Listy Biometryczne - Biometrical Letters
Vol.XXV (1988), No.1,2, 31-39

MISSING VALUE IMPUTATION IN MULTIVARIATE DATA USING THE SINGULAR VALUE DECOMPOSITION OF A MATRIX

W.J. Krzanowski

Department of Applied Statistics, University of Reading
Whiteknights, Reading, RG6 2AN, Berkshire, England

Summary

Complete data matrices are necessary for some multivariate techniques, so imputation of missing values is required in certain circumstances. Existing techniques are briefly reviewed, and a new method is proposed. This method is based on the singular value decomposition of a matrix, and does not make any distributional or structural assumptions about the data. It should therefore be suitable for a large variety of situations. The method is illustrated on a small data set from which values have been deleted at random.

1. INTRODUCTION

Many biometrical problems involve multivariate data. Such data can be arranged in an (nxp) data matrix X, the $(i,j)^{th}$ element of which gives the value observed for the j^{th} response (variable) on the i^{th} individual (case) in the sample. Common biometrical techniques used to analyse such data include principal component analysis; canonical variate or discriminant analysis (if the individuals in the sample form *a-priori* groups); canonical correlation analysis (if the variables form *a-priori* groups); and cluster analysis (if some partitioning of the sample is sought). To obtain full benefit from these analyses, the data matrix must be complete. When this is so, techniques such as principal component analysis, canonical variate analysis and canonical correlation analysis which depend on linear transformations of the original variables will provide scores on all transformed variables for all individuals in the sample. Plotting the scores against each other by using transformed variables as axes is a valuable aspect of the analysis, as it gives an

Key words: eigenvalues, eigenvectors, imputation, iterative computational scheme, missing values, singular value decomposition.

optimal view of the individuals in the sample with regard to some specific objective. Thus a plot of principal component scores gives a (low-dimensional) representation of the data values in such a way as to maximise between-individual variability; a plot of the canonical variate scores gives a (low-dimensional) representation of the data values in such a way as to highlight between-group differences relative to those within groups; while a plot of the canonical correlation scores shows the nature of the scatter of sample values that gives rise to a particular canonical correlation. Patterns highlighted by these plots give valuable insights into the data that have been collected.

Frequently, however, the process of data collection does not supply a complete data matrix. Some variables are not recorded (or the values are lost) for some individuals, thereby giving gaps in the original data. For example, in an agricultural experiment some results may not be available because animals die or because plants are damaged. Sometimes the fact that the variable is missing indicates that its true value is probably unusual, and in these circumstances any mechanical method of analysis may be very misleading. On the other hand, information about some variables may simply not be readily available. This may be so particularly if the relevance of the information is uncertain, as in exploratory work. How can we make the maximum use of the available data so as to obtain the best analysis?

The literature on the analysis of partially missing data is comparatively recent, as many of the proposed methods require very heavy computations that have only become manageable since computer power increased dramatically in the late 1970's. A complete account of currently available methods is given by Little and Rubin (1987). In general, there are two possible strategies for obtaining a multivariate analysis in the face of partially missing data:

(i) use the available data to estimate all the necessary parameters (e.g. mean vectors or dispersion matrices) that are required by the analysis, and proceed without further recourse to the raw data;

(ii) use the available data to estimate all missing values, impute these estimates into the data matrix and then conduct the analysis on the completed matrix.

The former strategy will usually lead to the main aspects of the analysis, but will not enable scores to be calculated for any individuals whose data vectors are incomplete, while the latter strategy will allow computation of all scores in addition to the remainder of the analysis.

We therefore focus attention exclusively on imputation methods in this paper. Existing methods are briefly reviewed in Section 2, a new method is proposed in Section 3 and some examples of its use are given in Section 4. This method is suitable primarily for those situations in which principal component analysis is applicable. Some comments about possible generalization to other multivariate situations are made in Section 5.

