Towards Multiple-Imputation-Proper Predictive Mean Matching

Philipp Gaffert* Florian Meinfelder† Volker Bosch‡

Abstract
Rubin (1987) introduced multiple imputation in a parametric Bayesian framework and considers it proper if the uncertainty of the imputation fully propagates. Augmenting it with a semiparametric concept like predictive mean matching (Rubin 1986, Little 1988) promises both, valid inferences and robustness against some model misspecifications. Although numerous multiple imputation predictive mean matching algorithms exist their theoretical properties remain largely unexplored. In this paper, we show why all of these algorithms are improper, but the one by Siddique & Belin 2008. On this exception we build a new algorithm and demonstrate its superiority in terms of coverages of frequentist confidence intervals within a comparative simulation study.

Key Words: Approximate Bayesian bootstrap, Distance-based donor selection, Hot deck imputation, Multiple imputation, Predictive mean matching, Proper imputation

1. Introduction

Imputation often precedes statistical analysis of missing data. '[T]o represent the uncertainty about which value to impute' Rubin proposed multiple imputation. It replaces each missing value by $M \geqslant 2$ imputed values and adds an extra component to the variance of any estimator through Rubin's combining rules (Rubin, 1987). In contrast to treating imputed values as if they were observed, multiple imputation inhibits overly progressive frequentist inference.

In the Bayesian framework, in which multiple imputation is set, draws from well defined distributions give the imputations. As these distributions do typically not reflect empirical distributions appropriately, implausible imputations occur in applications. Predictive mean matching, like other hot deck methods, imputes values from the observed part of the data set and hence guarantees plausible imputations. As a nearest neighbor technique predictive mean matching is also more robust to model misspecification (Schenker & Taylor, 1996, p. 429), namely nonlinear associations, heteroscedastic residuals, and deviations from normality (Morris et al., 2014, p. 4). Nonetheless, the quality of predictive mean matching imputations largely depends upon the availability of near donors; truncation settings typify its limits as shown in the 2009 Bamberg University PhD thesis by the second author.

Combining multiple imputation with predictive mean matching promises a robust imputation procedure yielding valid inferences and is thus highly appealing to practitioners. Consequently, a suchlike combination constitutes not only a feature but often the default of the imputation algorithms in all major statistical software programs (Morris et al., 2014, p. 3). In contrast, skepticism dominates the literature (see, e.g., Little & Rubin (2002, p. 69)). As an example, Morris et al. (2014, p. 5) note in the given context: '... there is thus no guarantee that Rubin's rules will be appropriate for inference.' This contrast between the uncertainty in theory and the prominence in applications motivated our work.

*Global Data Science, GfK SE, Nordwestring 101, 90419 Nuremberg, Germany
†Department of Statistics and Econometrics, Otto-Friedrich-University, Feldkirchenstrasse 21, 96052 Bamberg, Germany
‡Global Data Science, GfK SE, Nordwestring 101, 90419 Nuremberg, Germany

In this paper, as our main contribution, we elaborate one key deviation of multiple imputation predictive mean matching algorithms from the theory of multiple imputation. Knowing about this deviation we identify the algorithm by Siddique & Belin (2008) as, in the sense of Rubin (1987), the most multiple imputation proper one and thus as the foundation for our proposed algorithm. A simulation study demonstrates its empirical preeminence regarding coverages of frequentist confidence intervals.

2. Fully parametric multiple imputation

Let the data be n independent realizations of a p-dimensional normal random vector (Y, X). We suppose that the $n \times (p-1)$ matrix of covariates X is fully observed and define n_{mis} and $n_{obs} = n - n_{mis}$ as the number of missing values or recipients and the number of observed values or donors in the $n \times 1$ vector y, respectively. We assume ignorability for the missingness as $pr(y = missing) = \Phi\left(X^*\alpha + \eta\right)$, where X^* denotes the X matrix of covariates with a leading constant column; η independent normal noise; α a vector of parameters of length p; and Φ the normal cumulative distribution function. In the given setting the correct conditional imputation model is the linear model $y = X^*\beta + \varepsilon$ with β denoting a vector of parameters of length p and ε independent normal noise with zero mean and variance σ_ε^2. Fully parametric multiple imputation repeats the following steps $M \geqslant 2$ times to correctly reflect the uncertainty about the parameters of the imputation model (Little & Rubin, 2002, p. 216).

1. The posterior step. First draw from the posterior distribution of the residual variance $\tilde{\sigma}_\varepsilon^2 \mid y_i, X_i$ given by $\Gamma^{-1}\{n_{obs}/2, (y_i - X_i^*\hat{\beta})^{\mathrm{T}}(y_i - X_i^*\hat{\beta})/2\}$. Then draw from the posterior distribution of the intercept and slope parameters $\tilde{\beta} \mid y_i, X_i, \tilde{\sigma}_\varepsilon^2$ given by $N_p\{\hat{\beta}, \tilde{\sigma}_\varepsilon^2(X_i^{*\mathrm{T}}X_i^*)^{-1}\}$. y_i and X_i refer to the fully observed subset of the data, and $\hat{\beta}$ denotes the maximum likelihood parameter estimate.

