The Journal of Open Source Software

hoggorm: a python library for explorative multivariate statistics

Oliver Tomic[1], **Thomas Graff**[2], **Kristian Hovde Liland**[1], **and Tormod Næs**[3]

1 Norwegian University of Life Sciences, Ås, Norway **2** TGXnet, Norway **3** Nofima, Ås, Norway

DOI: 10.21105/joss.00980

Software

- Review
- Repository
- Archive

Submitted: 16 August 2018
Published: 11 July 2019

Summary

hoggorm is a python library for explorative analysis of multivariate data that implements statistical methods typically used in the field of chemometrics (Tormod Næs & Martens, 1988). Although hoggorm shares some statistical methods with the Python library scikit-learn for machine learning, it follows the chemometrics paradigm for data analysis where great attention is paid to understanding and interpretation of the variance in the data.

Currently (version 0.13.3), statistical methods implemented in hoggorm are: (I) principal component analysis (PCA) for analysis of single data arrays or matrices (Mardia, Kent, & Bibby, 1979); (II) principal component regression (PCR) (Tormod Næs & Martens, 1988) and (III) partial least squares regression (PLSR) (H. Wold, Martens, & Wold, 1983) for analysis of two data arrays. PLSR is provided in two versions; (a) PLS1 for multivariate independent data and a single response variable; (b) PLS2 for situations where the independent data and response data are both multivariate. PCA is an unsupervised method which compresses data into low dimensional representations that capture the dominant variation in the data. PCR uses the compressed features as a basis for regression, while PLSR uses supervised compression to capture the dominant co-varation between the data matrix and the target/response. Both PLS1, PLS2 and PCR posess a couple of useful properties: they easily handle situations where: (a) the multivariate independent data are short and wide, that is, data with few objects (instances) and many variables (features); (b) the multivariate independent data contain many highly correlated variables, thus providing stable models despite high correlations.

The hoggorm package provides access to an extended repertoire of interpretation tools that are integrated in PCA, PCR, PLS1 and PLS2. These including scores, loadings, correlation loadings, explained variances for calibrated and validated models (both for individual variables as well as all variables together). Scores are the objects' coordinates in the compressed data representation and can for instance be used to search for patterns or groups among the objects. Loadings are the variables' representations in the compressed space showing their contribution to the components. Finally, correlation loadings show how each variable correlates to the score vectors/components and how much of the variation in each variable is explained across components. Note that models trained with hoggorm may also be applied for prediction purposes, both for continuous and categorical variables, where appropriate.

Furthermore, hoggorm implements the matrix correlation coefficient methods RV (P. Robert & Escoufier, 1976) and RV2 (also known as modified RV) (Smilde, Kiers, Bijlsma, Rubingh, & Erk, 2009), as well as the similarity index for comparing coupled matrices index (SMI) (Indahl, Næs, & Liland, 2018). These methods can be used to quickly determine how much common information there is between two data matrices. Results from models trained with hoggorm may be visualised using the complementary plotting package hoggormplot (v0.13.2).

Acknowledgements

Both users and developers have made valuable contributions to improve the usability the hoggorm library. This includes reporting of bugs, testing various features and other forms of feedback. A complete list of contributors is provided at https://github.com/olivertomic/hoggorm/graphs/contributors. The authors are also thankful to reviewer Dr. Javier Sanchez Galan at the Universidad Tecnológica de Panama in Panama for testing and constructive feedback on requirements for publication of the hoggorm package at the Journal of Open Source Software.

References

Indahl, U., Næs, T., & Liland, K. (2018). A similarity index for comparing coupled matrices. *Journal of Chemometrics*. doi:10.1002/cem.3049

Mardia, K., Kent, J., & Bibby, J. (1979). *Multivariate analysis*. London: Academic Press.

P. Robert, & Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: The *RV*-coefficient. *Applied Statistics*, *25*, 257–265. doi:10.2307/2347233

Smilde, A., Kiers, H., Bijlsma, S., Rubingh, C., & Erk, M. van. (2009). Matrix correlations for high-dimensional data: The modified rv-coefficient. *Bioinformatics*, *25*, 401–405. doi:10.1093/bioinformatics/btn634

Tormod Næs, & Martens, H. (1988). Principal components regression in NIR analysis: Viewpoints, background details and selection of components. *Journal of Chemometrics*, *2*, 155–167. doi:10.1002/cem.1180020207

Wold, H., Martens, M., & Wold, S. (1983). Matrix pencils. Lecture notes in mathematics. In B. Kågström & A. Ruhe (Eds.), (pp. 286–293). Springer, Berlin, Heidelberg. doi:https://doi.org/10.1007/BFb0062108