A generic C++ implementation of the Pruned DPA for segmentation

A. Cleynen, M. Koskas, G. Rigail

March 27th, 2012
Guillem Rigaill [4]

Pruned dynamic programming for optimal multiple change-point detection

Algorithm for the segmentation of datasets:

- **Exact** with respect to given loss
- **Fast**: empirically in $n \log(n)$
- Returns optimal segmentation in 1 to $K_{\text{max}}$ segments
- Allows for a vast range of methods for the choice of $K$
CGH profile, Pruned DPA

- 1 to $K_{max} = 50$
- Runtime = 10.834s
- Lebarbier: [3], Lavielle: [2] :
  \[ pen(K) = \beta K \log \left( \frac{n}{K} \right) \]
- $K = 28$
1 to $K_{max} = 50$
Runtime = 10.834s

mBIC (Zhang and Siegmund [5]):
\[
pen(K) = \beta K \log \left( \frac{n}{K} \right) + g \left( \sum_{r \in \hat{m}(K)} \log n_r \right)
\]

$K = 42$
1 to $K_{\text{max}} = 50$

Runtime = 0.171s

SIC: $\text{pen}(K + 1) = \frac{1}{2} K \log n$

$K = 12$
CGH profile, PELT

- Runtime = 14.651s
- SIC: \( pen(K + 1) = \frac{1}{2}K \log n \)
- \( K = 17 \)
Comparison for K=17

Runtimes:
- PDPA = 3.621s
- CART = 0.062s
- PELT = 14.651s

Breakpoints
- PELT and PDPA: same breakpoints
- CART: only 9 out of 16 are identical
An example

Four-point signal

\[ Y_1 = 0 \quad Y_2 = 0.5 \quad Y_3 = 0.4 \quad Y_4 = -0.5 \]

Contrast \( \gamma(Y, \mu) = (Y - \mu)^2 \)

Segmentation in \( K=2 \) segments
Dynamic Programming approach:

- $\forall t \in \{1, \ldots, n\}$ compute cost of segmentation with last breakpoint $t$ as a function of last-segment parameter $\mu$
- $\forall t \in \{1, \ldots, n\}$ find minimum of cost function in $\mu$
- identify the minimum of those minimums
Main idea:

- If we add a new point, the values of $\mu$ change, but not the best candidates for last breakpoint

$\Rightarrow$ A beaten candidate can never become optimal again
Pruned DPA: The algorithm

Initialization:
\( \forall t \in \{1, n\} \) compute \( C_{1,t} \)

\[ \tau = K - 1 = 1 \]
\( (t = 1 \text{ Signal: } Y_1 = 0) \)

Cost function:
- \( h_{2,1}(\mu) = C_{1,1} = 0 \)

Set of intervals:
- \( S_{2,1}^1 = \mathbb{R} (= [-0.5; 0.5]) \)
New entry:

\[ t = 2 \]

Signal:
\[ Y_1 = 0 \quad Y_2 = 0.5 \]

Cost functions:
- \[ h_{2,2}^1(\mu) = h_{2,1}^1 + (Y_2 - \mu)^2 = 0.25 - \mu + \mu^2 \]
- \[ h_{2,2}^2(\mu) = C_{1,2} = 0.125 \]

Set of intervals:
- \[ S_{2,2}^1 = [0.146; 0.5] \]
- \[ S_{2,2}^2 = [-0.5; 0.146] \]
New entry:

Signal: \( Y_1 = 0 \quad Y_2 = 0.5 \quad Y_3 = 0.4 \)

Cost functions:

- \( h_{2,3}^1(\mu) = h_{2,2}^1 + (Y_3 - \mu)^2 = 0.41 - 1.8\mu + 2\mu^2 \)
- \( h_{2,3}^2(\mu) = h_{2,2}^2 + (Y_3 - \mu)^2 = 0.285 - 0.8\mu + \mu^2 \)
- \( h_{2,3}^3(\mu) = C_{1,3} = 0.14 \)

Set of intervals:

- \( S_{2,3}^1 = [0.190; 0.5] \)
- \( S_{2,3}^2 = \emptyset \)
- \( S_{2,3}^3 = [−0.5; 0.190] \)

\( \tau = 2 \) is discarded
Signal: \( Y_1 = 0 \quad Y_2 = 0.5 \quad Y_3 = 0.4 \quad Y_4 = -0.5 \)

Cost functions:

- \( h_{2,4}^1(\mu) = h_{2,3}^1 + (Y_4 - \mu)^2 = 0.66 - 0.8\mu + 3\mu^2 \)
- \( h_{2,4}^3(\mu) = h_{2,3}^3 + (Y_4 - \mu)^2 = 0.39 + \mu + \mu^2 \)
- \( h_{2,4}^4(\mu) = C_{1,4} = 0.62 \)

