zbMATH Open
THE FIRST RESOURCE FOR MATHEMATICS

**Josse, Julie**; **Pagès, Jérôme**; **Husson, François**
**Multiple imputation in principal component analysis.** (English) Zbl 1274.62409
Adv. Data Anal. Classif., ADAC 5, No. 3, 231-246 (2011).

Summary: The available methods to handle missing values in principal component analysis only provide point estimates of the parameters (axes and components) and estimates of the missing values. To take into account the variability due to missing values a multiple imputation method is proposed. First a method to generate multiple imputed data sets from a principal component analysis model is defined. Then, two ways to visualize the uncertainty due to missing values onto the principal component analysis results are described. The first one consists in projecting the imputed data sets onto a reference configuration as supplementary elements to assess the stability of the individuals (respectively of the variables). The second one consists in performing a principal component analysis on each imputed data set and fitting each obtained configuration onto the reference one with Procrustes rotation. The latter strategy allows to assess the variability of the principal component analysis parameters induced by the missing values. The methodology is then evaluated from a real data set.

**MSC:**

62H25 Factor analysis and principal components; correspondence analysis
62G09 Nonparametric statistical resampling methods

Cited in **5** Documents

**Keywords:**

principal component analysis; missing values; EM algorithm; multiple imputation; bootstrap; Procrustes rotation

**Software:**

R; missMDA

**Full Text:** DOI

**References:**

[1] Adams E, Walczak B, Vervaet C, Risha PG, Massart D (2002) Principal component analysis of dissolution data with missing elements. Int J Pharm 234: 169–178· doi:10.1016/S0378-5173(01)00966-8

[2] Bro R (1998) Multi-way analysis in the food industry–models, algorithms, and applications. Tech. rep., MRI, EPG and EMA, Proc ICSLP 2000

[3] Bro R, Kjeldahl K, Smilde AK, Kiers HAL (2008) Cross-validation of component models: A critical look at current methods. Anal Bioanal Chem 5: 1241–1251· doi:10.1007/s00216-007-1790-1

[4] Caussinus H (1986) Models and uses of principal component analysis. In: de Leeuw J, Heiser W, Meulman J, Critchley F (eds) Multidimensional data analysis. DSWO Press, Ram, pp 149–178

[5] Chateau F, Lebart L (1996) Assessing sample variability in the visualization techniques related to principal componenet analysis: Bootstrap and alternative simulation methods. In: COMPSTAT, pp 205–210

[6] Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Royal Stat Soc B 39: 1–38 · Zbl 0364.62022

[7] Denis JB (1991) Ajustements de modèles linéaires et bilinéaires sous contraintes linéaires avec données manquantes. Revue de Statistique Appliquée 39: 5–24

[8] Dray S (2008) On the number of principal components: A test of dimensionality based on measurements of similarity between matrices. Comput Stat Data Anal 52: 2228–2237 · Zbl 1452.62409 · doi:10.1016/j.csda.2007.07.015

[9] Escofier B, Pagès J (2008) Analyses factorielles simples et multiples, 4th edn. Economica, Paris

[10] Gabriel KR, Zamir S (1979) Lower rank approximation of matrices by least squares with any choice of weights. Technometrics 21: 236–246 · Zbl 0471.62004 · doi:10.1080/00401706.1979.10489819

[11] Golub GH, Van Loan CF (1996) Matrix computations, 3rd edn. Johns Hopkins University Press, Baltimore · Zbl 0865.65009

[12] Gower JC, Dijksterhuis GB (2004) Procrustes problems. Oxford University Press, New York · Zbl 1057.62044

[13] Greenacre M (1984) Theory and applications of correspondence analysis. Acadamic Press, London · Zbl 0555.62005

Edited by FIZ Karlsruhe, the European Mathematical Society and the Heidelberg Academy of Sciences and Humanities
© 2022 FIZ Karlsruhe GmbH

[14] Grung B, Manne R (1998) Missing values in principal component analysis. Chemiometr Intell Lab Syst 42: 125–139· doi:10.1016/S0169-7439(98)00031-8

