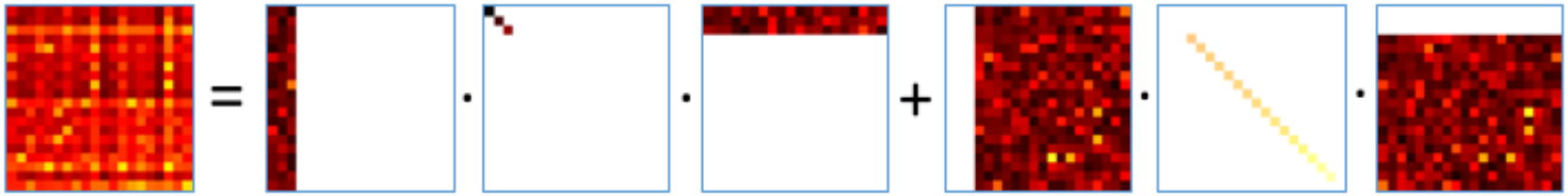
Rate-Optimal Perturbation Bounds for Singular Subspaces with Applications to High-Dimensional Statistics

Anru Zhang
Department of Statistics
University of Wisconsin – Madison

Introduction

- Focus: singular value decomposition (SVD)

$$X = U \cdot \Sigma_1 \cdot V^\top + U_\perp \cdot \Sigma_2 \cdot V_\perp^\top$$



- Due to perturbation,

$$\hat{X} = X + Z,$$

SVD is altered to

$$\hat{X} = \hat{U} \cdot \hat{\Sigma}_1 \cdot \hat{V}^\top + \hat{U}_\perp \cdot \hat{\Sigma}_2 \cdot \hat{V}_\perp^\top.$$

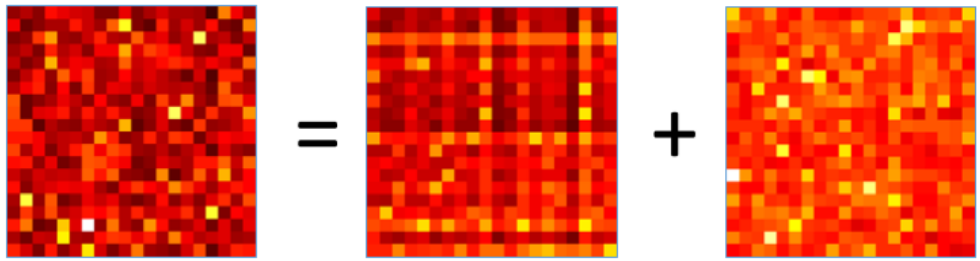
Introduction

small perturbation + large signal $\rightarrow$ close $\hat{V}$ to V (or $\hat{U}$ and U)

- **Problem: Perturbation Bounds on Singular Subspaces**
  - How to quantify the difference between $\hat{V}$ and V (or $\hat{U}$ and U)?
  - Is there any upper bounds for the difference?
  - Are U and $\hat{U}$, V and $\hat{V}$ equally different?
- **Motivation**: **spectral method**, which has been used in a wide range of modern high-dimensional statistical problems, utilize this property.

Application 1: Low-rank Matrix Denoising

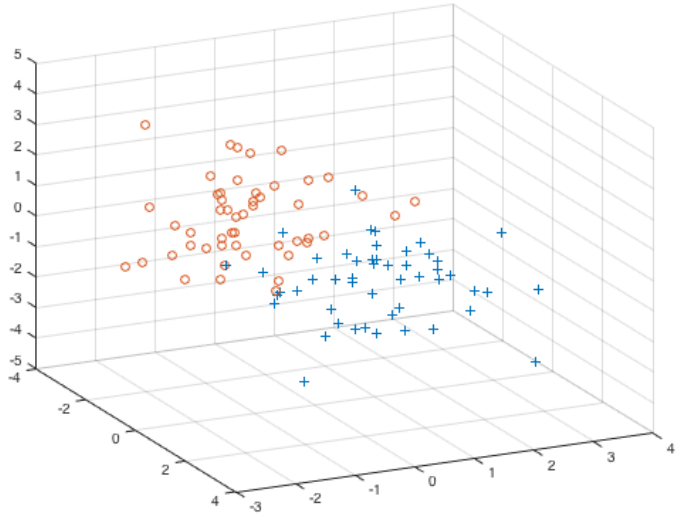
$$\hat{X} = X + Z,$$



X is approximately rank-r, $\quad Z \overset{iid}{\sim}$ sub-Gaussian$(0, \sigma^2)$

