Appl. Statist. (2007)
56, *Part* 1, *pp.* 63–78

Using a mixture model for multiple imputation in the presence of outliers: the 'Healthy for life' project

Michael R. Elliott

University of Michigan, Ann Arbor, USA

and Nicolas Stettler

Children's Hospital of Philadelphia, USA

[Received March 2005. Final revision June 2006]

Summary. We consider the problem of obtaining population-based inference in the presence of missing data and outliers in the context of estimating the prevalence of obesity and body mass index measures from the 'Healthy for life' study. Identifying multiple outliers in a multivariate setting is problematic because of problems such as masking, in which groups of outliers inflate the covariance matrix in a fashion that prevents their identification when included, and swamping, in which outliers skew covariances in a fashion that makes non-outlying observations appear to be outliers. We develop a latent class model that assumes that each observation belongs to one of K unobserved latent classes, with each latent class having a distinct covariance matrix. We consider the latent class covariance matrix with the largest determinant to form an 'outlier class'. By separating the covariance matrix for the outliers from the covariance matrices for the remainder of the data, we avoid the problems of masking and swamping. As did Ghosh-Dastidar and Schafer, we use a multiple-imputation approach, which allows us simultaneously to conduct inference after removing cases that appear to be outliers and to promulgate uncertainty in the outlier status through the model inference. We extend the work of Ghosh-Dastidar and Schafer by embedding the outlier class in a larger mixture model, consider penalized likelihood and posterior predictive distributions to assess model choice and model fit, and develop the model in a fashion to account for the complex sample design. We also consider the repeated sampling properties of the multiple imputation removal of outliers.

Keywords: Body mass index; Child; Community health centre; Latent class; Multiple-edit–multiple-imputation model; Obesity; Survey sampling

1. Introduction

Childhood obesity has become epidemic in the USA and is rapidly increasing throughout the developed and even the developing world (Hedley *et al.*, 2004; Kimm and Obarzanek, 2002). The increase in childhood obesity during the past 25 years has led policy makers to rank it as one of the most critical public health threats of the 21st century (Koplan *et al.*, 2004). Although the nationally representative sample of the National Health and Nutrition Examination Survey provides childhood overweight status by age group in non-Hispanic white, non-Hispanic black and Mexican-American children (Hedley *et al.*, 2004; Ogden *et al.*, 2002), data on other ethnic groups are lacking. Additionally, the magnitude of the problem of obesity among children living in medically underserved areas is unknown, yet it is an important factor to consider in directing scarce resources for the treatment and prevention of obesity.

Address for correspondence: Michael Elliott, Department of Biostatistics, University of Michigan School of Public Health, 1420 Washington Heights, Ann Arbor, MI 48109, USA.
E-mail: mrelliot@isr.umich.edu

 0035–9254/07/56063

In this context, the 'Healthy for life' survey obtained data from a probability sample of children using Health Service and Resource Administration (HSRA) supported community health centres at least once during calendar year 2001 (Stettler *et al.*, 2005). The purpose of the survey was to quantify the prevalence of paediatric obesity in medically underserved areas. Data were collected by abstraction of the clinically measured height and weight during the last visit to the health clinic in 2001. Because height data are collected only sporadically, nearly a quarter of the height and consequently the body mass index (BMI) data were missing; missingness was associated with age, since older children tended to grow more slowly and thus were less likely to have height recorded at a given visit. To reduce bias and inefficiency that are associated with a complete-case analysis, and to allow analysts to work with the data in a convenient manner, a multiple-imputation (MI) method was implemented in Stettler *et al.* (2005). However, the MI procedure was potentially problematic, because the data were overdispersed under the assumption of normality and included incorrectly recorded or abstracted elements. Failure to account for missing data and clerical errors in the height and weight HSRA data may have important clinical and public health repercussions. Without correction, differences in prevalence between centres with more or fewer clerical errors may be misinterpreted as an increased risk for obesity in children living in one area rather than insufficient standardization between centres. Standardization in measurement and transcription in multicentre studies is expensive, as it requires rigorous training and travelling. The method that we propose provides a post-data collection alternative to eliminate outliers when extensive training has not been possible before data collection.

The literature on outlier detection is voluminous: books in the field that provide an overview include Hawkins (1980) and Barnett and Lewis (1994). Using standard methods such as consideration of the Mahalanobis distance to identify multiple outliers in multivariate data is problematic (Campbell, 1980; Rousseeuw and van Zomeren, 1990; Hadi, 1992). 'Masking' prevents identification of outliers when a small cluster of observations inflates the empirical covariance matrix, whereas 'swamping' can make some observations appear to be outliers when true outliers pull the empirical covariance matrix away from non-outlier observations. Methods for simultaneously assessing outliers and accounting for missing data in an MI framework include Little and Smith (1987), Little (1988), Penny and Jollife (1999) and Ghosh-Dastidar and Schafer (2003).

The goal of our analysis is to obtain the distribution of the BMI in the HSRA paediatric population after removing outliers that are probably due to clerical errors. We begin by defining a mixture model for the joint distribution of Box–Cox-transformed and age- or gender-normalized ('z-score') height and weight data. The mixture model is defined by latent classes that have common means, conditional on age and health centre to accommodate the disproportional sample design, but that have differing covariances; the 'clerical error class' is the class with the largest covariance matrix determinant. We then use this mixture model to develop an MI algorithm that imputes latent variance class conditional on its posterior probability of membership; missing height z-score data are then imputed conditional on weight, health centre and latent variance class. The height and weight z-scores are then backtransformed to heights and weights on their original scales, and then used to compute the BMI. Subjects who were assigned to the clerical error class at a given imputation are dropped before the complete-data analysis of the observed and imputed data.

Two statistics summarizing the distribution of the BMI in the HSRA paediatric population after clerical error removal were of particular interest: the proportion of children above a fixed BMI cut point measure for obesity, and the 2.5- and 97.5-percentiles of the BMI distribution. In principle these could be obtained from the mixture model that was used to produce the MIs.

We used the MI approach to rely on the empirical distribution of the data to the largest extent possible, and to accommodate the complex sample design at the complete-data stage of analysis, further enhancing robustness. Another alternative to the MI approach would be to remove all subjects whose modal probability of class membership was in the outlier class before conducting analyses; the use of MI instead of 'one-off' outlier removal provides a simple method to obtain the obesity prevalence estimates within a variety of subdomains while stochastically eliminating potential clerical error outliers from the analysis, so that subjects with less extreme values can be 'partially' removed in proportion to the probability with which they belong to the clerical outlier class. Thus outliers are assessed on the basis of how an observation relates to a posterior predictive distribution that excluded the observation in estimation, yielding a discordancy or 'surprise' measure (Chaloner and Brant, 1988; Bayarri and Morales, 2003). Our work extends the 'multiple-edit–MI' model of Ghosh-Dastidar and Schafer (2003) by

(a) embedding the clerical error class in a larger mixture model that allows for differing degrees of height–weight overdispersion in the population,
(b) considering the Akaike information criterion AIC, Bayes information criterion BIC and posterior predictive distribution p-values to select among differing class sizes considered and
(c) developing the model in a fashion to account for the complex sample design.

