Nonparametric Markov chain bootstrap for multiple imputation

Li-Chun Zhang∗
Statistics Norway Kongensg. 6, P.O. Box 8131 Dep., 0033 Oslo, Norway

Received 28 June 2001; received in revised form 28 September 2002

Abstract

Multiple imputation is a statistical method for analyzing data with missing values. Nonparametric Markov chain bootstrap methods can be used to generate multiple imputations of both scalar and multivariate outcome variables, under the assumption that the data are missing completely at random, and nonparametric inference can be obtained using multiple implementation bootstrap. The nonparametric approach is useful when parametric settings are inappropriate or difficult. An extension of the Markov chain bootstrap method is discussed under a more complex nonresponse assumption.

© 2003 Elsevier B.V. All rights reserved.

Keywords: Bootstrap; Markov chain; Confidence density; Analysis of variance model

1. Introduction

Nonparametric Markov chain bootstrap methods for multiple imputation of scalar as well as multivariate outcome variables are proposed. Multiple imputation (Rubin, 1987) is most easily motivated from the Bayesian perspective. Let \( x_{\text{obs}} \) be the set of observed values and \( x_{\text{mis}} \) the set of missing values. The posterior density of a population quantity, denoted by \( \theta \), is given by

\[
p(\theta|x_{\text{obs}}) = \int_{X_{\text{mis}}} p(\theta|x_{\text{obs}}, x_{\text{mis}}) f(x_{\text{mis}}|x_{\text{obs}}) \, dx_{\text{mis}},
\]

∗ Tel.: +47-21-09-00-00; fax: +47-21-09-49-73.
E-mail address: lcz@ssb.no (L.-C. Zhang).
where \( p(\theta|x_{\text{obs}},x_{\text{mis}}) \) is the complete-data posterior density of \( \theta \), and \( f(x_{\text{mis}}|x_{\text{obs}}) \) the predictive density of the missing values. In a fully parametric setting with parameter vector \( \eta \), we have

\[
f(x_{\text{mis}}|x_{\text{obs}}) = \int f(x_{\text{mis}}|\eta,x_{\text{obs}}) p(\eta|x_{\text{obs}}) \, d\eta,
\]

where \( f(x_{\text{mis}}|\eta,x_{\text{obs}}) \) is the conditional density of \( x_{\text{mis}} \), and \( p(\eta|x_{\text{obs}}) \) is the posterior density of \( \eta \) given \( x_{\text{obs}} \). Multiple imputations are independent draws of \( X_{\text{mis}} \) from \( f(x_{\text{mis}}|x_{\text{obs}}) \). When direct generation of \( x_{\text{mis}} \) from \( f(x_{\text{mis}}|x_{\text{obs}}) \) is difficult, Tanner and Wong (1987) proposed data augmentation, which is an iterative algorithm related to Gibbs’ sampler.

It is possible to make nonparametric inference based on the multiple imputations, without resorting to the parametric posterior density \( p(\theta|x_{\text{obs}},x_{\text{mis}}) \). Let \( \hat{\theta}(x) \) be some well-defined complete-data nonparametric estimator, where \( x = (x_{\text{obs}},x_{\text{mis}}) \). Denote by \( x_{\text{mis}}^{(1)}, \ldots, x_{\text{mis}}^{(M)} \) a set of \( M \) multiple imputations. The multiple imputation estimator of \( \theta \) is given by

\[
\hat{\theta} = M^{-1} \sum_{m=1}^{M} \hat{\theta}(x^{(m)}), \quad \text{where } x^{(m)} = (x_{\text{obs}},x_{\text{mis}}^{(m)}).
\] (1)

More generally, the confidence density for a scalar \( \theta \) given \( x \) is defined as

\[
\pi(\theta|x) = \{ \partial \theta_{x}(z)/\partial z \}^{-1},
\]

where \( \theta_{x}(z) \) is the one sided level-\( z \) confidence limit of \( \theta \) based on \( x \). The 100\( z \)th percentile of \( \pi(\theta|x) \) is by definition \( \theta_{x}(z) \), so that the confidence density summarizes the one sided confidence limits of \( \theta \) in a density function. Efron (1993) shows how \( \pi(\theta|x) \) can be easily derived from the nonparametric bootstrap ABC method of confidence limits, and that \( \pi(\theta|x) \) is a good approximation to the posterior density of \( p(\theta|x) \), with an appropriate uninformative prior for the parameter vector \( \eta \). Efron (1994) defines multiple implementation bootstrap in analogy to parametric inference. Let \( \pi(\theta|x^{(1)}), \ldots, \pi(\theta|x^{(M)}) \) be the \( M \) confidence densities derived from \( x^{(1)}, \ldots, x^{(M)} \), respectively. The multiple imputation confidence density is given by

\[
\tilde{\pi}(\theta) = M^{-1} \sum_{m=1}^{M} \pi(\theta|x^{(m)}), \quad \text{where } x^{(m)} = (x_{\text{obs}},x_{\text{mis}}^{(m)}).
\] (2)

