Incremental Quantile Estimation for Massive Tracking

Fei Chen
Bell Labs
Lucent Technologies
600 Mountain Avenue
Murray Hill, NJ 07974
feic@bell-labs.com

Diane Lambert
Bell Labs
Lucent Technologies
600 Mountain Avenue
Murray Hill, NJ 07974
dl@bell-labs.com

José C. Pinheiro
Bell Labs
Lucent Technologies
600 Mountain Avenue
Murray Hill, NJ 07974
jcp@bell-labs.com

ABSTRACT

Data—call records, internet packet headers, or other transaction records—are coming down a pipe at a ferocious rate, and we need to monitor statistics of the data. There is no reason to think that the data are normally distributed, so quantiles of the data are important to watch. The probe attached to the pipe has only limited memory, though, so it is impossible to compute the quantiles by sorting the data. The only possibility is to incrementally estimate the quantiles as the data fly by. This paper provides such an incremental quantile estimator. It resembles an exponentially weighted moving average in form, processing and memory requirements, but it is based on stochastic approximation so we call it an *exponentially weighted stochastic approximation* or *EWSA*. Simulations show that the EWSA outperforms other kinds of incremental estimates that also require minimal main memory, especially when extreme quantiles are tracked for patterns of behavior that change over time. Use of the EWSA is illustrated in an application to tracking call duration for a set of callers over a three month period.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database Management—*database applications, data mining*; G.3 [**Mathematics of Computing**]: Probability and Statistics

General Terms

Customer profiles, customer relationship management, dynamic database, EWMA, equi-depth histograms, massive data, percentiles, sequential estimation, stochastic approximation, transaction data

1. BACKGROUND

Data—packet headers, network statistics, or transaction records—flow down a pipe too fast for a probe attached to a pipe to store in memory. Nonetheless, the statistics of the data have to be monitored as the data fly by, using only the limited data that can be kept in a buffer at any time. Linear statistics, like means, can easily be monitored in this way, but quantiles (or percentiles standardized to the 0-1 scale instead of 0-100 scale) may be more important than means to monitor when the data are skewed, and quantiles are not linear statistics. Quantiles are also the quantities that have to be monitored in a fixed-depth histogram in which the relative frequencies associated with the histograms bins or cells are fixed and the bin boundaries vary.

Two kinds of quantile monitoring problems can be distinguished: static and dynamic. In static monitoring, the goal is to estimate the quantiles that would be computed if all the data obtained so far could be held in memory and sorted. Static quantiles are unknowable only because there is not enough memory to hold all the data at once. They can be estimated incrementally, though, by processing the data that does fit in a buffer and then combining the buffer estimates incrementally.(See, for example, [10, 6, 9, 5].) In dynamic monitoring, the goal is not to reproduce the number that would be obtained for the entire database, but rather to estimate a quantile of the *current* behavior of the entity being tracked, especially when there is reason to suspect that the quantity being tracked is changing over time. For example, suppose the .99 quantile of call duration is being tracked and call durations have been increasing recently. Then the goal may be to estimate the .99 quantile of duration for current calls, which would be larger than the .99 quantile computed from past calls. Dynamic monitoring is important for provisioning, for example, where prediction rather than summarization of the past is the goal.

In either static or dynamic monitoring, an incremental estimate of a quantile must require little space and time to compute. Ideally, it should rely on only its previous estimate and the current set of measurements in the buffer and require only a few arithmetic operations. Section 2 describes such an estimate, which we call an *exponentially weighted stochastic approximation* (EWSA) estimate. A simulation study in Section 3 compares the performance of the EWSA to other incremental quantile estimators when measurements are either normal, t, or exponential and the parameters of the distributions are either constant over time (static, or stationary) or changing linearly over time (dynamic, or nonstationary). The EWSA has the best simulated performance, especially for dynamic monitoring when the quantile of interest is extreme (greater than .95 or smaller than .05). Section 4 applies the EWSA to tracking call duration for a set of customers over a three month period. Final thoughts and suggestions for further research are presented in Section 5.

2. INCREMENTAL QUANTILE ESTIMATES

A buffer can hold M observations, and at iteration n the observations are labelled $X_{n1}, \ldots, X_{nM}$. These M observations are considered as a random sample from a distribution F_n, which has q^{th} quantile Q_n. That is, $F_n(Q_n) = q$. The q^{th} quantile may be dynamic, in which case F_n and Q_n change over time, or the quantile may be static, in which case F_n and Q_n are constant. In either case, we want a recursive or incremental estimate of Q_n that can be computed knowing only the current set of M observations, the quantile estimate after the previous set of M observations (from the previously filled buffer), and a few tuning parameters. For the first set of observations, we can use the qn^{th} sorted largest observation to estimate Q_1. More refined initial estimates can be computed, though, if similar data have been monitored in the past. (See [2] for one possibility.)

