Ann Inst Stat Math (2019) 71:889–910
https://doi.org/10.1007/s10463-018-0664-y

Dimension reduction for kernel-assisted M-estimators with missing response at random

Lei Wang[1]

Received: 19 February 2017 / Revised: 8 April 2018 / Published online: 25 April 2018

Abstract To obtain M-estimators of a response variable when the data are missing at random, we can construct three bias-corrected nonparametric estimating equations based on inverse probability weighting, mean imputation, and augmented inverse probability weighting approaches. However, when the dimension of covariate is not low, the estimation efficiency will be affected due to the curse of dimensionality. To address this issue, we propose a two-stage estimation procedure by using the dimension-reduced kernel estimators in conjunction with bias-corrected estimating equations. We show that the resulting three kernel-assisted estimating equations yield asymptotically equivalent M-estimators that achieve the desirable properties. The finite-sample performance of the proposed estimators for response mean, distribution function and quantile is studied through simulation, and an application to HIV-CD4 data set is also presented.

Keywords Consistency and asymptotic normality · Dimension reduction · Kernel-assisted · M-estimators · Missing at random

1 Introduction

Let $Y_1, \dots, Y_n$ be n independent observations on a real-valued random variable Y with distribution function F, and $\theta = \theta(F)$ be a parameter of interest from a set Θ. Let θ_0, the true value of the parameter θ, be a unique solution of the following estimating equation

$$E\{\varphi(Y, \theta)\} = 0,$$

✉ Lei Wang
lwangstat@nankai.edu.cn

1 Institute of Statistics and LPMC, Nankai University, Tianjin 300071, China

where $\varphi(\cdot)$ is a known function. Huber (1981) introduced a flexible class of estimators, called "M-estimators," which are defined to be the solution of an estimating equation $n^{-1}\sum_{i=1}^{n}\varphi(Y_i,\theta)=0$. The M-estimators are generalizations of the usual maximum likelihood estimators, generalized method of moment estimators, location and scale estimators, and have become a very useful tool in statistical inference. The asymptotic properties of M-estimators with complete data are investigated extensively, and the results can be found in Huber (1981), Serfling (1981), Zhang (1995) and others.

In survey sampling, social science, epidemiology studies and many other statistical problems, Y often has missing values and there exists a d-dimensional auxiliary vector X which is observed for the entire sample. Let δ be the response status indicator for Y, where $\delta=1$ if Y is observed and $\delta=0$ otherwise. The probability of missingness is assumed only to depend on observed values X, i.e., $\Pr(\delta=1|X,Y)=\Pr(\delta=1|X)=\pi(X)$. In this case, it can be shown that $\delta\perp Y|X$ holds and then Y is called missing at random (MAR; Rubins 1976). Here, $\perp$ stands for conditional independence.

Under the MAR assumption, dropping observations with missing response data will result in a serious loss of efficiency and badly biased estimators. Hence, a variety of methods such as likelihood-based approaches, weighting, regression and multiple imputation are proposed to obtain unbiased estimators, see Cheng (1994), Ibrahim et al. (2005), Kim and Shao (2013) and the references therein for comprehensive literature reviews. In this paper, we focus on the following three types of nonparametric bias-corrected estimating equations. For example, to obtain unbiased estimators, Wang (2007) and Wooldridge (2007) proposed the nonparametric inverse probability weighting (IPW) M-estimators. However, the kernel-assisted IPW method may not have enough estimation efficiency, as it does not fully extract the information contained in the auxiliary variables. Alternatively, following Wang and Chen (2009) and Chen et al. (2015), estimating equations can be imputed nonparametrically based on all observed covariates X to improve estimation efficiency over the IPW M-estimators. Also, a combination of the kernel-assisted IPW and imputation methods is considered, which leads to the nonparametric augmented inverse probability weighting (AIPW) approach (Xue 2009; Wang et al. 2010; Chen et al. 2015). These three nonparametric methods overcome the difficulty with the misspecification of propensity $\pi(X)$ and expected estimating equation $m_\varphi(X,\theta)=E\{\varphi(Y,\theta)|X\}$ commonly encountered with parametric methods.

When the dimension of covariates is not low, due to the well-known curse of dimensionality, the performance of nonparametric estimation of $\pi(X)$ and $m_\varphi(X,\theta)$ is poor, which causes the kernel-assisted IPW, MI, AIPW M-estimators lose efficiency and seriously limits their application scope. To address this issue, one possible way is to reduce the dimension of the kernel regression estimators. In the presence of response $\mathbb{Y}$ and high-dimensional covariates X, a major research effort in the last two decades is to find an $S=BX$ such that $\mathbb{Y}\perp X|BX$, i.e., $\mathbb{Y}$ and X are conditionally independent given BX, where B is a $p\times d$ deterministic matrix with $p<d$. The matrix B is a basis of the central dimension reduction space of the regression of $\mathbb{Y}$ on X, and BX can be viewed as a parsimonious summary of X in the sense that it has lower dimension than X but carries all the information contained in X about $\mathbb{Y}$. Then, the high-dimensional X can be replaced by the low-dimensional S in kernel regression to obtain more efficient estimators. For example, Hu et al. (2014) considered regression estimation of

the response mean by dimension reduction to find S_δ, S_Y, $S_{\delta Y}$, respectively, where S_δ contains all information in X for missingness δ, i.e., $\delta \perp X|S_\delta$; S_Y contains all information in X for response Y, i.e., $Y \perp X|S_Y$; $S_{\delta Y}$ contains all information in X for both δ and Y, i.e., $Y \perp X|S_{\delta Y}$ and $\delta \perp X|S_{\delta Y}$. Li et al. (2017) studied the estimation of mean response based on regression using S_Y. Deng and Wang (2017) proposed the dimension reduction estimation based on regression and IPW methods for probability density using the same S_δ, S_Y, $S_{\delta Y}$ as in Hu et al. (2014).

In this paper, to make valid and efficient M-estimators, we first apply the sufficient dimension reduction (SDR) technique (Li 1991; Cook and Weisberg 1991; Cook 1994; Ma and Zhu 2012, 2013) to obtain a lower dimensional S instead of X in the kernel regression estimation of the propensity $\pi(X)$ as well as the expected estimating equation $m_\varphi(X, \theta)$, and then construct three efficient bias-corrected nonparametric estimating equations based on the IPW, MI, AIPW approaches. There are several novel contributions in the current article.

- (i) Different from Hu et al. (2014) and Li et al. (2017), we consider a general M-estimator, which includes population mean, moments, distribution function, quantile and many other marginal parameters of the response. The estimators proposed in Hu et al. (2014) and Li et al. (2017) can be seen as the special cases of our proposed dimension reduction M-estimators. The general dimension reduction conditions for M-estimators are studied and investigated in Sect. 2.
- (ii) We consider three different kernel-assisted estimating equations which can overcome the difficulty with the misspecification of propensity and expected estimating equation commonly encountered with parametric methods. To the best of our knowledge, none of these papers studied the kernel-assisted IPW, MI, AIPW M-estimators together. We investigate the impact of SDR in nonparametric estimation of parameters when a $\sqrt{n}$-consistent estimator $\hat{S}$ of S is used. It can be shown that the proposed estimators obtained by using $\hat{S}$ and S are asymptotically equivalent. We further show that the resulting three estimating equations yield asymptotically equivalent M-estimators which achieve the desirable properties of consistency and asymptotic normality.
- (iii) Simulation results for the response mean, distribution function, quantile show that the proposed methods not only have substantially accurate coverage probabilities, but also have efficient point estimators.

The rest of this article is organized as follows. After presenting three types of nonparametric estimating equations, we introduce our main idea and establish a number of asymptotic properties in Sect. 2. Three special kernel-assisted M-estimators for the response mean, distribution function and quantile are investigated in Sect. 3. Simulation studies are given in Sect. 4. Section 5 analyzes the AIDS Clinical Trials Group Protocol 175 data for illustration. All technical details are provided in “Appendix.”

2 Methodology and theory

Let (X_i, Y_i, δ_i), $i = 1, ..., n$, be independent and identically distributed realizations from (X, Y, δ), where X_i is a d-dimensional fully observed covariate vector, Y_i is a univariate response having missing values, and δ_i is a binary response indicator that equals 1 if and only if Y_i is observed. Throughout this paper, we make the MAR assumption such that $m_\varphi(X,\theta) = E\{\varphi(Y,\theta)|X,\delta=1\} = E\{\delta\varphi(Y,\theta)|X\}/E(\delta|X)$. Thus, the unspecified functions $\pi(X)$ and $m_\varphi(X,\theta)$ can be estimated by the kernel regression estimators as follows:

$$\hat{\pi}(X) = \frac{\sum_{j=1}^n \delta_j K_h(X-X_j)}{\sum_{j=1}^n K_h(X-X_j)} \text{ and } \hat{m}_\varphi(X,\theta) = \frac{\sum_{j=1}^n \delta_j \varphi(Y_j,\theta) K_h(X-X_j)}{\sum_{j=1}^n \delta_j K_h(X-X_j)}, \tag{1}$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$, $K(\cdot)$ is a symmetric kernel function and h is a bandwidth.

