Published in final edited form as:

Adv Stat Anal. 2017 July ; 101(3): 267–288. doi:10.1007/s10182-016-0283-y.

Prediction model-based kernel density estimation when group membership is subject to missing

Hua He¹, Wenjuan Wang², and Wan Tang^{3,*}

¹Department of Epidemiology, Tulane University School of Public Health and Tropical Medicine

²Brightech International, LLC

³Department of Biostatistics & Bioinformatics, Tulane University School of Public Health and Tropical Medicine

Abstract

The density function is a fundamental concept in data analysis. When a population consists of heterogeneous subjects, it's often of great interest to estimate the density functions of the subpopulations. Nonparametric methods such as kernel smoothing estimates may be applied to each subpopulation to estimate the density functions if there are no missing values. In situations where the membership for a subpopulation is missing, kernel smoothing estimates using only subjects with membership available are valid only under missing complete at random (MCAR). In this paper, we propose new kernel smoothing methods for density function estimates by applying prediction models of the membership under the missing at random (MAR) assumption. The asymptotic properties of the new estimates are developed, and simulation studies and a real study in mental health are used to illustrate the performance of the new estimates.

Keywords

density function; kernel smoothing estimate; missing at random (MAR); prediction model; mean score method

1 Introduction

The probability density function is a fundamental concept in understanding a univariate continuous outcome, and the kernel smoothing approach is widely used to estimate the density function (Parzen, 1962; Simonoff, 1996; Wand and Jones, 1995). If a population consists of heterogeneous subjects, it is often of great interest to estimate the density function for some subpopulations. If the outcome of interest and the membership are both available, the problem of kernel smoothing has been thoroughly studied. Wang (2008) generalized the kernel smoothing method to situations with missing values in the outcome but with the membership available for every subject. In practice, it is common that the membership for some subjects may not be known when they are initially sampled from the target population. For example, two-stage study designs (White, 1982) are widely used in

*Corresponding author: Wan Tang, wtang1@tulane.edu, phone: 1-504-988-1887, fax: 1-504-988-1706.

medical research, especially in diagnostic test studies. In a two-stage study, it is often the case that subjects are first sampled from a target population, and then their memberships are ascertained subsequently with some gold standard test. When the membership is defined by the onset of some kind of disease, and the gold standard test for diagnosis is too expensive or involves intensive procedures such as surgeries, subjects who are at lower risk of the disease may not be administered with the gold standard test for the disease and thus yield missing values in the membership for these subjects. For instance, in a study of depression among postpartum women conducted by Chaudron et al. (2010), a total of 419 postpartum women initially consented to participate in the study, and their demographic information and screening test results were collected, but the depression status was not known at initial visit for these subjects. Subsequently, only 198 subjects underwent the structural clinical interview for DSM-IV diagnosis (SCID) (Spitzer et al., 1994), which is in general viewed as the gold-standard test for diagnosing depression. Since SCID involves hours of interviews by a psychiatrist with the subjects, it is very expensive and inconvenient for the subjects to undergo the SCID diagnosis. So in depression studies, especially for elderly patients, it is common that subjects with very low likelihood of being depressed may not be administered with SCID, thus the memberships for the SCID diagnoses for those subjects are missing

In situations with the membership missing but with outcomes available for all subjects, the Naive approach that uses only those subjects with known membership in the analysis is valid only in the special situation when the membership is missing completely at random (MCAR), which means that the missingness of the membership does not depend on any observed and unobserved information. This is a very stringent assumption and it is often violated in practice. For example, in the postpartum depression study, mothers having low scores on screening test and low incomes were less likely to undergo the SCID diagnosis. In such situations, the missingness is very likely not MCAR. When the missingness of the membership does depend on some observed information, but not on unobserved information, it is called missing at random (MAR). If the missingness of the membership depends on some unobserved information, the missingness is called missing not at random (MNAR) (Little and Rubin, 2002). For example, if the decision to undergo SCID diagnosis depends on some preliminary tests and incomes, when both the tests and incomes are collected, the missingness of SCID can be treated as MAR. But if some tests or income information is not collected, we would assume that the missingness is MNAR. When adequate information regarding the membership is collected, it's reasonable to assume MAR. When the missingness of the membership is MAR, Tang et al. (2012) developed a kernel smoothing method to estimate the density function of subpopulations by incorporating the inverse probability weighting (IPW) approach into the kernel smoothing estimation. The IPW approach dates back to Horvitz and Thompson (1952), and was utilized to effectively correct the bias under the MAR assumption (Alonzo and Pepe, 2005; Alonzo et al., 2003; He et al., 2009; He and McDermott, 2012; Tang et al., 2012). The IPW approach relies on our knowledge regarding the missing mechanism. Although it has the ability to reduce the bias, the IPW approach suffers from several limitations. Firstly, the IPW approach requires that the probabilities of the membership being observed are bounded away from zero, i.e., every subject must have delegates with membership being observed. However, this may not be true in practice. It is common that the subjects with negative test results may not administer the

gold standard test, especially if the gold standard test is expensive and/or invasive. In other words, some subjects may have a very low or zero probability for their membership to be ascertained, and hence the IPW approach does not work in those situations. Secondly, when the probabilities for the membership to be ascertained are very small for some subjects, the IPW method yields very unstable estimates, as these subjects have very large weights in estimating the density function, and hence may lead to biased estimates and high variances in finite samples. Lastly, the IPW method uses only the subjects with the membership available in estimating the density function, while the subjects with missing memberships are ignored, even though some information on these subjects is utilized in modeling the missing mechanism. Excluding those subjects with missing memberships can yields less efficient estimates, especially when information that highly relates to the membership is collected. In such situations, the membership can be reliably predicted by subjects' information. Hence another natural, meaningful approach is to correct the bias by filling in the missing membership with predicted values.

In this paper, we develop new kernel smoothing methods to estimate the density function of a subpopulation based on a prediction model for the missing membership. For comparison purpose as well as a way to introduce notation, we first give a brief review of the IPW based kernel estimation in Section 2. We then propose the new approach based on a prediction model in Section 3. Bandwidth selection for the proposed smoothing approaches is discussed in Section 4. Simulation studies are carried out in Section 5 to examine the performance of the approach, with a real study example given in Section 6. The paper concludes with a discussion about pro and cons of the methods in Section 7.

2 Existing Methods

To simplify the notation, we consider the cases where there are only two subpopulations, say diseased and non-diseased subpopulations. It is straightforward to generate the methods we will discuss below to the situations with multiple subpopulations. Let $(\mathbf{x}_i, T_i, D_i)$ be a simple random sample from a population of interest, where D_i is a membership indicator of disease status (diseased and non-diseased), $\mathbf{x}_i$ is a vector of covariates, and T_i is a univariate outcome variable of interest. We are interested in estimating the density function of T_i for the diseased and non-diseased groups. Assume $D_i = 1$ for the diseased group, and $D_i = 0$ for the non-diseased group. In this paper, we focus on the density function estimates for the diseased group, as a similar argument can be applied to the non-diseased group. In the complete data case where D_i is observed for each subject, standard univariate kernel smoothing methods can be applied to the subsample of $D_i = 1$. Note that in this case, the covariates $\mathbf{x}_i$ are not used in estimating the density function of T_i .

Let $f(t)$ denote the probability density function of T_i for the group of $D_i = 1$. Assume that the second derivative of f is square integrable,

$$R(f'') = \int f''(t)^2 dt < \infty. \quad (2.1)$$

Let $K(\cdot)$ denote a symmetric density function with a finite second moment, i.e.,

$$K \geq 0, \int K(t)dt=1, K(-t)=K(t) \text{ and } \mu_2(K)=\int t^2 K(t)dt<\infty. \quad (2.2)$$

Also, let $g(t)$ be the density of T_i in the whole population and $m=\sum_{i=1}^n D_i$ be the number of subjects with $D_i=1$ in the sample.

