> Size matters <

Finding the most informative set of window lengths

Jefrey Lijffijt, Panagiotis Papapetrou and Kai Puolamäki

Aalto University, Finland and Birkbeck University of London, UK

> Problem setting

Event sequences often contain variability at different levels. Figure 1 gives an example of a sequence with multi-scale trends. Choosing the *length of a sliding window* is difficult yet important.

> Method

Let $X_\omega(i)$ be the sub-sequence of sequence X with window length ω starting at index i.

For a set of window sizes Ω and a set of indices I compute $f\left(X_\omega(i)\right)$, for all $\omega \in \Omega$, $i \in I$, where f is a statistic parameterized by the length of the window.

Define distance $d(\omega_s,\omega_t) = \sum_{i \in I}\left(f\left(X_{\omega_s}(i)\right) - f\left(X_{\omega_t}(i)\right)\right)^2$.

Now, we want to *find the set of k window lengths that explain most of the variation in X*. Or, equivalently:

$$\underset{R:R\subseteq\Omega,|R|=k}{\text{minimize}} \sum_{\omega_s \in \Omega} \min_{\omega_t \in R} d(\omega_s,\omega_t)$$

Optimization algorithm:

- Compute k-means clustering with Lloyd's algorithm
- Add the window lengths closest to each centroid
- Repeat *rep* times and choose the best solution

> Solution stability

We tested the stability of the optimization algorithm in a series of experiments using synthetic data.

1) Test the stability on a single Bernoulli sequence.
2) Test the stability over similar sequences.
3) Test the dependency on the event frequency.
4) Test how many data samples are required.

