Inference, Computation, & Dynamic Visualization for Convex Clustering

Genevera I. Allen

Departments of Statistics, Computer Science,
and Electrical and Computer Engineering, Rice University,
Jan and Dan Duncan Neurological Research Institute,
Baylor College of Medicine & Texas Children's Hospital.

September 9, 2018

1. Motivation & Background

2. Inference for Multivariate Means in Adaptive Data Analysis

3. Computation & Dynamic Clustering Visualization

Motivation: Clustering & Biclustering

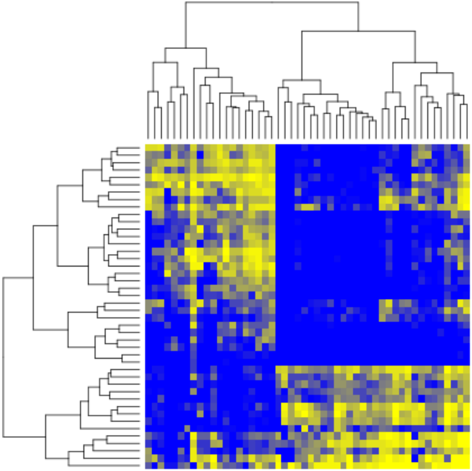
Clustering

Find groups of objects which are similar to each other.

Biclustering

Simultaneously find groups of features & observations.

- Cluster rows & columns of data matrix.



Clustering Approaches

The Good:

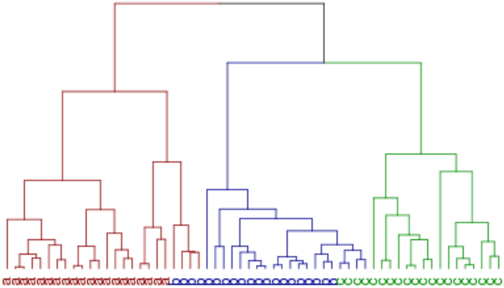
- Simple & Fast.
- Appealing Visualizations.
- Easy Interpretation.

The Bad:

- Local solutions.
- Instability.
- Tuning parameters.

The Ugly:

- How many clusters?
- Inference.



Hierarchical

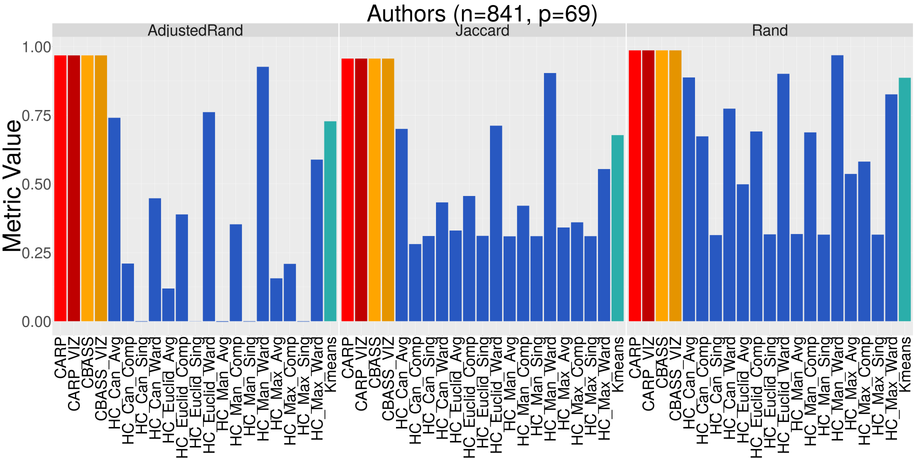
Convex Clustering & Biclustering

Why Convex?

- Global solution!
- Superior mathematical and statistical properties:
  - Consistency.
  - Stability.
  - Improved clustering performance.
- Data-driven selection of # of clusters.
- Inference?
- Fast Computation & Visualization?

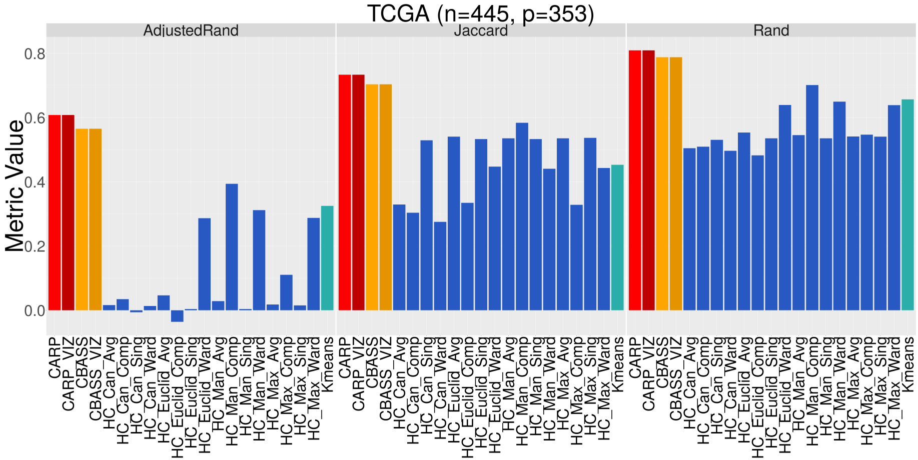
Pelckmans et al. 2005; Lindsten et al. 2011; Hocking et al. 2011; Chi & Lange 2013; Tan & Witten 2015; Chi, Allen & Baraniuk, 2017; Radchenko & Mukherjee, 2017

Clustering Accuracy



Author Data

Clustering Accuracy



TCGA Breast Cancer Data

Convex Clustering

$$\underset{\mathbf{u}}{\text{minimize}} \quad \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \lambda \sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

- $\mathbf{x}_i$ - each observation (p-vector).
- $\mathbf{u}_i$ - cluster centroid for each observation.