2.1 The EWMA Estimate

First, suppose there is just one set of M measurements $X_1, \ldots, X_M$ from a distribution F. If nothing is assumed about F, then the best estimate of the q^{th} quantile Q of F is the qM^{th} sorted observation if qM is an integer and a linear interpolation of the $\lfloor qM \rfloor^{th}$ and $(\lfloor qM \rfloor + 1)^{th}$ sorted observations otherwise, where $\lfloor x \rfloor$ is the greatest integer in x [7]. This *sample quantile* $\widehat{Q}_n$ is biased, but its bias is usually smaller than its standard error if M is large [3]. If M is small, then $\widehat{Q}_n$ tends to be badly biased, but it can still be a reasonable initial estimate of Q if the incremental estimate can recover from bad starting values.

If a large number M of observations is available at each iteration, then the sample quantile of each set is a reasonable quantile estimate. If the underlying distribution F_n is not changing (*i.e.*, it is stationary), then the sample quantiles can be combined iteratively by computing the *moving average* (MA) of the sample quantiles:

$$A_n = (1 - n^{-1})A_{n-1} + n^{-1}\widehat{Q}_n$$

where $\widehat{Q}_n$ is the sample quantile of the n^{th} set of measurements. The moving average estimate A_n has smaller standard error than the sample quantile $\widehat{Q}_n$ for the observations in batch n, but it does not have smaller bias than $\widehat{Q}_n$ when F_n is not changing with n.

If F_n changes with n, then the *exponentially weighted moving average* (EWMA)

$$A^*_n = (1 - w)A^*_{n-1} + w\widehat{Q}_n,$$

where $0 < w < 1$, is more appropriate. Because w is fixed, A^*_n "ages out" older observations and adapts to changing F_n. The larger w, the faster the aging. Good choices of w are often determined by experimentation, with typical values ranging between .01 and .1. EWMA updating is simple to understand and requires little memory beyond what is needed to compute the sample quantiles. An EWMA is as biased as a sample quantile for one set of M measurements, however, and it is not useful when the quantile estimate must be updated at each observation.

2.2 The SA Estimate

Tierney [9] proposed an incremental quantile estimate based on stochastic approximation [8] that, unlike the moving average and EWMA estimates, becomes less biased as data are collected and can be computed even if the quantile estimate has to be updated at every observation. In fact, Tierney showed that if the distribution F_n of the data does not change with n, then the stochastic approximation (SA) estimate S_n behaves nearly as well as the sample quantile that would be computed from all the data collected so far. Asymptotically, the two estimates are indistinguishable. He described the SA quantile estimate for the case $M = 1$, but the algorithm can be generalized to any M and that is the version presented here.

At the n^{th} set of observations, the SA quantile estimate S_n for the q^{th} quantile is defined by

$$S_n = S_{n-1} + \frac{w_n}{e_{n-1}}\left(q - \frac{\#_{i=1}^{M}\{X_{ni} \le S_{n-1}\}}{M}\right),$$

where $w_n = 1/n$, $\#A$ is the number of times that condition A is satisfied, and $e_n = \max(f_n, f_0/\sqrt{n})$, with f_0 an initial and f_n a current estimate of the density of X at the q^{th} quantile. In words, S_n adjusts the last estimate S_{n-1} by a factor that is proportional to the difference between the theoretical fraction q of observations less than the quantile and the fraction of the M observations less than or equal to the last estimate of the quantile. The smaller the estimated density at the last estimate of the quantile, the larger the adjustment. (The estimated density must be bounded from below to prevent the correction factor from "exploding"). The weight $w_n = 1/n$ converges to zero, so newer observations have little influence on the the estimate.

The initial density estimate f_0 is generally obtained from a preliminary sample or historical data. The incremental density estimate f_n is defined as

$$f_n = (1 - w_n)f_{n-1} + w_n\frac{\#_{i=1}^{M}\{|X_{ni} - S_{n-1}| \le c_n\}}{2c_nM},$$

where $c_n = 1/\sqrt{n}$.

