Depth based methods for estimation a conditional distribution in data streams

Daniel Kosiorowski[1,2]
[1] Department of Statistics, Cracow University of Economics, Cracow, Poland
[2] Corresponding author: Daniel Kosiorowski, e-mail: dkosioro@uek.krakow.pl

Abstracts

Estimation of a conditional distribution is a very important part for a variety of statistical procedures used in the modern Economics. This estimation is especially difficult in case of an economic data stream, i.e., when data are generated by the multidimensional non-stationary process of unknown form which may contain outliers. In this paper we propose an approach for a dynamic and robust estimation of the conditional distribution in the data streams. Our depth based proposals appeal to the ideas of adjusted Nadaraya-Watson estimator proposed by Hall et al. (1999) and depth induced k-nearest neighbors rule described by Paindavaine & Van Bever (2012). We show very promising statistical properties of our proposal in cases of selected nonlinear data streams generated using a general CHARME setting.

Keywords: data stream; robust procedure; depth function

1. Introduction

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 (see Huber, 2011).

Data stream as a concept is defined as *high-speed generated instances of data that challenge our computational systems to store, process, and reason about* (Gaber, 2012). Data stream analysis could be described as a new form of online data analysis that has challenged our computational capabilities. The streams are generated by smart phones, WIFI connected sensors, small computational devices, industrial robots, financial markets, shopping centers, the Internet. The 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 – see Fan & Yao (2005) or Franses & Van Dijk (2000). 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 (see Muthukrishan, 2006). In the Economics, we use by default stochastic methodological framework appealing to a nonlinear time series theory and generally consider different research tasks than in the Informatics. 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*** (e.g. between its mean value and its volatility or skewness) rather than a result of removal a noise from the stream (as in engineering).

Generally speaking existing approaches to the data stream analysis aim at

computationally feasible dynamic reduction of a dimension of the data which provides sufficient information for a decision maker. The approaches originating from the Informatics involve *Two phase techniques* (an online summary of data using microclusters), *Hoeffding bound-based techniques* (very fast machine learning), *Symbolic ApproXimation*, *Granularity – Based Techniques*. The approaches very often do not assume any probabilistic model for the data and if so they assume i.i.d. framework what is not appropriate for the Economics – for an overview see Gaber (2012) and references therein. In the context of the economic stream analysis we focus our attention on a general scheme to multi regime time series modeling presented by Stockis et al. (2010) called CHARME (*Conditional Heteroscedastic Autoregressive Mixture of Experts*). CHARME is an useful framework to modeling time series with switching regimes and includes as special cases many linear and nonlinear time-series, e.g. autoregressive, self-exciting threshold (SETAR), GARCH or SV models (see Franses & Van Dijk, 2000). CHARME allows for multivariate generalizations in an easy way. Within the CHARME setting – *to read a signal carried by the stream means to detect which of the regimes of the model generates the data.*

Let $\mathbf{X}_1 = (X_{11},...,X_{1d})$, $\mathbf{X}_2 = (X_{21},...,X_{2d})$,..., denote *d*-dimensional data stream $d \geq 1$. In the CHARME model a hidden Markov chain $\{Q_t\}$ in a finite set of states $\{1,2,...,K\}$ drives the dynamics of $\{\mathbf{X}_t\}$, and the model is defined:

$$\mathbf{X}_t = \sum_{k=1}^{K} S_{tk}(m_k(\mathbf{X}_{t-1},...,\mathbf{X}_{t-p}) + \sigma_k(\mathbf{X}_{t-1},...,\mathbf{X}_{t-p})\epsilon_t) + b_t\mathbf{\Theta}_t, \qquad (1)$$

with $S_{tk} = 1$ for $Q_t = k$ and $S_{tk} = 0$ otherwise, m_k, σ_k, $k = 1,...,K$, are unknown functions, ϵ_t are i.i.d. random variables with mean zero, the term $b_t\mathbf{\Theta}_t$ is the outlier component, b_t is an unobservable binary random variable indicating the occurrence of an outlier at time t, with $\mathbf{\Theta}_t$ being the associated (random) outlier.

We assume, Q_t changes its value only rarely, i.e., the observed process follows the same regime for a relative long time before the change of the regime occurs.