Nonparametric inference can be useful in situations where it is inappropriate or difficult to set up a fully parametric model \( f(x|\eta) \). To complete the analogy to the parametric setting, we need suitable nonparametric predictive density \( f(x_{\text{mis}}|x_{\text{obs}}) \). We propose a general approach of constructing nonparametric Markov chains targeted at the predictive density of \( X_{\text{mis}} \) under the MCAR-assumption, i.e. missing completely at random (Little and Rubin, 1987). For a scalar variable, the basic idea is as follows. Let \( X = (X_1, \ldots, X_n) \) be an independent and identically distributed (iid) sample. Let \( r_i = 1 \) indicate that \( x_i \) is observed, and \( r_i = 0 \) that \( x_i \) is missing. Let \( \mu \) be the expectation of \( X_i \) for \( i = 1, \ldots, n \). The MCAR-assumption implies that the residual \( \hat{x}_i = x_i - \mu \) has the same distribution whether \( x_i \) is missing or not. Any initial vector of \( x_{\text{mis}} \) yields an imputed
complete-data mean, denoted by $\bar{x}^*$, and correspondingly $\hat{x}_i = x_i - \bar{x}^*$ if $r_i = 1$. We update $x_{mis}$ by $\{x_i^*; r_i = 0\}$, where $x_i^* = \bar{x}^* + \hat{e}_i$ and $\hat{e}_i$ is drawn with equal probability and replacement from $\{\hat{e}_i; r_i = 1\}$. Iterations of this generate a nonparametric Markov chain of $X_{mis}$. On convergence, the marginal distribution of $X_{mis}$ is by construction an estimate of $f(x_{mis}|x_{obs})$. The same idea can be applied to multivariate data under the two-way analysis of variance (ANOVA) model. For $J$-dimensional data, with $i=1,\ldots,n$ and $j=1,\ldots,J$, we assume that $x_{ij} = \mu_{ij} + \alpha_i + \beta_j$ where $\sum_i \alpha_i = \sum_j \beta_j = 0$, and $x_{ij} = \mu_{ij} + \hat{\epsilon}_{ij}$ where the $\hat{\epsilon}_{ij}$'s are iid with zero mean. The MCAR-assumption implies that the distribution of $\hat{\epsilon}_{ij}$ remains the same whether $x_{ij}$ is missing or not. We therefore generate the Markov chain by repeated bootstrap resampling of $\{\hat{\epsilon}_{ij}; r_{ij} = 0\}$ from $\{\hat{\epsilon}_{ij}; r_{ij} = 1\}$.

The rest of the paper is organized as follows. In Section 2 we examine several Markov chain bootstrap methods for a scalar outcome variable, both with and without covariates. Conditions are given for the convergence of the Markov chains. It is shown that the proposed methods have similar properties as existing nonparametric methods, such as the so-called hot-deck and approximate Bayesian bootstrap (ABB) methods (Rubin and Schenker, 1986). In Section 3 we consider a method for the multivariate data under the ANOVA model, where both the hot-deck and ABB methods run into difficulties. We have not theoretically established the sufficient conditions for the convergence under the ANOVA model, but convergence can be assessed based on the Markov chains actually generated (Brooks and Roberts, 1998). The proposed method are illustrated using the Student Score Data. The results show both similarities and differences to those obtained by Efron (1994). In Section 4 we discuss an extension of the proposed methodology under a nonresponse assumption similar to that behind predictive mean matching (PPM) and local residual draw (LRD) imputation (e.g. Schenker and Taylor, 1996). This can be used for sensitivity analysis of a possible departure from the MCAR-assumption.

2. Methods for a scalar outcome variable

All techniques are described for single imputation of $X_{mis}$. Multiple imputations are obtained by independent applications of the same technique. It turns out that the Markov chain bootstrap methods have properties similar to existing methods, providing theoretical justification in the simple settings considered here. However, since the existing methods have the advantage of being noniterative, there is no need in practice to replace them by the new methods.

2.1. Simple random sample without covariates

Let $X=(X_1,\ldots,X_n)$ be an iid sample with mean $\mu$. Suppose that, due to nonresponse, only $n_1$ of the $n$ values are observed. Let $r_i$ be the nonresponse indicator of $x_i$, such that $\sum_{i=1}^n r_i = n_1$ and $\sum_{i=1}^n (1 - r_i) = n_0 = n - n_1$. The hot-deck method draws the $n_0$ components of $X_{mis}$ with equal probability and replacement from the $n_1$ values of $x_{obs}$. Because the hot-deck treats the distribution $f(x_{mis}|x_{obs})$ as known, the predictive
uncertainty in $X_{\text{mis}}$ is somewhat underestimated. Rubin and Schenker (1986) suggested the ABB method, drawing first a bootstrap sample of $n_1$ components from $x_{\text{obs}}$ with equal probability and replacement, denoted by $x_{\text{obs}}^*$. The $n_0$ values of $X_{\text{mis}}$ are then generated by the hot-deck method from $x_{\text{obs}}^*$. Let $s^2_r = \sum_{i=1}^n r_i(x_i - \bar{x}_r)^2/(n_1 - 1)$ be the observed sample variance, where $\bar{x}_r = \sum_{i=1}^n r_i x_i/n_1$ is the observed sample mean. Let $\bar{x}_m$ and $s_m^2$ be the sample mean and variance of the imputed values of $X_{\text{mis}}$. Let $E_s$ and $V_s$ denote expectation and variance with respect to the imputation method. Under the hot-deck method, we have, with $\tilde{s}_r^2 = (n_1 - 1)s^2_r/n_1$,