2.3 The EWSA Estimate

The only drawback to the SA estimate is that it gives little weight to new data so it cannot track changes over time. But, as introduced in [4], exponential weighting can be used with stochastic approximation so that more weight is given to more recent observations.

Exponential weighting is needed in both the quantile estimate and the density estimate in the stochastic approximation estimate. First, the last quantile estimate is adjusted by w/f_{n-1} (with w fixed) instead of by $(nf_{n-1})^{-1}$. Second, a nonvanishing neighborhood replaces the shrinking neighborhood with width $2n^{-1/2}$ in the density estimate. Third, the density estimate is updated with a fixed weight instead of the shrinking weight w_n.

The exponentially weighted stochastic approximation (EWSA) estimator S^*_n is computed recursively. First, initial values of f^*_0 and S^*_0 are either chosen using prior knowledge or obtained from a set of M observations $X_{01}, \ldots, X_{0M}$ as follows.

Initialization

1. Set the initial estimate S_0^* equal to the q^{th} sample quantile $\widehat{Q}_n$ of $X_{01}, \ldots, X_{0M}$.
2. Estimate the scale r_0^* of f_0^* by the interquartile range of $X_{01}, \ldots, X_{0M}$; *i.e.*, by the difference of the .75 and .25 sample quantiles. Then take $c_0^* = r_0^* M^{-1} \sum_{i=1}^{M} i^{-1/2}$.
3. Take $f_0^* = (2c_0^* M)^{-1} \max\{\#\{|X_{0i} - S_0^*| \leq c_0^*\}, 1\}$, which is the density of observations in a neighborhood of width $2c_0^*$ of S_0^*, unless the fraction of observations in the neighborhood is zero.

Next, suppose that M observations $X_{n1}, \ldots, X_{nM}$ are available to update S_{n-1}^* to S_n^*. Let $c = \sum_{i=M+1}^{2M} i^{-1/2}/M$, which is the average updating weight that the stochastic approximation estimator would assign to the next M observations. Then the quantile and density estimates are updated as follows.

Updating

1. $S_n^* = S_{n-1}^* + \frac{w}{f_{n-1}^*}(p - \frac{\#\{X_{ni} \leq S_{n-1}^*\}}{M})$.
2. $f_n^* = (1-w) f_{n-1}^* + \frac{w}{2c_{n-1}^* M} \#\{|X_{ni} - S_{n-1}^*| \leq c_{n-1}^*\}$.
3. Take r_n^* to be the difference of the current EWSA estimates for the .75 and .25 quantiles, and define the neighborhood size for the next updating step to be $c_n^* = r_n^* c$.

The neighborhood size c_n^* does not decrease to zero, so f_n^* converges to a positive number rather than to zero under weak conditions. Thus, $1/f_n^*$ should never "explode". For small n, however, the estimate f_n^* may be poor, and the EWSA estimate may behave no better than a sample quantile. A poor estimate of f_n^* may even take the EWSA estimate below the smallest possible observation or above the largest possible observation. That problem is avoided by forcing S_n^* to lie in the range of the data, if the range is known.

2.4 The Modified NCO Estimate

A single-pass algorithm for estimating quantiles for a large, static dataset is described in [5]. Their method uses a fixed number b of buffers of size k to obtain approximate quantiles for an entire database with pre-specified accuracy. The algorithm has three operations, `New`, `Collapse`, and `Output`, so we refer to it as the NCO procedure. Simply stated, `New` populates a buffer with k observations. `Collapse` takes $b \geq 2$ buffers, assigns each a weight proportional to the number of sets of observations that it represents and returns one "collapsed" buffer of k sorted values and a set of weights. The `New` and `Collapse` steps are repeated until all the data in the database has been considered. Finally, `Output` returns the approximate quantile estimate using the information from the final `Collapse` step.

The NCO algorithm is not designed to be used with stream data, but it can be adapted for that purpose. It then requires only two buffers: one for the new set of data and one for the current quantile estimates. We call this the modified NCO (MNCO) estimate. Because the estimate collapses sample quantiles, it cannot be used with $M = 1$ and may be severely biased for small M. (There would, of course, be no reason to use such a small M with a static database, for which the algorithm was designed.) At the n^{th} update, the buffer of current estimates has weight $(n-1)/n$, and the buffer of new data has weight $1/n$. As a result, new observations have little influence on the quantile estimate so the method is not appropriate for nonstationary data.

