Estimation of Conditional Distributions in Data Stream Using Depth Functions

Daniel Kosiorowski*

Cracow University of Economics, Cracow, Poland daniel.kosiorowski@uek.krakow.pl

With an appearance of huge data sets in the Economics several new challenges emerged for statisticians and econometricians. Nowadays statistical procedures dedicated for filtering, monitoring, predicting and mining an useful information from the data should fulfill high criteria not only in classical terms of consistency, unbiasedness, effectiveness etc. but also should be computationally feasible. Their complexity should not be bigger than $O(n^{2/3})$ and they should be at least moderate robust to outliers, inliers and missing data.

Data stream as a concept can be informally defined as *high-speed generated instances of data that challenge our computational systems to store, process, and reason about*. Data stream analysis differs from the classical econometric analysis. In case of classical stochastic process analysis we assume a fixed interval of time, say $[0,T]$. All our calculations concern this interval, so we infer on base of information consisted in this interval. In case of the data stream analysis we do not fix any interval. Each consecutive while denotes a new stochastic process analysis. The terminology originates from the theoretical Informatics, where the data streams were considered at first. We can indicate several specific features of the economic data stream analysis: **1.** Data are generated by a process exhibiting a nonlinear structure of dependence between the observations. **2.** Data streams usually exhibit several regimes. **3.** Data stream analysis is performed on base of a constantly updated sample – on base of a sliding window or windows (the windows may differ with respect to their length or probing frequency for purposes related to different time scales). **4.** The streams usually consist of a huge amount of multivariate observations containing outliers, which is not stored in computer memory and has to be processed online. **5.** A signal carried by the stream is observed at irregularly spaced time points. By the ***signal*** we mean a ***relation between numerical characteristics of the stream.***

A common application of conditional distribution estimation in the data stream analysis involves constructing prediction interval for the next observation in the stream. Motivated by this problem we suggest two new methods for robust estimation of conditional distribution function. Our proposals appeal to so called data depth concept – a very promising approach of the multivariate statistics involving multivariate generalizations of one dimensional techniques based on order statistics, ranks and measures of outlyingness.

Our first proposal originates from the idea of adjusted kernel estimator of the conditional density function. We adjust the kernel using statistical depth function. Our second method takes a form of the k- nearest neighbors density estimator where closeness (neighbors) are measured by means of the depth function. We show that our proposals have comparable statistical properties to estimators proposed in the literature in cases of one-dimensional data streams, but outperform them in cases of multidimensional data stream. The proposals are robust. We underline however several conceptual difficulties related to an understanding of the robustness of the nonparametric density estimator. We present several applications of our proposals in cases of real economic data streams.

Key Words: Data Stream, Conditional Distribution, Depth Function, Robust Estimator