Biometrika (2017), **104**, 3, *pp.* 583–596
Printed in Great Britain
doi: 10.1093/biomet/asx028
Advance Access publication 19 May 2017

Joint sufficient dimension reduction and estimation of conditional and average treatment effects

By MING-YUEH HUANG and KWUN CHUEN GARY CHAN

Department of Biostatistics, University of Washington, Seattle, Washington 98105, U.S.A.
myh0728@uw.edu kcgchan@u.washington.edu

Summary

The estimation of treatment effects based on observational data usually involves multiple confounders, and dimension reduction is often desirable and sometimes inevitable. We first clarify the definition of a central subspace that is relevant for the efficient estimation of average treatment effects. A criterion is then proposed to simultaneously estimate the structural dimension, the basis matrix of the joint central subspace, and the optimal bandwidth for estimating the conditional treatment effects. The method can easily be implemented by forward selection. Semiparametric efficient estimation of average treatment effects can be achieved by averaging the conditional treatment effects with a different data-adaptive bandwidth to ensure optimal undersmoothing. Asymptotic properties of the estimated joint central subspace and the corresponding estimator of average treatment effects are studied. The proposed methods are applied to a nutritional study, where the covariate dimension is reduced from 11 to an effective dimension of one.

Some key words: Forward selection; High-order kernel; Joint central subspace; Optimal bandwidth; Semiparametric efficiency; Undersmoothing.

1. Introduction

Investigating the causal effect of a treatment on an outcome is often the primary interest in medical and social studies. While randomization is the gold standard in identifying treatment effects, often only observational data are available. A major challenge in observational studies is confounding, where the treatment and the outcome of interest are associated with other pretreatment variables, potentially leading to seriously biased estimation of average treatment effects. A simple method of dealing with confounding is local matching. Under conditional independence, the distribution of outcomes in a specific group behaves just like that of a random sample from the group while conditioning on values of confounding variables. Thus, a consistent estimator can be obtained by averaging the differences between groups over the distribution of confounding variables. Hahn (1998) introduced a class of nonparametric imputation estimators that use local matching and showed that they are asymptotically efficient.

Although nonparametric imputation estimators are $n^{1/2}$-consistent under regularity conditions, the remainder terms depend on their biases and variances. In particular, the variance increases with the number of confounders, so a balancing score vector with a smaller dimension is always preferred. According to Rosenbaum & Rubin (1983), the propensity score is the coarsest balancing

score, and also has the smallest dimension. However, Hahn (1998) showed that projection onto the true propensity score can be inefficient. Another well-known balancing score is the prognostic score of Hansen (2008). Leacy & Stuart (2014) further combined propensity and prognostic scores to improve the classical matching estimator of average treatment effects. Unfortunately, estimators based directly on propensity and prognostic scores often cannot attain the semiparametric efficiency bound. Hence, finding a suitable balancing score vector with high efficiency and minimal dimension is important in practice.

There are two approaches to reducing the dimension of the covariate vector while keeping as much information as possible about the relationship between response and covariates. The first is variable selection, in which the main goal is to drop redundant variables: de Luna et al. (2011) first defined some subsets of confounding variables which are minimal in the sense that the treatment ceases to be unconfounded for any proper subset of these sets; they also showed that these subsets can reduce the efficiency bound for average treatment effects. Related methods are discussed in Vansteelandt et al. (2012). Instead of using the total subset selection procedure, another method of variable selection is penalized regression. Since confoundedness resides in the conditional distribution of the potential outcomes and treatment variable, given confounders, Ghosh et al. (2015) developed a lasso-type criterion to select redundant variables. The second approach is sufficient dimension reduction, which seeks a few linear combinations of confounders that retain the full information on confoundedness. In a related missing data problem, Hu et al. (2014) introduced effective balancing scores, which are the central subspaces of missing indicators and observed responses, given covariates. While the estimation of average treatment effect can often be regarded as two missing data problems, applying this approach twice would yield two estimates of central subspaces that could be highly collinear.

In this paper, we introduce a joint sufficient dimension reduction model on the propensity score as well as the conditional distributions of potential outcomes, and use the joint central subspace to form a semiparametric efficient estimator of average treatment effects. No stringent parametric model formulation is assumed in the dimension reduction framework. Further, while classical dimension reduction methods consider models on the joint distribution of treatment and potential outcomes, our approach focuses on the marginal distributions and can yield a balancing score with smaller dimension. The exclusion restrictions used in de Luna et al. (2011) for efficiency gains are not required, but the semiparametric efficiency bound is retained by our estimator.

Several approaches to sufficient dimension reduction can be extended to this causal inference problem. In a complete-data setting, these approaches include inverse regression (Li, 1991; Li & Wang, 2007; Zhu et al., 2010), average derivative and minimum average variance estimation (Zhu & Zeng, 2006; Xia, 2007; Wang & Xia, 2008; Yin & Li, 2011), and the semiparametric framework (Ma & Zhu, 2012, 2013; Huang & Chiang, 2017). Here we extend the work of Huang & Chiang (2017) and construct a crossvalidation-type least squares criterion to estimate the structural dimension and the basis matrix simultaneously. The bandwidth used in the semparametric estimator of the unspecified link function can be selected using the same criterion and attains the optimal rate for estimating the conditional treatment effects. The proposed model is flexible and the average treatment effects can be estimated efficiently.

Although local matching using the propensity score may not be semiparametrically efficient, inverse propensity score weighting with estimated weights has been shown to be efficient (Hirano et al., 2003), and many recent efforts have focused on improving the weighting estimators (Qin & Zhang, 2007; Imai & Ratkovic, 2014; Chan et al., 2016). In contrast to those estimators, our method provides a rate-optimal estimator of covariate-specific treatment effects, which is useful for personalized prediction and the study of heterogeneity.

2. Joint sufficient dimension reduction

2·1. *Notation and model construction*

Let $Y(0)$ and $Y(1)$ be the potential outcomes when an individual is assigned to the control and treatment groups, respectively, and let T be a binary treatment indicator. Since each unit is either treated or not treated, the observed outcome is $Y = TY(1) + (1-T)Y(0)$. In addition, a vector of covariates or confounders $X = (X_1, \ldots, X_p)^{\mathrm{T}}$ is observed for each subject, and we make the following conditional independence assumption.

