Australian & New Zealand Journal of Statistics

Aust. N. Z. J. Stat. **51**(3), 2009, 247–270 doi: 10.1111/j.1467-842X.2009.00541.x

KERNEL DENSITY ESTIMATION WITH MISSING DATA AND AUXILIARY VARIABLES

SUZANNE R. DUBNICKA

Kansas State University

Summary

In most parametric statistical analyses, knowledge of the distribution of the response variable, or of the errors, is important. As this distribution is not typically known with certainty, one might initially construct a histogram or estimate the density of the variable of interest to gain insight regarding the distribution and its characteristics. However, when the response variable is incomplete, a histogram will only provide a representation of the distribution of the observed data. In the AIDS Clinical Trial Study protocol 175, interest lies in the difference in CD4 counts from baseline to final follow-up, but CD4 counts collected at final follow-up were incomplete. A method is therefore proposed for estimating the density of an incomplete response variable when auxiliary data are available. The proposed estimator is based on the Horvitz–Thompson estimator, and the propensity scores are estimated nonparametrically. Simulation studies indicate that the proposed estimator performs well.

Key words: Horvitz–Thompson estimator; integrated squared error; missing at random; kernel smoothing.

1. Introduction

In clinical trials and other large studies, data are collected on many variables, although it is often one variable that is of particular interest. When analysing the data from such studies, one typically carries out inference on summary measures regarding the response variable of interest. For example, one may fit a linear model by estimating the regression coefficients, or conduct a hypothesis test regarding the mean response. Only limited information is gained by restricting analyses to these parameters and summary measures.

However, the probability density function of a variable can provide information that summary measures, such as means and variances, cannot. Knowledge of this density may be necessary for verifying assumptions of other statistical procedures, but the density itself may also be of direct interest. In particular, we learn something about how the response variable varies by creating a histogram of the data or obtaining a kernel density estimate, both of which approximate the true density. Therefore, an estimate of the response variable density can serve as a useful descriptive tool.

One common complication with large clinical trials is incomplete data. Many methods have been developed to allow the usual statistical analyses to be carried out in the presence of missing data. For example, the EM algorithm (Dempster, Laird & Rubin 1977) can be used to obtain maximum likelihood estimates when some data are missing, and multiple imputation provides a means for ‘filling in’ missing values with plausible values and employing the usual parametric methods with some adjustment (Little & Rubin 2002). In addition, weighted

Department of Statistics, 101 Dickens Hall, Manhattan, KS 66506, USA.
e-mail: dubnicka@stat.ksu.edu

Acknowledgments. The author would like to thank the reviewers and the technical editor for their thoughtful comments and attention to detail.

estimating equations (Robins, Rotnitzky & Zhao 1994) can be constructed in a wide variety of situations to estimate parameters when data are incomplete. Note that all of these methods are concerned with the estimation of unknown parameters.

In this paper, we are interested in obtaining a broader overview of the data than parameter estimates or other summary measures can provide. However, when a response variable is not completely observed, we can no longer use graphical techniques such as histograms to represent the density of the response variable. In such cases, a histogram would provide us only with an approximation of the density based on the observed data, rather than the full data. This approximation may be a biased estimate of the true density. Therefore, we are specifically concerned with the problem of estimating the density of a univariate response variable when the data are incomplete. These density estimates can be used for descriptive purposes, as described above, or for inference, such as testing the equality of densities for two or more treatment groups. We focus on the process of estimating the density with incomplete data in this paper. The important question of testing the equality of two or more densities with incomplete data will be addressed in a subsequent paper.

Consider the AIDS Clinical Trial Study protocol 175 (ACTG 175) (Hammer *et al*. 1996). ACTG 175 was a randomized clinical trial designed to compare four antiretroviral therapies for improving the immune systems of AIDS patients. CD4 counts were recorded for each patient at the beginning of the study (baseline), at 20 ± 5 weeks and at 96 ± 5 weeks (final follow-up), as depressed CD4 counts indicate impairment of the immune system. As positive changes in CD4 counts are thought to reflect more effective treatment, an important variable of interest is the difference in CD4 count from baseline to final follow-up. However, as with most long-term clinical trials, the data are incomplete. In particular, CD4 counts were not observed for 37% of the patients at final follow-up. However, data were recorded on several other baseline variables that are considered to be useful in predicting missingness.

The ACTG 175 data have been studied previously to determine whether there is a difference between the monotherapy and the other three treatments combined in terms of mean difference in CD4 counts from baseline to final follow-up (Davidian, Tsiatis & Leon 2005). In that analysis, three of the therapies were combined because they were shown to be the same in terms of mean difference in CD4 counts. However, the difference in CD4 counts from baseline to follow-up may differ in other ways that could be useful to doctors prescribing these drugs. An estimate of the density of CD4 counts would potentially provide additional useful information regarding these therapies. In this paper, we propose a method for estimating the density of the difference in CD4 counts from baseline to final follow-up under the assumption that CD4 counts at final follow-up are missing at random (MAR). The additional variables, referred to here as auxiliary variables, will be utilized to estimate the probability that the CD4 count is missing at final follow-up.

Recently, a method has been proposed for estimating the density of fish ages from one population for which only covariate information is observed, based on another population for which both the fish ages and covariate information are observed, under the assumption that the conditional distributions of the covariates given the response are the same for the two populations (Elmore, Hall & Troynikov 2006). In essence, the authors consider the situation in which the responses are missing by design, which is more restrictive than the missing-at-random assumption. However, the methods proposed in this paper should also apply to the problem of estimating fish ages, provided that the distribution of the fish ages is indeed the same for the two populations.

The notation used in this paper is consistent with notation commonly used in the missing-data literature. Let $y_1, \dots, y_n$ denote a random sample from a distribution with smooth density g, and let $\boldsymbol{x}_i = (x_{1i}, \dots, x_{pi})$ denote the auxiliary variables for the ith subject. Furthermore, let $r_i = 1$ if y_i is observed and $r_i = 0$ otherwise. As the responses are missing at random, the probability that y_i is observed depends only on the observed data: $\pi_i = Pr(r_i = 1 \mid y_i, \boldsymbol{x}_i) = Pr(r_i = 1 \mid \boldsymbol{x}_i)$. This probability is often called the propensity score (Rosenbaum & Rubin 1983). Our goal is to estimate the density g.

2. Horvitz–Thompson-type kernel density estimators

When the responses are completely observed, one would estimate g with the full-data kernel density estimator (KDE):

$$\hat{g}_h^{\mathrm{F}}(y) = \frac{1}{n} \sum_{i=1}^{n} K_h(y - y_i),$$

where $K_h(\cdot) = (1/h)K(\cdot/h)$ with K denoting the kernel and h the bandwidth. However, if some of the responses are missing, $\hat{g}_h^F(y)$ cannot be computed. A naive estimate of g when some y_i are missing is the complete case estimator, which simply estimates g with the data at hand:

$$\hat{g}_h^{\mathrm{CC}}(y) = \frac{1}{r} \sum_{i=1}^{n} r_i K_h(y - y_i),$$

where $r = \sum_{i=1}^{n} r_i$ is a random variable. However, it is well known that, unless data are missing completely at random, complete-case analyses generally result in biased estimators (Little & Rubin 2002). As the density estimator $\hat{g}_h^{\mathrm{F}}(y)$ is already a biased estimator of g, we would expect $\hat{g}_h^{\mathrm{CC}}(y)$ to produce a density estimator with greater bias. This suspicion is confirmed in the simulation studies.

In finite-population sampling, the Horvitz–Thompson estimator (Horvitz & Thompson 1952) is used to produce an unbiased estimator of the population total of a finite population under unequal probability sampling. In this spirit, when some y_i are missing at random and the true propensity scores $\pi_i = \pi(\boldsymbol{x}_i) = Pr(r_i \mid \boldsymbol{x}_i)$ are known, we consider a Horvitz–Thompson-type estimator of g:

$$\hat{g}_h^{\mathrm{HTt}}(y) = \frac{1}{n} \sum_{i=1}^{n} \frac{r_i}{\pi_i} K_h(y - y_i). \tag{1}$$

The superscript HTt indentifies this KDE as a Horvitz–Thompson-type estimator with true propensity scores. This estimator is reminiscent of weighted estimating equations (Robins *et al.* 1994; Davidian *et al.* 2005), which are also based on the Horvitz–Thompson estimator. In Section 4, the theoretical properties of $\hat{g}_h^{\mathrm{HTt}}(y)$ are discussed. Effectively, the bias of $\hat{g}_h^{\mathrm{HTt}}(y)$ is the same as that of the full-data KDE $\hat{g}_h^{\mathrm{F}}(y)$, but the mean integrated squared error (MISE) of $\hat{g}_h^{\mathrm{HTt}}(y)$ is greater than that of $\hat{g}_h^{\mathrm{F}}(y)$. This increase in MISE is to be expected because the estimator is essentially based on a smaller sample size.

As the true propensity scores π_i are typically known only in a few cases, such as when y_i is subject to missingness by design, we must estimate π_i. In practice, we would use the estimator $\hat{g}_h^{\mathrm{HTt}}(y)$ with π_i replaced with an estimate $\hat{\pi}_i$:

$$\hat{g}_h^{\mathrm{HT}}(y) = \frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\hat{\pi}_i}K_h(y - y_i). \tag{2}$$

The superscript HT refers to a Horvitz–Thompson-type estimator with estimated propensity scores.

Note that, asymptotically, the bias of $\hat{g}_h^{\mathrm{HT}}(y)$ is the same as that of $\hat{g}_h^{\mathrm{HTt}}(y)$, but its variance is actually smaller. This follows because $\hat{g}_h^{\mathrm{HT}}(y)$ uses all of the available data while $\hat{g}_h^{\mathrm{HTt}}(y)$ uses only data for completely observed subjects. As the MISE of an estimator is computed from its bias and variance, the MISE of the Horvitz–Thompson-type KDE is smaller when estimated propensity scores are used than then true propensity scores are used. See Sections 3 and 4 for further details.

The major disadvantage of $\hat{g}_h^{\mathrm{HTt}}(y)$, however, is that it is not a density. That is,

$$\int \hat{g}_h^{\mathrm{HT}t}(y)\,dy = \frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\pi_i}\int K(u)\,du = \frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\pi_i} \neq 1,$$

where $\int K(u)\,du = 1$ because K is a density function. To correct this problem, we propose a modified version of the Horvitz–Thompson KDE:

$$\hat{g}_h^{\mathrm{HTM}}(y) = \frac{1}{n^*}\sum_{i=1}^{n}\frac{r_i}{\hat{\pi}_i}K_h(y - y_i), \tag{3}$$

where $n^* = \sum_{i=1}^{n} r_i/\hat{\pi}_i$. Now, the modified Horvitz–Thompson-type KDE (3) with true π_is replacing $\hat{\pi}_i$s, denoted $\hat{g}_h^{\mathrm{HTMt}}(y)$, is a density, but its bias and MISE will differ somewhat from those of $\hat{g}_h^{\mathrm{HTt}}(y)$ and $\hat{g}_h^{\mathrm{F}}(y)$. Note that, unless $\hat{\pi}_i$ is very small and $r_i = 1, n^*$ does not tend to differ substantially from n in practice. Empirical and theoretical properties of the Horvitz–Thompson-type KDEs are discussed in Sections 3 and 4.