$$E_s[\bar{x}_m] = \bar{x}_r \quad \text{and} \quad V_s[\bar{x}_m] = \tilde{s}_r^2/n_0 \quad \text{and} \quad E_s[s_m^2] = \tilde{s}_r^2.$$  

Whereas, under the ABB method (Rubin and Schenker, 1986)

$$E_s[\bar{x}_m] = \bar{x}_r \quad \text{and} \quad V_s[\bar{x}_m] = (n - 1)\tilde{s}_r^2/(n_0 n_1) \quad \text{and} \quad E_s[s_m^2] = (n_1 - 1)\tilde{s}_r^2/n_1.$$  

The simple random (nonparametric) Markov chain bootstrap (SRMCB) method is defined as follows. At the $k$th iteration, let $x_{\text{mis}}(k) = \{x_{\text{obs}}^*(k); r_i = 0\}$ be the current values of $X_{\text{mis}}$ with mean $\bar{x}_m(k)$. Let $\tilde{x}_r(k) = (n_1 \bar{x}_r + n_0 \bar{x}_m(k))/n$ be the corresponding imputed sample mean. Let $\hat{e}_{i(k)} = x_i - \tilde{x}_r(k)$ be the estimated residual if $r_i = 1$. We update $x_{\text{obs}}^*(k+1)$ by

$$x_{\text{obs}}^*(k+1) = \tilde{x}_r(k) + \hat{e}_{i(k+1)}^*,$$

where $\hat{e}_{i(k+1)}^*$ is drawn with equal probability and replacement from $\{\hat{e}_{i(k)}; r_i = 1\}$. Since the estimated residuals for the missing values by construction have the same distribution as that of the observed values, the equilibrium marginal distribution of the Markov chain provides an estimate of $f(x_{\text{mis}}|x_{\text{obs}})$ under the MCAR-assumption. Let $s^2_{r(k)} = \sum_{i=1}^n r_i \hat{e}_{i(k)}^2/(n_1 - 1)$. In Appendix A it is shown that, for any $k > 0$ we have $s^2_{r(k)} > s^2_r$, and

$$E_s[\tilde{x}_r(k)] = E_s[\bar{x}_m(k)] = \bar{x}_r \quad \text{and} \quad \lim_{k \to \infty} E_s[s^2_{r(k)}] = s^2_r \left(1 + \frac{n_0}{n^2 - n_0}\right),$$

and

$$\lim_{k \to \infty} V_s[\bar{x}_m(k)] = \tilde{s}_r^2 \left(1 + \frac{1}{n_0 n^2 - n_0}\right) \quad \text{and} \quad \lim_{k \to \infty} E_s[s^2_m(k)] = s^2_r \left(1 + \frac{n_0}{n^2 - n_0}\right).$$

The SRMCB method is equivalent to the existing methods in terms of the imputed sample mean. For any sample of finite size, the variance of the SRMCB method is always larger than that of the hot-deck method, i.e. $V_s[\bar{x}_m(k)] > \tilde{s}_r^2/n_0$. The difference, however, is negligible for large samples. To compensate for the underestimation of the predictive variability, we may apply the approximate Bayesian Markov chain bootstrap (ABMCB). That is, we add an extra bootstrap step to obtain $x_{\text{obs}}^*$ from $x_{\text{obs}}$, and run the Markov chain based on $x_{\text{obs}}^*$ instead. It follows that, for large samples, the ABMCB method is equivalent to the ABB method.

2.2. Linear regression with completely observed covariates

Let $y_i$ be associated with covariates $x_i$ for $i = 1, \ldots, n$. We assume that the covariates are completely observed. Suppose that, due to nonresponse, only $n_1$ of the $n$
values of \( y = (y_1, \ldots, y_n) \) are observed. Let \( r_i \) be the nonresponse indicator for \( y_i \). Let 
\[
\beta_r = (\sum_{i=1}^{n} r_i x_i \beta_T)^{-1} (\sum_{i=1}^{n} r_i x_i y_i)
\]
be the least-square fit of the regression coefficients based on the observed data. We assume that the linear regression always contains the intercept, such that \( \hat{y}_i = \sum_{i=1}^{n} r_i \beta_T / n_1 = 0 \) where \( \epsilon = y_i - x_i^T \beta \) if \( r_i = 1 \). Let \( \beta^* = (\sum_{i=1}^{n} x_i^T)^{-1} (\sum_{i=1}^{n} r_i x_i y_i + \sum_{i=1}^{n} (1 - r_i) x_i y_i^*) \) be the least-square fit of the regression coefficients based on the imputed data. Let \( s_r^2 = \sum_{i=1}^{n} r_i \epsilon_i^2 / (n_1 - 1) \) be the sample variance of \( \{\epsilon_i; r_i = 1\} \), and 
\[
s_m^2 = \sum_{i=1}^{n} (1 - r_i) (\epsilon_i^*)^2 / (n_0 - 1).
\]
Then, with \( s_r^2 = (n_1 - 1)s_r^2/n_1 \),
\[
E_s[y_i^*] = x_i^T \beta_r \quad \text{and} \quad E_s[\beta^*] = \beta_r \quad \text{and} \quad E_s[s_m^2] = s_r^2.
\]
For the approximate Bayesian regression imputation method, an extra bootstrap step is added before applying the regression imputation method.

