Robust Monitoring of a Conditional Distribution in Economic Data Streams Using Statistical Depth Functions

Daniel Kosiorowski
Department of Statistics
Cracow University of Economics
Cracow, Poland
e-mail: daniel.kosiorowski@uek.krakow.pl

***Abstract*—Estimation of a conditional distribution is a building block of 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 a novel approach for robust monitoring conditional and unconditional distributions in the data streams. Our proposals are based on the idea of adjusted Nadaraya-Watson estimator proposed in [6] and they appeal to the so called data depth concept. We show very promising statistical properties of our proposals in cases of selected linear and nonlinear data streams models.**

Keywords-data stream; robust procedure; depth function

I. INTRODUCTION

An economic data stream could be informally defined as *a random sequence of observations of an undetermined length* – see [21], [16]. We should notice that in case of stochastic process analysis, say $\{X_t\}$, 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 [4], [10], [15]. 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 Informatics, where the data streams were considered at first. 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 – see [3], [4], [16]. 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 a different time scales). 4. The streams usually consist of a huge amount of multivariate observations containing outliers, which is not stored in computer memory. 5. A signal carried by the stream is observed at irregularly spaced time points and has to be processed on-line. By the *signal* we mean a relation between numerical characteristics of the stream rather than a result of removal a noise from the stream. Let $\mathbf{x}_1, \mathbf{x}_2, ...$ be an observed economic data stream.

A window $\mathbf{W}_{i,n}$ denotes the sequence of points ending at $\mathbf{x}_i$ of size n, i.e., $\mathbf{W}_{i,n} = (\mathbf{x}_{i-n+1}, ..., \mathbf{x}_i)$. Many approaches to data stream analysis is based on a monitoring various distance measures between distributions estimated from two or more windows – see [1], [12], [13]. In this paper, we study certain aspects of robust monitoring of a *one-dimensional economic data stream* using a moving window of a fixed length. We consider the following problems:

PROBLEM 1: We monitor a one-dimensional stream $X_1, X_2, ...$, and our aim is to detect changes in unconditional distribution of the X_i, on base of the moving window $\mathbf{W}_{i,n}$, $i = 1, 2, ...$, i.e., changes of $P(X_i \in A)$, $A \subset \mathbb{R}$, $i = 1, 2, ...$.

PROBLEM 2: We monitor a one-dimensional stream $X_1, X_2, ...$, and our aim is to detect changes in a conditional distribution of the X_{i+1}, conditioned on the observed window $\mathbf{W}_{i,n}$, $i = 1, 2, ...$, i.e., changes of $P(X_{i+1} \in A \mid \mathbf{W}_{i,n} = \mathbf{x})$, $A \subset \mathbb{R}$, $i = 1, 2, ...$.

In order to solve the above problems we focus our attention on the adjusted Nadaraya–Watson estimator of the conditional distribution proposed in [6]. The authors assumed that data are available in the form of strictly stationary stochastic process $\{(Y_i, \mathbf{X}_i)\}$, where Y_i is a scalar and $\mathbf{X}_i$ is a d-dimensional vector. They proposed two estimators for estimating the conditional distribution function $F(y \mid \mathbf{x}) \equiv P(Y_i \le y \mid \mathbf{X}_i = \mathbf{x})$, local logistic method and adjusted Nadaraya-Watson estimator, which have better statistical properties then known local and/or nonparametric approaches. Their proposals however are not robust. In the economic time series context, $\mathbf{X}_i$ typically denotes a vector of lagged values of a phenomenon Y_i, in which case $F(\cdot \mid \mathbf{x})$ is the *predictive distribution* of Y_i, given $\mathbf{X}_i = \mathbf{x}$ representing the past. Let $p_i = p_i(\mathbf{x})$, for $1 \le i \le n$, denote weights (functions of the data $\mathbf{x}_1, ..., \mathbf{x}_n$ as well as of $\mathbf{x}$) with the property that each $p_i \ge 0$, $\sum_{i=1}^{n} p_i = 1$ and

$$\sum_{i=1}^{n} p_i(\mathbf{x})(\mathbf{X}_i - \mathbf{x}) K_h(\mathbf{X}_i - \mathbf{x}) = 0 . \qquad (1)$$

In a spirit of ideas presented in [6] we can define following estimators of the unconditional and conditional densities

