ICML

Co-manifold learning with missing data

Gal Mishne, Eric C. Chi and Ronald R. Coifman

Department of Mathematics, Yale University
Department of Statistics, North Carolina State University

June 12, 2019

Yale

NC STATE UNIVERSITY

The Biclustering Problem

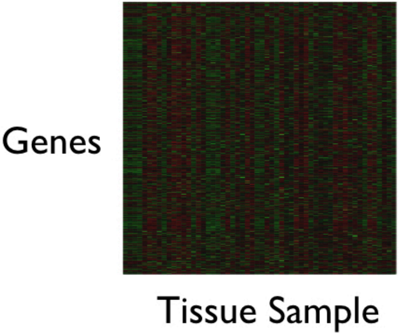
Task

Given a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, find subgroups of rows & columns that go together.

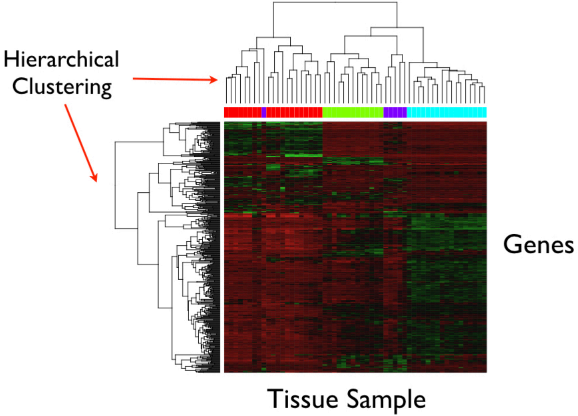
- **Text mining**: similar documents share a small set of highly correlated words.
- **Collaborative filtering**: likeminded customers share similar preferences for a subset of products
- **Cancer genomics**: subtypes of cancerous tumors share similar molecular profiles over a subset of genes

Cancer Genomics

- Lung cancer is heterogenous at the molecular level
- Which genes are driving lung cancer?
- These genes are potential drug targets
- Collect expression data



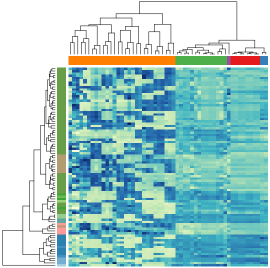
Simple Solution: Cluster Dendrogram



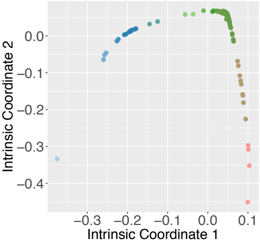
- Each dendrogram is constructed independently of multiscale structure in other dimension.

From Co-clustering to Co-Manifold Learning

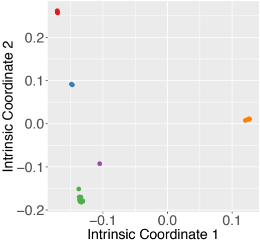
I would add that in many real-world applications there is no "true" fixed number of biclusters, i.e. the truth is a bit more continuous...
–Anonymous Referee 2



Clustered Dendrogram



New Row Coordinate System

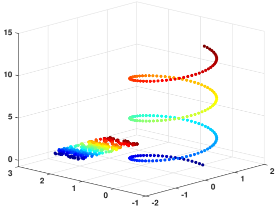


New Column Coordinate System

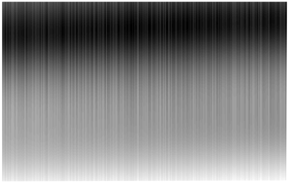
What if data matrices are not completely observed?

