Journal of the American Statistical Association

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

An Effective Semiparametric Estimation Approach for the Sufficient Dimension Reduction Model

Ming-Yueh Huang & Chin-Tsang Chiang

To cite this article: Ming-Yueh Huang & Chin-Tsang Chiang (2016): An Effective Semiparametric Estimation Approach for the Sufficient Dimension Reduction Model, Journal of the American Statistical Association, DOI: 10.1080/01621459.2016.1215987

To link to this article: http://dx.doi.org/10.1080/01621459.2016.1215987

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2017, VOL. 0, NO. 0, 1–15, Theory and Methods
https://doi.org/10.1080/01621459.2016.1215987

Taylor & Francis
Taylor & Francis Group

An Effective Semiparametric Estimation Approach for the Sufficient Dimension Reduction Model

Ming-Yueh Huang and Chin-Tsang Chiang

Institute of Applied Mathematical Sciences, National Taiwan University, Taipei, Taiwan

ABSTRACT

In the exploratory data analysis, the sufficient dimension reduction model has been widely used to characterize the conditional distribution of interest. Different from the existing approaches, our main achievement is to simultaneously estimate two essential elements, basis and structural dimension, of the central subspace and the bandwidth of a kernel distribution estimator through a single estimation criterion. With an appropriate order of kernel function, the proposed estimation procedure can be effectively carried out by starting with a dimension of zero until the first local minimum is reached. Meanwhile, the optimal bandwidth selector is ensured to be a valid tuning parameter for the central subspace estimator. An important advantage of this estimation technique is its flexibility to allow a response to be discrete and some of covariates to be discrete or categorical providing that a certain continuity condition holds. Under very mild assumptions, we further derive the uniform consistency of the introduced optimization function and the consistency of the resulting estimators. Moreover, the asymptotic normality of the central subspace estimator is established with an estimated rather than exact structural dimension. In extensive simulations, the developed approach generally outperforms the competitors. Data from previous studies are also used to illustrate the proposal. On the whole, our methodology is very effective in estimating the central subspace and conditional distribution, highly flexible in adapting diverse types of a response and covariates, and practically feasible in obtaining an asymptotically optimal and valid bandwidth estimator. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received July 2014
Accepted July 2016

KEYWORDS
Asymptotic normality; Central subspace; Consistency; Cross-validation estimation; Inverse regression estimation; Optimal bandwidth; Pseudo-least integrated squares estimation; Semiparametric efficiency bound; Semiparametric estimation; Structural dimension; Sufficient dimension reduction

1. Introduction

A crucial and challenging issue for the conditional distribution $F_Y(y|x)$ of a univariate response Y on p-dimensional covariates $X = (X_1, \ldots, X_p)^\top$ is to correctly identify the most informative covariate space. Without prior knowledge about the underlying distribution structure, one could reasonably invoke a general semiparametric formulation of the following form:

$$F_Y(y|x) = F(y, B^\top x), \tag{1.1}$$

where $F(y, u)$, $u = (u_1, \ldots, u_d)^\top$, is an unknown $(d+1)$-variate function and $B = (\beta_1, \ldots, \beta_d)$ is a $p \times d$ full-rank coefficient matrix. It is noteworthy that the considered model preserves full regression information through a set of linear combinations $B^\top X$ and requires no distributional assumption on the regression function $F(y, u)$. In addition to these features, the dimension d of covariate space plays an essential role in reflecting the degree of dimension reduction. Thus, a fully nonparametric regression is a direct implication when $d = p$ and several semiparametric regressions such as single-index distribution models of Cosslett (1983), Delecroix, Härdle, and Hristache (2003), and Hall and Yao (2005) are special cases with $d < p$. As shown by Zeng and Zhu (2010), model (1.1) is equivalent to the sufficient dimension reduction (SDR) model (see Li 1991; Cook 1998), that is $Y \perp\!\!\!\perp X|B^\top X$, in which $\perp\!\!\!\perp$ stands for conditional independence.

For the dimension-reduction problem, the primary research interest often focuses on estimating the SDR subspace $\mathcal{S}(B)$ spanned by the column vectors of B. When the smallest SDR subspace exists and is unique, such a subspace, which is termed the central subspace (CS) and is denoted by $\mathcal{S}(B_0)$ or, alternatively, by $\mathcal{S}_{Y|X}$ with $B_0 = (\beta_{01}, \ldots, \beta_{0d_0})$ being a $p \times d_0$ basis matrix and d_0 being the so-called structural dimension, is the major parameter of interest. As for the estimation of $\mathcal{S}_{Y|X}$, it includes the estimation of B_0 and the determination of d_0. Based on a low-dimensional estimation of the conditional mean $E[X|Y]$, Li (1991) first developed the slice inverse regression approach to estimate $\mathcal{S}(B_0)$ without relying on a high-dimensional estimation of the regression function $F(y, u_0)$, $u_0 = (u_{01}, \ldots, u_{0d_0})^\top$. Since the estimation might fail for the degeneration of $\mathrm{cov}(E[X|Y])$ in any direction orthogonal to $\mathcal{S}(B_0)$, Cook and Weisberg (1991) proposed the sliced average variance estimation to overcome this drawback. By efficiently arranging the empirical directions according Y and its independent copy, Li and Wang (2007) further presented the direction regression estimator as an alternative. In the last two decades, several approaches have also been intensively developed based on higher order moments. Different from the existing inverse regression approaches, Zhu, Yu, and Zhu (2010) devised a sparse-decomposition estimation to simultaneously estimate B_0 and determine d_0. As for the structural dimension

CONTACT Chin-Tsang Chiang chiangct@ntu.edu.tw Institute of Applied Mathematical Sciences, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei 10617, Taiwan.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

estimator of d_0, their simulation study showed that the proposed procedure outperforms the Bayesian information criterion procedure of Zhu, Miao, and Peng (2006). In these estimation techniques, the support $\mathcal{Y}$ of Y should be divided into appropriate slices and the resulting estimators might be sensitive toward the choice of slice numbers. To solve this problem, the selection of a slicing scheme can be easily implemented by the fusing method of Cook and Zhang (2014). Without such an operational limitation, Zhu, Zhu, and Feng (2010) further proposed a cumulative slicing estimation, which is based on the conditional moments of X given the cumulative counting process $N_y = \mathbf{1}(Y \leq y)$ over $\mathcal{Y}$. In conclusion, inverse regression approaches make no assumption on $F(y, u_0)$, enjoy the simplicity and computational efficiency in implementation, and can recover $\mathcal{S}_{Y|X}$ exhaustively through low-dimensional moments under some suitable conditions.

Without making assumptions on some conditional moments of X given Y, Zhu and Zeng (2006) used the Fourier transform of the gradient of $F_Y(y|x)$ to generate a candidate matrix for $\mathcal{S}_{Y|X}$ and presented an explicit CS estimator for data with multivariate normal covariates. By extending the minimum average variance estimation (MAVE) of Xia et al. (2002), Xia (2007) estimated $\mathcal{S}_{Y|X}$ by means of the double-kernel local linear smoothing technique, which was originally developed by Fan and Yim (2004) for estimating a fully nonparametric conditional density function. Since this approach is sensitive to outliers and unable to deal with a discrete response, Wang and Xia (2008) proposed the sliced regression estimation to reduce the effect of outliers and remove the limitation on the type of a response. Based on an ensemble of the MAVEs, Yin and Li (2011) further developed a general dimension reduction approach. Through intensive simulations, the family of characteristic functions was also demonstrated to have better performance than other families of transformations. However, the covariates of interest are required to be continuous in the initial estimation and the proposed CS estimator cannot achieve the $\sqrt{n}$-consistency for $d_0 \geq 4$. Furthermore, the MAVE-type estimators heavily rely on a high-dimensional smoothing technique and usually suffer from the "curse of dimensionality" (see Bellman 1961). Even the numerical precision can be iteratively improved via a low-dimensional kernel smoothing, the choice of optimal bandwidth selectors and the determination of optimal number of slices or sample functions are still open for further exploration. By borrowing the geometric tools, which consist of the constructions of the nuisance tangent space and its orthogonal complement, of Bickel et al. (1998) and Tsiatis (2006), Ma and Zhu (2012) derived the space of influence functions and proposed the regular asymptotic linear estimation for $\mathcal{S}_{Y|X}$. Their estimators are also shown to be the semiparametric counterparts of the most existing estimators. Recently, Ma and Zhu (2013) adopted the local coordinate system of the Grassmann manifold (see Borisenko and Nikolaevskiĭ 1991) to resolve the identifiability problem of B_0 and derived an efficient member, which is asymptotically equivalent to the pseudo maximum likelihood estimator, in the influence function family. Although the semiparametric efficient estimator of Ma and Zhu (2013) can avoid the use of a higher-order kernel function, the asymptotically valid but subjective bandwidths usually lead to poor performance for small samples. As for the determination of d_0, the sliced cross-validation criterion of Wang and Xia (2008) and the bootstrap procedure of Dong and Li (2010) are possible avenues in the context of semiparametric estimation. However, the first procedure needs an appropriate criterion to choose the related tuning parameters and the second one demands large memory space and prohibitive computation time.

In light of the fact that $E[N_y|X = x] = F(y, B_0^\top x)$ for all possible values of (x, y), the CS estimation can be reasonably cast as an integration of estimation for the central mean subspaces (CMS) of N_y over $\mathcal{Y}$. At first sight, the pseudo sum of integrated squares (PSIS) estimation of Chiang and Huang (2012) seems to be a useful strategy to estimate B_0 for a given d_0. However, a direct extension of this estimation criterion with the specification of a second-order kernel function might lead to multiple solutions and inconsistent estimators. To overcome these shortcomings, a new cross-validation version of the PSIS is proposed to simultaneously estimate the parameters (d_0, B_0, h_0), where h_0 is the optimal bandwidth that minimizes an asymptotic weighted mean integrated squared-error of the kernel distribution estimator. According to the best of our knowledge, the existing semiparametric SDR approaches invoke two separate procedures to estimate B_0 and determine d_0. By taking the minimum value of the proposed cross-validated PSIS over all possible d-dimensional basis directions and bandwidths, the resulting function of d is further shown to converge in probability to a strictly convex function with a unique minimizer d_0. In addition, a qth-order kernel function with $q > \max\{d/2 + 1, 2\}$ is designed for each given dimension d. With such a device, the $\sqrt{n}$-consistency and asymptotic normality of the CS estimator can be achieved by the use of an asymptotically optimal bandwidth estimator. It is noted that h_0 plays a major role in the kernel distribution estimator of $F(y, u_0)$ and is relegated as a tuning parameter in the estimators of d_0 and B_0. In contrast with subjective bandwidths in the semiparametric efficient estimator and the MAVE-type estimators, our bandwidth estimator usually avoids the ill-conditioning problem on the pseudo information matrix.

Although the proposed CS estimator cannot attain the semiparametric efficiency bound, its finite-sample performance is generally superior to those of the pseudo-maximum likelihood estimator and the semiparametric efficient estimator of Ma and Zhu (2013). Further, the estimation criterion for $\mathcal{S}_{Y|X}$ and $F(y, u_0)$ can be effectively carried out by a forward algorithm, in which the function of d is computed by starting with $d = 0$ until the first local minimum is reached. As for data with some discrete and/or categorical covariates, the presented SDR approach is still applicable when each of $B_0^\top X$ satisfies the continuity condition of Ichimura (1993) or Horowitz and Härdle (1996). Other attractive attributes of the proposal include:

1. In the spirit of the proposed cross-validation criterion, a new approach can be developed to simultaneously estimate the CMS and mean regression.
2. The asymptotic variance of the CS estimator is independent of the order of a specified kernel function and not affected by the variations of the structural dimension and bandwidth estimators.
3. The asymptotic behavior of the kernel distribution estimator is not sensitive to the asymptotic variation of the CS estimator.

The rest of this article is organized as follows. In Section 2, a cross-validation estimation criterion is proposed to estimate $F(y, B_0^\top x)$. The large sample properties of the CS and structural dimension estimators are further established in Section 3. Moreover, we assess the finite-sample performance of the proposed estimator and its competitors in Section 4 through a series of simulations. In Section 5, our methodology is also applied to data from the studies of house-price in Boston and red Vinho Verde, which is a type of Portuguese wine, preferences. Section 6 provides some concluding remarks and future research directions. As for the technical lemmas and the proofs of the main results, they are relegated to the Appendices in the supplementary materials.

2. Estimation for the SDR Model

In the rest of this article, E^{-i} is computed as a generic estimator E with the ith subject being deleted, $i = 1, \ldots, n$. For a given d_0, we outline the background of the pseudo-least integrated squares estimation (PLISE) for B_0. A cross-validation version of the pseudo sum of integrated squares (PSIS) is further proposed to estimate (B_0, d_0, h_0). Moreover, an effective computational algorithm is provided to carry out the estimation criterion.

2.1. Background of the PLISE

Intrinsically, the SDR model with a known d_0 can be regarded as a semiparametric regression with a finite-dimensional parameter B_0 and an infinite-dimensional parameter $F(y, u_0)$. To avoid the identifiability problem of B_0 in model (1.1), the local coordinate system of the Grassmann manifold is used to parameterize the basis into the form $B_0 = (I_{d_0}, C_0^\top)^\top$, where I_{d_0} is the $d_0 \times d_0$ identity matrix and $C_0 = (\gamma_{01}, \ldots, \gamma_{0d_0})$ is a $(p - d_0) \times d_0$ parameter matrix. Even the map $C_0 \mapsto \mathcal{S}(B_0)$ is one-to-one, it is not onto the Grassmann manifold $\mathrm{Gr}(d_0, \mathbb{R}^{d_0})$. Fortunately, this limitation can be resolved by choosing d_0 significant continuous covariates, labeled by $X_1, \ldots, X_{d_0}$, among all covariates. The coefficients in γ_{0j} are, thus, interpreted as the relative effects of $(X_{d_0+1}, \ldots, X_p)^\top$, compared to X_j, in the jth CS direction, $j = 1, \ldots, d_0$. In addition, a series of indicator or dummy variables, which take only the values of 0 or 1, is customarily used to define different levels of categorical covariates. When some of $B_0^\top X$ are linear combinations of discrete and/or categorical variables, there might be more than one CS up to orthogonal transformations. Without imposing any structure on $F(y, u_0)$ or smoothness condition on $B_0^\top X$, the developed approach cannot be directly adopted to the estimation of $F(y, B_0^\top X)$.

