Clusterpath:
an algorithm for clustering using convex fusion penalties

Toby Dylan Hocking
toby.hocking@inria.fr
http://cbio.ensmp.fr/~thocking/
joint work with Armand Joulin, Francis Bach, and Jean-Philippe Vert

Sierra team — INRIA Paris
Mines ParisTech / Curie Institute / INSERM

1 July 2011
1 Introduction: framing clustering as a convex optimization problem results in the clusterpath

2 Algorithms proposed to solve the clusterpath

3 Results, conclusions, and open questions
Existing clustering methods have some drawbacks

- Clustering: assign labels to $n$ points in $p$ dimensions $X \in \mathbb{R}^{n \times p}$.
- Methods:
  - K-means
  - Hierarchical
  - Mixture models
  - Spectral (Ng et al. 2001)
  - ...
- Issues:
  - Hierarchy
  - Convexity
  - Greediness
  - Stability
Our contributions:

- A new **convex** objective function for clustering.
- Interpretable, **hierarchical** clusterpath.
- Efficient algorithms.
- Clustering **performance** on par with spectral clustering.
The clusterpath relaxes a hard fusion penalty

\[
\min_{\alpha \in \mathbb{R}^{n \times p}} \frac{1}{2} \| \alpha - X \|_F^2
\]

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 continuous path of optimal \( \alpha \) obtained by varying \( t \). Related work: “fused lasso” Tibshirani and Saunders (2005), “grouping pursuit” Shen and Huang (2010), “sum of norms” Lindsten et al. (2011).
Let $X \in \mathbb{R}^{3 \times 2}$.

Approximate $X$ using $\alpha$:

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

Constrain the total distance between every pair of points:

$$\sum_{i<j} \| \alpha_i - \alpha_j \|_q \leq t$$

(grey lines)
Geometric interpretation of penalty with identity weights

Let $X \in \mathbb{R}^{3 \times 2}$.

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

Constrain the total distance between every pair of points:
$$\sum_{i < j} \| \alpha_i - \alpha_j \|_q \leq t$$

(grey lines)
Geometric interpretation of penalty with identity weights

- Let \( X \in \mathbb{R}^{3 \times 2} \).
- Approximate \( X \) using \( \alpha \):
  \[
  \min_{\alpha} \| \alpha - X \|_F^2
  \]
- Constrain the total distance between every pair of points:
  \[
  \sum_{i < j} \| \alpha_i - \alpha_j \|_q \leq t
  \]
  (grey lines)
Geometric interpretation of penalty with general weights

Let $X \in \mathbb{R}^{3 \times 2}$.

Approximate $X$ using $\alpha$:
$$\min_{\alpha} \lVert \alpha - X \rVert_F^2$$

Constrain the total area of boxes between every pair of points:
$$\sum_{i < j} \lVert \alpha_i - \alpha_j \rVert_q \omega_{ij} \leq t$$

(grey rectangles)
Let $X \in \mathbb{R}^{3 \times 2}$.

Approximate $X$ using $\alpha$:

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

Constrain the total area of boxes between every pair of points:

$$\sum_{i < j} \| \alpha_i - \alpha_j \|_q w_{ij} \leq t$$

(grey rectangles)
Geometric interpretation of penalty with general weights

Let $X \in \mathbb{R}^{3 \times 2}$.

Approximate $X$ using $\alpha$:
$$\min_\alpha \| \alpha - X \|_F^2$$

Constrain the total area of boxes between every pair of points:
$$\sum_{i<j} \| \alpha_i - \alpha_j \|_q w_{ij} \leq t$$

(grey rectangles)
Geometric interpretation of penalty with general weights

- Let $X \in \mathbb{R}^{3 \times 2}$.
- Approximate $X$ using $\alpha$:
  $$\min_{\alpha} \|\alpha - X\|_F^2$$
- Constrain the total area of boxes between every pair of points:
  $$\sum_{i < j} \|\alpha_i - \alpha_j\|_q w_{ij} \leq t$$
  (grey rectangles)
Let $X \in \mathbb{R}^{3 \times 2}$.

