4. Approximate solutions

- Basic dynamic programming algorithm: $O(n^2k)$ time for finding k segments from a sequence of length n
- Is this feasible?
- Heuristics?
- Top-down segmentation
- Randomized methods
- Pruning the set of possible boundary points

Top-down segmentation

- For $b = 1, \ldots, k-1$: find the best place for the bth segment boundary
- At stage b: the sequence is divided into b parts
- Go through the indices $i = 1, \ldots, n-1$ and see how much the error decreases if a new segment is started from i
- Select the best point and loop
- D. Hawkins, Point estimation of parametres of piecewise regression models. Journal of the Royal Statistical Society 1976.; V. Guralnik and J. Srivastava: Event detection from time series data. KDD 1999.

Example

$k = 3$, L_1 metric, sequence

0 0 0 0 0 0 0 0 0 0 0 1 2 3 3 3 3 3 3 3 3 3

Properties of top-down segmentation

- Does not always produce an optimal segmentation
- The best boundary for 2 segments is not necessarily among the optimal set of boundary points for k segments
- Example?
- Running time? $O(nk)$, considerably better than $O(n^2k)$
- Details?

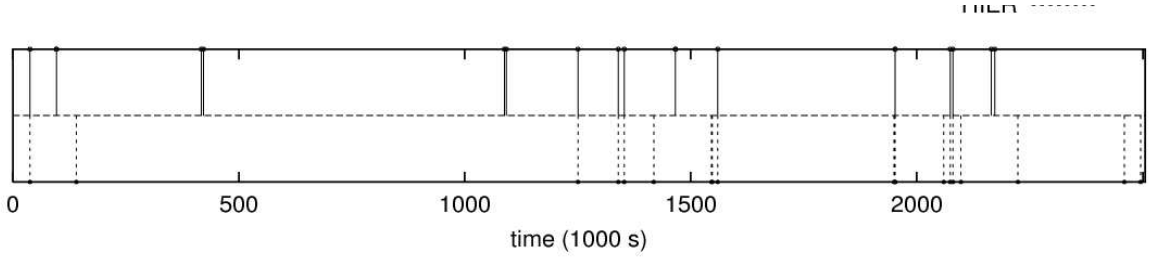


Fig. 5. Change points of the divisions up to 20 pieces generated by the dynamic programming (DP) and hierarchical (HIER) algorithms. The data set contains 46,662 events.

Randomized algorithms

- How to find a good k-segmentation?
- Start with a random k-segmentation
- Take one boundary point away and search for the best place for it
- Repeat until convergence
- Works quite well, almost optimal results
- Running time $O(nI)$, where I is the number of iterations
- Any provable properties
- Johan Himberg, Kalle Korpiaho, Heikki Mannila, Johanna Tikanmäki, and Hannu T.T. Toivonen. Time series segmentation for context recognition in mobile devices In The 2001 IEEE International Conference on Data Mining (ICDM'01), 203 - 210.

Pruning the set of possible boundary points

- If a part of sequence is locally very flat, it is not likely that the best segmentation will have a boundary in that part
- Idea: do some local preprocessing and leave only the points that seem useful
- Works quite well

5. Looking at all segmentations

- Total probability of data, given that there are k segments
- Posterior probability of a segment starting from a given point, fixed k
- Posterior probability of a segment starting from a given point, variable k

Best solution vs. all solutions

- $L(a_1, a_2, \ldots, a_n|p) = \prod_{i=1}^{n} p^{a_i}(1-p)^{1-a_i} = p^{n_1}(1-p)^{n_0}$
- Finding the best solution:

  $$\text{cost}(1, j, h) = \min_{l=i+1}^{j-1} \text{cost}(1, l, h-1) + \text{cost}(l+1, j, 1).$$

- In probability form:

  $$Pr(1, j, h) = \max_{l=i+1}^{j-1} Pr(1, l, h-1) L(l+1, j, 1).$$

- What about the total probability of all segmentations with h segments?

  $$TPr(1, j, h) = \sum_{l=i+1}^{j-1} TPr(1, l, h-1) L(l+1, j, 1).$$

Example

- Sequence 0 0 1 0; 2 segments
- $\log_2(1/3) = -1.59$, $\log_2(2/3) = -0.59$

Posterior probability of a segment boundary

- Assume there are k segments
- How likely is it a segment starts at position i?
- Optimal solution tells just one segmentation
- Others might be equally good
- How to estimate the probability of a segment boundary?

