Overlapping Variable Clustering with Statistical Guarantees

Marten Wegkamp

Department of Mathematics
Cornell University

CIRM Meeting in Mathematical Statistics
December 18, 2017

Happy Birthday

Meeting in honor of the 60th birthdays of Oleg Lepski and Alexandre Tsybakov

Joint work with

Florentina Bunea, Xin Bing and Yang Ning
Department of Statistical Science, Cornell University.

Based on

Overlapping Variable Clustering with Statistical Guarantees.
arXiv:1704.06977

Variable clustering

What is **variable** clustering ?

Observable: $\mathbf{X} = (X_1, \ldots, X_p)$ random vector in $\mathbb{R}^p$.

Data: $\mathbf{X}_1, \ldots, \mathbf{X}_n$ iid copies of $\mathbf{X}$.

✓ Goal of **variable** clustering:
Find sub-groups of similar **coordinates** of $\mathbf{X}$, using the data.

⊗ Goal different than data/point clustering:
Find sub-groups of similar observations $\mathbf{X}_i$, $1 \leq i \leq n$.

⊗ Data different than network clustering:
Network data is 0/1 adjacency matrix.

Overlapping Variable Clustering

- $\mathbf{X} = (X_1, \ldots, X_p) \in \mathbb{R}^p$: zero mean random vector.
- $G_k \subseteq \{1, \ldots, p\}$; $G_1 \cup \cdots \cup G_K = \{1, \ldots, p\}$

Goal: Find overlapping sub-groups in a random $\mathbf{X}$

(I) Define clusters G_k such that :

- All X_j with $j \in G_k$ are similar.
- G_k's may overlap.

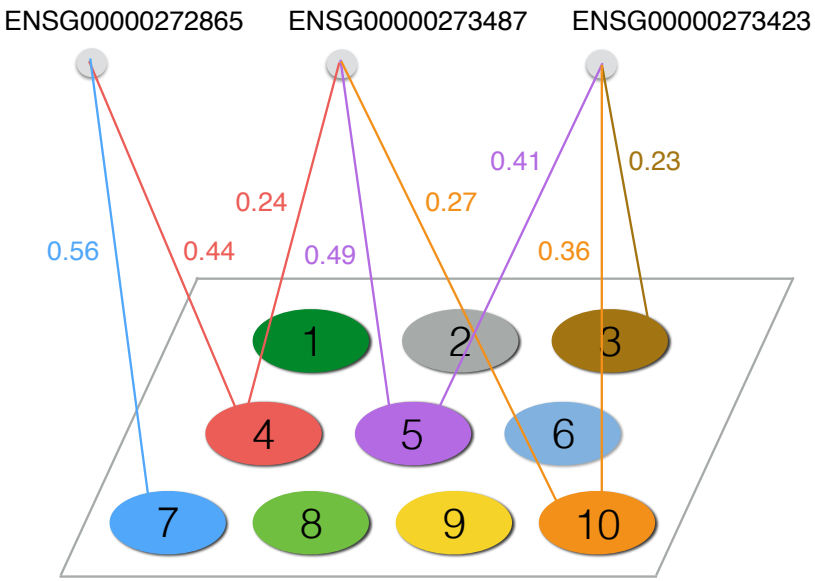
(II) Estimate clusters from $\mathbf{X}_1, \ldots, \mathbf{X}_n$ i.i.d. $\mathbf{X}$.

