*The Korean Journal of Applied Statistics* (2008)
**21**(6), 1109–1116

# Comparison of Shape Variability in Principal Component Biplot with Missing Values

Sang Min Shin [1] · Yong-Seok Choi[2] · Nae-Young Lee[3]

[1]Dept. of Statistics, Pusan National University; [2]Dept. of Statistics, Pusan National University; [3]Dept. of Nursing, Silla University

(Received October 2008; accepted October 2008)

**Abstract**

Biplots are the multivariate analogue of scatter plots. They are useful for giving a graphical description of the data matrix, for detecting patterns and for displaying results found by more formal methods of analysis. Nevertheless, when some values are missing in data matrix, most biplots are not directly applicable. In particular, we are interested in the shape variability of principal component biplot which is the most popular in biplots with missing values. For this, we estimate the missing data using the EM algorithm and mean imputation according to missing rates. Even though we estimate missing values of biplot of incomplete data, we have different shapes of biplots according to the imputation methods and missing rates. Therefore we propose a RMS(root mean square) for measuring and comparing the shape variability between the original biplots and the estimated biplots.

Keywords: Biplots, EM algorithm, mean imputation, principal component biplot, RMS, shape variability.

## 1. Introduction

The biplots, proposed by Gabriel (1971), are the multivariate analogue of scatter plots. They approximate the multivariate distribution of a sample in a few dimensions, typically two and they superimpose on this display representations of variables on which the samples are measured. In this way, the relations between the individual samples can be easily seen and as we shall see, they can also be related to values of the measurements. However when some values are missing in data matrix, most biplots are not directly applicable.

When the incomplete cases comprise only a small fraction of all cases(say, five percent or less) the case deletion may be a perfectly reasonable solution to the missing data problem. However in multivariate settings where missing values occur on more than one variable, the incomplete cases

This work was supported for two years by Pusan National University Research Grant.

[1]Graduate student, Dept. of Statistics, Pusan National University, Jangjeon-Dong, Geumjeong-Gu, Busan 609-735, Korea. E-mail: shyaman@pusan.ac.kr

[2]Corresponding author: Professor, Dept. of Statistics, Pusan National University, Jangjeon-Dong, Geumjeong-Gu, Busan 609-735, Korea. E-mail: yschoi@pusan.ac.kr

[3]Full time lecturer, Dept. of Nursing, Silla University, Guaebop-Dong, Sasang-Gu. Busan 617-736, Korea. E-mail: naeyoungle@silla.ac.kr

are often a substantial portion of the entire data set. If so, deleting them may be inefficient, causing large amounts of information to be discarded. Moreover, omitting them from the analysis will tend to introduce bias (Schafer, 1997). It is well known that mean imputation as a particularly simple form of imputation is to estimate missing values by the mean of the recorded values. But it has some undesirable properties. Since the sample size is effectively reduced by nonresponse, standard variance formulas will underestimate the true variance (Little and Rubin, 1987). The EM algorithm is a very general iterative algorithm for maximum likelihood estimation in complete data problems (McLachlan and Krishnan, 1997; Dempster *et al.*, 1977).

In this study, we have different shapes of biplots according to two imputation methods. In particular, we are interested on the shape variability of principal component biplot with missing values according to missing rates. For this, we introduce principal component biplot and missing mechanism in Section 2. And in Section 3, we introduce the EM algorithm and mean imputation for estimating the missing data and propose a RMS(root mean square) for measuring and comparing the shape variability in biplots. Finally, Section 5 provides an illustration.