Missing data scenario

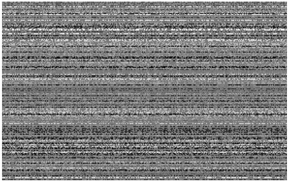
- Complete data: $\mathbf{X} \in \mathbb{R}^{n \times p}$
- Suppose we only get to observe $\Theta \subset \{1, \ldots, n\} \times \{1, \ldots, p\}$.
- Possibly by design: too expensive to collect / measure all *np* possible entries
- **Goal:** Recover row and column coordinate systems, not necessarily complete missing data



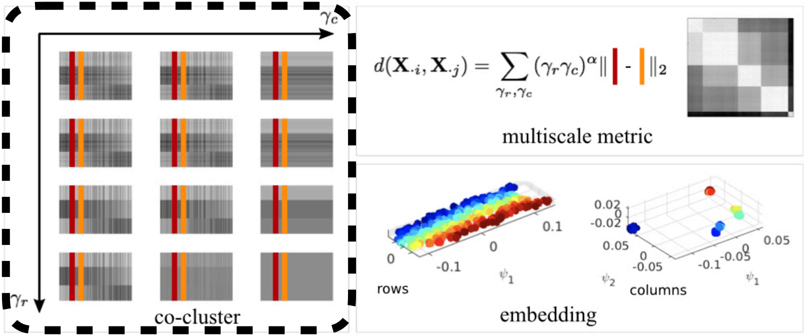
y **- helix**
z **- 2D plane**



$$\mathbf{X}[i,j] = \|y_i - z_j\|_2$$



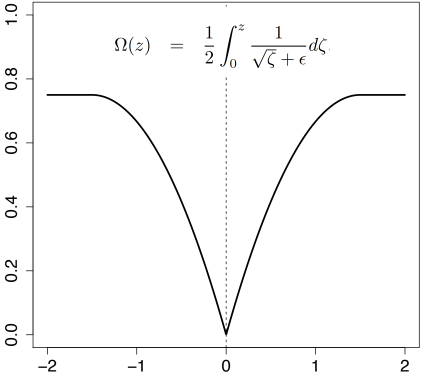
$$\mathcal{P}_\Theta(\mathbf{X}) = \begin{cases} \mathbf{X}[i,j] & (i,j) \in \Theta \\ 0 & \text{otherwise} \end{cases}$$



- Solve co-clustering-missing problem at multiple row and column scales
- Build multiscale row and column metrics
- Calculate non-linear embeddings

Step 1: Co-clustering an Incomplete Data Matrix

$$\min_{\mathbf{U}} F(\mathbf{U}) = \frac{1}{2}\|\mathcal{P}_\Omega(\mathbf{X}-\mathbf{U})\|_F^2 + \gamma_c \sum_{i<j} \Omega(\|\mathbf{U}_{\cdot i} - \mathbf{U}_{\cdot j}\|_2) + \gamma_r \sum_{k<l} \Omega(\|\mathbf{U}_{k\cdot} - \mathbf{U}_{l\cdot}\|_2)$$



- Folded concave penalty $\implies$ less bias towards 0

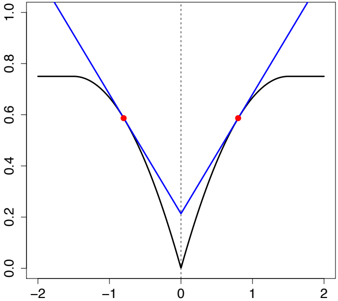
Step 1: Majorization-Minimization (MM)

$$G(\mathbf{U} \mid \mathbf{V}) = \frac{1}{2}\|\tilde{\mathbf{X}} - \mathbf{U}\|_F^2 + \gamma_c \sum_{i<j} \tilde{w}_{c,ij}\|\mathbf{U}_{\cdot i} - \mathbf{U}_{\cdot j}\|_2 + \gamma_r \sum_{k<l} \tilde{w}_{r,kl}\,\|\mathbf{U}_{k\cdot} - \mathbf{U}_{l\cdot}\|_2 + c$$

$$\tilde{\mathbf{X}} = \mathcal{P}_\Omega(\mathbf{X}) + \mathcal{P}_{\Omega^c}(\mathbf{V})$$

$$\tilde{w}_{c,ij} = \Omega'(\|\mathbf{V}_{\cdot i} - \mathbf{V}_{\cdot j}\|_2) \quad \text{and} \quad \tilde{w}_{r,kl} = \Omega'(\|\mathbf{V}_{k\cdot} - \mathbf{V}_{l\cdot}\|_2)$$

Can be solved with Convex Bi-clustering [Chi et al. 2017].



