Sparse CCA via Precision Adjusted Iterative Thresholding

Speaker : Semin Choi

Department of Statistics, Seoul National University, South Korea

January 10, 2017

Sparse CCA via Precision Adjusted Iterative Thresholding

- Let $(X, Y) \in \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$ denote random vectors with covariances (Σ_1, Σ_2) and cross-covariance Σ_{12}.
- CCA finds pairs of linear projections of the two views, $(a'X, b'Y)$ that are maximally correlated:

Proposition 1.

When Σ_{12} is of rank 1, the solution (up to sign jointly) of CCA problem is (θ, η) if and only if the covariance structure between X and Y can be written as

$$\Sigma_{12} = \lambda \Sigma_1 \theta \eta^T \Sigma_2$$

where $0 < \lambda \leq 1$, $\theta^T \Sigma_1 \theta = 1$ and $\eta^T \Sigma_2 \eta = 1$. In other words, the correlation between $a^T X$ and $b^T Y$ are maximized by corr($\theta^T X, \eta^T Y$), and λ is the canonical correlation between X and Y.

Sparse CCA via Precision Adjusted Iterative Thresholding

Proposition 2.

For general Σ_{12} with rank $r \geq 1$, the solution (up to sign jointly) of CCA problem is (θ_1, η_1) if and only if the covariance structure between X and Y can be written as

$$\Sigma_{12} = \Sigma_1 \left(\sum_{i=1}^{r} \lambda_i \theta_i \eta_i^T \right) \Sigma_2$$

where $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_r > 0$, $\theta_i^T \Sigma_1 \theta_j = \mathbb{I}(i = j) = \eta_i^T \Sigma_2 \eta_j$.

CAPIT : Iterative Thresholding

Algorithm 1 : CAPIT

Input : Sample covariance matrices $\hat{\Sigma}_{12}$;
Estimators of precision matrix $\hat{\Omega}_1, \hat{\Omega}_2$;
Initialization pair $\alpha^{(0)}, \beta^{(0)}$;
Thresholding level γ_1, γ_2.
Output : Canonical direction estimator $\alpha^{(\infty)}, \beta^{(\infty)}$.
Set $\hat{A} = \hat{\Omega}_1 \hat{\Sigma}_{12} \hat{\Omega}_2$;
repeat

- Right Multiplication: $\omega^{l,(i)} = \hat{A}\beta^{(i-1)}$;
- Left Thresholding : $\omega_{th}^{l,(i)} = T(\omega^{l,(i)}, \gamma_1)$;
- Left Normalization : $\alpha^{(i)} = \omega_{th}^{l,(i)} / \|\omega_{th}^{l,(i)}\|$;
- Left Multiplication : $\omega^{r,(i)} = \alpha^{(i)} \hat{A}$;
- Right Thresholding : $\omega_{th}^{r,(i)} = T(\omega_{th}^{r,(i)}, \gamma_2)$;
- Right Normalization : $\beta^{(i)} = \omega_{th}^{r,(i)} / \|\omega_{th}^{r,(i)}\|$;

until Convergence of $\alpha^{(i)}$ and $\beta^{(i)}$.

CAPIT : Iterative Thresholding

- CAPIT without thresholding = SVD-power method.
- $\Omega_1 \Sigma_{12} \Omega_2 \Rightarrow \Omega_1^{1/2} \Sigma_{12} \Omega_2^{1/2}$?
    - Let $\Sigma_1^{1/2} \theta_i = \theta_i'$ and $\Sigma_2^{1/2} \eta_i = \eta_i'$.
    - Then, $\Sigma_{12} = \Sigma_1^{1/2} \left(\sum_{i=1}^{r} \lambda_i \theta_i' \eta_i'^T \right) \Sigma_2^{1/2}$ and $\|\theta_i'\|_2 = \|\eta_i'\|_2 = 1$.
    - It is same as the original CCA algorithm.

