Learning algorithms and statistical software, with applications to bioinformatics

PhD defense of
Toby Dylan Hocking
toby.hocking@inria.fr
http://cbio.ensmp.fr/~thocking/

20 November 2012
Summary of contributions

- **Ch. 2:** clusterpath for finding groups in data, ICML 2011.
- **Ch. 3:** breakpoint annotations for smoothing model training and evaluation, HAL-00663790.
- **Ch. 4-5:** penalties for breakpoint detection in simulated and real signals, under review.
- Statistical software contributions in R:
  - **Ch. 7:** direct labels for readable statistical graphics, Best Student Poster at useR 2011.
  - **Ch. 8:** documentation generation to convert comments into a package for distribution, accepted in JSS.
  - **Ch. 9:** named capture regular expressions for extracting data from text files, talk for useR 2011, accepted into R-2.14.
Cancer cells show chromosomal copy number alterations

Spectral karyotypes show the number of copies of the sex chromosomes (X,Y) and autosomes (1-22).
Source: Alberts et al. 2002.

Normal cell with 2 copies of each autosome.

Cancer cell with many copy number alterations.
Copy number profiles of neuroblastoma tumors
Ch. 2: clusterpath finds groups in data

Ch 3: breakpoint annotations for smoothing model selection

Ch. 4–5: penalties for breakpoint detection
The clusterpath relaxes a hard fusion penalty

\[
\min_{\alpha \in \mathbb{R}^{n \times p}} \|\alpha - X\|_F^2 \\
\text{subject to} \sum_{i < j} 1_{\alpha_i \neq \alpha_j} \leq t
\]

Combinatorial! Relaxation:

\[
\sum_{i < j} \|\alpha_i - \alpha_j\|_{q, w_{ij}} \leq t
\]

The **clusterpath** is the path of optimal \( \alpha \) obtained by varying \( t \). Related work: “fused lasso” Tibshirani and Saunders (2005), “convex clustering shrinkage” Pelckmans et al. (2005), “grouping pursuit” Shen and Huang (2010), “sum of norms” Lindsten et al. (2011).
Choice of norm and weights alters the clusterpath

Take $X \in \mathbb{R}^{10 \times 2}$, solve

$$\min_{\alpha} \|\alpha - X\|_F^2$$

subject to

$$\Omega(\alpha)/\Omega(X) \leq 0.$$ 

Penalty with $\ell_q$ norm:

$$\Omega(Y) = \sum_{i < j} \|Y_i - Y_j\|_q w_{ij}$$

Weights:

$$w_{ij} = \exp(-\gamma \|X_i - X_j\|_2^2)$$
Choice of norm and weights alters the clusterpath

Take $X \in \mathbb{R}^{10 \times 2}$, solve

$$\min_{\alpha} \|\alpha - X\|_F^2$$

subject to

$$\Omega(\alpha)/\Omega(X) \leq 0.1.$$  

Penalty with $\ell_q$ norm:

$$\Omega(Y) = \sum_{i < j} \|Y_i - Y_j\|_q W_{ij}$$

Weights:

$$w_{ij} = \exp(-\gamma \|X_i - X_j\|_2^2)$$
Choice of norm and weights alters the clusterpath

Take $X \in \mathbb{R}^{10 \times 2}$, solve

$$\min_{\alpha} ||\alpha - X||_F^2$$

subject to

$$\Omega(\alpha)/\Omega(X) \leq 0.2.$$ 

Penalty with $\ell_q$ norm:

$$\Omega(Y) = \sum_{i<j} ||Y_i - Y_j||_q w_{ij}$$

Weights:

$$w_{ij} = \exp(-\gamma ||X_i - X_j||_2^2)$$
Choice of norm and weights alters the clusterpath

Take $X \in \mathbb{R}^{10 \times 2}$, solve

$$\min_\alpha ||\alpha - X||_F^2$$

subject to

$$\Omega(\alpha)/\Omega(X) \leq 0.3.$$  

Penalty with $\ell_q$ norm:

$$\Omega(Y) = \sum_{i<j} ||Y_i - Y_j||_q w_{ij}$$

Weights:

$$w_{ij} = \exp(-\gamma||X_i - X_j||^2_F)$$
Choice of norm and weights alters the clusterpath

Take $X \in \mathbb{R}^{10 \times 2}$, solve

$$
\min_{\alpha} \| \alpha - X \|_F^2 
$$

subject to

$$
\Omega(\alpha)/\Omega(X) \leq 0.4.
$$

Penalty with $\ell_q$ norm:

$$
\Omega(Y) = \sum_{i < j} \| Y_i - Y_j \|_q w_{ij}
$$

Weights:

$$
w_{ij} = \exp(-\gamma \| X_i - X_j \|_2^2)
$$
Choice of norm and weights alters the clusterpath