By allowing for more than two latent covariance classes, we accommodate overdispersion in the HSRA height–weight data relative to the total population beyond that induced by transcription errors, which may be of direct interest of itself. We also consider the repeated sampling properties of the MI procedure proposed.

Section 2 describes the model and provides details about the Gibbs sampling algorithm that was used to develop the MI procedure. Section 3 applies the MI procedure to the 'Healthy for life' survey data. Section 4 considers the proposed MI procedure in the repeated sampling context. Section 5 summarizes the results and considers future extensions of this approach.

2. Description of the model

We consider a complete-data mixture model: for the ith subject, $i = 1, \ldots, n$,

$$\mathbf{Z}_i | C_i = k \sim N_q(\boldsymbol{\mu}_i, \Sigma_k), \tag{1}$$

$$C_i \sim \text{MULTI}(1, p_1, \ldots, p_K)$$

where $\mathbf{Z}_i$ is a q-dimensional outcome of interest, $\mu_{ij} = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_j$, $j = 1, \ldots, q$, where $|\Sigma_1| < \ldots < |\Sigma_K|$. Thus we assume that each subject has a mean that depends on a set of p covariates $\mathbf{x}_i$, and a covariance that is given by his or her latent variance class membership, which is denoted by the unobserved latent variable C_i, where $C_i = K$ indicates that the ith subject belongs to the 'clerical error' class with the largest variability. Clearly the C_i are missing for all subjects; we also allow for some components of $\mathbf{Z}_i$ to be missing. This model assumes that transcription errors have a single common covariance, which is larger than the correctly transcribed data, but allows for variance heterogeneity within the outcome.

We next postulate the following independent, weakly informative priors for the model parameters:

$$p(\boldsymbol{\beta}) \sim N(0, V_\beta),$$

$$p(\Sigma_K) \sim \text{INV-WISHART}(2, S_K),$$

$$p(p_1, \ldots, p_K) \sim \text{DIRICHLET}(1, \ldots, 1),$$

$$p\{\log(\sigma_{jjk})\} \overset{\text{ind}}{\sim} N(0, s^2), \qquad k = 1, \ldots, K-1,$$

$$p(\rho_{lmk}) \overset{\text{ind}}{\sim} U(-1, 1), \qquad l = 1, \ldots, q-1, \quad m = 2, \ldots, q.$$

This model-based approach assumes that the missingness mechanism and sample design are fully ignorable in the sense of Rubin (1987). The missingness at random assumption holds if, conditional on the observed elements of $\mathbf{Z}_i$, the missingness status of the elements of $\mathbf{Z}_i$ is unrelated to their value. This assumption is untestable but is a weaker assumption than a standard complete-case analysis, which assumes a mechanism of data missing completely at random (unconditional independence between the missingness status of elements of $\mathbf{Z}_i$ and their value). By including covariates that account for the sample design in the regression of the mean, there is no need to incorporate the sample design further in the model if the latent class C_i is independent of the probability of selection. This assumption can be tested by considering the association between the posterior probabilities of class membership and the case weights. Even if the probability of class membership is associated with the probability of selection, the effect of this form of model misspecification may be less severe on the MI procedure if the complete-data analysis uses a robust procedure (i.e. includes case weights) in the analysis. Indeed, by using Taylor series linearization estimates of variance (Woodruff, 1971), we can account for stratification, clustering and unequal probability of selection in the sample design in the complete-data analysis. This is an example of 'uncongeniality' (Meng, 1994) in which the analyst assumes more than the imputer; if the analyst's assumptions are correct, the resulting inferences will usually be consistent, though conservative.

Details of the Gibbs sampler data augmentation algorithm that was used to obtain imputations of the missing elements of $\mathbf{Z}_i$ and the completely unobserved C_i for use in the MI are provided in Appendix A. In many settings it may be reasonable to constrain the correlations

$$\rho_{lmk} = \sigma_{lmk} / \sigma_{llk}\sigma_{mmk}$$

across the non-clerical-error covariance classes $k = 1, \ldots, K-1$ to be equal; hence we include this constraint as an option in the algorithm.

2.1. Multiple imputation

We simultaneously accommodate the missing height data and remove the clerical error outlier class from our inference by use of MI (Rubin, 1987; Schafer, 1997), which allows us to take m independent draws of $\mathbf{Z}^{\text{comp}}$ given by replacing the missing elements of $\mathbf{Z}$ with their imputed values from the Gibbs sampling procedure, to analyse by using standard complete-data procedures and to combine the results in a fashion that properly propagates the uncertainty in the imputation procedure and, in this setting, in the clerical error class assignment. In the finite population inference setting, assume that, if we had complete data, we would estimate a (scalar) finite population quantity T with a statistic $Q(\mathbf{Z}^{\text{comp}})$, and that the corresponding $\text{var}\{Q(\mathbf{Z}^{\text{comp}})\}$ would be estimated by using $\widehat{\text{var}}\{Q(\mathbf{Z}^{\text{comp}})\}$ that appropriately accounts for the sample design. We then obtain an MI point estimate of the underlying model parameter or finite population quantity Q as

$$\hat{Q} = m^{-1} \sum_{t=1}^{m} Q(\mathbf{Z}^{\text{comp}(t)}).$$

Inference is based on

$$V^{1/2}(\hat{Q} - Q) \sim t_\nu$$

where

$$V = U + (1 + m^{-1})B$$

for

$$U = m^{-1} \sum_{t=1}^{m} \widehat{\text{var}}\{Q(\mathbf{Z}^{\text{comp}(t)})\},$$

$$B = (m-1)^{-1} \sum_{t=1}^{m} \{\hat{Q} - Q(\mathbf{Z}^{\text{comp}(t)})\}^2,$$

$$\nu = (m-1)\left\{1 + \frac{U}{(1+m^{-1})B}\right\}^2.$$

Because we do not want inference about the underlying population to be based on the data elements that are clerical errors, we delete subjects who are assigned to the Kth latent class when computing $Q(\mathbf{Z}^{\text{comp}(t)})$.

2.2. Model identifiability

The finite mixture normal model, of which the complete-data model (1) is a special case that assumes equal means across mixture classes, is identifiable up to the permutation of the class assignments (Teicher, 1963); model (1) requires at least

$$p + K - 1 + K\frac{q(q+1)}{2}$$

complete observations to identify the equivalent number of parameters. For small numbers of observations, the likelihood may be maximized at $\hat{\Sigma}_{k-1} = \hat{\Sigma}_k$ if $\Sigma_{k-1} \approx \Sigma_k$, leading to aliasing and requiring that a model of size $K - 1$ be fitted instead. Missing elements of $\mathbf{Z}_i$ will weaken identification by making the class assignment for this subject less certain.

3. Application to 'Healthy for life' survey

The 'Healthy for life' survey was a cross-sectional survey of a representative sample of all children users of the 141 HRSA-supported community health centres in region 2 (New Jersey, New York, Puerto Rico (PR) and the Virgin Islands) and region 3 (Delaware, the District of Columbia, Maryland, Pennsylvania, Virginia and West Virginia), from January 1st to December 31st, 2001. A three-stage disproportionately stratified sampling scheme was designed to provide stable prevalence estimates by age group (2–5 years and 6–11 years) and, within each age group, by gender, race or ethnicity (non-Hispanic white, non-Hispanic black, non-Hispanic Asian and Hispanic), and region (US mainland urban, suburban and rural, PR urban and non-urban, and New York City Chinatown). A total of 30 centres were sampled, with date of birth, gender, race or ethnicity, height, weight and other medical information abstracted from the child's most recent 2001 visit for approximately 100 children aged 2–11 years from each centre. Details of the sampling scheme are available in Stettler *et al.* (2005).

