# FAST LEAST SQUARES IMPUTATION OF MISSING DATA

STEF VAN BUUREN

JAN L.A. VAN RIJCKEVORSEL



Stef van Buuren
Dept. of Psychometrics
University of Leiden
Wassenaarseweg 52
2333 AK Leiden

Jan. L. A. van Rijckevorsel
Dept. of Statistics
NIPG-TNO
Wassenaarseweg 56
2333 AL Leiden

Januari 1991

# FAST LEAST SQUARES IMPUTATION OF MISSING DATA

STEF VAN BUUREN
UNIVERSITY OF LEIDEN

JAN L.A. VAN RIJCKEVORSEL
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. Consistency is measured by a between-total variance ratio. The idea is that similar profiles obtain comparable imputations. The text outlines a solution for the optimization problem, describes relationships to the relevant psychometric theory and studies some properties of the method. Some examples are presented. The main application fields are in the analysis of survey data, rating scales and questionnaires.

Keywords: missing data, imputation, categorical variables, consistency, least squares, MISTRESS, multiple imputation

## Introduction

Missing data are common and costly. Data with up to 30% missing are no exception, and attempts to do something about missing data are as old as missing data themselves. Popular ways to accommodate for missing data—pairwise and listwise deletion—may amount to wasting labour-intensively collected material. An alternative is to fill in missing entries with 'appropriate' replacements. The advantage of this is that standard multivariate techniques can be applied to the completed data.

Sometimes, external information is available that may help. For example, suppose that a subject is unwilling to tell his age, thereby producing a missing value. If we have the actual person in front of us we may nevertheless infer his age by using other clues, and subsequently fill in our prediction as if it had been observed. Another source of external information could be a previous score on same question or test. Unfortunately, such situations are more an exception than a rule.

In general, if we want to complete the data we should look for other sources of information. An obvious alternative is to consider the data that *are* available for the subject, in combination with the data that are collected on other sample units. Under the assumption that observations with a similar response pattern are likely to score identically on any remaining, unobserved variables, we may try to interpolate missing values. This type of imputation strategy is known as *hot deck* imputation. The basic 'reasonable imputation assumption' is: objects with almost similar profiles have the same distribution on any missing responses. The idea is to borrow the observed score from a

closely related profile. Some arguments that are often mentioned in favour of hot deck methods are: the reduction of the response bias, the preservation of the distribution of the population, and—most important of all—the production of a complete data set. We may add that, particularly in large surveys, computational ease and speed are highly evaluated.

In this paper we consider a hot deck imputation technique based on the within-homogeneity of all variables simultaneously, so missing items do not have to be missing at random (=MAR). The imputation method is composed of two ingredients: the definition of a *donor variable* and the derivation of an *imputation rule*. The donor variable measures how much individual data profiles differ from each other. So a donor is not a sample unit or a data profile, but a latent variable. The imputation rule states how blank entries should be filled, given the values on the donor variable. As we will see, these two components are closely intertwined in the method. Changing the donor variable may cause a modification of the imputations. The converse is also true; changing an imputation has an effect on the donor variable.

Simultaneous homogeneity over all variables is expressed by the donor. The donor variable is equal to a weighted average of all variables. We use the familiar between-total sum of squares ratio to indicate how well the donor represents the total variation in the data. It follows that the 'best' donor variable is equal to the first principal component of the completed data.

The position of each observation on the donor reflects how much sample units have in common. The function of the donor is thus much like the partitioning of observations into homogeneous classes employed by traditional hot deck procedures. The difference with existing hot deck procedures is that all (categorical) variables act simultaneously as donor. Imputation is based on comparing donor scores. If an incomplete profile resides closely to a completely observed profile (in the sense that their donor scores differ little), then its missing entries can replaced by the known values of the observed unit. We measure 'closeness' by the squared Euclidean distance between donor scores. Mathematically, we look for imputations that minimize a sum-of-squared-distances function.

In practice, just one donor may not be representative. Primary reasons for using just one is that it is simple, and that it has some attractive analytical properties. If necessary, the extension to multiple, orthogonal donors is possible. We will indicate where this is appropriate.

The donor is the most homogeneous replacement for all variables simultaneously, and hence it is most homogeneous with regard to the complete and incomplete data. This satisfies the, according to Ford (1983), most important principle in the construction of any hot deck procedure: the 'imputation model' and the 'data model' must be the same. Both imputations and quantifications maximize the homogeneity of the completed data set. As such they are relevant to the observed as well as to the missing observations.

The present imputation method is similar to missing data estimation by the EM algorithm (cf. Little and Rubin, 1987) in that both methods optimize an objective function over the imputations. Moreover, both methods consist of the two main steps: an Expectation (E) step that completes an incomplete data matrix, and a Maximization (M) step that estimates the model parameters. However, there are also substantial differences. We use Least Squares instead of Maximum Likelihood, we do not make any distributional assumptions, and we provide discrete instead of fractional imputations.

We like to emphasize at this point that maximizing consistency is by no means the only valid or useful criterion to find missing information. Suppose that we are interested in demonstrating that two variables are independent of each other. In that case, it will be clear that maximizing consistency is a bad idea since it moves us further away from the independence model. A more natural alternative here would be to do the opposite, that is, to minimize homogeneity. There exists yet no unbiased technique nor a general purpose strategy for dealing with missing data. Multivariate optimality for imputation of missing data is about impossible to define without violating some statistical model or another. Nonetheless, *if* we believe that the observed data tell us something about the missing data—and this is a fundamental assumption of all hot deck methods—then maximally consistent replacements will be attractive in general.

Good reviews imputation techniques for categorical data are Kalton and Kasprzyk (1982) and the three volumes edited by Madow, Olkin and Rubin (1983). The annual proceedings of the section of the survey research methods of the American Statistical Association offer a continuing story on the handling of missing data in survey research. The primary source for Maximum Likelihood models for missing categorical data is Little and Rubin (1987). Handling missing data in experimental designs is discussed in Dodge (1985). For multiple imputation, in which not just one but many replacements are searched, see Rubin (1987). Hedges and Olkin (1983) give a selected and annotated bibliography on incomplete data. Ford (1983) summarizes many hot deck strategies. Little and Rubin (1990) provide a recent overview of missing data strategies in the social sciences.

The structure of this paper is as follows: first, we discuss a small imputation example. After this, we define the consistency measure and we introduce two loss functions, one for numerical, and one for categorical data. Throughout the paper we will almost exclusively deal with the categorical problem. Subsequently, we relate the consistency criterion to other psychometric theory and we indicate a number of similar approaches to missing data. Computational details of our method are given next. Practical use of the method is illustrated by some examples, one of them concerning multiple imputation. Finally, we summarize the main results and we discuss some practical implications and future work.

TABLE 1
Example data.

| person | income | age | car |
|---|---|---|---|
| 1 | *a* | young | jap |
| 2 | middle | middle | am |
| 3 | *b* | old | am |
| 4 | low | young | jap |
| 5 | middle | young | am |
| 6 | high | old | am |
| 7 | low | young | jap |
| 8 | high | middle | am |
| 9 | high | *c* | am |
| 10 | low | young | am |

## Example

In order to be able to grasp the nature of consistent imputations, we discuss the small, artificial data listed in Table 1. This table contains 10 observations on three categorical variables. 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, we find `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. We simply look for similar rows. This is the same as saying that variables must be as homogeneous as possible, i.e., measure the same thing. So here we end up with two homogeneous groups with three members each.

Since there are 3 missing values, each with 3 categories to choose from, the total number of different solutions is $3 \times 3 \times 3 = 27$. Table 2 lists the amount of consistency for each of these solutions. The exact definition of consistency, expressed as fit, can be found in the next section.

Because the example is deliberately easy and somewhat trivial, the most consistent solution '`l h o`' can be derived by eye-balling alone. In more realistic situations, eye-balling is usually not enough. First, the best solution may contain new, previously unobserved, profiles. It will be difficult to find such combinations. Second, because imputations depend on each other, optimal consistency becomes hard to detect for more than three or four missing values.

