Imputing Missing Values of Environmental Multi-Dimensional Vectors Using a Modified Roweis Algorithm

Natalia Nikolova*
Daniela Toneva-Zheynova**
Danko Naydenov***
Kiril Tenekedjiev****

***Nikola Vaptsarov Naval Academy, Varna 9026, V. Drumev 73 Str. Bulgaria**
E-mail: natalianik at gmail dot com
****Technical University – Varna, Varna 9010 Bulgaria**
E-mail:d_toneva at abv dot bg
***** Technical University – Varna, Varna 9010, Bulgaria**
E-mail: sky at eurorisksystems dot com
******Nikola Vaptsarov Naval Academy, Varna 9026 Bulgaria**
E-mail: Kiril dot Tenekedjiev at fulbrightmail dot org

***Abstract*:** We present an algorithm for *m*-dimensional visualization of *n* vectors which are *p*-dimensional (*p*>*m*) with missing values. It is based on the Principal Component solution of Factor analysis model. Roweis first offer a missing value solution in that context based on the Expectation minimization (EM) algorithm. Here we propose modified version of the Roweis algorithm which works better in small and medium counts of *n* and *p*. Once the missing values are imputed the visualization into 2- or 3-dimensional space can be done by any method e.g. classical orthogonal Factor Analysis or Multi-Dimensional Scaling.

Keywords: Factor analysis, Principle Component Solution, EM maximization, visualization, multi- dimensional scaling

1. INTRODUCTION

Every record in an environmental database can be represented by a *p*-dimensional column vector $\vec{x}_j = \left(x_1^{(j)}, x_2^{(j)}, \ldots, x_p^{(j)}\right)^T$, where T stands for the transpose operator. Since p is usually more than 20, then $\vec{x}_i$ cannot be visualized. It is easier to perform visualization in the m=2 or m=3 dimensional space. This paper presents two algorithms for the data reduction from the p-dimensional to the m-dimensional space. These are presented in the second section of this work.

In some cases, the records are not full vectors, and part of their coordinates are missing due to various reasons. One solution is to disregard any $\vec{x}_j$ which contain incomplete coordinates. Another option is to use whatever available for the calculation of any statistics required. Other alternatives for data imputation are the mean substitution, mean substitution for subgroups, indicator/dummy variable adjustment. All those suffer from numerous serious drawbacks, which is why they are not recommended.

A suitable solution for estimation of unobservable quantities is offered by a powerful expectation minimization algorithm from (Dempster, Laird, Rumin, 1977), proven in (Wu, 1983). Roweis used this algorithm for imputation of missing data in the context of factor analysis program (Roweis, 1997). In this paper, we present a modification of the Roweis algorithm for medium sized data points which quickly converge. The third section of this paper is devoted to the representation of this modified algorithm. All discussions in this paper focus on the case of environmental data analysis.

2. VISUALIZATION OF MULTI-DIMENSIONAL DATA

2.1 Multi-dimensional scaling

It is required to represent the p-dimensional data into a lower m-dimensional space. As a result, $\vec{y}_1, \vec{y}_2, \ldots, \vec{y}_n$ become the images of $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n$ in the m-dimensional space. If m=2 or m=3, it is possible to visualize data and find outliers in the dataset. If m=4 or m=5, then it is still possible to directly analyze the information by plotting each two coordinates, but that is not possible in the p-dimensional space, because each two coordinates are not enough informative. The so stated problem may be approached by multi-dimensional scaling (Duda, et al., 2001). The main condition is that if two vectors are similar/distant in the p-dimensional space, then their representation in the m-dimensional space should also be similar/distant. That means that measures of similarity are required. Metrics can serve as suitable measures of similarity.

Let $\delta_{k,j} = \delta\left(\vec{x}_k, \vec{x}_j\right)$ be a real-valued numerical function defined over two p-dimensional vectors. Each metrics must

have the following properties: a) Nonnegativity: $\delta_{k,j} \geq 0$; b) Reflexivity: $\delta_{k,j} = 0$ if and only if $\vec{x}_j = \vec{x}_k$; c) Symmetry: $\delta_{k,j} = \delta_{j,k}$; d) Triangle inequality: $\delta_{k,r} + \delta_{r,j} \geq \delta_{k,j}$.