We next describe three bias-corrected estimating equations using the nonparametric kernel estimators $\hat{\pi}(X)$ and $\hat{m}_\varphi(X,\theta)$ in (1) for handling missing data.

(i) Nonparametric IPW method: The kernel-assisted IPW estimating equation assigns each observed Y_i with weight proportional to the inverse of the estimated propensity $\hat{\pi}(X_i)$, i.e.,

$$\frac{1}{n}\sum_{i=1}^n \frac{\delta_i \varphi(Y_i,\theta)}{\hat{\pi}(X_i)} = 0. \tag{2}$$

(ii) Nonparametric MI method: For each X_i with $\delta_i = 0$, we use an estimated $\hat{m}_\varphi(X_i,\theta)$ to impute $m_\varphi(X_i,\theta)$, and the kernel-assisted MI estimating equation for θ is given by

$$\frac{1}{n}\sum_{i=1}^n \{\delta_i \varphi(Y_i,\theta) + (1-\delta_i)\hat{m}_\varphi(X_i,\theta)\} = 0. \tag{3}$$

(iii) Nonparametric AIPW method: We consider a combination of the nonparametric IPW and MI methods. This leads to the kernel-assisted AIPW estimating equation for θ as follows:

$$\frac{1}{n}\sum_{i=1}^n \Big[\frac{\delta_i \varphi(Y_i,\theta)}{\hat{\pi}(X_i)} + \Big\{1 - \frac{\delta_i}{\hat{\pi}(X_i)}\Big\}\hat{m}_\varphi(X_i,\theta)\Big] = 0. \tag{4}$$

However, with a multivariate X, due to the well-known curse of dimensionality, the nonparametric kernel estimators $\hat{\pi}(X)$ and $\hat{m}_\varphi(X,\theta)$ in (1) may not be efficient, which causes the kernel-assisted M-estimators obtained by (2–4) loses efficiency.

In this paper, we apply SDR to reduce the dimension of the kernel regression estimation, such that we can obtain more efficient nonparametric estimators of $\hat{\pi}(X)$ and $\hat{m}_\varphi(X,\theta)$, respectively. The main idea is that if we can find a $p \times d$ matrix B with p much smaller than d and $E(\delta|X)$ and $E\{\delta\varphi(Y,\theta)|X\}$ are functions of $S = BX$, then $\pi(X)$ and $m_\varphi(X,\theta)$ satisfy

$$\pi(X) = E(\delta|S) = \pi(S), \quad m_\varphi(X, \theta) = \frac{E\{\delta\varphi(Y, \theta)|S\}}{E(\delta|S)} = m_\varphi(S, \theta),$$

respectively, which are functions of S only. That is the information contained by X about $(\delta, \delta\varphi(Y, \theta))$ is summarized by S, i.e., $(\delta, \delta\varphi(Y, \theta)) \perp X|S$. However, the second condition, $\delta\varphi(Y, \theta) \perp X|S$, involves θ which is unknown so that we cannot use $\delta\varphi(Y, \theta)$ because it is unobservable. If we try a set of θ values in a plausible range of the parameter space under the condition $\delta\varphi(Y, \theta) \perp X|S$, then the resulting S may depend on θ, which is not desirable.

Consider a special case where $\varphi(Y, \theta) = Y - \theta$ for estimating response mean. Note that $\delta(Y - \theta) = \delta(Y - \theta_1) + \delta(\theta_1 - \theta)$ for any θ and a fixed θ_1. Thus, $(\delta, \delta\varphi(Y, \theta_1)) \perp X|S$ implies $(\delta, \delta\varphi(Y, \theta)) \perp X|S$ for any θ, which indicates that S is invariant to θ. Many marginal parameters, for example, the moments, distribution function and quantile of the response, have similar properties. This motivates us to impose a condition on $\varphi(Y, \theta)$ and then we can find $S = BX$ invariant to θ. Suppose that there exists θ_1 such that

$$\varphi(Y, \theta) = \Psi_\theta\big(\varphi(Y, \theta_1)\big) \quad \text{for any } \theta, \tag{5}$$

where Ψ_θ is a function depending on θ, and

$$\delta \perp X|BX, \quad \delta\varphi(Y, \theta_1) \perp X|BX, \tag{6}$$

for some B. Then we can find a B invariant to the value of θ and $\big(\delta, \delta\varphi(Y, \theta)\big) \perp X|BX$ holds for any θ. However, when $\varphi(Y, \theta)$ is complicate, the above conditions may not easily be satisfied. In this case, as suggested by one referee, we can apply SDR to find another $\tilde{B}$ such that $(\delta, Y) \perp X|\tilde{B}X$. Here, $\tilde{B}$ does not depend on θ and it can be shown that $(\delta, \delta\varphi(Y, \theta)) \perp X|\tilde{B}X$ holds for any θ.

Denote $g_1(Y_i, S_i, \delta_i, \theta) := \delta_i\varphi(Y_i, \theta)/\hat{\pi}(S_i)$, $g_2(Y_i, S_i, \delta_i, \theta) := \delta_i\varphi(Y_i, \theta) + (1 - \delta_i)\hat{m}_\varphi(S_i, \theta)$, $g_3(Y_i, S_i, \delta_i, \theta) := \delta_i\varphi(Y_i, \theta)/\hat{\pi}(S_i) + \{1 - \delta_i/\hat{\pi}(S_i)\}\hat{m}_\varphi(S_i, \theta)$, $i = 1, \ldots, n$. In applications, the low-dimensional S_i has to be estimated from the observed data and many SDR methods can be applied. Popular choices include sliced inverse regression (SIR; Li 1991), sliced average variance estimates (SAVE; Cook and Weisberg 1991), minimum average variance estimation (MAVE; Xia et al. 2002), and fusion-refinement procedure (Ding and Wang 2011). Under conditions (5–6), once we have an estimator $\hat{B}$ of B, we can obtain the dimension reduction kernel-assisted estimating equations $g_l(Y_i, \hat{S}_i, \delta_i, \theta)$, $l = 1, 2, 3$, with S_i replaced by $\hat{S}_i = \hat{B}X_i$, respectively, $i = 1, \ldots, n$. Theorem 1 shows that $\sum_{i=1}^n g_l(Y_i, \hat{S}_i, \delta_i, \theta)/n$ and $\sum_{i=1}^n g_l(Y_i, S_i, \delta_i, \theta)/n$, $l = 1, 2, 3$, are asymptotically equivalent. A sketched proof is given in "Appendix."

Theorem 1 *Under the conditions listed in "Appendix" and conditions* (5)-(6)*, as* $n \to \infty$*, we have* $\sqrt{n}\{\frac{1}{n}\sum_{i=1}^n g_l(Y_i, \hat{S}_i, \delta_i, \theta) - \frac{1}{n}\sum_{i=1}^n g_l(Y_i, S_i, \delta_i, \theta)\} = o_p(1)$ *and*

$$\sqrt{n}\Big\{\frac{1}{n}\sum_{i=1}^{n} g_l(Y_i, \hat{S}_i, \delta_i, \theta) - E\varphi(Y_i, \theta)\Big\} \xrightarrow{L} N(0, \sigma^2(\theta)), \quad l = 1, 2, 3,$$

with

$$\sigma^2(\theta) = Var\Big\{\frac{\delta\varphi(Y, \theta)}{\pi(S)}\Big\} - E\Big[\frac{1-\pi(S)}{\pi(S)}E^2\{\varphi(Y, \theta)|S\}\Big].$$

From Theorem 1, the kernel-assisted M-estimators based on $g_l(Y_i, \hat{S}_i, \delta_i, \theta)$, denoted as $\hat{\theta}_l$, can be obtained by solving the following estimating equations

$$n^{-1}\sum_{i=1}^{n} g_l(Y_i, \hat{S}_i, \delta_i, \theta) = 0,$$

where $l = 1, 2, 3$ denotes for IPW, MI and AIPW methods, respectively. Our next result is about the asymptotic behavior of $\hat{\theta}_l$ with a $\sqrt{n}$-consistent estimator $\hat{B}$ of B, $l = 1, 2, 3$. Its proof is given in "Appendix."

Theorem 2 *Under the conditions of Theorem 1, as $n \to \infty$, we have*

$$\sqrt{n}(\hat{\theta}_l - \theta_0) \xrightarrow{L} N(0, V(\theta_0)),$$

with $V(\theta_0) = \gamma^{-2}(\theta_0)\sigma^2(\theta_0)$ and $\gamma(\theta_0) = E\{\partial\varphi(Y, \theta)/\partial\theta\big|_{\theta=\theta_0}\}$.