2.1. Estimating propensity scores

To obtain an estimate of π_i, we will exploit the information contained in the auxiliary variables $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{pi})$, which are related to y_i. If $\boldsymbol{x}_i$ is discrete with small dimensionality, then a natural unbiased estimator of π_i is

$$\hat{\pi}_i = \hat{\pi}(\boldsymbol{x}_i) = \frac{\sum_{j=1}^{n} r_j I(\boldsymbol{x}_j = \boldsymbol{x}_i)}{\sum_{j=1}^{n} I(\boldsymbol{x}_j = \boldsymbol{x}_i)},$$

the empirical proportion based on the observed data (Qi, Wang & Prentice 2005), where I is the indicator function. If $\boldsymbol{x}_i$ contains continuous elements, then the Nadaraya–Watson (local mean) estimator (Nadaraya 1964; Watson 1964) can be used:

$$\hat{\pi}_{\mathrm{NW}i} = \hat{\pi}_{\mathrm{NW}}(\boldsymbol{x}_i) = \frac{\sum_{j=1}^{n} r_j K_{h_1}^*(\boldsymbol{x}_i - \boldsymbol{x}_j)}{\sum_{j=1}^{n} K_{h_1}^*(\boldsymbol{x}_i - \boldsymbol{x}_j)}, \tag{4}$$

where $K_{h_1}^*(\cdot) = K^*(\cdot/h_1)$ and K^* is a kernel function.

However, there are some conceptual disadvantages to using the Nadaraya–Watson estimator, given by (4), or any local linear or local polynomial regression estimator. In particular, the binary nature of r_i has been ignored. The motivation behind (4) is provided by the model

$$r_i = \pi(\boldsymbol{x}_i) + \varepsilon_i,$$

which cannot hold if $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. continuous random variables. However, $E(r_i \mid \boldsymbol{x}_i) = \pi(\boldsymbol{x}_i)$ as required. Furthermore, the local linear estimator of π_i is not guaranteed to lie in the interval (0, 1), which is clearly undesirable when estimating probabilities, although the Nadaraya–Watson estimator does not have this problem.

Local likelihoods (Tibshirani & Hastie 1987; Fan, Heckmann & Wand 1995) can provide estimates of the propensity scores that account for the binary nature of the response variable. The local likelihood approach to estimating the propensity scores is an extension of the ordinary likelihood approach whereby the relationship between $\pi(\boldsymbol{x})$ and $\boldsymbol{x} = (x_1, \ldots, x_p)$ is modelled locally instead of globally. That is, the local likelihood for our binary response variable is the weighted likelihood given by

$$l_{\boldsymbol{x}}(\alpha, \boldsymbol{\beta}) = \sum_{i=1}^{n} l_i(\alpha, \boldsymbol{\beta}) K_{h_1}(x_{1i} - x_1) \cdots K_{h_p}(x_{pi} - x_p),$$

where

$$l_i(\alpha, \boldsymbol{\beta}) = r_i \log[g^{-1}\{\alpha + \boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{x}_i - \boldsymbol{x})\}] + (1 - r_i) \log[1 - g^{-1}\{\alpha + \boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{x}_i - \boldsymbol{x})\}]$$

is the contribution to the log-likelihood from the ith observation with g being the chosen link function, and $K_{h_j}(\cdot)$ is a kernel function. Maximization of $l_{\boldsymbol{x}}(\alpha, \boldsymbol{\beta})$ at $\boldsymbol{x}$ with respect to $(\alpha, \boldsymbol{\beta}^{\mathrm{T}})$ provides local estimates $(\hat{\alpha}(\boldsymbol{x}), \hat{\boldsymbol{\beta}}^{\mathrm{T}}(\boldsymbol{x}))$, and the estimate of $\pi(\boldsymbol{x})$ at $\boldsymbol{x}$ is $\hat{\pi}(\boldsymbol{x}) = g^{-1}(\hat{\alpha}(\boldsymbol{x}))$. Thus, a local likelihood estimator of π_i is

$$\hat{\pi}_{\mathrm{LL}i} = \hat{\pi}_{\mathrm{LL}}(\boldsymbol{x}_i) = g^{-1}(\hat{\alpha}(\boldsymbol{x}_i)). \tag{5}$$

The link function would be chosen as in the case of ordinary (global) binary regression.

In choosing between Nadaraya–Watson estimates, given by (4), and local likelihood estimates, given by (5), of the propensity scores, one must consider the trade-off between explicitly accounting for the binary nature of the response and computational difficulty. Although local likelihood estimates explicitly account for the binary response, the computation of these estimates is more expensive than that of the Nadaraya–Watson estimates, owing to the need to maximize the local likelihood. Furthermore, one must also choose an appropriate link function. However, both the Nadaraya–Watson estimates and the local likelihood estimates of the propensity scores are more flexible than ordinary (global) binary regression estimates, because they allow for the propensity scores to be a non-monotone function of the auxiliary variables. Nottingham, Birch & Bodt (2000) give a nice discussion and example of this phenomenon. Nadaraya–Watson and local likelihood estimates, using the logit link, of the propensity scores are included in the simulation study of Section 3. For now, we simply note that, although the local likelihood estimates and the Nadaraya–Watson estimates will differ from one another, the corresponding Horvitz–Thompson-type KDEs perform similarly in the simulation study.

2.2. Bandwidth selection

A critical decision in kernel density estimation is the choice of the bandwidth or smoothing parameter h. Many different methods for bandwidth selection have been proposed, including normal optimal smoothing, least squares (unbiased) cross-validation (Rudemo 1982; Bowman 1984), biased cross-validation (Scott & Terrell 1987), direct plug-in rules (Park & Marron 1990; Sheather & Jones 1991), and smoothed cross-validation (Müller 1987; Staniswalis 1989; Hall, Marron & Park 1991). The Sheather–Jones plug-in method (Sheather & Jones 1991) was shown in a simulation study (Jones, Marron & Sheather 1996) to perform consistently well.

When some responses are missing, the choice of bandwidth is complicated further. One approach to bandwidth selection when some responses are missing is simply to use one of the above methods on the complete cases. In the simulation study in the next section, three of the above methods are applied to the complete cases: normal optimal smoothing, least squares cross-validation, and the Sheather-Jones plug-in method. Although rather naive, the simulation study will show that complete-case methods are reasonably effective.

As the variability in the complete cases may underestimate the variability of the full data, the bandwidth selection could be improved. Least squares cross-validation lends itself nicely to the incorporation of weights such as the inverse probability weights of the Horvitz–Thompson-type kernel density estimates. Thus, a modified least squares cross-validation bandwidth selection procedure is also worth investigating. Using the same approach as that for deriving the ordinary least squares cross-validation criterion, the weighted least squares cross-validation criterion

$$\mathrm{WLSCV}(h) = \int \hat{g}_h^{\mathrm{HTt}}(y)^2 dy - \frac{2}{n} \sum_{i=1}^{n} \frac{r_i}{\pi_i} \hat{g}_{h,-i}^{\mathrm{HTt}}(y_i),$$

where $\hat{g}_{h,-i}^{\mathrm{HTt}}(y)$ is the Horvitz–Thompson-type KDE computed without observation i, $\hat{g}_{h,-i}^{\mathrm{HTt}}(y) = 1/(n-1)\sum_{j \neq i}(r_i/\pi_i)K_h(y - y_i)$, is an unbiased estimator of the difference $\mathrm{MISE}\{\hat{g}_h^{\mathrm{HTt}}(y)\} - g(y)$. The bandwidth h is selected to minimize WLSCV (h). In practice, $\hat{g}_h^{\mathrm{HTt}}(y)$ is replaced by either $\hat{g}_h^{\mathrm{HT}}(y)$ or $\hat{g}_h^{\mathrm{HTM}}(y)$ with either Nadaraya–Watson or local likelihood estimates of the propensity scores.

3. Numerical results

3.1. Simulation study

A simulation study was conducted to compare the performance of the Horvitz–Thompson-type KDEs with that of the full-data KDE and the complete-case KDE. The Horvitz–Thompson-type estimator and the modified Horvitz–Thompson-type estimator with Nadaraya–Watson and local likelihood estimates of the propensity scores as well as with the true propensity scores were evaluated. For all KDEs, the normal kernel was used: $K_h(\cdot) = (1/h)\phi(\cdot/h)$. Four bandwidth selection methods were used for the Horvitz–Thompson-type KDE: normal optimal smoothing (Norm), least squares cross-validation (CV), Sheather–Jones plug-in (SJ) based on the complete cases, and weighted least squares cross-validation (CV-WTD). For the full-data and complete-case KDEs, the first three bandwidth selection procedures were employed, based on the full data and the complete cases, respectively.

We considered the case of one auxiliary variable and generated the pairs $\{(x_i,\ y_i)$: $i=1,\dots,n\}$ by first generating y_i distributed as F_Y and x_i^* distributed as F_{X^*}, independently, $i=1,\dots,n$. We then let $x_i=\rho y_i+\sqrt{1-\rho^2}x_i^*,\ -1<\rho<1$, so that y_i and x_i were correlated. To create an incomplete set of responses, we then generated $r_i \sim$ Bernoulli $(\pi_i), i=1,\dots,$ n, where logit $(\pi_i)=\beta_0+\beta_1 x_i,\ -\infty<\beta_0,\ \beta_1<\infty$. We evaluated the performance of the density estimators by means of integrated squared errors (ISEs).

3.2. Results

A moderate sample size of $n=300$ was considered. The responses $y_1,\dots,y_n$ were generated from one of three distributions based on a mixture of normal distributions: symmetric, bimodal and skewed. Normal mixtures were chosen because they are quite flexible and the ISE is easy to compute (Marron & Wand 1992). The symmetric distribution was $f_1(y)=\phi(y;\ 3/2, 3/2)$, the bimodal distribution was $f_2(y)=0.5\phi(y;\ 0,\ 1)+0.5\phi(y;\ 4,\ 1)$, and the skewed distribution was $f_3(y)=0.4\phi(y;\ 2,\ 2)+0.6\phi(y;\ 37/12,\ 10/9)$, where $\phi(y;\ \theta,\ \sigma)$ represents the normal density with mean θ and standard deviation σ. We generated $x_1^*,\dots,\ x_n^*$ from the standard normal distribution and set $\rho=0.5$ and 0.8. When generating $r_1,\dots,\ r_n$, we selected values of β_0 and β_1 to achieve 17% and 30% missing data; the exact values β_0 and β_1 depended on the distribution of the responses.