A next step in the analysis is to find a proper form for the metrics. One typical metrics is the Euclidean distance:

$$\delta_{k,j} = \sqrt{\sum_{i=1}^{p} \left(x_i^{(k)} - x_i^{(j)} \right)^2} \tag{1}$$

Another general class of metrics is the Minkovski *r*-distance:

$$\delta_{k,j} = \sum_{i=1}^{p} \left(\left| x_i^{(k)} - x_i^{(j)} \right|^r \right)^{1/r} \tag{2}$$

Here $r \geq 1$.The greater the value of r, the more the coordinate with the greatest discrepancy between $\vec{x}_j$ and $\vec{x}_k$ influences $\delta_{k,j}$. A special case is the Manhattan or city block distance, where r=1:

$$\delta_{k,j} = \sum_{i=1}^{p} \left| x_i^{(k)} - x_i^{(j)} \right| \tag{3}$$

If r is infinite, then the Hamilton (or Chebishev) distance is defined as

$$\delta_{k,j} = \max_i \left(\left| x_i^{(k)} - x_i^{(j)} \right| \right) . \tag{4}$$

There are other groups of measures of similarity, which depend on the data set analyzed. In that case, numerical characteristics need to be calculated for the coordinates, like mean, standard deviation, as well as covariance and correlation coefficients. The vector of sample mean values $\bar{\vec{x}}$ and the sample covariance matrix K are calculated as (18) and (19). The elements of the covariance are denoted $k_{i,r}$.Then the standard deviation of the i-th coordinate is calculated as $s_i = \sqrt{k_{i,i}}$.

If all these characteristics of data are calculated then normalized version of the Minkowski r-distance can be defined:

$$\delta_{k,j} = \sum_{i=1}^{p} \left(\left| \frac{x_i^{(k)} - x_i^{(j)}}{s_i} \right|^r \right)^{1/r} \tag{5}$$

Because of the resulting relative values, such normalized metrics eliminate the effect of large values. Another measure of distance, suitable for multi-dimensional normally distributed data is the Mahalonobis distance:

$$\delta_{k,j} = \left(\vec{x}_k - \vec{x}_j \right)^T K^{-1} \left(\vec{x}_k - \vec{x}_j \right) \tag{6}$$

In the m-dimensional space, the distance between $\vec{y}_j$ and $\vec{y}_k$ is denoted d_{kj}. As long as the main objective of MDS is to visualize the data in the target space, d_{kj} must be a Euclidean distance, because humans accept it as a dissimilarity measure. Finding the proper representation in the m-dimensional space depends on how close d_{kj} is to the original distance measure $\delta_{k,j}$. Each of these distances is only $n(n–1)/2$ in number, because the distance of a vector from itself is zero, and the distance between the j-th and k-th vectors coincides with the distance between the k-th and the j-th vector.

Next is necessary to define a criterion to decide whether or not one configuration is better than another. Possible criterion functions are

$$J_{ee} = \frac{\sum_{k<j} \left(d_{k,j} - \delta_{k,j} \right)^2}{\sum_{k<j} \delta_{k,j}^2} \text{ and } J_{ff} = \sum_{k<j} \left(\frac{d_{k,j} - \delta_{k,j}}{\delta_{k,j}} \right)^2 \tag{7}$$

But the most used criterion is named "stress":

$$J_{ef} = \frac{1}{\sum_{k<j} \delta_{k,j}} \sum_{k<j} \frac{\left(d_{k,j} - \delta_{k,j} \right)^2}{\delta_{k,j}} . \tag{8}$$

Whereas J_{ee} emphasizes large errors (regardless of whether the distance $\delta_{k,j}$ are large or small), J_{ff} emphasizes the large fractional errors (regardless of whether the errors $\left| d_{k,j} - \delta_{k,j} \right|$ are large or small). A useful compromise is J_{ef}, which emphasizes large products of error and fractional error. The optimal set of images in the target space may be found by minimizing the selected criterion function.

In (7) and (8), d_{kj} must be Euclidean distances, because that facilitates the visualization. The choice of the initial configuration of images in the target space is important for the multi-dimensional gradient optimization. It is recommended to initiate the optimization by selecting the m coordinates that have the highest variance, since they will somewhat represent the initial data set in the target space.

2.2 Exploratory factor analysis

Let the p-dimensional observation random vector $\vec{X} = \left(X_1, X_2, \ldots, X_p \right)^T$ contains n realizations $\left(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n \right)$ written in a p x n matrix. Let the mean vector and the covariance matrix of $\vec{X}$ be $\vec{\mu}$ and S respectively:

$$\vec{\mu} = \left(\mu_1, \mu_2, \ldots, \mu_p \right)^T \tag{9}$$

$$S = \begin{pmatrix} \sigma_1^2 & r_{1,2}\sigma_1\sigma_2 & \vdots & r_{1,n}\sigma_1\sigma_p \\ r_{1,2}\sigma_1\sigma_2 & \sigma_2^2 & \vdots & r_{2,n}\sigma_2\sigma_p \\ \ldots & \ldots & \vdots & \ldots \\ r_{1,n}\sigma_1\sigma_p & r_{2,n}\sigma_2\sigma_p & \vdots & \sigma_p^2 \end{pmatrix} \tag{10}$$

where μ_i and σ_i are the mean value and the standard deviation of the i-th coordinate of X, whereas $r_{i,k}$ is the correlation coefficient between the coordinates with numbers i and k.