Step 1: Majorization-Minimization (MM)

Majorization:

$$G(\mathbf{U} \mid \mathbf{V}) = \frac{1}{2}\|\mathbf{X} - \mathbf{U}\|_\mathrm{F}^2 + \gamma_c \sum_{i<j} \tilde{w}_{c,ij}\|\mathbf{U}_{\cdot i} - \mathbf{U}_{\cdot j}\|_2 + \gamma_r \sum_{k<l} \tilde{w}_{r,kl}\,\|\mathbf{U}_{k\cdot} - \mathbf{U}_{l\cdot}\|_2 + c$$

- $F(\mathbf{U}) = G(\mathbf{U} \mid \mathbf{U})$
- $F(\mathbf{U}) \leq G(\mathbf{U} \mid \mathbf{V})$ for all $\mathbf{U}$

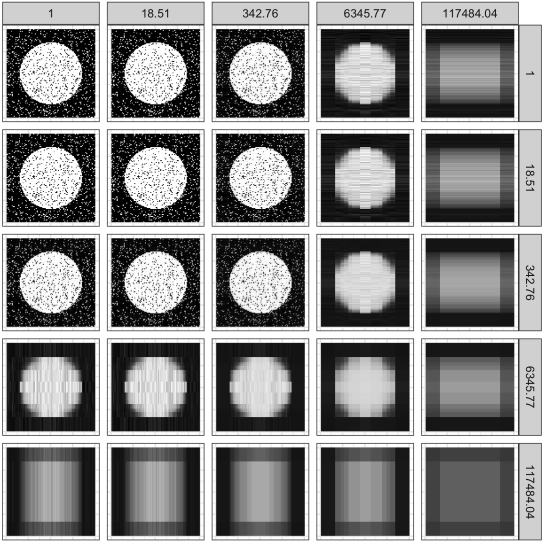
MM: Solve sequence of Convex Biclustering Problems

$$\mathbf{U}_{t+1} \quad = \quad \arg\min_{\mathbf{U}} G(\mathbf{U} \mid \mathbf{U}_t)$$

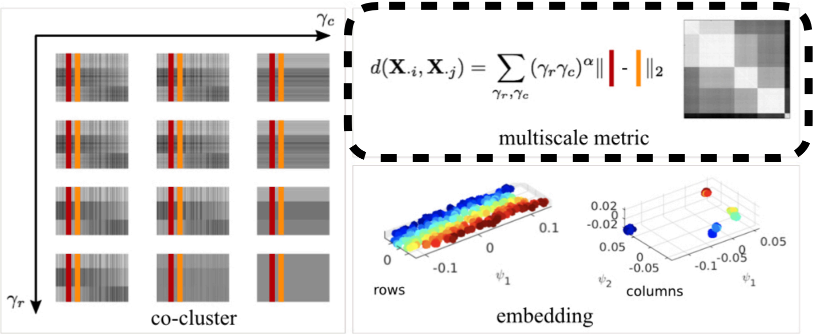
Proposition

Under suitable regularity conditions, the sequence $\mathbf{U}_t$ generated by Algorithm 1 has at least one limit point, and all limit points are d-stationary points of minimizing $F(\mathbf{U})$.

Step 1: Smoothing Rows and Columns at Different Scale



Co-Manifold Learning



- Solve co-clustering-missing problem at multiple row and column scales
- Build multiscale row and column metrics
- Calculate non-linear embeddings

Step 2: Multiscale metric

Intuition:

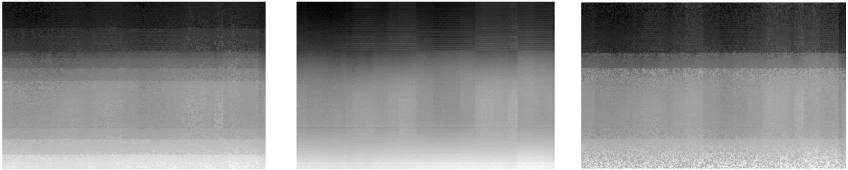
- Pair of rows are close over multiple scale $\to$ distance should be small
- Pair of rows are far apart over multiple scales $\to$ distance should be big

