Mathematics for Data Sciences **Sep 13th, 2011**

Lecture 3. Random Matrix Theory and PCA

Instructor: Yuan Yao, Peking University *Scribe: Tengyuan Liang; Bowei Yan*

1 Principal Component Analysis

In this first part, we will show how to get a principle component of high dimension data.

Let $X = [X_1|X_2|\cdots|X_n] \in \mathbb{R}^{p*n}$, $\hat{\Sigma}_n = \frac{1}{n}XX^T$ is the sample variance. To calculate the top k principal component, we need to

$$\min_{\beta,\mu,U} I := \sum_{i=1}^{n} \|x_i - (\mu + U\beta_i)\|^2 \tag{1}$$

where $U \in \mathbb{R}^{p*k}, UU^T = I_p, e^T U = 0$.

$$\frac{\partial I}{\partial \mu} = -2\sum_{i=1}^{n}(X_i - \mu - U\beta_i) = 0 \Rightarrow \hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i)$$

$$\frac{\partial I}{\partial \beta_i} = (X_i - \mu - U\beta_i)^T U = 0 \Rightarrow \beta_i = U^T(X_i - \mu)$$

Plug in the expression of $\hat{\mu}_n$ and β_i

$$\begin{aligned} I &= \sum_{i=1}^{n} \|X_i - \hat{\mu}_n - UU^T(X_i - \hat{\mu}_n)\|^2 \\ &= \sum_{i=1}^{n} \|X_i - \hat{\mu}_n - P_k(X_i - \hat{\mu}_n)\|^2 \\ &= \sum_{i=1}^{n} \|y_i - P_k(y_i)\|^2 \end{aligned}$$

where P_k is a projection operator.

So the original problem turns into

$$\begin{aligned} \min \sum_{i=1}^{n} \|y_i - P_k(y_i)\|^2 &= \min[tr(Y - P_kY)^T(Y - P_kY)] \\ &= \min[tr(I - P_k)Y^TY(I - P_k)] \\ &= \min tr[Y^TY(I - P_k)^2] \\ &= \min tr[Y^TY(I - P_k)] \\ &= \min[tr(Y^TY) - tr(Y^TYUU^T)] \\ &= \min[tr(Y^TY) - tr(U^TY^TYU)] \end{aligned}$$

Above we use $tr(AB) = tr(BA)$ and $P_k^2 = P_k$.

Since Y is considered as a constant, the problem above is equivalent to

$$\max_{UU^T=I_k} Var(U^TY) = \max_{UU^T=I_k} E[tr(U^TY^TYU)] = \max_{UU^T=I_k} tr(U^T\Sigma U) \tag{2}$$

Here we conclude that the principal component analysis is equivalent to find a k-affine space to approximate the space span by $\{X_i\}$.

2 Marcenko-Pastur Distribution of Random Matrix

Let $\mathrm{X} \in \mathbb{R}^{p*n}$, $X_i \sim \mathcal{N}(0, I_p)$.

When p fixed and $n \to \infty$

$$\widehat{\Sigma}_n = \frac{1}{n}\mathrm{XX}' \to \mathrm{I_p} \tag{3}$$

But when $\frac{p}{n} \to \gamma \neq 0$, the distribution of the eigenvalues of $\widehat{\Sigma}_n$ follows, if $\gamma \leq 1$,

$$\mu^{MP}(t) = \begin{cases} 0 & t \notin [a,b] \\ \frac{\sqrt{(b-t)(t-a)}}{2\pi\gamma t}dt & t \in [a,b] \end{cases} \tag{4}$$

and has an additional point mass $1 - 1/\gamma$ at the origin if $\gamma > 1$. Note that $a = (1-\sqrt{\gamma})^2, b = (1+\sqrt{\gamma})^2$.

3 Rank-1 Signal Model

Suppose $Y = X + \varepsilon$, where $X = \alpha u$.