We assume that the covariates $\mathbf{x}_i$ and the outcome variable T_i are always observed. Let R_i be the indicator of whether D_i is observed: $R_i=1$ if it is observed, and $R_i=0$ if otherwise. Thus, $(\mathbf{x}_i, T_i, R_i)$ is always observed, but D_i is only observed for those subject with $R_i=1$. The Naive approach would be to simply exclude the subjects with missing membership indicator D . If a standard kernel smoothing method is applied to the subgroup of those subjects with $R_i=1$ and $D_i=1$, we obtain the following Naive estimate

$$\tilde{f}_{\text{Naive}}(t)=\frac{\sum_{R_i=1 \text{ and } D_i=1} K_h(t-T_i)}{\sum_{R_i=1 \text{ and } D_i=1} 1}=\frac{\sum_{i=1}^n R_i D_i K_h(t-T_i)}{\sum_{i=1}^n R_i D_i}. \quad (2.3)$$

The Naive estimate is valid only in very stringent MCAR situations. Tang et al. (2012) developed kernel smoothing estimates based on an IPW approach under the MAR assumption, i.e.

$$R \perp D|(T, \mathbf{x}). \quad (2.4)$$

The missingness of the membership may depend on the membership of disease or not, and can only be dependent through the outcome and covariates. In other words, the information about the missing membership is fully captured by the outcome and the covariates. In practice, as long as enough information related to the membership is collected, the MAR assumption is reasonable.

The IPW estimate was defined as

$$\tilde{f}_{IPW}(t)=\frac{\sum_{R_i=1 \text{ and } D_i=1} \frac{1}{\pi_i} K_h(t-T_i)}{\sum_{R_i=1 \text{ and } D_i=1} \frac{1}{\pi_i}}=\frac{\sum_{i=1}^n \frac{D_i R_i}{\pi_i} K_h(t-T_i)}{\sum_{i=1}^n \frac{D_i R_i}{\pi_i}}, \quad (2.5)$$

where $\pi_i=\Pr(R_i=1|T_i, \mathbf{x}_i)$ is the probability that the group membership is observed for the i th subject. From (2.5), π_i cannot take value 0. In fact, in order to warrant a good performance of the IPW approach, π_i should be bounded away from 0, i.e.,

$$\pi_i > c > 0, \quad (2.6)$$

where c is a positive constant. In practice, we often need to model the missing mechanism and estimate π_i .

From (2.5), it is clear that the IPW method uses T_i only for those subjects with $R_i = 1$ and $D_i = 1$ in estimating the density function. Information regarding $\mathbf{x}$ and T for those subjects with unknown memberships are used only in modeling the missing mechanisms. When the disease mechanism is well understood and the relevant information regarding the disease is collected, it is possible to estimate the membership. In the next section, we will discuss how to estimate the density function by incorporating a prediction model of the missing membership into the kernel smoothing method.

3 Methods Based on Prediction Models

Let $d_i = \Pr(D_i = 1 | x_i, T_i)$ be the probability of the i th subject belonging to the diseased subpopulation given the observed covariates. Intuitively, one may want to replace the missing membership with the likelihood of being in diseased group. This naturally suggests the following mean score (MS) estimate

$$\tilde{f}_{MS}(t; h) = \frac{\frac{1}{n} \sum_{i=1}^n [D_i R_i + d_i (1 - R_i)] K_h(t - T_i)}{\frac{1}{n} \sum_{i=1}^n [D_i R_i + d_i (1 - R_i)]}. \quad (3.1)$$

Note that under the MAR assumption (2.4), the denominator of (3.1) provides a consistent estimate of the prevalence $\Pr(D_i = 1)$, while the numerator provides a consistent estimate of $\Pr(D_i = 1) E[K_h(t - T_i) | D_i = 1]$. Thus, (3.1) provides a consistent estimate of $E[K_h(t - T_i) | D_i = 1]$, which is the mean of the kernel estimate in the complete data case.

The MS method was proposed by Pepe et al. (1994) and Reilly and Pepe (1995), and utilized by Alonzo and Pepe (2005) to correct verification bias. In our MS estimate, the missing disease statuses are replaced by their respective mean scores (predicted probabilities of being diseased). More precisely, in addition to using the subjects with confirmed diseases ($D = 1$), we also use all subjects with missing disease memberships in estimating the density function of the diseased group. However, among the subjects with missing disease membership, those subjects with little likelihood of being in the diseased group contribute little to the density function estimate. As the likelihood of being diseased increases, the contribution to the density function estimate increases too. A subject with an estimated 100% likelihood of being diseased can contribute as much as a subject with a confirmed disease when estimating the density function.

In the literature of the correction of verification bias in diagnostic studies, one may also replace the disease statuses by their corresponding scores, whether observed or not (Alonzo

et al., 2003; Begg and Greenes, 1983; He et al., 2009). We may also follow a similar strategy and obtain the following estimate

$$\tilde{f}_{BG}(t;h) = \frac{\frac{1}{n} \sum_{i=1}^n d_i K_h(t - T_i)}{\frac{1}{n} \sum_{i=1}^n d_i}. \quad (3.2)$$

Since Begg and Greenes (1983) initially proposed the approach in the ROC analysis of diagnostic studies, we call this density function estimate the Begg-Greenes (BG) style kernel smoothing estimate. The BG approach uses all the subjects in estimating the density function, and the contribution to the estimate for a subject depends on the likelihood that the subject is diseased based on her/his test result and covariate information. Similar to the MS estimate (3.1), the denominator of (3.2) provides a consistent estimate of the prevalence and the numerator provides a consistent estimate of $E[K_h(t - T_i) | D_i = 1]$, and hence (3.2) itself provides an consistent estimate of $f_h(t) = E[K_h(t - T_i) | D_i = 1]$, which is not the density function $f(t)$ itself.

More precisely, we have the following asymptotic theory for the MS and BG estimates:

Theorem 1 For any fixed h , let $u_i = D_i R_i + d_i(1 - R_i)$, and $f_h(t) = E[K_h(t - T_i) | D_i = 1]$.

(a) Under assumptions (2.1), (2.2), and (2.4), at point t , as $n \rightarrow \infty$, we have

$$\sqrt{n} [\tilde{f}_{MS}(t;h) - f_h(t)] \rightarrow N(0, \sigma_1^2(h)), \quad (3.3)$$

where $\sigma_1^2(h) = \frac{1}{p^2} \text{Var}\{u_i [K_h(t - T_i) - f_h(t)]\}$.

(b) Under assumptions (2.1) and (2.2), at point t , as $n \rightarrow \infty$, we have

$$\sqrt{n} [\tilde{f}_{BG}(t;h) - f_h(t)] \rightarrow N(0, \sigma_2^2(h)), \quad (3.4)$$

where $\sigma_2^2(h) = \frac{1}{p^2} \text{Var}\{d_i [K_h(t - T_i) - f_h(t)]\}$.

A proof of Theorem 1 is provided in the Appendix.

Based on the asymptotic distribution of $\tilde{f}_{MS}(t, h)$ and $\tilde{f}_{BG}(t, h)$, the asymptotic bias and variance of the estimates can be derived and they are given by the following theorem.