Take $X \in \mathbb{R}^{10 \times 2}$, solve

$$\min_{\alpha} \|\alpha - X\|_F^2$$

subject to

$$\Omega(\alpha)/\Omega(X) \leq 0.5.$$  

Penalty with $\ell_q$ norm:

$$\Omega(Y) = \sum_{i < j} \|Y_i - Y_j\|_q w_{ij}$$

Weights:

$$w_{ij} = \exp(-\gamma \|X_i - X_j\|_2^2)$$
Choice of norm and weights alters the clusterpath

Take $X \in \mathbb{R}^{10 \times 2}$, solve

$$\min_{\alpha} \| \alpha - X \|_F^2$$

subject to

$$\Omega(\alpha)/\Omega(X) \leq 0.6.$$ 

Penalty with $\ell_q$ norm:

$$\Omega(Y) = \sum_{i<j} \| Y_i - Y_j \|_q w_{ij}$$

Weights:

$$w_{ij} = \exp(-\gamma\|X_i - X_j\|_2^2)$$
Choice of norm and weights alters the clusterpath

Take $X \in \mathbb{R}^{10 \times 2}$, solve

$$ \min_{\alpha} \| \alpha - X \|^2_F $$

subject to

$$ \Omega(\alpha)/\Omega(X) \leq 0.7. $$

Penalty with $\ell_q$ norm:

$$ \Omega(Y) = \sum_{i < j} \| Y_i - Y_j \|_q w_{ij} $$

Weights:

$$ w_{ij} = \exp(-\gamma \| X_i - X_j \|^2_2) $$
Choice of norm and weights alters the clusterpath

Take $X \in \mathbb{R}^{10 \times 2}$, solve

$$\min_{\alpha} \|\alpha - X\|_F^2$$

subject to

$$\Omega(\alpha)/\Omega(X) \leq 0.8.$$  

Penalty with $\ell_q$ norm:

$$\Omega(Y) = \sum_{i < j} \|Y_i - Y_j\|_q w_{ij}$$

Weights:

$$w_{ij} = \exp(-\gamma \|X_i - X_j\|_2^2)$$
Choice of norm and weights alters the clusterpath

Take $X \in \mathbb{R}^{10 \times 2}$, solve

$$\min_{\alpha} ||\alpha - X||_F^2$$

subject to

$$\Omega(\alpha)/\Omega(X) \leq 0.9.$$  

Penalty with $\ell_q$ norm:

$$\Omega(Y) = \sum_{i<j} ||Y_i - Y_j||_q w_{ij}$$

Weights:

$$w_{ij} = \exp(-\gamma ||X_i - X_j||_2^2)$$

<table>
<thead>
<tr>
<th>norm = 1</th>
<th>norm = 2</th>
<th>norm = $\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Diagram" /></td>
<td><img src="image2.png" alt="Diagram" /></td>
<td><img src="image3.png" alt="Diagram" /></td>
</tr>
</tbody>
</table>

Weights $\gamma = 0$

Weights $\gamma = 1$
Choice of norm and weights alters the clusterpath

Take \( X \in \mathbb{R}^{10 \times 2} \), solve

\[
\min_{\alpha} \|\alpha - X\|_F^2
\]

subject to

\( \Omega(\alpha)/\Omega(X) \leq 1 \).

Penalty with \( \ell_q \) norm:

\[
\Omega(Y) = \sum_{i<j} \|Y_i - Y_j\|_q w_{ij}
\]

Weights:

\[
w_{ij} = \exp(-\gamma \|X_i - X_j\|_2^2)
\]
Clusterpath learns a tree, even for odd cluster shapes

Comparison with other methods for finding 2 clusters.

Caveat: does not recover overlapping clusters, e.g. iris data, gaussian mixture.

- Theorem. No splits in the $\ell_1$ clusterpath with identity weights $w_{ij} = 1$. What about other situations?
- **Convex** and **hierarchical** clustering algorithms.
  - $\ell_1$ homotopy method $O(pn \log n)$.
  - $\ell_2$ active-set method $O(pn^2)$.
  - $\ell_\infty$ Franck-Wolfe algorithm.
- Implementation in R package clusterpath on R-Forge.
Ch. 2: clusterpath finds groups in data

Ch 3: breakpoint annotations for smoothing model selection

Ch. 4–5: penalties for breakpoint detection
How to detect breakpoints in $23 \times 575 = 13,225$ signals?
Which model should we use?

