Effect of the zero-mean assumption on the correlation coefficient

---

In the Python and Matlab wrappers, the templates T are forced to be zero-mean but the continuous data is not (*i.e.* each sliding window does not always have a mean that is very close to zero). We can write the data $u(t)$ as $u(t) = \bar{u}(t) + \delta u(t)$, where $\bar{u}$ is the mean of the data within the sliding window associated with the time t, and δu is the deviation from the mean. By construction, δu is a zero-mean signal.

The full definition of the correlation coefficient, where the mean is removed from each sliding window, is:

$$\mathrm{CC} = \frac{\sum_{i=1}^{N} \delta u(t_i) T(t_i)}{\sqrt{\sum_{i=1}^{N} \delta u(t_i)^2 \sum_{i=1}^{N} T(t_i)^2}} \tag{1}$$

Our definition of the correlation coefficient, where we assume the mean within each sliding window is already zero, is:

$$\begin{aligned} \mathrm{CC}_0 &= \frac{\sum_{i=1}^{N} (\bar{u} + \delta u(t_i)) T(t_i)}{\sqrt{\sum_{i=1}^{N} (\bar{u} + \delta u(t_i))^2 \sum_{i=1}^{N} T(t_i)^2}} \\ &= \frac{\bar{u} \sum_{i=1}^{N} T(t_i) + \sum_{i=1}^{N} \delta(t_i) T(t_i)}{\sqrt{\left(N\bar{u}^2 + 2\bar{u} \sum_{i=1}^{N} \delta u(t_i) + \sum_{i=1}^{N} \delta u(t_i)^2\right) \sum_{i=1}^{N} T(t_i)^2}} \end{aligned} \tag{2}$$

Because both T and δu are zero-mean, all the terms that are proportional to one of those vanish, and the last expression simplifies to:

$$\mathrm{CC}_0 = \frac{\sum_{i=1}^{N} \delta u(t_i) T(t_i)}{\sqrt{\left(\sum_{i=1}^{N} \delta u(t_i)^2 \sum_{i=1}^{N} T(t_i)^2\right) + \left(N\bar{u}^2 \sum_{i=1}^{N} T(t_i)^2\right)}} \tag{3}$$

Because:

$$\sqrt{\left(\sum_{i=1}^{N} \delta u(t_i)^2 \sum_{i=1}^{N} T(t_i)^2\right) + \left(N\bar{u}^2 \sum_{i=1}^{N} T(t_i)^2\right)} \geq \sqrt{\left(\sum_{i=1}^{N} \delta u(t_i)^2 \sum_{i=1}^{N} T(t_i)^2\right)},$$

Equation 3 shows that, if the templates are zero-mean, then the CCs might be biased toward lower values compared to when we explicitly remove the mean of the data within each sliding window.

When the data are highpass filtered, we could expect the mean to be zero, but the zero-mean assumption still breaks down when high amplitudes are present in the seismograms. In the following, we show an example from the Kaikoura earthquake (M7.8) to illustrate what impact the assumption's break down has on the CC

computation.

First, Figure 1 shows the spectral content of the data and the sliding window's duration (*i.e.* the template's duration) we use in this test.

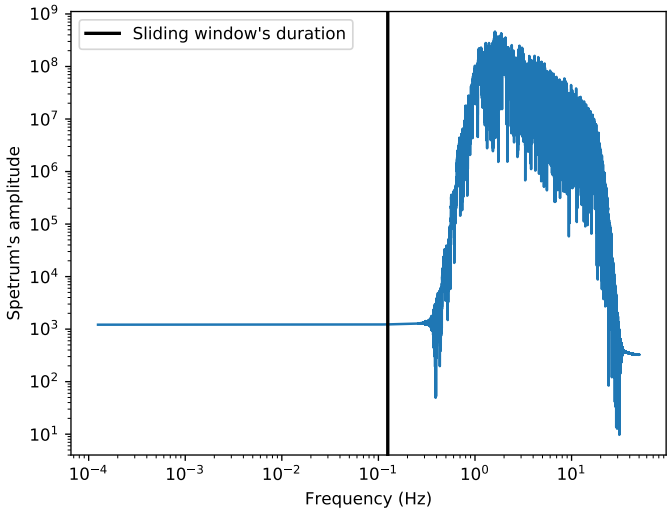

Figure 1: Spectrum of the seismogram: the data are highpass filtered such that periods lower then the sliding window's duration (=8s) are removed.

