Multilevel Multiple Imputation in presence of interactions, non-linearities and random slopes

Imputazione Multipla Multilivello in presenza di interazioni, non-linearità e pendenze casuali

Matteo Quartagno and James R. Carpenter

Abstract Multiple Imputation is a flexible tool to handle missing data that has been increasingly used in recent years. One of the conditions for its validity is that the two models used for (i) imputing and (ii) analysing the data need to be compatible. For example, when the partially observed data have a multilevel structure, both models need to reflect this. Choosing an appropriate imputation model is more complicated when data are missing in a variable included in the substantive multilevel analysis model as a covariate with a random slope, an interaction or a non-linear term. We propose an imputation method based on joint modelling of the partially observed variables. We factor this joint model in two parts: a joint multilevel distribution for the covariates, and a conditional multilevel distribution for the outcome given the covariates. We guarantee compatibility by using as the second term the substantive analysis model. We fit this model with a Gibbs sampler, and we use a Metropolis-Hastings step to accept/reject the proposed draws for the missing values, to guarantee that they are actual random draws from the desired distribution. Our proposed imputation approach is theoretically consistent with the substantive model, and we demonstrate the marked improvements this brings by simulation.

Abstract *L'imputazione multipla é uno strumento flessibile per gestire dati mancanti la cui popolarità è aumentata considerevolmente negli ultimi anni. Una delle condizioni necessarie per la sua validità è che i due modelli utilizzati per (i) imputare e (ii) analizzare i dati siano compatibili. Per esempio, quando il dataset parzialmente osservato ha una struttura multilivello, entrambi i modelli devono tenerne conto. Scegliere un modello di imputazione adeguato è più complicato quando i dati sono mancanti in una variabile che è inclusa nel modello d'analisi multilivello di interesse come una covariata con una pendenza casuale, un'interazione o un*

Matteo Quartagno
London School of Hygiene and Tropical Medicine, Keppel Street, e-mail: matteo.quartagno@lshtm.ac.uk

James R. Carpenter
London School of Hygiene and Tropical Medicine, Keppel Street e-mail: james.carpenter@lshtm.ac.uk

termine non-lineare. Proponiamo qui un metodo di imputazione basato sulla modellizzazione congiunta delle variabili parzialmente osservate. Fattorizziamo questo modello congiunto in due parti: una distribuzione congiunta per le covariate ed una distribuzione condizionata per la variabile risposta date le covariate. Garantiamo la compatibilità usando per questo secondo termine la stessa formulazione del modello d'analisi di interesse. Fittiamo questo modello con un campionatore di Gibbs, ed utilizziamo un passo Metropolis-Hastings per accettare/rifiutare i valori proposti per i dati mancanti, per garantire che siano effettive estrazioni casuali dalla distribuzione desiderata. Mostriamo con simulazioni che questo metodo performa in modo appropriato e supera una strategia di imputazione alternativa.

Key words: Missing Data, Multiple Imputation, Multilevel Model, Random Slope, Interactions, Non-linearities

1 Introduction

Multiple imputation (MI) is a missing data handling method that has become very popular in recent years, particularly in the world of medical and social research. Key reasons for its growing popularity include its flexibility, the possibility to use for the analysis step the same model of substantive scientific interest that we would have used on a fully observed dataset and the chance to make use of auxiliary variables to retrieve some information [8, 3].

A key role in MI is played by the *imputation model*, i.e. the model that is used to impute the missing data. In order for MI to lead to valid inference, this needs to be consistent with the substantive analysis model [2]. For example, if the partially observed dataset has a multilevel structure, this needs to be reflected in the imputation model as well as in the analysis model [6].

