Taming the tail through data depth

Sibsankar Singha, Marie Kratz and Sreekar Vadlamani *

Due to recent advances in high-dimensional statistics, there is a renewed interest in developing tools to better understand the geometric structure of datasets. Numerous multivariate statistical depth functions have been proposed to establish ranks and identify outliers in multivariate data. Most of the depth functions use a geometric approach, employing halfspaces, paraboloids, and projections to measure centrality from a global perspective. This results in an ordering of observations from the center outward.

The geometric tools used in the analysis of data depth offer nonparametric descriptions of a data set in a multidimensional space, making them quite useful for statistical inference problems (e.g. [7]), among which classification and regression (see e.g. [2, 3]), for learning theory (see e.g. [5]), outliers or anomaly detection (e.g. [8]) and multivariate risk analysis.

Our motivation lies in exploring asymptotics, specifically the examination of the behavior of depth-based multivariate quantiles as they approach extreme regions, both in terms of population measures and empirical data. Furthermore, our objective is to comprehend the connection between the extreme behavior of a probability measure (whether it exhibits a light or heavy tail) and the corresponding depth measures associated with it.

We focus on two prominent measures of data depth: halfspace depth, as described by [10], and spatial depth, introduced by [1]. The selection of these specific geometric measures stems from the core objective of our research. The fundamental question of characterising the tail behaviour of a probability measure using these geometric measures inherently addresses whether they capture essential aspects of the underlying probability distribution. In fact, it was demonstrated by [4] that geometric quantiles uniquely identify the underlying probability measure. However, the same does not hold true for halfspace depth, as shown by [6]. On the other hand, [9] established that halfspace depth uniquely identifies measures with finite support (e.g., empirical measures). This uniqueness property (under constraint for halfspace depths) indicates a direct correspondence between these two geometric measures of our interest and the underlying probability measures. It is therefore natural to look for clearer connection between the extremal behaviours of these geometric measures and their underlying probability measures. This motivates our study.

Considering practical applications, the questions regarding asymptotics become even more critical when examining sample versions of these two geometric measures. This forms the essence of the paper: We establish convergence rates for the sample versions and investigate the extreme behavior of the geometric measures based on the nature of the underlying distribution

*S. Singha is with TIFR–CAM email: sibsankar@tifrbng.res.in. M. Kratz is with ESSEC, Paris, email: kratz@essec.edu. S. Vadlamani is with TIFR–CAM, email: sreekar@tifrbng.res.in

References

[1] P. Chaudhury, Multivariate location estimation using extension of R-estimates through U-statistics type approach, *Journal of the American Statistical Association*, 91: 862–872, 1996.

[2] A. Cuevas, M. Febrero and R. Fraiman, On the use of the bootstrap for estimating functions with functional data, *Computational Statistics & Data Analysis*, 51(2): 1063–1074, 2007.

[3] M. Hallin, D. Paindaveine and M. Siman, Multivariate quantiles and multiple-output regression quantiles: From L_1 optimization to halfspace depth, *The Annals of Statistics*, 38(2): 635–669, 2010.

[4] V. I. Koltchinskii, M–estimation, convexity and quantiles, *The Annals of Statistics*, 25(2): 435–477, 1997.

[5] V. I. Koltchinskii and D. Panchenko, Empirical margin distributions and bounding the generalization error of combined classifiers, *The Annals of Statistics*, 30(1): 1–50, 2002.

[6] S. Nagy, *Halfspace depth does not characterize probability distributions*, *Statistical Papers*, 62: 1135–1139, 2021.

[7] R. Serfling, *Depth functions in nonparametric multivariate inference*, *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, AMS, 2002

[8] G. Staerman, E. Adjakossa, P. Mozharovskyi, V. Hofer, J. Sen Gupta and S. Clémençon, Functional anomaly detection: A benchmark study, *International Journal of Data Science and Analytics*, 1–17, 2022.

[9] A. J. Struyf and P. J. Rousseeuw, Halfspace depth and regression depth characterize the empirical distribution, *Journal of Multivariate Analysis*, 69(1): 135–153, 1999.

[10] J. Tukey, Mathematics and picturing data, *Proceedings of the International Congress on Mathematics*, 2: 523–531, 1975.