Published in final edited form as:
J Stat Plan Inference. 2012 March ; 142(3): 685–694. doi:10.1016/j.jspi.2011.09.009.

Kernel Smoothing Density Estimation when Group Membership is Subject to Missing

Wan Tang, **Hua He**, and **Douglas Gunzler**
Department of Biostatistics and Computational Biology, University of Rochester Medical Center, 601 Elmwood Ave, Box 630, Rochester, NY 14642, U.S.A

Abstract

Density function is a fundamental concept in data analysis. Nonparametric methods including kernel smoothing estimate are available if the data is completely observed. However, in studies such as diagnostic studies following a two-stage design the membership of some of the subjects may be missing. Simply ignoring those subjects with unknown membership is valid only in the MCAR situation. In this paper, we consider kernel smoothing estimate of the density functions, using the inverse probability approaches to address the missing values. We illustrate the approaches with simulation studies and real study data in mental health.

Keywords

kernel smoothing; density; membership missing

1 Introduction

As a fundamental concept in understanding univariate continuous outcomes, the probability density function is frequently used in data analysis. Among the nonparametric methods developed, the kernel smoothing estimate may be the most popular approach[7, 10, 13]. In the standard setting where a simple random sample (i.i.d.) from the target population is completely observed (complete data case), the problem of kernel smoothing has been thoroughly studied. Recently, some studies have generalized the kernel estimate to non-complete data situations. For example, estimates of density function were considered in the context of missing values in the variable of interest [14]. In this paper we discuss estimates of density function within the diagnostic test setting where some subjects have missing disease status.

In modern clinical trials, it is quite common that the subjects are not sampled directly from the target population. In studies of specific diseases, the case and control approach in which subjects are directly sampled from cases and controls can not be applied if the disease status is unknown for each subject. A common approach in such situations is to first recruit subjects from a larger population, including both diseased and non-diseased subjects, and then ascertain the disease status at a later time. Such a two-stage design [15] is common, especially in diagnostic test studies. If the disease status or membership for the disease and non-disease groups for each subject is confirmed, the problem of density estimate reduces to the standard setting with analysis based on the subsample consisting of those subjects belonging to the target population. However, if not all subjects are known for their true disease status, then we are facing the missing data problem.

There are many reasons for missing disease status and some of them may be related with the missing values themselves. For example, if the membership is the onset of some disease, and the gold standard test for diagnosis is too expensive or involves intensive procedure such as surgeries, some of the subjects at lower risk of the disease may choose not to take the standard test (in some cases, it is actually unethical to do so). For example, structural clinical interview for DSM-IV diagnosis (SCID)[11], which is in general viewed as the gold-standard for diagnosing depression, usually involves hours of interview of the doctor with the patient, making it expensive and inconvenient for the patients. Thus, it may not be practical to have each patient administered SCID.

This paper is motivated by problems from a real study. In a recent study of depression among postpartum women, accuracies of several screening tests for depression are assessed among postpartum mothers[3]. A total of 419 postpartum women initially agreed to participate in the study, and their demographic information and screening test results were collected. The depression status was not known in advance for these subjects. But only 198 subjects completed the subsequent SCID. Assessing test accuracy with the subject's disease status subject to missing is known as the verification bias problem in diagnostic test studies. Ignoring those subjects with missing disease status in general produces biased estimates. Although methods are available for correcting such bias, they focus on modeling the operating receiver characteristic (ROC) curves and/or the area under the ROC curve[1, 4, 9]. Since ROC curves are determined by the distributions of test results and the disease status, density estimates of the distributions of test outcomes for the diseased and non-diseased provide a direct alternative to the problem.

Since the membership of disease and non-disease groups is missing for some subjects, standard kernel smoothing methods do not apply. The naive approach that uses only those subjects with known group membership in the analysis is valid only in the special situation when the membership is missing completely at random (MCAR). As mentioned above, the missing group membership usually occurs in a systematic way, creating a basis for biased estimates when applying such a naive approach.

In the rest of the paper, we develop a kernel smoothing method for addressing the bias issue under missing group membership within the current context. For comparison purposes as well as a way to introduce notation, we first give a brief review of standard univariate kernel smoothing under complete data in Section 2. We then propose a new approach to address the limitations of traditional methods where the group membership is subject to missing in section 3. Bandwidth selection for the proposed smoothing approach 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 in Section 7.

2 Kernel Smoothing Estimate for Complete Data

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 groups of interest such as diseased and non-diseased groups in our context, $\mathbf{x}_i$ is a vector of covariates, and T_i is a univariate random variable such as a test outcome in our study. We are interested in estimating the density function of T_i among the different groups. Assume $D_i = 1$ for the group of interest, and $D_i = 0$ for the remaining subjects. In the case of complete data 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 the estimate of density function of T_i.

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

$$R(f^{''}) = \int f^{''}(t)^2 dt < \infty. \tag{1}$$

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

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

Also, let $m{=}\sum_{i=1}^{n} D_i$ be the number of subjects with $D_i = 1$ in the sample. The kernel density estimate of $f(t)$ at a point $T = t$ using the kernel function $K(\cdot)$ is defined as

$$\widehat{f}(t;h){=}\frac{1}{m}\sum_{i:D_i=1} K_h(t-T_i){=}\frac{\sum_{i=1}^{n} D_i K_h(t-T_i)}{\sum_{i=1}^{n} D_i}, \tag{3}$$

where the bandwidth $h > 0$ is a constant and $K_h(t){=}\frac{1}{h}K(\frac{t}{h})$. The kernel function $K(\cdot)$ generally gives more weight to observations closer to t. The choice of the bandwidth is closely related to the behavior of the kernel estimate. Larger bandwidths correspond to estimates that are more biased but less unstable, while smaller bandwidths yield estimates that are less biased, but more unstable. Bandwidth should be carefully selected to balance bias and variance.