- Target: X, U or V.
- Specific applications
  - Magnetic Resonance Imaging (MRI) (Candès, Sing-Long and Trzasko, 2012);
  - Relaxometry (Bydder and Du, 2006)
- Natural estimators for U, V: $\hat{U}$, $\hat{V}$, the first r singular vectors of $\hat{X}$.
- **Q: How do $\hat{U}$, $\hat{V}$ perform, respectively?**

Application 2: High-dimensional Clustering



- Observe n points $X_1, \ldots, X_n \in \mathbb{R}^p$, $p \geq n$.
- Each point belongs to one of two classes (Jin, Ke and Wang, 2015)

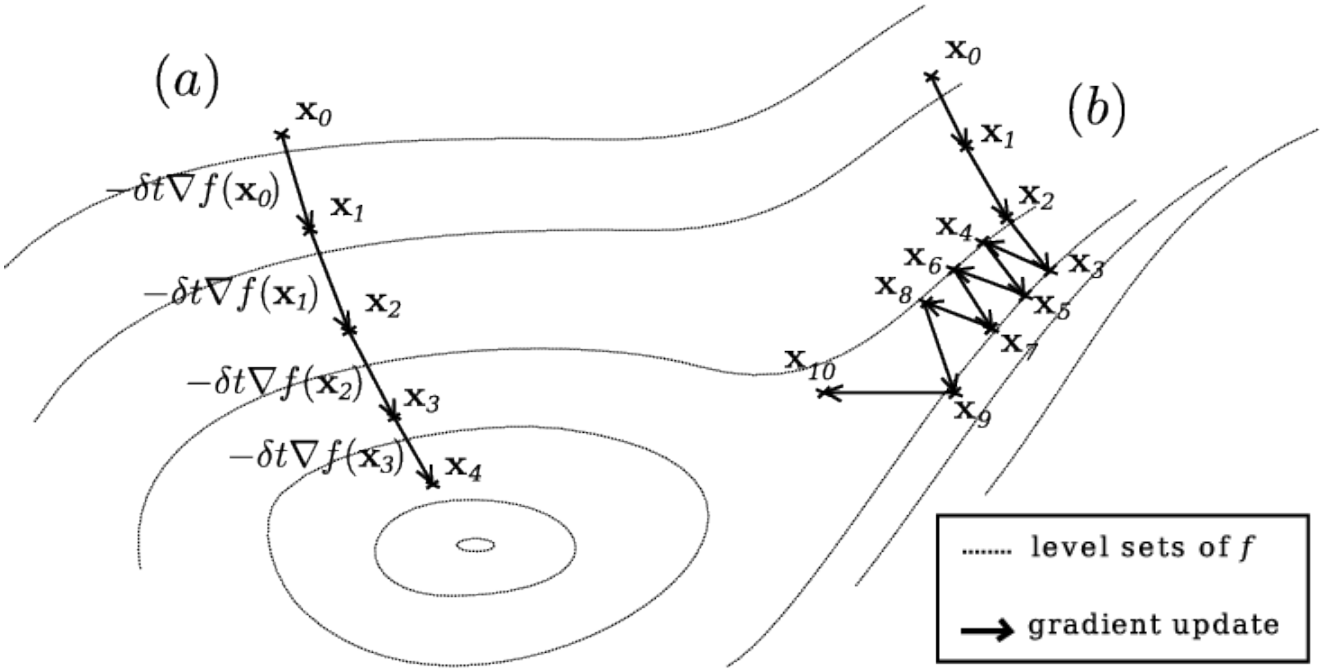
$$X_i = \mu l_i + \varepsilon_i \in \mathbb{R}^p, \quad i = 1, \ldots, n, \quad \varepsilon_i \overset{iid}{\sim} \text{sub-Gaussian}(0, \sigma^2 I_p),$$

$l_i \in \{-1, 1\}$ are labels; $\mu \in \mathbb{R}^p$ is the mean.