Assumption 1 (Unconfounded treatment assignment). We have that $T \perp\!\!\!\perp \{Y(0), Y(1)\} \mid X$, where $\perp\!\!\!\perp$ denotes independence.

This assumption is often made to identify the average treatment effect $\tau = E\{Y(1) - Y(0)\}$. Under Assumption 1, Hahn (1998) and Robins et al. (1994) derived the semiparametric efficiency bound for τ as

$$\sigma^2_{\mathrm{eff}} = E\left[\{m_1(X) - m_0(X) - \tau\}^2 + \frac{\sigma_1^2(X)}{\pi(X)} + \frac{\sigma_0^2(X)}{1-\pi(X)}\right],$$

where $\pi(X) = \mathrm{pr}(T = 1 \mid X)$ is the propensity score, $m_k(X) = E\{Y(k) \mid X\}$ is the conditional mean of the potential outcome, and $\sigma_k^2(X) = \mathrm{var}\{Y(k) \mid X\}$ $(k = 0, 1)$. Also, σ^2_{eff} can be shown to be the asymptotic variance of a nonparametric imputation estimator by directly using X as a balancing score. A balancing score with smaller dimension but the same efficiency is obtained if the conditional distributions of $Y(0)$, $Y(1)$ and T given X are captured by a lower-dimensional linear subspace of X. Therefore, we focus on finding $B^{\mathrm{T}}X$ such that

$$T \perp\!\!\!\perp X \mid B^{\mathrm{T}}X, \quad Y(0) \perp\!\!\!\perp X \mid B^{\mathrm{T}}X, \quad Y(1) \perp\!\!\!\perp X \mid B^{\mathrm{T}}X, \tag{1}$$

where B is a full-rank $p \times d$ parameter matrix with $d \leqslant p$; we call the column space of B a joint sufficient dimension reduction subspace. For simplicity, we will write $\mathrm{span}(B)$ for the column space of a matrix B. Obviously, (1) holds when $d = p$ and B is the $p \times p$ identity matrix, I_p. Thus it always covers the true model. Moreover, if $\mathrm{span}(B_1) \subset \mathrm{span}(B_2)$ and $\mathrm{span}(B_1)$ is a joint sufficient dimension reduction subspace, then $\mathrm{span}(B_2)$ will also be a joint sufficient dimension subspace. The most interesting parameter is therefore the joint sufficient dimension reduction subspace of smallest dimension, called the joint central subspace when it exists, which is unique up to an equivalence class as discussed in Remark 2. The corresponding basis matrix B_0 has dimension d_0. The existence and uniqueness of the joint central subspace can be ensured under some mild conditions, similar to the discussion of Cook (1998) on sufficient dimension reduction for univariate responses.

Alternatively, based on the classical literature on sufficient dimension reduction, one can also consider the model

$$\{T, Y(0), Y(1)\} \perp\!\!\!\perp X \mid B^{\mathrm{T}}X, \tag{2}$$

which is different from model (1). In fact, $\mathrm{span}(B_0)$ will be contained in the central subspace of (2). Since the average treatment effect involves only the marginal distributions of $\{T, Y(0), Y(1)\}$ and it is appealing to seek a balancing score with lower dimension, we consider model (1) instead of model (2).

Based on the definition of a joint central subspace, $B_0^{\mathrm{T}}X$ is obviously a balancing score and

$$T \perp\!\!\!\perp \{Y(0), Y(1)\} \mid B_0^{\mathrm{T}}X,$$

which ensures unbiased estimation of the average treatment effect. The main feature of this balancing score is that it creates both propensity and prognostic balance (Hansen, 2008; Leacy & Stuart, 2014), and we will show in § 3 that it attains the semiparametric efficiency bound in the estimation of τ.

Remark 1. Unlike sufficient dimension reduction tools such as sliced inverse regression (Li, 1991), we do not require an additional linearity assumption on the covariate distribution.

Remark 2. Under model (1), $\pi(B^{\mathrm{T}}X)$ and $F_k(y \mid B^{\mathrm{T}}X)$ $(k = 0, 1)$ remain the same for any basis matrix B with the same column space. In fact, all such B span the same space and are isomorphic up to a linear transformation. The parameter space of B is called the Grassmann manifold, or Grassmannian, denoted by $\mathrm{Gr}(d, \mathbb{R}^p)$. To avoid ambiguity, we follow Ma & Zhu (2013) in employing the local coordinate system of the Grassmannian and parameterize the basis by $B = (I_d, C)^{\mathrm{T}}$, where C is a $(p - d) \times d$ free parameter matrix. This parameterization is particularly useful in theoretical developments for characterizing the information matrix, and is not an additional model assumption or a restriction on $\mathrm{span}(B)$. Computation of the proposed estimator does not require fixing the reference variables in advance; see Remark 6.

Remark 3. Since the main parameters of interest are means of potential outcomes, one could also consider the joint central mean subspace, which is the smallest linear subspace with basis matrix B_M such that

$$T \perp\!\!\!\perp X \mid B_M^{\mathrm{T}}X, \quad Y(0) \perp\!\!\!\perp E\{Y(0) \mid X\} \mid B_M^{\mathrm{T}}X, \quad Y(1) \perp\!\!\!\perp E\{Y(1) \mid X\} \mid B_M^{\mathrm{T}}X. \tag{3}$$

The corresponding method in a complete-data setting can be found in Cook & Li (2002) and Xia (2008). Note that the distribution $T \perp\!\!\!\perp X \mid B_M^{\mathrm{T}}X$ remains the same as in the sufficient dimension reduction model because T is binary and its distribution is determined by its mean. Since (3) models the conditional means only and the mean is a functional of the distribution, one can verify that $\mathrm{span}(B_M) \subset \mathrm{span}(B_0)$. Moreover, by Theorems 2 and 3 of Rosenbaum & Rubin (1983), $B_M^{\mathrm{T}}X$ is also a balancing score, for which Proposition 1 below holds. Thus we obtain a balancing score with a possibly smaller dimension for the estimation of average treatment effects. However, a comparison of σ^2_{eff} and (8) in § 3·1 reveals that the efficiency bound will not be generally attained with use of $B_M^{\mathrm{T}}X$ as a balancing score; that is, in general there is a trade-off between a lower dimension and a smaller asymptotic variance. This trade-off is also discussed in Hu et al. (2014), who studied mean estimation in missing data. Furthermore, the current formulation is sufficient for the estimation of any conditional functionals, not just the conditional means, and does not require re-estimation of the central subspace for different functionals of interest; see Remark 7. Therefore, we consider model (1) so that all relevant information is kept.