3. COMPARING THE ESTIMATES

The behaviors of the EWMA, SA, EWSA, and MNCO incremental quantile estimators on simulated data are compared in this section. Simulated data from normal, t_2 (a t with two degrees of freedom), and exponential distributions with both stationary and nonstationary parameters are considered, giving a total of six distributions. The t_2 distribution has heavy tails and is prone to outliers. The exponential distribution is highly skewed. Three quantiles are estimated under each distribution, the .5 (median), the .9 and the .99, for batch or sample sizes of $M = 1, 5$, and 15, although only the SA and EWSA estimates can be computed when $M = 1$ (see Section 2). The sample quantile of the first batch is used to initialize each estimate for $M > 1$. For $M = 1$, the sample quantile of the first 10 observations is used as the initial estimate. The total number N of observations per run, counting all observations in all batches, varies with the batch size: $N = 1000, 3000$, and 4000 for $M = 1, 5$, and 15 respectively. There were 1000 runs for each of the 18 combinations of distribution and M.

The performance of each incremental estimate $\widehat{Q}_n$ at the n^{th} update under each scenario is measured by its empirical *root mean square error* (RMSE), or its average squared distance from the quantile Q_n of the distribution used to generate the data. The RMSE at updating step n is defined by

$$\begin{aligned} \text{RMSE}\left(\widehat{Q}_n\right) &= \left[\text{E}\left(\widehat{Q}_n - Q_n\right)^2\right]^{1/2} \\ &= \left[\text{Var}\left(\widehat{Q}_n\right) + \text{Bias}^2\left(\widehat{Q}_n\right)\right]^{1/2}, \end{aligned}$$

where $\text{E}(X)$ denotes the expected value of X and $\text{Var}(X)$ denotes the variance of X. The RMSE at n is estimated by averaging the squared difference between Q_n and $\widehat{Q}_n$ over the 1000 simulation runs and then taking the square root.

3.1 Stationary Data

3.1.1 *Normal*$(0,1)$ *Distribution*

Figure 1 displays the simulated RMSE curves for the four incremental estimates when observations are taken from a standard normal distribution, which has a median of 0, a .9 quantile of 1.282, and a .99 quantile of 2.326. The EWMA and EWSA estimates use $w = .05$.

When $M = 1$, the EWSA and SA estimates have similar RMSE curves for the median, but the EWSA is clearly better for estimating the .9 and .99 quantiles, as the three leftmost panels of Figure 1 show. A closer analysis (not given here) shows that the problem is that the SA does not recover as easily from poor initial estimates. This is not surprising

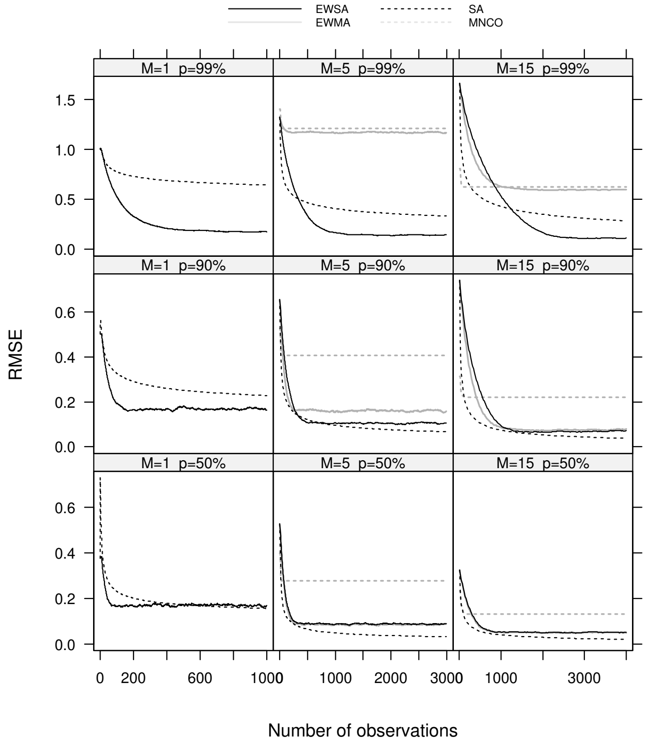

Figure 1: Simulated RMSE curves for the EWSA, SA, EWMA, and MNCO incremental quantile estimates for stationary, standard normal data. Each panel corresponds to a combination of sample size M and quantile.

