.0463)	0.8732(0.0670)
	400	0.9771(0.0181)	0.9455(0.0270)	0.6303(0.2872)	0.4312(0.1212)	0.9719(0.0242)	0.9274(0.0367)
	100	0.0773(0.2328)	0.0366(0.1327)	0.8863(0.1340)	0.7826(0.1743)	0.8864(0.0905)	0.7886(0.1717)
III	200	0.0914(0.2499)	0.0321(0.0964)	0.9327(0.0704)	0.8621(0.1025)	0.9404(0.0663)	0.8838(0.0654)
	400	0.0680(0.3119)	0.0377(0.1077)	0.9594(0.0443)	0.9169(0.0486)	0.9691(0.0292)	0.9388(0.0386)
	100	0.9306(0.0565)	0.8392(0.0814)	0.4982(0.2261)	0.0507(0.0672)	0.9095(0.0664)	0.7891(0.0985)
IV	200	0.9652(0.0294)	0.9061(0.0461)	0.6469(0.3655)	0.1312(0.3039)	0.9550(0.0434)	0.8805(0.0691)
	400	0.9812(0.0154)	0.9489(0.0241)	0.9208(0.1375)	0.4467(0.0901)	0.9768(0.0220)	0.9322(0.0304)
	100	0.5375(0.1131)	0.4737(0.0598)	0.5327(0.1612)	0.4082(0.0993)	0.8551(0.1498)	0.6816(0.2616)
V	200	0.5474(0.1396)	0.4943(0.0724)	0.5352(0.1703)	0.4496(0.0837)	0.9367(0.0490)	0.8191(0.1142)
	400	0.5523(0.1573)	0.5029(0.0474)	0.5529(0.1574)	0.4815(0.0564)	0.9706(0.0207)	0.9160(0.0453)
	100	0.7564(0.2157)	0.6060(0.1546)	0.5059(0.1108)	0.0829(0.1111)	0.7059(0.2609)	0.5352(0.1246)
VI	200	0.8629(0.1486)	0.7164(0.1593)	0.5126(0.1048)	0.1621(0.2709)	0.8149(0.2319)	0.6286(0.2008)
	400	0.9254(0.0737)	0.8339(0.1263)	0.5346(0.1126)	0.4744(0.0733)	0.8993(0.1102)	0.7639(0.1686)

simulated samples. From Table 1, we can see that PDEE-DR is the most robust and accurate method among the three PDEE-based methods across all six models. When $n = 400$ and $p = 6$, all the median values of $\hat{r}^2$'s from PDEE-DR are above 0.96 for Models I–V and 0.90 for Model VI, respectively. This agrees with the results from the classical dimension reduction methods. Li and Wang (2007) argued that when its conditions are satisfied, DR is the most accurate method among the family of all sufficient dimension reduction methods that are based on the first two inverse moments, including SIR and SAVE. We also observe that the performances of all the three methods improve reasonably with increasing sample sizes, except PDEE-SIR for Model III with $n = 400$, although the gains in estimation accuracy are not substantial in some cases. Also, the performances of PDEE-SIR and PDEE-DR are pretty robust as p increases, while there are some substantial differences for the performances of PDEE-SAVE with the increase of p.

PDEE-SAVE fails for Models I–IV and is outperformed by PDEE-DR for most models, except for Model III where the exact symmetric structures are designed for the best performance of PDEE-SAVE. Even for those models, PDEE-SAVE does not show an advantage over PDEE-DR. In general, we do not recommend the use of PDEE-SAVE for partial sufficient dimension reduction. This is because, as Li and Zhu (2007) pointed out, SAVE is very sensitive to the choice of the number of slices. Under a complicated structure with either a categorical or continuous W, PDEE-SAVE is generally inferior to PDEE-DR.

Because of the symmetry of the mean functions in Model III, we expect PDEE-SIR to fail, while both PDEE-SAVE and PDEE-DR perform reasonably well with three-dimensional $\mathbf{W}$. Models IV, V, and VI are of complex structure with multidimensional $\mathbf{W}$ and strong nonlinear trend, particularly for Model VI, PDEE-DR still performs reasonably well. In general, we recommend PDEE-DR and PDEE-SIR for partial sufficient dimension reduction. Compared with PDEE-SIR, PDEE-DR requires an extra constant variance condition, thus we might need to use PDEE-SIR as a complementary method to PDEE-DR.

5.1.2 Study II: BIC in Estimating the Structural Dimension. Table 2 reports the percentages of correctly identifying structural dimension for Models II–IV using the modified BIC-type criterion proposed in Section 3. The medians of the estimated structural dimensions over 200 replications are reported in Table 2 as well. We can see that BIC correctly selects the structural dimension most of the time if we choose a suitable partial discretization-expectation estimator adapting to the model. On the other hand, from the median values of the estimated structural dimensions, we also find that, when BIC cannot correctly select the structural dimension, it tends to overselect the dimensions rather than to underselect them. Thus, we may not lose some useful combinations of the predictors even when we used an estimator such as PDEE-SIR for Model III (the median value is $\hat{d} = 2$) that cannot correctly determine the dimension ($d = 1$) with large probability. Further, we observed that in some cases, when p is larger, the proportions of correctly estimating the dimension are also higher. Also this phenomenon depends on the models, estimation methods, and sample sizes. We have not had a clear explanation and thus leave it to further investigation.

