ROBUST STATISTICS FOR STREAMING DATA IN HILBERT SPACES

HERVÉ CARDOT
Institut de Mathématiques de Bourgogne, Université Bourgogne Franche-Comté, Dijon, France
e-mail: herve.cardot@u-bourgogne.fr

Online algorithms based on recursive approaches do not require to store all the data in memory and are useful tools to deal with streaming data, as well as massive datasets. They are extremely fast and allow for automatic update when the data are observed sequentially.

In Hilbert spaces, the mean vector and the covariance operator (or covariance matrix for finite-dimensional spaces) are classical indicators of central location and (multivariate) dispersion that can be estimated sequentially. However, outlying data may be hard to detect automatically for high dimensional data and both the mean vector and the covariance matrix can be highly affected by a small proportion of outlying observations.

We present robust indicators of central position and multivariate dispersion based on the geometric median and the median covariation matrix which are relevant tools to perform robust center estimation and robust PCA for variables taking values in separable Hilbert spaces. Such indicators, which can be expressed as the solutions of convex optimization problems, can be efficiently estimated in a recursive and very fast way, thanks to averaged stochastic gradient algorithms (see [4]). Some consistency results are proven in [3], [2] and [1]. Numerical experiments on simulated and real high dimensional datasets confirm the effectiveness of such online estimation procedures in comparison to more classical robust techniques that are generally not designed to deal with large samples of high dimensional data.

References