Convex fusion penalty shrinks centroids together!

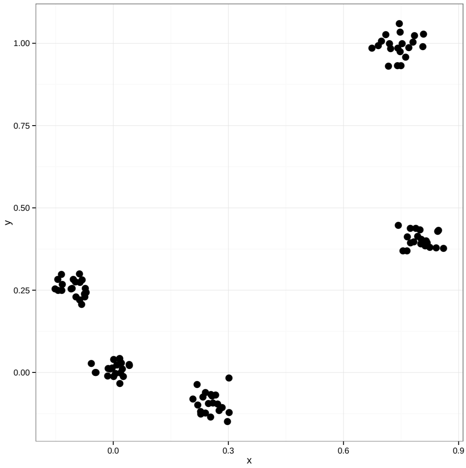
Pelckmans et al. 2005; Lindsten et al. 2011; Hocking et al. 2011; Chi & Lange 2013; Tan & Witten 2015.

Convex Clustering

$$\underset{\mathbf{u}}{\text{minimize}} \quad \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \lambda \sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

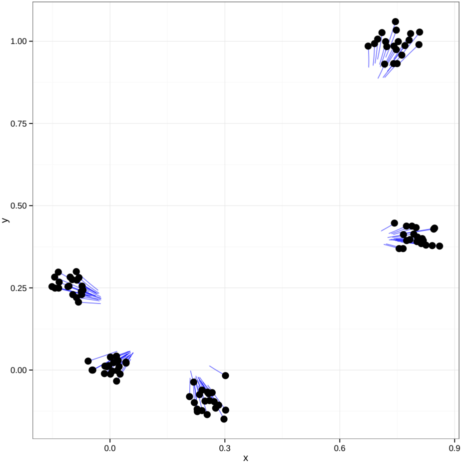
- λ controls BOTH cluster assignments & number of clusters.
  - $\lambda = 0$ - each observation is its own cluster.
  - λ larger - column means begin to coalesce together into clusters.
  - λ very large - all observations fused into one cluster.
- Algorithm: Alternating Minimization Algorithm.
- In R: `cvxclustr`.