The various combinations of response distribution, correlation, and percent missingness resulted in 12 different simulation situations. For each situation, the ISE of each density estimate, where ISE$\{\hat{g}_h(y)\}=\int(\hat{g}_h-g)^2$, was computed for each of the 501 datasets generated. The following density estimates were compared: full data (Full); complete case (CC); Horvitz–Thompson-type with true propensity scores (HT) and with Nadaraya–Watson estimates (HT-NW) and local likelihood estimates (HT-LL) of the propensity scores; and modified Horvitz–Thompson-type with true propensity scores (HTM) and with Nadaraya–Watson estimates (HTM-NW) and local likelihood estimates (HTM-LL) of the propensity scores. The logit link was used in the local likelihood approach to estimating propensity scores. Note that each of the aforementioned KDEs was computed using each of the four bandwidth selection procedures (Norm, CV, SJ, CV-WTD).

To minimize redundancy, we will present only a subset of the results of the 12 simulation situations, as the same general trend is observed in most situations. Specifically, we will present results for the situations in which 30% of the responses are missing and $\rho=0.8$. Differences in results obtained for other combinations of missingness and correlation will be noted when they exist. Furthermore, we primarily present KDEs with the Sheather–Jones bandwidth selection as it is consistently good for all three response variable distributions. Those interested in seeing the results for situations not presented here may contact the author.

First, we compared all eight KDE procedures. In all situations considered, the ISEs were consistently larger and more variable for the complete-case KDE than the ISEs for all other methods. This shows that using the naive complete-case approach will result in KDEs that are much more biased and variable estimates of the true density than the full-data KDE. Figure 1 presents boxplots of the ISEs of the eight methods for all three response variable distributions for the situation in which 30% of the responses are missing on average and the correlation

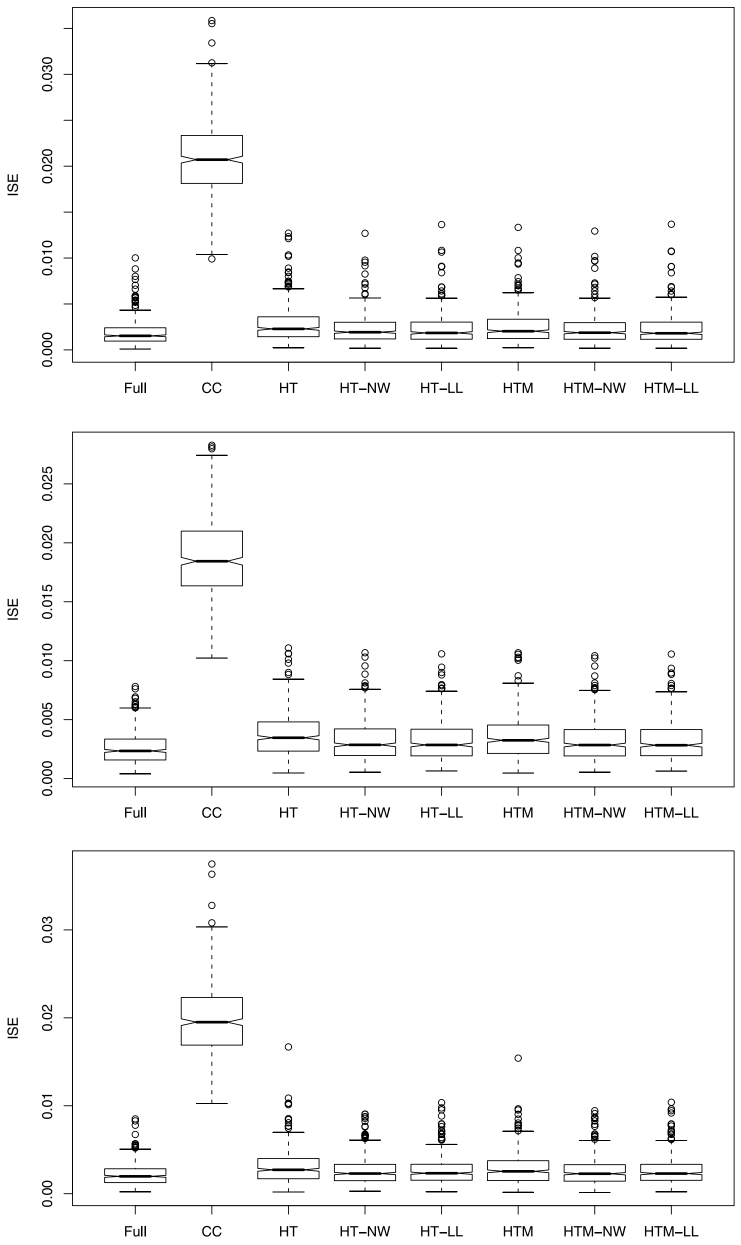

Figure 1. Boxplots of observed integrated squared errors of all kernel density estimators with Sheather–Jones bandwidth for (a) normal, (b) bimodal and (c) skewed distributions with 30% missing data and correlation between the response and auxiliary variable of $\rho = 0.8$. Based on 501 simulated datasets of size 300. Kernel density estimators: full data (Full); complete case (CC); Horvitz–Thompson with true propensity scores (HT), with Nadaraya–Watson estimates (HT-NW), and with local likelihood estimates (HT-LL) of the propensity scores; modified Horvitz–Thompson with true propensity scores (HTM), with Nadaraya–Watson estimates (HTM-NW), and with local likelihood estimates (HTM-LL) of the propensity scores.

between y and x is moderate ($\rho = 0.8$). Note that, although the local likelihood and Nadaraya–Watson estimates of the propensity scores differed from one another, the performances of the corresponding KDEs were approximately the same. The observed relationship among the ISEs of the different density estimates was also seen in the other situations, although the magnitude and variability of the ISEs differ somewhat.

We have chosen to present the results for KDEs using the Sheather–Jones bandwidth selection procedure, as it performs consistently well regardless of the response variable distribution. For example, boxplots of ISEs of the modified Horvitz–Thompson-type KDE with Nadaraya–Watson estimates of π_i are presented in Figure 2. When the response variable was normally distributed, the cross-validation bandwidth selection procedure produced larger and more variable ISEs, whereas the normal optimal and Sheather–Jones bandwidth selection procedures produced ISEs that had virtually the same distribution (Fig. 2a). This was also true when the response variable had a skewed distribution (Fig. 2c). However, when the response variable had a bimodal distribution, the Sheather–Jones bandwidth procedure produced substantially smaller ISEs (Fig. 2b). This trend was seen with the other KDEs and in the other combinations of missingness percentage and correlation.

To gain a better sense of how the Horvitz–Thompson-type KDEs compare with each other and with the full-data KDE, it is useful to reproduce Figure 1 without CC-KDE. Figure 3 presents boxplots for the full-data KDE and the six Horvitz–Thompson-type KDEs. From this plot, it can be seen that there was no significant difference between the median ISE for the four HT-type KDEs with estimated propensity scores. However, the median ISE of the full-data KDE was somewhat smaller than the median ISEs of all the HT-type KDEs, regardless of distribution, for 30% missingness and $\rho = 0.8$. Moreover, the variability of the ISEs of the full-data KDE was somewhat smaller. With some other combinations of percent missingness, ρ and response variable distribution, there was no significant difference between the median ISE of the full-data KDE and those of the HT-type KDEs with estimated propensity scores.

It is interesting to note from Figure 3 that the median ISEs for the HT-type KDEs with true propensity scores were actually larger than their counterparts with estimated propensity scores. The phenomenon is verified by comparing asymptotic variances in Section 4. Intuitively, the HT-type KDEs with true propensity scores only use the completely observed subjects to estimate the unknown density. However, the HT-type KDEs with estimated propensity scores incorporate all of the observed data through the estimation of the propensity scores.

It is also interesting to note that, for a given KDE and bandwidth procedure, the distributions of ISEs for $\rho = 0.5$ and for $\rho = 0.8$ were approximately the same with all else being equal. However, for a given KDE and bandwidth procedure, the ISEs were slightly larger and more variable for 30% missingness than for 17% missingness with all else being equal.

Finally, it is also important that a density estimator accurately represents the shape of the true density. To examine this, we created plots of the KDEs with the true densities. In all cases, the HT-type KDEs did accurately represent the shape of the true density. This is illustrated for the case of 30% missingness and $\rho = 0.8$ in Figure 4, which presents the estimated densities, using the modified HT-type KDE with Nadaraya–Watson estimates of the propensity scores, of the fits with median ISE, along with the true density, for each of the three response variable distributions. The three estimated densities in each plot represent the modified HT-type KDE using the three different bandwidth procedures: Norm, CV and SJ.