 ISBN: 978-1-61208-268-4

$$\tilde{f}(x)=\frac{1}{n}\sum_{i=1}^{n}p_i(x)K_h(x_i-x) \quad , \tag{2}$$

$$\tilde{g}(y\mid \mathbf{x})=\frac{h_1^{-1}\sum_{i=1}^{n}K_{h_1}^{1}(y_i-y)\,p_i(y,\mathbf{x})\mathbf{K}_h(\mathbf{x}_i-\mathbf{x})}{\sum_{i=1}^{n}p_i(\mathbf{x})\mathbf{K}_h(\mathbf{x}_i-\mathbf{x})} \quad , \tag{3}$$

where K is a kernel function (e.g., Gaussian), $K_h(\cdot)=h^{-1}K(\cdot/h)$, K_h^1 is one-dimensional Kernel, $\mathbf{K}_h$ denotes d-dimensional kernel, $d\geq 2$ (e.g., a product kernel or a kernel based on a norm of $\mathbf{x}$), h denotes a bandwidth.

It is well known that a crucial issue concerning the kernel density estimation involves an appropriate choice of the bandwidth h (i.e., providing a balance between unbiasedness, dispersion and computational complexity). In this paper, we "*robustify*" the above approach by the choice of weights $p_i(\mathbf{x})$ using adjusted sample depth function.

II. ROBUST DATA STREAM ANALYSIS

We understand the robustness of our proposals in a spirit of an approach presented in [5]. According to the authors a crucial property of an estimator is that it takes different values for different sample realizations. If a continuum of sample realizations is possible and the estimator is continuous in the sample, we expect a continuum of possible values for the estimator. We can look for the fraction of contamination for which this property is lost. In particular, we look for the fraction of outliers such that the estimator, or more specifically the measure of badness, can take only a finite number of different values despite a continuum of possible uncontaminated sample realizations. The statistical procedures, statistical decision rules, considered in this paper are functions of the estimators calculated on base of a moving window from the stream. In our proposals we use a very promising methodological approach of the multivariate analysis called *data depth concept* – see [14], [19], [23].

A *data depth* is a way to measure the "depth" or "outlyingness" of a given point with respect to a multivariate data cloud or its underlying distribution. A data *depth function* provides an order of the multivariate observations on base of their departure from the center. This ordering enables us for quantifying many complex multivariate features of the underlying distribution, including location, quantiles, scale, skewness and kurtosis. There are a variety of statistical depth functions known in the literature and implemented in the statistical software – see [2]. For the technical convenience purposes (vanishing value of the depth outside the convex hull of a sample) we further use so called *simplicial or Liu depth* – see [14], [19].

Let $\mathbf{X}^n=\{\mathbf{X}_1,...,\mathbf{X}_n\}$ be a random sample from the distribution $G(\cdot)$ in $\mathbb{R}^d$, $d\geq 1$. Let $I(\cdot)$ be the indicator function, that is, $I(A)=1$ if A occurs and $I(A)=0$ otherwise. Given the sample $\mathbf{X}^n$, the *sample simplicial depth* of $\mathbf{x}\in\mathbb{R}^d$ is defined as

$$D(\mathbf{x},\mathbf{X}^n)=\binom{n}{d+1}^{-1}\sum_{(*)}I(\mathbf{x}\in s[\mathbf{X}_{i_1},...,\mathbf{X}_{i_{d+1}}]), \tag{4}$$

where $(*)$ runs over all possible subsets of $\mathbf{X}^n$ of size $d+1$, $s[\mathbf{X}_{i_1},...,\mathbf{X}_{i_{d+1}}]$ is closed simplex with vertices $\mathbf{X}_{i_1},...,\mathbf{X}_{i_{d+1}}$.

When the distribution G is known, then the simplicial depth of $\mathbf{x}$ with respect to G is defined as $D(\mathbf{x},G)=P_G\{\mathbf{x}\in s[\mathbf{X}_1,...,\mathbf{X}_{d+1}]\}$, where $\mathbf{X}_1,...,\mathbf{X}_{d+1}$ are $d+1$ random observations from G. This depth is affine invariant and $D(\mathbf{x},\mathbf{X}^n)$ converges uniformly and strongly to $D(\mathbf{x},G)$. The affine invariance ensures that our proposed inference methods are coordinate–free, and the convergence of $D(\mathbf{x},\mathbf{X}^n)$ to $D(\mathbf{x},G)$ allows us to approximate $D(\mathbf{x},G)$ by $D(\mathbf{x},\mathbf{X}^n)$ when G is unknown. For our purposes it is useful to consider a rescaled version of the sample depth

$$\tilde{D}(\mathbf{x},\mathbf{X}^n)=D(\mathbf{x},\mathbf{X}^n)\Big/\sum_{i=1}^{n}D(\mathbf{x}_i,\mathbf{X}^n). \tag{5}$$

III. PROPOSALS

There is a long tradition in applying nonparametric methods in time series analysis involving nonparametric regression, runs tests for randomness, permutation tests or certain rank tests. However *nonparametric and robust analysis of a non-stationary time series* still seems to be a great challenge for the statistical and econometrical community – see [4], [9]. Our proposals concern decision making process basing on the stream and they aimed at detecting changes in certain very important for the decision makers properties of the stream – its unconditional and conditional distributions.