We define the regression SRMCMC method as follows. Let \( A_r = \sum_{i=1}^{n} r_i x_i \pi_i \) and 
\[
A_m = \sum_{i=1}^{n} (1 - r_i) x_i \pi_i \quad \text{and} \quad A = A_r + A_m.
\]
At the \( k \)th iteration, let \( y_i^{* (k)} \) be the current imputed value if \( r_i = 0 \). The corresponding imputed least-square fit is given by 
\[
\hat{\beta}(k) = A^{-1} (\sum_{i=1}^{n} r_i x_i y_i + \sum_{i=1}^{n} (1 - r_i) x_i y_i^{* (k)}).
\]
Let \( \hat{\epsilon}(k) = y_i - x_i^T \hat{\beta}(k) \) if \( r_i = 1 \). We update \( y_i^{* (k)} \) where \( r_i = 0 \) by
\[
y_i^{* (k+1)} = x_i^T \hat{\beta}(k) + \epsilon_i^{* (k+1)},
\]
where \( \epsilon_i^{* (k+1)} \) is drawn with equal probability and replacement from \( \{\hat{\epsilon}(k), r_i = 1\} \). It is shown in Appendix B that
\[
\lim_{k \to \infty} E_s[\hat{\beta}(k)] = \beta_r \quad \text{and} \quad E_s[s_m^2] = s_r^2 \left[ 1 + O \left( \frac{n_0}{n^2} \right) \right],
\]
provided the matrix \( A^{-1} (A_m - n_0 \bar{x}_m \bar{x}_m^T) \) is convergent.

The regression SRMCMC method is thus equivalent to the regression imputation method for large samples. For any sample of fixed size, however, the variability of the regression SRMCMC method is always larger than that of the regression imputation method, in the sense of \( s_r^2 > s_r^* \) for any \( k \) (Appendix B). Again, we obtain the regression ABMCMC method if we add an extra bootstrap of \( \{(y_i, x_i^T); r_i = 1\} \) before we start the Markov chain. For large samples, the regression ABMCMC method is equivalent to the approximate Bayesian regression imputation method.

3. Multivariate data under the ANOVA model

Consider a sample of \( J \)-variate observations, possibly after some transformation, arranged in a matrix \( (x_{ij}) \), where \( i = 1, \ldots, n \) and \( j = 1, \ldots, J \), assumed to satisfy a two-way ANOVA structure
\[
x_{ij} = \mu_{ij} + \epsilon_{ij}, \quad \text{where} \quad \mu_{ij} = \nu + \alpha_i + \beta_j \quad \text{and} \quad \sum_{i=1}^{n} \alpha_i = \sum_{j=1}^{J} \beta_j = 0.
\]
The \( e_{ij} \)'s are assumed to be iid with zero mean. Since we do not specify any parametric error distribution, the parametric scheme of multiple imputation described at the beginning of Section 1 is not applicable. Let \( r_{ij} \) be the nonresponse indicator of \( x_{ij} \). The MCAR-assumption implies that the distribution of \( e_{ij} \) remains the same whether \( x_{ij} \) is missing or not.

The hot-deck and ABB methods described in the previous section run into difficulties in this case. Both can be applied separately in each column of the data matrix. That is, the missing values in one column are drawn from the observed values in the same column. However, this is appropriate only if the columns are independent of each other. Otherwise, it would lead to biased estimation of any quantity which involves more than one column of the data matrix, such as the covariance matrix of the data.

We define the multivariate SRMCB method for the two-way ANOVA case as follows. At the \( k \)th iteration, let \( x_{ij(k)}^* \) be the current imputed value of \( x_{ij} \) if \( r_{ij} = 0 \). Let \((v_{ij(k)}, z_{ij(k)}, \beta_{ij(k)}) \) be the least-square fit based on \( \{x_{ij}; r_{ij} = 1\} \) and \( \{x_{ij}^*; r_{ij} = 0\} \). That is, \((v_{ij(k)}, z_{ij(k)}, \beta_{ij(k)}) \) minimize the sum of squares \( \sum_{i,j} r_{ij}(x_{ij} - \mu_{ij})^2 + \sum_{i,j} (1 - r_{ij})(x_{ij}^* - \mu_{ij})^2 \) under the restriction of \( \sum_{j} z_{ij} = \sum_{j} \beta_{ij} = 0 \). Let \( \hat{e}_{ij(k)} = x_{ij} - \mu_{ij(k)} \) if \( r_{ij} = 1 \), where \( \mu_{ij(k)} = v_{ij(k)} + z_{ij(k)} + \beta_{ij(k)} \). We update \( x_{ij(k+1)}^* \) if \( r_{ij} = 0 \) by

