Research Article

Statistics in Medicine

Received 26 September 2009, Accepted 23 September 2010 Published online 28 December 2010 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4124

Estimating propensity scores with missing covariate data using general location mixture models

Robin Mitra[a*†] **and Jerome P. Reiter**[b]

In many observational studies, analysts estimate causal effects using propensity scores, e.g. by matching, sub-classifying, or inverse probability weighting based on the scores. Estimation of propensity scores is complicated when some values of the covariates are missing. Analysts can use multiple imputation to create completed data sets from which propensity scores can be estimated. We propose a general location mixture model for imputations that assumes that the control units are a latent mixture of (i) units whose covariates are drawn from the same distributions as the treated units' covariates and (ii) units whose covariates are drawn from different distributions. This formulation reduces the influence of control units outside the treated units' region of the covariate space on the estimation of parameters in the imputation model, which can result in more plausible imputations. In turn, this can result in more reliable estimates of propensity scores and better balance in the true covariate distributions when matching or sub-classifying. We illustrate the benefits of the latent class modeling approach with simulations and with an observational study of the effect of breast feeding on children's cognitive abilities.

Keywords: latent class; matching; missing data; multiple imputation; observational studies; propensity score

1. Introduction

In many studies of causal effects, analysts use observational data in which the treatment and control conditions are not randomly assigned to subjects. Typically in such studies, the subjects in the treated group look different than those in the control group on several covariates. When these covariates are related to the outcome of interest, any observed differences in the two groups' outcome distributions may reflect the differences in the groups' covariates rather than only effects of the treatment [1, 2].

Analysts can reduce the bias that results from imbalanced covariate distributions, at least for observed covariates, using propensity scores [3, 4]. The propensity score for any subject, $e(x_i)$, is the probability that the subject receives the treatment given its vector of covariates x_i. That is, $e(x_i) = P(T_i = 1|x_i)$, where $T_i = 1$ if subject i receives treatment and $T_i = 0$ otherwise. When two large groups have the same distributions of propensity scores, the groups should have similar distributions of x [3]. Thus, by selecting control units whose propensity scores are similar to the treated units' propensity scores, analysts can create a matched control group whose covariates in x are similar to the treated group's covariates. Analysts then base inferences on the treated and matched control groups, thereby avoiding any bias that results from imbalanced covariate distributions in x in the two groups. In addition to one-to-one matching, analysts can make causal inferences by using full matching [5, 6], by sub-classifying on the propensity scores, [7, 8], and by using the the propensity scores for inverse probability weighted estimation [9, 10]. For a review of approaches to causal inference using propensity scores, see [11].

A key assumption in propensity score analyses is no unmeasured confounding, i.e. there does not exist an unobserved variable that affects the outcome and is differentially distributed in the treated and control groups. This assumption, known as strong ignorability [3], cannot be verified in observational studies; however, analysts can perform sensitivity checks to determine if unmeasured confounding could seriously impact conclusions [12].

[a] *School of Mathematics, University of Southampton, Southampton, SO17 1BJ, U.K.*
[b] *Department of Statistical Science, Duke University, Box 90251, Durham, NC 27708, U.S.A.*
[*] *Correspondence to: Robin Mitra, School of Mathematics, University of Southampton, Southampton, SO17 1BJ, U.K.*
[†] *E-mail: R.Mitra@soton.ac.uk*

Propensity scores are rarely known exactly and must be estimated from the data. Typically, this involves fitting regressions with T as the dependent variable and functions of x as the independent variables, and using the estimated probabilities as the propensity scores. See, for example, [13].

In this article, we consider a practical complication in propensity score approaches: estimating the scores when some covariate data are missing. There are several strategies in the literature for estimating propensity scores with missing covariates. The analyst could base propensity score estimation and subsequent causal inference only on the complete cases; however, this could result in biased estimates when the data are not missing completely at random (MCAR). Even when the data are MCAR, the complete case analysis generally results in reduced power. The analyst could estimate propensity scores within patterns of missing data; however, with many patterns there may not be sufficient data for accurate estimation. The analyst could apply the model-based approach of [14]. They use an EM algorithm to find the maximum likelihood estimates of the parameters in a general location model fit to the data (X_{obs}, T). After the algorithm converges, they estimate propensity scores as the predicted probabilities in the regression of T on X_{obs}.

We propose to estimate propensity scores using multiple imputation of missing data [15]. In this approach, the data analyst repeatedly imputes missing values by sampling from their posterior predictive distributions conditional on the observed covariate data. The analyst estimates propensity scores in each completed data set, and averages the propensity scores across data sets. The averaged scores could be used for matching, sub-classification, or inverse weighting. Alternatively, the analyst could use the propensity scores in each data set to estimate the treatment effect separately in each data set, and average the estimated treatment effects over the data sets. There is no consensus on which of these two approaches is more effective, as evidenced in the simulations of [16]. We tried both approaches in our analyses and found that matching on the average propensity score resulted in more accurate treatment effect estimates than matching within each data set and so report only the former sets of results.

Multiple imputation approaches have some advantages over maximum likelihood approaches. With completed data sets, analysts can easily pursue further modeling, such as sub-domain comparisons or regression adjustment to reduce residual imbalances [16, 17]. For example, analysts select the matched control set using the averaged propensity scores, perform the regression within each completed data set using only this control set and the treated units, and average the resulting coefficients—which differ because the imputed covariate values change over the completed data sets—using the combining rules of [15]. Additionally, the analyst's model for the propensity scores is not tied to the model for imputations, which provides flexibility in estimating propensity scores. For example, a specification of the propensity score model other than the one implicit in the model used for imputation could result in better balance on the completed-data covariates.

For imputation, we propose a general location model, i.e. the categorical variables follow a log-linear model and the continuous variables follow a multivariate normal distribution within each category, with a novel twist. We introduce a latent indicator variable that captures the notion, 'if we had complete data, these control units would be good candidates for close matches to the treated units'. More precisely, we assume that the control units are a mixture of units whose covariates are drawn from the same distributions as the treated units' covariates, and units whose covariates are drawn from different distributions. This formulation reduces the influence of control units outside the treated units' region of the covariate space on the estimation of parameters in the imputation model, which can result in more plausible imputations in the region where control and treated units' covariate distributions overlap most. Since propensity scores are estimated based on imputed values, better imputations can result in estimated propensity scores that are closer to their true values, and thus result in better balance in the true covariate distributions when matching or sub-classifying on those scores. The latent variable is never observed for control units. However, because all treated units are by definition in the treated units' covariate space, there is sufficient information to estimate the posterior distributions of the latent indicators using Markov Chain Monte Carlo techniques.

The remainder of the article is organized as follows. In Section 2, we illustrate the impact on covariate balance of using standard imputation models that generate implausible imputations. Most standard imputation models use the same parameter estimates for all units; we call these one class models. In Section 3, we describe a simple latent class mixture model and illustrate its improved performance over one class models in the settings of Section 2. We also compare the latent class model with the imputation-based approaches suggested by Nagin and Rosenbaum [18] and Qu and Lipkovich [19]. In Section 4, we present the general location latent class mixture model, which we utilize in Section 5 to handle missing covariate data in an observational study of the effect of breastfeeding on children's cognitive outcomes later in life. In Section 6, we conclude with general remarks about the approaches.

We illustrate the multiple imputation procedures using one-to-one matching primarily for computational convenience; that said, one-to-one matching is still often utilized for causal inference in medical studies e.g. [20–23].

We conjecture that the performance of full matching or sub-classification with missing covariate data also will be improved when estimating propensity scores with latent class imputation models, as these approaches can benefit from more accurate estimation of propensity scores in the region of covariate overlap. In all sections, we use the average treatment effect on the treated units as the causal estimand.