Theorem 2 Let $p = \Pr(D_i = 1)$ be the prevalence of the disease ($D_i = 1$) in the target population. (a) Under assumptions (2.1), (2.2), and (2.4), as $h \rightarrow 0$ and $nh \rightarrow \infty$, the asymptotic bias and variance of $\tilde{f}_{MS}(t, h)$ at point t are

$$\begin{aligned} \text{Bias}\{\tilde{f}_{MS}(t;h)\} &= \frac{1}{2}h^2\mu_2(K)f''(t)+o(h^2), \\ \text{Var}\{\tilde{f}_{MS}(t;h)\} &= \frac{g(t)R(K)}{np^2h}E[d_i^2+\pi_i(d_i-d_i^2)|T_i=t]+o\left(\frac{1}{nh}\right); \end{aligned} \quad (3.5)$$

(b) Under assumptions (2.1) and (2.2), as $h \rightarrow 0$ and $nh \rightarrow \infty$, the asymptotic bias and variance of $\tilde{f}_{BG}(t, h)$ at point t are

$$\begin{aligned} \text{Bias}\{\tilde{f}_{BG}(t;h)\} &= \frac{1}{2}h^2\mu_2(K)f''(t)+o(h^2), \\ \text{Var}\{\tilde{f}_{BG}(t;h)\} &= \frac{g(t)R(K)}{np^2h}E(d_i^2|T_i=t)+o\left(\frac{1}{nh}\right). \end{aligned} \quad (3.6)$$

A proof of Theorem 2 is also given in the Appendix. Note that the MS estimate requires MAR assumption, and the asymptotic variance depends on π_i , the probability of the membership being observed. But the BG estimate does not require the MAR assumption and the estimate does not involve the missing mechanism. From both Theorem 1 and Theorem 2, the $\tilde{f}_{MS}(t, h)$ and $\tilde{f}_{BG}(t, h)$ have the same asymptotic bias as that of the bias in complete data case. Because $d_i^2 < d_i$, the BG estimate generally has a smaller asymptotic variance compared to the MS method according to Theorem 2.

In practice, d_i is often unknown and needs to be estimated from some prediction models. The d_i can be estimated using either non-parametric models, such as generalized additive model,

$$g(\Pr(D_i=1|\mathbf{x}_i, T_i)) = \beta_0 + f_1(T_i) + f_2(\mathbf{x}_i), \quad (3.7)$$

or a parametric model, such as a logistic regression model,

$$\Pr(D_i=1|\mathbf{x}_i, T_i) = g(\mathbf{x}_i, T_i; \beta), \quad (3.8)$$

where $g(\cdot; \beta)$ is a known function indexed by some finite-dimensional parameters β . The parameter β in (3.8) can be estimated using the maximum likelihood method, i.e., based on the following score equation

$$\frac{1}{n} \sum_{i=1}^n R_i \frac{\frac{\partial}{\partial \beta} g(\mathbf{x}_i, T_i; \beta)}{g(\mathbf{x}_i, T_i; \beta)(1 - g(\mathbf{x}_i, T_i; \beta))} (D_i - g(\mathbf{x}_i, T_i; \beta)) = 0.$$

Note that unlike modelling the missing mechanism, the prediction model (3.8) or (3.7) itself has the missing value issue, as some of the responses are missing. However, under the MAR assumption (2.4), it is valid to estimate β based on the prediction model (3.8) or f_1 and f_2

using only the subset of subjects with D , T and $\mathbf{x}$ available since the missing responses depend only on T and $\mathbf{x}$, which are always observed. Thus, only subjects with verified memberships are used in the score equation, i.e., the sample size for model (3.8) is $\sum_{i=1}^n R_i$.

More generally, one may use estimating equations (EE) for the prediction model (3.8). In general, an EE for model (3.8) has the form:

$$\frac{1}{n} \sum_{i=1}^n \Psi_i(x_i; \beta) = \frac{1}{n} \sum_{i=1}^n R_i \frac{\partial}{\partial \beta} g(\mathbf{x}_i, T_i; \beta) V_i^{-1} (D_i - g(\mathbf{x}_i, T_i; \beta)) = 0, \quad (3.9)$$

where V_i can be any function of $\mathbf{x}_i$, T_i and β , and usually we may choose it as the variance of $D_i | \mathbf{x}_i, T_i$ i.e., $V_i = g(\mathbf{x}_i, T_i; \beta) (1 - g(\mathbf{x}_i, T_i; \beta))$.

For example, we may use the following logistic regression model to estimate d_i

$$\text{logit}[\Pr(D_i=1|\mathbf{x}_i, T_i)] = \beta \mathbf{x}_i.$$

To simplify the notation, we have subsumed the constant term of the logistic regression as well as T into the vector of covariates $\mathbf{x}$. The score equations based on the maximum likelihood method for the logistic model is the special case of (3.9) with $\Psi_i(\mathbf{x}_i; \beta) = R_i(D_i - g_i(\beta))\mathbf{x}_i$.

Plugging in the predicted values $\hat{d}_i$ in (3.1) and (3.2), the density function based on the MS approach can be estimated by

$$\hat{f}_{MS}(t; h) = \frac{\frac{1}{n} \sum_{i=1}^n [D_i R_i + \hat{d}_i (1 - R_i)] K_h(t - T_i)}{\frac{1}{n} \sum_{i=1}^n [D_i R_i + \hat{d}_i (1 - R_i)]} \quad (3.10)$$

and the estimate of the density function based on the BG approach is given by

$$\hat{f}_{BG}(t; h) = \frac{\frac{1}{n} \sum_{i=1}^n \hat{d}_i K_h(t - T_i)}{\frac{1}{n} \sum_{i=1}^n \hat{d}_i}. \quad (3.11)$$

For d'_{iS} that are estimated from a parametric model, the variation from estimated d'_{iS} should be taken into account when we estimate the density function. The following theorem gives the asymptotic properties of the kernel density estimates based on the MS and BG approaches.

Theorem 3 Assume (2.1), (2.2), (2.4), (3.8) and (3.9). Let $\mathbf{I} = -E \frac{\partial \Psi_i}{\partial \beta^T}$. For any fixed h , at point t , as $n \rightarrow \infty$, we have

(a)

$$\sqrt{n} [\hat{f}_{MS}(t; h) - f_h(t)] \rightarrow N(0, \sigma_3^2(h)), \quad (3.12)$$

where

$$\sigma_3^2(h) = \frac{1}{p^2} \text{Var}\{u_i[K_h(t - T_i) - f_h(t)] + (\mathbf{c} - f_h(t)\mathbf{d})\mathbf{I}^{-1}\Psi_i\}, \mathbf{c} = E[K_h(t - T_i)(1 - R_i)\frac{\partial g_i}{\partial \beta^T}(\beta)]$$

, and $\mathbf{d} = E\left[(1 - R_i)\frac{\partial g_i}{\partial \beta^T}(\beta)\right]$.

(b)

$$\sqrt{n} [\hat{f}_{BG}(t) - f_h(t)] \rightarrow N(0, \sigma_4^2(h)), \quad (3.13)$$

where

$$\sigma_4^2(h) = \frac{1}{p^2} \text{Var}\{d_i[K_h(t - T_i) - f_h(t)] + (\mathbf{c} - f_h(t)\mathbf{d})\mathbf{I}^{-1}\Psi_i\}, \mathbf{c} = E[K_h(t - T_i)\frac{\partial g_i}{\partial \beta^T}(\beta)],$$

and $\mathbf{d} = E\left[\frac{\partial g_i}{\partial \beta^T}(\beta)\right]$.

A proof of Theorem 3 is provided in the Appendix. Comparing (3.12) and (3.13) to (3.3) and (3.4), the variances for the asymptotic distributions of $\hat{f}_{MS}(t, h)$ and $\hat{f}_{BG}(t, h)$ involve one more term $(\mathbf{c} - f_h(t)\mathbf{d})\mathbf{e}^{-1}\Psi_i$ due to the variation of estimated $\hat{d}_i$. The following theorem gives the asymptotic bias and variance for the estimates of $\hat{f}_{MS}(t, h)$ and $\hat{f}_{BG}(t, h)$.