\[
x_{ij(k+1)}^* = \mu_{ij(k)} + \hat{e}_{ij(k+1)}^*,
\]

where \( \hat{e}_{ij(k+1)}^* \) is drawn with equal probability and replacement from \( \{\hat{e}_{ij(k)}; r_{ij} = 1\} \). Convergence of the Markov chain can be assessed based on empirical evidences just like the Markov chains under parametric settings. By construction, the equilibrium marginal distribution of \( \{x_{ij}; r_{ij} = 0\} \) yields an estimate of \( f(x_{\text{mis}}|x_{\text{obs}}) \). This gives us the multivariate analogy of the scalar SRMCB method. Again, to compensate for the underestimation of the predictive variability, we may apply the multivariate ABMCMC method, where we add an extra bootstrap step of \( \{(x_{ij}^T, r_i^T); i = 1, \ldots, n\} \) before starting the Markov chain.

Consider the Student Score Data (Table 1), where the scores in the parentheses are assumed to be missing at random (Efron, 1994). Let \( x_{ij} \) be the scores of the \( i \)th student, for \( i = 1, \ldots, 22 \) and \( j = 1, \ldots, 5 \). Suppose we are interested in estimating the eigenvalues \( \theta \) of the population covariance matrix of \( x_i \). Efron (1994) compares two methods. The first one is based on an ad hoc nonparametric estimate, denoted by \((\hat{v}, \hat{z}, \hat{\beta})\), obtained from minimizing the ‘observed’ sum of squares \( \sum_{i,j} r_{ij}(x_{ij} - \mu_{ij})^2 \) under the restriction of \( \sum_{j} x_{ij} = \sum_{j} \beta_{ij} = 0 \). This yields the imputed values \( \hat{x}_{ij} = \hat{v} + \hat{z}_i + \hat{\beta}_j \) where \( r_{ij} = 0 \), and the empirical covariance matrix \( \hat{\Sigma} \) of \( \{x_{ij}; r_{ij} = 1\} \) and \( \{\hat{x}_{ij}; r_{ij} = 0\} \). An estimate \( \hat{\theta} \) of \( \theta \) is given by the eigenvalues of \( \hat{\Sigma} \). The second method is based on the multiple implementation bootstrap (2), where the multiple imputations are generated using a semiparametric method. Based on each bootstrap sample of \( \{(x_{ij}^T, r_i^T); i = 1, \ldots, n\} \), we obtain as above the ad hoc estimates \( \hat{x}_{ij} \) and \( \hat{\Sigma} \). The complete data are now assumed to follow the normal distribution with expectation \( \hat{\mu} = (\sum_i r_i x_i + \sum_i (1 - r_i) \hat{x}_i)/n \) and covariance matrix \( \hat{\Sigma} \). The missing values are filled in by sampling from the corresponding conditional normal distribution. Efron (1994) finds that the confidence intervals produced by the second method “are somewhat too short in the upper direction”, and suggests the assumed normal distribution.
Table 1
The Student Score Data: 22 students have each taken 5 exams, labeled A–E. The missing A and E scores have been randomly selected, and indicated by the parentheses.