A window $\mathbf{W}_{i,n}$ denotes the sequence of points ending at $\mathbf{x}_i$ of size n, i.e., $\mathbf{W}_{i,n} = (\mathbf{x}_{i-n+1},...,\mathbf{x}_i)$. In the data stream analysis we consider a decision process basing on the statistics calculated from a moving window or windows from the stream.

PROBLEM TO SOLVE: Observing a one-dimensional stream $X_1, X_2,...$, we estimate a conditional distribution of the X_{i+1}, conditioned on the window $\mathbf{W}_{i,n}$, $i = 1,2,...$, i.e., $P(X_{i+1} \in A \mid \mathbf{W}_{i,n} = \mathbf{x})$, $A \subset \mathbb{R}$, in consecutive whiles $i = 1,2,...$.

In order to solve the above problem we focus our attention on the adjusted Nadaraya–Watson estimator of the conditional distribution proposed in by Hall et al. (1999). Our main aim is to "robustify" their approach using appropriate chosen weights induced by statistical depth function (see Serfling, 2006). However we should notice here several conceptual difficulties concerning understanding of a robustness of a nonparametric density estimator. If data are generated by a mixture of distributions then kernel estimator or k- nearest neighbor estimator tend to describe all parts of the mixture what could be treated as its advantage or its disadvantage depending on a point of view. As a breakdown of the density estimator we can take its unacceptable bias or variability in a fixed point or use certain global measure such as integrated mean squared error. In this

paper we evaluate robustness of a density estimator in terms of the *Hellinger* distance between the estimated density and a reference density for the most central part of the distribution support i.e. a part covering say 80% of the probability mass. For the centrality measurement we use statistical depth function.

2. Data depth concept

Statistical depth functions allow to measure centrality of any $\mathbf{x} \in \mathbb{R}^d$ with respect to (w.r.t.) a probability measure P over $\mathbb{R}^d$ or w.r.t. an empirical measure $P^{(n)}$ calculated basing on a sample $\mathbf{X}^n = \{\mathbf{x}_1,...,\mathbf{x}_n\}$. The larger the depth of $\mathbf{x}$, the more central $\mathbf{x}$ is w.r.t. to P. Liu's or *simplicial depth* is an example of the dept function:

$$D(\mathbf{x}, P) = P(\mathbf{x} \in S[\mathbf{X}_1, \dots, \mathbf{X}_{d+1}]), \quad (2)$$

where S is a simplex with vertices $\mathbf{X}_1,...,\mathbf{X}_{d+1}$ being i.i.d. P generated. If d-variate observations $\mathbf{X}_1,...,\mathbf{X}_n$ are available then sample versions of the depth are simply obtained by replacing P with the corresponding empirical distribution $P^{(n)}$.

For any statistical depth function and any $\alpha > 0$, the set $D_\alpha(P) = \{\mathbf{x} \in \mathbb{R}^d : D(\mathbf{x}, P) \geq \alpha\}$ is called the *depth region of order* α and represents points which *centrality measure* is not smaller than α. A sample depth provides a center outward ordering of the observations w.r.t. to the corresponding deepest point $\hat{\mathbf{m}}^{(n)}$. We can order a sample in such a way that $D(\mathbf{X}_{(1)}, P^{(n)}) \geq ... \geq D(\mathbf{X}_{(n)}, P^{(n)})$. Neglecting possible ties, it means that, in the depth sense, $\mathbf{X}_{(1)}$ is the nearest neighbor to $\hat{\mathbf{m}}^{(n)}$, $\mathbf{X}_{(2)}$ the second nearest closest etc. Statistical depth functions can be used to define neighbors of the deepest point $\hat{\mathbf{m}}^{(n)}$. However, in the context of well known *k-nearest neighbors rule* we require defining neighbors of any point $\mathbf{x} \in \mathbb{R}^d$. Following ideas of Paindavaine & Bever (2012) we can obtain this by *symmetrization* with respect to $\mathbf{x}$, i.e., we consider depth of a point $\mathbf{x}$ with respect to their empirical distribution $P_{\mathbf{x}}^{(n)}$ associated with the sample obtained by adding to the original observations $\mathbf{X}_1,...,\mathbf{X}_n$ their reflections $2\mathbf{x} - \mathbf{X}_1$, …, $2\mathbf{x} - \mathbf{X}_n$ w.r.t. $\mathbf{x}$. Further for our purposes related to the proposal 1 it is useful to consider a rescaled version of the sample depth

$$\widetilde{D}(\mathbf{x}, \mathbf{X}^n) = D(\mathbf{x}, \mathbf{X}^n) / \textstyle\sum_{i=1}^n D(\mathbf{x_i}, \mathbf{X}^n). \quad (3)$$