The factor model assumes that

$$\vec{x}_j = \vec{\mu} + L\vec{f}_j + \vec{\varepsilon}_j \text{, за } j=1, 2, \ldots, n \tag{11}$$

Here, L is a p x m matrix of factor loadings $l_{i,k}$ and $\vec{f}_i$ is an unobservable m-dimensional realization of the random variable $\vec{F}(F_1, F_2, \ldots, F_n)$ called a common factor. The vector $\vec{f}_j$ can be expanded:

$$\vec{f}_j = \left(f_1^{(j)}, f_2^{(j)}, \ldots, f_m^{(j)}\right)^T \text{ за } j=1, 2, \ldots, n \quad (12)$$

The common factors partly explain the variation of X, but not the variation of errors. The unexplained part of the variation of the coordinate X_i is modeled by the unobservable random variable E_j, which is called error or specific factor of X_i. The errors are organized in a p-dimensional random vector $\vec{\Xi} = \left(\Xi_1, \Xi_2, \ldots, \Xi_p\right)^T$, which has n realizations $\vec{\varepsilon}_j = \left(\varepsilon_1^{(j)}, \varepsilon_2^{(j)}, \ldots, \varepsilon_p^{(j)}\right)^T$.

There are many different factor models. The assumptions of the widespread orthogonal factor model (Johnson R., Wicherin D.,2007) are:

1) The expected vector of the common factors is an m-dimensional zero vector :

$$E\left(\vec{F}\right) = \vec{0}_{m\times 1};$$

2) The covariance matrix of the common factors is an m-dimensional identity matrix:

$$cov\left(\vec{F}\right) = E\left(\left(\vec{F} - E\left(\vec{F}\right)\right)\left(\vec{F} - E\left(\vec{F}\right)\right)^T\right) =$$

$$= E\left(\vec{F} \times \vec{F}^T\right) = I_{n\times m} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix} \quad (13)$$

3) The expected error vector is the p-dimensional zero vector:

$$E\left(\vec{\Xi}\right) = \vec{0}_{p\times 1};$$

4) The covariance matrix of the error vector is diagonal pxp matrix

$$cov\left(\vec{\Xi}\right) = E\left(\left(\vec{\Xi} - E\left(\vec{\Xi}\right)\right)\left(\vec{\Xi} - E\left(\vec{\Xi}\right)\right)^T\right) =$$

$$= E\left(\vec{\Xi} . \vec{\Xi}^T\right) = \Psi_{p\times p} = \begin{pmatrix} \psi_1^2 & 0 & \ldots & 0 \\ 0 & \psi_2^2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \psi_p^2 \end{pmatrix} \quad (14)$$

where ψ_i^2 is called the specific variance of X_i

5) The covariance of each error term ε_j and each factor F_k is zero :

$$cov\left(\varepsilon_j F_k\right) =$$

$$= E\left(\left(\varepsilon_j - E\left(\varepsilon_j\right)\right)\left(F_k - E\left(F_k\right)\right)^T\right) = E\left(\varepsilon_j F_k\right) = 0 ,$$

for j=1, 2, …, n and k=1, 2, …, m (15)

The orthogonal factor model implies certain structure of the covariance matrix of $\vec{X}$:

$$cov\left(\vec{X}\right) = E\left(\left(\vec{X} - E\left(\vec{X}\right)\right)\left(\vec{X} - E\left(\vec{X}\right)\right)^T\right) = LL^T + \Psi \quad (16)$$

In particular the diagonal

$$var\left(X_i\right) = E\left(\left(X_i - E\left(X_i\right)\right)^2\right) = \sigma_i^2 \Rightarrow r_i^2 = h_i^2 + \psi_i^2 \quad (17)$$

where

$$h_i^2 = l_{i1}^2 + l_{i2}^2 + \ldots + l_{im}^2 \quad (18)$$

is called the communality of X_i.

The covariance between the original variables and the compressed one in the factor form is the loading factor variable:

$$cov\left(\vec{X}, \vec{F}\right) = E\left(\left(\vec{X} - E\left(\vec{X}\right)\right)\left(\vec{F} - E\left(\vec{F}\right)\right)^T\right) =$$

$$= E\left(\left(\vec{X} - E\left(\vec{X}\right)\right)\vec{F}^T\right) = L \quad (19)$$

As it is already clear, the vectors $\vec{f}_j$ (j=1, 2, …, n) are compressed versions of $\vec{x}_j$ and can be visualized easier.

The factor analysis model can be identified by several different methods. Because of the Java restrictions we chose the Principle Component Solution.