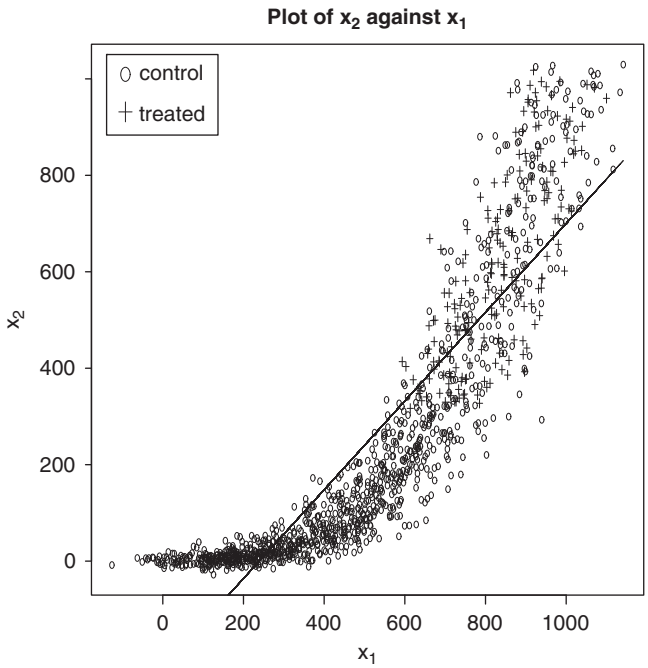

Figure 1. Scatter plot of x_2 against x_1 when a cubic relationship is present, illustrating the effects of using a poor imputation model.

There are many other approaches to handling missing covariate data in causal inference settings, for example doubly robust estimators [24, 25]. Reference [26] contains an extensive review of other approaches. We do not consider other methods further here.

2. Potential inadequacies of one class models

With high-dimensional covariate spaces, it is unfortunately all-too-easy to mis-specify the imputation models. Indeed, the difficulties of specifying models in high dimensions motivate propensity score methods in place of regression analysis for causal inference in the first place. We now demonstrate how mis-specified models can generate implausible imputations, which in turn can negatively impact covariate balance in propensity score matching. We use an obvious mis-specification in this example to clearly illustrate the impacts of implausible imputations on covariate balance.

We simulate two continuous covariates for $n=1200$ units as shown in Figure 1. The $n_T=200$ treated units tend to have larger values of x_1 and x_2 than the $n_C=1000$ control units. We introduce missing values in control units' x_2 data with a missing at random mechanism so that units with large values of x_1 are more likely to be missing x_2. Thus, there are many missing values among control units living in the same covariate space as the treated units. Approximately, 40 per cent of control units are missing x_2. The models used to create these simulations are in Appendix C of the supplementary material.

We impute missing x_2 using data augmentation [27] via the one class multivariate normal model. Specifically, we presume that $x_i \sim \mathrm{N}(\mu, \Sigma)$, with the non-informative prior distribution $p(\mu, \Sigma) \propto |\Sigma|^{-((p+1)/2)}$. Since only x_2 is missing in our simple example, imputations can be sampled directly from the posterior predictive distribution of $(x_2|x_1)$, which is a linear regression on x_1. For more complicated patterns of missing data, draws from the joint posterior distribution of the missing values can be obtained with Gibbs samplers. The draws of the missing values serve as multiple imputations.

Following [14], we do not control outcome data in the imputation model. This decision is made to stay consistent with the philosophy of propensity score matching: manipulations of the covariates and assigning of a control group are done without consideration of response variables. In this way, propensity score computations, and consequently any matching, subclassification, or weighting based on propensity scores, are not affected by assumptions about the outcome variable. That said, it can be advantageous to include the outcome in imputation models [28, 29]. One could easily modify any of the imputation models that follow to include outcome variables in the imputation models.

After $m=100\,000$ imputations, we estimate each unit's propensity score in each data set using a logistic regression with main effects for x_1 and x_2. Within any data set, this regression provides reasonable balance on the treated units' and

(possibly imputed) control units' covariates. We then average each unit's m propensity scores, and perform matching both with and without replacement based on the averaged propensity scores. We use a nearest neighbor matching algorithm in both settings. Generating 100 000 imputations took approximately 10 h of computing time on an Intel 2.8 GHz processor. We note that it is not necessary to set $m = 100\,000$ in genuine applications; we chose a very large m to essentially eliminate any sampling variability in the selected matches that could result from using a finite number of imputations. We discuss the role of m further in Section 5.

The one class multivariate normal imputation model implies a linear relationship between x_1 and x_2, which is clearly inappropriate as indicated by the estimated regression line in Figure 1. What can happen when using this regression model to impute the missing x_2? First, consider control units with actual covariate values of both x_1 and x_2 in the treated units' region of the covariate space and above the regression line; these are ideal candidates to be included in the matched control set. When based on the one class model, imputations of missing x_2 for these control units will tend to be lower than the actual values. As a result, these control units' completed data could be in a different space than the treated units' covariates. If propensity score matching is done with the completed data, these control units could be (incorrectly) excluded from the matched control set. Second, consider control units with values of x_1 similar to treated units' values of x_1 but with actual values of x_2 that are smaller than any treated units' values of x_2 (e.g. below the regression line); these are not good candidates for the matched control set. When based on the one class model, imputations of missing x_2 for these units will tend to be higher than their true x_2 values, so that their imputed values could be in the same region as the treated units' covariates. Therefore, they could be incorrectly selected as matched controls. We note that control units whose covariates are far away from the treated units' covariate space are not likely to be selected as matches, even with the model mis-specification.

Figure 2 displays the distributions of true x_1 and x_2 values for the treated and matched control units with multiple imputation using the one class model. These matched controls are also compared with the matched controls selected before introduction of missing x_2 values, which we call the original complete data. When matching without replacement, the lower quartiles for both x_1 and x_2 for the matched controls from the one class model are smaller than the lower quartiles for the treated units and matched controls selected from the original complete data. This is primarily because the model tends to impute missing x_2 values higher than their true values for control units just outside the treated units' covariate space. When matching with replacement, the imbalance is more noticeable, with disparities in both the lower and upper quartiles of the distributions for both covariates. The implausible imputations result in controls with dissimilar true values of x_1 and x_2 being selected multiple times to be in the matched set.

One might think that the problems resulting from poor imputations can be fixed by adjusting propensity score models to achieve the best possible balance on the completed-data covariates. Unfortunately, this may not resolve the problems. When an imputation model generates implausible imputations, acceptable balance on the imputed (and observed) covariates may not equate to acceptable balance on the true values of the covariates. For example, suppose that a treated unit has covariate values of $(x_1 = 1, x_2 = 2)$. Suppose that we consider two possible candidates for matches to this unit, one with $(x_1 = 1, x_2 = 2)$ and the other with $(x_1 = 1, x_2 = 0)$; clearly, record one is preferred. However, suppose that the treated unit and record two are missing x_2, and that we use a terrible imputation model that imputes $x_2^* = 10$ for the treated and control record. Then, the propensity score matching algorithm will select record 2, although it is a worse match in reality and although the propensity score matching is perfectly balanced on x_1 and the imputed value of x_2^*.

Of course, when confronted with these data in a genuine setting, a wise modeler would recognize the inadequacy of the multivariate normal model from exploratory data analysis and use some other imputation approach. Imputations from correct, or at least approximately correct, models can help analysts avoid these problems. However, in settings with many covariates, it is not always easy to diagnose model inadequacies. Furthermore, although unfortunate, many analysts default to multivariate normal imputation procedures, so that they may face the problems from imputation model mis-specification.