2. EXISTING METHODS OF IMPUTATION

Suppose that an individual with the value of X_j missing contains the values of other variables X_k, X_m, etc, that are correlated with X_j. The earliest suggested form of imputation was to estimate the missing value x_{ij} by $\bar{x}_j$, i.e. by the mean of the recorded values of X_j (taken within a specified group if appropriate). While this method is extremely simple, its disadvantages are that:

(i) variances and covariances are systematically underestimated (a natural consequence of imputing values at the centre of the distribution), and

(ii) no information on correlations between X_j and each of the other variables is used in forming the imputed values.

To get over disadvantage (i) we could impute the value $\bar{x}_j + \varepsilon_j$ for a missing x_{ij}, where ε_j is a random quantity having zero mean and variance equal to the variance of X_j. To get over disadvantage (ii) we must consider conditional rather the unconditional means.

A more promising form of imputation is thus to substitute means that depend on the variables recorded in incomplete rows of the data matrix. If the variables $X_1,\ldots,X_p$ are multivariate normal with mean μ and dispersion matrix Σ, then the missing values in a particular case have linear regressions on the observed variables, with regression coefficients that are well-known functions of μ and Σ. The method proposed by Buck (1960) first estimates μ and Σ as the sample mean vector and covariance matrix using just the complete cases, and then uses these estimates to calculate the linear regression of the missing variables on the recorded variables for each case. Substituting the observed values of the recorded variables for a case in the appropriate regression yields predictions (and hence imputed values) for the missing values in that case. Computations can be arranged systematically using the "sweep" operator (Little and Rubin, 1987). There is still some underestimation of variances and covariances, but much less than with the unconditional means estimation. It can again be corrected by adding a small random perturbation to the conditional mean before imputation.

Beale and Little (1975) considered maximum likelihood estimation of the parameters of a multivariate normal distribution when some data values are missing, and developed an iterative scheme. This scheme is an example of the more general E-M algorithm (Dempster, Laird and Rubin, 1977) consisting of two basic steps, the expectation (E) step and the maximisation (M) step. The E step requires missing values to be replaced by estimates of their expectations given the model and current parameter estimates, and in the multivariate normal case is just Buck's method described above. The M step requires the multivariate normal parameters to be re-estimated given these imputed values and the recorded data values; the relevant estimates are set out by Beale and Little (1975). Initial

parameter estimates from complete cases start the process off, and then imputed values and parameter estimates are successively refined until they stabilise. This scheme is thus an iterated version of Buck's method and produces imputed values as a by-product of the parameter estimation.

Note therefore that all the available imputation methods (apart from the crude imputation of means) rely for their justification on the assumption of multivariate normality. Little and Rubin (1987) claim that the multivariate normal assumption can be relaxed considerably without invalidating use of the method, while Little (1988) refines the E-M algorithm to allow robust estimation of the model parameters. However, what is currently unavailable in the literature is a perfectly general imputation scheme free of any distributional constraints, and the purpose of this paper is to propose such a method.

3. AN IMPUTATION SCHEME USING THE SINGULAR VALUE DECOMPOSITION OF THE DATA MATRIX

Consider the (nxp) data matrix X, where $n \geq p$. If $p>n$ then the (pxn) transpose X' should replace X in all the following and the roles of p and n should be interchanged. Any such matrix X can be decomposed via the singular value decomposition (Good, 1969) into the form

$$X = UDV' \qquad (1)$$

where $U'U = I_p$, $V'V = VV' = I_p$ and $D = \mathrm{diag}(d_1,\ldots,d_p)$ with $d_1 \geq d_2 \geq \ldots \geq d_p \geq 0$. The matrices X'X and XX' have the same eigenvalues, and the elements d_i are the square roots of these eigenvalues; the i^{th} column $v_i = (v_{i1},\ldots,v_{ip})'$ of the (pxp) matrix V is the eigenvector corresponding to the i^{th} largest eigenvalue d_i^2 of X'X; while the j^{th} column $u_i = (u_{1i},\ldots,u_{ni})'$ of the (nxp) matrix U is the eigenvector corresponding to the i^{th} largest eigenvalue d_i^2 of XX'. Decomposition (1) has its elementwise representation

$$x_{ij} = \sum_{t=1}^{p} u_{it} d_t v_{tj} \qquad (2)$$

Krzanowski (1987) used this representation as a basis for determining the dimensionality of a multivariate data set: if the data structure is essentially m-dimensional then the variation in the remaining (p-m) dimensions can be treated as random noise. The main features of the data can thus be supposed to lie in the space of the first m principal components. The correspondence between the quantities on the right-hand side of (2) and the principal axes of the data configuration suggests, therefore, the m-component model

$$x_{ij} = \sum_{t=1}^{m} u_{it} d_t v_{tj} + \varepsilon_{ij} \quad , \qquad (3)$$

where ε_{ij} is a residual term.