Let $\bar{\vec{x}}$ and K be the sample mean vector and the sample covariance matrix of the data $\left(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\right)$:

$$\bar{\vec{x}} = \frac{1}{n}\sum_{j=1}^{n} \vec{x}_j \quad (20)$$

$$K = \frac{1}{n-1}\sum_{j=1}^{n}\left(\vec{x}_j - \bar{\vec{x}}\right)\left(\vec{x}_j - \bar{\vec{x}}\right)^T \quad (21)$$

As long as K is symmetric and positively semi-definite, then there exist p real eigen values λ_i and p real eigen vectors $\vec{v}_i$ of K so that

$$K\vec{v}_i = \lambda_i \vec{v}_i, \text{ for } i=1, 2, \ldots, p \quad (22)$$

Here, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. Then the loading matrix can be formed column wise by the greatest "m" eigen values and eigen vectors as shown

$$L = \left[\sqrt{\lambda_1}\vec{v}_1, \sqrt{\lambda_2}\vec{v}_2, \ldots, \sqrt{\lambda_m}\vec{v}_m\right] \quad (23)$$

Then the specific variance of the i-coordinate can be calculated as

$$\Psi_i^2 = \sigma_i^2 - l_{i,1}^2 - l_{i,2}^2 - \ldots - l_{i,m}^2 \quad (24)$$

The goodness-of-fit criteria for the factor analysis model can be calculated as absolute ΔK_{abs}^{err} and relative error ΔK_{rel}^{err} matrices

$$\Delta K_{abs}^{err} = K - LL^T - \Psi \tag{25}$$

This matrix is sometimes called residual matrix

$$\Delta K_{rel}^{err} = \frac{\Delta K_{rel}^{err}}{K}, \tag{26}$$

where the division is realized element-wise for all $k_{i,r} \neq 0$.

The last goodness-of-fit measure is the degree of explained variance:

$$PVE = 100\left(\lambda_1 + \lambda_2 + \ldots + \lambda_m\right) \Big/ \left(\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_p^2\right), \% \tag{27}$$

The factor scores $\vec{f}_j$ can be estimated using either ordinary least square approach

$$\vec{f}_j^{olsq} = \left(L^T L\right)^{-1} L^T \left(\vec{x}_j - \bar{\vec{x}}\right) \text{ for } j=1, 2, \ldots, n \tag{28}$$

or using the weighted least square approach:

$$\vec{f}_j^{wlsq} = \left(L^T \Psi^{-1} L\right)^{-1} L^T \Psi^{-1} \left(\vec{x}_j - \bar{\vec{x}}\right) \text{ for } j=1, 2, \ldots, n \tag{29}$$

The last formula is useful when the variances of the coordinates X_i are substantially different from each other.

The whole algorithm of the data compression can be performed on data, which is normalized in the first place:

$$\vec{z}_j = \left(z_1^{(j)}, z_2^{(j)}, \ldots, z_p^{(j)}\right) \tag{30}$$

where

$$z_i^{(j)} = \frac{x_i^{(j)} - \bar{x}_i}{s_i}, \tag{31}$$

where $\bar{x}_i$ is the i-th element of $\bar{\vec{x}}$, and s_i is the i-th diagonal element of the matrix K.

3. IMPUTING MISSING VALUES

3.1 Approaches for imputation of missing values

Sometimes the record $\vec{x}_j$ of the data matrix $D_{miss} = \left(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\right)$ is not a full vector. Instead, part of the coordinates $x_1^{(j)}, x_2^{(j)}, \ldots, x_p^{(j)}$ are missing due to various reasons. There are two obvious solutions. The first one is to disregard any $\vec{x}_j$ which contain incomplete coordinates. This method is called listwise or case deletion (Beale, Little, 1975). Its main drawback is that the data quantity diminishes, and the results can be biased if there is a systematic reason not to report part of the information.

The second obvious solution is to use whatever available for the calculation of any statistics required. The method is called pairwise deletion (Acock, 2005). Its main drawback is that different formulae is used to calculate different parts of the result, and although the parts are calculated based on all the information available, the result could be impossible and contradictory. A classic example is the covariance matrix calculated by pairwise method, which has at least one negative eigen value that is impossible for any real sample of data.

There are other traditional methods for missing values imputation as mean substitution (Acock, 1989), mean substitution for subgroups (Acock, Demo, 1994), indicator/dummy variable adjustment (Cohen et al., 2003). Unfortunately, those suffer from serious and numerous drawbacks as falsely increasing the power of the statistical conclusions.

There are many reasons for the missing data especially when social surveys are involved. As long as the environmental data is gathered by direct measurements we have accepted the hypothesis that the data is missing at random (MAR) (Little, 1988).