Table 2. Proportions that BIC correctly estimates the structural dimension

		PDEE-SIR		PDEE-SAVE		PDEE-DR	
Model	n	$p = 6$	$p = 12$	$p = 6$	$p = 12$	$p = 6$	$p = 12$
II	100	0.65 (2)	0.925 (2)	0.695 (2)	0.615 (2)	0.71 (2)	0.975 (2)
	200	0.925 (2)	0.985 (2)	0.935 (2)	0.685 (2)	1 (2)	0.995 (2)
	400	0.995 (2)	0.995 (2)	0.625 (2)	0.06 (3)	1 (2)	0.8 (2)
III	100	0.3 (2)	0.055 (2)	0.96 (1)	0.59 (1)	0.87 (1)	0.04 (2)
	200	0.045 (2)	0.005 (2)	0.99 (1)	0.725 (1)	0.48 (2)	0 (2)
	400	0.01 (2)	0 (3)	1 (1)	0.875 (1)	0.595 (1)	0 (3)
IV	100	0.91 (2)	0.995 (2)	0.71 (2)	0.675 (2)	0.915 (2)	0.995 (2)
	200	1 (2)	1 (2)	0.765 (2)	0.73 (2)	1 (2)	0.98 (2)
	400	1 (2)	1 (2)	0.49 (3)	0.03 (3)	1 (2)	0.605 (2)

5.1.3 Study III: Comparison With PLSI. We now compare our proposed estimators with the "PLSI" proposed by Xia and Härdle (2006) for the following PLSI models with homoscedastic and heteroscedastic error:

$$\text{(VII)}\ Y = \theta W + 3\sin\left(\boldsymbol{\beta}_1^T\mathbf{X}/4\right) + 0.2\left[1 + \left(\boldsymbol{\beta}_2^T\mathbf{X}\right)^3\right]\epsilon, \quad (5.1)$$

$$\text{(VIII)}\ Y = \theta W + 3\sin\left(\boldsymbol{\beta}_1^T\mathbf{X}/4\right) + 0.2\epsilon. \quad (5.2)$$

Here $\mathbf{X}$, W, and ϵ are independent and follow $N(0, I_p)$, $N(0, 1)$, and $N(0, 1)$, respectively; $p = 6$; $\beta_1 = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}, 0, 0, 0)^T$; $\beta_2 = (0, 0, 0, 0, 1/\sqrt{10}, 3/\sqrt{10})^T$; and $\theta = 0.3$. We adopt these two models because PLSI is designed to fit homoscedastic models such as (VIII), though it is still applicable to heteroscedastic models such as (VII). Also note that for Model VII, the PLSI method can only identify the direction $\boldsymbol{\beta}_1$ in the mean function. Thus, in the simulation study for this model, we will only compare between the performances of PLSI for estimating θ and $\boldsymbol{\beta}_1$ and the performances of PDEE for identifying θ and the central subspace that contains $\boldsymbol{\beta}_1$. The sample size is $n = 200$. In addition to the median and IQR of the squared trance correlation coefficient r^2, we also report in Table 3 the median and IQR (in parentheses) of the estimators of θ, and these of the angles ($\angle$, in radians) between $\boldsymbol{\beta}_1$ and its estimators. The average CPU times consumed are also reported.

We can see that from Table 3, for estimating θ in both of the models, PDEE-SIR and PDEE-DR have similar performances, the biases are slightly larger, and IQR slightly smaller as compared with PLSI. Thus, we may consider that they perform comparably. For estimating $\boldsymbol{\beta}_1$ in Model VII, both PDEE-SIR and PDEE-DR work better than PLSI, as their resulting angles are smaller and their r^2's are larger. PDEE-SAVE performs the worst among all competitors. However, for $\boldsymbol{\beta}_1$ in Model VIII, PLSI is a good choice with smaller angle and larger r^2. These observations indicate that as PLSI is specifically designed for dealing with the estimation for mean function, it has

Table 3. Median (IQR) of the estimated parameters for Models VII and VIII

Method	θ	$\angle$	$\hat{r}^2$	Time (second)
		Model VII		
PDEE-SIR	0.2910(0.0630)	0.1670(0.0683)	0.9724(0.0231)	0.1525
PDEE-SAVE	0.3013(0.0872)	1.2152(0.9086)	0.1213(0.6212)	0.1823
PDEE-DR	0.2924(0.0612)	0.1637(0.0817)	0.9734(0.0271)	0.1755
PLSI	0.2978(0.0639)	0.3080(0.4729)	0.9081(0.3445)	2147.7
		Model VIII		
PDEE-SIR	0.2957(0.0196)	0.1059(0.0531)	0.9888(0.0120)	0.3946
PDEE-SAVE	0.2961(0.0330)	0.4702(0.9960)	0.7949(0.7945)	0.2086
PDEE-DR	0.2943(0.0200)	0.1108(0.0512)	0.9878(0.0115)	0.2009
PLSI	0.2983(0.0259)	0.0546(0.0461)	0.9970(0.0058)	11199

advantage for this purpose; otherwise, PDEE works better. On the other hand, we usually do not have prior information on model homoscedasticity, and further, it is obvious that PLSI is much more time consuming than the PDEE methods: the CPU time consumed of PLSI is more than 10,000 times of those of the PDEE methods. Thus, for robustness consideration, PDEE may be recommendable, as their performance is also competitive in the scenario that is not in favor of PDEE. Meanwhile, PDEE also performs well in estimating $\boldsymbol{\beta}_2$ whereas PLSI does not. Although for the reasons mentioned above, we do not report the estimation for $\boldsymbol{\beta}_2$ for the fairness of comparison.

