Published as: van Buuren, S. & van Rijckevorsel, J. L. A. (1993). Data augmentation and optimal scaling. In: Steyer, R., Wender, K. F., & Widaman, K. F. (eds.), *Psychometric Methodology: Proceedings of the 7th European meeting of the Psychometric Society, July 29-31, 1993, Trier*, pp. 80–84. Stuttgart; New York: Gustav Fischer Verlag.

DATA AUGMENTATION AND OPTIMAL SCALING

Stef van Buuren and Jan L. A. van Rijckevorsel

TNO Institute of Preventive Health Care, Leiden

Tanner and Wong (1987) propose an ingenious iterative scheme, called data augmentation, for identifying posterior densities of model parameters and multiple imputations of incomplete data matrices. A key ingredient of the method is to impute incomplete parts of the data by extensive sampling. We study how data augmentation can be applied in the context of optimal scaling with missing values. It turns out that a major issue is how to define proper predictive distributions from which the samples are to be drawn. We study the posterior distributions of recovered internal consistency under six consistency levels andtwo levels of missing data.

Keywords: missing data imputation, optimal scaling, EM, Bayesian inference, Gibbs sampling

Introduction

The MISTRESS computer program (van Buuren and van Rijckevorsel, 1991) finds imputations z for incomplete data y such that the internal consistency of the completed data $x = (y, z)$ is maximized. Consistency is measured

by the correlation ratio η^2 (cf. Guttman, 1941). We observed that in a number of situations, especially for low levels of η^2, the method does not produce reasonable imputations. As a matter of fact, if η^2 is near zero then MISTRESS will do more harm than simple random imputation. Reversely, for highly consistent data random imputation does not make sense since missing values can be well predicted from the observed information, and then it is reasonable to maximize consistency over the incomplete data parts. We therefore suspect that random and consistency imputation, i.e. MISTRESS, are most appropriate in respectively the low and high extremes of the consistency spectrum. For intermediate η^2 neither alternative is satisfactory; the random method ignores too much information, while imputation by maximizing consistency assumes more information than is actually present.

The purpose of this paper to gain insight into the circumstances in which it is appropriate to use random and consistency imputation. We are especially interested in the relationship between the true consistency of the data η_0^2 and the posterior distribution $p(\eta^2|y)$ of the consistency given the observed data y only. If all is well, then $E[p(\eta^2|y)] = \eta_0^2$ and $V[p(\eta^2|y)] \to 0$, where E and V are expectation and variance operators respectively. For random imputation and low η_0^2, we expect that $E[p(\eta^2|y)] = \eta_0^2$, but that $V[p(\eta^2|y)]$ will be large. For consistency imputation and (very) high η_0^2, we also expect that $E[p(\eta^2|y)] = \eta_0^2$, but now with $V[p(\eta^2|y)] \to 0$. For intermediate levels, we anticipate that $E[p(\eta^2|y)] \neq \eta_0^2$ for either method.

Optimal scaling

Optimal scaling refers to a set of methods that quantify categorical data by a transformation $q = t(x)$ such that some function $\omega = f(q)$ is maximized.

If we choose $\omega = \eta^2$ and if $t(x)$ belongs to the class of homomorphic mappings of x into $\mathbb{R}$ we obtain homogeneity analysis, the basic technique of the Gifi system (cf. Gifi, 1990). Since $t(x)$ and $f(q)$ are non-stochastic functions the density $p(\eta^2|x)$ is only nonzero at the maximizing value of η^2, while being zero everywhere else.

If some observations are missing then one approach is to adapt $f(q)$ such that only the observed data y are used. We use a different approach that leaves $f(q)$ unchanged but that imputes any missing cells with replacement values z. A straightforward way to model the innate uncertainty of such replacements is to impute the data m times, rather than once (cf. Rubin, 1987). The density $p(\eta^2|y)$ then reflects the variability of η^2 that is caused by the uncertainty of the missing entries.

Setting $\theta = \eta^2$ (or if one prefers $\theta = \{\eta^2, \phi_1, \ldots, \phi_k\}$ with ϕ's representing category quantifications or object scores) enable us to apply the DA algorithm. Some steps are still unclear though. Finding θ_l for given y and z_l, choosing m and checking for convergence are some points that need attention. Furthermore, a major issue is how the predictive distribution $p(z|y, \theta^t)$ is to be specified. We deal with the latter topic first.