2·2. *Simultaneous estimation for the basis and dimension of the joint central subspace*

Here we develop an estimation criterion for the joint central subspace with a random sample $\{(T_i, Y_i, X_i) : i = 1, \ldots, n\}$. First, we note that model (1) is equivalent to

$$E(T \mid X) = E(T \mid B^{\mathrm{T}}X),$$
$$E[1\{Y(k) \leqslant y\} \mid X] = E[1\{Y(k) \leqslant y\} \mid B^{\mathrm{T}}X] \quad (y \in \mathbb{R};\ k = 0, 1),$$

where $1(\cdot)$ represents the indicator function. That is, we consider semiparametric regression models $g(B^{\mathrm{T}}X)$ and $G(y, B^{\mathrm{T}}X)$, where g and G are unspecified, for responses T and $1\{Y(k) \leqslant y\}$ on their corresponding mean functions $\pi(X)$ and $F_k(y \mid X) = \mathrm{pr}\{Y(k) \leqslant y \mid X\}$ ($y \in \mathbb{R}$; $k = 0, 1$). However, since $1\{Y(0) \leqslant y\}$ and $1\{Y(1) \leqslant y\}$ are not always observed, we must be careful in using them as responses. Our key idea comes from the following proposition.

PROPOSITION 1. *Under Assumption 1 and model (1),*

$$\begin{aligned} E[1\{Y(k) \leqslant y\} \mid X] &= E\{1(Y \leqslant y) \mid T = k, X\} \\ &= E\{1(Y \leqslant y) \mid T = k, B_0^{\mathrm{T}}X\} \quad (y \in \mathbb{R};\ k = 0, 1). \end{aligned} \tag{4}$$

The first equality in (4) indicates that the regression problem can be solved by considering treatment and control groups separately. The second equality leads to a sieve approach for the estimation of the unknown link function. Let

$$\begin{aligned} \hat{\pi}(B^{\mathrm{T}}x) &= \frac{\sum_{j=1}^n T_j \mathcal{K}_h\{B^{\mathrm{T}}(X_j - x)\}}{\sum_{j=1}^n \mathcal{K}_h\{B^{\mathrm{T}}(X_j - x)\}}, \\ \hat{F}_k(y \mid B^{\mathrm{T}}x) &= \frac{\sum_{j=1}^n T_j^k (1 - T_j)^{1-k}\, 1(Y_j \leqslant y) \mathcal{K}_h\{B^{\mathrm{T}}(X_j - x)\}}{\sum_{j=1}^n T_j^k (1 - T_j)^{1-k} \mathcal{K}_h\{B^{\mathrm{T}}(X_j - x)\}} \quad (k = 0, 1) \end{aligned}$$

denote estimators for $\pi(B^{\mathrm{T}}x)$ and $F_k(y \mid B^{\mathrm{T}}x) = \mathrm{pr}(Y \leqslant y \mid T = k, B^{\mathrm{T}}X = B^{\mathrm{T}}x)$, where $\mathcal{K}_h(u) = \prod_{k=1}^d K(u_k/h)/h$ with $u = (u_1, \ldots, u_d)^{\mathrm{T}}$, h a positive bandwidth, and K a qth-order kernel function. The choice of q will be discussed in Remark 5. Now we may use Proposition 1 to construct a crossvalidation-type criterion for the estimation of the joint central subspace. Write $\langle f(y), g(y)\rangle_W = \int f^{\mathrm{T}}(y) W g(y)\, \mathrm{d}F_Y(y)$ and $\|f(y)\|_W = \langle f(y), f(y)\rangle_W^{1/2}$ for generic vector-valued functions f, g and matrix W, where $F_Y(y)$ is the marginal distribution of Y. Let (T^0, Y^0, X^0) be a future run, independent of the current data $\{(T_i, Y_i, X_i) : i = 1, \ldots, n\}$, and define the prediction risk

$$E\big\{\|Y_y^0 - \hat{F}(y \mid B^{\mathrm{T}}X^0)\|_{W^0}^2\big\} \tag{5}$$

where

$$\begin{aligned} Y_y^0 &= \{T^0, 1(Y^0 \leqslant y), 1(Y^0 \leqslant y)\}^{\mathrm{T}}, \\ \hat{F}(y \mid B^{\mathrm{T}}X) &= \{\hat{\pi}(B^{\mathrm{T}}X), \hat{F}_0(y \mid B^{\mathrm{T}}X), \hat{F}_1(y \mid B^{\mathrm{T}}X)\}^{\mathrm{T}}, \\ W^0 &= (1 - \pi) e_1 e_1^{\mathrm{T}} + (1 - T^0) e_2 e_2^{\mathrm{T}} + T^0 e_3 e_3^{\mathrm{T}}, \end{aligned}$$

with $\pi = \mathrm{pr}(T = 1)$ being the marginal probability of treatment and (e_1, e_2, e_3) the standard basis of $\mathbb{R}^3$. The weight $1 - \pi$ is used to treat the squared error as an integration over the distribution of T. Further, let $F(y \mid B^{\mathrm{T}}X) = \{\pi(B^{\mathrm{T}}X), F_0(y \mid B^{\mathrm{T}}X), F_1(y \mid B^{\mathrm{T}}X)\}^{\mathrm{T}}$. A simple calculation shows that the prediction risk in (5) can be decomposed into

$$\sigma_0^2 + b_0^2(B) + \mathrm{MISE}_B(h) + C(B, h), \tag{6}$$

where

$$\sigma_0^2 = E\{\|Y_y^0 - F(y \mid B_0^{\mathrm{T}} X^0)\|_{W^0}^2\}, \quad b_0^2(B) = E\{\|F(y \mid B_0^{\mathrm{T}} X^0) - F(y \mid B^{\mathrm{T}} X^0)\|_{W^0}^2\},$$
$$\mathrm{MISE}_B(h) = E\{\|F(y \mid B^{\mathrm{T}} X^0) - \hat{F}(y \mid B^{\mathrm{T}} X^0)\|_{W^0}^2\},$$
$$C(B,h) = 2E\{\langle F(y \mid B_0^{\mathrm{T}} X^0) - F(y \mid B^{\mathrm{T}} X^0), F(y \mid B^{\mathrm{T}} X^0) - \hat{F}(y \mid B^{\mathrm{T}} X^0)\rangle_{W^0}\}.$$

