36-710: Advanced Statistical Theory **Fall 2018**

Lecture 18: November 8

Lecturer: Alessandro Rinaldo *Scribes: Yue Li*

Today we continue talking about PCA in high dimensions. By applying PCA we have the empirical covariance matrix $\hat{\Sigma}$. Formally, we have

$$\hat{\Sigma} = \Sigma + E,$$

where $\hat{\Sigma}$ is a $d \times d$ PSD matrix, Σ is a $d \times d$ matrix representing the true covariance, and E is error term of perturbation.

Our goal is to estimate leading eigenvalue and eigenvector of Σ using $\hat{\Sigma}$.

- To estimate eigenvalues, the mail tool is Weyl's inequality

  $$\max_{j=1,2,\ldots,d} \left| \lambda_j(\Sigma) - \lambda_j(\hat{\Sigma}) \right| \leq \|E\|_{op} = \|\Sigma - \hat{\Sigma}\|_{op},$$

  where $\lambda_j(A)$ is the jth leading eigenvalue of A(A is PSD matrix).

- To estimate eigenspaces of Σ it is not enough to have $\max_{j=1,2,\ldots,d} \left| \lambda_j - \hat{\lambda}_j \right|$ to be small. ($\hat{\lambda}_j = \lambda_j(\hat{\Sigma})$). Below we introduce the definition and theorems related to the analysis of eigenspaces.

See [SS90] and [Bh13] for reference.

18.1 Distance between linear subspaces

Let $\mathcal{E}$ and $\mathcal{F}$ be d-dimensional linear subspaces of $\mathbb{R}^p$, and let $P_{\mathcal{E}}$ and $P_{\mathcal{F}}$ be projector matrices onto $\mathcal{E}$ and $\mathcal{F}$. Formally, $P_{\mathcal{E}}x \in \mathcal{E}$ satisfies $P_{\mathcal{E}}x = \arg\min_{\alpha \in \mathcal{E}} \|x - \alpha\|_2$, and similar for $P_{\mathcal{F}}x$.

Let $E \in \mathbb{R}^{p \times d}$ and $F \in \mathbb{R}^{p \times d}$ be orthogonal matrices whose column range is $\mathcal{E}$ and $\mathcal{F}$. Then it can be shown that $P_{\mathcal{E}} = EE^T$. Below we introduce the conception of canonical angle.

Remark: For a "one-dimensional" illustration, let $v_1, v_2 \in \mathbb{S}^{p-1}$. The angle between v_1 and v_2 is

$$\theta = \angle(v_1, v_2) = \cos^{-1}(|v_1^T v_2|).$$

More generally, $\cos(\theta) = \frac{v_1^T v_2}{\|v_1\| \|v_2\|}$.

Definition 18.1 *The canonical angle between $\mathcal{E}$ and $\mathcal{F}$ are $\theta_1, \theta_2, \ldots, \theta_d$, where $\theta_1 = \cos^{-1}(\sigma_1), \ldots, \theta_d = \cos^{-1}(\sigma_d)$, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$ are d singular values of $E^T F$ or $F^T E$.*

One way to think about this is, if we write $E^T F = U\cos(\Theta)V^T$, then $\Theta = \text{diag}(\theta_1, \ldots, \theta_d)$. Follows from a central result in Matrix Analysis known as the CS-decomposition.

Equivalent interpretation of kth canonical angle, $k = 1, 2, \ldots, d$:

- When $k = 1$, canonical angle
$$\theta_1 = \cos^{-1}\left(\max_{x\in\mathcal{E},\|x\|=1}\max_{y\in\mathcal{F},\|y\|=1}|x^T y|\right) = \cos^{-1}(x_1^{*T}y_1^*).$$
- For $k = 2, \ldots, d$, we have
$$\theta_k = \cos^{-1}\left(\max_{x\in\mathcal{E},\|x\|=1}\max_{y\in\mathcal{F},\|y\|=1}|x^T y|\right),$$
with $x^T x_{k-j}^* = y^T y_{k-j}^* = 0$, $j = 1, \ldots, k-1$.

Above is a kind of intuitive way of understanding canonical angle. Below is another definition of canonical angle which is more often used.

Definition 18.2 *An equivalent definition of canonical angle is*
$$\theta_j = \sin^{-1}(s_j), j = 1, 2, \ldots, d,$$
where $s_1 \geq s_2 \geq \cdots \geq s_d > 0$ are singular values of $P_{\mathcal{E}}P_{\mathcal{F}^\perp} = EE^T(I_p - FF^T)$, where $P_{\mathcal{F}^\perp} = I_p - P_{\mathcal{F}}$.

In other words, we can write $P_{\mathcal{E}}P_{\mathcal{F}^\perp} = U\sin(\Theta)V^T$, with $\Theta = \text{diag}(\theta_1, \ldots, \theta_d)$.

