Missing Data Imputation Toolbox v1.0

Methods for PCA model building with missing data

Abel Folch-Fortuny[1], Francisco Arteaga[2] and Alberto Ferrer[1]

[1]Departamento de Estadística e Investigación Operativa Aplicadas y Calidad,
Universitat Politècnica de Valencia (Spain)
{abfolfor@upv.es, aferrer@eio.upv.es}

[2]Department of Biostatistics and Investigation,
Universidad Católica de Valencia San Vicente Mártir (Spain)
francisco.arteaga@ucv.es

Summary

The Missing Data Imputation Toolbox is devoted to build principal component analysis (PCA) models when the data set **X** has missing values.

The inputs are the data set and the missing data method. If no method is specified trimmed scores regression (TSR) is used, since it represents the compromise solution between prediction quality, robustness against data structure and computation time [1].

The first step is to specify the number of principal components (PCs) to extract. To assess this, the program shows the scree plot and the cumulative % of explained variance of each component using a pair-wise estimation of the covariance matrix of **X**.

Once the number of PCs is selected, the program runs the correspondent method: trimmed scores regression (TSR), known data regression (KDR), KDR with principal component regression (KDR-PCR), KDR with partial least squares (KDR-PLS), projection to the model plane (PMP), iterative algorithm (IA) [2], modified nonlinear iterative partial least squares regression algorithm (NIPALS) [3] and data augmentation (DA) [4]. TSR, KDR, KDR-PCR and KDR-PLS are novel methods [1] adapted from their PCA model exploitation versions [5,6], as well as PMP [7].

The results are the original matrix with the imputed values, the mean and the covariance matrix of the imputed data, the number of iterations, the reconstruction of the entire original data set using the model, the method used and the computation time.

Installation

MDI toolbox needs no installation. Once the folder is included in the MATLAB search path list, the software is ready to use.

To set the path, type in the MATLAB command window:

```
>>addpath('newpath')
```

and include as 'newpath' the corresponding path of the MDI Toolbox folder (*e.g.* 'C:/Downloads/MDI Toolbox').

Finally, save the path typing:

```
>>savepath
```

License

This program is free software: you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

You should have received a copy of the GNU General Public License along with this program. If not, see <http://www.gnu.org/licenses/>.

Functions

Main function (mditoolbox.m)

INPUTS:

- Data matrix with missing data (**X**)
- Missing data method (**Method**)
    - Trimmed scores regression (**'TSR'**) (used if no method is specified)
    - Known data regression (**'KDR'**)
    - KDR with principal component regression (**'KDR-PCR'**)
    - KDR with partial least squares regression (**'KDR-PLS'**)
    - Projection to the model plane (**'PMP'**)

- Nonlinear iterative partial least squares (**'NIPALS'**)
- Iterative algorithm (**'IA'**)
- Data augmentation (**'DA'**)

OUTPUTS:

- Data structure with the results (**Results**)
  - Original data set with the imputed values (**X_imputed**)
  - Mean vector of the imputed **X** (**Mean**)
  - Covariances of the imputed **X** (**Covariances**)
  - Iterations of the missing data method (**Iterations**)
  - PCA reconstruction of **X** (**Xrec**)
  - Method specified (**Method**)
  - Computation time (**Computation_time**)

PROCEDURE:

The inputs are the data set with missing values (as NaNs), and the missing data method for the imputation. If no method is specified TSR is used, since it represents a good compromise solution between prediction quality, robustness against data structure and computation time.

The first step is to specify the number of principal components (PCs) to extract. To assess this, the program shows the scree plot and the % of explained variance of each component of the estimated covariance matrix of **X**. Then, the user introduces the number of components in the command window and presses ENTER.

Once the corresponding missing data method is applied, a data structure is obtained, containing the original matrix with the imputed values, the mean and the covariance matrix of the imputed data, the number of iterations, the reconstruction of the entire original dataset using the model, the method used and the computation time.

Pair-wise estimation of the covariance matrix (S_pairwise.m)

INPUTS:

- **X**: data matrix with missing data

OUTPUTS:

- Estimated covariance matrix (**Sout**)
- Estimated correlation matrix (**Rout**)

PROCEDURE:

The function estimates the covariance matrix using a pair-wise estimation, *i.e.* calculating the covariance between each pair of variables using the rows of the data matrix **X** in which both measurements are available. This matrix is used in the main function to calculate the estimated scree plot and the percentages of explained variance for each component.

