PSYCHOMETRIKA—VOL. 57, NO. 4, 567–580
DECEMBER 1992

IMPUTATION OF MISSING CATEGORICAL DATA BY MAXIMIZING INTERNAL CONSISTENCY

STEF VAN BUUREN AND JAN L. A. VAN RIJCKEVORSEL

DEPARTMENT OF STATISTICS
TNO INSTITUTE OF PREVENTIVE HEALTH CARE, LEIDEN

This paper suggests a method to supplant missing categorical data by "reasonable" replacements. These replacements will maximize the consistency of the completed data as measured by Guttman's squared correlation ratio. The text outlines a solution of the optimization problem, describes relationships with the relevant psychometric theory, and studies some properties of the method in detail. The main result is that the average correlation should be at least 0.50 before the method becomes practical. At that point, the technique gives reasonable results up to 10–15% missing data.

Key words: missing data, correlation ratio, optimal scaling.

Introduction

This paper discusses a technique to transform incomplete categorical data into complete data by imputing "appropriate" scores into the missing cells. Imputations adhere to the original coding of the data and will maximize the internal consistency of the completed data as measured by the squared correlation ratio η^2, a measure proposed by Guttman (1941). Though there are many ways to find imputations, maximally consistent imputations are particularly attractive because they are, in some sense, the most probable candidates that can be found from the observed data. Replacing missing data by maximizing consistency enhances the similarity among comparable rows in the data. In fact, the basic rationale of the technique is that imputations are sought that will match the nonmissing scores of similarly looking, but complete rows in the data.

The artificial data in Table 1 illustrates the concept of consistency imputation. The table contains 10 observations on three categorical variables: income, age, and type of car. There are three missing values, indicated by a, b, and c. The problem is to find replacement values that are reasonable in some way. For a this is easy; the most consistent estimate is "low," because this makes the profiles 1, 4, and 7 identical. A young owner of a Japanese car will have a low income simply because this is a recurring pattern. Moreover, the profile contains all Japanese cars in the data. Analogously, one finds "high" for b and "old" for c. Both imputations make the remaining two incomplete profiles identical to row 6. So, the missing scores are interpolated from other profiles. One simply looks for similar rows. This is the same as saying that variables must be as homogeneous as possible; that is, they measure the same thing.

For complete data, maximizing η^2 is equivalent to performing homogeneity analysis, also known as multiple correspondence analysis or dual scaling. These techniques quantify the rows and columns of a matrix of categorical data. Row quantifications are

We thank Anneke Bloemhoff of NIPG-TNO for compiling and making the Dutch Life Style Survey data available to use, and Chantal Houée and Thérèse Bardaine, IUT, Vannes, France, exchange students under the COMETT program of the EC, for computational assistance. We also thank Donald Rubin, the Editors and several anonymous reviewers for constructive suggestions.

Requests for reprints should be sent to Stef van Buuren, TNO Institute of Preventive Health Care, PO Box 124, 2300 AC Leiden, THE NETHERLANDS. Email: buuren@nipg.tno.nl.

0033-3123/92/1200-90299$00.75/0

TABLE 1

Example Data

Person	Income	Age	Car
1	*a*	young	jpn
2	middle	middle	am
3	*b*	old	am
4	low	young	jpn
5	middle	young	am
6	high	old	am
7	low	young	jpn
8	high	middle	am
9	high	*c*	am
10	low	young	am

often called *object scores*, while column quantifications are sometimes termed *category quantifications*. See the textbooks by Nishisato (1980), Greenacre (1984), and Gifi (1990) for more details. The essential ingredient of the present technique is the imputation rule that states how blank entries should be filled. Suppose one has some provisional object point for which a score is missing. This paper shows that the most consistent imputation for this object is the category whose category point is closest to the object. Therefore, the imputation rule compares object scores and category quantifications, and elects the category that is nearest to the object. Each missing value is considered in turn, and the steps to compute object scores, quantifications and imputations are iterated until maximal η^2 is found. The method is implemented in a SAS macro called MISTRESS, which is available from the UICSTAT archive at LISTSERV@UICVM.BITNET. (For additional information on how to obtain SAS macros from the archive please refer to Michael Friendly's computing announcement which appeared on pages 313–315 in the June 1992 issue of *Psychometrika*.)

The idea to maximize consistency by imputation was already put forward by Nishisato (1980, p. 201). Nishisato not only deals with maximal consistency, but also discusses minimal consistency, and how these two objectives might be used together. Until now, however, the problem of finding the accompanying imputations has not been solved satisfactorily. The present paper proposes a computationally feasible strategy to select the optimal category. A closely related method is due to Dear (1959) who decomposes the data into its known and unknown parts and uses the first principal component from the known data to estimate the unknown elements. The iterated principal components method proposed by Gleason and Staelin (1975, p. 238) optimizes the sum of p dominant eigenvalues of the correlation matrix of the completed data. Both the Dear method and the PCA method with $p = 1$ can be seen as numerical equivalents of the present technique. For discrete data, Gleason and Staelin (p. 244) suggest rounding imputed values to the nearest category score, but the resulting discretized variables are treated as numerical, which may not always be appropriate in practice. In an analysis of variance context, Hartley and Hocking (1971) identify the so-called (X, m, d) model in which one tries to find estimates for missing classifications on the experimental variables. This is a combined estimation and classification problem. They note some

difficulties with the model, but do not pursue the matter any further. Greenacre (1984, p. 237) uses a somewhat different approach. He does not replace the missing data themselves, but imputes "consistency optimizing" rounded estimates of marginal frequencies. Good reviews of imputation techniques for categorical data in general are Kalton and Kasprzyk (1982) and the three volumes edited by Madow, Olkin, and Rubin (1983). For multiple imputation, in which not just one but many replacements are searched, see Rubin (1987). Little and Rubin (1990) provide a recent overview of missing data strategies in the social sciences.