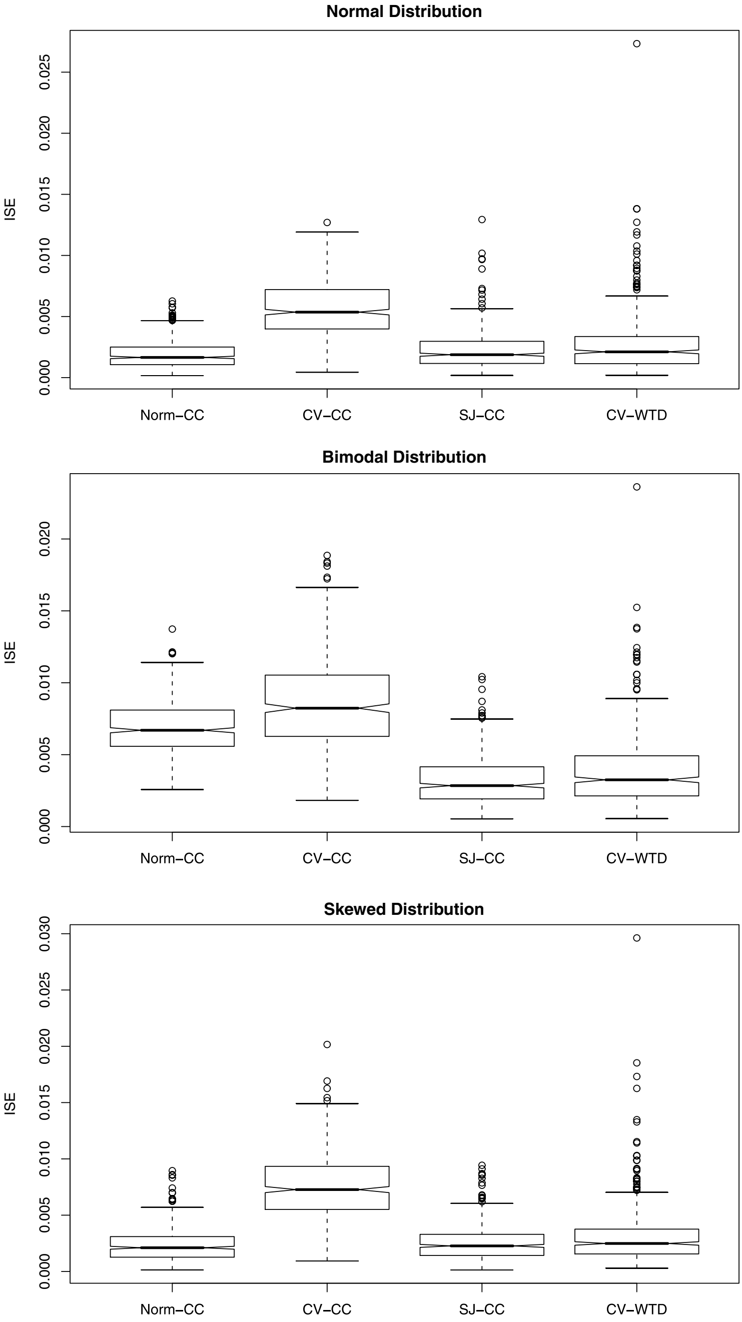

Figure 2. Boxplots of observed integrated squared errors of modified Horvitz–Thompson-type kernel density estimators, with Nadaraya–Watson propensity score estimates, comparing bandwidth procedures for (a) normal, (b) bimodal and (c) skewed distributions with 30% missing data and correlation between the response and auxiliary variable $\rho = 0.8$. Based on 501 simulated datasets of size 300. Bandwidth selection procedures: normal optimal smoothing (Norm), least squares cross-validation (CV), Sheather–Jones plug-in (SJ) based on the complete cases, and weighted least squares cross-validation (CV-WTD).

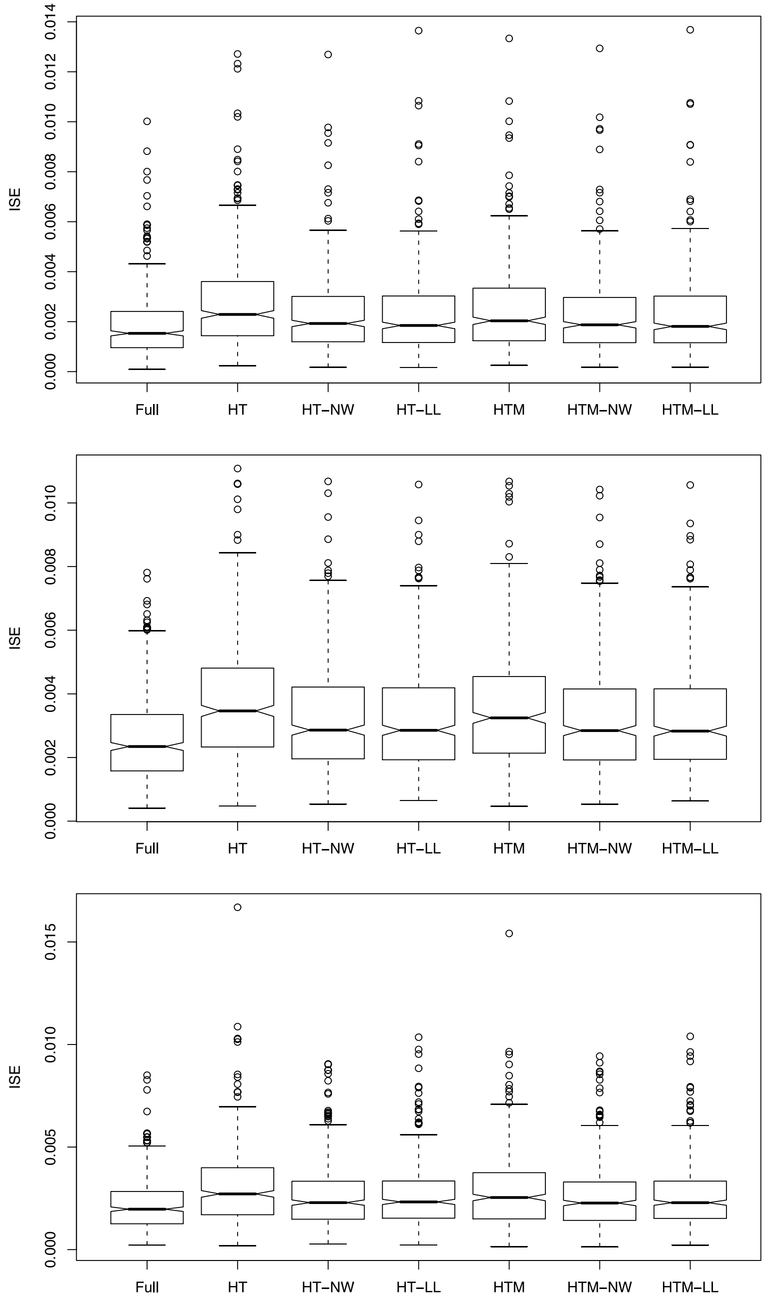

Figure 3. Boxplots of observed integrated squared errors of kernel density estimators with Sheather–Jones bandwidth for (a) normal, (b) bimodal and (c) skewed distributions with 30% missing data and correlation between the response and auxiliary variable of $\rho = 0.8$. Based on 501 simulated data sets of size 300. Kernel density estimators: full data (Full); Horvitz–Thompson with true propensity scores (HT), with Nadaraya–Watson estimates (HT-NW), and with local likelihood estimates (HT-LL) of the propensity scores; modified Horvitz–Thompson with true propensity scores (HTM), with Nadaraya–Watson estimates (HTM-NW), and with local likelihood estimates (HTM-LL) of the propensity scores.

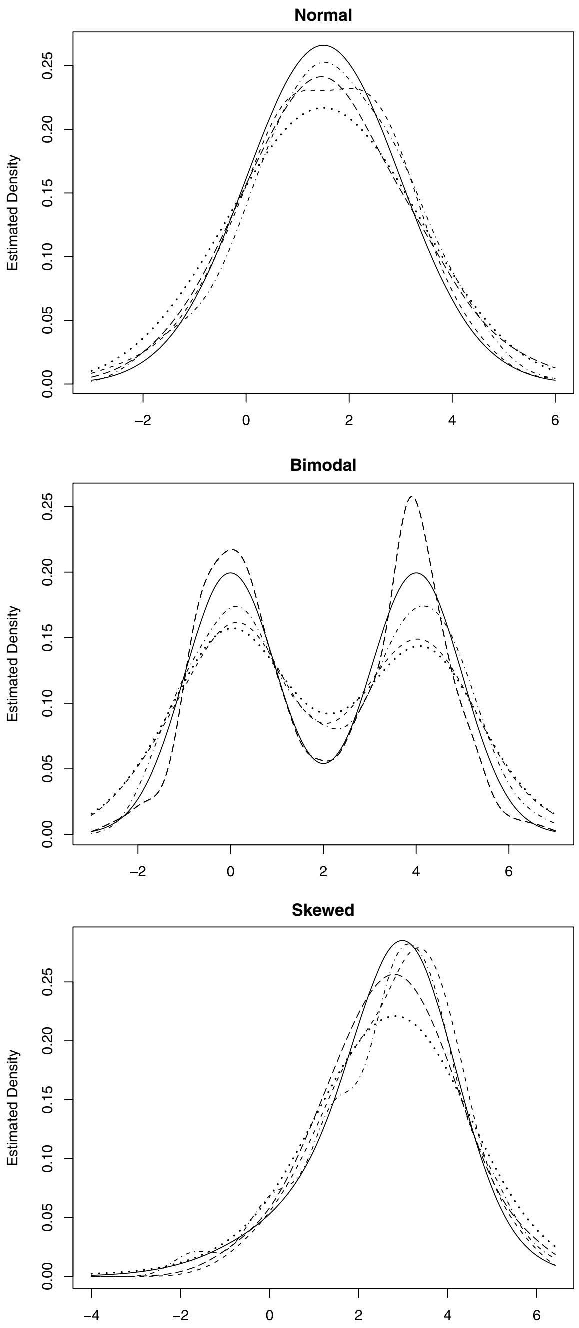

Figure 4. Plot of modified Horvitz–Thompson-type kernel density estimator with normal optimal smoothing (short-dashed line), least squares cross-validation (dotted line), Sheather–Jones plug-in (dot–dashed line), and weighted cross-validation (long-dashed line) bandwidth selectors having median integrated squared error, and true density (solid line) for (a) normal, (b) bimodal and (c) skewed distributions with 30% missing data and correlation between the response and auxiliary variable of $\rho = 0.8$. Based on 501 simulated datasets of size 300.

3.3. Example

To illustrate the proposed density estimators, we return to the AIDS Clinical Trials Group (ACTG) protocol 175 data. In ACTG 175, CD4 counts were recorded for patients at baseline, 20 ± 5 weeks, and final follow-up at 96 ± 5 weeks. One important variable is the difference in CD4 counts from baseline to final follow-up. Ultimately, researchers were interested in determining if one treatment was superior in terms of increase in CD4 counts. In addition, under proper randomization, the distribution of baseline CD4 counts should not differ based on treatment group. Therefore, one might simply consider comparing the four treatment CD4 counts at final follow-up. Here we will consider the change in CD4 counts from baseline to final follow-up and CD4 counts at final follow-up across treatment groups. The author is currently developing methods for comparing densities for the different treatment groups where data are incomplete.

In addition to CD4 counts, several auxiliary variables were collected at the beginning of and throughout the study. At baseline, the following variables were recorded for each participant: baseline CD4 count, weight, age, indicators of intravenous drug use, HIV symptoms, prior experience with antiretroviral therapy, hemophilia, sexual preference, gender and race, CD8 count (another measure of immune status), and Karnofsky score (an index that reflects a patient's ability to perform ordinary daily activities). After baseline but before final follow-up, intermediate measures of immune status were also collected, specifically CD4 and CD8 counts at 20 ± 5 weeks. Furthermore, patients were randomized to four antiretroviral regimens [zidovudine (ZDV), ZDV + didanosine (ddI), ZDV + zalcitabine, ddI] in equal proportions.

In the ACTG 175 data, all of the auxiliary variables were available for all 2139 study participants, but 37% of the CD4 counts at 96 ± 5 weeks were missing. Missingness at follow-up is often associated with baseline response and intermediate measures of response, related auxiliary variables and intervention group. Owing to the large number of auxiliary variables and the computational intensity of the methods considered here, we considered the same subset of auxiliary variables as Davidian *et al.* (2005), namely, weight, indicators of HIV symptoms and prior antiretroviral therapy, Karnofsky score, CD8 count and CD4 count at baseline and at 20 ± 5 weeks, and off-treatment status in addition to treatment group in estimating the propensity scores.