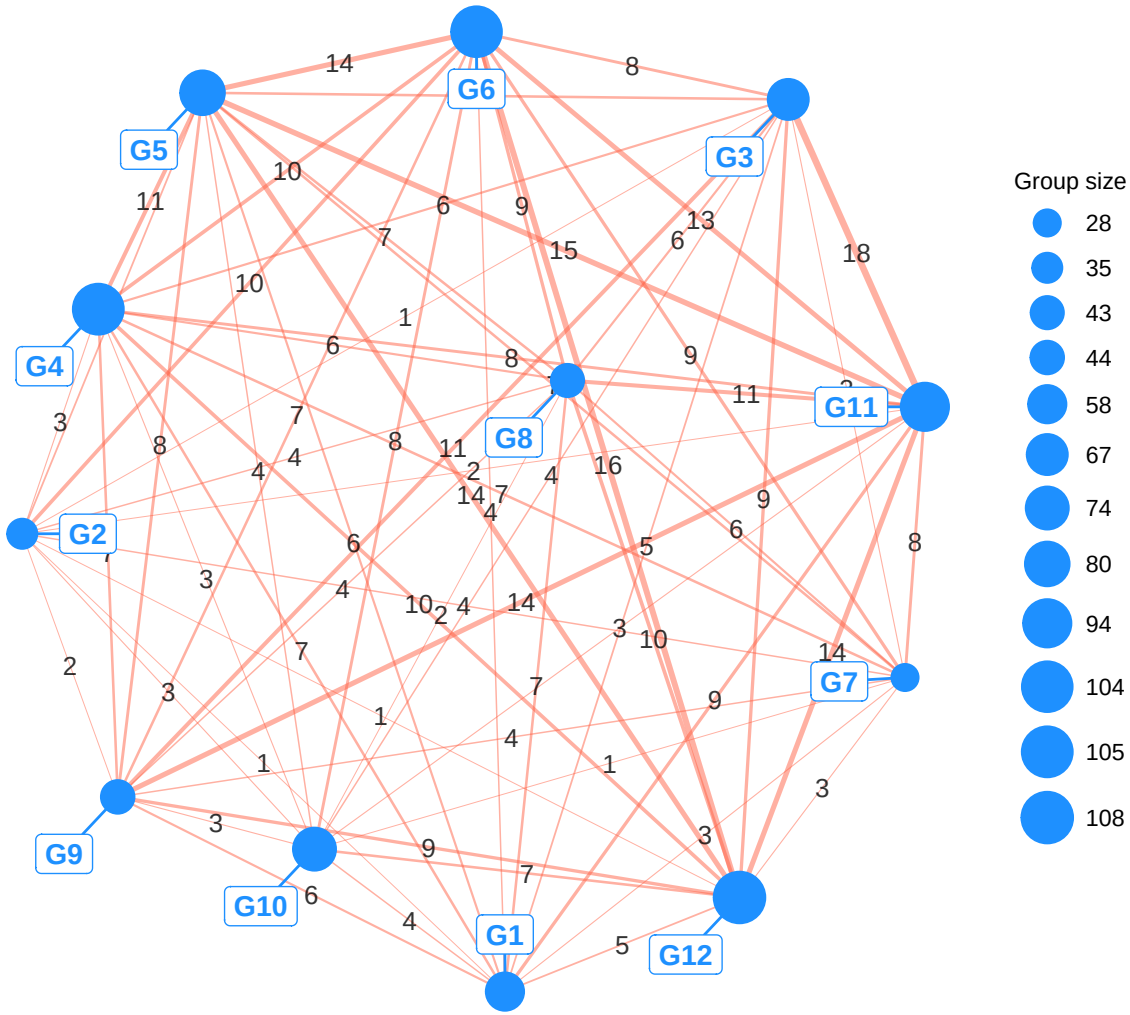
Applications in neuroscience (Craddock et al. 2012, 2013) and genetics (Jiang et al (2004), Wiwie et al. (2015).

- Ongoing work with Jishnu Das (Ragon Institute of MGH, MIT and Harvard)
- RNA-seq dataset of 285 blood platelet samples from patients with different malignant tumors (Best et al.).
- Extract only 500 Ensembl genes to verify if clusters correspond to biological knowledge [Gene Ontology functional annotation, Ashburner et al (2000)].

Co-clustering genes using expression profiles

- RNA-seq transcript level data; Blood platelet samples ($p = 500$) from $n = 285$ individuals.
- ENSG00000273487 and ENSG00000272865 both non-coding RNA: placed together in Cluster 4. ✓
- Each also placed in other clusters. Non-coding RNAs are pleiotropic (multiple functions). ✓
- Model allows for structural zeros. Genes not expressed across samples placed in a separate group. ✓





Nodes represent 12 groups. Numbers shown on the edge between two nodes represent number of genes shared by the two groups.

Our solution: sparse latent factor models

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}$$

- A is a row sparse allocation matrix.
- $\mathbf{Z} \in \mathbb{R}^K$: vector of zero mean latent variables with covariance matrix C.
- $\mathbf{E} \in \mathbb{R}^p$: noise with zero-mean and covariance matrix $\Gamma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$.
- $\mathbf{Z}$ and $\mathbf{E}$ are uncorrelated.
- K is unknown.

This model is not identifiable

- Would like to define: $G_k := \{j \in \{1, \ldots, p\} : \ A_{jk} \neq 0\}$.
- Issue: $AZ = AQQ^T Z$, for any orthogonal Q.

Identifiable sparse latent factor models

Allocation $p \times K$ matrix A satisfies

(i) $\sum_{k=1}^{K} |A_{ik}| \leq 1$, for each row $i \in \{1, \ldots, p\}$.

(ii) For every column $k \in \{1, \ldots, K\}$, there exist at least two rows $i \in \{1, \ldots, p\}$ such that $|A_{ik}| = 1$ while $A_{i\ell} = 0$ for all $\ell \neq k$ (pure variables).

