Technical Report

A note on "Missing-data theory in the context of exploratory data analysis"

Francisco Arteaga. francisco.arteaga@ucv.es

Departamento de Bioestadística e Investigación. Universidad Católica de Valencia San Vicente Mártir

Abstract

MEDA (J. Camacho. Missing data methods for exploratory data analysis, *Chemometrics and Intelligent Laboratory Systems*, 103 (2010) 8-18) is a tool for exploratory data analysis and for the interpretation of latent structures. MEDA is useful to infer the structure in the data and also to interpret the contributions of each latent variable. MEDA assists us to study the structure of a data set, based on the capability of each variable to rebuild each other, as the missing data methods do, this yields in a measure of the relation between pairs of variables in the A-dimensional sub-space of interest. MEDA is worked out from the N by M matrix of data, $\mathbf{X}$, but in this work we show that MEDA depends only on $\mathbf{S}$, the sample covariance matrix of $\mathbf{X}$, and provides one interpretation of that fact.

1. Introduction

In a recent contribution Camacho [1] proposes a method to study the structure of a data set, based on the capability of each variable to rebuild each other, as the missing data methods do. He argues that current methods for data exploration with latent variables, consisting of rotations or alternative approaches, search for a simplified representation that can miss part or the whole data structure. Their proposed method does not impose a simple structure, and does not depend on the normalization option of the data, as the rotation methods do.

MEDA is stated as an algorithm that evaluates the capability of each variable to re-build each other, as a measure of the relation between pairs of variables in the A-dimensional sub-space of interest, using a missing data method. For each dimension A, from 1 to $rank(\mathbf{X})$, MEDA yields in a squared matrix Q_A^2 with ones in the diagonal, and with a measure of the goodness of prediction of the l^{th} variable from the m^{th} variable in row m and column l, this is denoted by $Q_{A(m,l)}^2$.

MEDA can be seen as a substitute of rotation methods with better properties associated: it is more accurate than rotation methods in the detection of relationships between pairs of variables, it is robust to the overestimation of the number of PCs and it does not depend on the normalization of the loadings.

Camacho [1] calculates each $Q_{A(m,l)}^2$ value from the data matrix $\mathbf{X}$ and our objective in this work is to show that this is unnecessary, because $Q_{A(m,l)}^2$ can be calculated from $\mathbf{S}$, the sample covariance matrix of $\mathbf{X}$. This is an interesting property for a method that tries to analyse the relations among the variables for a data set: MEDA depends only on the sample covariance matrix.

2. MEDA depends only on the sample covariance matrix

MEDA uses OLS to estimate $\mathbf{X}$ from each variable $\mathbf{x}_m$ in the A-dimensional space of interest.

From the PCA model, $\mathbf{X} = \mathbf{T}^A(\mathbf{P}^A)^{\mathrm{t}} + \mathbf{E}^A$, if we call $\mathbf{X}^A = \mathbf{T}^A(\mathbf{P}^A)^{\mathrm{t}}$, and being $\mathbf{x}_m$ the m^{th} column of $\mathbf{X}$, if we apply OLS to rebuild $\mathbf{X}^A$, we have:

$$\widehat{\mathbf{X}}^A = \frac{\mathbf{x}_m \mathbf{x}_m^{\mathrm{t}}}{\|\mathbf{x}_m\|^2} \mathbf{X}^A \tag{1}$$

Other missing data methods can be used, but Arteaga and Ferrer [2] showed that the regression-based methods are statistically superior to the other methods (single-component projection, projection to the model plane, iterative imputation method). Arteaga and Ferrer [3] also showed that the different regression based methods can be seen as different approximations to the Known Data Regression method (this implies the use of OLS, as we do in equation (1)), when the submatrix of known data is ill conditioned. In our case, this submatrix is the column $\mathbf{x}_m$, no approximation is needed, and then KDR method is the natural choice.