Now suppose that model (3) holds for a specified value of m, but that the single observation x_{ij} is missing from the data matrix. Then x_{ij} is estimated by

$$\hat{x}_{ij}^{(m)} = \sum_{t=1}^{m} u_{it} d_t v_{tj} \qquad (4)$$

where the u_{it}, d_t, v_{tj} must be estimated from the rest of the data. The best estimates of these latter quantities will be those based on the maximal amount of data. Denote by $X^{(-i)}$ the data matrix obtained on deleting the i^{th} row from X, and by $X_{(-j)}$ the data matrix obtained on deleting the j^{th} column from X. Let the singular value decomposition of these matrices be as follows:

$$X^{(-i)} = \bar{U}\bar{D}\bar{V}' \text{ with } \bar{U} = (\bar{u}_{st}),\ \bar{V} = (\bar{v}_{st}) \text{ and } \bar{D} = \mathrm{diag}(\bar{d}_1,\ldots,\bar{d}_p) \qquad (5)$$

and

$$X_{(-j)} = \tilde{U}\tilde{D}\tilde{V}' \text{ with } \tilde{U} = (\tilde{u}_{st}),\ \tilde{V} = (\tilde{v}_{st}) \text{ and } \tilde{D} = \mathrm{diag}(\tilde{d}_1,\ldots,\tilde{d}_{p-1}) \qquad (6)$$

The maximum-data estimates of u_{it} and v_{tj} in (4) are clearly $\tilde{u}_{it}$ and $\bar{v}_{tj}$ respectively, while d_t can be estimated either by $\bar{d}_t$, $\tilde{d}_t$ or by some combination of the two. A suitable compromise seems to be $\sqrt{\bar{d}_t}\sqrt{\tilde{d}_t}$, whence an estimate of the missing value x_{ij} is given (cf Krzanowski, 1987, p.579) by

$$\hat{x}_{ij}^{(m)} = \sum_{t=1}^{m} \left[\tilde{u}_{it}\sqrt{\tilde{d}_t} \right] \left[\bar{v}_{tj}\sqrt{\bar{d}_t} \right] \quad . \qquad (7)$$

Finally, following the maximum-data precept, we use the highest value of m that we can. From (6) this is evidently p-1, so that the value which is to be imputed for x_{ij} will be

$$\hat{x}_{ij} = \sum_{t=1}^{p-1} \left[\tilde{u}_{it}\sqrt{\tilde{d}_t} \right] \left[\bar{v}_{tj}\sqrt{\bar{d}_t} \right] \quad . \qquad (8)$$

If there is more than one missing value in the data, an iterative scheme can be readily set up. Starting with initial estimates imputed for all the missing values, each missing value is re-estimated in turn using (8). Each of these estimates requires two singular value decompositions, namely those of $X^{(-i)}$ and $X_{(-j)}$ for the required i and j (using the current missing value estimates to "fill-out" X). However, singular value decomposition algorithms are computationally fast and readily available on standard software (e.g. NAG system, Numerical Algorithms Group, Oxford) so computing is not a problem. The process is iterated until stability is achieved in the imputed values. A simple initial estimate $\hat{x}_{ij}$ is provided by the mean $\bar{x}_j$ of the j^{th} variable; alternatively the initial value

$\bar{x}_j + \varepsilon_j$ as defined in the first paragraph of Section 2 is equally acceptable.

