Semiparametric dimension reduction estimation for mean response with missing data

BY ZONGHUI HU, DEAN A. FOLLMANN AND JING QIN

*Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases,
National Institutes of Health, Maryland 20892-7609, U.S.A.*

huzo@niaid.nih.gov dfollmann@niaid.nih.gov jqin@niaid.nih.gov

SUMMARY

Model misspecification can be a concern for high-dimensional data. Nonparametric regression obviates model specification but is impeded by the curse of dimensionality. This paper focuses on the estimation of the marginal mean response when there is missingness in the response and multiple covariates are available. We propose estimating the mean response through nonparametric functional estimation, where the dimension is reduced by a parametric working index. The proposed semiparametric estimator is robust to model misspecification: it is consistent for any working index if the missing mechanism of the response is known or correctly specified up to unknown parameters; even with misspecification in the missing mechanism, it is consistent so long as the working index can recover $E(Y|X)$, the conditional mean response given the covariates. In addition, when the missing mechanism is correctly specified, the semiparametric estimator attains the optimal efficiency if $E(Y|X)$ is recoverable through the working index. Robustness and efficiency of the proposed estimator is further investigated by simulations. We apply the proposed method to a clinical trial for HIV.

Some key words: Dimension reduction; Inverse probability weighting; Kernel regression; Missing at random; Robustness to model misspecification.

1. INTRODUCTION

With advances in technology, increasingly high dimensional variables are observable. This provides more information for exploring research goals, but imposes a substantial challenge for model specification. It is worth developing new statistical methods that make effective use of the data and are robust to model misspecification.

This paper is motivated by the problem of mean response estimation where missingness occurs to the response but multiple explanatory variables are observable. We consider a random sample (X_i, Y_i, δ_i) for $i = 1, \dots, n$, where Y is the response, $\delta_i = 1$ if Y_i is observed and $\delta_i = 0$ if Y_i is missing. The covariate X is always observed. The parameter of interest is the marginal mean $\theta = E(Y)$. Such a missing pattern can arise in different situations, one of which is the two-stage sampling scheme (Bose, 1943; White, 1982): when it is not feasible to obtain enough observations of Y , more observations are taken of the covariate X to make up for the lack of information on Y .

Regression is most efficient for mean response estimation if the relationship between Y and X is correctly specified (Yates, 1933; Matloff, 1981). It estimates $E(Y)$ by $n^{-1} \sum_{i=1}^n q(X_i)$ with q a parametric function for $E(Y|X)$, which can be inconsistent if q is misspecified.

One way to avoid model specification is through nonparametric regression (Cheng, 1994), which directly estimates $M(X) = E(Y|X)$ by nonparametric regression without specifying its

parametric form and estimates $E(Y)$ by $\hat{\theta}_{\text{NR}} = n^{-1} \sum_{i=1}^n \hat{M}(X_i)$, with $\hat{M}(\cdot)$ the nonparametric estimate of $M(\cdot)$. The advantage of $\hat{\theta}_{\text{NR}}$ is that it does not require any model specification under an ignorable missing mechanism. This nonparametric regression approach has been successfully adopted by [Aerts et al. \(2002\)](#), [Wang & Rao \(2002a, b\)](#) and [Wang & Dai \(2008\)](#), among others. However, as noted by [Aerts et al. \(2002\)](#), when the covariate is of high dimension, this approach requires multivariate nonparametric regression and is impeded by the curse of dimensionality ([Härdle et al., 2004](#)). In the following, $X \in \mathbb{R}^d$ represents a multi-dimensional covariate.

Throughout this paper, we will focus on the missing at random ([Rosenbaum & Rubin, 1983](#)) mechanism for nonresponse Y . That is, given the covariate X , missingness in Y is independent of the value of Y ,

$$\text{pr}(\delta = 1 \mid Y, X) = \text{pr}(\delta = 1 \mid X) = \pi(X), \quad (1)$$

where π is the selection probability. There are two main methods for dealing with data missing at random. One is inverse probability weighting estimation ([Horvitz & Thompson, 1952](#)), which estimates $E(Y)$ by $\hat{\theta}_{\text{IPW}} = n^{-1} \sum_{i=1}^n \delta_i Y_i / \pi(X_i)$. The other is augmented inverse probability weighting estimation ([Robins et al., 1994](#)), which estimates $E(Y)$ by

$$\hat{\theta}_{\text{AIPW}} = n^{-1} \sum_{i=1}^n \frac{\delta_i Y_i}{\pi(X_i)} - \frac{\delta_i - \pi(X_i)}{\pi(X_i)} q(X_i),$$

with $q(\cdot)$ a parametric function for the conditional mean $E(Y \mid \cdot)$. Performance of both estimators depends on correct model specification: $\hat{\theta}_{\text{IPW}}$ is consistent if π is correctly specified; $\hat{\theta}_{\text{AIPW}}$ is consistent if either π or q is correctly specified, and achieves optimal efficiency when both are correctly specified.

In this paper, we pursue a semiparametric approach, which reconciles the nonparametric regression approach and the model-based approaches. We adopt nonparametric regression to relax dependence on model specification while using prior model information about $E(Y \mid X)$ to improve efficiency. As the covariate contains information about the response, we use a parametric working index $S = S(X)$ to summarize this information. The conditional mean of Y given S is denoted by $m(S) = E(Y \mid S)$. As $E(Y) = E\{E(Y \mid S)\}$ for arbitrary S , $n^{-1} \sum_{i=1}^n m(S_i)$ is unbiased for $\theta = E(Y)$, where $m(S)$ is to be estimated by univariate nonparametric regression of Y versus S . In the following, we refer to S as the working index, $S(\cdot)$ as the working index function and denote this semiparametric dimension reduction estimator by $\hat{\theta}$. Compared to the nonparametric regression approach, use of a working index reduces the dimension in nonparametric regression from d to 1 and avoids the curse of dimensionality. Compared to the two model-based approaches, the semiparametric dimension reduction estimator has an extended robustness to model misspecification: $\hat{\theta}$ is consistent when either the selection probability is correctly specified or the working index is an atom of $E(Y \mid X)$ as defined below; it attains optimal efficiency when both are satisfied.

DEFINITION 1. *We say that $E(Y \mid X)$ is recoverable through $S = S(X)$ if $E(Y \mid X) = g(S)$ for some function g . We refer to $S = S(X)$ as an atom of $E(Y \mid X)$.*

2. SEMIPARAMETRIC DIMENSION REDUCTION ESTIMATION

Denote S as a continuous function from $\mathbb{R}^d$ to $\mathbb{R}$, such that $S = S(X)$ is univariate and $S_i = S(X_i)$. As $E\{E(Y \mid S)\} = E(Y)$, we have $E(Y) = E\{m(S)\}$ with $m(S) = E(Y \mid S)$, and

$$\hat{\theta} = n^{-1} \sum_{i=1}^n \hat{m}(S_i) \quad (2)$$

is an estimate of $\theta = E(Y)$, where $\hat{m}(s)$ is the estimate of $m(s)$ from nonparametric regression.

The transformation $\mathcal{S}: X \rightarrow \mathcal{S}(X)$ is a condensation of the information about Y contained in X to a single index, so as to reduce the dimension in the nonparametric regression for $m(\cdot)$. This idea looks similar to single index modelling (Ichimura, 1993), but the requirements for $\mathcal{S}$ are different. For single index modelling, the objective is to model the conditional mean $E(Y | X)$, and $\mathcal{S}$ has to satisfy the recoverability condition, $E(Y | X) = g\{\mathcal{S}(X)\}$ for some function g , to get consistent estimation. Semiparametric dimension reduction estimation (2) is consistent for any index S , as its objective is to estimate the marginal mean $E(Y)$.

We shall focus on the Nadaraya–Watson kernel regression (Nadaraya, 1964; Watson, 1964) and the local linear smoother (Fan, 1992) for nonparametric regression. By using the single index $S = \mathcal{S}(X)$ instead of the original covariate X , we estimate $m(S) = E(Y | S)$ by univariate nonparametric regression instead of $M(X) = E(Y | X)$ by multivariate nonparametric regression. However, we may lose the missing at random property; that is, $\text{pr}(\delta = 1 | y, s) = \text{pr}(\delta = 1 | s)$ is not ensured by (1). Accordingly, $m(s) = E(Y | s)$ may be different from $E(Y | s, \delta = 1)$. The nonparametric estimation of $m(s)$ cannot be simply performed over the complete cases $\{(X_i, Y_i) : \delta_i = 1\}$ but needs to account for the missingness.