Since the bias reduces to 0 as the bandwidth h approaches infinity, the bandwidth needs to be small to reduce bias. However, it also needs to be large enough to include sufficient subjects; otherwise, the variance may be very large and the estimates themselves may not even exist. More precisely, we let the bandwidth h be a non-random sequence of positive numbers such that

$$\lim_{n\to\infty} h{=}0 \text{ and } \lim_{n\to\infty} nh{=}\infty. \tag{4}$$

Under conditions (1),(2), and (4), the asymptotic bias and variance for the kernel density estimates in (3)[7] are given by

$$\begin{array}{c}\text{Bias}\{\widehat{f}(t;h)\}{=}f_h(t)-f(t){=}\frac{1}{2}h^2\mu_2(K)f^{''}(t){+}o(h^2),\\ \text{Var}\{\widehat{f}(t;h)\}{=}\frac{1}{mh}f(t)R(K){+}o(\frac{1}{mh}),\end{array} \tag{5}$$

where $f_h(t) = E\,[K_h(t - T_i)|D_i = 1]$ and $R(K) = \int K(z)^2 dz$. Taking both the bias and variance into consideration, the behavior of the estimate at each point is assessed by the mean squared error (MSE) of the estimate at the point. As a direct consequence of (5), the asymptotic MSE of the kernel density estimate in (3) is

$$\mathrm{MSE}\{\widehat{f}(t;h)\}=E\left[\widehat{f}(t;h)-f(t)\right]^2=\mathrm{Bias}\{\widehat{f}(t;h)\}^2+Var\{\widehat{f}(t;h)\}. \tag{6}$$

It is often desirable to assess the estimate over the entire real line. By integrating the MSE over the entire real line, the mean integrated squared error (MISE) provides such a measure to assess the overall accuracy of the estimate. As a direct consequence of (6), the asymptotic MISE of (3) is given by

$$\mathrm{MISE}\{\widehat{f}(t;h)\}=\int E\left[\widehat{f}(t;h)-f(t)\right]^2dt=\frac{1}{4}h^4\mu_2^2(K)R(f'')+\frac{1}{nh}R(K)+o\left(\frac{1}{nh}+h^4\right), \tag{7}$$

where $R(f'') = \int f''(t)^2 dt$. Optimal bandwidths should give rise small MISEs. Thus, by minimizing (7), we obtain the asymptotic optimal bandwidth $h_{AMISE}=\left[\frac{R(K)}{\mu_2^2(K)R(f'')n}\right]^{1/5}$.

3 Univariate Kernel Smoothing for Incomplete Data

In this section we discuss estimation of density function of T for the group $D = 1$ under missing group membership for some subjects. We assume that the covariates $\mathbf{x}_i$ and the variable of interest T_i are always observed. Let V_i be the indicator of whether D_i is observed; $V_i = 1$ if it is observed, and $V_i = 0$ if otherwise. Thus, $(\mathbf{x}_i, T_i, V_i)$ is always observed, but D_i is only observed for those subject with $V_i = 1$. We will extend the kernel density estimate (3) described in the last section to this situation.

The naive estimate would be simply applying the kernel smoothing estimate to the subgroup of those subjects who are known to be in the group defined by $V_i = 1$ and $D_i = 1$, i.e.,

$$\widetilde{f_{Naive}}(t)=\frac{\sum_{V_i=1 \text{ and } D_i=1}K_h(t-T_i)}{\sum_{V_i=1 \text{ and } D_i=1}1}. \tag{8}$$

This naive estimate is valid only under the very strong missing completely at random (MCAR) condition. In particular, it does not apply when the missingness of the group membership follows the missing at random (MAR) assumption.

Conditional on the observed outcomes T and covariates $\mathbf{x}$, assume that the membership observation indicator V is independent with the exact membership D, i.e.

$$V \perp D|(T,\mathbf{x}). \tag{9}$$

This MAR assumption is common in the literature of missing values. In the aforementioned diagnostic study examples, MAR is plausible if the decision of the administration of gold standard test depends on the observed test results and covariates. The inverse probability weighting (IPW) technique is commonly used to address missing values under MAR[5, 8]. Below, we apply this approach to address missing group membership in our setting.

Let $\pi_i = \Pr(V_i = 1|T_i, \mathbf{x}_i)$ be the probability that the group membership is observed for the ith subject. If π_i is known by design as in some two-stage studies, then those subjects with known group status in the group (V_i = 1 and D_i = 1) can be used for the density estimate with proper weighting. The idea of IPW is that each subject with known group membership is selected for verification (V_i = 1) with a probability π_i among similar subjects and thus should be weighted by the inverse of this probability in its contribution to the estimation. By applying this idea to the kernel density estimate in our setting, the density estimate with the inverse probability weight based on the known probabilities (IPWK) at a point $T = t$ is given by

$$\tilde{f}(t)=\frac{\sum_{V_i=1 \text{ and } D_i=1}\frac{1}{\pi_i}K_h(t-T_i)}{\sum_{V_i=1 \text{ and } D_i=1}\frac{1}{\pi_i}}=\frac{\sum_{i=1}^{n}\frac{D_iV_i}{\pi_i}K_h(t-T_i)}{\sum_{i=1}^{n}\frac{D_iV_i}{\pi_i}}. \tag{10}$$

Since V_i = 0 for those with unknown D_i, the estimate above is computable only if π_i is known. Further, the estimate is only based on those with known group membership (D_i = 1 and V_i = 1).