3 Response mean, distribution function and quantile

3.1 Response mean

For response mean, as discussed in Sect. 2, we have $(\delta, \delta\varphi(Y, \theta_1)) = (\delta, \delta Y)$ by setting $\theta_1 = 0$, which means that we only need to find a $p \times d$ matrix B such that $(\delta, \delta Y) \perp X|BX$. On the other hand, under the MAR assumption, if we have a matrix B such that $\delta Y \perp X|BX$, then it also can be shown that $\delta \perp X|BX$ holds (Hu et al. 2014). Hence, we apply SDR to search a matrix B satisfying

$$\delta Y \perp X|BX. \tag{7}$$

Let $\hat{B}$ be a root-n-consistent estimator of B defined in (7) obtained by SDR. Denote

$$\hat{\pi}(S) = \frac{\sum_{j=1}^{n}\delta_j K_h(S - \hat{S}_j)}{\sum_{j=1}^{n} K_h(S - \hat{S}_j)} \text{ and } \hat{m}_Y(S) = \frac{\sum_{j=1}^{n}\delta_j Y_j K_h(S - \hat{S}_j)}{\sum_{j=1}^{n}\delta_j K_h(S - \hat{S}_j)},$$

with $\hat{S}_j = \hat{B}X_j$ for $j = 1, \ldots, n$. The kernel-assisted IPW, MI, AIPW M-estimators for $\mu = E(Y)$ are given as follows:

$$\hat{\mu}_{ipw} = \sum_{i=1}^{n} \frac{\delta_i Y_i}{\hat{\pi}(\hat{S}_i)} \Big/ \sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}(\hat{S}_i)},$$

$$\hat{\mu}_{mi} = n^{-1} \sum_{i=1}^{n} \left\{ \delta_i Y_i + (1-\delta_i)\hat{m}_Y(\hat{S}_i) \right\},$$

$$\hat{\mu}_{aipw} = n^{-1} \sum_{i=1}^{n} \left[\frac{\delta_i Y_i}{\hat{\pi}(\hat{S}_i)} + \left\{ 1 - \frac{\delta_i}{\hat{\pi}(\hat{S}_i)} \right\} \hat{m}_Y(\hat{S}_i) \right]. \quad (8)$$

It can be seen that the estimators proposed in Hu et al. (2014) and Li et al. (2017) are the special cases of our proposed dimension reduction kernel-assisted M-estimators in (8). Corollary 1 describes the asymptotic properties of $\hat{\mu}_l$, where $l = 1, 2, 3$ denotes for the kernel-assisted IPW, MI and AIPW estimators, respectively.

Corollary 1 *Under the regularity conditions in "Appendix," assume that μ_0 is the true value, as $n \to \infty$, we have*

$$\sqrt{n}(\hat{\mu}_l - \mu_0) \xrightarrow{L} N(0, V^2), \quad l = 1, 2, 3,$$

with

$$V^2 = E\{\pi^{-1}(S)Var(Y|S)\} + Var\{E(Y|S)\}.$$

3.2 Distribution function

In this case, for any given y, we want to estimate $F(y) = \Pr(Y \leq y)$ and the corresponding function $\varphi(Y, \theta) = I(Y \leq y) - \theta$, where $I(\cdot)$ is the indicator function. We consider $(\delta, \delta\varphi(Y, \theta_1)) = (\delta, \delta I(Y \leq y))$ by setting $\theta_1 = 0$ and then only need to find a $p \times d$ matrix B such that $(\delta, \delta I(Y \leq y)) \perp X|BX$. However, this B depends on the given value y. On the other hand, it can be shown that B defined in (7) also satisfies $(\delta, \delta I(Y \leq y)) \perp X|BX$ for any given y (see Lemma 1 in "Appendix"), which is easy to implement. Define

$$\hat{m}_F(y|S) = \frac{\sum_{j=1}^{n} \delta_j I(Y_j \leq y) K_h(S - \hat{S}_j)}{\sum_{j=1}^{n} \delta_j K_h(S - \hat{S}_j)}.$$

Similarly, we have the kernel-assisted IPW, MI, AIPW M-estimators for $F(y)$ as follows:

$$\hat{F}_{ipw}(y) = \sum_{i=1}^{n} \frac{\delta_i I(Y_i \leq y)}{\hat{\pi}(\hat{S}_i)} \Big/ \sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}(\hat{S}_i)},$$

$$\hat{F}_{mi}(y) = n^{-1} \sum_{i=1}^{n} \Big[\delta_i I(Y_i \leq y) + (1-\delta_i)\hat{m}_F(y|\hat{S}_i)\Big],$$

$$\hat{F}_{aipw}(y) = n^{-1} \sum_{i=1}^{n} \Big\{\frac{\delta_i I(Y_i \leq y)}{\hat{\pi}(\hat{S}_i)} + \Big[1 - \frac{\delta_i}{\hat{\pi}(\hat{S}_i)}\Big]\hat{m}_F(y|\hat{S}_i)\Big\}. \tag{9}$$

Corollary 2 describes the asymptotic properties of $\hat{F}_l(y)$ for $l = 1, 2, 3$.

Corollary 2 *Under the regularity conditions in "Appendix," as $n \to \infty$, we have*

$$\sqrt{n}(\hat{F}_l(y) - F(y)) \xrightarrow{L} N(0, V^2(y)), \quad l = 1, 2, 3,$$

with

$$V^2(y) = E[\pi^{-1}(S)Var\{I(Y \leq y)|S\}] + E[E^2\{I(Y \leq y)|S\}] - F^2(y).$$

3.3 Quantile

Let ζ_τ be the τth quantile of the unknown distribution function F, where $0 < \tau < 1$, and $F^{'}(y)$ be the first-order derivative of F, i.e., the probability density function of $F(y)$. The proposed M-estimators for ζ_τ are defied as

$$\hat{\zeta}_{l\tau} = \inf\{y : \hat{F}_l(y) \geq \tau\}, \tag{10}$$

where $\hat{F}_l(y)$ is defined in (9) and $l = 1, 2, 3$ denotes for IPW, MI and AIPW, respectively. Since the asymptotic normality of the estimators $\hat{F}_l(y)$ is given by Corollary 2, we can show the asymptotic normality of $\hat{\zeta}_{l\tau}$ based on the well-know Bahadur's expression (Cheng 1994).

Corollary 3 *Under the conditions in Corollary 2, assume that $F^{'}(y)$ is bounded away from 0, as $n \to \infty$, we have*

$$n^{1/2}(\hat{\zeta}_{l\tau} - \zeta_\tau) \xrightarrow{L} N(0, V^2(\zeta_\tau)/\{F^{'}(\zeta_\tau)\}^2).$$

4 Simulation studies

In this section, we conduct simulation studies to examine the finite-sample performances of the kernel-assisted IPW, MI, AIPW M-estimators and some other estimators. In particular, we obtain the simulated bias and standard deviation (SD) of estimators, the standard error (SE) obtained by the bootstrap with replication size 200, and the coverage probability (CP) of the confidence intervals at the nominal level 95% based on asymptotic normality and bootstrap SE. All results are based on 1000 simulation replications and the sample sizes $n = 200$ and 500.

In the first simulation, we consider

$$Y_i = X_i^T \beta + \epsilon_i, \quad i = 1, ..., n,$$

where X_i's are from the 10-dimensional normal distribution with zero mean and identity covariance matrix, ϵ_i's are independently from N(0, 1), and ϵ_i's and X_i's are independent, the true parameter vector $\beta^T = (1, 1, \ldots, 1)$. In the second simulation, we consider

$$Y_i = \exp(X_i^T \beta_1) + X_i^T \beta_2 + \varepsilon_i, \quad i = 1, ..., n,$$

where X_i's are from a 10-dimensional normal distribution with mean 0 and covariance matrix Γ with $\Gamma_{jj} = 1$ and $\Gamma_{jj'} = 0.2^{|j'-j|}$, $1 \le j < j' \le 10$; ϵ_i's are from $N(0, \sigma^2(X_i))$ with $\sigma^2(X_i) = 1 + 0.5X_{i1}^2$; the true parameter vectors $\beta_1^T = (0.5, 0.5, 0.5, 0.5, 0.5, 0, \ldots, 0)$ and $\beta_2^T = (0, 0, 0, 0, 0, 1, \ldots, 1)$.

We generate δ_i from the Bernoulli distribution with probability $\pi(X_i)$ and consider four choices of $\pi(X_i)$:

M1: $\pi(X_i) = 1/[1 + \exp\{-0.6 - 0.2(X_{i1} + \cdots + X_{i5})\}]$;
M2: $\pi(X_i) = 1/[1 + \exp\{-0.2 - 0.4\exp(X_{i1} + \cdots + X_{i5})\}]$;
M3: $\pi(X_i) = 1 - \exp[-\exp\{0.1 + 0.2(X_{i1} + \cdots + X_{i5})\}]$;
M4: $\pi(X_i) = \Phi\{0.4 + 0.2(X_{i1} + \cdots + X_{i5})\}$;

where $\Phi(\cdot)$ is the standard normal distribution function. The coefficients in the propensity models are chosen so that the unconditional rates of missing data are between 20 and 40%.

To implement our proposed methods, we obtain an estimated B defined in (7) using the SIR method based on R package "dr" with 10-dimensional covariates X and the observed response $\mathbb{Y} = \delta Y$. The structural dimension p of B is determined by the BIC-type criterion (Zhu et al. 2010) as follows:

$$\hat{p} = \arg\max_{v=1,\ldots,d} \left[\frac{n}{2} \frac{\sum_{j=1}^{v} \{\log(\hat{\eta}_j + 1) - \hat{\eta}_j\}}{\sum_{j=1}^{d} \{\log(\hat{\eta}_j + 1) - \hat{\eta}_j\}} - 2n^{1/2} \frac{v(v+1)}{2d} \right],$$

where the dimension of X is d, $\hat{\eta}_1 \ge \hat{\eta}_2 \ge \ldots \ge \hat{\eta}_d \ge 0$ are the eigenvalues from an estimate of $\Sigma_x = \text{Var}\{E(X|Y)\}$. Along the line of Shao and Wang (2016), the nonparametric kernel regression estimator is computed using a Gaussian product kernel with bandwidth $h = 1.5\hat{\xi} n^{-1/3}$, where $\hat{\xi}$ is an estimated standard deviation of observations.