Different methods have been proposed recently for multilevel MI [1]. The most flexible of these is Joint Modelling Multiple Imputation (JM-MI), which consists in assuming a joint multivariate normal model for the partially observed data, and in fitting this model with a Bayesian (e.g. Gibbs) sampler to impute the missing data. A multilevel version of JM-MI was first introduced in [9], and later extended to allow for binary and categorical data [5] and for cluster-specific covariance matrices [10]. However, in some circumstances it is not possible to find a simple joint imputation model that is fully compatible with the analysis model; some examples include the imputation of variables that are included in the substantive analysis model as covariates with a random slope, an interaction or a non-linear (e.g. quadratic) term. Using a heteroscedastic imputation model can be useful to deal with random slopes and interactions, as it allows for cluster-specific associations between variables [7]. However, full compatibility is still not guaranteed.

Goldstein et al. (2014) proposed a fully bayesian approach that broadly consists in factoring the joint distribution in two terms: a joint model for the covariates of the analysis model and a conditional model for the outcome given the covariates, that usually corresponds with the substantive analysis model. Although it was proposed as a fully bayesian method, it can be used as a multiple imputation approach compatible with the substantive model. The advantage of this is that it allows auxiliary variables to be included.

The aim of this paper is to introduce substantive model compatible JM-MI, and to compare it with standard JM-MI, when the substantive analysis model includes a random slope, an interaction or a quadratic term. We illustrate the advantage of the newly proposed method by simulations.

2 Methods

Assume we have a partially observed dataset with individuals i nested in clusters j. We intended to collect data on three continuous variables Y, X_1 and X_2, but we end up with some missing data in each of the three variables. The substantive analysis model of scientific interest is a linear mixed model:

$$\begin{aligned} y_{i,j} &= (\beta_0 + u_{0,j}) + (\beta_1 + u_{1,j})x_{1,i,j} + \beta_2 x_{2,i,j} + \varepsilon_{i,j} \\ \begin{pmatrix} u_{0,j} \\ u_{1,j} \end{pmatrix} &\sim N(\mathbf{0}, \Sigma_u) \qquad \varepsilon_{i,j} \sim N(0, \sigma_e^2) \end{aligned} \tag{1}$$

In order to deal with missing data, we can use JM-MI. But what imputation model should we use?

2.1 JM-Hom: Homoscedastic Joint Modelling Imputation

One possibility is to assume a 3-variate normal joint model for the three variables:

$$\begin{cases} y_{i,j} = \alpha_0 + v_{0,j} + e_{0,i,j} \\ x_{1,i,j} = \alpha_1 + v_{1,j} + e_{1,i,j} \\ x_{2,i,j} = \alpha_2 + v_{2,j} + e_{2,i,j} \end{cases}$$
$$\begin{pmatrix} v_{0,j} \\ v_{1,j} \\ v_{2,j} \end{pmatrix} \sim N(\mathbf{0}, \Omega_u) \qquad \begin{pmatrix} e_{0,i,j} \\ e_{1,i,j} \\ e_{2,i,j} \end{pmatrix} \sim N(\mathbf{0}, \Omega_e) \tag{2}$$

This model can be easily fitted with a standard Gibbs sampler, creating K different imputed datasets. These are then analysed with (1) to obtain K copies of the parameter estimates that are finally combined with Rubin's rules.

This approach naturally extends to include binary or categorical variables. This is achieved by means of a latent normal variables approach, as outlined in [5].

2.2 JM-Het: Heteroscedastic Joint Modelling Imputation

Because of the presence of a random slope, Model (1) is not compatible with Model (2), i.e. the conditional distribution of Y given X_1 and X_2 derived from (2) is not (1). To overcome this issue, one possibility is to assume instead an heteroscedastic imputation model, similar to (2) but with random cluster-specific covariance matrices following an inverse Wishart distribution:

$$\Omega_{e,j} \sim IW(a,A) \tag{3}$$

This model makes an attempt at modelling cluster-specific associations between variables, by assuming cluster-specific covariance matrices at level 1. However, it is still not a fully compatible approach.

It can be fitted with a similar Gibbs sampler to the one used for model (2).