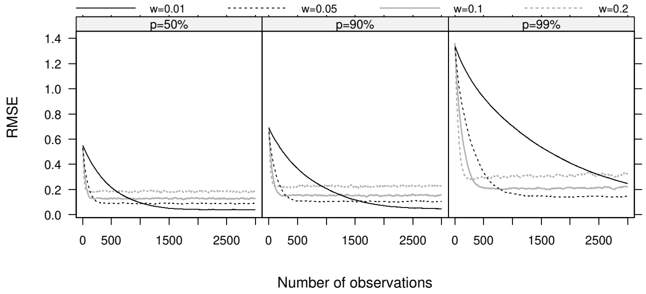

Figure 2: RMSE curves for the EWSA, SA, EWMA, and MNCO incremental quantile estimates for stationary, standard normal data, with sample size $M = 5$, corresponding to updating weights $w = .01, .05, .1$, and $.2$. Each panel corresponds to a different quantile.

because it does not age out old estimates, and poor initial estimates are more likely for larger quantiles. The EWSA RMSE curves for $M = 1$ stabilize around .17 for all three quantiles, but the SA RMSE curves stabilize at values that increase with the quantile. Both the EWSA and SA are nearly unbiased for the median (after about 200 updates, their absolute bias is less than .01). The SA is negatively biased for the larger quantiles, however, with the bias stabilizing at $-.06$ at the .9 quantile and at $-.51$ at the .99 quantile. The EWSA has negligible bias that stabilizes at less than .01 for the .9 and .99 quantiles. Note that the performance at the .1 and .01 quantiles would be the same as the performance at the .9 and .99 quantiles because the normal distribution is symmetric about the median.

For $M = 5$, the SA estimate has the lowest simulated RMSE curve for the median and .9 quantile, but the EWSA is considerably better for the .99 quantile (see the middle panels in Figure 1). The EWMA is as good as the EWSA for estimating the median, but considerably worse than either the EWSA or SA for the .9 and .99 quantiles. The MNCO has the worst RMSE curve among the four estimates. All four estimates are nearly unbiased at the median, but their biases increase with the quantile. The simulated bias for the .9 quantile stabilizes at $-.005$ for the SA, .007 for the EWSA, $-.115$ for the EWMA, and $-.101$ for the MNCO. The stabilized bias for the .99 quantile is $-.26$ for the SA, .02 for the EWSA, -1.16 for the EWMA, and -1.15 for the MNCO.

The large negative EWMA and MNCO biases reveal the basic problem with these estimates: they behave too much like the sample quantile for the batch size M rather than like the sample quantile for the total number of observations seen so far. The bias for a .99 sample quantile for a sample of size 5 from a standard normal distribution is about -1.168, which is about the same as the simulated bias for the EWMA and MNCO estimates.

The comparisons for $M = 15$ are similar to those for $M = 5$. (See the rightmost panels of Figure 1). The SA, EWSA, and EWMA estimates have similar RMSEs for the .5 and .9 quantiles, but the SA RMSE is slightly smaller. The EWMA is as good as the SA and EWSA here because the .5 and .9 sample quantiles for a sample of size 15 are nearly unbiased. The SA has smaller RMSE than the EWSA at the .99 quantile until about the 70^{th} update (corresponding to about 1000 observations), but then the SA performance degrades. The stabilized biases for the .99 quantile estimates are: $-.26$ (SA), .005 (EWSA), $-.59$ (EWMA), and $-.54$ (MNCO).

The choice of updating weight w affects the RMSE of the EWSA estimate, as Figure 2 shows for $M = 5$. The RMSE drops faster for larger w, but it also stabilizes at a larger value. The pattern is more pronounced for larger quantiles. The bias decreases faster for larger w because the quantile estimates move away from their initial values more quickly, but the variance is larger because the number of observations that affect the estimate is smaller with larger w. The larger variance accounts for the larger stabilized RMSE. In our applications, we have found that $w = .05$ gives the best trade-off between initial drop in bias and asymptotic variability.

3.1.2 *t_2 and Exponential Distributions*

A standard t_2 distribution (mean zero, scale one and two degrees of freedom) was used to simulate data that are prone to outliers. The median, .9 and .99 quantiles of the t_2 are 0, 1.89 and 6.96 respectively.