One way to deal with missingness in nonparametric regression is via the inverse probability weighted local linear smoother (Wang et al., 1998). At each s , it fits a weighted local linear regression centred at s with $\alpha = (\alpha_0, \alpha_1)^\top$ minimizing

$$\sum_{j=1}^n \frac{\delta_j}{\pi(X_j)} K_h(S_j - s) \{Y_j - \alpha_0 - \alpha_1(S_j - s)\}^2, \quad (3)$$

where $\alpha = \alpha(s)$ varies with s , $K_h(u) = h^{-1}K(u/h)$, K is the kernel function and is generally a symmetric density function and $h = h_n$ is the smoothing bandwidth. The estimate of $m(s)$ is $\hat{m}(s) = \hat{\alpha}_0$. At each s , α can be estimated from (3) through a weighted least-squares regression, which gives $\hat{m}(s) = \sum_{j=1}^n w_j y_j / \sum_{j=1}^n w_j$, with $w_j = \{\delta_j / \pi(X_j)\} K_h(S_j - s) \{A_{n,2} - A_{n,1}(S_j - s)\}$ and $A_{n,l} = n^{-1} \sum_{j=1}^n \{\delta_j / \pi(X_j)\} K_h(S_j - s) (S_j - s)^l$ for $l = 1, 2$.

If we restrict $\alpha_1 = 0$ in (3), we get the inverse probability weighted Nadaraya–Watson estimate of $m(s)$ as

$$\hat{m}(s) = \sum_{j=1}^n \frac{\delta_j}{\pi(X_j)} K_h(S_j - s) Y_j \bigg/ \sum_{j=1}^n \frac{\delta_j}{\pi(X_j)} K_h(S_j - s). \quad (4)$$

The semiparametric dimension reduction estimator of $E(Y)$ is then given by (2) with $\hat{m}(\cdot)$ from the inverse probability weighted nonparametric regression (3) or (4). When the selection probability is unknown, it is estimated by a working model $w(X)$ and $w(X)$ is used instead of $\pi(X)$. This will be illustrated in the later sections.

3. ASYMPTOTIC PROPERTIES

Let f_X be the density function of X and f_S the density of S . We assume the following regularity conditions.

Condition 1. The kernel function K is a symmetric density such that $\int u^2 K(u) du < \infty$.

Condition 2. The function $\pi(x)$ is bounded away from 0 for x in the support of f_X .

Condition 3. The function $f_X(x)$ is bounded away from 0 for x in the support of f_X .

In addition, we impose a restriction on the index function: $\mathcal{S}$ is continuous with $(\partial \mathcal{S} / \partial x_1, \dots, \partial \mathcal{S} / \partial x_d)^\top \neq 0$ for $x = (x_1, \dots, x_d)^\top$ in the support of f_X . Under this restriction

and Condition 3, $f_S(s)$ is well defined and bounded away from 0 for $s = \mathcal{S}(x)$ with x in the support of f_X .

We focus on the inverse probability weighted local linear smoother for nonparametric estimation. We first consider the case where the selection probability π is known.

LEMMA 1. *Under Conditions 1–3 as $n \rightarrow \infty$, $h \rightarrow 0$ and $nh \rightarrow \infty$, the inverse probability weighted local linear estimator of $m(s) = E(Y | S = s)$ has an asymptotic normal distribution $(nh)^{1/2}\{\hat{m}(s) - m(s) - h^2 B(s)\} \rightarrow N\{0, \sigma^2(s)\}$ with*

$$B(s) = \frac{1}{2}m''(s)\gamma_K + o(1), \quad \sigma^2(s) = \frac{\text{var}(Y | S = s)}{f_S(s)} E \left\{ \frac{1}{\pi(X)} \middle| \mathcal{S}(X) = s \right\} \tau_K + o(1).$$

Here $m''(\cdot)$ stands for the second derivative of $m(\cdot)$, γ_K and τ_K depend only on the kernel function K . Under Condition 1, $\gamma_K = \int u^2 K(u) du$ and $\tau_K = \int K^2(u) du$ are both finite. Accordingly, the semiparametric dimension reduction estimator is asymptotically normal.

THEOREM 1. *As $n \rightarrow \infty$, $n^{1/2}(\hat{\theta} - \theta - h^2 \mathcal{B})$ is asymptotically normally distributed with mean zero and variance $\text{var}\{m(S)\} + E\{\text{var}(Y | S)/\pi(X)\} + (nh)^{-1}\mathcal{V}$, with*

$$\mathcal{B} = \frac{1}{2}E\{m''(S)\}\gamma_K + o(1), \quad \mathcal{V} = E[\text{var}(Y | S)\{f_S(S)\pi(X)\}^{-1}]\tau_K + o(1). \quad (5)$$

The estimator $\hat{\theta}$ is consistent for any index S . The asymptotic bias $h^2 \mathcal{B}$ comes from nonparametric estimation of the conditional mean $m(\cdot) = E(Y | S = \cdot)$, which depends on the curvature of $m(\cdot)$ through its second- and higher-order derivatives. This asymptotic bias disappears if $m''(\cdot) \equiv 0$, as the $o(1)$ term is from higher-order derivatives of $m(\cdot)$. If $S = \mathcal{S}(X) = E(Y | X)$, then $m(\cdot)$ is simply the identity function

$$m(S) = E(Y | S) = E\{E(Y | X) | S\} = E(S | S) = S,$$

and $\hat{\theta}$ is not only consistent but asymptotically unbiased. In fact, any index function $S(x) = aE(Y | x) + b$, for any constant a and b with $a \neq 0$, shares this property.

For the nonparametric regression estimator $\hat{\theta}_{\text{NR}}$, a similar derivation gives the asymptotic bias as $(1/2)E\{h^\top \mathcal{H}_M(X)h\}\gamma_K$, where $h \in \mathbb{R}^d$ is the vector of bandwidths in multivariate nonparametric regression for estimating $M(X) = E(Y | X)$, and $\mathcal{H}_M$ is the Hessian matrix of $M(X)$, which consists of its second-order derivatives. This asymptotic bias is generally not zero unless $E(Y | X)$ is a linear function of X . Therefore, in applications where Y is nonlinearly related to X , with careful modelling for the working index so that $m(S) = E(Y | S)$ is close to linear, the semiparametric dimension reduction estimator improves over the nonparametric regression estimator in terms of asymptotic bias.

With inverse probability weighted Nadaraya–Watson (4) estimation of $m(\cdot)$, the asymptotic distribution of $\hat{\theta}$ is the same as in Theorem 1 except that its bias involves the first derivative of $m(\cdot)$ and equals $\mathcal{B}h^2 + 2E\{m'(S)f'(S)/f(S)\}h^2 + o(h^2)$. This bias does not vanish even for $m(\cdot)$ linear.

Remark 1. In most cases, working index $S = \mathcal{S}(X; \gamma)$ involves an unknown parameter vector γ . With maximum likelihood estimation of γ , $\hat{\gamma} - \gamma = O_p(n^{-1/2})$ and the semiparametric dimension reduction estimator has the same asymptotic normality as in Theorem 1.

Remark 2. The asymptotic variance consists of two parts. The first part $n^{-1}[\text{var}\{m(S)\} + E\{\text{var}(Y | S)/\pi(X)\}]$ reduces to $n^{-1}\text{var}(Y)$ if $\pi = 1$, and greater than $n^{-1}\text{var}(Y)$ if otherwise, which is the inevitable loss in efficiency due to missingness in Y . The second part

Table 1. *The asymptotic variances for estimating $\theta = E(Y)$ with π known*

$n \text{ var}(\hat{\theta}_{\text{IPW}})$	$\text{var}(Y) + E[\{\pi(X)^{-1} - 1\}E(Y^2 X)]$
$n \text{ var}(\hat{\theta}_{\text{AIPW}})$	$\text{var}(Y) + E[\{\pi(X)^{-1} - 1\}\text{var}(Y X)] + E[\{\pi(X)^{-1} - 1\}\{E(Y X) - q(X)\}^2]$
$n \text{ var}(\hat{\theta})$	$\text{var}(Y) + E[\{\pi(X)^{-1} - 1\}\text{var}(Y S)] + O\{(nh)^{-1}\}$
$n \text{ var}(\hat{\theta}_{\text{NR}})$	$\text{var}(Y) + E[\{\pi(X)^{-1} - 1\}\text{var}(Y X)] + O\{(nh^d)^{-1}\}$

$\hat{\theta}_{\text{IPW}}$, inverse probability weighting estimator; $\hat{\theta}_{\text{AIPW}}$, augmented inverse probability weighting estimator; $\hat{\theta}$, semiparametric dimension reduction estimator; $\hat{\theta}_{\text{NR}}$, nonparametric regression estimator.

