Learning smoothing models of copy number profiles using breakpoint annotations

10 May 2012
Toby Dylan Hocking
toby.hocking@inria.fr
http://cbio.ensmp.fr/~thocking/
Invited talk for the Evry Stat and Genome lab
Acknowledgements

Gudrun Schleiermacher
Isabelle Janoueix-Lerosey
Francis Bach
Jean-Philippe Vert
Introduction: how to detect breakpoints in copy number?

Breakpoint annotations: a new method for smoothing model selection and validation

Definitions and results on the neuroblastoma data with several common smoothing models

Conclusions and future work
Motivation: tumor genome copy number analysis

- Comparative genomic hybridization microarrays (aCGH) allow genome-wide copy number analysis since logratio is proportional to DNA copy number (Pinkel et al., 1998).
- Tumors often contain breakpoints, amplifications, and deletions at specific chromosomal locations that we would like to detect.
- Which genomic alterations are linked with good or bad patient outcome?
- To answer clinical questions like this one, we first need to accurately detect these genomic alterations.
aCGH neuroblastoma copy number data
Copy number profiles are predictive of progression in neuroblastoma


2 types of profiles:

- Numerical: entire chromosome amplification. **Good** outcome.
- Segmental: deletion 1p 3p 11q, gain 1q 2p 17q. **Bad** outcome. In this talk “breakpoints.”
But which model to use to detect breakpoints?

- GLAD: adaptive weights smoothing (Hupé et al., 2004)
- DNAcopy: circular binary segmentation (Venkatraman and Olshen, 2007)
- cghFLasso: fused lasso signal approximator with heuristics (Tibshirani and Wang, 2007)
- HaarSeg: wavelet smoothing (Ben-Yaacov and Eldar, 2008)
- flsa: fused lasso signal approximator path algorithm (Hoefling 2009)
- cghseg: pruned dynamic programming (Rigaill 2010)

How to define which model is best?
And how to choose the degree of smoothness?
Introduction: how to detect breakpoints in copy number?

Breakpoint annotations: a new method for smoothing model selection and validation

Definitions and results on the neuroblastoma data with several common smoothing models

Conclusions and future work
Which model segments these the best? (demo)
cghseg.k smoothing for lambda=10^{-3.48}
cghseg.k smoothing for lambda=10^{-1.58}

position on chromosome (mega base pairs)
cghseg.k smoothing for lambda=10^-2.47
Choose the smoothing model that minimizes error with respect to breakpoint annotations

number of annotations inconsistent with model smoothing

−5.00 −3.48 −2.47 −1.58 0.00

log10(smoothness parameter)
Introduction: how to detect breakpoints in copy number?

Breakpoint annotations: a new method for smoothing model selection and validation

Definitions and results on the neuroblastoma data with several common smoothing models

Conclusions and future work
Definition of breakpoints in smoothing models

Consider just one copy number profile.

For the \(d\) probes corresponding to chromosome \(c\), let

\[
p_1 \leq \ldots \leq p_d \in \mathbb{N} \quad \text{positions on chromosome } c
\]
\[
y_1, \ldots, y_d \in \mathbb{R} \quad \text{logratio measurements}
\]
\[
\hat{y}_1^\lambda, \ldots, \hat{y}_d^\lambda \in \mathbb{R} \quad \text{smoothed profile (solid blue lines)}
\]

Breakpoints predicted on chromosome \(c\):

\[
\hat{B}_c^\lambda = \{(p_j + p_{j+1})/2 \mid \hat{y}_j^\lambda \neq \hat{y}_{j+1}^\lambda, \forall j = 1, \ldots, d - 1\}
\]

Then, we define \(\hat{B}^\lambda\) to be the complete set of genomic breakpoints predicted by algorithm \(\lambda\), over all chromosomes \(c\) (dashed vertical blue lines).
Gudrun and Isabelle annotated 6 regions on each profile

<table>
<thead>
<tr>
<th>k</th>
<th>$\min = r_k$</th>
<th>$\max = \bar{r}_k$</th>
<th>chrom $c_k$</th>
<th>breakpoint</th>
<th>normal</th>
<th>(all)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00</td>
<td>125.00</td>
<td>1</td>
<td>103</td>
<td>464</td>
<td>567</td>
</tr>
<tr>
<td>2</td>
<td>0.00</td>
<td>93.30</td>
<td>2</td>
<td>110</td>
<td>464</td>
<td>574</td>
</tr>
<tr>
<td>3</td>
<td>0.00</td>
<td>91.00</td>
<td>3</td>
<td>43</td>
<td>531</td>
<td>574</td>
</tr>
<tr>
<td>4</td>
<td>0.00</td>
<td>50.40</td>
<td>4</td>
<td>35</td>
<td>534</td>
<td>569</td>
</tr>
<tr>
<td>5</td>
<td>53.70</td>
<td>135.01</td>
<td>11</td>
<td>107</td>
<td>464</td>
<td>571</td>
</tr>
<tr>
<td>6</td>
<td>24.00</td>
<td>81.20</td>
<td>17</td>
<td>175</td>
<td>388</td>
<td>563</td>
</tr>
<tr>
<td>(all)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>573</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2845</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3418</td>
</tr>
</tbody>
</table>

![Graph showing the chromosome profile with annotations and breakpoints](image-url)
Quantify agreement to breakpoint annotations

Let $b^k_i \subset \mathbb{N}$ be the set of breakpoint counts annotated in region $k$ for profile $i$, i.e. \{0\}, \{1, 2, \ldots \}.