Theorem 4 Under assumptions (2.1), (2.2), (2.4) and (3.9), as $h \rightarrow 0$ and $nh \rightarrow \infty$, we have

(a) the asymptotic bias and variance of $\hat{f}_{MS}(t, h)$ at point t are

$$\text{Bias}\{\hat{f}_{MS}(t; h)\} = \frac{1}{2} h^2 \mu_2(K) f''(t) + o(h^2), \quad (3.14)$$

$$\text{Var}\{\hat{f}_{MS}(t; h)\} = \frac{g(t)R(K)}{np^2 h} E[d_i^2 + \pi_i(d_i - d_i^2) | T_i = t] + o\left(\frac{1}{nh}\right). \quad (3.15)$$

(b) the asymptotic bias and variance of $\hat{f}_{BG}(t, h)$ at point t are

$$\begin{aligned} \text{Bias}\{\hat{f}_{BG}(t;h)\} &= \frac{1}{2} h^2 \mu_2(K) f''(t) + o(h^2), \\ \text{Var}\{\hat{f}_{BG}(t;h)\} &= \frac{g(t)R(K)}{np^2 h} E(d_i^2 | T_i=t) + o\left(\frac{1}{nh}\right). \end{aligned} \quad (3.16)$$

A proof of Theorem 4 is also given in the Appendix. By comparing Theorems 2 and 4, we see that the kernel density estimate with parameters known or estimated has the same asymptotic bias and variance. The variations associated with the estimation of the parameters are of order $1/n$, and thus are dominated by the main terms in the variance formula, which is of order $1/(nh)$. Based on Theorem 4, the BG estimate asymptotically has a smaller variance than the respective MS estimate.

4 Bandwidth Selection

It is obvious from Theorems 2 and 4 that the behaviors of the estimates are closely related to the bandwidth h we choose. If h is chosen too small, then the estimates are less biased, but more variant. On the other hand, if h is chosen too big, the bias would be large. When we select the bandwidth, we need to consider the trade off between bias and variance. The mean squared error, which is defined as the sum of variance and the square of the bias, is in general used as a summary statistic for overall performance of the estimator at given point t . Based on Theorems 2 and 4, we immediately have the following theorem:

Theorem 5 (a) the MSEs for $\tilde{f}_{MS}$ and $\hat{f}_{MS}$ at point t can be expressed as

$$\text{MSE}(t;h) = \frac{1}{4} h^4 \mu_2^2(K) f''(t)^2 + \frac{g(t)R(K)}{np^2 h} E[d_i^2 + \pi_i(d_i - d_i^2) | t] + o\left(\frac{1}{nh} + h^4\right). \quad (4.1)$$

(b) the MSEs of $\tilde{f}_{BG}$ and $\hat{f}_{BG}$ at point t can be expressed as

$$\text{MSE}(t;h) = \frac{1}{4} h^4 \mu_2^2(K) f''(t)^2 + \frac{g(t)R(K)}{np^2 h} E(d_i^2 | t) + o\left(\frac{1}{nh} + h^4\right). \quad (4.2)$$

The MSE assesses the behavior of the estimate at a single point t , with a smaller MSE indicating a better fit. Asymptotic local optimal bandwidths at a point t can be obtained by minimizing the corresponding MSE. To assess the behavior of the estimates with a common bandwidth over the entire range, the integration of MSE over the whole range (MISE) can be used for this purpose. The following theorem gives the MISE for the proposed kernel density estimates.

Theorem 6 (a) the MISEs for $\tilde{f}_{MS}$ and $\hat{f}_{MS}$ are given by

$$\text{MISE}(h) = \frac{1}{4} h^4 \mu_2^2(K) R(f'') + \frac{c_0}{np^2 h} R(K) + o\left(\frac{1}{nh} + h^4\right), \quad (4.3)$$

where $c_0 = \int E[d_i^2 + \pi_i(d_i - d_i^2)]g(t)dt$.

(b) the MISEs for $\tilde{f}_{BG}$ and $\hat{f}_{BG}$ are given by

$$\text{MISE}(h) = \frac{1}{4} h^4 \mu_2^2(K) R(f'') + \frac{c_1}{np^2 h} R(K) + o\left(\frac{1}{nh} + h^4\right), \quad (4.4)$$

where $c_1 = \int E(d_i^2 | t) g(t) dt = E(d_i^2)$.

Similar to the MSE, the smaller the MISE the better the estimate over the entire range. Optimal bandwidths can be selected by minimizing the MISE in (4.3) and (4.4) for the MS and BG estimates, respectively. It is easy to understand that the minimum can be achieved,

asymptotically, when $h^4 \mu_2^2(K) R(f'') = \frac{c_0}{np^2 h} R(K)$ for the MS estimate and

$h^4 \mu_2^2(K) R(f'') = \frac{c_1}{np^2 h} R(K)$ for the BG estimate. Thus, the asymptotic optimal bandwidths are given by

$$h_{\text{MISE}} = \left[\frac{c_0 R(K)}{\mu_2^2(K) R(f'') np^2} \right]^{1/5} \text{ and } \left[\frac{c_1 R(K)}{\mu_2^2(K) R(f'') np^2} \right]^{1/5} \quad (4.5)$$

respectively, for MS and BG estimates. In (4.5), $R(K)$ and $\mu_2(K)$ can be easily computed for a given kernel function. For the Epanechnikov kernel function

$$K(x) = \begin{cases} \frac{3}{4\sqrt{5}}(1 - \frac{1}{5}x^2) & \text{if } |x| < \sqrt{5} \\ 0 & \text{if otherwise} \end{cases},$$

we have $\mu_2(K) = 1$ and $R(K) = 0.6$. We can also compute an empirical estimate for c_0 and c_1 based on the fitted values of d_i and π_i . More precisely, c_0 and c_1 can be estimated by

$\frac{1}{n} \sum_{i=1}^n \left[(1 - \hat{\pi}_i) \widehat{d_i^2} + \widehat{d_i} \hat{\pi}_i \right]$ and $\frac{1}{n} \sum_{i=1}^n \widehat{d_i^2}$ respectively. The computation of $R(f'')$ is a bit involved since it involves the estimate of f'' . One common practice is to use the normal distribution as an initial choice for optimal bandwidth selection. If f follows a normal distribution with variance σ^2 , the corresponding $R(f'') \approx 0.212 \sigma^{-5}$. Under this approach, we first estimate the variance of the outcome to obtain an estimate of σ , and then substitute $0.212 \sigma^{-5}$ in place of $R(f'')$ in (4.5) to obtain an estimate of the optimal bandwidth. We may further estimate f'' using the bandwidth we obtained based on the normal distribution

assumption, and then substitute the estimate of $R(f'')$ in (4.5) to update the estimate of the optimal bandwidth.

5 Simulation Studies

Simulation studies are conducted to assess the behaviors of the proposed estimates. Similar to that in Tang et al. (2012), we consider the situation where T is a binormal, but truncated by $[-1, 1]$. More precisely, $T|D = 1$ ($T|D = 0$) follows $N(0, 1)$ ($N(1, 1)$) truncated on $[-1, 1]$. There are no other covariates $\mathbf{x}_i$ involved. The missingness of group membership D depends only on T , with

$$\pi_i = \Pr(D_i \text{ observed} | T_i) = \frac{\exp(\beta_0 + \beta_1 T_i)}{1 + \exp(\beta_0 + \beta_1 T_i)}. \quad (5.1)$$

Under (5.1), β_0 can be used to control the missing rate, while β_1 indicates the degree of deviation from MCAR. The missing indicator R_i is generated from a Bernoulli (π_i). In the simulation study, the prevalence of the disease in the population is fixed at .5, i.e., $p = .5$. The Epanechnikov kernel function is used for density estimates. We consider a moderate sample size of 200. From Section 2 and Section 3, the IPW method requires $\Pr(R|T, X)$, and the BG and MS methods require $\Pr(D|T, X)$. In most cases, the $\Pr(R|T, X)$ and $\Pr(D|T, X)$ are not known and need to be estimated. Hence in the simulation studies, we consider logistic regression models for both R and D , i.e., A logistic regression model

$$\text{logit}(\Pr(R_i | T_i)) = \beta_0 + \beta_1 T_i$$

is used to estimate $\hat{\pi}_i$ for the IPW method, and another logistic regression model

$$\text{logit}(\Pr(D_i | T_i)) = \alpha_0 + \alpha_1 T_i \quad (5.2)$$

is used to estimate $\hat{d}_i$ for the BG and MS methods. The functions we estimate in the simulations are $\hat{f}_{MS}$ and $\hat{f}_{BG}$ in (3.10) and (3.11), respectively for MS and BG methods.

