World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:1, No:8, 2007

Exponentially Weighted Simultaneous Estimation of Several Quantiles

Valeriy Naumov, and Olli Martikainen

Abstract**—**In this paper we propose new method for simultaneous generating multiple quantiles corresponding to given probability levels from data streams and massive data sets. This method provides a basis for development of single-pass low-storage quantile estimation algorithms, which differ in complexity, storage requirement and accuracy. We demonstrate that such algorithms may perform well even for heavy-tailed data.

Keywords**—**Quantile estimation, data stream, heavy-tailed distribution, tail index.

I. Introduction

QUANTILE estimation in data streams or massive data sets arises for an increasing number of applications. Examples of such applications include knowledge discovery and data mining, query optimization for large databases, network routing and traffic analysis, fraud detection, stock market analysis and digital surveys in astronomy.

Large number of low-storage methods has been developed for arbitrary quantile estimation from massive static data sets. Survey of low-storage quantile estimations see in [1] and [2]. We only make mention of several methods closely related to the subject of our paper. Quantile estimate proposed in [3] is based on stochastic approximation and utilizes incremental update of the density at the last estimate of the quantile. The P^2 algorithm [4] approximates the empirical quantile function using parabolic interpolation. It was extended for simultaneous estimation of several quantiles in [5] and [6]. Recently proposed sequential scoring algorithm provides a very accurate quantile estimate even at the tail region [7]. It uses simple linear approximation of ECDF apart from its tails, which are approximated using exponential curves.

A data stream is a real-time, continuous, ordered by arrival time or by timestamp, sequence of items [8]. Applications that monitor a non-stationary data stream in real-time must react quickly to unusual data values. While quantile estimation for massive static data sets can be computed accurately, for data streams the quantile in effect at a given time cannot be known precisely. In this case estimators that adapts to changing data, like exponentially weighted moving average, would be preferable.

Exponential weighted stochastic approximation (EWSA) proposed in [9] is a modification of stochastic approximation [3] for the quantile estimation in streaming data, which is collected in batches. As with stochastic approximation, EWSA has problems when estimating values near the tail of a density. The poor tail performance may even take the EWSA estimate below the smallest possible observation or above the largest possible observation. While it can be used with batches including one observation, its performance decreases for smaller batch sizes.

In this paper we propose exponentially weighted quantile estimators for streaming data, which are inspired by the P^2 concept [4]. We use linear and parabolic approximations of the empirical quantile function, but for better performance the tails are approximated by exponential curves.

II. Quantile Estimation

Let $X_1, X_2, \ldots, X_n$ be a sample of random variables, and $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$ be the ordered sequence of the observations. The estimate of a q-quantile can be obtained from the ordered sequence of the observations as $X_{(\lceil nq \rceil)}$, where $\lceil nq \rceil$ is the smallest integer greater than or equal to nq, for $0<q<1$. However, as the number of observations becomes large, limitations on sorting time and storage size make these methods unrealistic.

The estimate of a q-quantile can also be found from the empirical cumulative distribution function (ECDF), which can be expressed in the following way

$$\overline{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le t)$$

where

$$I(X_i \le t) = \begin{cases} 1 & \text{if } X_i \le t, \\ 0 & \text{if } X_i > t. \end{cases}$$

This implies the recursive version of ECDF given by

$$\overline{F}_n(t) = \left(1 - \frac{1}{n}\right)\overline{F}_{n-1}(t) + \frac{1}{n} I(X_i \le t).$$

Exponentially weighted moving average of an estimate can be easily obtained from its recursive version [12]. Replacing $1/n$ on the right hand side with a fixed constant u, $0 < u < 1$, gives

V. Naumov is with Lappeenranta University of Technology, PO Box 20, FIN-53850 Lappeenranta, Finland (corresponding author, phone: 358-50-3299222, fax: 358-5-6212899, e-mail: valeriy.naumov@lut.fi).