(iii) $C = \mathrm{Cov}(Z)$ with

$$\Delta(C) =: \min_{j \neq k} \left(\min\{C_{jj}, C_{kk}\} - |C_{jk}| \right) > 0.$$

Latent factor model

Remarks

- Flexible way to generate covariance matrices Σ with positive and negative values.
- $\Gamma = \text{cov}(\mathbf{E})$ may have distinct values on the diagonal.
- Condition on C is mild.
- Extends non-overlapping variable model, where each row of A is of the form (ii); see Bunea et al (2015, 2016).
- Spoiler alert: A is identifiable up to signed permutations.

Latent factor model

Remarks

- (i) allows each row $A_{j\cdot}$ to be sparse (to avoid that each X_i is associated with all latent factors).
- (ii) requires that some components of $\mathbf{X}$ are associated with one and only one latent factor (pure variables, pure nodes, anchor words).
- In non-overlapping clustering, *all* X_j's are pure variables.

Aim:

To cluster groups based on

- the dependence of $\mathbf{X}$ and its latent factor $\mathbf{Z}$ and
- the direction of their correlation.

Identifiable sparse latent variable models

The pure variable assumption

A pure variable X_j associates with only one latent factor Z_k.

Pure variables are crucial in building overlapping clusters

- Clusters $G_k := \{j \in \{1, \ldots, p\} : \ A_{ja} \neq 0\}$ are defined by unobserved Z_k (Z_k = (biological) function).
- A pure variable X_j is an observable proxy for a Z_k (Observable X_j performs function Z_k. It anchors G_k).

Identifiability

Definition

Pure variable set I is the index set of pure variables.

$$\begin{aligned} I &= I_1 \cup \cdots \cup I_K \qquad \text{(partition)} \\ I_k &= \{i \in [p] : |A_{ik}| = 1, A_{i\ell} = 0, \text{ for any } \ell \neq k\} \end{aligned}$$

Challenge:

- To find K
- To distinguish between the set I and its complement J.

Example

Let $C = \text{diag}(1, 2, 3)$ and

$$A = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1/2 & -1/2 & 0 \\ 2/3 & 1/6 & -1/6 \end{pmatrix}$$

$ACA^T =$

$$\begin{pmatrix} 1 & -1 & 1 & 0 & 0 & 0 & 0 & 1/2 & 2/3 \\ -1 & 1 & -1 & 0 & 0 & 0 & 0 & -1/2 & -2/3 \\ 1 & -1 & 1 & 0 & 0 & 0 & 0 & 1/2 & 2/3 \\ 0 & 0 & 0 & 2 & -2 & 0 & 0 & 1 & -1/3 \\ 0 & 0 & 0 & -2 & 2 & 0 & 0 & -1 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 3 & 3 & 0 & -1/2 \\ 0 & 0 & 0 & 0 & 0 & 3 & 3 & 0 & -1/2 \\ 1/2 & -1/2 & 1/2 & 1 & -1 & 0 & 0 & 3/4 & 1/6 \\ 2/3 & -2/3 & 2/3 & -1/3 & 1/3 & -1/2 & -1/2 & 1/6 & 7/12 \end{pmatrix}$$

Look at diagonal!

$ACA^T + \Gamma =$

$$\begin{pmatrix} * & -1 & 1 & 0 & 0 & 0 & 0 & 1/2 & 2/3 \\ -1 & * & -1 & 0 & 0 & 0 & 0 & -1/2 & -2/3 \\ 1 & -1 & * & 0 & 0 & 0 & 0 & 1/2 & 2/3 \\ 0 & 0 & 0 & * & -2 & 0 & 0 & 1 & -1/3 \\ 0 & 0 & 0 & -2 & * & 0 & 0 & -1 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & * & 3 & 0 & -1/2 \\ 0 & 0 & 0 & 0 & 0 & 3 & * & 0 & -1/2 \\ 1/2 & -1/2 & 1/2 & 1 & -1 & 0 & 0 & * & 1/6 \\ 2/3 & -2/3 & 2/3 & -1/3 & 1/3 & -1/2 & -1/2 & 1/6 & * \end{pmatrix}$$

Oops! $*$ are arbitrary positive numbers

Example

Option 1: Look for sparsity pattern
(works for non-negative A and diagonal C)

Pure versus impure

Option 2: Look for maxima

Proposition (Pure variable test)

For each row i, define

$$\begin{aligned} M_i &:= \max_{j \in [p] \setminus \{i\}} |\Sigma_{ij}| \\ S_i &:= \{j \in [p] \setminus \{i\} : |\Sigma_{ij}| = M_i\}. \end{aligned}$$