We study the performance of (a) the response mean μ, (b) the distribution function $F(y)$ at $y = 3$ and (c) the median $\zeta_{0.5}$, based on the following six estimators:

(i) the proposed dimension reduction kernel-assisted IPW, MI, AIPW M-estimators $\hat{\mu}_l$, $\hat{F}_l(y)$, $\hat{\zeta}_{l\tau}$ for $l = 1, 2, 3$ defined by (8)–(10), respectively;

Table 1 Results of response mean based on 1000 runs

Model		$n = 200$				$n = 500$			
		Bias	SD	SE	CP	Bias	SD	SE	CP
Example 1									
M1	$\hat{\mu}_{ipw}$	0.086	0.257	0.256	0.935	0.023	0.160	0.163	0.947
	$\hat{\mu}_{mi}$	0.035	0.245	0.242	0.944	0.006	0.158	0.152	0.954
	$\hat{\mu}_{aipw}$	0.022	0.246	0.242	0.940	− 0.004	0.158	0.152	0.954
	$\check{\mu}$	0.003	0.238	0.233	0.945	− 0.006	0.153	0.148	0.941
	$\tilde{\mu}$	0.341	0.294	0.287	0.774	0.342	0.184	0.183	0.532
	$\hat{\mu}^*_{ipw}$	0.215	0.189	0.187	0.746	0.225	0.119	0.118	0.501
M2	$\hat{\mu}_{ipw}$	0.070	0.256	0.255	0.933	0.002	0.151	0.168	0.965
	$\hat{\mu}_{mi}$	0.020	0.243	0.239	0.938	− 0.015	0.147	0.152	0.958
	$\hat{\mu}_{aipw}$	− 0.001	0.244	0.240	0.939	− 0.030	0.147	0.153	0.950
	$\check{\mu}$	0.005	0.238	0.232	0.932	− 0.001	0.140	0.148	0.965
	$\tilde{\mu}$	0.473	0.280	0.275	0.597	0.477	0.161	0.175	0.219
	$\hat{\mu}^*_{ipw}$	0.342	0.197	0.189	0.507	0.346	0.126	0.125	0.235
M3	$\hat{\mu}_{ipw}$	0.033	0.153	0.163	0.959	0.041	0.156	0.171	0.946
	$\hat{\mu}_{mi}$	0.016	0.151	0.153	0.941	0.010	0.150	0.153	0.945
	$\hat{\mu}_{aipw}$	0.006	0.151	0.153	0.946	− 0.006	0.151	0.153	0.947
	$\check{\mu}$	0.004	0.146	0.148	0.947	0.005	0.145	0.148	0.950
	$\tilde{\mu}$	0.351	0.185	0.183	0.510	0.502	0.172	0.179	0.181
	$\hat{\mu}^*_{ipw}$	0.345	0.186	0.187	0.547	0.331	0.121	0.118	0.187
M4	$\hat{\mu}_{ipw}$	0.141	0.262	0.260	0.906	0.046	0.169	0.172	0.939
	$\hat{\mu}_{mi}$	0.049	0.239	0.242	0.940	0.005	0.160	0.153	0.941
	$\hat{\mu}_{aipw}$	0.030	0.241	0.243	0.939	− 0.011	0.161	0.153	0.941
	$\check{\mu}$	− 0.002	0.231	0.232	0.944	− 0.007	0.153	0.148	0.936
	$\tilde{\mu}$	0.521	0.274	0.283	0.543	0.526	0.181	0.181	0.161
	$\hat{\mu}^*_{ipw}$	0.338	0.183	0.182	0.504	0.348	0.117	0.116	0.196
Example 2									
M1	$\hat{\mu}_{ipw}$	0.055	0.431	0.439	0.943	0.062	0.284	0.276	0.954
	$\hat{\mu}_{mi}$	− 0.040	0.407	0.396	0.919	− 0.007	0.274	0.261	0.946
	$\hat{\mu}_{aipw}$	− 0.048	0.407	0.396	0.917	− 0.010	0.275	0.261	0.945
	$\check{\mu}$	− 0.004	0.399	0.386	0.938	0.018	0.273	0.258	0.951
	$\tilde{\mu}$	0.519	0.569	0.534	0.893	0.532	0.373	0.355	0.724
	$\hat{\mu}^*_{ipw}$	− 0.533	0.367	0.358	0.554	− 0.529	0.232	0.225	0.317
M2	$\hat{\mu}_{ipw}$	0.129	0.448	0.437	0.960	0.097	0.283	0.269	0.946
	$\hat{\mu}_{mi}$	0.036	0.430	0.400	0.932	0.017	0.271	0.257	0.939
	$\hat{\mu}_{aipw}$	0.023	0.429	0.400	0.932	0.010	0.271	0.257	0.938
	$\check{\mu}$	0.006	0.422	0.392	0.931	− 0.002	0.266	0.253	0.938
	$\tilde{\mu}$	0.668	0.551	0.507	0.802	0.663	0.349	0.327	0.476
	$\hat{\mu}^*_{ipw}$	− 0.169	0.407	0.398	0.824	− 0.168	0.257	0.245	0.613

Table 1 continued

Model		$n = 200$				$n = 500$			
		Bias	SD	SE	CP	Bias	SD	SE	CP
M3	$\hat{\mu}_{ipw}$	0.162	0.459	0.463	0.962	0.126	0.286	0.278	0.931
	$\hat{\mu}_{mi}$	0.017	0.432	0.412	0.934	0.002	0.271	0.260	0.930
	$\hat{\mu}_{aipw}$	0.006	0.432	0.412	0.931	− 0.004	0.271	0.260	0.930
	$\check{\mu}$	0.014	0.422	0.399	0.921	− 0.001	0.265	0.253	0.932
	$\tilde{\mu}$	0.748	0.605	0.556	0.806	0.721	0.367	0.353	0.468
	$\hat{\mu}^*_{ipw}$	− 0.320	0.397	0.378	0.753	− 0.329	0.253	0.241	0.685
M4	$\hat{\mu}_{ipw}$	0.151	0.465	0.459	0.963	0.117	0.288	0.276	0.928
	$\hat{\mu}_{mi}$	− 0.012	0.439	0.407	0.926	− 0.019	0.273	0.257	0.932
	$\hat{\mu}_{aipw}$	− 0.023	0.439	0.406	0.923	− 0.025	0.274	0.257	0.928
	$\check{\mu}$	− 0.003	0.420	0.394	0.932	− 0.020	0.268	0.250	0.922
	$\tilde{\mu}$	0.752	0.603	0.556	0.788	0.725	0.375	0.355	0.479
	$\hat{\mu}^*_{ipw}$	− 0.381	0.407	0.396	0.680	− 0.378	0.262	0.258	0.554

SD standard deviation, *SE* standard error, *CP* coverage probability

(ii) the nonparametric IPW M-estimators $\hat{\mu}^*_{ipw}$, $\hat{F}^*_{ipw}(y)$, $\hat{\zeta}^*_{ipw\tau}$ defined by (2) with the 10-dimensional X_i's used to obtain the kernel regression estimators in (1);

(iii) the M-estimators $\tilde{\mu}$, $\tilde{F}(y)$, $\tilde{\zeta}_\tau$ based on the observed Y_i's, which are defined as the roots of the following equations $\sum_{i=1}^n \delta_i \varphi(Y_i, \theta)/\sum_{i=1}^n \delta_i = 0$;

(iv) the M-estimators $\check{\mu}$, $\check{F}(y)$, $\check{\zeta}_\tau$ when there is no missing data, which are defined as the roots of the following equations $\sum_{i=1}^n \varphi(Y_i, \theta)/n = 0$.

The simulation results are presented in Tables 1, 2 and 3. A few conclusions can be drawn from the simulation results.

(i) Bias. The proposed estimators $\hat{\mu}_l$, $\hat{F}_l(y)$, $\hat{\zeta}_{l\tau}$, $l = 1, 2, 3$, have negligible biases in all cases. Among these three estimators, the MI and AIPW M-estimators are comparable, and they perform better than the IPW M-estimator when $n = 200$. On the other hand, $\tilde{\mu}$, $\tilde{F}(y)$, $\tilde{\zeta}_\tau$ are biased, due to the fact that missing is not completely at random. As expected, the IPW estimators without dimension reduction $\hat{\mu}^*_{ipw}$, $\hat{F}^*_{ipw}(y)$, $\hat{\zeta}^*_{ipw\tau}$ have very poor performance.