Convex Clustering Solution Path



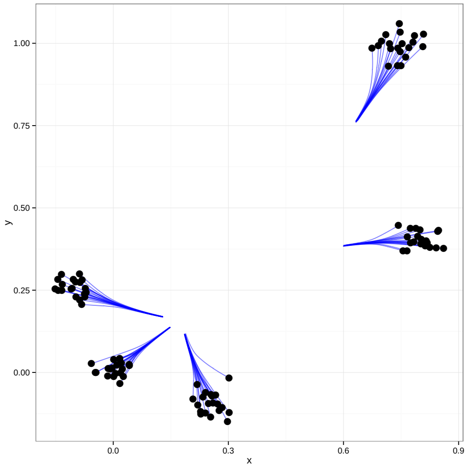
$\lambda = 0$

Convex Clustering Solution Path



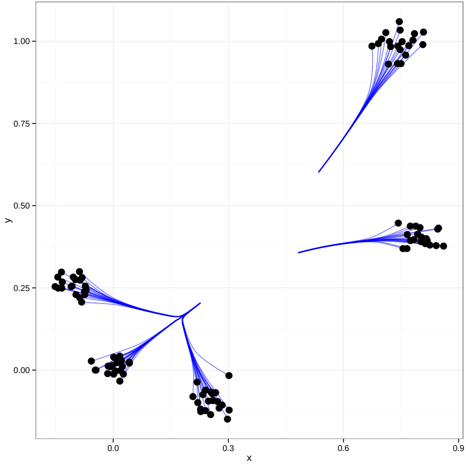
λ

Convex Clustering Solution Path



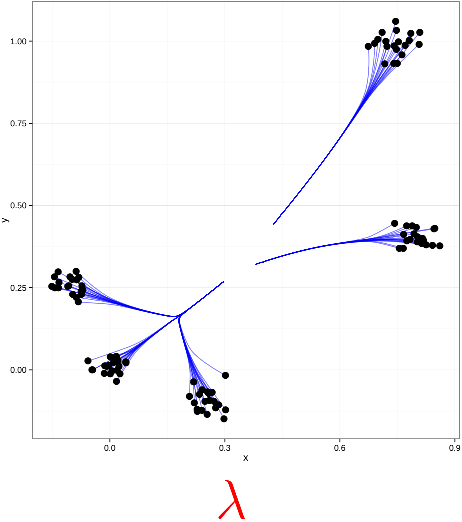
λ

Convex Clustering Solution Path

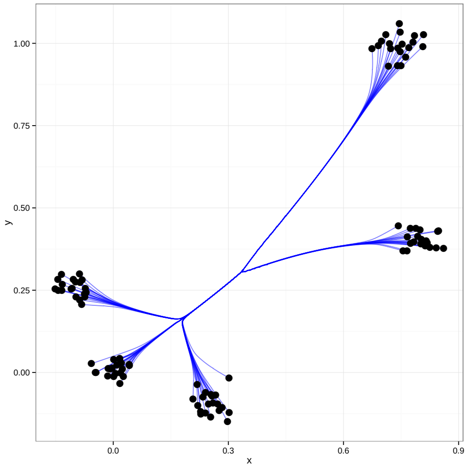


λ

Convex Clustering Solution Path

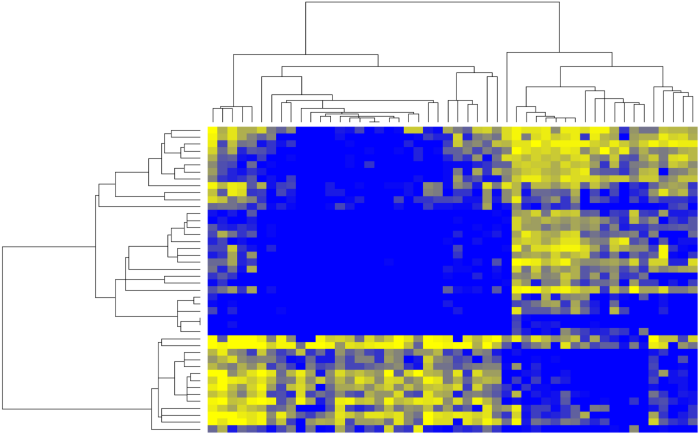


Convex Clustering Solution Path



λ

Convex Biclustering



Convex Biclustering

$$\underset{\mathbf{U}}{\text{minimize}} \quad \frac{1}{2}\|\mathbf{X} - \mathbf{U}\|_F^2 + \lambda\Big(\sum_{i<j} w_{ij}\|\mathbf{U}_{i\cdot} - \mathbf{U}_{j\cdot}\|_2 + \sum_{l<k} \tilde{w}_{lk}\|\mathbf{U}_{\cdot l} - \mathbf{U}_{\cdot k}\|_2\Big)$$

- Checkerboard-like pattern: every data point X_{ij} has its own bicluster centroid U_{ij}.
- Simultaneously fuses row centroids AND column centroids to yield biclusters!

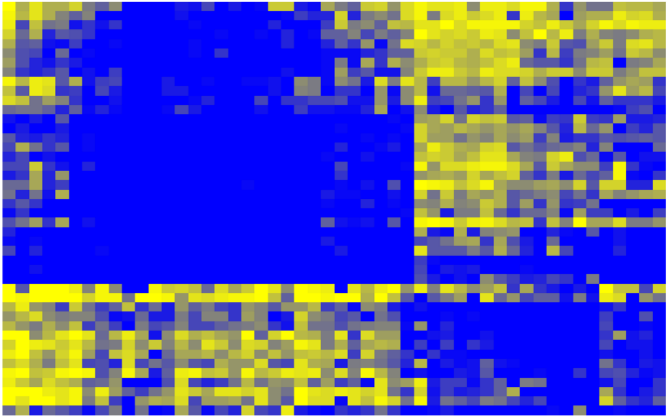
Chi, Allen, and Baraniuk, 2017

Convex Biclustering

$$\underset{\mathbf{U}}{\text{minimize}} \quad \frac{1}{2}\|\mathbf{X}-\mathbf{U}\|_F^2 + \lambda\left(\sum_{i<j} w_{ij}\|\mathbf{U}_{i\cdot}-\mathbf{U}_{j\cdot}\|_2 + \sum_{l<k} \tilde{w}_{lk}\|\mathbf{U}_{\cdot l}-\mathbf{U}_{\cdot k}\|_2\right)$$

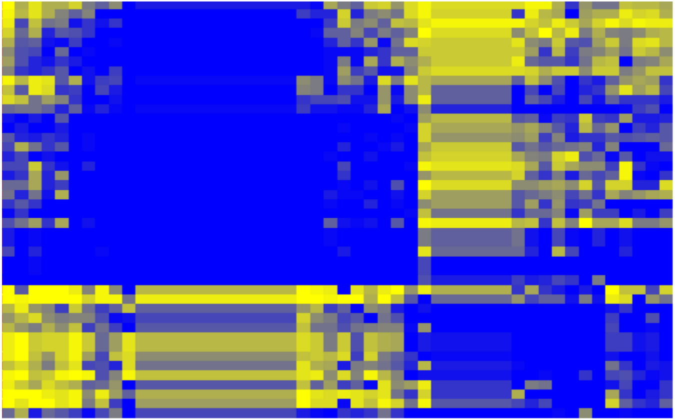
- λ controls BOTH bicluster assignments and # of biclusters.
- Weights similar to convex clustering.
  - Must sum to $1/\sqrt{p}$ and $1/\sqrt{n}$ to ensure the same fusion rate.
- Algorithm: Dystra-like Proximal Algorithm + AMA.
- In R: `cvxbiclustr`.