2. The imputation step. Draw n_{mis} times independently from the imputation model, i.e. $\tilde{y}_j \sim N(X_j^*\tilde{\beta}, \tilde{\sigma}_\varepsilon^2)$.

The next section introduces three nonparametric algorithms within the multiple imputation set-up that underlie our proposed algorithm.

3. Nonparametric multiple imputation

3.1 The approximate Bayesian bootstrap imputation

The approximate Bayesian bootstrap imputation was the first nonparametric multiple imputation algorithm (Rubin & Schenker, 1986, p. 131). In the posterior step the approximate Bayesian bootstrap imputation algorithm draws a bootstrap sample from the donors instead of parameters, and, in the imputation step, it draws from this bootstrap sample considering the integer bootstrap frequencies ω instead of from the conditional predictive distribution. However, Kim (2002, p. 472) shows that inferences are correct for $n_{obs} \to \infty$ only, because, just like the maximum likelihood estimator, the bootstrap estimator ignores the appropriate degrees of freedom correction (Davison & Hinkley, 1997, p. 22). Thus, for finite n_{obs} the total parameter variance is underestimated. Parzen et al. (2005, p. 973) show that multiplication of the total variance estimate for the mean with the following factor ϕ removes the bias.

$$\phi(n_{obs}, n_{mis}, M) = \frac{\frac{n^2}{n_{obs}} + \frac{n_{mis}}{M}\left(\frac{n-1}{n_{obs}} - \frac{n}{n_{obs}^2}\right)}{\frac{n^2}{n_{obs}} + \frac{n_{mis}}{M}\left(\frac{n-1}{n_{obs}} - \frac{n}{n_{obs}^2}\right) - \frac{nn_{mis}}{n_{obs}}\left(\frac{3}{n} + \frac{1}{n_{obs}}\right)} \geqslant 1. \qquad (1)$$

Some criticism about this correction factor stems from Demirtas et al. (2007).

3.2 Predictive mean matching

As opposed to the approximate Bayesian bootstrap, predictive mean matching (Rubin, 1986, p. 92) substitutes the draw from the conditional predictive distribution only, i.e. the imputation step.

1. Calculate the predictive mean for the n_{obs} elements of y as $\hat{y}_i = X_i^* \hat{\beta}$.
2. Calculate the predictive mean for the n_{mis} elements of y as $\tilde{y}_j = X_j^* \tilde{\beta}$.
3. Match each element of $\tilde{y}_j$ to the its respective closest element of $\hat{y}_i$.
4. Impute the observed y_i of the closest matches.

Algorithm 1: The following four steps constitute the predictive mean matching procedure by Little (1988, p. 292).

3.3 Distance-based donor selection

For the posterior step the distance-based donor selection algorithm by Siddique & Belin (2008), which Siddique & Harel (2009) later called MIDAS, employs bootstrapping as originally proposed by Heitjan & Little (1991, p. 18). Maximum likelihood estimation of the imputation model parameters on M independent bootstrap samples replaces the draws from the posterior distribution (Little & Rubin, 2002, p. 216). The unique feature of the algorithm by Siddique & Belin (2008) is that it reuses the donor's bootstrap frequencies for the imputation step. For recipient j donor i from the full donor pool is drawn with probability

$$v_{i,j} = f(\omega, \tilde{y}_i, \tilde{y}_j, \kappa) = \omega_i \tilde{d}_{i,j}^{-\kappa} / \sum_{i=1}^{n_{obs}} (\omega_i \tilde{d}_{i,j}^{-\kappa}). \tag{2}$$

ω denotes the nonnegative integer bootstrap frequencies of the donors, $\tilde{d}_{i,j}$ the scalar absolute distance between the predictive means of donor i and recipient j and κ a closeness parameter adjusting the importance of the distance. For $\kappa = 0$ the procedure is equivalent to the approximate Bayesian bootstrap, for $\kappa \to \infty$ the procedure becomes equivalent to nearest neighbor matching as in algorithm 1.

4. Why predictive mean matching is not multiple-imputation-proper

Using PMM for the multiple imputation of data sets causes the between variance of the parameter estimates of interest to suffer from attenuation bias.

To illustrate this situation, consider an analysis of variance example with $\psi = 1, \ldots, \Psi$ different predictor cells. Suppose that the incomplete variable Y in cell ψ is normally distributed with mean μ_ψ and variance σ_ψ^2. Furthermore, suppose that each of the Ψ cells contains a sufficient number of donors, say, five or more. Now, without loss of generality, let us examine at the recipients in the first cell. Parametric multiple imputation draws $M \geqslant 2$ times $\tilde{\sigma}_1^2$, then $\tilde{\mu}_1 \mid \tilde{\sigma}_1^2$, and then $\tilde{y}_{\psi=1} \mid (\tilde{\mu}_1, \tilde{\sigma}_1^2)$, which is efficient. A nonparametric alternative is an approximate Bayesian bootstrap imputation in cell $\psi = 1$ that proceeds as follows. It draws $M \geqslant 2$ times a bootstrap sample from the donors in the cell and draws values to impute from this bootstrap sample. The key element that these two proper

procedures have in common is that the distribution from which the imputed values are drawn varies over the multiple imputations. In the parametric case the parameters of the underlying normal distributions vary, and in the nonparametric case, the composition of the empirical distribution varies.