PROPOSAL 1: Let $W_{j,n}=\{x_{j-n},...,x_j\}$, denotes a window from the stream of length n in a time point $j=l,...$, and let g denotes a certain *fixed reference density*. In order to monitor the unconditional distribution of the stream, determined by density f, monitor *Hellinger* distance

$$d_j(\tilde{f}_j,g),\ j=l,..., \tag{6}$$

where

$$\tilde{f}_j(x)=\frac{1}{n}\sum_{i=1}^{n}K_h(x_{ij}-x)\tilde{D}(x,W_{j,n}), \tag{7}$$

is the adjusted kernel density estimate and K is a kernel function, $K_h(\cdot)=h^{-1}K(\cdot/h)$, $\tilde{D}(x,W_{j,n})$ denote the adjusted sample depth (5) of x, $x_{ij}\in W_{j,n}$, $i=1,...,n$, $j=l,...$.

PROPOSAL 2: Let $W_{j-N,n}=\{x_{j-N-n},...,x_{j-N}\}$,..., $W_{j-1,n}=\{x_{j-n-1},...,x_{j-1}\}$, $W_{j,n}=\{x_{j-n},...,x_j\}$ denote N windows from the stream each of length n, $j=l,...$,

$k, N \in \mathbb{N}$, $N \gg k$ and let g denotes a fixed reference density. Let $Y_j^N = \{x_{j-N},...,x_{j-1},x_j\} \equiv \{y_1^j,...,y_N^j\}$, $\mathbf{X}_j^N = \{(x_{j-k-N},...,x_{j-1-N}),...,(x_{j-k-1},...,x_{j-1})\} \equiv \{\mathbf{x}_1^j,...,\mathbf{x}_N^j\}$.

In order to monitor the conditional distribution of X_j determined by density f_j, given the small section of the past such as $(X_{j-1},...,X_{j-k})$, $k = 2,3$, we propose to monitor the *Hellinger* distance between the densities

$$d_j(\tilde{f}_j, g), \quad j = l,..., \tag{8}$$

where

$$\tilde{f}_j(y \mid (X_{j-1},...,X_{j-k}) = \mathbf{x}) = \frac{h_1^{-1}\sum_{i=1}^{N} K_{h_1}^1(y_i^j - y)\tilde{D}\left((y,\mathbf{x}),(Y_j^N,\mathbf{X}_j^N)\right)K_h(\mathbf{x}_i^j - \mathbf{x})}{\sum_{i=1}^{N}\tilde{D}\left(\mathbf{x},\mathbf{X}_j^N\right)K_h(\mathbf{x}_i^j - \mathbf{x})}, \tag{9}$$

is the adjusted kernel density estimate of $\tilde{f}_j$ 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 (5).

For the both proposals, in order to choose the bandwidths h we presently use a variant of cross-validation from [7] applied to the most central points in the window with respect to the reference sample, e.g., $\{y \in Y_j^N : D(y, Y^g) \geq \alpha\}$, where Y^g denote the reference sample. In (7) and (9) we propose to use *adjusted simplicial depth*, however it is possible to make use of other "more smoothly trimming" depth function, e.g., *projection depth*. For the computational convenience purposes we propose to choose the well known *Hellinger* or *Kolmogorov* distance as the distance between the density estimate and the reference density. For "regular distributions", it seems to be sufficient to approximate this distance using usual pointwise distances between the densities in say 100 – 1000 points. In case of a complete lack of the knowledge about the stream model (we need the reference densities) we propose to estimate the densities first by means of the statistics (7) or (9), eventually decompose the output density by means of a non hierarchical clustering algorithm (e.g., k-trimmed means) and then simulate the reference samples by means of the well known *inverse distribution function* method.