[15] Husson F, Josse J (2010) missMDA: Handling missing values with/in multivariate data analysis (principal component methods). http://www.agrocampus-ouest.fr/math/husson , http://www.agrocampus-ouest.fr/math/josse , R package version 1.2

[16] Ilin A, Raiko T (2010) Practical approaches to principal component analysis in the presence of missing values. J Mach Learn Res 11: 1957–2000 · Zbl 1242.62047

[17] Josse J, Pagès J, Husson F (2009) Gestion des données manquantes en analyse en composantes principales. J de la Société Française de Statistique 150: 28–51 · Zbl 1311.62091

[18] Josse J, Pagès J, Husson F (2011) Selecting the number of components in principal component analysis using cross-validation approximations (submitted)

[19] Kiers HAL (1997) Weighted least squares fitting using ordinary least squares algorithms. Psychometrica 62: 251–266 · Zbl 0873.62058 · doi:10.1007/BF02295279

[20] Kroonenberg PM (2008) Applied Multiway data analysis (chap.7). Wiley series in probability and statistics, New York

[21] Little RJA, Rubin DB (1987) 2002) Statistical analysis with missing data. Wiley series in probability and statistics, New York

[22] Milan M (1995) Application of the parametric bootstrap to models that incorporate a singular value decomposition. J Royal Stat Soc Ser C 44: 31–49 · Zbl 0821.62030

[23] Netflix (2009) Netflix challenge. http://www.netflixprize.com

[24] Nora-Chouteau C (1974) Une méthode de reconstitution et d'analyse de données incomplètes. PhD thesis, Université Pierre et Marie Curie

[25] Peres-Neto PR, Jackson DA, Somers KM (2005) How many principal components? stopping rules for determining the number of non-trivial axes revisited. Comput Stat Data Anal 49: 974–997 · Zbl 1429.62223 · doi:10.1016/j.csda.2004.06.015

[26] R Development Core Team (2009) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org , ISBN 3-900051-07-0

[27] Raiko T, Ilin A, Karhunen J (2007) Principal component analysis for sparse high-dimensional data. In: Neural Information Processing, pp 566–575

[28] Rubin DB (1987) Multiple imputation for non-response in survey. Wiley, New York

[29] Schafer JL (1997) Analysis of incomplete multivariate data. Chapman \& Hall/CRC, London

[30] Schafer JL, Olsen MK (1998) Multiple imputation for missing-data problems: A data analyst's perspective. Multivar Behav Res 33: 545–571· doi:10.1207/s15327906mbr3304$_5$

[31] Song J (1999) Analysis of incomplete high-dimensional multivariate normal data using a common factor model. PhD thesis, Dept. of Biostatistics, UCLA, Los Angeles

[32] Tanner MA, Wong WH (1987) The calculation of posterior distributions by data augmentation. J Am Stat Assoc 82: 805–811

[33] Timmerman ME, Kiers HAL, Smilde AK (2007) Estimating confidence intervals for principal component loadings: a comparaison between the bootstrap and asymptotic results. Br J Math Stat Psychol 60: 295–314· doi:10.1348/000711006X109636

[34] Tipping M, Bishop CM (1999) Probabilistic principal component analysis. J Royal Stat Soc B 61: 611–622 · Zbl 0924.62068 · doi:10.1111/1467-9868.00196

[35] van Buuren S (2007) Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res 16: 219–242 · Zbl 1122.62382 · doi:10.1177/0962280206074463

[36] Wold H (1966) Nonlinear estimation by iterative least squares procedures. In: David FN (eds) Research Papers in Statistics: Festschrift for Jerzy Neyman. Wiley, New York, pp 411–444

This reference list is based on information provided by the publisher or from digital mathematics libraries. Its items are heuristically matched to zbMATH identifiers and may contain data conversion errors. It attempts to reflect the references listed in the original paper as accurately as possible without claiming the completeness or perfect precision of the matching.