In light of the fact that $E[N_y|X = x] = F(y, B_0^\top x)$ for all possible values of (x, y), the SDR model can be naturally converted to the framework of mean regressions over $\mathcal{Y}$. To simplify the presentation, we define $F_{C_d}(y, u) = P(Y \le y|B_d^\top X = u)$, $F_Y(y) = P(Y \le y)$, $\langle g_1(\cdot), g_2(\cdot)\rangle_{L^2} = \int_{\mathcal{Y}} g_1(y)g_2(y)dF_Y(y)$, and $\|g(\cdot)\|_{L^2} = \sqrt{\langle g(\cdot), g(\cdot)\rangle_{L^2}}$ for any bounded functions $(g_1(y), g_2(y), g(y))$ and $B_d = (I_d, C_d^\top)^\top$ with $C_d = (\gamma_{d1}, \ldots, \gamma_{dd})$ being a $(p - d) \times d$ matrix. The following proposition further gives an insight into the rationale of our approach.

Proposition 1. For a given B_d, $F_{C_d}(y, u)$ minimizes $E[\|N. - G(\cdot, B_d^\top X)\|_{L^2}^2]$ over all $(d + 1)$-variate functions $G(y, u)$'s. Moreover, the basis matrix B of a SDR subspace minimizes $E[\|N. - F_{C_d}(\cdot, B_d^\top X)\|_{L^2}^2]$ over all B_d.

It follows from the first assertion in Proposition 1 that the conditional distribution $F_{C_d}(y, u)$ plays an essential role in the estimation of B. By the existence and uniqueness of $\mathcal{S}_{Y|X}$ and the second assertion, the strict inequality $E[\|N. - F_{C_{d_0}}(\cdot, B_{d_0}^\top X)\|_{L^2}^2] > E[\|N. - F(\cdot, B_0^\top X)\|_{L^2}^2]$ is ensured for all $B_{d_0} = (I_{d_0}, C_{d_0}^\top)^\top \ne B_0$.

In the spirit of the PLISE of Chiang and Huang (2012), $F_{C_d}(y, u)$ can be estimated by a kernel distribution estimator of the form

$$\widehat{F}_{C_d}(y, u) = \frac{\sum_{i=1}^n N_{iy}\mathcal{K}_{q,h_d}(B_d^\top X_i - u)}{\sum_{i=1}^n \mathcal{K}_{q,h_d}(B_d^\top X_i - u)}, \tag{2.1}$$

where $\mathcal{K}_{q,h_d}(u) = \prod_{k=1}^d K_q(u_k/h_{dk})/h_{dk}$, $h_d = (h_{d1}, \ldots, h_{dd})^\top$ is a positive-valued bandwidth vector, and $K_q(\upsilon)$ is a qth-order kernel function, that is, $\int K_q(\upsilon)d\upsilon = 1$, $\int \upsilon^\kappa K_q(\upsilon)d\upsilon = 0, \kappa = 1, \ldots, q - 1$, and $\int \upsilon^q K_q(\upsilon)d\upsilon < \infty$. Here, a qth-order kernel function $K_q(\upsilon)$ is specified to be twice continuously differentiable and symmetric with $q > \max\{d/2 + 1, 2\}$ and bounded support. Let $\widehat{F}_Y(y)$ stand for the empirical distribution of Y. The PLISE estimator $\widehat{B}_{d_0} = (I_{d_0}, \widehat{C}_{d_0})^\top$ of B_0 is defined as the minimizer of

$$\mathrm{PSIS}(C_{d_0}) = \frac{1}{n}\sum_{i=1}^n \int \left(N_{iy} - \widehat{F}_{C_{d_0}}\left(y, B_{d_0}^\top X_i\right)\right)^2 d\widehat{F}_Y(y). \tag{2.2}$$

In principle, the above PSIS penalizes discrepancies in high use areas of $\mathcal{Y}$ more severely than low use areas of $\mathcal{Y}$. To maintain the numerical stability, the smallest even integer is generally preferred for q in practical implementation. In the next section, we further establish the $\sqrt{n}$-consistency and asymptotic normality of $\widehat{B}_{d_0}$ under the regularity conditions and the bandwidth constraint

A0. h_{dk}, $k = 1, \ldots, d$, fall in the interval $H_{d,n} \triangleq (h_{dl}n^{-1/\max\{2d+2,d+4\}}, h_{du}n^{-1/4q})$ for some positive constants h_{dl} and h_{du}.

Since

$$E\left[\left\|N. - F_{C_d}\left(\cdot, B_d^\top X\right)\right\|_{L^2}^2\right] \times \begin{cases} > E\left[\left\|N. - F\left(\cdot, B_0^\top X\right)\right\|_{L^2}^2\right] & \text{for } \mathcal{S}(B_d) \not\supseteq \mathcal{S}_{Y|X}, \\ = E\left[\left\|N. - F\left(\cdot, B_0^\top X\right)\right\|_{L^2}^2\right] & \text{for } \mathcal{S}(B_d) \supseteq \mathcal{S}_{Y|X}, \end{cases} \tag{2.3}$$

the mean integrated squared-error risk fails to distinguish the true model from overfitted ones. By replacing B_{d_0} with B_d in (2.2), the resulting sample analogue of $E[\|N. - F_{C_d}(\cdot, B_d^\top X)\|_{L^2}^2]$ is infeasible for the determination of d_0. For this reason, a cross-validation version of $\mathrm{PSIS}(C_d)$ is developed to estimate B_0 and determine d_0. It is noted that the bandwidth vector h_d is treated as free parameters in our estimation criterion.

Remark 1. One alternative to the PLISE is the pseudo-maximum likelihood estimator $\widetilde{B}_{d_0} = (I_{d_0}, \widetilde{C}_{d_0}^\top)^\top$, a maximizer of the pseudo-log-likelihood function

$$p\ell(C_{d_0}) = \frac{1}{n}\sum_{i=1}^n \ln \widehat{f}_{C_{d_0}}\left(Y_i|B_{d_0}^\top X_i\right), \tag{2.4}$$

where

$$\widehat{f}_{C_{d_0}}(y|u_0) = \frac{\sum_{i=1}^n \frac{1}{h_y} K_q(\frac{Y_i - y}{h_y}) \mathcal{K}_{q,h_{d_0}}(B_{d_0}^\top X_i - u_0)}{\sum_{i=1}^n \mathcal{K}_{q,h_{d_0}}(B_{d_0}^\top X_i - u_0)} \tag{2.5}$$

is a kernel estimator for the conditional density $f_{C_{d_0}}(y|u_0)$ of Y on $B_{d_0}^\top X = u_0$ and h_y is a positive-valued bandwidth. Under some regularity conditions and the bandwidth constraint.

A0*. h_y and $h_{d_0 k}$, $k = 1, \ldots, d_0$, fall in the interval $H_n = (h_l n^{-1/(2d_0+4)}, h_u n^{-1/4q})$ for some positive constants h_l and h_u,

$\widetilde{B}_{d_0}$ can also attain the asymptotic semiparametric efficiency bound because the score function of $p\ell(C_{d_0})$ and that of Ma and Zhu (2013) are asymptotically equivalent. Compared with $\widehat{B}_{d_0}$, $\widetilde{B}_{d_0}$ is rather sensitive to outliers and cannot deal with a discrete response. For the bandwidth selection, one possible option is to use a maximizer of $p\ell_{\mathrm{cv}}(h) = \sum_{i=1}^n \ln \widehat{f}_{C_{d_0}}^{-i}(Y_i|\widetilde{B}_{d_0}^\top X_i)/n$. As shown by Hall (1987), the asymptotic properties of $\widehat{f}_{C_{d_0}}(y|u_0)$ are profoundly influenced by the tail properties of $K_q(\upsilon)$ and $f_{C_{d_0}}(y|u_0)$. A trimming sequence is often required to suppress the influence of negative and small positive values of $\widehat{f}_{C_{d_0}}(Y_i|B_{d_0}^\top X_i)$'s in (2.4). Currently, there is still no objective criterion for choosing such a tuning parameter.

2.2. Cross-Validation Estimation

For a new observation (X_0, Y_0), which is independent of a random sample $\{(X_i, Y_i)\}_{i=1}^n$, a simple calculation yields that

$$\begin{aligned} E\left[\left\|N_{0\cdot} - \widehat{F}_{C_d}\left(\cdot, B_d^\top X_0\right)\right\|_{L^2}^2\right] &= \sigma_0^2 + b_0^2(C_d) + E\Big[\Big\|\widehat{F}_{C_d}\left(\cdot, B_d^\top X_0\right) \\ &- F_{C_d}\left(\cdot, B_d^\top X_0\right)\Big\|_{L_2}^2\Big] + 2E\Big[\langle F_{C_d}\left(\cdot, B_d^\top X_0\right) \\ &- F\left(\cdot, B_0^\top X_0\right), \widehat{F}_{C_d}\left(\cdot, B_d^\top X_0\right) - F_{C_d}\left(\cdot, B_d^\top X_0\right)\rangle_{L^2}\Big], \end{aligned} \tag{2.6}$$

where $N_{0y} = \mathbf{1}(Y_0 \le y)$, $\sigma_0^2 = E[\|N_{0\cdot} - F(\cdot, B_0^\top X_0)\|_{L_2}^2]$, and $b_0^2(C_d) = E[\|F_{C_d}(\cdot, B_d^\top X_0) - F(\cdot, B_0^\top X_0)\|_{L_2}^2]$. By the fact that $b_0^2(C_d) = 0$ if and only if $\mathcal{S}(B_d) \supseteq \mathcal{S}_{Y|X}$, a distribution model $F_{C_d}(y, B_d^\top X)$ with $\mathcal{S}(B_d) \not\supseteq \mathcal{S}_{Y|X}$ can be identified to be incorrect whenever $b_0^2(C_d) > 0$. Let $f_{B_d^\top X}(u)$ and $F_X(x)$ stand for the respective density of $B_d^\top X$ and distribution of X, and $\mathrm{MISE}_{C_d}(h_d) = \iint (\mathcal{B}_{C_d}^2(y, B_d^\top x) + \mathcal{V}_{C_d}(y, B_d^\top x)) dF_X(x) dF_Y(y)$ with

$$\begin{aligned} \mathcal{B}_{C_d}(y,u) &= \frac{\int \upsilon^q K_q(\upsilon) d\upsilon}{q!} \sum_{k=1}^d h_{dk}^q \\ &\quad \times \sum_{\ell=0}^1 \frac{(-F_{C_d}(y,u))^{1-\ell} \partial_{u_k}^q F_{\ell, C_d}(y,u)}{F_{0,C_d}(y,u)}, \\ \mathcal{V}_{C_d}(y,u) &= \frac{(\int K_q^2(\upsilon) d\upsilon)^d F_{C_d}(y,u)(1 - F_{C_d}(y,u))}{n \prod_{k=1}^d h_{dk} f_{B_d^\top X}(u)}, \end{aligned}$$

and $F_{\ell, C_d}(y,u) = F_{C_d}^\ell(y,u) f_{B_d^\top X}(u)$,

$\ell = 0, 1$. The third term on the right-hand side of (2.6) can be shown to be asymptotically equivalent to $\mathrm{MISE}_{C_d}(h_d)$ (see Härdle and Marron (1985) and Härdle, Hall, and Marron 1988).

Proposition 2. Under (i) $\partial_u^q F_{C_d}(y,u)$ and $\partial_u^q f_{B_d^\top X}(u)$ are Lipschitz continuous in u with the Lipschitz constants being independent of (y, C_d) and (ii) $h_{dk} \to 0, k = 1, \ldots, d$, and $n \prod_{k=1}^d h_{dk} \to \infty$ as $n \to \infty$,

$$\begin{aligned} &E\left[\left\|\widehat{F}_{C_d}\left(\cdot, B_d^\top X_0\right) - F_{C_d}\left(\cdot, B_d^\top X_0\right)\right\|_{L_2}^2\right] \\ &\quad = \mathrm{MISE}_{C_d}(h_d)(1 + o(1)) = O\left(\sum_{k=1}^d h_{dk}^{2q} + \frac{1}{n \prod_{k=1}^d h_{dk}}\right). \end{aligned} \tag{2.7}$$

By (2.7), the optimal bandwidth of $\mathrm{MISE}_{C_0}(h_{d_0})$ can be derived to be $h_0 = O(n^{-1/(2q+d_0)})$, which satisfies the bandwidth constraint in assumption A0. In addition, the nonnegativity of $b_0^2(C_d)$ and Proposition 2 imply that $E[\|N_{0\cdot} - \widehat{F}_{C_d}(\cdot, B_d^\top X_0)\|_{L^2}^2] \ge \sigma_0^2 + O(n^{-2q/2q+d_0})$ and $(d_0, B_0, h_0) = \mathrm{argmin}_{\{(d,B_d,h_d)\}} E[\|N_{0\cdot} - \widehat{F}_{C_d}(\cdot, B_d^\top X_0)\|_{L^2}^2]$ for large enough n. According to these properties, we estimate (d_0, B_0, h_0) with a minimizer $(\widehat{d}, \widehat{C}, \widehat{h})$ of

$$\mathrm{CV}(d, C_d, h_d) = \frac{1}{n} \sum_{i=1}^n \int_{\mathcal{Y}} \left(N_{iy} - \widehat{F}_{C_d}^{-i}\left(y, B_d^\top X_i\right)\right)^2 d\widehat{F}_Y(y). \tag{2.8}$$

As in general semiparametric estimation, q and h_d usually play important roles in estimating B_0. For the minimization of $\mathrm{CV}(d, C_d, h_d)$, an asymptotically optimal bandwidth estimator $\widehat{h}$ of h_0 satisfies assumption A0 in probability and assures the $\sqrt{n}$-consistency of $\widehat{C}$. Without specifying an appropriate $K_q(\upsilon)$ in (2.8), $\widehat{C}$ cannot have this theoretical advantage.

Since the formulation in model (1.1) can be regarded as a set of nested semiparametric models indexed by d, the determination of d_0 is naturally transferred to the model selection problem. When $h_d = O(n^{-1/(2q+d)})$ and $q > \max\{d/2 + 1, 2\}$ for each given d, $E[\|\widehat{F}_{C_d}(\cdot, B_d^\top X_0) - F(\cdot, B_0^\top X_0)\|_{L^2}^2] = b_0^2(C_d) + E[\|\widehat{F}_{C_d}(\cdot, B_d^\top X_0) - F_{C_d}(\cdot, B_d^\top X_0)\|_{L^2}^2]$ can be shown to be an increasing function of d with the rate of convergence $O_p(n^{-2q/(2q+d)})$ for $\mathcal{S}(B_d) \supset \mathcal{S}_{Y|X}$. As a result, $\mathcal{S}_{Y|X}$ can be successfully estimated through (2.8) in a forward manner with respect to d. For classical parametric models, the cross-validation criterion, which is akin to the Akaike information criterion, is inconsistent in model selection because $E[\|F_{\widehat{C}}(\cdot, \widehat{B}^\top X_0) - F(\cdot, B_0^\top X_0)\|_{L^2}^2]$ is $O(n^{-1})$ for each SDR subspace $\mathcal{S}(B)$.