A number of well-established procedures exist for handling missing data in homogeneity analysis. These approaches either code missing responses as auxiliary categories or ignore them during the computations. Beside textbooks, Meulman (1982) is a readily accessible systematic study of the properties of three strategies for handling missing data. Van der Heijden and Escofier (1989) describe a number of other missing data options in multiple correspondence analysis.

The main difference between the current method and previous strategies is that imputations are constrained to stay within the original category coding, which is conceptually attractive. The corresponding optimization problem has been around for some time. It is solved by partitioning the loss into known and missing components and applying k-means to the missing part. Furthermore, the results on local minima, recovery and the amount of nonresponse provide critical conditions for the practical application of the technique.

Method

Let the data be coded into indicator vectors $g_j (j = 1, \ldots, m)$ of length k_j such that $q_{jk} = 1$ if the observation falls into category k of variable j and $g_{jk} = 0$ otherwise, and let y_j be a k_j vector of category quantifications, where k_j denotes the number of categories of the j-th variable. The expression $x_j = g_j' y_j$ then yields a quantified score x_j. Let the average of the m quantified scores $z = 1/m \sum x_j$ define the object scores. The total variation of the data can then be decomposed as

$$\sum_{j=1}^{m} x_j^2 = mz^2 + \sum_{j=1}^{m} (z - x_j)^2. \tag{1}$$

This is a between-within partitioning of the form $T = B + W$. The *correlation ratio*, introduced by Guttman (1941), denoted by η and defined by $\eta^2 = B/T$ measures how well z can be considered as a representative of each x_j. The ratio ranges from 0 to 1, and it is equal to 1 if all variables are proportional.

For complete data, η^2 can be optimized over $y_1, \ldots, y_m$. To identify the solution, $x_1, \ldots, x_m$ must have a zero mean and the variance of z must equal some positive constant. Procedures for finding optimal $y_1, \ldots, y_m$ are known as homogeneity analysis, multiple correspondence analysis, dual scaling, among others (see Gifi, 1990). These techniques usually consider one or more orthogonal sets of z's, with corresponding η's.

Let Ω denote the set of all nonmissing variables and let the symbol x_j^* stand for an imputed value. Obviously,

$$x_j = \begin{cases} x_j, & \text{if } j \in \Omega; \\ x_j^*, & \text{if } j \notin \Omega. \end{cases}$$

It is possible to partition the variation into three independent quadratic components:

$$\sum_{j=1}^{m} x_j^2 = mz^2 + \sum_{j \in \Omega} (z - x_j)^2 + \sum_{j \notin \Omega} (z - x_j^*)^2, \tag{2}$$

so that the squared consistency is again equal to $\eta^2 = B/T = mz^2/(\sum_{j=1}^{m} x_j^2)$. Since $T = B + W$, the maximum of η^2 coincides with the minimum of $W/T = 1 - \eta^2$. Maximal homogeneity among the imputed variables can be found by minimizing this W/T-ratio over $y_1, \ldots, y_m$ and over the imputations $x_1^*, \ldots, x_m^*$. The corresponding loss function is defined in terms of x_j and x_j^* and can be written as

$$\sigma(y_1, \ldots, y_m; g_1^*, \ldots, g_m^*) = \sum_{j \in \Omega} (z - g_j' y_j)^2 + \sum_{j \notin \Omega} (z - g_j^{*\prime} y_j)^2, \tag{3}$$

where g_j^* indicates an initially unknown indicator vector. Let $\sigma(\cdot)$ stand for $\sigma(y_1, \ldots, y_m; g_1^*, \ldots, g_m^*)$. Now, the imputation problem is where to impute the "1" in the missing vector g_j^*, which is a problem in combinatorial optimization. More specifically, this type of problem is known as *sum-of-squares partitioning*. See, for example, Späth (1985) for a detailed discussion. The so-called "k-means algorithm" is used here to solve it. For given j, the k-means algorithm minimizes $\sigma(\cdot)$ over y_j and g_j^*, simultaneously. Reasons to choose the k-means algorithm over other procedures like enumeration, dynamic programming techniques and simulated annealing are that the k-means strategy is simple and well-studied, can be applied to large data sets and has satisfactory performance characteristics (Milligan, 1980; Scheibler & Schneider, 1985). Moreover, the combination of homogeneity analysis and the k-means algorithm was used before for a different purpose in the GROUPALS clustering technique developed by van Buuren and Heiser (1989). The algorithm is, however, prone to local minima.

The k-means algorithm (see Hartigan, 1975) iteratively relocates classifications one by one. A slightly modified version of the standard algorithm is used here. The modification is that all nonmissing entries remain tied to their categories and are never relocated. The modified k-means algorithm starts with some initial imputation of the missing data, considers each imputation in turn and checks whether or not a change from the current category s to a new category t would decrease the loss. If so, the imputation will be relocated from s to t and the solution is updated accordingly. The process is repeated until no relocations exist that would reduce the loss.

The precise imputation rule follows directly from the loss function. Let d_s and d_t denote the number of observations in category s and t of variable j, and let y_s and y_t be the corresponding category quantifications. Suppose that observation i has a score z_i and that we move imputation g_i^* from s to t. Fisher (1958) showed that the new loss will be equal to

$$\sigma^*(\cdot) - \frac{d_s(z_i - y_s)^2}{d_s - 1} + \frac{d_t(z_i - y_t)^2}{d_t + 1},$$

where $\sigma^*(\cdot)$ denotes the current loss. Therefore, the imputation rule is: If the inequality

$$\frac{d_t(z_i - y_t)^2}{d_t + 1} < \frac{d_s(z_i - y_s)^2}{d_s - 1}$$

is true, relocate imputation s to t. Every relocation that adheres to this rule decreases the loss. The inequality usually holds if z_i is closer to y_t than to y_s, depending on the

size of d_s and d_t. After relocation, the corresponding y_s and y_t are updated. In general, it suffices to know the former weights y_s and y_t, the score z_i and the marginal frequencies d_s and d_t. The update for the donating category s becomes

$$\hat{y}_s := \frac{y_s d_s - z_i}{d_s - 1} = y_s + \frac{z_i - y_s}{d_s - 1}.$$

Likewise, for the receiving category t we obtain

$$\hat{y}_t := \frac{y_t d_t + z_i}{d_t + 1} = y_t + \frac{z_i - y_t}{d_t + 1}.$$

It is assumed that $d_s \geq 1$ so that division by zero does not occur. Both formulas are independent of the number of observations, which makes them very efficient updates, especially for large sample sizes.