Assume that the selection probability is continuous in T_i and $\mathbf{x}_i$. Furthermore, we assume that π_i is bounded below away from 0, i.e.,

$$\pi_i>c>0, \tag{11}$$

where c is a constant. This condition is necessary for the estimates to have good behaviors, since otherwise we may have some very large weights, yielding unstable estimates. Under the conditions (1),(2), and (4), and (11), we have the following

Theorem 1

The asymptotic bias and variance of $\tilde{f}$ are

$$\begin{aligned}Bias\{\tilde{f}(t;h)\}&=\tfrac{1}{2}h^2\mu_2(K)f''(t)+o(h^2),\\ Var\{\tilde{f}(t;h)\}&=\tfrac{c(t)}{nph}f(t)R(K)+o(\tfrac{1}{nh}),\end{aligned} \tag{12}$$

where p = Pr(D_i = 1) is the proportion of the group with D_i = 1 in the whole population and $c(t)=E\left[\frac{1}{\pi(T_i;\mathbf{x}_i)}|T_i=t\right]$.

It is clear that we obtain the same bias as the complete case, but with a larger variance. Under MCAR, π_i is a constant and hence $c(t)=\frac{1}{\pi}$. The asymptotic variance in this special case is $\frac{1}{np\pi h}f(t)R(K)$, which is consistent with the complete data case, since the actual sample size in this case is $np\pi$. In other words, the results in the theorem reduce to the complete case when there is no missing values in the group membership, i.e., π_i = 1.

In most studies other than those based on two-stage designs, the probabilities π_i are not known. For example, although physicians in our setting may make the decision for SCID assessment based on the subject's demographic, history of mental health and screening test results, it is quite rare that they make their decisions by modeling π_i and generating a random V_i based on the model. In such cases, the missing mechanism satisfies the MAR assumption, but with known &$pi;_i$. Although in observational studies, MAR may not be a correct model, it will hold approximately true, if sufficient information is included when modeling the weight function π_i. In either situations, we need to model and estimate π_i.

Since the indicator of observed group membership is binary, we can model π_i using logistic regression:

$$\text{logit}\,[\Pr(V_i=1|\mathbf{x}_i,T_i)]=\text{logit}(\pi_i)=\beta\mathbf{x}_i. \tag{13}$$

To simplify the notation, we have subsumed the constant term of the logistic regression as well as T into the vector of covariates $\mathbf{x}$. Given the above model, we can readily estimate β.

In particular, the MLE of β can be obtained by solving the following score equations:

$$\frac{1}{n}\sum_{i=1}^{n}s_i(\mathbf{x}_i,V_i;\beta)=\frac{1}{n}\sum_{i=1}^{n}(V_i-\pi_i(\beta))\mathbf{x}_i=0. \tag{14}$$

Note that unlike the density estimate, all subjects (whether D_i is observed or not) are used for estimating β in the above equations.

Denote the estimated probabilities of being selected for verification by $\widehat{\pi}_i=\frac{\exp(\widehat{\beta}\mathbf{x}_i)}{1+\exp(\widehat{\beta}\mathbf{x}_i)}$. By substituting the estimates into (10), we obtain the following IPW estimate with weight based on modeling of the missing mechanism (IPW):

$$\widehat{f}(t)=\frac{\sum_{V_i=1\text{ and }D_i=1}\frac{1}{\widehat{\pi}_i}K_h(t-T_i)}{\sum_{V_i=1\text{ and }D_i=1}\frac{1}{\widehat{\pi}_i}}=\frac{\sum_{i=1}^{n}\frac{D_i,V_i}{\widehat{\pi}_i}K_h(t-T_i)}{\sum_{i=1}^{n}\frac{D_i,V_i}{\widehat{\pi}_i}}. \tag{15}$$

Under the conditions (1),(2), and (4), and (11) and (14), we have the following

Theorem 2

The asymptotic bias and variance of $\hat{f}$ are

$$\begin{aligned}\text{Bias}\{\widehat{f}(t;h)\}&=\tfrac{1}{2}h^2\mu_2(K)f''(t)+o(h^2),\\ \text{Var}\{\widehat{f}(t;h)\}&=\tfrac{c(t)}{nph}f(t)R(K)+o(\tfrac{1}{nh}),\end{aligned} \tag{16}$$

where $c(t)=E\left[\frac{1}{\pi(T_i,\mathbf{x}_i)}|T_i=t\right]$.

Comparing Theorems 1 and 2, we see that $\tilde{f}$ and $\hat{f}$ have the same asymptotic bias and variance. These are the direct consequences of the following lemma which gives the asymptotic distributions of the respective estimates $\tilde{f}$ and $\hat{f}$. A proof of the lemma is provided in the appendix.

