A SUPERVISED CLASSIFICATION STRATEGY BASED ON THE NOVEL DIRECTIONAL DISTRIBUTION DEPTH FUNCTION

Edoardo Redivo[1] Cinzia Viroli[1]

[1] Department of Statistical Sciences, University of Bologna, (e-mail: edoardo.redivo@unibo.it, cinzia.viroli@unibo.it)

ABSTRACT: Statistical depth functions are a class of functions that provide a center-outward ordering of sample points in multidimensional space. In this work we introduce a novel depth function that is based on the cumulative distribution function along random directions, and is thus termed directional distribution depth. Some properties and a connection to the Mahalanobis depth when applied to sphered data are shown. The proposed depth is used as a basis for supervised classification using maximum depth classifiers and more flexible polynomial separators in the depth space. It is shown to be effective and competitive against other depth functions through simulated experiments and real data applications.

KEYWORDS: depth functions, random projection, supervised classification

1 Introduction

In multivariate analysis the identification of order statistics, quantiles and atypical patterns is very challenging due to the lack of an order among observations, which is instead natural in the real line $\mathbb{R}^1$ (Kong & Mizera, 2012; Serfling, 2002). To overcome this challenge, the most important line of research is rooted in the concept of statistical depth, which leads to a center-outward ordering of the sample points in $\mathbb{R}^p$ with $p \geq 2$. More specifically, a depth function is a function that can assign a real number to each point of in multivariate space, measuring the outlyingness of the point with respect to the barycenter, and can be used as a starting point for outlier detection, clustering, classification.

Popular depth functions are the Mahalanobis depth, which is based on the Mahalanobis distance (Mahalanobis, 1936), and the halfspace depth, which measures the depth of a point by the smallest probability of a halfspace that contains that same point. Liu *et al.* (1999) described different depth functions as valuable exploratory tools in multivariate analysis. Introducing some notation, let $\mathbf{X}$ be a multivariate random variable of order p with a probability distribution F: a data depth measures how deep (or central) a given value $\mathbf{x}$ of $\mathbf{X}$ is with respect to the data cloud or a given distribution function and is usually denoted as $D(\mathbf{x}, F)$. A simple example is the Mahalanobis depth, which is inversely proportional to the Mahalanobis distance: $MD(\mathbf{x}, F) = \left[1 + (\mathbf{x} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]^{-1}$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean vector and dispersion matrix of $\mathbf{X}$ and can be estimated from the data.

Zuo & Serfling (2000) reviewed some of the most popular depth functions and introduced some desirable properties that in their view can be define a proper depth function. More precisely, a depth function is a non-negative and bounded function, which is: (i) invariant to the coordinate system or to the scale of the underlying measurements (affine invariance); (ii) maximum at its center; (iii) monotonically decreasing when a point moves away from the deepest central point and (iv) it should approach zero as a point approaches infinity. Some other properties that can be attractive and that we will consider are consistency of the function based on sample data to a population counterpart, and computational feasibility, *i.e.*, it should be possible to compute the depth values of data points efficiently even for large p.

2 Directional Distribution Depth

Let $\mathbf{S}$ be a random vector of length p with a uniform distribution on the sphere, that is any of its realizations $\mathbf{s}$ is a direction belonging to the sphere ($\mathbb{S}^{p-1}$) and having unit norm ($||\mathbf{s}||_2 = 1$). The depth of a point is derived by projecting it along any direction and evaluating the cumulative distribution function of the univariate distribution of the projected data $\mathbf{S}^\top \mathbf{X}$. The resulting probability is transformed so that the depth is symmetric with respect to the median, defined as the deepest point. As a last step we take the expected value over all random direction. More precisely, the directional distribution depth is the mapping $\mathbb{R}^p \times \mathcal{F} \to [0, 1]$ defined as

$$D(\mathbf{x}, F) = E_{\mathbf{S}}\left[1 - 2|F_{\mathbf{S}^\top\mathbf{X}}(\mathbf{S}^\top \mathbf{x}) - 0.5|\right], \tag{1}$$

where $E_{\mathbf{S}}$ is the expectation with respect to the random vector $\mathbf{S}$, F is the probability distribution of the multivariate data and $F_{\mathbf{S}^\top\mathbf{X}}$ is the marginal probability distribution of the transformation $\mathbf{S}^\top \mathbf{X}$ evaluated at $\mathbf{S}^\top \mathbf{x}$. $F_{\mathbf{S}^\top\mathbf{X}}$ can be any (probabilistic or nonparametric) univariate distribution function differently parameterized along each direction. In this work we will focus and compare the depth based on the Gaussian distribution, on the *fgld* quantile function due to its large flexibility (Redivo *et al.*, 2023; Chakrabarty & Sharma, 2021) and the nonparametric kernel density estimation.