Data for a total of 3579 children were obtained from the participating centres. Population totals at the time of sampling were available only for the age groups of 1–4 and 5–12 years; thus known sampling fractions could be obtained only for children in these age categories, and thus samples of children who were 1–4 and 5–12 years old were initially drawn. Because the desired

age stratifications were ages 2–5 and 6–11 years, however, 720 1-year-old and 12-year-old children were dropped before analysis. An additional 351 cases were dropped because they lacked information on age, making it impossible to determine whether they were eligible for the analysis, whereas 23 were deleted because they lacked gender or both height and weight information, making it difficult to impute data meaningfully. Finally, three cases were removed because they lack information on weight, making the resulting missing data pattern monotonic and simplifying computations somewhat, whereas eight cases were removed because of known transcription errors. This yielded a total of 2474 cases that were available for analysis. Of these 606 were missing height data. Under Teicher (1963), sufficient data should be available to identify four classes of variability easily, the most that we consider below.

Because the populations of children within the age categories 1–4 and 5–12 years were known, the selection probability for each sampled child could be determined as $\pi_{sa} = 1/H_s \times n_{sa}/N_{sa}$, where H_s is the number of centres in region-by-size substratum s, n_{sa} is the number of children who were drawn in the ath age group from the centre sampled and N_{sa} is the number of children in the ath age group at the centre. Post-strata consisting of 12 r region by a age cells were then formed, and post-stratification adjustments

$$f_{ra} = \frac{\sum_{s \in r} N_{sa}}{\sum_{s \in r} 1/\pi_{sa}}$$

were computed so that the sum of the weighted sample matched known age–region totals. The final design weights that were used in the analysis were then given by $w_i = f_{ra}/\pi_{sa}$ for $i \in a$ and $s \subset r$. The mean design weight was 171.2, with a range from 28.8 to 659.1. The case weights are not used in the data augmentation algorithm that produces the MIs but are used in the complete-data analysis of the multiply imputed data to enhance robustness.

In keeping with the clinical literature for the analysis of the BMI (Cole *et al.*, 2000; Leonard *et al.*, 2004; Weiss *et al.*, 2004), we perform a 'preprocessing' step to transform the raw height and weight data to approximate normality via a Box–Cox-type transformation (Box and Cox, 1964). Denote weight (in kilograms) and height data (in metres) for the ith subject by Y_{i1} and Y_{i2}. Then

$$Z_{ij} = \frac{(Y_{ij}/M_{ij})^{L_{ij}} - 1}{L_{ij} S_{ij}}, \qquad i = 1, \ldots, n, \quad j = 1, 2,$$

where $L_{ij} = L_j(A_i, G_i)$, $M_{ij} = M_j(A_i, G_i)$ and $S_{ij} = S_j(A_i, G_i)$ are population parameters that are functions of the age A_i and gender G_i of the ith subject and are obtained from National Center for Health Statistics growth charts and treated as known (Cole, 1990, 1994). This yields height and weight z-scores that are used in the remainder of the analysis. Covariates $\mathbf{x}_i$ consist of an age-by-centre group dummy variable, to accommodate within-centre correlation along with any systematic association between BMI and the probability of selection.

A preliminary analysis suggested that correlations $\rho_k = \sigma_{12k}/\sigma_{11k}\sigma_{22k}$ across the non-clerical-error covariance classes $k = 1, \ldots, K-1$ were similar. Because this is consistent with correlation between height and weight z-scores being equal in the population regardless of the variance class to which a subject belonged, which is a reasonable assumption, we assume this constraint in our analysis.

For our prior distributions, we assumed

$$V_\beta = \begin{pmatrix} 1000 & 0 \\ 0 & 1000 \end{pmatrix},$$

$$p\{\log(\sigma_{jjk})\} \overset{\text{ind}}{\sim} N(0, 4), \qquad k = 1, \ldots, K-1,$$

and

$$S_K = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}.$$

We ran two Gibbs samplers and assessed convergence by using the Gelman–Rubin statistic $\hat{R}$, an adjusted ratio of between- and within-sequence variability to within-sequence variability (Gelman *et al.* (2004), pages 296–297). A 2000-draw burn-in was used for each chain. The two- and three-class models required 5000 draws for convergence; the four-class model required 20000 draws.

3.1. Results of model fit

Table 1 provides the 2.5-, 50- and 97.5-percentiles of the posterior distributions of the variance class fractions and variance parameters for two-, three- and four-class models under the assumption that $\rho_1 = \ldots = \rho_{k-1}$. The two-class model suggests that 5–8% of the entries are transcription errors. Both the three- and the four-class models suggest an extremely overdispersed group of 1–3% of entries that are probably transcription errors, along with a more moderately 'overdispersed' class of extremely large or extremely small children for their gender and height consisting of 5–10% of the population. The four-class model splits the overdispersed class in the three-class model into a moderately and highly overdispersed class.

3.2. Model checking and model choice

Table 1 shows AIC- (Akaike, 1978) and BIC- (Schwarz, 1978) measures under each of the three models, obtained by using the posterior mode of $(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\rho})$ where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_q^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\phi} = (\sigma^2_{111}, \ldots, \sigma^2_{qqK}, \rho_{12}, \rho_{q-1,q})$ and $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_K)$. Both criteria suggest that the three-class model provides the best fit to the data. We focus on constrained three-class models for the remainder of the analysis.

To test the distributional assumptions of the model, we use posterior predictive distributions (Gelman *et al.*, 1996). We consider the posterior predictive distribution of the χ^2-type statistic

Table 1. Results from the data model for the 'Healthy for life' study for two-, three- and four-variance class models†

k	p_k	σ^2_{11k}	σ^2_{22k}	ρ_k	*AIC*	*BIC*
2 class						
1	$0.936_{0.917,0.953}$	$1.43_{1.33,1.56}$	$1.19_{1.10,1.29}$	$0.70_{0.67,0.73}$	13657.25	14401.39
2	$0.064_{0.047,0.083}$	$12.71_{9.75,17.30}$	$18.77_{13.71,27.18}$	$0.73_{0.62,0.81}$		
3 class						
1	$0.912_{0.873,0.936}$	$1.43_{1.35,1.55}$	$1.14_{1.04,1.24}$	$0.70_{0.67,0.72}$	13616.95	14378.53
2	$0.072_{0.049,0.106}$	$3.88_{2.40,6.07}$	$12.34_{7.01,18.83}$	$0.70_{0.67,0.72}$		
3	$0.015_{0.007,0.029}$	$37.48_{21.14,83.88}$	$29.23_{15.23,64.03}$	$0.92_{0.63,0.98}$		
4 class						
1	$0.879_{0.717,0.918}$	$1.41_{1.26,1.54}$	$1.09_{0.98,1.22}$	$0.70_{0.67,0.72}$	13662.57	14441.59
2	$0.067_{0.011,0.216}$	$2.23_{1.00,3.76}$	$4.75_{1.22,8.82}$	$0.70_{0.67,0.72}$		
3	$0.043_{0.014,0.084}$	$5.11_{3.13,5.11}$	$17.04_{9.35,29.06}$	$0.70_{0.67,0.72}$		
4	$0.013_{0.006,0.026}$	$43.39_{23.31,98.60}$	$28.49_{15.26,65.17}$	$0.95_{0.73,0.99}$		