A powerful expectation minimization algorithm for estimation of unobservable quantities was proposed in the seminal paper (Dempster, Laird, Rumin, 1977) and proven in (Wu, 1983). In (Roweis, 1997), the algorithm was used for imputation of missing data in the context of factor analysis program. Here we present a modification of the Roweis algorithm for medium sized data points which quickly converge.

3.2 Setup of the modified Roweis algorithm

Given is $D_{miss} = \left(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\right)$ with p-dimensional vectors, where some data points are missing. Let's assume that part of the coordinates of $\vec{x}_j$ are missing. The missing coordinate indices in $\vec{x}_j$ can be organized in a set M_j with cardinality N_j:

$$M_j = \left\{m_1^{(j)}, m_2^{(j)}, \ldots, m_{N_j}^{(j)}\right\}, \text{ for } j=1, 2, \ldots, n \tag{32}$$

Some of the sets in M_j can be empty (N_j=0).

Let $\bar{M}_j$ be the complement of M_j with respect to the set $R=\{1, 2, \ldots, p\}$ of the indices in $\vec{x}_i$:

$$\bar{M}_j = R \setminus M_j = \left\{\bar{m}_1^{(j)}, \bar{m}_2^{(j)}, \ldots, \bar{m}_{p-N_j}^{(j)}\right\}, \text{ for } j=1, 2, \ldots, n.$$

Using M_j and $\bar{M}_j$ (called the sets of value presence of D_{miss}) it is easy to decompose $\vec{x}_j$ into missing part $\vec{z}_j$ and known part $\vec{y}_j$:

$$\vec{z}_j = \begin{pmatrix} \vec{z}_1^{(j)} \\ \vec{z}_2^{(j)} \\ \vdots \\ \vec{z}_{N_j}^{(j)} \end{pmatrix} = \begin{pmatrix} \vec{x}_{m_1^{(j)}}^{(j)} \\ \vec{x}_{m_2^{(j)}}^{(j)} \\ \vdots \\ \vec{x}_{m_{N_j}^{(j)}}^{(j)} \end{pmatrix}, \text{ for } j=1, 2, \ldots, n \tag{33}$$

$$\vec{y}_j = \begin{pmatrix} y_1^{(j)} \\ y_2^{(j)} \\ \vdots \\ y_{p-N_j^{(j)}}^{(j)} \end{pmatrix} = \begin{pmatrix} x_{\overline{m}_1^{(j)}}^{(j)} \\ \vec{x}_{\overline{m}_2^{(j)}}^{(j)} \\ \vdots \\ \vec{x}_{\overline{m}_{p-N_j}^{(j)}}^{(j)} \end{pmatrix}, \text{ for } j=1, 2, \ldots, n \tag{34}$$

EXAMPLE:

Let p=5 and for the third observation $\vec{x}_3$, the second and the fifth coordinates are missing (denoted with the NaN notation):

$$\vec{x}_3 = \begin{pmatrix} x_1^{(3)} \\ x_2^{(3)} \\ x_3^{(3)} \\ x_4^{(3)} \\ x_5^{(3)} \end{pmatrix} = \begin{pmatrix} -17 \\ NaN \\ 24 \\ 6 \\ NaN \end{pmatrix} \tag{35}$$

Then, N_3=2, $M_3 = \left\{m_1^{(3)}, m_2^{(3)}\right\} = \{2,5\}$, $p-N_3$=5–2=3, $\overline{M}_i = \left\{\overline{m}_1^{(3)}, \overline{m}_2^{(3)}, \overline{m}_3^{(3)}\right\} = \{1,3,4\}$, $\vec{z}_3 = \begin{pmatrix} x_2^{(3)} \\ x_5^{(3)} \end{pmatrix}$, and

$$\vec{y}_3 = \begin{pmatrix} x_1^{(3)} \\ x_3^{(3)} \\ x_4^{(3)} \end{pmatrix} = \begin{pmatrix} -17 \\ 24 \\ 6 \end{pmatrix}.$$

3.3 Building Explanatory Factor Model with Missing Values.

The explanatory factor model was derived according to (11) as:

$$\vec{x}_j = \vec{\mu} + L\vec{f}_j + \vec{\varepsilon}_j, \text{ for } j=1, 2, \ldots, n \tag{36}$$