O. Martikainen is with University of Oulu, PO Box 3000, FIN-90014 Oulu, Finland (e-mail: olli.martikainen@oulu.fi).

an exponentially weighted moving average of ECDF defined by

$$\hat{F}_n(t) = (1-u)\hat{F}_{n-1}(t) + uI(X_i \le t). \quad (1)$$

We use this estimator in our method for the exponentially weighted quantile estimation in large data sets and streaming data, which is inspired by the recursive version of the P^2 algorithm [4].

Proposed method comprises the initialization phase and then alternating the estimation phase with the interpolation phase. Suppose that m probability levels $0 < p_1 < p_2 < \ldots < p_m < 1$ are given. During the estimation phase ECDF is sequentially estimated at $m+2$ grid points $h_0 \le h_1 \le \ldots \le h_m \le h_{m+1}$ as shown in Figure 1. The estimation at internal grid points $h_1, h_2, \ldots, h_m$ uses formula (1), while the estimation at boundary grid points h_0 and h_{m+1} uses different procedure described later in this section. The method tries to keep the internal grid points close to the p_1 - quantile, p_2 - quantile, …, p_m - quantile. On this account the estimation phase ends when the value p_j^* of ECDF at an internal grid point h_j is sufficiently far of the probability level p_j. Then the position of the internal grid point h_j is adjusted during the interpolation phase. First p_j - quantile is estimated by interpolation of the empirical quantile function using its values $h_0, h_1, \ldots, h_{m+1}$ at points $p_0^*, p_1^*, \ldots, p_{m+1}^*$ as shown in Figure 2. After interpolation the current estimate of p_j - quantile becomes the grid point h_j.

Total number m of the probability levels employed in the method and its values depend of the number of estimated

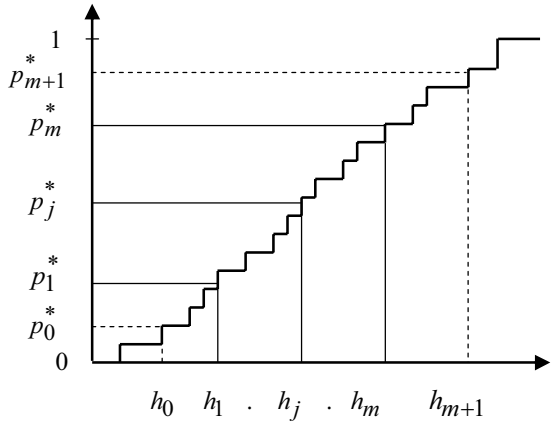

Fig. 1 The empirical cumulative distribution function and grid points

quantiles and can be selected in the same way as in the original P^2 algorithm [4]. For single q-quantile estimation, the P^2 algorithm uses one main probability level q and either two or four supplementary probability levels $q/2, (1+q)/2$ or $q/4, q/2, (1+q)/2$, $(3+q)/4$ respectively.

For simultaneous estimation of several quantiles $q_1 < q_2 < \ldots < q_k$, one may use either $m = k+2$ probability levels $q_1/2$, q_1, q_2, …, q_k, $(1+q_k)/2$ or $m = k+4$ probability levels $q_1/4, q_1/2, q_1, q_2, \ldots, q_k, (1+q_k)/2$, $(3+q_k)/4$. Raatikainen recommends in [5] and [6] to use P^2 algorithm with $m = 2k+1$ probability levels $q_1/2, q_1$, $(q_1+q_2)/2, q_2, \ldots, (q_{k-1}+q_k)/2, q_k, (1+q_k)/2$. In short, the more probability levels we use, the better the performance is, but time required for the calculations increases.

After selection of the probability levels, the quantile estimation begins by sorting the first $m+2$ observations. The internal grid points and corresponding values of ECDF then initialize to $h_j = X_{(j+1)}$ and $p_j^* = p_j$ respectively for $j = 1, \ldots, m$.

A. Internal Grid Points

The value of ECDF at each internal grid point is updated after every observation. As a new observation X comes in, the value of ECDF at the internal grid points is adjusted as

$$p_j^* := \begin{cases} (1-u)p_j^* + u & \text{if } X \le h_j, \\ (1-u)p_j^* & \text{if } X > h_j. \end{cases} \quad (2)$$

Ideally, each internal grid point h_i should be equal to the estimated p_i-quantile. If the value of ECDF at an internal grid point h_j is sufficiently far of the probability level p_j then the grid point is adjusted.

