Spectral Methods for Latent Variable Models

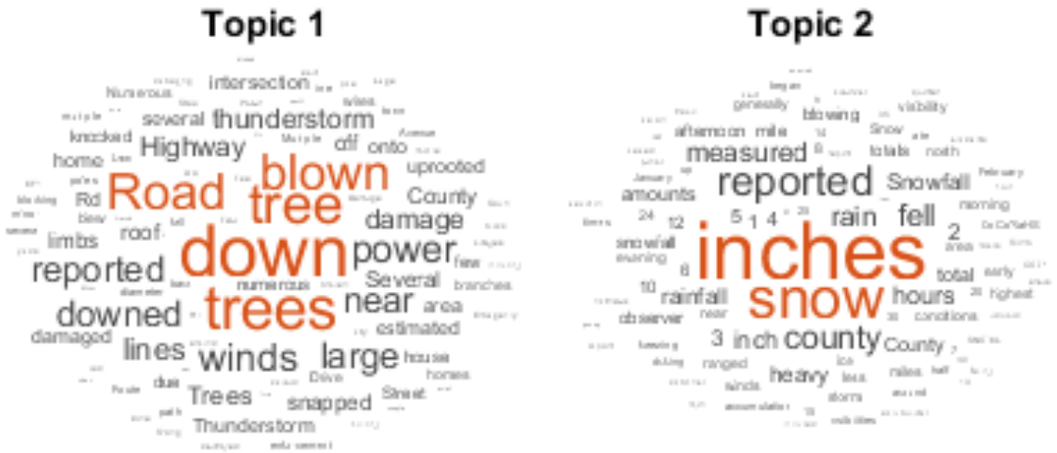
Kaizheng Wang

Department of ORFE
Princeton University

March 20th 2020

Data Diversity

Unstructured, heterogeneous and incomplete information:



Credit: https://www.mathworks.com/help/textanalytics/gs/getting-started-with-topic-modeling.html, https://www.alliance-scotland.org.uk/alliance-homepage-holding-people-networking-2017-01-3/, https://medicalxpress.com/news/2015-04-tumor-only-genetic-sequencing-misguide-cancer.html, https://www.nature.com/articles/nature21386/figures/1, https://viterbi-web.usc.edu/ soltanol/RSC.pdf, Dzenan Hamzic

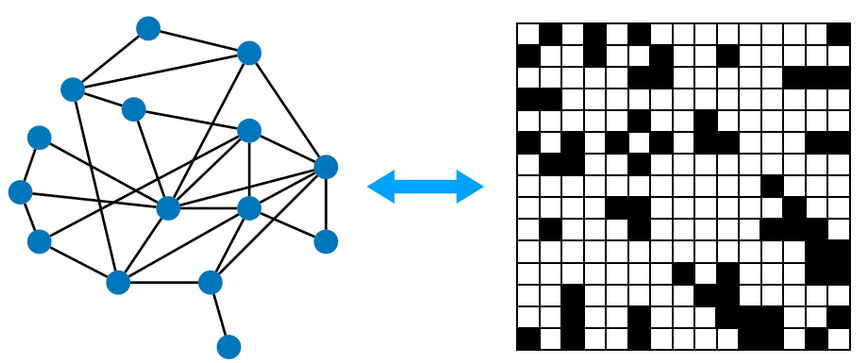
Matrix Representations

Object-by-feature ($n \times d$):

- **Texts:** document-term;
- **Genetics:** individual-marker;
- **Recomm. systems:** user-item.

Object-by-object ($n \times n$):

- **Networks:** adjacency matrices.

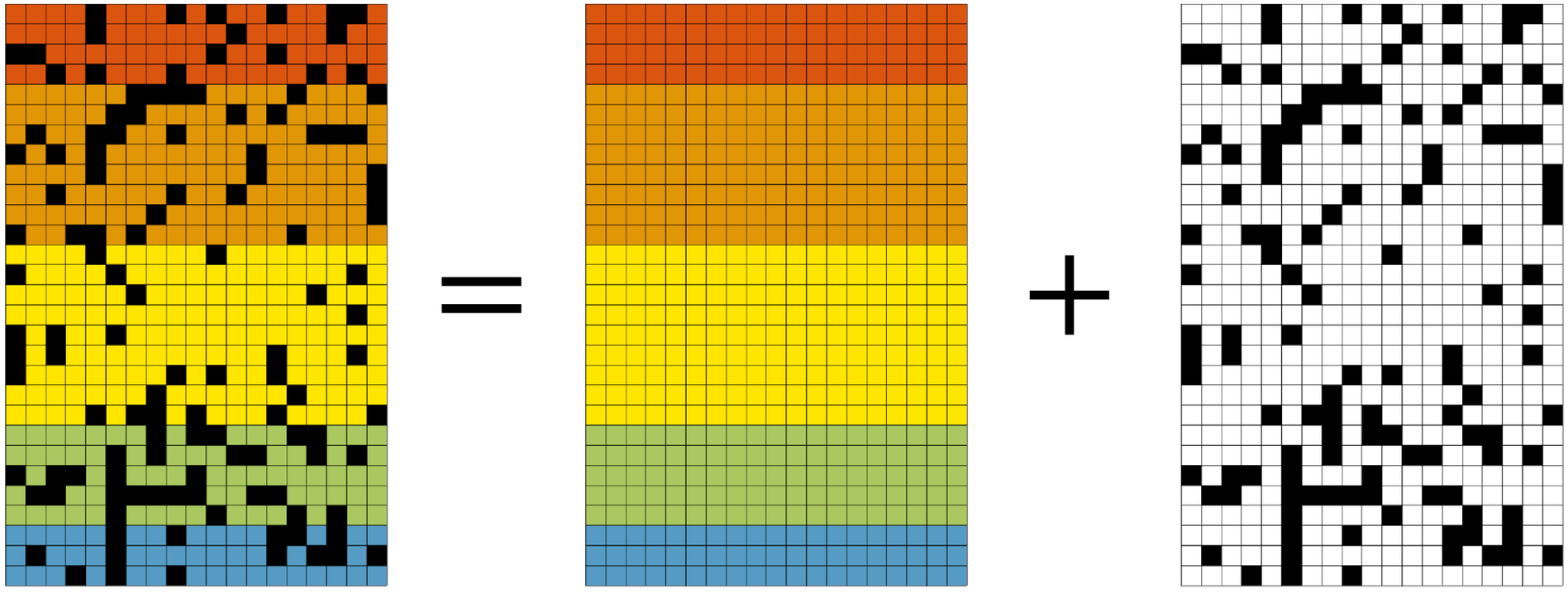


Credit (upper right): https://viterbi-web.usc.edu/ soltanol/RSC.pdf.

Matrix Representations

Common belief: **high ambient** dim. but **low intrinsic** dim.

Low-rank approximation:



Matrix Representations

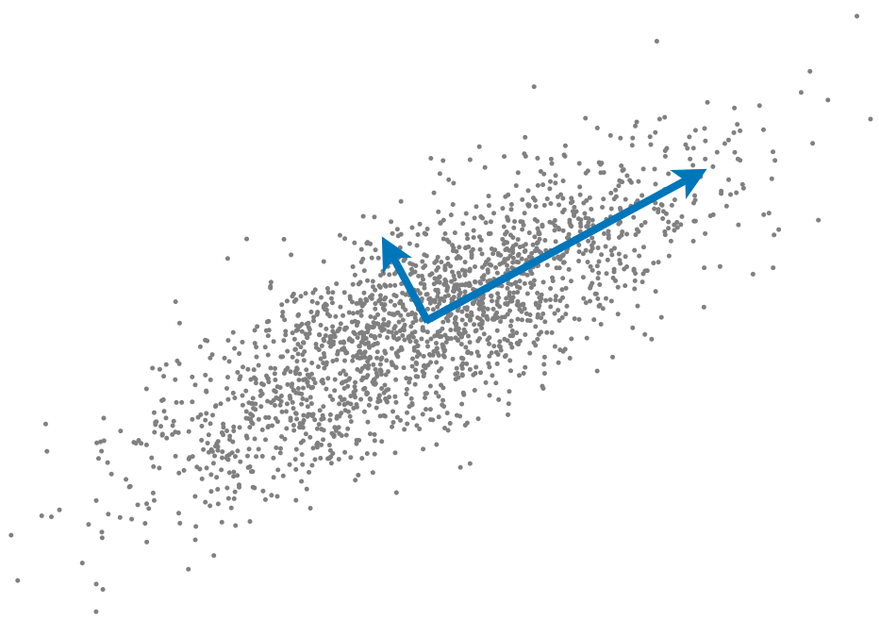
Low-dimensional embedding via **latent variables**:

$$\begin{pmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_n \end{pmatrix} = \begin{pmatrix} \boldsymbol{f}_1 \\ \vdots \\ \boldsymbol{f}_n \end{pmatrix} \quad \boldsymbol{B} \quad + \begin{pmatrix} \boldsymbol{E}_1 \\ \vdots \\ \boldsymbol{E}_n \end{pmatrix}.$$

