Online Covariance

by Joshua Burkholder

Given the following set of two-dimensional inputs:

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_{n-1}, y_{n-1}), (x_n, y_n)\}$$

Let n be the number of two-dimensional inputs, X represent the x dimension, Y represent the y dimension, $Cov_n(X,Y)$ be the biased sample covariance of the x and y dimensions for the first n two-dimensional inputs, $Cov_{n-1}(X,Y)$ be the biased sample covariance of the x and y dimensions for the first $n-1$ two-dimensional inputs, x_n be the x value of the n-th two-dimensional input, $\bar{x}_n$ be the sample mean of the x values for the first n two-dimensional inputs, y_n be the y value of the n-th two-dimensional input, and $\bar{y}_{n-1}$ be the sample mean of the y values for the first $n-1$ two-dimensional inputs. Then, the recurrence equation for the biased sample covariance (a.k.a. online covariance) is:

$$Cov_n(X,Y) = Cov_{n-1}(X,Y) - \frac{Cov_{n-1}(X,Y) - (x_n - \bar{x}_n)(y_n - \bar{y}_{n-1})}{n}$$

Note: The recurrence equation above also applies when computing the online covariance matrix:

$$\vec{\vec{\Sigma}}_n[j,k] = \vec{\vec{\Sigma}}_{n-1}[j,k] - \frac{\vec{\vec{\Sigma}}_{n-1}[j,k] - \left(\vec{\mathbf{x}}_n[j] - \overline{\vec{\mathbf{x}}[j]}_n\right)\left(\vec{\mathbf{x}}_n[k] - \overline{\vec{\mathbf{x}}[k]}_{n-1}\right)}{n}.$$

However, we will restrict ourselves to the online covariance computation of two-dimensional input in this post and explore the online covariance matrix computation of m-dimensional input in a later post.

Proof:

The definition of the biased sample covariance of the x and y dimensions for the first n two-dimensional inputs is defined as:

$$Cov_n(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{n}.$$

If we expand this definition, we have:

$$Cov_n(X,Y) = \frac{\sum_{i=1}^{n-1}(x_i - \bar{x}_n)(y_i - \bar{y}_n) + (x_n - \bar{x}_n)(y_n - \bar{y}_n)}{n}.$$

Since the recurrence equations for the sample mean of the x and y values are:

$$\bar{x}_n = \bar{x}_{n-1} - \frac{\bar{x}_{n-1} - x_n}{n} \text{ and } \bar{y}_n = \bar{y}_{n-1} - \frac{\bar{y}_{n-1} - y_n}{n},$$

then we have:

$$Cov_n(X,Y) = \frac{\sum_{i=1}^{n-1}\left(x_i - \left(\bar{x}_{n-1} - \frac{\bar{x}_{n-1} - x_n}{n}\right)\right)\left(y_i - \left(\bar{y}_{n-1} - \frac{\bar{y}_{n-1} - y_n}{n}\right)\right) + (x_n - \bar{x}_n)(y_n - \bar{y}_n)}{n}$$

$$Cov_n(X,Y) = \frac{\sum_{i=1}^{n-1}\left(x_i - \bar{x}_{n-1} + \frac{\bar{x}_{n-1} - x_n}{n}\right)\left(y_i - \bar{y}_{n-1} + \frac{\bar{y}_{n-1} - y_n}{n}\right) + (x_n - \bar{x}_n)(y_n - \bar{y}_n)}{n}$$

$$Cov_n(X,Y) = \frac{\sum_{i=1}^{n-1}\left(\begin{array}{l} x_i y_i - x_i \bar{y}_{n-1} + x_i\left(\frac{\bar{y}_{n-1} - y_n}{n}\right) - \bar{x}_{n-1} y_i + \bar{x}_{n-1}\bar{y}_{n-1} - \bar{x}_{n-1}\left(\frac{\bar{y}_{n-1} - y_n}{n}\right) \\ +\left(\frac{\bar{x}_{n-1} - x_n}{n}\right) y_i - \left(\frac{\bar{x}_{n-1} - x_n}{n}\right)\bar{y}_{n-1} + \left(\frac{\bar{x}_{n-1} - x_n}{n}\right)\left(\frac{\bar{y}_{n-1} - y_n}{n}\right) \end{array}\right) + (x_n - \bar{x}_n)(y_n - \bar{y}_n)}{n}$$