<table>
<thead>
<tr>
<th>Student</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>Student</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>63</td>
<td>65</td>
<td>70</td>
<td>63</td>
<td>12</td>
<td>42</td>
<td>60</td>
<td>54</td>
<td>49</td>
<td></td>
<td>(33)</td>
</tr>
<tr>
<td>2</td>
<td>53</td>
<td>61</td>
<td>72</td>
<td>64</td>
<td>73</td>
<td>13</td>
<td>(40)</td>
<td>63</td>
<td>53</td>
<td>54</td>
<td>(25)</td>
</tr>
<tr>
<td>3</td>
<td>51</td>
<td>67</td>
<td>65</td>
<td>65</td>
<td>(68)</td>
<td>14</td>
<td>(23)</td>
<td>55</td>
<td>59</td>
<td>53</td>
<td>(44)</td>
</tr>
<tr>
<td>4</td>
<td>(44)</td>
<td>69</td>
<td>53</td>
<td>53</td>
<td>53</td>
<td>15</td>
<td>(49)</td>
<td>49</td>
<td>45</td>
<td>48</td>
<td>(39)</td>
</tr>
<tr>
<td>5</td>
<td>(42)</td>
<td>69</td>
<td>61</td>
<td>55</td>
<td>45</td>
<td>16</td>
<td>17</td>
<td>53</td>
<td>57</td>
<td>43</td>
<td>51</td>
</tr>
<tr>
<td>6</td>
<td>(31)</td>
<td>49</td>
<td>62</td>
<td>63</td>
<td>62</td>
<td>17</td>
<td>39</td>
<td>46</td>
<td>46</td>
<td>32</td>
<td>(43)</td>
</tr>
<tr>
<td>7</td>
<td>44</td>
<td>61</td>
<td>52</td>
<td>62</td>
<td>(46)</td>
<td>18</td>
<td>48</td>
<td>38</td>
<td>41</td>
<td>44</td>
<td>33</td>
</tr>
<tr>
<td>8</td>
<td>49</td>
<td>41</td>
<td>61</td>
<td>49</td>
<td>(64)</td>
<td>19</td>
<td>46</td>
<td>40</td>
<td>47</td>
<td>29</td>
<td>(17)</td>
</tr>
<tr>
<td>9</td>
<td>30</td>
<td>69</td>
<td>50</td>
<td>52</td>
<td>45</td>
<td>20</td>
<td>30</td>
<td>34</td>
<td>43</td>
<td>46</td>
<td>18</td>
</tr>
<tr>
<td>10</td>
<td>(36)</td>
<td>59</td>
<td>51</td>
<td>45</td>
<td>51</td>
<td>21</td>
<td>(12)</td>
<td>30</td>
<td>32</td>
<td>35</td>
<td>21</td>
</tr>
<tr>
<td>11</td>
<td>(56)</td>
<td>40</td>
<td>56</td>
<td>54</td>
<td>(35)</td>
<td>22</td>
<td>(5)</td>
<td>26</td>
<td>15</td>
<td>20</td>
<td>(20)</td>
</tr>
</tbody>
</table>

Note: The 22 sets of scores are taken from Table 1.2.1 in Mardia et al. (1979).

Table 2
Estimates of the eigenvalues of the covariance matrix for the Student Score Data

<table>
<thead>
<tr>
<th>Method</th>
<th>Eigenvalue</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. 1</td>
</tr>
<tr>
<td>Based on least-square fit to the true complete data</td>
<td>650.2</td>
</tr>
<tr>
<td>$\hat{\theta}$ by the ad hoc nonparametric method</td>
<td>633.2</td>
</tr>
<tr>
<td>$\tilde{\theta}$ by the multivariate SRMCB method</td>
<td>637.9</td>
</tr>
<tr>
<td>$\tilde{\theta}$ by the weighted SRMCB method with $\delta = 0.7$</td>
<td>635.8</td>
</tr>
</tbody>
</table>

to be “a possible source of the difficulty”. We therefore only make comparison to the ad hoc nonparametric method below.

The multivariate SRMCB method is applied. The Markov chain converges quickly. We use $\{x_{ij}^{*}(100), r_{ij} = 0\}$ from 25 independent Markov chains as the multiple imputation. These give us the estimate $\hat{\theta}$ by (1) and the confidence density $\tilde{\pi}(\theta)$ by (2)—the dotted curves and lines in Fig. 1. For comparison we have also reproduced the confidence density based on the ad hoc $\hat{\theta}$. The difference between the two is very small for the maximum eigenvalue. Both the estimates and the confidence densities coincide closely with each other. The agreement is much closer than between the ad hoc method and the semiparametric multiple implementation bootstrap (Efron, 1994, Fig. 5). The results suggest (a) that $\hat{\theta}$ has small definitional bias for the maximum eigenvalue and (b) that the multivariate SRMCB method is quite satisfactory in terms of the predictive variability of $X_{mis}$. The definitional bias of the ad hoc method, however, is more noticeable for the other eigenvalues. Fig. 1 shows the details for the second largest eigenvalue. Table 2 contains the estimates of all the 5 eigenvalues. It appears that $\hat{\theta}$ is downwards biased for all of them. The SRMCB method works better than
Eigenvalue no. 1

50 100 150

Eigenvalue no. 2

500 1000 1500

Fig. 1. The Eigenvalue Problem of the Student Score Data: the largest eigenvalue (left panel) and the second largest eigenvalue (right panel). In both cases, the solid curve is the confidence density based on the ad hoc nonparametric bootstrap, and the vertical solid line marks \( \hat{\theta} \). The dotted curve \( \tilde{\pi}(\theta) \) is based on 25 imputations generated by the multivariate SRMCB method, and the vertical dotted line marks the corresponding \( \tilde{\theta} \). The dashed curve \( \tilde{\pi}(\theta) \) is based on 25 imputations generated by the weighted multivariate SRMCB method with \( \delta = 0.7 \), and the dashed vertical line marks the corresponding \( \tilde{\theta} \).

the ad hoc method, and the estimates based on the least-square fit of the true complete data are well within the 95% confidence intervals.

4. Discussion: an extension for sensitivity analysis

Several nonparametric Markov chain bootstrap methods for multiple imputation under the MCAR-assumption have been developed. Theoretical justifications are given in the case of nonresponse on a scalar outcome variable, both with and without covariates. The usefulness of the approach has been demonstrated for multivariate data under the two-way ANOVA model. We now discuss an extension under a more complex nonresponse assumption.

