# 1 Imputation strategies for multivariate data

Multiple imputation for multivariate data comes in two main flavors: joint modeling (JM) and fully conditional specification (FCS). With JM, imputations are drawn from an assumed joint multivariate distribution. Often a multivariate normal model is used for both continuous and categorical data, although other joint models have been proposed (see e.g. Olkin and Tate, 1961; Van Buuren and van Rijckevorsel, 1992; Schafer, 1997; Van Ginkel et al., 2007; Goldstein et al., 2009; Chen et al., 2011). Joint modeling imputations generated under the normal model are usually robust to misspecification of the imputation model (Schafer, 1997; Demirtas et al., 2008), although transformation towards normality is generally beneficial.

Contrary to JM, multiple imputation by means of FCS does not start from an explicit multivariate model. With FCS, multivariate missing data is imputed by univariately specifying an imputation model for each incomplete variable, conditional on a set of other (possibly incomplete) variables. The multivariate distribution for the data is thereby implicitly specified through the univariate conditional densities and imputations are obtained by iterating over the conditionally specified imputation models.

The general idea of using conditionally specified models to deal with missing data has been discussed and applied by many authors (see e.g. Kennickell, 1991; Raghunathan and Siscovick, 1996; Oudshoorn et al., 1999; Brand, 1999; Van Buuren et al., 1999; Van Buuren and Oudshoorn, 2000; Raghunathan et al., 2001; Faris et al., 2002; Van Buuren et al., 2006). Comparisons between JM and FCS have been made that indicate that FCS is a useful and flexible alternative to JM when the joint distribution of the data is not easily specified (Van Buuren, 2007) and that similar results may be expected from both imputation approaches (Lee and Carlin, 2010).

In this dissertation, new methodology based on FCS is introduced, although comparisons are occasionally made to imputation approaches that utilize some form of joint modeling. The choice for FCS is based on applicability, by avoiding the complex specification and estimation of multivariate models that observe different kinds of restrictions. Because the multidimensional imputation problem is split in multiple unidimensional imputation problems, it is relatively simple to specify imputation models that do not conform to standard multivariate distributions. Moreover, this flexibility in specifying univariate imputation models makes it much easier to adapt imputation models to accommodate for some form of restriction. As a result, the incomplete data can be more efficiently addressed and unique data features can be preserved. For example, in official statistics many restrictions are posed on survey or register data, such as bounds (no unrealistic human age), strict non-negativity (no negative incomes) and conditional restrictions (girls under twelve years of age are not allowed to have children, nor can they be married).

# 2 Current modeling practice

A straightforward implementation of FCS (for more detail on FCS, see Section 1) can be found in the MICE algorithm proposed by Van Buuren and Groothuis-Oudshoorn (2000, 2011). The MICE algorithm is a Markov Chain Monte Carlo

(MCMC) method, which becomes a Gibbs sampler in situations where the conditional densities are said to be compatible. Compatibility is reached when the joint multivariate distribution has the separate conditional distributions as its conditional densities. For the MICE algorithm, the joint distribution is only implicitly known and compatibility may be difficult to prove. In some situations, compatibility may not actually exists. However, in practice FCS seems to be robust when compatibility conditions are not met (Van Buuren et al., 2006). Recently, Bartlett et al. (2014) introduced a substantive model compatible FCS (SMC-FCS) that ensures that each covariate is imputed from a model which is compatible with the substantive model. This may be particularly of interest when the substantive analysis model contains non-linearities or interactions.

The MICE algorithm starts with randomly drawing imputations from the observed data. Subsequently, the variables are imputed in a variable-by-variable approach. A single iteration of the algorithm cycles through all incomplete variables.

The number of iterations for the MICE algorithm has to be carefully chosen. In most situations, a low number of iterations appears to be enough (Brand, 1999; Van Buuren et al., 1999), but slow convergence can occur if, for example, the amount of missing data is large or if there is high autocorrelation in the imputation chains. After imputation, convergence of the $m$ multiple imputation chains should be investigated.

The number of imputations is also of importance when doing multiple imputation. Usually, the default amount of imputations in software is set to be as low as three to five. Many authors have investigated the role of $m$ with regard to several criteria, such as the confidence interval, statistical power and the proportion of missingness attributable to the nonresponse (see e.g. Royston, 2004; Graham et al., 2007; Bodner, 2008; White et al., 2011). The work by these authors suggests that it may often be beneficial to set the amount of imputations much larger, although it comes at a cost in terms of data storage and computational time.