$$Cov_n(X,Y) = \frac{\left(\begin{array}{l} \sum_{i=1}^{n-1}(x_i y_i - x_i \bar{y}_{n-1} - \bar{x}_{n-1} y_i + \bar{x}_{n-1}\bar{y}_{n-1}) + \sum_{i=1}^{n-1}\left(x_i\left(\frac{\bar{y}_{n-1} - y_n}{n}\right) - \bar{x}_{n-1}\left(\frac{\bar{y}_{n-1} - y_n}{n}\right)\right) \\ +\sum_{i=1}^{n-1}\left(\left(\frac{\bar{x}_{n-1} - x_n}{n}\right) y_i - \left(\frac{\bar{x}_{n-1} - x_n}{n}\right)\bar{y}_{n-1} + \left(\frac{\bar{x}_{n-1} - x_n}{n}\right)\left(\frac{\bar{y}_{n-1} - y_n}{n}\right)\right) + (x_n - \bar{x}_n)(y_n - \bar{y}_n) \end{array}\right)}{n}$$

$$Cov_n(X,Y) = \frac{\left(\begin{array}{l} \sum_{i=1}^{n-1}(x_i - \bar{x}_{n-1})(y_i - \bar{y}_{n-1}) + \left(\frac{\bar{y}_{n-1} - y_n}{n}\right)\sum_{i=1}^{n-1}(x_i - \bar{x}_{n-1}) \\ +\left(\frac{\bar{x}_{n-1} - x_n}{n}\right)\sum_{i=1}^{n-1}\left(y_i - \bar{y}_{n-1} + \left(\frac{\bar{y}_{n-1} - y_n}{n}\right)\right) + (x_n - \bar{x}_n)(y_n - \bar{y}_n) \end{array}\right)}{n}$$

Since the biased sample covariance of the x and y dimensions for the first $n-1$ two-dimensional inputs is defined as:

$$Cov_{n-1}(X,Y)=\frac{\sum_{i=1}^{n-1}(x_i-\bar{x}_{n-1})(y_i-\bar{y}_{n-1})}{n-1},$$

then we also have:

$$\sum_{i=1}^{n-1}(x_i-\bar{x}_{n-1})(y_i-\bar{y}_{n-1})=(n-1)Cov_{n-1}(X,Y).$$

With this, we have:

$$Cov_n(X,Y)=\frac{\left((n-1)Cov_{n-1}(X,Y)+\left(\frac{\bar{y}_{n-1}-y_n}{n}\right)\sum_{i=1}^{n-1}(x_i-\bar{x}_{n-1})+\left(\frac{\bar{x}_{n-1}-x_n}{n}\right)\sum_{i=1}^{n-1}\left(y_i-\bar{y}_{n-1}+\left(\frac{\bar{y}_{n-1}-y_n}{n}\right)\right)+(x_n-\bar{x}_n)(y_n-\bar{y}_n)\right)}{n}$$

$$Cov_n(X,Y)=\frac{\left((n-1)Cov_{n-1}(X,Y)+\left(\frac{\bar{y}_{n-1}-y_n}{n}\right)\left(\sum_{i=1}^{n-1}(x_i)+\sum_{i=1}^{n-1}(-\bar{x}_{n-1})\right)+\left(\frac{\bar{x}_{n-1}-x_n}{n}\right)\left(\sum_{i=1}^{n-1}(y_i)+\sum_{i=1}^{n-1}(-\bar{y}_{n-1})+\sum_{i=1}^{n-1}\left(\frac{\bar{y}_{n-1}-y_n}{n}\right)\right)+(x_n-\bar{x}_n)(y_n-\bar{y}_n)\right)}{n}$$

$$Cov_n(X,Y)=\frac{\left((n-1)Cov_{n-1}(X,Y)+\left(\frac{\bar{y}_{n-1}-y_n}{n}\right)\left(\sum_{i=1}^{n-1}(x_i)-\bar{x}_{n-1}\sum_{i=1}^{n-1}(1)\right)+\left(\frac{\bar{x}_{n-1}-x_n}{n}\right)\left(\sum_{i=1}^{n-1}(y_i)-\bar{y}_{n-1}\sum_{i=1}^{n-1}(1)+\left(\frac{\bar{y}_{n-1}-y_n}{n}\right)\sum_{i=1}^{n-1}(1)\right)+(x_n-\bar{x}_n)(y_n-\bar{y}_n)\right)}{n}$$

$$Cov_n(X,Y)=\frac{\left((n-1)Cov_{n-1}(X,Y)+\left(\frac{\bar{y}_{n-1}-y_n}{n}\right)\left(\sum_{i=1}^{n-1}(x_i)-\bar{x}_{n-1}(n-1)\right)+\left(\frac{\bar{x}_{n-1}-x_n}{n}\right)\left(\sum_{i=1}^{n-1}(y_i)-\bar{y}_{n-1}(n-1)+\left(\frac{\bar{y}_{n-1}-y_n}{n}\right)(n-1)\right)+(x_n-\bar{x}_n)(y_n-\bar{y}_n)\right)}{n}$$

