Math 690: Topics in Data Analysis and Computation

Lecture notes for Fall 2017

9.5 and 9.7

Scribed by Yixin Lin, Shen Yan

September 5

Last time we saw PCA from two perspectives: linear algebra, and probability (population covariance matrix).

Why can PCA be solved by eigenproblem? Recall we want to

$$\max_{w_1,\cdots,w_d,w_k^T w_l=\delta_{kl}} \sum_k w_k^T S w_k$$

Solved by the eigendecomposition of the matrix S.

Courant-Fischer Minimax Theorem For any Hermitian or real-symmetric matrix, we know it has n real eigenvalues with eigenvectors forming a orthonormal basis.

Let A be an $n \times n$ Hermition matrix. It has n eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$.

$$\lambda_k(A) = \sup_{\dim(V)=k} \inf_{\|v\|=1, v\in V} v^* A v$$

$$\lambda_k(A) = \inf_{\dim(V)=n-k+1} \sup_{\|v\|=1, v\in V v^* Av} v^* A v$$

Proof: $A = U\Lambda U^*$, verify for each k.

(Ex) We mentioned that PCA can be viewed as maximizing the projected variation, which is equivalent to minimizing the residual after projection.

$$P_{w_k} = w_k w_k^T \text{ is the projection matrix}$$

$$\min_{w_1,\cdots,w_d,w_k^T w_l=\delta_{kl}} \sum_{k=1}^d \sum_{i=1}^n \|x_i - P_{w_k} w_i\|^2$$

Hint for exercise: $\|(I - ww^T)x_i\|^2$

Population covariance matrix $\Sigma = \mathbb{E}x_i x_i^T = ?$, can't compute the integral. Approximate it with $S = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$.

Covariance estimation Given $\{x_i\}_{i=1}^n \sim^{\text{iid}} P_x$ in $\mathbb{R}^p$. Our goal is to estimate

$$\mu = \mathbb{E}x_i, \Sigma = \mathbb{E}(x_i - \mu)(x_i - \mu)^T$$

as the sample mean and covariance

$$\hat{\mu} = \frac{1}{n}\sum_i x_i, \hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^n (x_i - \hat{\mu})(x_i - \hat{\mu})^T$$

Why do we use these as estimators? They are unbiased, but other statistics may be unbiased as well. The reason is that these are max-likelihood estimators(MLE) when the data is distributed as Gaussian. Note that $\frac{1}{n-1}\sum_{i=1}^n (x_i-\hat{\mu})(x_i-\hat{\mu})^T$ is the unbiased estimator, while $\frac{1}{n}\sum_{i=1}^n (x_i-\hat{\mu})(x_i-\hat{\mu})^T$ is the MLE.

Proof sketch: suppose $x_i \sim N(\mu, \Sigma)$. Then

$$p(x_i) = \frac{\exp[-(x_i-\mu)^T\Sigma^{-1}(y_i-\mu)/2]}{(2\pi)^{p/2}|\Sigma|^{1/2}}$$

$$\log p(x_i|\mu,\Sigma) = -\frac{(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)}{2} - \frac{1}{2}\log|\Sigma| + c$$

$$\log p(\{x_i\}_{i=1}^n|\mu,\Sigma) = \log\prod_{i=1}^n p(x_i|\mu,\Sigma)$$

$$= \sum_{i=1}^n \Big\{ -\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) - \frac{1}{2}\log|\Sigma| + c\Big\}$$

$$= nc - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^n (x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

$\max_\mu \Rightarrow \hat{\mu}^{\text{MLE}} =$ sample mean.

$$= nc - \frac{n}{2}\log|\Sigma| - \frac{n}{2}\text{Tr}(\Sigma^{-1}S) \text{ where } S = \frac{1}{n}\sum_{i=1}^n (x_i-\hat{\mu})(x_i-\hat{\mu})^T$$

$$= |\Sigma| - \frac{n}{2}\text{Tr}(\Sigma^{-1}S)$$

$$\max_\Sigma -\frac{1}{2}\log|\Sigma| - \frac{1}{2}\text{Tr}(\Sigma^{-1}S)$$

$$\max_\Sigma \frac{1}{2}\log|\Sigma^{-1}| - \frac{1}{2}\text{Tr}(\Sigma^{-1}S)$$

$$\max_\Sigma c + \frac{1}{2}\log|\Sigma^{-1}S| - \frac{1}{2}\text{Tr}(\Sigma^{-1}S)$$

Let $B = \Sigma^{-1}S$. $\log|B| - \text{Tr}(B) = \sum_{i=1}^p \log(x_i) - \lambda i$ can be purely written in the eigenvalues of B, which results in $\lambda_i = 1$ means that $B = I$ (is identity). Exercise.