$n \times d$	$n \times r$	$r \times d$	$n \times d$
samples	**latent coordinates**	latent bases	noises

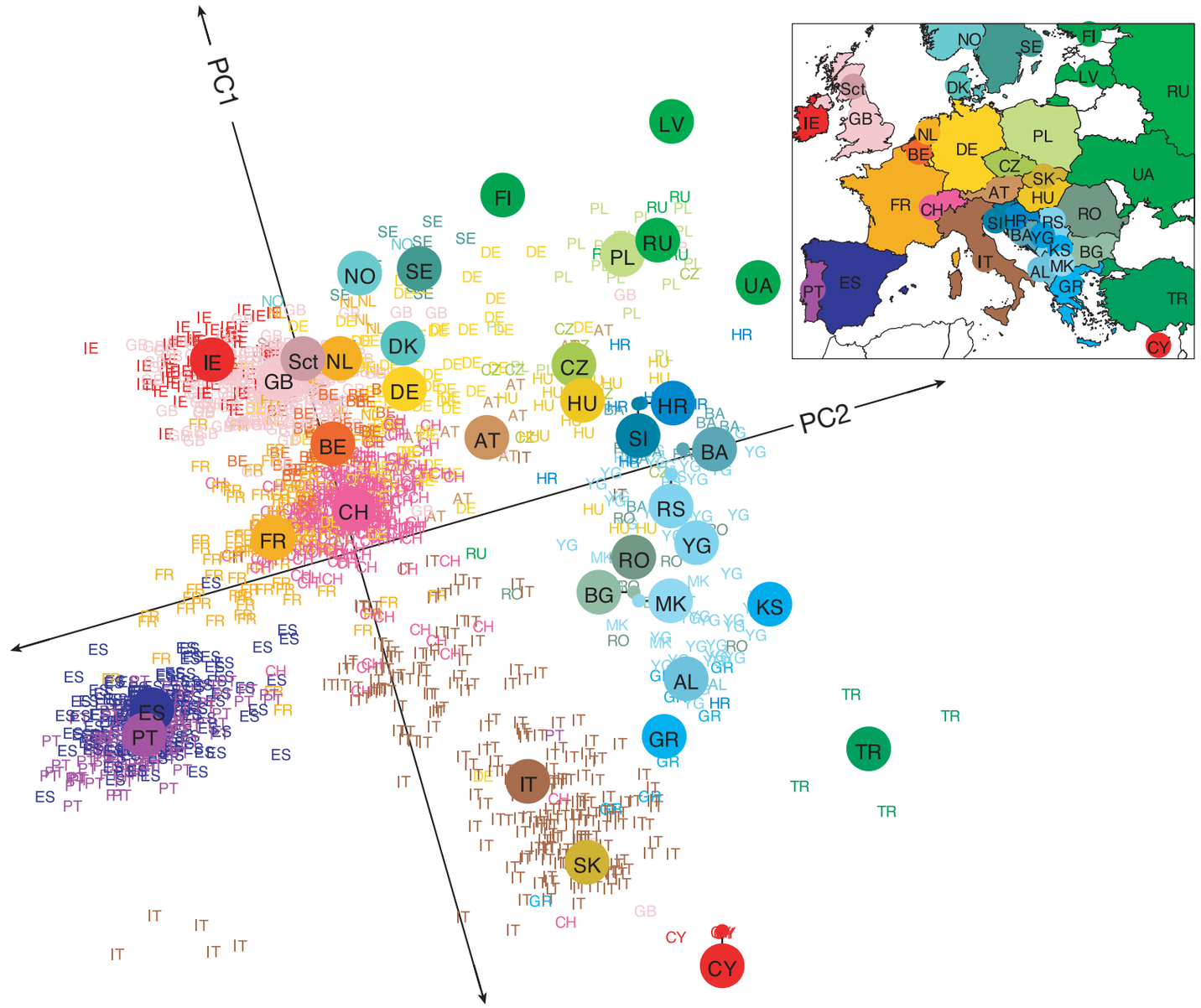
Principal Component Analysis (PCA)

— truncated SVD



Example: Genes Mirror Geography within Europe

Novembre et al. (2008), Nature.
n = 1387 individuals and d = 197146 SNPs;
Figure 1a: 2-dim. embedding vs. labels.



Outline

- Distributed PCA and linearization of eigenvectors
- An ℓ_p theory for spectral methods
- Summary and future directions

Distributed PCA and linearization of eigenvectors

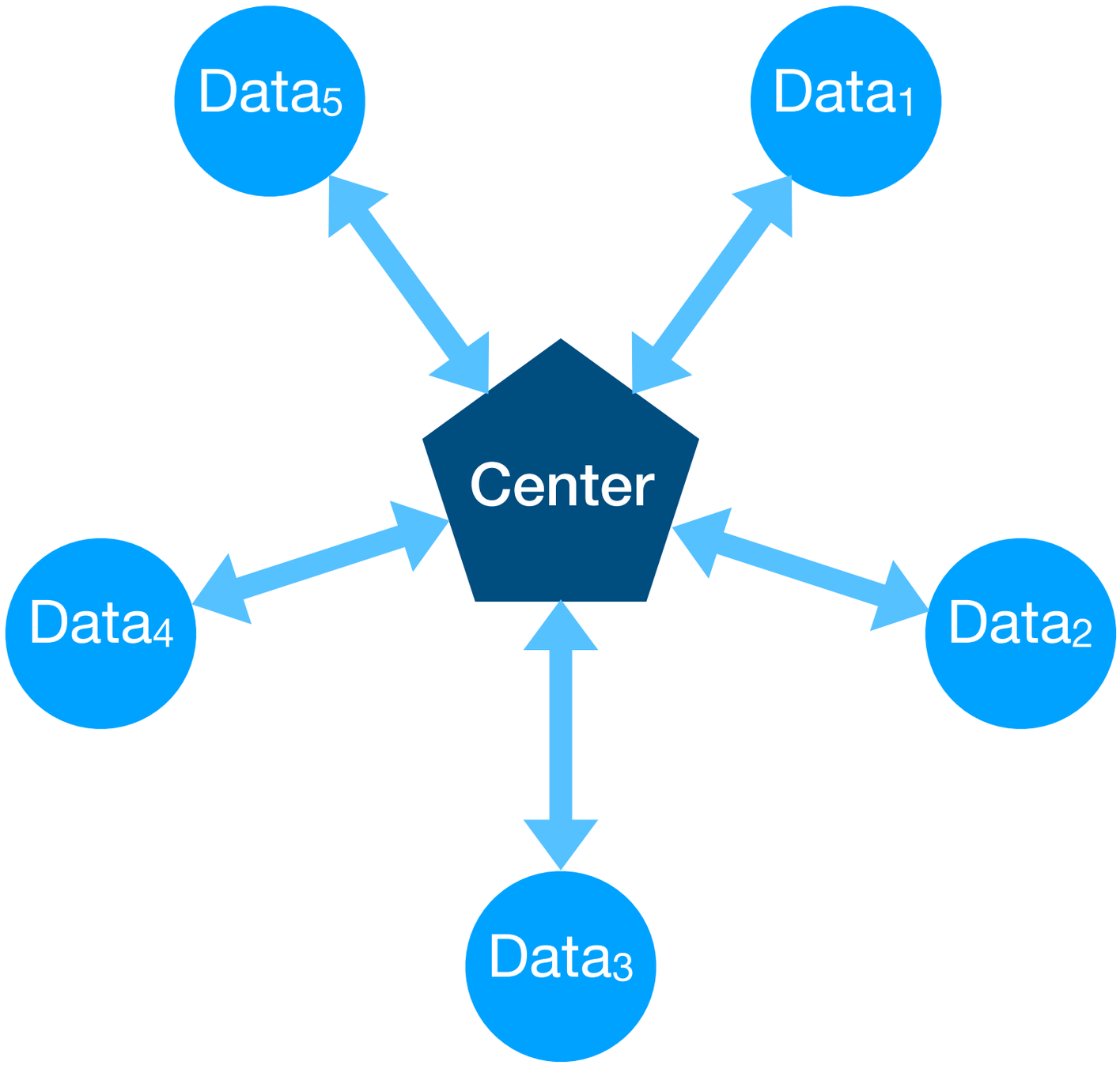
Principal Component Analysis

Data: $\{\boldsymbol{X}_i\}_{i=1}^n \subseteq \mathbb{R}^d$ i.i.d., $\mathbb{E}\boldsymbol{X}_i = \mathbf{0}$, $\mathbb{E}(\boldsymbol{X}_i\boldsymbol{X}_i^\top) = \boldsymbol{\Sigma}$.

Goal: estimate the principal subspace
spanned by the K leading eigenvectors of $\boldsymbol{\Sigma}$.

PCA: $\boldsymbol{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)^\top \xrightarrow{\text{SVD}} \hat{\boldsymbol{U}} = (\hat{\boldsymbol{u}}_1, \cdots, \hat{\boldsymbol{u}}_K) \in \mathbb{R}^{d\times K}$

Distributed PCA



Distributed PCA

m local machines in total, each has n samples.