$(n^2h)^{-1}E[\text{var}(Y | S)\{f_S(S)\pi(X)\}^{-1}]\tau_K$ comes from nonparametric estimation of $m(\cdot)$ and diminishes as $n^2h \rightarrow \infty$. Therefore, the asymptotic variance is determined by the first part, which can be rewritten as $n^{-1}(\text{var}(Y) + E[\{\pi(X)^{-1} - 1\}\text{var}(Y | S)])$. Table 1 presents the asymptotic variances of the estimators discussed in this paper.

When $E(Y | X)$ is correctly specified and used for $S(X)$ and $q(X)$, both $\hat{\theta}$ and $\hat{\theta}_{\text{AIPW}}$ reach the same optimal efficiency $n^{-1}(\text{var}(Y) + E[\{\pi(X)^{-1} - 1\}\text{var}(Y | X)])$. Both are more efficient than $\hat{\theta}_{\text{IPW}}$, which uses only the model information for π and only the data information from the completely observed subjects, while $\hat{\theta}$ and $\hat{\theta}_{\text{AIPW}}$ additionally use the model information for $E(Y | X)$ and the covariate information from all the subjects. The estimators $\hat{\theta}$ and $\hat{\theta}_{\text{AIPW}}$ are also more efficient than the nonparametric regression estimator for large d .

When there is insufficient information to fully specify the parametric form of $E(Y | X)$, any misspecification in q adversely affects the efficiency of $\hat{\theta}_{\text{AIPW}}$, causing an increase in the asymptotic variance by $E[\{\pi(X)^{-1} - 1\}\{E(Y | X) - q(X)\}^2]$. However, for $\hat{\theta}$, so long as the index captures the core information about $E(Y | X)$ in the sense that $S(X)$ is an atom of $E(Y | X)$, it holds that $E[\{\pi(X)^{-1} - 1\}\text{var}(Y | S)] = E[\{\pi(X)^{-1} - 1\}\text{var}(Y | X)]$ and the optimal efficiency is maintained. Thus, $\hat{\theta}$ is less demanding than $\hat{\theta}_{\text{AIPW}}$ of the specification in $E(Y | X)$ regarding asymptotic efficiency.

When the selection probability $\pi(x; \beta)$ is correctly specified but with unknown parameter β , a maximum likelihood estimation based on binomial log likelihood $\sum_{i=1}^n \delta_i \log\{\pi(x_i; \beta)\} + (1 - \delta_i) \log\{1 - \pi(x_i; \beta)\}$ satisfies $\hat{\beta} - \beta = o_p(n^{-1/2}) = o_p(1)$. The following corollary studies the semiparametric dimension reduction estimator $\hat{\theta}_{\hat{\pi}}$ with the estimated selection probability $\hat{\pi} = \pi(x; \hat{\beta})$.

COROLLARY 1. *If the estimation of β in the correctly specified selection probability $\pi(x; \beta)$ satisfies the condition $\hat{\beta} - \beta = o_p(1)$, $\hat{\theta}_{\hat{\pi}}$ has the same asymptotic normality as in Theorem 1.*

Once the index is chosen, the optimal bandwidth h should minimize the mean squared error. From Theorem 1,

$$E(\hat{\theta} - \theta)^2 \approx h^4 E^2\{m''(S)\}\gamma_K^2/4 + (n^2h)^{-1}E[\text{var}(Y | S)\{f_S(s)\pi(X)\}^{-1}]\tau_K + n^{-1}[\text{var}\{m(S)\} + E\{\text{var}(Y | S)/\pi(X)\}].$$

The optimal bandwidth corresponds to $\partial E(\hat{\theta} - \theta)^2/\partial h = 0$, which gives

$$h_{\text{opt}} = (E[\text{var}(Y | S)\{f_S(S)\pi(X)\}^{-1}]\tau_K[E\{m''(S)\}\gamma_K]^{-2})^{1/5}n^{-2/5}.$$

We can estimate $m''(s)$ and $\text{var}(Y | s)$ by fitting a third- or higher-degree polynomial regression for $m(s)$, estimate $f_S(s)$ by kernel density estimation, and estimate $\pi(x)$ by logistic regression. We then apply the plug-in method to estimate h_{opt} . The plug-in method has been shown to be stable and efficient (Sheather & Jones, 1991; Fan & Marron, 1992) relative to data-driven bandwidth

selection. Simulation studies indicate that the semiparametric dimension reduction estimator is not very sensitive to bandwidth, which makes bandwidth selection not critical.

In Theorem 1 and Corollary 1, we have studied the semiparametric dimension reduction estimator with the selection probability known or correctly specified. Next, we investigate its robustness to model misspecification. Let $w(x)$ denote the working selection probability, which may be different from $\pi(x)$.

THEOREM 2. *The estimator $\hat{\theta}$ has the following double robustness to model misspecification:*
 (i) *when w is correctly specified for the selection probability, $\hat{\theta}$ is consistent for any index;*
 (ii) *when w is misspecified for the selection probability, $\hat{\theta}$ is consistent if the index is an atom of $E(Y | X)$.*

Consistency under condition (i) follows from Theorem 1 and Corollary 1. For the consistency under (ii), we give a brief proof under the inverse probability weighted Nadaraya–Watson nonparametric estimation of $m(S) = E(Y | S)$. The proof under the inverse probability weighted local linear nonparametric estimation is similar but notationally complicated.

Proof of Theorem 2. At any given s , a nonparametric estimate of $m(s)$ is given by (4) with $\pi(X_j)$ replaced by $w(X_j)$. By the law of large numbers, the numerator of (4) converges almost surely to

$$E \left\{ \frac{\delta Y}{w(X)} K_h(S - s) \right\} = E \left\{ \frac{\pi(X)E(Y | X)}{w(X)} K_h(S - s) \right\}, \quad (6)$$

and the denominator converges almost surely to

$$E \left\{ \frac{\pi(X)}{w(X)} K_h(S - s) \right\} = f_S(s) E \left\{ \frac{\pi(X)}{w(X)} \middle| S(X) = s \right\} + O(h^2). \quad (7)$$

Under condition (ii), $E(Y | X) = g(S)$ for some function g , so the numerator (6) becomes

$$E \left\{ \frac{\pi(X)g(S)}{w(X)} K_h(S - s) \right\} = f_S(s)g(s) E \left\{ \frac{\pi(X)}{w(X)} \middle| S(X) = s \right\} + O(h^2).$$

Therefore, $\hat{m}(s)$ converges almost surely to $g(s)$, where $m(s) = E(Y | s) = E\{E(Y | X) | s\} = g(s)$. In other words, $\hat{m}(s)$ is consistent for $m(s)$. Consequently, $\hat{\theta}$ is consistent for $E\{m(S)\} = E(Y)$. $\square$

In summary, when π is known or can be correctly specified, there is little concern about the selection of S , as $\hat{\theta}$ is consistent under any working index. However, it is most efficient if $S(X)$ is an atom of $E(Y | X)$. When there is insufficient information to ensure the correct specification of π , specification of the working index is crucial but not unique: $\hat{\theta}$ is consistent so long as $S(X)$ is an atom of $E(Y | X)$. Therefore, the best choice for the working index is an atom of $E(Y | X)$. The working index does not need to be fully specified for $E(Y | X)$, as the nonparametric regression of $m(S)$ further picks up the unknown link g , the remaining information about the response not covered by the index. For example, if the conditional mean response is $E(Y | X) = \beta_0 + \beta_1 \exp(\alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3)$, then $S(X) = \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3$ is an atom of $E(Y | X)$ and a candidate for the index. In fact, for any one-to-one transformation $\mathcal{L}$, $\mathcal{L} \circ S$ is also a candidate for the index, as $\mathcal{L} \circ S(X)$ is again an atom of $E(Y | X)$. Thus, specification of the complete parametric form of $E(Y | X)$ is not necessary.