Covariance estimation: asymptotic consistency How well does $S \approx E$ as $\lim_n \to \infty$. (Assume $\mu = 0$.)

$$S = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$$

1. Law of large numbers: $S \to^{n\to+\infty} \mathbb{E}S = \Sigma$

2. rate $\sim n^{-1/2}, S_{p\times p}$

$$\mathbb{E}|S_{kl} - \Sigma_{kl}|^2 \leq \frac{c}{n}$$

This result comes from an element-wise LLN:

$$S_{kl} = \frac{1}{n}\sum_{i=1}^{n} x_i(k)x_i(l) \to^{\text{dist}} N(0, \frac{c}{n})$$

Using Frobenius norm: $\|A\|_{\text{Fro}} = (\sum_{i,j} A_{ij}^2)^{1/2}$

$$\mathbb{E}\|S - \Sigma\|_{\text{Fr}} = \mathbb{E}\sum_{k,l} |S_{kl} - \Sigma_{kl}|^2 \leq \frac{c}{n} \cdot p^2$$

What if p is large (e.g. $p \approx n$ or $p \gg n$)? This may be large...

However, if the p is large but the true covariance matrix Σ has low rank, then there may be no curse of dimensionality. For example, consider $\Sigma = uu^T$, where $u = (1, 0, 0, \cdots, 0)$, $x_i = \alpha_i u$, $\alpha_i = N(0, 1)$. In this case, $\mathbb{E}x_i x_i^T = uu^T$, $S = (\frac{1}{n}\sum \alpha_i^2)uu^T$, by LLN we always get the same convergence rate regardless of p.

Noisy PCA: What we observe is that the noisy patches $y_i = x_i + z_i$ for some $x_i \sim P_x$, clean patches, and $z_i \sim N(0, \sigma^2 I)$

Goal: $\Sigma_x = \mathbb{E}x_i x_i^T$. Estimate Σ_x or the principle components of Σ_x.

$$S_y = \frac{1}{n}\sum_{i=1}^{n} y_i y_i^T$$

$$\mathbb{E}S_y = \Sigma_y = \Sigma_x + \alpha^2 I_p, \text{ exercise}$$

If we're in the classical case, where p is fixed, $\hat{\Sigma}_y, \hat{\sigma} \Rightarrow \hat{\sigma}_x = \hat{\sigma}_y - \sigma^2 I$. Inconsistency of $p \approx n, p \gg n$.

September 7

Setting up:

$$y_i = X_i + z_i \text{ (i.e. info + noise)}$$

$$z_i \sim N(0, \sigma^2 I), x_i \sim P_x$$

$$\Sigma_y = \Sigma_x + \sigma^2 I$$

$$\Sigma_x \to \Sigma_y$$

In this case, eigenvectors of the 2 matrices are the same, and eigenvalues only differ by σ^2

$$S_y = \frac{1}{n}\sum_i (x_i + z_i)(x_i + z_i)^T = S_x \text{``+''} S_z$$

In this case though, top few eigenvectors are only 'consistent", i.e. having positive correlation; and eigenvalues increase.

In the class we have a demo showing the eigenvalue distribution of S_y, where $x_i = \alpha_i u, \alpha_i \sim N(0, 1)$, and z_i is also normal.

Null case: consider only z

$$S = \frac{1}{n}\sum_i z_i z_i^T, z_i \sim N(0, I_p)$$

Question: what is eig(S) like?

(White) Marcenko-Pastur Law '67: $P, n \to +\infty, P/n \to \gamma > 0$. Distribution of eigenvalues of S converges to limiting density:
When $\gamma \leq 1$

$$p_{MP}(t) = \frac{\sqrt{(t-a)(b-t)}}{2\pi\gamma t}, a < t < b, a = (1-\sqrt{\gamma})^2, b = (1+\sqrt{\gamma})^2$$

When $\gamma > 1$

$$p_{MP}(t) + (1 - \frac{1}{\gamma})\delta_0(t)$$

This is because when $P > n$, at least $P - n$ of the eigenvalues are 0.

The mathematical formulation:

Definition: The empirical spectral density (ESD) =

$$\frac{1}{p}\sum_{i=1}^{p}\delta_{\lambda_i}$$

for eigenvalues of S $\lambda_1, \cdots, \lambda_p$

The convergence of $ESD_S(S) \to^{weakly} p_{MP}(t)$ almost surely

Remark: when $\gamma \to 0^+$, this density converges to δ_1 (population case). when $\gamma \to +\infty$, this density will be roughly a semicircle centered at γ

Remark: Convergence to limit density is fast

$$p, \sim 10^2 \text{ , the approximation is OK}$$

Remark: "White MP"

$$z_i \sim N(0, I_p)$$

"Colored MP"

$$\text{when } z_i \sim N(0, \Sigma), \Sigma \neq I$$

For example, if Σ has 2 eigenvalues d_1, d_2, the the limiting distribution of eig(S) when $P/n \to 0$ is like $\frac{1}{2}d_1 + \frac{1}{2}d_2$.