1. **PCA in parallel**: the ℓ-th machine conducts

$$\boldsymbol{X}^{(\ell)} \in \mathbb{R}^{n\times d} \xrightarrow{\text{SVD}} \hat{\boldsymbol{U}}^{(\ell)} \in \mathbb{R}^{d\times K}$$

and sends $\hat{\boldsymbol{U}}^{(\ell)}$ to the central server;

2. **Aggregation**: $\{\hat{\boldsymbol{U}}^{(\ell)}\}_{\ell=1}^{m} \longrightarrow \hat{\boldsymbol{U}} \in \mathbb{R}^{d\times K}$.

Related works: Mcdonald et al. 2009; Zhang et al. 2013; Lee et al., 2015; Battey et al., 2015; Qu et al. 2002; El Karoui and d'Aspremont 2010; Liang et al. 2014.

Center of Subspaces

How to find $\hat{\boldsymbol{U}} \in \mathcal{O}_{d\times K}$ that best summarizes $\{\hat{\boldsymbol{U}}^{(\ell)}\}_{\ell=1}^{m}$?

Center of Subspaces

How to find $\hat{\boldsymbol{U}} \in \mathcal{O}_{d\times K}$ that best summarizes $\{\hat{\boldsymbol{U}}^{(\ell)}\}_{\ell=1}^{m}$?

- **Subspace distance**:

$$\rho(\boldsymbol{V}, \boldsymbol{W}) = \|\boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{W}\boldsymbol{W}^\top\|_{\mathrm{F}}.$$

- **Least squares**:

$$\hat{\boldsymbol{U}} = \operatorname*{argmin}_{\boldsymbol{V}\in\mathcal{O}_{d\times K}} \sum_{\ell=1}^{m} \rho^2(\boldsymbol{V}, \hat{\boldsymbol{U}}^{(\ell)}).$$

- **Algorithm**:

$$(\hat{\boldsymbol{U}}^{(1)}, \cdots, \hat{\boldsymbol{U}}^{(m)}) \in \mathbb{R}^{d\times mK} \xrightarrow{\text{SVD}} \hat{\boldsymbol{U}} \in \mathcal{O}_{d\times K}.$$

Theoretical Results

Assume that $\boldsymbol{X}_i$ is sub-Gaussian, i.e. $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i\|_{\psi_2} \lesssim 1$.

Define the effective rank and condition number as

$$r = \mathrm{Tr}(\boldsymbol{\Sigma})/\lambda_1, \qquad \kappa = \lambda_1/(\lambda_K - \lambda_{K+1}).$$

Theoretical Results

Assume that $\boldsymbol{X}_i$ is sub-Gaussian, i.e. $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i\|_{\psi_2} \lesssim 1$.

Define the effective rank and condition number as

$$r = \mathrm{Tr}(\boldsymbol{\Sigma})/\lambda_1, \qquad \kappa = \lambda_1/(\lambda_K - \lambda_{K+1}).$$

Theorem (FWWZ, AoS 2019)

There exists constant C such that when $n \geq C\kappa^2\sqrt{Kr}$,

$$\left\| \|\hat{\boldsymbol{U}}\hat{\boldsymbol{U}}^\top - \boldsymbol{U}\boldsymbol{U}^\top\|_{\mathrm{F}} \right\|_{\psi_1} \lesssim \underbrace{\kappa\sqrt{\frac{Kr}{mn}}}_{\text{variance}} + \underbrace{\kappa^2\frac{\sqrt{Kr}}{n}}_{\text{bias}}.$$

Theoretical Results

Assume that $\boldsymbol{X}_i$ is sub-Gaussian, i.e. $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i\|_{\psi_2} \lesssim 1$.

Define the effective rank and condition number as

$$r = \mathrm{Tr}(\boldsymbol{\Sigma})/\lambda_1, \qquad \kappa = \lambda_1/(\lambda_K - \lambda_{K+1}).$$

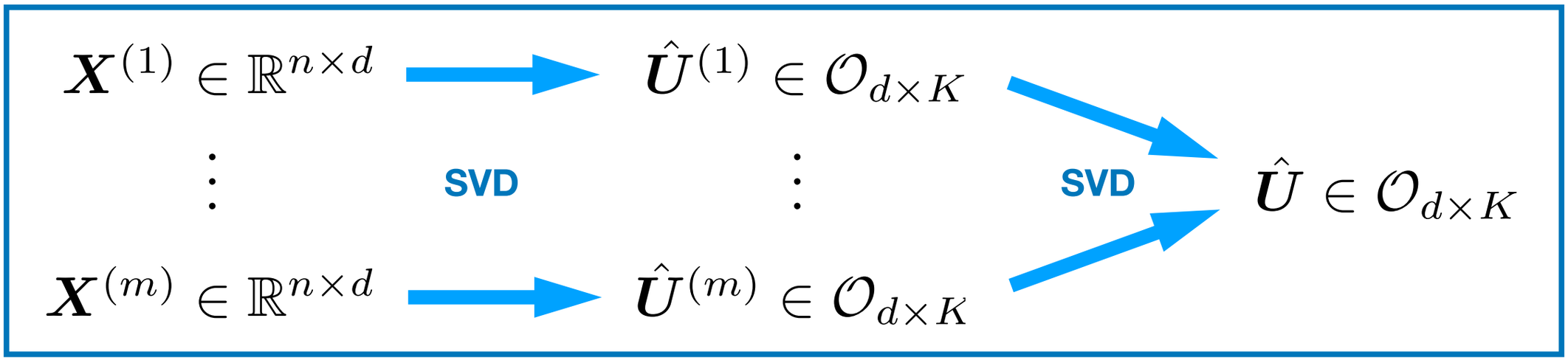
Theorem (FWWZ, AoS 2019)

There exists constant C such that when $n \geq C\kappa^2\sqrt{K}r$,

$$\left\| \|\hat{\boldsymbol{U}}\hat{\boldsymbol{U}}^\top - \boldsymbol{U}\boldsymbol{U}^\top\|_{\mathrm{F}} \right\|_{\psi_1} \lesssim \underbrace{\kappa\sqrt{\frac{Kr}{mn}}}_{\text{variance}} + \underbrace{\kappa^2\frac{\sqrt{K}r}{n}}_{\text{bias}}.$$

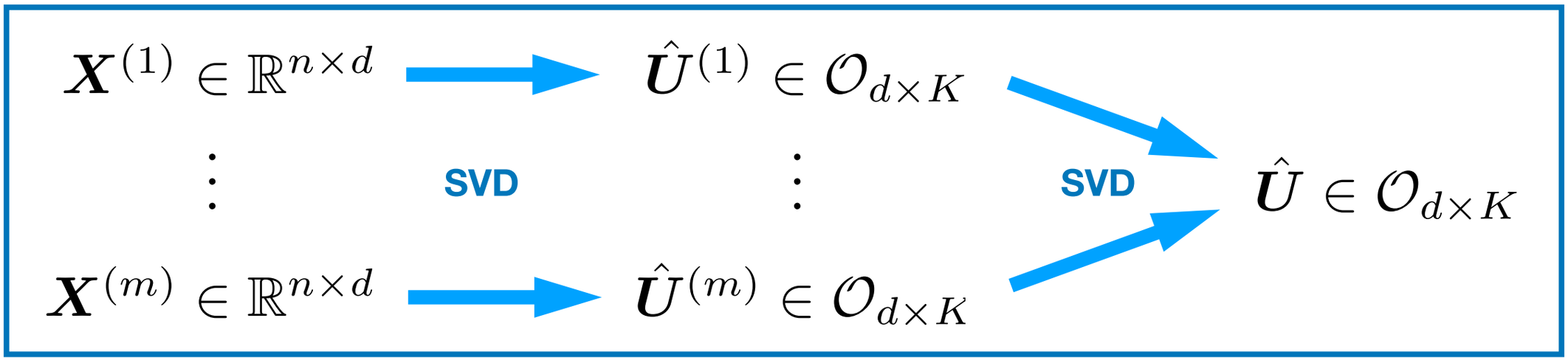
- If $m \lesssim n/(\kappa^2 r)$, distributed PCA is optimal.
- The condition $n \geq C\kappa^2\sqrt{K}r$ cannot be improved.

Analysis of Aggregation



$\hat{\boldsymbol{U}}$: eigenvectors of $\frac{1}{m}\sum_{\ell=1}^{m}\hat{\boldsymbol{U}}^{(\ell)}\hat{\boldsymbol{U}}^{(\ell)\top}$.

Analysis of Aggregation



$\hat{\boldsymbol{U}}$: eigenvectors of $\frac{1}{m}\sum_{\ell=1}^{m}\hat{\boldsymbol{U}}^{(\ell)}\hat{\boldsymbol{U}}^{(\ell)\top}$.

Averaging **reduces variance** but **retains bias**.