Lemma 3

Under the conditions (1),(2), and (4), and (11) and (14). For fixed h, we have (a)

$$\sqrt{n}\left[\tilde{f}(t)-f_h(t)\right]\to N(0,\sigma_1^2), \tag{17}$$

where $\sigma_1^2=E\left\{\frac{1}{p\pi_i}[K_h(t-T_i)-f_h(t)]^2|D_i=1\right\}$, and (b)

$$\sqrt{n}\left[\widehat{f}(t)-f_h(t)\right]\to N(0,\sigma_2^2), \tag{18}$$

where $\sigma_2^2=Var\left(\frac{D_iV_i}{p\pi_i}[K_h(t-T_i)-f_h(t)]+(\mathbf{c}_2-\mathbf{c}_1 f_h\,(t)\,I^{-1}(\beta)s(\mathbf{x}_i,\beta)\right)$, $\mathbf{c}_1 = E\,[(1-\pi_i)\,\mathbf{x}_i|D_i = 1]$ and $\mathbf{c}_2 = E\,[(1-\pi_i)\,K_h(t-T_i)\mathbf{x}_i|D_i = 1]$.

It is straightforward to prove Theorem 1 based on the fact that $\sigma_1^2=\frac{1}{\pi(T_i)h}f(t)R(K)+o(h^{-1})$ and (17). Theorem 2 is based on (18) and the fact that $\sigma_2^2=\frac{1}{\pi(T_i)h}f(t)R(K)+o(h^{-1})$. It should be pointed out that although the expression for the variance has extra terms, IPW with estimated missing probabilities in general has slightly better behavior than IPWK, even when the selection probabilities π_i are known (see the simulation study in Section 5 for details). Note that a similar phenomenon in regression analysis is well known.

4 Bandwidth Selection

It is clear from Theorems 1 and 2 that the behaviors of the estimates are closely related to the bandwidth h used. If h is too small, then the estimates are not stable. On the other hand, if h is too big, the bias can be large. We must trade o3 between bias and variance in selection of the bandwidths. As in the complete data cases, we can assess the qualities of the estimates by their mean squared errors. Based on Theorems 1 and 2, we immediately have the following

Theorem 4

The MSE of $\tilde{f}$ and $\hat{f}$ at point t both equal to

$$MSE(t;h)=\frac{1}{4}h^4\mu_2^2(K)f''(t)^2+\frac{c(t)}{nph}f(t)R(K)+o(\frac{1}{nh}+h^4). \tag{19}$$

Since MSE = Bias^2 +Variance, the proof is straightforward.

The MSE assesses the behavior of the estimate at a single point t, with a smaller MSE indicating better fit. Asymptotically optimal bandwidth at a point t can be obtained by minimizing the corresponding MSE. To assess the behavior of the estimate with a common bandwidth over the entire range, we need to assess the integrated MSE (MISE), which is the integration of MSE over the whole range. Following from Theorem 2, we have:

Corollary 5

The MISE for $\tilde{f}$ and $\hat{f}$ are given by

$$MISE(t;h)=\frac{1}{4}h^4\mu_2^2(K)R(f'')+\frac{c_o}{nph}R(K)+o(\frac{1}{nh}+h^4), \tag{20}$$

where $c_0=\int c(t)f(t)dt=E(\frac{1}{\pi_i})$.

As in the case of MSE, the smaller the MISE the better the fit. Optimal bandwidths can be selected by minimizing the MISE (20). It is easy to see that the minimum can be achieved, asymptotically, when $\frac{c_0}{nph}R(K)=h^4\mu_2^2(K)R(f'')$. Thus, the asymptotic optimal bandwidth is given by:

$$h_{MISE}=\left[\frac{c_oR(K)}{\pi_2^2(K)R(f'')np}\right]^{1/5}. \tag{21}$$

In the formula for the optimal bandwidth (21), $R(K)$ and $\mu_2(K)$ can be easily computed. 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 μ_2 (K) = 1 and $R(K)=\frac{3}{5}$. The term c_0 is related to the missing rate, and it can be estimated roughly by the sample missing rate. The computation of $R(f'')$ is a bit involved since it involves the estimate of f''. One approach is to use the normal distribution as a basis for bandwidth selection. If f follows a normal distribution with variance σ^2, the corresponding $R(f'') \approx 0.212\ \sigma^4$. Under this approach, we estimate first the variance of the variable to obtain an estimate of σ and then substitute 0.212 σ^4 in the place of $R(f'')$ in (21) to obtain an estimate of the optimal bandwidth. We may further estimate f'' using the bandwidth we obtained based on the normal distribution approach, and then substitute the estimate of $R(f'')$ in (21) to obtain an estimate of the optimal bandwidth.

5 Simulation Studies

Simulation studies were performed in a couple of different scenarios to assess the behaviors of the estimates. We considered the situation where T is a binormal, i.e., T is normally distributed in both groups. However, to focus on a finite range, we used a normal distribution truncated on [−1, 1]. More precisely, $T|D = 1$ ($T|D = 0$) follows standard normal $N(0, 1)$ ($N(1, 1)$) truncated on [−1, 1]. In addition, the missingness of group membership D depends only on T, i.e., no other covariates $\mathbf{x}_i$ were involved.

We assumed a missing probability model based on logistic regression, i.e., $\pi=\Pr(\text{D observed}|T)=\frac{\exp(\beta_0+\beta_1T)}{1+\exp(\beta_0+\beta_1T)}$. Following this model, β_0 can used to control the missing rate, while β_1 indicates the degree of deviation from MCAR. In the simulation study, the proportion of the first group in the whole population was fixed at .5, i.e., p = .5. The Epanechnikov kernel function was used for density estimates.

As pointed out in [1], sample sizes as large as 1000 are common in diagnostic test studies, and asymptotic theory can be applied. In the simulation study, to assess the behaviors of the estimates, especially how they would change with bandwidth, under small to moderate small sample sizes, we set the sample size at 200. Note also that under our setting, there are roughly 100 in each group, however, memberships of only part of them are confirmed.