In general it holds that using a higher $m$ is always better. This does not necessarily mean that outcomes from resulting analyses will be better. In fact, Schafer (1997) suggests that resources can often be better spent and Schafer and Olsen (1998) indicate that in most situations there is only little advantage to analyzing more than a few imputed datasets. To save computation time and resources, Van Buuren (2012) suggests to set $m = 5$ during model building and to increase $m$ only for the 'actual' imputation stage. However, with computers becoming increasingly faster and data storage solutions becoming more accommodative of large datasets, one can imagine that today's drawbacks in performing more imputations are becoming increasingly less important in the future.

# References

Bartlett, J. W., Seaman, S. R., White, I. R., Carpenter, J. R., and for the Alzheimer's Disease Neuroimaging Initiative* (2014). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, Advance online publica-

tion. Available from: http://smm.sagepub.com/content/early/2014/03/31/0962280214521348.abstract.

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling*, 15(4):651–675.

Brand, J. P. (1999). *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets*. Erasmus MC: University Medical Center Rotterdam.

Chen, H. Y., Xie, H., and Qian, Y. (2011). Multiple imputation for missing values through conditional semiparametric odds ratio models. *Biometrics*, 67(3):799–809.

Demirtas, H., Freels, S. A., and Yucel, R. M. (2008). Plausibility of multivariate normality assumption when multiply imputing non-gaussian continuous outcomes: a simulation assessment. *Journal of Statistical Computation and Simulation*, 78(1):69–84.

Faris, P. D., Ghali, W. A., Brant, R., Norris, C. M., Galbraith, P. D., and Knudtson, M. L. (2002). Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses. *Journal of Clinical Epidemiology*, 55(2):184–191.

Goldstein, H., Carpenter, J., Kenward, M. G., and Levin, K. A. (2009). Multilevel models with multivariate mixed response types. *Statistical Modelling*, 9(3):173–197.

Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007). How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention Science*, 8(3):206–213.

Kennickell, A. B. (1991). Imputation of the 1989 survey of consumer finances: Stochastic relaxation and multiple imputation. Technical report, mimeo, Board of Governors of the Federal Reserve System.

Lee, K. J. and Carlin, J. B. (2010). Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology*, 171(5):624–632.

Olkin, I. and Tate, R. (1961). Multivariate correlation models with mixed discrete and continuous variables. *The Annals of Mathematical Statistics*, 32(2):448–465.

Oudshoorn, C. G. M., Buuren, S., and Rijckevorsel, J. L. A. (1999). *Flexible multiple imputation by chained equations of the AVO-95 survey*. TNO Prevention and Health Leiden.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology*, 27(1):85–96.

Raghunathan, T. E. and Siscovick, D. S. (1996). A multiple-imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives. *Applied Statistics*, 45(3):335–352.

Royston, P. (2004). Multiple imputation of missing values. *Stata Journal*, 4:227–241.

Schafer, J. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall/CRC.

Schafer, J. L. and Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate behavioral research*, 33(4):545–571.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, 16(3):219–242.

Van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Chapman & Hall/CRC.

Van Buuren, S., Boshuizen, H., and Knook, D. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18(6):681–694.

Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C., and Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation*, 76(12):1049–1064.

Van Buuren, S. and Groothuis-Oudshoorn, C. (2011). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3):1–67.

Van Buuren, S. and Oudshoorn, C. (2000). Multivariate imputation by chained equations: Mice v1. 0 user's manual. *The Netherlands: TNO Report PG/VGZ/00.038. Netherlands Organization for applied scientific research.*

Van Buuren, S. and van Rijckevorsel, J. L. (1992). Imputation of missing categorical data by maximizing internal consistency. *Psychometrika*, 57(4):567–580.

Van Ginkel, J. R., Van der Ark, L. A., and Sijtsma, K. (2007). Multiple imputation for item scores when test data are factorially complex. *British Journal of Mathematical and Statistical Psychology*, 60(2):315–338.

White, I., Royston, P., and Wood, A. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399.