Matrix reduced-rank approximation through cross-validation

Marisol García-Peña[1], Sergio Arciniegas-Alarcón[2], Johannes Forkman[3] and Diego Duarte[4]

[1]*Pontificia Universidad Javeriana, Bogotá, Colombia*
[2]*Universidade Federal da Bahia, Salvador, Brasil*
[3]*Swedish University of Agricultural Sciences, Uppsala, Sweden*
[4]*Universidad de La Sabana, Chía, Colombia.*

Abstract

Low-rank approximation of data matrices is used in many statistical methods, including principal component analysis, and is commonly obtained through singular value decomposition (SVD) by retaining the leading singular values and their associated principal components. A key challenge is determining the optimal rank of the approximation. Cross-validation for rank selection was pioneered by Wold [8] and Eastment and Krzanowski [4, 7]. Several alternative approaches have been proposed, see for example [2, 3, 6]. More recently, Arciniegas-Alarcón et al. [1] introduced a leave-one-out cross-validation procedure in which each observation is omitted in turn, imputed, and subsequently predicted using an SVD-based low-rank approximation. The optimal rank is then selected as the one that minimises the prediction error in a least-squares sense.

This study extends the method of Arciniegas-Alarcón et al. [1] in two respects. First, leave-group-out cross-validation is adopted in place of leave-one-out cross-validation since the latter may be asymptotically inconsistent [5] and thus could have lower predictive ability. Second, several alternative imputation methods are evaluated. The proposed approaches are assessed through simulation studies and applications to real genotype-by-environment data matrices. The results indicate that the most effective strategy combines an initial imputation based on either projection onto the model plane or an approximate Bayesian bootstrap with subsequent fitting of an SVD-derived multiplicative model.

Keywords

Low-rank approximation, Missing data imputation, Principal component analysis, Rank selection.

Acknowledgements

The first author thanks the Pontificia Universidad Javeriana for the support through project 21243. The fourth author thanks the Faculty of Engineering. The High Performance Computing Center ZINE was employed the simulation study.

References

[1] Arciniegas-Alarcón, S., García-Peña M., Krzanowski, W.J. and Rengifo, C. (2024). Cross-validation to select the optimum rank for a reduced-rank approximation to multivariate data. *Journal of Crop Improvement 38*, 344-–367.

[2] Bro, R., Kjeldahl, K., Smilde, AK, Kiers, HA. (2008). Cross-validation of component models: a critical look at current methods. *Analytical and Bioanalytical Chemistry 390*, 1241–51.

[3] Camacho, J., Ferrer, A. (2014). Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: practical aspects. *Chemomometrics and Intelligent Laboratory Systems 131*, 37—50.

[4] Eastment, H.T. and Krzanowski, W.J. (1982). Cross-validatory choice of the number of components from a principal component analysis. *Technometrics 24*, 73–77.

[5] Eshghi, P. (2014). Dimensionality choice in principal component analysis via cross-validatory methods. *Chemometrics and Intelligent Laboratory Systems 130*, 6—13.

[6] Josse, J., Husson, F. (2012). Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics & Data Analysis 56*, 1869–1879.

[7] Krzanowski, W.J. and Kline, P. (1995). Cross-validation for choosing the number of important components in principal component analysis. *Multivariate Behavioral Research 30*, 149–165.

[8] Wold, S. (1978). Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics 20*, 397–405.