†The values are the posterior median and 95% posterior predictive interval in the subscripts for the proportion of population p_k in variance class k, variance of weight σ^2_{11k}, variance of height σ^2_{22k} and correlation between height and weight ρ_k; variance class K is assumed to consist of clerical error outliers. The correlation is assumed to be equal for classes $k = 1, \ldots, K-1$ ('true data').

$$S = n^{-1} \sum_{i=1}^{n} \sum_{k=1}^{K} I(C_i = k) \tilde{Z}_{ki}^2$$

where

$$\tilde{Z}_{ki}^2 = \begin{cases} (\mathbf{Z}_i - \boldsymbol{\mu}_i)^{\mathrm{T}} \Sigma_k^{-1} (\mathbf{Z}_i - \boldsymbol{\mu}_i) & \text{if } Z_{i2} \text{ is observed,} \\ (Z_{i1} - \mu_{i1})^2 / \Sigma_{11k} & \text{if } Z_{i2} \text{ is missing.} \end{cases}$$

If the number of classes is sufficient, the assumption of normality within class will hold, at least approximately, and S^{obs} and S^{rep} will correspond. For the three-class model, $P(S^{\mathrm{obs}} < S^{\mathrm{rep}} | y) = 0.46$, suggesting that the distributional assumptions of the model are reasonable.

To assess the assumption that the latent class membership probabilities are independent of the selection probabilities, we determined the Spearman correlations between the posterior medians of latent variance class probability membership for the three-class constrained model $\hat{\pi}_{ki}$ and the inverse of the case weight $1/w_i$: Spearman's ρ is -0.022 for $k = 1$ ($p = 0.27$), 0.008 for $k = 2$ ($p = 0.69$) and 0.012 for $k = 3$ ($p = 0.54$). This suggests that including the centres as fixed effects has been sufficient to remove design effects from the model.

3.3. Multiple imputation

To assure independence between draws of $\mathbf{Z}^{\mathrm{comp}}$, we retained the imputation from every 250 draws from the Gibbs sampler for a total of 20 imputations. Fig. 1 plots the observed and imputed height and weight z-scores for the first four imputations; draws assigned to the 'outlier' class are denoted in red. Fig. 1 highlights the overrepresentation of imputed values in the outlier class, which is consistent with clerical errors being associated with missing data. The uncertainty that is associated in assigning outliers at the edge of the non-outlier distribution is accounted for, as well as the fact that the error mechanism that generates outliers will also create erroneous values that are assigned to the central part of the distribution by chance.

3.4. Estimation of obesity rate and body mass index extremes

For each imputed data set, the observed weight and observed or imputed height z-score were backtransformed to weight and height measures, and the BMI was computed as weight/height^2; the BMI was then itself z-score transformed as in Cole (1990). A subject was classified as obese if their BMI z-score for their age exceeded the 95-percentile of a reference population (Kuczmarski *et al.*, 2000). The standard complete-case data analysis to estimate the obesity prevalence utilized design-based procedures that account for stratification by region, clustering by centre and unequal probability of selection by age–centre cell in the sample design; in particular, fully weighted means and Taylor series linearization estimates of variance (Woodruff, 1971) were calculated by using procedure `SURVEYMEAN` in SAS version 8.2 (SAS Institute, 2001).

Table 2 gives the percentage of the population of patients that is obese, by age and region: under a complete-case analysis, the standard MI for height without outlier mixture model and MI under the outlier mixture model.

The standard MI analysis suggests that children missing height data appeared somewhat more likely to be heavier for their age and gender in the US regions, and somewhat underweight for their age and gender in the urban PR centres. Accounting for the effect of the missingness also reduced the confidence intervals by about 5–10% relatively to the complete-case data, which is consistent with the 25% missingness of the height data and the fact that the length of the confidence interval is $O(n^{-1/2})$.

The MI mixture analysis suggested that the outliers may have caused the obesity rate to be biased upwards under a standard MI analysis: if height data are missing and an older child is

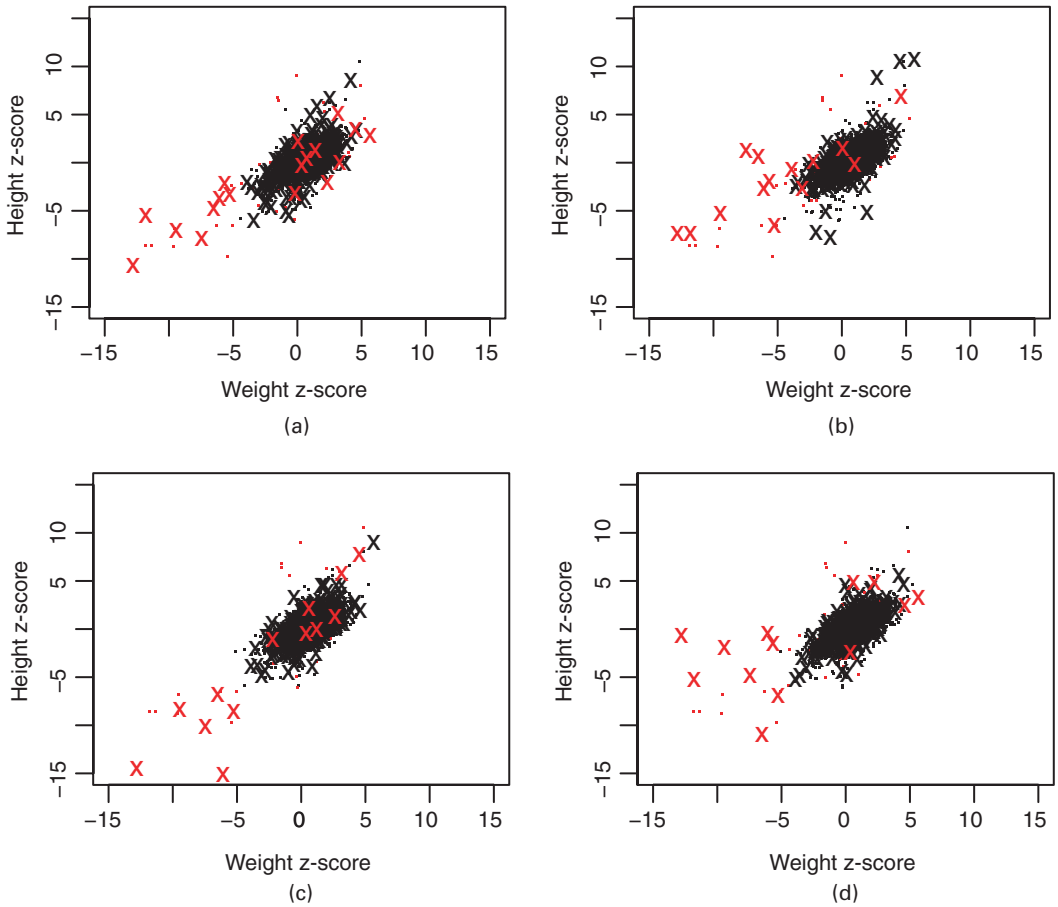

Fig. 1. Weight *versus* observed (·) and imputed (×) height *z*-scores for (a) imputation 1, (b) imputation 2, (c) imputation 3 and (d) imputation 4: black values are assigned to non-clerical-error variance class clusters and red values are assigned to clerical error variance class clusters

