Published as: van Buuren, S., van Rijckevorsel, J. L. A., & Rubin, D. B. (1993). Multiple imputation by splines. In: Bulletin of the International Statistical Institute, Contributed II, pp. 503–504.

MULTIPLE IMPUTATION BY SPLINES

Stef van Buuren[1], Jan L.A. van Rijckevorsel[1] and Donald B. Rubin[2]

[1] TNO Institute of Preventive Health Care, P.O. Box 124, 2300 AC Leiden, Netherlands
[2] Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138

Résumé
La texte propose une combinaison de remplissage multiples des données manquantes avec une transformation à la distribution normale multivariable. Le but de la méthode est la creation des remplissages universelles celles qui tiennes compte des relations non-linéaires. La théorie de la corrélation maximale suggére une transformation économique avec des fonctions B-splines.

A practical difficulty in the application of multiple imputation is the creation of imputations that are reasonable under a large variety of scientific models. Since one typically likes to evade conclusions that are induced by the artificial nature of imputations, it is desirable that imputations are generated independently of any strong scientific hypotheses and models that may confront the data. Seen in this light, the process of finding imputations is a kind of statistical interpolation that preserves the structure in the data as well as the uncertainty about this structure.

This paper describes an extension of the general-purpose multivariate normal imputation (NI) model proposed by Rubin and Schafer (1990) to the non-normal case. Since empirical relations between variables are not always linear, it is hoped that non-normal imputation (NNI) models are often closer to the data than NI-models, and therefore lead to imputations that are more realistic. The NNI-method alternates two steps, one of which is application of the NI-model itself. These steps are: 1. transform the data to approximate multivariate normality given the imputed data, and 2. impute the data given multivariate normality. The method is explained in more detail below.

Let $X = (X_1, X_2, \ldots, X_m)$ be a set of m random variables, where each variable $X_j = (X_j^{\mathrm{obs}}, X_j^{\mathrm{mis}})$ may be partially observed, with $j = 1,\ldots,m$. The problem is to find $P(X)$, the unconditional multivariate density of X, using multiple imputation. Let t denote an iteration counter. Assuming missing at random and i.i.d. multivariate normality, Rubin and Schafer (1990) propose to repeat the following sequence of Gibbs sampler iterations:

$$
\begin{gathered}
\text{For } X_1\text{: draw imputations } X_1^{t+1} \text{ from } P(X_1 \mid X_2^t, X_3^t, \ldots, X_m^t) \\
\text{For } X_2\text{: draw imputations } X_2^{t+1} \text{ from } P(X_2 \mid X_1^{t+1}, X_3^t, \ldots, X_m^t) \\
\vdots \\
\text{For } X_m\text{: draw imputations } X_m^{t+1} \text{ from } P(X_m \mid X_1^{t+1}, X_2^{t+1}, \ldots, X_{m-1}^{t+1})
\end{gathered}
$$

i.e., condition each time on the most recently drawn values of all other variables. At each substep, a linear regression model like $X_1 = X_2^t b_{12} + X_3^t b_{13} + \ldots + X_m^t b_{1m} + \varepsilon_1$ with $\varepsilon_1 \sim N(0, \sigma_1^2)$, is used to draw imputations using a procedure as given in Rubin (1987, p. 167). The b's can be estimated efficiently by the SWEEP operator (cf. Little and Rubin, 1987, p. 112), where they have a noninformative prior. In addition, the algorithm converges in one step for monotone patterns of missing data.

One way to extend this method to non-normal variables is to replace the regression model by an additive model like $X_1 = f_{12}(X_2^t) + f_{13}(X_3^t) + \ldots + f_{1m}(X_m^t) + \varepsilon_1$ in which the *f*'s are smoothing splines having zero expectation. Hastie and Tibshirani (1990, p. 129) show that it is possible to construct Bayesian probability intervals around the posterior mean of the predictors, thus yielding a distribution from which proper imputations can be drawn. Determination of the probability intervals, however, requires inversion of an unstructured $n \times n$ matrix, n being the number of observations. Because the *f*'s must also be found by the iterative backfitting algorithm, it is likely that the computational burden is too high to be of any practical value within the present problem.