3. Latent class mixture model

The simulations in Section 2 demonstrate that implausible imputations can negatively impact covariate balance. They further illustrate that inappropriate use of a one class model can lead to these problems. In this section, we propose an approach that attempts, in some sense, to mitigate these problems automatically through latent class mixture modeling. We note that Beunckens *et al.* [30] also use latent class models for multiple imputation, but not in the context of propensity score matching.

The motivation underlying the use of latent class models in this context is as follows. In matching contexts, ideally we want to select controls that look like the treated units on relevant covariates. When covariate data are missing, we are unsure which control units are in this region of potential matches. However, if we knew which control units were in the potential match region, we could toss out the control units outside the potential match region and, therefore, fit

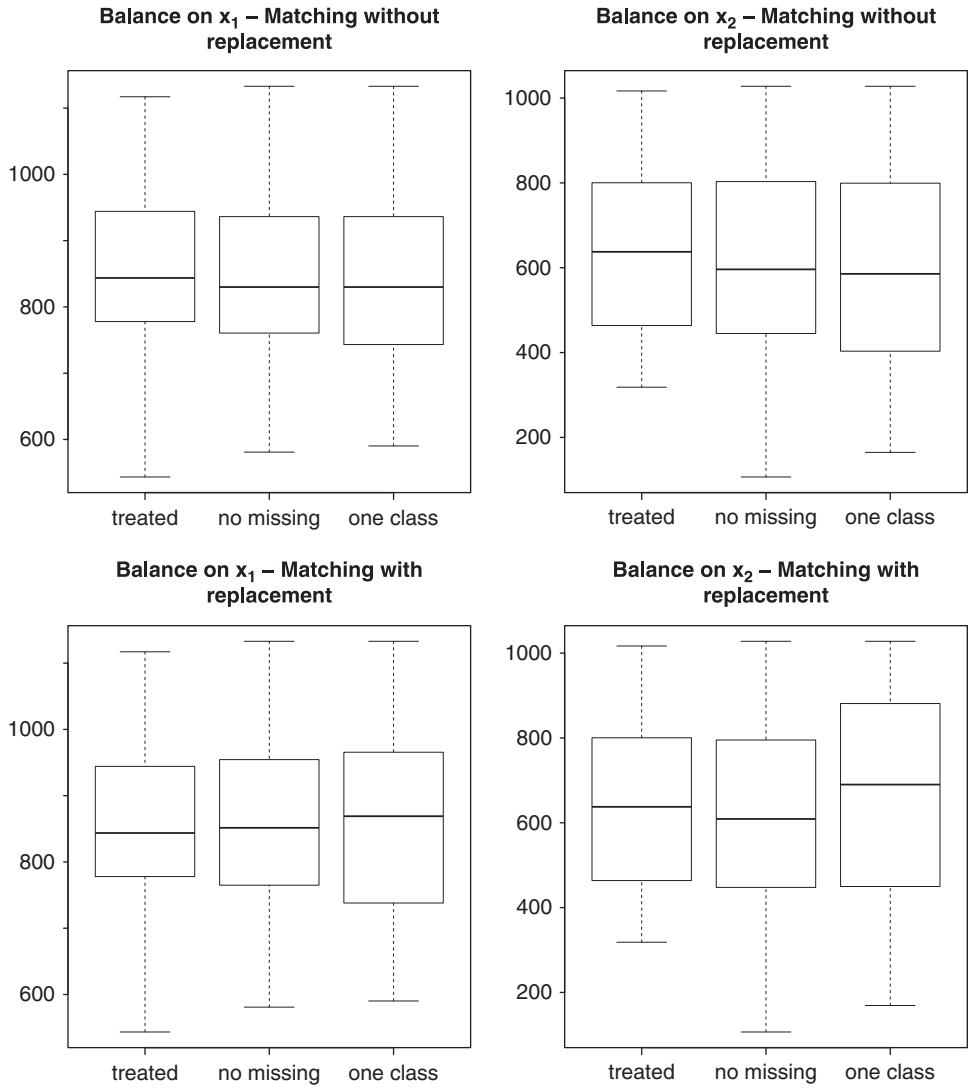

Figure 2. Box plots of x_1 and x_2 for the treated units, matched control units selected from the original complete data (before introduction of missing values in x_2), and matched control units from the one class model in the model mis-specification simulation design. Matching is performed both without and with replacement.

imputation models using only the relevant covariate space. In this way, imputation of missing covariates in the treated units' covariate space would not be affected by outlying controls, as happened in Section 2. Since we do not know which control units are in the potential matched region, we introduce a latent class indicator such that one class corresponds to units lying in the potential matched region and the other class corresponds to all other units. With this two-class model, by definition we know the latent class indicators for all treated units, so that there is information to estimate the distribution of the latent class indicators for the control units.

The latent class mixture model is specified so that the missing data for treated and control units estimated to have the same covariate distributions are imputed based on the same parameter values. Imputing with a common parameter value can be preferable to imputing treated and control units with the same covariate distributions separately. In effect, the parameters from the common distribution are estimated with the data from both the treated and potential matched controls, which reduces variance in the imputations and hence can improve matches in any one data set compared with estimating parameters separately for each treatment group.

When the covariate distribution of the control group has smaller variance than that of the treated group, latent class models still can provide more plausible imputations than one class models. A one class model would impute control units' missing values from a distribution based on the entire data set, which would have a variance for the control units that is too large, leading to implausible imputations. A latent class model has the potential to distinguish the control units' distribution from the treated units' distribution, so that the imputations will effectively be from two separate models.

This can result in more plausible imputations for the control units. Of course, arguably this situation is problematic for propensity score inference, since many treated records will not have reasonably close matches in the data set. Analysts may have to adopt other inferential strategies, e.g. regression modeling with flexible functional forms [31].

We now demonstrate that the latent class model can improve the problems seen in Section 2. We begin by describing a latent class mixture model for continuous variables only. We present a general location latent class mixture model for continuous and categorical variables in Section 4.

3.1. Latent class model for continuous data only

For each unit i, let $z_i \in \{0, 1\}$ be the latent class indicator, where $z_i = 1$ corresponds to unit i lying in the treated units' covariate space and $z_i = 0$ otherwise. We model each unit's covariate data conditional on z_i with class-specific parameters, so that,

$$x_i | z_i \sim \mathrm{N}(\mu^{z_i}, \Sigma^{z_i}). \tag{1}$$

The distribution of the latent class indicators conditional on treatment is

$$p(z_i = 1 | T_i = 0) = \pi^*, \tag{2}$$

$$p(z_i = 1 | T_i = 1) = 1. \tag{3}$$

As in the one class model we place non-informative priors on $(\mu^{z_i}, \Sigma^{z_i})$, so that

$$p(\mu^{z_i}, \Sigma^{z_i}) \propto |\Sigma^{z_i}|^{-((p+1)/2)}. \tag{4}$$

This can lead to an improper posterior when $z_i = z_j$ for all (i, j). However, this possibility is rare in practice. If this does occur, a one class model may be adequate. Analysts can adopt the approach of [32], also recommended by Wasserman [33], and use a data-dependent prior distribution that restricts imputation of z_i so that sufficient numbers of units are in both classes. We place a Beta prior distribution on π^*, $p(\pi^*) = \mathrm{Be}(a, b)$, where (a, b) are specified hyper-parameters. Common choices for (a, b) include $a = b = 1$, implying a uniform prior for π^*, and $a = b = 0.5$ for Jeffrey's prior.

With this model specification, the full conditional distributions are available in closed form. It is straightforward to sample from the joint posterior distribution of all unknowns using a Gibbs sampler, thus creating multiple imputations of the missing covariate values.