- Variance: controlled by Davis-Kahan:

$$\|\hat{\boldsymbol{U}}^{(\ell)}\hat{\boldsymbol{U}}^{(\ell)\top}-\boldsymbol{U}\boldsymbol{U}^{\top}\|_{\mathrm{F}} \lesssim \|(\hat{\boldsymbol{\Sigma}}^{(\ell)}-\boldsymbol{\Sigma})\boldsymbol{U}\|_{\mathrm{F}}/\Delta.$$

- Bias: how large it is?

Linearization of Eigenvectors

Theorem (FWWZ, AoS 2019)

$$\|\hat{\boldsymbol{U}}^{(\ell)}\hat{\boldsymbol{U}}^{(\ell)\top} - [\boldsymbol{U}\boldsymbol{U}^\top + f(\hat{\boldsymbol{\Sigma}}^{(\ell)} - \boldsymbol{\Sigma})]\|_{\mathrm{F}} \lesssim [\|(\hat{\boldsymbol{\Sigma}}^{(\ell)} - \boldsymbol{\Sigma})\boldsymbol{U}\|_{\mathrm{F}}/\Delta]^2,$$

$f: \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ is a **linear** functional determined by $\boldsymbol{\Sigma}$.

Linearization of Eigenvectors

Theorem (FWWZ, AoS 2019)

$$\|\hat{\boldsymbol{U}}^{(\ell)}\hat{\boldsymbol{U}}^{(\ell)\top} - [\boldsymbol{U}\boldsymbol{U}^\top + f(\hat{\boldsymbol{\Sigma}}^{(\ell)} - \boldsymbol{\Sigma})]\|_{\mathrm{F}} \lesssim [\|(\hat{\boldsymbol{\Sigma}}^{(\ell)} - \boldsymbol{\Sigma})\boldsymbol{U}\|_{\mathrm{F}}/\Delta]^2,$$

$f: \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ is a **linear** functional determined by $\boldsymbol{\Sigma}$.

More precise than **Davis-Kahan**:

$$\|\hat{\boldsymbol{U}}^{(\ell)}\hat{\boldsymbol{U}}^{(\ell)\top} - \boldsymbol{U}\boldsymbol{U}^\top\|_{\mathrm{F}} \lesssim \|(\hat{\boldsymbol{\Sigma}}^{(\ell)} - \boldsymbol{\Sigma})\boldsymbol{U}\|_{\mathrm{F}}/\Delta.$$

PCA has small bias:

$$\|\mathbb{E}(\hat{\boldsymbol{U}}^{(\ell)}\hat{\boldsymbol{U}}^{(\ell)\top}) - \boldsymbol{U}\boldsymbol{U}^\top\|_{\mathrm{F}} \lesssim [\|(\hat{\boldsymbol{\Sigma}}^{(\ell)} - \boldsymbol{\Sigma})\boldsymbol{U}\|_{\mathrm{F}}/\Delta]^2.$$

Summary

Theoretical guarantees for **distributed PCA**:

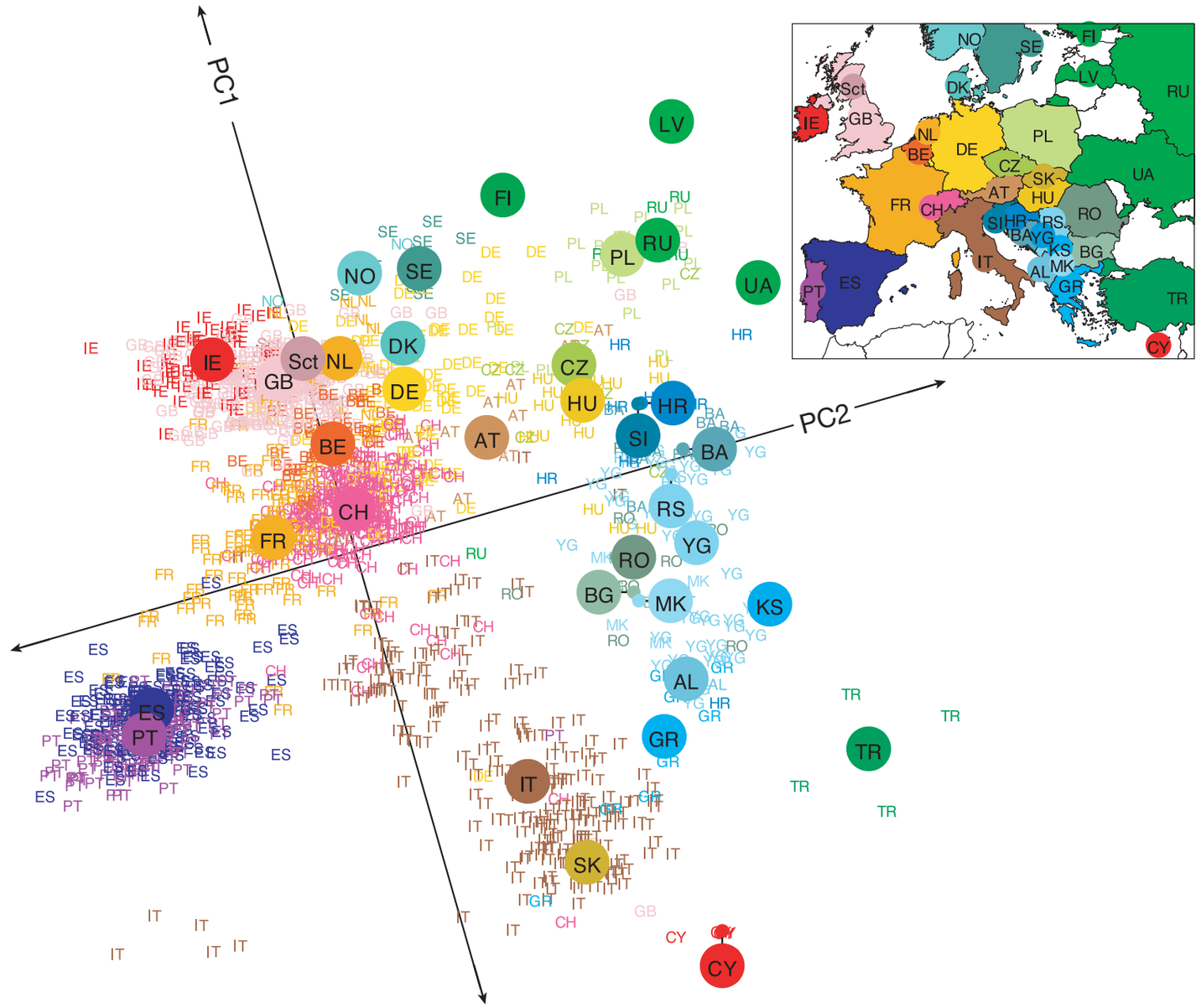
- **Bias** and variance of PCA;
- **Linearization** of eigenvectors, high-order Davis-Kahan.

Paper (alphabetical order):

- Fan, Wang, **Wang** and Zhu. Distributed estimation of principal eigenspaces. **The Annals of Statistics**, 2019.

Example: Genes Mirror Geography within Europe

Novembre et al. (2008), Nature.
n = 1387 individuals and d = 197146 SNPs;
Figure 1a: 2-dim. embedding vs. labels.



A Pipeline for Spectral Methods

1. Similarity matrix construction

e.g. Gram $\boldsymbol{X}\boldsymbol{X}^\top$, adjacency $\boldsymbol{A}$;

2. Spectral decomposition

get r eigen-pairs $\left\{\lambda_j, \boldsymbol{u}_j\right\}_{j=1}^{r}$;

3. r-dim. embedding

e.g. using the rows of $\left(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_r\right)$;

4. Downstream tasks

e.g. visualization.

Ext.: { robust, probabilistic, sparse, nonnegative } PCA.

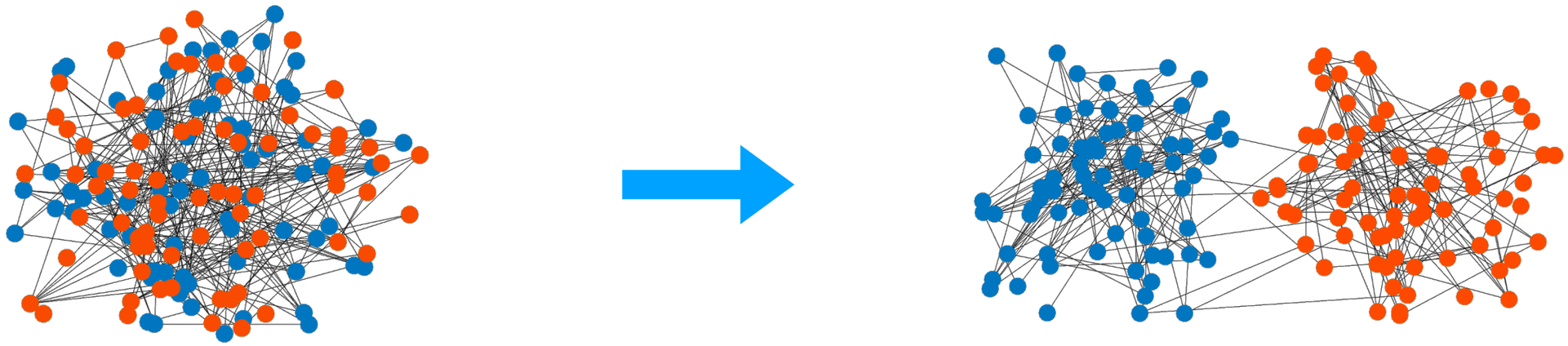
Pearson (1901), Hotelling (1933), Schölkopf (1997), Tipping and Bishop (1999), Shi and Malik (2000), Ng et al. (2002), Belkin and Niyogi (2003), Von Luxburg (2007)