Take the Student Score Data. Instead of missing completely at random, suppose that the probability of missing depends on the underlying mean \( \mu_{ij} \) such that, for \( j = 1 \) or 5,

\[
f(e_{ij} = e | r_{ij} = 1, \mu_{ij} = \mu) = f(e_{ij} = e | r_{ij} = 0, \mu_{ij} = \mu) = f(e | \mu).
\]

That is, the residuals are only exchangeable conditional on the same value of \( \mu_{ij} \). This is similar to the assumption behind PPM or LRD imputation. In the linear regression setting, the PPM method draws the imputed value \( y_j^* \) from \( \{y_i; r_i = 1 \text{ and } x_i^T \hat{\beta} \approx x_j^T \hat{\beta}\} \), whereas the LRD method imputes \( y_j^* = x_j^T \hat{\beta} + e_j^* \) where \( e_j^* \) is drawn from \( \{e_i; r_i = 1 \text{ and } x_i^T \hat{\beta} \approx x_j^T \hat{\beta}\} \). See Schenker and Taylor (1996) for more discussions.
We have experimented with a scheme motivated by the following smoother:

\[
f(\varepsilon_{ij} | \mu_{ij}) = \int_{\mu} w(\mu - \mu_{ij}) f(\varepsilon | \mu) \, d\mu, \quad \text{where} \quad \int_{\mu} w(\mu - \mu_{ij}) \, d\mu = 1.
\]

At the \(k\)th iteration, we calculate as before \(\hat{\varepsilon}_{ij(k)}\) if \(r_{ij} = 1\). We update \(x^*_{ij(k+1)} = \mu_{ij(k)} + \hat{\varepsilon}^*_{ij(k+1)}\) if \(r_{ij} = 0\), where \(\hat{\varepsilon}^*_{ij(k+1)}\) is drawn from \(\{\hat{\varepsilon}_{ij(k)}; r_{ij} = 1\}\) with the probability

\[
P[\hat{\varepsilon}^*_{ij(k+1)} = \hat{\varepsilon}_{st(k)}] = \frac{g(\hat{\varepsilon}_{ij(k)} - \mu_{st(k)})}{\sum_{p,q;rpq=1} g(\hat{\varepsilon}_{ij(k)} - \mu_{pq(k)})}.
\]

where \(g(z) = \delta |z|\) and \(0 < \delta \leq 1\), i.e. the weighted multivariate SRMCB method. With \(\delta = 1\), this is the multivariate SRMCB method. With \(\delta < 1\) higher probability is assigned to points with predictive means closer to the one being imputed. Points with predictive means far away will only have a negligible chance of being selected.

The weighted SRMCB can practically be considered as a nonparametric LRD method, with the SRMCB of Section 3 as its global version. There are of course more elaborate ways of drawing the residuals locally. For instance, Schenker and Taylor (1996) develop an adaptive method. We notice that, on the one hand, the number of possible donors vary from one missing value to another with the adaptive method, whereas all the observed points are possible donors with the simple scheme here. On the other hand, for their application, Schenker and Taylor (1996) set the adaptive method such that all possible donors have the same selection probability, whereas here this probability depends on the difference between the predictive means of the missing value and a possible donor. In the small sample situation of the Student Score Data we feel that it is unlikely that these two ways of drawing the residuals will give very different results. We have therefore only used our simple scheme for the illustration below.

The dashed curve and line in Fig. 1 and the last row of Table 2 are obtained by the weighted SRMCB method with \(\delta = 0.7\). It is interesting to notice that as \(\delta\) varies from 1 to 0.7, the point estimates as well as confidence densities resemble more and more the ad hoc nonparametric method. This shows us how the results might vary under one possible departure from the MCAR-assumption.

Acknowledgements

The author thanks two anonymous referees for constructive comments on an previous version of this work.

Appendix A. Results for the SRMCB method

At the \(k\)th iteration, we have

\[
E_\ast[\hat{\varepsilon}^*_{ij(k)} | \tilde{x}_{(k-1)}] = \sum_{i=1}^{n} \frac{n_i (x_i - \tilde{x}_{(k-1)})}{n_1} = \tilde{x}_r - \tilde{x}_{(k-1)},
\]

and

\[
E_\ast[\tilde{x}_{(k)} | \tilde{x}_{(k-1)}] = n^{-1} (n_1 \tilde{x}_r + n_0 E_\ast[\tilde{x}_{m(k)} | \tilde{x}_{(k-1)}])
\]

\[
= n^{-1} (n_1 \tilde{x}_r + n_0 \tilde{x}_{(k-1)} + n_0 E_\ast[\hat{\varepsilon}_{ik} | \tilde{x}_{(k-1)}]) = \tilde{x}_r.
\]
Meanwhile, we have \((n_1 - 1)s_r^2 = \sum_{i=1}^{n} r_i(x_i - \bar{x}_r + \bar{x}_r - \bar{x}_{(k)})^2 = (n_1 - 1)s_r^2 + n_1(\bar{x}_r - \bar{x}_{(k)})^2\).

