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 **2** TGXnet **3** Nofima

DOI: 10.21105/joss.00896

Software

- Review
- Repository
- Archive

Submitted: 16 August 2018
Published: 16 August 2018

License
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC-BY).

Summary

hoggorm is a python library for explorative analysis of multivariate data that implements statistical methods typically used in the field of chemometrics (Martens and Næs 1988). Although hoggorm shares some statistical methods with the Python library scikit-learn for machine learning, hoggorm follows the chemometrics paradigm for data analysis where great attention is paid to understanding and interpretation of the variance in the data. For this purpose, hoggorm provides access to typical interpretation tools, such as scores, loadings, correlation loadings, explained variances for calibrated and validated models (both for individual variables as well as all variables together). Note that models trained with hoggorm may also be applied for prediction purposes, both for continuous and categorical variables, where appropriate. Currently (version 0.12.0), statistical methods implemented in hoggorm are: (I) principal component analysis (PCA) for analysis of single data arrays or matrices (Mardia, Kent, and Bibby 1979); (II) principal component regression (PCR) (Martens and Næs 1988) and (III) partial least squares regression (PLSR) (Wold 1982) 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. Furthermore, hoggorm implements the matrix correlation coefficient methods RV (P. Robert and Escoufier 1976), RV2 (also known as modified RV) (Smilde et al. 2009) as well as the similarity index for comparing coupled matrices index (SMI) (U.G. Indahl 2018).

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

References

Mardia, K.V., J.T. Kent, and J.M. Bibby. 1979. *Multivariate Analysis*. London: Academic Press.

Martens, Harald, and Tormod Næs. 1988. "Principal Components Regression in Nir Analysis." *Journal of Chemometrics* 2:155–67.

P. Robert, and Y. Escoufier. 1976. "A Unifying Tool for Linear Multivariate Statistical Methods: The *RV*-Coefficient." *Applied Statistics* 25:257–65.

Smilde, Age K., Henk A. L. Kiers, S. Bijlsma, C. M. Rubingh, and M. J. van Erk. 2009. "Matrix Correlations for High-Dimensional Data: The Modified Rv-Coefficient." *Bioinformatics* 25:401–5.

U.G. Indahl, K.H. Liland, T. Næs. 2018. "A Similarity Index for Comparing Coupled Matrices." *Journal of Chemometrics*. https://doi.org/10.1002/cem.3049.

Wold, H. 1982. "Systems Under Indirect Observation." In, edited by K.G. Jøreskog and H. Wold. Amsterdam: North Holland.