Remark 2. With a slight modification of the cross-validation criterion in (2.8), B_0 and d_0 can be estimated by separate procedures. More precisely, a minimizer $(\widehat{B}_d, \widehat{h})$ of $\mathrm{CV}_d(C_d, h_d) \triangleq \mathrm{CV}(d, C_d, h_d)$ is first sought for each given d and an estimator $\widehat{d}^*$ of d_0 is subsequently derived by minimizing the following cross-validation sum of integrated squares:

$$\mathrm{CV}^*(d) = \frac{1}{n} \sum_{i=1}^n \int_{\mathcal{Y}} \left(N_{iy} - \widetilde{F}_{\widehat{C}_d}^{-i}\left(y, \widehat{B}_d^\top X_i\right)\right)^2 d\widehat{F}_Y(y), \tag{2.9}$$

where $\widetilde{F}_{\widehat{C}_d}^{-i}(y, \widehat{B}_d^\top X_i)$'s are computed as $\widehat{F}_{\widehat{C}_d}^{-i}(y, \widehat{B}_d^\top X_i)$'s with $K_2(\upsilon)$ substituting for $K_4(\upsilon)$. Compared with $\mathrm{CV}(d, C_d, h_d)$, $\mathrm{CV}^*(d)$ can avoid the numerical instability caused by the use of a higher-order kernel function. However, the performance of $\widehat{d}^*$ is much poorer than that of $\widehat{d}$ for small samples. Since $\mathrm{CV}_d(C_d, h_d)$ might have multiple minimizers for $d > d_0$, the $\sqrt{n}$-consistency of $\widehat{B}_d$

is still in doubt. Some complications and difficulties are also anticipated for the consistency of $\widehat{d}^*$ to d_0.

Remark 3. For the conditional mean $E[Y|X]$, a mean dimension reduction subspace is defined as $\mathcal{S}(B_M)$ such that$Y \perp\!\!\!\perp E[Y|X]|B_M^\top X$, where B_M is a $p \times d$ full-rank coefficient matrix. It follows that

$$Y = g(B_M^\top X) + \varepsilon \tag{2.10}$$

with $g(u)$ being an unknown d-variate function and $E[\varepsilon|X = x] = 0$ for each x. Hereinafter, the minimum mean dimension reduction subspace is termed the CMS and is spanned by the column vectors of a $p \times d_0$ basis matrix B_{M0}. For a given $p \times d$ matrix B_{Md}, the conditional mean $g_{C_{Md}}(u) = E[Y|B_{Md}^\top X = u]$ is naturally estimated by a simple kernel estimator

$$\widehat{g}_{C_{Md}}(u) = \frac{\sum_{j=1}^n Y_j \mathcal{K}_{q,h_d}(B_{Md}^\top X_j - u)}{\sum_{j=1}^n \mathcal{K}_{q,h_d}(B_{Md}^\top X_j - u)}. \tag{2.11}$$

Let h_{M_0} be the minimizer of $\mathrm{MISE}_{C_{M0}}(h_{d_0})$ with

$$\mathrm{MISE}_{C_{Md}}(h_d) = \int \left(\mathcal{B}^2_{M,C_{Md}}\left(B_{Md}^\top x\right) + \mathcal{V}_{M,C_{Md}}\left(B_{Md}^\top x\right)\right) dF_X(x),$$

$$\mathcal{B}_{M,C_{Md}}(u) = \frac{\int v^q K_q(v)dv}{q!} \sum_{k=1}^d h_{dk}^q \sum_{\ell=0}^1 \left(\frac{-g_{C_{Md}}(y,u)}{g_{0,C_{Md}}(u)}\right)^{1-\ell} \times \left(\partial_{u_k}^q g_{\ell,C_{Md}}(u)\right)^\ell,$$

$$\mathcal{V}_{M,C_{Md}}(u) = \frac{(\int K_q^2(v)dv)^d Var(Y|B_{Md}^\top X = u)}{n\prod_{k=1}^d h_{dk} f_{B_{Md}^\top X}(u)},$$

and $g_{\ell,C_{Md}}(u) = g^\ell_{C_{Md}}(u) f_{B_{Md}^\top X}(u)$, $\ell = 0, 1$.

In the spirit of the cross-validation criterion in (2.8), we can also estimate (d_0, B_{M0}, h_{M_0}) with a minimizer $(\widehat{d}_M, \widehat{C}_M, \widehat{h}_M)$ of

$$\mathrm{CV}_M(d_M, C_{Md}, h_d) = \frac{1}{n}\sum_{i=1}^n \left(Y - \widehat{g}^{-i}_{C_{Md}}\left(B_{Md}^\top X_i\right)\right)^2, \tag{2.12}$$

provided that the second moment of ε exists.

2.3. Computational Algorithm

Let $\mathrm{vec}(\cdot)$ be the vectorization operation that stacks the columns of a matrix, $\bar{N}_y = \sum_{i=1}^n N_{iy}/n$, $\mathcal{S}_1(d, C_d, h_d) = \partial_{\mathrm{vec}(C_d)}\mathrm{CV}(d, C_d, h_d)$, $\mathcal{S}_2(d, C_d, h_d) = \partial_{h_d}\mathrm{CV}(d, C_d, h_d)$, $\mathcal{I}_1(d, C_d, h_d) = \partial_{\mathrm{vec}(C_d)}\mathcal{S}_1 \quad (d, C_d, h_d)$, and $\mathcal{I}_2(d, \quad C_d, h_d) = \partial_{h_d}\mathcal{S}_2(d, C_d, h_d)$. The minimization of $\mathrm{CV}(d, C_d, h_d)$ in (2.8) can be carried out by the following steps:

Step 1. Set $\mathrm{CV}(0) = \sum_{i=1}^n \int_{\mathcal{Y}} (N_{iy} - \bar{N}_y^{-i})^2 d\widehat{F}_Y(y)/n$ and $\mathrm{CV}(p+1) = +\infty$.

Step 2. Compute $(\widehat{C}_d, \widehat{h}_d) = \mathrm{argmin}_{\{(C_d,h_d)\}}\mathrm{CV}(d, C_d, h_d)$ and $\mathrm{CV}(d) = \mathrm{CV}(d, \widehat{C}_d, \widehat{h}_d)$, $d = 1, \ldots, p$, by the line search Newton-CG method (see Jorge and Stephen 2006) with the following modification for the nonlinear minimization:

Step 2.1. Start with searching for $C_d^{(0)}$ and $h_d^{(0)} \propto n^{-1/(2q+d)}$ such that $\mathcal{I}_1(d, C_d^{(0)}, h_d^{(0)})$ and $\mathcal{I}_2(d, C_d^{(0)}, h_d^{(0)})$ are nonsingular.

Step 2.2. Repeat

$$\begin{pmatrix}\mathrm{vec}(C_d^{(m+1)})\\ h_d^{(m+1)}\end{pmatrix} = \begin{pmatrix}\mathrm{vec}(C_d^{(m)})\\ h_d^{(m)}\end{pmatrix} - \begin{pmatrix}\alpha_{1m}\mathcal{I}_1^{-1}(d, C_d^{(m)}, h_d^{*(m)})\mathcal{S}_1(d, C_d^{(m)}, h_d^{(m)})\\ \alpha_{2m}(\mathcal{I}_2(d, C_d^{(m+1)}, h_d^{(m)}) + E_d^{(m)})^{-1}\mathcal{S}_2(d, C_d^{(m+1)}, h_d^{(m)})\end{pmatrix}$$

until the step lengths α_{1m} and α_{2m} satisfy the strong Wolfe conditions, where $h_d^{*(m)} = h_d^{(m)} n^{\varepsilon_{1m}}$ with $\varepsilon_{1m} \in 1/(2q+d) - (1/\max\{2d+2, d+4\}, 1/4q)$ being close to zero such that $\mathcal{I}_1(d, C_d^{(m)}, h_d^{*(m)})$ is nonsingular, and $E_d^{(m)} = 0$ if $\mathcal{I}_2(d, C_d^{(m)}, h_d^{(m)})$ is nonsingular and $E_d^{(m)} = n^{-\varepsilon_{2m}}\lambda_d^{(m)} I_d$ with $\lambda_d^{(m)}$ being the absolute value of the smallest nonzero eigenvalue of $\mathcal{I}_2(d, C_d^{(m)}, h_d^{(m)})$ and $\varepsilon_{2m} > 0$ if $\mathcal{I}_2(d, C_d^{(m)}, h_d^{(m)})$ is singular.

Step 3. Implement Step 2 until $\mathrm{CV}(d) \geq \mathrm{CV}(d-1)$ and estimate $F_Y(y|x)$ by $\widehat{F}_Y(y)$ if $d = 1$ and $\widehat{F}_{\widehat{C}_{d-1}}(y, \widehat{B}_{d-1}^\top x)$ otherwise.

By the specification of $(h_d^{*(m)}, E_d^{(m)})$ in Step 2.2, the pseudo-information matrices are ensured to be nonsingular. In practical implementation, Steps 2.1–2.2 can also be replaced by the conjugate gradient algorithm of Fletcher and Reeves (1964), which is appealing because no matrix operation is involved. However, the nonlinear optimization problem is still inevitable in the minimization of $\mathrm{CV}(d, C_d, h_d)$. For a large d_0 or an extremely large p or n, the performance of this scheme tends to be computationally inefficient and intensive. In application, a large structural dimension is commonly associated with the curse of dimensionality and an extremely large number of covariates or sample size is usually related to the high-dimensionality of big data. An efficient computational algorithm is, thus, necessary to overcome the existing difficulties.

3. Asymptotic Properties

The consistency of $(\widehat{d}, \widehat{C}, \widehat{h})$ to (d_0, C_0, h_0) is established based on the uniform consistency of $\mathrm{CV}(d, C_d, h_d)$ to

$$\mathrm{ECV}(d, C_d, h_d) = \begin{cases}\sigma_0^2 + \mathrm{MISE}_{C_d}(h_d) & \text{for } \mathcal{S}(B_d) \supseteq \mathcal{S}_{Y|X},\\ \sigma_0^2 + b_0^2(C_d) + \mathrm{MISE}_{C_d}(h_d) & \text{for } \mathcal{S}(B_d) \not\supseteq \mathcal{S}_{Y|X}.\end{cases}$$

Due to a random dimension of $\mathcal{S}(\widehat{B})$, the related large sample properties are assessed via the projection matrix $P_{\widehat{C}}$, where $P_{C_d} = B_d(B_d^\top B_d)^{-1}B_d^\top$ is the orthogonal projection operator onto $\mathcal{S}(B_d)$. Both $P_{\widehat{C}}$ and $P_{\widehat{C}_{d_0}}$ are further shown to have the same asymptotic distribution.

3.1. Notations and Assumptions

Let $(\cdot)^{\otimes}$ stand for the kronecker power of a vector, $\|\cdot\|$ be the Frobenius norm of a matrix, $\check{X} = (\widetilde{X}^\top, \ldots, \widetilde{X}^\top)^\top \in \mathbb{R}^{(p-d)d}$ with $\widetilde{X} = (X_{d+1}, \ldots, X_p)^\top$. In Appendix A, the partial derivatives $\partial_{\mathrm{vec}(C_d)}\widehat{F}_{C_d}(y, B_d^\top x)$ and $\partial^2_{\mathrm{vec}(C_d)}\widehat{F}_{C_d}(y, B_d^\top x)$ are shown to converge uniformly to $F^{[1]}_{C_d}(y, x) = \sum_{\ell=0}^1 (-F_{C_d}(y, B_d^\top x))^\ell F^{[1]}_{1-\ell,C_d}(y, x)/F_{0,C_d}(y, B_d^\top x)$

and

$$F^{[2]}_{C_d}(y,x)=\sum_{\ell_1,\ell_2=0}^{1}2^{\ell_1}\left(\frac{-F^{[2-\ell_1]}_{0,C_d}(y,x)}{F_{0,C_d}(y,B_d^\top x)}\right)^{\ell_1+\ell_2}\times\frac{F^{[(2-\ell_1)(1-\ell_2)]}_{1,C_d}(y,x)}{F_{0,C_d}(y,B_d^\top x)}\ \text{a.s.},$$

where $F^{[m]}_{\ell,C_d}(y,x)=\partial^m_{\mathrm{vec}(C_d)}(F^{\ell}_{C_d}(y,B_d^\top x)E[(\check{X}-\check{x})^{\otimes m}|B_d^\top X=B_d^\top x]f_{B_d^\top X}(B_d^\top x))$, $\ell=0,1$, $m=0,1,2$. According to these properties, the corresponding pseudo-score vector and information matrix of $\mathrm{CV}(d,C_d,h_d)$ can be derived to be asymptotically equivalent to

$$S_{C_d}=\int_{\mathcal{Y}}\left(N_y-F_{C_d}\left(y,B_d^\top X\right)\right)F^{[1]}_{C_d}(y,X)dF_Y(y)$$

and

$$V_{C_d}=E\left[\int_{\mathcal{Y}}\left(\left(F^{[1]}_{C_d}(y,X)\right)^{\otimes 2}-\left(N_y-F_{C_d}(y,B_d^\top X)\right)F^{[2]}_{C_d}(y,X)\right)dF_Y(y)\right].$$

It is straightforward to have $E[S_{C_0}]=0$ and $V_{C_0}=E[\int_{\mathcal{Y}}(F^{[1]}_{C_0}(y,X))^{\otimes 2}dF_Y(y)]$.

In the theoretical development, the following regularity conditions are further imposed:

A1. $\partial^{q+2}_u F_{C_d}(y,u)$, $\partial^{q+k}_u E[(\check{X}-\check{x})^{\otimes k}|B_d^\top X=u]$, and $\partial^{q+2}_u f_{B_d^\top X}(u)$, $\ell=0,1$, $k=1,2$, are Lipschitz continuous in (u,C_d) with the Lipschitz constants being independent of (y,x).

A2. $\inf_{\{(u,B_d)\}} f_{B_d^\top X}(u)>0$.

A3. $\inf_{\{d<d_0,B_d\}} b_0^2(C_d)>0$.

A4. V_{C_d} is nonsingular.

For the convergence of kernel estimators, the smoothness conditions are drawn in assumption A1. In addition, assumption A2 is necessary for the uniform consistency and assumptions A3 and A4 assure the existence of a unique minimizer of $\mathrm{ECV}(d,C_d,h_d)$ and the asymptotic normality of $P_{\widehat{C}}$.

3.2. Consistency and Asymptotic Normality

Since the fourth term on the right-hand side of (2.6) is equal to zero for $\mathcal{S}(B_d)\supseteq\mathcal{S}_{Y|X}$ and $O(b_0(C_d)\sqrt{\mathrm{MISE}_{C_d}(h_d)})$ for $\mathcal{S}(B_d)\not\supseteq\mathcal{S}_{Y|X}$, the different convergence rates of $|\mathrm{CV}(d,C_d,h_d)-\mathrm{ECV}(d,C_d,h_d)|$ are established.