	fully parametric	PMM	ABB imputation
Posterior step	Draw $(\tilde{\beta}, \tilde{\sigma}_v^2)$ from the imputation model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + v$, with heteroscedastic residuals, i.e., $var(v \mid \psi = 1) = \sigma_{v,1}^2, \ldots, var(v \mid \psi = 4) = \sigma_{v,4}^2$.		Within each of the $\Psi = 4$ cells draw a bootstrap sample of the $n_{obs,1}$, …, $n_{obs,4}$ donors
Imputation step	Draw from the normal imputation model: $\tilde{y}_j \mid (\tilde{\beta}, \tilde{\sigma}_v^2, x_1, x_2)$	As within each cell the predicted means $\hat{\tilde{y}}_\psi$ are identical, algorithm 1 draws $n_{mis,\psi}$ values from $n_{obs,\psi}$, i.e., a simple random hot-deck imputation within the cell	Within each cell, draw $n_{mis,\psi}$ values from the *bootstrapped* $n_{obs,\psi}$, i.e., a simple random hot-deck imputation within the *bootstrapped* cell

Table 1: The algorithms of proper fully parametric imputation, proper approximate Bayesian bootstrap (ABB) imputation, and PMM are compared. The underlying data situation involves two binary predictors (x_1, x_2), one incomplete variable y, and normal noise v. The two predictors form $\Psi = 4$ cells: $\psi(x_1 = 0, x_2 = 0) = 1, \ldots, \psi(x_1 = 1, x_2 = 1) = 4$. Ignorability is assumed. In the imputation step, PMM is very similar to ABB imputation, but it ignores the bootstrap. Because ABB imputation is approximately proper, PMM must attenuate the between imputation variance.

PMM proceeds in a considerably different manner. The recipients and the donors in cell $\psi = 1$ end up having exactly the same predicted mean[1]. Choosing the nearest neighbor ultimately consists of making a random draw from the donors in cell $\psi = 1$. This may be valid once, but the procedure is the same for all $m = 1, \ldots, M$ imputations. It thereby mimics the simple random hot-deck (Lillard et al., 1982, p. 15), which is known to underestimate the between variance component because it partly omits the posterior step. Table 1 schematically presents this reasoning.

It appears to be surprising that although PMM contains a draw from the estimated distribution of the intercept and slope parameters β (see section 2 and algorithm 1), the parameter uncertainty does not propagate. In this regard, the above example is deceptive. Therefore, consider another example. For simplicity, suppose that there are two normal orthogonal predictors x_1, x_2. Now, the definition of the relevant donors is less clear than in the previous example, where it appeared obvious that all donors of $\psi = 1$ are suitable. The job of the β is simply to define the relevant 'cell'. Drawing $\tilde{\beta}$ is an important task, because the cell definition is not certain and must thus vary over the multiple imputations. Figure 1 displays the effect of varying β coefficients on the cell definition.

However, PMM then goes wrong. The cells are defined, i.e., we have conditioned on $\tilde{\beta}$, and all PMM does is make a random draw from the cell or even take the nearest one. It thereby ignores parameter uncertainty to a large extent. To be precise, the $\tilde{\beta}$ define the mean of the cell; however, the uncertainty in estimating the residual variance parameter σ_v^2 from the imputation model (see section 2) remains unconsidered. In any given cell, we observe a distribution of units in a sample, which suffers from sampling error. Thus,

[1]This is only true if type-2 matching is applied, which slightly differs from algorithm 1. For details see van Buuren (2012, p. 71).

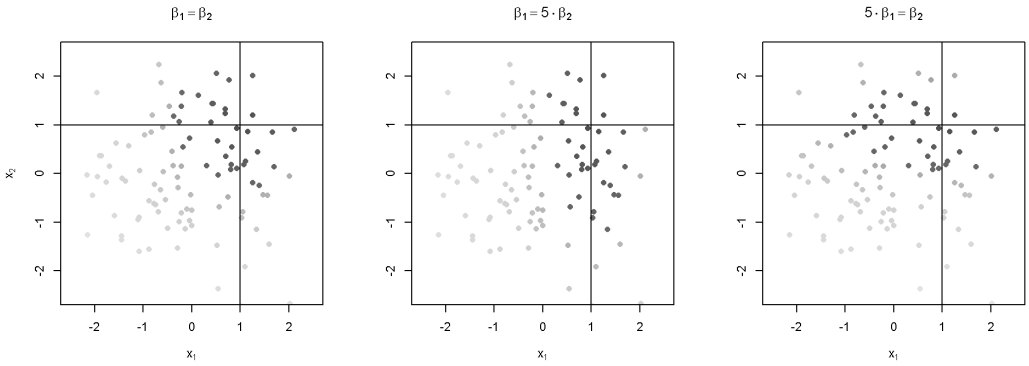

Figure 1: The plots show 100 random draws from a bivariate normal distribution with zero correlation. The shading indicates distances in the predictive means to one recipient $P_0(x_1 = 1, x_2 = 1)$. Different draws from the estimated distribution of the β parameters can alter the definition of the cell from which the donor is drawn. Considering distances, not frequencies, the cell is a circle in the left plot, a long ellipse in the middle plot and a wide ellipse in the right plot.

what is needed is some type of approximate Bayesian bootstrap imputation algorithm *after* conditioning on the $\tilde{\beta}$ parameters.