5.2 Real Data Analyses

In this section, we consider two datasets: NHAMES III data and the Boston housing data. PDEE-SIR and the group dimension reduction estimator are applied to estimate the partial (mean) dimension reduction subspaces for further analyses.

5.2.1 NHAMES III Data. The alcoholism study (Pfeiffer and Bura 2008) we discussed in Section 1 fits exactly in the context of the partial dimension reduction. The aim of this study was to classify men aged 40 years or older into two groups: heavy drinkers and abstainers, combining nine serum biomarkers to build a screening device, while controlling age. The predictors are hematocrit, sodium, chloride, phosphorus, uricacid, blood glucose, blood urea nitrogen, alkaline phosphatase, albumin, and age; age is included as a predictor since it affects both the drinking pattern and the values of the nine biomarkers. This study suggests strong age effects for drinking pattern. Specifically, for men aged 72 years or older, only 5% are heavy drinkers, comparing to 63% for the younger group consisting of men aged between 40 and 71. Hence, when building the screening device based on serum biomarkers, it is more sensible not to mix age with the biomarkers. We take W = age and $\mathbf{X}$ as the nine biomarkers. The goal of this study is to search for $\boldsymbol{\beta}^T\mathbf{X}$ such that $Y \perp\!\!\!\perp \mathbf{X}|(\boldsymbol{\beta}^T\mathbf{X}, W)$, which is exactly a problem of the inference of the partial central subspace $\mathcal{S}^{(W)}_{Y|\mathbf{X}}$.

We apply PDEE-SIR to infer about the partial central subspace $\mathcal{S}^{(W)}_{Y|\mathbf{X}}$. The BIC criterion we discussed in Section 3 yields $\hat{d} = 1$, and the resulting estimated direction is $\hat{\boldsymbol{\beta}}^T = (-0.706,\ 0.0186,\ -0.0065,\ -0.1486,\ -0.066,\ 0.0000032,\ 0.0274,\ 0.0015,\ -0.151)$. Also, the direction seems not to be included in variance and thus, as was commented in Section 1, the GDR (Li, Li, and Zhu 2010) can thus be modified to infer about the partial central mean subspace, regarding W itself as a projection in the real line and then $\theta = 1$ for this projection. In other words, we consider a subspace with structural dimension 1 about W. Since the constant variance conditions required by the partial SAVE (Shao, Cook, and Weisberg 2009) and the partial DR (Li and Wang 2007) are not satisfied for this data, we did not apply PDEE via those two approaches.

Figure 1(a) shows the receiver operating characteristic (ROC) curves for our screening score of the composite measure of the nine biomarkers and the one by the GDR. They both perform similarly and the area under curve (AUC) values are also very close (0.773 and 0.775, respectively). Figure 1(b) and 1(c) shows the ROC curves for both methods while conditioning on age ≤ 42 and age ≥ 72, respectively. From the three ROCs, we cannot say that any method dominates the other. The composite biomarkers from these two methods may be considered to perform similarly here. However, our method is computationally more efficient since our approach avoids nonparmetric smoothing.

5.2.2 Boston Housing Data. In this subsection, we revisit a frequently studied dataset and obtain some new observations. The Boston Housing dataset was originally analyzed by Harrison and Rubinfeld (1978). It contains information collected by the U.S. Census Service concerning housing in the area of Boston. The data consist of 14 variables or features and 506 data points. Variables are per capita crime rate by town (crime rate); proportion of residential land zoned for lots over 25,000 sq. ft. (zn); proportion of nonretail business acres per town (indus); Charles River dummy variable (1 if tract bounds river; 0 otherwise) (chas); nitric oxide concentration (parts per 10 million) (nox); average number of rooms per dwelling (rm); proportion of owner-occupied units built prior to 1940 (age); weighted distances to five Boston employment centers (dis); index of accessibility to radial highways (rad); full value property tax per $10,000 (tax); pupil-teacher ratio by town (ptratio); $(B - 0.63)^2$, where B is the proportion of blacks by town; percentage of lower status of the population (lstat); and median value of the owner-occupied homes in $1000's (medv). The logarithm of medv is taken as the dependent predictor, others are taken as the independent predictors.

This dataset has been analyzed several times in the literature for the dimension reduction purposes by treating all the predictors equally in estimating the central subspace such as Chen and Li (1998), Zhou and He (2008), and Chen, Zhou, and Cook (2010). As suggested by Wang et al. (2010), the predictor