(ii) Standard deviation. The SDs of the proposed estimators $\hat{\mu}_l$, $\hat{F}_l(y)$, $\hat{\zeta}_{l\tau}$, $l = 1, 2, 3$ are smaller than the SDs of $\tilde{\mu}$, $\tilde{F}(y)$, $\tilde{\zeta}_\tau$, respectively, and become smaller when the mean response rate or the sample size is larger. It also can be seen that the SDs of the MI and AIPW M-estimators are comparable, and smaller than the SDs of the IPW M-estimator.

(iii) Standard error. The bootstrap variance estimator works well under all cases because the values of SE are rather close to those of SD.

(iv) Coverage probability. The coverage probabilities based on the proposed M-estimators are all close to the nominal level 0.95, and are quite comparable with the M-estimators $\check{\mu}$, $\check{F}(y)$, $\check{\zeta}_\tau$ assuming no missing data. The coverage probabilities of $\tilde{\mu}$, $\tilde{F}(y)$, $\tilde{\zeta}_\tau$ do not perform well in most of cases, because of their biases.

Table 2 Results of distribution function at $y = 3$ based on 1000 runs

Model		$n = 200$				$n = 500$			
		Bias	SD	SE	CP	Bias	SD	SE	CP
Example 1									
M1	$\hat{F}_{ipw}$	0.002	0.030	0.031	0.952	0.002	0.019	0.019	0.938
	$\hat{F}_{mi}$	0.005	0.030	0.030	0.938	0.004	0.019	0.019	0.933
	$\hat{F}_{aipw}$	0.005	0.029	0.030	0.939	0.003	0.019	0.019	0.935
	$\check{F}$	− 0.001	0.027	0.027	0.947	0.001	0.018	0.017	0.943
	$\tilde{F}$	0.030	0.036	0.036	0.900	− 0.030	0.023	0.023	0.782
	$\hat{F}^*_{ipw}$	− 0.311	0.035	0.036	0	− 0.310	0.023	0.022	0
M2	$\hat{F}_{ipw}$	0.001	0.029	0.030	0.949	0.005	0.018	0.019	0.946
	$\hat{F}_{mi}$	0.005	0.028	0.028	0.933	0.006	0.018	0.018	0.932
	$\hat{F}_{aipw}$	0.006	0.028	0.028	0.932	0.007	0.018	0.018	0.932
	$\check{F}$	0.001	0.027	0.027	0.950	0.002	0.017	0.017	0.944
	$\tilde{F}$	− 0.040	0.035	0.035	0.810	− 0.041	0.022	0.022	0.557
	$\hat{F}^*_{ipw}$	− 0.259	0.035	0.036	0	− 0.260	0.022	0.022	0
M3	$\hat{F}_{ipw}$	− 0.001	0.030	0.031	0.947	0.002	0.018	0.019	0.949
	$\hat{F}_{mi}$	0.005	0.029	0.029	0.932	0.004	0.018	0.018	0.938
	$\hat{F}_{aipw}$	0.005	0.029	0.029	0.933	0.004	0.018	0.018	0.937
	$\check{F}$	0.001	0.027	0.027	0.946	0.001	0.017	0.017	0.946
	$\tilde{F}$	− 0.040	0.036	0.036	0.843	− 0.040	0.022	0.023	0.632
	$\hat{F}^*_{ipw}$	− 0.300	0.037	0.035	0	− 0.301	0.022	0.022	0
M4	$\hat{F}_{ipw}$	− 0.001	0.029	0.032	0.964	0.001	0.019	0.020	0.958
	$\hat{F}_{mi}$	0.005	0.029	0.030	0.945	0.003	0.019	0.019	0.942
	$\hat{F}_{aipw}$	0.005	0.028	0.029	0.945	0.003	0.018	0.019	0.943
	$\check{F}$	0.001	0.027	0.027	0.947	0.001	0.017	0.017	0.951
	$\tilde{F}$	− 0.040	0.034	0.037	0.840	− 0.040	0.024	0.023	0.595
	$\hat{F}^*_{ipw}$	− 0.317	0.035	0.035	0	− 0.318	0.022	0.022	0
Example 2									
M1	$\hat{F}_{ipw}$	0.001	0.040	0.041	0.943	− 0.001	0.025	0.025	0.947
	$\hat{F}_{mi}$	0.007	0.039	0.039	0.941	0.004	0.024	0.024	0.943
	$\hat{F}_{aipw}$	0.008	0.039	0.039	0.939	0.004	0.024	0.024	0.946
	$\check{F}$	0.001	0.034	0.034	0.938	0.001	0.022	0.021	0.950
	$\tilde{F}$	− 0.037	0.043	0.043	0.858	− 0.037	0.027	0.027	0.726
	$\hat{F}^*_{ipw}$	− 0.257	0.035	0.034	0	− 0.259	0.022	0.022	0

Table 2 continued

Model		$n = 200$				$n = 500$			
		Bias	SD	SE	CP	Bias	SD	SE	CP
M2	$\hat{F}_{ipw}$	− 0.006	0.038	0.039	0.955	− 0.002	0.023	0.024	0.949
	$\hat{F}_{mi}$	0.002	0.038	0.037	0.947	0.002	0.023	0.023	0.952
	$\hat{F}_{aipw}$	0.003	0.038	0.037	0.937	0.003	0.023	0.023	0.946
	$\check{F}$	− 0.001	0.034	0.034	0.948	0.001	0.021	0.021	0.952
	$\tilde{F}$	− 0.056	0.041	0.041	0.716	− 0.053	0.026	0.026	0.460
	$\hat{F}^*_{ipw}$	− 0.216	0.035	0.034	0	− 0.215	0.022	0.021	0
M3	$\hat{F}_{ipw}$	− 0.003	0.040	0.040	0.954	− 0.004	0.024	0.025	0.950
	$\hat{F}_{mi}$	0.009	0.039	0.038	0.936	0.004	0.024	0.024	0.947
	$\hat{F}_{aipw}$	0.009	0.039	0.038	0.936	0.004	0.024	0.024	0.943
	$\check{F}$	0.002	0.034	0.034	0.936	0.001	0.022	0.021	0.938
	$\tilde{F}$	− 0.052	0.043	0.043	0.782	− 0.053	0.027	0.027	0.505
	$\hat{F}^*_{ipw}$	− 0.254	0.034	0.035	0	− 0.254	0.022	0.021	0
M4	$\hat{F}_{ipw}$	− 0.005	0.039	0.041	0.956	− 0.007	0.025	0.025	0.931
	$\hat{F}_{mi}$	0.008	0.038	0.039	0.942	0.002	0.025	0.024	0.941
	$\hat{F}_{aipw}$	0.009	0.038	0.039	0.942	0.002	0.025	0.024	0.937
	$\check{F}$	− 0.001	0.033	0.034	0.947	− 0.001	0.022	0.021	0.929
	$\tilde{F}$	− 0.056	0.043	0.043	0.748	− 0.058	0.028	0.027	0.464
	$\hat{F}^*_{ipw}$	− 0.270	0.033	0.034	0	− 0.271	0.022	0.021	0

SD standard deviation, *SE* standard error, *CP* coverage probability

It also can be seen that the confidence intervals based on $\hat{\mu}^*_{ipw}$, $\hat{F}^*_{ipw}(y)$, $\hat{\zeta}^*_{ipw\tau}$ have poor coverage rates.

In conclusion, the simulation results suggest that the proposed methods not only have efficient point estimates, but also have substantially accurate coverage probabilities. Among three proposed methods, the MI and AIPW are recommended in practice.

5 Real data

In this section, we illustrate the proposed method using data (Hammer et al. 1996) collected on 2139 HIV positive patients enrolled in AIDS Clinical Trials Group Protocol 175 (ACTG 175). In this HIV clinical trial, the patients were randomized into four arms to receive the respective antiretroviral regimen: (1) zidovudine or ZDV with 532 subjects; (2) didanosine or ddi with 522 subjects; (3) ZDV + ddi with 524 subjects; and (4) ZDV + zalcitabine with 561 subjects. Let response Y_s be the CD4 count at 96 $\pm$ 5 weeks which receiving the sth antiretroviral regimen, $s = 1, ..., 4$. There are six continuous baseline covariates: age, weight, CD4 cell counts at baseline and 20 $\pm$ 5 weeks, and CD8 cell counts at baseline and 20 $\pm$ 5 weeks. We think it is reasonable