If we call $\mathbf{x}_l^A$ the l^{th} column of $\mathbf{X}^A$, from equation (1), $\mathbf{x}_m$ is used to re-build $\mathbf{x}_l^A$ as:

$$\hat{\mathbf{x}}_l = \frac{\mathbf{x}_m \mathbf{x}_m^{\mathrm{t}}}{\|\mathbf{x}_m\|^2} \mathbf{x}_l^A \tag{2}$$

The estimation error for the l^{th} column of $\mathbf{X}$ is $\hat{\mathbf{e}}_l = \mathbf{x}_l - \hat{\mathbf{x}}_l$, and we can calculate $\|\hat{\mathbf{e}}_l\|^2$ as:

$$\|\hat{\mathbf{e}}_l\|^2 = \|\mathbf{x}_l\|^2 - 2\mathbf{x}_l^t \frac{\mathbf{x}_m \mathbf{x}_m^{\mathrm{t}}}{\|\mathbf{x}_m\|^2} \mathbf{x}_l^A + \left(\mathbf{x}_l^A\right)^{\mathrm{t}} \frac{\mathbf{x}_m \mathbf{x}_m^{\mathrm{t}}}{\|\mathbf{x}_m\|^2} \mathbf{x}_l^A \tag{3}$$

From equation (3), the index of goodness of prediction proposed by Camacho [1]:

$$Q_{A(m,l)}^2 = 1 - \frac{\|\hat{\mathbf{e}}_l\|^2}{\|\mathbf{x}_l\|^2} \tag{4}$$

becomes:

$$Q_{A(m,l)}^2 = 2\mathbf{x}_l^{\mathrm{t}} \frac{\mathbf{x}_m \mathbf{x}_m^{\mathrm{t}}}{\|\mathbf{x}_m\|^2 \|\mathbf{x}_l\|^2} \mathbf{x}_l^A - \left(\mathbf{x}_l^A\right)^{\mathrm{t}} \frac{\mathbf{x}_m \mathbf{x}_m^{\mathrm{t}}}{\|\mathbf{x}_m\|^2 \|\mathbf{x}_l\|^2} \mathbf{x}_l^A \tag{5}$$

As $\left(\mathbf{x}_l^A\right)^{\mathrm{t}} \mathbf{x}_m = \left(\mathbf{x}_l^A\right)^{\mathrm{t}} \mathbf{x}_m^A$, equation (5) can be written as:

$$Q_{A(m,l)}^2 = 2\mathbf{x}_l^{\mathrm{t}} \frac{\mathbf{x}_m (\mathbf{x}_m^A)^{\mathrm{t}}}{\|\mathbf{x}_m\|^2 \|\mathbf{x}_l\|^2} \mathbf{x}_l^A - \left(\mathbf{x}_l^A\right)^{\mathrm{t}} \frac{\mathbf{x}_m^A (\mathbf{x}_m^A)^{\mathrm{t}}}{\|\mathbf{x}_m\|^2 \|\mathbf{x}_l\|^2} \mathbf{x}_l^A = 2\frac{S_{m,l} S_{m,l}^A}{S_{m,m} S_{l,l}} - \frac{\left(S_{m,l}^A\right)^2}{S_{m,m} S_{l,l}} \tag{6}$$

yielding in:

$$Q_{A(m,l)}^2 = \frac{S_{m,l}^A}{S_{m,m} S_{l,l}} \left(2S_{m,l} - S_{m,l}^A\right) \tag{7}$$

Being $\mathbf{S}$ the covariance matrix for $\mathbf{X}$, $\mathbf{S}^A$ the covariance matrix for $\mathbf{X}^A$, and the subindices indicating the row and the column in each matrix.

Equation (7) implies that the $\mathbf{Q}_A^2$ matrix can be derived directly from $\mathbf{S}$, because $\mathbf{S}^A = \mathbf{P}^A \mathbf{\Theta}_A (\mathbf{P}^A)^{\mathrm{t}}$, with $\mathbf{\Theta}_A$ the A by A diagonal matrix with the A greatest eigenvalues of $\mathbf{S}$, in descending order, in its diagonal, and $\mathbf{P}^A$ are the associated eigenvectors, arranged as columns. This shows that MEDA depends only on the covariance matrix, and offers a method to calculate the associated $\mathbf{Q}_A^2$ matrix.

3. Consequences of the new expression for MEDA

The fact that MEDA depends only on the covariance matrix of the data set is an interesting property for a method that tries to analyse the relations among the variables for a data set, but it also has some practical consequences.