3. Proposals of robust estimation of the conditional distribution

Let $(Y, \mathbf{X})$ with $y \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^d$ be a random vector with joint density $f(y, \mathbf{x})$ and $f_X(\mathbf{x})$ be the marginal density of $\mathbf{X}$, then the **conditional density** $g(Y \mid \mathbf{X} = \mathbf{x}) = f(y, \mathbf{x}) / f_X(\mathbf{x})$, can be estimated by inserting a kernel density or k-nearest neighbors density estimator in the nominator and denominator of $g(y \mid \mathbf{x})$.

PROPOSAL 1: Following Hall et al. (1999), let $W_{j-N,n} = \{x_{j-N-n},...,x_{j-N}\}$,…, $W_{j,n} = \{x_{j-n},...,x_j\}$ be N windows from the stream each of length n, $j = l,...$, $k, N \in \mathbb{N}$, $N \gg k$ and let g be a reference density. In a process of estimating the conditional distribution of X_j determined by f_j, given the past $(X_{j-1},...,X_{j-k})$, $k = 2,3$, in a while $j = l,...$, we propose to calculate

$$\tilde{f}_j(y \mid (X_{j-1},...,X_{j-k}) = \mathbf{x}) = \\ = \frac{h_1^{-1}\sum_{i=1}^{N} K_{h_1}^1(y_i^j - y)\tilde{D}\left((y,\mathbf{x}),(Y_j^N,\mathbf{X}_j^N)\right)K_h(\mathbf{x}_i^j - \mathbf{x})}{\sum_{i=1}^{N}\tilde{D}\left(\mathbf{x},\mathbf{X}_j^N\right)K_h(\mathbf{x}_i^j - \mathbf{x})}, \quad (4)$$

where $\mathbf{X}_j^N = \quad = \{(x_{j-k-N},...,x_{j-1-N}),...,(x_{j-k-1},...,x_{j-1})\} \equiv \{\mathbf{x}_1^j,...,\mathbf{x}_N^j\}$, $\{y_1^j,...,y_N^j\} \equiv$ $Y_j^N = \{x_{j-N},...,x_{j-1},x_j\}$, $\tilde{f}_j$ is the adjusted kernel density estimate of f_j and $K_{(\cdot)}$ is univariate or multivariate kernel, $\mathbf{K}_h(\cdot) = h^{-1}\mathbf{K}(\cdot/h)$, $\tilde{D}(\cdot,\cdot)$ is the adjusted sample depth (e.g., simplicial depth for small and moderate data sets or approximate projection depth w.r.t. a reference sample for big and huge data sets – see Zuo & Shao, 2012).

PROPOSAL 2: With the same notation and framework as in the proposal 1 and following ideas in Paindavaine & Van Bever (2012) to estimate **the conditional distribution** of $X_j \equiv Y_j$ determined by f_j, given the past $(X_{j-1},...,X_{j-k})$, $k = 2,3$, in a while $j = 1,...$, calculate the depth induced k-nearest neighbor estimate of $\tilde{g}^j(Y \mid \mathbf{X} = \mathbf{x}) = f^j(y,\mathbf{x}) / f_X^j(\mathbf{x})$, by inserting into nominator and denominator depth induced k-nearest neighbor density estimates calculated basing on N windows from the stream each of length n, $W_{j-N,n} = \{x_{j-N-n},...,x_{j-N}\},\ldots,W_{j,n} = \{x_{j-n},...,x_j\}$,

$$f_n(\mathbf{z}) = k \Big/ P^{(n)}\left(S_{\mathbf{z},\|\mathbf{z}-\mathbf{x}_{(k)}(\mathbf{z})\|}\right), \quad (5)$$

where $\mathbf{x}_{(k)}(\mathbf{z})$ is k-th depth nearest neighbor of $\mathbf{z}$ within enhanced sample $\mathbf{x}_1,...,\mathbf{x}_n, 2\mathbf{z}-\mathbf{x}_1,...,2\mathbf{z}-\mathbf{x}_n$, $P^{(n)}$ is the empirical measure calculated from windows $W_{j-N,n},\ldots,W_{j,n}$, and $S_{\mathbf{x},\varepsilon}$ is a closed ball with a center in $\mathbf{z}$ and with a radius $\varepsilon > 0$.