As $h \to 0$ and $nh^d \to \infty$, it is shown in the Supplementary Material that the last two terms of (6) converge to zero and are dominated by the first two terms. Note that $b_0^2(B) \geqslant 0$ and $b_0^2(B) = 0$ when $\mathrm{span}(B)$ is a joint sufficient dimension reduction subspace. Hence, the minimum of the prediction risk must occur when $\mathrm{span}(B) \supseteq \mathrm{span}(B_0)$. Moreover, since our model has a nested structure, the prediction risk decreases with the working dimension d when $d \leqslant d_0$. On the other hand, when $d \geqslant d_0$ and $\mathrm{span}(B) \supseteq \mathrm{span}(B_0)$, the prediction risk has an asymptotic order of $\sigma_0^2 + O_{\mathrm{p}}\{n^{-2q/(2q+d)}\}$, which increases with d. Therefore, for large enough n, the minimal prediction risk occurs at the joint central subspace B_0. A formal result is stated as follows.

PROPOSITION 2. *Under Assumption 1 and model (1), the joint central subspace B_0 and the optimal bandwidth $h_0 = c_{d_0} n^{-1/(2q+d_0)}$ minimize the prediction risk in (5) as $h \to 0$, $nh^d \to \infty$ and $n \to \infty$, where the constant c_{d_0} is given in the Supplementary Material.*

The proof of Proposition 2 is given in the Supplementary Material. According to Proposition 2 and the fact that the prediction risk is asymptotically convex in d, we obtain our estimator through the following algorithm.

Step 1. For $d = 0$, calculate

$$\mathrm{CV}_0 = \frac{1}{n}\sum_{i=1}^{n}\left[(T_i - \bar{T})^2(1-\bar{T}) + \sum_{k=0}^{1} T_i^k (1-T_i)^{1-k} \int \{1(Y_i \leqslant y) - \hat{F}_k(y)\}^2 \,\mathrm{d}\hat{F}_Y(y)\right],$$

where $\bar{T} = n^{-1}\sum_{i=1}^n T_i$, $\hat{F}_Y(y)$ is the empirical distribution of $(Y_1, \ldots, Y_n)$, and $\hat{F}_k(y) = n^{-1}\sum_{i=1}^n T_i^k (1-T_i)^{1-k} 1(Y_i \leqslant y)/\{\bar{T}^k(1-\bar{T})^{1-k}\}$ $(k = 0, 1)$.

Step 2. For $d \geqslant 1$, let $(\hat{B}_d, \hat{h}_d)$ be the minimizer of $\mathrm{CV}(B,h)$ among all $p \times d$ matrices B and positive h, where

$$\mathrm{CV}(B,h) = \frac{1}{n}\sum_{i=1}^{n}\Bigg[\{T_i - \hat{\pi}^{-i}(B^{\mathrm{T}} X_i)\}^2(1-\bar{T}) + \sum_{k=0}^{1} T_i^k (1-T_i)^{1-k} \int \{1(Y_i \leqslant y) - \hat{F}_k^{-i}(y \mid B^{\mathrm{T}} X_i)\}^2 \,\mathrm{d}\hat{F}_Y(y)\Bigg];$$

here the superscript $-i$ indicates the estimator based on a sample with the ith subject deleted. Then calculate $\mathrm{CV}_d = \mathrm{CV}(\hat{B}_d, \hat{h}_d)$.

Step 3. Repeat Step 2 until $d = \hat{d}$ with $\mathrm{CV}_{\hat{d}+1} > \mathrm{CV}_{\hat{d}}$. The proposed estimator is $(\hat{B}, \hat{h}) = (\hat{B}_{\hat{d}}, \hat{h}_{\hat{d}})$.

We will show that $\mathrm{CV}(B,h)$ converges uniformly to the prediction risk as $n\to\infty$ in the proof of the following theorem, and the proposed algorithm can simultaneously estimate the joint central subspace and the structural dimension consistently. In summary, our proposed method, which is easily implemented in practice, can simultaneously estimate the basis matrix, the structural dimension of the joint central subspace, and the optimal bandwidth. The asymptotic properties of the proposed estimators are established in the following theorem.

THEOREM 1. *Suppose that Assumption* 1 *and Conditions* A1–A5 *in the Supplementary Material are satisfied. Then* $\mathrm{pr}(\hat{d}=d_0)\to 1$, $\hat{h}=O_{\mathrm{p}}\{n^{-1/(2q+d_0)}\}$, *and*

$$n^{1/2}\mathrm{vec}(\hat{B}-B_0)1(\hat{d}=d_0)=n^{-1/2}\sum_{i=1}^{n}\xi_{B_0}(T_i,Y_i,X_i)+o_{\mathrm{p}}(1)\to N(0,\Sigma_{B_0})$$

in distribution as $n\to\infty$, *where* vec *is the columnwise matrix vectorization operator,* $\xi_{B_0}(T,Y,X)=V^{-1}(B_0)S(B_0)$ *and* $\Sigma_{B_0}=V^{-1}(B_0)E\{S^{\otimes 2}(B_0)\}V^{-1}(B_0)$, *with* $S(B)$ *and* $V(B)$ *as defined in the Supplementary Material.*

Remark 4. The proposed criterion selects the basis matrix and the structural dimension of the joint central subspace simultaneously, which is different from most existing sufficient dimension reduction methods such as inverse regression (Zhu et al., 2010), minimum average variance estimation (Yin & Li, 2011) and semiparametric approaches (Ma & Zhu, 2012, 2013). Moreover, the bandwidth chosen by this criterion is the rate-optimal bandwidth in terms of the mean integrated squared error, which will be further discussed in § 2·3.