On a computational level, the best accuracy of prediction seems to be achieved when the entries in different columns of X are comparable in size and there is relatively little variation among the d_i. The most stable procedure is thus one in which the mean m_j and standard deviation s_j of column j (j=1,...,p) are first found from the values present in that column. Existing entries x_{ij} of X are then standardised to $x'_{ij} = (x_{ij} - m_j)/s_j$, estimates $\hat{x}'_{rt}$ of missing values are found by applying (8) to the standardised data, and then the final imputed values $\hat{x}_{rt}$ are obtained from $\hat{x}_{rt} = m_t + s_t \hat{x}'_{rt}$.

4. EXAMPLES

To illustrate the performance of the technique, let us consider a simple multivariate data set. Table 1 presents the data relating to 20

Table 1. Data for twenty samples of soil

Sample	x_1	x_2	x_3	x_4	x_5
1	77.3	13.0	9.7	1.5	6.4
2	82.5	10.0	7.5	1.5	6.5
3	66.9	20.6	12.5	2.3	7.0
4	47.2	33.8	19.0	2.8	5.8
5	65.3	20.5	14.2	1.9	6.9
6	83.3	10.0	6.7	2.2	7.0
7	81.6	12.7	5.7	2.9	6.7
8	47.8	36.5	15.7	2.3	7.2
9	48.6	37.1	14.3	2.1	7.2
10	61.6	25.5	12.9	1.9	7.3
11	58.6	26.5	14.9	2.4	6.7
12	69.3	22.3	8.4	4.0	7.0
13	61.8	30.8	7.4	2.7	6.4
14	67.7	25.3	7.0	4.8	7.3
15	57.2	31.2	11.6	2.4	6.5
16	67.2	22.7	10.1	3.3	6.2
17	59.2	31.2	9.6	2.4	6.0
18	80.2	13.2	6.6	2.0	5.8
19	82.2	11.1	6.7	2.2	7.2
20	69.7	20.7	9.6	3.1	5.9

samples of soil given by Kendall (1980, p.20); percentages of snd content (x_1), silt content (x_2), clay content (x_3), organic matter (x_4) and pH (x_5) are shown for each of the samples.

Note that here $x_1 + x_2 + x_3 = 100$; applying any regression-based technique to these raw data would incur multicollinearity problems, but the singular-value approach of this paper can be applied directly without any computational drawbacks. Missing values were introduced into this data

set at random as follows. Pseudo-random numbers lying between 0 and 1 were generated on the computer using the NAG package (Numerical Algorithms Group, Oxford). For a fixed value of r $(0<r<1)$, if the $(5i+j)^{th}$ random number was less than r then the element in the $(i+1,j)^{th}$ position of the data matrix was deleted ($i = 0,\ldots,19$; $j = 1,\ldots,5$). The expected proportion of missing values in the data is thus r. The technique was tried for various values of r up to 0.3 and gave satisfactory results. A summary of the outcome for one of these cases, $r = 0.2$, is given in Table 2.

Table 2. Estimates of missing values introduced into the data of Table 1 by random deletion

Matrix elements deleted	Data values in these positions	Estimates based on simple means	Estimates obtained from equation (8)
(1,3)	9.7	11.36	8.59
(2,1)	82.5	65.86	81.33
(3,5)	7.0	6.66	6.71
(4,2)	33.8	22.15	36.52
(6,3)	6.7	11.36	9.22
(6.4)	2.2	2.51	1.74
(10,1)	61,6	65.86	59.14
(10,4)	1.9	2.51	3.48
(12,3)	8.4	11.36	10.57
(12,4)	4.0	2.51	2.99
(15,1)	57.2	65.86	58.41
(16,3)	10.1	11.36	8.85
(16,5)	6.2	6.66	6.89
(17,3)	9.6	11.36	11.86
(17,4)	2.4	2.51	3.66
(18,1)	80.2	65.86	76.03
(18,3)	6.6	11.36	10.27

The expected number of missing values in this case is 20; in the actual execution of the procedure, 17 values were deleted from the data matrix. Column 1 of Table 2 gives the positions of the matrix elements that were selected for deletion while column 2 gives the true values in these positions. Column 3 then shows the simple estimate of each value given by the mean of all non-missing values in the corresponding column of X, while column 4 shows the estimates obtained from the method of Section 3 above. The improvements of the latter over the former are obvious; in only 5 cases are the estimates of column 4 further from the true values than those of column 3 (and then only marginally so), while in many of the remaining cases the column 4 estimate is substantially better than the column 3 estimate. Comparable results were obtained in all other simulations that were conducted.