Trimmed score method (pcambtri.m)

INPUTS:

- **X**: data matrix with missing data

OUTPUTS:

- **X**: data matrix with imputed data

PROCEDURE:

The function estimates replaces the missing values in **X** by the mean of the available values of the corresponding value. This method is known as the trimmed score method [5]. This initial imputation is used in the column-wise k-fold (*ckf*) algorithm.

Column-wise k-fold (ckf) algorithm (ckf.m)

INPUTS:

- **xcs**: data matrix
- **T**: scores matrix **xcs**
- **P**: loadings matrix of **xcs**

OUTPUTS:

- **cumpress**: cumuluative sum of squares of the prediction error (PRESS).

PROCEDURE:

This function performs a PCA cross-validation using the recently proposed [8] columnm-wise k-fold (*ckf*) algorithm. This function uses the original data set imputed using trimmed score method. More details on this method can be found in [8].

Regression-based missing data methods: TSR (pcambtsr.m), KDR (pcambkdr.m), KDR-PCR (pcambpcr.m), KDR-PLS (pcambpls.m), PMP (pcambpmp.m)

INPUTS:

- Data matrix with missing data (**X**)
- Number of components (**Method**)

OUTPUTS:

- Original data set with the imputed values (**X_imputed**)
- Mean vector of the imputed **X** (**Mean**)
- Covariances of the imputed **X** (**Covariances**)
- Iterations of the missing data method (**Iterations**)
- PCA reconstruction of **X** (**Xrec**)

PROCEDURE:

The first step is to identify which are the observed and missing values, to define the pattern of missingness. Then the data is mean centred taking into account only the measured values.

The iterative process starts with an initial mean imputation and, in each step, calculates the new imputation based on the key matrix **L**. The details on both the imputation and the key matrix can be found in the original paper [1]. The loop stops when either it reaches the convergence or exceeds the number of iterations (by default 5000).

The function returns the original matrix with the imputed values, the mean and the covariance matrix of the imputed data, the number of iterations and the reconstruction of the entire original dataset using the model.

IA (pcambia.m)

INPUTS:

- Data matrix with missing data (**X**)
- Number of components (**Method**)

OUTPUTS:

- Original data set with the imputed values (**X_imputed**)
- Mean vector of the imputed **X** (**Mean**)
- Covariances of the imputed **X** (**Covariances**)
- Iterations of the missing data method (**Iterations**)
- PCA reconstruction of **X** (**Xrec**)

PROCEDURE:

The first step is to identify which are the observed and missing values, to define the pattern of missingness. Then the data is mean centred taking into account only the measured values.

The iterative process starts with an initial mean imputation and, in each step, calculates the new imputation based on the original IA algorithm proposed by Walczak and Massart [2]. The loop stops when either it reaches the convergence or exceeds the number of iterations (by default 10000).

The function returns the original matrix with the imputed values, the mean and the covariance matrix of the imputed data, the number of iterations and the reconstruction of the entire original dataset using the model.

Modified NIPALS (pcambnipals.m)

INPUTS:

- Data matrix with missing data (**X**)
- Number of components (**Method**)

OUTPUTS:

- Original data set with the imputed values (**X_imputed**)
- Mean vector of the imputed **X** (**Mean**)
- Covariances of the imputed **X** (**Covariances**)
- Iterations of the missing data method (**Iterations**)
- PCA reconstruction of **X** (**Xrec**)

PROCEDURE:

The first step is to identify which are the observed and missing values, to define the pattern of missingness. Then the data is mean centred taking into account only the measured values.

The iterative process starts with an initial mean imputation and, in each step, calculates the new imputation based on the original modified NIPALS algorithm [3]. The loop stops when either it reaches the convergence or exceeds the number of iterations (by default 5000).

The function returns the original matrix with the imputed values, the mean and the covariance matrix of the imputed data, the number of iterations and the reconstruction of the entire original dataset using the model.

DA (DataAugmentation.m)

INPUTS:

- Data matrix with missing data (**X**)
- Number of independent chains (**M**) (10 by default)
- Length of each chain (**CL**) (100 by default)
- Number of components (**Method**)

OUTPUTS:

- Original data set with the imputed values (**X_imputed**)
- Mean vector of the imputed **X** (**Mean**)
- Covariances of the imputed **X** (**Covariances**)
- PCA reconstruction of **X** (**Xrec**)

PROCEDURE:

The first step is to identify which are the observed and missing values, to define the pattern of missingness. Then the data is mean centred taking into account only the measured values.