Remark 5. We use different bandwidths h_d for the working dimension d in the proposed algorithm, and we show in the proof of Theorem 1 that $\hat{h}_d=O_{\mathrm{p}}\{n^{-1/(2q+d)}\}$. Coupled with the restriction in Condition A3, the order of the kernel function should satisfy $q>\max\{2,(d+2)/2\}$. Since we always use a symmetric kernel function, whose order will be even, and require the order to be as small as possible, a suitable choice is $q=\max\{4,2\lfloor(d+6)/4\rfloor\}$ for each working dimension d, where $\lfloor\cdot\rfloor$ denotes the floor function.

Remark 6. Grassmanian optimization algorithms in Edelman et al. (1999) and Adragni et al. (2012) can be used in Step 2 without fixing the reference variables a priori. Those algorithms are a modification of conventional gradient-based algorithms, which consider movements along geodesics based on a metric defined in the tangent space of the Grassmann manifold. Since the optimization is nonconvex, a reliable initial value is needed. We suggest using the kernel-based method of Fukumizu & Leng (2014), as it is the fastest method known to date. That method is consistent but does not attain a $n^{1/2}$ convergence rate. Given this initial value, Step 2 can be implemented in a slightly different manner. The initial value could first be transformed into a local coordinate representation by Gaussian elimination, where the reference variables are chosen to be those with the largest columnwise coefficients in absolute value for numerical stability. Then the optimization can be performed with respect to free parameters in the Euclidean local coordinate system. We have compared these two methods and different initial values through simulations reported in the Supplementary Material, and found them to yield similar performance.

2·3. *Estimation of conditional effects*

An important advantage of our proposed criterion is that it selects the bandwidth simultaneously with the estimation of the joint central subspace, and the selected bandwidth $\hat{h}$ minimizes the

mean integrated squared error asymptotically. In particular, the bandwidth achieves the optimal rate for estimating conditional regression functions. Hence, we may use the bandwidth directly to obtain estimators of the conditional effects $E\{Y(k) \mid X = x\}$:

$$\hat{E}\{Y(k) \mid X = x\} = \frac{\sum_{i=1}^n Y_i T_i^k (1 - T_i)^{1-k} \mathcal{K}_{\hat{h}}\{\hat{B}^{\mathrm{T}}(X_i - x)\}}{\sum_{i=1}^n T_i^k (1 - T_i)^{1-k} \mathcal{K}_{\hat{h}}\{\hat{B}^{\mathrm{T}}(X_i - x)\}} \quad (k = 0, 1). \tag{7}$$

At the dimension reduction stage, we suggest using at least a fourth-order kernel function to ensure the large-sample properties of the estimated joint central subspace, as discussed in Remark 5. However, in practice the negative weights of a higher-order kernel can be detrimental to the stability of the resulting estimators. One possible way to obtain a more stable estimate is by using a second-order kernel function in (7) and substituting an adjusted bandwidth $\hat{h}^* = \hat{h}^{(2q+\hat{d})/(4+\hat{d})}$, so that the resulting convergence rate of $\hat{h}^*$ is optimal with respect to a mean integrated squared error based on a second-order kernel function. In our numerical experiments we have found that the finite-sample performance of (7) is much better if a second-order kernel and an adjusted bandwidth have been used.

To estimate the variance of the estimated conditional effects, an infinitesimal jackknife estimator can be applied. The idea is to perturb the empirical weight $1/n$ in the original estimator by a small amount ε and then take $\varepsilon \to 0$. More precisely, if we write $\hat{E}\{Y(k) \mid X = x\}$ as a function of n variables

$$Q_k(w_1, \ldots, w_n) = \frac{\sum_{i=1}^n w_i Y_i T_i^k (1 - T_i)^{1-k} \mathcal{K}_{\hat{h}}\{\hat{B}^{\mathrm{T}}(X_i - x)\}}{\sum_{i=1}^n w_i T_i^k (1 - T_i)^{1-k} \mathcal{K}_{\hat{h}}\{\hat{B}^{\mathrm{T}}(X_i - x)\}}$$

evaluated at $w = (w_1, \ldots, w_n) = (1/n, \ldots, 1/n)$, the infinitesimal jackknife estimator of variance is $n^{-2} \sum_{i=1}^n \hat{D}_{k,i}^2$, where $\hat{D}_{k,i}$ is the derivative of Q_k with respect to w evaluated at $(1/n, \ldots, 1/n)$, i.e., for $k = 0$ or 1,

$$\hat{D}_{k,i} = \frac{Y_i T_i^k (1 - T_i)^{1-k} \mathcal{K}_{\hat{h}}\{\hat{B}^{\mathrm{T}}(X_i - x)\}}{\sum_{i=1}^n T_i^k (1 - T_i)^{1-k} \mathcal{K}_{\hat{h}}\{\hat{B}^{\mathrm{T}}(X_i - x)\}} - T_k \frac{T_i^k (1 - T_i)^{1-k} \mathcal{K}_{\hat{h}}\{\hat{B}^{\mathrm{T}}(X_i - x)\}}{\sum_{i=1}^n T_i^k (1 - T_i)^{1-k} \mathcal{K}_{\hat{h}}\{\hat{B}^{\mathrm{T}}(X_i - x)\}}.$$

To estimate the variance of $\hat{E}\{Y(1) - Y(0) \mid X = x\}$, we can directly apply the infinitesimal jackknife estimator $n^{-2} \sum_{i=1}^n (\hat{D}_{1,i} - \hat{D}_{0,i})^2$.

Remark 7. The estimator (7) can be extended to estimate $E[g\{Y(k)\} \mid X = x]$ for real-valued functions g in the following way:

$$\hat{E}[g\{Y(k)\} \mid X = x] = \frac{\sum_{i=1}^n g(Y_i) T_i^k (1 - T_i)^{1-k} \mathcal{K}_{\hat{h}}\{\hat{B}^{\mathrm{T}}(X_i - x)\}}{\sum_{i=1}^n T_i^k (1 - T_i)^{1-k} \mathcal{K}_{\hat{h}}\{\hat{B}^{\mathrm{T}}(X_i - x)\}} \quad (k = 0, 1).$$

Model (1) guarantees that $E[g\{Y(k)\} \mid X = x] = E[g\{Y(k)\} \mid B_0^{\mathrm{T}} X = B_0^{\mathrm{T}} x]$ for an arbitrary function g.