Initialization by Coordinate Thresholding

Algorithm 2 (CAPIT : Initialization by Coordinate Thresholding)

Input : Sample covariance matrices $\hat{\Sigma}_{12}$;
Estimators of precision matrix $\hat{\Omega}_1, \hat{\Omega}_2$;
Thresholding level t_{ij}.
Output : Initializer $\alpha^{(0)}$ and $\beta^{(0)}$.
Set $\hat{A} = \hat{\Omega}_1 \hat{\Sigma}_{12} \hat{\Omega}_2$;

1. Coordinate selection : pick the index sets B_1 and B_2 of the coordinates of θ and η respectively as follows,
   $B_1 = \{i : \max_j |\hat{a}_{ij}|/t_{ij} \geq \sqrt{\frac{\log p_1}{n}}\}$,
   $B_2 = \{j : \max_i |\hat{a}_{ij}|/t_{ij} \geq \sqrt{\frac{\log p_2}{n}}\}$;
2. Reduced SVD : compute the leading pair of singular vectors $(\alpha^{(0),B_1}, \beta^{(0),B_2})$ on the submatrix $\hat{A}_{B_1,B_2}$;
3. Zero-padding procedure : construct the initializer $(\alpha^{(0)}, \beta^{(0)})$ by zero-padding $(\alpha^{(0),B_1}, \beta^{(0),B_2})$ on index sets B_1^c and B_2^c respectively,
   $\alpha^{(0)}_{B_1} = \alpha^{(0),B_1}, \alpha^{(0)}_{B_1^c} = 0, \beta^{(0)}_{B_2} = \beta^{(0),B_2}, \beta^{(0)}_{B_2^c} = 0$

The Single Canonical Pair Model

- We propose a probabilistic model of (X, Y), so that the canonical directions (θ, η) are explicitly modeled in the joint distribution of (X, Y).

The Single Canonical Pair Model

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_1 & \lambda\Sigma_1\theta\eta^T\Sigma_2 \\ \lambda\Sigma_2\eta\theta^T\Sigma_1 & \Sigma_2 \end{pmatrix} \right) \quad (1)$$

with $\Sigma_1 > 0, \Sigma_2 > 0, \theta^T\Sigma_1\theta = \eta^T\Sigma_2\eta = 1$ and $0 < \lambda \leq 1$.

Convergence Rates

- We consider the idea of data splitting.
- Suppose we have $2n$ i.i.d. copies $(X_i, Y_i)_{1 \leq i \leq 2n}$.
- $\hat{\Sigma}_{12} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i^T$.
- The reason for data splitting is that we can write the matrix $\hat{A}$ in an alternative form :

$$\hat{A} = \frac{1}{n}\sum_{i=1}^{n} \tilde{X}_i \tilde{Y}_i^T$$

  where $\tilde{X}_i = \hat{\Omega}_1 X_i$ and $\tilde{Y}_i = \hat{\Omega}_2 Y_i$ for all $i = 1, ..., n$.
- Conditioning on $(X_i, Y_i)_{n+1 \leq i \leq 2n}$, the transformed data $(\tilde{X}_i, \tilde{Y}_i)_{1 \leq i \leq n}$ are still i.i.d.

Convergence Rates

- Conditioning on $(X_i, Y_i)_{n+1\leq i\leq 2n}$, the expectation of $\hat{A}$ is $\lambda\alpha\beta^T$ where $\alpha = \hat{\Omega}_1\Sigma_1\theta$ and $\beta = \hat{\Omega}_2\Sigma_2\eta$.
- We consider the loss function $L(a,b)^2 = 2|\sin\angle(a,b)|^2$
- It is easy to calculate that

$$L(a,b) = \left\| \frac{aa^T}{\|a\|^2} - \frac{bb^T}{\|b\|^2} \right\|_F$$

Convergence Rates

To achieve statistical consistency, we need some assumptions on the interesting part (θ, η) and nuisance part $(\Sigma_1, \Sigma_2, \lambda)$.

Assumption A - Sparsity Condition on (θ, η):

We assume θ and η are in the weak l_q ball, with $0 \leq q \leq 2$. i.e.

$$|\theta_{(k)}|^q \leq s_1 k^{-1}, \ |\eta_{(k)}|^q \leq s_2 k^{-1},$$

where $\theta_{(k)}$ is the k-th largest coordinate by magnitude. Let $p = p_1 \vee p_2$ and $s = s_1 \vee s_2$.