Set of intervals:

- \( S_{2,4}^1 = \emptyset \)
- \( S_{2,4}^3 = [-0.5; 0.190] \)
- \( S_{2,4}^4 = [0.190; 0.5] \)

\( \tau = 1 \) is discarded
Last step: minimization

- \( H_{K,t}(\mu) = \min_{\{K-1<\tau<t\}} \{ h^{\tau}_{K,t}(\mu) \} \)
- \( H_{K,t}(\mu) = \begin{cases} 
  0.39 + \mu + \mu^2 & \text{for } \mu \in [-0.5; 0.190] \\
  0.62 & \text{for } \mu \in [0.190; 0.5] 
\end{cases} \)
- \( C_{K,t} = \min_{\mu} \{ H_{K,t}(\mu) \} \)

\[ C_{K,n} (= C_{2,4}) = 0.14 \]
\[ \tau = 3 \]
\[ \mu = -0.5 \]
From the original DPA to the Pruned DPA

Original DPA: segment additivity $\Theta(Kn^2)$

$$C_{k,t} = \min_{\{k-1<\tau<t\}} \left\{ C_{k-1,\tau} + \min_{\mu} \left\{ \sum_{i=\tau+1}^{t} \gamma(Y_i, \mu) \right\} \right\}$$

Pruned DPA: point additivity

$$C_{k,t} = \min_{\mu} \left\{ \min_{\{k-1<\tau<t\}} \left\{ C_{k-1,\tau} + \sum_{i=\tau+1}^{t} \gamma(Y_i, \mu) \right\} \right\}$$

$$= \min_{\mu} \left\{ \min_{\{k-1<\tau<t\}} \left\{ C_{k-1,\tau} + \sum_{i=\tau+1}^{t-1} \gamma(Y_i, \mu) + \gamma(Y_t, \mu) \right\} \right\}$$
Generic C++ implementation
Performances on real datasets

Figure: Chromosome 1, positive strand of S. cerevisiae (yeast)
Performances on real datasets
Performances on real datasets

Neuroblastoma copy number data
Result from Hocking et al. [1]

<table>
<thead>
<tr>
<th>Method</th>
<th>errors</th>
<th>FP</th>
<th>FN</th>
<th>Timing(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDPA-Lavielle</td>
<td>2.2</td>
<td>0.6</td>
<td>11.6</td>
<td>2.10</td>
</tr>
<tr>
<td>Fused Lasso ($\lambda = f(K)$)</td>
<td>6.7</td>
<td>3.6</td>
<td>18.5</td>
<td>0.08</td>
</tr>
<tr>
<td>Circular Binary Segmentation (SD)</td>
<td>11.5</td>
<td>7.6</td>
<td>32.2</td>
<td>51.62</td>
</tr>
<tr>
<td>Fused Lasso ($\lambda = cste$)</td>
<td>16.0</td>
<td>12.7</td>
<td>36.6</td>
<td>0.04</td>
</tr>
<tr>
<td>Circular Binary Segmentation (default)</td>
<td>40.5</td>
<td>49.3</td>
<td>0.5</td>
<td>1.78</td>
</tr>
<tr>
<td>PDPA-mBIC</td>
<td>40.9</td>
<td>49.4</td>
<td>0.0</td>
<td>1.47</td>
</tr>
</tbody>
</table>

**Table:** Comparison of a few segmentation methods on a real data set. Tuning parameters are learned by Leave-one-out.

More methods are compared in [1]
Conclusions and Perspectives

**Conclusion:** PDPA is a fast and exact algorithm that allows the use of:

- a large range of data type (CGH, Seq-data, etc)
- a large range of possible contrasts (Quadratic, Poisson, etc)
- a large range of methods for the choice of K (mBIC, Lavielle, AIC, etc)

**Perspectives:**

- Application to real datasets for the discovery of new transcripts, etc.
- Theoretical proof of the complexity of the algorithm
- Implementation of Ridge-type penalties
The End

Thank you!
References

Toby Dylan Hocking.
Learning smoothing models of copy number profiles using breakpoint annotations.
http://hal.inria.fr/hal-00663790.

Marc Lavielle.
Using penalized contrasts for the change-point problem.

E. Lebarbier.
Detecting multiple change-points in the mean of gaussian process by model selection.

Guillem Rigaill.
Pruned dynamic programming for optimal multiple change-point detection.

Nancy R Zhang and David O Siegmund.
A modified bayes information criterion with applications to the analysis of comparative
genomic hybridization data.
PMID: 17447926.
Runtime Comparisons, PELT-PDPA