2.3 JM-SMC: Substantive Model Compatible Joint Modelling Imputation

In order to define an imputation model fully compatible with (1), following along the lines of [4], we can factorise the joint distribution of the three variables in two terms: (i) a joint model for the two covariates and (ii) a conditional model for the outcome given the covariates. This way, we can make sure that the conditional model for the outcome corresponds to (1):

$$\begin{gathered}
\begin{cases} x_{1,i,j} = \alpha_1 + v_{1,j} + e_{1,i,j} \\ x_{2,i,j} = \alpha_2 + v_{2,j} + e_{2,i,j} \end{cases} \\
\begin{pmatrix} v_{1,j} \\ v_{2,j} \end{pmatrix} \sim N(\mathbf{0},\Omega_u) \qquad \begin{pmatrix} e_{1,i,j} \\ e_{2,i,j} \end{pmatrix} \sim N(\mathbf{0},\Omega_e) \\
y_{i,j} = (\beta_0 + u_{0,j}) + (\beta_1 + u_{1,j})x_{1,i,j} + \beta_2 x_{2,i,j} + \varepsilon_{i,j} \\
\begin{pmatrix} u_{0,j} \\ u_{1,j} \end{pmatrix} \sim N(\mathbf{0},\Sigma_u) \qquad \varepsilon_{i,j} \sim N(0,\sigma_e^2)
\end{gathered} \tag{4}$$

In order to impute from this model, an additional Metropolis-Hastings step within the Gibbs sampler is needed to impute the missing X_1 and X_2 values: imputations are drawn from a proposal distribution and accepted or rejected depending on the value of the Metropolis ratio. If using a symmetrical proposal distribution, this is simply the ratio of the likelihood of the model with the new proposed imputed value over the likelihood of the model with the previous imputed value.

This method naturally extends to allow for interactions or non-linearities in the conditional model for the outcome given the covariates in (4). Hence, it is possible to impute compatibly with the analysis model at the only cost of having to know the functional form of the substantive model in advance.

2.4 Software

We fit and impute from all three models (2), (3) and (4) using functions jomo and jomo.smc from our R package jomo, freely available on CRAN. The substantive model (1) is fitted with the R package lme4, and the results are combined with Rubin's rules as implemented in the mitml package.

3 Simulations

To illustrate the improvements that random coefficient compatible multiple imputation brings, in a base-case scenario we generate 1000 multilevel datasets, each constituted of 6000 observations, equally divided in 60 clusters, on three variables Y, X_1 and X_2. X_1 and a latent normal Z are generated from a bivariate normal distribution. X_2 is then created as a binary variable that takes the value 1 when $Z > 0$. The data-generating mechanism for the outcome is the following:

$$y_{i,j} = (0.5 + u_{0,j}) + (1 + u_{1,j})x_{1,i,j} - 0.3x_{2,i,j} + \varepsilon_{i,j}$$
$$\begin{pmatrix} u_{0,j} \\ u_{1,j} \end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \qquad \varepsilon_{i,j} \sim N(0,1)$$

We assume that the desired analysis model is (1). We fit this model on the fully observed data (FD) and store all the parameter estimates. Then, for the fixed effect parameters β_0, β_1 and β_2 we calculate the mean, the empirical and model based standard errors and the coverage level across 1000 simulations. We additionally report the mean of the three variance components: the random intercept variance σ^2_{u0}, the random slope variance σ^2_{u1} and the residual variance σ^2_e.

We then make around 35% of the data on X_1 and X_2 missing at random conditional on the outcome Y, and we re-analyse the data using the complete records (CR). Finally, we handle missing data with the three different MI strategies presented in the previous section.

We investigate four additional scenarios:

- A scenario where X_2 is 3-level categorical;
- One with an additional continuous variable X_3, highly correlated with X_1, but not included either in the data generating process for Y or in the substantive analysis model (i.e. an auxiliary variable);
- One where Y is generated from a model with a quadratic effect on X_1;
- One where there is an interaction between X_1 and X_2.

3.1 Results

Table 1 shows the base-case simulation results. While Complete Records (CR) estimates are strongly biased, because of the dependence of the missingness mechanism from the outcome Y, all imputation methods are preferable; however, JM-Hom also leads to biased estimates and marked undercoverage for most parameters. Inference on the fixed effect parameters after imputation with JM-Het is affected by smaller biases, and leads to good coverage levels. However, bias is larger in the estimation of the variance components. Finally, JM-SMC leads to unbiased parameter estimates and good coverage levels.