For categorical data, Wilks procedure—filling in the average—boils down to selecting the modal category. The corresponding solutions in the example are '`l l y`' and '`h h y`'. These imputations have consistencies of 0.70104 and

TABLE 2
Consistency of all possible imputations for Table 1.

| *a* | *b* | *c* | fit | *a* | *b* | *c* | fit | *a* | *b* | *c* | fit |
|---|---|---|---|---|---|---|---|---|---|---|---|
| l | l | y | .70104 | m | l | y | .63594 | h | l | y | .61671 |
| l | l | m | .77590 | m | l | m | .72943 | h | l | m | .66458 |
| l | l | o | .76956 | m | l | o | .72636 | h | l | o | .65907 |
| l | m | y | .78043 | m | m | y | .70106 | h | m | y | .70106 |
| l | m | m | .84394 | m | m | m | .77839 | h | m | m | .74342 |
| l | m | o | .84394 | m | m | o | .77839 | h | m | o | .74342 |
| l | h | y | .78321 | m | h | y | .73319 | h | h | y | .68827 |
| l | h | m | .84907 | m | h | m | .80643 | h | h | m | .74193 |
| l | h | o | .84964* | m | h | o | .80949 | h | h | o | .74198 |

TABLE 3
Donor scores and scale values for the optimal imputation.

| | | | | donor scores | | variable | | scale values | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | initial | final | | | initial | final |
| 1 | *l* | y | j | -1.43 | -1.33 | inc | low | -1.13 | -1.15 |
| 2 | m | m | a | 0.79 | 0.66 | | middle | 0.41 | 0.33 |
| 3 | *h* | o | a | 1.02 | 1.00 | | high | 1.06 | 0.98 |
| 4 | l | y | j | -1.41 | -1.33 | | | | |
| 5 | m | y | a | 0.04 | -0.01 | age | young | -0.96 | -0.92 |
| 6 | h | o | a | 1.11 | 1.00 | | middle | 0.92 | 0.79 |
| 7 | l | y | j | -1.41 | -1.33 | | old | 1.07 | 1.00 |
| 8 | h | m | a | 1.05 | 0.92 | | | | |
| 9 | h | *o* | a | 1.02 | 1.00 | car | jap | -1.41 | -1.33 |
| 10 | l | y | a | -0.58 | -0.59 | | am | 0.63 | 0.57 |

0.68827 respectively, which illustrates the well known fact that Wilks method tends to discard between-groups variance.

The obvious difficulty with categorical data is that distances between profiles cannot be easily derived; it makes for example little sense to substract Japanese from American cars. We deal with categorical variables by first transforming them into numerical data by quantifying each category separately. Subsequently, the resulting numerical variables combine into the donor. Larger differences between profiles result in larger 'distances' on the donor. Table 3 lists those donor scores for each observation and the scale values of the categories, both before and after imputation.

The initial solution ignores missing data, an option known as 'missing passive' (cf. Gifi, 1990: p. 136). In this case, the donor values for subjects 1, 3, and 9 are based on two, instead of three, observed categories. As will be shown below, a scale value for a category is equal to the average of all donor scores that fall into that category. For example, the initial value of `low` is equal to $(-1.41 - 1.41 - 0.58)/3 = -1.33$, i.e., the average donor score of observations 4, 7, and 10.

Let us now try to impute the incomplete entry in profile 1. The initial donor score of the profile is $-1.43$. To complete the data, we may pick any of the three income categories. The scale values of these categories are $-1.13$,

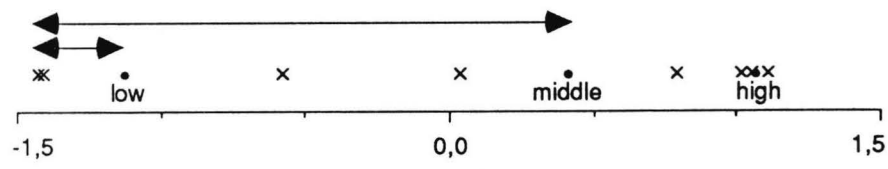


FIGURE 1
Joint plot of 10 object points (×) and the categories of income (•).

0.41, and 1.06. The joint scale of the initial donor scores and the category points is plotted in Figure 1.

The most consistent imputation is that category whose scale value is closest to the donor. Here, $-1.13$ is closest to $-1.43$, so we choose **low**. Apparently, subject 1 has most in common with profiles 4, 7, and 10. So, when compared to the other two income classes, the low income group is most similar to profile 1. Consequently, we borrow the replacement value from this group. Since middle and high income groups are more distinct, imputing **middle** and **high** will increase the within-groups variance more than necessary, and so, these values should not be used as stand-ins.

We execute the same steps for the missing data in rows 3 and 9. After all missing entries have received an initial imputation, new donor scores are computed, but now using the completed data. The entire process is repeated until the consistency of the solution does not rise anymore. The values of the final solution are also given in Table 3.

## Construction of the donor

Let $x_j$ $(j = 1, \ldots, m)$ denote $m$ completely observed variables and let the random variable $z$ contain their average, i.e. $z = 1/m \sum x_j$. The total variation of the data can be decomposed as

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

This is a between-within partitioning of the form $T = B + W$. The *correlation ratio*, denoted by $\eta$ and defined by $\eta^2 = B/T$ measures how well the average can be considered as a representative of each $x_j$. The ratio ranges from 0 to 1. It is equal to zero if variables add up to zero. The coefficient equals 1 if all variables are identical.

The donor variable $z$ tells us something about the similarity among profiles that belong to distinct replications. Let $z_i$ for $i = 1, \ldots, n$ denote the score of the $i$-th profile on the donor $z$. The difference between $z_i$ and $z_{i'}$ is equal to some distance norm between profile $i$ and profile $i'$. The donor variable $z$ defines a metric in which observational units can be represented. We use this metric to compare different data profiles.

The correlation ratio $\eta$ can also be interpreted as a measure of how well the entity $z_i - z_{i'}$ reflects the multivariate differences between rows $i$ and $i'$ over all $x_j$. Obviously, the larger $\eta$ becomes the better the difference $z_i - z_{i'}$ portrays the similarity between $i$ and $i'$. We might say that $z$ is a satisfactory donor variable in this case. However, if variables are negatively correlated, or if only a subset of variables is highly correlated, then $\eta$ may take on seriously inflated values. Consequently, $z$ gives misleading information and we are interested in other definitions of $z$ that give higher $\eta$'s.

A better alternative then is to weigh each variable by a weight $a_j$ before computing the average. Suppose that $x_j$ has zero mean and that the average of the weighted variables is given by $z = 1/m \sum x_j a_j$. The decomposition now reads

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

which is again of the form $T = B + W$. Note that $\eta$ does not only depend on $x_j$, but also on $a_j$, so we maximize $\eta$ over $a_1, \ldots, a_m$. In order to avoid the trivial outcome where $z = 0$ and $a_j = 0$ for $j = 1, \ldots, m$, it is convenient to require that both $z$ and $x_j$ are unit normalized. Other choices that result in basically the same solution are also possible. This amounts to classical principal component analysis. The first principal component then acts as the donor variable. It has the pleasant property that it corresponds to the highest possible $\eta$.

Defining a donor variable for categorical data is slightly more complicated. Essentially, we represent $x_j$ by a binary indicator vector $g_j$ of length $k_j$, $a_j$ by a quantification vector $y_j$, also of length $k_j$, while the quantified variable $g_j{}'y_j$ has zero mean. Integer $k_j$ denotes the number of categories of the $j$-th variable. We represent each score as a vector $g_j$ such that

$$g_{jk} = \begin{cases} 1, & \text{if the observation falls into category } k \text{ of variable } j; \\ 0, & \text{otherwise.} \end{cases}$$

Suppose that we assign a category weight, or scale value, $y_{jk}$ to every category. We collect these weights into a column vector $y_j$. The scalar expression $g_j{}'y_j$ then yields a quantified score. The average of the $m$ scaled categorical variables is $z = 1/m \sum g_j{}'y_j$. The between-within partitioning can now be written as

$$\sum_{j=1}^{m} (g_j{}'y_j)^2 = mz^2 + \sum_{j=1}^{m} (z - g_j{}'y_j)^2.$$

Like above, we assume that $z$ is unit normalized, so the between variation is standardized to $m$. The vectors $y_1, \ldots, y_m$ contain the free parameters. Procedures for finding optimal $y_1, \ldots, y_m$ are known as homogeneity analysis, multiple correspondence analysis, dual scaling and others (see Gifi, 1990).

These techniques usually consider several—mostly two—orthogonal sets of $z$'s, with corresponding $\eta$'s. We define the donor variable $z$ as the set of numbers that maximizes $\eta$.