Results

Occurrence of Local Minima

A recovery study on artificial data was carried out to investigate the existence of local minima, and used 10 datasets, each consisting of 100 subjects, 7 normally distributed variables with bivariate correlations that varied systematically from $r = 0.00$ to $r = 0.90$ with a step size of 0.10. The data were discretized into 5 categories using an optimal coding scheme preserving the normal distribution as closely as possible (van Rijckevorsel & de Leeuw, 1992) and 5% random missing data were created. Two methods to find a starting allocation of missing entries are used: random and passive. The random procedure imputes a category randomly drawn with a probability proportional to the observed marginal frequencies; the passive method uses the "missing passive" option (Gifi, 1990) to compute an initial configuration and subsequently allocates each observation to a category that is closest. After these initial imputations are found, the method iterates over $z, y_1, \ldots, y_m$ and $g_1^*, \ldots, g_m^*$ until the difference between two consecutive values for $\sigma(\cdot)$ is less than 1.0E−7. We computed 25 replications per condition, for a total of $10 \times 2 \times 25 = 500$ analyses.

Convergence usually occurred after about 10-20 iterations. Lower levels of consistency produce local minima. Figure 1 graphs $\log(\sigma_{\max} - \sigma_{\min})$ versus the number of different minima of the 25 replications per level, for both random and passive starts. The most striking feature of the plot is the difference between the starting methods. Passive starts always produces less minima and induces more similar fits. For correlation levels 0.00–0.40, both methods find less minima as the data become more consistent. Beyond $r = 0.40$, the passive starting method invariably generates the same solution, and the corresponding points are not plotted. The random method, however, keeps producing about 15 different minima, no matter what the internal consistency of the data is. Passive is clearly superior to random, both in terms of the number of minima and in terms of fit.

Also of interest is whether the obtained solution is close to the global optimum. Since the actual globally optimal solution is not known, this question cannot be answered directly. Taking the best fitting solution over all 50 replications within a correlation level as a provisional global optimum, one finds the solutions obtained by the passive method beyond $r = 0.40$ are very close or equal to the provisional global minimum in terms of fit. Keep in mind that these results apply to a situation of 5% missing data. If the actual percentage is higher, the "safe" correlation level is also

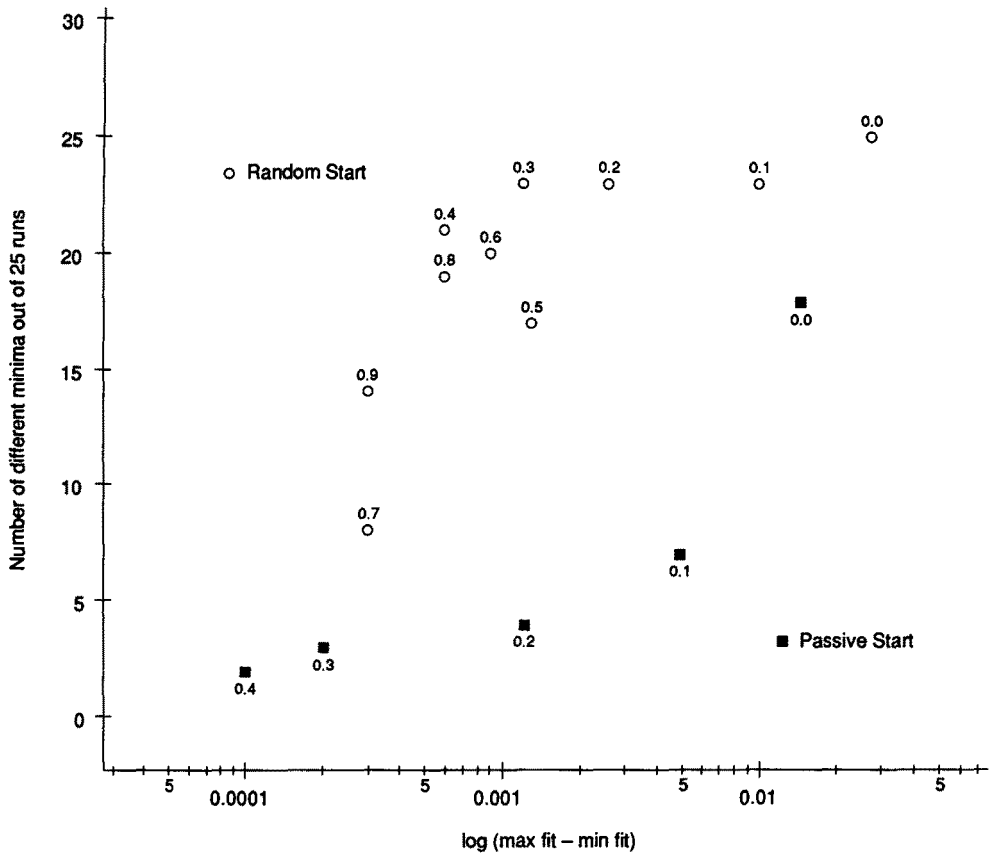

FIGURE 1.
Fit differences versus number of minima for two starting methods and 10 correlation levels.

likely to rise. Other factors that might influence the number of local minima are the sample size and the number of categories.