Table 1 Base-case scenario: mean and coverage level of fixed effect parameter estimates and mean of variance component estimates over 1000 simulations. We compare Full Data (FD), Complete records (CR), JM imputation with a homoscedastic (JM-Hom) or heteroscedastic (JM-Het) imputation model and substantive model compatible JM-MI (JM-SMC).

	β_0		β_1		β_2		σ^2_{u0}	σ^2_{u1}	σ^2_e
Method	Mean	Cov	Mean	Cov	Mean	Cov	Mean	Mean	Mean
True value	0.50	0.95	1.00	0.95	-0.30	0.95	1.00	1.00	1.00
FD	0.49	0.94	1.00	0.93	-0.30	0.96	0.99	1.01	1.00
CR	1.08	0.00	0.89	0.82	-0.25	0.72	0.57	0.79	0.92
JM-Hom	0.40	0.88	1.07	0.85	-0.28	0.91	1.12	0.54	1.25
JM-Het	0.44	0.92	1.04	0.92	-0.29	0.94	1.09	0.93	1.05
JM-SMC	0.49	0.94	1.00	0.93	-0.30	0.95	0.99	1.01	1.00

Figure 1 pools the results across all the five simulation scenarios into a single panel. This is in terms of relative bias and coverage level for all the fixed effect parameter estimates. While JM-SMC always leads to negligible bias and coverage very close to the nominal level, JM-Hom and JM-Het are prone to bias in the estimation of some parameters in all scenarios. This is particularly serious for the two scenarios with an interaction and a quadratic effect. CR estimates are again always the most seriously biased, because of the missing data mechanism.

Finally, Figure 2 compares the level-2 variance component estimates from the three MI methods. Once again, JM-SMC is the only method leading to unbiased parameter estimates, while JM-Hom is the worst imputation method. JM-Het consistently overestimates the random intercept variance and gives biased estimates of the random slope variance.

4 Conclusions

We have investigated the behaviour of a new substantive model compatible MI strategy to deal with missing data in a multilevel dataset, and compared it with two existing multilevel imputation strategies. In particular, we have showed that when the analysis model of scientific interest includes a random slope, an interaction or a non-linearity, our proposed new method is the only one able to take this into account during the imputation, leading to correct inference.

The only additional price to pay for using this method, is that the precise functional form of the substantive model needs to be known in advance of the imputation process. Further research will investigate what is the best approach to take when model selection has to be performed along the imputation. Future work will also explore ways to impute level-2, i.e. cluster-level, variables within this framework.

All the imputation models presented in this paper can be fitted with functions jomo and jomo.smc in the R package jomo. This allows for binary and survival outcomes as well.

In conclusion, while standard JM-MI remains a valuable, and more flexible, method for the imputation of simple multilevel dataset, substantive model compatible JM-MI is preferable in presence of partially observed covariates with a random slope, an interaction or a non-linear term.

Acknowledgements This work was supported by the Medical Research Council, grant numbers MC UU 12023/21 and MC UU 12023/29.