5. Existing ideas to make predictive mean matching proper

PMM has recently been under suspicion for underestimating the between variance. Van Buuren (2012, p. 71) and Morris et al. (2014, p. 7) criticize the selection of the nearest neighbor of algorithm 1. Selecting the nearest neighbor is a special case of general k-nearest-neighbor selection (Heitjan & Little, 1991, p. 16), which is typically applied in current statistical software programs (see table 3). An adaptive procedure for choosing the optimal k exists (Schenker & Taylor, 1996, p. 442), but software implementations of this procedure are lacking. The attenuation bias argument is that $k = 1$ leads to selecting the same donor repeatedly across imputations. The insight of section 4 is that once the cell is defined, the bootstrap frequencies are necessary to correctly reflect the between variance. The nearest neighbor selection function, however, is unable to fully capture the variance of bootstrap frequencies $var(\omega_i)$. If the nearest donor receives a bootstrap frequency that is larger than zero, then it will be selected. The exact value of the bootstrap frequency is irrelevant. It is easily found that $var(\omega_i) \geqslant var\{I(\omega_i)\}$, where I is a function that indicates whether ω_i is larger than zero. Therefore, the nearest neighbor selection is not compatible with the necessary bootstrap step. This finding underpins the criticism by van Buuren (2012) and Morris et al. (2014).

In addition to the nearest neighbor selection, van Buuren (2012, p. 71) and Morris et al. (2014, p. 7) criticize the very popular match type 2 (see table 3). In the discussion of match types, three different types can be distinguished. Type 1 refers to the matching of $\hat{y}_i$ to $\hat{\tilde{y}}_j$, as in algorithm 1. By contrast, type 2 refers to the matching of $\hat{\tilde{y}}_i$ to $\hat{\tilde{y}}_j$ (Heitjan & Little, 1991, p. 19). Type 3 refers to a procedure in which two sets of parameters, denoted by $\{(\tilde{\beta}_1, \tilde{\sigma}^2_{v,1}), (\tilde{\beta}_2, \tilde{\sigma}^2_{v,2})\}$, are drawn from the posterior distribution, one for the donors and one for the recipients, and $\hat{\tilde{y}}_i \mid (\tilde{\beta}_1, \tilde{\sigma}^2_{v,1})$ is then matched to $\hat{\tilde{y}}_j \mid (\tilde{\beta}_2, \tilde{\sigma}^2_{v,2})$ (Royston & White, 2011; Harrell, 2015). The criticism relates to the one predictor case, where type-2 matching linked with $k = 1$ causes the M multiple imputations to be identical

and therefore, prevents the uncertainty associated with parameter estimation from being propagated; again, this is an attenuation bias argument.

The insight from section 4 reveals that the M multiple imputations are identical only because the algorithm lacks the necessary bootstrapping. The parametric imputation step as in section 2 is conditioned on one set of parameters drawn in the posterior step, as in the case of type 2. Other match types alter the cell definition and are an engineering trick that treat the symptom, which occurs in the special case of one predictor, but do not cure the disease of effectively omitting the posterior step. Consequently, the discussion on match types is dispensable, and the use of type-2 matching should be advocated for.

6. The proposed algorithm

6.1 Revisiting the MIDAS algorithm

In contrast to algorithm 1 and all other PMM implementations (see table 3), the MIDAS algorithm proposed by Siddique & Belin (2008), which has been introduced in section 3.3, explicitly combines the two steps that are required based on the insights of section 4. The parameters $\tilde{\beta}$ and κ define the cell. The larger κ is, the smaller is the cell. The uncertainty involved in estimating β is correctly considered, and κ is not an estimate. However, because the within cell distribution has sampling error, equation (2) involves the bootstrap frequencies. The MIDAS algorithm is thus a major improvement in terms of multiple imputation theory, although its inventors have not been aware of this fact (Juned Siddique, personal communication 2016; Thomas R. Belin, written communication 2017). The proposed algorithm 2 largely builds on MIDAS. Although other PMM algorithms (i.e., distance functions) could easily be adjusted to deploy the bootstrap frequencies in the imputation step, very recent and yet unpublished research by Anna Poehlmann a graduate student from the University of Bamberg, Germany indicates that MIDAS significantly outperforms all known alternatives.

6.2 Making predictions for recipients and donors

The magnitude of the error, which is caused by partly omitting the posterior step, depends on the magnitude of the variance of the imputation model parameter estimates that is in turn inversely proportional to the number of available donors. Consequently, the MIDAS algorithm will be particularly beneficial when n_{obs} is small. In small samples, however, the influence of a single data point on the model parameter estimates can be considerable. Because model estimation implies minimizing the distance from the model to the donor data, the model is, by construction, closer to the donors than to the recipients, particularly for small n_{obs}, i.e., residuals systematically differ between donors and recipients. Consequently, the expectation of the residual variance added to the recipients is too small. Although this implementation is still the most common, Gelman & Hill (2011) and Meinfelder & Schnapp (2015) estimate the parameters on the full set of observations by using previously imputed values for y_j. These algorithms make in-sample predictions for both the donors and the recipients. By contrast, the proposed algorithm 2 makes only out-of-sample predictions by estimating the β parameters with the leave-one-out principle.