Definition 18.3 *The quantity $\|\sin\Theta(\mathcal{E},\mathcal{F})\|_F$ is a metric over all d-dimensional linear sub-spaces in $\mathbb{R}^p$.*

And we have equivalent expressions
$$\|\sin\Theta\|_F^2 = \|P_{\mathcal{E}}P_{\mathcal{F}^\perp}\|_F^2 = \|P_{\mathcal{F}}P_{\mathcal{E}^\perp}\|_F^2 = \frac{1}{2}\|P_{\mathcal{E}} - P_{\mathcal{F}}\|_F^2.$$

18.2 Davis-Khan Theorem

In this part we introduce one important result in matrix perturbation.

Theorem 18.4 *(Davis-Khan) Let Σ and $\hat\Sigma$ be symmetric $p \times p$ matrix with eigenvalues*
$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p, \quad \hat\lambda_1 \geq \hat\lambda_2 \geq \cdots \geq \hat\lambda_p.$$
If the following conditions hold:

1. *Integers r, s and d satisfy $1 \leq r \leq s \leq p$, $d = s - r + 1$,*
2. *$V, \hat V \in \mathbb{R}^{p\times d}$ with columns given by eigenvectors of Σ and $\hat\Sigma$ respectively corresponding to $\{\lambda_j, j = r, \ldots, s\}$ and $\{\hat\lambda_j, j = r, \ldots, s\}$.*
3. *Let $\delta = \inf\left\{|\hat\lambda - \lambda|, \lambda \in [\lambda_s, \lambda_r], \hat\lambda \in \left(-\infty, \hat\lambda_{s+1}\right] \cup \left[\hat\lambda_{r-1}, +\infty\right)\right\}$, where $\hat\lambda_0 = +\infty$, $\hat\lambda_{p+1} = -\infty$.*

4. $\mathcal{E} = range(V)$, $\mathcal{F} = range(\hat{V})$.

Then, if $\delta > 0$,

$$\|\sin\Theta(\mathcal{E},\mathcal{F})\|_F \le \frac{\|\hat{\Sigma}-\Sigma\|_F}{\delta}.$$

Remarks:

1. In [YWS14], better bound is achieved by replacing $\|\cdot\|_F$ by $\|\cdot\|_{op}$:

$$\|\sin\Theta(\mathcal{E},\mathcal{F})\|_F \le \frac{2\min\left\{\sqrt{d}\|\hat{\Sigma}-\Sigma\|_{op}, \|\hat{\Sigma}-\Sigma\|_F\right\}}{\min\{\lambda_{r-1}-\lambda_r, \lambda_s-\lambda_{s+1}\}}.$$

   Note that the eigengap part does not depend on data.

2. The paper [YWS14] also shows there exists an orthogonal matrix $\hat{O} \in \mathbb{R}^{d\times d}$, such that

$$\|\hat{V}\hat{O}-V\|_F \le \frac{2^{3/2}\min\left\{\sqrt{d}\|\hat{\Sigma}-\Sigma\|_{op}, \|\hat{\Sigma}-\Sigma\|_F\right\}}{\min\{\lambda_{r-1}-\lambda_r, \lambda_s-\lambda_{s+1}\}}.$$

3. In many applications $s = r = j$ and $d = 1$. Then the bound becomes

$$\sin\Theta(v_j,\hat{v}_j) \le \frac{\|\hat{\Sigma}-\Sigma\|_F}{\min\{\lambda_{j-1}-\lambda_j, \lambda_j-\lambda_{j+1}\}},$$

   where v_j and $\hat{v}_j$ are the jth leading eigenvector of Σ and $\hat{\Sigma}$. And by the conclusion in [YWS14],

$$\min_{\epsilon\in\{-1,+1\}} \|\epsilon\hat{v}_j - v_j\|_2 \le 2^{3/2}\frac{\|\hat{\Sigma}-\Sigma\|_{op}}{\min\{\lambda_{j-1}-\lambda_j, \lambda_j-\lambda_{j+1}\}}.$$

18.3 PCA in high-dimensions

For PCA in high dimensions, references can be found in [JL09].

In the case of spiked covariance model, we have

$$\Sigma = \theta vv^T + I_p, \text{ where } \theta > 0, v \in \mathbb{S}^{p-1}.$$

Then the eigenvalues of Σ are $(1_\theta, 1, \ldots, 1)$. For the leading vector, we have

$$\max_{\epsilon\in\{-1,+1\}} \|\epsilon\hat{v}-v\|_2^2 = 2 - 2|\hat{v}^Tv| \le 2 - 2(\hat{v}^Tv)^2 = \|\hat{v}\hat{v}^T - vv^T\|_F^2 = 2\sin^2(\angle(\hat{v},v)).$$

Using Davis-Khan theorem, we can get bound on this, on which we will be continuing next time.

References

[SS90] STEWART, G.W. and SUN, J., "Matrix Perturbation Theory," *Academic Press*, 1990.

[Bh13] BHATIA, R., "Matrix analysis," *Springer Science & Business Media*, 2013, Vol.169.

[YWS14] Yu, Y., Wang, T. and Samworth, R.J., "A useful variant of the DavisKahan theorem for statisticians," *Biometrika*, 2014, 102(2), pp.315-323.

[JL09] Johnstone, Iain M. and Arthur Yu Lu, "On consistency and sparsity for principal components analysis in high dimensions." *Journal of the American Statistical Association* 104, no. 486 (2009): 682-693.