Convex Biclustering Solution Path



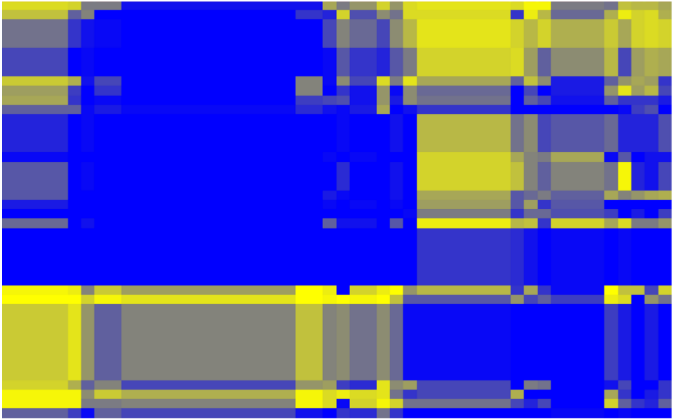
$\lambda = 0$

Convex Biclustering Solution Path



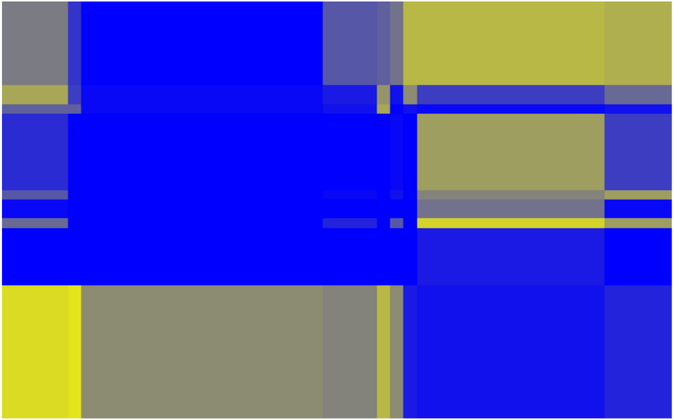
λ

Convex Biclustering Solution Path



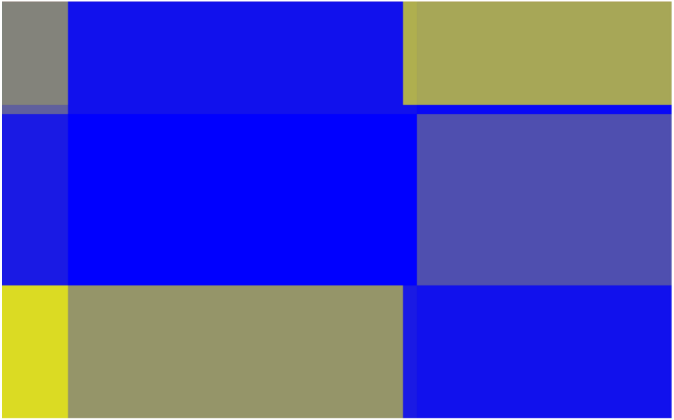
λ

Convex Biclustering Solution Path



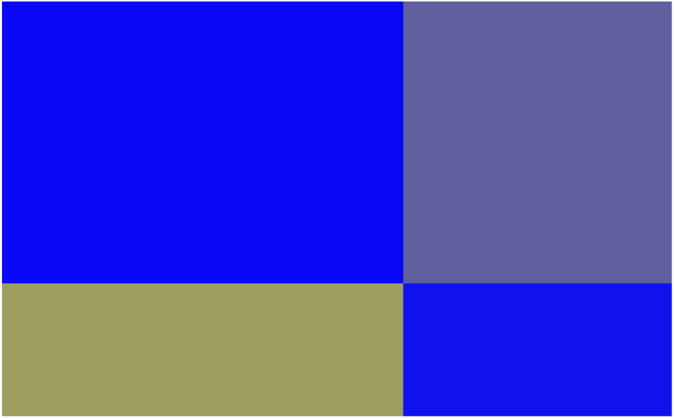
λ

Convex Biclustering Solution Path



λ

Convex Biclustering Solution Path



λ

Convex Biclustering Solution Path

λ

Advantages

The Good:

- Global solution!
  - Stable, reproducible results.
- One tuning parameter.
  - λ controls BOTH # of clusters & cluster assignments.
  - Can select in data-driven manner - **Cross Validation**!
- Statistical Consistency.

The Bad:

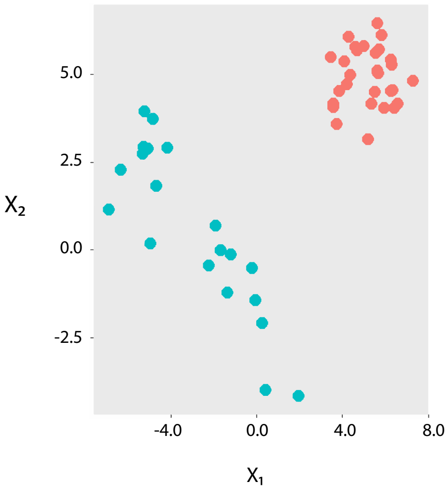
- Inference.
- Nested family of clustering solutions?
- Slower iterative algorithms to find solution.

Chi, Allen, and Baraniuk, 2017; Tan & Witten 2015; Radchenko & Mukherjee, 2017

Inference for Clustering

- Are there true clusters in my data?
- How many clusters?

Our Approach: Inference for cluster means.



Heller and Ghahramani (2005); Liu et al. (2008); Kimes et al. (2017); Huang et al. (2015); Hyun et al. (2016)

Background & Objective

Classical Inference for Multivariate Means:

- One sample:
  - $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - $\mathrm{H}_0 : \quad \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{vs.} \quad \mathrm{H}_\mathrm{A} : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$
  - Hotelling's T^2: $\mathrm{T}^2 = (\mathbf{X} - \hat{\mu})^\mathrm{T} \hat{\mathbf{S}}^{-1} (\mathbf{X} - \hat{\mu}) \sim \frac{\mathrm{p}(\mathrm{n}-1)}{\mathrm{n}-\mathrm{p}} \mathrm{F}_{\mathrm{p},\mathrm{n}-\mathrm{p}}$.
- Two sample:
  - $\mathbf{X}_1 \sim \mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ & $\mathbf{X}_2 \sim \mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$
  - $\mathrm{H}_0 : \quad \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{vs.} \quad \mathrm{H}_\mathrm{A} : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$
  - Hotelling's 2-Sample T^2:
    $\mathrm{T}^2 = \frac{\mathrm{n}_1 \mathrm{n}_2}{\mathrm{n}_1 + \mathrm{n}_2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\mathrm{T} \hat{\mathbf{S}}_{\mathrm{pooled}}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \sim \frac{\mathrm{p}(\mathrm{n}_1 + \mathrm{n}_2 - 2)}{\mathrm{n}_1 + \mathrm{n}_2 - \mathrm{p} - 1} \mathrm{F}_{\mathrm{p},\mathrm{n}_1 + \mathrm{n}_2 - \mathrm{p} - 1}$.