For given A and its induced pure variable set I, we have

$$i \in I \iff M_i = \max_{k \in [p] \setminus \{j\}} |\Sigma_{kj}| \text{ for all } j \in S_i.$$

Look for maxima in $(\Sigma_{ij})_{i \neq j}$

$$\begin{pmatrix}
* & -1 & 1 & 0 & 0 & 0 & 0 & 1/2 & 2/3 \\
-1 & * & -1 & 0 & 0 & 0 & 0 & -1/2 & -2/3 \\
1 & -1 & * & 0 & 0 & 0 & 0 & 1/2 & 2/3 \\
0 & 0 & 0 & * & -2 & 0 & 0 & 1 & -1/3 \\
0 & 0 & 0 & -2 & * & 0 & 0 & -1 & 1/3 \\
0 & 0 & 0 & 0 & 0 & * & 3 & 0 & -1/2 \\
0 & 0 & 0 & 0 & 0 & 3 & * & 0 & -1/2 \\
1/2 & -1/2 & 1/2 & 1 & -1 & 0 & 0 & * & 1/6 \\
2/3 & -2/3 & 2/3 & -1/3 & 1/3 & -1/2 & -1/2 & 1/6 & *
\end{pmatrix}$$

We find $\mathcal{I} = \{\{1,2,3\},\{4,5\},\{6,7\}\}$ and $J = \{8,9\}$.

Identifiability

Theorem

For any $\mathbf{X}$ generated by a sparse factor model,
we can construct the pure variable set I and its partition
$\mathcal{I} =: \{I_1, \ldots, I_K\}$ uniquely from $\Sigma = \mathrm{Cov}(\mathbf{X})$,
up to label permutations.

Identifiability

Theorem

- There exists a unique matrix A, up to a signed permutation, such that $\mathbf{X} = A\mathbf{Z} + \mathbf{E}$.
- The associated overlapping clusters $G_1, \ldots, G_K$ are identifiable, up to label switching.

The pure variables assumption is necessary.

Estimation

Outline

- Estimation of I and its partition
- Estimation of A_I.
- Estimation of A_J.

Estimation of I

Algorithm idea

- Use the constructive characterization of I at the population level.
- Replace Σ by the sample covariance $\widehat{\Sigma}$.
- Allow for tolerance $\delta =: \|\widehat{\Sigma} - \Sigma\|_\infty$ when comparing maxima.
- Algorithm returns partition $\widehat{\mathcal{I}}$ (and $\widehat{K}$ and $\widehat{I}$).

Estimation of the partition of pure variables

Conditions

- **X** is subGaussian with subGaussian constant σ^2.
  (This implies $\delta = O(\sigma^2\sqrt{(\log p)/n})$.)
- $\Delta(C) := \nu > 2\max\left(2\delta, \sqrt{2\delta\|C\|_\infty}\right)$

Definition (nearly pure variable set)

$$
\begin{aligned}
J_1 &= \{i \in J : \text{there exists } k \text{ such that } |A_{ik}| \geq 1 - 4\delta/\nu\} \\
J_1^k &= \{i \in J_1 : |A_{ik}| \geq 1 - 4\delta/\nu\}\,.
\end{aligned}
$$

Hence $\{J_1^1, \ldots, J_1^K\}$ forms a partition of J_1.

Estimation of the partition of pure variables

Theorem

Under the above conditions, with high probability,

(a) $\widehat{K} = K$.

(b) $I \subseteq \widehat{I} \subseteq I \cup J_1$.

Moreover, there exists a label permutation π, such that

(c) $I_{\pi(k)} \subseteq \widehat{I}_k \subseteq I_{\pi(k)} \cup J_1^{\pi(k)}$ for each $k \in [K]$.

Estimation of the partition of pure variables

Minimal recovery mistakes: no conditions on A

Pure	$(1, 0, 0, 0, 0, 0)$	In, correct.
Quasi Pure	$(0.99, 0.01, 0, 0, 0, 0)$	In, slight mistake.
Impure	$(0.25, 0.25, 0.001, 0.099, 0.2, 0.2)$	Out, correct.

Exact recovery: conditions on A

Pure	$(1, 0, 0, 0, 0, 0)$	In, correct.
Quasi Pure	$(0.99, 0.01, 0, 0, 0, 0)$	Not allowed.
Impure	$(0.25, 0.25, 0.001, 0.099, 0.2, 0.2)$	Not allowed.
Impure	$(0.25, 0.25, 0.1, 0.1, 0.3, 0.2)$	Out, correct.