3.2. Performance in simulations

We now apply the latent class model in the setting of Section 2. We also examine a scenario in which the one class model is appropriate in order to illustrate the effect on covariate balance when the latent class model is inefficient compared with the one class model. For each scenario, we run the Gibbs sampler for 100 000 iterations after a burn-in period of 1000 runs, thus creating 100 000 multiply imputed data sets. It takes about 10 h of computing time to create the 100 000 imputed data sets on a 2.8 GHz Intel processor. We choose $m = 100\,000$ for the illustration because, given a missing data pattern, with this large m the set of matched controls would be the same for any repeated run of the imputations; thus, using $m = 100\,000$ essentially guarantees that the results are not affected by variability from finite m. In each completed data set, we estimate the propensity scores using the logistic regressions described in Section 2. We then compute the average propensity scores for each unit across the 100 000 data sets, and match treated to control units using nearest neighbor matching, both with and without replacement.

Figure 3 summarizes the covariate balance on x_1 and x_2 for the simulation with model mis-specification for the latent class and one class multiple imputation approaches. These results are based on the same data set and same one class imputations used in Section 2; matching is done only once for each method. For both matching with and without replacement, true covariate balance is slightly improved when using the latent class model as compared with using the one class model. Notably, obtaining more plausible imputations of x_2 not only helps to balance x_2, it results in better balance on x_1.

In the model mis-specification scenario, the latent class model is not the correct model for $f(x_2|x_1)$. However, as evident from Figure 1, using a linear model for imputations is not unreasonable for units lying in the treated units' region of the covariate space. This points to a general advantage of adding the latent indicators: assumptions of linearity or other simplifications, while possibly inappropriate over the whole covariate space, may be reasonable on a smaller region where the treated units lie.

Of course, covariate balance is an intermediate step in causal inference. The ultimate goal is to estimate treatment effects. We therefore simulate a response variable, y, with a simple response surface, namely

$$y_i = x_{i1} + x_{i2} + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(0, 1). \tag{5}$$

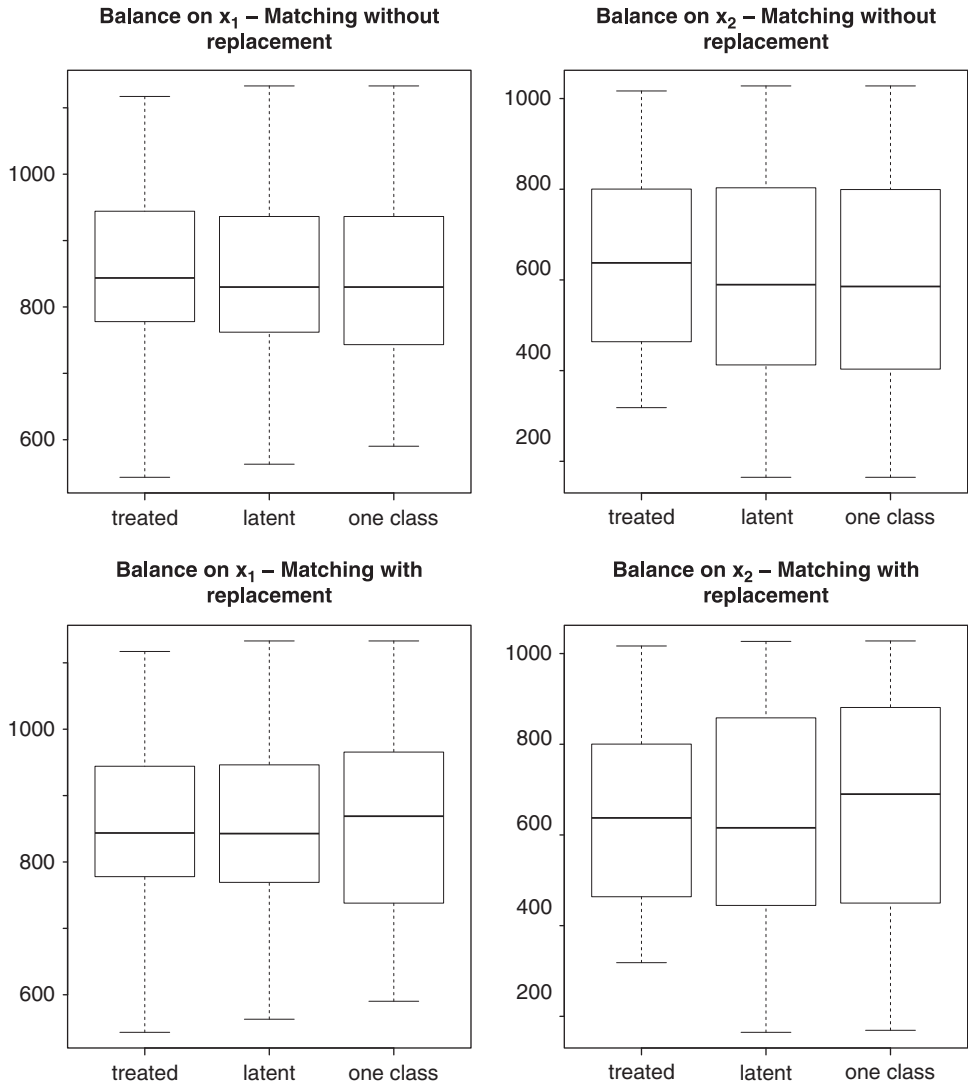

Figure 3. Box plots for x_1 and x_2 for treated and matched control units in the model mis-specification simulation. Here, 'latent' indicates controls selected via the general location mixture model, and 'one class' indicates controls selected via the one class model.

Here, the treatment effect $\tau = 0$. We estimate τ with $\hat{\tau} = \bar{y}_T - \bar{y}_{MC}$, where $\bar{y}_T$ is the sample mean of y in the treated group and $\bar{y}_{MC}$ is the sample mean of y in the matched control group.

We create three data sets, labeled 'Rep 1,' 'Rep 2,' and 'Rep 3,' by simulating three sets of outcomes from (5) given the true values of x_1 and x_2. We introduce missing values in the control units' data as described in Section 2. We also introduce missing values in the treated units' x_2 data. Specifically, we consider scenarios with 0, 10, and 30 per cent of treated units' x_2 data MCAR. In each data set, we apply the latent class and one class approaches to create $m = 100\,000$ imputed data sets, from which we estimate the propensity scores as described in Section 2. We construct the matched control set using nearest neighbor matching both with and without replacement.

In addition to the one class and latent class approaches, we examine two alternative imputation-based approaches to handling missing data in propensity score estimation. The first approach implements the suggestion from [18]: use the one class model to create one completed data set, and include indicators for the missing data patterns in the models used to estimate propensity score models from this single completed data set. We label this HNR. The second approach implements the suggestion from [19]: use the one class model to create m completed data sets (where $m = 100\,000$ in our setting), and include indicators for the missing data patterns in the models used to estimate propensity score models from these completed data sets. We label this QL.

Table I summarizes the estimated treatment effects for three simulated data sets after applying the four methods of handling missing data. The table also includes the estimated treatment effects before introduction of any missing data, which we call $\hat{\tau}_{com}$. In general, the estimated treatment effects for the latent class model track $\hat{\tau}_{com}$ more closely than

Table I. Three replications of treatment effect estimates when matching with and without replacement in the model mis-specification scenario for various approaches to handling missing data via imputation.