Recovery of Missing Information

Another issue is how well the method recovers the "true" data. In practice one never has access to the true data, and only by creating artificial blanks within the complete data may we know them. A way to measure the amount of recovery in this case is the Q_α statistic proposed by Gleason and Staelin (1975). If h_{ij} represents the raw data, h^*_{ij} represents the imputed data, p denotes the proportion of missing values and σ_j^2 denotes the variance of the complete data with sample size n, then

$$Q_\alpha = \left(\sum_{j \notin \Omega} \frac{(h_{ij} - h^*_{ij})^2}{\sigma_j^2 nmp} \right)^{1/2} \tag{4}$$

measures the dissimilarity between the true and the imputed values for some method α. For mean substitution, Gleason and Staelin show that the expected value of Q_α is equal to 1.

The same 10 complete matrices used in the local minima simulation study were considered and from these 25 incomplete matrices were generated with 5% and 10% missing values for each consistency level. Next, Q_α was computed for these 250 ma-

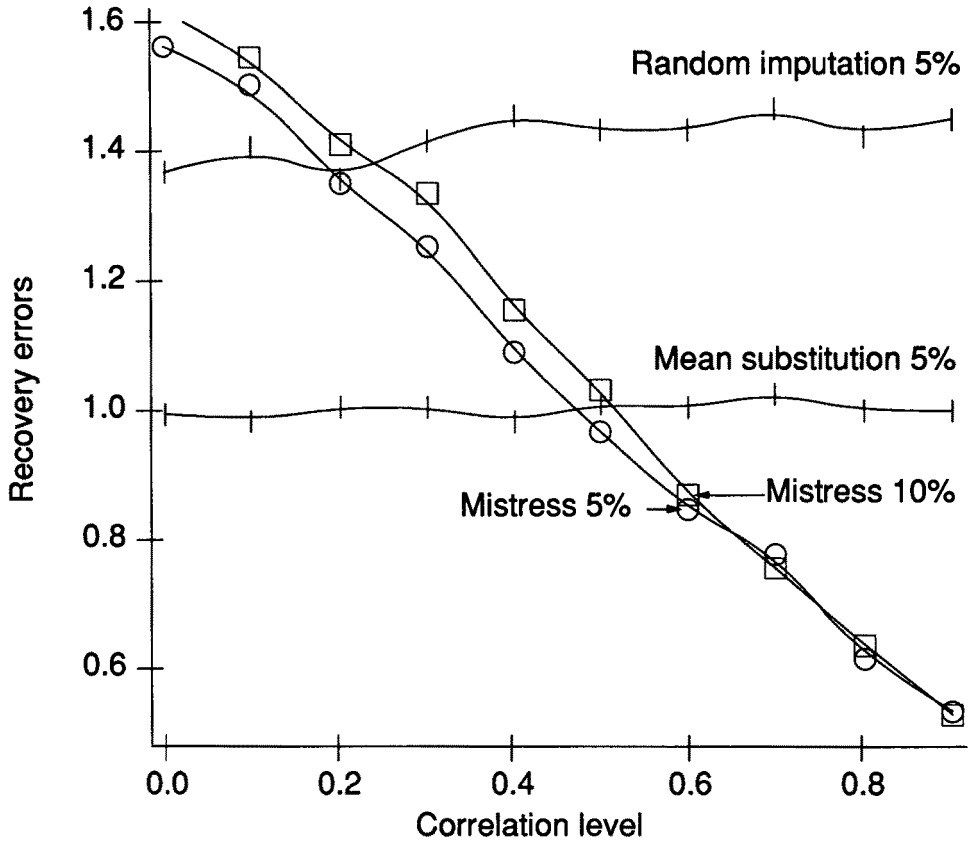

FIGURE 2.
Correlation level versus recovery errors.

trices using three imputation methods: random imputation, mean substitution, and consistency imputation. Mean substitution for categorical data reduces to imputing the modal category.

The four curves in Figure 2 contain the averages over 25 replications of Q_α. As expected, the curve for mean substitution hovers around "1". It will be clear that random imputation is inferior to mean substitution. For low levels of correlation, consistency imputation is worse than random and introduces too much structure in the data. Beyond correlations of 0.20, the method is better than random and beyond 0.50 it recovers the data better than mean substitution. This result agrees with the curve for the Dear method found by Gleason and Staelin (1975), and with the findings of van Buuren and van Rijckevorsel (1992), who compared three other imputations methods and found that maximizing consistency is best above correlations of 0.50. Gleason and Staelin found still better recovery for their PCA method, but they consider not one, but many independent components. The same could have been done here, but it would not maximize consistency.

Amount of Missing Data

A further practical point concerns the amount of missing data. Since the method optimizes η^2 over the missing data, increasing the number of unknown entries results in higher values of η^2. The more missing data, the more consistency obtained. The magnitude of bias towards the consistency model depends on the amount of missing

data. Optimization techniques tend to find those imputations that are most favorable for the model at hand. Bias towards the model is not a problem of MISTRESS only, but a characteristic of any optimization method that attempts to estimate missing data.

The question of how much bias is acceptable remains. In the context of dual scaling, Nishisato (1980, p. 203) advocates the use of minimal and maximal correlation ratios η^2_{min} and η^2_{max}. One obtains η^2_{min} by inserting those responses that will *minimize* the internal consistency. The difference $\eta^2_\delta = \eta^2_{\text{max}} - \eta^2_{\text{min}}$ increases with the amount of missing data, and the idea is that there might be some critical point beyond which η^2_δ becomes significant and the analysis should be aborted. Nishisato and Ahn (in press) suggest the computation of empirical 95% confidence bands for η^2_{max} and η^2_{min} to see if they overlap. If they do not, η^2_δ is considered to be significantly different from zero.