- **Goal: recover labels l.**

Other Applications

- In addition, spectral method is often applied to find a **"warm start"** for more delicate iterative algorithms.
  - phase retrieval (Cai, Li and Ma, 2016)
  - matrix completion (Sun and Luo, 2015)
  - community detection (Jin, 2015)



Other Applications

Other applications of **spectral methods** include

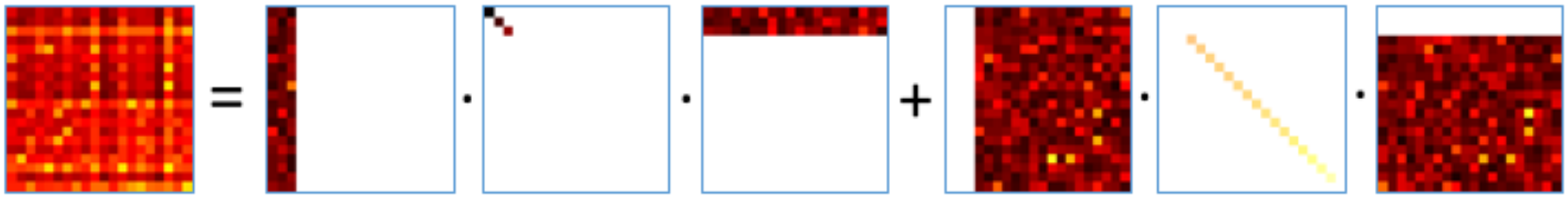
- community detection
- matrix completion
- principle component analysis
- canonical correlation analysis
- ...

Specific practices include

- collaborative filtering (the Netflix problem)
- multi-task learning
- system identification
- sensor localization
- ...

Problem Formulation

$$X = U \cdot \Sigma_1 \cdot V^\top + U_\perp \cdot \Sigma_2 \cdot V_\perp^\top$$



$$\hat{X} = X + Z, \quad \hat{X} = \hat{U} \cdot \hat{\Sigma}_1 \cdot \hat{V}^\top + \hat{U}_\perp \cdot \hat{\Sigma}_2 \cdot \hat{V}_\perp^\top$$

- **Target:**

  Measure the difference between $\hat{V}$ and V ($\hat{U}$ and U)

$\sin\Theta$ Distance of Singular Sub-spaces

Definition of $\sin\Theta$ distances:

- Suppose $V^\top \hat{V}$ have singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$.
- Define the sine principle angles as

$$\sin\Theta(V, \hat{V}) = \text{diag}(\sqrt{1-\sigma_1^2}, \ldots, \sqrt{1-\sigma_r^2}).$$

- Quantitative measure of distance: $\|\sin\Theta(\hat{V}, V)\|$ and $\|\sin\Theta(\hat{V}, V)\|_F$.

Good properties:

- Triangular inequality $\rightarrow$ indeed a distance;
- Many other distances are equivalent $\rightarrow$ convenient to use.

Classic Results of Perturbation Bounds

- The Perturbation bounds: develop the upper bound for

$$\|\sin\Theta(V,\hat{V})\|, \quad \|\sin\Theta(U,\hat{U})\|, \quad \|\sin\Theta(V,\hat{V})\|_F, \quad \|\sin\Theta(U,\hat{U})\|_F.$$

- This problem has been widely studied in the literature (Davis and Kahan, 1970; Wedin, 1972; Weyl, 1912; Stewart, 1991, 2006; Yu et al., 2015; Fan, Wang and Zhong, 2016).
- Classical tools:
  - Davis and Kahan (1970): eigenvectors of symmetric matrices;
  - Wedin (1972): singular vectors for asymmetric matrices.

Classic Result: Wedin's $\sin\Theta$ Theorem

$$X = U \cdot \Sigma_1 \cdot V^\top + U_\perp \cdot \Sigma_2 \cdot V_\perp^\top$$

$$\hat{X} = \hat{U} \cdot \hat{\Sigma}_1 \cdot \hat{V}^\top + \hat{U}_\perp \cdot \hat{\Sigma}_2 \cdot \hat{V}_\perp^\top$$

Wedin's $\sin\Theta$ Theorem (1972) states that if $\sigma_{\min}(\hat{\Sigma}_1) - \sigma_{\max}(\Sigma_2) = \delta > 0$,

$$\max\left\{\|\sin\Theta(V, \hat{V})\|, \|\sin\Theta(U, \hat{U})\|\right\} \leq \frac{\max\left\{\|Z\hat{V}\|, \|\hat{U}^\top Z\|\right\}}{\delta}.$$

- joint upper bound for both $\hat{U}$ and $\hat{V}$;
- may be sub-optimal.