Setting $\beta_0 = 0$ and $\beta_1 = 0.5$, we computed the integrated squared bias, integrated variance, as well as the MISE for bandwidths that varied from 0.01 to 1 based on a Monte Carlo size of 1000. Shown in Figure 1 are the MISEs as a function of bandwidth for the four estimates. When the bandwidth increased, the bias increased, while the variance decreased. Both IPW estimates, whether based on the known or estimated selection probabilities, behaved better as compared to the naive methods. The IPW approaches had a comparable amount of bias as compared to the estimate based on the complete data. But the naive method had a much larger bias. The variances of the IPW methods and those of the naive methods were comparable, although as expected the estimate based on the complete data had a smaller variance. Similar to the regression setting, density estimates obtained based on estimated probabilities behaved slightly better than those based on the true (known in simulation) probabilities, although the difference was small. Since $\mu_2(K) = 1$, $R(K) = .6$, the variance of the standard normal truncated by [−1,] is $1-\frac{2\varphi(1)}{\Phi(1)-\Phi(-1)}\approx 0.29$, where φ and Φ are the density and cumulative distribution functions of the standard normal. It follows that the bandwidth suggested by (20) would be $\left[\frac{.6/.5}{.212(.29)^{-5/2}\times 500}\right]^{1/5}\approx 0.303\,23$. Based on the simulation results, the minimum of the MISE was achieved around $h = 0.335$ for the IPW estimates, confirming that the formula (20) is reliable for computing the optimal bandwidth.

From Figure 1, we see that the naive method produces larger biases as compared to the IPW estimates, as the former did not address MAR. To assess the effect of missing data mechanism, we considered different values of β_1. 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 larger β_1 indicating more deviations from MCAR. Since the missing rate was roughly 50%, we obtained from (21) a rough estimate 0.3 as the asymptotic optimal bandwidth by substituting 2 for c_0 in (21) and computing $R(f'')$ under the known distribution with the sample size 200.

The following plot contains the mean of the smoothed curves with a Monte Carlo size of 1000. The mean curves for the IPW approaches are almost identical to that of the estimate based on the complete data. It is clear from the plot that as the missingness deviates from MCAR, the amount of bias from the naive methods increased. However, the IPW methods still provided good estimates.

6 Study of Depression in Elderly Primary-care Patients

We now illustrate our proposed methodology using the baseline data from a real longitudinal study on depression in elderly patients (age 65 and over) recruited from primary-care practices in Monroe County, New York. In addition to depression status determined by SCID, other information collected include demographic variables, the Hamilton Depression Rating Scale (HAM-D) for depression, a 24-item observer-rated scale designed to measure the severity of depressive symptoms [16], and the total score on the Cumulative Illness Rating Scale (CIRS), a reliable and valid measure of medical burden that quantifies the amount of pathology in each organ system [6]. Among the 708 patients enrolled, 249 patients were classified as having depression and the remaining 459 patients were declared as depression-free. Although the total score of HAM-D is inherently discrete, it was often treated as a continuous outcome because of its large range. In this example, we will estimate its density function among depressed patients.

Data for both the SCID and the HAM-D were collected from all participating patients in this study; therefore, similar to [4] we used a subset that resembled data that would be obtained from a two-phase design. Hence, we can assess the behaviors of the estimates by comparing them to the estimates based on the complete data. In this subset, the HAM-D results were available for all patients, but the SCID diagnoses were available only for certain patients selected according to the following mechanism:

$$\Pr(\text{SCID available})=0.15+0.55\, I[\text{CIRS}>7]+0.30 I[\text{Age}<75]. \tag{22}$$

Thus, the verification mechanism preferentially selected the patients who were under the age of 75, with a relatively high cumulative illness burden. Using this mechanism, 394 of the 708 patients (55.6%) were selected for SCID verification of the depression diagnosis. A logistic regression model with age and CIRS as the predictors was used to model the missing data mechanism.

The true density function is not known in this real data example; however, since there is no missing values in HAM-D and SCID in the data set, we used the estimate based on the complete data set as a reference in assessing the quality of the estimate under missing data. To model the missing data mechanism, we assumed that the missingness was related to CIRS and age, and applied the following generalized linear model with a logistic link to estimate the relationship:

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

Thus, the model used for the missing mechanism is not exactly the one we used to generate the missing values (the predictors are not dichotomized in the regression).

Shown in the plot below are the estimates of the density using the complete data set of all subjects who were depressed (complete), naive estimate (naive), and IPW estimates (IPW for IPW with estimated missing probabilities and IPWK for IPW with known missing probabilities). The IPW estimates are in general closer to, as compared with the naive estimate, the estimate based on the complete data. Note that although the estimated weight function was not based on the exact model for generating the missing values, bias was greatly reduced under the proposed approach (comparing to the naive method). It is also interesting to note that the IPW estimate using the known missing data probabilities (IPWK) is not as good as the one based on the estimated weight (IPW). Such a difference is also observed in various other settings involving IPW estimates, including parametric regressions[12] and U-statistic estimates[4].

7 Discussion

In this paper we generalized the kernel smoothing density estimates for diagnostic test outcomes when the true disease status is subject to missing. Through the study, it is clear that the naive approach that ignores the missing values of disease status and uses only those with known disease membership is not valid. When the missing mechanism is well understood and characterized by the MAR mechanism, the proposed methodology works well in reducing the bias. Note that the information in X_i are only used in the modeling of the missing mechanism (π_i). If the relationship between the disease status D_i and (T_i, X_i) is well understood, then by including this additional model on the disease mechanism we may obtain more efficient estimates.