Since the sample mean for the first $n-1$ x and y values are defined as:

$$\bar{x}_{n-1} = \frac{\sum_{i=1}^{n-1} x_i}{n-1} \text{ and } \bar{y}_{n-1} = \frac{\sum_{i=1}^{n-1} y_i}{n-1},$$

then we also have:

$$\sum_{i=1}^{n-1} x_i = \bar{x}_{n-1}(n-1) \text{ and } \sum_{i=1}^{n-1} y_i = \bar{y}_{n-1}(n-1).$$

With that, we have:

$$Cov_n(X,Y) = \frac{\left((n-1)Cov_{n-1}(X,Y) + \left(\frac{\bar{y}_{n-1} - y_n}{n}\right)\left(\bar{x}_{n-1}(n-1) - \bar{x}_{n-1}(n-1)\right) + \left(\frac{\bar{x}_{n-1} - x_n}{n}\right)\left(\bar{y}_{n-1}(n-1) - \bar{y}_{n-1}(n-1) + \left(\frac{\bar{y}_{n-1} - y_n}{n}\right)(n-1)\right) + (x_n - \bar{x}_n)(y_n - \bar{y}_n)\right)}{n}$$

$$Cov_n(X,Y) = \frac{\left((n-1)Cov_{n-1}(X,Y) + \left(\frac{\bar{y}_{n-1} - y_n}{n}\right)\left(\cancel{\bar{x}_{n-1}(n-1)} - \cancel{\bar{x}_{n-1}(n-1)}\right) + \left(\frac{\bar{x}_{n-1} - x_n}{n}\right)\left(\cancel{\bar{y}_{n-1}(n-1)} - \cancel{\bar{y}_{n-1}(n-1)} + \left(\frac{\bar{y}_{n-1} - y_n}{n}\right)(n-1)\right) + (x_n - \bar{x}_n)(y_n - \bar{y}_n)\right)}{n}$$

$$Cov_n(X,Y) = \frac{\left((n-1)Cov_{n-1}(X,Y) + \left(\frac{\bar{x}_{n-1} - x_n}{n}\right)\left(\frac{\bar{y}_{n-1} - y_n}{n}\right)(n-1) + (x_n - \bar{x}_n)(y_n - \bar{y}_n)\right)}{n}$$

$$Cov_n(X,Y) = \frac{\left((n-1)Cov_{n-1}(X,Y) + (n-1)\left(\frac{\bar{x}_{n-1} - x_n}{n}\right)\left(\frac{\bar{y}_{n-1} - y_n}{n}\right) + x_n y_n - x_n \bar{y}_n - \bar{x}_n y_n + \bar{x}_n \bar{y}_n\right)}{n}$$

Since the recurrence equation for the sample mean of the y values is:

$$\bar{y}_n = \bar{y}_{n-1} - \frac{\bar{y}_{n-1} - y_n}{n},$$

then we have:

$$Cov_n(X,Y) = \frac{\left(\begin{array}{l} (n-1)Cov_{n-1}(X,Y) + (n-1)\left(\frac{\overline{x}_{n-1} - x_n}{n}\right)\left(\frac{\overline{y}_{n-1} - y_n}{n}\right) \\ +x_n y_n - x_n\left(\overline{y}_{n-1} - \frac{\overline{y}_{n-1} - y_n}{n}\right) - \overline{x}_n y_n + \overline{x}_n\left(\overline{y}_{n-1} - \frac{\overline{y}_{n-1} - y_n}{n}\right) \end{array} \right)}{n}$$

$$Cov_n(X,Y) = \frac{\left(\begin{array}{l} (n-1)Cov_{n-1}(X,Y) + (n-1)\left(\frac{\overline{x}_{n-1} - x_n}{n}\right)\left(\frac{\overline{y}_{n-1} - y_n}{n}\right) \\ +x_n y_n - x_n\overline{y}_{n-1} + x_n\left(\frac{\overline{y}_{n-1} - y_n}{n}\right) - \overline{x}_n y_n + \overline{x}_n\overline{y}_{n-1} - \overline{x}_n\left(\frac{\overline{y}_{n-1} - y_n}{n}\right) \end{array} \right)}{n}$$