As an alternative, let us investigate the consequences of two simplifying assumptions. First, we assume that only one transformation per variable is needed, e.g. for variable 1 we require that $f_{21}(X_1) = f_{31}(X_1) = \ldots = f_{m1}(X_1) \equiv f_1(X_1)$. Also, we assume that f_1^{-1} exists so that $f_1^{-1}(f_1(X_1)) = X_1$. The idea is now to find $f_1, \ldots, f_m$ such that $f_1(X_1)$ becomes as normal as possible, apply the Rubin-Schafer method to generate imputations, transform back the imputations by f_1^{-1}, re-estimate the transformation towards multivariate normality, and so on. Although we have not attempted to provide a formal proof, we expect that such a sequence converges with fixed data to a unique density $P(X)$. Given an initial estimate for COV$[f_1(X_1^t), f_2(X_2^t), \ldots, f_m(X_m^t)]$, the first substep of the Gibbs sampler sequence is then

1a. Draw imputations $f(X_1^{t+1})$ from $P(f_1(X_1) \mid f_2(X_2^t), f_3(X_3^t), \ldots, f_m(X_m^t))$
1b. Transform imputations back to the original scale by $f^{1-1}(f_1(X_1^{t+1}))$
1c. Re-estimate COV$[f_1(X_1^{t+1}), f_2(X_2^t), \ldots, f_m(X_m^t)]$.

One point, the determination of the *f*'s, remains still unclear, but works on attempts to transform to multivariate normality are relevant. For the bivariate problem, Lancaster proved the following theorem: If a bivariate distribution of (X_1, X_2) can be obtained from the bivariate normal by separate transformations on X_1 and X_2, the correlation in the transformed distribution cannot exceed r, the correlation in the normal distribution (cf. Kendall and Stuart, 1979, p. 599). Consequently by maximizing the correlation between $f_1(X_1)$ and $f_2(X_2)$, we are, in some sense, trying to produce a bivariate normal distribution by operating upon the individual variables.

For m variables, the situation is more complicated. See de Leeuw (1988) for a description of the conditions that should be met to attain multivariate normality. These conditions are known as 'linearizability of all bivariate regressions'. Assuming linearity of regressions, maximization of the largest eigenvalue of the correlation matrix of transformed variables induces multivariate normality. If the underlying distribution is already Gaussian, the identity transformation is optimal (Koyak, 1987). Restricting the *f*'s to the class of linear B-splines functions has been pioneered in this context by van Rijckevorsel (1987) and gives an efficient algorithm. Related theoretical work implying the impossibility of finding such transformations in general is Holland (1973).

Combining the Rubin-Schafer imputation method with a transformation step towards multivariate normality should prove a reasonable estimate of $P(X)$. Some enhancements are worthwhile studying: taking advantage of monotone missing data patterns, and temporarily restricting the transformation of the dependent variable to identity so that no inverse transformations are needed for missing values that violate the monotone pattern. For normal data, we expect that the coverage probabilities of the resulting inferences will be approximately the same for the NI-model and the NNI-model. For non-normal data, coverages of the NNI-model should be superior to the NI-model, but whether the differences are relevant in practice remains to be studied.

BIBLIOGRAPHY

de Leeuw, J. (1988). Multivariate analysis with linearizable regressions. *Psychometrika*, 53, 437–454.

Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized additive models*. Chapman and Hall, London.

Holland, P.W. (1973). Covariance stabilizing transformations. *Annals of Statistics*, 1, 84–92.

Kendall, M.G. and Stuart, A. (1979). *The advanced theory of statistics, Vol. 2, 4th ed*. Macmillan, New York.

Koyak, R.A. (1987). On measuring internal dependence in a set of random variables. *Annals of Statistics*, 15, 1215–1228.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical analysis with missing data*. Wiley, New York.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley, New York.

Rubin, D.B. and Schafer, J.L. (1990). *Efficiently creating multiple imputations for incomplete multivariate normal data*. ASA 1990 Proceedings of the Statistical Computing Section (pp. 83–88).

van Rijckevorsel, J.L.A. (1987). *The application of fuzzy coding and horseshoes in multiple correspondence analysis*. DSWO Press, Leiden.