- 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)
- GADA: sparse Bayesian learning (Pique-Regi et al., 2008)
- flsa: fused lasso signal approximator path algorithm (Hoefling 2009)
- cghseg: pruned dynamic programming (Rigaill 2010)
- PELT: pruned exact linear time (Killick et al., 2011)

... and how to select the smoothing parameter in each model?
575 copy number profiles, each annotated in 6 regions

[Graph with labeled regions and annotations]

annotation
- >0breakpoints
- 0breakpoints

position on chromosome (mega base pairs)
Not enough breakpoints

![Graph showing not enough breakpoints](image-url)
Too many breakpoints
Good agreement with annotated regions
Select the best model using the breakpoint annotations

Breakpoint detection training errors for 3 models of data(neuroblastoma, package="neuroblastoma").

Idea: for several smoothing parameters $\lambda$, calculate the annotation error function $E(\lambda)$, (black line) then select the model with least error. (black dot)

$$\hat{\lambda} = \arg \min_{\lambda} E(\lambda).$$
PELT/cghseg show the best breakpoint detection

ROC curves for breakpoint detection training errors of each model, by varying the smoothness parameter $\lambda$.

Open circle shows smoothness $\lambda$ selected using annotations.
Few annotations required for a good breakpoint detector

Annotated profiles in global model training set

Percent of correctly predicted annotations on test set profiles

glad.lambdabreak
dnacopy.sd
flsa.norm
cghseg.k, pelt.n

Percent of correctly predicted annotations on test set profiles
Contributions in chapter 3: data, algorithms, comparison, code

Hocking et al. Learning smoothing models using breakpoint annotations. HAL-00663790, Jan 2012.

- data(neuroblastoma, package="neuroblastoma") on CRAN: 575 annotated profiles.
- Model training and evaluation protocols based on breakpoint annotations.
- Quantitative comparison of 17 breakpoint detection models on the neuroblastoma data set.
- Free/open-source GUIs for creating annotation databases.

More accurate breakpoint detection?
The cghseg.k/pelt.n least squares model

For a signal $y \in \mathbb{R}^d$, we define $\hat{y}^k$, the maximum likelihood model with $k \in \{1, \ldots, d\}$ segments as

$$\arg \min_{\mu \in \mathbb{R}^d} \|y - \mu\|_2^2$$

subject to

$$k - 1 = \sum_{j=1}^{d-1} 1_{\mu_j \neq \mu_{j+1}}.$$

Lavielle (2005): select the number of segments using the penalty

$$k^*(\lambda) = \arg \min_k \lambda kd + \|y - \hat{y}^k\|_2^2.$$ 

We use the tradeoff $\lambda$ which maximizes agreement with a database of breakpoint annotations.

But why this particular penalty? Should we normalize using the variance of the signal $y$?
The cghseg.k/pelt.n least squares model

For a signal $y \in \mathbb{R}^d$, we define $\hat{y}^k$, the maximum likelihood model with $k \in \{1, \ldots, d\}$ segments as

$$\arg \min_{\mu \in \mathbb{R}^d} ||y - \mu||_2^2$$

subject to $k - 1 = \sum_{j=1}^{d-1} 1_{\mu_j \neq \mu_{j+1}}$.

Lavielle (2005): select the number of segments using the penalty

$$k^*(\lambda) = \arg \min_k \lambda kd + ||y - \hat{y}^k||_2^2.$$ 

We use the tradeoff $\lambda$ which maximizes agreement with a database of breakpoint annotations.

But why this particular penalty? Should we normalize using the variance of the signal $y$?
Ch. 2: clusterpath finds groups in data

Ch 3: breakpoint annotations for smoothing model selection

Ch. 4–5: penalties for breakpoint detection
Simulated signals with varying sampling density

Can we learn a parameter $\lambda$ on the first signal, and use it for accurate breakpoint detection on the second?
Exact breakpoint error curves for 2 signals

For each signal $i$, calculate the breakpoint error function $BErr_i : \{1, \ldots, k_{\text{max}}\} \rightarrow \mathbb{R}^+$. 
Model selection error curves suggest $\alpha = 1/2$

For each signal $i$, select the number of segments using

$$k_i^{\alpha}(\lambda) = \arg\min_k \lambda d_i^{\alpha} + ||y_i - \hat{y}_i^k||_2^2.$$  

**Question:** which penalty exponent $\alpha$ is best?

Plot the error curves $E_i^{\alpha}(\lambda) = \text{BErr}_i[k_i^{\alpha}(\lambda)]$.