Adjustment of the internal grid points is controlled by small non-negative parameters $\delta_1^-, \delta_2^-, \ldots, \delta_m^-$ and $\delta_1^+, \delta_2^+, \ldots, \delta_m^+$ satisfying the following inequalities:

$$\delta_j^- < p_j - p_{j-1},\ \delta_j^+ < p_{j+1} - p_j,\ 1 \le j \le m,$$

where $p_0 = 0$ and $p_{m+1} = 1$. If the value of ECDF at an internal grid point h_j is below the probability level p_j by more than δ_j^- or it is above the probability level p_j by more than δ_j^+, then the grid point is adjusted using appropriate interpolation.

With the linear interpolation new value for the grid point h_j is given by

$$h_j := h_j + (h_{j+1} - h_j)\frac{p_j - p_j^*}{p_{j+1}^* - p_j^*}, \text{ if } p_j \ge p_j^*, \quad (3)$$

$$h_j := h_j + (h_j - h_{j-1})\frac{p_j - p_j^*}{p_j^* - p_{j-1}^*}, \text{ if } p_j \le p_j^*. \quad (4)$$

More accurate results can be obtained using the parabolic interpolation as shown in Figure 2. With parabolic interpolation new value for the grid point h_j is given by

$$h_j := h_j + \frac{p_j - p_j^*}{p_{j+1}^* - p_{j-1}^*} \cdot \left((p_j - p_{j-1}^*) \frac{h_{j+1} - h_j}{p_{j+1}^* - p_j^*} + (p_{j+1}^* - p_j) \frac{h_j - h_{j-1}}{p_j^* - p_{j-1}^*} \right) \quad (5)$$

All grid points must always be in a non-decreasing order. Therefore, if the parabolic interpolation predicts a value, which will make grid point h_j less than h_{j-1} or greater than h_{j+1}, then the parabolic prediction is ignored and a linear interpolation is used.

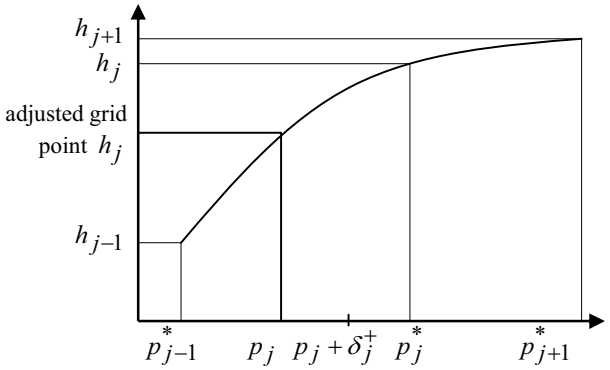

Fig. 2 Parabolic interpolation of the empirical quantile function

For simplicity, instead of $2m$ controlling parameters $\delta_1^-, \delta_2^-, \ldots, \delta_m^-$, $\delta_1^+, \delta_2^+, \ldots, \delta_m^+$, we can use a single parameter by setting all δ_j^- and δ_j^+ equal to a nonnegative parameter δ, such that $\delta < \min_{0 \le j \le m} (p_{j+1} - p_j)$. Another possible approach is to select σ, $0 \le \sigma < 1$, and use δ_j^- and δ_j^+ defined by

$\delta_j^- = \sigma(p_j - p_{j-1})$, $\delta_j^+ = \sigma(p_{j+1} - p_j)$, $1 \le j \le m$.

With zero value of δ or σ all internal grid points will be updated after every new observation.

B. Boundary Grid Points

For the described interpolation to perform we need to specify the boundary grid points h_0 and h_{m+1}, and the values of the empirical distributed function p_0^* and p_{m+1}^*, which are used for the adjustment of the leftmost and the rightmost internal grid points.