## Imputation

Until here all results apply to complete data. We now discuss missing data. Let $\Omega$ denote the set of all *observed* variables and let the symbol $x_j^\star$ stand for an imputed value. Obviously,

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

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

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

so that the consistency of the data is again equal to

$$\eta^2 = \frac{B}{T} = \frac{mz^2}{\sum_{j=1}^{m} (x_j a_j)^2}.$$

Since $T = B + W$ the maximum of $\eta^2$ 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 $z$, $a_1, \ldots, a_m$ *and* over the imputations $x_1^\star, \ldots, x_m^\star$. The corresponding loss function can be written as

$$\sigma(z; a_1, \ldots, a_m; x_1^\star, \ldots, x_m^\star) = \sum_{j \in \Omega} (z - x_j a_j)^2 + \sum_{j \notin \Omega} (z - x_j^\star a_j)^2.$$

In the same way we derive the loss function for discrete data as

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

Let $\sigma(\cdot)$ stand for $\sigma(z; y_1, \ldots, y_m; g_1^\star, \ldots, g_m^\star)$. Maximal $\eta$ is obtained by minimizing $\sigma(\cdot)$ over $z$, $y_1, \ldots, y_m$ and $g_1^\star, \ldots, g_m^\star$. The imputation problem is where to impute the '1' in the missing vector $g_j^\star$. This is a combinatorial optimization problem.

Since larger $\eta$ lead to more consistent imputations, it seems logical to look for imputations that will maximize $\eta$. We thus strike two flies at one blow:

imputations will not only amplify the structure of multivariate row differences as summarized by $z$, but, at the same time they cause $z$ to be a more adequate composite of those differences. This principle induces imputations such that similar looking units become even more alike while plainly different units grow even more distinct under imputation. Dependencies in the data are thus extrapolated to the missing entries.

One might consider minimizing $\sigma_{\text{passive}} = \sum_{j\in\Omega}(z - g_j{}'y_j)^2$ only, thus by ignoring all missing data. This approach is known as 'missing deleted' or 'missing passive' (Gifi, 1990). Since quadratic terms are always positive, it follows that $\sigma_{\text{passive}} \leq \sigma(\cdot)$. Another possibility, also to be found in Gifi and known as 'missing multiple', is a crude imputation method which defines a new category for each missing value. The main problem with this approach is that it introduces categories with a single observation. Because we need more parameters to fit the same problem, it follows that $\sigma_{\text{multiple}} \leq \sigma(\cdot)$. Note that Gifi's use of the term 'missing multiple' bears no relation with Rubin's multiple imputation.

## Maximizing consistency by imputation

The search for scores that maximize consistency is deeply rooted in psychometrics and the following results are mainly due to this development. We defined a measure for consistency as

$$\eta^2 = \frac{mz^2}{\sum(x_j a_j)^2} = 1 - \frac{\sum(z - x_j a_j)^2}{\sum(x_j a_j)^2}.$$

It is known that $\eta^2$ is proportional to the largest eigenvalue $\lambda_+^2$ of the correlation matrix $R$ of (quantified) variables, i.e.,

$$\eta^2 = \frac{1}{m}\lambda_+^2.$$

We also know that $\eta^2$ is equal to the averaged squared correlations between the quantified variables and donor variable, i.e.,

$$\eta^2 = \frac{1}{m}\sum_{j=1}^{m} r^2_{z,g_j{}'y_j}.$$

The average correlation among variables,

$$\bar{r} = \frac{1}{2m(m-1)}\sum_{j<j'} r_{jj'},$$

is a well known measure for internal consistency. This measure is also used for categorical data, if computed from the optimal scores instead of the raw data. The average correlation is proportional with Cronbach's $\alpha$, defined by

$$\alpha = \frac{m}{m-1} \frac{\sum_{j \neq j'} r_{jj'}}{\sum_{j,j'=1}^{m} r_{jj'}}.$$

This is a very popular statistic in item analysis and questionnaire research. The relationship between these two indices is (Cronbach, 1951)

$$\bar{r} = \frac{\alpha}{m - (m-1)\alpha}.$$

Lord (1958) demonstrates that

$$\eta^2 = \frac{1}{1 + (m-1)(1-\alpha)},$$

so maximizing $\eta$ over the missing data also maximizes Cronbach's $\alpha$, $\bar{r}$, $\lambda_+^2$ and related measures.

In a well-readable paper, Gleason and Staelin (1975) propose the redundancy index

$$\varphi = \sqrt{\frac{\sum_{j<j'} r_{jj'}^2}{m(m-1)}},$$

that measures the average level of correlation among the variables. The index is bounded by $0 \leq \varphi \leq 1$ and it is much related to the consistency measures given above. We observed the obvious relationship between $\eta^2$ and $\varphi$ for 10 different consistency levels and 5% missing data. A total of 25 replications, each with random missing data, occur within each level.

The data used in Figure 2 are artificial data with systematically varying degrees of consistency. An amount of 5% randomly distributed missing data were created for each replication. Both $\eta^2$ and $\varphi$ increase with the consistency of the data, and both do so in roughly the same way.

Maximizing $\eta$ also influences the value of the $\chi^2$-statistic. For categorical variables, there exist $q = \sum_{j=1}^{m}(k_j - 1)$ independent solutions for $\sigma(\cdot)$. Let these solutions be ordered by their correlation ratios such that $\eta_s \geq \eta_{s+1}$ for $s = 1, \ldots, q-1$. It is known that

$$\chi^2 = n \sum_{s=1}^{q} \eta_s^2,$$

where $n$ is the number of observations. Since the imputation technique maximizes $\eta_1^2$ the value of the total $\chi^2$ will increase compared to random imputation. Imputation introduces a departure from the independence model. Note

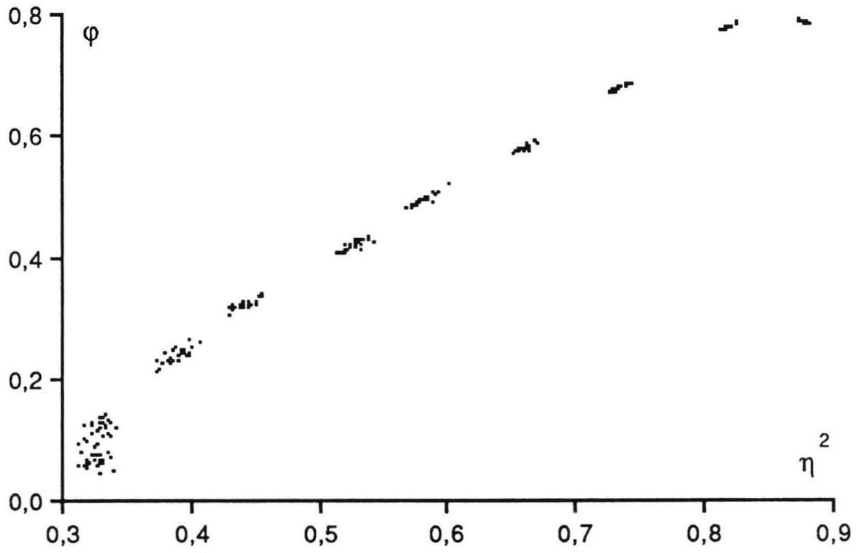


FIGURE 2
Fit ($\eta^2$) vs. correlation level ($\varphi$) for 10 consistency levels.

that maximizing the total $\chi^2$ per se is also possible by taking all $q$ solutions into account simultaneously.

Finally, it is known that maximizing $\eta^2$ will maximize the linearity of bivariate regressions. An obvious advantage is that missing data are replaced such that it becomes more reasonable to describe the data by any linear model. We refer to de Leeuw (1988) for more details.

We finish this section with the following. The idea to maximize consistency by imputation is not entirely new. Gleason and Staelin (1975) replace correlations between numerical variables by estimates that maximize the consistency of the completed data. This method is a modification of the imputation techniques proposed earlier by Dear (1959) and Buck (1960). Gleason and Staelin treat categorical data by an ad hoc rounding procedure (p. 244). Unlike the numerical case, they do not present any simulation results for their discrete imputation method. 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 they do not pursue the matter any further. Nishisato (1980) wants to impute and quantify categorical data, just like in this paper, but does not present a solution to the problem of selecting the optimal category to be imputed.

A difficulty with imputing categorical data in general is that one has a limited set of donor categories to choose from and no distance measure between them. One can quantify categories and use the Euclidean distances—as we do—or try to find margins that optimize consistency. Greenacre (1984: p. 237) does the latter by imputing 'consistency optimizing' rounded estimates of marginal frequencies.