## 2. Principal Component Biplot with Missing Values

### 2.1. Principal component biplot

Consider the $N \times p$ complete data matrix $\mathbf{X} = \{x_{ij}\}$, $i = 1, \ldots, N$; $j = 1, \ldots, p$, with rank $r$. Subtracting out the mean of each variables such as $\overline{x_j} = \sum_{i=1}^{N} x_{ij}/N$, we obtain a new $N \times p$ data matrix $\mathbf{Y} = (x_{ij} - \overline{x_j})$, $i = 1, \ldots, N$; $j = 1, \ldots, p$, which is called variables-centered. Now, we noted that the principal component analysis based on the singular value decomposition of $\mathbf{Y}$ such as

$$\mathbf{Y} = \mathbf{U}\mathbf{D}_\lambda\mathbf{V} = \sum_{t=1}^{r} \lambda_t \mathbf{u}_t \mathbf{v}_t', \tag{2.1}$$

where $\mathbf{U}$ is an $N \times r$ matrix whose columns are the $N \times 1$ eigenvectors $\mathbf{u}_t$, $\mathbf{V}$ is an $p \times r$ matrix whose columns are the $p \times 1$ eigenvectors $\mathbf{v}_t$ of $\mathbf{Y}\mathbf{Y}'$ and $\mathbf{D}_\lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$ with singular values, $\lambda_1 \geq \cdots \geq \lambda_r > 0$. And let

$$\mathbf{U} = \mathbf{A}$$

and let

$$\mathbf{W} = \mathbf{V}\mathbf{D}_\lambda.$$

Then the data matrix $\mathbf{Y}$, which is written as the form of the singular value decomposition (2.1), can be factorized as

$$\mathbf{Y} = \mathbf{A}\mathbf{W}'. \tag{2.2}$$

If $\mathbf{Y}$ can be satisfactorily approximated by a rank $s$ $(1 \leq s \leq r)$ matrix $\mathbf{Y}_{(s)}$, it may allow an useful approximate visual inspection of $\mathbf{Y}$ itself. Specially taking $s = 2$ gives a rank two approximation to (2.2)

$$\mathbf{Y}_{(2)} = \mathbf{A}_{(2)}\mathbf{W}_{(2)}', \tag{2.3}$$

where $\mathbf{A}_{(2)}$ and $\mathbf{W}_{(2)}$ are $N \times 2$ and $p \times 2$ matrices respectively. Gabriel (1971) calls $\mathbf{A}_{(2)}$ and $\mathbf{W}_{(2)}$ the factors of $\mathbf{Y}_{(2)}$ for biplotting and $\mathbf{Y}_{(2)}$ is called principal component biplot.

As in principal component analysis, the goodness-of-fit means a measure of how satisfactorily $\mathbf{Y}_{(2)}$ approximates $\mathbf{Y}$ in written as

$$\text{fit} = \frac{\lambda_1^2 + \lambda_2^2}{\sum_{t=1}^{r} \lambda_t^2}.$$

### 2.2. Missing mechanism

Biplot, described in Section 2.1, is a useful approach in the exploring data analysis. However unfortunately, when some values are missing in data matrix $\mathbf{X}$, most biplots are not directly applicable. Therefore the purpose of this section is to obtain the basis for the analysis of incomplete data of principal component biplot.

For this, we need to understand the missing mechanism. Let $\mathbf{X}$ denote an $N \times p$ data matrix with missing observations and let $\mathbf{C}$ denote an $N \times p$ missing data indicator matrix, such that $c_{ij} = 1$ if $x_{ij}$ is missing and 0 otherwise. A full model for the data and the missing mechanism specifies a distribution $f(\mathbf{X} \mid \theta)$ for $\mathbf{X}$, indexed by unknown parameter $\theta$ and a distribution $f(\mathbf{C} \mid \mathbf{X}, \psi)$ for $\mathbf{C}$, given $\mathbf{X}$ indexed by unknown parameter $\psi$. Write $\mathbf{X} = (\mathbf{X}_{obs}, \mathbf{X}_{mis})$, where $\mathbf{X}_{obs}$ represents the observed values of $\mathbf{X}$ and $\mathbf{X}_{mis}$ represents the missing values. Rubin (1976) defined the missing data as MCAR(missing completely at random) if

