Statistical Properties of Convex Clustering

Kean Ming Tan
University of Washington

August 10, 2015
Convex Clustering

\[
\mathbf{X} = \begin{array}{cccccc}
\end{array}
\]

Features

Observations

1 2 \ldots \ldots \ldots \ldots \ldots p

1
2
\cdot
\cdot
n
Convex Clustering

\[ \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \]

Observations

Features
Recent interest in formulating estimators as the solutions to convex optimization problems:
  - efficient algorithms give convergence to global optimum.
  - optimality conditions fully characterize estimators.
Recent interest in formulating estimators as the solutions to convex optimization problems:
  - efficient algorithms give convergence to global optimum.
  - optimality conditions fully characterize estimators.

Clustering is a hard problem:
  - non-convex.
  - greedy algorithms do not achieve global optimum.
Convex Clustering

- Recent interest in formulating estimators as the solutions to convex optimization problems:
  - efficient algorithms give convergence to global optimum.
  - optimality conditions fully characterize estimators.

- Clustering is a hard problem:
  - non-convex.
  - greedy algorithms do not achieve global optimum.

- How about a convex formulation for clustering?
A convex optimization problem, for $q \geq 1$ and $\lambda \geq 0$:

$$\minimize_{\hat{U} \in \mathbb{R}^{n \times p}} \frac{1}{2} \sum_{i=1}^{n} \|X_i - \hat{U}_i\|_2^2 + \lambda \sum_{i < i'} \|\hat{U}_i - \hat{U}_{i'}\|_q$$

**Regularization Term:**
Encourages rows of $\hat{U}$ to be identical.

**Definition:**
The $i$th and $i'$th observations are in same cluster if and only if $\hat{U}_i = \hat{U}_{i'}$.

---

Role of Tuning Parameter $\lambda$

$\lambda = 0$, 10 clusters
Role of Tuning Parameter $\lambda$

$\lambda = 0.3$, 9 clusters
Role of Tuning Parameter $\lambda$

$\lambda = 0.4$, 7 clusters
Role of Tuning Parameter $\lambda$

$\lambda = 0.52, 6$ clusters
Role of Tuning Parameter $\lambda$

$\lambda = 0.6$, 5 clusters
Role of Tuning Parameter $\lambda$

$\lambda = 0.65$, 4 clusters
Role of Tuning Parameter $\lambda$

$\lambda = 0.67, 1$ clusters
Algorithm

Standard algorithms can be used to obtain the global optimum of the convex clustering problem.

. . . for instance, alternating directions method of multipliers.

Most of the existing literature on convex clustering has focused on algorithms, rather than statistical properties or empirical performance.
For $y \sim N_n(\mu, \sigma^2 I)$, the degrees of freedom of $\hat{\mu}$ is defined as

$$\sum_{i=1}^{n} \frac{\text{Cov}(\hat{\mu}_i, y_i)}{\sigma^2}.$$ 

**Question:** Can we derive an unbiased estimator for the degrees of freedom of convex clustering, for a given value of $q$ and $\lambda$?
Unbiased Estimators for Degrees of Freedom

Assume that each observation is independent $N_p(\mu_k, \sigma^2 I)$.

**Lemma:** For $q = 1$, number of unique elements in $\hat{U}$.

**Lemma:** For $q = 2$, a complicated expression!

**Application:** Use BIC to select $\lambda$, i.e. to determine $\#$ of clusters.
Prediction Consistency

Under certain assumptions, convex clustering’s error in estimating the true cluster means decreases to zero as $n, p \to \infty$. 
Connection to $k$-means Clustering

$k$-means clustering with 2 clusters:

\[ \min_{\mu_1, \mu_2, C_1, C_2} \sum_{i \in C_1} \|X_i - \mu_1\|^2_2 + \sum_{i \in C_2} \|X_i - \mu_2\|^2_2 \]
Connection to \(k\)-means Clustering

\(k\)-means clustering with 2 clusters:

\[
\begin{align*}
\text{minimize}_{\mu_1, \mu_2, C_1, C_2} & \sum_{i \in C_1} \|X_{i.} - \mu_1\|_2^2 + \sum_{i \in C_2} \|X_{i.} - \mu_2\|_2^2 \\
\end{align*}
\]