$$Cov_n(X,Y) = \frac{\left(\begin{array}{l} (n-1)Cov_{n-1}(X,Y) + (n-1)\left(\frac{\overline{x}_{n-1} - x_n}{n}\right)\left(\frac{\overline{y}_{n-1} - y_n}{n}\right) \\ +x_n y_n - x_n\overline{y}_{n-1} - \overline{x}_n y_n + \overline{x}_n\overline{y}_{n-1} + (x_n - \overline{x}_n)\left(\frac{\overline{y}_{n-1} - y_n}{n}\right) \end{array} \right)}{n}$$

$$Cov_n(X,Y) = \frac{\left(\begin{array}{l} (n-1)Cov_{n-1}(X,Y) + (n-1)\left(\frac{\overline{x}_{n-1} - x_n}{n}\right)\left(\frac{\overline{y}_{n-1} - y_n}{n}\right) \\ +(x_n - \overline{x}_n)(y_n - \overline{y}_{n-1}) + (x_n - \overline{x}_n)\left(\frac{\overline{y}_{n-1} - y_n}{n}\right) \end{array} \right)}{n}$$

Since the recurrence equation for the sample mean of the x values is:

$$\overline{x}_n = \overline{x}_{n-1} - \frac{\overline{x}_{n-1} - x_n}{n}$$
$$\overline{x}_n = \frac{n\overline{x}_{n-1}}{n} + \frac{-\overline{x}_{n-1} + x_n}{n}$$
$$\overline{x}_n = \frac{(n-1)\overline{x}_{n-1} + x_n}{n},$$

then we have:

$$Cov_n(X,Y)=\frac{\left(\begin{array}{l}(n-1)Cov_{n-1}(X,Y)+(n-1)\left(\frac{\bar{x}_{n-1}-x_n}{n}\right)\left(\frac{\bar{y}_{n-1}-y_n}{n}\right)\\+(x_n-\bar{x}_n)(y_n-\bar{y}_{n-1})+\left(x_n-\left(\frac{(n-1)\bar{x}_{n-1}+x_n}{n}\right)\right)\left(\frac{\bar{y}_{n-1}-y_n}{n}\right)\end{array}\right)}{n}$$

$$Cov_n(X,Y)=\frac{\left(\begin{array}{l}(n-1)Cov_{n-1}(X,Y)+(n-1)\left(\frac{\bar{x}_{n-1}-x_n}{n}\right)\left(\frac{\bar{y}_{n-1}-y_n}{n}\right)\\+(x_n-\bar{x}_n)(y_n-\bar{y}_{n-1})+\left(\frac{nx_n}{n}+\frac{-(n-1)\bar{x}_{n-1}-x_n}{n}\right)\left(\frac{\bar{y}_{n-1}-y_n}{n}\right)\end{array}\right)}{n}$$

$$Cov_n(X,Y)=\frac{\left(\begin{array}{l}(n-1)Cov_{n-1}(X,Y)+(n-1)\left(\frac{\bar{x}_{n-1}-x_n}{n}\right)\left(\frac{\bar{y}_{n-1}-y_n}{n}\right)\\+(x_n-\bar{x}_n)(y_n-\bar{y}_{n-1})+\left(\frac{-(n-1)\bar{x}_{n-1}+(n-1)x_n}{n}\right)\left(\frac{\bar{y}_{n-1}-y_n}{n}\right)\end{array}\right)}{n}$$

$$Cov_n(X,Y)=\frac{\left(\begin{array}{l}(n-1)Cov_{n-1}(X,Y)+(n-1)\left(\frac{\bar{x}_{n-1}-x_n}{n}\right)\left(\frac{\bar{y}_{n-1}-y_n}{n}\right)\\+(x_n-\bar{x}_n)(y_n-\bar{y}_{n-1})-(n-1)\left(\frac{\bar{x}_{n-1}-x_n}{n}\right)\left(\frac{\bar{y}_{n-1}-y_n}{n}\right)\end{array}\right)}{n}$$