Let's define CSUM_0 the cumulative sum of the squared data under the zero-mean assumption (Eq. 4), and CSUM the cumulative sum of the squared data with the mean removed within each sliding window (Eq. 5).

$$\mathrm{CSUM}_0(t_n) = \sum_{i=1}^{N} \left(\delta u(t_{n+i}) + \bar{u}(t_n)\right)^2 \tag{4}$$

$$\mathrm{CSUM}(t_n) = \sum_{i=1}^{N} \delta u(t_{n+i})^2 \tag{5}$$

The relative difference between the CCs computed with our definition (Eq. 4) and the full definition (Eq. 5) is given by their ratio:

$$\begin{aligned}
\frac{\mathrm{CC}(t_n)}{\mathrm{CC}_0(t_n)} &= \frac{\sum_{i=1}^{N} \delta u(t_{n+i}) T(t_i)}{\sum_{i=1}^{N} \delta u(t_{n+i}) T(t_i)} \times \frac{\sqrt{\sum_{i=1}^{N} (\bar{u} + \delta u(t_i))^2 \sum_{i=1}^{N} T(t_i)^2}}{\sqrt{\sum_{i=1}^{N} \delta u(t_i)^2 \sum_{i=1}^{N} T(t_i)^2}} \\
&= \sqrt{\frac{\sum_{i=1}^{N} \left(\bar{u}(t_n) + \delta u(t_{n+i})\right)^2}{\sum_{i=1}^{N} \delta u(t_{n+i})^2}} \\
&= \sqrt{\frac{\mathrm{CSUM}_0(t_n)}{\mathrm{CSUM}(t_n)}}
\end{aligned} \tag{6}$$

As we can see on Equation 6, the relative difference between our definition of the CC and the usual one only involves the ratio of the cumulative sums. Figure 2 shows that even though very large differences can be seen between the cumulative sum computed with Equation 4 and with Equation 5, their ratio stays close to 1 (maximum around ≈ 1.025).