Table 2. Percentage obese by age and by region†

Region	*% obese for age 2–5 years and the following models:*			*% obese for age 6–11 years and the following models:*		
	CC	*MI(1)*	*MI(2)*	*CC*	*MI(1)*	*MI(2)*
All	$23.0_{19.2,26.7}$	$23.5_{19.9,27.1}$	$22.7_{19.1,26.2}$	$22.8_{20.0,25.6}$	$22.7_{20.3,25.2}$	$22.7_{20.2,25.2}$
Urban	$22.0_{17.0,27.1}$	$21.8_{17.0,26.6}$	$21.6_{16.7,26.5}$	$23.0_{18.7,27.4}$	$23.0_{19.0,27.1}$	$23.4_{19.2,27.6}$
Suburban	$21.0_{14.0,28.0}$	$23.5_{16.8,30.3}$	$22.3_{15.4,29.2}$	$24.0_{17.7,30.3}$	$24.2_{18.7,29.7}$	$23.6_{18.1,29.1}$
Rural	$19.4_{13.0,25.8}$	$22.6_{15.9,29.2}$	$20.8_{15.2,26.5}$	$28.6_{22.0,35.1}$	$25.2_{20.4,30.1}$	$25.2_{19.9,30.5}$
New York City Chinatown	$16.2_{3.7,28.7}$	$15.8_{3.6,27.9}$	$15.6_{3.8,27.3}$	$18.2_{6.3,30.0}$	$17.8_{7.1,28.5}$	$18.1_{6.8,29.4}$
PR urban	$21.1_{7.5,34.6}$	$19.7_{7.2,32.1}$	$18.9_{6.4,31.4}$	$21.3_{9.1,33.4}$	$20.8_{8.9,32.8}$	$20.5_{8.9,32.1}$
PR other	$27.4_{17.3,37.5}$	$27.0_{17.2,36.8}$	$25.8_{16.2,35.5}$	$18.9_{13.0,24.8}$	$18.6_{12.7,24.4}$	$18.3_{12.6,24.1}$

†CC, complete-case analysis; MI(1), standard MI for the height without outlier mixture model; MI(2), MI under the outlier mixture model. 95% confidence intervals are given in the subscripts.

Table 3. Empirical estimates of 2.5- and 97.5-percentiles of BMI, by age and by region†

Region	*Percentiles for age 2–5 years and the following models:*			*Percentiles for age 6–11 years and the following models:*		
	CC	*MI(1)*	*MI(2)*	*CC*	*MI(1)*	*MI(2)*
All	(13.9,22.8)	(13.8,22.8)	(13.8,22.7)	(13.3,30.5)	(13.1,30.2)	(13.4,30.2)
Urban	(13.5,22.7)	(13.6,22.7)	(13.7,22.5)	(13.7,29.3)	(13.5,29.3)	(13.7,29.3)
Suburban	(13.9,22.3)	(13.9,22.8)	(13.9,22.5)	(13.6,32.6)	(13.7,31.0)	(13.6,31.7)
Rural	(13.6,22.3)	(13.4,23.4)	(13.5,22.6)	(12.3,31.0)	(12.2,31.4)	(13.0,30.8)
New York City Chinatown	(14.1,23.3)	(13.8,23.3)	(13.8,23.2)	(13.4,26.0)	(13.3,26.0)	(13.6,26.0)
PR urban	(13.5,27.1)	(13.6,24.2)	(13.7,23.7)	(12.8,27.5)	(12.8,27.5)	(12.8,27.2)
PR other	(14.0,22.8)	(13.9,22.8)	(13.9,22.9)	(11.8,28.1)	(11.8,28.3)	(12.3,28.2)

†CC, complete-case analysis; MI(1), standard MI for the height without outlier mixture model; MI(2), MI under the outlier mixture model.

incorrectly noted as younger, the resulting weight z-score would be extremely large, probably yielding a large BMI after height imputation, and potentially classifying a non-obese child as obese; the reverse is true if a younger child is incorrectly noted as older. Since children are more likely than not to be non-obese, the net effect of age transcription errors should be to inflate rates of obesity among younger children, and to deflate to a much lesser degree obesity rates among older children.

Table 3 presents the estimated population 2.5- and 97.5-percentiles for the BMI under a complete-case analysis, the standard MI for height without outlier mixture model and MI under the outlier mixture model.

The standard MI analysis suggests that the complete-case analysis underestimated both 2.5 and 97.5 BMI percentiles for 2–5-year-old children in US suburban and rural areas, and overestimated 97.5-percentiles for PR urban 2–5-year-old children and 6–11-year-old suburban US children.

The MI mixture analysis suggested that younger children appear to have large BMI outliers and older children small BMI outliers, and outliers of both types appear to be somewhat disproportionately located in the US rural and PR urban centres. The tendency for the younger children to have large BMI outliers and older children to have small BMI outliers is also consistent with clerical errors being due to age, rather than incorrect transcriptions of height or weight directly.

4. Simulation study

For a simulation study, we generated data under model (1) where $\boldsymbol{\mu}_i \equiv \mathbf{0}$,

$$\Sigma_k = \sigma_k \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

for $k < K$ and

$$\Sigma_K = \sigma_K \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and $n = 500$. We assumed a simple random sampling. In addition, we deleted elements of Z_{i2} under a missingness at random mechanism, where

$$P(M_i = 1 | Z_{i1} = z_{i1}, C_i = k) = \begin{cases} \dfrac{\exp(z_{i1}/\sqrt{\sigma_k})}{1+\exp(z_{i1}/\sqrt{\sigma_k})} & \text{if } k < K, \\ 0 & \text{if } k = K \end{cases}$$

for the missingness indicator M_i, so that large values of Z_{i1} tended to be associated with missing Z_{i2} unless the observation was a member of the outlier class.

We considered four latent variance class models in each of 200 simulations:

(a) $K = 2$, $\sigma_1 = 1$, $\sigma_2 = 100$, $\rho = 0.5$ and $\mathbf{p} = (0.98\ 0.02)^{\mathrm{T}}$ (model A);
(b) $K = 2$, $\sigma_1 = 1$, $\sigma_2 = 100$, $\rho = 0.5$ and $\mathbf{p} = (0.90\ 0.10)^{\mathrm{T}}$ (model B);
(c) $K = 4$, $\sigma_1 = 0.25$, $\sigma_2 = 1$, $\sigma_3 = 9$, $\sigma_4 = 100$, $\rho = 0.5$ and $\mathbf{p} = (0.245\ 0.245\ 0.245\ 0.02)^{\mathrm{T}}$ (model C);
(d) $K = 4$, $\sigma_1 = 0.25$, $\sigma_2 = 1$, $\sigma_3 = 9$, $\sigma_4 = 100$, $\rho = 0.5$ and $\mathbf{p} = (0.225\ 0.225\ 0.225\ 0.10)^{\mathrm{T}}$ (model D).

