Deriving a moving-window method for calculating slope of the line of best-fit for arbitrary window size in $O(n)$ time

Peter Boothe, who uses LaTeX for thinking

15 September 2016

The formula for the slope of a line of best fit (let's call that slope β) in a sample is:

$$\beta = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

We would like to calculate successive values of β from previous ones in a system where the x_i are just successive integers. For simplicity, let's use $x_i = i$ and calculate β from k to $n+k$, which we will write as $\beta_{k,n+k}$.

The first thing to attack is the denominator, because it ends up just being a constant we can calculate once. To see this, we note that when $x_i = i$ we can calculate the average x value between k and $n+k$ to be

$$\overline{x_{k,n+k}} = \frac{n+2k}{2}$$

Pleasantly, $\forall k$, it is true that

$$\sum_{i=k}^{n+k} \left(i - \frac{n+2k}{2} \right)^2 = \sum_{i=1}^{n} \left(i - \frac{n+1}{2} \right)^2$$

the proof of which is left as an exercise for the reader. Therefore, let us make a simplifying definition of:

$$\sigma_n^2 = \sum_{i=1}^{n} \left(i - \frac{n+1}{2} \right)^2$$

We use σ_n^2 because this is the variance (the square of the standard deviation) of the line $y = x$ in our window of size n, and σ is the traditional symbol for standard deviation[1]. Also, because it is a constant, we need only calculate its value once for a given n, and we will then have its value for every window of size n.

[1] This identity between the equation and the variance for $y = x$ underpins a geometrical version of the exercise that we left for the reader.

Now, let us focus our attention on $\beta_{k,n}$, that is, the slope of best fit for a window of data of size n that is offset from 1 by k indices. Because summations in math are inclusive of both ends of their range, and loops in programming languages are generally inclusive of the bottom end and exclusive of the top end, we'll have to be careful about off-by-one errors.

$$\beta_{k,n} = \frac{\sum_{i=k+1}^{n+k}\left(i - \frac{n+2k+1}{2}\right)\left(y_i - \overline{y_{k+1,n+k}}\right)}{\sigma_n^2}$$

We would like to find a closed-form equation, that can be calculated in constant time, for the difference between two successive values of β. Then, we could use this equation to bootstrap the calculation of the β of the first window into a calculation of β for every subsequent window, so that we might calculate every possible window (of a certain fixed size) of β on a dataset of size N in $O(N)$ time. This would speed up the "performance degradation" check in signal searcher, as well as put it on a firm mathematical footing.[2]

Therefore, what we are looking for is the simplest possible way of calculating $\beta_{k+1,n} - \beta_{k,n}$.

To begin, let's ask Mathematica for help, and ignore the σ_n^2 in the denominator while we do so. In the Mathematica code in bold, "Distribute" is a command that distributes a multiplication across an addition, and "FullSimplify" attempts to combine and cancel terms to the best of Mathematica's ability.

When we use these tools in Mathematica, we get:

Define our function:

```
betaslope[k_, n_]:=Sum[Distribute[(i − (n + 2k + 1)/2)(Subscript[y, i] − OverBar[Subscript[y, k, k + n]])],
    {i, k + 1, n + k}]
```

Verify that it looks the way we want it to:

```
betaslope[k, n]
```

$$\sum_{i=1+k}^{k+n}\left(-i\overline{y_{k,k+n}} - \tfrac{1}{2}(-1-2k-n)\overline{y_{k,k+n}} + iy_i + \tfrac{1}{2}(-1-2k-n)y_i\right)$$

Verify that the subtraction of the two terms (before simplification) looks correct:

```
betaslope[k + 1, n] − betaslope[k, n]
```

$$-\sum_{i=1+k}^{k+n}\left(-i\overline{y_{k,k+n}} - \tfrac{1}{2}(-1-2k-n)\overline{y_{k,k+n}} + iy_i + \tfrac{1}{2}(-1-2k-n)y_i\right) + \sum_{i=2+k}^{1+k+n}\left(-i\overline{y_{1+k,1+k+n}} - \tfrac{1}{2}(-1-2(1+k)-n)\overline{y_{1+k,1+k+n}} + iy_i + \tfrac{1}{2}(-1-2(1+k)-n)y_i\right)$$