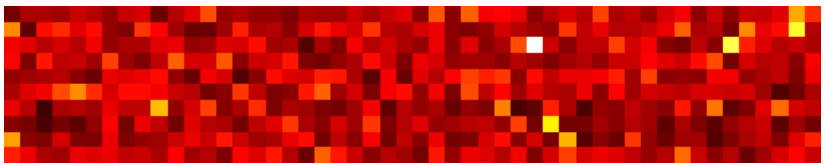


Figure: Intuitively, estimating V is more difficult than U for the matrix above.

Unilateral Perturbation Bound

- Decompose

$$Z = \begin{bmatrix} U & U_\perp \end{bmatrix} \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} \begin{bmatrix} V^\top \\ V_\perp^\top \end{bmatrix}.$$

$$Z_{11} = U^\top Z V, \quad Z_{21} = U_\perp Z V^\top, \quad Z_{12} = U^\top Z V_\perp, \quad Z_{22} = U_\perp Z V_\perp.$$

Define $z_{ij} := \|Z_{ij}\|$ for $i, j = 1, 2$.

Theorem (Unilateral Perturbation Bound (Cai & Z. 2016))

Denote $\alpha := \sigma_{\min}(U^\top \hat{X} V), \beta := \sigma_{\max}(U_\perp^\top \hat{X} V_\perp)$. *If* $\alpha^2 > \beta^2 + z_{12}^2 \wedge z_{21}^2$, *then*

$$\|\sin\Theta(V, \hat{V})\| \le \frac{\alpha z_{12} + \beta z_{21}}{\alpha^2 - \beta^2 - z_{21}^2 \wedge z_{12}^2} \wedge 1,$$

$$\|\sin\Theta(U, \hat{U})\| \le \frac{\alpha z_{21} + \beta z_{12}}{\alpha^2 - \beta^2 - z_{21}^2 \wedge z_{12}^2} \wedge 1.$$

Remark

- Since $\alpha > \beta$,

$$\text{if } z_{12} > z_{21}, \quad \frac{\alpha z_{12} + \beta z_{21}}{\alpha^2 - \beta^2 - z_{21}^2 \wedge z_{12}^2} > \frac{\alpha z_{21} + \beta z_{12}}{\alpha^2 - \beta^2 - z_{21}^2 \wedge z_{12}^2},$$

  vice versa.

- When $\alpha \gg \max(\beta, \|Z\|)$, the upper bound is approximately

$$\|\sin\Theta(V, \hat{V})\| \le \frac{z_{12}}{\alpha}, \quad \|\sin\Theta(U, \hat{U})\| \le \frac{z_{21}}{\alpha}.$$

  In contrast, Wedin's $\sin\Theta$ law only leads to

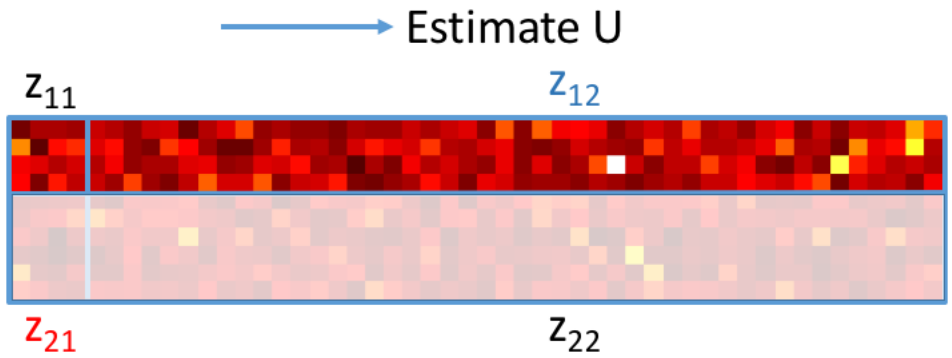
$$\|\sin\Theta(V, \hat{V})\| \le \frac{\|Z\|}{\alpha}, \quad \|\sin\Theta(U, \hat{U})\| \le \frac{\|Z\|}{\alpha}.$$

- The upper bound in Frobenius norm $\sin\Theta$ norm can be derived similarly.