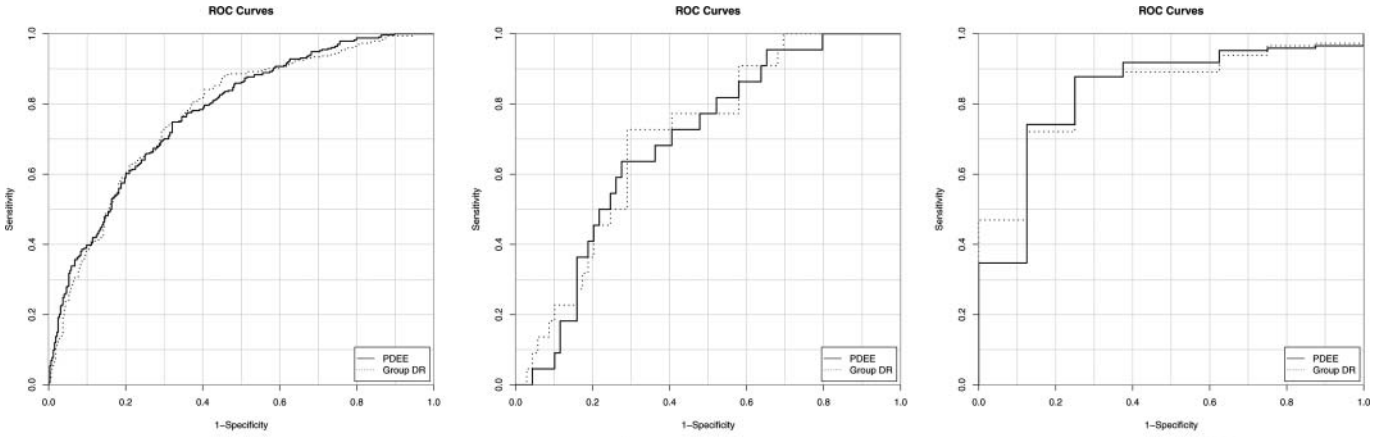

Figure 1. NHANES data: receiver operating characteristic curves for the derived composite biomarkers from PDEE and Group DR.

chas does not have impact for the housing price and is excluded from our analysis. When we use SIR to identify the central subspace, both BIC and sequential test methods yield $\hat{d} = 3$ as its dimension. The R^2 value, which will be defined below, is 0.88. However, as Sentürk and Müller (2005) pointed out, crime rate plays an important role on the housing price, and should be treated discriminately, which also agrees with common sense. Thus, we do not put it in the combinations of the predictors instead treating it as a special predictor in modeling. When using the partial dimension reduction and GDR method, crime rate is regarded as W. $\mathbf{X}$ is the vector of the other 11 predictors. Hence we identify the space spanned by the linear combinations of $\mathbf{X}$, $\boldsymbol{\beta}^T\mathbf{X}$, such that $Y \perp\!\!\!\perp \mathbf{X} | (\boldsymbol{\beta}^T\mathbf{X}, W)$.

We apply PDEE-SIR to infer about the partial central subspace $\mathcal{S}^{(W)}_{Y|\mathbf{X}}$. The number of slices is taken to be five as in the simulation studies. The dimension $\hat{d} = 2$ of the partial central subspace is determined by the BIC criterion. GDR is also applied, and the BIC criterion also infers that the dimension of the relevant subspace is $\hat{d} = 2$. The estimated directions from both methods are reported in Table 4. To check the fitting effects, we fit regression models nonparametrically with predictor $(\hat{\boldsymbol{\beta}}^T X, W)$, where $\hat{\boldsymbol{\beta}}$ is the estimator for $\boldsymbol{\beta}$ by PDEE-SIR or GDR. We adopt r^2 to measure the fitting effects, where $r^2 = (\text{SST} - \text{SSE})/\text{SST}$, $\text{SST} = \sum(y_i - \bar{y})^2$, $\text{SSE} = \sum(y_i - \hat{y}_i)^2$ and $\hat{y}_i$ are the fitted response values.

From Table 4, we see that the reported values of r^2 are 0.9622 and 0.9139, respectively, for PDEE-SIR and GDR, both are larger than the 0.88 from SIR while treating all predictors indiscriminately. This finding suggests that it is desirable to treat crime rate specially. Moreover, PDEE-SIR helps make a better fit than GDR.

6. FURTHER DISCUSSION

Other than the application to model (1.2), our method may also be applied to the following model considered by Li, Zhu, and Zhu (2011):

$$Y = \gamma Z + \psi(\boldsymbol{\beta}^T\mathbf{X}, W) + \epsilon, \tag{6.1}$$

where $\psi(.)$ denotes an unknown smooth function, γ is an unknown parameter, and $\boldsymbol{\beta}$ is an unknown $p \times d$ orthonormal matrix with $d \le p$. The main interest therein is to estimate γ with the aid of partial dimension reduction estimation of $\boldsymbol{\beta}$ in a consistent and link-free fashion, since $\text{Span}(\boldsymbol{\beta}) = \mathcal{S}^{(W,Z)}_{Y|\mathbf{X}}$. In their article, they only deal with the case where both W and Z are categorical since there is no existing partial dimension reduction method available to handle continuous Z or W. In contrast, our method can easily deal with the problems with continuous Z and (or) W. PDEE can again estimate β consistently with the root-n convergence rate and then the asymptotic normality of a least squares estimator of γ could be derived in the way described in Section 4. Future research along this direction is under way.

Another issue is that of handling the dimension q of $\mathbf{W}$. It is clear that our method without iteration is limited to handling small or moderate dimension q. This is because it involves a high-dimensional integral or average that may affect estimation accuracy. Although our method has already partly avoided the curse of dimensionality without nonparametric smoothing, the high-dimensional integral is still a big issue in practice. This is the cost we have to pay for using this new methodology. This is also the reason why we suggest a marginal average in the estimation procedure. How to handle large q deserves further study.