In addition, $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathrm{I_p})$, $\alpha \sim \mathcal{N}(0, \sigma_X^2)$, so $Y \sim \mathcal{N}(0, \sigma_X^2 uu' + \sigma_\varepsilon^2 I_p)$.

Define $SNR = \frac{\sigma_X^2}{\sigma_\varepsilon^2}$, we aim to show how SNR affect the result of PCA when p is large. Let $Y = [Y_1, Y_2, \cdots, Y_n] \in \mathbb{R}^{p\times n}$

For simplicity, denote $\Sigma = \sigma_X^2 uu' + \sigma_\varepsilon^2 I_p$, then $Y_i = \Sigma^{\frac{1}{2}} Z_i$, where $Z_i \sim \mathcal{N}(0, I_p)$.

$$\widehat{\Sigma}_n = \frac{1}{n}\mathrm{YY}' = \Sigma^{\frac{1}{2}} \cdot \frac{1}{n}\mathrm{ZZ}' \cdot \Sigma^{\frac{1}{2}} = \Sigma^{\frac{1}{2}} \cdot S_n \cdot \Sigma^{\frac{1}{2}} \tag{5}$$

$$S_n = \frac{1}{n}\mathrm{ZZ}' \sim MP distribution \tag{6}$$

We can do a Similarity Transformation to matrix $\Sigma^{\frac{1}{2}} \cdot S_n \cdot \Sigma^{\frac{1}{2}}$, that is $\Sigma^{-\frac{1}{2}} \cdot \Sigma^{\frac{1}{2}} S_n \Sigma^{\frac{1}{2}} \cdot \Sigma^{\frac{1}{2}}$, this new matrix is $S_n \cdot \Sigma$, which has the same eigenvalues as the $\Sigma^{\frac{1}{2}} \cdot S_n \cdot \Sigma^{\frac{1}{2}}$.

Then we focus on $S_n \cdot \Sigma$, suppose one of its eigenvalue is λ and the corresponding eigenvector is v, then:

$$S_n \Sigma v = \lambda v, \tag{7}$$

Notice that

$$\hat{\Sigma}_n \hat{v} = \lambda \hat{v},$$

where

$$c\hat{v} = \Sigma^{1/2} v, \quad c^2 = v'\Sigma v = \sigma_X^2 (u'v)^2 + \sigma_\epsilon^2. \tag{8}$$

Plug in the expression of Σ

$$S_n(\sigma_X^2 uu' + \sigma_\varepsilon^2 I_p)v = \lambda v \tag{9}$$

Rearrange the term with u to one side, we got

$$(\lambda I_p - \sigma_\varepsilon^2 S_n)v = \sigma_X^2 S_n uu'v \tag{10}$$

Assuming that $\lambda I_p - \sigma_\varepsilon^2 S_n$ is reversible, then multiple its reversion at both sides of the equality, we get,

$$v = \sigma_X^2 \cdot (\lambda I_p - \sigma_\varepsilon^2 S_n)^{-1} \cdot S_n u(u'v) \tag{11}$$

Multiply by u' at both side,

$$u'v = \sigma_X^2 \cdot u'(\lambda I_p - \sigma_\varepsilon^2 S_n)^{-1} S_n u \cdot (u'v) \tag{12}$$

that is, if $u'v \neq 0$,

$$1 = \sigma_X^2 \cdot u'(\lambda I_p - \sigma_\varepsilon^2 S_n)^{-1} S_n u \tag{13}$$

For SVD $S_n = W\Lambda W'$, where Λ is diagonal, $W \cdot W' = W' \cdot W = I_p$, $W = [W_1, W_2, \cdots, W_n] \in \mathbb{R}^{p\times p}, \alpha = [\alpha_1, \alpha_2, \cdots, \alpha_n] \in \mathbb{R}^{p\times 1}$, in which W_i is the corresponding eigenvector, then $u = \sum_{i=1}^p \alpha_i W_i = W \cdot \alpha$, then, $\alpha = W'u$, and,

$$1 = \sigma_X^2 \cdot u'[W(\lambda I_p - \sigma_\varepsilon^2 \Lambda)^{-1} W'][W\Lambda W']u = \sigma_X^2 \cdot (u'W)(\lambda I_p - \sigma_\varepsilon^2 \Lambda)^{-1} \Lambda (W'u) \tag{14}$$