$$f(\mathbf{C} \mid \mathbf{X}_{obs}, \mathbf{X}_{mis}, \psi) = f(\mathbf{C} \mid \psi),$$

for all $\mathbf{X}_{obs}$ and $\mathbf{X}_{mis}$; that is, missingness does not depend on the observed or missing values of $\mathbf{X}$. Rubin also defined a weaker condition for the missing mechanism, calling the missing data MAR(missing at random) if

$$f(\mathbf{C} \mid \mathbf{X}_{obs}, \mathbf{X}_{mis}, \psi) = f(\mathbf{C} \mid \mathbf{X}_{obs}, \psi),$$

for all $\mathbf{X}_{obs}$ and $\mathbf{X}_{mis}$; that is, missingness does not depend on the missing values but may depend on observed values in the data set. The MCAR and MAR are the ignorable mechanisms for likelihood-based inference (Little, 1988).

Now, we explain the ideas for decomposing data matrix when it contains missing values under MAR assumption. Let $N \times p$ data matrix $\mathbf{X}$ with missing observations be normally distributed with mean $\mu$ and covariance matrix $\Sigma$, then we have $k$ subsets $\mathbf{R}_i$, $i = 1, \ldots, k$; $1 < k \leq 2^p - 1$. $\mathbf{R}_i$ denotes the $n_i \times p$, $i = 1, \ldots, k$; $\sum_{i=1}^{k} = N$ sub-matrix of $\mathbf{X}$ which has the observations on the same variables only.

If there are $p_i$ characteristics, $j_1, \ldots, j_{p_i}$, say in the $i^{th}$ set of observations, then $\mathbf{B}_i$ would be a $p_i \times p$ matrix a 1 in the $(\alpha, p_\alpha)$ position of observed value of $\mathbf{R}_i$, $\alpha = 1, \ldots, p_i$ and zeros elsewhere. In the same way, $\mathbf{C}_i$ would be a $(p - p_i) \times p$ matrix a 1 in the $(\beta, p_\beta)$ position of the missing value of $\mathbf{R}_i$, $\beta = 1, \ldots, (p - p_i)$ and zeros elsewhere. Then $\mathbf{B}_i$ and $\mathbf{C}_i$ satisfied the properties as follows:

$$\mathbf{I}_p = \mathbf{B}_i'\mathbf{B}_i + \mathbf{C}_i'\mathbf{C}_i, \quad \mathbf{B}_i\mathbf{B}_i' = \mathbf{I}_{p_i}; \ \mathbf{C}_i\mathbf{C}_i' = \mathbf{I}_{(p-p_i)}. \tag{2.4}$$

Next, let $\mathbf{x}_{ij}'$, $i = 1, \ldots, k$; $j = 1, \ldots, n_i$ be a row vector of $\mathbf{R}_i$, then by (2.4)

$$\mathbf{x}_{ij} = \mathbf{B}_i'(\mathbf{B}_i\mathbf{x}_{ij}) + \mathbf{C}_i'(\mathbf{C}_i\mathbf{x}_{ij}) = \mathbf{B}_i'\mathbf{Z}_{ij} + \mathbf{C}_i'\mathbf{M}_{ij}, \quad i = 1, \ldots, k; \ j = 1, \ldots, n_i,$$

where $\mathbf{Z}_{ij}$ denotes the observed data, $\mathbf{M}_{ij}$ denotes the missing data. So, likelihood function of $\mathbf{Z}_{ij}$ is

$$L(\theta) = \left( \prod_{i=1}^{k} |\mathbf{B}_i \Sigma \mathbf{B}_i'|^{-\frac{1}{2} n_i} \right) \exp\left( -\frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\mathbf{Z}_{ij} - \mathbf{B}_i \mu)' (\mathbf{B}_i \Sigma \mathbf{B}_i')^{-1} (\mathbf{Z}_{ij} - \mathbf{B}_i \mu) \right)$$

and likelihood function of the raw data matrix $\mathbf{X}$ is

$$L_c(\theta) = |\Sigma|^{-\frac{1}{2} N} \exp\left( -\frac{1}{2} \sum_{i=1}^{N} (\mathbf{x}_i - \mu)' \Sigma^{-1} (\mathbf{x}_i - \mu) \right).$$