In simplest approach boundary grid points amount the minimum and the maximum of the observations so far, and corresponding values of the empirical distributed function are set to $p_0^* = 0$ and $p_{m+1}^* = 1$. During the initialization phase the boundary grid points are set to $h_0 = X_{(1)}$ and $h_{m+1} = X_{(m+2)}$. If a new observation is less than the current minimum h_0, then the observation becomes the minimum, and if an observation is greater than the current maximum h_{m+1}, then the observation becomes the maximum. With this approach the position of the left boundary grid point may only decrease while the position of the right boundary grid point may only increase. This is not suitable for non-stationary data streams and does not work well for heavy-tailed data.

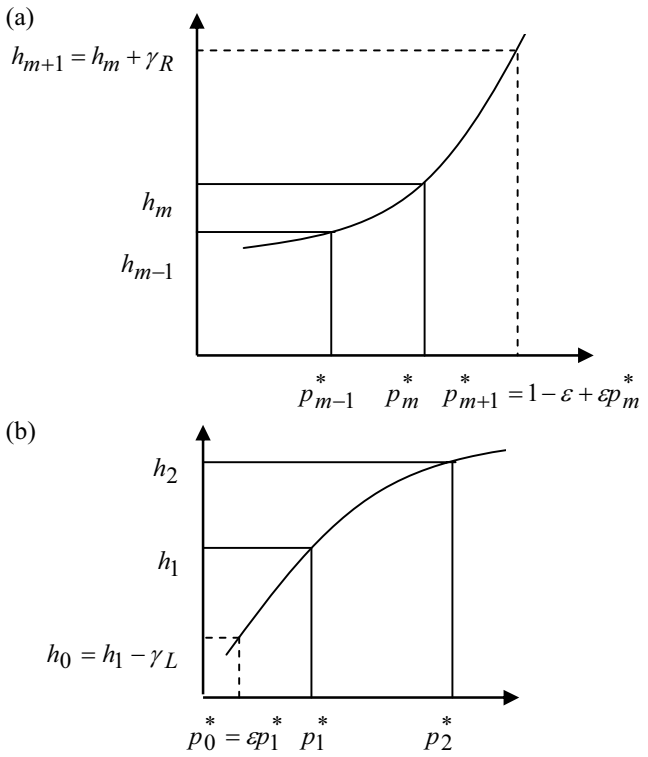

Fig. 3 Approximation of the right (a) and left (b) tails of the empirical quantile function, $\varepsilon = 1/e \approx 0.36788$

Instead of the placement of boundary grid points at the current minimum and maximum we propose another approach. In proposed method the left or right boundary grid points and corresponding values of the empirical distributed function are calculated only when there is a need in adjustment of the leftmost or the rightmost internal grid points respectively. The left boundary grid point $h_0 = h_1 - \gamma_L$ estimates the conditional expected observation below the threshold h_1, and the right boundary grid point $h_{m+1} = h_m + \gamma_R$ estimates the conditional expected observation over the threshold h_m. The values of the empirical distributed function at the boundary grid points are given by $p_0^* = \varepsilon\, p_1^*$ and $p_{m+1}^* = 1 - \varepsilon + \varepsilon\, p_m^*$, where $\varepsilon = 1/e \approx 0.36788$ as shown in Figure 3. These settings are

based on Breiman's approximation of ECDF by the exponential curves [13] given by

$$P(X \le t) \approx p_1^* \exp\left(\frac{t-h_1}{\gamma_L}\right),\ t < h_1, \quad (6)$$

$$P(X \le t) \approx 1-(1-p_m^*)\exp\left(\frac{h_m-t}{\gamma_R}\right),\ t > h_m. \quad (7)$$

C. Parameters γ_L and γ_R

Parameter γ_R estimates the conditional expectation of the excess $X-h_m$ given that $X > h_m$. It can be calculated using exponentially weighted moving average. But this estimator is unstable for heavy-tailed data and we propose its modification, which is based on the idea presented in [11]. It requires the knowledge of the right tail index, whose estimate we denote as ζ_R. Calculation of the estimators γ_R and ζ_R is controlled by parameters v, w, $0 < v < 1$, $0 < w < 1$, and $\kappa > 1$.