Setting $\beta_0 = 0$ and $\beta_1 = 1$ in (5.1), in this case, about 53% subjects have missing membership. We computed the integrated squared bias, the integrated variance, as well as the MISE for bandwidths from 0.01 to 1 based on a Monte Carlo size of 1000. We considered the Naive, IPW, MS and BG methods in the simulations. Shown in Figure 1 are the MISEs as a function of bandwidth for the four estimates. For any given h , the MS and BG estimates have uniformly better performance than the Naive and IPW methods in terms of MISE.

To compute the optimal bandwidth based on (4.5), we use a normal distribution as a reference to compute $R(f'')$. Since the variance of the standard normal truncated by $[-1, 1]$ is

$1 - \frac{2\phi(1)}{\Phi(1) - \Phi(-1)} \approx 0.29$, where ϕ and Φ are the density and cumulative distribution functions of the standard normal, $R(f'') = 0.212 \times 0.29^{-5/2} = 4.681$ for the optimal bandwidth computation. Since $\mu_2(K) = 1$, $R(K) = 0.66$, $E(1/\pi_i) = 2.15$, $c_0 = 0.37$, and $c_1 = 0.27$, it follows that the optimal bandwidth based on the asymptotic formulas for the complete-data, Naive, IPW, MS, and BG estimates are 0.27, 0.30, 0.31, 0.25 and 0.23, respectively. According to the simulation, the minimum values of MISE for these estimates occur at 0.27, 0.41, 0.32, 0.27 and 0.27, respectively, which indicate that the asymptotic optimal bandwidth formulas provide good estimates for bandwidths even for moderate sample sizes such as 200, except for the Naive method.

Figure 2 presents the kernel density function estimates under different scenarios. To assess the effects of missing data mechanisms, we vary the value of β_1 to reflect different missing mechanisms and the amount of missing values. By fixing $\beta_0 = 0$, but varying β_1 from 0 to 1, the missing mechanism changed from MCAR ($\beta_1 = 0$) to MAR, with a larger β_1 indicating more deviations from MCAR. The amount of the missing memberships ranges from 50% to 53%. For each of the methods, the estimated optimal bandwidth based on the asymptotic formula mentioned above is used. The resultant curves are calculated based on the mean of the estimates with a Monte Carlo size of 1000. The Naive method produces larger bias in general when $\beta_1 \neq 0$. It is clear from the results that as the missingness deviates from MCAR, the amount of bias for the Naive method increased as expected. The curves for the IPW, MS and BG estimations are almost identical to the one based on the complete data. However, the IPW estimates have larger variation than the MS and BG estimates. This is evidenced by the larger MISE for the IPW method in Figure 1 and comparable bias for the IPW, BG and MS methods in Figure 2.

6 Study of Depression in Elderly Primary-care Patients

We will now illustrate the proposed approaches using the baseline data from a longitudinal study on depression in elderly patients recruited from primary-care practices in Monroe County, New York. A brief description of the study can be found in Tang et al. (2012). A total of 708 patients were enrolled and underwent a comprehensive diagnostic assessment for depression using the Structural Clinical Interview for DSM-IV (SCID). Among the 708 subjects, 249 patients were classified as having depression and the remaining 459 patients were declared as non-depressed. We are interested in the density function of medical burden, measured by Cumulative Illness Rating Scale (CIRS) (Linn et al., 1968), among the depressed patients. Data for both the CIRS and the SCID were collected for all participating patients in this study; therefore, we used randomly selected subsets of these data that resembled data which would have been obtained from a two-stage study design. In these subsets, the CIRS results are available for all subjects, but the SCID diagnoses are available only for certain patients randomly selected according to the following mechanism:

$$\Pr(\text{SCID available}) = 0.15 + 0.55I[\text{CIRS} > 7] + 0.30I[\text{Age} < 75] \quad (6.1)$$

Using this mechanism, approximately 56% of patients would be selected for SCID verification of the depression diagnosis, subjects younger than 75 but with CIRS greater than 75 are more likely selected to have SCID diagnosis. Since the complete data is available, it can be treated as a reference in assessing the performance of the new kernel density function estimates under MAR assumption.

In addition to the CIRS, other information collected as part of this study include the Hamilton Depression Rating Scale (HAM-D), a 24-item observer-rated scale designed to measure the severity of depressive symptoms. To better predict the SCID diagnosis, both CIRS and HAM-D, as well as age are included in the model as predictors. That is, the prediction probabilities of depression are estimated using the following logistic model

$$\text{logit}[\Pr(D=1|\text{CIRS}, \text{age}, \text{Ham-D})] = \beta_0 + \beta_1 \text{CIRS} + \beta_2 \text{Age} + \beta_3 \text{Ham-D}. \quad (6.2)$$

Since Ham-D is an excellent screening tool in for depression, we expect that the model should predict the SCID depression pretty well. We notice that there are some missing values for the SCID diagnosis, but under MAR, we can still make valid inferences using only the subjects with observed SCID diagnoses (6.2).

As in general, the missing probabilities needed for the IPW method are estimated from a model; we estimated π_i based on the following logistic model for the missing mechanism:

$$\text{logit}[\Pr(\text{SCID available}|\text{CIRS}\&\text{age})] = \beta_0 + \beta_1 \text{CIRS} + \beta_2 \text{Age}. \quad (6.3)$$

The Epanechnikov kernel is used for all the density function estimates. The optimal bandwidths are estimated based on the asymptotic formulas using a normal reference for f'' . The optimal bandwidths are 2.77, 3.02, 3.26, 2.49, and 2.32 for the complete data, the Naive, IPW, MS and BG estimates, respectively. Shown in Figure 3 are the estimates of the density function using different methods. Compared to the estimated density function using the complete data, the Naive method performs the worst, as expected. Both the MS and BG approaches perform very well, especially the BG approach. The estimated density functions are almost identical to those based on the complete data. The IPW method performed worse than the MS and BG methods, but better than the Naive method. This may be because the model for the missing mechanism uses continuous CIRS and AGE instead of $\text{CIRS} > 7$ and $\text{Age} < 75$ as predictors, as in (6.1), or because the prediction model yields very good estimates of d_i due to the inclusion of the Ham-D.

7 Discussion

In this paper we have developed the MS and BG methods for kernel density function estimation by incorporating the prediction probabilities into the kernel smoothing approach when the membership is missing. This study illustrates that the Naive approach, which uses only those subjects with known memberships such as disease status, may yield invalid

estimates of the density function. When the mechanism for the missing membership is well understood, the IPW method can be applied to correct the bias. However, we may lose some power in the estimates because information from subjects with missing memberships are discarded except for the modeling of the missing mechanism itself. On the other hand, when the disease process is well understood and enough information related to the disease is collected, we can obtain very good estimates of the disease probabilities. In this case, the proposed MS and BG methods proposed in this paper work very well in terms of reducing the bias and estimate efficiency. In situations that if the information about the disease process is inadequate, but the mechanism for the missing membership is well understood, the IPW methods will perform better. Thus, depending on one's knowledge of the missing mechanism and/or the disease process, one may select either the IPW or the MS/BG methods; theoretical and simulation results suggest that we use MS/BG methods when the disease process is well understood, and use the IPW method if the missing mechanism is well understood. Note also that in general, the MAR assumption is required for the methods, especially if parameters need to be estimated from the observed data. In cases where the $d'_i s$ are known, MAR is required for the BG estimation.