For each simulation we determined the following:

(a) the choice of K under the AIC- and BIC-criteria;
(b) the mean and 95% confidence interval for $p = P(Z_{i2} < Z_{2_{(0.9)}})$ where $Z_{2_{(0.9)}}$ is the 90-percentile for Z_2 (0.82 for models A and B, and 1.86 for models C and D) and $\rho_k = \rho$ for $k < K$
(i) under a complete-case analysis,
(ii) under a standard MI procedure that did not account for outliers and
(iii) under the mixture MI procedure using the value of K that was selected under the AIC-criteria and treating the K mixture class as containing transcription-type outliers to be deleted.

Table 4 shows the results of the simulation study. The AIC- and BIC-criteria performed well for the two-class model but tended to mix the two smallest non-outliers classes for the four-class model; this tendency was more pronounced in BIC than in AIC, and when the proportion of outliers was smaller. When the fraction of outliers was small, both imputation methods correctly estimated the proportion of the Z_{i2} observations above the 90-percentile; when it was large, standard imputation overestimated the fraction belonging to the 90-percentile and above by approximately 30%. The non-outlier correlation was more sensitive to the missingness mechanism and presence of outliers than was the estimate of the proportion above the 90-percentile. As the proportion of outliers increased they overwhelmed the estimation of the common non-outlier correlation; standard imputation corrected this only to a very modest degree. The estimate of the common correlation was essentially unbiased under all four scenarios under the mixture imputation, and the coverage was approximately correct despite the tendency to underestimate the size of the model. This was also true for the proportion above the 90-percentile, with the partial exception of the four-class simulation with 2% clerical outlier contamination: the 8% of simulations in which the BIC-criterion incorrectly suggested the two-class model lead to biased estimation of the proportion of subjects above the true 10-percentile.

A procedure which iteratively removed outliers based on the z-statistic, retaining observations if z-statistics for both variables lie between $z_{0.025/n} = -3.89$ and $z_{1-0.025/n} = 3.89$ and repeating until all observations are retained, reduced the mean-squared error relative to a standard imputation procedure but did not substantially improve the bias or coverage (the results are not shown).

5. Discussion

We have described a method using a latent class model for variability that simultaneously accounts for missing data and clerical error outliers that should be removed. We applied our

Table 4. Simulation study†

	Results for the following simulations:			
	A	*B*	*C*	*D*
AIC				
$K=2$	98	94	1	0
$K=3$	2	5	57	57
$K=4$	0	1	42	43
BIC				
$K=2$	100	100	8	0
$K=3$	0	0	87	97
$K=4$	0	0	3	3
CC				
Mean $\hat{p}$	0.073	0.104	0.079	0.109
Mean-square error $\hat{p}$	9.76×10^{-4}	3.84×10^{-4}	7.40×10^{-4}	4.12×10^{-4}
Nominal 95% coverage $\hat{p}$	62	93	72	95
Mean $\hat{\rho}$	0.21	0.06	0.31	0.11
Mean-square error $\hat{\rho}$	0.163	0.235	0.066	0.171
Nominal 95% coverage $\hat{\rho}$	20	4	25	3
IMP(1)				
Mean $\hat{p}$	0.106	0.131	0.106	0.131
Mean-square error $\hat{p}$	3.15×10^{-4}	12.11×10^{-4}	2.17×10^{-4}	10.13×10^{-4}
Coverage $\hat{p}$	98	74	98	70
Mean $\hat{\rho}$	0.23	0.06	0.33	0.13
Mean-square error $\hat{\rho}$	0.149	0.228	0.057	0.162
Nominal 95% coverage $\hat{\rho}$	22	5	32	4
IMP(2)				
Mean $\hat{p}$	0.099	0.099	0.094	0.100
Mean-square error $\hat{p}$	2.84×10^{-4}	3.02×10^{-4}	4.97×10^{-4}	2.22×10^{-4}
Nominal 95% coverage $\hat{p}$	98	98	94	98
Mean $\hat{\rho}$	0.49	0.49	0.50	0.49
Mean-square error $\hat{\rho}$	0.003	0.003	0.005	0.006
Nominal 95% coverage $\hat{\rho}$	94	95	94	94

†Simulation A, $K=2$, $\sigma_1=1$, $\sigma_2=100$ and $\mathbf{p}=(0.98\ 0.02)^{\mathrm{T}}$; simulation B, $K=2$, $\sigma_1=1$, $\sigma_2=100$ and $\mathbf{p}=(0.90\ 0.10)^{\mathrm{T}}$; simulation C, $K=4$, $\sigma_1=0.25$, $\sigma_2=1$, $\sigma_3=9$, $\sigma_4=100$ and $\mathbf{p}=(0.245\ 0.245\ 0.245\ 0.02)^{\mathrm{T}}$; simulation D, $K=4$, $\sigma_1=0.25$, $\sigma_2=1$, $\sigma_3=9$, $\sigma_4=100$ and $\mathbf{p}=(0.225\ 0.225\ 0.225\ 0.10)^{\mathrm{T}}$. For all simulations, $p=0.1$ and $\rho=0.5$. CC, complete-case analysis; IMP(1), standard MI procedure; IMP(2), mixture MI procedure using the value of K selected under the AIC-criteria. The results are based on 200 independent simulations.

method to generate an unbiased estimate of the prevalence of obesity among children aged 2–11 years who receive care at HSRA-supported community health centres in HSRA regions II and III at least once during calendar year 2001. By developing our method in the context of an MI framework, we allowed estimation to proceed using standard design-based methods for complete-case analyses. We showed that failing to account for the outliers that are caused by clerical errors leads to a modest overestimation of rates of obesity and widths of confidence intervals, particularly among subpopulations for which clerical errors were more likely. The results of our analysis suggested that including transcription errors may have led to modest overestimates of the prevalence of obesity among younger children in selected subregions, but for most subdomains their effect appears to be minimal.

Our method also suggested evidence for at least two classes of variability in the HSRA-served population aged 2–11 years: a normative variance class containing 90–95% of the population and a class overdispersed by a factor of 2–4 containing 5–10% of the population. These overdispersion classes are now accounted for in the modelling procedure and are of clinical interest as well as they are indicative of obesity or malnutrition 'clustering'. A richer data set could determine whether overdispersion might be related to individual or health centre covariates such as income, health status or other socio-economic factors.

The method that was considered here has obviously been tailored to the application, in that we assume that a distinct class of clerical error outliers exists, which we would remove if we could be certain which they were. The method could be used simply to identify cases for further consideration, which, because of issues about returning to the health centres for further data collection, was not considered to be practical here. The model does rely on the transcription or clerical error class being overdispersed relative to the correctly transcribed data, but, if the clerical errors are such that the resulting data are representative of the true population, they presumably have little effect on inference.