Let a new observation X be greater then the rightmost internal grid point h_m, and Y_R and Z_R be defined by

$$Y_R = X - h_m,\ Z_R = \frac{Y_R}{\kappa\gamma_R}.$$

For small value of Y_R, when $Z_R \le 1$, the parameter γ_R is adjusted using exponentially weighted moving average given by

$$\gamma_R := (1-w)\gamma_R + wY_R. \quad (8)$$

For large Y_R, when $Z_R > 1$, the adjustment of γ_R consists of two steps. First we calculate the value of $\chi_R = (1-v)\zeta_R + v\ln(Z_R)$. If it is smaller then one, we update the estimate ζ_R of the right tail index using the following version of Hill's estimator [10]:

$$\zeta_R := (1-v)\zeta_R + v\ln(Z_R). \quad (9)$$

Finally we adjust the parameter γ_R as

$$\gamma_R := (1-w)\gamma_R + w\frac{\kappa\gamma_R}{1-\zeta_R}. \quad (10)$$

Parameter γ_L estimates the conditional expectation of the shortfall $h_1 - X$ given that $X < h_1$. When a new observation X smaller then the leftmost internal grid point h_1 comes in, we calculate Y_L and Z_L as

$$Y_L = h_1 - X,\ Z_L = \frac{Y_L}{\kappa\gamma_L}.$$

If $Z_L \le 1$, the new value for γ_L is given by

$$\gamma_L := (1-w)\gamma_L + wY_L. \quad (11)$$

Otherwise, we first calculate the value of $(1-v)\xi_L + v\ln(Z_L)$, and if it is smaller then one, we update the estimate ζ_L of the left tail index as

$$\zeta_L := (1-v)\zeta_L + v\ln(Z_L). \quad (12)$$

Then we adjust parameter γ_L as

$$\gamma_L := (1-w)\gamma_L + w\frac{\kappa\gamma_L}{1-\zeta_L}. \quad (13)$$

In our experiments, during the initialization phase the estimators ζ_L and ζ_R are set to zero, and the estimators γ_L and γ_R are set to $X_{(2)} - X_{(1)}$ and $X_{(m+2)} - X_{(m+1)}$ respectively.

III. Simulation Results

In this section we present the results of our simulation study of two algorithms. The first algorithm (EWLF) utilizes linear interpolation of the empirical quantile function, while the second algorithm (EWPF) uses parabolic approximation.

We use generated samples with 10,000,000 observations for the standard normal, the chi-square with 1 degree of freedom, and two heavy-tailed distributions: the standard Cauchy and the Pareto distribution with tail index $\alpha = 1.2$. We simultaneously estimate $m = 15$ quantiles corresponding to eleven main probability levels 0.001, 0.01, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95, 0.99, 0.999 and four supplementary probability levels 0.00025, 0.0005, 0.9995 and 0.99975. We also consider performance of the EWLF algorithm with $m = 25$ probability levels, when additionally 10 supplementary probability levels are inserted between each two adjacent main probability levels. For each quantile being estimated we calculate the quantile estimate averaged over 100 runs and the mean square error with regards to the true quantile value (MSE). Results presented in tables below are obtained with the following values of control parameters: $\delta = u = w = 0.00001$, $v = 0.0001$, and $\kappa = 10$.