## Computation

Finding optimal imputations for the numerical case corresponds to minimizing $\sigma(z; a_1, \ldots, a_m; x_1^\star, \ldots, x_m^\star)$ over $z$, $a_1$, $\ldots$, $a_m$ and $x_1^\star$, $\ldots$, $x_m^\star$ for given $x_1$, $\ldots$, $x_m$. We alternate over different subsets of parameters by Alternating Least Squares. It will be clear that the loss is minimized over $z$ if we compute the unit-normalized version of $z := 1/m \sum_{j=1}^m x_j a_j$.

The problem of finding optimal imputations for the $j$-th variable is equal to minimizing $(z - x_j^\star a_j)^2$ over $x_j^\star$. This seems simple—divide $z$ by $a_j$ and normalize the result—but there is a complication. The problem is that unrestricted minimization of $(z - x_j^\star a_j)^2$ generates imputations with ridiculously large magnitudes. This happens if there are two imputed values within the same observation. These values both increase boundlessly in order to blow up the correlation. The result is that after standardization the observed values are practically reduced to zero, while the imputations account for all variation. We prevent this type of degeneracy by requiring that the variances of the observed and the imputed values are to be equal.

Finding the minimum over $a_1, \ldots, a_m$ is equivalent to solving $m$ separate bivariate regression problems. Since each of the three substeps lowers the loss over a different set of parameters, alternating these steps leads to an overall minimum.

For discrete data, the problem is to find the minimum of $\sigma(\cdot)$ over $z$, $y_1, \ldots, y_m$ and $g_1^\star, \ldots, g_m^\star$. The minimum over $z$ is easily found by setting $z := 1/m \sum_{j=1}^m g_j' y_j$ and normalizing the result.

The binary nature of $g_j^\star$ turns the imputation step into a combinatorial optimization problem. For given $j$, we must minimize $\sigma(\cdot)$ over $y_j$ and $g_j^\star$ simultaneously. This type of problem frequently occurs in cluster analysis; it is known as the *sum-of-squares partitioning* problem. See for example Späth (1985) for a detailed treatment. We use a modified version of the so-called $k$-means algorithm. The $k$-means algorithm iteratively relocates classifications one by one. The obvious modification is that all nonmissing entries remain tied to their categories and are never relocated.

The modified $k$-means algorithm works as follows. Suppose we start with some initial imputation of the missing data. We examine each imputation one after another and perform a check 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 will be updated accordingly. The process is repeated until no relocations occur anymore.

The 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 weights. Suppose that observation $i$ has a score $z_i$ and that we move imputation $g_i^\star$ from $s$ to $t$. Fisher (1958) shows that the new loss will be equal to

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

where $\sigma^\star(\cdot)$ denotes the current loss. The imputation rule therefore 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 then 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 $y_s$.

It will be clear that, just because we have moved an imputation from one category to another, the corresponding weights $y_s$ and $y_t$ are no longer optimal. It is easy to see by applying the familiar least squares formula that—like in normal homogeneity analysis—the weights that minimize $\sigma(\cdot)$ will always be equal to the centroid of all $z$ belonging to that category. It is of course possible to recompute the two centroids simply by averaging over all objects in those categories, but there exists a much more efficient way to update. 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 $d_t$. Using these quantities, the new centroid 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

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

We assume that $d_s \geq 1$, i.e., each category has at least one observation, so that division by zero does not occur. Both formula's are independent of the number of observations. This makes them very efficient updates, especially if the number of observations if large. Alternating the relocation step and the update step defines the $k$-means algorithm. Each step lowers the loss, so alternating the steps also lowers the overall loss.

Since multi-dimensional versions of the $k$-means algorithm are readily available, generalizing to more than one donor is straightforward. If we require that the donors are orthogonal to each other, then differences among profile points can be measured in squared Euclidean distances. We use the Gram-Schmidt method to obtain orthogonalized donors.

The method is implemented in a computer program called MISTRESS, written in the C programming language. Since the computationally most demanding operation is the normalization of $z$, the program is quite fast. Timings for problem sizes of 100, 500, 1000, and 2500 rows by 6 columns and 8% missing data on a standard Macintosh II are 12, 50, 110, and 270 seconds respectively. The program spends about 70% of the time to find the initial solution.

## Local minima

It is well-known that the regular $k$-means algorithm is very sensitive to local minima. In our case, this problem is likely to be less severe since the missing entries are the only candidates for the relocating operation. Consequently, the number of alternative, suboptimal solutions is usually considerably less than in cluster analysis.

A small recovery study on artificial data investigates the existence of local minima. It refers to 10 datasets, each consisting of 100 subjects, 7 variables with 5 categories each, and 5% random missing data. The average intercorrelation is systematically varied from $\overline{r} = 0.00$ to $\overline{r} = 0.90$ with a step size of 0.10.

Two methods to find a starting allocation of missing entries are used: random and passive. The random procedure imputes a category that is randomly drawn with a probability proportional to the observed marginal frequencies. The passive method uses the 'missing passive' solution and allocates each observation to a category that is closest is the sense defined above. After the initial imputations are found, the method iterates over $z$, $y_1, \ldots, y_m$ and $g_1^\star, \ldots, g_m^\star$ until the difference between two consecutive values for $\sigma(\cdot)$ is less than 1.0E−7. We compute 25 replications per condition, so we have a total of $10 \times 2 \times 25 = 500$ analyses.

Convergence usually occurs after about 10–20 iterations, somewhat depending on the amount of intercorrelation—we found that medium levels need fewer iterations—and the starting method. If local minima do not occur, all points within one correlation level coincide. Especially the lower levels levels have multiple solutions in our case, thus demonstrating the existence of local minima.

A second question of interest is how diverse those local minima are, i.e. are the locally optimal solutions alike, or are they entirely different? A simple statistic in this respect is the difference between the maximum and minimum

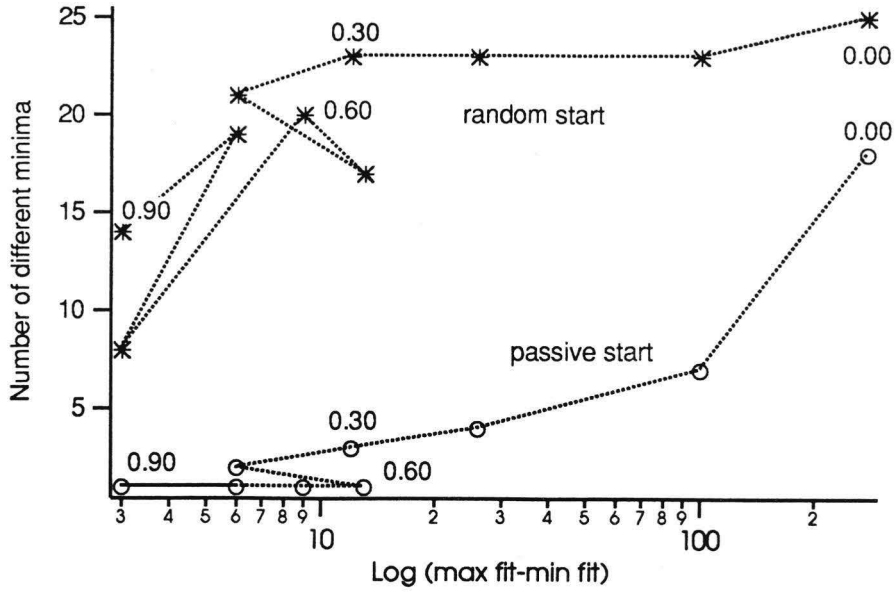


FIGURE 3
Log of fit differences vs. number of minima for two starting methods and 10 correlation levels.

fit values of the 25 replications within the same level. If this bandwidth is equal to zero, no local minima occur.

Figure 3 graphs $\log(\sigma_{\max} - \sigma_{\min})$ versus the number of different minima per level, for both random and passive starts. Each point is based on 25 replications. Points beyond a correlation level of 0.40 are not plotted for the passive method since all 25 replications appeared to be identical. The lower correlation levels yield many different solutions for both random and passive starts. The number of distinct solutions found for the passive start rapidly decreases as the internal correlation goes up. In constrast, the random starting method keeps producing about 15 local minima. As the data become more consistent, the range of fits decreases for both methods, though the pattern for the passive method is much more outspoken. The use of passive starts beyond $\overline{r} = 0.40$ generates exactly the same solution for all 25 replications. Passive is clearly superior to random, both in terms of the number of minima and in terms of fit.

