Higher Order Zonoids

Marc-Arthur Diaye (University Paris 1 Pantheon-Sorbonne)

Gleb Koshevoy (CEMI RAS and Poncelet Laboratory (IMU and CNRS))

1. Introduction

Big data is characterized by the use of large (and smart) data sets that are structured or not. Such data have the values that the users can extract from. From this standpoint, being able to visualize the data in a smart and understandable manner is a step toward creating a market value from the data. The issue is not new and there exist already several data-visualization related "software" like Tableau and more recently Tensor Flow. However data visualization is commonly analyzed through the traditional prism of finding some "sufficient" statistics that provide the main informations from the original data set. These "sufficient" statistics could be of many forms: graphs, statistical quantities,... We argue in this project that another direction could be to first reduce the original data set. More precisely we present a new approach to visualize, order and evaluate data and hierarchical structures on data. For this we generalize the notion of lift-zonoid, introduced in [2], to higher zonoids and establish important stability properties of lift-zonoids to higher zonoids.

2. Random set

A *random closed set* **S** is a map **S**: (O,F,P)→ C, where (O,F,P) is the probability space, C is the space of closed convex sets of d-dimensional vector space $\mathbf{R}^d$, and, for any compact set B of $\mathbf{R}^d$, the set of elements o of O, such that the intersection of **S**(o) and B is non-empty, belongs to F. A random convex closed set **S** is said to be *integrable*, if there exists an integrable random vector s such that s belongs to **S** a.s. This vector s is called an integrable *selection* of **S**. The expectation of a random closed set **S**, *E(**S**)*, is defined to be the convex closure of the expectations of all its integrable selections.

If x is a random vector in $\mathbf{R}^d$, then the *segment* $[0,x]$ with end-points being the origin and x is a random convex body. Then the expectation of such a segment *E([0,x])* is called the *zonoid* of x and is denoted by Z_x. The zonoid does not uniquely determine the distribution of x. It is possible to achieve the uniqueness by uplifting x into $\mathbf{R}^{d+1}$. For this, consider the segment $[0,(1,x)]$ in $\mathbf{R}^{d+1}$, and call $E([0,(1,x)])=\underline{Z}_x$ the *lift zonoid* of x.

The following properties of the lift zonoid are established in [2]:
(i). The lift zonoid uniquely determines the underlying distribution.
(ii). The Central Limit Theorem is valid for the lift zonoid.

For a random closed set **S**, the lift zonoid of **S**, denoted by $\underline{Z(S)}$, is defined to be the convex closure of the lift zonoids of all its integrable selections. In general, there exists a random set, such that the lift zonoid does not uniquely determine its distribution.

3. Higher order zonoid and data reduction

We consider the following subclass of convex closed sets.

Definition. A *k order zonoid is the lift-zonoid of a random set* **S**: $(O,F,P) \rightarrow Z_{k-1}$, where Z_{k-1} *is the space of k-1 order zonoids, and a 0 zonoids is a random vector.*

We establish the following characterization as generalization of [1] and [2].

Theorem. A *k order zonoid is uniquely determine the underlying random set.*

As an application for data analysis, one can regard a *k* order zonoid as follows: suppose at the ground level we have data vectors of $\mathbf{R}^d$ and a collection of distributions supported at these vectors, than at the level one we have lift zonoids of these distributions and a collection of measures on these lift zonoids, and so on, than at the level k-1 we have k-1 order zonoids and a distribution on them, and finally we have k order zonoid for the latter distribution. Then the theorem says that this top level k order zonoid uniquely determine the whole pyramid of distributions. Since k order zonoid is a convex body, we define an ordering on k zonoids as the set inclusion. This order generalizes second order stochastic dominance for the random vectors (0 zonoids). The volume of a k order zonoid generalized Gini measure for one-dimensional distributions.

References

[1] M.-A. Diaye, G. Koshevoy and I. Molchanov, Lift zonoid for random sets (in preparation)

[2] G. Koshevoy and K. Mosler, Lift zonoids, random convex hulls and the variability of random vectors, Bernoulli 4 (1998), 377-399.