Theorem 1. Suppose that assumptions A0–A2 are satisfied. Then,

$$\sup_{\{(d,C_d,h_d)\}}\frac{|\mathrm{CV}(d,C_d,h_d)-\mathrm{ECV}(d,C_d,h_d)|}{\mathrm{MISE}_{C_d}(h_d)}=o(1)\ \ a.s.\ \text{for}\ \mathcal{S}(B_d)\supseteq\mathcal{S}_{Y|X}\tag{3.1}$$

and

$$\sup_{\{(d,C_d,h_d)\}}\frac{|\mathrm{CV}(d,C_d,h_d)-\mathrm{ECV}(d,C_d,h_d)|}{b_0(C_d)\sqrt{\mathrm{MISE}_{C_d}(h_d)}}=O(1)\ \ a.s.\ \text{for}\ \mathcal{S}(B_d)\not\supseteq\mathcal{S}_{Y|X}.\tag{3.2}$$

Proof. See Appendix B. □

The asymptotic distributions of $\widehat{C}_{d_0}$ and $P_{\widehat{C}_{d_0}}$ are further given in the following theorem:

Theorem 2. Suppose that assumptions A0–A4 are satisfied. Then,

$$\sqrt{n}\mathrm{vec}(\widehat{C}_{d_0}-C_0)\xrightarrow{d}N(0,E[(\mathrm{vec}(A_0))^{\otimes 2}])\ \text{as}\ n\longrightarrow\infty\tag{3.3}$$

and

$$\sqrt{n}\mathrm{vec}(P_{\widehat{C}_{d_0}}-P_{C_0})\xrightarrow{d}N(0,\Sigma_0)\ \text{as}\ n\longrightarrow\infty,\tag{3.4}$$

where $\mathrm{vec}(A_0)=V_{C_0}^{-1}S_{C_0}$ and $\Sigma_0=E[(\mathrm{vec}((I_p-P_{C_0})(I_{d_0},A_0^\top)^\top(B_0^\top B_0)^{-1}B_0^\top+B_0(B_0^\top B_0)^{-1}(I_{d_0},A_0^\top)(I_p-P_{C_0})))^{\otimes 2}]$.

Proof. See Appendix B. □

For any qth-order kernel function $K_q(\upsilon)$ with $q>\max\{d_0/2+1,2\}$, the asymptotic distribution of $\widehat{C}_{d_0}$ is independent of q, and so is the asymptotic distribution of $P_{\widehat{C}_{d_0}}$. To maintain a better numerical stability, the smallest even order is preferred in practical implementation.

In addition, we establish the consistency of $(\widehat{d},\widehat{C},\widehat{h})$ to (d_0,C_0,h_0) with the aid of Theorems 1 and 2.

Theorem 3. Suppose that assumptions A0–A3 are satisfied. Then, for any $\varepsilon_1,\varepsilon_2>0$,

$$P\left(\widehat{d}=d_0,\|\widehat{C}-C_0\|<\varepsilon_1,\max_{\{1\le k\le d_0\}}\left|\frac{\widehat{h}_k}{h_{0k}}-1\right|<\varepsilon_2\right)\longrightarrow 1\ \text{as}\ n\longrightarrow\infty.\tag{3.5}$$

Proof. See Appendix B. □

Let $E_{0\varepsilon_2}=\{\widehat{d}=d_0,\max_{\{1\le k\le d_0\}}|\widehat{h}_k/h_{0k}-1|<\varepsilon_2\}$. By $\mathbf{1}(E_{0\varepsilon_2})+\mathbf{1}(E^c_{0\varepsilon_2})=1$, $P(\|\sqrt{n}\mathrm{vec}(P_{\widehat{C}}-P_{C_0})\mathbf{1}(E^c_{0\varepsilon_{12}})\|>\varepsilon)\le P(E^c_{0\varepsilon_2})\ \forall\varepsilon>0$, and Theorem 3, one has

$$\sqrt{n}\mathrm{vec}(P_{\widehat{C}}-P_{C_0})=\sqrt{n}\mathrm{vec}(P_{\widehat{C}_{d_0}}-P_{C_0})\mathbf{1}(E_{0\varepsilon_2})+o_p(1).\tag{3.6}$$

As a result, both $P_{\widehat{C}}-P_{C_0}$ and $P_{\widehat{C}_{d_0}}-P_{C_0}$ have the same asymptotic distribution.

Theorem 4. Suppose that assumptions A0–A4 are satisfied. Then,

$$\sqrt{n}\mathrm{vec}(P_{\widehat{C}}-P_{C_0})\xrightarrow{d}N(0,\Sigma_0)\ \text{as}\ n\longrightarrow\infty.\tag{3.7}$$

It can be observed from Theorems 2 and 4 that the asymptotic normality of $P_{\widehat{C}}$ is not affected by the variations of $\widehat{d}$ and $\widehat{h}$. By using the Taylor expansion, we further have

$$\widehat{F}_{\widehat{C}}\left(y,\widehat{B}^\top x\right)-F\left(y,B_0^\top x\right)=\left(\partial_{\mathrm{vec}(C_{d_0})}\widehat{F}_{C^*_{d_0}}\left(y,B^{*\top}_{d_0}x\right)\right)\mathrm{vec}(\widehat{C}_{d_0}-C_0)\times\mathbf{1}(E_{0\varepsilon_2})+\left(\widehat{F}_{C_0}\left(y,B_0^\top x\right)-F\left(y,B_0^\top x\right)\right)\mathbf{1}(E_{0\varepsilon_2})+\left(\widehat{F}_{\widehat{C}}\left(y,\widehat{B}^\top x\right)-\widehat{F}_{C_0}\left(y,B_0^\top x\right)\right)\mathbf{1}\left(E^c_{0\varepsilon_2}\right),\tag{3.8}$$

where $\mathrm{vec}(C^*_{d_0})$ lies on the line segment between $\mathrm{vec}(\widehat{C}_{d_0})$ and $\mathrm{vec}(C_0)$. Moreover, the convergence rates of the first and third terms on the right-hand side of (3.8) can be shown to be $O_p(n^{-1/2})$ and $o_p(n^{-1/2})$ by Lemma 1 and Theorems 2 and 3. The following theorem is, thus, a direct consequence of Lemma 2 and the above properties.

Theorem 5. Suppose that assumptions A0–A4 are satisfied. Then,

$$\sup_{\{(y,x)\}} |\widehat{F}_{\widehat{C}}(y, \widehat{B}^\top x) - F\left(y, B_0^\top x\right) - \frac{1}{n}\sum_{i=1}^{n} \xi_{i,C_0}^{[1]}(y,x)| = O_p\left(\frac{1}{\sqrt{n}}\right), \tag{3.9}$$

where $\xi_{i,C_d}^{[1]}(y,x)$'s are defined in Appendix A.

Remark 4. Let $S_{M,C_{M0}} = (Y - g(B_{M0}^\top X))g_{C_{M0}}^{[1]}(X)$ and $V_{M,C_{M0}} = E[(g_{C_{M0}}^{[1]}(X))^{\otimes 2}]$ with

$$\begin{aligned} g_{\ell,C_{Md}}^{[m]}(x) &= \partial_{\mathrm{vec}(C_{Md})}^{m}\Big(E\Big[(\check{X} - \check{x})^{\otimes m} | B_{Md}^\top X \\ &= B_{Md}^\top x\Big] g_{C_{Md}}^{\ell}\left(B_{Md}^\top x\right) f_{B_{Md}^\top X}(B_{Md}^\top x)\Big), \end{aligned}$$

and

$$\begin{aligned} g_{C_{Md}}^{[1]}(x) &= \frac{1}{g_{C_{Md}}(B_{Md}^\top x)} \sum_{\ell=0}^{1} \Big(- g_{C_{Md}}\left(B_{Md}^\top x\right)\Big)^{\ell} g_{1-\ell,C_{Md}}^{[1]}(x), \\ &\ell = 0, 1, m = 0, 1, 2. \end{aligned}$$

Following the proofs of Theorems 1–4, we can also derive the consistency of $(\widehat{d}_M, \widehat{C}_M, \widehat{h}_M)$ to $(d_0, C_{M0}, h_{M0}) = \mathrm{argmin}_{\{(d,C_{Md},h_d)\}} \{E[(g_{C_{Md}}(B_{Md}^\top X_0) - g(B_{M_0}^\top X_0))^2] + \mathrm{MISE}_{C_{Md}}(h_d)\}$ and the asymptotic normality of $P_{\widehat{C}_M}$.

Theorem 6. Under assumptions A0, A2, (B1) $\partial_u^{q+2} g_{C_{Md}}(u)$, $\partial_u^{q+k} E[(\check{X} - \check{x})^{\otimes k} | B_{Md}^\top X = u]$, and $\partial_u^{q+2} f_{B_{Md}^\top X}(u)$, $\ell = 0, 1$, $k = 1, 2$, are Lipschitz continuous in (u, C_{Md}) with the Lipschitz constants being independent of x, (B2) $\inf_{\{d<d_0\}} E[(g_{C_{Md}}(B_{Md}^\top X) - g(B_{M0}^\top X))^2] > 0$, and (B3) $V_{M,C_{M0}}$ is nonsingular,

$$\begin{aligned} &P\left(\widehat{d}_M = d_{M0}, \|\widehat{C}_M - C_{M0}\| < \varepsilon_1, \max_{\{1\le k\le d_0\}} \left|\frac{\widehat{h}_{Mk}}{h_{M0k}} - 1\right| < \varepsilon_2\right) \\ &\quad \longrightarrow 1 \quad \forall \varepsilon_1, \varepsilon_2 > 0 \end{aligned} \tag{3.10}$$

and

$$\sqrt{n}\mathrm{vec}(P_{\widehat{C}_M} - P_{C_{M0}}) \xrightarrow{d} N(0, \Sigma_{M0}) \text{ as } n \longrightarrow \infty, \tag{3.11}$$

where $\mathrm{vec}(A_{M0}) = V_{M,C_{M0}}^{-1} S_{M,C_{M0}}$ and $\Sigma_{M0} = E[(\mathrm{vec}((I_p - P_{C_{M0}})(I_{d_0}, A_{M0}^\top)^\top (B_{M0}^\top B_{M0})^{-1} B_{M0}^\top + B_{M0}(B_{M0}^\top B_{M0})^{-1}(I_{d_0}, A_{M0}^\top)(I_p - P_{C_{M0}})))^{\otimes 2}]$.

4. Monte Carlo Simulations

The numerical experiments were performed on a Unix workstation with AMD Opteron 6134 and 64 GB RAM. To assure numerical stability, the simulations were based on 1000 replications. Since all of structural dimension estimates are smaller than 6, only a fourth-order kernel function, for example, $K_4(u) = (105/64)(1 - 3u^2)(1 - u^2)^2 \mathbf{1}(|u| \le 1)$, is required in our estimation. For the assessments of generic basis estimator $B_{ge} = (I_{d_{ge}}, C_{ge}^\top)^\top$ and distribution estimator $\widehat{F}_{C_{ge}}(y, B_{ge}^\top x)$, we adopted the estimation accuracies $\triangle(B_{ge}, B_0) = ||P_{C_{ge}} - P_{C_0}||_2$ and $\widehat{\mathrm{MISE}}(\widehat{F}_{C_{ge}}) = \sum_{i=1}^{n} \int_{\mathcal{Y}} (\widehat{F}_{C_{ge}}(y, B_{ge}^\top X_i) - F(y, B_0^\top X_i))^2 d\widehat{F}_Y(y)/n$, where $||\cdot||_2$ stands for the spectral norm of a matrix. The adverbs "slightly," "quite," and "substantially" are further used to describe the magnitude differences within [0.05, 0.1], (0.1, 0.15], and (0.15, 1], respectively.

4.1. Finite-Sample Performance of $(\widehat{d}, \widehat{C}, \widehat{h})$

The finite-sample performance of $(\widehat{d}, \widehat{C}, \widehat{h})$ was first assessed through a class of simulations with a mixture of discrete and continuous covariates. In this simulation scenario, the first eight covariates $(X_1, \ldots, X_8)^\top$ of $X = (X_1, \ldots, X_{10})^\top$ were designed to follow a multivariate normal distribution with mean of zero, standard deviation of one, and pairwise correlations of 0.2 or 0.5. Conditioning on $(X_1, \ldots, X_8)^\top = (x_1, \ldots, x_8)^\top$, X_9 and X_{10} were independently generated from Bernoulli and Poisson distributions with the respective parameters $|x_1|/(|x_1| + |x_2|)$ and $\sum_{k=1}^{8} |x_k|$. The relationship between Y and X was further set as follows:

$$\text{M1. } Y = 2\beta_{01}^\top X + (\beta_{02}^\top X)^2 + \varepsilon \text{ with } \varepsilon \sim N(0, 0.25),$$

where $B_0 = (\beta_{01}, \beta_{02})$ with $\beta_{01} = (1, 0, 1, 0, 0, \ldots, 0)^\top$ and $\beta_{02} = (0, 1, 0, 1, 0, \ldots, 0)^\top$, and $\beta_{01}^\top X$ and $\beta_{02}^\top X$ satisfy the continuity condition of Ichimura (1993). In such a setup, the linearity and/or constant variance conditions are violated and fully nonparametric regression estimators are infeasible. Due to these facts, comparisons were made for the PILSE $\widehat{B}_{d_0}$, which is computed based on (2.8) with $d = d_0$, the pseudo maximum likelihood estimator $\widetilde{B}_{d_0}$, and the semiparametric efficient estimator $\check{B}_{d_0}$ of Ma and Zhu (2013) with a given d_0. In addition, the proposed estimator $\widehat{B}$ was compared with the semiparametric efficient estimator $\check{B}$ with an estimated structural dimension. Although the semiparametric counterparts of inverse regression estimators are applicable for the designed covariates, only the semiparametric efficient estimator was used as a benchmark in the simulation investigation.