The parameters that has to be evaluated are $\vec{\mu}$, L and Ψ whereas the unobservables are $\vec{f}_i$, for i=1, 2, …, n and the missing data set $S_{z,miss} = \{\vec{z}_1, \vec{z}_2, \ldots, \vec{z}_n\}$. $S_{z,miss}$ does not form a matrix because the dimensions of $\vec{z}_j$ are different. The factor values can be organized in a compressed data matrix:

$$D_F = \left(\vec{f}_1, \vec{f}_2, \ldots, \vec{f}_n\right) \tag{37}$$

Following the basic ideas of the EM-algorithm we are looking for to minimize the criterion formed as expectation value of the squared Euclidean norm of the errors:

$$\begin{aligned} &Q\left(\vec{\mu}, L, S_{z,miss}, D_F\right) = \\ &= E\left(\left\|\vec{\Xi}\right\|^2\right) = \sum_{j=1}^{n}\left\|\varepsilon_j\right\|^2 = \frac{1}{n}\sum_{j=1}^{n}\left\|\vec{x}_j^{(imp)} - \vec{\mu} - L\vec{f}_j\right\|^2 = \\ &= \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{p}\left(x_i^{(j)(imp)} - \mu_i - \sum_{k=1}^{m} l_{i,k} f_k^{(j)}\right)^2 \end{aligned} \tag{38}$$

The minimizing of Q and the estimation of the unobservables is done by the following numerical procedure:

Modified Roweis algorithm:

1) Find the initial values for the parameters $\vec{\mu}$ and S as pairwise estimates of the sample mean $\overline{\vec{x}}$ and K:

$$\overline{x}_i = \sum_{\substack{j=1 \\ i\in\overline{M}_j}}^{n} x_i^{(j)} \Big/ \sum_{\substack{j=1 \\ i\in\overline{M}_j}}^{n} 1, \text{ for } i=1,2,\ldots,p$$

$$\overline{\vec{x}} = \left(\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_p\right)^T$$

$$k_{i,r} = \sum_{\substack{j=1 \\ i\in\overline{M}_j \\ r\in\overline{M}_j}}^{n} \left(x_i^{(j)} - \overline{x}_i\right)\left(x_r^{(j)} - \overline{x}_r\right) \Big/ \left(\sum_{\substack{j=1 \\ i\in\overline{M}_j \\ r\in\overline{M}_j}}^{n} 1 - 1\right), \text{ for } i=1,2,\ldots,p$$

and $r=1,2,\ldots,p$

$K=\{k_{j,r}\}$, for i=1,2,…,p and r=1,2,…, p

2) Impute the missing values using mean substitution:

$$x_i^{(j)(imp)} = \begin{cases} \overline{x}_i^{(j)} & \text{за } i \in M_j \\ x_i^{(j)} & \text{за } i \in \overline{M}_j \end{cases}, \text{ for } j=1, 2,\ldots, n \text{ and } i=1,2,\ldots,p$$

Then $\vec{x}_j^{(imp)} = \left(x_1^{(j)(imp)}, x_2^{(j)(imp)}, \ldots, x_p^{(j)(imp)}\right)^T$

Form the imputed current data matrix:

$$D_{imp} = \left(\vec{x}_1^{(imp)}, \vec{x}_2^{(imp)}, \ldots, \vec{x}_n^{(imp)}\right)$$

3) Normalize the imputed data into current normalized imputed matrix:

$$D_{nimp} = \left(\vec{x}_1^{(nimp)}, \vec{x}_2^{(nimp)}, \ldots, \vec{x}_n^{(nimp)}\right)$$

where

$$x_i^{(j)(nimp)} = \frac{\left(x_i^{(j)(imp)} - \overline{x}_i\right)}{\sqrt{k_{i,i}}}, \text{ for } i=1,2,\ldots,p \text{ and } j=1,2,\ldots,n$$

Then $\vec{x}_j^{(nimp)} = \left(x_1^{(j)(nimp)}, x_2^{(j)(nimp)}, \ldots, x_p^{(j)(nimp)}\right)^T$

4) Form a normalized zero data matrix the same as D_{nimp}, where the missing positions in the original data matrix D_{miss} are substituted with zeros:

$$D_{0n} = \left(\vec{x}_1^{(0n)}, \vec{x}_2^{(0n)}, \ldots, \vec{x}_n^{(0n)}\right)$$

where

$$x_i^{(j)(0n)} = \begin{cases} 0 & \text{for } i \in M_j \\ x_i^{(j)(nimp)} & \text{for } i \in \overline{M}_j \end{cases}, \text{ for } j=1,\ 2,\ldots,\ n \text{ and } i=1,2,\ldots,p$$

Then $\vec{x}_j^{(0n)} = \left(x_1^{(j)(0n)}, x_2^{(j)(0n)}, \ldots, x_p^{(j)(0n)}\right)^T$

5) Calculate the maximum likelihood estimates of the sample mean value vector and the sample covariance matrix of D_{nimp}

$$\overline{\vec{x}}^{(nimp)} = \sum_{j=1}^{n} \vec{x}_j^{(nimp)} / n$$

$$K^{(nimp)} = \sum_{j=1}^{n} \left(\vec{x}_j^{(nimp)} - \overline{\vec{x}}^{(nimp)}\right)\left(\vec{x}_j^{(nimp)} - \overline{\vec{x}}^{(nimp)}\right)^T / (n-1)$$