Estimation of the allocation submatrix A_I

$$A = \begin{bmatrix} A_I \\ A_J \end{bmatrix}.$$

Signed subpartitions

- A_I consists of rows with $K-1$ 0's and one entry ± 1.
- For any $i \in \widehat{I}_k$, we set $\widehat{A}_{ik} = \pm 1$ and $\widehat{A}_{i\ell} = 0$ for all $\ell \neq k$.
- Sign of $\widehat{A}_{ik}$?
  - (1) Pick an arbitrary element $i \in \widehat{I}_k$ and set $\widehat{A}_{ik} = 1$.
  - (2) For any other $j \in \widehat{I}_k$, $j \neq i$, set $\widehat{A}_{jk} = 2\mathbb{1}\{\widehat{\Sigma}_{ij} > 0\} - 1$.
- Obtain two subgroups $\widehat{I}_k^1$ and $\widehat{I}_k^2$, each consisting of elements with same sign.

Example

Let $C = \text{diag}(1, 2, 3)$ and

$$A = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1/2 & -1/2 & 0 \\ 2/3 & 1/6 & -1/6 \end{pmatrix}$$

$$\begin{pmatrix} * & -1 & 0 & 0 & 0 & 0 & 1/2 & 2/3 \\ -1 & * & 0 & 0 & 0 & 0 & -1/2 & -2/3 \\ 0 & 0 & * & -2 & 0 & 0 & 1 & -1/3 \\ 0 & 0 & -2 & * & 0 & 0 & -1 & 1/3 \\ 0 & 0 & 0 & 0 & * & 3 & 0 & -1/2 \\ 0 & 0 & 0 & 0 & 3 & * & 0 & -1/2 \\ 1/2 & -1/2 & 1 & -1 & 0 & 0 & * & 1/6 \\ 2/3 & -2/3 & -1/3 & 1/3 & -1/2 & -1/2 & 1/6 & * \end{pmatrix}$$

Example

$$\Sigma_{1:2,\cdot} = \begin{pmatrix} * & -1 & 0 & 0 & 0 & 0 & 1/2 & 2/3 \\ -1 & * & 0 & 0 & 0 & 0 & -1/2 & -2/3 \end{pmatrix}$$

$I = \{1,2,3,4,5,6\}$, $I_1 = \{1,2\}$, $I_2 = \{3,4\}$, $I_3 = \{5,6\}$.

- Take $1 \in I_1$, set $A_{1,1} = +1$.
- Take $2 \in I_1$, set $A_{2,1} = -1$ since $\Sigma_{12} = -1 < 0$.
- Set $I_1^1 = \{1\}$, $I_1^2 = \{2\}$.

Example

$$\Sigma_{3:4,\cdot} = \begin{pmatrix} 0 & 0 & * & -2 & 0 & 0 & 1 & -1/3 \\ 0 & 0 & -2 & * & 0 & 0 & -1 & 1/3 \end{pmatrix}$$

- Take $3 \in I_2$, set $A_{3,2} = 1$.
- Take $4 \in I_2$, set $A_{4,2} = -1$ since $\Sigma_{3,4} = -2 < 0$.
- Set $I_2^1 = \{3\}$, $I_2^2 = \{4\}$.

Example

$I = \{1, 2, 3, 4, 5, 6\}$, $I_1 = \{1, 2\}$, $I_2 = \{3, 4\}$, $I_3 = \{5, 6\}$.

- ...
- Set $I_3^1 = \{5, 6\}$, $I_3^2 = \emptyset$.
- Set

$$A_I = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

Estimation of the allocation submatrix A_J

- We estimate the matrix A_J row by row.
- Rearrange Σ and A as

$$\Sigma = \begin{bmatrix} \Sigma_{II} & \Sigma_{IJ} \\ \Sigma_{IJ} & \Sigma_{JJ} \end{bmatrix} \text{ and } A = \begin{bmatrix} A_I \\ A_J \end{bmatrix}.$$

so

$$\Sigma = \begin{bmatrix} \Sigma_{II} & \Sigma_{IJ} \\ \Sigma_{JI} & \Sigma_{JJ} \end{bmatrix} = \begin{bmatrix} A_I C A_I^T & A_I C A_J^T \\ A_J C A_I^T & A_J C A_J^T \end{bmatrix} + \begin{bmatrix} \Gamma_{II} & 0 \\ 0 & \Gamma_{JJ} \end{bmatrix}.$$

Example

Let $C = \text{diag}(1, 2, 3)$.

$$A = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1/2 & -1/2 & 0 \\ 2/3 & 1/6 & -1/6 \end{pmatrix}$$