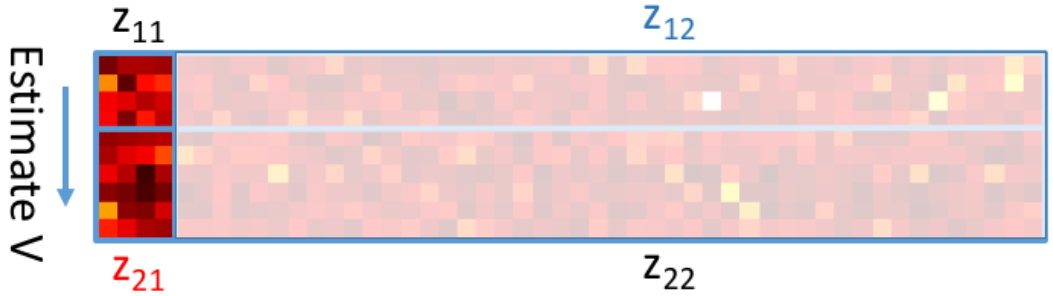
Idea Behind

Assume $U = \begin{bmatrix} I_r \\ 0 \end{bmatrix}$, $V = \begin{bmatrix} I_r \\ 0 \end{bmatrix}$. Let us take a look at $\hat{X}$.

- When estimating U, z_{12} becomes "signal" while z_{21} becomes "noise."



- When estimating V, z_{12} becomes "noise" while z_{21} becomes "signal."



Lower Bound

Theorem (Perturbation Lower Bound)

Define the class of $p_1 \times p_2$ rank-r matrices and perturbations,

$$\mathcal{F}_{r,\alpha,\beta,z_{21},z_{12}} = \Big\{(X,Z) : \mathit{rank}(X) = r,$$

$$\sigma_{\min}(U^\intercal \hat{X} V) \geq \alpha, \|Z_{22}\| \leq \beta, \|Z_{12}\| \leq z_{12}, \|Z_{21}\| \leq z_{21}\Big\}.$$

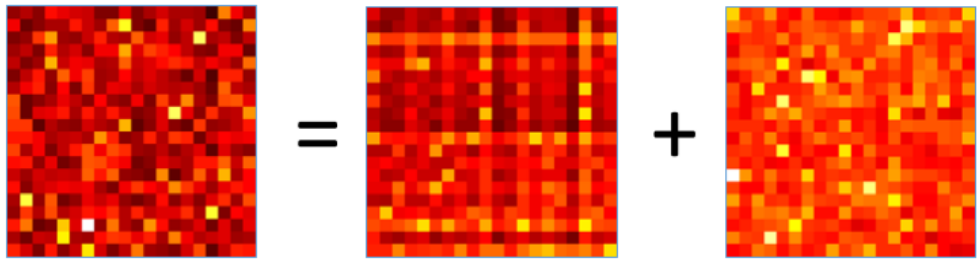
Provided that $\alpha^2 > \beta^2 + z_{12}^2 + z_{21}^2$, $r < \frac{p_1 \wedge p_2}{2}$,

$$\inf_{\tilde{V}} \sup_{(X,Z)\in\mathcal{F}_{\alpha,\beta,z_{21},z_{12}}} \left\|\sin\Theta(V,\tilde{V})\right\| \geq \frac{1}{2\sqrt{10}} \left(\frac{\alpha z_{12} + \beta z_{21}}{\alpha^2 - \beta^2 - z_{12}^2 \wedge z_{21}^2} \wedge 1 \right),$$

$$\inf_{\tilde{U}} \sup_{(X,Z)\in\mathcal{F}_{\alpha,\beta,z_{21},z_{12}}} \left\|\sin\Theta(U,\tilde{U})\right\| \geq \frac{1}{2\sqrt{10}} \left(\frac{\alpha z_{21} + \beta z_{12}}{\alpha^2 - \beta^2 - z_{12}^2 \wedge z_{21}^2} \wedge 1 \right).$$