Background & Objective

Inference on multivariate means in adaptive data analysis?

Selective Inference:

- Inference on Means after Clustering (our focus).
- Inference on Means after Dimension Reduction, Outlier Removal, Feature Selection, etc.

Major Challenge!: Need to decompose randomness in $\mathbf{X}$ due to Hotelling's T^2 and all residual randomness independent of Hotelling's T^2 (Multivariate!).

Selective Inference Literature: *Lee et al. (2016); Fithian et al. (2015); Tian and Taylor (2015); Tibshirani et al. (2016); Hyun et al. (2016)*

New Data Decomposition for Hotelling's T^2

Step 1: New representations of T^2 that are univariate (principal angles!).

Theorem

Let $\mathbf{X} - \boldsymbol{\mu}_0 = \mathbf{U}\,\mathbf{D}\,\mathbf{V}^{\mathrm{T}}$ be the SVD of the data centered by the null mean, $\bar{\mathbf{X}}$ the sample mean, $\hat{\mathbf{S}}$ the sample covariance matrix, and define

$$\theta = \arccos\left(\sqrt{\mathbf{1}_{\mathrm{n}}^{\mathrm{T}}\,\mathbf{V}\,\mathbf{V}^{\mathrm{T}}\,\mathbf{1}_{\mathrm{n}}}\right).$$

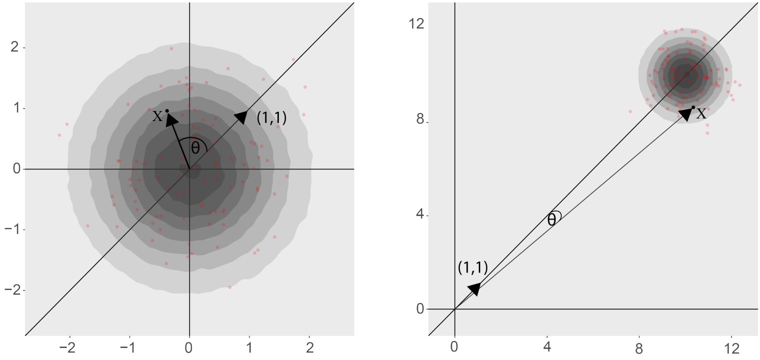
Then, Hotelling's T^2 test-statistic can be written as:

$$\mathrm{T}^2 = (\mathrm{n}-1)\cot^2(\theta)$$

- Hotelling's two-sample T^2 can also be written in terms of a principal angle.

New Data Decomposition for Hotelling's T^2

Step 1: New representations of T^2 that are univariate (principal angles!).



New Data Decomposition for Hotelling's T^2

Step 2: Data decomposition in terms of T^2.

Main Theorem (Paraphrased)

Let $\mathbf{X} = \mathbf{U}\,\mathbf{D}\,\mathbf{V}^{\mathrm{T}}$ be the SVD and $\theta = \arccos\left(\sqrt{\mathbf{1}_{\mathrm{n}}^{\mathrm{T}}\,\mathbf{V}\,\mathbf{V}^{\mathrm{T}}\,\mathbf{1}_{\mathrm{n}}}\right)$ as before. Then,

(a)

$$\mathbf{X} = \mathbf{U}\,\mathbf{D}\,(\boldsymbol{\Gamma}\cos(\theta) + \boldsymbol{\Lambda}\sin(\theta) + \boldsymbol{\Omega})$$

where $\boldsymbol{\Gamma}$, $\boldsymbol{\Lambda}$, and $\boldsymbol{\Omega}$ are random matrices;

(b) θ is independent of $\boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \boldsymbol{\Omega}, \mathbf{U}, \mathbf{D}$.

- Similar decomposition for Hotelling's two-sample T^2 statistic.

New Data Decomposition for Hotelling's T^2

Step 3: Use these new decompositions to conduct selective inference by deriving exact null distributions for the following tests:

$$\mathrm{H}_0: \quad \boldsymbol{\mu}_{\mathrm{k}} = \boldsymbol{\mu}_0 \;\; \text{vs.} \;\; \mathrm{H}_{\mathrm{A}}: \boldsymbol{\mu}_{\mathrm{k}} \neq \boldsymbol{\mu}_0 \;\Big|\; \text{Convex Clustering Solution}$$

(Confidence regions for cluster means)

$$\mathrm{H}_0: \quad \boldsymbol{\mu}_{\mathrm{k}} = \boldsymbol{\mu}_{\mathrm{j}} \;\; \text{vs.} \;\; \mathrm{H}_{\mathrm{A}}: \boldsymbol{\mu}_{\mathrm{k}} \neq \boldsymbol{\mu}_{\mathrm{j}} \;\Big|\; \text{Convex Clustering Solution}$$

(Test whether two clusters are truly separate)

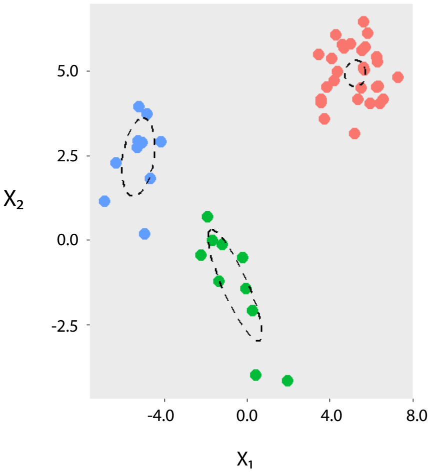
Skipping the details ...