Table 4. Estimated directions by PDEE-SIR and GDR, and R^2 where W = crime rate

Method	R^2	lstat	age	nox	rad	tax	ptratio	b	dis	zn	rm	indus
PDEE-SIR	**0.9622**	−0.0058	−0.0063	−0.5907	0.0494	−0.0018	−0.0661	0.0005	−0.1778	0.0100	0.7825	−0.0078
		−0.0343	−0.0028	−0.9564	0.0228	−0.0007	−0.0278	0.0005	−0.0055	−0.0032	−0.2873	0.0134
Group DR	**0.9139**	0.0535	0.0010	−0.3968	−0.0142	0.0005	0.0402	−0.5307	0.0384	−0.0007	−0.0056	0.7900
		0.0629	−0.0068	0.6188	0.04	−0.0021	−0.0521	−0.7297	−0.0955	−0.0015	−0.0137	−0.2589

APPENDIX

Proof of Proposition 1

Let $\boldsymbol{\alpha}$ be an orthogonal basis of $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$. Hence, we have $Y \perp\!\!\!\perp \mathbf{X}|(\boldsymbol{\alpha}^T\mathbf{X}, \mathbf{W})$. From the definition of conditional distribution of Y and $\mathbf{X}$ when both $\boldsymbol{\alpha}^T\mathbf{X}$ and $\mathbf{W}$ are given, we can see easily that it is equivalent to that for all $\mathbf{t}$, the conditional distribution of them when both $\boldsymbol{\alpha}^T\mathbf{X}$ and $W(\mathbf{t})$ are given. That is, $Y \perp\!\!\!\perp \mathbf{X}|(\boldsymbol{\alpha}^T\mathbf{X}, \mathbf{W})$ is equivalent to that for all $\mathbf{t}$, $Y \perp\!\!\!\perp \mathbf{X}|(\boldsymbol{\alpha}^T\mathbf{X}, W(\mathbf{t}))$. This is because, by the conditional independence, for any $\mathbf{t}$, the conditional distributions have the following equalities

$$\begin{aligned}
&P(Y \le y, \mathbf{X} \le \mathbf{x}|\boldsymbol{\alpha}^T\mathbf{X} \le \boldsymbol{\alpha}^T\mathbf{x}, \mathbf{W} \le \mathbf{t}) \\
&= \frac{P(Y \le y, \mathbf{X} \le \mathbf{x}, \boldsymbol{\alpha}^T\mathbf{X} \le \boldsymbol{\alpha}^T\mathbf{x}, \mathbf{W} \le \mathbf{t})}{P(\boldsymbol{\alpha}^T\mathbf{X} \le \boldsymbol{\alpha}^T\mathbf{x}, \mathbf{W} \le \mathbf{t})} \\
&= \frac{P(Y \le y, \boldsymbol{\alpha}^T\mathbf{X} \le \boldsymbol{\alpha}^T\mathbf{x}, \mathbf{W} \le \mathbf{t})}{P(\boldsymbol{\alpha}^T\mathbf{X} \le \boldsymbol{\alpha}^T\mathbf{x}, \mathbf{W} \le \mathbf{t})} \\
&\quad \times \frac{P(Y \le y, \boldsymbol{\alpha}^T\mathbf{X} \le \boldsymbol{\alpha}^T\mathbf{x}, \mathbf{W} \le t)}{P(\boldsymbol{\alpha}^T\mathbf{X} \le \boldsymbol{\alpha}^T\mathbf{x}, \mathbf{W} \le \mathbf{t})} \\
&= \frac{P(Y \le y, \boldsymbol{\alpha}^T\mathbf{X} \le \boldsymbol{\alpha}^T\mathbf{x}, I(\mathbf{W} \le \mathbf{t}) = 1)}{P(\boldsymbol{\alpha}^T\mathbf{X} \le \boldsymbol{\alpha}^T\mathbf{x}, I(\mathbf{W} \le \mathbf{t}) = 1)} \\
&\quad \times \frac{P(\mathbf{X} \le \mathbf{x}, \boldsymbol{\alpha}^T\mathbf{X} \le \boldsymbol{\alpha}^T\mathbf{x}, I(\mathbf{W} \le \mathbf{t}) = 1)}{P(\boldsymbol{\alpha}^T\mathbf{X} \le \boldsymbol{\alpha}^T\mathbf{x}, I(\mathbf{W} \le \mathbf{t}) = 1)} \\
&= \frac{P(Y \le y, \mathbf{X} \le \mathbf{x}, \boldsymbol{\alpha}^T\mathbf{X} \le \boldsymbol{\alpha}^T\mathbf{x}, I(\mathbf{W} \le \mathbf{t}) = 1)}{P(\boldsymbol{\alpha}^T\mathbf{X} \le \boldsymbol{\alpha}^T\mathbf{x}, I(\mathbf{W} \le \mathbf{t}) = 1)} \\
&= P(Y \le y, \mathbf{X} \le \mathbf{x}|\boldsymbol{\alpha}^T\mathbf{X} \le \boldsymbol{\alpha}^T\mathbf{x}, I(\mathbf{W} \le \mathbf{t}) = 1) \\
&= P(Y \le y, \mathbf{X} \le \mathbf{x}|\boldsymbol{\alpha}^T\mathbf{X} \le \boldsymbol{\alpha}^T\mathbf{x}, \mathbf{W}(\mathbf{t}) = 1).
\end{aligned}$$

Thus, $\mathcal{S}_{Y|\mathbf{X}}^{(W(t))} \subseteq \mathcal{S}_{Y|\mathbf{X}}^{(W)}$ and $\bigcup_{\mathbf{t}} \mathcal{S}_{Y|\mathbf{X}}^{(W(t))} = \mathcal{S}_{Y|\mathbf{X}}^{(W)}$. □