Step 1: Fill in $\mathbf{X}$ over multiple γ_r, γ_c scales: $\tilde{\mathbf{X}}^{(r,c)} = \mathcal{P}_\Theta(\mathbf{X}) + \mathcal{P}_{\Theta^c}(\mathbf{U}(\gamma_r, \gamma_c))$

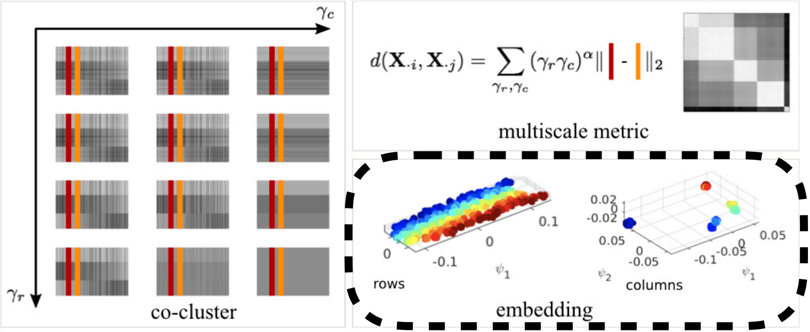
Step 2: Take weighted combination over all scales of pairwise distances

$$d(\mathbf{X}_{i\cdot}, \mathbf{X}_{j\cdot}) = \sum_{r,c} (\gamma_r \gamma_c)^\alpha \|\tilde{\mathbf{X}}_{i\cdot}^{(r,c)} - \tilde{\mathbf{X}}_{j\cdot}^{(r,c)}\|_2$$

- α tunable to emphasize local versus global structure



Co-Manifold Learning



- Solve co-clustering-missing problem at multiple row and column scales
- Build multiscale row and column metrics
- Calculate non-linear embeddings

Step 3: Spectral Embedding

Example: Diffusion Map (Coifman & Lafon, 2006)

- Construct an affinity matrix

$$\mathbf{A}[i,j] \quad = \quad \exp\{-d^2(\mathbf{X}_{i\cdot}, \mathbf{X}_{j\cdot})/\sigma^2\}$$

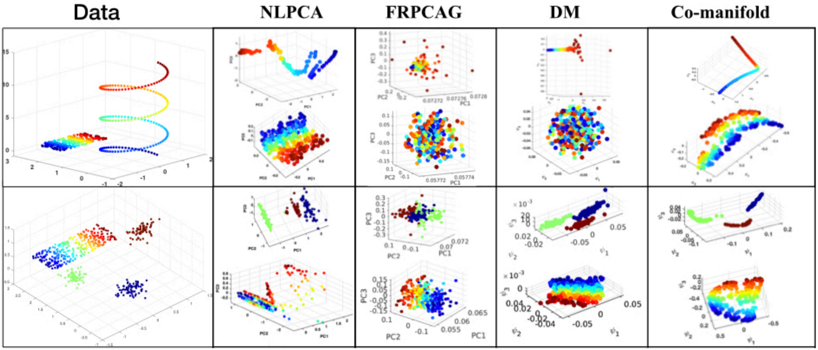
- Compute row-stochastic matrix

$$\mathbf{P} \quad = \quad \mathbf{D}^{-1}\mathbf{A}, \qquad \mathbf{D}[i,i] = \sum_j \mathbf{A}[i,j]$$

- Eigendecomposition of $\mathbf{P}$: keep first d eigenvalues and eigenvectors
- Mapping Ψ embeds the rows into the Euclidean space $\mathbb{R}^d$:

$$\Psi : \mathbf{X}_{i\cdot} \to \left(\lambda_1\psi_1(i), \lambda_2\psi_2(i), \ldots, \lambda_d\psi_d(i)\right)^{\mathsf{T}}.$$

Some Examples



Nonlinear Uncoupled | **Linear Coupled** | **Nonlinear Uncoupled** | **Nonlinear Coupled**

Some Examples

Quantitative evaluation via clustering

Lung500

Linkage