An ℓ_p theory for spectral methods

- **Network analysis and Wigner-type matrices**
- Mixture model and Wishart-type matrices

Community Detection and SBM

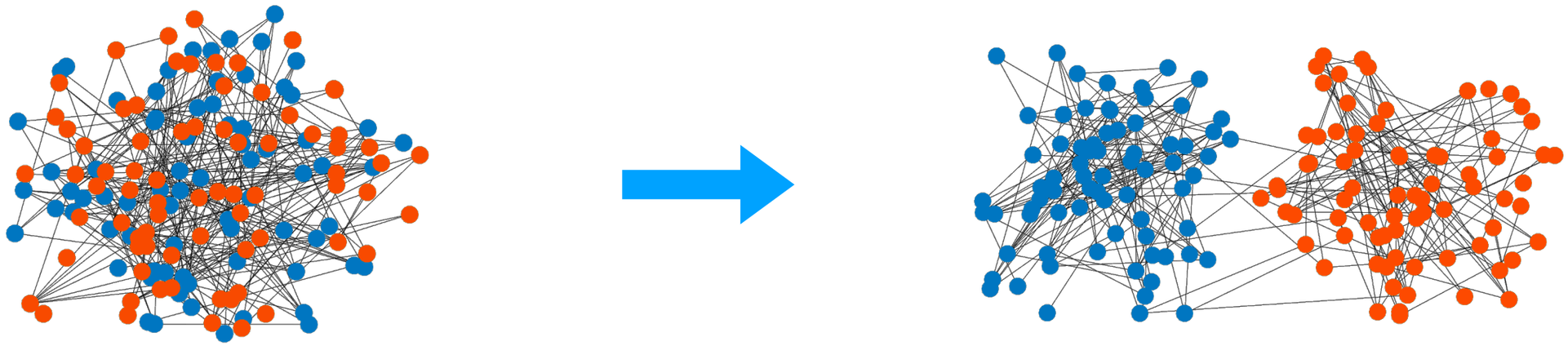
Community detection in networks:



Credit: Yuxin Chen.

Community Detection and SBM

Community detection in networks:



Stochastic Block Model (Holland et al., 1983)

Symmetric adjacency matrix $\boldsymbol{A} \in \{0,1\}^{n\times n}$, $|J| = |J^c| = \frac{n}{2}$:

$$\mathbb{P}(A_{ij} = 1) = \begin{cases} p, & \text{if } i,j \text{ or } i,j \\ q, & \text{if } i,j \text{ or } i,j \end{cases}.$$

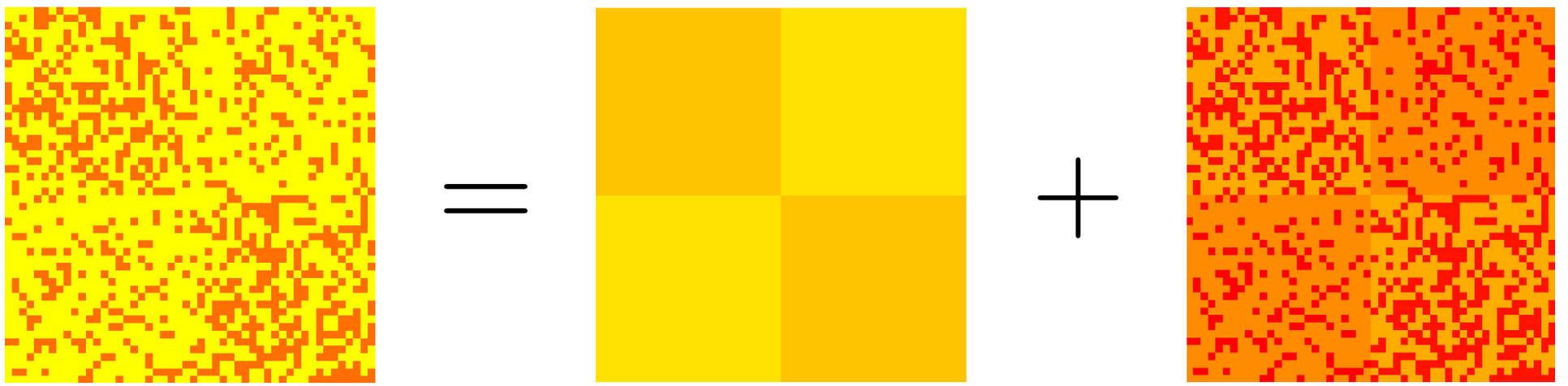
McSherry (2001), Coja-Oghlan (2006), Rohe et al. (2011), Mossel et al. (2013), Massoulie (2014), Lelarge et al. (2015), Chin et al. (2015), Abbe et al. (2016), Zhang and Zhou (2016).

Credit: Yuxin Chen.

Community Detection and SBM

$$\mathbb{E}\boldsymbol{A} = \begin{pmatrix} p\mathbf{1}_{J,J} & q\mathbf{1}_{J,J^c} \\ q\mathbf{1}_{J^c,J} & p\mathbf{1}_{J^c,J^c} \end{pmatrix} = \frac{p+q}{2n}\mathbf{1}\mathbf{1}^\top + \frac{p-q}{2n}\begin{pmatrix} \mathbf{1}_J \\ -\mathbf{1}_{J^c} \end{pmatrix}(\mathbf{1}_J^\top \quad -\mathbf{1}_{J^c}^\top).$$

The 2nd eigenvector $\bar{\boldsymbol{u}} = \frac{1}{\sqrt{n}}(\mathbf{1}_J - \mathbf{1}_{J^c})$ reveals (J, J^c).



Credit: Yuxin Chen.

Community Detection and SBM

$$\mathbb{E}\boldsymbol{A} = \begin{pmatrix} p\mathbf{1}_{J,J} & q\mathbf{1}_{J,J^c} \\ q\mathbf{1}_{J^c,J} & p\mathbf{1}_{J^c,J^c} \end{pmatrix} = \frac{p+q}{2n}\mathbf{1}\mathbf{1}^\top + \frac{p-q}{2n}\begin{pmatrix} \mathbf{1}_J \\ -\mathbf{1}_{J^c} \end{pmatrix}(\mathbf{1}_J^\top \quad -\mathbf{1}_{J^c}^\top).$$

The 2nd eigenvector $\bar{\boldsymbol{u}} = \frac{1}{\sqrt{n}}(\mathbf{1}_J - \mathbf{1}_{J^c})$ reveals (J, J^c).

Spectral method:

$$\boldsymbol{A} \xrightarrow{\text{SVD}} \text{the 2nd eigenvector } \boldsymbol{u} \longrightarrow \operatorname{sgn}(\boldsymbol{u})$$

To recover (J, J^c), we need $\boldsymbol{u} \approx \bar{\boldsymbol{u}}$ in a uniform way.

Classical ℓ_2 bounds (Davis and Kahan, 1970) are too loose!

Optimality of Spectral Method

Theorem (AFWZ, AoS 2020+)

Let $a \neq b$ and $\mathbb{P}(A_{ij} = 1) = \begin{cases} \frac{a \log n}{n}, & \text{if } i, j \text{ or } i, j \\ \frac{b \log n}{n}, & \text{if } i, j \text{ or } i, j \end{cases}$.

- Exact recovery w.h.p. when $\left(\sqrt{a} - \sqrt{b}\right)^2 > 2$;
- Error rate $n^{-(\sqrt{a}-\sqrt{b})^2/2}$ when $\left(\sqrt{a} - \sqrt{b}\right)^2 \leq 2$.
- Optimality (Abbe et al., 2016; Zhang and Zhou, 2016).

Optimality of Spectral Method

Theorem (AFWZ, AoS 2020+)

Let $a \neq b$ and $\mathbb{P}(A_{ij} = 1) = \begin{cases} \frac{a \log n}{n}, & \text{if } i, j \text{ or } i, j \\ \frac{b \log n}{n}, & \text{if } i, j \text{ or } i, j \end{cases}$.

- Exact recovery w.h.p. when $\left(\sqrt{a} - \sqrt{b}\right)^2 > 2$;
- Error rate $n^{-(\sqrt{a}-\sqrt{b})^2/2}$ when $\left(\sqrt{a} - \sqrt{b}\right)^2 \leq 2$.
- Optimality (Abbe et al., 2016; Zhang and Zhou, 2016).