Replace $W'u = \alpha$, then,

$$1 = \sigma_X^2 \cdot \sum_{i=1}^p \frac{\lambda_i}{\lambda - \sigma_\varepsilon^2 \lambda_i} \alpha_i^2 \tag{15}$$

where $\sum_{i=1}^p \alpha_i^2 = 1$. Since W is a random orthogonal basis on a sphere, α_i will concentrate on its mean $\alpha_i = \frac{1}{\sqrt{q}}$. According to the fact that p is large enough($\sim \infty$), due to Law of Large Numbers(LLN) and $\lambda \sim \mu^{MP}$(λ_i can be thought sampled from the μ^{MP}), the equation (12) can be thought of as the Expected Value (Monte-Carlo Integration), then equation (12) can be written as,

$$1 = \sigma_X^2 \cdot \frac{1}{p} \sum_{i=1}^p \frac{\lambda_i}{\lambda - \sigma_\varepsilon^2 \lambda_i} \sim \sigma_X^2 \cdot \int_a^b \frac{t}{\lambda - \sigma_\varepsilon^2 t} d\mu^{MP}(t) \tag{16}$$

For convenience, assume without loss of generosity that $\sigma_\varepsilon^2 = 1$, that is the noise volatility is 1. Now we unveil the story of the ratio γ, do the integration in equation (13), we got,

$$1 = \sigma_X^2 \cdot \int_a^b \frac{t}{\lambda - t} \frac{\sqrt{(b-t)(t-a)}}{2\pi\gamma t} dt = \sigma_X^2 \cdot \frac{1}{4\gamma}[2\lambda - (a+b) - 2\sqrt{|(\lambda-a)(b-\lambda)|}] \tag{17}$$

3.1 About the Eigenvalue: Phase-Transition

- If $\lambda \in [a, b]$, then $\widehat{\Sigma}_n$ has eigenvalue λ within $\text{supp}(\mu^{MP})$, so it is undistinguishable from the noise S_n.
- If $\lambda \geq b$, PCA will pick up the top eigenvalue as non-noise. So $\lambda = b$ is the phase transition where PCA works to pop up correct eigenvalue. Then plug in $\lambda = b$ in equation (14), we get,

$$1 = \sigma_X^2 \cdot \frac{1}{4\gamma}[2b - (a+b)] = \frac{\sigma_X^2}{\sqrt{\gamma}} \Leftrightarrow \sigma_X^2 = \sqrt{\frac{p}{n}} \tag{18}$$

  So, in order to make PCA works, we need to let $SNR \geq \sqrt{\frac{p}{n}}$

3.2 Eigenvector

We know that if PCA works good and noise doesn't dominate the effect, the innerproduct $|u'\hat{v}|$ should be close to 1. On the other hand, from RMT we know that if the top eigenvalue λ is merged in the M. P. distribution, then the top eigenvector computed is purely random and $|u'\hat{v}| = 0$, which means that from $\hat{v}$ we can know nothing about the signal u. We now study the phase transition of top-eigenvector.

