Cluster Analysis: Unsupervised Learning via Supervised Learning with a Non-convex Penalty

WEI PAN

1Division of Biostatistics, School of Public Health
University of Minnesota

Australian Statistical Conference, Sydney, July 2014

Joint work with Xiaotong Shen and Binghui Liu.
Outline

- Problem
- New methods: Pan, Shen and Liu (2013, *JMLR*)
  Shen, Pan and Zhu (2012, *JASA*): TLP
- Numerical Results: simulated and real data
- Summary
Clustering Analysis

• Given data $X = (x'_1, ..., x'_n)'$ with $x_i = (x_{i1}, x_{i2}, ..., x_{ip})'$, find centroids $\mu_i$ for each $x_i$;
  Clustering: many $\mu_i$’s are equal!

• Most algorithms specify a few $\mu_i$’s, then try to estimate them. K-means, (Gaussian) mixture models, ...

• Here, we specify $n$ $\mu_i$’s, over-parametrized!
  Main idea: group $\mu_i$’s by penalization!
New Methods

- A general framework: like regression,

\[ \hat{\mu} = \arg \min_{\mu} \frac{1}{2} \sum_{i=1}^{n} L(x_i - \mu_i) + \lambda \sum_{i<j} h(\mu_i - \mu_j), \]

where \( L() \) is a loss, \( h() \) is a grouping or fusion penalty.

- LS-\( L_1 \) (or Lasso) (Tibshirani 1996):

\[ \frac{1}{2} \sum_{i=1}^{n} ||x_i - \mu_i||_2^2 + \lambda \sum_{i<j} ||\mu_i - \mu_j||_1, \]

where \( ||.||_q \) is the \( L_q \)-norm.

- Ours: TLP (Shen et al 2012) is defined as

\[ \text{TLP}(\alpha; \tau) = \min(|\alpha|, \tau), \]

where \( \tau \) is a tuning parameter.
• A key property:

\[ \frac{\text{TLP}(\alpha; \tau)}{\tau} \rightarrow L_0(\alpha) = I(\alpha \neq 0) \]

as \( \tau \rightarrow 0^+ \).

Figure 1: TLP.
• Ours: a group TLP (gTLP) penalty

\[ gTLP(\mu_i - \mu_j; \tau) = TLP(||\mu_i - \mu_j||_2; \tau). \]

better than \( L_q \)-norm for \( q \geq 1 \).

• Summary: Lasso- and gTLP-based \textbf{PRclust}:

\[
\hat{\mu} = \arg \min_{\mu} \frac{1}{2} \sum_{i=1}^{n} ||x_i - \mu_i||_2^2 + \lambda \sum_{i<j} ||\mu_i - \mu_j||_1, \quad (1)
\]

\[
\hat{\mu} = \arg \min_{\mu} \frac{1}{2} \sum_{i=1}^{n} ||x_i - \mu_i||_2^2 + \lambda \sum_{i<j} TLP(||\mu_i - \mu_j||_2; \tau), \quad (2)
\]

A cluster: \( x_i \)'s with equal \( \hat{\mu}_i \).

• Computing: Not separable, no coordinate-descent algorithm!

• Alternative: quadratic penalty method via reparametrization
\( \theta_{ij} = \mu_i - \mu_j \) for \( 1 \leq i < j \leq n \); new objective functions:

\[
S_L(\mu, \theta) = \frac{1}{2} \sum_{i=1}^{n} ||x_i - \mu_i||_2^2 + \frac{\lambda_1}{2} \sum_{i<j} ||\mu_i - \mu_j - \theta_{ij}||_2^2 + \\
\lambda_2 \sum_{i<j} ||\theta_{ij}||_1,
\]

(3)

\[
S(\mu, \theta) = \frac{1}{2} \sum_{i=1}^{n} ||x_i - \mu_i||_2^2 + \frac{\lambda_1}{2} \sum_{i<j} ||\mu_i - \mu_j - \theta_{ij}||_2^2 + \\
\lambda_2 \sum_{i<j} \text{TLP}(||\theta_{ij}||_2; \tau).
\]

(4)

- gTLP: non-convex; use difference of convex programming ...
- Then apply coordinate-descent
- Property: finite and monotone convergence to a local minimizer.
• An advantage of PRclust: use a model selection criterion in regression;
  GCV (Golub et al 1979);
  GDF based on data perturbation (Ye 1998; Shen and Ye 2002).
Results

- Simulation cases: case I, $n = 50 + 50$;

Figure 2: The first simulated data set in a) Case I, b) Case II and c) Case VI.
Figure 3: GDF in K-means.
Figure 4: Solution paths of $\hat{\mu}_{i,1}$ for a) PRclust (with gTLP), b) PRclust2, c) PRclust with the Lasso penalty and d) HTclust for
Figure 5: Solution paths of $\hat{\mu}_{i,1}$ for PRclust-$L_q$ with a) $q = 1$, b) $q = 2$ and c) $q = \infty$ for the first simulated dataset in Case I.
Summary

• Non-convex (e.g. TLP) grouping penalty: better in separating clusters than convex (e.g. $L_q$-norm) grouping penalties!

• A group penalty (e.g. gTLP) is better than a non-group one (e.g. TLP or Lasso).

• Clustering: like regression or supervised learning?! techniques from the latter, e.g. model selection criteria, ...

• Extensions and applications: on-going
Acknowledgment: This research was supported by NIH.

Thank you!