An alternative strategy is used here. Rather than focusing on the difference between η^2_{max} and η^2_{min}, concentrate on the difference between η^2_{max} and the true consistency, η^2_{true}, which is simply equal to the consistency of the complete data. By creating four nonresponse rates (5, 10, 15, and 20 percent), and applying MISTRESS to the incomplete data, four estimates of η^2_{max} are obtained. Imputation bias can then be quantified as $\eta^2_{\text{max}} - \eta^2_{\text{true}}$. To get an idea what amount of bias could be reasonably tolerated, it is useful to see whether η^2_{max} falls within a confidence interval of η^2_{true}. A simple but convenient bootstrap procedure is used (as outlined by Meulman, 1982, p. 67) to compute 80% confidence intervals. First, take 10 random samples *with replacement* of size n (n being the number of observations) from the rows of the complete data. Second, compute the consistency for each bootstrap sample. Third, neglect the most extreme values and take the rest as a crude estimate of the 80% confidence interval. Note that the very act of drawing with replacement may make the data more homogeneous (rows will be repeated now and then), so that on the whole, bootstrap intervals will be slightly biased upwards. This effect is particularly visible in the lower consistency levels for which the bootstrap interval may not even contain η^2_{true}. Compared to 90% or 95% confidence bands, the use of 80% intervals leads to smaller intervals and more conservative results. For larger confidence bands, it is less likely that the hypothesis $\eta^2_{\text{true}} = \eta^2_{\text{max}}$ will be rejected.

The same 10 complete data sets were analyzed as before and the average of η^2_{max} was computed over three replications for every combination of nonresponse and consistency level. Figure 3 portrays the results as a scatterplot of true versus obtained consistency. The gray surface indicates the 80% confidence regions for η^2_{true}. Lines connect all points that have the same nonresponse level. As expected, η^2_{max} increases if more imputations are to be found. The curves for 5% and 10% missing data reside within the 80% confidence band. This means that the variation caused by imputation compares to the bootstrap stability. The 15% curve stays above the confidence region until $\eta^2_{\text{true}} = 0.60$. This result suggests that for imputing 15% missing data one needs at least a consistency of 0.60. Below the crossover point MISTRESS invents too much structure, and should therefore not be applied. The same story can be told for 20% missing data, except that the crossover point is now at to 0.75.

Dutch Life Style Survey

This example is taken from the Dutch Life Style Survey (Leef Situatie Onderzoek) conducted by the Netherlands Bureau of Census. The data were collected at different time points during the years 1977 through 1986. Not all questions were posed at each occasion, and so the data contains many systematic missing entries. The analysis sample consists of 7332 individuals. Persons responded to the question whether one of the following conditions applied: dirty (D), heavy (H), risky (R), stench (S), and noise

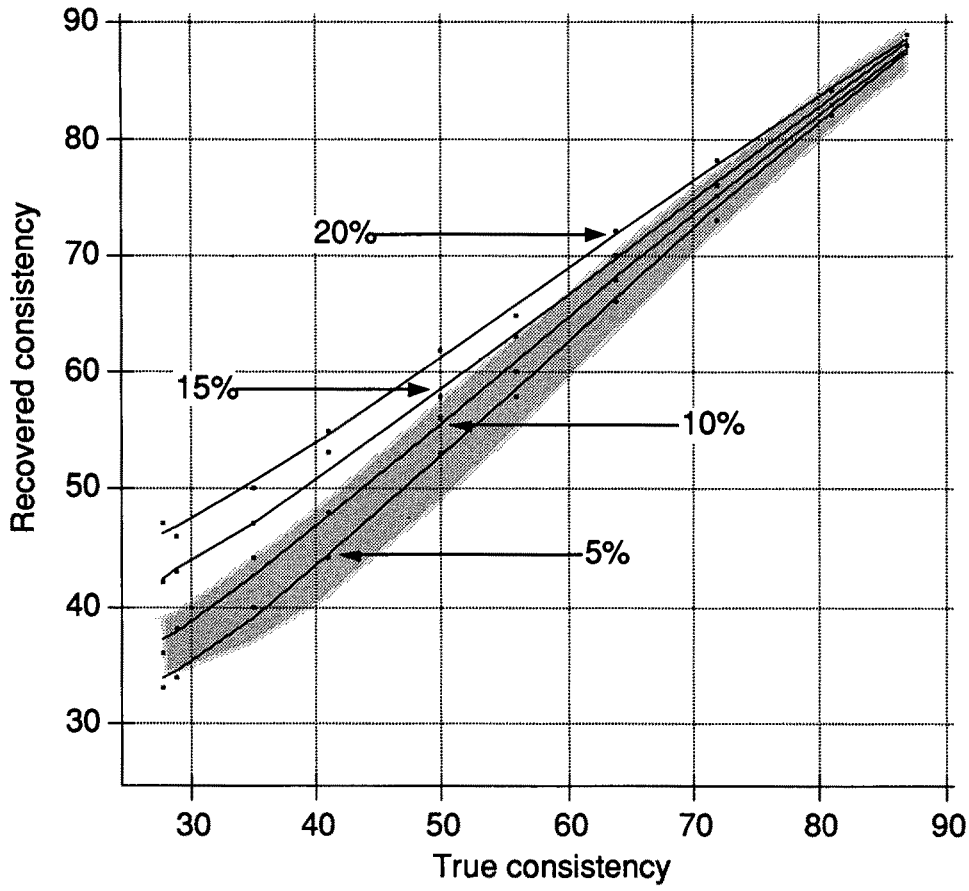

FIGURE 3.
True consistency (H) versus obtained consistency (V) for 4 levels of missing data.

(N). For 5750 people, the type of job is also known, classified into 7 categories: management (MAN), administrative (ADM), commercial (COM), scientific (SCI), service (SER), agrarian (AGR), and industrial (IND). The classification by profession is missing for 7332 − 5750 = 1582 observations.

The data and the imputation results, ordered by row scores, are presented in Table 2. The most consistent imputations are in boldface. All workers experiencing at least three or more adverse conditions are assigned to the group of industrial workers. Three out of 10 incomplete profiles with 2 annoyance scores are assigned to farmers. The 816 persons working in a clean environment are all assigned to the management group. Labor conditions are consistent with the type of work people do, and this relationship is automatically taken into account when searching for maximally homogeneous imputation.