As a special case, $E(Y | X)$ is an atom of itself. Therefore, as in the regression method and the augmented inverse probability weighting estimation procedure, a practical way to find a working index is to target a working model for $E(Y | X)$. However, it is rarely feasible to have an exact

Table 2. Estimation of $\theta = E(Y)$: the selection probability is known in I, correctly specified in II and misspecified in III. In (a) $q(X) = S(X)$ is correctly specified for $E(Y | X)$, in (b) is approximately an atom of $E(Y | X)$ and in (c) is function of only X_1

$w(X)$			$\hat{\theta}_{\text{IPW}}$	$\hat{\theta}_{\text{AIPW}}$			$\hat{\theta}$		
				(a)	(b)	(c)	(a)	(b)	(c)
I.	$\pi(X)$	bias	-0.14	0.02	-0.06	-0.10	-0.02	0.04	-0.01
		SD	1.61	0.50	0.87	1.04	0.48	0.52	0.97
		MSE	2.61	0.25	0.77	1.09	0.23	0.26	0.93
II.	$\pi(X; \hat{\beta})$	bias	-0.20	0.02	-0.07	-0.10	-0.02	0.04	-0.01
		SD	1.22	0.49	0.78	0.97	0.48	0.51	0.96
		MSE	1.53	0.24	0.61	0.95	0.23	0.26	0.91
III.	misspecified	bias	4.04	0.01	3.21	2.88	-0.01	0.05	0.33
		SD	14.53	0.50	11.31	8.81	0.48	0.50	1.52
		MSE	227.04	0.25	138.00	85.74	0.23	0.26	2.42

SD, standard deviation; MSE, mean squared error.

specification for $E(Y | X)$. In the following simulations, we investigate the performance of $\hat{\theta}$ where $E(Y | X)$ is misspecified, so that $S(X; \hat{\gamma})$ is only an approximate atom of $E(Y | X)$. The simulations indicate that $\hat{\theta}$ is quite tolerant of such model misspecification.

4. SIMULATION STUDIES

We study the numerical performance of the semiparametric dimension reduction estimator $\hat{\theta}$. In the first simulation, we compare it with the inverse probability weighting estimator $\hat{\theta}_{\text{IPW}}$ and the augmented inverse probability weighting estimator $\hat{\theta}_{\text{AIPW}}$ in terms of efficiency and robustness. In the second simulation, we additionally compare $\hat{\theta}$ with the nonparametric regression estimator $\hat{\theta}_{\text{NR}}$ in the setting of high-dimensional covariates. Each simulation is repeated 1000 times with $n = 500$.

In the first simulation, the response Y is simulated from $Y | X \sim N\{(X_1 + X_2)^3, 1\}$, with $X = (X_1, X_2)^T$, and X_1 and X_2 from independent $N(0, 1)$. The selection probability is $\pi(X) = 1 - 1/\{1 + \exp(1 - 3X_1 + 2X_2^2)\}$. To investigate how the working selection probability affects the estimation, three working models are considered: (I) $w(X) = \pi(X)$ is known; (II) $w(X) = \pi(X; \beta)$ is correctly specified with unknown β ; and (III) $w(X)$ is misspecified as logit $\{w(X)\} = (1, X)^T \beta$. Additionally, three working models $S(X; \gamma)$ for $E(Y | X)$ are considered: $S(X; \gamma)$ is a cubic function of X_1 and X_2 ; $S(X; \gamma)$ is a linear function of X_1 and X_2 ; and $S(X; \gamma)$ is a cubic function of only X_1 . Maximum likelihood estimation is applied for β , and a complete case maximum likelihood estimation is applied for γ . For the augmented inverse probability weighting estimation, we take $q(X) = S(X)$; that is, we evaluate $\hat{\theta}_{\text{AIPW}}$ and $\hat{\theta}$ assuming the same knowledge about $E(Y | X)$. Thus $q = S$ is correctly specified for $E(Y | X)$ by the first working index but misspecified by the other two working indices.

Table 2 presents the Monte Carlo bias, standard deviation and mean squared error. Below are observations under the three working selection probabilities.

Under case I where $w = \pi$ is known, all three estimators are consistent and have negligible biases. The estimator $\hat{\theta}_{\text{IPW}}$ has substantially larger standard deviation than $\hat{\theta}_{\text{AIPW}}$ and $\hat{\theta}$, the rationale for which was discussed in Remark 2. When $q(X) = S(X)$ is correctly specified for $E(Y | X)$, both $\hat{\theta}_{\text{AIPW}}$ and $\hat{\theta}$ should achieve the same optimal efficiency according to Remark 2, and we observe small standard deviations for both. When $q(X) = S(X)$ is misspecified for $E(Y | X)$, $\hat{\theta}_{\text{AIPW}}$ has a big increase in standard deviation as compared to that under correct

Table 3. Estimation of $\theta = E(Y)$: the selection probability is correctly specified in I and misspecified in II; $q(X) = S(X)$ is correctly specified for $E(Y | X)$ in (a) and an approximate atom of $E(Y | X)$ in (b)

$w(X)$			$\hat{\theta}_{IPW}$	$\hat{\theta}_{AIPW}$		$\hat{\theta}$		$\hat{\theta}_{NR}$
				(a)	(b)	(a)	(b)	
I:	$\pi(X, \hat{\beta})$	Bias	-0.05	-0.05	-0.69	-0.54	0.60	0.61
		SD	5.34	5.34	10.06	3.19	3.16	7.04
		MSE	28.48	28.45	101.63	10.49	10.31	49.91
II:	Misspecified	Bias	6.47	-0.05	-17.03	-0.79	0.57	0.61
		SD	8.37	5.34	11.50	3.33	3.20	7.04
		MSE	111.83	28.49	422.16	11.72	10.54	49.91

SD, standard deviation; MSE, mean squared error.

specification, which is due to the fact that $\hat{\theta}_{AIPW}$ reaches the optimal efficiency only when $q(X)$ is correctly specified for $E(Y | X)$ and is less efficient otherwise. For $\hat{\theta}$, with $S(X)$ linear in X_1 and X_2 , $S(X)$ is misspecified for $E(Y | X)$ but roughly recovers $E(Y | X)$ through a cubic function, $\hat{\theta}$ thus has a close to optimal efficiency in accordance with Remark 2. This last observation shows the robustness of $\hat{\theta}$ to index specification in terms of optimal efficiency.

Under case II where w is correctly specified with β estimated, all estimators are consistent. Based on Corollary 1, estimation of the selection probability does not affect the efficiency of $\hat{\theta}$, so we observe similar standard deviations in II as in I. However, estimation of the selection probability improves the efficiency of both $\hat{\theta}_{IPW}$ and $\hat{\theta}_{AIPW}$ when $q(X)$ is misspecified; a counter-intuitive phenomenon discussed in [Robins et al. \(1994\)](#).

Under case III where w is misspecified, $\hat{\theta}_{IPW}$ is very biased. For $\hat{\theta}_{AIPW}$, it has big bias when $q(X)$ is misspecified for $E(Y | X)$. For $\hat{\theta}$, with $S(X)$ a linear function in X_1 and X_2 , $S(X)$ is misspecified for $E(Y | X)$ but roughly recovers $E(Y | X)$, we thus observe a negligible bias. This observation demonstrates the double robustness of $\hat{\theta}$ in Theorem 2. In addition, we observe that, with w moderately misspecified and the working index as a cubic function of only X_1 , which is far from $E(Y | X)$, $\hat{\theta}$ still has quite small bias and standard deviation, indicating that $\hat{\theta}$ is quite tolerant in misspecification of both working models.

In the second simulation, Y is simulated from $Y | X \sim N\{\exp(X_1 + X_2 + X_3 + X_4 + X_5), 2\}$, with $X = (X_1, \dots, X_5)^T$ and X_i from independent $N(0, 1)$. The selection probability is $\pi(X) = 1 - 1/\{1 + \exp(1 + X_1 + X_2 + X_3 + X_4 + X_5)\}$. For $\hat{\theta}_{NR}$, no model specification is needed. For $\hat{\theta}$, two working selection probabilities are considered: $w(X, \beta)$ is correctly specified with unknown β and $w(X)$ is misspecified as $w(X) = \sum_{i=1}^n \delta_i/n$. Two working models $S(X; \gamma)$ for $E(Y | X)$ are considered: (I) $S(X; \gamma) = \exp(X^T \gamma)$ is correctly specified for $E(Y | X)$ and (II) $S(X; \gamma) = (1, X)^T \gamma$ is an approximate atom of $E(Y | X)$. Maximum likelihood estimation is applied for the estimation of β and γ . For $\hat{\theta}_{AIPW}$, we let $q(X) = S(X)$. Simulation outcomes are reported in Table 3.