It is well known (see Tsybakov, 2012) that a crucial issue in the kernel density estimation is a correct choice of a bandwidth h in (4). For the first proposal, in order to choose the bandwidths h we used at the beginning a variant of cross-validation proposed by Hall et al. (2004) and applied to the most central points in the window w.r.t. the reference sample, e.g., $\{y \in Y_j^N : D(y,Y^g) \geq \alpha\}$, where Y^g denote the reference sample. But due to the computational complexity of procedure we decided to use a „*dynamic" rule of thumb*" $h_{opt}^i = MAD\left\{W_{i,n}\right\} \cdot n^{-1/4}$, where MAD denotes the median absolute deviation $i = 1,...$. For the second proposal we started with the sample simplicial depth calculated for each consecutive window, but due to its computational complexity we decided to use approximate projection depth calculated from the reference samples and related to the regimes of the considered CHARME model.

4. Properties of the proposals – Monte Carlo studies

In order to check small sample performance of the proposals we 500 times generated samples of 5000 obs. from several data stream models. We considered windows of a fixed length of 100–500 obs. and samples without and with up to 15% of the additive outliers (AO). Our simulations involved among other two SETAR models (for details see Franses & Van Dijk, 2000) defined by

$$X_{t+1} = \begin{cases} 1+0.9X_t + \varepsilon_{t+1} & X_{t-1} \le 3 \\ 5-0.9X_t + \varepsilon_{t+1} & X_{t-1} > 3 \end{cases}, \quad Y_{t+1} = \begin{cases} 1+0.9Y_t + \varepsilon_{t+1} & Y_{t-1} \le 3 \\ 10-0.9Y_t + \varepsilon_{t+1} & Y_{t-1} > 3 \end{cases},$$

where $\varepsilon_t \sim$ were i.i.d. Student t(3) distribution generated.

We considered several CHARME schemes consisted of two AR(1)-GARCH(1,1) models $X_t = 5 + 0.1X_{t-1} + \varepsilon_t$, $\varepsilon_t = \sigma_t Z_t$ $\sigma^2 = 1 + 0.1\sigma_{t-1}^2 + 0.75X_{t-1}^2$, where $Z_t \sim$ N(0,1), skewed T(4) Student distr., skewed normal distr., skewed GED distribution.

Fig. 1: Functional boxplot for cond. density est. AR(1)-GARCH(1,1) + 5% additive outliers – proposal 1.

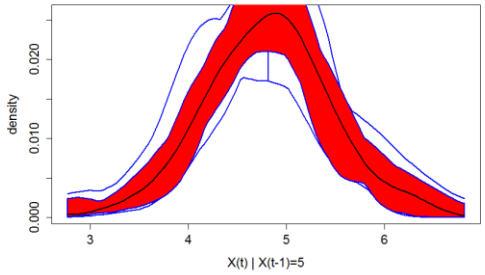


Fig. 2: Pointwise medians and MAD's for cond. density est. AR(1)-GARCH(1,1) + 5% additive outliers – proposal 1.

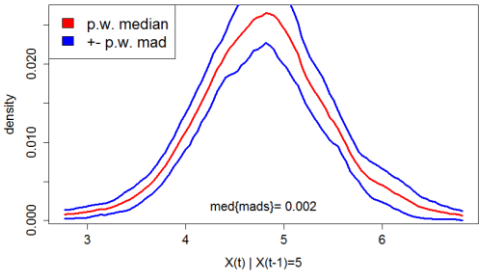


Fig. 3: Functional boxplot for cond. density est. AR(1)-GARCH(1,1) + 15% additive outliers – proposal 1.