Ask Mathematica to maximally distribute terms.

```
Distribute[betaslope[k + 1, n]] − Distribute[betaslope[k, n]]
```

$$\tfrac{1}{2}(-1-2k-n)n\overline{y_{k,k+n}} + \tfrac{1}{2}n(1+2k+n)\overline{y_{k,k+n}} - \tfrac{1}{2}(-1-2(1+k)-n)n\overline{y_{1+k,1+k+n}} - \tfrac{1}{2}n(3+2k+n)\overline{y_{1+k,1+k+n}} - \sum_{i=1+k}^{k+n} iy_i + \sum_{i=2+k}^{1+k+n} iy_i - \sum_{i=1+k}^{k+n}\tfrac{1}{2}(-1-2k-n)y_i +$$

[2] It works right now, but I can't guarantee it won't miss a few small instances of performance degradation that happen to span the measurement boundaries in just the right way.

$\sum_{i=2+k}^{1+k+n} \frac{1}{2}(-1-2(1+k)-n)y_i$

Ask Mathematica to simplify the result as much as possible.

FullSimplify[Distribute[betaslope[*k* + 1, *n*]] − Distribute[betaslope[*k*, *n*]]]

$-\sum_{i=1+k}^{k+n} iy_i + \sum_{i=2+k}^{1+k+n} iy_i - \sum_{i=1+k}^{k+n} \frac{1}{2}(-1-2k-n)y_i + \sum_{i=2+k}^{1+k+n} \frac{1}{2}(-1-2(1+k)-n)y_i$

Hooray! We already see that to update the slope, we don't have to keep track of the moving average! Mathematica has declined to work sensibly with the bounds of summation, and also declined to move terms from the inside of the summation to the outside, so we have to do the remaining steps by hand.

$$-\sum_{i=1+k}^{k+n} iy_i + \sum_{i=2+k}^{1+k+n} iy_i - \sum_{i=1+k}^{k+n} \frac{1}{2}(-1-2k-n)y_i + \sum_{i=2+k}^{1+k+n} \frac{1}{2}(-1-2(1+k)-n)y_i$$

$$\left(-\sum_{i=1+k}^{k+n} iy_i + \sum_{i=2+k}^{1+k+n} iy_i\right) + \left(-\sum_{i=1+k}^{k+n} \frac{1}{2}(-1-2k-n)y_i + \sum_{i=2+k}^{1+k+n} \frac{1}{2}(-1-2(1+k)-n)y_i\right)$$

Let's work on each of the parenthesized expressions in turn. First,

$$-\sum_{i=1+k}^{k+n} iy_i + \sum_{i=2+k}^{1+k+n} iy_i$$

$$\left(-(1+k)y_{1+k} + -\sum_{i=2+k}^{k+n} iy_i\right) + \left(\sum_{i=2+k}^{k+n} iy_i + (1+k+n)y_{1+k+n}\right)$$

$$-(1+k)y_{1+k} + \left(-\sum_{i=2+k}^{k+n} iy_i + \sum_{i=2+k}^{k+n} iy_i\right) + (1+k+n)y_{1+k+n}$$

$$-(1+k)y_{1+k} + (1+k+n)y_{1+k+n}$$

Hooray! An $O(1)$ closed form! Now for the second expression...

$$-\sum_{i=1+k}^{k+n} \frac{1}{2}(-1-2k-n)y_i + \sum_{i=2+k}^{1+k+n} \frac{1}{2}(-1-2(1+k)-n)y_i$$

$$-\frac{1}{2}(-1-2k-n)\sum_{i=1+k}^{k+n} y_i + \frac{1}{2}(-1-2(1+k)-n)\sum_{i=2+k}^{1+k+n} y_i$$

$$-\frac{1}{2}(-1-2k-n)\left(y_{1+k} + \sum_{i=2+k}^{k+n} y_i\right) + \frac{1}{2}(-3-2k-n)\left(\sum_{i=2+k}^{k+n} y_i + y_{1+k+n}\right)$$