Thus,
\[
E_*(\bar{x}_r - \bar{x}_{(k)})^2 = V_*[\bar{x}_{(k)}] = E_*[V_*[\bar{x}_{(k)}|\bar{x}_{(k-1)}]] + V_*[E_*[\bar{x}_{(k)}|\bar{x}_{(k-1)}]] \\
= E_* \left[ V_* \left[ \frac{n_0}{n} \bar{x}_{m(k)}|\bar{x}_{(k-1)} \right] \right] + V_*[\bar{x}_r] \\
= \frac{n_0}{n^2} \left( 1 - \frac{1}{n_1} \right) E_*[s_{r(k-1)}^2] \\
= \frac{n_0}{n^2} \left( 1 - \frac{1}{n_1} \right) E_* \left[ s_r^2 + \frac{n_1}{n_1 - 1} (\bar{x}_r - \bar{x}_{(k-1)})^2 \right] \\
= \left( 1 - \frac{1}{n_1} \right) \left[ \sum_{j=1}^{k} \left( \frac{n_0}{n^2} \right)^j \right] s_r^2 + \left( \frac{n_0}{n^2} \right)^k (\bar{x}_r - \bar{x}_{(0)})^2 \\
\Rightarrow \lim_{k \to \infty} \left( 1 - \frac{1}{n_1} \right) \left[ \sum_{j=1}^{k} \left( \frac{n_0}{n^2} \right)^j \right] s_r^2 + \left( \frac{n_0}{n^2} \right)^k (\bar{x}_r - \bar{x}_{(0)})^2 \\
\Rightarrow \lim_{k \to \infty} \left( \frac{n_0}{n_1 - 1} - \frac{n_0}{n_1} \right) s_r^2.
\]

It follows that \(\lim_{k \to \infty} E_*[s_{r(k)}^2] = s_r^2 + n_0 s_r^2/(n^2 - n_0) \to \infty s_r^2.

### Appendix B. Results for the regression SRMCB method

At the \(k\)th iteration, let \(b_{(k)} = \beta_{(k)} - \beta_r\). We have \(\hat{\alpha}_{(k)} = \epsilon_i - x_i^T b_{(k)}\) if \(r_i = 1\), where \(\epsilon_i = y_i - x_i^T \beta_r\). Thus, \(E_*[y_{i(k+1)}^*|\beta_{(k)}] = x_i^T \beta_r + (x_i - \bar{x}_r)^T b_{(k)}\) if \(r_i = 0\). We have, for \(B = A^{-1}(A_m - n_0 \bar{x}_m \bar{x}_r^T)\),
\[
E_*[\beta_{(k+1)}|\beta_{(k)}] = A^{-1} \left\{ \sum_{i=1}^{n} r_i x_i (x_i^T \beta_r + \epsilon_i) + \sum_{i=1}^{n} (1 - r_i) x_i E_*[y_{i(k+1)}^*|\beta_{(k)}] \right\} \\
= \beta_r + B b_{(k)}.
\]

It follows that \(E_*[\beta_{(k)}|\beta_{(0)}] = \beta_r + B^k b_{(0)} \to \beta_r\) provided the matrix \(B\) is convergent. That is, if \(|\lambda| < 1\) for any eigenvalue \(\lambda\) of \(B\). Let \(\mu\) and \(\Sigma\) be the mean vector and the covariance matrix of \(x_i\). Asymptotically, as \(n \to \infty\), we have \(n_0/n \to p_0\), and \(A / n \to \Sigma + \mu \mu^T\), and \(A_m/n \to \Sigma + \mu \mu^T\), and \(\bar{x}_m \bar{x}_r \to \mu \mu^T\). It follows that, for large samples, the convergence of the regression SRMCB method is determined by the matrix \(p_0(\Sigma + \mu \mu^T)^{-1}\Sigma\). Meanwhile, we have \(s_{r(k)}^2 > s_r^2\) for any \(k\), since
\[
(n_1 - 1)s_{r(k)}^2 = \sum_{i=1}^{n} r_i(y_i - x_i^T \beta_r + x_i^T \beta_r - x_i^T \beta_{(k)})^2 \\
= (n_1 - 1)s_r^2 + (\beta_r - \beta_{(k)})^T A_r (\beta_r - \beta_{(k)}).
\]
For large $n$ and $k$, $E_*(s^2_r(\k)) = s^2_r + O(V_*(\k))$. Since $y^*_i(\k) = x_i^T \beta_r + e_i(\k) + e^*_i(\k)$ where $e_i(\k) = x_i^T b(\k)$ and $V_* \left( \sum_{i=1}^n (1 - r_i) x_i (e_i(\k) + e^*_i(\k)) \right) = O(n_0)$, we have

$$V_*(\beta(\k)) = V_* \left[ A^{-1} \left( \sum_{i=1}^n (1 - r_i) x_i (e_i(\k) + e^*_i(\k)) \right) \right] = A^{-1} O(n_0)(A^{-1})^T = O \left( \frac{n_0}{n^2} \right).$$

References