As all of the Horvitz–Thompson-type KDEs performed similarly in the simulation study, we will focus on the modified Horvitz–Thompson-type estimate $\hat{g}_h^{\mathrm{HTM}}(y)$ (2) with Nadaraya–Watson estimates of the propensity scores $\pi_{\mathrm{NW}i}$ (4) and Sheather–Jones bandwidth for illustrative purposes. All computations were carried out in R (R Development Core Team 2006).

The propensity scores were estimated using the contributed R package `np` (Hayfield & Racine 2008), as it easily allows for the incorporation of continuous and categorical (nominal and ordinal) variables. The functions `npregbw` and `npreg` were used to obtain bandwidths and the propensity score estimates, respectively. Once the propensity score estimates were obtained, the modified Horvitz–Thompson-type density estimate was computed using R functions written by the author. Sheather–Jones bandwidths were computed on the complete cases using the function `h.select` of the contributed R package `sm` (Bowman & Azzalini 2005).

First, we consider the difference in CD4 counts from baseline to final follow-up. The Horvitz–Thompson-type density estimate is depicted in Figure 5, along with the complete-case

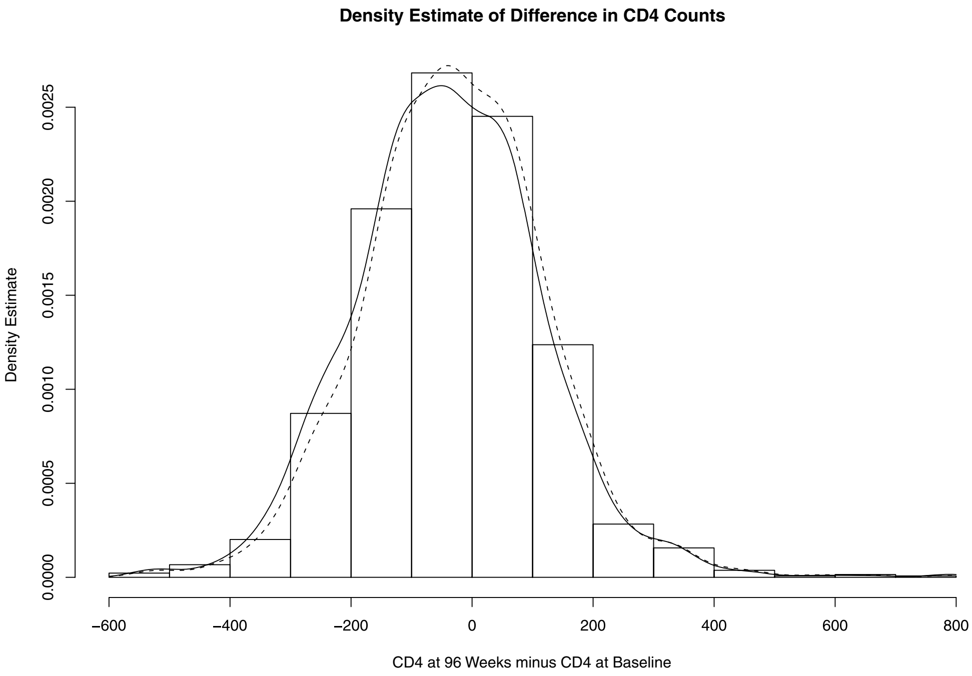

Figure 5. Modified Horvitz–Thompson-type kernel density estimate $\hat{g}_h^{\mathrm{HTM}}(y)$ of the difference in CD4 counts from baseline to final follow-up (solid line) with the complete-case kernel density estimate (dashed line) overlayed on the histogram of observed CD4 counts at 96 ± 5 weeks.

density estimate and a histogram of the observed differences. The two estimates are quite similar, and closely follow the shape of the histogram, although the complete-case density estimate is slightly more peaked. This may lead us to believe that the distribution of differences in CD4 counts from baseline to final follow-up for patients with missing CD4 counts at final follow-up is similar to that of those with completely observed data. In addition, the complete-case density estimate is shifted slightly to the right of the Horvitz–Thompson-type density estimate. Although not a large difference, this might indicate that the differences in CD4 counts for missing patients were somewhat less than those for the observed patients. As the patients with unobserved CD4 counts were probably the ones who dropped out of the study because of the advanced stage of their illness, they would probably have experienced less success with any of the treatments. The Horvitz–Thompson-type density estimate seems to account for this.

Next, we turn our attention to the CD4 counts at final follow-up. As the observed CD4 counts at 96 ± 5 weeks were positive with a fair amount of mass near zero, and ordinary kernel density estimates are notoriously subject to boundary bias, it was necessary to adjust our density estimates. Several approaches can be employed to adjust for boundary bias, including a transformation technique and the use of boundary kernels. A simple approach that is quite effective is the reflection technique. Applying the reflection technique (Silverman 1986) to the modified Horvitz–Thompson density estimator results in the following density

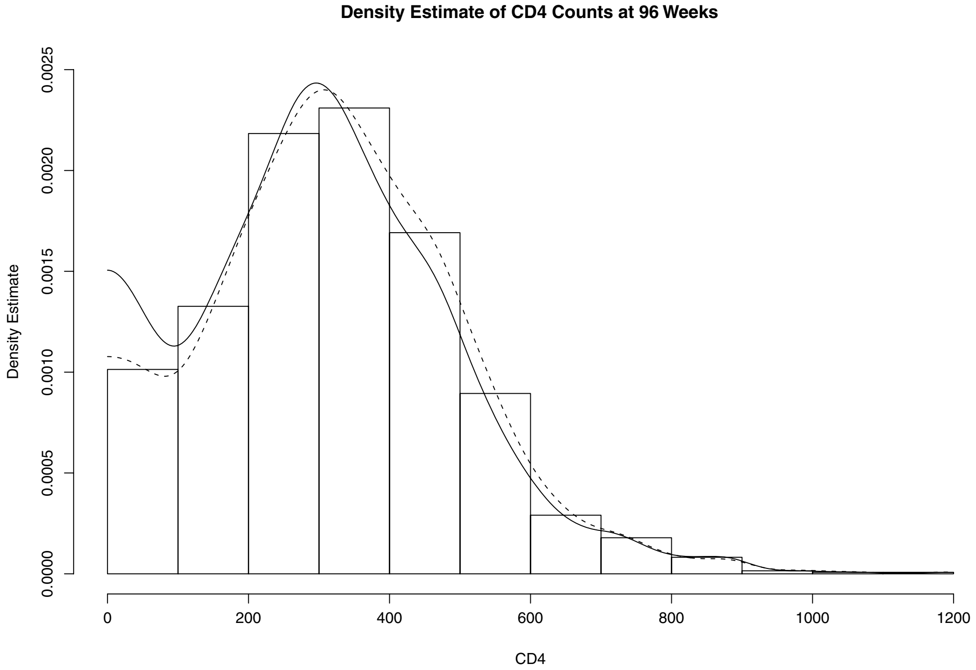

Figure 6. Modified Horvitz–Thompson-type kernel density estimate $\hat{g}_h^{\mathrm{HTM}}(y)$ of CD4 counts at 96 ± 5 weeks (solid line) with complete-case kernel density estimate (dashed line) overlayed on the histogram of observed CD4 counts at 96 ± 5 weeks.

estimator:

$$\hat{g}_h^{\mathrm{HTM-R}}(y) = \begin{cases} \hat{g}_h^{\mathrm{HTM}}(y) + \hat{g}_h^{\mathrm{HTM}}(-y), & y \geqslant 0 \\ 0, & y < 0. \end{cases}$$

The proposed density estimator $\hat{g}_h^{\mathrm{HTM-R}}(y)$ is depicted in Figure 6 along with the complete-case density estimator, also using the reflection technique, and a histogram of the observed CD4 counts at 96 ± 5 weeks. Notice that the complete-case density estimate closely outlines the histogram of the observed CD4 counts, as one would expect. The proposed density estimator, however, puts more mass on the smaller CD4 counts, reflecting the fact that the patients with unobserved CD4 counts at 96 ± 5 weeks (37% of the 2139) were more likely to be those patients whose illness had progressed and who therefore had lower CD4 counts.

4. Theoretical details

In this section, we provide some theoretical results for $\hat{g}_h^{\mathrm{HTt}}(y)$ and $\hat{g}_h^{\mathrm{HT}}(y)$ with Nadaraya–Watson propensity score estimates. In particular, we provide bias and variance results for each and show how the two are related.

We first consider the bias and variance of $\hat{g}_h^{\mathrm{HTt}}(y)$ from (1). In Appendix I, it is shown that the expectation of this Horvitz–Thompson-type KDE is the same as that of the full-data KDE: $\mathrm{E}\{\hat{g}_h^{\mathrm{HTt}}(y)\} = \mathrm{E}\{\hat{g}_h^{\mathrm{F}}(y)\}$. Thus, the Horvitz–Thompson-type KDE, using the true propensity scores, has the same bias as the full-data KDE. Furthermore, it is shown that the variance of this Horvitz–Thompson-type KDE, using true propensity scores, is larger than that of the full-data KDE:

$$\mathrm{var}\{\hat{g}_h^{\mathrm{HTt}}(y)\} = \mathrm{var}\{\hat{g}_h^{\mathrm{F}}(y)\} + \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{E}_{\boldsymbol{x}_i}\left[\left(\frac{1-\pi_i}{\pi_i}\right)\mathrm{E}_{y_i|\boldsymbol{x}_i}\{K_h^2(y-y_i)\}\right]. \tag{6}$$

This implies that the MISE of HT-KDE is greater than that of the full-data KDE. This increase in MISE is to be expected because the estimator is essentially based on a smaller sample size. See Appendix I for details.