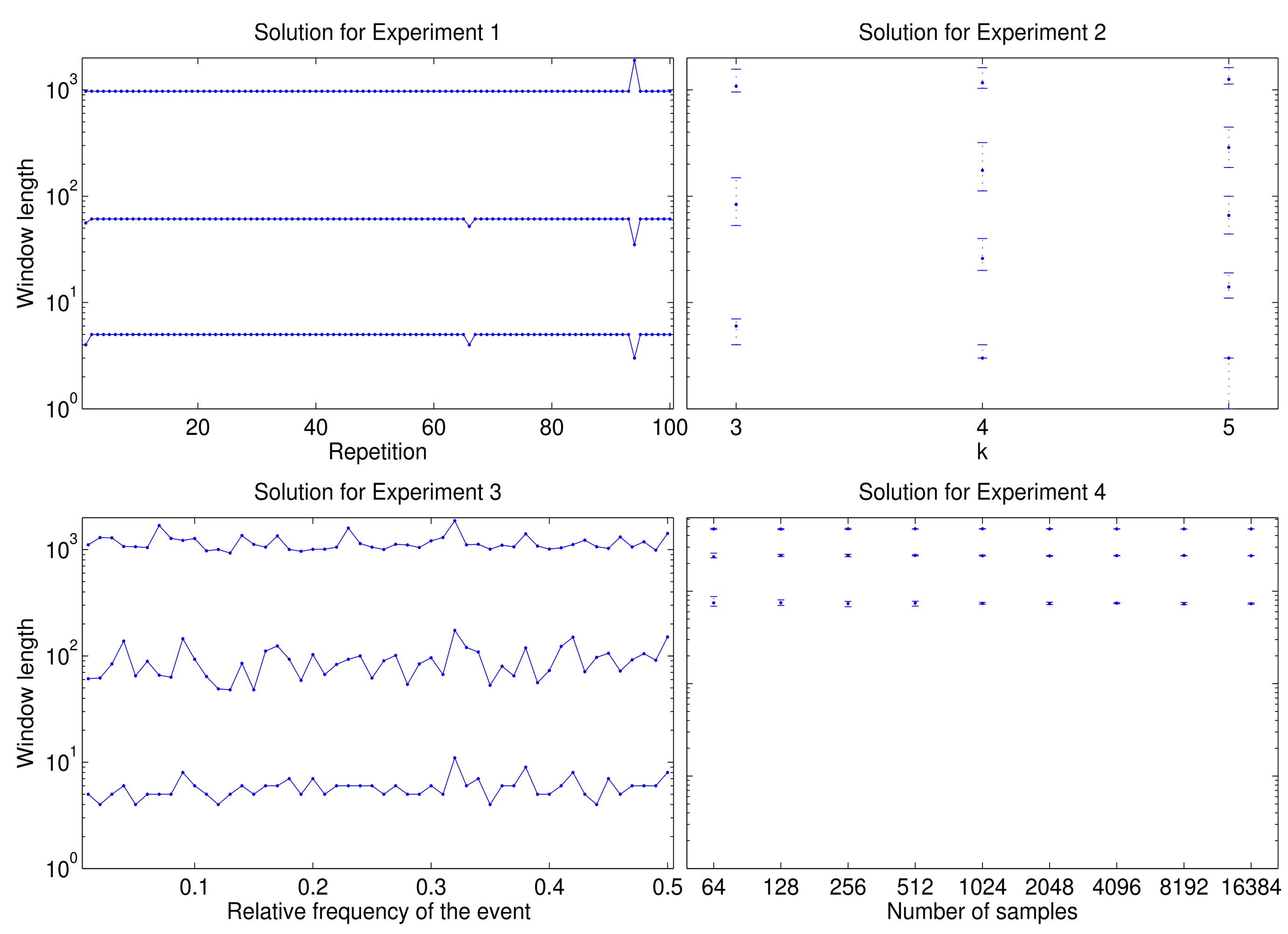

Figure 2: Results from the four experiments. Experiments 1 and 2 show that the optimization algorithm is highly robust. Experiment 3 shows that, when tracing relative event frequencies, the solution is independent of the absolute frequency. Finally, experiment 4 shows that 1,000 samples is sufficient to have almost no uncertainty.

> Solution

We propose to use an *optimized set of window lengths* that summarizes all other possibly interesting window lengths.

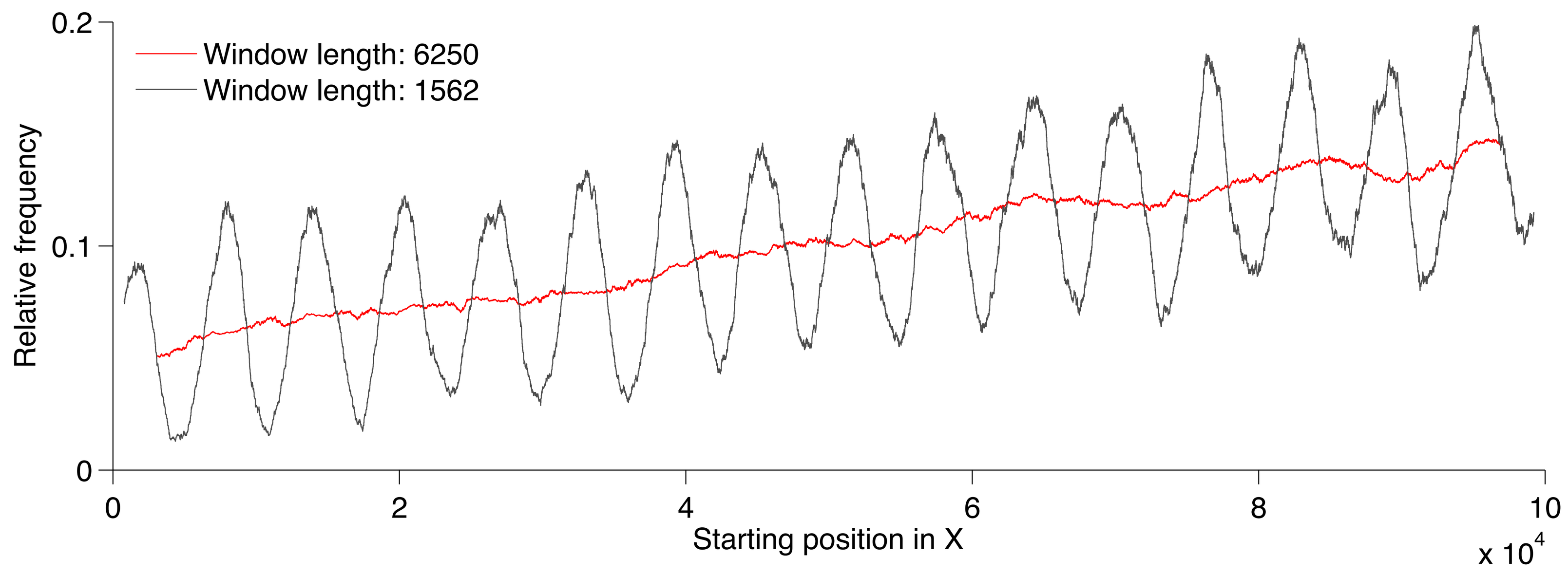

Figure 1: Example of the relative frequency of an event in an event sequence. We observe that the sequence contains two trends: a slow increase (red line) and a rhythmic component (grey line).