	Without replacement			With replacement		
Estimate	Rep. 1	Rep. 2	Rep. 3	Rep. 1	Rep. 2	Rep. 3
Original complete data	30.6	42.7	22.8	14.9	18.5	18.1
0 per cent missing treated						
Latent class	43.0	52.6	34.8	16.6	12.1	18.2
One class	52.4	62.7	54.6	−34.1	−34.2	−6.2
HNR	634.7	634.9	526.3	−17.3	5.7	3.4
QL	634.7	634.9	526.3	−17.3	5.7	3.4
10 per cent missing treated						
Latent class	44.3	55.1	35.1	26.6	26.8	34.2
One class	46.1	61.8	50.4	−22.1	−19.7	−14.1
HNR	216.4	266.2	272.6	−27.8	35.2	0.5
QL	227.4	244.1	249.1	−31.2	10.6	−11.2
30 per cent missing treated						
Latent class	43.8	45.9	38.5	18.7	−2.4	13.0
One class	42.0	49.6	34.7	−2.7	−42.0	6.6
HNR	120.4	123.3	62.6	−72.0	−26.9	−66.2
QL	117.3	90.1	46.5	−73.3	−92.2	−61.7

The true treatment effect equals zero. Across the three replications, the $\mathrm{SE}(\bar{Y}_{\mathrm{T}}) \approx 22$.

the other methods of handling the missing data. For example, the other methods often result in $\hat{\tau}<0$, whereas this never happens for $\hat{\tau}_{\mathrm{com}}$ and happens once for the latent class model.

For the matching without replacement scenarios, the latent class model outperforms the one class model with 0 and 10 per cent missing data in all three data sets; the two methods have similar performances with 30 per cent missing data. Both the latent class and one class models dominate QL and HL when matching without replacement, although the advantage decreases with the percentage of missing data for treated units. The relatively poor performances of QL and HNR for small fractions of missing data and matching without replacement result because these methods often select controls without missing data, since the propensity score model includes an indicator for missing x_2 and most treated units have a zero value for that indicator. Many of these control records appear to be reasonable matches based on the imputed values of x_2, but in reality are poor matches based on the true values of x_2.

For the matching with replacement scenarios, no method dominates across all the data sets. The estimated treatment effects for the latent class model appear to be the most stable: for any scenario, their variance across data sets tends to be smallest. The estimates for the one class model are almost always negative, and they vary widely across data sets in the 0 and 30 per cent missing data scenarios. With no missing data in the treated units, HNR and QL tend to result in estimated treatment effects closest to $\tau=0$. As noted previously, in this setting they tend to select matches with complete data. Because control units with observed (x_1, x_2) similar to those for treated units can be used repeatedly as matches, HNR and QL effectively avoid the problems from mis-specified imputation models by not selecting control records with imputed x_2. Their performance deteriorates as the amount of missing treated data increases, because the matching algorithm increasingly selects records with poorly imputed values of x_2.

One might ask what happens when the one class model is correct for the covariates, but imputations are done with the latent class model. In this case, the latent class model estimates the parameters for the treated/matched class using only a fraction of the control units, whereas the one class model appropriately uses all control units. This loss of efficiency in parameter estimation results in greater imputation variance, which could worsen the quality of matches with respect to true covariate balance.

To explore this scenario, we add a simulation in which (x_1, x_2) have a linear relationship. The $n_{\mathrm{T}}=200$ treated units tend to have larger values of x_1 and x_2 than the $n_{\mathrm{C}}=1000$ controls. This simulation design is summarized in Figure 4. As in the previous simulations, we introduce missing values in control units' x_2 data with a missing at random mechanism so that units with large values of x_1 are more likely to be missing x_2, and we introduce 0, 10, or 30 per cent missing values in x_1 using an MCAR mechanism. The models used to create these simulations are in Appendix C of the supplementary material. We simulate the response surface as in (5) and estimate treatment effects accordingly. We create $m=100\,000$ completed data sets for each of three simulated data sets using the latent class, one class, HNR, and QL methods.

Table II summarizes the results from the correctly specified simulation scenario. For without replacement matching, the latent class and one class estimates are very similar, and both dominate the QL and HNR methods by wide margins. For matching with replacement, the one class estimates tend to be closer to $\tau=0$ than the latent class estimates. The absolute magnitudes of the difference tend to be less than those in Table I. For 0 per cent missing data, the QL and

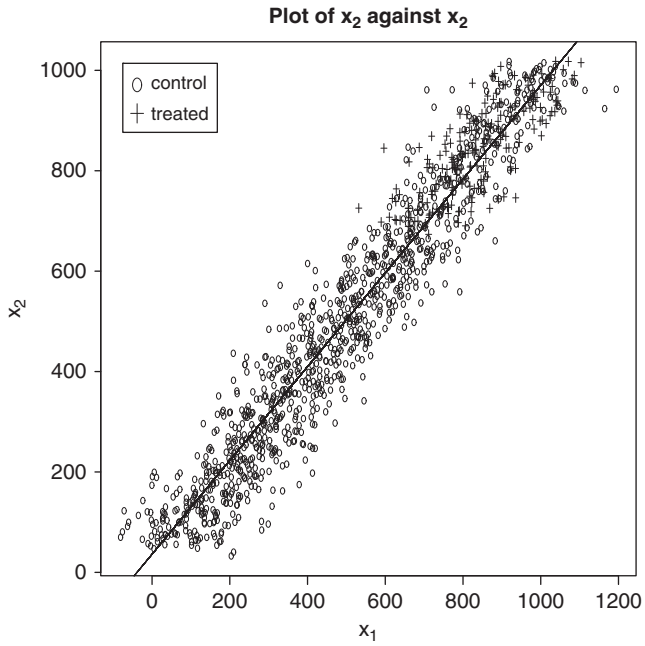

Figure 4. Scatter plot of x_2 against x_1 when a linear relationship is present and the one class model holds.

Table II. Three replications of treatment effect estimates when matching with and without replacement in the correct specification scenario for various approaches to handling missing data via imputation.

	Without replacement			With replacement		
Estimate	Rep. 1	Rep. 2	Rep. 3	Rep. 1	Rep. 2	Rep. 3
Original complete data	14.4	18.8	39.9	−0.7	1.2	1.8
0 per cent missing treated						
Latent class	18.5	25.5	45.9	−6.7	−11.4	−22.3
One class	16.5	22.8	44.9	3.8	−4.9	−4.2
HNR	500.1	499.7	479.6	6.7	2.6	1.4
QL	500.1	499.7	479.6	6.7	2.6	1.4
10 per cent missing treated						
Latent class	18.2	25.9	44.8	−8.1	−9.1	−21.6
One class	15.0	25.6	44.5	6.3	−0.2	−14.0
HNR	176.3	218.0	199.3	−8.1	16.2	5.8
QL	164.8	202.7	203.9	−24.8	9.8	0.09
30 per cent missing treated						
Latent class	15.7	23.2	45.7	−3.6	−13.1	−21.9
One class	22.0	26.4	47.8	14.3	−0.02	6.0
HNR	102.9	99.6	98.6	−24.3	8.8	−32.1
QL	96.9	83.4	103.3	−33.5	−33.9	−34.3

The true treatment effect equals zero. Across the three replications, the $\mathrm{SE}(\bar{Y}_T) \approx 14$.

HNR estimates tend to be closer to $\tau = 0$ than the latent class estimates; however, this is no longer the case with 30 per cent missing data. Again, differences across methods are small compared with those seen in Table I.

The simulation results suggest that the latent class model can help analysts to avoid bias from poor matches caused by implausible imputations, without substantial penalties when they are inefficient compared with one class models.

4. General location latent class mixture model

When covariates include both categorical and continuous variables, the general location model is often used for imputation of missing data. In this section, we extend the general location model to include latent indicators for propensity score matching.

4.1. One class general location model

Let X be an $n \times p$ matrix of covariate data for the individuals in the study comprising q continuous variables, $W = (W_1, \ldots, W_q)$, and r categorical variables, $V = (V_1, \ldots, V_r)$, where $q + r = p$. Each V_j takes on d_j distinct values. Thus, each unit can be classified into one of $D = \prod_{j=1}^{r} d_j$ cells of an r-dimensional contingency table.