3. Efficient estimation of average treatment effects

3·1. *Semiparametric efficiency bound and the efficient estimator*

One should note that σ^2_{eff} is the efficiency bound for the nonparametric model without any specification of the forms of $\pi(X)$, $F_0(y \mid X)$ and $F_1(y \mid X)$. Since model (1) imposes a multi-indices structure on these distribution functions, some efficiency gain might be expected. However, we have found that the dimension reduction structure (1) is ancillary for the estimation of τ, so knowledge of the joint central subspace does not reduce the asymptotic variance bound.

Theorem 2. *Under model* (1)*, the semiparametric efficiency bound of* τ *is* σ^2_{eff}.

According to Rosenbaum & Rubin (1983) and Hansen (2008), the average treatment effects τ can be consistently estimated through balancing scores or prognostic scores. In fact, if $b(X)$ satisfies

$$T \perp\!\!\!\perp \{Y(0), Y(1)\} \mid b(X),$$

then $\tau = E[E\{Y \mid T = 1, b(X)\} - E\{Y \mid T = 0, b(X)\}]$, and τ can be estimated by

$$\hat{\tau}_b = \frac{1}{n}\sum_{i=1}^{n}\big[\hat{m}_1\{b(X_i)\} - \hat{m}_0\{b(X_i)\}\big]$$

where

$$\hat{m}_k\{b(x)\} = \frac{\sum_{j=1}^{n} Y_j T_j^k (1 - Tj)^{1-k} \mathcal{K}_\varsigma\{b(X_j) - b(x)\}}{\sum_{j=1}^{n} T_j^k (1 - Tj)^{1-k} \mathcal{K}_\varsigma\{b(X_j) - b(x)\}}.$$

Since $\hat{m}_k\{b(x)\}$ is an estimator of $m_k\{b(x)\} = E\{Y \mid T = k, b(X) = b(x)\}$, we can follow the proof of Hahn (1998) for nonparametric imputation estimators and obtain the asymptotic variance of $\hat{\tau}_b$ as

$$E\left(\big[m_1\{b(X)\} - m_0\{b(X)\} - \tau\big]^2 + \frac{\sigma_1^2\{b(X)\}}{\pi\{b(X)\}} + \frac{\sigma_0^2\{b(X)\}}{1 - \pi\{b(X)\}}\right) \tag{8}$$

where $\sigma_k^2\{b(X)\} = \text{var}\{Y(k) \mid b(X)\}$, under some regularity conditions. Under model (1) and with $b(X) = B_0^{\mathrm{T}} X$, the asymptotic variance attains the semiparametric efficiency bound σ^2_{eff} of Hahn (1998).

A simple estimator of τ is

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\{\hat{m}_1(\hat{B}^{\mathrm{T}} X_i) - \hat{m}_0(\hat{B}^{\mathrm{T}} X_i)\},$$

where $\hat{B}$ is an estimator of B_0. In this study, we further show that the asymptotic variance of $\hat{B}$ does not affect the asymptotic behaviour of $\hat{\tau}$ and that $\hat{\tau}$ is semiparametrically efficient.

Although the estimator of the average treatment effect is an average of conditional treatment effects, one requires a different bandwidth to attain optimal undersmoothing. We first provide the

asymptotic distribution of the proposed estimator for a range of bandwidths satisfying Condition A6 in the Supplementary Material; then a data-adaptive method for choosing the bandwidth is discussed in § 3·2.

THEOREM 3. *Suppose that Assumption 1 and Conditions* A1–A6 *in the Supplementary Material are satisfied. Then* $n^{1/2}(\hat{\tau} - \tau) \to N(0, \sigma^2_{\text{eff}})$ *in distribution as* $n \to \infty$.

In practice, the semiparametric efficiency bound can be estimated by a direct plug-in estimator

$$\hat{\sigma}^2_{\text{eff}} = \frac{1}{n}\sum_{i=1}^{n}\left[\{\hat{m}_1(\hat{B}^{\mathrm{T}}X_i) - \hat{m}_0(\hat{B}^{\mathrm{T}}X_i) - \hat{\tau}\}^2 + \frac{\hat{\sigma}^2_1(\hat{B}^{\mathrm{T}}X_i)}{\hat{\pi}(\hat{B}^{\mathrm{T}}X_i)} + \frac{\hat{\sigma}^2_0(\hat{B}^{\mathrm{T}}X_i)}{1 - \hat{\pi}(\hat{B}^{\mathrm{T}}X_i)}\right]$$

where

$$\hat{\sigma}^2_k(B^{\mathrm{T}}x) = \frac{\sum_{i=1}^n T_i^k(1-T_i)^k Y_i^2 K_\varsigma\{B^{\mathrm{T}}(X_i - x)\}}{\sum_{i=1}^n T_i^k(1-T_i)^k K_\varsigma\{B^{\mathrm{T}}(X_i - x)\}} - \hat{m}^2_k(B^{\mathrm{T}}x) \quad (k = 0, 1).$$

The bandwidth ς can be replaced by that used to estimate $\hat{\tau}$.

Remark 8. A slight variation of $\hat{\tau}$ can be constructed in the spirit of Cheng (1994). Since the potential outcomes are only partially unobservable, Cheng (1994) suggested imputing an unobserved value with its conditional expectation. More precisely, the estimator is

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\big[T_iY_i + (1-T_i)\hat{m}_1\{b(X_i)\}\big] - \big[(1-T_i)Y_i + T_i\hat{m}_0\{b(X_i)\}\big]\Big).$$

By paralleling the proof of Cheng (1994), one can show that the asymptotic distribution of this estimator is the same as that of $\hat{\tau}$, and hence neither is better in general. In our simulation studies, we have found that this alternative estimator has slightly smaller bias than $\hat{\tau}$ and a very similar standard deviation.