Application: Matrix Denoising

$$\hat{X} = X + Z,$$



X is rank-r, $Z \overset{iid}{\sim}$ sub-Gaussian$(0, 1)$

- Target: U or V.
- Natural estimators for U, V: $\hat{U}$, $\hat{V}$, the first r singular vectors of $\hat{X}$.
- **Q: How do $\hat{U}$, $\hat{V}$ perform, respectively?**

- The r-th singular value of X, $\sigma_r(X)$, is a good characterization for the difficulty of this problem.
- Applying the perturbation bound, we obtain

Theorem

Suppose $X = U \cdot \Sigma \cdot V^\top \in \mathbb{R}^{p_1 \times p_2}$ is of rank-r. Then

$$E\left\|\sin\Theta(V,\hat{V})\right\|^2 \le \frac{C(p_2\sigma_r^2(X)+p_1p_2)}{\sigma_r^4(X)} \wedge 1,$$

$$E\left\|\sin\Theta(U,\hat{U})\right\|^2 \le \frac{C(p_1\sigma_r^2(X)+p_1p_2)}{\sigma_r^4(X)} \wedge 1.$$

Define the following class of low-rank matrices

$$\mathcal{F}_{r,t} = \{X \in \mathbb{R}^{p_1 \times p_2} : \text{rank}(X) = r, \sigma_r(X) \geq t\}.$$

Theorem (Lower Bound)

If $r \leq \frac{p_1}{16} \wedge \frac{p_2}{2}$, *then*

$$\inf_{\tilde{V}} \sup_{X \in \mathcal{F}_{r,t}} E\|\sin\Theta(V, \tilde{V})\|^2 \geq c\left(\frac{p_2 t^2 + p_1 p_2}{t^4} \wedge 1\right),$$

$$\inf_{\tilde{V}} \sup_{X \in \mathcal{F}_{r,t}} E\|\sin\Theta(U, \tilde{U})\|^2 \geq c\left(\frac{p_1 t^2 + p_1 p_2}{t^4} \wedge 1\right).$$

To sum up,

$$\inf_{\tilde{V}} \sup_{X \in \mathcal{F}_{r,t}} E\|\sin\Theta(V, \tilde{V})\|^2 \asymp \left(\frac{p_2 t^2 + p_1 p_2}{t^4} \wedge 1\right),$$

$$\inf_{\tilde{V}} \sup_{X \in \mathcal{F}_{r,t}} E\|\sin\Theta(U, \tilde{U})\|^2 \asymp \left(\frac{p_1 t^2 + p_1 p_2}{t^4} \wedge 1\right).$$

Some interesting facts

- Results for estimating X (Gavish and Donoho, 2014)

$$\inf_{\tilde{X}} \sup_{X \in \mathcal{F}_{r,t}} E \frac{\|\tilde{X} - X\|^2}{\|X\|^2} \asymp c\left(\frac{p_1 + p_2}{t^2} \wedge 1\right).$$

Thus,

$$\inf_{\tilde{X}} \sup_{X \in \mathcal{F}_{r,t}} E \frac{\|\tilde{X} - X\|^2}{\|X\|^2} \asymp \inf_{\tilde{U}} \sup_{X \in \mathcal{F}_{r,t}} E\|\sin\Theta(\tilde{U}, U)\| + \inf_{\tilde{V}} \sup_{X \in \mathcal{F}_{r,t}} E\|\sin\Theta(\tilde{V}, V)\|.$$

- When $p_2 \gg p_1$, $(p_1 p_2)^{1/2} \ll t^2 \ll p_2$,

$$\inf_{\tilde{V}} \sup_{X \in \mathcal{F}_{r,t}} E\|\sin\Theta(\hat{V}, V)\| \geq c, \quad \inf_{\tilde{X}} \sup_{X \in \mathcal{F}_{r,t}} E \frac{\|\tilde{X} - X\|^2}{\|X\|^2} \geq c.$$