Comparing $\hat{\theta}$ with $\hat{\theta}_{NR}$, we see that $\hat{\theta}_{NR}$ is more variable than $\hat{\theta}$ due to the curse of dimensionality. Even with a misspecified selection probability, $\hat{\theta}$ still shows an overall better performance than $\hat{\theta}_{NR}$. We observe similar performance of $\hat{\theta}$ as in the first simulation relative to $\hat{\theta}_{IPW}$ and $\hat{\theta}_{AIPW}$. With $S(X)$ being an approximate atom of $E(Y | X)$, $\hat{\theta}$ has negligible bias and maintains small standard deviation whether w is correctly specified or misspecified, which again demonstrates its robustness to model misspecification. In addition, we see that under the correct specification of both working models, though both $\hat{\theta}$ and $\hat{\theta}_{AIPW}$ should be of the same optimal efficiency, $\hat{\theta}$ is less variable than $\hat{\theta}_{AIPW}$. Further investigation reveals that there are occasionally extreme Y

Table 4. Estimation of $\theta = E(Y)$: the selection probability is known in I and is misspecified as constant in II; $q(X) = S(X)$ is correctly specified for $E(Y | X)$ in (a) and an approximate atom of $\pi(X)$ in (b)

$w(X)$			$\hat{\theta}_{IPW}$	$\hat{\theta}_{AIPW}$		$\hat{\theta}$	
				(a)	(b)	(a)	(b)
I.	$\pi(X)$	Bias	-0.19	-0.06	-0.27	-0.06	-0.13
		SD	1.40	0.49	1.74	0.48	0.92
		MSE	1.99	0.24	3.09	0.23	0.86
II.	1	Bias	-1.03	-0.06	-1.6	-0.06	-0.22
		SD	0.45	0.48	0.47	0.48	0.58
		MSE	1.26	0.23	2.78	0.23	0.39

SD, standard deviation; MSE, mean squared error.

values, which causes the parameter estimates of γ to deviate from the truth and consequently $q(X) = S(X)$ to deviate from $E(Y | X)$. These extreme Y values adversely affect the efficiency of $\hat{\theta}_{AIPW}$. However, their impact on $\hat{\theta}$ is dampened considerably by the nonparametric regression.

5. REMARKS ON INDEX FUNCTION SELECTION

Based on Theorems 1 and 2, it is best to have the working index as an atom of $E(Y | X)$ to ensure both consistency and optimal efficiency. However, for attaining consistency alone, there is another possibility. In the proof of Theorem 2, if index $S = S(X)$ is an atom of $\pi(X)/w(X)$ in the sense that $\pi(X)/w(X) = p(S)$ for some function p , then the denominator (7) becomes $f_S(s)p(s) + O(h^2)$ and the numerator (6) becomes $f_S(s)p(s)E\{E(Y | X) | S(X) = s\} + O(h^2)$, where $E\{E(Y | X) | S(X) = s\} = m(s)$. Thus $\hat{m}(s)$ is consistent for $m(s)$ and $\hat{\theta}$ is accordingly consistent for $E(Y)$. Therefore, an index $S = S(X)$ that is an atom of $\pi(X)/w(X)$ also leads to consistent estimation.

A special case of the above condition is $w = 1$ and $S(X)$ an atom of $\pi(X)$, i.e. $\pi(X) = p\{S(X)\}$ for some function p . Under such an index, $\text{pr}(\delta = 1 | Y, S) = E\{E(\delta | Y, X) | S\} = p(S)$, so Y is missing at random with respect to S . Consequently, $m(S) = E(Y | S) = E(Y | S, \delta = 1)$ can be estimated by the complete-case nonparametric regression (3) or (4) with $w = 1$, i.e. without inverse probability weighting. If the index is $S(X) = \pi(X)$, then the semiparametric dimension reduction estimation reduces to a smoothed version of the propensity score method.

The following is a simple simulation where a working index is related to π . Data are generated as in the first simulation with each estimation repeated 500 times. Two working selection probabilities are considered: (I) $w(X) = \pi(X)$ and (II) $w(X) = 1$. Two working indices are considered: $S(X; \gamma)$ is correctly specified for $E(Y | X)$; $S(X; \gamma)$ is from logistic regression $\text{logit}\{p(\delta = 1 | X)\} = X^T \gamma$ with $S(X) = X^T \hat{\gamma}$, which is roughly an atom of $\pi(X)$. For $\hat{\theta}_{AIPW}$, we let $q(X) = S(X)$. Simulation results are in Table 4. We see that with $S(x)$ an approximate atom of $\pi(x)$ and $w(x)$ misspecified, $\hat{\theta}$ has negligible bias while $\hat{\theta}_{IPW}$ and $\hat{\theta}_{AIPW}$ have big biases. However, the standard deviation of $\hat{\theta}$ under (b) is a bit larger than that under (a), indicating a loss of estimation efficiency under this working index.

The attraction of using an atom of π as a working index is that no inverse probability weighting is needed in the nonparametric estimation of $m(\cdot)$, which is appealing when $\pi(X)$ is close to zero over a certain region of X and inverse probability weighting may inflate the estimation variance. Except for this situation, the best choice for the working index is still an atom of $E(Y | X)$, which projects X onto a direction that efficiently captures the information about Y and allows efficient estimation of $E(Y)$. This property is not possessed by other types of working indices.

Since $E(Y | X)$ is a special case of its own atom, any working model for $E(Y | X)$ can be used for the index function, usually by linear regression or generalized linear regression. Semiparametric least square estimation (Ichimura, 1993; Delecroix et al., 2003) can also be considered for estimating the working index. By semiparametric least square estimation, a parametric form of $S(X) = v_\gamma(X)$ is specified first, which usually takes a linear form $v_\gamma(X) = X^\top \gamma$ unless prior knowledge about the relationship between X and Y suggests a more complex form. Then, an iterative algorithm is adopted for the estimation of γ by minimizing $E[Y - g\{v_\gamma(X)\}]^2$ with the link function g unspecified and nonparametrically estimated. However, a computationally extensive procedure like semiparametric least squares is not necessary for finding the working index in semiparametric dimension reduction estimation.

6. APPLICATION

In an HIV clinical trial, 2139 HIV positive patients were involved (Hammer et al., 1996; Davidian et al., 2005). The patients were randomized into four arms to receive the respective antiretroviral regimen: (i) zidovudine or ZDV; (ii) didanosine or ddi; (iii) ZDV + ddi and (iv) ZDV + zalcitabine. We are interested in the difference between the monotherapy (i) and the combined therapies (ii)–(iv) in terms of the CD4 counts at 96 ± 5 weeks post therapy. The response Y is the CD4 count at 96 ± 5 weeks. The covariate X is the vector of six baseline variables: weight, HIV symptoms, prior experience with antiretroviral therapy, Karnofsky score, CD4 count and CD8 count. Due to death and dropout, there were 39% monotherapy recipients with missing observations in Y and 37% combined therapy recipients with missing observations in Y . However, all patients were observed for baseline covariate X .

Let Z denote the random assignment: $Z = 1$ if receiving combined therapy and $Z = 0$ if receiving monotherapy. The difference of interest is $\Delta = \mu^{(1)} - \mu^{(0)}$, where $\mu^{(1)} = E(Y | Z = 1)$ is the mean response under the combined therapy and $\mu^{(0)} = E(Y | Z = 0)$ the mean response under the monotherapy. In this study, it is reasonable to assume that the baseline covariate predicts the missingness in Y ; that is, Y is missing at random. Denote $w^{(0)}(X)$ and $w^{(1)}(X)$ as the working models for the probabilities of observing Y in the two groups.

We apply semiparametric dimension reduction to the estimation of $\mu^{(0)}$ and $\mu^{(1)}$. Denote $S^{(z)}$ as the parametric working index function, so that $S^{(z)} = S^{(z)}(X)$ summarizes the information contained in X about the response Y in group z . Denote $m^{(z)}(S) = E(Y | S, Z = z)$ as the conditional mean of Y in group z given $S^{(z)} = S^{(z)}(X)$, where the superscript of S is omitted for notational simplicity. It follows that $\mu^{(z)} = E\{m^{(z)}(S)\}$. It is intuitive that $S^{(0)}(\cdot)$ may be different from $S^{(1)}(\cdot)$, $m^{(0)}(\cdot)$ different from $m^{(1)}(\cdot)$ and $w^{(0)}(\cdot)$ different from $w^{(1)}(\cdot)$. They are thus estimated within each group separately.