Table 1 displays the means and standard deviations of 1000 CS estimates with a given d_0 for the sample sizes (n) of 100, 200, and 400, and the correlations (ρ) of 0.2 and 0.5. Due to the poor performance of separate-bandwidths in the cross-validation version of (2.4), $\widetilde{B}_{d_0}$ was computed by using bandwidths of the form $(h_y, h_{d_0}^\top)^\top \propto (\widehat{\sigma}(Y), \widehat{\sigma}(\widetilde{\beta}_1^\top X), \ldots, \widehat{\sigma}(\widetilde{\beta}_{d_0}^\top X))^\top$, where $\widehat{\sigma}$ represents the sample standard deviation. As for the computation of $\check{B}_{d_0}$ and $\check{B}$, we adopted the theoretically valid bandwidths of Ma and Zhu (2013). The biases of compared estimators are generally small except for $\widetilde{B}_{d_0}$ in the settings with $n = 100$. Further, the variations of these competitors become small as n gets large and are less sensitive toward the designed values of ρ. Even for relatively large sample sizes (e.g., $n = 400$), the variation of $\widehat{B}_{d_0}$ is substantially smaller than those of $\widetilde{B}_{d_0}$ and $\check{B}_{d_0}$. Basically, the poor finite-sample performance of $\check{B}_{d_0}$ is mainly caused by the use of subjective bandwidths, which are only valid in asymptotic sense. An extremely large sample size is usually required to assure that $\widetilde{B}_{d_0}$ and $\check{B}_{d_0}$ can attain this asymptotic efficiency bound. Moreover, the above findings are confirmed by the accuracy measures $\triangle(\widehat{B}_{d_0}, B_0)$, $\triangle(\widetilde{B}_{d_0}, B_0)$, and $\triangle(\check{B}_{d_0}, B_0)$ in Table 4. The empirical insights from this investigation also help practitioners understand why our SDR approach is developed based on a cross-validation version of

Table 1. The means (standard deviations) of 1000 PILSEs ($\widehat{\beta}_{d_0}$), pseudo-maximum likelihood estimates ($\widetilde{\beta}_{d_0}$), and semiparametric efficient estimates ($\check{\beta}_{d_0}$) under model M1.

ρ	n	$\widehat{\gamma}_{d_0 1}$	$\widehat{\gamma}_{d_0 2}$	$\widetilde{\gamma}_{d_0 1}$	$\widetilde{\gamma}_{d_0 2}$	$\check{\gamma}_{d_0 1}$	$\check{\gamma}_{d_0 2}$
0.2	100	1.00 (0.048)	0.00 (0.051)	1.10 (0.099)	− 0.02 (0.100)	1.01 (0.175)	0.01 (0.200)
		0.00 (0.047)	1.00 (0.052)	− 0.02 (0.091)	1.04 (0.113)	0.01 (0.174)	1.02 (0.185)
		1.00 (0.048)	0.00 (0.054)	1.10 (0.110)	− 0.02 (0.112)	1.02 (0.184)	0.01 (0.189)
		0.00 (0.049)	0.99 (0.054)	− 0.01 (0.091)	1.04 (0.118)	0.01 (0.185)	1.01 (0.184)
		0.00 (0.040)	0.00 (0.043)	0.01 (0.078)	0.01 (0.091)	0.00 (0.230)	0.01 (0.248)
		0.00 (0.039)	0.00 (0.046)	0.01 (0.085)	0.00 (0.093)	0.01 (0.222)	0.01 (0.230)
		0.00 (0.057)	0.00 (0.065)	− 0.05 (0.100)	0.02 (0.119)	− 0.01 (0.130)	0.00 (0.138)
		0.00 (0.019)	0.00 (0.023)	− 0.01 (0.036)	0.00 (0.046)	− 0.01 (0.164)	0.00 (0.160)
	200	1.00 (0.034)	0.00 (0.037)	1.08 (0.068)	− 0.01 (0.060)	1.03 (0.153)	0.02 (0.170)
		0.00 (0.036)	1.00 (0.035)	− 0.01 (0.054)	1.05 (0.065)	0.00 (0.149)	1.02 (0.167)
		1.00 (0.035)	0.00 (0.036)	1.08 (0.069)	− 0.01 (0.059)	1.03 (0.148)	0.00 (0.180)
		0.00 (0.035)	1.00 (0.037)	− 0.01 (0.056)	1.04 (0.064)	0.00 (0.145)	1.02 (0.164)
		0.00 (0.028)	0.00 (0.028)	0.01 (0.045)	0.01 (0.048)	0.01 (0.178)	0.01 (0.209)
		0.00 (0.027)	0.00 (0.029)	0.01 (0.045)	0.00 (0.049)	0.02 (0.174)	0.01 (0.216)
		0.00 (0.043)	0.00 (0.044)	− 0.03 (0.066)	0.01 (0.069)	0.00 (0.106)	0.00 (0.124)
		0.00 (0.013)	0.00 (0.013)	0.00 (0.020)	0.00 (0.019)	0.00 (0.153)	− 0.01 (0.157)
	400	1.00 (0.024)	0.00 (0.020)	1.00 (0.033)	− 0.02 (0.031)	1.03 (0.088)	0.01 (0.115)
		0.00 (0.025)	1.00 (0.020)	− 0.02 (0.032)	1.00 (0.030)	0.00 (0.085)	1.02 (0.098)
		1.00 (0.024)	0.00 (0.021)	1.00 (0.034)	− 0.02 (0.029)	1.03 (0.093)	0.01 (0.116)
		0.00 (0.024)	1.00 (0.020)	− 0.02 (0.033)	1.00 (0.031)	0.00 (0.082)	1.02 (0.095)
		0.00 (0.019)	0.00 (0.016)	0.00 (0.022)	0.00 (0.019)	0.00 (0.089)	0.01 (0.115)
		0.00 (0.019)	0.00 (0.015)	0.00 (0.022)	0.00 (0.020)	0.01 (0.094)	0.01 (0.114)
		0.00 (0.029)	0.00 (0.025)	0.00 (0.042)	0.01 (0.039)	0.00 (0.070)	0.00 (0.076)
		0.00 (0.010)	0.00 (0.007)	0.00 (0.008)	0.00 (0.007)	0.00 (0.097)	0.00 (0.091)
0.5	100	1.00 (0.062)	− 0.01 (0.065)	1.14 (0.164)	− 0.04 (0.151)	1.02 (0.166)	0.00 (0.176)
		− 0.01 (0.065)	1.00 (0.065)	− 0.03 (0.135)	1.09 (0.181)	0.00 (0.155)	1.01 (0.178)
		1.00 (0.061)	− 0.01 (0.069)	1.14 (0.146)	− 0.05 (0.156)	1.03 (0.172)	0.02 (0.173)
		− 0.01 (0.062)	1.00 (0.066)	− 0.03 (0.125)	1.08 (0.183)	0.01 (0.152)	1.01 (0.175)
		0.00 (0.052)	0.00 (0.057)	0.02 (0.106)	0.01 (0.121)	0.02 (0.194)	0.02 (0.222)
		0.00 (0.054)	0.00 (0.055)	0.02 (0.108)	0.01 (0.126)	0.02 (0.185)	0.01 (0.212)
		0.00 (0.064)	0.00 (0.069)	− 0.07 (0.146)	0.04 (0.176)	0.00 (0.136)	0.01 (0.149)
		0.00 (0.026)	0.00 (0.026)	0.00 (0.058)	0.00 (0.063)	− 0.02 (0.165)	− 0.01 (0.163)
	200	1.00 (0.045)	0.00 (0.044)	1.09 (0.095)	− 0.03 (0.080)	1.04 (0.139)	0.00 (0.165)
		− 0.01 (0.046)	1.00 (0.044)	− 0.02 (0.081)	1.06 (0.094)	0.00 (0.139)	1.02 (0.161)
		1.00 (0.046)	0.00 (0.044)	1.10 (0.092)	− 0.03 (0.075)	1.04 (0.140)	0.00 (0.163)
		− 0.01 (0.047)	1.00 (0.047)	− 0.02 (0.078)	1.06 (0.091)	0.01 (0.138)	1.03 (0.160)
		0.00 (0.035)	0.00 (0.034)	0.01 (0.060)	0.01 (0.059)	0.02 (0.152)	0.01 (0.192)
		0.00 (0.036)	0.00 (0.034)	0.01 (0.060)	0.01 (0.060)	0.03 (0.157)	0.02 (0.193)
		0.00 (0.048)	0.00 (0.046)	− 0.04 (0.080)	0.03 (0.089)	0.00 (0.121)	0.00 (0.134)
		0.00 (0.015)	0.00 (0.014)	0.00 (0.027)	0.00 (0.023)	− 0.01 (0.152)	0.00 (0.153)
	400	1.00 (0.029)	0.00 (0.021)	1.07 (0.069)	− 0.01 (0.036)	1.03 (0.083)	0.01 (0.082)
		− 0.01 (0.027)	1.00 (0.024)	− 0.01 (0.055)	1.09 (0.052)	0.01 (0.080)	1.02 (0.076)
		1.00 (0.030)	0.00 (0.022)	1.07 (0.071)	− 0.01 (0.034)	1.04 (0.086)	0.01 (0.079)
		0.00 (0.029)	1.00 (0.022)	− 0.01 (0.056)	1.09 (0.051)	0.01 (0.077)	1.02 (0.076)
		0.00 (0.024)	0.00 (0.018)	0.01 (0.043)	0.01 (0.028)	0.01 (0.075)	0.01 (0.093)
		0.00 (0.024)	0.00 (0.017)	0.01 (0.042)	0.01 (0.026)	0.01 (0.076)	0.01 (0.090)
		0.00 (0.030)	0.00 (0.024)	− 0.03 (0.058)	0.02 (0.039)	− 0.01 (0.076)	0.00 (0.071)
		0.00 (0.012)	0.00 (0.008)	0.01 (0.018)	0.00 (0.009)	0.01 (0.091)	0.00 (0.077)

the PSIS rather than the pseudo-likelihood function and the semiparametric efficient estimating equations.

In Table 2, we present the proportions of 1000 structural dimension estimates. It is revealed in this table that the proportions of $\widehat{d} = d_0$ quickly approach to one for adequate sample sizes (e.g., $n = 200$) and, thus, $\triangle(\widehat{B}, B_0)$ and $\triangle(\widehat{B}_{d_0}, B_0)$ in Table 4 are rather close to each other. Our unreported numerical results further indicate that the proportions of $\widehat{d}^* = d_0$, where $\widehat{d}^*$ is a minimizer of $\mathrm{CV}^*(d)$ in (2.9), are lower than those of $\widehat{d} = d_0$ in the most cases. In addition, to check the adequacy of a fitted model $\widehat{F}_{\widehat{C}_d}(y, \widehat{B}_d^\top x)$ in Step 2 of the proposed computational algorithm, the single-indexing cross-validation procedure (see Chiang and Huang 2012) can be applied to determine the structural dimension estimator $\widehat{d}_{\mathrm{SIC}}$. The proportions of $\widehat{d}_{\mathrm{SIC}} = d_0$ are shown to be substantially lower than those of $\widehat{d} = d_0$ for the sample sizes of 100 and 200. In conjunction with the semiparametric efficient estimation, the bootstrap procedure of Dong and Li (2010) with 500 replicates was also implemented to determine the structural dimension estimator $\check{d}$. The numerical results are only provided for the settings with $n = 100$ because of a heavy computational load (median computation times of 7081 and 7522 seconds for $\rho = 0.2$ and $\rho = 0.5$). It is shown that the proportions of $\check{d} = d_0$ are substantially lower than those of $\widehat{d} = d_0$. As expected, the values of $\triangle(\check{B}, B_0)$ are substantially larger than those of $\triangle(\widehat{B}, B_0)$. For the performance of $\widehat{h}$, the values of $(\widehat{h}_k/h_{0k})$'s in Table 3 are quite close to one for adequate sample sizes. We observe from Table 4 that the difference between $\triangle(\widehat{B}, B_0)$ and $\triangle(\widehat{B}_{d_0}, B_0)$ strongly relies on the correctness of $\widehat{d}$, whereas the difference between $\widehat{\mathrm{MISE}}(\widehat{F}_{\widehat{C}})$ and $\widehat{\mathrm{MISE}}(\widehat{F}_{\widehat{C}_{d_0}})$ is not so sensitive to this factor. The satisfactory performance of $\widehat{F}_{\widehat{C}}(y, \widehat{B}^\top x)$ can be partially explained by the intrinsic feature of the proposed estimation criterion.

Table 2. The proportions of 1000 structural dimension estimates under model M1.

ρ	n	0	1	2	3	4	5+
		Proportions of $\widehat{d}\,(d_0 = 2)$					
0.2	100	0.000	0.018	0.981	0.001	0.000	0.000
	200	0.000	0.000	1.000	0.000	0.000	0.000
	400	0.000	0.000	1.000	0.000	0.000	0.000
0.5	100	0.000	0.011	0.987	0.002	0.000	0.000
	200	0.000	0.000	1.000	0.000	0.000	0.000
	400	0.000	0.000	1.000	0.000	0.000	0.000
		Proportions of $\widehat{d}_{\mathrm{SIC}}\,(d_0 = 2)$					
0.2	100	0.000	0.544	0.337	0.077	0.023	0.019
	200	0.000	0.214	0.700	0.077	0.009	0.000
	400	0.000	0.013	0.987	0.000	0.000	0.000
0.5	100	0.000	0.404	0.388	0.126	0.063	0.019
	200	0.000	0.227	0.685	0.077	0.008	0.003
	400	0.000	0.036	0.964	0.000	0.000	0.000
		Proportions of $\check{d}\,(d_0 = 2)$					
0.2	100	0.000	0.000	0.576	0.424	0.000	0.000
0.5	100	0.000	0.000	0.514	0.482	0.004	0.000

4.2. A Comparison with the Ensemble Approach

Three examples from the simulation studies of Li (1992) and/or Wang and Xia (2008) were borrowed to compare the most competitive refined MAVE ensemble $\bar{B}$, which is based on the family of characteristic functions. The numerical study of Yin and Li (2011) showed that the ensemble approach outperforms some representative inverse regression and semiparametric approaches. Due to this fact, unnecessary comparisons are avoided herein. The second scenario aims to assess the performance of our approach and this ensemble approach via the following regression models:

M2. $Y = \left(B_0^\top X\right)^{-1} + \varepsilon$ with $\varepsilon \sim N(0, 0.04)$,

M3. $Y = \cos(2X_1) - \cos X_2 + \varepsilon$ with $\varepsilon \sim N(0, 0.04)$,

M4. $Y = X_1/(0.5 + (X_2 + 1.5)^2) + X_3^2\varepsilon$ with $\varepsilon \sim N(0, 1)$,

where the corresponding bases in models M2, M3, and M4 are $(1, 1, 1, 1, 0, \ldots, 0)^\top \in \mathbb{R}^{10}$, $((1, 0, 0, \ldots, 0)^\top, (0, 1, 0, \ldots, 0)^\top) \in \mathbb{R}^{10\times 2}$, and $((1, 0, 0, 0, \ldots, 0)^\top, (0, 1, 0, 0, \ldots, 0)^\top, (0, 0, 1, 0, \ldots, 0)^\top) \in \mathbb{R}^{10\times 3}$. In this investigation, three types of distributions (normal, uniform, and mixture normal) were considered for the distribution of X. It is noted that $\widehat{B}$ is $\sqrt{n}$-consistent for any finite d_0, whereas $\bar{B}$ cannot be $\sqrt{n}$-consistent for $d_0 \geq 4$.