Table 3 Results of quantile at $\tau = 0.5$ based on 1000 runs

Model		$n = 200$				$n = 500$			
		Bias	SD	SE	CP	Bias	SD	SE	CP
Example 1									
M1	$\hat{\zeta}_{ipw}$	0.061	0.321	0.335	0.927	− 0.002	0.201	0.211	0.945
	$\hat{\zeta}_{mi}$	0.016	0.310	0.327	0.935	− 0.016	0.200	0.205	0.935
	$\hat{\zeta}_{aipw}$	− 0.001	0.311	0.327	0.935	− 0.026	0.200	0.205	0.937
	$\check{\zeta}$	0.010	0.288	0.296	0.936	− 0.016	0.185	0.188	0.924
	$\tilde{\zeta}$	0.347	0.363	0.367	0.824	0.332	0.234	0.233	0.695
	$\hat{\zeta}^*_{ipw}$	0.480	0.285	0.286	0.102	0.489	0.187	0.184	0
M2	$\hat{\zeta}_{ipw}$	0.013	0.329	0.333	0.923	− 0.019	0.196	0.213	0.956
	$\hat{\zeta}_{mi}$	− 0.043	0.319	0.320	0.925	− 0.034	0.194	0.202	0.947
	$\hat{\zeta}_{aipw}$	− 0.062	0.321	0.320	0.917	− 0.049	0.194	0.202	0.941
	$\check{\zeta}$	− 0.019	0.299	0.294	0.924	0.001	0.182	0.189	0.956
	$\tilde{\zeta}$	0.487	0.350	0.355	0.684	0.512	0.214	0.221	0.366
	$\hat{\zeta}^*_{ipw}$	0.479	0.287	0.287	0.094	0.501	0.184	0.186	0
M3	$\hat{\zeta}_{ipw}$	0.071	0.327	0.333	0.926	0.019	0.203	0.215	0.943
	$\hat{\zeta}_{mi}$	− 0.008	0.320	0.320	0.934	− 0.009	0.200	0.203	0.940
	$\hat{\zeta}_{aipw}$	− 0.025	0.318	0.321	0.935	− 0.021	0.200	0.203	0.938
	$\check{\zeta}$	− 0.017	0.296	0.292	0.935	0.004	0.188	0.187	0.934
	$\tilde{\zeta}$	0.466	0.366	0.356	0.720	0.505	0.221	0.228	0.402
	$\hat{\zeta}^*_{ipw}$	0.486	0.296	0.297	0.068	0.501	0.186	0.183	0
M4	$\hat{\zeta}_{ipw}$	0.126	0.331	0.343	0.916	0.027	0.202	0.218	0.944
	$\hat{\zeta}_{mi}$	0.037	0.326	0.331	0.925	− 0.008	0.195	0.204	0.935
	$\hat{\zeta}_{aipw}$	0.016	0.329	0.332	0.923	− 0.022	0.195	0.204	0.938
	$\check{\zeta}$	0.017	0.305	0.298	0.929	− 0.002	0.181	0.187	0.946
	$\tilde{\zeta}$	0.550	0.356	0.362	0.653	0.525	0.227	0.229	0.365
	$\hat{\zeta}^*_{ipw}$	0.512	0.294	0.299	0.097	0.494	0.184	0.187	0
Example 2									
M1	$\hat{\zeta}_{ipw}$	0.034	0.340	0.373	0.960	0.039	0.230	0.227	0.921
	$\hat{\zeta}_{mi}$	− 0.035	0.330	0.364	0.956	− 0.011	0.224	0.225	0.921
	$\hat{\zeta}_{aipw}$	− 0.043	0.337	0.365	0.953	− 0.014	0.225	0.225	0.924
	$\check{\zeta}$	− 0.008	0.290	0.312	0.952	0.011	0.200	0.196	0.927
	$\tilde{\zeta}$	0.316	0.384	0.409	0.878	0.324	0.253	0.258	0.763
	$\hat{\zeta}^*_{ipw}$	1.003	0.304	0.300	0.099	0.999	0.195	0.196	0
M2	$\hat{\zeta}_{ipw}$	0.069	0.346	0.364	0.941	0.041	0.225	0.221	0.925
	$\hat{\zeta}_{mi}$	0.006	0.339	0.352	0.949	− 0.009	0.220	0.218	0.933
	$\hat{\zeta}_{aipw}$	− 0.009	0.340	0.353	0.950	− 0.016	0.220	0.218	0.934
	$\check{\zeta}$	0.008	0.305	0.311	0.940	− 0.011	0.199	0.195	0.931

Table 3 continued

Model		n = 200				n = 500			
		Bias	SD	SE	CP	Bias	SD	SE	CP
	$\tilde{\zeta}$	0.488	0.386	0.388	0.751	0.473	0.245	0.245	0.526
	$\hat{\zeta}^*_{ipw}$	0.996	0.298	0.299	0.097	0.989	0.201	0.200	0
M3	$\hat{\zeta}_{ipw}$	0.067	0.373	0.378	0.923	0.082	0.227	0.233	0.934
	$\hat{\zeta}_{mi}$	− 0.040	0.370	0.364	0.917	− 0.010	0.226	0.229	0.945
	$\hat{\zeta}_{aipw}$	− 0.049	0.373	0.365	0.918	− 0.015	0.227	0.229	0.947
	$\check{\zeta}$	0.002	0.319	0.313	0.928	0.002	0.192	0.200	0.953
	$\tilde{\zeta}$	0.485	0.404	0.412	0.783	0.505	0.261	0.260	0.510
	$\hat{\zeta}^*_{ipw}$	0.988	0.302	0.298	0.095	1.003	0.189	0.195	0
M4	$\hat{\zeta}_{ipw}$	0.087	0.358	0.386	0.952	0.082	0.227	0.233	0.934
	$\hat{\zeta}_{mi}$	− 0.031	0.349	0.375	0.952	− 0.010	0.226	0.229	0.945
	$\hat{\zeta}_{aipw}$	− 0.041	0.353	0.376	0.945	− 0.015	0.227	0.229	0.947
	$\check{\zeta}$	0.012	0.307	0.315	0.940	0.002	0.192	0.200	0.953
	$\tilde{\zeta}$	0.512	0.418	0.416	0.770	0.505	0.261	0.260	0.510
	$\hat{\zeta}^*_{ipw}$	0.976	0.307	0.303	0.101	0.986	0.199	0.197	0

SD standard deviation, *SE* standard error, *CP* coverage probability

Table 4 Estimates (with standard errors in parentheses) for response mean of CD4 data

Regimen	$\hat{\mu}_{ipw}$	$\hat{\mu}_{mi}$	$\hat{\mu}_{aipw}$	$\tilde{\mu}$
(1)	274.87 (12.45)	275.06 (12.07)	274.67 (12.10)	287.62 (13.13)
(2)	338.97 (11.97)	339.46 (11.99)	339.48 (11.98)	341.26 (13.34)
(3)	342.30 (13.01)	343.26 (12.96)	343.02 (12.96)	354.82 (14.18)
(4)	323.84 (11.84)	323.88 (11.98)	323.63 (11.95)	328.79 (13.01)

to assume that, given the six baseline covariates, the missing data propensity does not depend on the CD4 count at 96 ± 5 weeks. Thus, the missing is MAR (Deng and Wang 2017). Due to death and dropout, Y_s has missing values, but X values are fully observed. Specifically, the nonresponse rate of Y_s is about 39.66, 36.21, 35.69 and 37.43%, respectively.

For the HIV study, the CD4 cell count is of prime interest which decreases as HIV progresses. We compute the response mean of Y_s based on $\hat{\mu}_{ipw}$, $\hat{\mu}_{mi}$, $\hat{\mu}_{aipw}$ and $\tilde{\mu}$ for $s = 1, ..., 4$. The point estimates and their standard errors based on the bootstrap are reported in Table 4. It can be seen that the proposed three estimates and their standard errors are close, and smaller than $\tilde{\mu}$ in all cases. In addition, we know that the last three antiretroviral regimens perform better than the first one, since all these estimates indicate that the last three regimens have significantly higher CD4 counts at 96 ± 5 weeks than the first regimen. Also, we compute the distribution function $\Pr(Y_s \leq y)$ for $s = 1, ..., 4$ at $y = 100, 200, \ldots, 700$, based on $\hat{F}_{ipw}$, $\hat{F}_{mi}$, $\hat{F}_{aipw}$ and

Table 5 Estimates (with standard errors in parentheses) for distribution function of CD4 data

y	$\hat{F}_{ipw}$	$\hat{F}_{mi}$	$\hat{F}_{aipw}$	$\tilde{F}$
Regimen (1)				
100	0.182 (0.031)	0.180 (0.031)	0.182 (0.031)	0.162 (0.029)
200	0.348 (0.038)	0.347 (0.038)	0.348 (0.037)	0.318 (0.038)
300	0.573 (0.037)	0.574 (0.036)	0.574 (0.036)	0.539 (0.039)
400	0.785 (0.030)	0.785 (0.030)	0.786 (0.030)	0.763 (0.033)
500	0.913 (0.020)	0.913 (0.020)	0.914 (0.020)	0.900 (0.023)
600	0.965 (0.013)	0.966 (0.013)	0.966 (0.013)	0.960 (0.016)
700	0.990 (0.007)	0.990 (0.007)	0.990 (0.007)	0.988 (0.009)
Regimen (2)				
100	0.077 (0.019)	0.078 (0.020)	0.078 (0.020)	0.078 (0.020)
200	0.190 (0.030)	0.191 (0.030)	0.192 (0.030)	0.189 (0.032)
300	0.437 (0.036)	0.438 (0.036)	0.439 (0.036)	0.432 (0.038)
400	0.678 (0.034)	0.679 (0.034)	0.678 (0.033)	0.673 (0.036)
500	0.852 (0.026)	0.850 (0.026)	0.850 (0.026)	0.847 (0.028)
600	0.935 (0.018)	0.933 (0.018)	0.933 (0.018)	0.931 (0.020)
700	0.961 (0.014)	0.959 (0.014)	0.958 (0.014)	0.958 (0.016)
Regimen (3)				
100	0.088 (0.022)	0.088 (0.022)	0.089 (0.022)	0.080 (0.021)
200	0.219 (0.032)	0.218 (0.032)	0.220 (0.032)	0.196 (0.030)
300	0.423 (0.036)	0.422 (0.035)	0.423 (0.035)	0.389 (0.036)
400	0.624 (0.035)	0.623 (0.034)	0.623 (0.034)	0.593 (0.037)
500	0.835 (0.027)	0.834 (0.027)	0.833 (0.027)	0.819 (0.030)
600	0.952 (0.015)	0.950 (0.016)	0.950 (0.016)	0.947 (0.018)
700	0.973 (0.012)	0.972 (0.012)	0.971 (0.012)	0.970 (0.013)
Regimen (4)				
100	0.091 (0.021)	0.092 (0.021)	0.093 (0.021)	0.088 (0.022)
200	0.240 (0.030)	0.240 (0.030)	0.241 (0.029)	0.236 (0.032)
300	0.458 (0.035)	0.458 (0.035)	0.458 (0.035)	0.453 (0.037)
400	0.714 (0.032)	0.714 (0.032)	0.713 (0.032)	0.707 (0.034)
500	0.854 (0.025)	0.854 (0.025)	0.853 (0.025)	0.846 (0.027)
600	0.940 (0.017)	0.939 (0.017)	0.939 (0.017)	0.932 (0.019)
700	0.975 (0.010)	0.974 (0.011)	0.974 (0.011)	0.969 (0.013)