On the other hand,

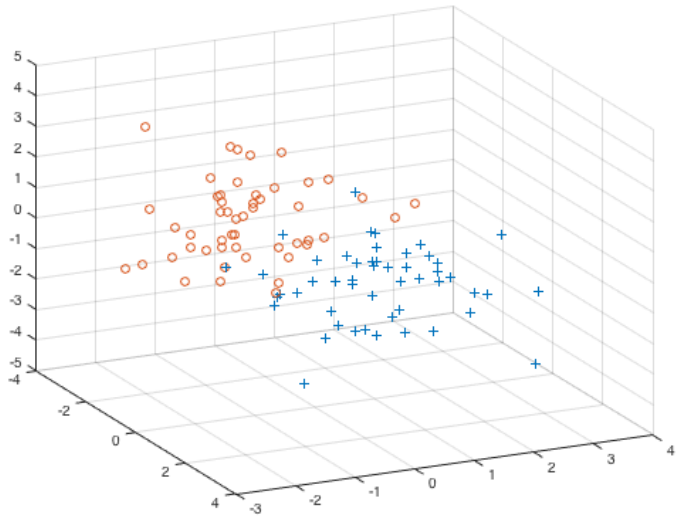
$$E\left\|\sin\Theta(\hat{U}, U)\right\|^2 \to 0.$$

Simulation Results

(p_1, p_2, r, t)	$\|\sin\Theta(\hat{U}, U)\|^2$	$\|\sin\Theta(\hat{V}, V)\|^2$
(10, 100, 2, 15)	0.0669	0.3512
(10, 100, 2, 30)	0.0139	0.1120
(20, 100, 5, 20)	0.0930	0.2711
(20, 100, 5, 40)	0.0195	0.0770
(20,1000, 5, 30)	0.0699	0.5838
(20, 1000, 10, 100)	0.0036	0.1060
(200, 1000, 10, 50)	0.0797	0.3456
(200, 1000, 50, 100)	0.0205	0.1289

Table: Average losses in spectral $\sin\Theta$ distances for both the left and right singular space changes after Gaussian noise perturbations.

Application 2: High-dimensional Clustering



- Observations: $X_1, \ldots, X_n \in \mathbb{R}^p$, $p \geq n$.
- Each point belongs to one of two classes.

$$X_i = \mu l_i + \varepsilon_i, \quad i = 1, \ldots, n, \quad \varepsilon_i \overset{iid}{\sim} \text{sub-Gaussian}(0, \sigma^2).$$

$l_i \in \{-1, 1\}$ are labels; $\mu \in \mathbb{R}^p$ is the mean.

- **Goal**: recover labels l.

- Suppose $\hat{u} \in \mathbb{R}^p$, $\hat{v} \in \mathbb{R}^n$ are the first left, right singular vector of

$$[X_1 X_2 \cdots X_n] \in \mathbb{R}^{p \times n}$$

- Method: in this simple model, we recover l by

$$\hat{l} = \text{sgn}(\hat{v}).$$

- Reason:
  - $\hat{u}$ contains information of μ $\rightarrow$ less important;
  - $\hat{v}$ contains information of l $\rightarrow$ more important.

  Good match to the unilateral perturbation bound.

For any label estimator $\tilde{l}$, define the misclassification rate

$$\mathcal{M}(\tilde{l}, l) = \frac{1}{n} \max \left\{ \sum_{i=1}^{p} 1\{\tilde{l}_i \neq l_i\}, \sum_{i=1}^{p} 1\{\tilde{l}_i \neq -l_i\} \right\}.$$

Theorem (Misclassification Rate)

Suppose $p \geq n$. When $\|\mu\|_2 \geq C(p/n)^{\frac{1}{4}}$,

$$E\mathcal{M}(\hat{l}, l) \leq \frac{C}{\|\mu\|_2^2} + \frac{Cp}{n\|\mu\|_2^4}.$$