Let $\mathbf{Z}$ be the matrix of all $\mathbf{Z}_{ij}$ and $\mathbf{M}$ be the matrix of all $\mathbf{M}_{ij}$, then log-likelihood function of $\mathbf{X}$ is

$$\log L_c(\theta) = \log L(\theta) + \log P(\mathbf{M} \mid \mathbf{Z}, \theta).$$

## 3. The EM Algorithm and Shape Variability in Biplots with Missing Values

### 3.1. The EM algorithm

Now since the mean imputation is a particular simple form of imputation, in this section, we introduce only the EM algorithm. In McLachlan and Krishnan (1997) and Dempster *et al.* (1977), for unknown parameter $\theta$ on the $(t+1)^{th}$ iteration of EM algorithm, the E-step requires the calculation of

$$Q\left(\theta \mid \theta^{(t)}\right) = E_{\theta^{(t)}}\left(\log L_c(\theta) \mid \mathbf{Z}\right) = \log L(\theta) + H\left(\theta \mid \theta^{(t)}\right),$$

where

$$Q\left(\theta \mid \theta^{(t)}\right) = \int \log L_c(\theta) P\left(\mathbf{M} \mid \mathbf{Z}, \theta^{(t)}\right) d\mathbf{M}$$

and

$$H\left(\theta \mid \theta^{(t)}\right) = \int \log P\Big(\mathbf{M} \mid \mathbf{Z}, \theta\Big) P\left(\mathbf{M} \mid \mathbf{Z}, \theta^{(t)}\right) d\mathbf{M}.$$

Next, the M-step requires the maximization of $Q(\theta \mid \theta^{(t)})$ with respect to $\theta$. That is, we choose $\theta^{(t+1)}$ such that

$$Q\left(\theta^{(t+1)} \mid \theta^{(t)}\right) \geq Q\left(\theta \mid \theta^{(t)}\right),$$

for all $\theta$. The E-step and M-step are alternated repeatedly until the difference

$$L\left(\theta^{(t+1)}\right) - L\left(\theta^{(t)}\right),$$

changes by an arbitrarily small amount. Hence convergence must be obtained with a sequence of likelihood values that are bounded above. In Srivastava (2002), if $\hat{\mu}_0$ and $\hat{\Sigma}_0$ is the initial values for $\mu$ and $\Sigma$, respectively, then we can estimate the missing values $\mathbf{M}$ and data matrix $\mathbf{X}$ as following:

$$\hat{\mathbf{M}}_{ij} = \mathbf{C}_i \hat{\mu}_0 + (\mathbf{C}_i \hat{\Sigma}_0 \mathbf{B}_i')(\mathbf{B}_i \hat{\Sigma}_0 \mathbf{B}_i')^{-1} (\mathbf{Z}_{ij} - \mathbf{B}_i \hat{\mu}_0),$$
$$\hat{\mathbf{x}}_{ij} = \mathbf{B}_i' \mathbf{Z}_{ij} + \mathbf{C}_i' \hat{\mathbf{M}}_{ij}, \quad i = 1, \ldots, k; \; j = 1, \ldots, n_i.$$

### 3.2. Shape variability in biplots with missing values

Choi *et al.* (2005) provided a study on the biplots' variability using the Procrustes analysis. They used the Procrustes statistic for measuring the similarity of tow biplots and their study is also related with the shape variability, widely. In this section, we consider the shape variability in biplots with missing values. As noted in Section 1, even though we estimate missing values of biplot of incomplete data we have different shapes of biplots according to the imputation methods and missing rates. Therefore we should compare the shapes of biplots of complete data with those of incomplete data. For this, we propose a RMS as one of the measurements of shape variability.

