Latent Dissonance via Efficient Quantile Estimation

Eric Joraskie Nick Kantack

***Abstract*—This whitepaper introduces a method of estimating the quantiles of a distribution that is observed in a stream fashion. Specifically, this method requires no queuing of data but instead updates the mean, variance, and quantile estimates as each datum is processed. We prove that this estimation algorithm is an unbiased estimator of the true statistics and has a time complexity that is linear in the desired quantile resolution.**

I. Mean Estimator

When a queue of N data points with mean $\langle x \rangle$ loses a datum (x_-) and gains a datum (x_+) the new mean $\langle x \rangle_2$ can be tracked via

$$\langle x \rangle_2 = \langle x \rangle + \frac{x_+ - x_-}{N} \tag{1}$$

If we make the approximation that $x_- = \langle x \rangle$, we can obtain the unbiased estimator of the new mean $\langle x \rangle_2$ of

$$\langle x \rangle_2 \approx \langle x \rangle \frac{N-1}{N} + x_+ \frac{1}{N} \tag{2}$$

This is a recognizable exponential moving average. Furthermore, $\langle x \rangle_2$ is a linear combination of data (x_+) each of which has an expectation of $\langle x \rangle$, therefore, the expectation of $\langle x \rangle_2$ is $\langle x \rangle$ by linearity of expectations, and thus we have an unbiased estimator.

II. Variance Estimator

Suppose that σ^2 is the variance before the data queue of the previous section loses x_- and gains x_+. The new variance σ_2^2 is altered by three influences.

1) The residual from x_- is lost
2) The residual from x_+ is added
3) The residuals between the queue ends notice the change in mean

The first contribution is clearly $-(x_- - \langle x \rangle)/N$, and the second is clearly $(x_+ - \langle x \rangle_2)/N$. To find the third contribution, we examine the change in σ^2 induced by shifting the mean from $\langle x \rangle$ to $\langle x \rangle + \delta x$. Defining

$$\widetilde{\sigma}^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \langle x \rangle)^2 \tag{3}$$

and

$$\widetilde{\sigma}_2^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - (\langle x \rangle + \delta x))^2 \tag{4}$$

we can solve for $\widetilde{\sigma}_2^2 - \widetilde{\sigma}^2$.

$$\begin{aligned} \widetilde{\sigma}_2^2 = \frac{1}{N} \sum_{n=1}^{N} \left(x_n^2 - 2x_n \langle x \rangle + \langle x \rangle^2 \right) + \\ \frac{1}{N} \sum_{n=1}^{N} \left(\delta x^2 + \delta x \langle x \rangle - x_n \delta x \right) \end{aligned} \tag{5}$$

$$\widetilde{\sigma}_2^2 = \widetilde{\sigma}^2 + \delta x^2 + \delta x \langle x \rangle - \frac{\delta x}{N} \sum_{n=1}^{N} x_n \tag{6}$$

$$\widetilde{\sigma}_2^2 - \widetilde{\sigma}^2 = \delta x^2 \tag{2}$$

Now we can combine all three changes to express σ_2^2 as a function of σ^2. Noting that δx in this case $(x_+ - x_-)/N$ and thus

$$\sigma_2^2 = \sigma^2 + \left(\frac{x_+ - x_-}{N} \right)^2 + \frac{(x_+ - \langle x \rangle_2)^2 - (x_- - \langle x \rangle)^2}{N} \tag{7}$$

We can again applying the approximation $x_- \approx \langle x \rangle_2$ for an unbiased estimator of σ_2^2.

III. Quantile Estimators

Suppose that n equally spaced quantiles $(q_1, q_2, ..., q_n)$ are to be estimated for the queue. We adopt an algorithm for updating the quantile positions stored in the vector $\boldsymbol{q}$. When a datum is observed which falls between two quantiles, the space between those two quantiles is reduced by a standard amount s while the spaces between all other quantile pairs is increased by $s/(n+1)$ (s is chosen to be small relative to σ, such as $s = \sigma/10$). This is accomplished by the following rule.

$$\begin{aligned} q_k &\leftarrow q_k - \frac{n-k}{n+1} s \quad \text{if} \quad x_+ < q_k \\ q_k &\leftarrow q_k + \frac{k+1}{n+1} s \quad \text{if} \quad x_+ \geq q_k \end{aligned} \tag{8}$$

We can prove that under the condition described above, the resulting quantile are an unbiased estimator of the real quantiles of the queue. To prove this, we examine a specific, arbitrary quantile q_k. Let $P(x)$ be the unknown probability density function of the data. The expectation of the movement of q_k after a new datum is

$$\langle \Delta q_k \rangle = s \left[-\frac{n-k}{n+1} \int_{-\infty}^{x_k} P(x) dx \; + \frac{k+1}{n+1} \int_{x_k}^{\infty} P(x) dx \right] \tag{9}$$

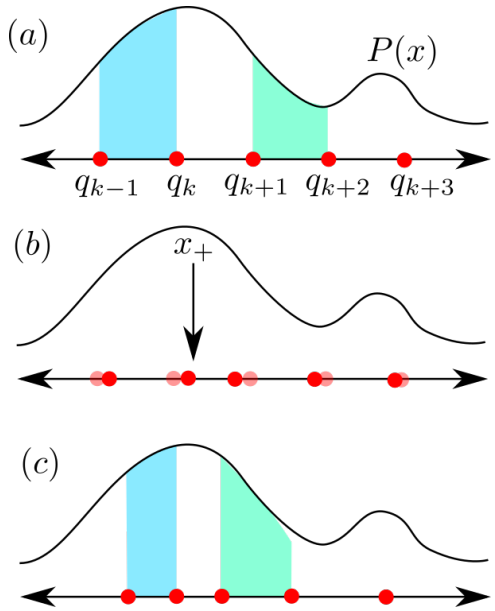

Fig. 1. (a) Initially, the mean and variance are used to position rough estimates for the location of quantiles. These estimates will be incorrectly placed, and thus the populations between quantiles (blue and green regions) will be unequal. (b) When a new datum arrives, the immediate quantiles to either side are shifted inward by a small amount (e.g. $\sigma/10$). Further off quantiles are also shifted inwards, although by an amount that scales down linearly with how many quantiles separate it from the datapoint. (c) In the steady state, quantiles will drift so that the populations between quantiles are approximately equal.

If all quantiles are correctly placed, then

$$\int_{-\infty}^{x_k} P(x)dx = \frac{k+1}{n+1} \tag{10}$$

and

$$\int_{x_k}^{\infty} P(x)dx = \frac{n-k}{n+1} \tag{11}$$

Substituting into (9), we obtain

$$\langle \Delta q_k \rangle = s\left[-\frac{n-k}{n+1}\frac{k+1}{n+1} + \frac{k+1}{n+1}\frac{n-k}{n+1}\right] = 0 \tag{12}$$

Therefore, this serves as an unbiased estimator of the true quantiles.