There are some problems left unanswered for our methods. Most notable is the study of boundary effects. It is well known that many smoothing methods do not behave well without adjustment near the boundaries. Our approach is no exception, as indicated by the relatively poor behaviors of the estimates near the boundaries. Further investigation is necessary to address such biases.

Another aspect that needs future study is the effect of MAR on the estimates. Although MAR is commonly assumed in the literature and satisfied by many studies in practice, missing not at random (MNAR) may arise in some studies. When enough information is collected (via covariates) and included in modeling 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 MNAR? All these need future studies.

Acknowledgments

This research was supported in part by Award Number R21 DA027521-01 from the National Institute On Drug Abuse. 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.

References

1. Alonzo TA, Pepe MS, Lumley T. Estimating disease prevalence in two-phase studies. Biostatistics. 2003; 4:313–326. [PubMed: 12925524]
2. Carroll, RJ.; Ruppert, D.; Stefanski, LA. Measurement Error in Nonlinear Models. London: Chapman and Hall; 1995.
3. Chaudron LH, Szilagyi PG, Tang W, Anson E, Talbot NL, Wadkins HIM, Tu X, Wisner KL. Accuracy of Depression Screening Tools for Identifying Postpartum Depression Among Urban Mothers. Pediatrics. 2010:e609–e617. [PubMed: 20156899]
4. He H, Lyness JM, McDermott MP. Direct estimation of the area under the receiver operating characteristic curve in the presence of verification bias. Stat Med. 2009; 28(3):361–376. [PubMed: 18680124]
5. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association. 1952; 47:663–685.
6. Linn BS, Linn MW, Gurel L. Cumulative illness rating scale. Journal of the American Geriatrics Society. 1968; 16:622–626. [PubMed: 5646906]
7. Parzen, Emanuel. On estimation of a probability density function and mode. Ann Math Statist. 1962; 33:1065–1076.
8. Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. Journal of the American Statistical Association. 1995; 90:122–129.
9. Rotnitzky A, Faraggi D, Schisterman E. Doubly robust estimation of the area under the receiver-operating characteristic curve in the presence of verification bias. Journal of the American Statistical Association. 2006; 101:1276–1288.
10. Simonoff, JS. Smoothing Methods in Statistics. Springer; New York: 1996.
11. Spitzer, RL.; Gibbon, M.; Williams, JBW. Structured Clinical Interview for Axis I DSM-IV Disorders. Biometrics Research Department, New York State Psychiatric Institute; 1994.
12. Tsiatis, Anastasios A. Springer Series in Statistics. Springer; New York: 2006. Semiparametric theory and missing data.
13. Wand, MP.; Jones, MC. Monographs on Statistics and Applied Probability. Vol. 60. Chapman and Hall Ltd; London: 1995. Kernel smoothing.
14. Wang, Qihua. Probability density estimation with data missing at random when covariables are present. J Statist Plann Inference. 2008; 138(3):568–587.
15. White H. Maximum likelihood estimation of misspecified models. Econometrica. 1982; 50(1):1–25.

16. Williams JBW. A structured interview guide for the Hamilton Depression Rating Scale. Archives of General Psychiatry. 1988; 45:742–747. [PubMed: 3395203]

Appendix

In this appendix we give a proof of Lemma 3 using estimating equation techniques. Note that under MAR, we have

$$E\left(\frac{D_i V_i}{\pi_i}\right)=p, \tag{23}$$

and

$$E\left[\frac{D_i V_i}{\pi_i} K_h(t-T_i)\right]=E[K_h(t-T_i)|D_i=1]=pf_h(t). \tag{24}$$

When the parameters of the missing mechanism, β, is known and the true value is used, we can obtain the estimate $\tilde{f}(t)$ by solving the following estimating equations

$$\begin{aligned}\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\left[\frac{D_iV_i}{\pi_i}K_h(t-T_i)-pf_h(t)\right]=0\\ \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\left(\frac{D_iV_i}{\pi_i}-p\right)=0\end{aligned} \tag{25}$$

If β is estimated from the data using the logistic model (14), the estimate $\hat{f}(t)$ can be obtained by solving the system of estimating equations consisting of (25) and (14), and we can use the technique of stacking estimating equations. Since these estimating equations are unbiased, the estimates $\tilde{f}(t)$ and $\hat{f}(t)$ are consistent estimates of $f_h(t)$. See appendix A.3 of [2] for details about estimating equations.