Proof of Theorem 1

Let P be the projection onto $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$, $P_{M(t)}$ be the projection onto $\text{Span}\{M(\mathbf{t})\}$, and P_M be the projection onto $\text{Span}\{M\}$. From Proposition 1, $\text{Span}\{M(\mathbf{t})\} = \mathcal{S}_{Y|\mathbf{X}}^{(W(t))} \subseteq \mathcal{S}_{Y|\mathbf{X}}^{(W)}$, for any $\mathbf{t}$. Let $\boldsymbol{\nu} \perp \mathcal{S}_{Y|\mathbf{X}}^{(W)}$, then, $\boldsymbol{\nu} \perp \text{Span}\{M(\mathbf{t})\}$ for all $\mathbf{t} \in \mathbb{R}_{\mathbf{T}}^q$. Hence, $M\boldsymbol{\nu} = \text{E}(M(\mathbf{T})\boldsymbol{\nu}) = 0$. Thus $\text{Span}\{M\} = \text{Span}(\text{E}\{M(\mathbf{T})\}) \subseteq \mathcal{S}_{Y|\mathbf{X}}^{(W)}$.

We now show that $\mathcal{S}_{Y|\mathbf{X}}^{(W)} \subseteq \text{Span}\{M\}$. Equivalently, we show that

$$\begin{aligned}
&P\{\mathbf{X} \le x, Y \le y|(P_M\mathbf{X}, \mathbf{W})\} \\
&\quad = P\{\mathbf{X} \le x|(P_M\mathbf{X}, \mathbf{W})\}P\{Y \le y|(P_M\mathbf{X}, \mathbf{W})\}, \qquad \text{(A.1)}
\end{aligned}$$

for all $x \in \mathbb{R}^p$ and $y \in \mathbb{R}^1$. Suppose that $\boldsymbol{\xi} \perp\!\!\!\perp \text{Span}\{M\}$, then $\boldsymbol{\xi}^T M\boldsymbol{\xi} = \boldsymbol{\xi}^T\text{E}\{M(\mathbf{T})\}\boldsymbol{\xi} = 0$, which implies that $\boldsymbol{\xi}^T M(\mathbf{T})\boldsymbol{\xi} = 0$ almost surely with respect to $F(.)$ conditionally on the support of $\mathbf{W}$, where $F(.)$ is the cumulative distribution function of $\mathbf{T}$. Hence, $\text{Span}\{M(\mathbf{t})\} \subseteq \text{Span}\{M\}$ on a subset A of $\mathbb{R}_{\mathbf{W}}^q$ with $F(A) = 1$. So, $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|(P_M\mathbf{X}, W(\mathbf{t}))$, for $\mathbf{t} \in A$. Therefore, we have $P\{\mathbf{X} \le x, Y \le y|(P_M\mathbf{X}, W(\mathbf{t}))\} = P\{\mathbf{X} \le x|(P_M\mathbf{X}, W(\mathbf{t}))\}P\{Y \le y|(P_M\mathbf{X}, W(\mathbf{t}))\}$ for any $\mathbf{t} \in A$. Also, since

$$\sigma(\mathbf{W}) = \bigcup_{\mathbf{t} \in \mathbb{R}_{\mathbf{W}}^q} \{\mathbf{W} \le \mathbf{t}\},$$

where $\sigma(\mathbf{W})$ is σ-field associated with $\mathbf{W}$ and $\mathbb{R}_{\mathbf{W}}^q$ is the support of $\mathbf{W}$. We have

$$Y \perp\!\!\!\perp \mathbf{X}|(P_M\mathbf{X}, \mathbf{W}).$$

And this verifies (A.1). □

Proof of Proposition 2

By similar argument as that of Li, Wen, and Zhu (2008) in the proof of theorem 3.2, under condition (a), we have

$$M_{l_n,n} - M = [\text{E}_{l_n}M(\mathbf{T}) - \text{E}M(\mathbf{T})] + \frac{1}{n}\sum_{i=1}^{n} U_{i,n} + \frac{1}{l_n}\sum_{i=1}^{l_n} R_n(\mathbf{T}_i), \qquad \text{(A.2)}$$