Extensions of this approach are possible. The presumed cause for the missingness—failure to collect height data because of recent visits or slower growth among older children—is probably unrelated to the height of the child after conditioning on the weight, centre, gender and age, the last two of which are incorporated in the z-score transformation. Hence the missingness mechanism is arguably ignorable. We could weaken this restriction by postulating a missingness mechanism for height that involved non-identifiable parameters that are a function of height and by conducting a sensitivity analysis that is a function of these parameters. A more fully design consistent model would cross-classify the dispersion classes by the probability of selection, as Elliott and Sammel suggest in their discussion of Patterson *et al.* (2002). Third, rather than the somewhat *ad hoc* method of fixing the number of classes K via a penalized likelihood method and treating it as known, a fully Bayesian method that accommodates uncertainty in the number of classes could be implemented by adding a prior distribution for the total number of classes and adding a model choice step to the Gibbs routine via a product space search (Carlin and Chib, 1995) or reversible jump (Green, 1995) step. Finally, an interesting approach might be to model the clerical error mechanism itself, for example, as transpositions of digits, dropping or adding of digits, misreadings (e.g. '7' as '2') or misinterpretations of units (e.g. pounds as kilograms), rather than as a 'black box' as in the current multiple-edit–MI approaches.

Acknowledgements

The authors acknowledge Daniel Heitjan, Professor of Biostatistics, University of Pennsylvania Medical Centre, along with the Joint Editor, Associate Editor and two reviewers, for their helpful suggestions for the manuscript. This research was supported in part by National Institute of Heart, Lung, and Blood grant R01-HL-068987-01 and National Center for Research Resources grant K23-RR-16073. The authors also thank Dr Steven Auerbach of the HSRA and the community health centre staff who made the 'Healthy for life' project possible.

Appendix A: Data augmentation algorithm

A Gibbs sampling routine that accommodates missing data proceeds by imputing the missing data conditional on the latest draw of the parameters, then drawing the parameters from their posterior distributions conditional on both the observed and the latest draw of imputed data (Li, 1988). Here we consider the missing data not only to include the missing data that in principle could have been observed but also the latent class assignment. Denoting $\Sigma_i I(C_i = k)$ by C^k, the conditional draws are as follows.

(a) Order the elements of $\mathbf{Z}_i$ into the observed components $\mathbf{Z}_i^{\text{obs}}$ and $\mathbf{Z}_i^{\text{mis}}$, with corresponding reordered model parameters conditional on membership in the k variance class $\boldsymbol{\mu}_i^{\text{obs}}$, $\boldsymbol{\mu}_i^{\text{mis}}$, Σ_k^{obs}, Σ_k^{mis} and $\Sigma_k^{\text{mis,obs}}$, the covariance between the unobserved and observed components. Draw $\mathbf{Z}_i^{\text{imp}}$ from a normal distribution with mean $\boldsymbol{\mu}_i^{\text{mis}} + \Sigma_k^{\text{mis,obs}}(\Sigma_k^{\text{obs}})^{-1}(\mathbf{Z}_i^{\text{obs}} - \boldsymbol{\mu}_i^{\text{obs}})$ and variance $\Sigma_k^{\text{mis}} - \Sigma_k^{\text{mis,obs}}(\Sigma_k^{\text{obs}})^{-1} \times \Sigma_k^{\text{obs,mis}}$. $\mathbf{Z}_i^{\text{comp}}$ is then given by replacing $\mathbf{Z}_i^{\text{mis}}$ with $\mathbf{Z}_i^{\text{imp}}$ and returning to the original order of $Z_{i1}, \ldots, Z_{iq}$.

(b) Draw latent class indicator C_i from a multinomial distribution of size 1 and K cells, with probability

$$\pi_{ki} = \frac{p_k |\Sigma_k|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{Z}_i^{\text{comp}} - \boldsymbol{\mu}_i)^{\mathrm{T}} \Sigma_k^{-1} (\mathbf{Z}_i^{\text{comp}} - \boldsymbol{\mu}_i)\}}{\sum_k p_k |\Sigma_k|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{Z}_i^{\text{comp}} - \boldsymbol{\mu}_i)^{\mathrm{T}} \Sigma_k^{-1} (\mathbf{Z}_i^{\text{comp}} - \boldsymbol{\mu}_i)\}}.$$

(c) Draw qth mean $\boldsymbol{\beta}_q$ from a bivarate normal distribution with mean $(\Sigma_{i:i\in q} D_i)^{-1} \Sigma_{i:i\in q} D_i \mathbf{d}_i$ and covariance $(\Sigma_{i:i\in q} D_i)^{-1}$, where

$$D_i = \sum_k I(C_i = k) \Sigma_k^{-1} + V_\beta^{-1}$$

and

$$\mathbf{d}_i = \Big\{ \sum_k I(C_i = k) \Sigma_k^{-1} + V_\beta^{-1} \Big\} \mathbf{Z}_i^{\text{comp}}.$$

(d) Draw latent class marginal probability p_k from a Dirichlet distribution with parameters $C^1 + 1, \ldots, C^K + 1$.

(e) Draw the inverse of the outlier covariance matrix Σ_K^{-1} from an inverse Wishart distribution, with $\text{df} = C^K + 2$ and scale $\{\Sigma_i I(C_i = K)(\mathbf{Z}_i^{\text{comp}} - \boldsymbol{\mu}_i)(\mathbf{Z}_i^{\text{comp}} - \boldsymbol{\mu}_i)^{\mathrm{T}} + S_K\}^{-1}$.

(f) Draw the remainder of the covariance parameters $\boldsymbol{\phi} = (\sigma^2_{111}, \ldots, \sigma^2_{qq(K-1)}, \rho_{12}, \ldots, \rho_{q-1,q})^{\mathrm{T}}$ by using a Metropolis algorithm. Draw a proposal $\boldsymbol{\phi}^*$ from a $(q(K-1) + q(q-1)/2)$-variate normal distribution centred at the current draw $\boldsymbol{\phi}$ and covariance $-cH^{-1}$; accept $\boldsymbol{\phi}^*$ with probability r, where

$$r = \min\left[\frac{\exp\Big\{ -\sum_k \sum_j -\frac{1}{8} \log(\sigma^*_{jjk})^2 + \sum_{i=1}^{n} \sum_{k=1}^{K-1} I(C_i = k)\, l_i(\boldsymbol{\beta}, \boldsymbol{\phi}_k^*) \Big\}}{\exp\Big\{ -\sum_k \sum_j -\frac{1}{8} \log(\sigma_{jjk})^2 + \sum_{i=1}^{n} \sum_{k=1}^{K-1} I(C_i = k)\, l_i(\boldsymbol{\beta}, \boldsymbol{\phi}_k) \Big\}}, 1 \right]$$

for

$$l_i(\boldsymbol{\beta}, \boldsymbol{\phi}_k) = -\log(2\pi) - \tfrac{1}{2} \log|\Sigma_k| - \tfrac{1}{2}(\mathbf{Z}_i^{\text{comp}} - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^{\mathrm{T}} \Sigma_k^{-1} (\mathbf{Z}_i^{\text{comp}} - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta});$$

otherwise retain the current $\boldsymbol{\phi}$. H is given by $\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\phi}) / \boldsymbol{\phi}\boldsymbol{\phi}^{\mathrm{T}}$ for

$$l(\boldsymbol{\beta}, \boldsymbol{\phi}) = \sum_i \sum_k I(C_i = k)\, l_i(\boldsymbol{\beta}, \boldsymbol{\phi}_k),$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ are evaluated at their maximum likelihood estimate, and c is a tuning parameter to adjust the acceptance rate.