A third question is whether the obtained solution is close to the global optimum. Since we do not know the actual globally optimal solution, this question cannot be answered right away. If we take the best fitting solution of the random method as a provisional global optimum, we find that the solution obtained by the passive method for $\overline{r} = 0.40$ and higher is close to, or equal to the provional global minimum in terms of fit. So the solution provided by the passive method is near the global optimum.

The simulation study indicates that above an average intercorrelation of, say 0.30–0.40, distinct local minima hardly occur anymore, assuming we start from a good initial imputation. We must 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 likely to rise.

## Multiple imputation

A drawback of any imputation method that imputes a single value is that the precision of the imputations is unknown, i.e. the variance is not estimated. In MISTRESS one could say that the imputation variance is equal to zero since there is only one imputation that maximizes consistency. This shows much confidence in the appropriateness of consistency as a criterion and in the reliability of the data. According to Rubin: "It is of no use looking for the 'best' or 'most appropriate' imputation. Such a thing simply doesn't exist." (cf. Rubin, 1987). What is best for one model doesn't work for another. So one has to make a distinction between the optimal value in terms of the one closest to the real, but unobserved, value and an imputation that is best in some model sense. Such values coincide if we succeed in finding that only model that generated the data; a desirable but rarely attained state of affairs, as every data analyst knows. Only in simulation studies, where indeed reality is artificially simulated and thus grossly simplified, one can hope for and achieve the coincidence of such imputations.

A different approach is to estimate the variance of imputation by generating not one, but several, say 3 to 5, completed matrices. Imputations are to be drawn from a posterior predictive distribution, or from decent approximations thereof. The spread of the imputations then conveys roughly how imputations vary. Rubin (1987) shows for a large class of statistical models that, after a model is separately fitted on each completed data matrix, simple pooling procedures can be used to obtain unbiased estimates of model parameters and the associated variances. The individual imputations do not have to be very precise, as long as together they estimate the variance. Because multiple imputation involves a lot of work, it is worth the effort if it concerns a large body of data that is to be used by several researchers applying different models and different subsets of the data on various occasions. See also Schnell (1986: p. 227).

There exist various sampling methods to compute prediction densities, like the Gibbs sampler (Gelfand et. al., 1990). In psychometrics, the combination of multiple imputation and sampling in various combinations is discussed by Rubin (1990). For categorical data multiple imputations are to be drawn from a predictive distribution of categories. One can define such a distribution in several ways. The dominant distinction lies between *implicit* and *explicit* models. If we use a specified distribution to this purpose like the normal we

use an explicit model. Often there exists no proper argument to select an explicit model, and thus an implicit model, or implicit distribution is used. The most implicit model is the traditional hot deck method where the value of the preceding observation is imputed. Multivariate simultaneous consistency is a less implicit model.

Because there is only one optimal imputation per missing value it is impossible to generate multiple imputations by just maximizing consistency. MISTRESS yields a crisp 'all-or-none' predictive distribution for each incomplete response pattern, which is not very useful in the context of multiple imputation. The technique simply imputes the most likely category, i.e. the one with the highest probability (in terms of consistency). For multiple imputation, we must have some way to even out the predictive category distribution so that all categories are candidates for impution, though with varying probabilities. It would require another paper to discuss MISTRESS as a way to create posterior predictive distributions of missing data. Here we only mention some possibilities of doing so as a way to apply the method.

Suppose we obtain a single set of imputations by maximizing consistency, i.e. the solution that imputes the nearest category. By approximating posterior predictive distributions, $\Pr(X_{\text{mis}} \mid X_{\text{obs}})$ we introduce a Bayesian aspect. Let $p_{ijk}$ denote the probability that category $k$ of variable $j$ is the imputation for object $i$, then we can specify the following:

a) *a density distribution based on inverse distances*
We assume that $p_{ijk}$ is inversely related to the squared distance between the scale value $y_{jk}$ and the donor score $x_i$. More precisely, we define

$$\overline{p}_{ijk} = \begin{cases} \frac{1}{t}, & \text{if } (x_i - y_{jk})^2 \leq t; \\ \frac{1}{(x_i - y_{jk})^2}, & \text{otherwise,} \end{cases}$$

so that,

$$p_{ijk} = \frac{\overline{p}_{ijk}}{\sum_{k=1}^{k_j} \overline{p}_{ijk}}$$

are properly scaled probabilities. A leveling parameter $t$ prevents unrealistic, large probalities caused by division by a small denominator. Selecting a proper leveling parameter requires some experimentation with the data, and depends on the shape of the predictive distribution. Choosing $t$ such that the probability of drawing the closest category is on the average not larger than 0.50 or 0.60 seems a sensible rule of the thumb.

b) *a density distribution based on multiple donors*
We mentioned that multiple orthogonal $z$'s with corresponding $y$'s can be considered as well. We can use each column of $Z$ to generate a successive, separate imputation instead of using the first column of $Z$ with the largest

consistency only. The columns of $Z$ are ordered by their respective contribution to the overall consistency, denoted by $\eta_s^2$. The inverse of the relative contribution

$$p_{ijk_s} = \frac{\eta_s^2}{\sum_{s=1}^{q} \eta_s^2}$$

defines the probability that an imputation is sampled from the imputations corresponding to the $s$-th column of $Z$. These probabilities are independent of $i$ and $j$. The range of potential categories to be chosen is then restricted by their occurrence in one of the $q$ imputation values. If the same category value occurs more than once as an imputation the probabilities for different $s$ add up. One could sample as many times as one likes but just as many draws as the number of columns of $Z$ seems reasonable.

c) a *density distribution based on conditional frequencies*
One of the oldest methods assumes that $p_{ijk}$ is proportional to the observed frequency of category $k$ of variable $j$. This is a very simple way to define a predictive distribution, but it uses only univariate information. If we crosstabulate the data, each cell corresponds to a possible response pattern, and one may use the conditional frequencies instead. Note that this way of deriving the distribution will only be effective if the cells in the multidimensional crosstabulation contain a sufficient number of observations. In practice, this implies that the number of variables is limited. If an observation is missing on two or more variables, we can alternate over the imputations.

These alternatives yield different predictive distributions. We do not know how this effect the results. We expect that differences will be relatively small, but more research is needed to confirm this idea. In the next section, we apply the option of inverse donor distances, with in this case quite satisfactory results.

## 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 1977–1986. The data are compiled and made available to us by Anneke Bloemhoff of NIPG-TNO. As is often the case in large surveys, not all questions were posed at each occasion. Consequently, when taken together, the data contains many systematic missing entries. This example illustrates how MISTRESS can be used to find imputations for those unknown values.

The analysis sample consists of 7332 individuals. For each person, we have scores on five labour conditions. These are labeled dirty (D), heavy (H), risky (R), stench (S) and noise (N). Each subject responded whether the attribute was applicable to his, or her, job. For a subgroup of 5750 people we

TABLE 4
Single imputation LSO table (**FAT** = imputation).