5. COMMENT

Thus the method appears to be highly promising. Further investigation is clearly necessary, but the main point in its favour is the lack of dependence on any assumptions about the data (whether structural or distributional). The singular value decomposition is applicable to any numerical data matrix, so the method should be valid for a very wide range of practical situations.

The context of the above development has been that of a single unstructured data matrix X, and hence the situations in which the method is naturally applicable are those in which principal component analysis is appropriate. In some situations there is an additional a-priori structure imposed on the data matrix. For example, the n individuals in the sample may have come from k separate populations (with n_i individuals coming from the i^{th} such population and $\sum_{i=n}^{k} n_i = n$). The (nxp) data matrix may thus be subdivided by rows into k submatrices X_i, the i^{th} of which is $(n_i xp)$. A canonical variate analysis would seek an optimal representation of the data to highlight differences between the k populations. Alternatively, the p measured variables may fall into two a-priori groups of p_1 and p_2 variables respectively, and a canonical correlation analysis would seek to explore the relationships between these two sets of variables. In this case the data matrix X may be partitioned into the form $(X_1 : X_2)$, where X_i is (nxp_i) for $i = 1,2$.

If imputation of missing values is required in these situations, a simple-minded approach would be to treat each of the submatrices X_i separately by the method described in this paper. However, a referee has pointed out that such an approach will not take into account the structure of data implicit in the corresponding multivariate technique. Both canonical variate and canonical correlation analysis involve the eigenvalues and eigenvectors of one matrix with respect to another, rather than just those of a single matrix. Consequently, a missing value imputation scheme should be based on a more general singular value decomposition (see, e.g., Rao and Mitra, 1971, p.7). Such an approach also needs further investigation.

REFERENCES

Beale, E.M.L., Little, R.J.A. (1975). Missing values in multivariate statistical analysis. *J.R. Statist. Soc.* B 37, 129-146.

Buck, S.F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *J.R. Statist. Soc.* B 22, 302-306.

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood

estimation from incomplete data via the EM algorithm (with discussion). *J.R. Statist. Soc.* B 39, 1-38.

Good, I.J. (1969). Some applications of the singular value decomposition of a matrix. *Technometrics* 11, 823-831.

Kendal, M.G. (1980). *Multivariate Analysis* (2nd Ed.). London: Charles Griffin & Co.

Krzanowski, W.J. (1987). Cross-validation in principal component analysis. *Biometrics* 43, 575-584.

Little, R.J.A. (1988). Robust estimation of the mean and covariance matrix from data with missing values. *Applied Statistics* 37, 23-38.

Little, R.J.A., Rubin, D.B. (1987). *Statistical Analysis with Missing Data.* New York: Wiley.

Rao, C.R., Mitra, S.K. (1971). *Generalized Inverse of Matrices and its Applications.* New York: Wiley.

Received 26 June 1988: revised 6 February 1989

Uzupełnianie brakujących obserwacji w danych wielowymiarowych przy użyciu dekompozycji pewnej macierzy według wartości osobliwych

Streszczenie

Niektóre techniki wielowymiarowe wymagają, aby macierz danych była kompletna. W pewnych warunkach niezbędne jest więc szacowanie brakujących wartości. Praca zawiera krótki przegląd istniejących metod oraz proponuje nową. Metoda ta oparta jest na dekompozycji pewnej macierzy według wartości osobliwych i nie wymaga żadnych założeń dotyczących rozkładu lub struktury danych. Z tego względu powinna być odpowiednia dla wielu różnych sytuacji. Przedstawiono wykorzystanie metody dla małego zbioru danych, z którego pewne wartości zostały usunięte w sposób losowy.

AR ZGiSAD 294/89 n. 350