Brief comments on computational issues with multiple imputation
James Carpenter and Mike Kenward
September 2008

## 1 Introduction

Multiple Imputation (MI) is a Monte-Carlo (i.e. stochastic) method for parameter estimation in partially observed datasets. It is now being increasingly used in practice. This document highlights some of the computational issues that may arise.

MI is usually performed under the assumption that the mechanism causing the missing data is 'Missing At Random'. We discuss the practical implications of this elsewhere (Carpenter and Kenward, 2008). Here we note that although this assumption may be plausible, it cannot be verified from the data at hand, and therefore the analysis of a partially observed data set under the MAR assumption can never have the same status as the analysis of the fully observed dataset would have had. Thus, it is helpful to present any analysis carried out on the partially observed data (e.g. using MI) alongside the analysis based on only those units/individuals with no missing data (so called 'Complete Cases' (CC)), to see how the conclusions differ. If there should be differences, it is then important, if possible, to provide an explanation for these, as this increases confidence in any conclusions drawn.

Of course MI and CC analysis can often be different because the mechanism causing the missing data is MAR, and CC is only generally valid if data are MCAR. In this case it is useful to explain the visible ways in which the mechanism departs from MCAR, and how this distorts the results of a CC analysis.

The purpose of this document is to briefly mention some computational issues which could invalidate the results of an MI analysis. These should be checked when preparing a MI analysis for publication.

## 2 Convergence

In order to impute completed datasets, the imputation model is typically fitted to all the observed data (partial, as well as complete, cases). This is usually done by specifying (either explicitly, or implicitly[1]) a joint model for the observed data where the partially observed variables are responses.

Most software packages set the imputation model in a Bayesian framework, then fit it using Markov Chain Monte Carlo (MCMC) methods[2], and draw the imputed data from the resulting posterior distributions of the unobserved measurements.

Fitting models using MCMC methods requires care. First the simulation process needs to be run a number of times till it converges to the correct distribution (termed the 'burn in'). Second, the simulation process needs to be run a number of times between drawing imputations, so that — informally speaking — the imputed data sets are sufficiently different.

There is a large literature on fitting Bayesian models by MCMC methods (Gilks *et al.*, 1996), and this needs to be kept in mind when performing MI. However, many software packages downplay this aspect, focusing instead on specifying the imputation model.

[1] Such as using the chained equations approach, also referred to as the full conditional specification

[2] These are simulation based methods for obtaining parameter estimates in Bayesian models

However, with complex imputation models and/or large data sets, the default settings in software packages may be inappropriate. If possible, graphical diagnostics should be examined to check the process is working satisfactorily.

## 3 Distributional assumptions

In statistical modelling, considerable care is usually taken to ensure that the distributional model for the data is adequate. For example, binary data is rarely analysed using a linear regression. However, the imputation model is often based on a multivariate normal distribution. In other words, a mix of quantitative, ordinal and categorical data will be treated as joint multivariate normal for the purpose of imputation.

The extent to which this will invalidate the results is difficult to quantify, as it depends critically on the main analysis and the pattern and extent of missing data. A number of simulation studies have appeared which suggest that it is not as misleading as it might appear when:

1. the extent of missing information in a non-quantitative variables is not too great.

   In this case, multiple imputation is allowing the inclusion in the analysis of a large number of individuals with a few missing observations, so that very accurate predictions of their missing data are not necessary

2. for binary/ordinal data the probabilities are not too close to 0 or 1.

   In this setting, logistic/probit analysis is reasonably close to linear regression.

In practice, one should check whether the results of MI are resting inappropriately on these assumptions. Analysts should be on the look out for binary or ordinal variables with a high proportion of identical values, especially if their effect changes markedly between the CC and MI analysis.

An attraction of MI is that data can be transformed before imputation to be more nearly normal, and back transformed after imputation for the analysis of interest. Some recent work, informally reported to us, confirms that this generally reduces bias and improves statistical properties. Both skew quantitative and ordinal variables can benefit from transformation before imputation. Where possible, categorical variables can benefit from re-ordering so they are closer to ordinal. In practice the problem of ill specified imputation distributions is also likely to be greater when the aim of the analysis is to estimate properties of distributions such as means and percentiles, rather than regression coefficients.

This problem may also be alleviated by using the so called 'chained equations' or 'full conditional specification' imputation methods. These methods use appropriate conditional models for each variable (i.e. logistic for binary variables and so on). Quantitative variables can still benefit from transformation to approximate normality.

## 4 Perfect prediction

Handling the imputation of a large number of categorical variables can be tricky (what ever approach is used) because of problems of perfect prediction. If near perfect prediction occurs, fitted values for a particular subset of the data become very close to 1 (or 0), and their standard errors also become

very large. This can result in a non-trivial fraction of the imputed data being inappropriate. This may result in increased variance of parameter estimates from CC to MI, and unexpected changes in direction.

While these problems are easy to spot in more simple problems, this is not the case in complex settings. Further many software packages suppress the information needed to spot this. Some work has been done to address this issue semi-automatically during imputation; nevertheless analysts need to be aware of the potential for errors in these situations.

## 5 Out of sample imputation

This final issue occurs when the majority of the missing values occur in a particular area of the data set (for example at later waves in a longitudinal study, or among particular socio-economic groups).

In this case, the imputation model, which is by necessity fitted to the observed data, may not be an appropriate choice for imputation. The problem is analogous to making predictions 'out of the original sample' in regression. For instance, we estimate the relationship between age and lung function in adults, and then us this to impute missing lung function in children.

Again, most software packages used for imputation will not flag this up as an issue, so the analyst needs to be aware of it. The issue is clearly more acute the higher the proportion of missing data, and the greater the focus in the model of interest on contrasts between well observed and partially observed parts of the dataset.

## 6 Implications

It is not always realised that MI involves statistical modelling that is an order of magnitude more complicated that the original model of interest, because it always involves (either explicitly or implicitly) a joint model for the data.

Fitting such models is not straightforward: there are a number of pitfalls, of which the above are the most common in our experience.

The extent to which these pitfalls will mislead depends on both the model of interest and the degree of missing information. In practice therefore, it is a good to

1. compare the results of MI and the CC analysis, to check that differences between them are plausible given the scentific context and the likely missing data mechanisms, and
2. consider checking the points discussed above, possibly redoing the analysis in a different package to give greater confidence in the results.

One issue we have not touched upon here that is also important, is that the structure of the model of interest (response, interactions, non-linearities, hierarchical aspects) needs to be included in the imputation model for valid results (King *et al.*, 2001).

Although these caveats may appear discouraging, we believe multiple imputation has a key role to play in the analysis of large, complex datasets with missing data. However, like all powerful tools, users need to be aware of the risks and how to minimse them, if they are to reap the full benefits.

## References

Carpenter, J. R. and Kenward, M. G. (2008) *Missing data in clinical trials — a practical guide*. Birmingham: National Health Service Co-ordinating Centre for Research Methodology. Free from http://www.pcpoh.bham.ac.uk/publichealth/methodology/projects/RM03_JH17_MK.shtml.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996) *Markov chain Monte-Carlo in practice*. London: Chapman and Hall.

King, G., Honaker, J., Joseph, A. and Scheve, K. (2001) Analyzing incomplete political science data: an alternative algorithm for multiple imputation. *The American Political Science Review*, **95**, 49–69.