Let $\hat{\mathbf{X}}$ be the estimates of $\mathbf{X}$ in Section 3.1, then the factorization of the centroid matrix $\hat{\mathbf{Y}}$ by (2.3) is

$$\hat{\mathbf{Y}}_{(2)} = \hat{\mathbf{A}}_{(2)}\hat{\mathbf{W}}'_{(2)},$$

where $\hat{\mathbf{A}}_{(2)}$ and $\hat{\mathbf{W}}_{(2)}$ are $N \times 2$ matrix and $p \times 2$ matrix, respectively. And let $\hat{\mathbf{E}}$ be a $(N+p) \times 2$ biplot coordinates matrix of $\hat{\mathbf{Y}}$, such that

$$\hat{\mathbf{E}} = \left(\hat{\mathbf{A}}'_{(2)}, \hat{\mathbf{W}}'_{(2)}\right)'.$$

In the same way, if data matrix $\mathbf{X}$ is complete, then we can obtain the biplot coordinates matrix $\mathbf{E}$. Then using the vectorize operator $\mathrm{vec}(\hat{\mathbf{E}})$ of $\hat{\mathbf{E}}$ and $\mathrm{vec}(\mathbf{E})$ of $\mathbf{E}$, we can propose a measure of the shape variability as RMS such that

$$\mathrm{RMS} = \sqrt{\frac{1}{2(N+p)}\left\|\mathrm{vec}(\hat{\mathbf{E}}) - \mathrm{vec}(\mathbf{E})\right\|}.$$

We note that if the value of RMS is nearly 0, the shape variability of two biplots is very small. In fact, Dryden and Mardia (1998) provided a measure of RMS' type of full Procrustes distance from each configurations to the full Procrustes mean in Procrustes analysis.

## 4. Illustration

We will illustrate principal component biplot by using the census tract data from Johnson and Wichern (1998, Table 8.5). In fact, the data set consists of 5 variables measured. Observations from adjacent census tracts are likely to be correlated. That is, 14 observations may not constitute a random sample. So we are interest in optimal 2-dimensional biplot. Thus the traditional principal component biplot is given in Figure 4.1. In biplot, we use the 3 alphabet initials and numbers for the variable markers and observation markers, respectively.

Note that the variable marker MVH(median value home), which is at left of Figure 4.1, is extraneous for the other variables. And an interior angles of TOP(total population), TOE(total employment), HSE(health services employment) are narrow, so these variables have high correlation. 6, 8, 9 tracks have a higher score for these variables. Whereas 2, 3, 7, 10, 13, *etc.*, which have lower score for these variables. In this case, the goodness-of-fit of approximation is 86.41%.

Now, we deleted three observations under MAR assumption(missing rate: 4.29%) by art and estimate missing values using two methods, EM algorithm and mean imputation. The principal component biplots of each method are given in Figure 4.2. There is little variability of variable markers on both cases. But the case using mean imputation has a little variability of observation