The MSE of $\hat{g}_h^{\mathrm{HTt}}(y)$ is given by

$$\mathrm{MSE}\{\hat{g}_h^{\mathrm{HTt}}(y)\} = \mathrm{MSE}\{\hat{g}_h^{\mathrm{F}}(y)\} + \frac{1}{n}\mathrm{E}\left[\left\{\frac{1-\pi(\boldsymbol{x}_1)}{\pi(\boldsymbol{x}_1)}\right\}K_h^2(y-y_1)\right],$$

and the MISE is $\int \mathrm{MSE}\{\hat{g}_h^{\mathrm{HTt}}(y)\}$, which does not have the compact representation that the MISE of $\hat{g}_h^{\mathrm{F}}(y)$ does. A large-sample representation of the MSE can be shown to be

$$\mathrm{MSE}\{\hat{g}_h^{\mathrm{HTt}}(y)\} = \frac{1}{nh}R(K)g(y)\mathrm{E}\left\{\frac{1}{\pi(\boldsymbol{x}_1)}\middle| y\right\} + \frac{1}{4}h^4\mu_2(K)^2 g''(y)^2 + o\{(nh)^{-1} + h^4\},$$

provided that the expectation exists, where $R(K) = \int K^2(y)\,dy$ and $\mu_2(K) = \int y^2 K(y)\,dy$. Thus, unlike the asymptotic MISE (AMISE) of $\hat{g}_h^{\mathrm{F}}(y)$, the AMISE of $\hat{g}_h^{\mathrm{HTt}}(y)$ depends on the unknown density g directly:

$$\mathrm{AMISE}\{\hat{g}_h^{\mathrm{HTt}}(y)\} = \frac{1}{nh}R(K)\int g(y)\mathrm{E}\left\{\frac{1}{\pi(\boldsymbol{x}_1)}\middle| y\right\}dy + \frac{1}{4}h^4\mu_2(K)^2 R(g'').$$

(See Appendix I for details.) Furthermore, the AMISE depends on the propensity scores through the conditional expectation E $\{1/\pi(\boldsymbol{x}_1)\,|\,y\}$. Given the complexity of AMISE, we are not able to easily compute estimates of the asymptotically optimal bandwidth.

In Appendix II, it is shown that $\hat{g}_h^{\mathrm{HTt}}(y)$ and $\hat{g}_h^{\mathrm{HT}}(y)$ with Nadaraya–Watson propensity score estimates are related through the following expression:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\hat{\pi}_i}K_h(y-y_i) = \frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\pi_i}K_h(y-y_i) + \frac{1}{n}\sum_{i=1}^{n}\left(1-\frac{r_i}{\pi_i}\right)\mathrm{E}\{K_h(y-y_i)\,|\,\boldsymbol{x}_i\} + o_p(1). \tag{7}$$

The primary implications of this relationship are that $\hat{g}_h^{\mathrm{HT}}(y)$ has the same bias as $\hat{g}_h^{\mathrm{F}}(y)$ and $\hat{g}_h^{\mathrm{HTt}}(y)$ and that $\mathrm{var}\{\hat{g}_h^{\mathrm{HT}}(y)\}$ is given by

$$\mathrm{var}\{\hat{g}_h^{\mathrm{HT}}(y)\} = \mathrm{var}\{\hat{g}_h^{\mathrm{HTt}}(y)\} - \mathrm{var}\left[\frac{1}{n}\sum_{i=1}^{n}\left(1-\frac{r_i}{\pi_i}\right)\mathrm{E}\{K_h(y-y_i)\,|\,\boldsymbol{x}_i\}\right], \tag{8}$$

asymptotically. The variance follows from the fact that the covariance of the two terms of (7) is the negative of the variance of the second term. Therefore, the variance of HT-KDE with estimated propensity scores is less than that of HT-KDE with true propensity scores. This, in turn, implies that the MISE of HT-KDE with estimated propensity scores is less than that of HT-KDE with true propensity scores, which is confirmed in the simulation study.

The fact that the variance of the Horvitz–Thompson-type kernel density estimate is smaller when estimated propensity scores are used than when true propensity scores are used may seem counterintuitive. One should recall that only the complete cases are used in estimating the density when true propensity scores are used. When the propensity scores are estimated, however, all of the observed data are used in the estimation of the density. It seems that the incorporation of the auxiliary data for the incomplete cases allows us to improve the efficiency with which we estimate the unknown density.

5. Discussion

In this paper, we have proposed a method for estimating the density of a response variable, which may be missing at random, when auxiliary data are available. The KDE is based on the Horvitz–Thompson estimator and assumes that the response variable is missing at random. Simulation studies showed that the proposed density estimator performs much better, in terms of ISE, than the complete-case density estimator. In addition, the proposed density estimator performs almost as well as the density estimator that utilizes the full data set, as if none of the values were missing.

Based on the simulation study, we prefer the modified Horvitz–Thompson-type KDE $\hat{g}_h^{\mathrm{HTM}}(y)$ (3) with Nadaraya–Watson estimates of the propensity scores $\hat{\pi}_{\mathrm{NW}i}$ (4) over the other Horvitz–Thompson-type density estimators considered, because it is indeed a density function and the Nadaraya–Watson probability estimates are somewhat easier to obtain than the local likelihood estimates. Furthermore, the Sheather–Jones bandwidth selection procedure based on the observed responses performs the best over a wide variety of true densities.

In practice, the modified Horvitz–Thompson-type KDE has also been shown to be effective. In particular, the modified Horvitz–Thompson-type KDE appropriately adjusted the complete-case KDE to reflect the fact that HIV patients with low CD4 counts are more likely to miss visits late in the study owing to a decline in their health. This also suggests that, if auxiliary variables that are highly correlated with the response are available, this modified Horvitz–Thompson-type KDE may be appropriate when the response variable is not missing at random (non-ignorably missing). This is currently under investigation.

Another issue that requires further investigation is bandwidth selection for the Nadaraya–Watson estimator of propensity scores. In the simulation study, the `sm.regression` function of the `sm` package (Bowman & Azzalini 2005) was used to calculate the Nadaraya–Watson estimates of the propensity scores. The default method selects a bandwidth based on the approximate degrees of freedom, which has a default value of 6 for the case of one auxiliary variable (covariate). Other methods available in the `sm` package are cross-validation and an AIC-based method proposed by Hurvich, Simonoff & Tsai (1998). A very small simulation study did not indicate any differences among the three bandwidth selectors in terms of their effect on the Horvitz–Thompson-type KDEs. However, a more thorough investigation is planned.

Future research includes developing methods to test for the equality of densities, and methods for estimating the conditional density of the response variable given the auxiliary variables when responses are not completely observed. The test for the equality of densities involves comparing the Horvitz–Thompson-type KDEs for the different groups with an overall Horvitz–Thompson-type KDE for the combined data. This test will be useful in determining if the distribution of CD4 counts differs for the four treatment groups. In addition, the estimation of the conditional density of y given $\boldsymbol{x}$ when y may be missing at random is a relatively straightforward extension of the methods proposed in this paper. Furthermore, Hyndman, Bashtannyk & Grunwald (1996) have proposed an improved KDE for the conditional distribution of y given $\boldsymbol{x}$ that requires estimation of the conditional mean of y given $\boldsymbol{x}$. The methods of this paper, along with those proposed by Kennedy (2007), can be used to extend the improved conditional density estimator of Hyndman *et al.* (1996) when the response variable is missing at random. Such an estimator and the related issue of bandwidth selection are topics of future research.

Appendix I: Results for HT-KDE with true propensity scores

The expectation of the Horvitz–Thompson-type KDE$\hat{g}_h^{\mathrm{HTt}}(y)$ in (1) is

$$\begin{aligned}
\mathrm{E}\{\hat{g}_h^{\mathrm{HTt}}(y)\} &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left\{\frac{r_i}{\pi_i}K_h(y-y_i)\right\}\\
&= \mathrm{E}_{\boldsymbol{x}_i}\left[\mathrm{E}_{y_i|\boldsymbol{x}_i}\left\{K_h(y-y_i)\mathrm{E}_{r_i|y_i,\boldsymbol{x}_i}\left(\frac{r_i}{\pi_i}\right)\right\}\right]\\
&= \mathrm{E}\{K_h(y-y_i)\} = \mathrm{E}\{\hat{g}_h^{\mathrm{F}}(y)\}.
\end{aligned}$$

The third equality holds because Pr $(r_i = 1 \mid y_i, \boldsymbol{x}_i) = $ Pr $(r_i \mid \boldsymbol{x}_i) = \pi_i$ owing to the missing-at-random assumption.

The variance of $\hat{g}_h^{\mathrm{HTt}}(y)$ can be found by first noting that

$$\begin{aligned}
\mathrm{var}\left\{\frac{r_i}{\pi_i}K_h(y-y_i)\right\} &= \mathrm{E}_{\boldsymbol{x}_i}\left[\mathrm{var}_{r_i,y_i|\boldsymbol{x}_i}\left\{\frac{r_i}{\pi_i}K_h(y-y_i)\right\}\right]\\
&\quad + \mathrm{var}_{\boldsymbol{x}_i}\left[\mathrm{E}_{r_i,y_i|\boldsymbol{x}_i}\left\{\frac{r_i}{\pi_i}K_h(y-y_i)\right\}\right],
\end{aligned}$$

where

$$\begin{aligned}
\mathrm{E}_{r_i,y_i|\boldsymbol{x}_i}\left\{\frac{r_i}{\pi_i}K_h(y-y_i)\right\} &= \mathrm{E}_{y_i|\boldsymbol{x}_i}\left[\mathrm{E}_{r_i|y_i,\boldsymbol{x}_i}\left\{\frac{r_i}{\pi_i}K_h(y-y_i)\right\}\right]\\
&= \mathrm{E}_{y_i|\boldsymbol{x}_i}\{K_h(y-y_i)\},
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{var}_{r_i,y_i|\boldsymbol{x}_i}\left\{\frac{r_i}{\pi_i}K_h(y-y_i)\right\} &= \mathrm{E}_{y_i|\boldsymbol{x}_i}\left[\mathrm{var}_{r_i|y_i,\boldsymbol{x}_i}\left\{\frac{r_i}{\pi_i}K_h(y-y_i)\right\}\right]\\
&\quad + \mathrm{var}_{y_i|\boldsymbol{x}_i}\left[\mathrm{E}_{r_i|y_i,\boldsymbol{x}_i}\left\{\frac{r_i}{\pi_i}K_h(y-y_i)\right\}\right]\\
&= \mathrm{E}_{y_i|\boldsymbol{x}_i}\left\{\frac{\pi_i(1-\pi_i)}{\pi_i^2}K_h^2(y-y_i)\right\}\\
&\quad + \mathrm{var}_{y_i|\boldsymbol{x}_i}\{K_h(y-y_i)\}\\
&= \frac{1-\pi_i}{\pi_i}\mathrm{E}_{y_i|\boldsymbol{x}_i}\{K_h^2(y-y_i)\}\\
&\quad + \mathrm{var}_{y_i|\boldsymbol{x}_i}\{K_h(y-y_i)\}.
\end{aligned}$$

Simplifying, we have

$$\begin{aligned}\operatorname{var}\left\{\frac{r_i}{\pi_i}K_h(y-y_i)\right\} &= \mathrm{E}_{\boldsymbol{x}_i}\left[\left(\frac{1-\pi_i}{\pi_i}\right)\mathrm{E}_{y_i|\boldsymbol{x}_i}\left\{K_h^2(y-y_i)\right\}\right]\\ &\quad+\mathrm{E}_{\boldsymbol{x}_i}[\operatorname{var}_{y_i|\boldsymbol{x}_i}\{K_h(y-y_i)\}]+\operatorname{var}_{\boldsymbol{x}_i}[\mathrm{E}_{y_i|\boldsymbol{x}_i}\{K_h(y-y_i)\}]\\ &= \mathrm{E}_{\boldsymbol{x}_i}\left[\left(\frac{1-\pi_i}{\pi_i}\right)\mathrm{E}_{y_i|\boldsymbol{x}_i}\left\{K_h^2(y-y_i)\right\}\right]+\operatorname{var}\{K_h(y-y_i)\}.\end{aligned}$$

Finally, the variance of (1) is

$$\operatorname{var}\left\{\hat{g}_h^{\mathrm{HTt}}(y)\right\}=\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}_{\boldsymbol{x}_i}\left[\left(\frac{1-\pi_i}{\pi_i}\right)\mathrm{E}_{y_i|\boldsymbol{x}_i}\left\{K_h^2(y-y_i)\right\}\right]+\frac{1}{n^2}\sum_{i=1}^{n}\operatorname{var}\{K_h(y-y_i)\},$$

which is the result given in (6).