Compared with the structural dimension estimator $\bar{d}$ of Yin and Li (2011), we can observe from Tables 5 and 7 that the proportions of $\widehat{d} = d_0$ are slightly higher for (model, distribution of X, sample size) of (M2, mixture normal, 100), (M4, normal, 100), (M4, mixture normal, 100), and (M4, mixture normal, 200), quite higher for (M3, uniform, 100), substantially higher for (M3, normal, 100), (M3, mixture normal, 100), (M4, normal, 200), and (M4, uniform, 200), slightly lower for (M2, normal, 100), (M2, normal, 200), and (M2,uniform,200), quite lower for (M2, uniform, 100), and almost comparable for the rest settings. A relatively poor performance in the reported cases is mainly caused by a particularly high proportion of extreme values of Y. For $B_0^\top x$ near zero, the right and left tails of $F(y, B_0^\top x)$ might be underestimated and overestimated, respectively. As a result, the proposed estimation procedure tends to have a higher structural dimension estimator. The use of an appropriate trimming function, which is naturally absorbed into the weight function $\widehat{F}_Y(y)$ in (2.8), would be helpful to alleviate this problem.

Compared with $\triangle(\bar{B}, B_0)$, it is shown in Tables 6 and 7 that the mean values of $\triangle(\widehat{B}, B_0)$ are slightly smaller for (M3, normal, 200), (M4, uniform, 100), and (M4, mixture normal, 100), quite smaller for ((M2, mixture normal, 100) and (M4, normal, 100), substantially smaller for (M3, normal, 100), (M3, mixture normal, 100), (M4, normal, 200), (M4, uniform, 200), and (M4, mixture normal, 200), slightly larger for (M2, normal, 200), (M3, uniform, 200), and (M4, uniform, 400), and comparable for other settings. In the unreported simulation results, the median values of $\triangle(\widehat{B}, B_0)$ are almost smaller than those of $\triangle(\bar{B}, B_0)$ even the proportions of $\widehat{d} = d_0$ are lower than those of $\bar{d} = d_0$. Generally speaking, $\triangle(\widehat{B}, B_0)$ and $\triangle(\bar{B}, B_0)$ are strongly affected by the correctness of their structural dimension estimators $\widehat{d}$ and $\bar{d}$. As for the computational speed, our approach is comparable with the ensemble approach for the sample sizes of 100 and 200 but takes a relatively long time for the sample size of 400. For an incorrect $\widehat{d}$, the difference between $\triangle(\widehat{B}, B_0)$ and

Table 3. The means (standard deviations) of 1000 estimated bandwidths and 1000 ratios of estimated bandwidth to optimal bandwidth under model M1.

ρ	n	$\widehat{h}_1$	h_{01}	$\widehat{h}_1/h_{01}$	$\widehat{h}_2$	h_{02}	$\widehat{h}_2/h_{02}$
0.2	100	1.38 (0.079)	1.54	0.90 (0.052)	1.33 (0.097)	1.62	0.82 (0.059)
	200	1.35 (0.071)	1.42	0.95 (0.050)	1.15 (0.067)	1.25	0.92 (0.054)
	400	1.32 (0.058)	1.31	1.00 (0.044)	1.09 (0.057)	1.09	1.00 (0.052)
0.5	100	1.45 (0.082)	1.47	0.99 (0.056)	1.34 (0.090)	1.42	0.94 (0.063)
	200	1.41 (0.073)	1.41	1.00 (0.052)	1.17 (0.065)	1.12	1.04 (0.058)
	400	1.38 (0.058)	1.37	1.01 (0.043)	1.06 (0.053)	1.03	1.03 (0.051)

Table 4. The means (standard deviations) of 1000 estimation accuracies under model M1.

ρ	n	$\widehat{\mathrm{MISE}}(\widehat{F}_{\widehat{C}})$	$\widehat{\mathrm{MISE}}(\widehat{F}_{\widehat{C}_{d_0}})$	$\triangle(\widehat{B}, B_0)$	$\triangle(\widehat{B}_{d_0}, B_0)$	$\triangle(\widetilde{B}_{d_0}, B_0)$	$\triangle(\check{B}, B_0)$	$\triangle(\check{B}_{d_0}, B_0)$
0.2	100	0.03 (0.013)	0.03 (0.008)	0.09 (0.133)	0.07 (0.042)	0.14 (0.069)	0.78 (0.289)	0.30 (0.109)
	200	0.02 (0.006)	0.02 (0.005)	0.05 (0.026)	0.05 (0.026)	0.09 (0.030)	*	0.26 (0.107)
	400	0.01 (0.002)	0.01 (0.002)	0.03 (0.017)	0.03 (0.017)	0.06 (0.018)	*	0.13 (0.084)
0.5	100	0.03 (0.012)	0.03 (0.007)	0.10 (0.116)	0.09 (0.054)	0.19 (0.097)	0.79 (0.312)	0.28 (0.104)
	200	0.02 (0.004)	0.02 (0.004)	0.06 (0.033)	0.06 (0.033)	0.12 (0.041)	*	0.23 (0.103)
	400	0.01 (0.002)	0.01 (0.002)	0.04 (0.019)	0.04 (0.019)	0.08 (0.021)	*	0.12 (0.061)

Table 5. The proportions of 1000 structural dimension estimates under models M2–M4 and 1000 dimension estimates of the CMS under model M3.

			Proportions of $\widehat{d}$					
Model	Dist. of X	n	0	1	2	3	4	5
M2 ($d_0 = 1$)	Normal	100	0.000	0.860	0.100	0.028	0.007	0.005
		200	0.000	0.910	0.087	0.003	0.000	0.000
		400	0.000	0.994	0.006	0.000	0.000	0.000
	Uniform	100	0.000	0.769	0.201	0.028	0.002	0.000
		200	0.000	0.916	0.081	0.003	0.000	0.000
		400	0.000	0.991	0.009	0.000	0.000	0.000
	Mixture	100	0.000	0.957	0.043	0.000	0.000	0.000
	Normal	200	0.000	0.973	0.027	0.000	0.000	0.000
		400	0.000	0.999	0.001	0.000	0.000	0.000
M3 ($d_0 = 2$)	Normal	100	0.000	0.056	0.941	0.003	0.000	0.000
		200	0.000	0.000	1.000	0.000	0.000	0.000
		400	0.000	0.000	1.000	0.000	0.000	0.000
	Uniform	100	0.000	0.044	0.931	0.025	0.000	0.000
		200	0.000	0.000	0.994	0.006	0.000	0.000
		400	0.000	0.000	1.000	0.000	0.000	0.000
	Mixture	100	0.000	0.120	0.879	0.000	0.001	0.000
	Normal	200	0.000	0.001	0.999	0.000	0.000	0.000
		400	0.000	0.000	1.000	0.000	0.000	0.000
M4 ($d_0 = 3$)	Normal	100	0.000	0.053	0.155	0.462	0.213	0.117
		200	0.000	0.006	0.018	0.808	0.149	0.019
		400	0.000	0.000	0.058	0.840	0.081	0.021
	Uniform	100	0.000	0.033	0.161	0.414	0.257	0.135
		200	0.000	0.000	0.009	0.897	0.094	0.000
		400	0.000	0.000	0.013	0.923	0.060	0.004
	Mixture	100	0.000	0.051	0.115	0.503	0.227	0.104
	Normal	200	0.000	0.001	0.148	0.793	0.045	0.012
		400	0.000	0.004	0.013	0.925	0.054	0.004
			Proportions of $\widehat{d}_M$					
Model	Dist. of X	n	0	1	2	3	4	5
M3 ($d_{M0} = 2$)	Normal	100	0.023	0.187	0.716	0.054	0.013	0.007
		200	0.000	0.020	0.944	0.036	0.000	0.000
		400	0.000	0.000	0.979	0.021	0.000	0.000
	Uniform	100	0.000	0.061	0.741	0.135	0.042	0.021
		200	0.000	0.000	0.971	0.029	0.000	0.000
		400	0.000	0.000	0.999	0.001	0.000	0.000
	Mixture	100	0.048	0.184	0.739	0.024	0.001	0.004
	Normal	200	0.000	0.032	0.938	0.030	0.000	0.000
		400	0.000	0.000	0.981	0.019	0.000	0.000

$\triangle(\widehat{B}_{d_0}, B_0)$ is rather apparent, whereas that between $\widehat{\mathrm{MISE}}(\widehat{F}_{\widehat{C}})$ and $\widehat{\mathrm{MISE}}(\widehat{F}_{\widehat{C}_{d_0}})$ is negligible.

In the designs of models M2 and M3, the corresponding $\mathcal{S}(B_0)$ and $\mathcal{S}(B_{M0})$ are easily shown to be the same. Due to extreme values around the origin of $B_0^\top X$ in model M2, one can expect poor results of the estimation criterion in (2.12) for B_{M0}. Thus, our investigation only focuses on the setting of model M3. In Table 5, all the proportions of $\widehat{d} = d_0$ are higher than those of $\widehat{d}_M = d_{M0}(= d_0)$. In addition, the means and variances of $\triangle(\widehat{B}, B_0)$ are smaller than those of $\triangle(\widehat{B}_M, B_{M0})$ except for the uniform distribution of X. As shown in Table 6, the conclusions in Section 4.1 and above can also be drawn for the difference between $\triangle(\widehat{B}_M, B_{M0})$ and $\triangle(\widehat{B}_{Md_0}, B_{M0})$ and that between $\widehat{\mathrm{MSE}}(\widehat{g}_{\widehat{C}_M})$ and $\widehat{\mathrm{MSE}}(\widehat{g}_{\widehat{C}_{Md_0}})$.

5. Applications

The developed approach for the SDR model was applied to the studies of house-price in Boston and red Vinho Verde preferences. In our data analysis, continuous covariates were standardized to have mean of zero and standard deviation of one. A random weighted bootstrap technique of Kosorok (2008) for semiparametric models was further employed to estimate the asymptotic distribution of $\widehat{B}$. Independent of collected data, exchangeable random weights were repeatedly generated 500 times based on iid Gamma(4, 2) random variables, which have better numerical results than others (see Lo 1991; Chiang and Huang 2012).

5.1. A Study of House-Price in Boston

The house-price data were collected by the U.S. Census Service from the 1970 census in the area of Boston. It consists of 14 attributes in each of 506 census tracts. The primary goal of this study focuses on identifying a correct regression model to quantify the influence of potential explanatory variables on the median value of owner-occupied homes (medv). From previous data analyses (see Breiman and Friedman 1985; Fan and Huang 2005; Chiang and Huang 2012; Kong and Xia 2012, among others), the variables of most interest are the logarithm of percentage of lower income status of the population (lstat), average number of rooms per dwelling (*rm*), pupil–teacher ratio by town (*ptratio*), weighted distances to five Boston employment centers (dis), and logarithm of full-value property-tax rate per $1000 (tax).

Table 6. The means (standard deviations) of 1000 estimation accuracies and the median computation times (seconds) of 1000 CS estimates under models M2–M4.

Model	Dist. of X	n	$\widehat{\text{MISE}}(\widehat{F}_{\widehat{C}})$	$\widehat{\text{MISE}}(\widehat{F}_{\widehat{C}_{d_0}})$	$\triangle(\widehat{B}, B_0)$	$\triangle(\widehat{B}_{d_0}, B_0)$	Time
M2	Normal	100	0.03 (0.010)	0.04 (0.008)	0.19 (0.331)	0.11 (0.152)	31
		200	0.01 (0.003)	0.01 (0.002)	0.17 (0.264)	0.08 (0.029)	111
		400	0.01 (0.003)	0.01 (0.002)	0.06 (0.077)	0.06 (0.022)	1151
	Uniform	100	0.03 (0.007)	0.03 (0.007)	0.29 (0.390)	0.08 (0.027)	28
		200	0.01 (0.004)	0.01 (0.002)	0.13 (0.265)	0.05 (0.025)	113
		400	0.01 (0.002)	0.01 (0.002)	0.06 (0.108)	0.05 (0.017)	741
	Mixture	100	0.02 (0.006)	0.02 (0.006)	0.09 (0.195)	0.07 (0.024)	23
	Normal	200	0.02 (0.003)	0.02 (0.003)	0.09 (0.153)	0.05 (0.023)	112
		400	0.01 (0.002)	0.01 (0.002)	0.05 (0.015)	0.04 (0.014)	1003
M3	Normal	100	0.05 (0.008)	0.05 (0.007)	0.26 (0.200)	0.21 (0.082)	67
		200	0.03 (0.003)	0.03 (0.003)	0.13 (0.053)	0.13 (0.053)	239
		400	0.01 (0.001)	0.01 (0.001)	0.08 (0.035)	0.08 (0.035)	938
	Uniform	100	0.04 (0.009)	0.04 (0.008)	0.45 (0.201)	0.34 (0.132)	67
		200	0.02 (0.004)	0.02 (0.004)	0.22 (0.080)	0.15 (0.061)	170
		400	0.01 (0.003)	0.01 (0.003)	0.12 (0.033)	0.12 (0.033)	600
	Mixture	100	0.05 (0.010)	0.05 (0.010)	0.27 (0.276)	0.20 (0.079)	67
	Normal	200	0.03 (0.003)	0.03 (0.003)	0.10 (0.049)	0.10 (0.039)	249
		400	0.01 (0.002)	0.01 (0.002)	0.09 (0.027)	0.09 (0.027)	885
M4	Normal	100	0.12 (0.013)	0.12 (0.012)	0.78 (0.279)	0.46 (0.169)	93
		200	0.08 (0.007)	0.07 (0.007)	0.46 (0.259)	0.32 (0.118)	342
		400	0.03 (0.004)	0.03 (0.003)	0.33 (0.302)	0.20 (0.093)	1380
	Uniform	100	0.13 (0.014)	0.13 (0.014)	0.84 (0.210)	0.59 (0.140)	98
		200	0.07 (0.007)	0.07 (0.007)	0.43 (0.218)	0.37 (0.114)	338
		400	0.03 (0.005)	0.02 (0.004)	0.41 (0.198)	0.34 (0.101)	1052
	Mixture	100	0.13 (0.012)	0.13 (0.012)	0.75 (0.330)	0.46 (0.138)	90
	Normal	200	0.08 (0.007)	0.08 (0.007)	0.37 (0.260)	0.23 (0.086)	315
		400	0.03 (0.004)	0.03 (0.004)	0.29 (0.218)	0.19 (0.084)	1292
M3	Normal	100	0.04 (0.013)	0.03 (0.011)	0.42 (0.375)	0.24 (0.172)	
		200	0.03 (0.015)	0.03 (0.012)	0.19 (0.211)	0.15 (0.103)	
		400	0.02 (0.012)	0.02 (0.007)	0.14 (0.135)	0.12 (0.067)	
	Uniform	100	0.03 (0.007)	0.03 (0.006)	0.42 (0.348)	0.24 (0.120)	
		200	0.02 (0.007)	0.02 (0.005)	0.17 (0.151)	0.15 (0.067)	
		400	0.01 (0.002)	0.01 (0.002)	0.08 (0.041)	0.08 (0.028)	
	Mixture	100	0.04 (0.017)	0.03 (0.011)	0.38 (0.379)	0.22 (0.173)	
	Normal	200	0.04 (0.017)	0.03 (0.014)	0.18 (0.225)	0.14 (0.119)	
		400	0.03 (0.015)	0.02 (0.010)	0.12 (0.133)	0.10 (0.064)	

Table 7. The proportions of 1000 structural dimension estimates, the means (standard deviations) of 1000 estimation accuracies, and the median computation times (seconds) of 1000 CS estimates under models M2–M4.

