Forum 2012
13 NOVEMBRE 2012
A l'Ecole Polytechnique

digiteo
Recherche en sciences & technologies de l'information

Toby Dylan HOCKING, Guillem RIGAILL, Francis BACH, Jean-Philippe VERT

Nº 41

Interactively learning change-point detection models

ABSTRACT / RÉSUMÉ

- Change-point detection models have parameters that control the number of change-points.
- It is often easy to visually identify regions with or without changes.
- Plot and annotate a few signals, then learn a model for accurate change-point detection in the other un-annotated signals.

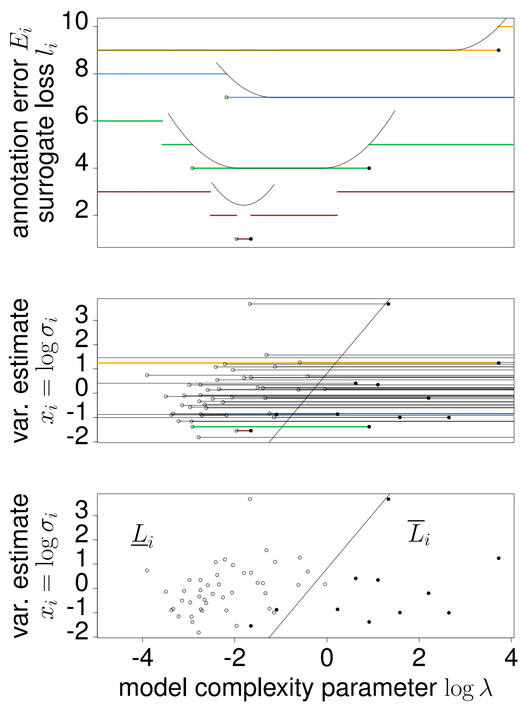
1 Interactive annotation web site

- Goal: detect changes in a database of neuroblastoma tumors.
- Left: user-specific change-point and copy number annotations.
- Right: data export to UCSC genome browser for visualization.

2 Learning problem and convex relaxation

- Model: maximum likelihood segmentation $\hat{y}_i^k$ using a pruned dynamic programming algorithm (Rigaill 2010).
- For each signal y_i, choose the number of segments k using

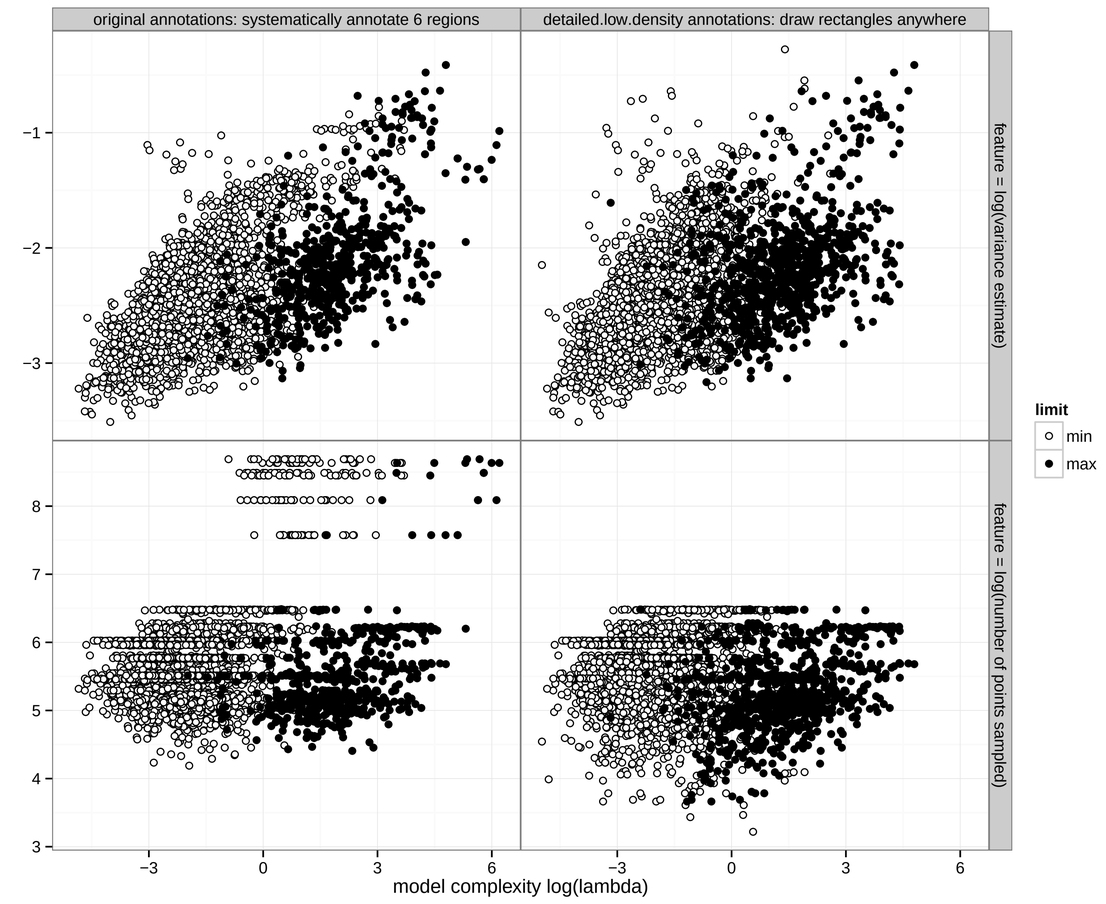
$$k_i^*(\lambda) = \underset{k \in \{1,\ldots,k_{\max}\}}{\arg\min} \ \lambda k + ||y_i - \hat{y}_i^k||_2^2.$$



- Top: For each signal i, calculate a non-convex error curve E_i using its change-point annotations (colored curves).
- Relax to obtain its convex surrogate loss l_i (black curves).
- Middle: Identify the target interval $[\underline{L}_i, \overline{L}_i]$ of minimum error and plot it against a feature x_i.
- Geometric interpretation: find a line $\underline{L}_i < f(x_i) < \overline{L}_i$ that intersects each target interval.
- Bottom: just plot the limits $\underline{L}_i, \overline{L}_i$, and find a separating line.

3 Results on the neuroblastoma data

- Two annotation data sets (panels from left to right) based on `data(neuroblastoma,package="neuroblastoma")` from CRAN.
- In both, model complexity depends on signal variance (top) and number of points sampled (bottom).



Below: learning model complexity $f(x_i)$ as a function of signal features (log.s.log.d, L1-reg) detects changes better than learning a constant λ (cghseg.k).

4 Conclusions and availability

- Fast implementation using accelerated proximal gradient methods (FISTA, Beck and Teboulle 2009).
- State-of-the-art change-point detection performance.
- Live web site at `https://bioviz.rocq.inria.fr/plotter/`
- Source code `https://gforge.inria.fr/projects/breakpoints/`
- Interval regression software at `http://cbio.ensmp.fr/~thocking`

Contact : Toby.Hocking@inria.fr