Let k be the number of categories for some variable. If an observation is missing on this variable, then all k categories are candidates for imputation. We use a multinomial distribution with probabilities $p_1, \ldots, p_k$ to draw imputations from these candidates. The probabilities may differ for each missing entry, and below we specify in what way $p_1, \ldots, p_k$ depend on y and θ^t. In homogeneity analysis, differences between rows are summarized by a few orthogonal latent variables, often called dimensions, and values on these dimensions are called object scores. Also, category quantifications can be represented in the same

metric. In general, objects lie close to categories on which they score, and vice versa. This suggest that for an object i with missing data the category that is most proximate should receive the highest probability of being imputed, while far categories should have low probabilities. More formally, let o_i be the object score and let c_j be the j-th category quantification for $j = 1, \ldots, k$. We then look for a monotone function $p_{ij} = g(o_i - c_j)$ that transforms k differences into probabilities. There are many ways to do this. In the sequel, we assume that the difference $o_i - c_j$ is distributed normally with mean 0 and variance σ^2. Then $\tilde{p}_{ij} = d(o_i - c_j)$, with $d(\cdot)$ defined as the normal density function, is the usual normal probability. Standardizing p_{ij} as $p_{ij} = \tilde{p}_{ij} / \sum_j^k \tilde{p}_{ij}$ so that $\sum_j^k p_{ij} = 1$ yields the desired parameters for the predictive multinomial distribution from which we draw the imputation.

The width of the predictive distribution depends on σ^2. If we choose a high value of σ^2 (say $\sigma^2 = 100$), then all probabilities will be approximately equal to $1/k$, so in this case we are doing uniform, or random, imputation. At the other extreme, setting $\sigma^2 = 0.01$ accomplishes that the closest categories is always selected over all other, i.e. one probability will approach unity while all other will be zero. Selection of the closest category is the same as performing MISTRESS, and so for low σ^2 we are doing consistency imputation. The method thus covers both random and consistency imputation.

We now turn to some miscellaneous issues. As to the choice of m, the number of parallel solutions, Tanner and Wong (1987) advise to set m low initially, increasing it with successive iterations. Typical values for m range from 20 to as high as 1600. Our experience is that the algorithm still converges for an initial m between 3 and 5.

Gelfand et al. (1990) discuss convergence issues of the algorithm. It appears to be difficult to develop automated convergence assessments. Current practice includes making overlay plots and see if the estimated densities $p(\theta^t|y)$ and $p(\theta^{t+1}|y)$ are visually indistinguishable, and monitoring the 25%, 50% and 75% percentiles for convergence. We do the latter.

Finally, drawing θ_l for given y and z_l simply comes down to calculating the fit, object scores and category quantification in the usual way, given the augmented data (y, z_l). Thus, if no missing data are present, the algorithm produces m identical homogeneity analyses.

Data augmentation

We estimate the posterior distribution $p(\eta^2|y)$ by the data augmentation (DA) algorithm proposed by Tanner and Wong (1987). The method is a simulation technique that is designed to produce simulated posterior distributions of model parameters. These distributions are possibly non-normal, and the data may be incomplete. The DA algorithm has already generated a considerable amount of research. See Gelfand and Smith (1990), Gelfand, Hills, Racine-Poon and Smith (1990), Wei and Tanner (1990) and Zeger and Rezaul Karim (1991). Rubin (in press) introduces the method in psychometrics.

Let the observed data y be augmented by the latent data z. Suppose we are interested in deriving the posterior density $p(\theta|y)$. It is usually not possible to calculate this density directly in the presence of missing data, but finding the density $p(\theta|y, z)$, i.e. the conditional distribution θ given the augmented data

$x = (y, z)$, is often easier. The relation between $p(\theta|y)$ and $p(\theta|y, z)$ is given by the standard identity

$$p(\theta|y) = \int_Z p(\theta|y, z)p(z|y)dz,$$

where $p(z, y)$ denotes the predictive density of the latent data z given y. On the other hand, $p(z|y)$ generally depends on θ by

$$p(z|y) = \int_\Theta p(z|y, \theta)p(\theta|y)d\Theta.$$

Tanner and Wong (1987) showed that by iterative sampling from $p(\theta|y, z)$ and $p(z|y, \theta)$ the average of $p(\theta|y, z)$ over the augmented data patterns will converge to the $p(\theta|y)$. More precisely, the general data augmentation consist of the following steps:

1. Construct some initial estimate θ^0. Choose m as the numbers of samples per iteration. Choose a convergence parameter δ. Set iteration counter $t = 0$.

2. For each sample $l = 1, \ldots, m$ do

    a. Draw imputations z_l from the predictive distribution $p(z|y, \theta^t)$

    b. Draw estimates θ_l from the $p(\theta_l|y, z_l)$, using the augmented data.