The sparsity level s_1 and s_2 satisfy the following condition,

$$s = o\left(\left(\frac{n}{\log p}\right)^{\frac{1}{2} - \frac{q}{4}}\right)$$

Convergence Rates

Assumption B - General Conditions on $(\Sigma_1, \Sigma_2, \lambda)$:

- (a) We assume there exist constants w and W, such that

$$0 < w \leq \lambda_{\min}(\Sigma_i) \leq \lambda_{\max}(\Sigma_i) \leq W < \infty$$

  for $i = 1, 2$.

- (b) In order that the signals do not vanish, we assume the canonical correlation is bounded below by a positive constant C_λ, i.e. $0 < C_\lambda \leq \lambda \leq 1$.
- (c) Moreover, we require that estimators $(\hat{\Omega}_1, \hat{\Omega}_2)$ are consistent in the sense that

$$\xi_\Omega = \|\hat{\Omega}_1 \Sigma_1 - I\| \vee \|\hat{\Omega}_2 \Sigma_2 - I\| = o(1),$$

  with probability at least $1 - O(p^{-2})$.

Convergence Rates

Theorem 1(Convergence Rates)

Assume the Assumptions A and B hold. Let $(\alpha^{(k)}, \beta^{(k)})$ be the sequence from Algorithm 1, with the initializer $(\alpha^{(0)}, \beta^{(0)})$ calculated by Algorithm 2. The thresholding levels are

$$t_{ij}, \quad \gamma_1 = c_1\sqrt{\frac{\log p}{n}}, \quad \gamma_2 = c_2\sqrt{\frac{\log p}{n}}$$

for sufficiently large constants (t_{ij}, c_1, c_2). Then with probability at least $1 - O(p^{-2})$, we have

$$L(\alpha^{(k)}, \theta)^2 \vee L(\beta^{(k)}, \eta)^2 \leq C\left(s\left(\frac{\log p}{n}\right)^{1-q/2} + \|(\hat{\Omega}_1\Sigma_1 - I)\theta\|^2 \vee \|(\hat{\Omega}_2\Sigma_2 - I)\eta\|^2\right)$$

for all $k = 1, 2, ..., K$ with $K = O(1)$ and some constant $C > 0$.

Data-Driven Thresholding

$$t_{ij} = \frac{20\sqrt{2}}{9}\left(\sqrt{\|\hat{\Omega}_1\|\hat{\omega}_{2,jj}} + \sqrt{\|\hat{\Omega}_2\|\hat{\omega}_{1,ii}} + \sqrt{\hat{\omega}_{1,ii}\hat{\omega}_{2,jj}} + \sqrt{8\|\hat{\Omega}_1\|\|\hat{\Omega}_2\|/3}\right)$$

$$\gamma_1 = (0.17 \min_{i,j} t_{ij}\|\hat{\Omega}_2\|^{1/2} + 2.1\|\hat{\Omega}_2\|^{1/2}\|\hat{\Omega}_1\|^{1/2} + 7.5\|\hat{\Omega}_2\|)\sqrt{\frac{\log p}{n}}$$

$$\gamma_2 = (0.17 \min_{i,j} t_{ij}\|\hat{\Omega}_1\|^{1/2} + 2.1\|\hat{\Omega}_1\|^{1/2}\|\hat{\Omega}_2\|^{1/2} + 7.5\|\hat{\Omega}_1\|)\sqrt{\frac{\log p}{n}}$$

$$\delta_1 = \delta_2 = 0.08 w^{1/2} \min_{i,j} t_{ij} \qquad \text{(in the next page)}$$

Outline of Proof for Convergence Rates

- Construction of the Oracle Sequence :

$$H_1 = \left\{ k : |\alpha_k| \geq \delta_1 \sqrt{\frac{\log p_1}{n}} \right\}, \quad H_2 = \left\{ k : |\beta_k| \geq \delta_2 \sqrt{\frac{\log p_2}{n}} \right\}$$

  and $L_1 = H_1^c$, $L_2 = H_2^c$.