			Proportions of $\bar{d}$							
Model	Dist. of X	n	0	1	2	3	4	5+	$\triangle(\bar{B}, B_0)$	Time
M2 ($d_0 = 1$)	Normal	100	0.000	0.934	0.060	0.006	0.000	0.000	0.23 (0.200)	59
		200	0.000	1.000	0.000	0.000	0.000	0.000	0.10 (0.032)	178
		400	0.000	1.000	0.000	0.000	0.000	0.000	0.06 (0.018)	538
	Uniform	100	0.000	0.914	0.082	0.004	0.000	0.000	0.25 (0.238)	60
		200	0.000	0.998	0.002	0.000	0.000	0.000	0.10 (0.051)	185
		400	0.000	1.000	0.000	0.000	0.000	0.000	0.05 (0.016)	534
	Mixture	100	0.000	0.908	0.088	0.004	0.000	0.000	0.23 (0.251)	62
	Normal	200	0.000	1.000	0.000	0.000	0.000	0.000	0.08 (0.028)	184
		400	0.000	1.000	0.000	0.000	0.000	0.000	0.05 (0.014)	532
M3 ($d_0 = 2$)	Normal	100	0.000	0.060	0.764	0.136	0.032	0.008	0.53 (0.299)	71
		200	0.000	0.000	1.000	0.000	0.000	0.000	0.19 (0.057)	185
		400	0.000	0.000	1.000	0.000	0.000	0.000	0.11 (0.032)	516
	Uniform	100	0.000	0.078	0.804	0.098	0.020	0.000	0.49 (0.304)	72
		200	0.000	0.000	1.000	0.000	0.000	0.000	0.14 (0.044)	209
		400	0.000	0.000	1.000	0.000	0.000	0.000	0.08 (0.023)	522
	Mixture	100	0.000	0.024	0.536	0.260	0.134	0.046	0.69 (0.334)	70
	Normal	200	0.000	0.000	0.994	0.006	0.000	0.000	0.17 (0.090)	210
		400	0.000	0.000	1.000	0.000	0.000	0.000	0.09 (0.024)	496
M4 ($d_0 = 3$)	Normal	100	0.000	0.016	0.305	0.357	0.205	0.117	0.91 (0.151)	126
		200	0.000	0.000	0.344	0.554	0.092	0.010	0.71 (0.284)	211
		400	0.000	0.000	0.106	0.852	0.040	0.002	0.39 (0.267)	498
	Uniform	100	0.000	0.008	0.272	0.456	0.200	0.064	0.88 (0.170)	92
		200	0.000	0.000	0.156	0.696	0.118	0.030	0.63 (0.270)	209
		400	0.000	0.000	0.014	0.944	0.042	0.000	0.34 (0.186)	496
	Mixture	100	0.000	0.018	0.268	0.430	0.186	0.098	0.85 (0.211)	92
	Normal	200	0.000	0.000	0.146	0.710	0.124	0.020	0.57 (0.297)	210
		400	0.000	0.000	0.028	0.951	0.021	0.000	0.26 (0.182)	496

Table 8. The CS and CMS estimates (standard errors).

X	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$
lstat	1	0	0
rm	0	1	0
ptratio	0	0	1
dis	−0.24 (0.100)	−0.47 (0.108)	0.05 (0.163)
tax	1.13 (0.237)	−0.44 (0.386)	−3.02 (0.107)
X	$\widehat{\beta}_{M1}$	$\widehat{\beta}_{M2}$	$\widehat{\beta}_{M3}$
lstat	1	0	0
rm	0	1	0
ptratio	0	0	1
dis	−0.01 (0.205)	−0.45 (0.307)	0.19 (0.225)
tax	1.14 (0.262)	−0.48 (0.309)	−2.92 (0.227)

By applying the model checking procedure of Chiang and Huang (2012), the single-index distribution model was examined to be incorrect. To reappraise the effects of these predictors on the conditional distribution of house price, Huang and Chiang (2016) further proposed a more flexible semiparametric model of the form:

$$F_Y(y|x) = F\left(y, \beta_0^\top(y)x\right), \tag{5.1}$$

where $X = (\text{lstat, rm, ptratio, dis, tax})^\top$ and $\beta_0(y) = (1, \beta_{02}(y), \beta_{03}(y), \beta_{04}(y), \beta_{05})^\top$ is the corresponding y-varying and y-invariant coefficient vector. Moreover, the authors pointed out the inadequacy of model applications in the former works. To spot the true SDR model, a more thorough investigation was performed to identify $\mathcal{S}_{Y|X}$. In this data analysis, the structural dimension is determined to be 3 and the estimated CS directions, along with the bootstrap standard errors, are presented in Table 8. Further, the dimension of the CMS is determined to be 3 and the closeness between $\widehat{B}_M$ and $\widehat{B}$ is evidenced by their extremely high canonical correlations of $(1, 1, 0.978)$. It follows that $\mathcal{S}(B_{M0})$ can succeed in recovering $\mathcal{S}(B_0)$ exhaustively. In light of the fact that $E[N_y|X = x] = F(y, B_0^\top x)$ for all possible values of (x, y), $\mathcal{S}_{Y|X}$ can be shown to be an integration of the CMS of N_y over $\mathcal{Y}$. When model (5.1) is valid, $\mathcal{S}_{Y|X}$ turns out to be the same with the space spanned by $\{\beta_0(y) : y \in \mathcal{Y}\}$. By applying the eigen-decomposition technique to the estimator $\int_{\mathcal{Y}} \widehat{\beta}(y)\widehat{\beta}^\top(y)dy$ of $\int_{\mathcal{Y}} \beta_0(y)\beta_0^\top(y)dy$, the structural dimension is determined to be 3 by the Bayesian information criterion procedure of Zhu, Miao, and Peng (2006). However, the canonical correlations of $(1, 0.412, 0.062)$ between the estimated directions and $\widehat{B}$ imply that the varying linear-index might fail to capture the feature of $\{N_y : y \in \mathcal{Y}\}$. In future research, a multiphase transformation model with finite change points seems to be another potential avenue.

It is shown in Table 8 that the estimated CS directions $\widehat{\beta}_1^\top X$, $\widehat{\beta}_2^\top X$, and $\widehat{\beta}_3^\top X$ are determined mainly by (lstat, dis, tax), (rm, dis), and (ptratio, tax). The projection plots of medv versus each of $\widehat{\beta}_1^\top X$, $\widehat{\beta}_2^\top X$, and $\widehat{\beta}_3^\top X$, and the scatterplots of paired linear-indices are further displayed in Figure 1. Apparently, the plot of medv versus $\widehat{\beta}_1^\top x$ exhibits a strongly decreasing trend, while that of medv versus $\widehat{\beta}_3^\top x$ indicates a pattern of heteroscedasticity. In the plot of medv versus $\widehat{\beta}_2^\top x$, the part with $\widehat{\beta}_2^\top x > 0$ is approximately linear but that with $\widehat{\beta}_2^\top x < 0$ is approximately parabolic. In addition, we observe a possible nonlinear confounding between $\widehat{\beta}_1^\top x$ and $\widehat{\beta}_2^\top x$, a slightly negative association between $\widehat{\beta}_3^\top x$ and $\widehat{\beta}_1^\top x$, and a slightly positive association between $\widehat{\beta}_3^\top x$ and $\widehat{\beta}_2^\top x$.

5.2. A Study of Red Vinho Verde Preferences

The second application is to reanalyze 1599 red wine samples, which were collected between May 2004 and February 2007 in the Minho region of Portugal. In this dataset, an ordinal response related to preferences of sensory assessors (sens) is ranked in a scale ranging from 3 to 8 with 0 representing poor quality and 10 representing excellent quality. For the wine certification, the most commonly used physicochemical tests include alcohol (alcoh), fixed acidity (facid), volatile acidity (vacid), citric acid (cacid), residual sugar (sugar), chlorides (chlor), free sulfur dioxide (fsdiox), total sulfur dioxide (tsdiox), density (dens),

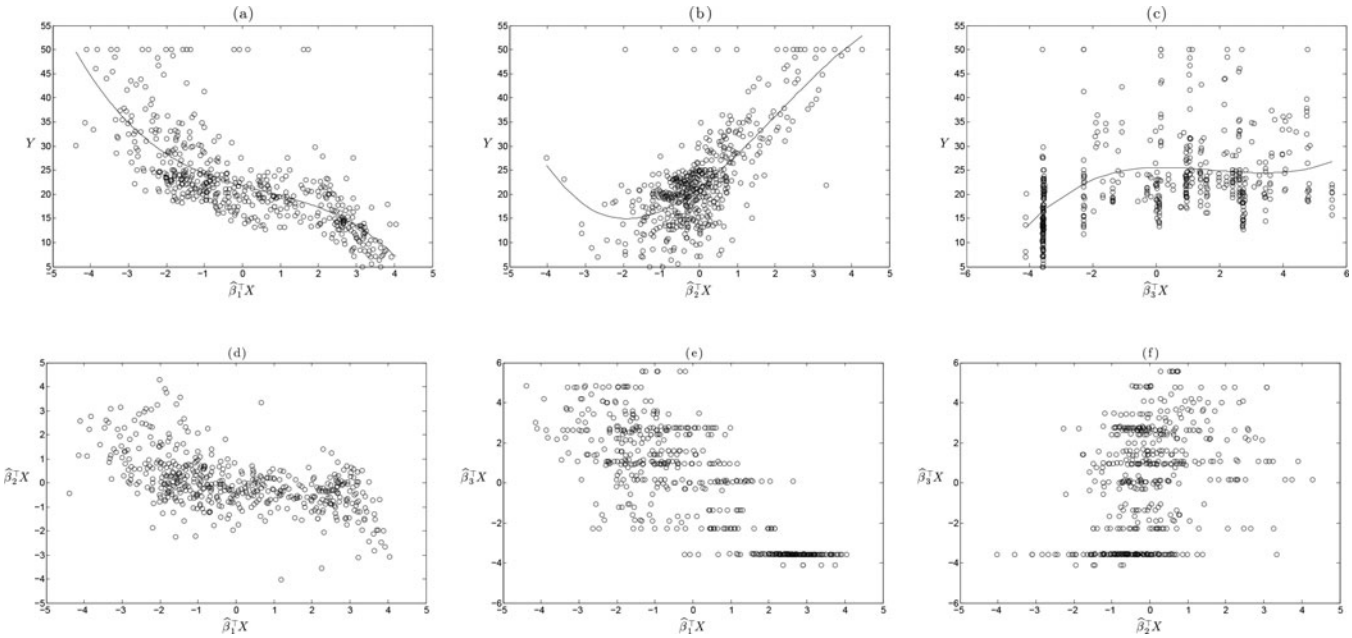

Figure 1. The projection plots of medv vs. each of linear-indices and the scatterplots of paired linear indices.

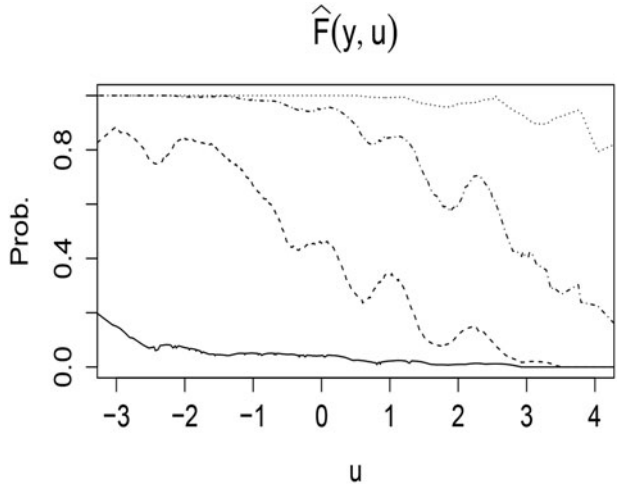

Figure 2. The estimated curves of $F(4, u)$ (solid curve), $F(5, u)$ (dashed curve), $F(6, u)$ (dotted–dashed curve), and $F(7, u)$ (dotted curve).

pH (pH), and sulphates (sulph). In the literature, the sensory analysis and these attributes have been recognized as essential elements to stratify wines and improve wine making.

As in the context of generalized linear models, it does not make sense to talk of "distance" between pairs of categorical numbers. However, Cortez et al. (2009) modeled the ordinal response sens in a continuous scale. To solve this problem, the threshold approach, in which the observable ordinal response is treated as a categorical version of the unobserved underlying continuous variable, of Edwards and Thurstone (1952) is another possibility in application. A more general model formulation in (1.1) becomes of great interest to describe the effects of physicochemical tests on such a response. By using the proposed cross-validation estimation criterion, the structural dimension is determined to be 1 and the CS direction is estimated to be (alcoh + 0.32 facid−0.59 vacid−0.12 cacid + 0.14 sugar−0.30 chlor + 0.15 fsdiox−0.34 tsdiox−0.24 dens + 0.02 pH + 0.50 sulph) with the bootstrap standard errors of (0.131, 0.065, 0.057, 0.036, 0.034, 0.047, 0.062, 0.112, 0.079, 0 .047) for the corresponding parameter estimates. It is shown that (alcoh,facid,sugar,fsdiox,sulph) and (vacid,cacid,chlor, tsdiox,dens) have significant positive and negative effects on the conditional distribution of sens. The above conclusion indicates that a support vector machine regression of Cortez et al. (2009) is likely to overfit the analyzed data.