The frequencies of the observed data are also pictured in Figure 4 by a slightly smoothed graphical analogue of the cross-tabulation in Table 2. The job classes and nuisance patterns are scaled by the consistency maximizing scores obtained by MISTRESS, with blue collar jobs relatively close together on one side and well-separated from white collar jobs on the other side of the X-axis. The conditions with few or no nuisance parameters are consistently separated from the rest on the Y-axis. The intuitive interpretation of the most consistent imputation is that it should disfigure the

TABLE 2

Single Imputation LSO Table (**Bold** is an Imputation)

Labor Conditions	Professional Category							z
D H R S N	MAN	ADM	COM	SCI	SER	AGR	IND	
1 1 1 1 1	1	1	1	6	2	5	64 **19**	3.18
1 1 1 1 0	0	0	0	6	3	7	11 **10**	2.68
1 0 1 1 1	1	1	0	3	3	1	21 **6**	2.58
0 1 1 1 1	0	0	0	1	1	1	**3**	2.48
1 1 1 0 1	0	1	1	1	1	3	61 **23**	2.47
1 1 0 1 1	0	1	1	4	6	4	50 **15**	2.41
1 0 1 1 0	0	1	2	1	0	2	8 **5**	2.08
0 1 1 1 0	0	1	0	0	0	1	2 **1**	1.97
1 1 1 0 0	1	1	0	9	2	9	51 **22**	1.96
1 1 0 1 0	0	1	1	9	2	20	13 **20**	1.90
0 0 1 1 1	0	3	2	2	0	0	12 **4**	1.88
1 0 1 0 1	0	0	2	1	1	1	20 **12**	1.87
1 0 0 1 1	4	3	2	6	3	2	46 **32**	1.81
0 1 1 0 1	0	0	1	2	1	0	8 **7**	1.76
1 1 0 0 1	2	6	4	6	9	12	88 **32**	1.70
0 1 0 1 1	0	0	0	1	0	0	5 **2**	1.70
0 0 1 1 0	0	0	0	1	1	0	2 **3**	1.38
1 0 1 0 0	0	1	0	1	4	3	14 **10**	1.37
1 0 0 1 0	0	2	1	2	2	13	17 **10**	1.31
0 1 1 0 0	0	0	3	6	0	1	10 **9**	1.26
0 1 0 1 0	0	1	0	3	4	0	4 **3**	1.20
1 1 0 0 0	2	6	16	21	38	81	95 **81**	1.19
0 0 1 0 1	1	16	3	10	6	2	15 **14**	1.17
0 0 0 1 1	3	19	6	16	6	0 **29**	28	1.00
1 0 0 0 1	8	11	6	20	14	10 **48**	103	0.99
0 1 0 0 1	2	4	12	19	21	4 **16**	40	0.89
0 0 1 0 0	3	5	7	27	12 **19**	2	29	0.16
1 0 0 0 0	4	15	28	32	25 **96**	60	104	0.09
0 0 0 1 0	6	22	3	27	8 **25**	1	17	0.09
0 1 0 0 0	2	12	58	115	87 **71**	16	80	-0.02
0 0 0 0 1	21	133	40	132	54 **122**	3	125	-0.11
0 0 0 0 0	157 **816**	843	373	916	349	54	324	-0.91
total	218 **816**	1100	573	1406	665 **333**	318 **93**	1470 **340**	

surface in Figure 4 as little as possible. The frequencies of the imputed data are depicted in Figure 5.

The imputations follow a curved and peaked range of frequencies from the origin {white collar, no nuisance} up to the far upper corner {blue collar, maximal nuisance}. Though the analysis extracts the major trend from the data, the albeit "reasonable" imputations are nevertheless very "single": all missing data with the non-nuisance

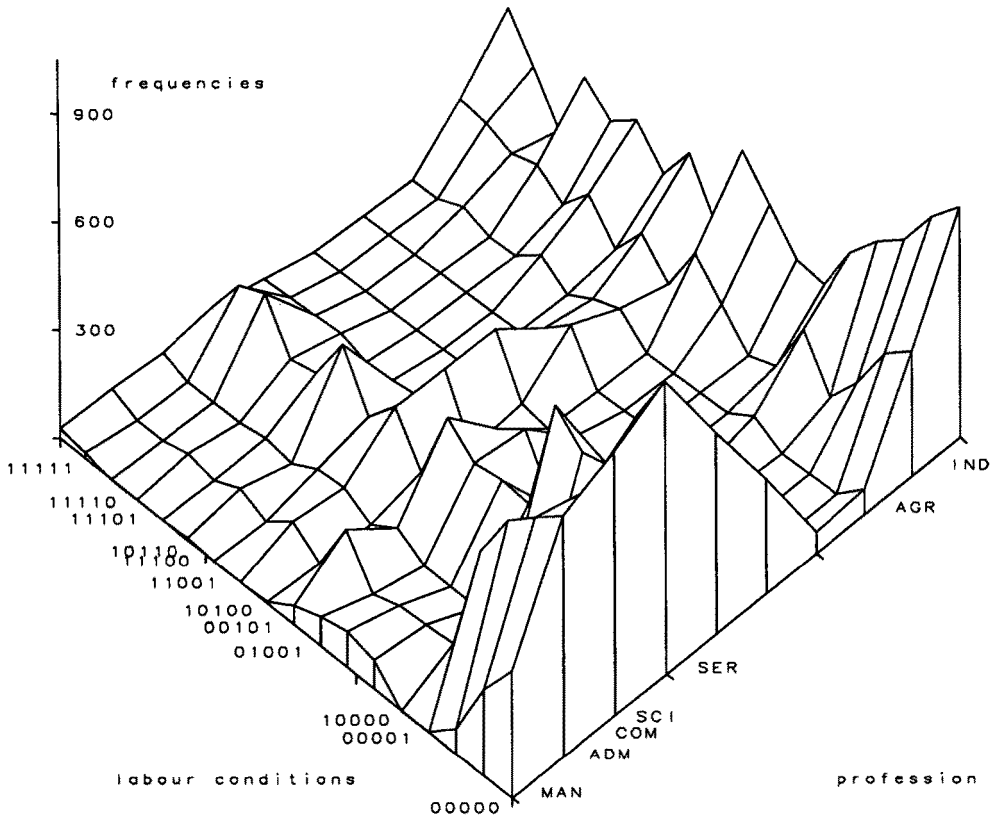

FIGURE 4.
Observed frequencies (Z) versus job classes (X) versus labor conditions (Y).

pattern are attributed to managers. In practice, this solution is a bit peculiar since other white collar workers would also have been reasonable candidates. The effect is a consequence of maximizing consistency, and it is most likely to become a problem if the number of different response profiles is small. A possible strategy will now be sketched, based on multiple imputation, how one may handle such situations.