To obtain the maximum likelihood estimates under the model for an efficient draw of $\boldsymbol{\phi}$, or as an alternative to the Markov chain Monte Carlo procedure, an EM algorithm (Dempster *et al.*, 1977) for fitting model (1) when missing data are present is available in Elliott (2006). Alternatively, to obtain a less efficient H by using the Markov chain Monte Carlo algorithm only, the posterior distribution of an unconstrained model can be obtained by drawing all values of Σ_k as in step (e). Estimate H by replacing $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ with their posterior means, where the common value of ρ is simply estimated as the mean of the posterior means of ρ_k weighted by the posterior mean of π_k; or, the unconstrained results can be used directly, if preferred.

Because the likelihood is unchanged under the permutation of the class labels k, the labelling of the kth class can change over the length of a single Gibbs chain (Stephens, 2000). To ensure that the kth class is labelled consistently throughout, the Markov chain Monte Carlo draws were discarded unless $|\Sigma_1| < \ldots < |\Sigma_K|$; this was a relatively rare event, required for fewer than 1% of the simulations.

References

Akaike, H. (1978) A Baysian analysis of the minimum AIC procedure. *Ann. Inst. Statist. Math.*, **30**, 9–14.

Barnett, V. and Lewis, T. (1994) *Outliers in Statistical Data*, 3rd edn. New York: Wiley.

Bayarri, M. J. and Morales, J. (2003) Bayesian measures of surprise for outlier detection. *J. Statist. Planng Inf.*, **111**, 3–22.

Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations (with discussion). *J. R. Statist. Soc.* B, **26**, 211–252.

Campbell, N. A. (1980) Robert procedures in multivariate data analysis: I, robust covariance estimation. *Appl. Statist.*, **29**, 231–237.

Carlin, B. P. and Chib, S. (1995) Bayesian model choice via Markov chain Monte Carlo methods. *J. R. Statist. Soc.* B, **57**, 473–484.

Chaloner, K. and Brant, R. (1988) A Bayesian approach to outlier detection and residual analysis. *Biometrika*, **75**, 651–659.

Cole, T. J. (1990) The LMS method for constructing normalized growth standards. *Eur. J. Clin. Nutrn*, **44**, 45–60.

Cole, T. J. (1994) Growth charts for both cross-sectional and longitudinal data. *Statist. Med.*, **13**, 2477–2492.

Cole, T. J., Bellizzi, M. C., Flegal, K. M. and Dietz, W. H. (2000) Establishing a standard definition for child overweight and obesity worldwide: international survey. *Br. Med. J.*, **320**, 1240–1245.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1–38.

Elliott, M. R. (2006) Multiple imputation in the presence of outliers. *Technical Report 06-59*. Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004) *Bayesian Data Analysis*, 2nd edn. Boca Raton: Chapman and Hall–CRC.

Gelman, A., Meng, X.-L. and Stern, H. S. (1996) Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statist. Sin.*, **6**, 733–807.

Ghosh-Dastidar, M. and Schafer, J. L. (2003) Multiple edit multiple imputation for multivariate continuous data. *J. Am. Statist. Ass.*, **98**, 807–817.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Hadi, A. S. (1992) Identifying multiple outliers in multivariate data. *J. R. Statist. Soc.* B, **54**, 761–771.

Hawkins, D. M. (1980) *Identification of Outliers*. London: Chapman and Hall.

Hedley, A. A., Ogden, C. L., Johnson, C. L., Carroll, M. D., Curtin, L. R. and Flegal, K. M. (2004) Prevalence of overweight and obesity among US children, adolescents, and adults. *J. Am. Med. Ass.*, **291**, 2847–2850.

Kimm, S. Y. and Obarzanek, E. (2002) Childhood obesity: a new pandemic of the new millennium. *Pediatrics*, **110**, 1003–1007.

Koplan, J. P., Liverman, C. T. and Kraak, V. A. (eds) (2004) *Preventing Childhood Obesity: Health in the Balance*. Washington DC: National Academies Press.

Kuczmarski, R. J., Ogden, C. L., Grummer-Strawn, L. M., Flegal, K. M., Guo, S. S., Wei, R., Mei, Z., Curtin, L. R., Roche, A. F. and Johnson, C. L. (2000) CDC growth charts: United States. In *Advance Data from Vital and Health Statistics*, no. 314. Hyattsville: National Center for Health Statistics.

Leonard, M. B., Feldman, H. I., Shults, J., Zemel, B., Foster, B. J. and Stallings, V. A. (2004) Long-term, high-dose glucocorticoids and bone mineral content in childhood glucocorticoid-sensitive nephrotic syndrome. *New Engl. J. Med.*, **351**, 868–875.

Li, K. H. (1988) Imputation using Markov Chains. *J. Statist. Computn Simuln*, **30**, 57–79.

Little, R. J. A. (1988) Robust estimation of the mean and covariance matrix from data with missing values. *Appl. Statist.*, **37**, 23–38.

Little, R. J. A. and Smith, P. J. (1987) Editing and imputation for quantitative survey data. *J. Am. Statist. Ass.*, **82**, 58–68.

Meng, X. L. (1994) Multiple imputation inferences with uncongenial sources of input (with discussion). *Statist. Sci.*, **9**, 538–573.

Ogden, C. L., Flegal, K. M., Carroll, M. D. and Johnson, C. L. (2002) Prevalence and trends in overweight among US children and adolescents, 1999-2000. *J. Am. Med. Ass.*, **288**, 1728–1732.

Patterson, B. H., Dayton, C. M. and Graubard, B. I. (2002) Latent class analysis of complex sample survey data: application to dietary data (with discussion). *J. Am. Statist. Ass.*, **97**, 721–729.

Penny, K. I. and Jolliffe, I. T. (1999) Multivariate outlier detection applied to multiply imputed laboratory data. *Statist. Med.*, **18**, 1879–1895.

Rousseeuw, P. J. and van Zomeren, B. C. (1990) Unmasking multivariate outliers and leverage points (with comments). *J. Am. Statist. Ass.*, **85**, 633–651.

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

SAS Institute (2001) *SAS Version 8.2*. Cary: SAS Institute.

Schafer, J. L. (1997) *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.

Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.
Stephens, M. (2000) Dealing with label switching in mixture models. *J. R. Statist. Soc.* B, **62**, 795–809.
Stettler, N., Elliott, M. R., Kallan, M. Auerbach, S. B. and Kumanyika, S. K. (2005) High prevalence of pediatric overweight in medically underserved areas. *Pediatrics*, **116**, 381–388.
Teicher, H. (1963) Identifyability of finite mixtures. *Ann. Math. Statist.*, **34**, 1265–1269.
Weiss, R., Dziura, J., Burgert, T. S., Tamborlane, W. V., Taksali, S. E., Yeckel, C. W., Allen, K., Lopes, M., Savoye, M., Morrison, J., Sherwin, R. S. and Caprio, S. (2004) Obesity and the metabolic syndrome in children and adolescents. *New Engl. J. Med.*, **350**, 2362–2374.
Woodruff, R. S. (1971) A simple method for approximating the variance of a complicated estimate. *J. Am. Statist. Ass.*, **66**, 411–414.