where $U_{i,n} = \frac{1}{l_n}\sum_{j=1}^{l_n}\phi(\mathbf{X}_i, Y_i, \mathbf{W}_i, \mathbf{T}_j)$, $n = 1, 2, \ldots$. The first term of (A.2) has order $O_p(l_n^{-\frac{1}{2}})$, which is no greater than $O_p(n^{-\frac{1}{2}})$ since $l_n = O_p(n)$. The norm of the third term is bounded from above by $\sup_{\mathbf{t}\in\mathbb{R}_{\mathbf{T}}^q}\|R_n(\mathbf{t})\| = o_p(n^{-\frac{1}{2}})$. So, we only need to show that the second term in (A.2), $S_n = \frac{1}{n}\sum_{i=1}^{n} U_{i,n}$, is of order $O_p(n^{-\frac{1}{2}})$. Since $\text{vec}(S_n) = \frac{1}{nl_n}\sum_{i=1}^{n}\sum_{j=1}^{l_n}\text{vec}(\phi(\mathbf{X}_i, Y_i, \mathbf{W}_i, \mathbf{T}_j))$ has mean 0 and variance matrix $\frac{1}{n^2}\frac{1}{l_n^2}\sum_{i=1}^{n}\sum_{j=1}^{l_n}\sum_{i'=1}^{n}\sum_{j'=1}^{l_n}\text{E}[\text{vec}(\phi(\mathbf{X}_i, Y_i, \mathbf{W}_i, \mathbf{T}_j))\,\text{vec}(\phi(\mathbf{X}_{i'}, Y_{i'}, \mathbf{W}_{i'}, \mathbf{T}_{j'}))^T]$, which can be reduced to $\frac{1}{nl_n}\text{E}[\text{vec}(\phi(\mathbf{X}, Y, \mathbf{W}, \mathbf{T})\,\text{vec}(\phi(\mathbf{X}, Y, \mathbf{W}, \mathbf{T}))^T] + \frac{l_n-1}{nl_n}\text{E}[\text{vec}(\phi(\mathbf{X}, Y, \mathbf{W}, \mathbf{T}_1)\,\text{vec}(\phi(\mathbf{X}, Y, \mathbf{W}, \mathbf{T}_2))^T]$, where $\mathbf{T}_1 \perp\!\!\!\perp (\mathbf{X}, Y, W)$, $(\mathbf{T}_1, \mathbf{T}_2) \perp\!\!\!\perp (\mathbf{X}, Y, W)$, and $\mathbf{T}_1 \perp\!\!\!\perp \mathbf{T}_2$. By similar argument as of Li, Wen, and Zhu (2008) in the proof of theorem 3.2, we have

$$\begin{aligned}
\text{var}(\text{vec}(S_n)) = &\frac{1}{n}\text{E}\left[\text{vec}(\phi(\mathbf{X}, Y, \mathbf{W}, \mathbf{T}_1)\,\text{vec}(\phi(\mathbf{X}, Y, \mathbf{W}, \mathbf{T}_2))^T\right] \\
&+ O\left((nl_n)^{-1}\right).
\end{aligned}$$

By lemma A.1 of Li, Wen, and Zhu (2008), we can then show that the second term of (A.2) is of order $O_p(n^{-\frac{1}{2}})$. □

Proof of Theorem 2

By a similar argument as that of Li, Wen, and Zhu (2008) in the proof of theorem 3.2, under condition (a), we have

$$M_{n,n} - M = [\text{E}_n M(\widetilde{\mathbf{W}}) - \text{E}M(\widetilde{\mathbf{W}})] + \frac{1}{n}\sum_{i=1}^{n} U_{i,n} + \frac{1}{n}\sum_{i=1}^{n} R_n(\mathbf{W}_i), \qquad \text{(A.3)}$$

where $U_{i,n} = \frac{1}{n}\sum_{j=1}^{n}\phi(\mathbf{X}_i, Y_i, \mathbf{W}_i, \mathbf{W}_j)$, $n = 1, 2, \ldots$. The first term of (A.3) has order that admits a linear representation, and the norm of the third term is bounded from above by $o_p(n^{-\frac{1}{2}})$ under condition (b). Hence, we need to show that $\frac{1}{n}\sum_{i=1}^{n} U_{i,n}$ is an asymptotically linear estimator for the asymptotic normality. With an argument similar to that of Zhu et al. (2010), $\frac{1}{n}\sum_{i=1}^{n} U_{i,n}$ can be written as a V-statistic and can be approximated by an U-statistic as follows: $\frac{1}{n}\sum_{i=1}^{n} U_{i,n} = \frac{1}{n(n-1)}\sum_{j<i}[\phi(\mathbf{X}_i, Y_i, W_i, W_j) + \phi(\mathbf{X}_j, Y_j, W_j, W_i)] + o_p(n^{-\frac{1}{2}}) = U_n + o_p(n^{-\frac{1}{2}})$, where U_n is a second-order U-statistic. Thus, U_n can be approximated by its projection $\hat{U}_n = \text{E}(U_n \mid \mathbf{X}_i, Y_i, \mathbf{W}_i) = \frac{1}{n}\sum_{i=1}^{n}[\text{E}(\phi(\mathbf{X}_i, Y_i, \mathbf{W}_i, \widetilde{\mathbf{W}})|(\mathbf{X}_i, Y_i, \mathbf{W}_i)) + \text{E}(\phi(\mathbf{X}, Y, \mathbf{W}, \mathbf{W}_i)|(\mathbf{X}_i, Y_i, \mathbf{W}_i))]$, which admits a linear representation (see Serfling 1980). Also, notice that $U_n = \hat{U}_n + o(\frac{\log n}{n})$ almost surely. By the Lindeberg–Levy central limit theorem, we then have the desired result. □

[Received July 2011. Revised October 2012.]