the estimator $\tilde{F}$ only using the observed Y_s data. The results are reported in Table 5 and the conclusions are similar.

6 Summary

In this paper, three nonparametric M-estimators are developed to handle missing response under the MAR assumption. When the dimension of covariate is high, we show that the dimension of kernel estimators of propensity $\hat{\pi}(X)$ and expected estimating equation $\hat{m}_\varphi(X, \theta)$ can be reduced to alleviate the curse of dimensionality under some conditions. Consistency and asymptotic normality of the proposed estimators are established. Simulation studies and a real data analysis illustrate that the proposed methods have good performance with finite-sample size. In conclusion, the proposed dimension reduction kernel-assisted M-estimators are most appealing in studies with high-dimensional covariates.

Empirical likelihood (EL; Owen 1988; Qin and Lawless 1994) is a competitive and powerful method for constructing confidence intervals, which is also a broadly applicable platform for nonparametric and semiparametric inferences. The proposed dimension reduction estimating equations can be extended to the empirical likelihood approach in the presence of auxiliary information to construct confidence intervals for M-estimators. Throughout this paper, we assume the data are MAR, however, in many applications the data are missing not at random or nonignorable missing. Extension to this case will also be a topic of our future research.

Acknowledgements We are grateful to the Editor, the Associate Editor and one anonymous referee for their insightful comments and suggestions on this article, which have led to significant improvements. This work was supported by the National Natural Science Foundation of China (11501208) and Fundamental Research Funds for the Central Universities.

Appendix

(C1) The true value θ_0 is the unique root of $n^{-1}\sum_{i=1}^n g_l(Y_i, S_i, \delta_i, \theta) = 0$, $n^{-1}\sum_{i=1}^n g_l(Y_i, S_i, \delta_i, \theta)$ is differentiable at $\theta = \theta_0$ for $l = 1, 2, 3$ with $\sum_{i=1}^n \partial g_l(Y_i, S_i, \delta_i, \theta_0)/\partial\theta \neq 0$.

(C2) The function $\varphi(Y, \theta)$ is monotone and continuous in θ, $E|\varphi(Y, \theta)| < \infty$, $\partial\varphi(Y, \theta)/\partial\theta$ is continuous at $\theta = \theta_0$; $E|\partial\varphi(Y, \theta_0)/\partial\theta| < \infty$, $E\{\varphi^2(Y, \theta)|S\} < \infty$.

(C3) The kernel $K(\cdot)$ is bounded and has compact support, and is of order $m \geq 2$, i.e., $\int K(s_1, ..., s_d)ds_1 \cdots ds_d = 1$, $\int s_j^t K(s_1, ..., s_d)ds_1 \cdots ds_d = 0$, and $\int s_j^m K(s_1, ..., s_d)ds_1 \cdots ds_d \neq 0$ for any $j = 1, ..., d$ and $t = 1, ..., m-1$.

(C4) The function $\pi(S)$ and the S-density function $f(S)$ have continuous and bounded partial derivatives with respect to S up to order m, and $\pi(S)$ are bounded away from 0 and 1.

(C5) The function $m_\varphi(S, \theta)$ is twice continuously differentiable in the neighborhood of S; has bounded partial derivatives up to order m.

(C6) As $n \to \infty$, $nh^{2d} \to \infty$, $nh^d/\log n \to \infty$, $nh^{2m} \to 0$, and the estimator $\hat{B}$ obtained by SDR is a root-n consistent estimator of B.

Proof of Theorem 1 For $g_2(Y_i, \hat{S}_i, \delta_i, \theta)$, note that

$$\frac{1}{n}\sum_{i=1}^n g_2(Y_i, \hat{S}_i, \delta_i, \theta) = \frac{1}{n}\sum_{i=1}^n \{\delta_i\varphi(Y_i, \theta) + (1-\delta_i)m_\varphi(S_i, \theta)\} + \frac{1}{n}\sum_{i=1}^n (1-\delta_i)\{\hat{m}_\varphi(\hat{S}_i, \theta) - m_\varphi(S_i, \theta)\},$$

where $S_i = BX_i$ and $\hat{S}_i = \hat{B}X_i$. Define $G(S) = f(S)\pi(S)$ and

$$\hat{G}_n(S) = \frac{1}{n}\sum_{j=1}^{n}\delta_j K_h(S_j - S).$$

Let $\Delta_n(\hat{S}_i, S_i) = \hat{G}_n(\hat{S}) - G(S_i)$. Then,

$$\frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\{\hat{m}_\varphi(\hat{S}_i,\theta) - m_\varphi(S_i,\theta)\} = A_{n1} + A_{n2} - A_{n3},$$

where

$$A_{n1} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(1-\delta_i)K_h(\hat{S}_j - \hat{S}_i)\frac{\delta_j\{\varphi(Y_j,\theta) - m_\varphi(S_j,\theta)\}}{G(S_i)},$$
$$A_{n2} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(1-\delta_i)K_h(\hat{S}_j - \hat{S}_i)\frac{\delta_j\{m_\varphi(S_j,\theta) - m_\varphi(S_i,\theta)\}}{G(S_i)},$$
$$A_{n3} = \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\{\hat{m}_\varphi(\hat{S}_i,\theta) - m_\varphi(S_i,\theta)\}\frac{\Delta_n(\hat{S}_i, S_i)}{G(S_i)}.$$

Using the fact $\delta\varphi(Y,\theta) \perp X|BX$, we can show that

$$E[\delta_j\{\varphi(Y_j,\theta) - m_\varphi(S_j,\theta)\}|X_j] = 0.$$

As in Wang and Chen (2009), we can prove that

$$A_{n1} = \frac{1}{n}\sum_{i=1}^{n}\delta_i\{\pi^{-1}(S_i) - 1\}\{\varphi(Y_i,\theta) - m_\varphi(S_i,\theta)\} + o_p(n^{-1/2}),$$

and $A_{n2} = o_p(n^{-1/2})$. Using the arguments in Andrews (1995) and $\|\hat{B} - B\| = O_p(n^{-1/2})$, it leads to

$$\sup_i \left|\hat{m}_\varphi(\hat{S}_i,\theta) - m_\varphi(S_i,\theta)\right| = o_p(n^{-1/4}),$$

such that $A_{n3} = o_p(n^{-1/2})$. Thus, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}g_2(Y_i,\hat{S}_i,\delta_i,\theta) =& \frac{1}{n}\sum_{i=1}^{n}\{\delta_i\varphi(Y_i,\theta) + (1-\delta_i)m_\varphi(S_i,\theta)\} \\ &+ \frac{1}{n}\sum_{i=1}^{n}\delta_i\{\pi^{-1}(S_i) - 1\}\{\varphi(Y_i,\theta) - m_\varphi(S_i,\theta)\} \\ &+ o_p(n^{-1/2}).\end{aligned}$$

It leads to

$$\frac{1}{n}\sum_{i=1}^{n} g_2(Y_i, \hat{S}_i, \delta_i, \theta) \to E\varphi(Y, \theta).$$

Furthermore, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{g_2(Y_i, \hat{S}_i, \delta_i, \theta) - E\varphi(Y, \theta)\} \to N(0, V(\theta)^2).$$

For $g_1(Y_i, \hat{S}_i, \delta_i, \theta)$, we have

$$\frac{1}{n}\sum_{i=1}^{n} g_1(Y_i, \hat{S}_i, \delta_i, \theta) = \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{\delta_i\varphi(Y_i, \theta)}{\pi(S_i)} + \frac{\delta_i\varphi(Y_i, \theta)\{\pi(S_i) - \hat{\pi}(\hat{S}_i)\}}{\pi^2(S_i)}\Big] + \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i\varphi(Y_i, \theta)\{\pi(S_i) - \hat{\pi}(\hat{S}_i)\}^2}{\pi^2(S_i)\hat{\pi}(\hat{S}_i)}.$$