For each given $y \in \{4, 5, 6, 7\}$, the conditional distribution estimate $\widehat{F}(y, u)$ (Figure 2) exhibits a decreasing trend over u. A widely applicable concordance index of Harrell et al. (1982) is, thus, adopted to assess the discriminability of $\beta_0^\top X$ on sens. Naturally, the concordance index, which is denoted by C_{index} and defined as $P(\beta_0^\top X_i > \beta_0^\top X_j | Y_i > Y_j)$, can be estimated by $\widehat{C}_n = \sum\sum_{i\neq j} I(\widehat{\beta}^\top X_i > \widehat{\beta}^\top X_j, Y_i > Y_j) / \sum\sum_{i\neq j} I(Y_i > Y_j)$. An estimate 0.80 of C_{index}, along with the bootstrap standard error of 0.020, shows that our classification rule has a high discriminatory power. In addition, the Cohen's Kappa $\kappa = (\sum_{k=3}^{8} \pi_{kk} - \sum_{k=3}^{8} \pi_{k+}\pi_{+k})/(1 - \sum_{k=3}^{8} \pi_{k+}\pi_{+k})$, where $\pi_{km} = P(\text{actual class} = k, \text{predicted class} = m)$, $\pi_{k+} = P(\text{actual class} = k)$, and $\pi_{+m} = P(\text{predicted class} = m)$, $k, m = 3, \ldots, 8$, is used to assess the strength of agreement between actual and predicted classes. A practical calibration rule is to predict the sensory class of a wine with the highest probability among $\{\widehat{p}_k(\widehat{\beta}^\top x) : \widehat{p}_k(\widehat{\beta}^\top x) = \widehat{F}_{\widehat{C}}(k, \widehat{\beta}^\top x)) - \widehat{F}_{\widehat{C}}(k-1, \widehat{\beta}^\top x), k = 3, \ldots, 8\}$. Based on the sample analog of κ, this measure is estimated to be 0.32 with the bootstrap standard error of 0.033. According to an empirical rule of Le (1998), a poor reproducibility can be concluded. In fact, a prediction rule that is well calibrated is not necessary to have a high discriminatory capacity, and vice versa. Thus, other potential covariates should be taken into account in a future study.

6. Concluding Remarks and Discussion

An innovative semiparametric estimation approach is developed to estimate the SDR model. The most distinctive advantage of our proposal is that the CS and conditional distribution are simultaneously estimated through an estimation criterion. An effective algorithm is further proposed to implement the estimation criterion. In addition, an asymptotically optimal bandwidth estimator is shown to satisfy the bandwidth constraint for the theoretical results and is demonstrated to have satisfactory performance in the numerical investigation. Since our estimation is mainly based on a single optimization function, it greatly simplifies the derivation of large-sample properties. A major achievement in this direction is to establish the consistency of $(\widehat{d}, \widehat{C}, \widehat{h})$ to (d_0, C_0, h_0) and the $\sqrt{n}$-consistency of $P_{\widehat{C}}$.

According to the empirical experience of Wang and Xia (2008), the finite-sample performance of inverse regression approaches might be poor for $d_0 \geq 3$. Besides, a comparable CS estimator of Yin and Li (2011) cannot achieve the $\sqrt{n}$-consistency for $d_0 \geq 4$. In contrast, our approach does not suffer from these methodological and theoretical shortcomings. The proposed CS estimator generally outperforms its competitors in a series of representative simulations. It is noted that only a fourth-order kernel function is required for $d_0 \leq 6$ in the developed approach. As for a higher structural dimension, a fairly large sample size is necessary to alleviate the numerical instability caused by the use of a higher-order kernel function.

In the proposed estimation criterion, there should exist at least d_0 continuous covariates as the candidates of $\{X_1, \ldots, X_{d_0}\}$. Under the continuity of the CS directions, our methodology allows a response to be discrete and some covariates to be discrete and/or categorical. In the perspective of Chiaromonte, Cook, and Li (2002), one might also invoke the partial SDR model to explain the dimension reduction of continuous covariates in each subpopulation or stratum, which is identified by discrete and/or categorical variables. In particular, the cross-validation criterion in (2.8) can be adapted to this model formulation in a straightforward manner. As one can see, data with discrete and/or categorical variables are frequently occurring in biomedical and social sciences. For the SDR model with some CS directions of such covariates, $\mathcal{S}_{Y|X}$ is not identifiable even the local coordinate system of the Grassmann manifold is adopted. Without imposing any structure on the conditional distribution of Y on X (see Bierens and Hartog 1988) or X on Y (see Cook and Li 2009), it is impractical to explain and estimate $\mathcal{S}_{Y|X}$.

In the simulation study, the designed models were found to have sparse representations. A challenging task is to explore nuisance covariates and enhance the prediction precision. To

the best of our knowledge, the existing approaches such as the adaptive least absolute shrinkage and selection operator could be properly modified to identify significant covariates. However, how to select an objective regularization or tuning parameter in a theoretically valid manner is still an open problem. Recently, high-dimensional data have been increasingly encountered in many fields of applications such as genomics, proteomics, bioinformatics, and biomedical imaging, among others. On the SDR model with high-dimensional covariates, Zhu, Miao, and Peng (2006) established the related large-sample properties of a sliced inverse regression estimator. In such a setup, it would be worthwhile to investigate the asymptotic behavior of the presented estimator. For a large d_0 or an extremely large p or n, the problems of slow speed and high cost are anticipated in performing the proposed computational algorithm, in which the nonlinear optimization is inevitable. To partly improve the computational efficiency, a feasible strategy is to use one of inverse regression estimators as an initial estimator whenever the regularity conditions are satisfied. This major technological challenge remains for future research.

In the study of Boston house-price data, a semiparametric model with possibly y-varying linear index was detected to be inappropriate. Furthermore, the CS of model (1.1) cannot precisely capture the local feature of the conditional distribution, that is, the CMS of $E[N_y|X]$ over $\mathcal{Y}$. On the basis of our observation, it might be able to characterize such a complicated structure through a multi-phase transformation model of the form:

$$F_Y(y|x) = \sum_{\ell=1}^{s+1} F_\ell\left(y, B_{0\ell}^\top x\right) \mathbf{1}(\tau_{\ell-1} \le y < \tau_\ell), \tag{6.1}$$

where $\tau_0 < \tau_1 < \cdots < \tau_s < \tau_{s+1}$ are the change points with τ_0 and τ_{s+1} being the respective lower and upper bounds of $\mathcal{Y}$, $B_{0\ell}$ is a $p \times d_{0\ell}$ full-rank basis matrix, and F_ℓ is an unknown $(d_{0\ell}+1)$-variate function, $\ell = 1, \ldots, s+1$, with $\mathcal{S}(B_{0\ell}) \neq \mathcal{S}(B_{0(\ell+1)})$, $\mathcal{S}(B_{0\ell}) \subseteq \mathcal{S}(B_0)$, and $\mathcal{S}(B_0) = \cup_{\ell=1}^{s+1}\mathcal{S}(B_\ell)$. An important task of this research is to estimate $\mathcal{S}(B_{0\ell})$, F_ℓ, $\ell = 1, \ldots, s+1$, and $\{\tau_1, \ldots, \tau_s\}$. As alternatives to model (6.1), one could also invoke multi-phase density and hazards models.

Supplementary Materials

The online supplement contains the appendices for the article.

Acknowledgments

The research of the corresponding author was partially supported by the Ministry of Science and Technology grants 99-2118-M-002-003 and 100-2118-M-002-005-MY2 (Taiwan). The authors are grateful to Dr. Shao-Hsuan Wang for some numerical computations used in this article. We would also like to thank the associate editor and three reviewers for many constructive comments on this article .

References

Bellman, R. (1961), *Adaptive Control Process: A Guide Tour*, Princeton, NJ: Princeton. [2]

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1998), *Efficient and Adaptive Estimation for Semiparametric Models*, New York: Springer. [2]

Bierens, H. J., and Hartog, J. (1988), "Nonlinear Regression With Discrete Explanatory Variables, With an Application to the Earnings Function," *Journal of Econometrics*, 38, 269–299. [13]

Borisenko, A. A., and Nikolaevskiĭ, Y. A. (1991), "Grassmann Manifolds and Grassmann Image of Submanifolds," *Uspekhi Matematicheskikh Nauk*, 46, 41–83, 240. [2]

Breiman, L., and Friedman, J. H. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation," *Journal of the American Statistical Association*, 80, 580–619. [10]

Chiang, C. T., and Huang, M. Y. (2012), "New Estimation and Inference Procedures for a Single-Index Conditional Distribution Model," *Journal of Multivariate Analysis*, 111, 271–285. [2,3,8,10]

Chiaromonte, F., Cook, R. D., and Li, B. (2002), "Sufficient Dimension Reduction in Regression With Categorical Predictors," *Annals of Statistics*, 30, 475–497. [13]

Cook, R. D. (1998), *Regression Graphics*, New York: Wiley. [1]

Cook, R. D., and Li, L. (2009), "Dimension Redution in Regressions With Exponential Family Predictors," *Journal of Computational and Graphical Statistics*, 18, 774–791. [13]

Cook, R. D., and Weisberg, S. (1991), Comment on "Sliced Inverse Regression for Dimension Reduction" by K.C. Li, *Econometrica*, 86, 328–332. [1]

Cook, R. D., and Zhang, X. (1991), "Fused Estimators of the Central Subspace in Sufficient Dimension Reduction," *Journal of the American Statistical Association*, 109, 815–827. [2]

Cortez, P., Cordeira, A., Almeida, F., Matos, T., and Reis, J. (2009), "Modeling Wine Preferences by Data Mining From Physicochemical Properties," *Decision Support Systems*, 45, 547–553. [13]

Cosslett, S. R. (1983), "Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model," *Journal of the American Statistical Association*, 51, 765–782. [1]

Delecroix, M., Härdle, W., and Hristache, M. (2003), "Efficient Estimation in Conditional Single-Index Regression," *Journal of Multivariate Analysis*, 86, 213–226. [1]

Dong, Y., and Li, B. (2010), "Dimension Reduction for Non-Elliptically Distributed Predictors: Second-Order Moments," *Biometrica*, 97, 279–294. [2,8]

Edwards, A., and Thurstone, L. (1952), "An Internal Consistency Check for Scale Values Determined by the Method of Successive Intervals," *Psychometrika*, 17, 169–180. [13]

Fan, J., and Huang, T. (2005), "Profile Likelihood Inferences on Semiparametric Varying-Coefficient Partially Linear Models," *Bernoulli*, 11, 1030–1057. [10]

Fan, J., and Yim, T. H. (2004), "A Crossvalidation Method for Estimating Conditional Densities," *Biometrika*, 91, 819–834. [2]

Fletcher, R., and Reeves, C. M. (1964), "Function Minimization by Conjugate Gradient," *Comput. J.*, 7, 149–154. [5]

Hall, P. (1987), "On Kullback–Leibler Loss and Density Estimation," *Annals of Statistics*, 15, 1491–1519. [4]

Hall, P., and Yao, Q. (2005), "Approximating Conditional Distribution Functions Using Dimension Reduction," *Annals of Statistics*, 33, 1404–1421. [1]

Härdle, W., Hall, P., and Marron, J. S. (1988), "How Far are Automatically Chosen Regression Smoothing Parameters From Their Optimum?" *Journal of the American Statistical Association*, 83, 86–101. [4]

Härdle, W., and Marron, J. S. (1985), "Optimal Bandwidth Selection in Nonparametric Regression Function Estimation," *Annals of Statistics*, 13, 1465–1481. [4]

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982), "Evaluating the Yield of Medical Tests," *Journal of the American Medical Association*, 247, 2543–2546. [13]

Horowitz, J. L., and Härdle, W. (1996), "Direct Semiparametric Estimation of Single-Index Models With Discrete Covariates," *Journal of the American Statistical Association*, 91, 1632–1640. [2]

Huang, M. Y., and Chiang, C. T. (2016), "Estimation and Inference Procedures for Semiparametric Distribution Models With Varying Linear-Index," *Scandinavian Journal of Statistics*, doi:10.1111/sjos.12258. [12]

Ichimura, H. (1993), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models," *Journal of Econometrics*, 58, 71–120. [2,7]

Jorge, N., and Stephen, J. W. (2006), *Numerical Optimization*, New York: Springer. [5]

Kong, E., and Xia, Y. (2012), "A Single-Index Quantile Regression Model and Its Estimation," *Econometric Theory*, 28, 730–768. [10]

Kosorok, M. R. (2008), *Introduction to Empirical Processes and Semiparametric Inference*, New York: Springer. [10]

Le, C. T. (1998), *Applied Categorical Data Analysis*, New York: Wiley. [13]

Li, B., and Wang, S. (2007), "On Directional Regression for Dimension Reduction," *Journal of the American Statistical Association*, 102, 997–1008. [1]

Li, K. C. (1991), "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*, 86, 316–342. [1]

——— (1992), "On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma," *Journal of the American Statistical Association*, 87, 1025–1039. [9]

Lo, A. Y. (1991), "Bayesian Bootstrap Clones and a Biometry Function," *Sankhyā*, A53, 320–333. [10]

Ma, Y., and Zhu, L. P. (2012), "A Semiparametric Approach to Dimension Reduction," *Journal of the American Statistical Association*, 107, 168–179. [2]

——— (2013), "Efficient Estimation in Sufficient Dimension Reduction," *Annals of Statistics*, 41, 250–268. [2,4,7]

Tsiatis, A. A. (2006), *Semiparametric Theory and Missing Data*, New York: Springer. [2]

Wang, H., and Xia, Y. (2008), "Sliced Regression for Dimension Reduction," *Journal of the American Statistical Association*, 103, 811–821. [2,9,13]

Xia, Y. (2007), "A Constructive Approach to the Estimation of Dimension Reduction Directions," *Annals of Statistics*, 35, 2654–2690. [2]

Xia, Y., Tong, H., Li, W. K., and Zhu, L. X. (2002), "An Adaptive Estimation of Dimension Reduction Space," *Journal of the Royal Statistical Society*, Series B, 363–410. [2]

Yin, X., and Li, B. (2011), "Sufficient Dimension Reduction Based on an Ensemble of Minimum Average Variance Estimators," *Annals of Statistics*, 39, 3392–3416. [2,9,13]

Zeng, P., and Zhu, Y. (2010), "An Integral Transform Method for Estimating the Central Mean and Central Subspaces," *Journal of Multivariate Analysis*, 101, 271–290. [1]

Zhu, L. P., Yu, Z., and Zhu, L. X. (2010), "A Sparse Eigen-Decomposition Estimation in Semiparametric Regression," *Computational Statistics & Data Analysis*, 54, 976–9866. [1]

Zhu, L. P., Zhu, L. X., and Feng, Z. H. (2010), "Dimension Reduction in Regressions Through Cumulative Slicing Estimation," *Journal of the American Statistical Association*, 105, 1455–1466. [2]

Zhu, L. X., Miao, B., and Peng, H. (2006), "On Sliced Inverse Regression With High-Dimensional Covariates," *Journal of the American Statistical Association*, 101, 630–643. [2,12,14]

Zhu, Y., and Zeng, P. (2006), "Fourier Methods for Estimating the Central Subspace and the Central Mean Subspace in Regression," *Journal of the American Statistical Association*, 101, 1638–1651. [2]