Theorem (Very Paraphrased)

Null distribution is proportional to a truncated F-distribution.

Inference for Convex Clustering: Toy Example

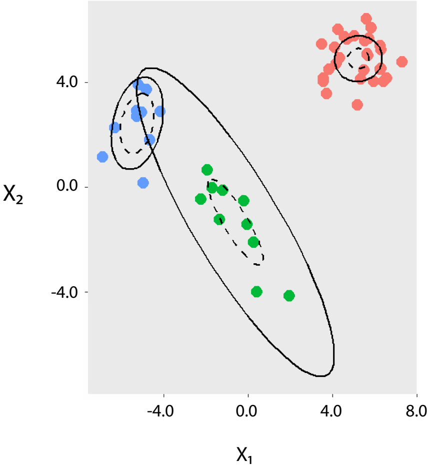
- Confidence ellipsoids for cluster means.
  - Naive: dashed lines.
- Two sample test for equality of cluster green and blue means.
  - Naive: p-value = 1.683517e-07



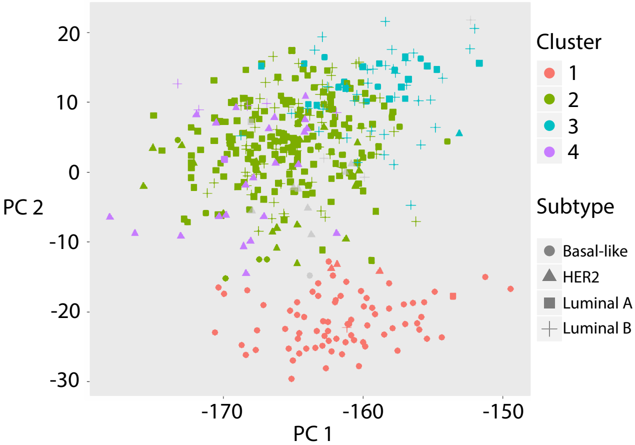
Inference for Convex Clustering: Toy Example

- Confidence ellipsoids for cluster means.
  - Naive: dashed lines.
  - Ours: solid lines.

- Two sample test for equality of cluster green and blue means.
  - Naive:
    p-value = 1.683517e-07
  - Ours:
    p-value = 0.1598832



Inference for Convex Clustering: Breast Cancer Example



- TCGA Breast Cancer Gene Expression Data (log-transformed RNASeq).
- $n = 445$ patients with known subtypes & $p = 353$ genes with known BRCA mutations.

Inference for Convex Clustering: Breast Cancer Example

Comparison Cluster	P-value	Holm Corrected Threshold
1 (Basal) vs 3 (Lum. B)	1.23e-11	8.33e-3
1 (Basal) vs 2 (Lum. A)	6.77e-4	1.00e-2
1 (Basal) vs 4 (HER2)	1.19e-3	1.25e-2
3 (Lum. B) vs 4 (HER2)	2.00e-3	1.67e-2
2 (Lum. A) vs 4 (HER2)	8.03e-3	2.50e-2
2 (Lum. A) vs 3 (Lum. B)	0.17	0.05

Our Objective

Watch your data form clusters & biclusters!

Goal

- Dendograms & Clusterheatmaps.
- Convex clustering & biclustering solution paths.

Problems:

- Potential fissions.
  - *Hocking et al. 2011; Tan & Witten 2015*
- Need exact λ where all fusions occur.
  - Existing algorithms solve for one λ at a time.
  - LAR / Path algorithm for Generalized Lasso doesn't work for convex clustering problem.

Computationally way too slow!

Our Objective

Watch your data form clusters & biclusters!

Goal

- Dendograms & Clusterheatmaps.
- Convex clustering & biclustering solution paths.

Our Approach: **Algorithmic Regularization Paths**

- Quickly approximate clustering solution path at a very fine resolution.
  - *Hu, Allen, & Chi, 2017*

Algorithmic Regularization Paths for Clustering

Classical Regularization Paths

Start: Each observation is its own cluster & no regularization.

Do: Increase the regularization level (λ) by a tiny amount.

Do: Solve the optimization problem at λ.

Iterate the AMA updates until convergence.

Stop: All observations fused to one cluster.

Output: Solution at each λ as the Clustering Path.

Algorithmic Regularization Paths for Clustering

Idea

Start: Each observation is its own cluster & no regularization.

Do: Perform one iterate of the AMA.

