Allgemeines Statistisches Archiv 88, 133–135
© Physica-Verlag 2004, ISSN 0002-6018

Introduction: The geometry of data

By Karl Mosler

In multivariate analysis, data often cannot be fitted by normal or, more general, elliptically symmetric distributions. Then the classical parametric methodology, which draws on normality or ellipticity, fails. Recently, nonparametric methods have been developed that exploit the geometrical structure of the data in an explicit way and make use of the analyst's geometric intuition. They include the description of multivariate data by trimmed regions and generalized box-plots, the classification of observations and, in particular, the identification of outliers by data driven notions of centrality, further, statistical tests based on data depths and multivariate notions of sign and rank, and nonparametric procedures to reduce dimension. Some of them employ special geometric notions, like zonoids and other convex bodies, and use tools from computational geometry, such as point-to-hyperplane duality and line sweep in dual arrangements. All these developments are bound to recent advances in computing.

Modern geometrical data analysis begins in the mid-seventies with Tukey (1975), who, focussing on the combinatorial geometry of data, puts forward the halfspace depth and its deepest point, the Tukey median, which is affine equivariant and robust. Barnett (1976) provides a first systematic treatment of multivariate order statistics. Further seminal papers are Oja (1983), introducing an affine equivariant median based on the minimization of simplicial volumes, and Brown and Hettmansperger (1987), using the Oja median to define multivariate notions of rank and (vector valued) quantiles. The subsequent literature divides into at least three streams which partially overlap: One vein of research, dealing with directed hyperplanes and exploiting geometric duality, uses multivariate ranks and vector valued quantiles to construct sign and rank tests. Another one considers convex bodies (simplices and, especially, zonotopes) that are generated by the data and investigates procedures based on the minimization of their volumes. A third vein develops special notions of data depth and uses them in diverse descriptive and inferential applications.

Given a data cloud or a probability distribution, a depth function measures how central a point is located in the cloud (or the distribution). The upper level sets of a depth function form a family of central regions by which the underlying distribution may be characterized. In the last years, these notions have been unified and put into a general context of theory and applications; see Liu *et al.* (1999), Zuo and Serfling (2000) and Mosler (2002). Current theoretical research on data depth focuses on continuity properties of the depth function and the depth central regions and, related to them, approximation and computation of depth functions and depth contours; further, on robustness and asymptotic behaviour and on equivariances other than affine equivariance; applications include quality control, outlier identification, discrimination, and classification.

This Special Issue collects six research articles which cover some of this research.

In the first, Gleb Koshevoy, Jyrki Möttönen and Hannu Oja deal with multivariate extensions of mean difference and mean deviation. They consider averages of volumes of simplices and use them to construct median type estimates and sign tests for one and several samples location problems. Different representations of their objective functions are obtained through duality and the notions are related to zonotopes and lift-zonotopes generated by the data. A number of applications illustrates the theory.

Classical tools of robust multivariate analysis are the minimum volume ellipsoid and the minimum covariance determinant (Rousseeuw, 1985). Each of them provides a robust estimate of a 'central part' of data. As an alternative robust estimator, Claudia Becker and Sebastian Paris Scholz present another convex body minimizer, the minimum volume zonoid estimator (MZE). They report preliminary results on the comparative behaviour of these three estimators. The data sets are rather small since no efficient algorithm is known so far to calculate the MZE.

The remaining four papers deal with different aspects of the theory and application of data depth functions. Firstly, Rainer Dyckerhoff considers the projection property, which is a key property of an affine invariant convex depth. The projection property says that the value of the depth amounts to the infimum of depths over all univariate projections. This property is very useful for the approximation and computation of the depth value. Also, by invoking this property, a multivariate depth can be constructed from a univariate one. Dyckerhoff investigates the projection property in detail and gives many examples.

Not only location and dispersion, but also the dependency of a distribution are reflected by data depth. In his paper, Mario Romanazzi measures dependency in two ways: Firstly, by comparing volumes of depth level sets under the given distribution with the volumes of the same level sets under an independent distribution with the given marginals. Secondly, by plotting the depth values for the given distribution against the depth values of the proper independent distribution. He presents results in parametric distributions and also provides a nonparametric approach.

The complement of a depth central region is an outlying region. By this, measures of outlyingness and depth functions are closely related. Yijun Zuo introduces an outlying function, which is the expectation of a weighted L^p-distance, and a related depth, called weighted L^p-depth. The set of deepest points is the weighted L^p-median. Zuo investigates the robustness of the depth and the median and demonstrates that the median has high breakdown point. It has also bounded influence curve if the weight function is properly chosen.

Finally, Regina Liu, Kesar Singh and Julie H. Teng apply data depth to the construction of control charts in quality control, in particular nonparametric multivariate moving average (MA) charts. Their main idea is to represent each observation by the rank which is induced by a proper depth function. The depth based MA charts are applied to monitor airline performance data and compared with conventional charts.

References

Barnett, V. (1976). The ordering of multivariate data (with discussion). *Journal of the Royal Statistical Society Series A* **139** 318–352.

Brown, B.M., Hettmansperger, T.P. (1987). Affine invariant rank methods in the bivariate location model. *Journal of the Royal Statistical Society Series B* **49** 301–310.

Liu, R., Parelius, J., Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference (with discussions). *Annals of Statistics* **27** 783–858.

Mosler, K. (2002). *Multivariate Dispersion, Central Regions and Depth: The Lift Zonoid Approach.* Springer, New York.

Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics and Probability Letters* **1** 327–332.

Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications, Vol. 8* (W. Grossmann, G. Pflug, I. Vincze, W.Wertz, eds.), 283–297. Reidel, Dordrecht.

Tukey, J.W. (1975). Mathematics and picturing data. In *Proceedings of the 1974 International Congress of Mathematicians, Vol. 2* (R.D. James, ed.) 523–531. Vancouver.

Zuo, Y., Serfling, R. (2000). General notions of statistical depth function. *Annals of Statistics* **28** 461–482.

Karl Mosler
Seminar für Wirtschafts- und Sozialstatistik
Universität zu Köln
50923 Köln