| labour conditions | professional category | | | | | | | donor score |
|---|---|---|---|---|---|---|---|---|
| **DHRSN** | **MAN** | **ADM** | **COM** | **SCI** | **SER** | **AGR** | **IND** | |
| 11111 | 1 | 1 | 1 | 6 | 2 | 5 | 64 **19** | 3.18 |
| 11110 | 0 | 0 | 0 | 6 | 3 | 7 | 11 **10** | 2.68 |
| 10111 | 1 | 1 | 0 | 3 | 3 | 1 | 21 **6** | 2.58 |
| 01111 | 0 | 0 | 0 | 1 | 1 | 1 | 3 | 2.48 |
| 11101 | 0 | 1 | 1 | 1 | 1 | 3 | 61 **23** | 2.47 |
| 11011 | 0 | 1 | 1 | 4 | 6 | 4 | 50 **15** | 2.41 |
| 10110 | 0 | 1 | 2 | 1 | 0 | 2 | 8 **5** | 2.08 |
| 01110 | 0 | 1 | 0 | 0 | 0 | 1 | 2 **1** | 1.97 |
| 11100 | 1 | 1 | 0 | 9 | 2 | 9 | 51 **22** | 1.96 |
| 11010 | 0 | 1 | 1 | 9 | 2 | 20 | 13 **20** | 1.90 |
| 00111 | 0 | 3 | 2 | 2 | 0 | 0 | 12 **4** | 1.88 |
| 10101 | 0 | 0 | 2 | 1 | 1 | 1 | 20 **12** | 1.87 |
| 10011 | 4 | 3 | 2 | 6 | 3 | 2 | 46 **32** | 1.81 |
| 01101 | 0 | 0 | 1 | 2 | 1 | 0 | 8 **7** | 1.76 |
| 11001 | 2 | 6 | 4 | 6 | 9 | 12 | 88 **32** | 1.70 |
| 01011 | 0 | 0 | 0 | 1 | 0 | 0 | 5 **2** | 1.70 |
| 00110 | 0 | 0 | 0 | 1 | 1 | 0 | 2 **3** | 1.38 |
| 10100 | 0 | 1 | 0 | 1 | 4 | 3 | 14 **10** | 1.37 |
| 10010 | 0 | 2 | 1 | 2 | 2 | 13 | 17 **10** | 1.31 |
| 01100 | 0 | 0 | 3 | 6 | 0 | 1 | 10 **9** | 1.26 |
| 01010 | 0 | 1 | 0 | 3 | 4 | 0 | 4 **3** | 1.20 |
| 11000 | 2 | 6 | 16 | 21 | 38 | 81 | 95 **81** | 1.19 |
| 00101 | 1 | 16 | 3 | 10 | 6 | 2 | 15 **14** | 1.17 |
| 00011 | 3 | 19 | 6 | 16 | 6 | 0 **29** | 28 | 1.00 |
| 10001 | 8 | 11 | 6 | 20 | 14 | 10 **48** | 103 | 0.99 |
| 01001 | 2 | 4 | 12 | 19 | 21 | 4 **16** | 40 | 0.89 |
| 00100 | 3 | 5 | 7 | 27 | 12 **19** | 2 | 29 | 0.16 |
| 10000 | 4 | 15 | 28 | 32 | 25 **96** | 60 | 104 | 0.09 |
| 00010 | 6 | 22 | 3 | 27 | 8 **25** | 1 | 17 | 0.09 |
| 01000 | 2 | 12 | 58 | 115 | 87 **71** | 16 | 80 | -0.02 |
| 00001 | 21 | 133 | 40 | 132 | 54 **122** | 3 | 125 | -0.11 |
| 00000 | 157 **816** | 843 | 373 | 916 | 349 | 54 | 324 | -0.91 |
| | 218 **816** | 1100 | 573 | 1406 | 665 **333** | 318 | **93**.0 1470 **340** | |

also know the type of job, 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 results for single imputation, ordered by donor scores, are presented in Table 4.

The majority of employees does not work in any of the disturbing circumstances. Most discomfort is experienced by blue collar workers like labourers, farmers and service personnel. All workers experiencing at least three or more adverse conditions are assigned to the group of industrial workers. So, under maximal consistency, we expect that people with many job-related harassments are labourers. 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. This is done because this group is by far the most outspoken group.

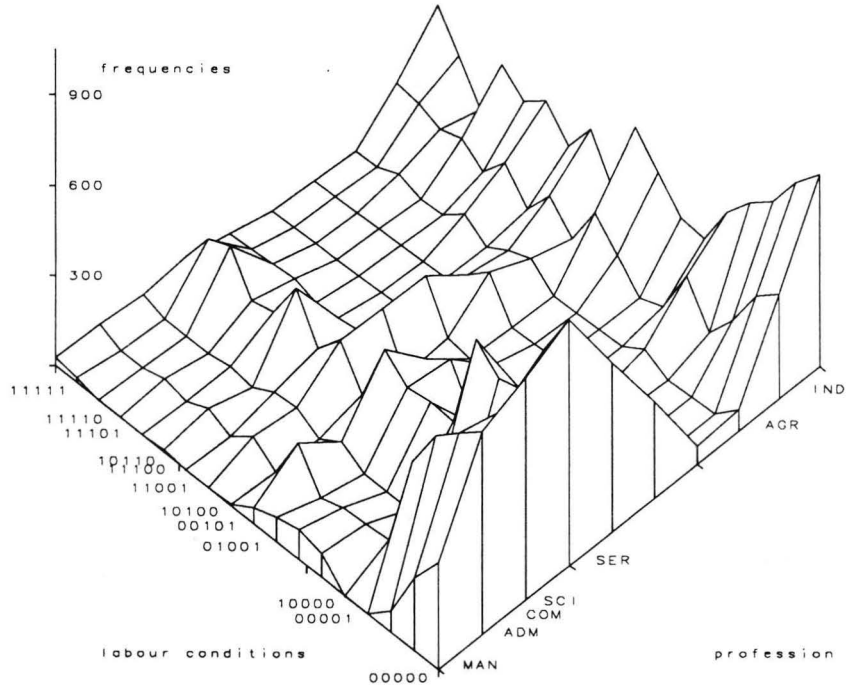


FIGURE 4
Observed frequencies ($Z$) vs. job classes ($X$) vs. labour conditions ($Y$).

The analysis shows that it is possible to find categorical imputations such that major trend in the data are extrapolated. Clearly, labour conditions are consistent with the type of work people do. 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 4. The plot shows job classes on the $X$-axis, labour conditions on the $Y$-axis, and vertically on the $Z$-axis the frequencies are shown. 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 interpretation is that, based on nuisance patterns, we have two homogeneous subgroups of jobs: blue collar and white collar jobs. A similar reasoning is to be applied to labour conditions, although they do not fall apart into two groups. The conditions with few or none nuisance parameters are somewhat separated from the rest on the $Y$-axis. A consistent subset of white collar jobs experiencing hardly any nuisance in labour conditions is thus located in

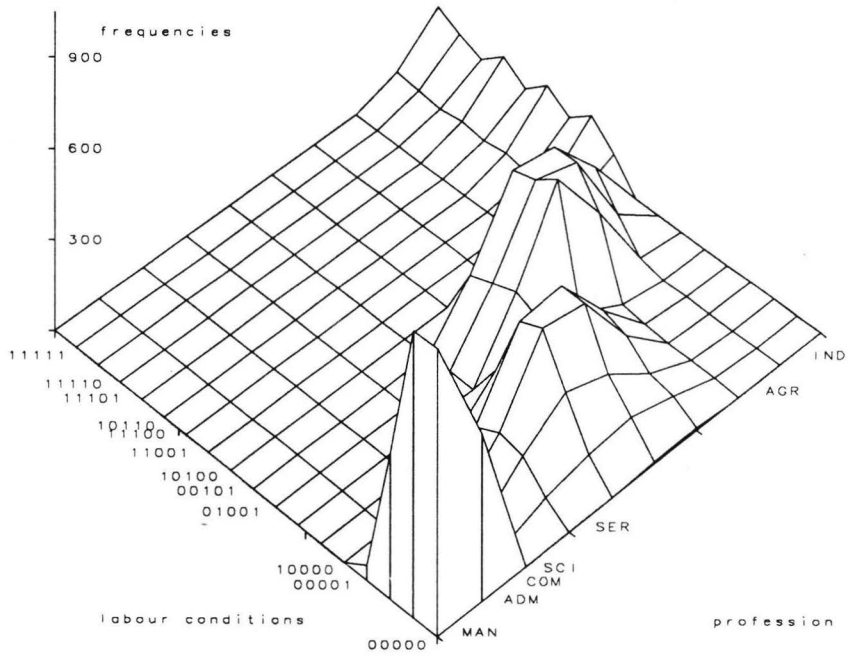


FIGURE 5
Single imputation frequencies ($Z$) vs. job classes ($X$) vs. labour conditions ($Y$).

the lower corner pointing towards us. An intuitive interpretation of the most consistent imputation is that it should disfigure the landscape in Figure 4 as little as possible. Like in Figure 4, we can picture the frequencies of the imputed data. See 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}. The albeit 'reasonable' imputations are nevertheless very 'single'. All missing data with the non-nuisance pattern are singularly attributed to managers. This is a bit peculiar since other white collar workers are also 'reasonable' candidates.