The boundary effect is a common problem for kernel estimate. It is well known that most non-parametric smoothing methods suffer from the boundary problem and usually behave poorly without adjustment at boundary points. The MS/BG approach is no exception, as indicated by the relatively poor behaviors of the estimates near the boundary. Further investigation on the boundary effect for the new approaches is needed.

Another aspect that needs future study is the effect of MAR on the estimates. In the special case that the $d'_i s$ are known, the BG method can handle missing not at random (MNAR) data, but in general we need to estimate the prediction model, which itself still requires the MAR assumption. When enough information is collected (via covariates) and included in the model for the missingness of disease status, the MAR model may be approximately true. But, will the MAR assumption affect the behavior of the estimates, and if so, to what extent? More generally, how do we deal with the MNAR? All questions require future studies.

Acknowledgments

This research was supported in part by NIH grants R33 DA027521 and R01GM108337. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute On Drug Abuse or the National Institutes of Health.

The authors would also like to thank Jeffrey M. Lyness, M.D. for providing the data used in Section 6.

Appendix

In this appendix, we give proofs for Theorem 1 - 4.

Proof of Theorem 1. We first show the asymptotic distribution of $\tilde{f}_{MS}(t, h)$ in Theorem 1(a). Let $u_i = D_i R_i + d'_i(1 - R_i)$ and $f_{ih}(t) = E[K_{ih}(t - T_i) | D_i = 1]$, as defined in (3.3), based on (3.1), we have

$$\begin{aligned}
& \sqrt{n}[\bar{f}_{MS}(t;h) - f_h(t)] \\
&= \sqrt{n} \left[\frac{\frac{1}{n} \sum_{i=1}^n u_i K_h(t - T_i)}{\frac{1}{n} \sum_{i=1}^n u_i} - f_h(t) \right] \\
&= \sqrt{n} \left[\frac{\frac{1}{n} \sum_{i=1}^n u_i K_h(t - T_i)}{\frac{1}{n} \sum_{i=1}^n u_i} - \frac{\frac{1}{n} \sum_{i=1}^n u_i K_h(t - T_i)}{p} + \frac{\frac{1}{n} \sum_{i=1}^n u_i K_h(t - T_i)}{p} - f_h(t) \right] \\
&= \frac{1}{p} \frac{1}{\sqrt{n}} \left[\frac{\sum_{i=1}^n u_i K_h(t - T_i)}{\sum_{i=1}^n u_i} \sum_{i=1}^n (p - u_i) + \sum_{i=1}^n [u_i K_h(t - T_i) - p f_h(t)] \right].
\end{aligned}$$

(7.1)

For any given h , at point t , as $n \rightarrow \infty$, by the the Weak Law of Large Numbers (WLLN), we have

$$\frac{\sum_{i=1}^n u_i K_h(t - T_i)}{\sum_{i=1}^n u_i} \rightarrow f_h(t). \quad (7.2)$$

By the Central Limit Theory (CLT), we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (p - u_i) \rightarrow N(0, \text{Var}(u_i)) \quad (7.3)$$

and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n [u_i K_h(t - T_i) - p f_h(t)] \rightarrow N(0, \text{Var}(u_i K_h(t - T_i))). \quad (7.4)$$

By applying Slutsky's theorem and with the consideration of the correlation between (7.3) and (7.4), we have

$$\sqrt{n} \left[\tilde{f}_{MS}(t;h) - f_h(t) \right] \rightarrow N(0, \sigma_1^2),$$

where $\sigma_1^2 = \frac{1}{p^2} \text{Var}(u_i K_h(t - T_i) - u_i f_h(t))$. The asymptotic distribution of $\tilde{f}_{MS}(t, h)$ in Theorem 1(a) has been proved.

Replacing $u_i = D_i R_i + d_i(1 - R_i)$ by $u_i = d_i$, and applying a similar argument for the proof of Theorem 1(a), we can prove Theorem 1 (b).

Proof of Theorem 2. Let $f(t)$ be the density function for the diseased population, and let $z_i = (T_i - t)/h$, then

$$\begin{aligned} \text{Bias} \left[\tilde{f}_{MS}(t;h) \right] &= E \left[\tilde{f}_{MS}(t;h) - f(t) \right] = E \left[\tilde{f}_{MS}(t;h) - f_h(t) + f_h(t) - f(t) \right] = E \left[\tilde{f}_{MS}(t;h) - f_h(t) \right] + E[f_h(t) - f(t)]. \end{aligned} \quad (7.5)$$

Based on Theorem 1, we have

$$E \left[\tilde{f}_{MS}(t;h) - f_h(t) \right] = 0 \text{ as } n \rightarrow \infty.$$

Since both $f_h(t)$ and $f(t)$ are defined for the diseased population, we have

$$\begin{aligned} f_h(t) - f(t) &= \int K_h(T_i - t) f(T_i) dT_i - f(t) \\ &= \int \frac{1}{h} K(z_i) f(t + h z_i) h dz_i - f(t) \\ &= \int K(z_i) \left[f(t) + f'(t) h z_i + \frac{1}{2} f''(t) h^2 z_i^2 + o(h^2) \right] dz_i - f(t) \\ &= \frac{1}{2} h^2 \mu_2(K) f''(t) + o(h^2). \end{aligned} \quad (7.6)$$

Combining (7.5) and (7.6), we have $\text{Bias} \left[\tilde{f}_{MS}(t;h) \right] = \frac{1}{2} h^2 \mu_2(K) f''(t) + o(h^2)$.

Next, we derive the variance of $\tilde{f}_{MS}(t, h)$. Let $w(t) = E(\pi_i d_i + d_i^2(1 - \pi_i) | T_i = t)$, $z_i = (T_i - t)/h$ and $g(t)$ be the population density function of T . Based on Theorem 1, the asymptotic variance for $\tilde{f}_{MS}(t)$ is

$$\begin{aligned}
& \text{Var}(\tilde{f}_{MS}(t)) \\
&= \frac{1}{np^2} \text{Var}[(K_h(t - T_i) - f_h(t))(D_i R_i \\
&\quad + d_i(1 - R_i))] \\
&= \frac{1}{np^2} E[(K_h(t - T_i) - f_h(t))(D_i R_i + d_i(1 - R_i))]^2 \\
&= \frac{1}{np^2} E[(K_h(t - T_i) - f_h(t))^2 (D_i R_i \\
&\quad + d_i^2(1 - R_i))] \\
&= \frac{1}{np^2} E\{E[(K_h(t - T_i) - f_h(t))^2 (D_i R_i \\
&\quad + d_i^2(1 - R_i)) | T_i] \\
&= t, x_i \\
&= x\} = \frac{1}{np^2} E[(K_h(t - T_i) - f_h(t))^2 (\pi_i d_i + d_i^2(1 - \pi_i))] = \frac{1}{np^2} E\{E[(K_h(t - T_i) - f_h(t))^2 (\pi_i d_i \\
&\quad + d_i^2(1 - \pi_i)) | T_i] \\
&= t\} \\
&= \frac{1}{np^2} E[(K_h(t - T_i) - f_h(t))^2 w(T_i)], \\
&= \frac{1}{np^2} \int \left(\frac{1}{h^2} K^2(z_i)\right. \\
&\quad + o\left(\frac{1}{h^2}\right) w(t \\
&\quad + h z_i) g(t \\
&\quad + h z_i) h d z_i \\
&= \frac{g(t) w(t)}{n h p^2} \int K^2(z_i) d z_i \\
&\quad + o\left(\frac{1}{n h}\right) \\
&= \frac{g(t) w(t)}{n h p^2} \int K^2(z_i) d z_i + o\left(\frac{1}{n h}\right).
\end{aligned}$$