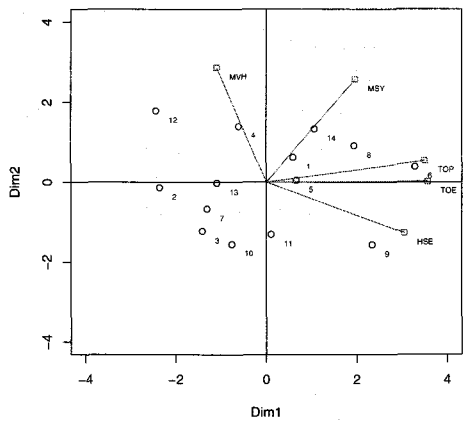


**Figure 4.1.** Principal component biplot of census track data

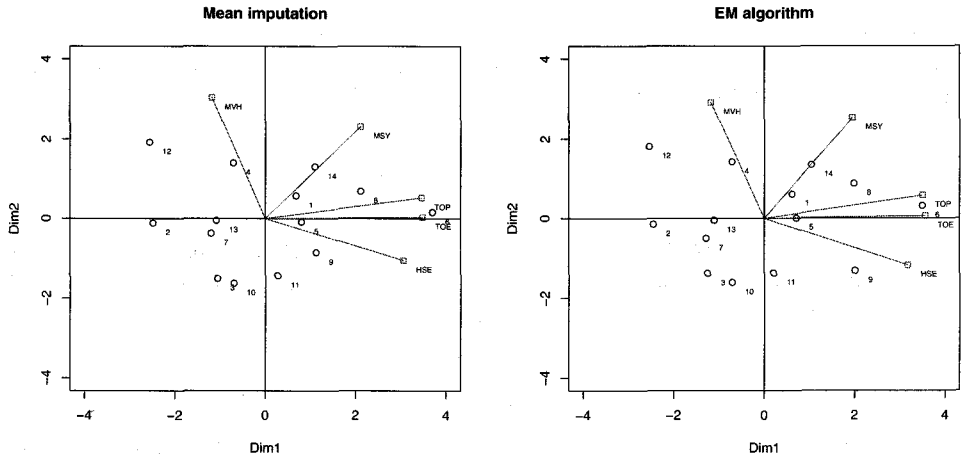


**Figure 4.2.** Principal component biplot of missing rate 4.29%

markers, which move toward the origin. In these cases, the goodness-of-fits of approximation are 84.97% and 87.26%, respectively.

When we increase the missing rate at 10.00%, biplot using mean imputation has some variability comparing biplot of raw data. Specially, the degree of moving of observation markers toward origin is increased and an interior angle of TOP and TOE is larger than an angle of raw data. Thus the correlation of these variables is lower than that of raw data. In this case, the goodness-of-fit of approximation is getting lower to 80.46%.

On the other hand, in the case using EM algorithm an interior angle of TOP and TOE is smaller than that of raw data. And so the correlation of these variables is higher than that of raw data. And observation marker is located similar to raw data. Specially, we take note of that the goodness-of-fit of approximation is getting higher to 87.25%.