The latter observation leads automatically to the possibility of multiple imputation, where the mass of frequencies is more equally spread over other 'reasonable' candidates. The data are completed five times by drawing imputations randomly from the predictive distributions based on inverse distances, defined in the preceding section. The leveling parameter is set on $t = 0.06$, which corresponds to equalizing all $p_{ijk}$ within a range of $\sqrt{t} \approx 0.25$ from $x_i$. This means that in this case 12% of all probabilities are truncated in order to

TABLE 5
Multiple imputation LSO table (**FAT** = imputation)

| labour conditions | professional category | | | | | | | | | | | | | | | | | | donor score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **DHRSN** | **MAN** | | | | **ADM** | | | | **COM** | | | | **SCI** | | | | **SER** | | | | **AGR** | | | | **IND** | | | | |
| 11111 | 1 | 2 | | | 1 | 1 | | | 1 | 2 | | | 6 | 1 | | | 2 | 2 | | | 5 | 5 | | | 64 | 5 | | | 3.18 |
| 11110 | 0 | | | | 0 | 1 | | | 0 | 1 | | | 6 | | | | 3 | | | | 7 | 2 | | | 11 | 5 | | | 2.68 |
| 10111 | 1 | 1 | | | 1 | | | | 0 | | | | 3 | | | | 3 | | | | 1 | 1 | | | 21 | 3 | | | 2.58 |
| 01111 | 0 | | | | 0 | | | | 0 | | | | 1 | | | | 1 | | | | 1 | | | | 3 | | | | 2.48 |
| 11101 | 0 | 2 | | | 1 | 2 | | | 1 | 3 | | | 1 | 1 | | | 1 | 2 | | | 3 | 6 | | | 61 | 7 | | | 2.47 |
| 11011 | 0 | 2 | | | 1 | | | | 1 | 2 | | | 4 | 1 | | | 6 | 1 | | | 4 | 3 | | | 50 | 6 | | | 2.41 |
| 10110 | 0 | | | | 1 | | | | 2 | | | | 1 | 1 | | | 0 | | | | 2 | 2 | | | 8 | 2 | | | 2.08 |
| 01110 | 0 | | | | 1 | | | | 0 | | | | 0 | | | | 0 | | | | 1 | | | | 2 | | | | 1.97 |
| 11100 | 1 | 1 | | | 1 | 1 | | | 0 | 2 | | | 9 | 1 | | | 2 | 2 | | | 9 | 5 | | | 51 | 11 | | | 1.96 |
| 11010 | 0 | | | | 1 | | | | 1 | 1 | | | 9 | 1 | | | 2 | 1 | | | 20 | 5 | | | 13 | 12 | | | 1.90 |
| 00111 | 0 | | | | 3 | | | | 2 | | | | 2 | | | | 0 | | | | 0 | 1 | | | 12 | 2 | | | 1.88 |
| 10101 | 0 | | | | 0 | | | | 2 | | | | 1 | 1 | | | 1 | 1 | | | 1 | 4 | | | 20 | 7 | | | 1.87 |
| 10011 | 4 | 1 | | | 3 | 2 | | | 2 | 1 | | | 6 | 1 | | | 3 | 1 | | | 2 | 7 | | | 46 | 18 | | | 1.81 |
| 01101 | 0 | | | | 0 | | | | 1 | | | | 2 | 1 | | | 1 | | | | 0 | 2 | | | 8 | 4 | | | 1.76 |
| 11001 | 2 | 1 | | | 6 | 1 | | | 4 | 1 | | | 6 | 1 | | | 9 | 2 | | | 12 | 9 | | | 88 | 17 | | | 1.70 |
| 01011 | 0 | | | | 0 | | | | 0 | | | | 1 | | | | 0 | | | | 0 | | | | 5 | 1 | | | 1.70 |
| 00110 | 0 | | | | 0 | | | | 0 | | | | 1 | | | | 1 | | | | 0 | | | | 2 | 2 | | | 1.38 |
| 10100 | 0 | | | | 1 | | | | 0 | | | | 1 | | | | 4 | | | | 3 | 3 | | | 14 | 7 | | | 1.37 |
| 10010 | 0 | | | | 2 | | | | 1 | | | | 2 | | | | 2 | | | | 13 | 2 | | | 17 | 7 | | | 1.31 |
| 01100 | 0 | | | | 0 | | | | 3 | | | | 6 | | | | 0 | | | | 1 | 3 | | | 10 | 6 | | | 1.26 |
| 01010 | 0 | | | | 1 | | | | 0 | | | | 3 | | | | 4 | | | | 0 | 1 | | | 4 | 2 | | | 1.20 |
| 11000 | 2 | 1 | | | 6 | 1 | | | 16 | | | | 21 | 1 | | | 38 | 2 | | | 81 | 39 | | | 95 | 37 | | | 1.19 |
| 00101 | 1 | | | | 16 | | | | 3 | | | | 10 | | | | 6 | | | | 2 | 7 | | | 15 | 6 | | | 1.17 |
| 00011 | 3 | | | | 19 | | | | 6 | | | | 16 | | | | 6 | 1 | | | 0 | 14 | | | 28 | 14 | | | 1.00 |
| 10001 | 8 | 1 | | | 11 | | | | 6 | 1 | | | 20 | 1 | | | 14 | 1 | | | 10 | 22 | | | 103 | 22 | | | 0.99 |
| 01001 | 2 | | | | 4 | | | | 12 | | | | 19 | | | | 21 | 1 | | | 4 | 10 | | | 40 | 5 | | | 0.89 |
| 00100 | 3 | | | | 5 | 1 | | | 7 | 3 | | | 27 | 2 | | | 12 | 12 | | | 2 | 1 | | | 29 | | | | 0.16 |
| 10000 | 4 | 4 | | | 15 | 7 | | | 28 | 14 | | | 32 | 13 | | | 25 | 51 | | | 60 | 4 | | | 104 | 2 | | | 0.09 |
| 00010 | 6 | 1 | | | 22 | 3 | | | 3 | 3 | | | 27 | 4 | | | 8 | 13 | | | 1 | 1 | | | 17 | | | | 0.09 |
| 01000 | 2 | 3 | | | 12 | 6 | | | 58 | 11 | | | 115 | 15 | | | 87 | 32 | | | 16 | 2 | | | 80 | 1 | | | -0.02 |
| 00001 | 21 | 6 | | | 133 | 11 | | | 40 | 28 | | | 132 | 31 | | | 54 | 42 | | | 3 | 3 | | | 125 | 1 | | | -0.11 |
| 00000 | 157 | 314 | | | 843 | 276 | | | 373 | 98 | | | 916 | 86 | | | 349 | 32 | | | 54 | 5 | | | 324 | 5 | | | -0.91 |
| | 218 | 341 | | | 1100 | 314 | | | 573 | 172 | | | 1406 | 164 | | | 665 | 200 | | | 318 | 171 | | | 1470 | 220 | | | |

force drawing from more than one category. Although we still use the rather heuristic definition of a predictive distribution, in this example multiple imputation seems to work quite well. The multiple imputations are shown in Table 5 and in Figure 6.

The listed imputation frequencies are the average over five multiple imputation. Because of rounding errors, not all imputations exactly add up to the marginal frequencies. Comparing Figures 5 and 6, it is obvious that the multiple imputations are more spread over jobs and nuisance patterns. Both imputations, single and multiple, follow the same gradient from {white collar, no nuisance} on the bottom to {blue collar, maximal nuisance} on top.

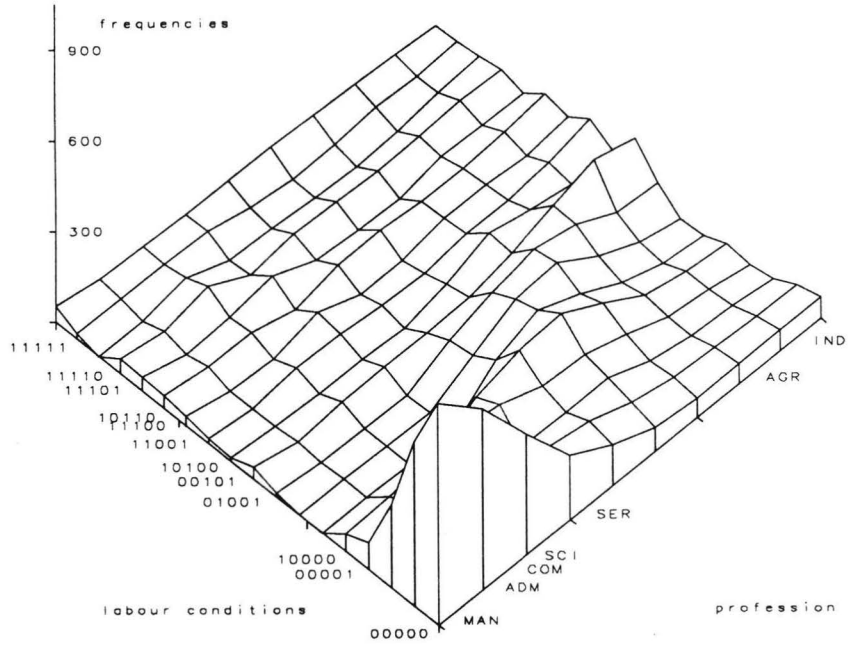


FIGURE 6
Average frequencies based on multiple imputation ($Z$) vs. job classes ($X$) vs labour conditions ($Y$).