Key ingredients:

- **Entrywise** linear approximation $\boldsymbol{u} = \boldsymbol{A}\boldsymbol{u}/\lambda \approx \boldsymbol{A}\bar{\boldsymbol{u}}/\bar{\lambda}$;
- Weighted sum of independent Bernoulli variables.

ℓ_∞ Analysis: Linearization

Theorem (Linear approximation)

$$\mathbb{P}(\sqrt{n}\|\boldsymbol{u} - \boldsymbol{A}\bar{\boldsymbol{u}}/\bar{\lambda}\|_\infty < \varepsilon_n) > 1 - n^{-3}.$$

ℓ_∞ Analysis: Linearization

Theorem (Linear approximation)

$$\mathbb{P}(\sqrt{n}\|\boldsymbol{u} - \boldsymbol{A}\bar{\boldsymbol{u}}/\bar{\lambda}\|_\infty < \varepsilon_n) > 1 - n^{-3}.$$

General results (AF**W**Z, AoS 2020+)

- Singular vectors of **Wigner-type** matrices:
  - Symmetric, independent entries above the diagonal;
  - Rectangular, independent entries;
- $\min_{\boldsymbol{O}\in\mathcal{O}_{r\times r}} \|\boldsymbol{U}\boldsymbol{O} - \boldsymbol{A}\bar{\boldsymbol{U}}\bar{\boldsymbol{\Lambda}}^{-1}\|_{2,\infty} \ll \|\bar{\boldsymbol{U}}\|_{2,\infty}$;
- Applications: synchronization, matrix completion, (inference).

A General ℓ_∞ Theory

Merits and demerits of ℓ_∞ analysis

Characterizes individual objects precisely

- Results and tools apply to ncvx opt. (M**W**CC, FoCM 2019);

Requires strong signals for **uniform** control:

- e.g. degree $\gtrsim \log n$ for SBM.

Successor: ℓ_p analysis with $p < \infty$

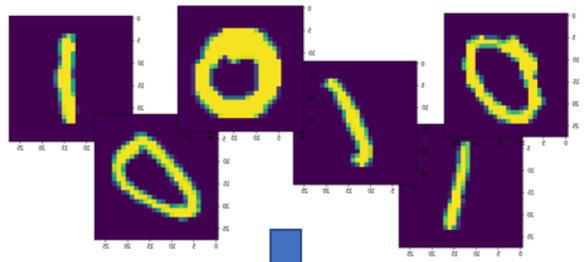
- Controls a **vast majority** of the entries.

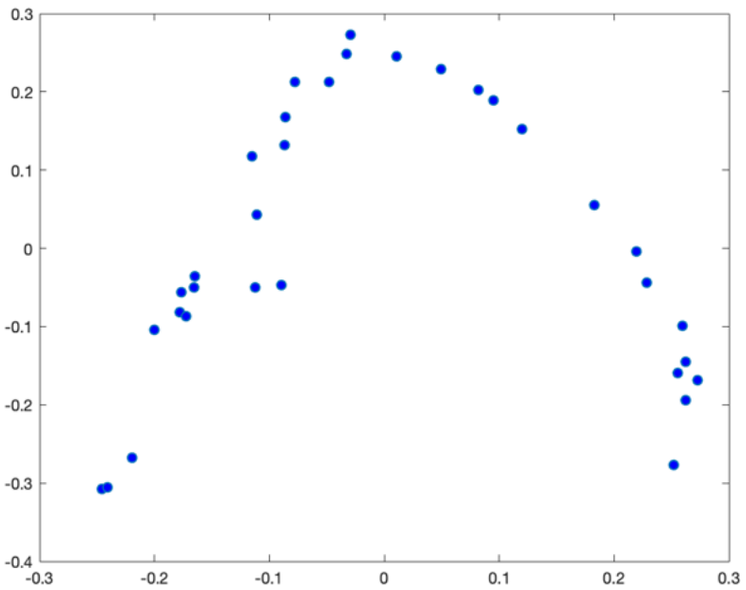
An ℓ_p theory for spectral methods

- Network analysis and Wigner-type matrices
- **Mixture model and Wishart-type matrices**

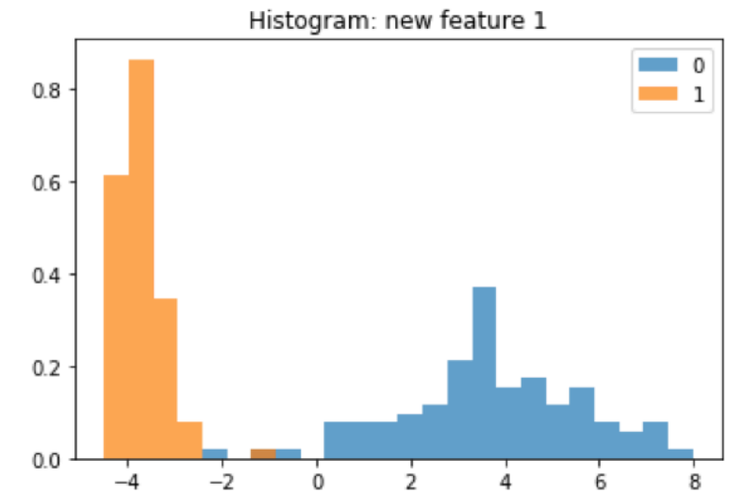
Dimensionality Reduction

Data

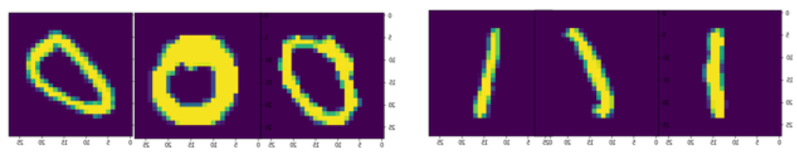




Embedding



Structures



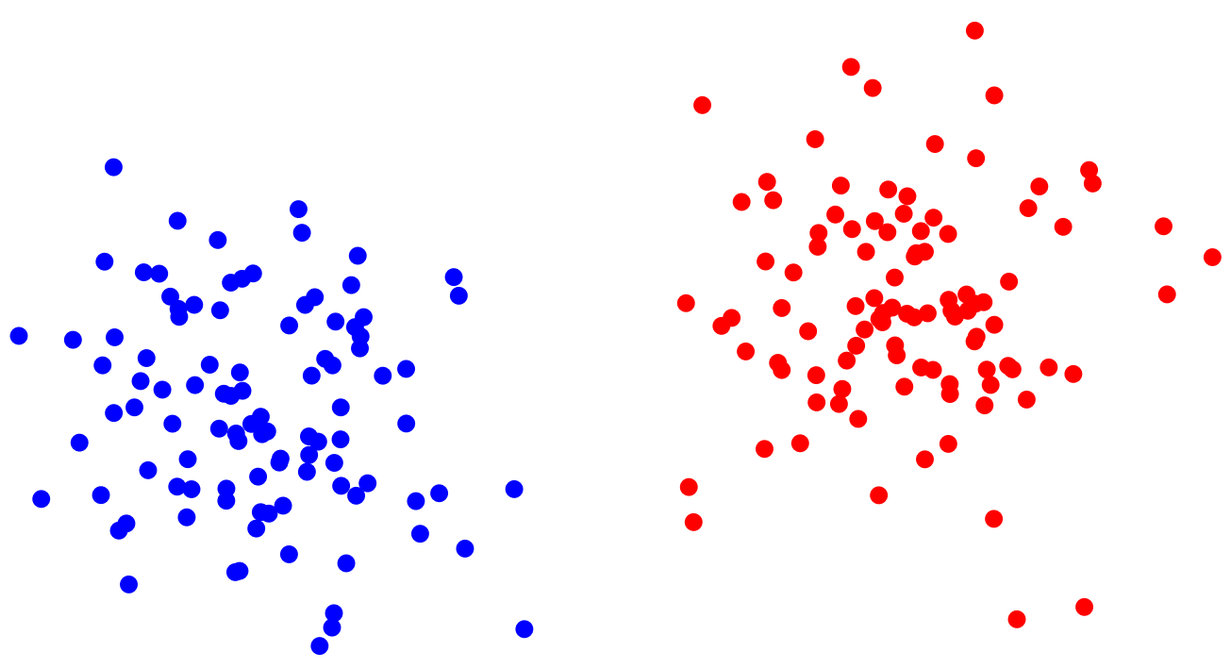
Gaussian Mixture Model

Heteroscedastic model

$$\ell_i \in \{\pm 1\}, \quad \boldsymbol{X}_i | \ell_i \sim \mathcal{N}(\ell_i \boldsymbol{\mu}, \boldsymbol{\Sigma}_i), \quad \boldsymbol{\Sigma}_i \preceq \boldsymbol{\Sigma}.$$

$\boldsymbol{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)^\top \in \mathbb{R}^{n \times d}$

Low-rank: $\mathbb{E}(\boldsymbol{X} \,|\, \boldsymbol{\ell}) = \boldsymbol{\ell}\boldsymbol{\mu}^\top$.