It then follows that

$$\operatorname{MSE}\left\{\hat{g}_h^{\mathrm{HTt}}(y)\right\}=\operatorname{MSE}\left\{\hat{g}_h^{\mathrm{F}}(y)\right\}+\frac{1}{n}\mathrm{E}\left[\left\{\frac{1-\pi(\boldsymbol{x}_1)}{\pi(\boldsymbol{x}_1)}\right\}K_h^2(y-y_1)\right] \tag{9}$$

$$\begin{aligned}&=\frac{1}{n}\mathrm{E}\left\{\frac{1}{\pi(\boldsymbol{x}_1)}K_h^2(y-y_1)\right\}-\frac{1}{n}[\mathrm{E}\{K_h(y-y_1)\}]^2\\ &\quad+[\mathrm{E}\{K_h(y-y_1)\}-g(y)]^2.\end{aligned} \tag{10}$$

Thus, the only difference between the MSEs of $\hat{g}_h^{\mathrm{HTt}}(y)$ and $\hat{g}_h^{\mathrm{F}}(y)$ results from the differences in the second term of (9).

Under assumptions (a1) and (b1)–(b3) below, we can derive a large-sample approximation of $\operatorname{MSE}\{\hat{g}_h^{\mathrm{HTt}}(y)\}$. We need only derive this large-sample approximation of the first term of (10), as approximations of the other terms are the same as in the full-data case. Let $g(\boldsymbol{x}_1, y_1)=g(y_1\,|\,\boldsymbol{x}_1)f(\boldsymbol{x}_1)$ denote the joint pdf of $\boldsymbol{x}_1$ and y_1. Then

$$\begin{aligned}\frac{1}{n}\mathrm{E}\left\{\frac{1}{\pi(\boldsymbol{x}_1)}K_h^2(y-y_1)\right\} &= \frac{1}{n}\int\!\!\int\frac{1}{\pi(\boldsymbol{x}_1)}K_h^2(y-y_1)g(\boldsymbol{x}_1,y_1)\,dy_1\,d\boldsymbol{x}_1\\ &= \frac{1}{nh}\int\!\!\int\frac{1}{\pi(\boldsymbol{x}_1)}K^2(z)g(y-zh|\boldsymbol{x}_1)f(\boldsymbol{x}_1)\,dz\,d\boldsymbol{x}_1\\ &= \frac{1}{nh}\int\!\!\int\frac{1}{\pi(\boldsymbol{x}_1)}K^2(z)f(\boldsymbol{x}_1)\{g(y\,|\,\boldsymbol{x}_1)+o(1)\}\,dz\,d\boldsymbol{x}_1\\ &= \frac{1}{nh}R(K)\int\frac{1}{\pi(\boldsymbol{x}_1)}g(\boldsymbol{x}_1,y)\,d\boldsymbol{x}_1+o((nh)^{-1})\\ &= \frac{1}{nh}R(K)\int\frac{1}{\pi(\boldsymbol{x}_1)}g(y)f(\boldsymbol{x}_1\,|\,y)\,d\boldsymbol{x}_1+o((nh)^{-1})\\ &= \frac{1}{nh}R(K)\mathrm{E}\left\{\frac{1}{\pi(\boldsymbol{x}_1)}\middle|\,y\right\}g(y)+o((nh)^{-1}).\end{aligned}$$

Note that the second equality follows from a change of variable and the third equality follows from a Taylor series expansion. The desired result then follows.

Appendix II: Results for HT-KDE with estimated propensity scores

The derivation of the relationship between $\hat{g}_h^{\mathrm{HTt}}(y)$ and $\hat{g}_h^{\mathrm{HT}}(y)$ in (7) follows many of the steps in the proof of theorem 2 of Qi *et al.* (2005). Note that $\hat{\pi}_i = \hat{\pi}(\boldsymbol{x}_i)$ can be calculated from continuous and/or discrete random variables using the Nadaraya–Watson estimator or the local likelihood estimator. We provide the proof in the case of continuous auxiliary variables using the Nadaraya–Watson estimator. Let d denote the number of continuous auxiliary variables, and let K^* be a sth-order kernel. The following regularity conditions regarding the propensity scores and Nadaraya–Watson estimator are needed:

(a1) The selection probability $\pi(\boldsymbol{x}) \geqslant \epsilon$, where $\epsilon > 0$.
(a2) The selection probability $\pi(\boldsymbol{x})$ has s continuous and bounded partial derivatives with respect to the continuous components of $\boldsymbol{x}$ a.e.
(a3) $nh_1^{2d} \to \infty$ and $nh_1^{2s} \to 0$ as $n \to \infty$.
(a4) The probability density function $f(\boldsymbol{x})$ of $\boldsymbol{x}$ and the conditional probability density function $f(\boldsymbol{x} \mid r)$ of $\boldsymbol{x} \mid r$ are bounded away from 0 and have r continuous and bounded partial derivatives with respect to the continuous components of $\boldsymbol{x}$ a.e.
(a5) The sth-order kernel K^* is a piecewise-smooth function of $\boldsymbol{x}$ satisfying $\int K^*(u)\,du = 1$, $\int u^m K^*(u)\,du = 0$ for $m = 0, 1, \ldots, (s-1)$, $\int u^s K^*(u)\,du \neq 0$, $\int K^{*2}(u)\,du < \infty$.

In addition, the following regularity conditions are needed regarding the KDE and the response distribution:

(b1) The kernel function K is symmetric about 0 with $\int K^2(y)\,dy = R(K) < \infty$, $\int y^2 K(y)\,dy = \mu_2(K) < \infty$, and $\int |y|^3 K(y)\,dy < \infty$.
(b2) $h \to 0$ and $nh \to \infty$ as $n \to \infty$.
(b3) The pdf $g(y)$ of y and the conditional pdf $g(y \mid \boldsymbol{x})$ are such that the second derivatives of each are continuous, square integrable and ultimately monotone.

We start by noting that (2) can be written as

$$\frac{1}{n}\sum_{i=1}^{n} \frac{r_i}{\hat{\pi}_i} K_h(y - y_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{r_i}{\pi_i} K_h(y - y_i) - \frac{1}{n}\sum_{i=1}^{n} r_i \left(\frac{\hat{\pi}_i - \pi_i}{\pi_i^2}\right) K_h(y - y_i) + o_p(1)$$

using a Talyor series expansion of $1/\hat{\pi}_i$ about $1/\pi_i$. Thus, it is necessary to show that

$$\frac{1}{n}\sum_{i=1}^{n} r_i \left(\frac{\hat{\pi}_i - \pi_i}{\pi_i^2}\right) K_h(y - y_i) = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{r_i - \pi_i}{\pi_i}\right) \mathrm{E}\{K_h(y - y_i) \mid \boldsymbol{x}_i\} + o_p(1), \quad (11)$$

which is the negative of the second term of the right-hand side of (7).

Now, the left-hand side of (11) can be written

$$\frac{1}{n}\sum_{i=1}^{n} \left(\frac{r_i - \pi_i}{\pi_i^2}\right) (\hat{\pi}_i - \pi_i) K_h(y - y_i) + \frac{1}{n}\sum_{i=1}^{n} \frac{\pi_i}{\pi_i^2} (\hat{\pi}_i - \pi_i) K_h(y - y_i). \quad (12)$$

Denote the two terms of (12) by D_{1n} and D_{2n}, respectively. We must show that $D_{1n} \xrightarrow{p} 0$ and that D_{2n} equals the right-hand side of (11).

Step 1. Show $D_{1n}\xrightarrow{p}0$. Let $\hat{f}(\boldsymbol{x}_i)=(nh_1^d)^{-1}\sum_{j=1}^n K_{h_1}^*(\boldsymbol{x}_i-\boldsymbol{x}_j)$. Then

$$\begin{aligned}D_{1n} &= \frac{1}{n}\sum_{i=1}^n \frac{(r_i-\pi_i)}{\pi_i^2}\left\{\frac{(nh_1^d)^{-1}\sum_{j=1}^n r_j K_{h_1}^*(\boldsymbol{x}_i-\boldsymbol{x}_j)}{\hat{f}(\boldsymbol{x}_i)}-\pi_i\right\}K_h(y-y_i)\\ &= \frac{1}{n^2}\sum_{j=1}^n\sum_{i=1}^n \frac{(r_i-\pi_i)(r_j-\pi_i)K_{h_1}^*(\boldsymbol{x}_i-\boldsymbol{x}_j)K_h(y-y_i)}{\pi_i^2 h_1^d\hat{f}(\boldsymbol{x}_i)}\\ &= S_{1n}-S_{2n}+o_p(1),\end{aligned}$$

where the last equality follows from the Taylor series expansion of $1/\hat{f}(\boldsymbol{x}_i)$ about $1/f(\boldsymbol{x}_i)$,

$$S_{1n}=\frac{1}{n^2}\sum_{j=1}^n\sum_{i=1}^n \frac{(r_i-\pi_i)(r_j-\pi_i)K_{h_1}^*(\boldsymbol{x}_i-\boldsymbol{x}_j)K_h(y-y_i)}{\pi_i^2 h_1^d f(\boldsymbol{x}_i)}$$

and

$$S_{2n}=\frac{1}{n^2}\sum_{j=1}^n\sum_{i=1}^n \frac{(r_i-\pi_i)(r_j-\pi_i)K_{h_1}^* K_h(y-y_i)\{\hat{f}(\boldsymbol{x}_i)-f(\boldsymbol{x}_i)\}}{\pi_i^2 h_1^d f^2(\boldsymbol{x}_i)}.$$

Through tedious calculations, we can show that $\mathrm{E}(S_{1n})\to 0$ and var $(S_{1n})\to 0$ as $n\to\infty$, so that $S_{1n}\xrightarrow{p}0$. Similar calculations give E $(S_{2n})\to 0$ and var $(S_{2n})\to 0$ as $n\to\infty$, so that $S_{2n}\xrightarrow{p}0$ also. Therefore, $D_{1n}\xrightarrow{p}0$.