Denote $\{X_j^{(z)}, Y_j^{(z)}, \delta_j^{(z)}\}$ as the data in group z . We fit a logistic regression $\text{logit}\{w_j^{(z)}\} = (1, X_j^{(z)})^\top \beta^{(z)}$ for the probability of $\delta_j^{(z)} = 1$ and let $w_j^{(z)} = (1, X_j^{(z)})^\top \hat{\beta}^{(z)}$ with $\hat{\beta}^{(z)}$ the maximum likelihood estimate of $\beta^{(z)}$. We fit a complete case linear regression $Y_j^{(z)} = (1, X_j^{(z)})^\top \gamma^{(z)} + e_j$, where e_j is the random error, and let $S_j^{(z)} = S^{(z)}(X_j^{(z)}) = (1, X_j^{(z)})^\top \hat{\gamma}^{(z)}$ with $\hat{\gamma}^{(z)}$ the maximum likelihood estimate of $\gamma^{(z)}$. Estimation of $m^{(z)}(s)$ is from the inverse probability weighted local linear regression, more specifically, through the minimization of

$$\sum_j \frac{\delta_j^{(z)}}{w_j^{(z)}} K_h(S_j^{(z)} - s) \left\{ Y_j^{(z)} - \alpha_0 - \alpha_1 (S_j^{(z)} - s) \right\}^2,$$

with the nonparametric estimate as $\hat{m}^{(z)}(s) = \hat{\alpha}_0$. The bandwidth is selected by the plug-in method and estimated within each group separately.

Table 5. Estimation of the therapy effect difference $\Delta = \mu^{(1)} - \mu^{(0)}$ with the point estimates, the standard errors and the 95% confidence intervals

Method	Estimate	SE	95% confidence interval
IPW	58.19	10.33	[37.94, 78.44]
AIPW	61.91	8.83	[44.60, 79.22]
SDR	62.42	9.02	[44.74, 80.10]

IPW, inverse probability weighting estimation; AIPW, augmented inverse probability weighting estimation; SDR, semiparametric dimension reduction estimation; SE, standard error.

The mean response estimates are

$$\hat{\mu}^{(0)} = N^{-1} \sum_{i=1}^N \hat{m}^{(0)}(S_i), \quad \hat{\mu}^{(1)} = N^{-1} \sum_{i=1}^N \hat{m}^{(1)}(S_i), \quad (8)$$

where $N = 2139$ is the total number of subjects over the two groups. Due to randomization, all the subjects share the same baseline distribution. The mean response estimation (8) has a causal interpretation: $m^{(0)}(S_i) = E\{Y | S = S^{(0)}(X_i)\}$ would have been the mean response had everyone with baseline X_i been assigned to the monotherapy, thus all baseline covariate information is included in the estimation of $\mu^{(0)}$; similarly in the estimation of $\mu^{(1)}$.

Table 5 presents the estimates of $\Delta = \mu^{(1)} - \mu^{(0)}$ with the standard errors and the 95% confidence intervals. Also reported are the results from the inverse probability weighting estimation and the augmented inverse probability weighting estimation, where the working selection probabilities $w^{(z)}$ and $q^{(z)}(X) = S^{(z)}(X)$ are used.

All three estimations indicate that the combined therapy has significantly higher CD4 counts at 96 ± 5 weeks than the monotherapy. In addition, results from the semiparametric dimension reduction estimation and the augmented inverse probability weighting estimation are close in both point estimates and standard errors. This may be because the specified working models $w^{(z)}$ are very close to the true selection probability $\pi^{(z)}$ and $S^{(z)}(X)$ is very close to the true conditional mean $E(Y | X, Z = z)$. As discussed in §3, with both π and $E(Y | X)$ correctly specified, the semiparametric dimension reduction estimator and the augmented inverse probability weighting estimator have asymptotically equivalent performance: both are consistent and attain the same optimal efficiency, we thus observe similar results.

7. CONCLUSION

In nonresponse data, covariates contain two types of information: the missing pattern of the response and the conditional mean of the response. The proposed semiparametric dimension reduction estimation makes use of both pieces of information with the second piece of information used through a parametric working index. Use of nonparametric regression along the direction of the working index leads to robustness to model misspecification and eliminates the curse of dimensionality. The semiparametric dimension reduction estimation is most appealing in studies with high-dimensional covariates, which can be generalized to the estimation of other marginal parameters, for example, the marginal moments and the marginal distribution function of the response.

ACKNOWLEDGMENT

We thank the referees, the associate editor and the editor for their helpful comments. We also thank Professor Marie Davidian for providing the HIV clinical trial data. This work was part of the authors' official duties at the National Institutes of Health.

APPENDIX

Proofs

Proof of Lemma 1. Let $\bar{m}(s, u) = m(s) + m'(s)(u - s)$, $\alpha^* = (nh)^{1/2}[\alpha_0 - m(s), h\{\alpha_1 - m'(s)\}]^t$, $S_i^* = \{1, (S_i - s)/h\}^t$. If α minimizes (3), then α^* minimizes

$$\sum_{i=1}^n \frac{\delta_i}{\pi_i} K_h(S_i - s) \{Y_i - \bar{m}(s, S_i) - (nh)^{-1/2} \alpha^{*t} S_i^*\}^2,$$

where $\pi_i = \pi(X_i)$. Equivalently, α^* minimizes the normalized function

$$l_n(\alpha^*) = \frac{1}{2} \sum_{i=1}^n \frac{\delta_i}{\pi_i} K_h(S_i - s) [\{Y_i - \bar{m}(s, S_i) - (nh)^{-1/2} \alpha^{*t} S_i^*\}^2 - \{Y_i - \bar{m}(s, S_i)\}^2].$$

Let $W_n(\alpha^*) = \partial l_n(\alpha^*)/\partial \alpha^*$ and $A_n(\alpha^*) = \partial^2 l_n(\alpha^*)/\partial \alpha^{*2}$. The evaluations of $W_n(\alpha^*)$ and $A_n(\alpha^*)$ at $\alpha^* = 0$ are

$$W_n = -(nh)^{-1/2} \sum_{i=1}^n \frac{\delta_i}{\pi_i} K_h(S_i - s) \{Y_i - \bar{m}(s, S_i)\} S_i^*, \quad (A1)$$

$$A_n = (nh)^{-1} \sum_{i=1}^n \frac{\delta_i}{\pi_i} K_h(S_i - s) S_i^* S_i^{*t}. \quad (A2)$$

Taylor expansion of $l_n(\alpha^*)$ around $\alpha^* = (0, 0)^T$ gives

$$\begin{aligned} l_n(\alpha^*) &= W_n^t \alpha^* + \frac{1}{2} \alpha^{*t} A_n \alpha^* + O_p\{(nh)^{-1/2}\} \\ &= W_n^t \alpha^* + \frac{1}{2} \alpha^{*t} \Sigma_s \alpha^* + O_p\{(nh)^{-1/2}\} + o_p(h) \end{aligned}$$

as $A_n = \Sigma_s + o_p(h)$ and

$$\Sigma_s = f_S(s) \begin{pmatrix} \gamma_0 & \gamma_1 \\ \gamma_1 & \gamma_2 \end{pmatrix},$$

where $\gamma_l = \int u^l K(u) du$, and under the regularity Condition 1, $\gamma_0 = 1$, $\gamma_1 = 0$ and $\gamma_2 < \infty$.

By the quadratic approximation of Fan et al. (1995), as $n \rightarrow \infty$, $h \rightarrow 0$, we have

$$\hat{\alpha}^* = -\Sigma_s^{-1} W_n + o_p(h).$$

Similarly to Fan et al. (1995),

$$\begin{aligned} E(W_n) &= \frac{1}{2} (nh^5)^{1/2} m''(s) f_S(s) \begin{pmatrix} \gamma_2 \\ \gamma_3 \end{pmatrix} + O\{(nh^7)^{1/2}\}, \\ \text{var}(W_n) &= h^{-1} E \left(\frac{\delta}{\pi^2} K_h^2(S - s) \{Y - m(S)\}^2 S^* S^{*t} \right) + o(h) \\ &= f_S(s) \text{var}(Y | S = s) E\{\pi(X)^{-1} | S(X) = s\} \begin{pmatrix} \tau_0 & \tau_1 \\ \tau_1 & \tau_2 \end{pmatrix} + o(h), \end{aligned}$$

where $\tau_l = \int u^l K^2(u) du$, and under the regularity Condition 1, $\tau_0 = \int K^2(u) du \leq 1$. It can be further shown, by checking Lyapounov's condition and using the Cramér-Wold device, that $\hat{\alpha}^*$ follows an asymptotic normal distribution with mean $\Sigma_s^{-1} E(W_n)$ and variance $\Sigma_s^{-1} \text{var}(W_n) \Sigma_s^{-1}$. Lemma 1 thus follows. $\square$

Proof of Theorem 1. Write $n^{1/2}(\hat{\theta} - \theta)$ as $n^{1/2}R_n + n^{1/2}V_n + n^{1/2}T_n$, where $R_n = n^{-1} \sum_{i=1}^n \{m(S_i) - \theta\}$, $V_n = n^{-1} \sum_{i=1}^n E[\hat{m}(S_i) - m(S_i) | \{(X_j, Y_j) : \delta_j = 1\}]$, and $T_n = n^{-1} \sum_{i=1}^n (\hat{m}(S_i) - m(S_i) - E[\hat{m}(S_i) - m(S_i) | \{(X_j, Y_j) : \delta_j = 1\}])$. Obviously, $n^{1/2}R_n$ converges in distribution to $N[0, \text{var}\{m(S)\}]$.