Hence, the asymptotic variance of $\tilde{f}_{MS}(t)$ is

$$\frac{g(t)R(K)}{nhp^2}E[d_i^2 + \pi_i(d_i - d_i^2)|T_i=t] + o\left(\frac{1}{nh}\right), \text{ with } R(K) = \int K^2(t)dt.$$

Let $w(t) = E(d_i^2|T_i=t)$. Based on Theorem 1, with a similar argument, the asymptotic variance for $\tilde{f}_{BO}(t)$ can be derived as below:

$$\begin{aligned} & \frac{1}{np^2} \text{Var}[(K_h(t - T_i) - f_h(t))d_i] \\ &= \frac{1}{np^2} E[(K_h(t - T_i) - f_h(t))^2 d_i^2] = \frac{1}{np^2} E[(K_h(t - T_i) - f_h(t))^2 w(T_i)] \\ &= \frac{g(t)R(K)}{np^2} E[d_i^2|T_i=t] + o\left(\frac{1}{nh}\right). \end{aligned}$$

Proof of Theorem 3. We first show the asymptotic distribution of $\hat{f}_{MS}(t, h)$ in Theorem 3(a). Suppose we have a prediction model (3.8) and the parameters are estimated from (3.9). Let $\hat{\beta}$ be the estimate of β . Based on the Taylor expansion of (3.9) at β , we have

$$\hat{\beta} - \beta = \frac{1}{n} \mathbf{I}^{-1} \sum_{i=1}^n \Psi_i + o\left(\frac{1}{n}\right),$$

where $\mathbf{I} = -E\left[\frac{\partial \Psi_i}{\partial \beta^T}\right]$. If $\hat{\beta}$ are estimated from the score equation, $\mathbf{I}$ is the Fisher information matrix.

Since

$$\sqrt{n} [\hat{f}_{MS}(t) - f_h(t)] = \sqrt{n} [\hat{f}_{MS}(t) - \tilde{f}_{MS}(t) + \tilde{f}_{MS}(t) - f_h(t)],$$

the asymptotic distribution of $\sqrt{n} [\tilde{f}_{MS}(t) - f_h(t)]$ is already given in Theorem 1(a). We will focus on deriving the asymptotic distribution of $\sqrt{n} [\hat{f}_{MS}(t) - \tilde{f}_{MS}(t)]$.

Let $\hat{u}_i = D_i R_i + \hat{d}_i(1 - R_i) = D_i R_i + g(x_i; \hat{\beta})(1 - R_i)$. Based on (3.10) and (3.1), by applying WLLE, we have

$$\begin{aligned}
& \sqrt{n} \left[\hat{f}_{MS}(t) - \tilde{f}_{MS}(t) \right] \\
&= \sqrt{n} \left[\frac{\frac{1}{n} \sum_{i=1}^n \hat{u}_i K_h(t - T_i)}{\frac{1}{n} \sum_{i=1}^n \hat{u}_i} - \frac{\frac{1}{n} \sum_{i=1}^n u_i K_h(t - T_i)}{\frac{1}{n} \sum_{i=1}^n u_i} \right] = \sqrt{n} \left[\frac{\frac{1}{n} \sum_{i=1}^n \hat{u}_i K_h(t - T_i)}{\frac{1}{n} \sum_{i=1}^n \hat{u}_i} \right. \\
&\quad \left. - \frac{\frac{1}{n} \sum_{i=1}^n \hat{u}_i K_h(t - T_i)}{\frac{1}{n} \sum_{i=1}^n u_i} + \frac{\frac{1}{n} \sum_{i=1}^n \hat{u}_i K_h(t - T_i)}{\frac{1}{n} \sum_{i=1}^n u_i} \right. \\
&\quad \left. - \frac{\frac{1}{n} \sum_{i=1}^n u_i K_h(t - T_i)}{\frac{1}{n} \sum_{i=1}^n u_i} \right] = \frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{u}_i K_h(t - T_i) \left[\frac{1}{\frac{1}{n} \sum_{i=1}^n \hat{u}_i} \right. \\
&\quad \left. - \frac{1}{\frac{1}{n} \sum_{i=1}^n u_i} \right] + \frac{1}{\sqrt{np}} \sum_{i=1}^n K_h(t - T_i) (\hat{u}_i - u_i) + o(1) \\
&= I + II + o(1).
\end{aligned}$$

The second term

$$\begin{aligned}
II &= \frac{1}{\sqrt{np}} \sum_{i=1}^n K_h(t - T_i) (\hat{u}_i - u_i) \\
&= \frac{1}{\sqrt{np}} \sum_{i=1}^n K_h(t - T_i) \left[\frac{\partial u_i}{\partial \beta^T} (\hat{\beta} - \beta) + o\left(\frac{1}{n}\right) \right] \\
&= \frac{1}{n} \left[\sum_{i=1}^n K_h(t - T_i) \frac{\partial u_i}{\partial \beta^T} \right] \frac{\sqrt{n}}{p} (\hat{\beta} - \beta) + o(1) \\
&= E \left(K_h(t - T_i) \frac{\partial u_i}{\partial \beta^T} \right) \frac{\mathbf{I}^{-1}}{\sqrt{np}} \sum_{i=1}^n \Psi_i + o(1),
\end{aligned}$$

where $\mathbf{I} = -E \left[\frac{\partial \Psi_i}{\partial \beta^T} \right]$.

For the first term, by Slutsky's theorem,

$$\frac{1}{\frac{1}{n}\sum_{i=1}^n \hat{u}_i} - \frac{1}{\frac{1}{n}\sum_{i=1}^n u_i} = -\frac{1}{p^2} E \left[\frac{\partial u_i}{\partial \beta^T}(\beta) \right] (\hat{\beta} - \beta) + o(1).$$

Thus,

$$\begin{aligned} I &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{1}{n} \sum_{i=1}^n \hat{u}_i K_h(t) \\ &\quad - T_i \left[\frac{1}{\frac{1}{n}\sum_{i=1}^n \hat{u}_i} \right. \\ &\quad \left. - \frac{1}{\frac{1}{n}\sum_{i=1}^n u_i} \right] = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[(u_i \right. \\ &\quad \left. + \frac{\partial u_i}{\partial \beta^T}(\hat{\beta} \right. \\ &\quad \left. - \beta)) K_h(t \right. \\ &\quad \left. - T_i \right] \left(\right. \\ &\quad \left. - \frac{1}{p^2} \left[\frac{\partial u_i}{\partial \beta^T}(\beta) \right] (\hat{\beta} \right. \\ &\quad \left. - \beta) \right) \\ &\quad + o(1) \\ &= -\frac{f_h(t)}{p} E \left[\frac{\partial u_i}{\partial \beta^T}(\beta) \right] \sqrt{n}(\hat{\beta} \\ &\quad - \beta) + o(1) \\ &= -\frac{f_h(t)}{p} E \left[\frac{\partial u_i}{\partial \beta^T}(\beta) \right] \frac{1}{\sqrt{n}} \mathbf{I}^{-1} \sum_{i=1}^n \Psi_i + o(1). \end{aligned}$$