Example

$$\Sigma = \begin{pmatrix} * & -1 & 0 & 0 & 0 & 0 & 1/2 & 2/3 \\ -1 & * & 0 & 0 & 0 & 0 & -1/2 & -2/3 \\ 0 & 0 & * & -2 & 0 & 0 & 1 & -1/3 \\ 0 & 0 & -2 & * & 0 & 0 & -1 & 1/3 \\ 0 & 0 & 0 & 0 & * & 3 & 0 & -1/2 \\ 0 & 0 & 0 & 0 & 3 & * & 0 & -1/2 \\ 1/2 & -1/2 & 1 & -1 & 0 & 0 & * & 1/6 \\ 2/3 & -2/3 & -1/3 & 1/3 & -1/2 & -1/2 & 1/6 & * \end{pmatrix}$$

Example

$$\Sigma_{I,I} = \begin{pmatrix} * & -1 & 0 & 0 & 0 & 0 \\ -1 & * & 0 & 0 & 0 & 0 \\ 0 & 0 & * & -2 & 0 & 0 \\ 0 & 0 & -2 & * & 0 & 0 \\ 0 & 0 & 0 & 0 & * & 3 \\ 0 & 0 & 0 & 0 & 3 & * \end{pmatrix}$$

Read off $C_{11} = 1, C_{22} = 2, C_{33} = 3$!

Example

We found

$$A_I = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

Example

$$\Sigma_{I,J} = \begin{bmatrix} 1/2 & 2/3 \\ -1/2 & -2/3 \\ 1 & -1/3 \\ -1 & 1/3 \\ 0 & -1/2 \\ 0 & -1/2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} * C * A_J^T$$

Example

$$\begin{bmatrix} 1/2 & 2/3 \\ 1/2 & 2/3 \\ 1 & -1/3 \\ 1 & -1/3 \\ 0 & -1/2 \\ 0 & -1/2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} * C * A_J^T$$

Sign multiplication

Example

$$\begin{bmatrix} 1/2 & 2/3 \\ 1 & -1/3 \\ 0 & -1/2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} * C * A_J^T$$

Averaging

Example

$$\begin{bmatrix} 1/2 & 2/3 \\ 1 & -1/3 \\ 0 & -1/2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} \beta_{11} & \beta_{21} \\ \beta_{12} & \beta_{22} \\ \beta_{13} & \beta_{33} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 1/2 & 2/3 \\ 1/2 & -1/6 \\ 0 & -1/6 \end{bmatrix}$$

Solving

Example

We find the matrix

$$\begin{pmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \\ 2/3 & -1/6 & -1/6 \end{pmatrix}$$

which is A up to a sign permutation of the 2nd column.

Estimation of the allocation sub-matrix A_J

- $\Sigma_{IJ} = A_I C A_J^T$
- $\theta^j =: CA_{j\cdot}$ for each $j \in J$
- $\theta^j_k =: \frac{1}{|I_k|} \sum_{i \in I_k} A_{ik} \Sigma_{ij}$
- $C_{kk} = \frac{1}{|I_k|(|I_k|-1)} \sum_{i,j \in I_k, i \neq j} |\Sigma_{ij}|$
- $C_{km} = \frac{1}{|I_k||I_m|} \sum_{i \in I_k, j \in I_m} |\Sigma_{ij}|$

Estimation of the allocation submatrix A_J

Estimation of a "non-pure" row $A_{j\cdot} =: \beta$

- Under the model, $\beta \in \mathbb{R}^K$ satisfies:
  - $\beta = C^{-1}\theta$;
  - β sparse;
  - $\|\beta\|_1 \leq 1$.
- Available: estimators $\widehat{\theta}$ and $\widehat{C}$.
  Crucial ingredient: estimated pure variable set.
- Construct $\widehat{\Omega}$ to estimate C^{-1}. Build pre-estimate $\bar{\beta} = \widehat{\Omega}\widehat{\theta}$.
- Final estimate is $\widehat{\beta}$, sparse projection of $\bar{\beta}$.

Motivation of the proposed estimator of Ω.

The decomposition

$$\begin{aligned} \bar{\beta}^j - \beta^j &= \widehat{\Omega}(\widehat{\theta}^j - \theta^j) + (\widehat{\Omega} - \Omega)\theta^j \\ &= \widehat{\Omega}(\widehat{\theta}^j - \theta^j) + (\widehat{\Omega}C - I)\beta^j, \end{aligned}$$

implies

$$\begin{aligned} \|\bar{\beta}^j - \beta^j\|_\infty &\leq \|\widehat{\Omega}\|_{\infty,1}\|\widehat{\theta}^j - \theta^j\|_\infty + \|\widehat{\Omega}C - I\|_\infty\|\beta^j\|_1 \\ &\leq \|\widehat{\Omega}\|_{\infty,1}\|\widehat{\theta}^j - \theta^j\|_\infty + \|\widehat{\Omega}C - I\|_\infty. \end{aligned}$$

Hence $\widehat{\Omega}$ should render small values for $\|\widehat{\Omega}\|_{\infty,1}$ and $\|\widehat{\Omega}C - I\|_\infty$.