## Contingency tables

This example compares some aspects of the treatment of missing data in loglinear analysis to the present method. We use the 2 × 2 × 2 table given in Little and Rubin (1987: p. 187). The data pertain to a partly real life, partly artificial example made up by Little and Rubin. There exist three dichotomized variables: survival (S), type of clinic (C) and amount of prenatal care (P). Type of clinic is unknown for 255 observations (= 26%), which means that 8.8% of the observations in the three-way table is missing.

Table 6 indicates that the preferred loglinear model for the table based on the 715 complete observations is [SC, PC], which means that type of clinic is related to survival and to the amount of prenatal care. Moreover, within the same clinic, survival and prenatal care are not related. Because deletion

TABLE 6
Chi-square and $p$-values under EM and MISTRESS imputation.

| model | completely classified | $p$ | imputation by EM | $p$ | imputation by MISTRESS | $p$ |
|---|---|---|---|---|---|---|
| [SP, SC, PC] | 0.044 | 0.834 | 0.057 | 0.810 | 4.77 | 0.029 |
| [SC, PC] | 0.083 | 0.959 | 0.031 | 0.984 | 9.76 | 0.008 |
| [SP, SC] | 169.469 | 0.000 | 0.002 | 0.999 | 355.16 | 0.000 |
| | $n = 715$ | | $n = 970$ | | $n = 970$ | |

of the association [SP] does not alter the fit, the more parsimonious model [SC, PC] is preferred. Model [SP, SC] does not fit at all.

The second pair of columns in Table 6 contains the $\chi^2$-values that measure the difference between the expected values under the three loglinear models and the imputed contingency tables (cf. Little and Rubin, 1987: p. 190–191). These values are not statistically significant, so all models fit the data. This is caused by, amongst others, the fact that the EM algorithm finds the most favorable imputations *given* the specific loglinear model. In general, loglinear models will fit better as more missing observations are added. In most cases, this will preserve—and even emphasize—the structure among variables as described by the loglinear model.

However, things can also go less well. Observe that model [SP, SC] now fits the imputed table ($p = 0.999$). For the completely classified table this model does not fit at all ($p = 0.000$), so filling in missing data brought about some real change. But this is a hazardous aspect of EM: suppose that we really had the 255 missing observations as in Little and Rubin, and that we applied EM. Then, we would have been pleased to find a $\chi^2$-value as low as 0.002, and we would have had little reason to question the validity our model. If we compute correlations we see what has happened: the original correlation—actually a $\phi$-coefficient here—of the omitted [PC] effect is equal to $-0.4924$, which is substantial. After EM, it is $-0.0130$! Imputation vitiated the correlation. Rubin (personal communication) shows that in this case multiple or single imputation makes no substantial difference in estimation of the model parameters for the loglinear models.

The same catch, though in the opposite direction, holds for MISTRESS. Because MISTRESS optimizes a different, almost reverse criterion, the imputed tables do not fit the loglinear model as well as the ones produces by EM. None of models fit to the imputed data. On the other hand, the $\chi^2$-statistic clearly signals the important [PC] interaction.

Both the EM algorithm and our method have the same basic weakness: if the model is wrong, imputations will be wrong. If the model prescribes that a certain interaction does not exist, then EM will do everything to make this true. In the above case, it makes a correlation of 0.49 disappear. Analogously, MISTRESS overemphasizes tiny correlations. Generally speaking, loglinear analysis stresses absence of particular interaction, while maximizing

consistency emphasizes presence of overall interaction. If we suspect that the analysis results are heavily biased by imputation, we should use these properties to our advantage.

## Conclusion

The technique proposed in this paper is fairly simple. For categorical data, it is a way of selecting the proper category to be imputed. In the context of maximizing consistency, this seems to be new. The method optimizes a well-defined and widespread criterion. Additionally, it is fast, flexible and of high practical value. Few assumptions are needed. The method stays close to the data.

It is possible to simulate various hot deck strategies. For example, by (over)weighting one of the variables we simulate a single donor variable. The method then evolves into a traditional hot deck method. In the same way, it is possible to rule out specific variables from the donor. Careful selection of variables may drastically improve the quality of the solution. Non-ignorable models, in which the pattern of nonresponse depends on the values of the data, can be evaluated by adding an indicator variable for the nonresponse distribution for each variable. Mixes of continuous and discrete data can also be analyzed. Since imputations are determined for each variable separately, mixing does not present any new problems.

There are also situations in which the method will perform less satisfactorily. The main concern is the amount of intercorrelation. If the magnitude of all correlations is below 0.20 then the method may generate imputations that overemphasize small correlations. In this case, random imputation or unconditional mean imputation often work better. It seems preferable to use MISTRESS here only in combination with a resampling method, like the bootstrap, in order to estimate the variability of consistency. If the average intercorrelation is exceeds 0.30 then recovery will be fine in general. Not only will the imputations be based on sufficient information, but also local minima become less of a problem.

A second cautionary note concerns imputation itself. However attractive the idea may seem, we must never forget that once after we have completed the data, they are partly artificial. The main pitfal is to analyze the filled-in data as if they were real, and thus overstate precision. The sagacious researcher will set up a subconscious alert that signals any pecularities that might result from imputation. According to Dempster and Rubin (1983), the entire idea of imputation carries one great seductive danger: "it can lull the user into the pleasurable state of believing that the data are complete after all."

We conclude with some words on applications and perspectives. The number of variables or observations hardly influences the computational efficiency of the method. Therefore, the technique can be used with large data

matrices. The main application field of MISTRESS is the analysis of surveys, rating scales and questionnaires. Furthermore, the relationship with Cronbach's $\alpha$ makes it attractive for dealing with missing data in psychological testing. It is easy and cheap to reiterate MISTRESS in any fashion, combined with bootstrapping, multiple imputation and the like.

The most spectacular application of the method is multiple imputation. It is ambiguous to call it an application. As a matter of fact one might devote another study to MISTRESS and multiple imputation. Analytical problem number one is to decide how implicit predictive densities are to be derived within the present framework. To this purpose, we mentioned in this paper only some intuitive possibilities. This of course needs further study. It is particularly interesting to examine the shape of the predictive density function under varying levels of consistency. If the consistency equals zero, a trivial possibility, the predictive distribution should be uniform. Conversely, if the consistency approaches unity, it should have zero variance. And then, we are back at MISTRESS.

## REFERENCES

Buck, S.F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society*, *B22*, 302–306.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297—334.

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

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

Dempster, A.P. and Rubin, D.B. (1983). Overview. In: W.G. Madow, I. Olkin and D.B. Rubin (Eds.), *Incomplete data in sample surveys, Vol. 2, Theory and annotated bibliography*, 3–10. Academic Press, New York.

Dodge, Y. (1985). *Analysis of experiments with missing data.* Wiley, New York.

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

Ford, B.L. (1983). An overview of hot deck procedures. In: W.G. Madow, I. Olkin and D.B. Rubin (Eds.), *Incomplete data in sample surveys, Vol. 2, Theory and annotated bibliography*. Academic Press, New York.

Gelfand, A.E. , Hills, S.E. , Racine-Roon, A. and Smith, A.F.M. (1990). *Illustration of Bayesian inference in normal data models using Gibbs sampling.* Unpublished manuscript, University of Connecticut.

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

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

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

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

Hedges, B. and Olkin, I (1983). Selected annotated bibliography. In: W.G. Madow, I. Olkin and D.B. Rubin (Eds.), *Incomplete data in sample surveys, Vol. 2, Theory and annotated bibliography*, 3–10. Academic Press, New York.

Kalton, G. and Kasprzyk, D. (1982). Imputing for missing survey responses. *Proceedings of the Section of Survey Research Methods, 1982, American Statistical Association*, 22–33.

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

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

Lord, F.M. (1958). Some relations between Guttman's principal components of scale analysis and other psychometric theory. *Psychometrika, 23*, 291–296.

Madow, W.G., Olkin I. and Rubin, D.B. (Eds.), (1983). *Incomplete data in sample surveys, Vol. 1, 2 and 3.* Academic Press, New York.

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

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

Rubin, D.B. (1990). *EM and beyond*. Manuscript, to appear in Psychometrika.

Schnell, R. (1986). *Missing Data Probleme in der empirischen Sozialforschung* [Missing data problems in the social sciences]. Inaugural dissertation, Ruhr Universität, Bochum.

Späth, H. (1985). *Cluster dissection and analysis*. Wiley, New York.