It follows that

$$\begin{aligned} &\sqrt{n} \left[\hat{f}_{MS}(t) \right. \\ &\quad \left. - f_h(t) \right] \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \frac{f_h(t)}{p} (p - u_i) + \frac{1}{p} [u_i K_h(t - T_i) - p f_h(t)] - \frac{f_h(t)}{p} E \left[\frac{\partial u_i}{\partial \beta^T}(\beta) \right] \mathbf{I}^{-1} \Psi_i \right. \\ &\quad \left. + \frac{E \left(K_h(t - T_i) \frac{\partial u_i}{\partial \beta^T}(\beta) \right)}{p} \mathbf{I}^{-1} \Psi_i \right\} \\ &\quad + o(1) \rightarrow N(0, \sigma_3^2), \end{aligned}$$

where

$$\sigma_3^2 = \frac{1}{p^2} \text{Var} \left(u_i K_h(t - T_i) - f_h(t) u_i + \left(E \left(K_h(t - T_i) \frac{\partial u_i}{\partial \beta^T} \right) - f_h(t) E \left(\frac{\partial u_i}{\partial \beta^T} \right) \right) \mathbf{I}^{-1} \Psi_i \right).$$

Let $\mathbf{c} = E \left[K_h(t - T_i)(1 - R_i) \frac{\partial g_i}{\partial \beta^T}(\beta) \right]$ and $\mathbf{d} = E \left[(1 - R_i) \frac{\partial g_i}{\partial \beta^T}(\beta) \right]$. Since

$$\frac{\partial u_i}{\partial \beta^T} = (1 - R_i) \frac{\partial g_i}{\partial \beta^T}, \text{ we have}$$

$$\sigma_3^2 = \frac{1}{p^2} \text{Var} \{ u_i [K_h(t - T_i) - f_h(t)] + (\mathbf{c} - f_h(t)\mathbf{d})\mathbf{e}^{-1} \Psi_i \}.$$

Theorem 3(b) can be proved similarly by replacing u_i by d_i and $\hat{u}_i$ by $\hat{d}_i$ in the above arguments.

Proof of Theorem 4. Based on Theorem 3, the asymptotic bias for both $\hat{f}_{MS}(t)$ and $\hat{f}_{BG}(t)$ are $f_h(t) - f(t)$. Thus, the bias follows from the proof of Theorem 2.

The proof for the asymptotic variance also follows similarly to that of Theorem 2.

Let $w(t) = E[\pi_i d_i + d_i^2(1 - \pi_i) | T_i = t]$, $w_1(t) = E\{u_i(\mathbf{c} - f_h(t)\mathbf{d})\mathbf{e}^{-1} \Psi_i | T_i = t\}$, and $w_2(T_i) = E[(\mathbf{c} - f_h(t)\mathbf{d})\mathbf{e}^{-1} \Psi_i]^2$. Based on Theorem 3, the asymptotic variance for $\hat{f}_{MS}(t)$ can be derived as below:

$$\begin{aligned} & \frac{1}{np^2} \text{Var} [u_i(K_h(t - T_i) - f_h(t)) + (\mathbf{c} - f_h(t)\mathbf{d})\mathbf{e}^{-1} \Psi_i] \\ &= \frac{1}{np^2} E[(K_h(t - T_i) - f_h(t))(D_i R_i + d_i(1 - R_i)) + (\mathbf{c} - f_h(t)\mathbf{d})\mathbf{e}^{-1} \Psi_i]^2 \\ &= \frac{1}{np^2} E[(K_h(t - T_i) - f_h(t))^2(D_i R_i + d_i^2(1 - R_i)) + 2(K_h(t - T_i) - f_h(t))(D_i R_i \\ & \quad + d_i(1 - R_i))(\mathbf{c} - f_h(t)\mathbf{d})\mathbf{e}^{-1} \Psi_i + ((\mathbf{c} - f_h(t)\mathbf{d})\mathbf{e}^{-1} \Psi_i)^2] \\ &= \frac{1}{np^2} E[(K_h(t - T_i) - f_h(t))^2 w(T_i) + 2(K_h(t - T_i) - f_h(t))w_1(T_i) + w_2(T_i)] = \frac{1}{np^2} \\ & \quad \int \left(\frac{1}{h^2} K^2(z_i) \right) w(t + h z_i) g(t + h z_i) h d z_i + o\left(\frac{1}{nh}\right) = \frac{g(t)w(t)}{nhp^2} \int K^2(z_i) d z_i + o\left(\frac{1}{nh}\right). \end{aligned}$$

Hence, the asymptotic variance of $\hat{f}_{MS}(t, h)$ is

$$\frac{g(t)R(K)}{nhp^2} E[d_i^2 + \pi_i(d_i - d_i^2) | T_i = t] + o\left(\frac{1}{nh}\right), \text{ where } R(K) = \int K^2(t) dt.$$

Similarly, we can prove the asymptotic variance of $\hat{f}_{BG}(t, h)$ in (3.16).

References

- Alonzo TA, Pepe MS. Assessing accuracy of a continuous screening test in the presence of verification bias. *Journal of the Royal Statistical Society, Series C*. 2005; 54:173–190.
- Alonzo TA, Pepe MS, Lumley T. Estimating disease prevalence in two-phase studies. *Biostatistics*. 2003; 4:313–326. [PubMed: 12925524]
- Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*. 1983; 39:207–215. [PubMed: 6871349]
- Chaudron L, Szilagyi P, Tang W, Anson E, Talbot N, Wadkins H, Tu X, Wisner K. Accuracy of depression screening tools for identifying postpartum depression among urban mothers. *Pediatrics*. 2010; 125(3):e609–e617. [PubMed: 20156899]
- He H, Lyness JM, McDermott MP. Direct estimation of the area under the receiver operating characteristic curve in the presence of verification bias. *Statistics in Medicine*. 2009; 28(3):361–376. [PubMed: 18680124]
- He H, McDermott M. A robust method for correcting verification bias for binary tests. *Biostatistics*. 2012; 13(1):32–47. [PubMed: 21856650]
- Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*. 1952; 47:663–685.
- Linn BS, Linn MW, Gurel L. Cumulative illness rating scale. *Journal of the American Geriatrics Society*. 1968; 16:622–626. [PubMed: 5646906]
- Little, RJA., Rubin, DB. *Statistical Analysis with Missing Data*. Second. Hoboken: Wiley; 2002.
- Parzen E. On estimation of a probability density function and mode. *Ann Math Statist*. 1962; 33:1065–1076.
- Pepe MS, Reilly M, Fleming TR. Auxiliary outcome data and the mean score method. *Journal of Statistical Planning and Inference*. 1994; 42:137–160.
- Reilly M, Pepe MS. A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*. 1995; 82:299–314.
- Simonoff, JS. *Smoothing Methods in Statistics*. Springer; New York: 1996.
- Spitzer, RL., Gibbon, M., Williams, JWB. *Structured Clinical Interview for Axis I DSM-IV Disorders*. Biometrics Research Department, New York State Psychiatric Institute; 1994.
- Tang W, He H, Gunzler D. Kernel smoothing density estimation when group membership is subject to missing. *Journal of Statistical Planning and Inference*. 2012; 142(3):685–694. [PubMed: 22116738]
- Wand, MP., Jones, MC. *Kernel smoothing*, volume 60 of *Monographs on Statistics and Applied Probability*. Chapman and Hall Ltd; London: 1995.
- Wang Q. Probability density estimation with data missing at random when covariables are present. *J Statist Plann Inference*. 2008; 138(3):568–587.
- White H. Maximum likelihood estimation of misspecified models. *Econometrica*. 1982; 50(1):1–25.

Figure 1.

Mean integrated square error curves across 1000 realizations for the kernel density estimates based on complete data, the Naive method, IPW, MS and BG methods, represented by solid, dashed, dotted, dotdash and longdash lines, respectively.

Figure 2.

The true curve and the fitted curves based on the mean of fitted values across 1000 realizations are represented by solid, dashed, dotted, dotdash, longdash and twodash lines based on the true curve, the complete data, the Naive, IPW, MS and BG methods, respectively, for $\beta_1 = 0, .2, .4, .6, .8$, and 1.

Figure 3.

The fitted density curves of CIRS score for depressed patients are represented by solid (complete), dashed (naive), dotted (IPW), dotdash (MS) and longdash (BG) lines.