Suppose that it is possible to define for each missing value a multinomial distribution $\Pr(X_{\text{mis}} \mid X_{\text{obs}})$ containing probabilities p_k that, conditional on the observed data, state how likely it is that category k should be imputed for X_{mis}. One possibility to find such a distribution would be to relate the distances between objects and categories to probabilities. Let $\Pr(X_{\text{mis}} \mid X_{\text{obs}})$ be chosen such that the category with highest p_k is the most consistent imputation. This most consistent imputation is appropriate only if p_k itself is relatively high, but of course, we may also find that all candidates are equally probable. The predictive distribution will then be relatively flat and the highest p_k is not particularly representative for the distribution. In that case, a method that simply picks the largest probability tends to generate a lot of unreasonable imputations. A sensible alternative would then be to draw a number of samples from the entire predictive distribution $\Pr(X_{\text{mis}} \mid X_{\text{obs}})$, and use these samples to generate a number of completed matrices, which are pooled in some way afterwards. Such procedures take the uncertainty induced by the imputations into account, and are known as multiple imputation (Rubin, 1987) and data augmentation (Tanner & Wong, 1987). See Rubin (1991) for an overview.

There are many possibilities to combine multiple imputation and maximizing con-

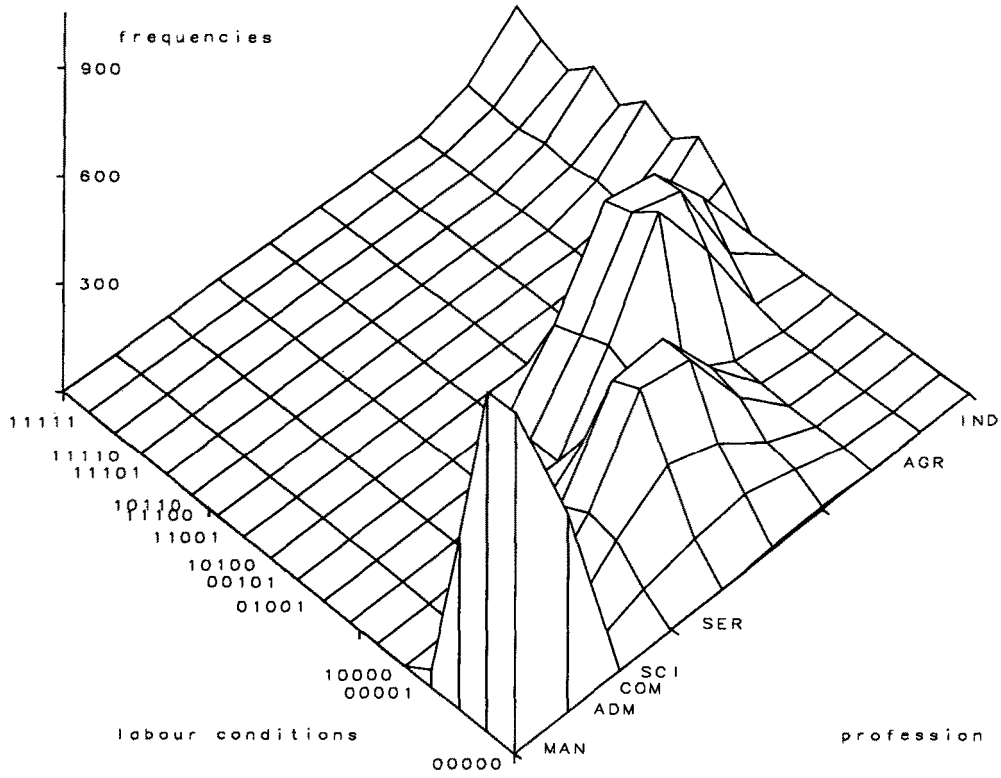

FIGURE 5.
Imputation frequencies (Z) versus job classes (X) versus labor conditions (Y).

sistency. For each missing entry, one can simply define Pr ($X_{\text{mis}} \mid X_{\text{obs}}$) as a set of k probabilities proportional to the inverse of the distances between the provisional object point and the k category quantifications. Using this definition, the data were re-imputed five times. Figure 6 is a graphical representation of the average imputation counts. Comparing Figures 5 and 6, both imputations, single and multiple, follow the same gradient from white collar, no nuisance on the bottom to blue collar, maximal nuisance on top. It is obvious that the multiple imputations are more spread over jobs and nuisance patterns and that the surface is more like that of Figure 4.

Discussion

The method optimizes a well-known and widespread criterion: η^2 is proportional to the largest eigenvalue of the correlation matrix, the average correlation, the average of the squared correlations between z and x, and Cronbach's α. The method stays close to the data, and also few assumptions are needed. Furthermore, given a sufficient amount of consistency, local minima appear to be not much of a problem. Since the number of variables or observations hardly influences the computations, the technique can be used with large data matrices.