IV. PROPERTIES OF THE PROPOSALS

Similarly as in [6] we compared the proposed statistics with various estimators of the unconditional $f(\cdot)$ and conditional density function $f(\cdot|\cdot)$ through several simulated models of the data streams, involving independent observations, nonlinear time series and time series models exhibiting several regimes including TAR with trend (*threshold autoregressive model*), and CHARME (*conditional autoregressive mixture of models*) – for details of the models see [4] and [18]. As benchmark estimators we used kernel density estimator with normal kernel, k- nearest neighbors' density estimator, and the adjusted Nadaraya-Watson estimator originally proposed in [6]. For each simulated sample, the performance of the estimators used in the proposals was evaluated in terms of the *mean absolute deviation error*, *integrated mean square error* and visually by means of *functional boxplot* (i.e., the estimated densities were the observations) – see [17]. For example, in order to investigate finite window properties of the proposals we 500 times generated samples, each of length 10000 observations, from the time series model CHARME consisted of two AR-GARCH sub-models or consisted of three AR or SV sub-models). We used moving window of a fixed length of 100 obs. We considered streams with and without up to 5% of the additive outliers – for details see [15]. The unconditional and conditional densities of the sub-models, which comprised on the used CHARME model, were closer or more distant from each other according to the *Hellinger* distance. One of the densities was treated as a *null hypothesis*; subsequent densities represented the *alternative hypotheses*. Next we calculated kernel density estimates of the proposed statistics (8) and (9) under null and alternative hypotheses. Significant differences of the distributions (e.g., location shifts) of the proposals under null and alternative hypotheses indicated their good discriminative properties – their usefulness in the monitoring of the economic data stream. The estimated distributions of the statistics were similar for different density families of submodels – what give us a hope for their universal consistency (i.e., distributions of the statistics are independent from the underlying distributions). Results of the simulations were quite promising especially in cases of the data streams containing outliers. However we had to cope with the crucial issues of appropriate and computationally feasible bandwidth choice and weights calculation. In the simulations we used the cross-validation approach from [7] applied to the most central points (for which sample simplicial depth function takes value higher than a certain prefixed threshold) and an approximate depth calculation algorithm implemented in [2]. We presently study a possibility of an application new promising approaches to approximate calculation of the sample depth proposed recently in [21] and [22]. We implement our ideas in [2].

Fig. 1 presents a part of the results of the simulation studies of small samples properties of our proposals. Fig. 1 presents the functional boxplot for 100 density estimates obtained on base of 100-obs. samples drawn from Student t distribution with 5 degree of freedom (left), and the functional boxplot for 100 conditional density estimates obtained on base of 100-obs. samples drawn from bivariate normal distribution with mean vector (0,0) and covariance matrix consisted of rows (10,3) and (3,2) – conditional density of the first coordinate under the condition that second coordinate equals 1 (right). Each of the samples consisted of up to 5% additive outliers. The estimates obtained by means our proposals were not affected by the

 ISBN: 978-1-61208-268-4

outliers – we therefore conclude that they are quite promising in the context of robust analysis of the economic stream. The proposals need further studies of the issues concerning the bandwidth choice and tuning the depth based weights adjusting the kernel estimates (7) and (9).

V. CONCLUSIONS

We proposed two depth based statistics for the robust monitoring of unconditional and conditional distributions of the data stream. Results obtained so far are quite promising in the context of robust analysis of the data stream. They are robust to outliers being sensitive to the major changes of the stream at the same time. 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). We actually intensively study the possibility to overcome these substantial difficulties.

ACKNOWLEDGMENT

The author thanks for financial support from Polish National Science Center grant UMO-2011/03/B/HS4/01138.

