ACTA BIOL. SZEGED. 42, pp. 55-58 (1997)

# TREATMENT OF MISSING DATA IN PRINCIPAL COMPONENT ANALYSIS

Zs. GUBA, L. SZATHMÁRY and L. ALMÁSI

*Department of Human Biology, Kossuth Lajos University, Debrecen H-4010, P.O.B. 6, Hungary*

(Received: December 20, 1996)

## Abstract

Different methods for the substitution of missing data were applied in order to establish how incomplete data affect the correlation pattern of skull measurements. Principal component analysis was first performed on skulls from a real anthropological sample with no missing values. Ten, twenty and thirty per cent of the data were then dropped and the missing data were replaced by three methods: mean substitution, pairwise deletion and DEAR's method. Finally, principal component analysis was carried out on the samples with replaced values. The rotated factor matrices indicated that DEAR's method is most acceptable as concerns the correlation pattern of variables. With this method, considerable changes in the pattern of the measurements were not found after substitution of thirty per cent of the data.

*Key words*: missing data, PC analysis, correlation pattern.

## Introduction

A historical anthropological sample involves more or less fragmentary human skeletal remains. Because of the limited number of human findings from a particular historical period, it would be useful to analyse as many individuals as possible. The number of individuals analysed with multivariate methods can be enlarged by replacing the missing measurements of fragmentary bones. Several missing value methods are used in multivariate analysis. This study compares three methods to examine what percentage of the data can be replaced so that the conclusions based on principal component analysis remain correct. For this purpose, data from a real anthropological sample were used.

## Material and method

The sample contains 13 measurements on 33 male skulls from a cemetery (Alattyán-Tulát, Hungary) (Table 1) originating from the Avar Period (WENGER 1957).

In the first step, principal component analysis with the varimax rotation method was carried out on the complete sample. 10, 20 and 30 per cent of the data were then dropped, and the missing values were replaced by three methods: mean substitution, pairwise deletion and DEAR's PC method.

*Table 1.* Dimensions Used (MARTIN, 1928).

| MARTIN no. | Dimensions |
|---|---|
| 1 | Maximum cranial length |
| 5 | Basi-nasal length |
| 8 | Maximum breadth |
| 9 | Minimum cranial breadth |
| 17 | Basion-bregma height |
| 20 | Basion-nasion height |
| 45 | Bizygomatic breadth |
| 48 | Upper facial height |
| 51 | Orbital breadth |
| 52 | Orbital height |
| 54 | Nasal breadth |
| 55 | Nasal height |
| 66 | Bigonial breadth |

In mean substitution, sample means for the variables are first calculated on the basis of all available sample values. These mean values are then used for the missing values.

With pairwise deletion of missing data, each correlation coefficient is computed, using all cases with valid data for the two variables being correlated.

In DEAR's PC method, the data matrix is replaced by a standardized matrix, where 0 is used for missing values. The coefficients of the first principal component are then obtained; these are the eigenvectors of unit length associated with the largest eigenvalue of the product matrix. Any missing value in the standardized matrix is replaced by the nearest point on the first principal component. After all missing values have been estimated, the standardized matrices are transformed back to their original units (DEAR, 1959, cf. CHAN and DUNN, 1972).

Next, PC analyses with the varimax rotation method were performed on the 9 samples obtained. The results are assessed on the basis of rotated factor matrices.

## Results

### *The complete sample*

The means and standard deviation of each variable analysed for the complete sample are shown in Table 2.

*Table 2.* Means and standard deviations of the variables examined.

| MARTIN no. | Mean | SD |
|---|---|---|
| 1 | 186.8 | 6.79 |
| 5 | 102.4 | 4.26 |
| 8 | 147.4 | 5.67 |
| 9 | 99.5 | 4.73 |
| 17 | 130.3 | 5.99 |
| 20 | 112.9 | 4.64 |
| 45 | 137.1 | 4.49 |
| 48 | 69.9 | 3.61 |
| 51 | 43.0 | 2.21 |
| 52 | 34.0 | 2.10 |
| 54 | 26.4 | 2.02 |
| 55 | 53.0 | 3.27 |
| 66 | 101.8 | 7.47 |

On the basis of the rotated factor matrix of the complete data (Table 3), 75.1 per cent of the total variance can be explained by 5 factors.

*Table 3.* Variables belonging to the factors in the rotated factor matrix of the complete sample.

| Factor | Eigenvalue | Variables |
|---|---|---|
| I | 3.84 | 9, 17, 20 |
| II | 1.70 | 1, 5, 51, 54 |
| III | 1.64 | 8, 45 |
| IV | 1.49 | 48, 55 |
| V | 1.09 | 52, 66 |

As concerns the arrangement of variables, we consider the affinity of the variables 9, 17 and 20 loaded in the first factor to be the most important feature of this sample. Not only for its definitely high eigenvalue, but also because it is not common for one breadth and two height dimensions to be weighted into the same factor. In contrast, the connection of the variable pairs 8-45, 48-55 and 52-66 may be regarded as a general property of the human skull. The second factor is a combination of two length and two breadth dimensions.