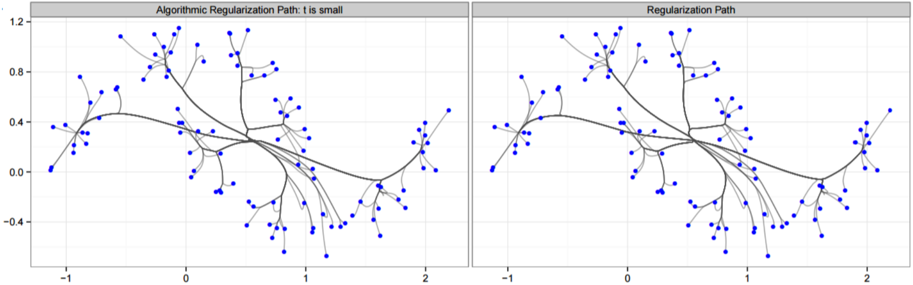
Do: Increase the regularization level by a tiny amount.

Stop: All observations fused to one cluster.

Output: Iterates as the Algorithmic Clustering Path.

Clustering Path Equivalence

Clustering Path Equivalence for small t:



Very Fast!

Clustering Path Equivalence

Theorem

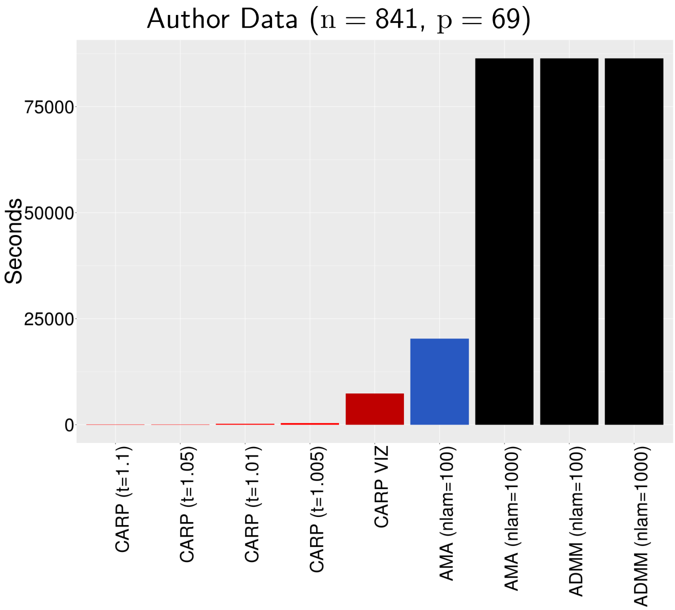
The algorithmic convex clustering path, $\tilde{\mathbf{U}}_t(k)$, is equivalent to the convex clustering path, $\hat{\mathbf{U}}(\lambda)$, as the step size $t \rightarrow 1$:

$$d_H(\hat{\mathbf{U}}(\lambda), \tilde{\mathbf{U}}_t(k)) \rightarrow 0.$$

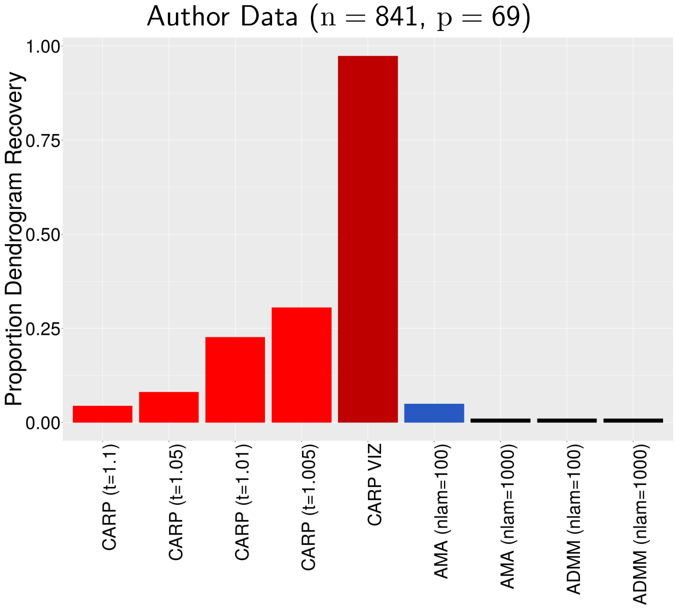
where $d_H(\hat{\mathbf{U}}(\lambda), \tilde{\mathbf{U}}_t(k))$ is the Hausdorff distance:

$$d_H(\hat{\mathbf{U}}(\lambda), \tilde{\mathbf{U}}_t(k)) = \max\left\{ \max_{k} \min_{\lambda} \; ||\, \mathbf{U}(\lambda) - \tilde{\mathbf{U}}_t(k)||_F^2, \right.$$
$$\left. \max_{\lambda} \min_{k} \; ||\, \mathbf{U}(\lambda) - \tilde{\mathbf{U}}_t(k)||_F^2 \right\}.$$