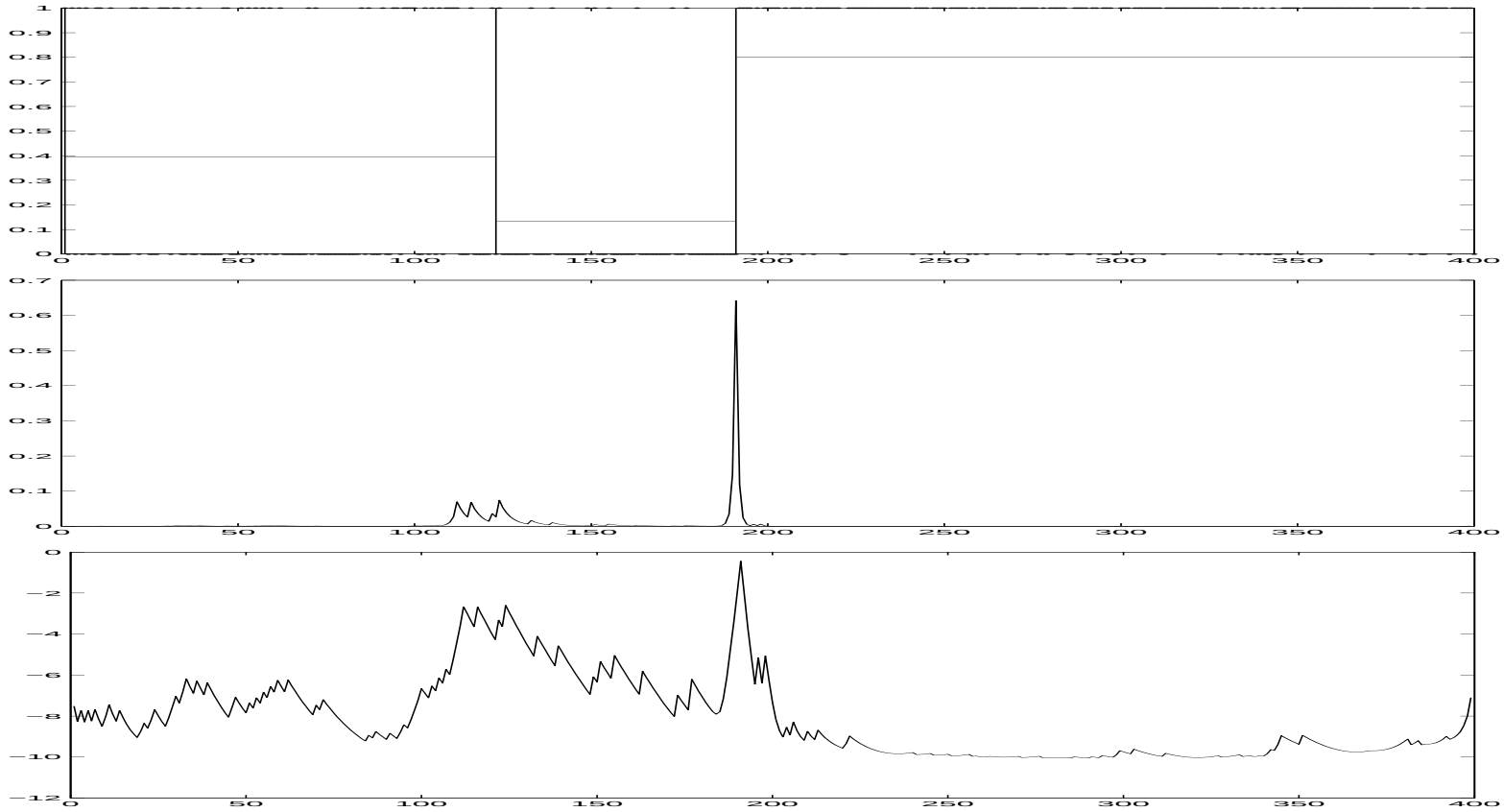
Posterior probability of a segment boundary

- $BP(i|k) =$ the probability of a segment starting from position 1, given that there are k segments
- $$BP(i|k) = \frac{\sum_{b=1}^{k-1} TPr(1,i,b)TPr(i+1,n,k-b)}{TPr(1,n,k)}$$
- A segment starts at i iff for some b with $0 < b < k$ there are b segments before i and $k - b$ segments after i
- The total probability of such segmentations is $\sum_{b=1}^{k-1} TPr(1,i,b)TPr(i+1,n,k-b)$
- The conditional probability of a segmentation with a segment starting at i is the above

Example, continuation

- What is the probability that a segment starts in position 2 of sequence 0 0 1 0?

Model, optimal segmentation, and posterior prob



Model, optimal segmentation, and posterior prob