3. Compute the mixture $p(\theta^{t+1}|y) = 1/m \sum_l^k p(\theta_l|y, z_l)$.

4. If $|p(\theta^t|y) - p(\theta^{t+1}|y)| < \delta$ stop. Else set $t = t + 1$ and go to 2.

Results

We applied the data augmentation algorithm in a small simulation study in order to investigate the effect of consistency on imputation. We generated 12 data sets of 100 cases and 5 normally distributed variables, under six levels

of consistency. Within one consistency level all correlation equaled a constant, which were respectively 0.00, 0.20, 0.40, 0.60, 0.80 and 1.00. Subsequently, all variables were discretized into 5 categories using a uniform coding function, and we created two levels of random missing data (5% and 20%).

We applied the algorithm to each of the data sets, using three methods: random imputation ($\sigma^2 = 100$), normal imputation ($\sigma^2 = 1$) and consistency imputation ($\sigma^2 = 0.01$). The total number of solutions is therefore 2 (%missing) $\times$ 6 (levels) $\times$ 3 (methods) = 36. The solutions for consistency imputation were actually computed by the MISTRESS computer program since MISTRESS is faster and better avoids local minima. For each solution, we computed the means and variance of the estimated distribution $p(\eta^2|y)$. The means are given in Table 1.

Method	Consistency level					
	00	20	40	60	80	100
$\sigma^2 = 100$	32	38	47	61	74	91
$\sigma^2 = 1$	35	40	51	64	77	99
$\sigma^2 = 0.01$	38	44	54	68	81	100
η_0^2	33	40	51	66	80	100

Table 1: 100 times the average of the recovered fits (5% missing data).

As expected, random imputation ($\sigma^2 = 100$) works best for completely inconsistent data; the average of the estimated η^2 is very close to the true consistency η_0^2 as computed from the complete data. From levels 0.20–0.60, imputation

based on the normal model ($\sigma^2 = 1$) recovers the true consistency best. Beyond a level of 0.60, MISTRESS ($\sigma^2 = 0.01$) should be preferred.

The variances for these solution are given in Table 2.

Method	Consistency level					
	00	20	40	60	80	100
$\sigma^2 = 100$	11	11	25	19	15	27
$\sigma^2 = 1$	7	7	8	7	9	18
$\sigma^2 = 0.01$	00	00	00	00	00	00

Table 2: 100.000 times the variances of the recovered fits (5% missing data).

It will be clear that random imputation is inferior to normal or consistency imputation; it always leads to larger variances, and so we are less sure about the obtained average consistency. Note that consistency imputation has no variance at all since it invariably comes up with the same solution (which may of course be wrong as we saw in Table 1). There is also a slight that the variance increases with consistency. For random imputation, this phenomenon is explained by the fact that for high levels of consistency is not very appropriate, thus leading to very different solutions.

For 20% missing data, differences between the methods are much more pronounced, but the conclusions remain the same. The results are given in Table 3.

Method	Consistency level					
	00	20	40	60	80	100
$\sigma^2 = 100$	29	30	39	47	56	70
$\sigma^2 = 1$	40	44	51	61	71	98
$\sigma^2 = 0.01$	55	55	65	75	84	100
η_0^2	33	40	53	66	78	100

Table 3: 100 times the average of the recovered fits (20% missing data).

Conclusion

Data augmentation can be used in conjunction with optimal scaling if we assume a stochastic component in the scaling model. In this paper, we chose the missing part of the data as an unknown random component.

We studied the distribution of consistency under various imputation methods. It appears that for absolutely inconsistent data, random imputation recovers the true consistency best. For intermediate levels (correlations between 0.20–0.60), we can use the normal imputation model, while beyond 0.60 consistency imputation becomes practical.

Since the three imputation methods can be seen as special cases of a more general method with various σ^2 it would be interesting to find an optimal value of σ^2. This is an area for further research.

REFERENCES

Box, G.E.P. and Tiao, G.C. (1973). Bayesian inference in statistical analysis. Addison-Wesley Pub. Co, Reading, Mass. Republished in 1992 by Wiley, New York.

Gelfand,A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85* , 398–409.

Gelfand, A.E., Hills, S.E., Racine-Poon, A. and Smith, A.F.M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, *85* , 972–985.

Gifi, A. (1990). *Nonlinear multivariate analysis.* Wiley, New York.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In: P. Horst et al. (Eds.), *The prediction of personal adjustment*, 317–348. Social Science Research Council, New York.

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

Rubin, D.B. (in press). EM and beyond. To appear in Psychometrika.

Tanner, M.A. and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*, 528–550.

Van Buuren, S. and Van Rijckevorsel, J.L.A. (1991). *Fast least squares imputation of missing data.* PRM 01-91, Dept. of Psychometrics, University of Leiden.

Wei, C.G. and Tanner, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, *85*, 699–704.

Zeger, A.L. and Rezaul Karim, M. (1991). Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association*, *86*, 79–86.