Let $f = \{f_d : d = 1, \ldots, D\}$ be the resulting set of cell counts, assuming an appropriate (e.g. anti-lexicographical) ordering of cells. We assume that f has a multinomial distribution with probability vector $\pi = \{\pi_d : d = 1, 2, \ldots D\}$. Within each cell d, we assume that W follows a multivariate normal distribution, $p(W|\mu_d, \Sigma) = N(\mu_d, \Sigma)$. Here, μ_d is the q-vector of means for cell d, and Σ is the $q \times q$ covariance matrix assumed equal for all d. We use a Dirichlet prior distribution for π with pre-specified hyper-parameters $\alpha = (\alpha_1, \ldots, \alpha_D)$. We use a non-informative prior distribution on (μ, Σ), i.e. $p(\mu, \Sigma) \propto |\Sigma|^{-((q+1)/2)}$.

In many applications, D is quite large, possibly exceeding n. With large D, many cells are empty or sparsely populated. To allow estimation of the parameters, analysts can restrict π and $\mu = (\mu_1, \ldots, \mu_D)$. For π, a typical approach is to use log-linear constraints. Specifically, let C be a $D \times s$ matrix such that $s \leqslant D$. The log-linear model requires π to satisfy $\log(\pi) = C\lambda$. Typically, C contains main effects for each V_j and possibly interactions among selected (V_i, V_j). For μ, analysts can specify a linear model on the categorical variables. This model frequently mimics the structure of C, including main effects and interactions among $V_1, \ldots, V_q$. Analysts may need to transform components of W so that the assumption of multivariate normality within cells are more reasonable, for example by using Box–Cox transformations [34].

Analysts can use a Gibbs sampler to sample from the joint posterior distribution of unknowns. A convenient approach for obtaining posterior draws of π is Bayesian iterative proportional fitting; see, [35, Ch. 4] and [36]. Conditional on parameter draws, missing categorical data are imputed from multinomial distributions and missing continuous data are imputed from multivariate normal distributions.

4.2. Adding the latent class indicators

As in Section 3.1, let z_i be the latent class indicator for each unit i. We model the distribution of latent class indicators as in (2) and (3), with the same considerations for the prior distribution on π^*. Given the latent class indicators $z = \{z_1, \ldots, z_N\}$, we can partition the data into two groups. Let $X = (X^0, X^1)$, where $X^0 = \{x_i : z_i = 0,\ i = 1, \ldots, n\}$ and $X^1 = \{x_i : z_i = 1,\ i = 1, \ldots, n\}$ correspond to covariates for units belonging to latent classes 0 and 1, respectively. As in Section 4.1, we can further partition the data into its continuous and categorical components, $X^0 = (V^0, W^0)$ and $X^1 = (V^1, W^1)$.

Essentially, the mixture model specifies separate general location models for X^0 and X^1. Let $\theta^0 = (\pi^0, \mu^0, \Sigma^0)$ and $\theta^1 = (\pi^1, \mu^1, \Sigma^1)$ be the parameters of the general location model for X^0 and for X^1, respectively. Then,

$$p(X|\theta^*, z) = p(X^0|\theta^0)p(X^1|\theta^1) \tag{6}$$

where $p(X^0|\theta^0)$ and $p(X^1|\theta^1)$ are modeled as described in Section 4.1. Cell counts are still modeled with multinomial distributions, but now cell probabilities depend on latent class membership. Similarly, the continuous data are still modeled as multivariate normal—after possible transformations—but the mean and covariance matrix depend on the latent class.

As in Section 4.1, we use Dirichlet prior distributions for π^1 and non-informative prior distributions for $(\mu^1, \Sigma)^1)$, and similarly for (π^0, μ^0, Σ^0). As in Section 3.1, with non-informative prior distributions, sufficient numbers of units are required in both classes to estimate the parameters.

The full conditional distributions for all unknowns are available in closed form. We describe in detail the data augmentation steps needed to impute missing values in Appendix A of the supplementary material. R code implementing this approach is provided at http://www.soton.ac.uk/∼rm1s07/robin_rcode.html.

5. Application to study breast feeding

We now apply the latent class model to impute missing covariates and perform propensity score matching in a study of the effect of breast feeding on child's cognitive development. The data are a subset of the 1979 National Longitudinal Survey of Youth, commonly referred to as the NLSY79. This longitudinal survey, begun in 1979, interviewed a nationally representative sample of 12 686 young men and women in the U.S. aged 14–22 years at that time. This cohort was interviewed annually until 1994, and biannually after then. After 1986, detailed information on children born to women in the NLSY79 were collected.

5.1. Description of variables

The response variable, y, is the Peabody individual assessment test math score (PIATM) administered to children at 5 or 6 years of age. The treatment variable is breast feeding duration, which is measured in weeks. We dichotomize this variable into a control condition, $<$24 weeks, and a treatment condition, $\geqslant$24 weeks. The 24 week cutoff corresponds to the number that has been given by the American Academy of Pediatrics [37] and the World Health Organization as a minimum standard for breast feeding duration. There are other ways to define the treatment variable, and the analysis could be repeated with different cut points on the breast feeding duration variable. We do not pursue these here. Additionally, we cannot determine from these data whether or not the mother used breast feeding exclusively.

We use 14 potentially relevant background covariates. These include five categorical variables: the child's race (Hispanic, black or other), the mother's race (Hispanic, black, asian, white, Hawaiian/Pacific Islander/American Indian, or other), child's sex, and two variables indicating whether the spouse or grandparents were present at birth. They also include nine continuous variables: the number of years between 1979 and when the mother gave birth, mother's intelligence as measured by an armed forces qualification test, mother's highest educational attainment, child's birth weight, the number of days that the child spent in hospital, the number of days that the mother spent in hospital, the number of weeks that the mother worked in the year prior to giving birth, the number of weeks the child was born premature, and family income.

We applied Box–Cox transformations [34] to several continuous variables to improve the assumption of normality; see Appendix B of the supplementary material for details. We also categorize the number of weeks the child was born premature into three levels: not preterm (zero weeks), moderately preterm (one to four weeks), and very preterm (five or more weeks), with cut points determined from guidelines of the March of Dimes (http://www.marchofdimes.com). The categorization was used because weeks preterm has a very large spike at zero weeks, as seen in its histogram displayed in Figure 6 in Appendix B of the supplementary material. Finally, we categorize the number of weeks that the mother worked in the year prior to giving birth into four levels: not worked at all, worked between 1 and 47 weeks, worked 48–51 weeks, and worked all 52 weeks. This variable has a distinct U-shaped histogram, which would be difficult to capture with a normal model; see Figure 7 in Appendix B of the supplementary material.[‡]

We include only first born children in the analysis to avoid complications due to birth order and family nesting. In addition, we discard 506 units with missing breast feeding duration and 4977 units with a missing PIATM. Excluding these units is reasonable under missing at random assumptions, which may not be true in practice. We do not consider other methods for handling the missing treatment indicators and missing outcome data in the analysis here, as the cases with complete treatment and outcome data suffice for our purposes: to examine the implications for covariate balance and treatment effect estimation of using the latent class imputation model. The resulting data comprise 2388 youths, of whom 370 are treated. Of these, 1306 have complete data on all covariates, of whom 216 are treated. Three covariates were completely observed in the study, and nine covariates had missing data rates of less than 10. The two covariates with the largest rates of missing data were family income (22.4 per cent) and the number of weeks that the mother worked in the year prior to giving birth (23.1 per cent).

Several covariates in the available data are clearly imbalanced. To illustrate, we focus on three variables. Figure 5 summarizes the distribution of transformed mother's intelligence score and years of education for observed treated and control units, and Table III displays the proportion of treated and control units in each level of child's race. Treated units tend to have higher mother's intelligence scores, more mother's years of education, and lower proportions of Hispanics and blacks. Because of these imbalances, we seek to do propensity score estimation and matching in the presence of the missing data.