Despite these assets, the method should be applied with care. A main practical issue is the amount of intercorrelation. If the magnitude of the correlations is below 0.20, then the method may generate estimates that are even worse than random imputation. In this case, unconditional mean imputation often works better. It seems preferable to use MISTRESS here only in combination with a resampling method, like the bootstrap, to estimate the variability of consistency. In the recovery study, maximizing

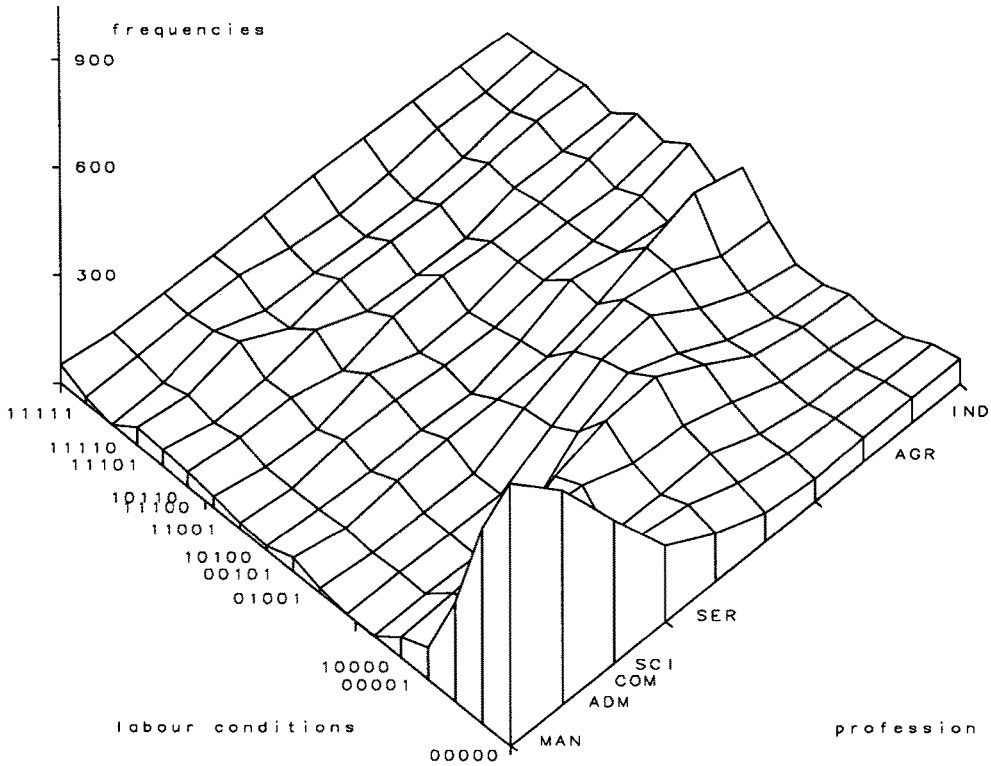

FIGURE 6.
Average counts based on multiple imputation (Z) versus job classes (X) versus labor conditions (Y).

consistency beats mean substitution only above average correlations of 0.50. Correlations of this magnitude can be found in, for example, psychological testing, surveys and longitudinal research. The fact that MISTRESS does not work well with inconsistent data is not the fault of the method itself, but is due to the lack of fit between the data and the consistency model. Inappropriate models simply yield bad imputations.

References

Dear, R. E. (1959). *A principal component missing data method for multiple regression models* (SP-86). Santa Monica, CA: System Development Corporation.

Fisher, W. D. (1958). On grouping for maximum homogeneity. *Journal of the American Statistical Association, 53*, 789–798.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester: Wiley.

Gleason, T. C., & Staelin, R. (1975). A proposal for handling missing data. *Psychometrika, 40*, 229–252.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. New York: Academic Press.

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* (pp. 319–348). New York: Social Science Research Council.

Hartigan, J. A. (1975). *Clustering algorithms*. New York: Wiley.

Hartley, H. O., & Hocking, R. R. (1971). The analysis of incomplete data. *Biometrics, 27*, 783–808.

Kalton, G., & Kasprzyk, D. (1982). Imputing for missing survey responses. *Proceedings of the Section of Survey Research Methods, 1982* (pp. 22–23). Alexander, VA: American Statistical Association.

Little, R. J. A., & Rubin, D. B. (1990). The analysis of social science data with missing values. In J. Fox & T. Scott Long (Eds.), *Modern methods of data analysis* (pp. 374–409). London: Sage.

Madow, W. G., Olkin, I., & Rubin, D. B. (Eds.). (1983). *Incomplete data in sample surveys* (Vols. 1–3). New York: Academic Press.

Meulman, J. (1982). *Homogeneity analysis of incomplete data*. Leiden: DSWO Press.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation of fifteen clustering algorithms. *Psychometrika, 45*, 325–342.

Nishisato, S. (1980). *Analysis of categorical data: Dual scaling and its applications*. Toronto: University of Toronto Press.

Nishisato, S., & Ahn, H. (in press). When not to analyze data: Decision making on missing responses in dual scaling. *Annals of Operations Research*.

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

Rubin, D. B. (1991). EM and beyond. *Psychometrika, 56*, 241–254.

Scheibler, D., & Schneider, W. (1985). Monte Carlo tests of the accuracy of cluster analysis algorithms. *Multivariate Behavioral Research, 20*, 283–304.

Späth, H. (1985). *Cluster dissection and analysis*. Chichester: Ellis Horwood.

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

van Buuren, S., & Heiser, W. J. (1989). Clustering n objects into k groups under optimal scaling of variables. *Psychometrika, 54*, 699–706.

van Buuren, S., & van Rijckevorsel, J. L. A. (1992). Data augmentation and optimal scaling. In R. Steyer, K. F. Wender, & K. F. Widaman (Eds.), *Psychometric Methodology. Proceedings of the 7th European Meeting of the Psychometric Society in Trier* (80–84). Stuttgart and New York: Gustav Fischer Verlag.

van der Heijden, P. G. M., & Escofier, B. (1989). *Multiple correspondence analysis with missing data*. Unpublished manuscript, University of Leiden, Department of Psychometrics and Research Methods.

van Rijckevorsel, J. L. A., & de Leeuw, J. (1992). Some results about the importance of knot selection in nonlinear multivariate analysis. *Statistica Applicata: Italian Journal of Applied Statistics, 4*.

Manuscript received 12/26/90
Final version received 1/6/92