It is convenient to study $|u'v|^2$ first and then translate back to $|u'\hat{v}|^2$. Using the equation (8),

$$1 = |v'v| = \sigma_X^4 \cdot v'uu'S_n(\lambda I_p - \sigma_\varepsilon^2 S_n)^{-2} S_n uu'v = \sigma_X^4 \cdot (|v'u|)[u'S_n(\lambda I_p - \sigma_\varepsilon^2 S_n)^{-2} S_n u](|u'v|) \tag{19}$$

$$|u'v|^{-2} = \sigma_X^4 [u'S_n(\lambda I_p - \sigma_\varepsilon^2 S_n)^{-2} S_n u] \tag{20}$$

Using the same trick as the equation (11),

$$|u'v|^{-2} = \sigma_X^4 [u'S_n(\lambda I_p - \sigma_\varepsilon^2 S_n)^{-2} S_n u] \sim \sigma_X^4 \cdot \int_a^b \frac{t^2}{(\lambda - \sigma_\varepsilon^2 t)^2} d\mu^{MP}(t) \tag{21}$$

and assume that $\lambda > b$,

$$|u'v|^{-2} = \sigma_X^4 \cdot \int_a^b \frac{t^2}{(\lambda - \sigma_\varepsilon^2 t)^2} d\mu^{MP}(t) = \frac{\sigma_X^4}{4\gamma}(-4\lambda + (a+b) + 2\sqrt{(\lambda-a)(\lambda-b)} + \frac{\lambda(2\lambda - (a+b))}{\sqrt{(\lambda-a)(\lambda-b)}} \tag{22}$$

from which it can be computed that (using $\lambda = (1+R)(1+\frac{\gamma}{R})$ which is computed above, where $R = SNR = \frac{\sigma_X^2}{\sigma_\varepsilon^2}$)

$$|u'v|^2 = \frac{1 - \frac{\gamma}{R^2}}{1 + \gamma + \frac{2\gamma}{R}}.$$

Using the relation

$$u'\hat{v} = u'\left(\frac{1}{c}\Sigma^{1/2} v\right) = \frac{\sqrt{1+R}}{c}(u'v)$$

where the second equality uses $\Sigma^{1/2}u = \sqrt{1+R}u$, and with the formula for c^2 above, we can compute

$$(u'\hat{v})^2 = \frac{1+R}{1+R(u'v)^2}(u'v)^2$$

in terms of R. Note that this number holds under the condition that $R > \sqrt{\gamma}$.

To sum up, according to [**Johnstone06**] or [**Nadakuditi10**], For eigenvalue,

$$\lambda_{max}(\widehat{\Sigma}_n) \to \begin{cases} (1+\sqrt{\gamma})^2 = b & \sigma_X^2, \leq \sqrt{\gamma} \\ (1+\sigma_X^2)(1+\frac{\gamma}{\sigma_X^2}), & \sigma_X^2 > \sqrt{\gamma} \end{cases} \tag{23}$$

which implies that if signal energy is small, top eigenvalue of sample covariance matrix never pops up from random matrix ones; only if the signal energy is beyond the phase transition threshold $\sqrt{\gamma}$, top eigenvalue can be separated from random matrix eigenvalues. However, even in the latter case it is a biased estimation.

For eigenvector,

$$|\langle u, v_{max}\rangle|^2 \to \begin{cases} 0 & \sigma_X^2, \leq \sqrt{\gamma} \\ \frac{1-\frac{\gamma}{\sigma_X^4}}{1+\frac{\gamma}{\sigma_X^2}}, & \sigma_X^2 > \sqrt{\gamma} \end{cases} \tag{24}$$

which means the same phase transition phenomenon: if signal is of low energy, PCA will tell us nothing about the true signal and the estimated top eigenvector is orthogonal to the true direction u; if the signal is of high energy, PCA will return a biased estimation which lies in a cone whose angle with the true signal is bounded by

$$\frac{1-\frac{\gamma}{\sigma_X^4}}{1+\frac{\gamma}{\sigma_X^2}}.$$

3.3 Further Results

When $\frac{log(p)}{n} \to 0$, we need to add more restrictions on $\widehat{\Sigma}_n$ in order to estimate it faithfully. There are typically three kinds of restrictions.

- Σ sparse
- Σ^{-1} sparse, also called–Precision Matrix
- banded structures (e.g. Toeplitz) on Σ or Σ^{-1}

Recent developments can be found by Bickel, Tony Cai, Tsybakov, Wainwright et al.