5.2. Complete case simulation

We first evaluate the performance of the latent class model at achieving true covariate balance in a simulation involving the 1306 complete cases. Although this is a much smaller sample size, we can introduce missing data, run the model, and examine covariate balance using the true data. We introduce missing values by randomly sampling with replacement from the missing data patterns present in the original data set. This results in 717 units with fully observed covariates; the remainder have some missing data. For the latent class imputation model, we use a main effects only log-linear model for the categorical variables. We use the transformed continuous variables that were suggested by the Box–Cox procedure, and relate the within-category means using a linear model with main effects of the categorical variables. We run the Gibbs sampler for 200000 iterations after discarding an initial 5000 as burn-in. We estimate propensity scores with a main effects logistic regression, and create the matched control set by nearest neighbor matching without

[‡]*Supporting information may be found in the online version of this article.*

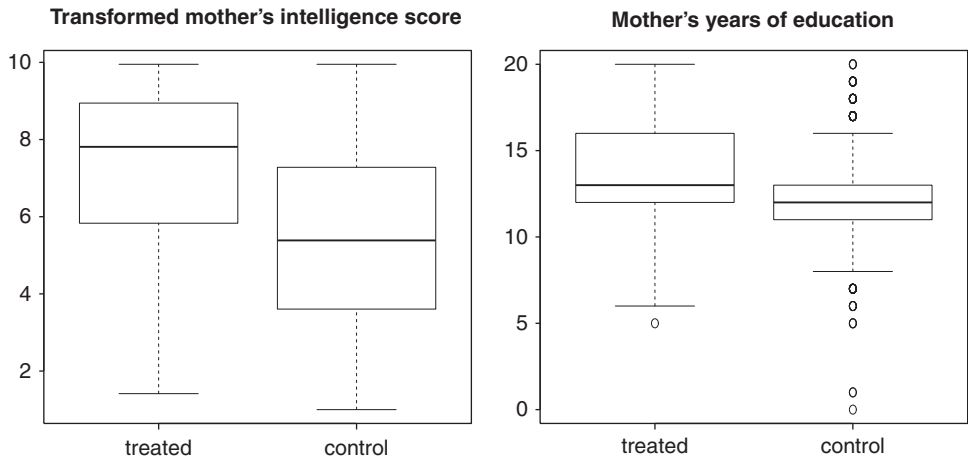

Figure 5. Box plots of the transformed mother's intelligence score and mother's years of education, respectively, for treated and control units before matching.

Table III. Distribution of child's race for treated and control units before matching.

Race	Treated	Control
Hispanic	0.1378	0.1903
Black	0.1108	0.2844
Other	0.7514	0.5253

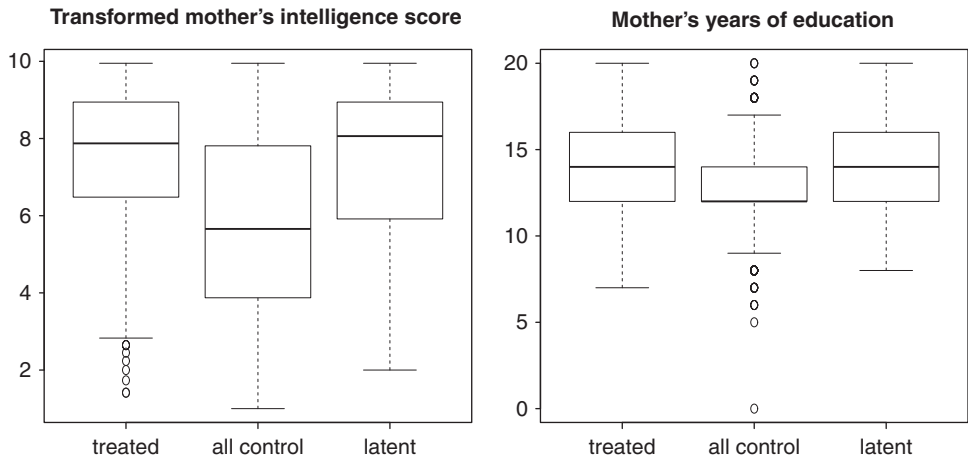

Figure 6. Boxplots of true values for two covariates for the treated units, full control reservoir, and matched control units selected using the general location mixture model (labeled 'latent') in the simulation involving the complete cases.

replacement. It took 9 days of computing time to generate this many imputed data sets on a 2.8 GHz Intel processor. We did not optimize the coding and used R for implementation; undoubtedly, computing time would be drastically reduced with efficient coding in a programming language like C.

Figure 6 displays the distributions of transformed mother's intelligence score and years of education for the treated units, full control reservoir, and matched control units selected using the latent class model. For both variables, the imbalance has been greatly reduced after imputation and matching. We present more detailed examination of balance on these covariates, for this simulation, in Figures 8 and 9 in Appendix B of the supplementary material.

We also compare proportions of child's race for treated and control units before and after matching in Table IV. Once again, covariate imbalance is greatly reduced here. Similar examinations with other variables indicate that the latent class model results in a well-balanced matched control set with respect to these covariates.

Table IV. True joint distribution of two covariates for the treated units, full control reservoir, and matched control units selected via the general location mixture model (labeled 'latent') in the simulation involving the complete cases.

Race	Treated	All controls	Latent
Hispanic	0.1528	0.1844	0.1296
Black	0.0926	0.2697	0.1111
Other	0.7546	0.5459	0.7593

5.3. Application to the full data

We now apply the latent class model on the original data set of 2388 units. Similar restrictions are imposed on the cell probabilities and within cell means as in the simulation involving the complete cases. We again run the Gibbs sampler for 200 000 iterations with an additional burn-in of 5000 iterations. It took 16 days of computing time to generate this many imputed data sets on a 2.8 GHz Intel processor. The treatment effect estimates listed here should be viewed as illustrative; a more thorough analysis would investigate the sensitivity of results to the assumption that PIATM is missing at random and to potential unmeasured confounding. We perform matching as described in the complete case simulation.

We estimate the treatment effect with $\bar{Y}_{\mathrm{T}} - \bar{Y}_{\mathrm{MC}} = -0.059$, with a standard error computed using the matched pairs of 0.92. For a discussion of approaches to estimating standard errors from propensity score matching, see [38] for matching without replacement and [39] for matching with replacement contexts. This is noticeably different than the treatment effect estimate based on all controls, which is 5.23 (two-sample SE = 0.74). The treatment effect after matching is thus significantly closer to zero. Similar results were obtained by Der *et al.* [40], who used a regression approach to infer that the effect of breast feeding is minimal.

For comparison, we also used the one class model to impute missing values. The estimated treatment effect is 0.96 (matched pairs SE = 0.89). Thus, there is approximately a one point (and one standard error) difference in the treatment effect estimates from the two imputation approaches. The difference is modest primarily because, on average across imputations of z_i, approximately 85 per cent of control units are imputed to lie in the latent class for the treated/matched control region, so that there is not much difference between the two models. We note, however, that certain outlying controls, e.g. one mother spent many more weeks in the hospital than others, are never in the latent class for treated units. Hence, these outlying data points will not unduly influence the estimated regression coefficients for units in the region of potential matches. This is not the case with the one class model, which does not separate those units out as part of the estimation process. This points to another general advantage of using the latent indicators: the latent class model can moderate the impact of outlying units on imputations for units in the region of plausible matches.

We also repeated the complete case simulation with the one class model. It was difficult to distinguish a clear winner on covariate balance between the one class and latent class models. Both greatly improved covariate balance over use of the full control reservoir.