Estimation of $\Omega = C^{-1}$

$$(\widehat{\Omega}, \widehat{t}) = \min_{t \in \mathbb{R}^+,\ \Omega \in \mathbb{R}^{\widehat{K} \times \widehat{K}}} t,$$

subject to

$$\Omega = \Omega^T,\ \|\Omega \widehat{C} - I\|_\infty \leq \lambda t,\ \|\Omega\|_{\infty,1} \leq t,$$

- New estimator:
  Uses $\|\cdot\|_{\infty,1}$ instead of more standard $\|\cdot\|_1$.

- We avoid common assumptions such as
  - condition number of C is bounded
  - number of clusters K is known / bounded.

Estimation of β^j

$$\widehat{\beta}^j = \arg\min_{\beta \in \mathbb{R}^{\widehat{K}}} \|\beta\|_1$$

subject to

$$\|\beta - \bar{\beta}^j\|_\infty \leq \mu,$$

- Solution of LP is sparse and properly scaled.
- Stacking $\widehat{\beta}^j$ over all rows $j \in \widehat{J}$ produces $\widehat{A}_{\widehat{J}}$.
- Merging $\widehat{A}_{\widehat{I}}$ with $\widehat{A}_{\widehat{J}}$ produces our final estimator $\widehat{A}$ of A.

Choice of tuning parameters

- $\delta' = (8\|C\|_\infty/\nu - 3)\,\delta$
- $\lambda = 2\delta'$
- $\mu = 5\|C^{-1}\|_{\infty,1}\delta'$

Estimation of the allocation submatrix A_J

Theorem

Let λ and μ be as defined above. Let $\mathcal{H}_K$ denote the hyperoctahedral group of the signed permutation matrices. Then, for all $1 \leq q \leq \infty$ and $1 \leq i \leq p$,

$$\min_{P \in \mathcal{H}_K} \left\| \widehat{A}_{i\cdot} - (AP)_{i\cdot} \right\|_q \quad \leq \quad 10 (\|A_{i\cdot}\|_0)^{1/q} \|C^{-1}\|_{\infty,1} \delta'$$

with probability larger than $1 - c_1 (n \vee p)^{-c_2}$, for positive constants c_1, c_2, provided $(2\mu + 4\delta/\nu) < 1$.

- $q = +\infty$ leads to inference on support recovery of $\beta^j = A_{j\cdot}$.
- Quality of estimating a sparse vector depends on the interplay between its sparsity and the behavior of the appropriate Gram matrix (here $C = \mathbb{E}[ZZ^T]$).
- The concept of ℓ_q-sensitivity, introduced by Gautier and Tsybakov (2011) and Belloni, Rosenbaum and Tsybakov (2017), is the most general characterization of this interplay to date. It facilitates a link between the ℓ_q-norm of sparse vectors β and the ℓ_∞-norm of the product between the Gram matrix and β, uniformly over vectors β of sparsity s, ranging over a collection of cones.

- In our context, that of a square, *invertible* matrix C, the reciprocal of the ℓ_∞-sensitivity of C becomes essentially $\|C^{-1}\|_{\infty,1}$, which indeed links $\|\beta\|_\infty$ to $\|C\beta\|_\infty$.
- The quantities $(s_i)^{1/q}\|C^{-1}\|_{\infty,1}$ provide concrete substitutes of the reciprocals of the ℓ_q-sensitivities of C, and all of our rates coincide with the minimax rates obtained by Belloni, Rosenbaum and Tsybakov (2017).

Estimation of the overlapping groups

Signed group structure

Estimate the group structure $\widehat{\mathcal{G}} = \{\widehat{G}_1, \ldots, \widehat{G}_{\widehat{K}}\}$ from the columns of $\widehat{A}$:

$$\begin{aligned} \widehat{G}_k &= \{i \in [p] : |\widehat{A}_{ik}| \neq 0\} \\ &= \{\widehat{G}_k^1, \widehat{G}_k^2\} \\ &= \Big\{\{i \in \widehat{G}_k : \widehat{A}_{ik} > 0\}, \{i \in \widehat{G}_k : \widehat{A}_{ik} < 0\}\Big\}. \end{aligned}$$

We quantify the misclassification proportion within each group $\widehat{G}_a$ by

$$\mathrm{GFPP}(\widehat{\mathrm{G}}_\mathrm{a}) := \frac{|(\mathrm{G}_\mathrm{a})^\mathrm{c} \cap \widehat{\mathrm{G}}_\mathrm{a}|}{|(\mathrm{G}_\mathrm{a})^\mathrm{c}|}, \quad \mathrm{GFNP}(\widehat{\mathrm{G}}_\mathrm{a}) := \frac{|\mathrm{G}_\mathrm{a} \cap (\widehat{\mathrm{G}}_\mathrm{a})^\mathrm{c}|}{|\mathrm{G}_\mathrm{a}|}.$$