Step 2. Show

$$D_{2n}=\frac{1}{n}\sum_{i=1}^n\left(\frac{r_i-\pi_i}{\pi_i}\right)\mathrm{E}\{K_h(y-y_i)\,|\,\boldsymbol{x}_i\}+o_p(1). \tag{13}$$

Rewriting D_{2n} and using the Taylor series expansion of $1/\hat{f}(\boldsymbol{x}_i)$ about $1/f(\boldsymbol{x}_i)$, we obtain

$$\begin{aligned}D_{2n} &= \frac{1}{n}\sum_{i=1}^n\frac{1}{\pi_i}\left\{\frac{(nh_1^d)^{-1}\sum_{j=1}^n K_{h_1}^*(\boldsymbol{x}_i-\boldsymbol{x}_j)}{\hat{f}(\boldsymbol{x}_i)}-\pi_i\right\}K_h(y-y_i)\\ &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n\frac{(r_j-\pi_i)K_{h_1}^*(\boldsymbol{x}_i-\boldsymbol{x}_j)}{\pi_i h_1^d\hat{f}(\boldsymbol{x}_i)}K_h(y-y_i)\\ &= \frac{1}{n^2}\sum_{j=1}^n\sum_{i=1}^n\frac{(r_j-\pi_i)K_{h_1}^*(\boldsymbol{x}_i-\boldsymbol{x}_j)}{\pi_i h_1^d f(\boldsymbol{x}_i)}K_h(y-y_i) \qquad (14)\end{aligned}$$

$$+\frac{1}{n^2}\sum_{j=1}^n\sum_{i=1}^n\frac{(r_j-\pi_i)K_{h_1}^*(\boldsymbol{x}_i-\boldsymbol{x}_j)\{\hat{f}(\boldsymbol{x}_i)-f(\boldsymbol{x}_i)\}}{\pi_i h_1^d f^2(\boldsymbol{x}_i)}K_h(y-y_i)+o_p(1). \tag{15}$$

Let the terms (14) and (15) be denoted by D_{2n1} and D_{2n2}, respectively, and let

$$D_{2n1}'=D_{2n1}-\frac{1}{n}\sum_{j=1}^n\left(\frac{r_j-\pi_j}{\pi_j}\right)\mathrm{E}\{K_h(y-y_i)\,|\,\boldsymbol{x}_i\}.$$

After tedious calculations, we can show that var $(D'_{2n1}) \to 0$ and var $(D_{2n2}) \to 0$ as $n \to \infty$, which imply that $D'_{2n1} \xrightarrow{p} 0$ and $D_{2n2} \xrightarrow{p} 0$, repectively. Thus, (13) is proved, and the desired result (7) follows.

We now need to derive the asymptotic variance of $\hat{g}_h^{\mathrm{HT}}(y)$ given in (8). Let A_n and B_n denote the first and second terms, respectively, on the right-hand side of (7). Then the variance of $\hat{g}_h^{\mathrm{HT}}(y)$ is $\mathrm{var}(A_n) + \mathrm{var}(B_n) + 2\,\mathrm{cov}(A_n, B_n)$. The variance of A_n was derived in Appendix 1. Let $q(\boldsymbol{x}_i) = \mathrm{E}\{K_h(y - y_i) \mid \boldsymbol{x}_i\}$. The variance of B_n is

$$\begin{aligned}
\mathrm{var}(B_n) &= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{E}_{\boldsymbol{x}_i}\left[\mathrm{var}_{r_i|\boldsymbol{x}_i}\left\{\left(1 - \frac{r_i}{\pi_i}\right) q(\boldsymbol{x}_i)\right\}\right] \\
&\quad + \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}_{\boldsymbol{x}_i}\left[\mathrm{E}_{r_i|\boldsymbol{x}_i}\left\{\left(1 - \frac{r_i}{\pi_i}\right) q(\boldsymbol{x}_i)\right\}\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{E}_{\boldsymbol{x}_i}\left\{\frac{q^2(\boldsymbol{x}_i)}{\pi_i^2}\mathrm{var}_{r_i|\boldsymbol{x}_i}(r_i)\right\} + 0 \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{E}_{\boldsymbol{x}_i}\left\{\left(\frac{1-\pi_i}{\pi_i}\right) q^2(\boldsymbol{x}_i)\right\}.
\end{aligned}$$

The covariance of A_n and B_n is

$$\mathrm{cov}(A_n, B_n) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{cov}\left(\frac{r_i}{\pi_i}K_h(y - y_i), \left(1 - \frac{r_i}{\pi_i}\right) q(\boldsymbol{x}_i)\right);$$

by independence, the other covariance terms are zero. Let $A'_i = (r_i/\pi_i)K_h(y - y_i)$ and $B'_i = (1 - r_i/\pi_i)q(\boldsymbol{x}_i)$. Then $\mathrm{cov}(A'_i, B'_i) = \mathrm{E}(A'_i B'_i) - \mathrm{E}(A'_i)\,\mathrm{E}(B'_i) = \mathrm{E}(A'_i B'_i)$ as $\mathrm{E}(B'_i) = 0$. Then

$$\begin{aligned}
\mathrm{E}(A'_i B'_i) &= E_{\boldsymbol{x}_i}\left(\mathrm{E}_{y_i|\boldsymbol{x}_i}\left[E_{r_i|y_i,\boldsymbol{x}_i}\left\{\frac{r_i}{\pi_i}\left(1 - \frac{r_i}{\pi_i}\right) K_h(y - y_i) q(\boldsymbol{x}_i)\right\}\right]\right) \\
&= \mathrm{E}_{\boldsymbol{x}_i}\left[q(\boldsymbol{x}_i)\mathrm{E}_{y_i|\boldsymbol{x}_i}\{K_h(y - y_i)\} E_{r_i|\boldsymbol{x}_i}\left\{\frac{r_i}{\pi_i}\left(1 - \frac{r_i}{\pi_i}\right)\right\}\right] \\
&= \mathrm{E}_{\boldsymbol{x}_i}\left\{q^2(\boldsymbol{x}_i)\left(-\frac{\pi_i(1-\pi_i)}{\pi_i^2}\right)\right\} \\
&= -\mathrm{E}_{\boldsymbol{x}_i}\left\{\left(\frac{1-\pi_i}{\pi_i}\right) q^2(\boldsymbol{x}_i)\right\}.
\end{aligned}$$

Therefore, cov $(A_n, B_n) = -\mathrm{var}(B_n)$, which gives the desired result, (8).

References

Bowman, A.W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71**, 353–360.

Bowman, A.W. & Azzalini, A. (2005). *sm: Smoothing methods for nonparametric regression and density estimation*. R package version 2.1-0; Ported to R by B. D. Ripley up to version 2.0 and later versions by Adrian W. Bowman and Adelchi Azzalini/. Available from URL: http://www.stats.gla.ac.uk/∼adrian/sm

Davidian, M., Tsiatis, A.A. & Leon, S. (2005). Semiparametric estimation of treatment effect in a pretest–posttest study with missing data. *Statist. Sci.* **20**, 261–301.

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. Ser. B Statist. Methodol.* **39**, 1–22.

Elmore, R.T., Hall, P. & Troynikov, V.S. (2006). Nonparametric density estimation from covariate information. *J. Amer. Statist. Assoc.* **101**, 701–711.

Fan, J., Heckmann, N.E. & Wand, M.P. (1995). Local polynomial kernel regression for generalized linear models and quasi likelihood functions. *J. Amer. Statist. Assoc.* **90**, 141–150.

Hall, P., Marron, J.S. & Park, B.U. (1991). Smoothed cross-validation. *Probab. Theory Related Fields* **92**, 1–20.

Hammer, S.M., Katzenstein, D.A., Hughes, M.D. et al. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England J. Medicine* **335**, 1081–1090.

Hayfield, T. & Racine, J.S. (2008). Nonparametric econometrics: The np package. *J. of Statist. Software* **27**. Available from URL: http://www.jstatsoft.org/v27/i05/.

Horvitz, D.G. & Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47**, 663–685.

Hurvich, C.M., Simonoff, J.S. & Tsai, C.L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. R. Statist. Soc. Ser. B Statist. Methodol.* **60**, 271–293.

Hyndman, R.J., Bashtannyk, D.M. & Grunwald, G.K. (1996). Estimating and visualizing conditional densities. *J. Comput. Graph. Statist.* **5**, 315–336.

Jones, M.C., Marron, J.S. & Sheather, S.J. (1996). A brief survey of bandwidth selection for density estimation. *J. Amer. Statist. Assoc.* **91**, 401–407.

Kennedy, K.F. (2007). A weighted estimating equation approach to local linear regression with missing covariate data (MSc Thesis). Kansas State University, Manhattan, Kansas, USA.

Little, R.J.A. & Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd edn. Hoboken, John Wiley & Sons.

Marron, J.S. & Wand, M.P. (1992). Exact mean integrated squared error. *Ann. Statist.* **20**, 712–736.

Müller, H.-G. (1987). Empirical bandwidth choice for nonparametric kernel regression by means of pilot estimators. *Statist. Decisions* **Supplement no. 2**, 193–206.

Nadaraya, E.A. (1964). On estimating regression. *Theory Probab. Appl.* **10**, 186–190.

Nottingham, Q.J., Birch, J.B. & Bodt, B.A. (2000). Local logistic regression: an application to army penetration data. *J. Statist. Comput. Simul.* **66**, 35–50.

Park, B.U. & Marron, J.S. (1990). Comparison of data-driven bandwidth selectors. *J. Amer. Statist. Assoc.* **85**, 66–72.

Qi, L., Wang, C.Y. & Prentice, R.L. (2005). Weighted estimators for proportional hazards regression with missing covariates. *J. Amer. Statist. Assoc.* **100**, 1250–1263.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. Available from URL: http://www.R-project.org

Robins, J.M., Rotnitzky, A. & Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89**, 846–866.

Rosenbaum, P.R. & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9**, 65–78.

Scott, D.W. & Terrell, G.R. (1987). Biased and unbiased cross-validation in density estimation. *J. Amer. Statist. Assoc.* **82**, 1131–1146.

SHEATHER, S.J. & JONES, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Statist. Soc. Ser. B Statist. Methodol.* **53**, 683–690.

SILVERMAN, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall/CRC.

STANISWALIS, J.G. (1989). Local bandwidth selection for kernel estimates. *J. Amer. Statist. Assoc.* **86**, 284–288.

TIBSHIRANI, R. & HASTIE, T. (1987). Local likelihood estimation. *J. Amer. Statist. Assoc.* **82**, 559–567.

WATSON, G.S. (1964). Smooth regression analysis. *Sankhyā, Series A* **26**, 359–372.