Proof of Lemma 3 (a)

Let $\psi_1=\frac{D_iV_i}{\pi_i}K_h(t-T_i)-pf_h(t)$, and $\psi_2=\frac{D_iV_i}{\pi_i}-p$, then $\sqrt{n}\left(\tilde{f}(t)-f_h(t)\right)\to N(0,\sigma^2)$ where the asymptotic variance σ^2 equals to the (1, 1) term of the matrix $A^{-1}BA^{-T}$, where

$A=E\begin{pmatrix}\frac{\partial\psi_1}{\partial f_h} & \frac{\partial\psi_1}{\partial p}\\ \frac{\partial\psi_2}{\partial f_h} & \frac{\partial\psi_2}{\partial p}\end{pmatrix}=\begin{pmatrix}-1 & -f_h(t)\\ 0 & -1\end{pmatrix}$ and

$$\begin{aligned}B=&E\begin{pmatrix}\psi_1^2 & \psi_1\psi_2\\ \psi_1\psi_2 & \psi_2^2\end{pmatrix}\\ =&\begin{pmatrix}E\left[\frac{D_iV_i}{\pi_i}K_h(t-T_i)\right]^2-[pf_h(t)]^2 & E\left[\frac{D_iV_i}{\pi_i}K_h(t-T_i)\frac{D_iV_i}{\pi_i}-pf_h(t)\frac{D_iV_i}{\pi_i}\right]\\ E\left[\frac{D_iV_i}{\pi_i}K_h(t-T_i)\frac{D_iV_i}{\pi_i}-pf_h(t)\frac{D_iV_i}{\pi_i}\right] & E\left(\frac{D_iV_i}{\pi_i}\right)^2-p^2\end{pmatrix}.\end{aligned}$$

Let $e=E(\frac{1}{\pi}|D=1)$, $g=E(\frac{K}{\pi}|D=1)$, and $E(\frac{K^2}{\pi}|D=1)$, then by simple matrix computation, $\sigma^2 = p(ef^2 - 2gf + s)$. It is straightforward to verify that

$p(ef^2-2gf+s)=Var\left[\frac{D_iV_i}{p\pi_i}(K_h(t-T_i)-f_h(t))\right].$

Proof of Lemma 3 (b)

Let $\Psi_3 = (V_i - \pi_i\,(\beta))\mathbf{x}_i$. Using the technique of stacking estimating equation, it follows that $\sqrt{n}\left(\widehat{f}(t) - f_h(t)\right) \to N(0, \sigma^2)$ where the asymptotic variance σ^2 is the (1, 1) term of the matrix $A^{-1}BA^{-T}$, where $A=E\begin{pmatrix} \frac{\partial\psi_1}{\partial f_h} & \frac{\partial\psi_1}{\partial p} & \frac{\partial\psi_1}{\partial\theta^T} \\ \frac{\partial\psi_2}{\partial f_h} & \frac{\partial\psi_2}{\partial p} & \frac{\partial\psi_2}{\partial\theta^T} \\ \frac{\partial\Psi_3}{\partial f_h} & \frac{\partial\Psi_3}{\partial p} & \frac{\partial\Psi_3}{\partial\theta^T} \end{pmatrix}=E\begin{pmatrix} -1 & -f_h(t) & -\frac{1-\pi_i}{\pi_i}K_h(t-T_i)D_iV_i\mathbf{x}_i^T \\ 0 & -1 & -\frac{1-\pi_i}{\pi_i}D_iV_i\mathbf{x}_i^T \\ 0 & 0 & \pi_i(\beta)\,(1-\pi_i(\beta))\mathbf{x}_i\mathbf{x}_i^T \end{pmatrix}$ and $B=E\begin{pmatrix} \psi_1^2 & \psi_1\psi_2 & \psi_1\Psi_3^T \\ \psi_1\psi_2 & \psi_2^2 & \psi_2\Psi_3^T \\ \psi_1\Psi_3 & \psi_2\Psi_3 & \Psi_3\Psi_3^T \end{pmatrix}$.

Let $\mathbf{c}_1 = E\,[(1-\pi_i)\,\mathbf{x}_i|D_i = 1]$ and $\mathbf{c}_2 = E\,[(1-\pi_i)\,K_h(t-T_i)\mathbf{x}_i|D_i = 1]$, then

$$B=\begin{pmatrix} ph-p^2f_h(t) & pg-p^2f_h(t) & pc_2^T \\ pg-p^2f_h(t) & pe-p^2 & pc_1^T \\ p\mathbf{c}_2 & p\mathbf{c}_1 & E\,[\pi_i(\beta)\,(1-\pi_i(\beta))\mathbf{x}_i\mathbf{x}_i^T] \end{pmatrix}.$$

Through tedious algebraic computation it can be proved that

$$\sigma^2=Var\left(\frac{D_iV_i}{p\pi_i}[K_h(t-T_i)-f_h(t)]+(\mathbf{c}_2-\mathbf{c}_1f_h(t))\,I^{-1}(\beta)\,s(\mathbf{x}_i,\beta)\right).$$

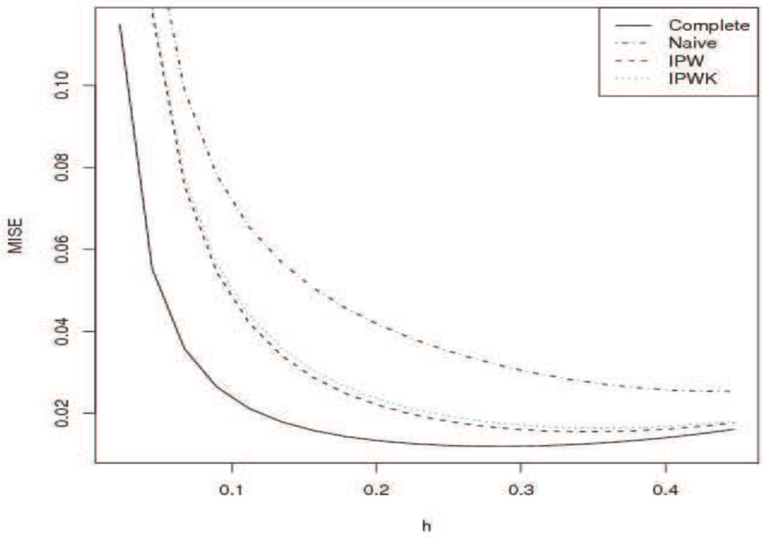

Figure 1.