*The incomplete samples (Table 4)*

*Table 4.* Variables belonging to the factors in the rotated factor matrices of the incomplete samples.

| % of missing data | DEAR's PC method | | | Mean substitution | | | Pairwise deletion | | |
|---|---|---|---|---|---|---|---|---|---|
| | Factor | Eigenvalue | Variables | Factor | Eigenvalue | Variables | Factor | Eigenvalue | Variables |
| 10 | | | | | | | | | |
| | I | 4.19 | 1, 8, 45, 54 | I | 3.69 | 5, 51, 1, 9 | I | 3.98 | 8, 54, 45 |
| | II | 1.86 | 5, 9, 17, 20 | II | 1.95 | 8, 45, 54 | II | 2.07 | 51, 66, 52 |
| | III | 1.54 | 51, 52, 66 | III | 1.63 | 17, 20 | III | 1.68 | 17, 20 |
| | IV | 1.44 | 48, 55 | IV | 1.45 | 48, 55 | IV | 1.51 | 5, 9, 1 |
| | | | | V | 1.04 | 52, 66 | V | 1.01 | 55, 48 |
| 20 | | | | | | | | | |
| | I | 4.81 | 51, 52, 66 | I | 3.75 | 8, 54, 45, 1 | I | 4.40 | 66, 51, 52, 9 |
| | II | 1.73 | 1, 5, 9, 17, 20 | II | 2.00 | 66, 51, 52, 9 | II | 2.18 | 8, 54, 45, 1 |
| | III | 1.41 | 8, 45, 54 | III | 1.62 | 17, 20, 5 | III | 1.69 | 17, 20, 5 |
| | IV | 1.26 | 48, 55 | IV | 1.40 | 48, 55 | IV | 1.49 | 48, 55 |
| 30 | | | | | | | | | |
| | I | 4.96 | 51, 52, 66 | I | 3.42 | 8, 45, 54, 1 | I | 4.27 | 8, 45, 54, 1 |
| | II | 1.62 | 1, 8, 45, 54 | II | 2.03 | 66, 51, 52 | II | 2.44 | 51, 66, 52 |
| | III | 1.35 | 5, 9, 17, 20 | III | 1.70 | 17, 5, 9 | III | 1.93 | 9, 5 |
| | IV | 1.30 | 48, 55 | IV | 1.35 | 55, 48 | IV | 1.53 | 55, 48 |
| | | | | V | 1.01 | 20 | V | 1.07 | 20, 17 |

Variables 9, 17 and 20 remain together only if DEAR's PC method is used to replace missing values. With either mean substitution or pairwise deletion, no affinity of these three variables can be observed at all.

The connection between variables 1, 5, 51 and 54 (the second factor originally) breaks up in every case. Variable 51 tends to connect with variables 52 and 66. Variables 1 and 54 are loaded into the same factor usually.

The variable pair 8 and 45 always remain together, but with a relation to other variables in the case of analyses of incomplete samples.

The arrangement of variables 48 and 55 is stable.

It is striking that the eigenvalues of special groups of variables have changed. While the eigenvalue of the factor belonging to variables 9, 17 and 20 is the highest in the matrix of the complete data, it is considerably lower in that of the incomplete samples.

## Conclusions

Of the three missing value methods used, DEAR's PC method gives the most acceptable results in a PC analysis. With this method the arrangement of variables in the rotated factor matrix proved quite correct, even in the case of 30 per cent of missing values, but the importance of the factors concluded from their eigenvalues must be treated with caution.

The other two missing value methods (mean substitution and pairwise deletion) provide acceptable results only for the variables for which a close relation is presumed to exist in each human skull, but not for the arrangement of variables presumed to be special in the sample examined.

## References

CHAN, L. S. and DUNN, O. J. (1972): The Treatment of Missing Values in Discriminant Analysis-I. The Sampling Experiment. - JASA *67*, 473-477.

DEAR, E.A. (1959): Principal Component Missing Data Method for Multiple Regression Models. SD Corp. - Technical Report SP-86.

MARTIN, R. (1928): Lehrbuch der Anthropologie. - Fisher, Jena, 2 Aufl. 2. Bd.

WENGER, S. (1957): Données ostéométriques sur le matériel anthropologique du cimetière d'Alattyán-Tulát, provenant de l'époque avare. - Crania Hungarica *2*, 1-55.