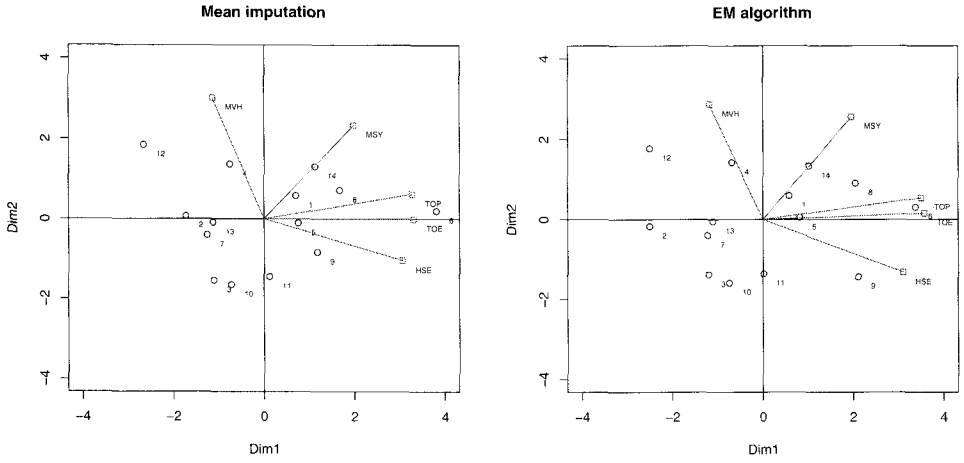


**Figure 4.3.** Principal component biplot of missing rate 10.00%

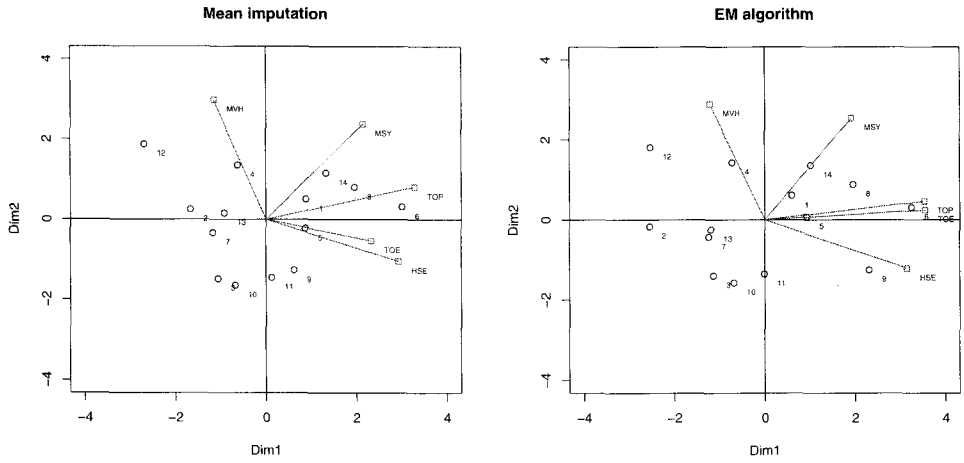


**Figure 4.4.** Principal component biplot of missing rate 14.29%

Finally, we increase the missing rate at 14.29%, then biplot using mean imputation has large shape variability. TOE and HSE have high correlation, but that the correlation of TOE and TOP is lower than that of raw data. However in the case using EM algorithm, an interior angle of TOP and TOE is smaller than that of raw data. And so the correlation of these variables is higher than that of raw data.

Specially, track 9, 13 is moved over in the case of mean imputation. But it is located similar to raw data, in the case of EM algorithm. In these cases, the goodness-of-fits of approximation are 81.07% and 85.19%, respectively.

In conclusion, the goodness-of-fit of approximation of biplot using mean imputation is partially decreasing according to the increment of the missing rate. However in the case using EM algorithm, the goodness-of-fit of approximation is similar to that of raw data. Moreover, we measure the shape

**Table 4.1.** Comparison of Goodness-of-fit

| Imputation method | missing rate | | | |
|---|---|---|---|---|
| | 0.00% | 4.29% | 10.00% | 14.29% |
| EM algorithm | 86.41% | 87.26% | 87.25% | 85.19% |
| Mean imputation | 86.41% | 84.87% | 80.46% | 81.07% |

**Table 4.2.** Comparison of RMS

| Imputation method | missing rate | | |
|---|---|---|---|
| | 4.29% | 10.00% | 14.29% |
| EM algorithm | 0.007 | 0.005 | 0.007 |
| Mean imputation | 0.051 | 0.064 | 0.146 |

variability as RMS given Table 4.2, then in the case using EM algorithm, there is little variability. But we have some relatively variabilities in the case using mean imputation.

## References


Choi, Y. S., Hyun, G. H. and Yun, W. J. (2005). Biplots' variability based on the Procrustes analysis, *Journal of the Korean Data Analysis Society*, **7**, 1925–1933.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Dryden, I. L. and Mardia, K. V. (1998). *Statistical Shape Analysis*, John Wiley & Sons, Chichester.

Gabriel, K. R. (1971). The biplot graphics display of matrices with application to principal component analysis, *Biometrika*, **58**, 453–467.

Johnson, R. A. and Wichern, D. W. (1998). *Applied Multivariate Statistical Analysis*, Prentice-Hall, New York.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values, *American Statistical Association*, **83**, 1198–1202.

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

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*, Wiley, New York.

Rubin, D. B. (1976). Inference and missing data, *Biometrika*, **63**, 581–592.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, Chapman & Hall/CRC, London.

Srivastava, M. S. (2002). *Methods of Multivariate Statics*, Wiley, New York.