6) Find the first $m(<p)$ eigen values λ_r ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$)and the their corresponding eigen vectors $\vec{v}_r$ for $K^{(nimp)}$. Form the matrices $L = \{l_{i,r}\}$ and $L^{inv} = \{l_{r,i}^{inv}\}$ for $i=1,2,\ldots,p$ and $r=1,2,\ldots,m$:

$$L = \left(\sqrt{\lambda_1}\vec{v}_1, \sqrt{\lambda_2}\vec{v}_2, \ldots, \sqrt{\lambda_m}\vec{v}_m\right)$$

and

$$L^{inv} = \begin{pmatrix} \vec{v}_1^T / \sqrt{\lambda_1} \\ \vec{v}_2^T / \sqrt{\lambda_2} \\ \ldots \\ \vec{v}_m^T / \sqrt{\lambda_m} \end{pmatrix}$$

7) Update the missing values in D_{nimp} by minimization of $\left\|L\vec{f}_j - \vec{x}_j^{(nimp)}\right\|$ according to the missing values $\vec{z}_j^{(nimp)}$. Here $\vec{f}_j$ can be substituted with $L^{inv}\vec{x}_j^{(nimp)}$. So the criterion to be minimized becomes $\left\|LL^{inv}\vec{x}_j^{(nimp)} - \vec{x}_j^{(nimp)}\right\|$. For each element with non-empty M_j the problem can be restated in an ordinary least square minimization towards $\vec{z}_j^{(nimp)} = \left(z_1^{(j)(nimp)}, z_2^{(j)(nimp)}, \ldots, z_{N_j}^{(j)(nimp)}\right)^T$:

$$T_j\vec{z}_j^{(nimp)} = \vec{t}_j$$

The $r^{\text{-th}}$ column of the matrix $T_j^* = \left\{t_{i,k}^{*(j)}\right\}$ for $i=1,2,\ldots,p$ and $k=1,2,\ldots,N_j$ is the column number $m_r^{(j)}$ of the matrix LL^{inv}. The matrix $T_j = \left\{t_{i,k}^{(j)}\right\}$ for $i=1,2,\ldots,p$ and $k=1,2,\ldots,N_j$ is derived from T_j^* by the formula:

$$t_{i,k}^{(j)} = \begin{cases} t_{i,k}^{*(j)} & \text{за } i \in \overline{M}_j \\ t_{i,k}^{*(j)} - 1 & \text{за } i \in M_j \end{cases}, \text{ for } i=1,2,\ldots,p \text{ and } k=1,2,\ldots,N_i$$

The p-dimensional vector $\vec{t}_j$ is equal to:

$$\vec{t}_j = \vec{x}_j^{(0n)} - LL^{inv}\vec{x}_j^{(0n)}$$

The derived system with p equations and N_j unknowns can be solved and $\vec{z}_j^{(nimp)}$ identified.

Update D_{nimp} by:

$$x_{m_i^{(j)}}^{(j)(nimp)} = \begin{cases} x_{m_i^{(j)}}^{(j)(nimp)} & \text{за } i \in M_j \\ z_i^{(j)(nimp)} & \text{за } i \in \overline{M}_j \end{cases}, \text{ for } j=1,\ 2,\ldots,\ n\ ; i=1,2,\ldots,N_i \text{ and } M_i \neq \varnothing$$

8) Restore D_{inp} using:

$$x_i^{(j)(imp)} = x_i^{(j)(nimp)}\sqrt{k_{i,i}} + \mu_i \text{ for } j=1,\ 2,\ldots,\ n \text{ and } i=1,2,\ldots,p$$

9) Update $\overline{\vec{x}}^{(imp)}$ and $K^{(imp)}$ using the maximum likelihood estimates of data in D_{imp}

$$\overline{\vec{x}}^{(imp)} = \sum_{j=1}^{n} \vec{x}_j^{(imp)} / n$$

$$K^{(imp)} = \sum_{j=1}^{n} \left(\vec{x}_j^{(imp)} - \overline{\vec{x}}^{(imp)}\right)\left(\vec{x}_j^{(imp)} - \overline{\vec{x}}^{(imp)}\right)^T / (n-1)$$

10) Update D_{nimp} by redoing 3).