To derive the distribution of V_n , we write $\hat{m}(s)$ from (3) as

$$\hat{m}(s) = \frac{n^{-1} \sum_{j=1}^n w_j Y_j}{n^{-1} \sum_{j=1}^n w_j}.$$

At any given s , $A_{n,l}$ converges almost surely to $E\{\delta K_h(S-s)(S-s)^l/\pi(X)\} = f_S(s)\gamma_l h^l + O(h^{l+1})$ for $l = 0, 1, 2$, and $n^{-1} \sum_{j=1}^n w_j$ converges almost surely to $f^2(s)\gamma_2 h^2 + O(h^4)$. Denote U_i as the i th summand in V_n , then $U_i = E[n^{-1} \sum_{j=1}^n w_j \{Y_j - m(S)\} (n^{-1} \sum_{j=1}^n w_j)^{-1} | \{(X_j, Y_j) : \delta_j = 1\}]$, for $i = 1, \dots, n$. It follows that $n^{1/2} V_n$ is equivalent to $n^{1/2} U_n$, with

$$\begin{aligned} U_n &= \frac{1}{n} \sum_{j=1}^n \frac{\delta_j}{\pi(X_j)} E \left[\frac{K_h(S_j - S) \{A_{n,2} - A_{n,1}(S_j - S)\} \{Y_j - m(S)\}}{n^{-1} \sum_{j=1}^n w_j} \middle| (X_j, Y_j) : \delta_j = 1 \right] \\ &= \frac{1}{n} \sum_{j=1}^n \frac{\delta_j Y_j}{\pi(X_j)} E \left[\frac{K_h(S_j - S) \{A_{n,2} - A_{n,1}(S_j - S)\}}{n^{-1} \sum_{j=1}^n w_j} \middle| S_j \right] \end{aligned} \quad (\text{A3})$$

$$\begin{aligned} &- \frac{1}{n} \sum_{j=1}^n \frac{\delta_j}{\pi(X_j)} E \left[\frac{K_h(S_j - S) m(S) \{A_{n,2} - A_{n,1}(S_j - S)\}}{n^{-1} \sum_{j=1}^n w_j} \middle| S_j \right] \\ &= \frac{1}{n} \sum_{j=1}^n \frac{\delta_j \{Y_j - m(S_j)\}}{\pi(X_j)}. \end{aligned} \quad (\text{A4})$$

The second equation is true as $S_j = S(X_j)$ and X_j always observed. The last equation is true as the conditional expectation in (A3) is $1 + o(h^2)$ and the conditional expectation in (A4) is $m(S_j)\{1 + o(h^2)\}$. Thus $n^{1/2} U_n$ converges in distribution to $N[0, E\{\text{var}(Y | S)/\pi(X)\}]$.

For T_n , $E(T_n) = 0$ and $nE(T_n^2) \leq E[\{\hat{m}(S) - m(S)\}^2] = O\{h^4 + (nh)^{-1}\}$, thus $n^{1/2} T_n = o_p(1)$ and its contribution to the asymptotic distribution is negligible. Since $n^{1/2} R_n$ and $n^{1/2} U_n$ are both normally distributed and independent of each other, $n^{1/2}(\hat{\theta} - \theta)$ is asymptotically normally distributed with mean zero and variance $\text{var}\{m(S)\} + E\{\text{var}(Y | S)/\pi(X)\}$.

In the above derivation, the terms involving h are neglected. Next, we look into the effects of these terms. For the asymptotic bias, $E(R_n) = E(T_n) = 0$ and the term containing h is only from V_n ; that is $E(V_n) = E(E[\hat{m}(S) - m(S) | \{(X_j, Y_j) : \delta_j = 1\}]) = E\{\hat{m}(S) - m(S)\}$, which equals $E[E\{\hat{m}(S) - m(S) | S\}] = h^2 \mathcal{B}$. For the asymptotic variance, the leading term containing h is from T_n . As $n\{\text{var}(T_n)\} = nE(T_n^2) = E[\{\hat{m}(S) - m(S)\}^2] - E(E^2[\hat{m}(S) - m(S) | \{(X_j, Y_j) : \delta_j = 1\}])$, where the first term equals $E\{h^4 B^2(S) + (nh)^{-1} \sigma^2(S)\}$ and the second term equals $E(U_n^2) = n^{-1} E[\delta \{Y - m(S)\}^2 / \pi^2(X)] = O(n^{-1})$, it follows that $n\{\text{var}(T_n)\} = (nh)^{-1} \mathcal{V} + o(1)$. Theorem 1 is proved. $\square$

Proof of Remark 1. With $\hat{S} = S(X; \hat{\gamma})$, the semiparametric dimension reduction estimator is denoted by $\hat{\theta}_{\hat{\gamma}} = n^{-1} \sum_{i=1}^n \hat{m}_{\hat{\gamma}}(\hat{S}_i)$, where

$$\hat{m}_{\hat{\gamma}}(\hat{S}_i) = n^{-1} \sum_{j=1}^n \frac{\delta_j}{\pi_j} K_b(\hat{S}_i - \hat{S}_j) Y_j / n^{-1} \sum_{j=1}^n \frac{\delta_j}{\pi_j} K_b(\hat{S}_i - \hat{S}_j). \quad (\text{A5})$$

Here for notational simplicity, we assume the inverse probability weighted Nadaraya–Watson nonparametric estimation (4) for $m(s)$. Since $\hat{S} \approx S(X; \gamma) + \partial^T S(X; \gamma) / \partial \gamma (\hat{\gamma} - \gamma)$, it follows that $\hat{S}_i - \hat{S}_j = S_i - S_j + \nabla S(X_i, X_j)(\hat{\gamma} - \gamma)$ with $\nabla S(X_i, X_j) = \partial^T S(X_i; \gamma) / \partial \gamma - \partial^T S(X_j; \gamma) / \partial \gamma$ and

$$K_b(\hat{S}_i - \hat{S}_j) = b^{-1} K \left(\frac{S_i - S_j}{b} + \frac{\nabla S(X_i, X_j)(\hat{\gamma} - \gamma)}{b} \right). \quad (\text{A6})$$

In (A6), $\nabla S(X_i, X_j)$ can be written as

$$\partial^T S(X_i; \gamma) / \partial \gamma - E\{\partial^T S(X_i; \gamma) / \partial \gamma\} - [\partial^T S(X_j; \gamma) / \partial \gamma - E\{\partial^T S(X_j; \gamma) / \partial \gamma\}].$$

It is obvious that $E[\partial^\top \mathcal{S}(X; \gamma)/\partial \gamma - E\{\partial^\top \mathcal{S}(X; \gamma)/\partial \gamma\}] = 0$. If the index function $\mathcal{S}$ satisfies the regularity condition, for example, $E\{|\partial^\top \mathcal{S}(X; \gamma)/\partial \gamma - E\{\partial^\top \mathcal{S}(X; \gamma)/\partial \gamma\}|^{(10+\delta)}\} < \infty$ for any $\delta > 0$, then by the Borel–Cantelli lemma, $|\partial^\top \mathcal{S}(X; \gamma)/\partial \gamma - E\{\partial^\top \mathcal{S}(X; \gamma)/\partial \gamma\}| < n^{1/(10+\delta)}$ with probability 1 and $|\nabla \mathcal{S}(X_i, X_j)| < 2 \times n^{1/(10+\delta)}$. With $\hat{\gamma} - \gamma = O_p(n^{-1/2})$, the second term inside the kernel function K of (A6) is $O_p(n^{-\delta/(10(10+\delta))})$ at bandwidth $b \sim n^{-2/5}$ and goes to 0 as $n \rightarrow \infty$. It thus follows from (A5) and (A6) that $\hat{\theta}_{\hat{\gamma}}$ is asymptotically equivalent as $\hat{\theta}$; that is, estimation of γ does not affect the asymptotic distribution of the proposed semiparametric estimator. $\square$

Proof of Corollary 1. We prove by showing that $\hat{m}_{\hat{\pi}}(s)$, the estimator of $m(s)$ under $\hat{\pi} = \pi(X, \hat{\beta})$, has the same asymptotic distribution as $\hat{m}(s)$ in Lemma 1. Then $\hat{\theta}_{\hat{\pi}} = n^{-1} \sum_{i=1}^n \hat{m}_{\hat{\pi}}(S_i)$ has the same asymptotic normality as $\hat{\theta}$ following the proof of Theorem 1.

