Online Covariance

by Joshua Burkholder

Given the following set of two-dimensional inputs:



Let  be the number of two-dimensional inputs,  represent the  dimension,  represent the  dimension,  be the biased sample covariance of the  and  dimensions for the first  two-dimensional inputs,  be the biased sample covariance of the  and  dimensions for the first  two-dimensional inputs,  be the  value of the -th two-dimensional input,  be the sample mean of the  values for the first  two-dimensional inputs,  be the  value of the -th two-dimensional input, and  be the sample mean of the  values for the first  two-dimensional inputs. Then, the recurrence equation for the biased sample covariance (a.k.a. online covariance) is:



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

.

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 -dimensional input in a later post.

Proof:

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

.

If we expand this definition, we have:

.

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

 and ,

then we have:



Since the biased sample covariance of the  and  dimensions for the first  two-dimensional inputs is defined as:

,

then we also have:

.

With this, we have:



Since the sample mean for the first   and  values are defined as:

 and ,

then we also have:

 and .

With that, we have:



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

,

then we have:



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



then we have:



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



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

,

,

and



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 ]