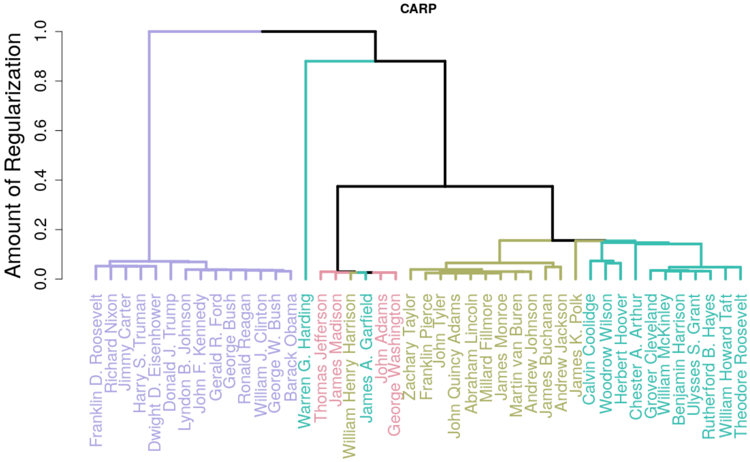
Timing Comparisons



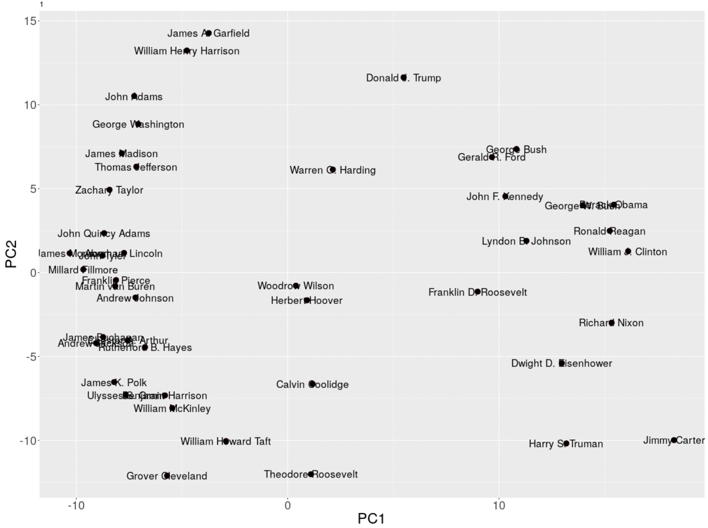
Timing Comparisons



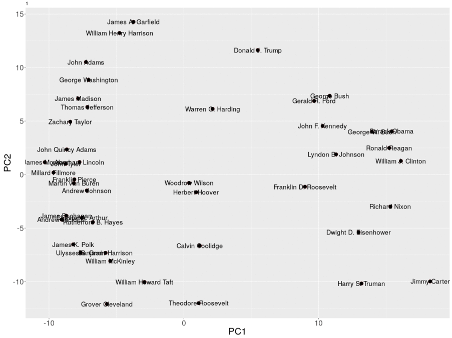
Visualization Results

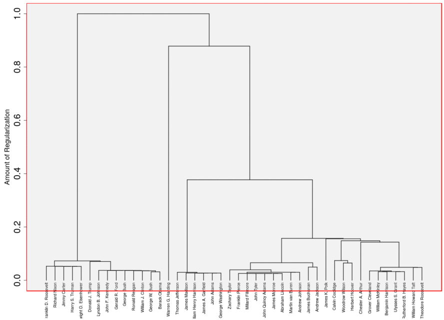


Visualization Results

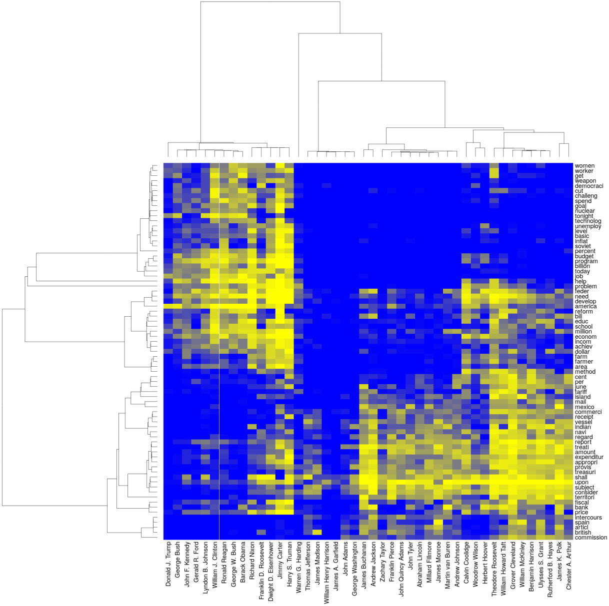


Visualization Results





Visualization Results



Summary

Summary

1. Convex Clustering & Biclustering have many advantages.
2. Developed valid inference procedures for convex clustering.
   - Novel data decomposition for Hotelling's T^2.
   - Applicable to a variety of adaptive data analysis techniques.
3. Developed a fast algorithm to compute cluster solution path.
   - Novel approach: Algorithmic Regularization Paths.
4. Developed interactive & dynamic visualizations for clustering and biclustering.

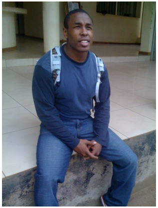
clustRviz

Coming soon!

Acknowledgments & References

John Nagorski,
PhD, Statistics, Rice University

Frederick Campbell,
PhD, Statistics, Rice University



Acknowledgments & References

E. C. Chi, G. I. Allen, and R. Baraniuk, "Convex Biclustering", **73**:1, 10-19, *Biometrics*, 2017.

Y. Hu, E. C. Chi, and G. I. Allen, "ADMM Algorithmic Regularization Paths for Sparse Statistical Machine Learning", In *Splitting Methods in Communication and Imaging, Science and Engineering*, R. Glowinski, W. Yin, and S. Osher (eds), 2017.

J. Nagorski, M. Weylandt, and G. I. Allen, "Dynamic Visualization and Fast Computation of the Solution Path for Convex Clustering", *Preprint*, 2018.

F. Campbell and G. I. Allen, "Inference for Multivariate Means in Adaptive Data Analysis", *Working Paper*, 2018.

Thank You!