TABLE I
QUANTILE COMPARISON

P	True	EWLF, *m*=15 Avg. est. MSE	EWLF, *m*=25 Avg. est. MSE	EWPF, *m*=15 Avg. est. MSE
Normal				
0.001	-3.0902	-3.1544 6.75e-02	-3.1218 3.77e-02	-3.0092 8.39e-02
0.01	-2.3264	-2.5086 0.182	-2.4140 8.80e-02	-2.3045 2.42e-02
0.05	-1.6449	-1.7967 0 .152	-1.7173 7.27e-02	-1.6453 6.55e-03
0.10	-1.2816	-1.4079 0.126	-1.3478 6.63e-02	-1.2878 7.16e-03
0.25	-0.6745	-0.7613 8.69e-02	-0.7243 5.00e-02	-0.6815 6.67e-03
0.50	0	-0.00048 2.88e-03	-0.00064 2.80e-03	-0.00083 3.06e-03
0.75	0.6745	0.7611 8.67e-02	0 .72374 4.93e-02	0.6801 6.67e-03
0.90	1.2816	1.4097 0.128	1.3490 6.74e-02	1.2880 7.58e-03
0.95	1.6449	1.7997 0.155	1.7202 7.54e-02	1.6478 6.11e-03
0.99	2.3264	2.5145 0.188	2.4201 9.42e-02	2.3103 1.94e-02
0.999	3.0902	3.1548 6.67e-02	3.1240 3.71e-02	3.0203 7.20e-02
Chi-square				

0.001	0.000002	0.000004 2.95e-06	0.000003 1.72e-06	0.000002 5.69e-07
0.01	0.00016	0.00043 2.75e-04	0.00030 1.35e-04	0.00016 6.98e-06
0.05	0.00393	0.00747 3.55e-03	0.00560 1.68e-03	0.00404 1.44e-04
0.10	0.01579	0.02779 1.20e-02	0.02177 5.99e-03	0.01615 4.49e-04
0.25	0.1015	0.1607 5.92e-02	0 .1325 3.10e-02	0.1037 2.36e-03
0.50	0.4549	0.6387 0.184	0.5523 9.74e-02	0.4636 9.15e-03
0.75	1.3233	1.7207 0.397	1.5324 0.209	1.3395 1.75e-02
0.90	2.7055	3.2582 0.553	2.9736 0 .268	2.7109 1.43e-02
0.95	3.8415	4.5396 0.698	4.1532 0 .312	3.8169 3.22e-02
0.99	6.6349	7.6200 0.986	7.0791 0 .446	6.4727 0.17
0.999	10.828	11.059 0 .255	10.865 0 .124	10.173 0.66
Pareto				
0.001	1.000834	1.000775 8.55e-05	1.000783 7.97e-05	1.000830 5.92e-05
0.01	1.00841	1.00870 3.41e-04	1.00850 1.98e-04	1.00841 1.84e-04
0.05	1.0437	1.0471 3.44e-03	1.0448 1.21e-03	1.0437 4.67e-04
0.10	1.0918	1.1049 1.32e-02	1.0966 4.88e-03	1.0918 6.31e-04
0.25	1.2709	1.3495 7.86e-02	1.3031 3.22e-02	1.2709 1.44e-03
0.50	1.7818	2.1596 0.378	1.9408 0 .159	1.7819 3.64e-03
0.75	3.1748	4.8020 1.63	3.8365 0 .662	3.1762 1.32e-02
0.90	6.8129	12.1895 5.37	8.7965 1.98	6.8169 4.83e-02
0.95	12.139	26.765 14.6	17.210 5.07	12.146 0.151
0.99	46.416	136.49 90.3	83.800 37.4	46.618 2.10
0.999	316.23	600.06 293	525.78 217	317.09 26.9
Cauchy				
0.001	-318.31	-1857.7 1974	-1238.4 1279	-321.88 33.89
0.01	-31.821	-448.05 448	-166.43 139	-32.034 1.93
0.05	-6.3138	-64.959 59.8	-20.799 14.5	-6.3215 .104
0.10	-3.0777	-22.051 19.1	-7.7654 4.69	-3.0804 2.85e-02
0.25	-1.0000	-5.2893 4.30	-2.1657 1.17	-1.0008 7.82e-03
0.50	0	-0.00795 5.25e-02	-0.00038 5.50e-03	-0.00012 3.83e-03
0.75	1.0000	5.2258 4.24	2.1635 1.16	1.0001 7.88e-03
0.90	3.0777	21.625 18.5	7.7650 4.69	3.0781 2.54e-02
0.95	6.3138	62.798 57.4	20.786 14.5	6.3182 0.112
0.99	31.821	414.95 409	165.54 137	32.025 2.00
0.999	318.31	1601.9 1632	1184 1100	320.50 34.2

The EWLF algorithm provides satisfactory results for the normal and the chi-square distributions, but it does not perform well for heavy-tailed distributions. The EWPF algorithm accurately estimates arbitrary quantiles from all considered distributions. The performance of the algorithms can be further improved by increasing the number of probability levels. On the other hand, increasing control parameters u, v, w and δ we may achieve faster convergence at the price of the accuracy worsening.