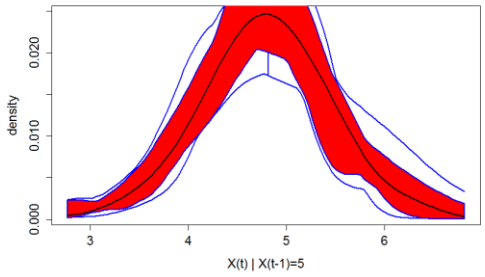
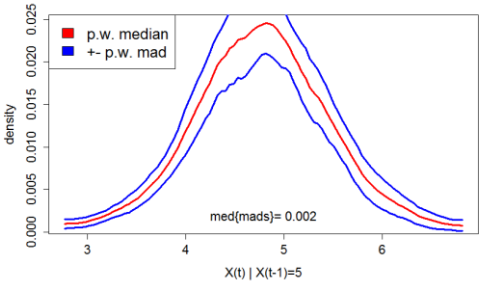


Fig. 4: Pointwise medians and MAD's for cond. density est. AR(1)-GARCH(1,1) + 15% additive outliers – proposal 1.



Results of the simulations were quite promising. Figures 1–4 summarize the results for proposal 1 and 500 density estimates from the streams generated by CHARME consisted of two AR(1)-GARCH(1,1) with up to 5% and 10% of additive outliers. The estimates were proportional to the reference density – so we obtained an argument for robustness of our proposal. Our second proposal turned out to be more locally robust however for small number of neighbors the estimator behaved unstable. For number of neighbors close to one third of the window length we obtained good results. Generally

we conclude that both proposals are robust to minor and sensitive to major changes of the stream. However obtaining correct tuning constants issue needs further studies.

5. Conclusions and future studies

Most of the robust and nonparametric multivariate statistical procedures are computationally very intensive and has to cope with so called "*curse of dimensionality*" (i.e., sparsity of the data in many dimensions). If the complexity of a procedure exceeds $O(n^{3/2})$, then the procedure is too complex for the analysis of the huge data sets and in particular for monitoring the economic data stream. We actually intensively study issues related to overcoming these substantial difficulties.

REFERENCES

1. Aggerwal Ch. C., (2007), Data Streams – Models and Algorithms, Springer, New York,

2. Kosiorowski, D., Bocian, M., Węgrzynkiewicz, A., Zawadzki, Z. (2012), Depth Procedures, R package {depthproc}, https://r-forge.r-project.org/projects/depthproc/.

3. Fan, J. Yao, Q., Nonlinear Time Series, Springer, New York, 2005.

4. Franses P. H., Van Dijk, D., (2000) Non-linear Time Series Models in Empirical Finance, Cambridge: Cambridge University Press.

5. Gaber, M., M. (2012), Advances in Data Stream Mining, *WIREs DMKD*, 2, 79–85.

6. Hall, P., Rodney, C. L. and Yao, Q., Methods for Estimating a Conditional Distribution Function. *Journal of the American Statistical Association*, vol. 94, 1999, 154-163.

7. Hall, P., Racine, J., Li, Q, Cross-Validation and the Estimation of Conditional Probability Densities, *Journal of the American Statistical Association*, vol. 99, 1015-1026.

8. Huber, P. (2011). Data Analysis, Wiley, New York.

9. Kosiorowski, D., Student Depth in Robust Economic Data Stream Analysis, Colubi A. (Ed.) Proceedings COMPSTAT'2012, ISI/IASC, 2012, 437 – 449.

10. Kosiorowski, D., Snarska, M., (2013) Robust Monitoring of a Multivariate Data Stream, unpublished manuscript, https://r-forge.r-project.org/projects/depthproc/

11. Muthukrishan, S. (2006), Data Streams: Algorithms and Applications, Now Publishers.

12. Paindavaine D., Van Bever G. (2012) Nonparametrically Consistent Depth-Based Classifiers, Ecore Discussion Paper, 2012/36, Universite Libre de Bruxells.

13. Serfling, R. (2006), Depth Functions in Nonparametric Multivariate Inference, In: Liu R.Y., Serfling R., Souvaine D. L. (Eds.): Series in DMTCS, AMS, vol. 72, 1 - 15.

14. Stockis, J-P.,Franke, J., Kamgaing, J. T., On Geometric Ergodicity of CHARME Models, *Journal of the Time Series Analysis*, vol. 31, 2010, 141 – 152.

15. Szewczyk, W. (2010), Streaming Data (2010), *Wiley Interdisciplinary Rev.: CS*, vol. 3.

16. Tsybakov, A. B. (2010), Introduction to Nonparametric Estimation, Springer, New York.

17. Shao, W., Zuo,Y. (2012). Simulated Annealing for Higher Dimensional Projection Depth. Computational Statistics and Data Analysis, vol. 56, 4026–4036.