6. Concluding remarks

When analysts estimate treatment effects with missing covariate data using propensity score approaches, using latent class mixture models for imputations can result in more plausible imputations, and hence better covariate balance when matching on the scores, than using a one class model. Essentially, the latent class model allows the data analyst to estimate imputation models specifically for the region of interest. In this way, control units that have minimal relevance for treatment effect estimation also have minimal relevance for imputation of missing data. Even when one class models are correct, we anticipate that the loss in efficiency from using the latent class model will be minor, particularly when compared with the reductions in bias achievable from avoiding implausible imputations. This was borne out in the simulations in Section 3, and to some extent in the breast feeding study in which the one class and latent-class approaches gave results with only modest differences.

We used a large number of imputations in both the simulated and genuine data examples. We found that, when matching on averaged propensity scores, as $m \to \infty$ the propensity score for each record converges on one value, so that the selected matched control set stabilizes, and there is no additional variance from using a finite number of imputations. In genuine settings, researchers may not be willing to wait for the imputation program to generate a large number of imputed data sets. One option is to monitor the composition of the matched set, and stop the imputations when it does not change after a large number of imputations. Another option is to use modest m and live with the additional variance.

This issue arises regardless of the choice of the latent class, one class, or QL approach to imputation. Further study is needed to document to impacts of finite m and characterize the trade off in accuracy and computational costs.

There is flexibility in the mixture modeling approach. The model might include more than one class for the treated and matched control units, or more than one class for the other control units. The advantages of additional classes is the potential to model the distributions more accurately, and hence to generate more plausible imputations, than with only two classes. For example, it is well known that mixture distributions can approximate complex distributional features, such as multi-modality and long tails, that single normal distributions may struggle to capture. The primary disadvantage of adding classes is computational: one has to ensure that the Markov Chain Monte Carlo algorithms converge, which takes longer with increased numbers of parameters. We note that label switching, which is a sticky problem in mixture modeling with many components, is not a major concern here, because the goal is imputation rather than interpretation. We also note that when the model has more than one class for the treated units, the analyst does not know the class membership of each treated unit, so that these component indicators must be estimated from the data as well. Inefficiencies akin to those seen in the correct model simulation in Section 3 can arise when unnecessary classes are included in the model. This leads to questions of how to specify the latent classes, which could be explored with semi-parametric approaches such as Dirichlet process mixture models.

The latent class approach also could be implemented in a chained equations approach to multiple imputation, as is used in MICE in Stata and R and IVEWARE in SAS. Each conditional distribution could include a latent class indicator, and estimation would proceed akin to the approach in [41]. However, it may be necessary to run a Gibbs sampler for each conditional model at each cycle of chained equations. A chained equation approach could enable greater flexibility in modeling compared with the two-class general location latent mixture model.

Acknowledgements

The authors would like to thank Professor Jennifer Hill who provided us with the data from the breast feeding study analyzed in Section 5. This research was supported by the National Science Foundation (NSF-ITR-0427889).

References

1. Cochran WG, Chambers SP. The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A* (*General*) 1965; **128**(2):234–266.
2. Rubin D. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974; **66**(5):688–701.
3. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**:41–55.
4. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 1985; **39**(1):33–38.
5. Rosenbaum PR. A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society*, *Series B—Methodological* 1991; **53**:597–610.
6. Stuart EA, Green KM. Using full matching to estimate causal effects in nonexperimental studies: examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology* 2008; **44**(2):395–406.
7. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 1984; **79**:516–524.
8. Hullsiek KH, Louis TA. Propensity score modeling strategies for the causal analysis of observational data. *Biostatistics* (*Oxford*) 2002; **3**(2):179–193.
9. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**(5):550–560.
10. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* 2004; **23**(19):2937–2960.
11. D'Agostino Jr RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine* 1998; **17**:2265–2281.
12. Rosenbaum PR. Sensitivity analysis for certain permutation inferences in matched observational studies (Corr: V75 p396). *Biometrika* 1987; **74**:13–26.
13. Woo MJ, Reiter JP, Karr AF. Estimation of propensity scores using generalized additive models. *Statistics in Medicine* 2008; **27**(19): 3805–3816.
14. D'Agostino Jr RB, Rubin DB. Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association* 2000; **95**(451):749–759.
15. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York, 1987.
16. Hill J. Reducing bias in treatment effect estimation in observational studies suffering from missing data. Columbia University Institute for Social and Economic Research and Policy (ISERP). *Working Paper 04-01*, 2004;
17. Hill JL, Reiter JP, Zanutto EL. A comparison of experimental and observational data analyses. In *Applied Bayesian Modeling and Causal Inference from an Incomplete-Data Perspective*, Gelman A, Meng XL (eds). Wiley: New York, 2004.
18. Haviland A, Nagin D, Rosenbaum P. Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods* 2007; **12**(3):247–267.

19. Qu Y, Lipkovich I. Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. *Statistics in Medicine* 2009; **28**(9):1402–1414.
20. Cho YB, Lee K, Suh K, Kim Y, Yoon J, Lee H, Hahn S, Park B. Maternal smoking during pregnancy and birthweight: a propensity score matching approach. *Journal of Gastroenterology and Hepatology* 2007; **22**(10):1643–1649.
21. da Veiga PV, Wilder RP. Maternal smoking during pregnancy and birthweight: a propensity score matching approach. *Maternal and Child Health Journal* 2008; **12**(2):194–203.
22. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine* 2008; **27**(12):2037–2049.
23. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment group in propensity-score matched samples. *Statistics in Medicine* 2009; published online in Wiley InterScience.http://www.interscience.wiley.com.
24. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 1994; **89**:846–866.
25. Liang H, Wang S, Robins JM, Carroll RJ. Estimation in partially linear models with missing covariates. *Journal of the American Statistical Association* 2004; **99**(466):357–367.
26. Ibrahim JG, Chen MH, Lipsitz SR, Herring AH. Missing-data methods for generalized linear models: a comparative review. *Journal of the American Statistical Association* 2005; **100**(469):332–346.
27. Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* 1987; **82**:528–540.
28. Little RJA. Regression with missing *X*'s: a review. *Journal of the American Statistical Association* 1992; **87**:1227–1237.
29. Moons KGM, Donders RART, Stijnen T, Harrell Jr FE. Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology* 2006; **59**(10):1092–1101.
30. Beunckens C, Molenberghs G, Verbeke G, Mallinckrodt C. A latent-class mixture model for incomplete longitudinal gaussian data. *Biometrics* 2008; **64**:96–105.
31. Hill JL. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 2010; DOI: 10.1198/jcgs.2010.08162.
32. Diebolt J, Robert CP. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society. Series B* (*Methodological*) 1994; **56**(2):363–375.
33. Wasserman L. Asymptotic inference for mixture models by using data-dependent priors. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*) 2000; **62**(1):159–180.
34. Box GEP, Cox DR. An analysis of transformations. *Journal of the Royal Statistical Society. Series B* (*Methodological*) 1964; **26**(2):211–252.
35. Schafer JL. *Analysis of Incomplete Multivariate Data*. Chapman & Hall: London, 1997.
36. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*. Chapman & Hall: London, 1995.
37. Chantry CJ, Howard CR, Auinger P. Full breastfeeding duration and associated decrease in respiratory tract infection in US children. *Pediatrics* 2006; **117**(2):425–432.
38. Austin PC. Type I error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. *The International Journal of Biostatistics* 2009; **5**(1). DOI: 10.2202/1557-4679.1146.
39. Hill J, Reiter JP. Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine* 2006; **25**:2230–2256.
40. Der G, Batty GD, Deary IJ. Effect of breast feeding on intelligence in children: prospective study, sibling pairs analysis, and meta-analysis. *British Medical Journal* 2006; **333**. DOI: 10.1136/bmj.38978.699583.55.
41. Raghunathan TE, Lepkowski JM, van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology* 2001; **27**:85–96.