6.3 A flexible closeness parameter

The closeness parameter κ in equation (2) determines the influence of the imputation model, i.e., of the conditionality on X, on the donor selection. In contrast to Siddique & Belin (2008), who simply employ a fixed value, we argue that κ could reflect the goodness

1. Obtain bootstrap frequencies ω_i for the donors.
2. Draw $\tilde{\beta}$ from a weighted least-squares regression with the weights ω_i and calculate the according coefficient of determination $\hat{\tilde{R}}^2$.
3. Calculate the elements of the $n_{mis} \times n_{obs}$ distance matrix using the leave-one-out principle as follows: $\hat{\tilde{\varphi}}_{i,j} = |(x_{\underline{i}} - x_{\underline{j}})\tilde{\beta}_{-i}|$. Here, $x_{\underline{i}}$ denotes the row vector of X_i for the ith donor, $x_{\underline{j}}$ denotes the row vector of X_j for the jth recipient, and $\tilde{\beta}_{-i}$ denotes the weighted least-squares parameter vector from the donor sample without the ith row.
4. Calculate the closeness parameter as follows:

$$\hat{\tilde{\kappa}}(\hat{\tilde{R}}^2) = \left\{ 50\hat{\tilde{R}}^2 / \left(1 + \epsilon - \hat{\tilde{R}}^2\right) \right\}^{3/8}, \tag{3}$$

   where ϵ is a very small positive scalar number used to ensure real results for $\hat{\tilde{R}}^2 = 1$.
5. Insert ω_i, $\hat{\tilde{\varphi}}_{i,j}$, and $\hat{\tilde{\kappa}}$ from above into equation (2) and draw the donors.
6. Repeat the above steps $M \geqslant 2$ times, apply Rubin's rules, and multiply the total variances of the means by the correction from equation (1). Substitute n_{obs} with n_{eff} from equation (4), and thus, n with $n_{\mathit{eff}} + n_{mis}$.

Algorithm 2: This touched-up version of the MIDAS algorithm is named *midastouch*, which is implemented in the `R::mice` package (van Buuren & Groothuis-Oudshoorn, 2011).

of fit of the imputation model such that $\partial\kappa/\partial R^2 > 0$. The rationale is that the probability of drawing a distant donor should decrease as the imputation model quality increases, as in equation (3). Its functional form is the inverse of the form of the sales response to advertising function presented by Little (1970, p. B472). Siddique & Belin (2008, p. 88) state that reasonable values for κ lie within the range $[0, 10]$, and they found in a simulation study that in a setting with $R^2 = 0.29$, the ideal value for κ is 3 (Siddique & Belin, 2008, p. 98). Equation (3) reflects these findings as follows:

$$\kappa(R^2 = 0) = 0, \quad \kappa(R^2 = 0.9) \approx 10, \quad \kappa(R^2 = 0.29) \approx 3$$

6.4 Fixing the attenuation bias of the approximate Bayesian bootstrap imputation

Because equation (2) generalizes the approximate Bayesian bootstrap imputation, it also suffers from the underestimation of the total variance for finite n_{obs} (Kim, 2002). Applying the correction factor ϕ from equation (1) appears to be the most obvious solution. It applies directly to the k-nearest-neighbor distance function[2] if conducted on the bootstrapped donor sample. The available donors for each recipient, however, are no longer n_{obs}, but rather k, which causes a slight adjustment in equation (1): n_{obs} must be substituted by k, and n must be substituted by $k + n_{mis}$. After conditioning on the bootstrap frequencies, all donors have the same probability of being drawn. This is different for the MIDAS algorithm and for algorithm 2, because the drawing probabilities depend on the distance to the recipient. Therefore, we propose replacing n_{obs} in equation (1) with a measure of the effective donor sample size for each recipient $n_{j,\mathit{eff}}$ (Kish, 1965, p. 427), which is expressed as follows: $n_{j,\mathit{eff}} = n_{j,obs}^2 / \sum_i (w_{i,j}/\omega_i)^2$ (Bosch, 2005, p. 5). $w_{i,j}$ and ω_i denote the drawing probabilities from equation (2) and the bootstrap frequencies, respectively. Averaging over all recipients and the M imputed data sets yields

$$n_{\mathit{eff}} = \frac{1}{M n_{mis}} \sum_{m=1}^{M} \sum_{j=1}^{n_{mis}} \left[\sum_{i=1}^{n_{obs}} \left\{ \hat{\tilde{\varphi}}_{i,j,m}^{-\hat{\tilde{\kappa}}_m} \Big/ \sum_{i=1}^{n_{obs}} (\omega_{i,m} \hat{\tilde{\varphi}}_{i,j,m}^{-\hat{\tilde{\kappa}}_m}) \right\}^2 \right]^{-1} \tag{4}$$

Variance correction factors for parameters other than the mean do not yet exist; for linear regression parameters, Wu (1986, p. 1280) offers a starting point.