- Then, we define the oracle version of $\hat{A}$: $\hat{A}^{\text{ora}} = \begin{pmatrix} \hat{A}_{H_1 H_2} & 0 \\ 0 & 0 \end{pmatrix}$.

- We construct the oracle initializer $(\alpha^{(0),\text{ora}}, \beta^{(0),\text{ora}})$ based on an oracle version of Algorithm 2 with the sets B_1 and B_2 replaced by $B_1^{\text{ora}} = B_1 \cap H_1$ and $B_2^{\text{ora}} = B_2 \cap H_2$.

- Feeding the oracle initializer $(\alpha^{(0),\text{ora}}, \beta^{(0),\text{ora}})$ and the matrix $\hat{A}^{\text{ora}}$ into Algorithm 1, we get the oracle sequence $(\alpha^{(k),\text{ora}}, \beta^{(k),\text{ora}})$.

Outline of Proof for Convergence Rates

1. We are going to bound $L(\hat{\alpha}^{\text{ora}}, \alpha)$ and $L(\hat{\beta}^{\text{ora}}, \beta)$ where $(\hat{\alpha}^{\text{ora}}, \hat{\beta}^{\text{ora}})$ is the first pair of singular vectors of $\hat{A}^{\text{ora}}$.
2. Show that the oracle sequence $(\alpha^{(k),\text{ora}}, \beta^{(k),\text{ora}})$ converges to $(\hat{\alpha}^{\text{ora}}, \hat{\beta}^{\text{ora}})$ after finite steps of iterations.
3. Show that the estimating sequence $(\alpha^{(k)}, \beta^{(k)})$ and the oracle sequence $(\alpha^{(k),\text{ora}}, \beta^{(k),\text{ora}})$ are identical with high probability.

Outline of Proof for Convergence Rates

Lemma 1

Under Assumptions A and B, we have

$$L(\hat{\alpha}^{\text{ora}}, \alpha)^2 \vee L(\hat{\beta}^{\text{ora}}, \beta)^2 \leq C\left(s\left(\frac{\log p}{n}\right)^{1-q/2} + \|\theta - \alpha\|^2 \vee \|\eta - \beta\|^2\right)$$

with probability at least $1 - O(p^{-2})$ for some constant $C > 0$.

- Let $A^{\text{ora}} = \begin{pmatrix} A_{H_1 H_2} & 0 \\ 0 & 0 \end{pmatrix}$ and $(\alpha^{\text{ora}}, \beta^{\text{ora}})$ be the first singular vectors of A^{ora}.
- $L(\hat{\alpha}^{\text{ora}}, \alpha) \leq L(\hat{\alpha}^{\text{ora}}, \alpha^{\text{ora}}) + L(\alpha^{\text{ora}}, \alpha)$

Outline of Proof for Convergence Rates

Lemma 2

Under Assumptions A and B, we have

$$L(\alpha^{(k),\text{ora}}, \hat{\alpha}^{\text{ora}})^2 \leq C\left(s\left(\frac{\log p}{n}\right)^{1-q/2} + \|\theta - \alpha\|^2\right)$$

for all $k \geq 1$ with probability at least $1 - O(p^{-2})$ for some constant $C > 0$.

Lemma 3

Under Assumptions A,B and the current choice of (γ_1, γ_2), $(\alpha^{(k),\text{ora}}, \beta^{(k),\text{ora}}) = (\alpha^{(k)}, \beta^{(k)})$ for all $k = 1, \ldots, K$, $K = O(1)$, with probability at least $1 - O(p^{-2})$.

Convergence Rates

- $\mathcal{G}_{q_0}(s_0, p) = \left\{\Omega = (\omega_{ij})_{p\times p} : \max_j |\omega_{j(k)}|^{q_0} \le s_0 k^{-1} \text{for all } k\right\}$ for $0 \le q_0 \le 1$.