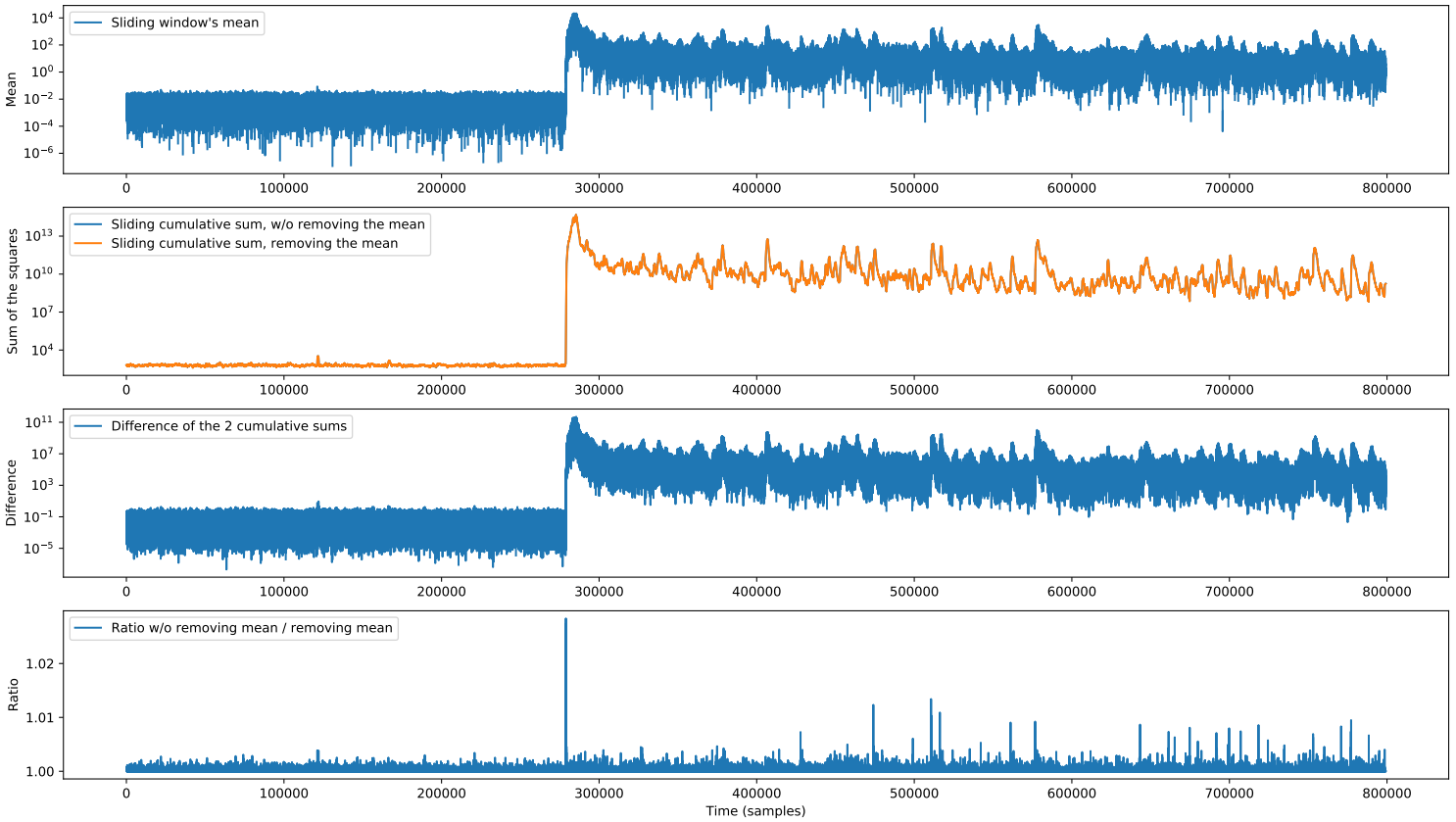

Figure 2: From top to bottom: **First panel:** sliding mean; strong deviations from the zero-mean assumptions are observed during and after the Kaikoura earthquake. **Second panel:** Sliding cumulative sum of the squares of the data *without* removing the mean within each sliding window (*blue* curve), and removing the mean within each sliding window (*orange* curve). The two curves are not distinguishable at this scale. **Third panel:** Difference of the two sliding cumulative sums (the orange curve was subtracted to the blue one, *i.e.* Eq. 4 - Eq. 5). The difference reaches very large numbers, but this doesn't provide insightful information on how it should impact the CC computation. **Fourth panel:** Ratio of the two cumulative sums (blue curve / orange curve, *i.e.* Eq. 6).

The ratio in Equation 6 can be turned into a relative error by doing $\mathcal{E}(t) = \left(\frac{\mathrm{CC}(t)}{\mathrm{CC}_0(t)} - 1\right) \times 100\%$. In that case, the maximum error we make is $\max_t \{\mathcal{E}\} = (\sqrt{1.025} - 1) \times 100\% \approx 1.2\%$. An example based on a single earthquake might not be enough to prove that the error is always small, but we expect the error to be small whenever the typical order of magnitude of δu is much greater than the mean $\bar{u}$. In mathematical terms, this corresponds to the cases when the first statistical moment of the continuous data (the mean) is much smaller than the (square root of the) second statistical moment of the continuous data (the standard deviation). As an illustration of that statement, we show on Figure 3 the sliding standard deviation and the sliding mean;

we can observe that the standard deviation is always several orders of magnitude higher than the mean. We further state that the situation where $\frac{\text{std}}{\text{mean}} \gg 1$ is the typical situation for seismic signals, when the data are highpass filtered.

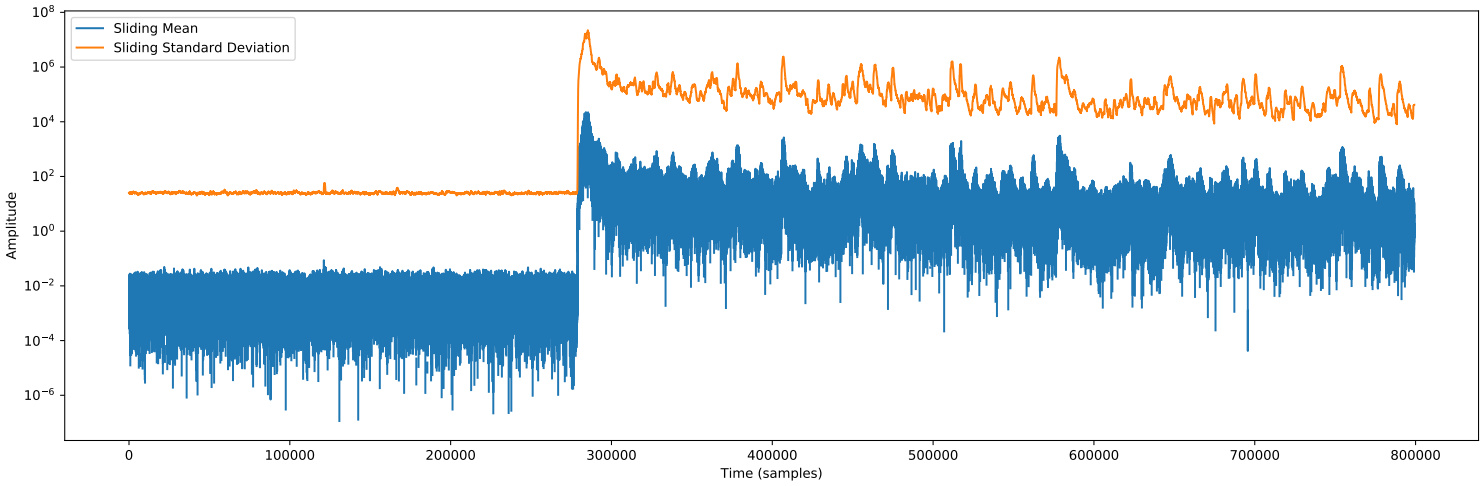

Figure 3: Comparison of the sliding mean against the sliding standard deviation. Even though the mean sometimes shifts considerably away from zero, it is always several orders of magnitude lower than the standard deviation. The contribution of the sliding mean in the correlation coefficient is then always negligible compared to the sliding standard deviation's.

This leads us to the final conclusion that:

- the zero-mean assumption only bias the correlation coefficients toward zero, and thence does not favor false detection (false positives),
- the zero-mean assumption affects little the precision of the correlation coefficient computation (error of $\approx 1.2\%$ on a M7.8 earthquake).