REFERENCES

Bura, E., and Cook, R. D. (2001), "Extending Sliced Inverse Regression: The Weighted Chi-Squared Test," *Journal of the American Statistical Association*, 96, 996–1003. [240]

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997), "Generalized Partially Linear Single-Index Models," *Journal of the American Statistical Association*, 92, 477–489. [238,240]

Chen, C. H., and Li, K. C. (1998), "Can SIR be as Popular as Multiple Linear Regression," *Statistica Sinica*, 8, 289–316. [243]

Chen, X., Zhou, C. L., and Cook, R. D. (2010), "Coordinate-Independnet Sparse Sufficient Dimension Reduction and Variable Selection," *The Annals of Statistics*, 38, 3696–3723. [243]

Chiaromonte, F., Cook, R. D., and Li, B. (2002), "Sufficient Dimension Reduction in Regressions With Categorical Predictors," *The Annals of Statistics*, 30, 475–497. [237,238,239]

Cook, R. D. (1998), *Regression Graphics*, New York: Wiley. [237]

Cui, X., Härdle, W., and Zhu, L. X. (2011), "Generalized Single-Index Models: The EFM Approach," *The Annals of Statistics*, 39, 1658–1688. [240]

Härdle, W., Liang, H., and Gao, J. T. (2000), *Partially Linear Models*, Heidelberg: Physica-Verlag. [238]

Harrison, D., and Rubinfeld, D. L. (1978), "Hedonic Prices and the Demand for Clean Air," *Journal of Environmental Economics and Management*, 5, 81–102. [243]

Li, B., Cook, R. D., and Chiaromonte, F. (2003), "Dimension Reduction for the Conditional Mean in Regressions With Categorical Predictors," *The Annals of Statistics*, 31, 1636–1668. [237]

Li, B., and Dong, Y. (2009), "Dimension Reduction for Non-Elliptically Distributed Predictors," *The Annals of Statistics*, 37, 1272–1298. [241]

Li, B., and Wang, S. (2007), "On Directional Regression for Dimension Reduction," *Journal of the American Statistical Association*, 102, 997–1008. [239,242,243]

Li, B., Wen, S. Q., and Zhu, L. X. (2008), "On a Projective Resampling Method for Dimension Reduction With Multivariate Responses," *Journal of the American Statistical Association*, 103, 1177–1186. [245]

Li, K. C. (1991), "Sliced Inverse Regression for Dimension Reduction" (with discussion), *Journal of the American Statistical Association*, 86, 316–342. [237,238,240]

Li, L., Li, B., and Zhu, L. X. (2010), "Groupwise Dimension Reduction," *Journal of the American Statistical Association*, 105, 1188–1201. [237,243]

Li, L., Zhu, L., and Zhu, L. X. (2011), "Inference on Primary Parameter of Interest With Aid of Dimension Reduction Estimation," *Journal of the Royal Statistical Society,* Series B, 73, 59–80. [238,240,244]

Li, Y., and Zhu, L. X. (2007), "Asymptotics for Sliced Average Variance Estimation," *The Annals of Statistics*, 35, 41–69. [238,242]

Pfeiffer, R. M., and Bura, E. (2008), "A Model Free Approach to Combining Biomarkers," *Biometrical Journal*, 50, 558–570. [237,241,243]

Sentürk, D., and Müller, H. G. (2005), "Covariate Adjusted Correlation Analysis via Varying Coefficient Models," *Scandinavian Journal of Statistics*, 32, 365–383. [244]

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, New York: Wiley. [245]

Shao, Y., Cook, R. D., and Weisberg, S. (2009), "Partial Central Subspace and Sliced Average Variance Estimation," *Journal of Statistical Planning and Inference*, 139, 952–961. [239,243]

Stute, W., and Zhu, L. X. (2005), "Nonparametric Checks for Single-Index Models," *The Annals of Statistics*, 33, 1048–1083. [238]

Wang, J. L., Xue, L., Zhu, L. X., and Chong, Y. S. (2010), "Estimation for a Partial-Linear Single-Index Model," *The Annals of Statistics*, 30, 475–497. [238,240,243]

Wen, X., and Cook, R. D. (2007), "Optimal Sufficient Dimension Reduction in Regressions With Categorical Predictors," *Journal of Statistical Planning and Inference*, 137, 1961–1978. [237,238]

Xia, Y., and Härdle, W. (2006), "Semi-Parametric Estimation of Partially Linear Single-Index Models," *Journal of Multivariate Analysis*, 97, 1162–1184. [238,242]

Yin, X. (2005), "Non-Parametric Estimation of Direction in Single-Index Models With Categorical Predictors," *Australian and New Zealand Journal of Statistics*, 47, 141–161. [241]

Yu, Y., and Ruppert, D. (2002), "Penalized Spline Estimation of Partially Linear Single-Index Models," *Journal of the American Statistical Association*, 97, 1042–1054. [238]

Yu, Z., Zhu, L. X., and Wen, X. (2012), "On Model Free Conditional Coordinate Test for Regressions," *Journal of Multivariate Analysis*, 109, 61–72. [238]

Zhou, J., and He, X. (2008), "Dimension Reduction Based on Constrained Canonical Correlation and Variable Filtering," *The Annals of Statisitcs*, 36, 1649–2668. [243]

Zhu, L. P., Wang, T., Zhu, L. X., and Ferré, L. (2010), "Sufficient Dimension Reduction Through Discretization-Expectation Estimation," *Biometrika*, 97, 295–304. [238,240,245]

Zhu, L. X. (2005), *Nonparametric Monte Carlo Tests and Their Applications*, New York: Springer. [238]

Zhu, L. X., and Fang, K. T. (1996), "Asymptotics for the Kernel Estimates of Sliced Inverse Regression," *The Annals of Statistics*, 24, 1053–1067. [239]

Zhu, L. X., Miao, B. Q., and Peng, H. (2006), "Sliced Inverse Regression With Large Dimensional Covariates," *Journal of the American Statistical Association*, 101, 630–643. [240]

Zhu, L. X., and Ng, K. W. (1995), "Asymptotics for Sliced Inverse Regression," *Statistica Sinica*, 5, 727–736. [238,239]