Spectral Method

Recall $\mathbb{E}(\boldsymbol{X} \mid \boldsymbol{\ell}) = \boldsymbol{\ell}\boldsymbol{\mu}^\top$.

Spectral method

1. Get the hollowed Gram matrix

$$\boldsymbol{G} = \mathcal{H}(\boldsymbol{X}\boldsymbol{X}^\top) \in \mathbb{R}^{n \times n};$$

2. $\boldsymbol{G}$ $\xrightarrow{\text{SVD}}$ the 1st eigenvector $\boldsymbol{u}$ $\longrightarrow$ $\operatorname{sgn}(\boldsymbol{u})$.

Hollowing

- improves concentration;
- helps tackle heteroscedasticity.

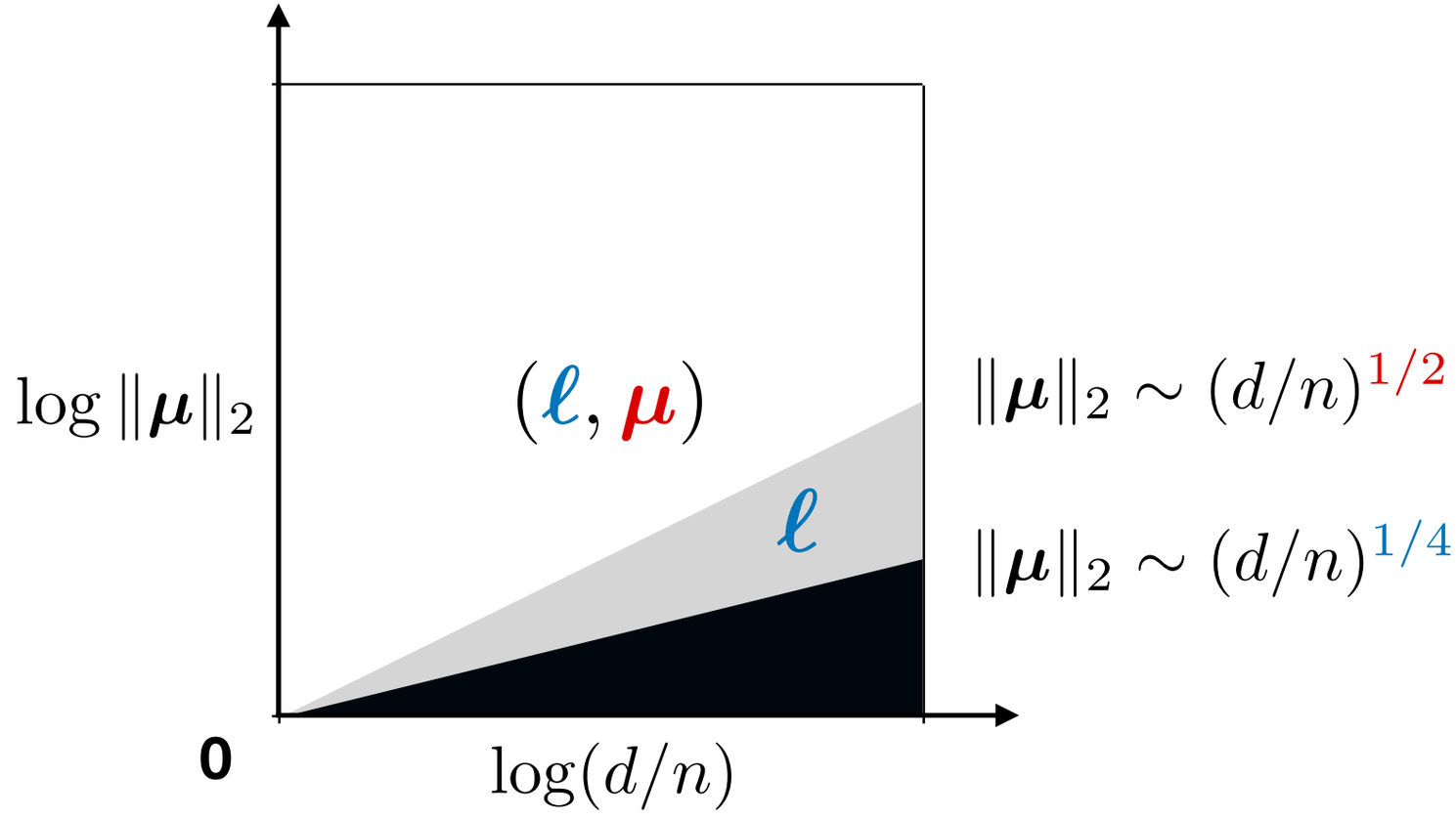
Related works: Montanari and Sun 2018, Ndaoud 2018, Cai et al. 2019.

Theoretical Challenges

- **Dependency**: $\mathcal{H}(\boldsymbol{X}\boldsymbol{X}^\top)$ has **Wishart**-type distribution.
  - SBM: **Wigner**-type adjacency matrix.
- **High dimensionality**: most existing results require $d \lesssim n$.
  - **Clustering** vs. **parameter est.**: $\frac{1}{2}N(\boldsymbol{\mu}, \boldsymbol{I}_d) + \frac{1}{2}N(-\boldsymbol{\mu}, \boldsymbol{I}_d)$

Theoretical Challenges

- **Dependency**: $\mathcal{H}(\boldsymbol{X}\boldsymbol{X}^\top)$ has **Wishart**-type distribution.
  - SBM: **Wigner**-type adjacency matrix.
- **High dimensionality**: most existing results require $d \lesssim n$.
  - **Clustering** vs. **parameter est.**: $\frac{1}{2}N(\boldsymbol{\mu}, \boldsymbol{I}_d) + \frac{1}{2}N(-\boldsymbol{\mu}, \boldsymbol{I}_d)$



ℓ_p Analysis: Linear Approximation

$\boldsymbol{u}$ is the 1st eigenvector of $\boldsymbol{G} = \mathcal{H}(\boldsymbol{X}\boldsymbol{X}^\top) \in \mathbb{R}^{n\times n}$,

$$\mathrm{SNR} = \frac{\|\boldsymbol{\mu}\|_2^4}{\|\boldsymbol{\Sigma}\|_2\|\boldsymbol{\mu}\|_2^2 + \|\boldsymbol{\Sigma}\|_{\mathrm{F}}^2/n} \gg 1,$$

$\bar{\boldsymbol{X}} = \boldsymbol{\ell}\boldsymbol{\mu}^\top$, $(\bar{\lambda}, \bar{\boldsymbol{u}})$ is the 1st eigen-pair of $\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^\top$.

ℓ_p Analysis: Linear Approximation

$\boldsymbol{u}$ is the 1st eigenvector of $\boldsymbol{G} = \mathcal{H}(\boldsymbol{X}\boldsymbol{X}^\top) \in \mathbb{R}^{n\times n}$,

$$\mathrm{SNR} = \frac{\|\boldsymbol{\mu}\|_2^4}{\|\boldsymbol{\Sigma}\|_2\|\boldsymbol{\mu}\|_2^2 + \|\boldsymbol{\Sigma}\|_{\mathrm{F}}^2/n} \gg 1,$$

$\bar{\boldsymbol{X}} = \boldsymbol{\ell}\boldsymbol{\mu}^\top$, $(\bar{\lambda}, \bar{\boldsymbol{u}})$ is the 1st eigen-pair of $\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^\top$.

Theorem

If $2 \le p \lesssim \mathrm{SNR}$, there exists $\varepsilon_n \to 0$ s.t.

$$\mathbb{P}\left(\|\boldsymbol{u} - \boldsymbol{G}\bar{\boldsymbol{u}}/\bar{\lambda}\|_p < \varepsilon_n\|\bar{\boldsymbol{u}}\|_p\right) > 1 - e^{-p}.$$

- Applies to **RKHS (dim. = ∞)**: kernel PCA;
- Adaptivity: large $\mathrm{SNR} \Rightarrow$ large $p \Rightarrow$ strong result.

ℓ_p Analysis: Corollaries

When $p \asymp \mathrm{SNR} \gtrsim \log n$: $\|\cdot\|_p \asymp \|\cdot\|_\infty$ and

$$\mathbb{P}\left(\|\boldsymbol{u} - \boldsymbol{G}\bar{\boldsymbol{u}}/\bar{\lambda}\|_\infty < \frac{\varepsilon_n}{\sqrt{n}} \right) > 1 - \frac{1}{n^3}.$$

ℓ_p Analysis: Corollaries

When $p \asymp \mathrm{SNR} \gtrsim \log n$: $\|\cdot\|_p \asymp \|\cdot\|_\infty$ and

$$\mathbb{P}\left(\|\boldsymbol{u} - \boldsymbol{G}\bar{\boldsymbol{u}}/\bar{\lambda}\|_\infty < \frac{\varepsilon_n}{\sqrt{n}} \right) > 1 - \frac{1}{n^3}.$$

Corollary (optimal clustering, $\boldsymbol{\Sigma} = \boldsymbol{I}_d$)

- Exact recovery w.h.p. when $\mathrm{SNR} > 2\log n$;
- Error rate $e^{-\mathrm{SNR}/[2+o(1)]}$ when $1 \ll \mathrm{SNR} \leq 2\log n$.