REFERENCES

[1] Aggerwal Ch. C. (ed.), Data Streams – Models and Algorithms, Springer, New York, 2007.
[2] Bocian, M. Kosiorowski, D., Węgrzynkiewicz, A., Zawadzki, Z. Depth Procerures R package {depthproc}, 2012, https://r-forge.r-project.org/projects/depthproc/ [retrieved: Feb. 2013]
[3] Donoho, D., High-dimensional Data Analysis: The Curses and Blessings of Dimensionality, Manuscript, 2000, http://www-stat.stanford.edu/~donoho/Lectures/AMS2000/Curses.pdf
[4] Fan, J. Yao, Q., Nonlinear Time Series: Nonparametric and Parametric Methods, Springer, New York, 2005.
[5] Genton M. G., Lucas A., Comprehensive Definitions of Breakdown Points for Independent and Dependent Observations, Journal of the Royal Statistical Society Series B, 2003, 65, 81 – 84.
[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, pp. 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, pp. 1015-1026.
[8] Hahsler, M., Dunhamr, H. M., EMM: Extensible Markov Model for Data Stream Clustering in R, Journal of Statistical Software, vol. 35, 2010, pp. 2 – 31.
[9] Härdle, W., Hautsch, N. and Overbeck, L. Applied Quantitative Finance, 2nd edition, Springer, Heidelberg, 2009.
[10] Jacod, J., Shiryaev, A.N., Limit Theorems for Stochastic Processes, Second ed., Springer-Verlag, New York, 2003.
[12] Kosiorowski, D., Student Depth in Robust Economic Data Stream Analysis, Colubi A. (Ed.) Proceedings COMPSTAT'2012, ISI/IASC, 2012, pp. 437 – 449.
[13] Kosiorowski, D., Snarska, M., Robust Monitoring of a Multivariate Data Stream, 2013, unpublished, https://r-forge.r-project.org/projects/depthproc/ [retrieved: Feb. 2013]
[14] Li, J., Liu, R. Y. New Nonparametric Tests of Multivariate Locations and Scales Using Data Depth. Statistical Science, vol. 19, 2004, pp. 686 – 696.
[15] Maronna, R. A., Martin, R. D., Yohai, V. J., Robust Statistics - Theory and Methods. Chichester: John Wiley & Sons Ltd., 2006.
[16] Muthukrishan, S., Data Streams: Algorithms and Applications, Now Publishers, 2006.
[17] Ramsay, J. O., Hooker, G., Graves, S., Functional Data Analysis with R and Matlab, New York, Springer, 2009.
[18] Shalizi C. R., Kontorovich, A., Almost None of the Theory of Stochastic Processes A Course on Random Processes, 2007, http://www.stat.cmu.edu/~cshalizi/almost-none/ [Feb. 2013]
[19] Serfling, R., Depth Functions in Nonparametric Multivariate Inference, In: Liu R.Y., Serfling R., Souvaine D. L. (Eds.): Series in Discrete Mathematics and Theoretical Computer Science, AMS, vol. 72, 2006, pp. 1 - 15.
[20] Stockis, J-P.,Franke, J., Kamgaing, J. T., On Geometric Ergodicity of CHARME Models, Journal of the Time Series Analysis, vol. 31, 2010, pp. 141 – 152.
[21] Szewczyk, W., Streaming Data, Wiley Interdisciplinary Rev.: Computational Statistics, vol. 3, 2010, [retrieved: Feb. 2013]
[22] Torti, F., Perrotta, D., Atkinson, A. C, Riani, M., Benchmark Testing of Algorithms for Very Robust Regression, Computational Statistics and Data Analysis, vol. 56, 2012, pp. 2501–2512.
[23] Shao, W., Zuo,Y. (2012). Simulated Annealing for Higher Dimensional Projection Depth. Computational Statistics and Data Analysis, vol. 56, 2012, pp. 4026–4036.

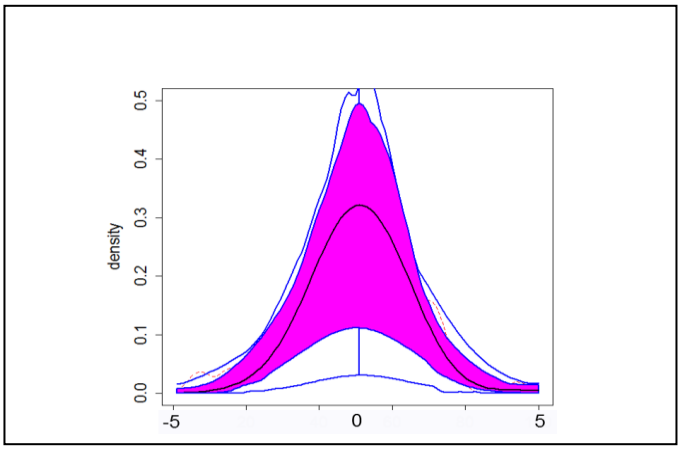

Figure 1. Functional boxplots for 100 density estimates obtained on base of 100-obs. samples drawn from Student t with 5 degree of freedom consisted of 5% outliers (left), and for 100 conditional density estimates obtained on base of 100-obs. samples drawn from bivariate normal distribution with mean vector (0,0) and covariance matrix consisted of rows (10,3) and (3,2) (right). Each sample consisted of 5% outliers. Our own calculations using {fda} R package.

 ISBN: 978-1-61208-268-4