Theorem 1. *Given whatever model choice of $F_{\mathbf{S}}$, the depth defined in (1) is a proper depth function in the sense of the definition given by Zuo & Serfling (2000).*

An interesting property closely related to the proposed depth function is that the average squared distance of univariate projections from the mean, applied to sphered data, is proportional to the Mahalanobis distance in the original multivariate space:

$$E_{\mathbf{S}}\left[\left(\mathbf{S}^{\top}\tilde{\mathbf{x}}-\mathbf{S}^{\top}\tilde{\boldsymbol{\mu}}\right)^2\right]=\frac{1}{p}(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}),$$

where $\tilde{\mathbf{x}}$ and $\tilde{\boldsymbol{\mu}}$ are respectively the point and center transformed via the sphering matrix. Next we adapt our depth definition to sample data. Let $\mathbf{X}_n$ be a sample of size n from $\mathbf{X}$, without loss of generality we assume it to sphered. Let $\mathbf{s}_B$ be a set of B random directions. Then the sample version of the directional distribution depth for a generic point $\mathbf{x}_i$ is

$$D_n(\mathbf{x}_i,F)=\frac{\sum_{b=1}^{B}\left[1-2|\hat{F}_{\mathbf{s}_b^{\top}\mathbf{X}_n}(\mathbf{s}_b^{\top}\mathbf{x}_i)-0.5|\right]}{B}, \tag{2}$$

This quantity is strongly consistent with respect to its population counterpart, that is as $n\rightarrow\infty$ and $B\rightarrow\infty$, $D_n(\mathbf{x},F)\xrightarrow{a.s.}D(\mathbf{x},F)$.

3 Application to Supervised Classification

We apply to proposed depth function to supervised classification by allocating a new observation to the class with the maximum depth among the K populations (Ghosh & Chaudhuri, 2005). The performance of the proposed depth (with its three distribution estimators) is evaluated through a simulation study, comparing it to maximum depth classifiers based on other depth definitions (Mahalanobis, projection, simplicial and halfspace) and to linear and quadratic discriminant analysis. The simulation comprises three distributional scenarios: with Gaussian data classifiers based on the directional distribution depth perform similarly well to those based on data generating normal model; with t-distributed data, linear discriminant analysis performs the best, being quite robust to the heavier tails, with the distributional depth classifiers lagging shortly behind; with skewed data our depth performs generally better than the alternatives, being the only one that can accommodate non-elliptical data, which is assumed by the Mahalanobis depth and the discriminant analysis methods. Throughout the simulations classifiers based on the halfspace depth have substantially worse results, and this is probably due to the difficulty in computing the depth, with only an approximation being available in higher dimensions, where the resulting classifier suffers the most.

We also applied depth based classifiers to commonly used benchmark data sets. Here we have considered polynomial separators for the classes in the depth space, in contrast to the quadrant bisector line implicitly assumed by the maximum depth classifier. This method is called DD-classifier and has been introduced in Li *et al.* (2012). The DD-classifier based on the new depth is able to achieve competitive accuracies (measured through mean accuracy in repeated training-testing splits) even against K-nearest neighbours and SVM.

References

CHAKRABARTY, T. K., & SHARMA, D. 2021. A Generalization of the Quantile-Based Flattened Logistic Distribution. *Annals of Data Science*, **8**(3), 603–627.

GHOSH, A. K., & CHAUDHURI, P. 2005. On Maximum Depth and Related Classifiers. *Scandinavian Journal of Statistics*, **32**(2), 327–350.

KONG, L., & MIZERA, I. 2012. Quantile Tomography: Using Quantiles With Multivariate Data. *Statistica Sinica*, **22**(4), 1589–1610.

LI, J., CUESTA-ALBERTOS, J. A., & LIU, R. Y. 2012. DD-Classifier: Nonparametric Classification Procedure Based on DD-Plot. *Journal of the American Statistical Association*, **107**(498), 737–753.

LIU, R. Y., PARELIUS, J. M., & SINGH, K. 1999. Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *Annals of Statistics*, **27**(3), 783–858.

MAHALANOBIS, P.C. 1936. On the Generalized Distance in Statistics. *Proceedings of the National Institute of Science of India*, **2**, 49–55.

REDIVO, E., VIROLI, C., & FARCOMENI, A. 2023. Quantile-based distribution functions and their use for classification, with application to naïve Bayes Classifiers. *Statistics and Computing*.

SERFLING, R. 2002. Quantile functions for multivariate analysis: approaches and applications. *Statistica Neerlandica*, **56**(2), 214–232.

ZUO, Y., & SERFLING, R. 2000. General notions of statistical depth function. *The Annals of Statistics*, **28**(2), 461 – 482.