3·2. *Bandwidth selection*

As indicated in Theorem 3, the bandwidth ς used in the nonparametric imputation should be smaller than the classical optimal bandwidth with rate $n^{-1/(2q+d)}$, so an important issue in practice is how to select a proper bandwidth. Häggström & de Luna (2014) suggested minimizing the conditional mean squared error of $\hat{\tau}$, which is of the form

$$E\left(\left[\hat{\tau} - \frac{1}{n}\sum_{i=1}^{n}\{m_1(B_0^{\mathrm{T}}X_i) - m_0(B_0^{\mathrm{T}}X_i)\}\right]^2 \;\middle|\; X_1, \ldots, X_n\right).$$

In our simulation experiments we have found that the bandwidth ς_k which minimizes the sample analogue of the conditional mean squared error

$$E\left[\left\{\frac{1}{n}\sum_{i=1}^{n}\hat{m}_k(B_0^{\mathrm{T}}X_i) - \frac{1}{n}\sum_{i=1}^{n}m_k(B_0^{\mathrm{T}}X_i)\right\}^2 \;\middle|\; X_1, \ldots, X_n\right] \quad (k = 0, 1) \tag{9}$$

leads to a slightly better estimator for τ. The main difference is that Häggström & de Luna (2014) used a local linear regression instead of a Nadaraya–Watson estimator to estimate the conditional effects. Since we directly adopt the local constant smoothing estimator and the estimated optimal bandwidth in the dimension reduction stage, a separate criterion might be helpful for alleviating the boundary effects of the Nadaraya–Watson estimator. Following the proof of Häggström & de Luna (2014), we can show that (9) is asymptotically equivalent to

$$\frac{1}{n}\int \frac{\sigma_k^2(B_0^{\mathrm{T}}x)f_X(x)}{\pi^k(B_0^{\mathrm{T}}x)\{1-\pi(B_0^{\mathrm{T}}x)\}^{1-k}}\,\mathrm{d}x + \frac{\sigma_K^2}{n^2\varsigma^{d_0}}\int \frac{\sigma_k^2(B_0^{\mathrm{T}}x)}{\pi^k(B_0^{\mathrm{T}}x)\{1-\pi(B_0^{\mathrm{T}}x)\}^{1-k}}\,\mathrm{d}x$$
$$+\ \varsigma^{2q}\left(\frac{\mu_{q,K}}{q!}\right)^2\left(\int \frac{D^q_{B^{\mathrm{T}}x}[m_k(B_0^{\mathrm{T}}x)\pi^k(B_0^{\mathrm{T}}x)\{1-\pi(B_0^{\mathrm{T}}x)\}^{1-k}f_{B_0^{\mathrm{T}}X}(B_0^{\mathrm{T}}x)]}{\pi^k(B_0^{\mathrm{T}}x)\{1-\pi(B_0^{\mathrm{T}}x)\}^{1-k}}\right.$$
$$\left.-\frac{m_k(B_0^{\mathrm{T}}x)D^q_{B^{\mathrm{T}}x}[\pi^k(B_0^{\mathrm{T}}x)\{1-\pi(B_0^{\mathrm{T}}x)\}^{1-k}f_{B_0^{\mathrm{T}}X}(B_0^{\mathrm{T}}x)]}{\pi^k(B_0^{\mathrm{T}}x)\{1-\pi(B_0^{\mathrm{T}}x)\}^{1-k}}(1_d,\ldots,1_d)\,\mathrm{d}x\right)^2,$$

where $\mu_{q,K} = \int v^q K(v)\,\mathrm{d}v$, $\sigma_K^2 = \int K^2(v)\,\mathrm{d}v$, $D^q f$ is the qth-order derivative of f in tensor form, $1_d = (1,\ldots,1)^{\mathrm{T}}$ with dimension d, and the optimal bandwidth is asymptotically equivalent to $O_{\mathrm{p}}\{n^{-2/(2q+d_0)}\}$. According to Condition A6 in the Supplementary Material, we require $q > 3d_0/2$ to ensure the asymptotic normality of $\hat{\tau}$ if the estimated optimal bandwidth is used. However, we find that the second-order kernel still works very well in practice. In addition, consistency can be guaranteed under the weaker assumptions that $h \to 0$ and $nh^{d_0} \to \infty$ as $n \to \infty$.

4. Application

In this section, we demonstrate our proposed method by applying it to the 2007–2008 National Health and Nutrition Examination Survey, the main goal of which was to investigate the health and nutrition statuses of children and adults in the United States. We focus on a subset of the data (Kohn et al., 2014) and study whether participation in the National School Lunch or School Breakfast programme would lead to an increase in body mass index for children and youths aged 4 to 17. The dataset contains 2330 children and youths, of whom 1284, 55%, participated in the school meal programme. The covariates are child age, X_1, child gender, X_2, black race, X_3, Hispanic race, X_4, family above 200% of the federal poverty level, X_5, participation in the Special Supplemental Nutrition Program for Women, Infants, and Children, X_6, participation in the Food Stamp Program, X_7, a childhood food security measurement, X_8, health insurance coverage, X_9, gender of survey respondent, X_{10}, and age of survey respondent, X_{11}.

The estimated structural dimension of the joint central subspace is 1 and the estimated linear index is shown in Table 1. Based on this balancing score, the estimated average difference in body mass index between participants and nonparticipants is 0·297 with a standard error of 0·206. Furthermore, the 95% confidence interval $(-0{\cdot}127, 0{\cdot}721)$ indicates an insignificant difference in body mass index between participants and nonparticipants, which is consistent with the conclusion that Chan et al. (2016) reached through weighting estimators. Therefore, participation in the school meal programme does not seem to be correlated with excessive food consumption.

Table 1. *Estimated linear index for the 2007–2008 National Health and Nutrition Examination Survey data*

	Estimate (SE)		Estimate (SE)
X_1	1		
X_2	0·006 (0·0025)	X_7	0·016 (0·0025)
X_3	0·004 (0·0021)	X_8	0·020 (0·0028)
X_4	0·031 (0·0039)	X_9	0·002 (0·0013)
X_5	−0·030 (0·0039)	X_{10}	0·029 (0·0015)
X_6	0·000 (0·0027)	X_{11}	0·965 (0·0002)

SE, standard error.