MISE (Complete: solid, Naive: dashed, IPW: dotted, IPWK: dotdash)

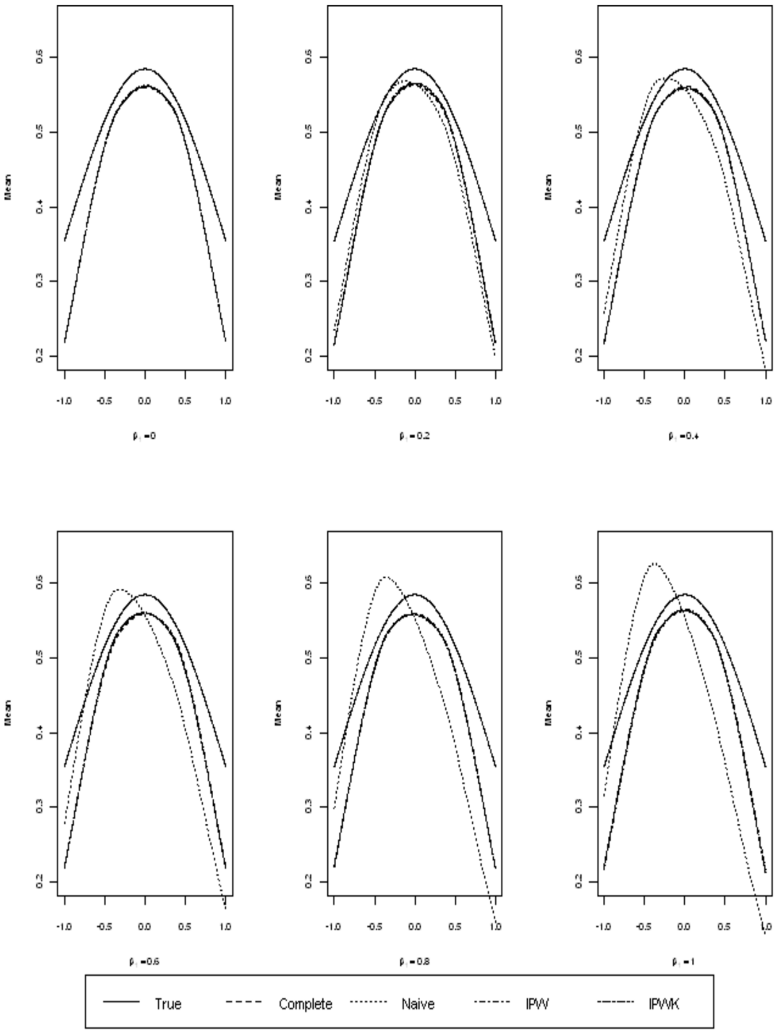

Figure 2.

Means for β_1 = 0, .2, .4, .6, .8, and 1. (True: solid, Complete: dashed, Naive: dotted, IPW: dotdash, IPWK: twodash)

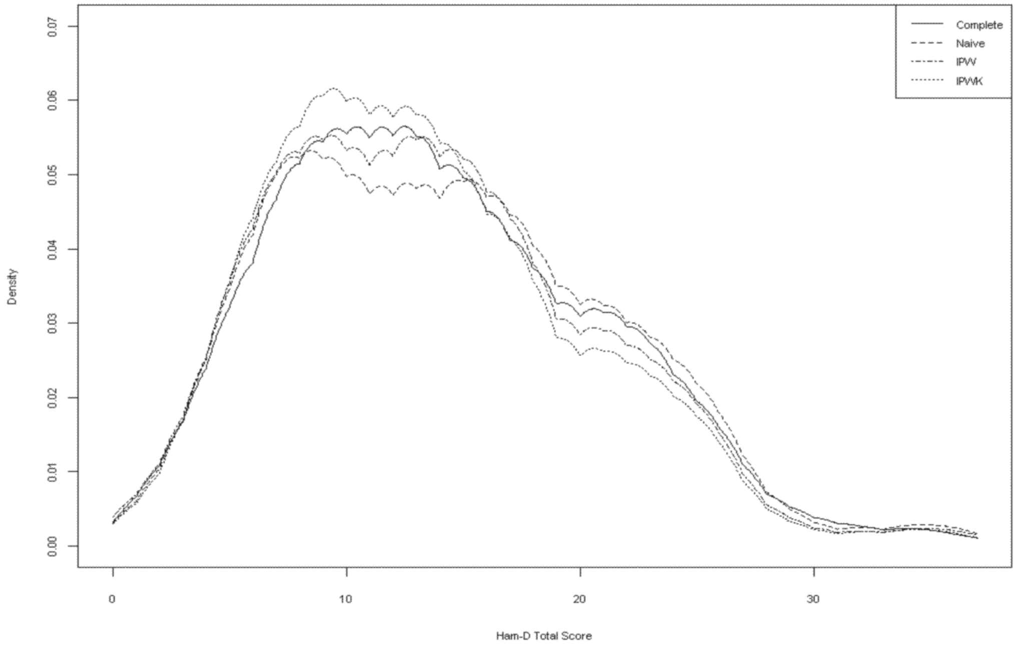

Figure 3.

Density estimate of Ham-D taotal score among depressed patients: complete (solid), naive (dashed), IPW (dotted), IPWK(dotdash)