Convex Clustering with \(q = 0\):

\[
\begin{align*}
\text{minimize}_{U \in \mathbb{R}^{n \times p}} & \quad \frac{1}{2} \sum_{i=1}^{n} \|X_{i.} - U_{i.}\|_2^2 + \lambda \sum_{i < i'} \|U_{i.} - U_{i'.}\|_0 \\
\end{align*}
\]
Connection to \( k \)-means Clustering

**\( k \)-means clustering with 2 clusters:**

\[
\text{minimize} \quad \sum_{i \in C_1} \| X_i - \mu_1 \|^2_2 + \sum_{i \in C_2} \| X_i - \mu_2 \|^2_2
\]

**Convex Clustering with \( q = 0 \):**

\[
\text{minimize} \quad \frac{1}{2} \sum_{i \in C_1} \| X_i - \mu_1 \|^2_2 + \frac{1}{2} \sum_{i \in C_2} \| X_i - \mu_2 \|^2_2 + \lambda \cdot |C_1| \cdot (n - |C_1|)
\]
Connection to $k$-means Clustering

$k$-means clustering with 2 clusters:

$$\min_{\mu_1, \mu_2, C_1, C_2} \sum_{i \in C_1} \|X_i - \mu_1\|^2_2 + \sum_{i \in C_2} \|X_i - \mu_2\|^2_2$$

Convex Clustering with $q = 0$:

$$\min_{\mu_1, \mu_2, C_1, C_2} \frac{1}{2} \sum_{i \in C_1} \|X_i - \mu_1\|^2_2 + \frac{1}{2} \sum_{i \in C_2} \|X_i - \mu_2\|^2_2 + \lambda \cdot |C_1| \cdot (n - |C_1|)$$

**Regularization Term:**
Encourage size of the clusters to be unbalanced
Connection to Single Linkage Clustering

- Associated with every convex optimization problem is an equivalent\(^2\) **dual problem**.

\(^2\)if certain conditions are satisfied . . . and they usually are
Connection to Single Linkage Clustering

- Associated with every convex optimization problem is an equivalent\(^2\) **dual problem**.
- The dual problem for convex clustering . . .

\(^2\text{if certain conditions are satisfied . . . and they usually are}
Connection to Single Linkage Clustering

- Associated with every convex optimization problem is an equivalent\(^2\) dual problem.
- The dual problem for convex clustering . . .
- . . . is almost identical to the dual problem for single linkage clustering!!!

\[^2\text{if certain conditions are satisfied . . . and they usually are}\]
Simulation Studies: Mixture of Gaussians

Fig 2. Simulation results for Gaussian clusters with $K = 2$, $n = p = 30$, averaged over 200 data sets, for two noise levels $\sigma = \{1, 2\}$. Colored lines correspond to single linkage clustering (---), average linkage hierarchical clustering (---), $k$-means clustering (---), convex clustering with $q = 1$ (---), and convex clustering with $q = 2$ (---).
Bottom Line: Why Convex Clustering?

- I’m a big fan of convexity . . . *when it’s useful*.
- Not clear that convex clustering is useful!
Bottom Line: Why Convex Clustering?

- I’m a big fan of convexity . . . **when it’s useful**.
- Not clear that convex clustering is useful!
  - Can obtain global optimum.
  - Can estimate degrees of freedom.
  - Can establish prediction consistency.

Bottom Line: Why Convex Clustering?

- I’m a big fan of convexity . . . *when it’s useful*.
- Not clear that convex clustering is useful!
  - Can obtain global optimum.
  - Can estimate degrees of freedom.
  - Can establish prediction consistency.
  - Essentially the same as single linkage clustering.
  - Similar to $k$-means clustering.
  - Underwhelming empirical performance.

[arXiv](http://arxiv.org/abs/1503.08340)
I’m a big fan of convexity . . . **when it’s useful**.

Not clear that convex clustering is useful!

+ Can obtain global optimum.
+ Can estimate degrees of freedom.
+ Can establish prediction consistency.
  - Essentially the same as single linkage clustering.
  - Similar to $k$-means clustering.
  - Underwhelming empirical performance.