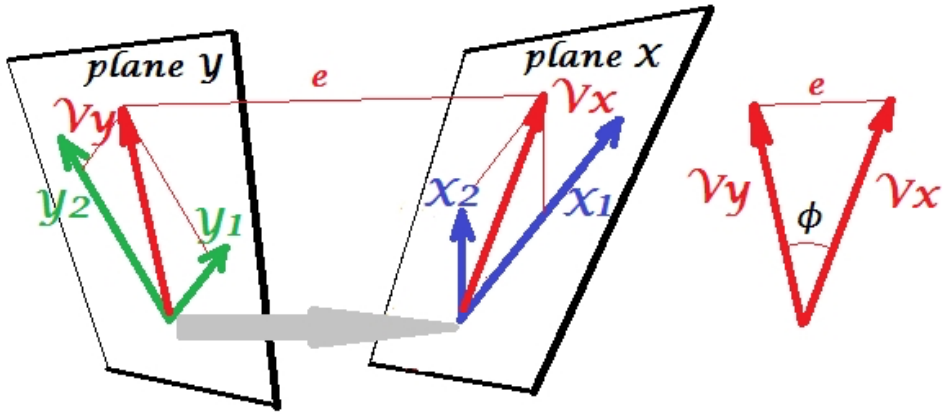
Moreover, $\|\mu\|_2 \geq C(p/n)^{1/4}$ is necessary since

Theorem (Lower Bound)

Suppose $p \geq n$,

$$\inf_{\hat{l}} \sup_{\mu:\|\mu\|_2 \geq c(p/n)^{\frac{1}{4}}} E\mathcal{M}(\tilde{l}, l) \geq \frac{1}{4}.$$

Application 3: Canonical Correlation Analysis (CCA)



- Two sets of random variables with joint distribution

$$\mathrm{Cov}\begin{pmatrix} X \\ Y \end{pmatrix} \sim \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix}.$$

- n observations

$$[X_1, \ldots, X_n] \in \mathbb{R}^{p_1 \times n}, \quad [Y_1, \ldots, Y_n] \in \mathbb{R}^{p_2 \times n}.$$

- **Canonical Correlation Analysis (CCA)** searches for the pairs of canonical correlation directions with maximized correlation.

- In short,

$$S = \Sigma_X^{-1/2} \Sigma_{XY} \Sigma_Y^{-1/2} \approx U \Sigma_1 V^\top.$$

Canonical correlation directions:

$$A = \Sigma_X^{-1/2} U, \quad B = \Sigma_Y^{-1/2} V.$$

- To estimate, we calculate

$$\hat{S} = \hat{\Sigma}_X^{-1/2} \hat{\Sigma}_{XY} \hat{\Sigma}_Y^{-1/2} \approx \hat{U} \hat{\Sigma}_1 \hat{V}^\top + \hat{U}_\perp \hat{\Sigma}_2 \hat{V}_\perp^\top.$$

Sample Canonical correlation directions:

$$\hat{A} = \hat{\Sigma}_X^{-1/2} \hat{U}, \quad \hat{B} = \hat{\Sigma}_Y^{-1/2} \hat{V}.$$

Theorem (Unilateral Upper Bound for CCA)

Whenever $\sigma_r^2(S) \geq C((p_1p_2)^{\frac{1}{2}} + p_1 + p_2^{\frac{3}{2}}/n^{\frac{1}{2}})$, *with high probability*

$$\max_O E_{X^*}\|(\hat{A}O)^\top X^* - A^\top X^*\|_2^2 \leq \frac{Crp_1}{n\sigma_r^2(S)} + \frac{Crp_1p_2}{n^2\sigma_r^4(S)}.$$

$$\max_O E_{Y^*}\|(\hat{B}O)^\top Y^* - B^\top Y^*\|_2^2 \leq \frac{Crp_2}{n\sigma_r^2(S)} + \frac{Crp_1p_2}{n^2\sigma_r^4(S)}.$$

- When $p_2 \gg p_1$, $\frac{p_2}{n} \gg \sigma_r^2(S) \gg \frac{(p_1p_2)^{\frac{1}{2}}}{n}$,
  no consistent estimator for B;
  $\hat{A}$ is a consistent estimator of A.
- This interesting phenomena again shows the merit of our proposed unilateral perturbation bound.

Other Applications...

The proposed perturbation bound can be potentially used in other applications...

- Community detection
- Multidimensional scaling (MDS)
- Matrix completion
- Cross-covariance matrix estimation
- ...

Reference

- Cai, T. T., & Zhang, A. (2018). Rate-Optimal Perturbation Bounds for Singular Subspaces with Applications to High-Dimensional Statistics. *Annals of Statistics*, 43, 102-138.

Thank you for your attention!