The t_2 RMSE curves (not shown here) for the EWMA, SA, EWSA and MNCO estimates are qualitatively similar to the

RMSE curves for normal data. The RMSEs under the t_2 distribution are considerably larger for the .9 and .99 quantiles, though. For $M = 1$, the stabilized SA and EWSA RMSEs under the t_2 distribution are .77 and .29 for the .9 quantile and 4.74 and 2.28 for the .99 quantile. The stabilized SA and EWSA RMSEs under the normal distribution are .23 and .17 for the .9 quantile and .64 and .18 for the .99 quantile. The EWSA is again the best estimate for the .99 quantile and SA is the best estimate for the median. The EWSA is a better estimate of the .9 quantile under the t_2 distribution for $M = 1$, but the SA is better for larger M. The MNCO is the worst estimate in all cases in which it can be used, and the EWMA performs reasonably well only when the corresponding sample quantile is not badly biased. Again, analyses not given here show that $w = .05$ gives the best trade-off between drop in bias and longrun variability.

An exponential distribution with a mean of one was used to simulate skewed data. Its median, .9 and .99 quantiles are .69, 2.30, and 4.61 respectively. The relative performance of the estimates under the exponential is similar to that in the normal and t_2 simulations, with the RMSEs falling between those for the normal and the t_2. Again, $w = .05$ is the best choice of updating weight among those considered.

3.2 Nonstationary Data

The SA and the MNCO estimates assign decreasing weights to the current observations, which is appropriate only if the distribution generating the data remains the same. The EWSA and EWMA, on the other hand, use constant updating weights, which allows them to adapt to some amount of nonstationarity. This section reports the simulation results for nonstationary data.

3.2.1 Normal Distribution With Drift

A normal distribution with variance one and a mean of $.006n$ at update n was used to simulate data with a linear trend in the mean. (This induces a linear drift of rate .006 in the quantiles.) The same mean was used to generate all observations in the same update. A weight of $w = .05$ was used for the EWSA and EWMA. Figure 3 displays the resulting RMSE curves for the EWMA, SA, EWSA and MCNO estimators.

As would be expected, the SA and MNCO estimates break down under nonstationarity. In fact, their RMSE curves increase linearly with update number because their biases become more negative as the target quantile drifts upward. The SA and MNCO RMSEs in the $M = 15$ panels are smaller because the number of simulated updates decreases with M in our simulation.

The EWSA and EWMA, on the other hand, adapt to the nonstationarity, and their RMSE curves stabilize. The stabilized RMSEs are larger under nonstationarity than under stationarity, though. For example, the stabilized EWSA RMSEs for the median, .9 and .99 quantiles under the normal drift model are .23, .27 and .54, compared to a stabilized RMSE of about .17 for all three quantiles under the stationary normal model. The RMSEs relative to the quantile being estimated do not increase under nonstationarity, though. The EWMA again behaves poorly at the .99 quantile because it has a large negative bias. At the median for all M and at the .9 quantile for $M = 5$ or 15, the EWMA is competitive with the EWSA. Overall, the EWSA estimate has the best RMSE curve, with the additional advantage that it can be used with $M = 1$.

3.2.2 t_2 and Exponential Distributions with Drift

The same drifting mean was used to simulate nonstationary t_2 data, giving quantiles that increase linearly at a rate of .006 per update. The RMSE curves, which are not shown here, are qualitatively similar to those in Figure 3, so the conclusions about the performance of the estimates are the same. The SA and MNCO break down under nonstationarity for heavy tailed data, and the EWSA and EWMA adapt to the nonstationarity. The EWSA outperforms the EWMA at the .99 quantile for all M and and at the .9 quantile for $M = 5$. The stabilized EWSA and EWMA RMSEs tend to be larger when the mean drifts, but the RMSE relative to the quantile being estimated is not sensitive to the drift in mean.

The nonstationary exponential model added .006 to the mean at each update, starting at a mean of 1.006. The data also become more variable as the mean increases, because the mean equals the standard deviation for an exponential distribution, and the target exponential quantiles are multiplied by a factor of 1.006 at each update. For example, the quantiles after 1000 updates are 600% larger than the orig-