11) Update L and L^{inv} by redoing 6) but remember the result of L in L_{new}

12) Calculate the matrices of absolute and relative deviations between L and L_{new}:

$$L_{abs} = \left\{l_{i,r}^{abs}\right\} \text{ for } i=1,2,\ldots,p \text{ and } r=1,2,\ldots,m$$

where

$l_{i,r}^{abs} = \left| l_{i,r} - l_{i,r}^{new} \right|$ for i=1,2,…,p and r=1,2,…, m

and

$L_{rel} = \left\{ l_{i,r}^{rel} \right\}$ for i=1,2,…,p and r=1,2,…, m

where

$$l_{i,r}^{rel} = \begin{cases} l_{i,r}^{abs} / \left| l_{i,r}^{new} \right| & \text{if } l_{i,r}^{new} \neq 0 \\ \infty & \text{if } l_{i,r}^{new} = 0 \text{ and } l_{i,r}^{abs} \neq 0 \\ 0 & \text{if } l_{i,r}^{new} = 0 \text{ and } l_{i,r}^{abs} = 0 \end{cases}, \text{ for } i=1,2,\ldots,p$$

and r=1,2,…, m

13) If all elements of L_{abs} and L_{rel} are less than 1% then go to 14) else go to 7)

14) Upgrade L=L_{new}.

15) Update D_{imp} by redoing 10).

16) Find the compressed matrix:

$$D_{compr} = \left(\vec{f}_1, \vec{f}_2, \ldots, \vec{f}_n \right)$$

where the factor analysis model without errors is used for finding the factor score:

$$\vec{f}_j = L^{inv} \vec{x}_j^{(nimp)} \text{ for } j=1,2,\ldots,n$$

17) Perform final check of consistency for the imputed values of D_{imp} by repeating for each coordinate i=1,2,...,n:

– form original data sequence $ss = \left\{ x_i^{(j)} \right\}$ for $i \in \bar{M}_j$ and j=1,2,…,n

– calculate ss_{min}, ss_{max}, $ss_{0.05}$, $ss_{0.25}$, $ss_{0.75}$ and $ss_{0.95}$ which are six quantiles of ss

– correct the imputed values using the following rules:

if $x_i^{(j)(imp)} < ss_{min}$ then

$x_i^{(j)(imp)} = ss_{0.05} + rand \times \left(ss_{0.25} - ss_{0.05} \right)$

if $x_i^{(j)(imp)} > ss_{\max}$ then

$x_i^{(j)(imp)} = ss_{0.95} - rand \times \left(ss_{0.95} - ss_{0.75} \right)$.

Here *rand* is an instance of an evenly distributed random variable in the closed interval (0; 1).

In the above algorithm 17) and 9) are optional, and so is the division with $\sqrt{k_{i,i}}$ in 3).

4. CONCLUSION

The results from the modified Roweis algorithm can be plugged into the visualization procedures given in section 2. The implementation of the presented ideas is in an Internet based tool written in Java. Its functionality is explained in the accompanying paper (Nikolova et al., 2012).

ACKNOWLEDGMENTS

This paper has been supported by the UPGRADE Black Sea SCENE project (project No. 226592, Seventh Framework Program), and the INPORT project (project No. DVU01-0031, Bulgarian National Science Fund).

REFERENCES

Acock, A.C. (1989). Measurement Error in Secondary Data Analysis, In K. Namboodiri, R. Corwin (Eds.) *Research in Sociology of Education and Socialization*, Vol. 8, pp. 201-230, Greenwich, CT: Jai Press

Acock, A.C. (2005). Working with Missing Values, *Journal of Marriage and Family*, Vol. 67, No. 4, pp. 1012-1016

Acock, A.C., and Demo, D. (1994) *Family Diversity and Well-being*, Thousand Oaks, CA: Sage

Beale, E.M.L., and Little, R.J.A. (1975) Missing Values in Multivariate Analysis, *Journal of the Royal Statistical Society, Series B (methodological)*, Vol. 37, No. 1, pp. 129-145.

Cohen, J., Cohen, P., West, S., and Aiken, L. (2003). *Applied Multiple Regression/Correction Analysis for the Behavioral Sciences*, Third Edition, Mahwah, NJ: Erlbaum

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*, Vol. 39, pp. 1-39

Duda, R., Hart, P., and Stork, D. (1973). *Pattern Classification and Scene Analysis*. A Wiley Interscience Publication.

Johnson R., and Wicherin D. (2007). *Applied Multivariate Statistical Analysis*, 6th ed., Prentice Hall

Little, R. J.A. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values, *Journal of the American Statistical Association*, Vol. 83, No. 404, pp. 1198-1202

Roweis, S. (1997). EM Algorithms for PCA and SPCA, *Neural Informatics Processing Systems, NIPS'1997*, pp. 626-632

Nikolova, N.D., Toneva-Zheynova, D., Mednikarov, B., and Tenekedjiev, K. (2012) Application of an Internet-based Tool for Visualization of Multi-Dimensional Objects with Missing Data in Maritime Education, Proc. International Conference: Expanding Frontiers – Challenges and Opportunities in Maritime Education and Training, 12-18 October, St. John, NL, Canada (in print)

Wu, J. (1983) On the Convergence Properties of the EM Algorithm, *The Annals of Statistics*, Vol. 11, No. 1, pp. 95-103