IV. Conclusion

In this paper we propose new method for simultaneous estimation of several quantiles in massive data sets and streaming data. By selecting particular interpolation method for the empirical quantile function one can obtain different algorithms for quantile estimation. We demonstrate that the usage of the parabolic interpolation can provide acceptable accuracy of the quantile estimation in large stationary data sets. In our further experiments we plan to apply proposed method for the quantile estimation in non-stationary data sets and during these experiments find optimal values of the control parameters.

References

[1] C. Hurley and R. Modarres, "*Low-storage quantile estimation*," Computational Statistics, vol. 10, no. 4, 1995, pp. 311–325.
[2] A.M. Law and W.D. Kelton, "*Simulation Modeling and Analysis*," 3rd ed. New York:McGraw-Hill, 2000.
[3] L. Tierney, "A space-efficient recursive procedure for estimating a quantile of an unknown distribution," *SIAM J. on Scientific and Statistical Computing*, vol. 4, no. 4, 1983, pp. 706–711.
[4] R. Jain and I. Chlamtac, "The P^2 algorithm for dynamic calculation of quantiles and histograms without storing observations," *Communications of the ACM*, vol. 28, no. 10, 1985, pp. 1076–1085.
[5] K.E.E. Raatikainen, "Simultaneous estimation of several percentiles", *Simulation*, vol. 49, no. 4, 1987, pp. 159–164.
[6] K.E.E. Raatikainen, "Sequential procedure for simultaneous estimation of several percentiles," *Trans. of the Society for Computer Simulation*, vol. 7, no. 1, 1990, pp. 21–44.
[7] J.C. Liechty, D.K.J. Lin and J.P. McDermott, "Single-pass low-storage arbitrary quantile estimation for massive datasets", *Statistics and Computing*, vol. 13, 2003, pp. 91–100.
[8] L. Golab and M.T. Özsu, "Issues in data stream management," *ACM SIGMOD Record*, vol. 32, No. 2, 2003, pp. 5–14.
[9] F. Chen, D. Lambert and J.C. Pinheiro, "Incremental quantile estimation for massive tracking," in *Proc. 6th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Boston, 2000, pp. 516–522.
[10] B.M. Hill, "A simple general approach to inference about the tail of a distribution," *Annals of Statistics*, vol. 3, no. 5 , 1975, pp. 1163–1174.
[11] L. Peng, "Estimating the mean of the heavy tailed distribution," *Statistics & Probability Letters*, vol. 52, no. 3, 2001, pp. 31–40.
[12] B. Abraham and J. Ledolter, "*Statistical Methods for Forecasting*," New York:John Wiley & Sons, 1983.
[13] L. Breiman, C.J. Stone and C. Kooperberg, "Robust confidence bounds for extreme upper quantiles," *J. Statist. Comput. Simul.*, vol. 37, 1990, pp. 127–149.

Conf. on Knowledge Discovery and Data Mining, Boston, 2000, pp. 516–522.
[10] B.M. Hill, "A simple general approach to inference about the tail of a distribution," *Annals of Statistics*, vol. 3, no. 5 , 1975, pp. 1163–1174.
[11] L. Peng, "Estimating the mean of the heavy tailed distribution," *Statistics & Probability Letters*, vol. 52, no. 3, 2001, pp. 31–40.
[12] B. Abraham and J. Ledolter, "*Statistical Methods for Forecasting*," New York:John Wiley & Sons, 1983.
[13] L. Breiman, C.J. Stone and C. Kooperberg, "Robust confidence bounds for extreme upper quantiles," *J. Statist. Comput. Simul.*, vol. 37, 1990, pp. 127–149.