Using the similar arguments in Wang (2007), we can prove that

$$\frac{1}{n}\sum_{i=1}^{n} g_1(Y_i, \hat{S}_i, \delta_i, \theta) = \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{\delta_i\varphi(Y_i, \theta)}{\pi(S_i)} + \Big\{1 - \frac{\delta_i}{\pi(S_i)}\Big\} m_\varphi(S_i, \theta)\Big],$$

which leads to

$$\frac{1}{n}\sum_{i=1}^{n} g_1(Y_i, \hat{S}_i, \delta_i, \theta) \to E\varphi(Y, \theta),$$

and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{g_1(Y_i, \hat{S}_i, \delta_i, \theta) - E\varphi(Y, \theta)\} \to N(0, V(\theta)^2).$$

For $g_3(Y_i, \hat{S}_i, \delta_i, \theta)$, it can be seen that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} g_3(Y_i, \hat{S}_i, \delta_i, \theta) &= \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{\delta_i\varphi(Y_i, \theta)}{\pi(S_i)} + \Big\{1 - \frac{\delta_i}{\pi(S_i)}\Big\} m_\varphi(S_i, \theta)\Big] \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\Big[\Big\{\frac{\delta_i}{\hat{\pi}(\hat{S}_i)} - \frac{\delta_i}{\pi(S_i)}\Big\}\Big\{\varphi(Y_i, \theta) - m_\varphi(S_i, \theta)\Big\}\Big] \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\Big[\Big\{1 - \frac{\delta_i}{\hat{\pi}(\hat{S}_i)}\Big\}\Big\{\hat{m}_\varphi(\hat{S}_i, \theta) - m_\varphi(S_i, \theta)\Big\}\Big].\end{aligned}$$

Using the similar arguments for $g_1(Y_i, \hat{S}_i, \delta_i, \theta)$ and $g_2(Y_i, \hat{S}_i, \delta_i, \theta)$, it can be proved that the last two terms on the right side of the above equation are $o_p(1)$. The proof is completed. □

Proof of Theorem 2 By Taylor expansion, there exists θ_l^* between $\hat{\theta}_l$ and $\hat{\theta}_0$, $l = 1, 2, 3$, such that

$$\frac{1}{n}\sum_{i=1}^{n} g_l(Y_i, \hat{S}_i, \delta_i, \hat{\theta}_l) = \frac{1}{n}\sum_{i=1}^{n} g_l(Y_i, \hat{S}_i, \delta_i, \theta_0) + \frac{1}{n}\sum_{i=1}^{n} \frac{\partial g_l(Y_i, \hat{S}_i, \delta_i, \theta_l^*)}{\partial\theta}(\hat{\theta}_l - \hat{\theta}_0).$$

Since $\sum_{i=1}^{n} g_l(Y_i, \hat{S}_i, \delta_i, \hat{\theta}_l) = 0$, we have

$$\sqrt{n}(\hat{\theta}_l - \theta_0) = -\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g_l(Y_i, \hat{S}_i, \delta_i, \theta_l^*)}{\partial\theta}\right\}^{-1}\frac{1}{n}\sum_{i=1}^{n} g_l(Y_i, \hat{S}_i, \delta_i, \theta_0).$$

Similar to Theorem 1, as $n \to \infty$, it can be proved that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g_l(Y_i, \hat{S}_i, \delta_i, \theta_l^*)}{\partial\theta} \to E\left\{\frac{\varphi(Y, \theta_0)}{\partial\theta}\right\}.$$

The proof is completed. □

Lemma 1 *Assume that* $P(\delta = 1|X) > 0$ *and* $P(Y = 0|X) = 0$. *For any given y, it can be verified that*

$$\mathcal{S}_{\delta I(Y\le y)|X} \subseteq \mathcal{S}_{\delta Y|X},$$

where $\mathcal{S}$ *denotes for the central subspace (Cook 1994).*

Proof of Lemma 1 Suppose that B is a basis of $\mathcal{S}_{\delta Y|X}$, such that $\delta Y \perp X|BX$. Then, we have $\Pr(\delta Y = 0|X) = \Pr(\delta Y = 0|BX)$ and $\Pr(\delta Y \le y|X) = \Pr(\delta Y \le y|BX)$. Note $\Pr(\delta Y \le y|X) = \Pr(\delta = 1, Y \le y|X) + I(y \ge 0)\Pr(\delta = 0|X)$ and $\Pr(\delta Y = 0|X) = \Pr(\delta = 1, Y = 0|X) + \Pr(\delta = 0|X) = \Pr(\delta = 0|X)$. We have

$$\begin{aligned}
\Pr(\delta I(Y\le y) = 1|X) &= \Pr(\delta = 1, Y \le y|X)\\
&= \Pr(\delta Y \le y|X) - I(y \ge 0)\Pr(\delta = 0|X)\\
&= \Pr(\delta Y \le y|X) - I(y \ge 0)\Pr(\delta Y = 0|X)\\
&= \Pr(\delta Y \le y|BX) - I(y \ge 0)\Pr(\delta Y = 0|BX)\\
&= \Pr(\delta I(Y \le y) = 1|BX).\\
\Pr(\delta I(Y\le y) = 0|X) &= \Pr(\delta = 0|X) + \Pr(\delta = 1, Y \ge y|X)\\
&= \Pr(\delta Y = 0|X) + \Pr(\delta Y \ge y|X) - I(y \le 0)\Pr(\delta = 0|X)\\
&= \Pr(\delta Y \ge y|X) + I(y \ge 0)\Pr(\delta Y = 0|X)\\
&= \Pr(\delta Y \ge y|BX) + I(y \ge 0)\Pr(\delta Y = 0|BX)\\
&= \Pr(\delta I(Y \le y) = 0|BX).
\end{aligned}$$

□

References

Andrews, D. W. (1995). Nonparametric kernel estimation for semiparametric models. *Econometric Theory*, *11*, 560–586.

Chen, X., Wan, A. T., Zhou, Y. (2015). Efficient quantile regression analysis with missing observations. *Journal of the American Statistical Association*, *110*, 723–741.

Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association*, *89*, 81–87.

Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, *89*, 177–189.

Cook, R. D., Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction". *Journal of the American Statistical Association*, *86*, 28–33.

Deng, J., Wang, Q. (2017). Dimension reduction estimation for probability density with data missing at random when covariables are present. *Journal of Statistical Planning and Inference*, *181*, 11–29.

Ding, X., Wang, Q. (2011). Fusion-refinement procedure for dimension reduction with missing response at random. *Journal of the American Statistical Association*, *106*, 1193–1207.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundaker, H., Schooley, R. T., Haubrich, R. H., et al. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *The New England Journal of Medicine*, *335*, 1081–1089.

Hu, Z., Follmann, D. A., Wang, N. (2014). Estimation of mean response via effective balancing score. *Biometrika*, *101*, 613–624.

Huber, P. J. (1981). *Robust statistics*. New York: Wiley.

Ibrahim, J. G., Chen, M. H., Lipsitz, S. R., Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, *100*, 332–346.

Kim, J. K., Shao, J. (2013). *Statistical methods for handling incomplete data*. London: Chapman and Hall/CRC.

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, *86*, 316–327.

Li, Y., Wang, Q., Zhu, L., Ding, X. (2017). Mean response estimation with missing response in the presence of high-dimensional covariates. *Communications in Statistics-Theory and Methods*, *46*, 628–643.

Ma, Y., Zhu, L. (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association*, *107*, 168–179.

Ma, Y., Zhu, L. (2013). A review on dimension reduction. *International Statistical Review*, *81*, 134–150.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, *75*, 237–249.

Qin, J., Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, *22*, 300–325.

Rubins, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.

Serfling, R. J. (1981). *Approximation theorems of mathematical statistics*. New York: Wiley.

Shao, J., Wang, L. (2016). Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika*, *103*, 175–187.

Wang, D., Chen, S. X. (2009). Empirical likelihood for estimating equations with missing values. *The Annals of Statistics*, *37*, 490–517.

Wang, L., Rotnitzky, A., Lin, X. (2010). Nonparametric regression with missing outcomes using weighted kernel estimating equations. *Journal of the American Statistical Association*, *105*, 1135–1146.

Wang, Q. (2007). M-estimators based on inverse probability weighted estimating equations with response missing at random. *Communications in Statistics-Theory and Methods*, *36*, 1091–1103.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, *141*, 1281–1301.

Xia, Y., Tong, H., Li, W. K., Zhu, L. X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B*, *64*, 363–410.

Xue, L. (2009). Empirical likelihood confidence intervals for response mean with data missing at random. *Scandinavian Journal of Statistics*, *36*, 671–685.

Zhang, B. (1995). M-estimation and quantile estimation in the presence of auxiliary information. *Journal of Statistical Planning and Inference*, *44*, 77–94.

Zhu, L. P., Zhu, L. X., Ferre, L., Wang, T. (2010). Sufficient dimension reduction through discretization-expectation estimation. *Biometrika*, *97*, 295–304.