Depth and depth-based classification with R package ddalpha

O. Pokotylo[a], **P. Mozharovskyi**[b], **R. Dyckerhoff**[c], and **S. Nagy**[d]

[a]alexey.pokotylo@gmail.com

[b]Center for Research in Economics and Statistics
National School for Statistics and Information Analysis
Campus de Ker-Lann - Rue Blaise Pascal - 35172 Bruz, France
pavlo.mozharovskyi@ensai.fr

[c]Institute of Econometrics and Statistics
University of Cologne
Albertus-Magnus-Platz - 50923 Cologne, Germany
rainer.dyckerhoff@statistik.uni-koeln.de

[d]Department of Probability and Mathematical Statistics
Charles University
Sokolovská 83 - Praha 8 - CZ-186 75, Czech Republic
nagy@karlin.mff.cuni.cz

Keywords : Data depth, Supervised classification, DD-plot, Outsiders, Visualization, Functional classification, DDalpha.

Following the seminal idea of Tukey (1975), data depth is a function that measures how close an arbitrary point of the space is located to an implicitly defined center of a data cloud. Having undergone theoretical and computational developments, it is now employed in numerous applications with classification being the most popular one. The R package ddalpha is a software directed to fuse experience of the applicant with recent achievements in the area of data depth and depth-based classification.

ddalpha provides an implementation for exact and approximate computation of most reasonable and widely applied notions of data depth in multivariate and functional (Nagy, Gijbels & Hlubinka, 2017) settings. These can be further used in the depth-based classifiers implemented in the package, where the $DD\alpha$-procedure (Lange, Mosler & Mozharovskyi, 2014) is in the main focus; see also Mozharovskyi, Mosler & Lange (2015). The package is expandable with user-defined custom depth methods and separators. The implemented functions for depth visualization and the built-in benchmark procedures may also serve to provide insights into the geometry of the data and the quality of pattern recognition.

Except for efficient and fast implementation of the $DD\alpha$-procedure, R package ddalpha suggests other classification techniques that can be employed in the DD-plot: the original polynomial separator by Li, Cuesta-Albertos & Liu (2012) and the depth-based kNN-classifier proposed by Vencalek (2011). Certain depth functions, especially those based on data geometry, vanish beyond the convex hull of the sample and thus can yield zero depth for distant observations during their classification (the so-called “outsiders”). For this case, ddalpha offers a number of outsider treatments and a mechanism for their management. Further, ddalpha possesses tools for immediate classification of functional data in which the measurements are first brought onto a finite dimensional basis, and then fed to the depth-based classifier. In addition, the componentwise classification technique by Delaigle, Hall & Bathia (2012) is implemented.

R package ddalpha implements various depth functions and classifiers for multivariate and functional data under one roof. Among others, ddalpha implements zonoid depth and efficient exact halfspace depth. All depths in the package are implemented for any dimension $d \geq 2$; except for the projection depth all implemented algorithms are exact, and supplemented by their approximating versions to deal with the increasing computational burden for large samples and higher dimensions. In addition, the package contains 50 multivariate and 5 functional ready-to-use classification problems and data generators for a palette of distributions.

Most of the functions of the package are programmed in C++, in order to be fast and efficient. The package has a module structure, which makes it expandable and allows user-defined custom depth methods and separators. ddalpha employs boost (package BH; Eddelbuettel, Emerson & Kane, 2016), a well known fast and widely applied library, and resorts to Rcpp (Eddelbuettel, Francois, Allaire, Ushey, Kou, Bates & Chambers, 2016) allowing for simple calls of R functions from C++.

R package ddalpha (Pokotylo, Mozharovskyi, Dyckerhoff & Nagy, 2018) is accessible on CRAN, for a comprehensive overview of its functionality see Pokotylo, Mozharovskyi & Dyckerhoff (2016).

References

Delaigle, A., Hall, P. and Bathia, N. (2012). Componentwise classification and clustering of functional data. *Biometrika*, **99**, 299-313.

Eddelbuettel, D., Emerson, J.W. and Kane, M.J. (2016). BH: boost C++ header files. R package version 1.60.0-2, https://CRAN.R-project.org/package=BH.

Eddelbuettel, D., Francois, R., Allaire, J., Ushey, K., Kou, Q., Bates, D. and Chambers, J. (2016). Rcpp: Seamless R and C++ integration. R package version 0.12.5, https://CRAN.R-project.org/package=Rcpp.

Lange, T., Mosler, K. and Mozharovskyi, P. (2014). Fast nonparametric classification based on data depth. *Statistical Papers*, **55** 49-69.

Li, J., Cuesta-Albertos, J.A. and Liu, R.Y. (2012). DD-classifier: nonparametric classification procedure based on DD-plot. *Journal of the American Statistical Association*, **107**, 737-753.

Mozharovskyi, P., Mosler, K. and Lange, T. (2015). Classifying real-world data with the *DDα*-procedure. *Advances in Data Analysis and Classification*, **9**, 287-314.

Nagy, S., Gijbels, I. and Hlubinka, D. (2017). Depth-based recognition of shape outlying functions. *Journal of Computational and Graphical Statistics*, **26**, 883-893.

Pokotylo, O., Mozharovskyi, P. and Dyckerhoff, R. (2016). Depth and depth-based classification with R-package ddalpha. arXiv:1608.04109 [stat.CO].

Pokotylo, O., Mozharovskyi, P., Dyckerhoff, R. and Nagy, S. (2018). ddalpha: depth-based classification and calculation of data depth. R package version 1.3.1.1, https://CRAN.R-project.org/package=ddalpha.

Tukey, J.W. (1975). Mathematics and the Picturing of Data. In James, R. (ed.) *Proceedings of the International Congress of Mathematicians, volume 2*, pp. 523-531. Canadian Mathematical Congress.

Vencalek, O. (2011). *Weighted Data Depth and Depth Based Discrimination.* Ph.D. thesis, Charles University, Prague.