7. Simulation study

7.1 Simulation settings

We present a simulation study to assess the magnitude of both, the identified shortcomings of the existing predictive mean matching algorithms and our proposed improvements. To give a full picture we compare algorithm 2 to all major statistical software programs as listed by Morris et al. (2014), but Solas for technical reasons. Furthermore we compare it to two benchmark algorithms, a fully parametric one utilizing the additional information of a normal likelihood and a fully improper one that treats the maximum likelihood parameter estimates as if they were the true parameters.

For simplicity we refer to the multivariate normal setting presented above and set all off-diagonal elements of the correlation matrix equal. To recognize different challenges in real-world applications we set up a full factorial design considering the following four binary factors. We distinguish missing always completely at random versus missing always

[2] k-nearest-neighbor selection means that the drawing probability for the k nearest donors is k^{-1}, and zero for all others.

			950‰ confidence interval coverages			
			$n_{obs} = 10$		$n_{obs} = 200$	
Ref	Software	Predictive mean matching command	$\hat{\mu}$	$\hat{\beta}_1$	$\hat{\mu}$	$\hat{\beta}_1$
Proposed algorithm (algorithm 2)						
1	*R::mice*	`method="midastouch"`	936	961	945	955
2		with correction factor ϕ	973	–	972	–
3	*R::mice*	`method="midastouch",kappa=3`	931	961	946	945
4		with correction factor ϕ	960	–	978	–
Predictive mean matching software listed by Morris et al. (2014, p. 3)						
5	*R::mice*	`method="pmm"`	605	899	941	959
6	*R::Hmisc*	`aregImpute`	515	872	936	959
7	*R::BaBooN*	`BBPMM`	686	781	937	958
8	*R::mi*	`.pmm`	573	664	908	913
9	*SAS::proc mi*	`regpmm`	487	841	928	943
10	*SAS::MIDAS*	`MIDAS`	899	967	937	954
11	*SPSS*	`multiple imputation /impute scalemodel=PMM`	640	659	907	911
12	*Stata*	`mi impute pmm`	616	652	907	911
13	*Stata*	`ice, match`	443	727	935	958
Benchmark algorithms						
14	*R::mice*	parametric: `method="norm"`	962	959	946	958
15	*R*	Fully ignoring between-variance	382	468	877	912

Table 2: 1-8,14,15, R Core Team (2015); 5,14, van Buuren & Groothuis-Oudshoorn (2011); 6, Harrell (2015), unable to cope with small sample sizes; 7, Meinfelder & Schnapp (2015); 8, Gelman & Hill (2011); 9,10, SAS Institute Inc. (2015); 10, Siddique & Harel (2009); 11, IBM Corp. (2015); 12,13, StataCorp. (2015); 13, Royston & White (2011).

at random (Mealli & Rubin, 2015, p. 998) and define the latter as $pr(y = missing) = \Phi\left[(1/4)\left\{X_1 + N(0,3)\right\}\right]$; $p-1 = 1$ covariate versus $p-1 = 8$ covariates; $R^2 = 0$ versus $R^2 = 0.75$; and $n_{obs} = 10$ versus $n_{obs} = 200$. Furthermore, we fix $M = 25$, $n_{mis} = 100$, all marginal means at zero, all marginal variances at one and the number of Monte Carlo simulations at $n_{sim} = 250$ for each combination.

7.2 Simulation results

We focus on the estimates of both, the mean of y, denoted as $\hat{\mu}$, and the regression coefficient of X_1 in the linear regression model of y on X^*, denoted as $\hat{\beta}_1$. Utilizing the multiple imputation variance estimator by Rubin (1987), which is unbiased in our setting (Yang & Kim, 2016, p. 246), and the appropriate degrees of freedom (Barnard & Rubin, 1999) we construct 95% frequentist confidence intervals. For each simulation run we note whether or not such a confidence interval covers the true parameter value. We present the key results in table 7.1 and the details in the supplementary material. For each cell in table 7.1 we average the coverages over $2^{(4-1)}n_{sim} = 2000$ simulation runs.

The most striking result is that the MIDAS algorithm by Siddique & Belin (2008) and Siddique & Harel (2009) outperforms all predictive mean matching algorithms of the major statistical software programs. Its advantage is especially large when the the uncertainty

about the imputation model parameters is considerable, i.e. when the number of donors is small and thus diminishes when the number of donors increases. This result strongly supports the findings in section 4.

With one predictor only, i.e. $p - 1 = 1$, some algorithms perform as badly as the bad benchmark that does not propagate parameter uncertainty at all. All these algorithms, rows 8, 11, 12 in table 7.1, rely on both, type-2 matching and the deterministic hot-deck. This attenuation bias buttresses the criticism by van Buuren (2012). Although the MIDAS algorithm by Siddique & Belin (2008) and Siddique & Harel (2009) involves type-2 matching it outperforms the bad benchmark.

7.3 The proposed algorithm

Especially the results for the small donor sample size $n_{obs} = 10$ indicate that our proposed touching up of the MIDAS algorithm by Siddique & Belin (2008) and Siddique & Harel (2009) leads to a considerable improvement. This seems to be true for all means, more specifically, the out-of-sample predictions for the donors from proposition 1, compare row 10 to row 3; the modified closeness parameter from equation (3), compare row 3 to row 1; and the application of the correction factor ϕ from equations (1) and (4), compare rows 1 and 3 to rows 2 and 4, all comparisons in table 7.1.