The iterative process starts with an initial multiple imputation for each missing value (100 by default) and calculates the posterior distribution of mean and covariance of the imputed values. Then, based on the new parameters, performs again the multiple imputation [4]. The loop stops when either it reaches the maximum runs specified (by default 10).

The function returns the original matrix with the imputed values, the mean and the covariance matrix of the imputed data, the number of iterations and the reconstruction of the entire original dataset using the model.

Data

Three .mat files are provided as examples of data sets with missing values. These data was used in [1] to perform the comparative study. In each .mat file there are several data matrices:

- **X**: complete data set
- **Round*x*_*y*MD**: data sets with missing values. The ***y*** stands for the percentage of missing values (from 10% to 70%), and the ***x*** stands for the replicate (from 1 to 10). There are 70 different data sets per .mat file.

Olive Oil data set (DataOliveOil.mat)

This data set consists of the percentage composition of eight fatty acids: palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic and eicosenoic, found in the lipid fraction of 572 Italian olive oils. In the original data set [9] there are nine collection areas from three different regions of Italy. To reduce dimensionality, this data set is built with 75 randomly chosen wines from South Apulia.

NIR Diesel data set (DataDiesel.mat)

NIR Diesel data set is obtained from the Eigenvector Research Inc. data library (http://www.eigenvector.com/data/index.html). The data set contains the NIR spectra of several diesel fuels (N = 40) obtained at the Southwest Research Institute (SWRI) on a project sponsored by the U.S. Army [10]. The fuels were originally scanned from the wavelength 750 nm to 1550 nm in 2 nm increments (K = 401 variables).

Simulated data set (DataSimulated.mat)

The Simulated one is a three-component multivariate data set [11,12]. This data set has ten variables (K = 10) and a hundred samples (N = 100), and follows a multivariate normal distribution with zero means and unit variances. The highest eigenvalue of the correlation matrix is forced to be 4.0, the second one 3.0 and the last one 2.0, explaining 40%, 30% and 20% of the variance from the ten variables, respectively.

Example of analysis

An example of analysis can be found in the article where this toolbox is described [13].

There are methods, *e.g.* DA and PMP, which are unable to converge with some of the example data sets. More details in [1].

References

[1] A. Folch-Fortuny, F. Arteaga, A. Ferrer, PCA model building with missing values: new methods and a comparative study, Chemometrics and Intelligent Laboratory Systems 146 (2015) 77-88.

[2] B. Walczak, D.L. Massart, Dealing with missing data Part I, Chemometrics and Intelligent Laboratory Systems 58 (2001) 15-27.

[3] S. Wold, C. Albano, W.J. Dunn, K. Esbensen, S. Hellberg, E. Johansson, M. Sjöstrom. In H. Martens and H. Russwurm, Jr. (Editors), Food Research and Data Analysis, Applied Science Publishers, London, 1983, pp. 183-185.

[4] J.L. Schafer, Analysis of Incomplete Multivariate Data, CRC Press, New York, 1997.

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

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

[7] P.R.C. Nelson, P.A. Taylor, J.F. MacGregor, Missing data methods in PCA and PLS: Score calculations with incomplete observations, Chemometrics and Intelligent Laboratory Systems 35 (1996) 45-65.

[8] E. Saccenti, J. Camacho, On the use of the observation-wise k-fold operation in PCA cross-validation, Journal of Chemometrics 29 (2015) 467–478.

[9] M. Forina, C. Armanino, S. Lanteri, E. Tiscornia, Classification of Olive Oils from their Fatty Acid Composition, in H. Martens, H.Jr Russwurm Eds, Food Research and Data Analysis, Applied Science Pub, London, 1983, pp. 189-214

[10] S.A. Hutzler and G.B. Bessee, Remote Near-Infrared Fuel Monitoring System, Interim Report, U.S. Army TARDEC Fuels and Lubricants Research Facility, Southwest Research Institute, San Antonio, United States, 1997.

[11] F. Arteaga, A. Ferrer, How to simulate normal data sets with the desired correlation structure, Chemometrics and Intelligent Laboratory Systems 101 (2010) 38-42.

[12] F. Arteaga, A. Ferrer, Building covariance matrices with the desired structure, Chemometrics and Intelligent Laboratory Systems 127 (2013) 80-88.

[13] A. Folch-Fortuny, F. Arteaga, A. Ferrer, Missing Data Imputation Toolbox for MATLAB, submitted.