Set

$$\begin{aligned} J_1 &:= \{i \in J : \exists a \text{ with } |A_{ia}| \geq 1 - 4\delta/\nu\} \\ J_2 &:= \{i \in J : \text{for any } a \text{ with } A_{ia} \neq 0,\ |A_{ia}| > (2\mu) \vee (4\delta/\nu)\} \\ J_3 &:= J \setminus (J_1 \cup J_2). \end{aligned}$$

Estimation of the overlapping groups

Theorem

Under same conditions, with high probability, we have:

(a) $\text{supp}(A_{J_2}) \subseteq \text{supp}(\widehat{A}) \subseteq \text{supp}(A)$,
$\text{sgn}(\widehat{A}_{\widehat{S}}) = \text{sgn}(A_{\widehat{S}})$.

(b) Let $s_j^a = 1\{|A_{ja}| \neq 0\}$ and $t_j^a = 1\{|A_{ja}| \leq (2\mu) \vee (4\delta/\nu)\}$.

$$\text{GFPP}(\widehat{G}_a) = 0; \quad \text{GFNP}(\widehat{G}_a) \leq \frac{\sum_{j \in J_1 \cup J_3 \setminus J_1^a} t_j^a}{\sum_{j \in J} s_j^a + |I_a|}.$$

Related work

Large literature on Non-Negative Matrix Factorization (NMF)

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}; \qquad \mathbf{X}, A, \mathbf{Z} \text{ non-negative matrices.}$$

Goal of NMF different than ours: find $\widetilde{A}$ and $\widetilde{\mathbf{Z}}$ with $\|\mathbf{X} - \widetilde{A}\widetilde{\mathbf{Z}}\| \leq \epsilon$.

In NMF, the pure variable assumption is needed for:

- Identifiability of A, provided $\mathbf{E} = 0$ (Donoho and Stodden, 2007).
- Identifiability in topic models (count data), Arora et al (2013): columns of $\mathbf{X}$ and A sum to 1; $\mathbf{E} = 0$.
- Polynomial time NMF algorithms: Arora et al (2012, 2013); Bittorf et al (2013). Other restrictions on matrices needed.

Network data

Model-based clustering methods for network data.

- For instance, mixed membership stochastic block-models: Airoldi et al (2008), Zhang et al (2008), Lei and Zhu (2014), Abbe and Sandon (2015), Lei and Rinaldo (2015), Guédon and Vershynin (2016), Le et al (2016).
- Data is of different nature: We observe an $p \times p$ binary matrix, with independent Bernoulli entries and we model the mean of this matrix ($\Gamma = 0$).
- Bayesian approaches: (Airoldi, Blei, Fienberg and Xing, 2008)
- Spectral clustering: Zhang et al (2015), He et al (2015) require many pure nodes.

Topic model

Model-based clustering methods for topic model.

- We observe a multinomial distribution, and we postulate $\mathbb{E}[\mathbf{X}] = A\mathbf{Z}$ with non-zero means, and $\mathbb{E}(\mathbf{X}\mathbf{X}^T) = A(\mathbf{Z}\mathbf{Z}^T)A^T + \mathbb{E}[\mathbf{E}\mathbf{E}^T]$.
- Large mostly CS literature: Arora et al. (2012, 2013), Blei (2012), Ke (2016), ...
- Computationally feasible solution: NMF under a separability assumption (anchor words).
- Anchor words are vertices of a simplex.
- Given K, find K vertices among rows of $\mathbb{E}(\mathbf{X}\mathbf{X}^T)$.

Related work

- Large literature on latent factor models for a different problem: **dimension reduction in covariance estimation**.
- $\Sigma = ACA^T + \Gamma$: "low rank + sparse" decomposition.
- Includes sparse PCA.
- Different problem, different identifiability questions: Identifiability of AA^T rather than that of A.

Summary

- We have introduced a flexible latent factor model to handle overlapping variable clusters.
- A with both + and - allows for a more refined cluster interpretation.
- We verified identifiability of the assignment matrix A in presence of noise $\mathbf{E}$.
- We proposed a new, fast algorithm to find the number of clusters and the pure variables.
- Sparse regression method is used to find the coefficients of the non-pure variables.
- Method works well with statistical guarantees for data generated for X sub-Gaussian; immediate extensions to elliptical copula factor models.
- Future work: extensions to topic models, fMRI data and networks.

Bon anniversaire!