Nonparametric estimation of heavy-tailed density by dependent and independent data of WWW-traffic

MARKOVICH, Natalia *Institute of Control Sciences Russian Academy of Sciences, Russia*,
markovic@ipu.rssi.ru

The analysis of multivariate data begins usually with the estimation of marginal distributions. WWW-traffic contains a wide range of data: independent and heavy-tailed distributed data (file sizes, sizes and durations of sub-sessions etc), dependent (or long-range dependent) and heavy-tailed distributed data (video conference data and packet counts per unit time in Ethernet traffic). Hence, one has to estimate marginal distributions both by independent and dependent random variables.

Nonparametric estimation of the probability density function is considered. For independent identically distributed (i.i.d.) data three approaches are considered. These are combined parametric-nonparametric method, kernel estimators with variable and non-variable bandwidths and re-transformed estimators. The latter estimators can be found, for example, in [1]-[4].

For dependent data the kernel estimator with non-variable bandwidth is considered. It is mentioned, that the bias of this estimate is the same as for i.i.d. data, but the variance is larger. Important is that the variance consists on the variance of kernel estimator based on independent observations (the first term) and the term reflecting the dependence structure of the data (the second term) [5]. The variance of all kernel estimates of independent data has the order $\sim 1/(nh)$, where n is the sample size and h is the bandwidth of the kernel estimator. This well known standard result cannot be improved.

At the same time, the second term may provide the kernel estimate with a large variance if the data are long-range dependent. We observe such data in Internet. Indeed, the correlation structure of the data cannot be changed. However, the bandwidth h can make the second term less. At the same time, it were not correctly to reduce only this part of the variance. It is shown, how to select such a h that reduces a mean squared error of the kernel estimate.

Since dependent data may be heavy-tailed it is reasonable to improve the behavior of the kernel estimator with non-variable bandwidth at infinity (i.e. the tail domain) by means of the preliminary transformation of the data. The accuracy of such re-transformed kernel estimates is not worse than that for a standard kernel estimate considered before. One can use the same transformations as in the i.i.d. case [2].

The exposition is accompanied with examples motivated by application to WWW-traffic measurements.

References

[1] Hall, P. (1992) *On global properties of variable bandwidth density estimators*, Annals of Statistics. 20, 2, 762–778.

[2] Maiboroda, R.E., Markovich, N.M. (2004). Estimation of heavy-tailed probability density function with application to Web data. *Computational Statistics*, 19, 569–592.

[3] Markovitch, N.M., Krieger, U.R. (2002a). The estimation of heavy-tailed probability density functions, their mixtures and quantiles. *Computer Networks*, 40,(3), 459–474.

[4] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, New York.

[5] Wand, M.P., Jones, M.C. (1995). *Kernel smoothing*. Chapman & Hall, New York.