- Kmeans (Lu and Zhou 2016);
- Spectral (Vempala and Wang 2004, Jin et al. 2017, Ndaoud 2018, Löffler et al. 2019);
- SDP (Mixon et al. 2016, Royer 2017, Fei and Chen 2018, Giraud and Verzelen 2018, Chen and Yang 2018).

Summary

Sharpness of **spectral methods**.

- Abbe, Fan, **Wang** and Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. **The Annals of Statistics**, 2020+.
- Abbe, Fan and **Wang**. An ℓ_p analysis of kernel PCA and contextual network analysis. Manuscript.

Extensions

- ranking (CFM**W**, AoS 2019), topic models, ncvx optimization;
- statistical inference based on linear representation.

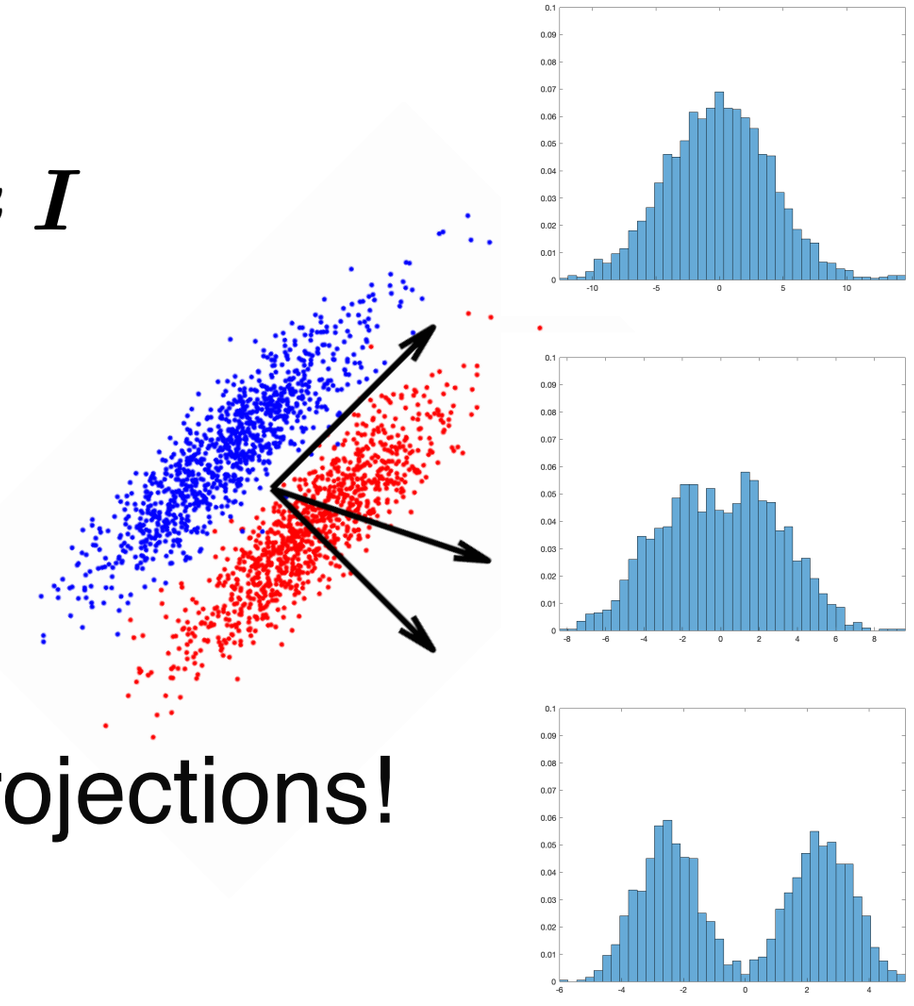
(alphabetical orders)

Finding a Needle in a Haystack

Spectral methods are **powerful** but **not omnipotent**.

$\frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \boldsymbol{\Sigma})$: covariance $\boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{\Sigma}$

- Max variance $\neq$ useful
- PCA: $\|\boldsymbol{\mu}\|_2^2 / \|\boldsymbol{\Sigma}\|_2 \gg 1$ or $\boldsymbol{\Sigma} \approx \boldsymbol{I}$

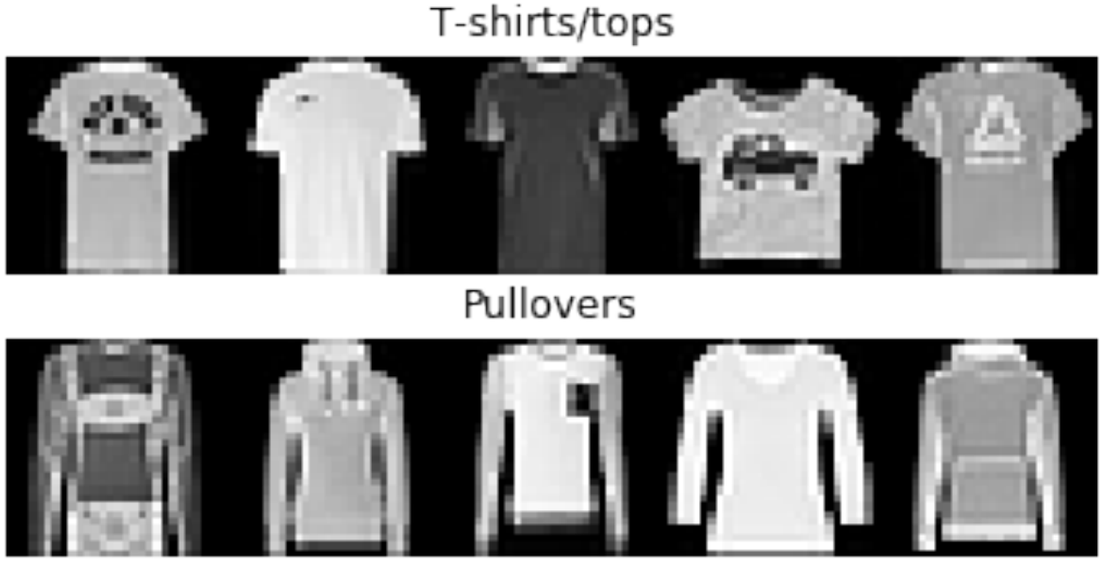


Seek for clustering-friendly projections!

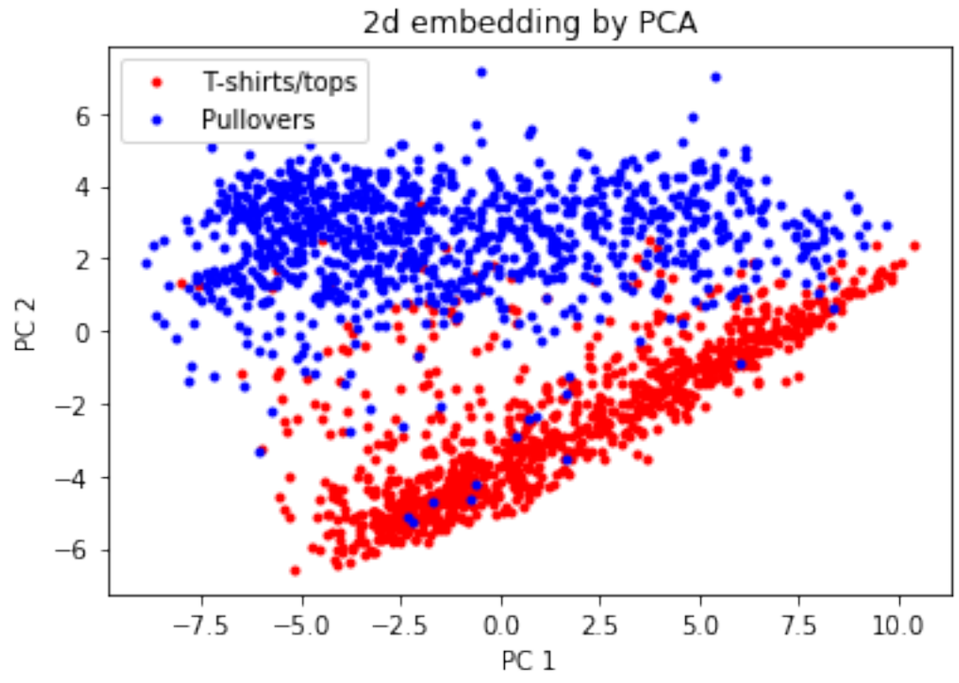
Example: Fashion-MNIST

70000 fashion products, 10 categories (Xiao et al. 2017).

- T-shirts/tops
- Pullovers



Visualization by PCA



A CURE for Clustering Problems

Clustering via Uncoupled REgression (CURE):

Wang, Yan and Diaz. Efficient clustering for stretched mixtures: landscape and optimality. Submitted.

- **Clustering** -> **classification**;
- **Stat.** and **comp.** guarantees under mixture models.

Q & A

Thank you!