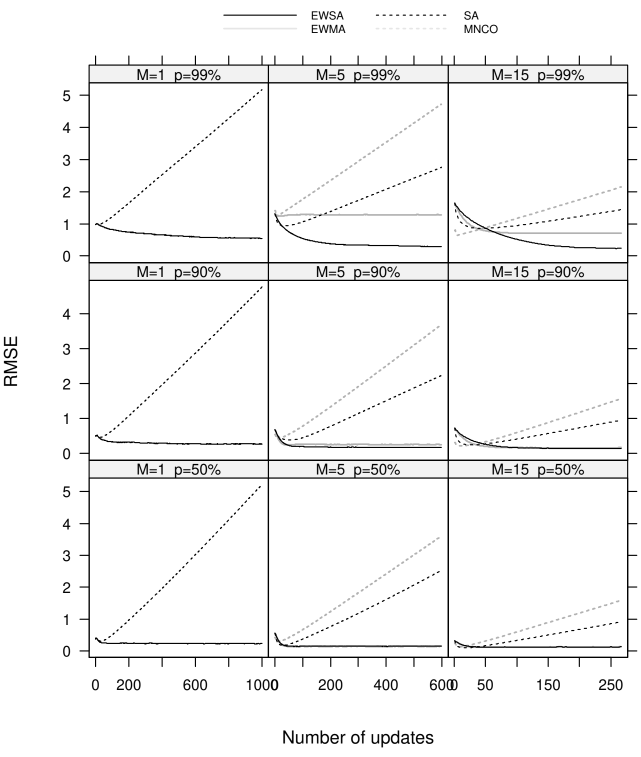

Figure 3: RMSE curves for the EWSA, SA, EWMA, and MNCO incremental quantile estimates for a normal process with a mean that increases by .006 at each update, starting from .006. Each panel corresponds to a combination of sample size and quantile.

inal quantiles, so the exponential nonstationarity is more severe than the normal and t_2 nonstationarity.

The RMSEs of the estimators under exponential nonstationarity are similar to the RMSEs under normal and t_2 nonstationarity, but the SA and MNCO estimates deteriorate faster and the stabilized EWSA and EWMA RMSEs are worse. After 1000 updates, the SA RMSE for the .99 quantile is 27.7 under exponential nonstationarity, compared to 5.2 for the normal nonstationary model and 9.3 for the t_2. The stabilized EWSA RMSE ranges from .13 for the median with $M = 15$ to 7.8 for the .99 quantile with $M = 1$. The conclusions about normal and t_2 nonstationary also apply to exponential nonstationarity: EWSA is best able to handle nonstationarity at all sample sizes and quantiles.

4. AN APPLICATION: TRACKING CALL DURATION

The cost of establishing service for a new telecommunications customer is so high and competition for customers is so fierce, that service providers are asking to have up-to-date information on each caller on their network. This information is always fresh for marketing analyses and it enables the kinds of real-time analyses that are needed for fraud detection (See [1].) In many of these applications, especially fraud, extreme behavior is most interesting, so the goal is to track extreme percentiles of at least some aspects of calling behavior. To keep the information as up-to-date as possible, we use $M = 1$ in these applications. Thus, only the SA and EWSA estimates can be used.

Here we focus on tracking the .99 quantile of call duration for a random sample of 146 callers who made between 1150 and 3400 completed calls during peak hours over a three month period, where peak hours are defined to be 9:00 a.m. to 8:00 p.m. Monday through Friday. To evaluate the performance of the EWSA and SA estimates, we assume that the call durations for each customer are stationary over the three month period, so the target quantile to be estimated is the .99 sample quantile Q_i computed from all the completed peak hour calls made by customer i during the three month period. Because there is a wide range in target quantiles across customers, we use the *absolute relative error*, $100|\widehat{Q}_i - Q_i|/Q_i$, to evaluate performance at each call n for each customer i. The median and .25 and .75 quantiles of the absolute relative error at call n across all customers that made at least n calls are shown in Figure 4.

As in the simulations described in Section 3, the EWSA clearly outperforms the SA estimate for the .99 quantile. The EWSA median absolute relative error stabilizes at about 5%, while the stabilized SA median absolute relative error is about ten times larger. After about 500 calls, the .75 quantile curve for EWSA stays below the .25 quantile curve for SA, suggesting that the EWSA is better than the SA for a majority of customers. The .75 quantile of the EWSA relative error stabilizes at 12%, which is about one third the stabilized .25 quantile of the SA relative error of 35%. In this application, then, the EWSA is a reliable and efficient estimate for tracking call duration behavior.