Let $\hat{b}^k_i(\lambda)$ be the number of breakpoints in this region predicted by the model with smoothness parameter $\lambda$.

The error function for one profile $i$ and one region $k$:

$$E^{k}_{i}(\lambda) = \begin{cases} 0 & \text{if } \hat{b}^k_i(\lambda) \in b^k_i \\ 1 & \text{otherwise} \end{cases}$$

The local error function for one profile $i$:

$$E^{local}_{i}(\lambda) = \sum_{k} E^{k}_{i}(\lambda)$$

The global error function for all profiles:

$$E^{global}(\lambda) = \sum_{i} E^{local}_{i}(\lambda)$$
Questions we can answer using these data

1. Can we find a model consistent with all the annotations?
2. Is it better to learn a global smoothness parameter $\hat{\lambda}$ or a profile-specific $\hat{\lambda}_i$?
3. How much time do we need to spend creating the annotation database?
Run each algorithm with a range of parameters, select model by minimizing annotation error

Result 1: no model is consistent with all the annotations, but cghseg.k has the lowest training error.
Details of the cghseg.k model

1. Segment each signal \( y \in \mathbb{R}^d \) using pruned DP to get the smoothed signal \( \hat{\beta}(y, k) \in \mathbb{R}^d \) for \( k \in \{1, \ldots, k_{\text{max}}\} \) segments.

2. For any \( \lambda > 0 \), define optimal number of segments

\[
k^*(y, \lambda) = \arg \min_{k \in \{1, \ldots, k_{\text{max}}\}} \lambda k + \|y - \hat{\beta}(y, k)\|^2 / d \tag{1}
\]

3. For any \( \lambda > 0 \), define the optimal smoothing

\[
\hat{y}^\lambda = \hat{\beta}(y, k^*(y, \lambda)) \tag{2}
\]

4. Discretize \( \lambda \in \{10^{-5}, \ldots, 10^0\} \) and calculate \( \hat{y}^\lambda \) for each.
Profile-specific smoothness parameters can decrease the training error, but is that good?

<table>
<thead>
<tr>
<th></th>
<th>cghseg.k</th>
<th>flsa.norm</th>
<th>dnacopy.sd</th>
<th>glad.lambdabreak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent incorrectly predicted annotations in training set</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
|                | ![Graph](image)

### Breakpoints

- More breakpoints: 2.2, 4.8, 11.5, 13.4
- Fewer breakpoints

---

**Legend:**
- **Errors**
- **False positive**
- **False negative**

---

**Note:**
- The graph illustrates the relationship between the log10(smoothing parameter) and the percentage of incorrectly predicted annotations in the training set for different models and profiles.
- The smoothing parameter affects the model's ability to detect breakpoints accurately.
- Higher values indicate fewer breakpoints, which can affect the model's performance.

---

**Statistic:**
- **Errors**
- **False positive**
- **False negative**

---

**Profiles:**
- Profile 375
- Profile 362
Annotations can be used for model selection and validation.

Set of all profiles and annotations

<table>
<thead>
<tr>
<th>train</th>
<th>test</th>
</tr>
</thead>
</table>

- BIC,
- AIC,
- etc.
- CV
- Annotations
- Selected model,
  degree of smoothness

Fit model with selected degree of smoothness on test profiles, evaluate using test annotations.
Cross-validation on regions and profiles

