Homework: Online Statistics And Regressions

May 6, 2019

In this homework set, we will compare online (or "update") algorithms in performing statistics and regressions.

1 Math and Coding

Load the `sparse_narrow` books from the class website and form mid prices. Join all three and then add data columns with time-based EWMAs of the mid prices at a 300 second characteristic time. Add data columns with exponentially-weighted moving returns based on them. Now do the same using count-based exponentially-weighted moving tuned to roughly the same center of mass.

Add columns with time- and count-based exponentially-weighted moving standard deviations of returns, using the online variance formula

$$\mu_{n+1} = \mu_n + (1 - e^{-\lambda(t_{n+1}-t_n)})(x_{n+1} - \mu_n)$$
$$\sigma^2_{n+1} = e^{-\lambda(t_{n+1}-t_n)} \left(\sigma^2_n + (x_{n+1} - \mu_n)(1 - e^{-\lambda(t_{n+1}-t_n)})(x_{n+1} - \mu_n)\right)$$

Write a function `full_regression()` with arguments `y`, `X` and `w` that computes the coefficients β of a weighted linear regression given a vector y and the design matrix X, with the assumption that y is in the first column. You can base this on built-in weighted regression routines because you will only be using it to check your work below.

Next write a function `regression_dispersion()` with the same arguments that computes the dispersion matrix.

Now write a function `regression_update()` with arguments `beta_previous`, `dispersion_previous`, `y_new`, `x_new`, `y_old`, `x_old` , `f` that computes new

coefficients β_{n+1} of an exponentially-weighted moving regression with $\lambda =$`f` using update formulas derived from the Sherman-Morrison identity[1].

Check that your code for `full_regression()` and `regression_update()` is correct by comparing to built-in routines.

Now you can add columns with time- and count-based exponentially-weighted moving regression betas of ETH-USD versus BTC-USD and ETH-BTC. Use the standard deviations and regression coefficients you obtained to compute exponentially-weighted moving correlations as well.

2 Analysis

Study and contrast the statistics of returns and regressions computed in time- versus count-based EWM windows.

[1] It is best to force the x variables to be matrices, rather than letting them be vectors. This avoids problems with silent dimensional collapse in dot products.