$$\frac{1+2k+n}{2}\left(y_{1+k}+\sum_{i=2+k}^{k+n} y_i\right)+\frac{-3-2k-n}{2}\left(\sum_{i=2+k}^{k+n} y_i+y_{1+k+n}\right)$$

$$\frac{1+2k+n}{2}y_{1+k}+\frac{1+2k+n}{2}\sum_{i=2+k}^{k+n} y_i+\frac{-3-2k-n}{2}\sum_{i=2+k}^{k+n} y_i+\frac{-3-2k-n}{2}y_{1+k+n}$$

$$\frac{1+2k+n}{2}y_{1+k}+\left(\frac{1+2k+n}{2}+\frac{-3-2k-n}{2}\right)\sum_{i=2+k}^{k+n} y_i+\frac{-3-2k-n}{2}y_{1+k+n}$$

$$\frac{1+2k+n}{2}y_{1+k}+(-1)\sum_{i=2+k}^{k+n} y_i+\frac{-3-2k-n}{2}y_{1+k+n}$$

$$\frac{1+2k+n}{2}y_{1+k}+-\sum_{i=2+k}^{k+n} y_i+\frac{-3-2k-n}{2}y_{1+k+n}$$

This is a mild bummer: we can't simplify it any further (or, at least, I can't) and there's still a summation symbol in there. However, if we keep track of what this sum was for the previous calculation, then updating it for this current calculation should be as simple as a single subtraction and a single addition. So while it looks bad, it's also $O(1)$ after everything has been calculated for the first window, so we will use this result. Putting these two halves together, we get:

$$\left(-\sum_{i=1+k}^{k+n} iy_i+\sum_{i=2+k}^{1+k+n} iy_i\right)+\left(-\sum_{i=1+k}^{k+n}\frac{1}{2}(-1-2k-n)y_i+\sum_{i=2+k}^{1+k+n}\frac{1}{2}(-1-2(1+k)-n)y_i\right)$$

$$(-(1+k)y_{1+k}+(1+k+n)y_{1+k+n})+\left(\frac{1+2k+n}{2}y_{1+k}+-\sum_{i=2+k}^{k+n} y_i+\frac{-3-2k-n}{2}y_{1+k+n}\right)$$

This can be simplified a little by making sure each y_i only appears once and collecting their factors:

$$(-(1+k)y_{1+k}+(1+k+n)y_{1+k+n})+\left(\frac{1+2k+n}{2}y_{1+k}+-\sum_{i=2+k}^{k+n} y_i+\frac{-3-2k-n}{2}y_{1+k+n}\right)$$

$$\left(-(1+k)+\frac{1+2k+n}{2}\right)y_{1+k}+\left(1+k+n+\frac{-3-2k-n}{2}\right)y_{1+k+n}+-\sum_{i=2+k}^{k+n} y_i$$

$$\left(\frac{-2-2k}{2}+\frac{1+2k+n}{2}\right)y_{1+k}+\left(\frac{2+2k+2n}{2}+\frac{-3-2k-n}{2}\right)y_{1+k+n}+-\sum_{i=2+k}^{k+n} y_i$$

$$\left(\frac{-1+n}{2}\right)y_{1+k}+\left(\frac{-1+n}{2}\right)y_{1+k+n}+-\sum_{i=2+k}^{k+n} y_i$$

$$\frac{n-1}{2}\left(y_{1+k}+y_{1+k+n}\right)+-\sum_{i=2+k}^{k+n} y_i$$

Hooray!

Finally, we now know that

$$\beta_{k+1,n}-\beta_{k,n}=\frac{\frac{n-1}{2}\left(y_{1+k}+y_{1+k+n}\right)+-\sum_{i=2+k}^{k+n} y_i}{\sigma_n^2}$$

The fact that this formula contains only terms which are $O(1)$ to calculate and sums which are $O(1)$ to update means that we can calculate $\beta_{k+1,n}-\beta_{k,n}$ in $O(1)$ time. If we can calculate each successive difference in $O(1)$ time, then we can calculate $\beta_{k,n+k}$ for all possible k and fixed window size n in time $O(n+k)$.