$$Cov_n(X,Y)=\frac{\left(\begin{array}{l}(n-1)Cov_{n-1}(X,Y)+\cancel{(n-1)\left(\frac{\bar{x}_{n-1}-x_n}{n}\right)\left(\frac{\bar{y}_{n-1}-y_n}{n}\right)}\\+(x_n-\bar{x}_n)(y_n-\bar{y}_{n-1})-\cancel{(n-1)\left(\frac{\bar{x}_{n-1}-x_n}{n}\right)\left(\frac{\bar{y}_{n-1}-y_n}{n}\right)}\end{array}\right)}{n}$$

$$Cov_n(X,Y)=\frac{(n-1)Cov_{n-1}(X,Y)+(x_n-\bar{x}_n)(y_n-\bar{y}_{n-1})}{n}$$

$$Cov_n(X,Y)=\frac{nCov_{n-1}(X,Y)-Cov_{n-1}(X,Y)+(x_n-\bar{x}_n)(y_n-\bar{y}_{n-1})}{n}$$

$$Cov_n(X,Y)=\frac{nCov_{n-1}(X,Y)}{n}+\frac{-\left(Cov_{n-1}(X,Y)-(x_n-\bar{x}_n)(y_n-\bar{y}_{n-1})\right)}{n}$$

$$Cov_n(X,Y)=Cov_{n-1}(X,Y)-\frac{Cov_{n-1}(X,Y)-(x_n-\bar{x}_n)(y_n-\bar{y}_{n-1})}{n}$$

Therefore, the recurrence equation for the biased sample covariance (a.k.a. online covariance) is:

$$Cov_n(X,Y) = Cov_{n-1}(X,Y) - \frac{Cov_{n-1}(X,Y) - (x_n - \bar{x}_n)(y_n - \bar{y}_{n-1})}{n}$$

Note: We can manipulate this recurrence equation such as that we also have:

$$Cov_n(X,Y) = Cov_{n-1}(X,Y) - \frac{Cov_{n-1}(X,Y) - (x_n - \bar{x}_{n-1})(y_n - \bar{y}_n)}{n},$$

$$Cov_n(X,Y) = Cov_{n-1}(X,Y) - \frac{Cov_{n-1}(X,Y) - \left(\frac{n-1}{n}\right)(x_n - \bar{x}_{n-1})(y_n - \bar{y}_{n-1})}{n},$$

and

$$Cov_n(X,Y) = \frac{(n-1)\left(Cov_{n-1}(X,Y) + \frac{(x_n - \bar{x}_{n-1})(y_n - \bar{y}_{n-1})}{n}\right)}{n}$$

Reference:
http://en.wikipedia.org/wiki/Algorithms_for_calculating_variance

---

Example of C++ code that computes the online covariance:

```
// Filename: main.cpp
#include <iostream>
#include <iomanip>

int main () {

        double x;
        double y;
        double n = 0;
        double mean_x = 0;  // mean of the x values
        double mean_y = 0;  // mean of the y values
        double cov = 0;     // covariance of the x and y values
        double prev_mean_x; // previous mean of the x values
        double prev_mean_y; // previous mean of the y values
        double prev_cov;    // previous covariance of the x and y values

        if ( std::cin >> x && std::cin >> y ) {
                ++n;
                mean_x = x;
                mean_y = y;
                cov = 0;
                while ( std::cin >> x && std::cin >> y ) {
                        prev_mean_x = mean_x;
                        prev_mean_y = mean_y;
                        prev_cov = cov;
                        ++n;
                        mean_x = prev_mean_x - ( prev_mean_x - x ) / n;
                        mean_y = prev_mean_y - ( prev_mean_y - y ) / n;
                        cov = prev_cov - ( prev_cov - ( x - mean_x ) * ( y - prev_mean_y ) ) / n;
                }
        }

        std::cout << "n:      " << n << '\n';
        std::cout << "mean_x: " << std::setprecision( 17 ) << mean_x << '\n';
        std::cout << "mean_y: " << std::setprecision( 17 ) << mean_y << '\n';
        std::cout << "cov:    " << std::setprecision( 17 ) << cov    << '\n';

}
```

Example of data.txt:

```
-281.189        612.083
974.663         -24.0965
25.8526         401.539
.               .
.               .
.               .
```

Command Line:

```
g++ -o main.exe main.cpp -std=c++11 -march=native -O3 -Wall -Wextra -Werror -static
./main.exe < data.txt
```

Note: Mathematica's `Covariance[]` function computes the ***unbiased*** sample covariance matrix, not the ***biased*** sample covariance matrix; therefore, the biased sample covariance matrix is computed in Mathematica as:

```
( ( Length[ list ] - 1 ) / Length[ list ] ) * Covariance[ list ]
```