Then pick the parameter with minimal breakpoint error (dots)

$$\hat{\lambda}_i^{\alpha} = \arg\min_\lambda E_i^{\alpha}(\lambda).$$
Train and test error curves suggest $\alpha = 1/2$

$n = 8$ signals $i$ considered, with $d_1 = 70, \ldots, d_8 = 70000$ points sampled, and a penalty with term $d_i^{\alpha}$.

$$\text{train}(\alpha) = \min_{\lambda} \sum_{i=1}^{n} E_i^{\alpha}(\lambda),$$

$$\text{test}(\alpha) = \sum_{i \neq j} E_i^{\alpha}(\hat{\lambda}_j^{\alpha}).$$
Suggested penalties do not work on real data

\[ k_i(\lambda) = \arg \min_k \lambda k \sqrt{d_i} + ||y_i - \hat{y}_i^k||^2_2 \]  

(1)

For each signal \( i \), normalize for

- number of points sampled \( d_i \),
- signal length in base pairs \( l_i \),
- and estimated noise \( \hat{s}_i \).

\[ k_i(\lambda) = \arg \min_k \lambda k \hat{s}_i^2 \sqrt{d_i/l_i} + ||y_i - \hat{y}_i^k||^2_2 \]  

(2)

<table>
<thead>
<tr>
<th>exponents</th>
<th>annotation error</th>
</tr>
</thead>
<tbody>
<tr>
<td>points</td>
<td>length</td>
</tr>
<tr>
<td>cghseg.k</td>
<td></td>
</tr>
<tr>
<td>(1)</td>
<td>1/2</td>
</tr>
<tr>
<td>(2)</td>
<td>1/2</td>
</tr>
</tbody>
</table>
In real data, we only have weak annotations
Model selection and error curves for 2 signals

Optimal number of segments $z^*_i(L) = \arg \min_k \exp(L) k + \| y_i - \hat{y}_i^k \|_2^2$.

![Graph showing model selection and error curves for 2 signals](image_url)
Target interval $[\underline{L}_i, \overline{L}_i]$ for 2 signals
Target interval $[L_i, \bar{L}_i]$ for all signals
Limit point representation, find a separator

\[ \text{variance estimate } \log \hat{s}_i \]

\[ \text{penalty exponent } L \]

\[ \text{limit} \]

\[ \bullet \overline{L}_i \]

\[ \circ \underline{L}_i \]
Max margin regression line

The diagram illustrates the relationship between the variance estimate \( \log \hat{s}_i \) and the penalty exponent \( L \). The data points are marked with circles, where filled circles represent \( L_i \) and open circles represent \( \bar{L}_i \). The limit lines are shown to help visualize the trend as the penalty exponent changes.
Learning the penalty function

- For every signal $i$, calculate a variance estimate feature $x_i = \log \hat{s}_i$.
- Predict model complexity using an affine function
  
  $$f(x_i) = w \log \hat{s}_i + \beta.$$  

- Equivalent to learning a penalty function
  
  $$z_i^*[f(x_i)] = \arg \min_k ||y_i - \hat{y}_i^k||_2^2 + \exp[f(x_i)] k$$
  
  $$= \arg \min_k ||y_i - \hat{y}_i^k||_2^2 + \hat{s}_i w e^\beta k.$$
Original data: annotate the same regions
Detailed data: draw rectangles around regions
Optimal model complexity depends on variance

\[
\text{arg min}_{\beta \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^{n} l_i (\mathbf{w}' \mathbf{x}_i + \beta) + \gamma \| \mathbf{w} \|_1
\]
Optimal model complexity depends on variance

\[
\arg \min_{\beta \in \mathbb{R}, w \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^{n} l_i(w'x_i + \beta) + \gamma ||w||_1
\]
Error estimated using 10-fold cross-validation

\[ f(x_i) = w_1 \log \hat{s}_i + w_2 \log d_i + \beta \]

- \text{cghseg.k}: \ w_1 = 0, \ w_2 = 1, \text{ learn } \beta \text{ using grid search to minimize the annotation error } E_i.
- \text{log.s.log.d}: \text{ learn } \beta, \ w_1, \ w_2 \text{ by by minimizing the un-regularized } (\gamma = 0) \text{ surrogate loss } l_i.
- \text{L1-reg}: \text{ variance estimate, signal size, model error, chromosome indicator features } x_i \in \mathbb{R}^{117}, \text{ CV to choose the degree of } \ell_1 \text{ regularization } \gamma.
Conclusions and future work

- Penalties suggested by theory and simulations are suboptimal in real data.
- Annotations $+$ convex optimization $=$ learned penalties.
- Learned penalties show good change-point detection.
- Learning more general penalty functions?
- Active learning strategy for picking the next signal to annotate?
- Do annotation-guided models help to predict clinical patient outcome?
Acknowledgements

Labs:
- INRIA-Sierra
- INSERM U830
- Institute Curie bioinformatics (U900)

Coauthors:
- Armand Joulin
- Isabelle Janoueix-Lerosey
- Gudrun Schleiermacher
- Olivier Delattre
- Guillem Rigaill

Thank you!