Approximate $X$ using $\alpha$:

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

Constrain the total area of boxes between every pair of points:

$$\sum_{i<j} \|\alpha_i - \alpha_j\|_q w_{ij} \leq t$$

(grey rectangles)
Geometric interpretation of penalty with general weights

- Let $X \in \mathbb{R}^{3 \times 2}$.
- Approximate $X$ using $\alpha$:
  $$\min_\alpha \| \alpha - X \|_F^2$$
- Constrain the total area of boxes between every pair of points:
  $$\sum_{i < j} \| \alpha_i - \alpha_j \|_q w_{ij} \leq t$$
  (grey rectangles)
Let $X \in \mathbb{R}^{3 \times 2}$.

Approximate $X$ using $\alpha$:

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

Constrain the total area of boxes between every pair of points:

$$\sum_{i < j} \| \alpha_i - \alpha_j \|_q w_{ij} \leq t$$

(grey rectangles)
Geometric interpretation of penalty with general weights

- Let $X \in \mathbb{R}^{3 \times 2}$.
- Approximate $X$ using $\alpha$:
  $$\min_{\alpha} \| \alpha - X \|_F^2$$
- Constrain the total area of boxes between every pair of points:
  $$\sum_{i<j} \| \alpha_i - \alpha_j \|_q w_{ij} \leq t$$
  (grey rectangles)
Let $X \in \mathbb{R}^{3 \times 2}$.

Approximate $X$ using $\alpha$:

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

Constrain the total area of boxes between every pair of points:

$$\sum_{i < j} \| \alpha_i - \alpha_j \|_q w_{ij} \leq t$$

(grey rectangles)
Let $X \in \mathbb{R}^{3 \times 2}$.

Approximate $X$ using $\alpha$:

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

Constrain the total area of boxes between every pair of points:

$$\sum_{i<j} ||\alpha_i - \alpha_j||_q w_{ij} \leq t$$

(grey rectangles)
Let $X \in \mathbb{R}^{3\times2}$.

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

Constrain the total area of boxes between every pair of points:
$$\sum_{i<j} \|\alpha_i - \alpha_j\|_q w_{ij} \leq t$$

(grey rectangles)
Choice of norm and weights alters the clusterpath

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

\[
\min_\alpha \| X - \alpha \|_F^2 \\
\text{subject to} \\
\Omega(\alpha)/\Omega(X) \leq 1,
\]

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

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

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

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

subject to

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

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

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

$\text{norm} = 1$

$\text{norm} = 2$

$\text{norm} = \infty$

Toby Dylan Hocking et al. (INRIA Paris)

Clusterpath for hierarchical convex clustering

1 July 2011
Choice of norm and weights alters the clusterpath

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

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

subject to

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

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

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

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

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

subject to

\[
\Omega(\alpha)/\Omega(X) \leq 0.7,
\]

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

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

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

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

subject to

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

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

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

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

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

subject to

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

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

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

\begin{align*}
\text{norm} = 1 & \quad \text{norm} = 2 & \quad \text{norm} = \infty \\
\begin{array}{|c|c|c|}
\hline
& \Omega(\alpha)/\Omega(X) \leq 0.5 & \\
\hline
0 = \gamma & \infty & 1 \\
\hline
\end{array}
\end{align*}
Choice of norm and weights alters the clusterpath

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

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

subject to

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

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

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

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

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

subject to

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

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

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

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

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

subject to

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

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

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

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

\[
\min_{\alpha} \| X - \alpha \|_F^2 \\
\text{subject to} \\
\Omega(\alpha)/\Omega(X) \leq 0.1,
\]

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

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

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

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

subject to

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

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

$$w_{ij} = \exp(-\gamma \|X_i - X_j\|_2^2)$$
1 Introduction: framing clustering as a convex optimization problem results in the clusterpath