Corollary 1(Convergence Rates)

Assume the Assumptions A and B holds, $\Omega_i \in \mathcal{G}_{q_0}(s_0, p_i), i = 1, 2, \|\Omega_i\|_{l_1} \le w^{-1}$ and $s_0^2 = O\big((n/\log p)^{1-q_0}\big)$. $\hat{\Omega}_i$ is obtained by applying CLIME procedure in Cai et al. (2011). Then, with probability at least $1 - O(p^{-2})$, we have

$$L(\alpha^{(k)}, \theta)^2 \vee L(\beta^{(k)}, \eta)^2 \le C\left(s\left(\frac{\log p}{n}\right)^{1-q/2} + s_0^2\left(\frac{\log p}{n}\right)^{1-q_0}\right)$$

for all $k = 1, 2, ..., K$ with $K = O(1)$ and some constant $C > 0$.

Minimax Lower Bound

$$\mathcal{F}_q^{p_1,p_2}(s_1,s_2,C_\lambda)=\left\{\begin{array}{c} N(0,\Sigma):\Sigma \text{ is specified in (1) }, \lambda\in(C_\lambda,1) \\ \Sigma_i=I_{p_i\times p_i}, i=1,2, \\ |\theta|_{(k)}^q\in s_1k^{-1}, |\eta|_{(k)}^q\leq s_2k^{-1}, \text{ for all } k. \end{array}\right\}.$$

Theorem 2:Minimax lower bound for known variance

For any $q\in[0,2]$, we assume that $s_i\left(\frac{n}{\log p_i}\right)^{q/2}=o(p_i)$ for $i=1,2$ and $\log p_1\asymp\log p_2$. Moreover, we also assume $s\left(\frac{\log p}{n}\right)^{1-q/2}\leq c_0$ for some constant $c_0>0$. Then we have

$$\inf_{(\hat\theta,\hat\eta)}\sup_{P\in\mathcal{F}}\mathbb{E}_P\left(L^2(\hat\theta,\theta)\vee L^2(\hat\eta,\eta)\right)\geq Cs\left(\frac{\log p}{n}\right)^{1-q/2}$$

where $\mathcal{F}=\mathcal{F}_q^{p_1,p_2}(s_1,s_2,C_\lambda)$ and C is a constant only depending on q and C_λ.

Minimax Lower Bound

$$\mathcal{F}_{q,q_0}^{p_1,p_2}(s_0, s_1, s_2, C_\lambda, w, W) = \left\{ \begin{array}{c} N(0,\Sigma) : \Sigma \text{ is specified in } (1)\,, \lambda \in (C_\lambda, 1) \\ \Sigma_i^{-1} \in \mathcal{G}_{q_0}(s_0, p_i), W^{-1} \le \lambda_{\min}(\Sigma_i^{-1}), \|\Sigma_i^{-1}\|_{l_1} \le w^{-1}, \\ |\theta|_{(k)}^q \in s_1 k^{-1}, |\eta|_{(k)}^q \le s_2 k^{-1}, \text{ for all } k. \end{array} \right.$$

- Note that $\mathcal{F}_q^{p_1,p_2}(s_1, s_2, C_\lambda) \subset \mathcal{F}_{q,q_0}^{p_1,p_2}(s_0, s_1, s_2, C_\lambda, w, W)$.
- The lower bound is same as above.

Corollary 2: Minimax rate

Under the assumptions in Corollary 1 and Theorem 2 and assume $n = o(p^h)$ for some $h > 0$. we have

$$\inf_{(\hat{\theta},\hat{\eta})} \sup_{P \in \mathcal{F}} \mathbb{E}_P \Big(L^2(\hat{\theta}, \theta) \vee L^2(\hat{\eta}, \eta) \Big) \asymp s \left(\frac{\log p}{n} \right)^{1-q/2}$$

for $\mathcal{F} = \mathcal{F}_{q,q_0}^{p_1,p_2}(s_0, s_1, s_2, C_\lambda, w, W)$, provided that $s_0^2 \left(\frac{\log p}{n} \right)^{1-q_0} \le Cs \left(\frac{\log p}{n} \right)^{1-q/2}$ for some constant $C > 0$.