leave-one-out  n/1-fold  n/2-fold  n/3-fold

annotated region

tumor/copy number profile

set

train

test
Leave-one-out cross-validation on regions

<table>
<thead>
<tr>
<th></th>
<th>Global errors</th>
<th>Global FP</th>
<th>Global FN</th>
<th>Local errors</th>
<th>Local FP</th>
<th>Local FN</th>
<th>Timings seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td>cghseg.k</td>
<td>2.2</td>
<td>0.6</td>
<td>11.6</td>
<td>11.1</td>
<td>13.0</td>
<td>7.0</td>
<td>2.10</td>
</tr>
<tr>
<td>flsa.norm</td>
<td>6.7</td>
<td>3.6</td>
<td>18.5</td>
<td>14.8</td>
<td>15.2</td>
<td>10.5</td>
<td>0.08</td>
</tr>
<tr>
<td>dnacopy.sd</td>
<td>11.5</td>
<td>7.6</td>
<td>32.2</td>
<td>14.4</td>
<td>12.3</td>
<td>26.9</td>
<td>51.62</td>
</tr>
<tr>
<td>glad.haarseg</td>
<td>11.8</td>
<td>12.6</td>
<td>8.0</td>
<td>20.0</td>
<td>23.8</td>
<td>2.3</td>
<td>29.51</td>
</tr>
<tr>
<td>glad.lambdabreak</td>
<td>14.1</td>
<td>12.3</td>
<td>23.0</td>
<td>15.5</td>
<td>15.1</td>
<td>18.0</td>
<td>14.44</td>
</tr>
<tr>
<td>flsa</td>
<td>16.0</td>
<td>12.7</td>
<td>36.6</td>
<td>14.4</td>
<td>15.9</td>
<td>10.3</td>
<td>0.04</td>
</tr>
<tr>
<td>dnacopy.alpha</td>
<td>18.4</td>
<td>21.9</td>
<td>2.6</td>
<td>25.5</td>
<td>30.8</td>
<td>1.1</td>
<td>25.93</td>
</tr>
<tr>
<td>glad.MinBkpWeight</td>
<td>25.2</td>
<td>30.0</td>
<td>4.3</td>
<td>23.9</td>
<td>28.1</td>
<td>4.7</td>
<td>40.88</td>
</tr>
<tr>
<td>*glad.default</td>
<td>27.4</td>
<td>33.3</td>
<td>1.2</td>
<td>27.4</td>
<td>33.3</td>
<td>1.2</td>
<td>1.13</td>
</tr>
<tr>
<td>dnacopy.prune</td>
<td>27.9</td>
<td>31.9</td>
<td>17.1</td>
<td>31.1</td>
<td>35.8</td>
<td>14.8</td>
<td>35.17</td>
</tr>
<tr>
<td>*dnacopy.default</td>
<td>40.5</td>
<td>49.3</td>
<td>0.5</td>
<td>40.5</td>
<td>49.3</td>
<td>0.5</td>
<td>1.78</td>
</tr>
<tr>
<td>*cghseg.mBIC</td>
<td>40.9</td>
<td>49.4</td>
<td>0.0</td>
<td>40.9</td>
<td>49.4</td>
<td>0.0</td>
<td>1.47</td>
</tr>
<tr>
<td>*cghFLasso</td>
<td>80.8</td>
<td>97.2</td>
<td>0.0</td>
<td>80.8</td>
<td>97.2</td>
<td>0.0</td>
<td>0.14</td>
</tr>
</tbody>
</table>

* no smoothness parameter to select using annotations.

Result 2: local models trained on 1 profile overfit with many false positives.

So how many annotations are needed for a good model?
Cross-validation on regions and profiles

<table>
<thead>
<tr>
<th>leave-one-out</th>
<th>n/1-fold</th>
<th>n/2-fold</th>
<th>n/3-fold</th>
</tr>
</thead>
<tbody>
<tr>
<td>tumor/copy number profile</td>
<td>test</td>
<td>train</td>
<td>test</td>
</tr>
<tr>
<td>annotated region</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
How many annotations for a good generalization error?

Result 3: only a few annotations needed to get low breakpoint detection error.
Can we find a model consistent with all the annotations? No, but cghseg.k has the lowest training error.

Is it better to learn a profile-specific $\hat{\lambda}_i$, or to share information between profiles and learn a global smoothness parameter $\hat{\lambda}$? It is beneficial to share information between profiles, since the local model overfits with many false positives.

How much time do we need to spend creating the annotation database? Not much, since only a few profiles need to be annotated in order to accurately generalize to un-annotated profiles.
Introduction: how to detect breakpoints in copy number?

Breakpoint annotations: a new method for smoothing model selection and validation

Definitions and results on the neuroblastoma data with several common smoothing models

Conclusions and future work
Conclusions

- Annotations can be quickly produced using a GUI.
- Only a few annotations are necessary to get a good global smoothing model.
- Annotations are useful for:
  - Selecting the model and degree of smoothness.
  - Validating the model generalization ability.
  - Collaborating with biologists.

Availability:

- Learning smoothing models of copy number profiles using breakpoint annotations, HAL preprint 00663790.
- CRAN: `library(bams) “breakpoint annotation model smoothing”` to reproduce figures and tables from the article.
- CRAN: `data(neuroblastoma,package="neuroblastoma")`
- PyPI: `annotate_regions` GUI.
Future work and questions

- Annotations to quantify accuracy of copy number calling models.
- Joint modeling of logratio, annotations, and clinical outcome?
- Can other modeling strategies lower the breakpoint detection error?
- Supervised, interactive smoothing model building using breakpoint annotations.