2 Algorithms proposed to solve the clusterpath

3 Results, conclusions, and open questions
Objective separable on each dimension $\alpha^k \in \mathbb{R}^n$:

$$\min_{\alpha} f_1(\alpha) = \sum_{k=1}^{p} \sum_{i=1}^{n} (X_{ik} - \alpha_{ik})^2 + \lambda \sum_{i<j} |\alpha_{ik} - \alpha_{jk}| w_{ij} = \sum_{k=1}^{p} \min_{\alpha^k} f_1(\alpha^k)$$

LARS-like path-following algorithm (Hoefling et al. 2009), checks for split of a cluster of size $n_C$ by solving a max-flow problem: $O(n_C^3)$.
An efficient path algorithm for the $\ell_1$ clusterpath

- Objective separable on each dimension $\alpha^k \in \mathbb{R}^n$:

$$\min_{\alpha} f_1(\alpha) = \sum_{k=1}^{p} \sum_{i=1}^{n} (X_{ik} - \alpha_{ik})^2 + \lambda \sum_{i<j} |\alpha_{ik} - \alpha_{jk}| \omega_{ij} = \sum_{k=1}^{p} \min_{\alpha^k} f_1(\alpha^k)$$

- LARS-like path-following algorithm (Hoefling et al. 2009), checks for split of a cluster of size $n_C$ by solving a max-flow problem: $O(n_C^3)$.

- **Theorem**: the $\ell_1$ clusterpath with $\omega_{ij} = 1$ is strictly agglomerative, so no need to check for splits.

- $n - 1$ joins on each of $p$ dimensions. Each join only costs $O(\log n)$ using queues and linked lists: $O(pn \log n)$. 

Toby Dylan Hocking et al. (INRIA Paris)  
Clusterpath for hierarchical convex clustering  
1 July 2011
We propose dedicated solvers for each norm

<table>
<thead>
<tr>
<th>Norm</th>
<th>Properties</th>
<th>Algorithm</th>
<th>Complexity</th>
<th>Problem sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>piecewise linear path</td>
<td></td>
<td>$O(pn \log n)$</td>
<td>large $\approx 10^5$</td>
</tr>
<tr>
<td></td>
<td>separable</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>rotation invariant active-set</td>
<td></td>
<td>$O(n^2 p)$</td>
<td>medium $\approx 10^3$</td>
</tr>
<tr>
<td>$\infty$</td>
<td>piecewise linear Frank-Wolfe</td>
<td>unknown*</td>
<td></td>
<td>medium $\approx 10^3$</td>
</tr>
</tbody>
</table>

*each iteration of complexity $O(n^2 p)$. 
The spectral clusterpath

Standard spectral clustering:
- Pairwise similarities: $W_{ij} = \exp(-\gamma \|X_i - X_j\|^2_2)$.
- Normalized Laplacian: $L = D - W$.
- K-means on the first few eigenvectors of $L$.

2 possible problems with spectral clustering:
- Hard-thresholding when picking the first few eigenvectors.
- K-means on eigenvectors.

We propose:
- Soft-thresholding eigenvectors based on eigenvalues.
- Clusterpath on eigenvectors.
1 Introduction: framing clustering as a convex optimization problem results in the clusterpath

2 Algorithms proposed to solve the clusterpath

3 Results, conclusions, and open questions
Typical results for several clustering algorithms

Weighted $\ell_2$ clusterpath with $\gamma = 2$. 
Conclusions: our contributions

- We proposed a new family of convex objective functions for clustering.
- Using $w_{ij} = 1$ with the $\ell_1$ norm, we have hierarchical clustering and complexity $O(pn \log n)$.
- Performance of weighted $\ell_2$ similar to spectral clustering.
- Free, open-source R/C++/Python optimization software available: http://clusterpath.r-forge.r-project.org
Open questions

- Benefit of relaxing the eigenvector thresholding in spectral clustering. Generality?
- Open question: hierarchical clusterpath for which other weights and norms?
- Learning the weights and number of clusters automatically?
Open questions

- Benefit of relaxing the eigenvector thresholding in spectral clustering. Generality?
- Open question: hierarchical clusterpath for which other weights and norms?
- Learning the weights and number of clusters automatically?

Thank you, any questions?
You can contact me directly at toby.hocking@inria.fr.