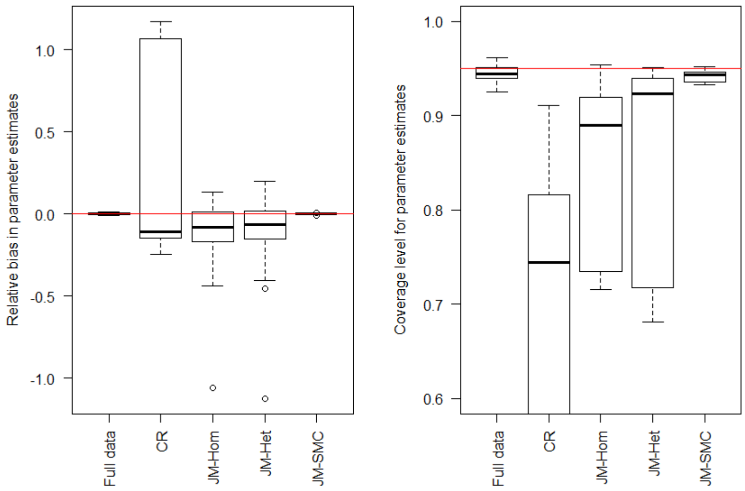

Fig. 1 Boxplots summarising results of five simulation scenarios. We compare relative bias (left panel) and coverage level (right panel) of fixed effect parameter estimates. The red lines indicate 0% relative bias and 95% coverage level. We compare Full Data (FD), Complete records (CR) and the three MI strategies.

References

1. Audigier V., White I.R., Debray T., Jolani S., Quartagno M., Carpenter J.R., van Buuren S., Resche-Rigon M. Stat Science. In press (2018)
2. Bartlett J. W., Seaman S.R., White I.R., Carpenter J.R., for the Alzheimers Disease Neuroimaging Initiative*. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. Kenward MG, ed. Statistical Methods in Medical Research. 24(4):462-487 (2015)
3. Carpenter, J. R., Kenward M. G.: Multiple Imputation and its Application. Wiley, Chichester (2013)
4. Goldstein H., Carpenter J.R., Browne W.J., Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. Journal of RSS Series A. 177(2), 553-564 (2014)
5. Goldstein H., Carpenter J.R., Kenward M.G., and Levin K.A. Multilevel models with multivariate mixed response types. Statistical Modelling. 9(3), 173 – 197 (2009).
6. Lüdtke, O., Robitzsch, A. and Grund, S., Multiple imputation of missing data in multilevel designs: a comparison of different strategies, Psychological Methods. 22(1): 141–165 (2017)
7. Quartagno M., Carpenter J.R., Multiple Imputation of IPD Meta-analysis: allowing for heterogeneity and studies with missing covariates. Statistics in Medicine. 35(17), 2938–54 (2016)
8. Rubin, D.: Multiple Imputation for non-response in Surveys - a phenomenological bayesian approach to nonresponse. Wiley, New York (1987)
9. Schafer, J. L., and R. M. Yucel. Computational Strategies for Multivariate Linear Mixed-Effects Models with Missing Values. Journal of Computational and Graphical Statistics. 11(2) , 437–57 (2002).
10. Yucel, R. M. Random-Covariances and Mixed-Effects Models for Imputing Multivariate Multilevel Continuous Data. Statistical modelling. 11(4), 351-370 (2011)

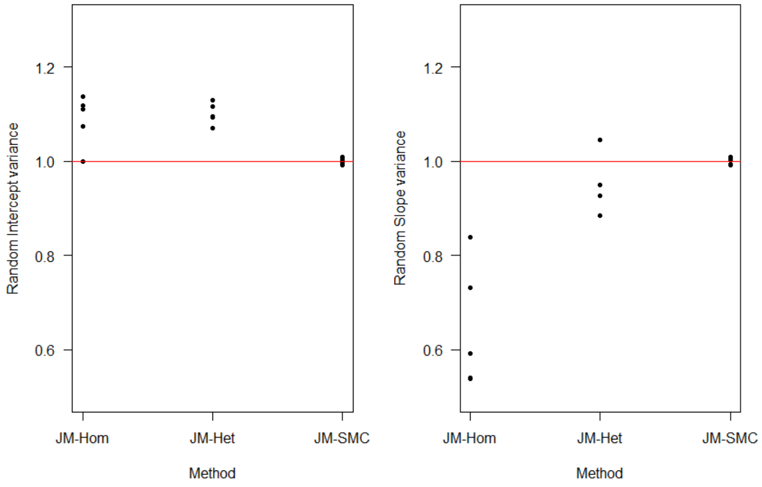

Fig. 2 Comparison of random intercept (left panel) and slope (right panel) variance estimates with the three MI strategies. Each point represents a different scenario. The red line indicates the correct value of 1.