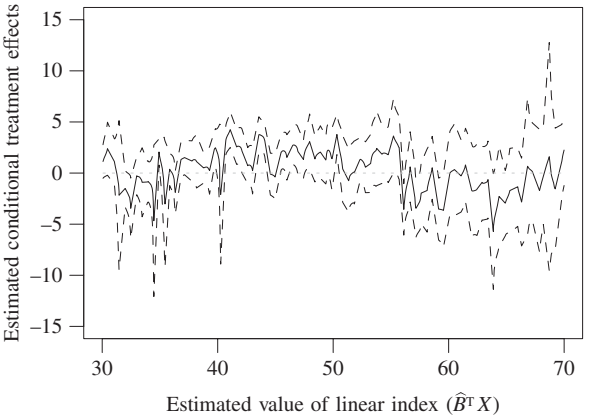

Fig. 1. Estimated difference in body mass index between participants and nonparticipants of the school meal programme plotted against the estimated linear index for the 2007–2008 National Health and Nutrition Examination Survey data. The solid line represents the estimated conditional effects and the dashed lines represent pointwise 95% confidence limits.

Figure 1 plots the estimated difference in body mass index between the two groups as a function of the estimated linear index. The standard errors are obtained using the infinitesimal jackknife. In general, the difference in average body mass index between the two groups is insignificant. However, the participants tend to have slightly higher body mass index than non-participants when the linear index lies between 40 to 50, which is a reason behind the slightly positive average treatment effects.

5. DISCUSSION

Recently Ma & Zhu (2013) introduced an efficient estimating equation, which is obtained through a likelihood approach, to achieve the semiparametric efficiency bound of the central subspace. However, the efficiency bound is derived under a fixed dimension and, in practice, the true structural dimension is unknown. Our proposed estimator estimates the structural dimension and the basis matrix simultaneously.

In observational studies, continuous treatments or exposures are also common. In the literature, a generalized propensity score has been introduced to estimate continuous treatment effects;

see, for example, Imbens (2000). Since our proposed model can adapt the propensity and outcome regression jointly, it would be interesting to extend this modelling approach to continuous treatment regimes.

Acknowledgement

The authors thank the editor, an associate editor, two reviewers, and Dr Mary Lou Thompson for their helpful comments and suggestions. The authors were partially supported by the National Heart, Lung, and Blood Institute of the U.S. National Institutes of Health.

Supplementary material

Supplementary material available at *Biometrika* online includes a comparison of several alternative estimation criteria for the joint central subspace, additional simulation results, and the proofs of Proposition 2 and Theorems 1–3.

References

Adragni, K. P., Cook, R. D. & Wu, S. (2012). Grassmannoptim: An R package for Grassmann manifold optimization. *J. Statist. Software* **50**, 1–18.
Chan, K. C. G., Yam, S. C. P. & Zhang, Z. (2016). Globally efficient nonparametric inference of average treatment effects by empirical balancing calibration weighting. *J. R. Statist. Soc.* B **78**, 673–700.
Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *J. Am. Statist. Assoc.* **89**, 81–7.
Cook, R. D. (1998). *Regression Graphics*. New York: Wiley.
Cook, R. D. & Li, B. (2002). Dimension reduction for conditional mean in regression. *Ann. Statist.* **30**, 455–74.
de Luna, X., Waernbaum, I. & Richardson, T. S. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika* **98**, 861–75.
Edelman, A., Arias, T. A. & Smith, S. T. (1999). The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.* **20**, 303–53.
Fukumizu, K. & Leng, C. (2014). Gradient-based kernel dimension reduction for regression. *J. Am. Statist. Assoc.* **109**, 359–70.
Ghosh, D., Zhu, Y. & Coffman, D. L. (2015). Penalized regression procedures for variable selection in the potential outcomes framework. *Statist. Med.* **34**, 1645–58.
Häggström, J. & de Luna, X. (2014). Targeted smoothing parameter selection for estimating average causal effects. *Comp. Statist.* **29**, 1727–48.
Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66**, 315–31.
Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika* **95**, 481–8.
Hirano, K., Imbens, G. W. & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71**, 1161–89.
Hu, Z., Follmann, D. A. & Wang, N. (2014). Estimation of mean response via the effective balancing score. *Biometrika* **101**, 613–24.
Huang, M.-Y. & Chiang, C. T. (2017). An effective semiparametric estimation approach for the sufficient dimension reduction model. *J. Am. Statist. Assoc.* to appear, doi:10.1080/01621459.2016.1215987.
Imai, K. & Ratkovic, M. (2014). Covariate balancing propensity score. *J. R. Statist. Soc.* B **76**, 243–63.
Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87**, 706–10.
Kohn, M. J., Bell J. F. Grow, M. G. & Chan, K. C. G. (2014). Food insecurity, food assistance and weight status in US youth: New evidence from NHANES 2007–08. *Pediatric Obesity* **9**, 155–66.
Leacy, F. P. & Stuart, E. A. (2014). On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: A simulation study. *Statist. Med.* **33**, 3488–508.
Li, B. & Wang, S. (2007). On directional regression for dimension reduction. *J. Am. Statist. Assoc.* **102**, 997–1008.
Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with Discussion). *J. Am. Statist. Assoc.* **86**, 316–42.
Ma, Y. & Zhu, L. (2012). A semiparametric approach to dimension reduction. *J. Am. Statist. Assoc.* **107**, 168–79.
Ma, Y. & Zhu, L. (2013). Efficient estimation in sufficient dimension reduction. *Ann. Statist.* **41**, 250–68.

QIN, J. & ZHANG, B. (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *J. R. Statist. Soc.* B **69**, 101–22.
ROBINS, J. M., ROTNITZKY, A. & ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.* **89**, 846–86.
ROSENBAUM, P. R. & RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.
VANSTEELANDT, S., BEKAERT, M. & CLAESKENS, G. (2012). On model selection and model misspecification in causal inference. *Statist. Meth. Med. Res.* **21**, 7–30.
WANG, H. & XIA, Y. (2008). Sliced regression for dimension reduction. *J. Am. Statist. Assoc.* **103**, 811–21.
XIA, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *Ann. Statist.* **35**, 2654–90.
XIA, Y. (2008). A multiple-index model and dimension reduction. *J. Am. Statist. Assoc.* **103**, 1631–40.
YIN, X. & LI, B. (2011). Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *Ann. Statist.* **39**, 3392–416.
ZHU, L. P., ZHU, L. X. & FENG, Z. H. (2010). Dimension reduction in regressions through cumulative slicing estimation. *J. Am. Statist. Assoc.* **105**, 1455–66.
ZHU, Y. & ZENG, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *J. Am. Statist. Assoc.* **101**, 1638–51.

[*Received on* 17 *January* 2016. *Editorial decision on* 9 *April* 2017]