Acknowledgement

Many of the presented findings are elaborated in more detail in the 2017 Bamberg University PhD thesis by the first author. The authors are grateful to Trivellore Raghunathan, Juned Siddique, and Donald Rubin for their valuable comments, which have helped to improve this paper. And most of all we wish to thank Susanne Raessler who has guided us the way to the field and who will always be in our hearts.

REFERENCES

ANDRIDGE, R. R. & LITTLE, R. J. A. (2010). A review of hot deck imputation for survey non-response. *Int. Statist. Rev.* **78**, 40–64.

BARNARD, J. & RUBIN, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika* **86**, 948–55.

BOSCH, V. (2005). A generalized measure of effective sample size. Technical report, GfK AG, Nuremberg, Germany.

DAVISON, A. C., & HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press,.

DEMIRTAS, H., ARGUELLES, L. M., CHUNG, H. & HEDEKER, D. (2007). On the performance of bias-reduction techniques for variance estimation in approximate Bayesian bootstrap imputation. *Comp. Statist. Data Anal.* **51**, 4064–68.

GELMAN, A. & HILL, J. (2011). Opening windows to the black box. R package version 1.0. *J. Statist. Software* **40**, 1–31.

GREENBERG, E. (2013). *Introduction to Bayesian Econometrics*. Cambridge: Cambridge University Press, 2nd ed.

HARRELL, F. E. (2015). *Hmisc: Harrell Miscellaneous*. R package version 3.16-0.

HEITJAN, D. F. & LITTLE, R. J. A. (1991). Multiple imputation for the fatal accident reporting system. *J. R. Statist. Soc.* C **40**, 13–29.

IBM CORP. (2015). *IBM SPSS Statistics for Windows, version 23.0*. Armonk, NY: IBM Corp.

KIM, J. K. (2002). A note on approximate Bayesian bootstrap imputation. *Biometrika* **89**, 470–77.

KISH, L. (1965). *Survey Sampling*. New York: Wiley.

LILLARD, L., SMITH, J. P. & WELCH, F. (1982). What do we really know about wages? The importance of nonreporting and census imputation. Technical report, Rand Corporation, Santa Monica, CA.

LITTLE, J. D. C. (1970). Models and managers: The concept of a decision calculus. *Management science* **16**, B466–85.

LITTLE, R. J. A. (1988). Missing-data adjustments in large surveys. *J. Bus. Econ. Statist.* **6**, 287–96.

LITTLE, R. J. A. & RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*. Hoboken, NJ: Wiley, 2nd ed.

MEALLI, F. & RUBIN, D. B. (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika* **102**, 995–1000.

MEINFELDER, F. & SCHNAPP, T. (2015). *BaBooN Bayesian Bootstrap Predictive Mean Matching*. R package version 0.2-0.

MORRIS, T. P., WHITE, I. R. & ROYSTON, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC medical research methodology* **14**, 1–13.

PARZEN, M., LIPSITZ, S. R. & FITZMAURICE, G. M. (2005). A note on reducing the bias of the approximate Bayesian bootstrap imputation variance estimator. *Biometrika* **92**, 971–74.

R CORE TEAM (2015). *A Language and Environment for Statistical Computing, version 3.2.2*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, https://www.R-project.org.

ROYSTON, P. & WHITE, I. R. (2011). Multiple imputation by chained equations (MICE): implementation in Stata. *J. Statist. Software* **45**, 1–20.

RUBIN, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *J. Bus. Econ. Statist.* **4**, 87–94.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

RUBIN, D. B. & SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J. Am. Statist. Assoc.* **81**, 366–74.

SAS INSTITUTE INC. (2015). *SAS software University Edition, version 9.4*. Cary, NC: SAS Institute Inc.

SCHENKER, N. & TAYLOR, J. M. G. (1996). Partially parametric techniques for multiple imputation. *Comp. Statist. Data Anal.* **22**, 425–46.

SIDDIQUE, J. & BELIN, T. R. (2008). Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statist. Med.* **27**, 83–102.

SIDDIQUE, J. & HAREL, O. (2009). MIDAS: a SAS macro for multiple imputation using distance-aided selection of donors. *J. Statist. Software* **29**, 1–18.

STATACORP. (2015). *Stata Statistical Software: Release 14*. College Station, TX: StataCorp LP.

VAN BUUREN, S. (2012). *Flexible Imputation of Missing Data*. Boca Raton, FL: Chapman & Hall/CRC.

VAN BUUREN, S. & GROOTHUIS-OUDSHOORN, K. (2011). mice: Multivariate Imputation by Chained Equations in R. R package version 2.22, *J. Statist. Software* **45**, 1–67.

WU, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14**, 1261–95.

YANG, S. & KIM, J. K. (2016). A note on multiple imputation for method of moments estimation. *Biometrika* **103**, 244–51.