> Burstiness of words

We computed the optimal sets of window lengths for several bursty and non-bursty words in Jane Austen's *Pride & Prejudice*. Burstiness is estimated by computing the MLE for the Weibull distribution on the inter-arrival times of a word.

Frequency	Non-bursty	Index	Bursty	Index
Low [39–41]	met, rest, right, help	1–4	write, de, william, read	5–8
Medium [175–228]	time, soon, other, only	9–12	lady, has, can, may	13–16
High [600–1666]	with, not, that, but	17–20	you, is, my, his	21–24

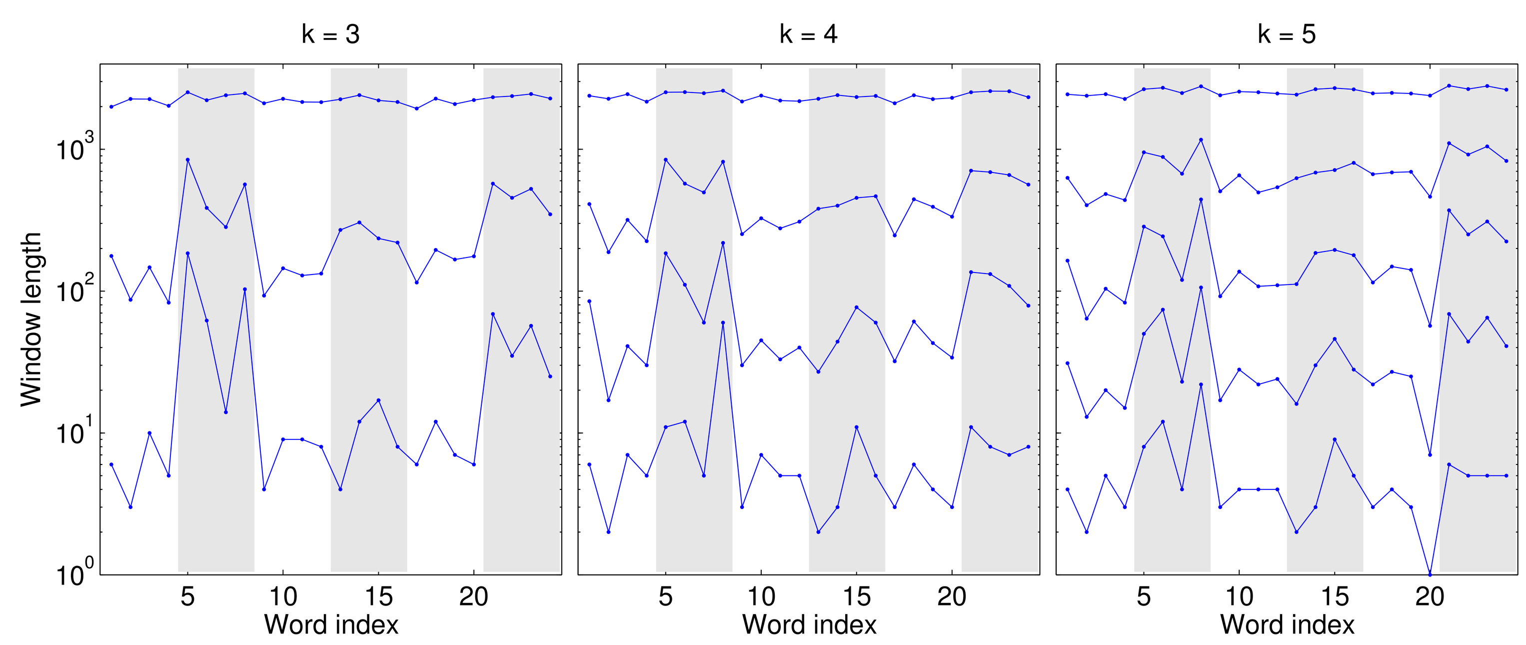

Figure 3: Bursty words give longer window lengths, because the scale structure is less gradual then for uniformly distributed words.

> Other experiments

Type/token ratio throughout several novels of Charles Dickens.

Frequency of (di-)nucleotides in DNA of Homo Sapiens and Canis Familiaris.

> Acknowledgements

This work has been funded by the Finnish Centre-of-Excellence in Algorithmic Data Analysis (ALGODAN) and the Finnish Doctoral Programme in Computational Sciences (FICS). We thank Heikki Mannila for useful feedback and discussions.

ECML-PKDD, Bristol, UK, September 24-28, 2012