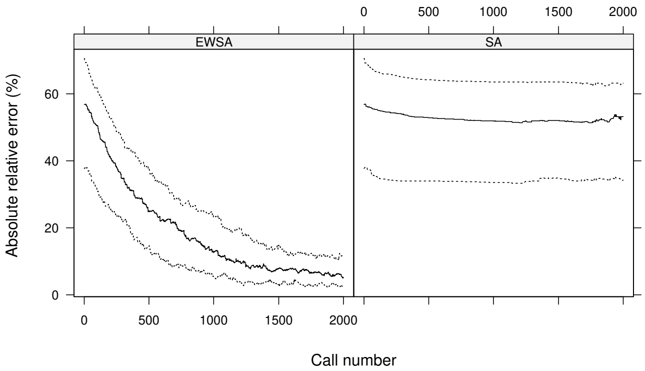

Figure 4: Absolute relative error curves for the SA and EWSA estimates of the .99 call duration quantiles. An updating weight of $w = .05$ is used for both estimates. The three curves in each panel refer to the .25, .5, and .75 pointwise quantiles of the absolute relative errors across customers. The .25 and .75 pointwise quantile curves are drawn in dashed lines, while a solid line is used for the pointwise median curves.

5. CONCLUSIONS

Increasingly, real-time applications involve massive amounts of data that are collected through a pipe and cannot be held in main memory. The goal may be to summarize the data collected so far (static distribution), or to understand the behavior at the current time or to predict the behavior at the next observation (dynamic distribution). Extreme behavior is often the most interesting in tracking applications, in which case quantiles are more useful summaries than the mean and standard deviation. This paper describes a space-efficient, computationally simple incremental estimate for quantiles: the exponentially weighted stochastic approximation (EWSA) estimate. The EWSA can be used with stream data collected in batches of any size, including one observation at a time. Simulations suggest that the EWSA outperforms other space-efficient, incremental estimates for extreme quantiles, especially when the data are not stationary, and the EWSA is as good as other estimators otherwise. In the application to tracking the .99 quantile of call duration call-by-call in Section 4, the EWSA was better than the only competing procedure that we are aware of, which is the stochastic approximation estimate.

This paper has focused on estimation from stream data, which is collected in batches. The EWSA estimate can also be used with very large, static databases that require incremental computations, though. The data would need to be split into subsets, and the subsets would be treated sequentially. Much larger batch sizes than the ones considered here could be used. It would be interesting to compare the EWSA in this context to the NCO and other quantile estimates designed for static databases. Other performance metrics not considered here, such as processing time, would also need to be considered. Approximate bounds on the error in an EWSA estimate in this context can be based on the distributional properties of the data.

Further research is needed to design methods for choosing optimal updating EWSA weights as a function of the quantile fraction q and the distribution of the data. Simulation results suggest that a weight of .05 is adequate for stationary data, but a different weight may be better for nonstationarity and for different distributions and parameter values. Presumably, more nonstationarityy requires a larger w, but we do not know how fast a drift the EWSA can accommodate for different kinds of distributions. More investigation is also needed on optimal batch sizes for EWSA estimation.

6. REFERENCES

[1] M. H. Cahill, D. Lambert, J. C. Pinheiro, and D. X. Sun. Detecting fraud in the real world. Technical report, Bell Labs, Lucent Technologies, 2000.

[2] F. Chen, D. Lambert, J. C. Pinheiro, and D. X. Sun. Reducting transaction databases, without lagging behind the data or losing information. Technical report, Bell Labs, Lucent Technologies, 2000.

[3] H. A. David. *Order Statistics*. Wiley, New York, NY, 2nd edition, 1981.

[4] D. Lambert. Sequential percentile estimation. U.S. Ballot Comments on the ISO/IEC Ballot on SC21 N6677, DIS 10164-11, Workload Monitoring Function, reference SC21 N6677, 1992.

[5] G. S. Manku, S. Rajagopalan, and B. G. Lindsay. Approximate medians and other quantiles in one pass and with limited memory. In *SIGMOD 1998*, pages 426–435, 1998.

[6] J. Munro and M. Paterson. Selection and sorting with limited storage. *Theoretical Computer Science*, 12:315–323, 1980.

[7] J. Pfanzagl. *Contributions to Applied Statistics (dedicated to Arthur Linder)*, chapter Investigating the Quantile of an Unknown Distribution, pages 111–126. Birkhauser Verlag, Basel, 1974.

[8] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–427, 1951.

[9] L. Tierney. A space-efficient recursive procedure for estimating a quantile of an unknown distribution. *SIAM Journal on Scientific and Statistical Computing*, 4:706–711, 1983.

[10] B. Weide. Space-efficient on-line selection algorithms. In *Computer Science and Statistics: Proceedings of the Eleventh Annual Symposium on the Interface*, pages 308–311, Raleigh, 1978. Institute of Statistics, North Carolina State University.