The estimator for $m(s)$ from (3) is $\hat{m}_{\hat{\pi}}(s) = \hat{\alpha}_{\hat{\pi},0}^*$, with $\alpha_{\hat{\pi}}^* = (\alpha_{\hat{\pi},0}^*, \alpha_{\hat{\pi},1}^*)^\top$ minimizing

$$l_n(\alpha^*) = W_n(\hat{\pi})^\top \alpha_{\hat{\pi}}^* + \frac{1}{2} \alpha_{\hat{\pi}}^{*\top} A_n(\hat{\pi}) \alpha_{\hat{\pi}}^* + O_p\{(nh)^{-1/2}\},$$

where

$$W_n(\hat{\pi}) = -(nh)^{-1/2} \sum_{i=1}^n \frac{\delta_i}{\hat{\pi}_i} K_h(S_i - s) \{Y_i - \bar{m}(s, S_i)\} S_i^*,$$

$$A_n(\hat{\pi}) = (nh)^{-1} \sum_{i=1}^n \frac{\delta_i}{\hat{\pi}_i} K_h(S_i - s) S_i^* S_i^{*\top}.$$

Algebraic computation shows that $A_n(\hat{\pi}) = A_n + G_n\{1 + o_p(1)\}$, with A_n defined in (A2) and $G_n = (nh)^{-1} \sum_{i=1}^n \{\delta_i/\pi_i\} K_h(S_i - s) S_i^* S_i^{*\top} (\hat{\pi}_i - \pi_i)/\pi_i$.

With $\hat{\beta} - \beta = o_p(1)$, it follows that $E\{(\hat{\pi} - \pi)/\pi | X\} = c_1(X)o(1)$ and $\text{var}\{(\hat{\pi} - \pi)/\pi | X\} = c_2(X)o(1)$ for some function c_1 and c_2 . Thus $E(G_n) = E[h^{-1} K_h(S - s) S^* S^{*\top} E\{(\hat{\pi} - \pi)/\pi | X\}] = o(1)$ and

$$\begin{aligned} \text{var}(G_n) &= n^{-1} \text{var} \left\{ \frac{\delta}{\pi} \frac{1}{h} K_h(S - s) S^* S^{*\top} \frac{\hat{\pi} - \pi}{\pi} \right\} \\ &= n^{-1} E \left[\text{var} \left\{ \frac{\delta}{\pi} \frac{1}{h} K_h(S - s) S^* S^{*\top} \frac{\hat{\pi} - \pi}{\pi} \middle| X \right\} \right] \\ &\quad + n^{-1} \text{var} \left[E \left\{ \frac{\delta}{\pi} \frac{1}{h} K_h(S - s) S^* S^{*\top} \frac{\hat{\pi} - \pi}{\pi} \middle| X \right\} \right] \end{aligned}$$

is $o\{(nh)^{-1}\}$. Therefore, $G_n = o_p(1)$ and $A_n(\hat{\pi}) = A_n(\pi) + o_p(1)$, which is $\Sigma_S + o_p(h) + o_p(1)$ with Σ_S as defined in the proof of Lemma 1. Then, by quadratic approximation of Fan et al. (1995), we have $\hat{\alpha}_{\hat{\pi}}^* = \Sigma_S^{-1} W_n(\hat{\pi}) + o_p(h)$.

By the linearization technique of Wang et al. (1998), we have $W_n(\hat{\pi}) = W_n - H_n + R_n\{1 + o_p(1)\}$, where W_n is as defined in (A1), $H_n = (nh)^{-1/2} \sum_{i=1}^n [\delta_i/\pi_i K_h(S_i - s) \{Y_i - \bar{m}(s, S_i)\} S_i^* (\hat{\pi}_i - \pi_i)/\pi_i]$ and $R_n = (nh)^{-1/2} \sum_{i=1}^n [\delta_i/\pi_i K_h(S_i - s) \{Y_i - \bar{m}(s, S_i)\} S_i^* (\hat{\pi}_i - \pi_i)^2/\pi_i^2]$. Similar to the above computations on G_n , it can be shown that $E(H_n + R_n) = O(nh^3)$, $\text{var}(H_n + R_n) = o(1)$. It thus follows that $\hat{m}_{\hat{\pi}}(s)$ has the same asymptotic normality as $\hat{m}(s)$ in Lemma 1. $\square$

REFERENCES

- AERTS, M., CLAESKENS, G., HENS, N. & MOLENBERGHS, G. (2002). Local multiple imputation. *Biometrika* **89**, 375–88.
 BOSE, C. (1943). Note on the sampling error in the method of double sampling. *Sankhya* **6**, 332–3.
 CHENG, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *J. Am. Statist. Assoc.* **89**, 81–7.
 DAVIDIAN, M., TSIATIS, A. A. & LEON, S. (2005). Semiparametric estimation of treatment effect in a pretest-posttest study with missing data. *Statist. Sci.* **20**, 261–301.

- DELECROIX, M., HÄRDLE, W. & HRISTACHE, M. (2003). Efficient estimation in conditional single-index regression. *J. Mult. Anal.* **86**, 213–26.
- FAN, J. (1992). Design-adaptive nonparametric regression. *J. Am. Statist. Assoc.* **87**, 998–1004.
- FAN, J., HECKMAN, N. E. & WAND, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Am. Statist. Assoc.* **90**, 141–50.
- FAN, J. & MARRON, J. S. (1992). Best possible constant for bandwidth selection. *Ann. Statist.* **20**, 2057–70.
- HAMMER, S. M., KATZENSTEIN, D. A., HUGGIES, M. D., GUNDAKER, H., SCHOOLEY, R. T., HAUBRICH, R. H., HENRY, W. K., LEDERMAN, M. M., PHAIR, J. P., NIU, M., HIRSH, M. S. & MERIGAN, T. C. (1996). A trial comparing nucleotide monotherapy with combined therapy in HIV-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New Engl. J. Med.* **335**, 1081–90. On behalf of the AIDS Clinical Trials Group, Study 175 study team.
- HÄRDLE, W., MÜLLER, M., SPERLICH, S. & WERWATZ, A. (2004). *Nonparametric and Semiparametric Models*. Berlin: Springer.
- HORVITZ, D. G. & THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Assoc.* **47**, 663–85.
- ICHIMURA, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Economet.* **58**, 71–120.
- MATLOFF, N. S. (1981). Use of regression functions for improved estimation of means. *Biometrika* **68**, 685–9.
- NADARAYA, E. A. (1964). On estimating regression. *Theory Prob. Appl.* **9**, 141–2.
- 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**, 846–66.
- ROSENBAUM, P. R. & RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **60**, 211–3.
- SHEATHER, S. J. & JONES, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Statist. Soc. B* **53**, 683–90.
- WANG, C. Y., WANG, S., GUTIERREZ, R. G. & CARROLL, R. J. (1998). Local linear regression for generalized linear models with missing data. *Ann. Statist.* **26**, 1028–50.
- WANG, Q. & DAI, P. (2008). Semiparametric model-based inference in the presence of missing responses. *Biometrika* **95**, 721–34.
- WANG, Q. & RAO, J. N. K. (2002a). Empirical likelihood-based inference in linear errors-in-covariate models with validation data. *Biometrika* **89**, 345–58.
- WANG, Q. & RAO, J. N. K. (2002b). Empirical likelihood-based inference under imputation for missing response data. *Ann. Statist.* **30**, 896–924.
- WATSON, G. S. (1964). Smooth regression analysis. *Sankhya* **26**, 359–72.
- WHITE, J. E. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *Am. J. Epidemiol.* **115**, 119–28.
- YATES, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Emp. J. Exp. Agric.* **1**, 129–42.

[Received March 2009. Revised September 2009]