Supplement

	match types		k-nearest-neighbor		parameter	predictions of	
Ref.	available	specify by	default	specify by	uncertainty	donors	recipients
1-4	2	-	n_{obs}	-	ABB	o.o.s.	o.o.s.
5	1	-	5	`donors=#`	parametric	i.s.	o.o.s.
6	1, 2, 3	`pmmtype=#`	3	`kclosest=#`	bootstrap	i.s.	o.o.s.
7	2	-	1	-	BB	i.s.	i.s.
8	2	-	1	-	parametric	i.s.	i.s.
9	1	-	5	-	parametric	i.s.	o.o.s.
10	2	-	n_{obs}	-	ABB	i.s.	o.o.s.
11	2	-	1	-	parametric	i.s.	o.o.s.
12	2	-	1	`knn(#)`	parametric	i.s.	o.o.s.
13	1, 2, 3	`matchtype=#`	3	`matchpool(#)`	parametric	i.s.	o.o.s.

Table 3: Characteristics of existing PMM software implementations (Morris et al., 2014, p. 3). The references (Ref.) refer to table 7.1. Abbreviations are: approximate Bayesian bootstrap (ABB), Bayesian bootstrap (BB), in sample (i.s.), and out of sample (o.o.s).

	950‰ confidence interval coverages													
	Response mechanism				Number of covariates				Coefficient of determination				Overall	
	MACAR		MAAR		$p-1=1$		$p-1=8$		$R^2=0$		$R^2=0.75$			
Ref.	$\hat{\mu}$	$\hat{\beta}_1$	$\hat{\mu}$	$\hat{\beta}_1$	$\hat{\mu}$	$\hat{\beta}_1$	$\hat{\mu}$	$\hat{\beta}_1$	$\hat{\mu}$	$\hat{\beta}_1$	$\hat{\mu}$	$\hat{\beta}_1$	$\hat{\mu}$	$\hat{\beta}_1$
Proposed algorithm (algorithm 2)														
1	967	960	905	962	948	958	924	964	976	962	896	960	936	961
2	991	–	955	–	972	–	974	–	988	–	958	–	973	–
3	968	959	894	963	951	961	911	961	970	950	892	972	931	961
4	985	–	934	–	963	–	956	–	990	–	929	–	960	–
Predictive mean matching software listed by Morris et al. (2014, p. 3)														
5	732	910	477	887	598	836	611	961	700	950	509	847	605	899
6	587	900	442	844	464	776	565	968	564	934	465	810	515	872
7	771	794	600	768	658	650	714	912	764	831	607	731	686	781
8	704	685	442	642	396	351	750	976	583	639	564	688	573	664
9	616	840	358	841	436	724	539	957	557	855	418	827	487	841
10	960	971	838	963	873	953	925	981	954	957	844	977	899	967
11	718	675	561	643	396	352	883	966	604	631	675	687	640	659
12	704	667	528	637	396	351	836	953	583	624	650	680	616	652
13	579	731	309	723	446	500	440	954	575	978	312	476	443	727
Benchmark algorithms														
14	964	956	960	961	970	946	954	971	962	948	962	969	962	959
15	479	469	285	466	396	351	367	585	313	438	451	498	382	468

Table 4: Coverages for $n_{obs} = 10$ split by the three remaining binary factors.

	950‰ confidence interval coverages													
	Response mechanism				Number of covariates				Coefficient of determination				Overall	
	MACAR		MAAR		$p-1=1$		$p-1=8$		$R^2=0$		$R^2=0.75$			
Ref.	$\hat{\mu}$	$\hat{\beta}_1$	$\hat{\mu}$	$\hat{\beta}_1$	$\hat{\mu}$	$\hat{\beta}_1$	$\hat{\mu}$	$\hat{\beta}_1$	$\hat{\mu}$	$\hat{\beta}_1$	$\hat{\mu}$	$\hat{\beta}_1$	$\hat{\mu}$	$\hat{\beta}_1$
Proposed algorithm (algorithm 2)														
1	948	960	942	950	942	954	948	956	953	959	936	951	945	955
2	974	–	969	–	966	–	977	–	960	–	983	–	972	–
3	952	949	940	942	948	934	944	956	957	933	935	957	946	945
4	982	–	973	–	975	–	980	–	987	–	968	–	978	–
Predictive mean matching software listed by Morris et al. (2014, p. 3)														
5	947	965	934	952	936	961	945	956	942	959	940	958	941	959
6	939	967	932	950	921	961	950	956	930	960	942	957	936	959
7	940	966	933	949	933	959	940	956	939	962	934	953	937	958
8	908	929	907	897	874	866	942	960	895	902	920	924	908	913
9	931	953	925	932	918	933	938	952	927	935	929	950	928	943
10	939	959	934	949	930	952	943	956	940	946	933	962	937	954
11	907	927	906	895	874	866	939	956	892	901	921	921	907	911
12	908	927	906	895	874	866	940	956	893	902	921	920	907	911
13	943	966	926	950	932	959	937	957	936	960	933	956	935	958
Benchmark algorithms														
14	950	966	942	949	945	959	947	956	951	957	942	958	946	958
15	886	927	868	896	875	866	879	957	843	903	911	920	877	912

Table 5: Coverages for $n_{obs} = 200$ split by the three remaining binary factors.