In reference [1] Camacho shows, in the appendix, the symmetry for $\mathbf{Q}_A^2$ and that, when $A = rank(\mathbf{X})$, $\mathbf{Q}_A^2$ matches the squared correlation matrix of $\mathbf{X}$. From equation (7) both properties become trivial:

The symmetry of matrices $\mathbf{S}$ and $\mathbf{S}^A$ imply the symmetry of $\mathbf{Q}_A^2$.

If $A = rank(\mathbf{X})$, then $S_{m,l}^A = S_{m,l}$, and equation (7) becomes: $Q_{A(m,l)}^2 = \left(\frac{S_{m,l}}{\sqrt{S_{m,m}}\sqrt{S_{l,l}}}\right)^2 = \rho_{m,l}^2$.

Being $\rho_{m,l}$ the coefficient of correlation between $\mathbf{x}_m$ and $\mathbf{x}_l$, and this implies the equivalence between $\mathbf{Q}_A^2$ and the squared correlation matrix, when $A = rank(\mathbf{X})$.

If we denote $\mathbf{\Omega}^A$ the sample covariance matrix for $\mathbf{E}^A$, is easy to see that $\mathbf{S} = \mathbf{S}^A + \mathbf{\Omega}^A$. This equation can be seen element-wise as $S_{m,l} = S_{m,l}^A + \Omega_{m,l}^A$, and equation (7) can be rewritten as:

$$Q^2_{A(m,l)} = \frac{\left(S_{m,l} - \Omega^A_{m,l}\right)\left(S_{m,l} + \Omega^A_{m,l}\right)}{S_{m,m}S_{l,l}} = \frac{\left(S_{m,l}\right)^2 - \left(\Omega^A_{m,l}\right)^2}{S_{m,m}S_{l,l}} = \rho^2_{m,l} - \frac{\left(\Omega^A_{m,l}\right)^2}{S_{m,m}S_{l,l}} \quad (8)$$

Equation (8) guarantees that $Q^2_{A(m,l)}$ is always less than or equal to $\rho^2_{m,l}$.

In equation (8), $\Omega^A_{m,l} = \sum_{a=A+1}^{rank(\mathbf{X})} \lambda_a p_{a,m} p_{a,l}$, being $p_{a,m}$ and $p_{a,l}$ the m^{th} and the l^{th} positions, respectively, for the a^{th} loading vector, $\mathbf{p}_a$. Notice that $\lambda_a p_{a,m} p_{a,l}$ can be interpreted as the contribution of the a^{th} principal component to the covariance $S_{m,l}$, and equation (8) can be written as:

$$Q^2_{A(m,l)} = \rho^2_{m,l} - \left(\frac{\sum_{a=A+1}^{rank(\mathbf{X})} \lambda_a p_{a,m} p_{a,l}}{\sqrt{S_{m,m}S_{l,l}}}\right)^2 \quad (9)$$

Equation (9) offers us an interesting interpretation of MEDA: $Q^2_{A(m,l)}$ is the squared correlation between $\mathbf{x}_m$ and $\mathbf{x}_l$, minus the squared of the cumulated contribution of the not extracted principal components on this correlation.

Another interesting consequence of equation (9) is that MEDA depends on the scale of **X**, because the square of the cumulated contribution of the not extracted principal components, on each correlation, does not coincide with the cumulated squared contributions of the not extracted principal components, on each correlation.

4. Conclusions

We have shown that MEDA depends only on the covariance matrix **S** and presents a useful expression for its calculation.

We use the new expression to show the symmetry for $\mathbf{Q}^2_A$ and the overlap between $\mathbf{Q}^2_A$ and the squared correlation matrix when the sub-space dimension is $A = rank(\mathbf{X})$.

The new expression is also useful to interpret $Q^2_{A(m,l)}$ as the difference between the squared correlation $\rho^2_{m,l}$ and the squared of the cumulated contribution of the not extracted principal components on this correlation.

Finally, the new expression shows us that MEDA depends on the scale of **X**.

5. References

[1] J. Camacho, Missing-data theory in the context of exploratory data analysis, *Chemometrics and Intelligent Laboratory Systems* 103 (2010) 8-18.

[2] F. Arteaga, A. Ferrer, Dealing with missing data in MSPC: several methods, different interpretations, some examples. *Journal of Chemometrics* 16 (2002) 408-418.

[3] F. Arteaga, A. Ferrer, Framework for regression-based missing data imputation methods in on-line MSPC. *Journal of Chemometrics* 19 (2005) 439-447.