Proof of Marcenko-Pasture Density

Recall first the proof of CLT, we prove that the characteristic function of $\frac{1}{n}\sum_i^n x_i$ converges to that of normal distribution.

$$\mathbb{E}_{x\sim p}e^{i\xi x} = \phi(\xi)$$

$$\mathbb{E}\exp[i\xi\frac{1}{\sqrt{n}}(x_1 + \cdots + x_n)] = \prod_i(\mathbb{E}\exp[i\xi\frac{x_i}{n})] = \text{ Taylor series approximation } \to \phi(\xi)_{\text{Gaussian}}$$

Similarly here, we consider the Stieltjes Transform of μ: $p(t)$ probabilistic density $d\mu(t) = p(t)dt$, $m_\mu(z) = \int_{-\infty}^{+\infty}\frac{1}{t-z}d\mu(t), I_m(z) > 0$, and we hope to prove the convergence in the sense of Stieltjes Transform.

Lemma 1: A sequence of probability measures $\mu_n \to \mu$ if and only if $m_n(z) \to m(z), \forall\xi, Im(z) >$

0, $m_n(\xi)$ is the Stieltjes Transform of μ_n.
Lemma 2: MP Equation: consider $m(\xi)$, the Stieltjes Transform of the MP density. Then

$$z+\frac{1}{m(z)}=\frac{1}{1+\gamma m(z)} \quad (*)$$

Proof of Lemma 2:
Inversion formula for Stieltjes Transform

$$\lim_{b\to 0+} Im(m_\mu(t+ib))=\pi\cdot p(t)$$

The solution of (*)

$$m(z)=\frac{-(z+\gamma-1)\pm\sqrt{\cdots}}{2z\delta}$$

Verify that $Im(m(t+ib))\to\pi\cdot p_{MP}(t)$
Back to the main theorem:
It suffices to show that $m_n(z)$ on the limit of $n,p\to\infty,\cdots$ satisfies (*):

$$m_n(z)=\int_{\mathbb{R}}\frac{1}{t-z}ESD_s(t)dt$$

$$\frac{1}{p}\sum_{i=1}^{p}\frac{1}{\lambda_i-z}=\frac{1}{p}Tr(S-zI)^{-1}$$

Recall that $S=\frac{1}{n}\sum_i z_iz_i^T=\sum_{i=1}^n x_ix_i^T, x_i=\frac{z_i}{\sqrt{n}}$, somehow the notation changed from S to B, but they are the same thing.
Identity:

$$I+z(B-zI)^{-1}=B(B-zI)^{-1}$$

"$I+\frac{z}{B-z}=\frac{B}{B-z_i}$"
Apply $\frac{1}{p}Tr(\cdot)$ on both sides:

$$1+z\frac{1}{p}Tr(B-z)^{-1}=\frac{1}{p}Tr(B(B-z)^{-1})$$

$$=\frac{1}{p}\sum_i Tr(x_ix_i^T(B-zI)^{-1})=\frac{1}{p}\sum_i x_i^T(B-zI)^{-1}x_i$$

Consider

$$x_i^T(B-zI)^{-1}x_i=x_i^T(x_ix_i^T+B_{(i)}-zI)^{-1}x_i$$

$$B_{(i)}=\sum_{j\neq i}x_jx_j^T$$

Using Sherman-Morrison formula, where both A and $A+pq^T$ are invertible:

$$(A+pq^T)^{-1}=A^{-1}-\frac{A^{-1}pq^TA^{-1}}{1+q^TA^{-1}p},$$

we have that (verify this)

$$x_i^T(x_ix_i^T+B_{(i)}-zI)^{-1}x_i=\frac{x_i^T(B_{(i)}-zI)^{-1}x_i}{1+x_i^T(B_{(i)}-zI)^{-1}x_i}.$$

Now use two facts which will not be proved here: firstly, the random variable

$$x_i^T(B_{(i)} - zI)^{-1}x_i \sim \mathbb{E}_{x_i} x_i^T(B_{(i)} - zI)^{-1}x_i = \frac{1}{n}\mathbf{Tr}(B_{(i)} - zI)^{-1},$$

where $\mathbb{E}_{x_i}$ means taking expectation over x_i only, and the second equality is by the definition of x_i; secondly,

$$\frac{1}{n}\mathbf{Tr}(B_{(i)} - zI)^{-1} = \gamma\frac{1}{p}\mathbf{Tr}(B_{(i)} - zI)^{-1} \sim \gamma m_n(z)$$

where the second approximation is by that the quantity $\frac{1}{p}\mathbf{Tr}(B_{(i)} - zI)^{-1}$ is the Stieltjes transform of the spectral density of $B_{(i)}$, which turns out to be close to the Stieltjes transform of the spectral density of B, namely $m_n(z)$, thanks to the stability of Stieltjes transform under rank-1 perturbation of the matrix. Putting together, the r.h.s becomes

$$\sim \frac{1}{p}n\frac{\gamma m_n(z)}{1 + \gamma m_n(z)} = \frac{m_n(z)}{1 + \gamma m_n(z)},$$

and this proves that $m_n(z)$ asymptotically satisfies the equation

$$1 + z m_n(z) = \frac{m_n(z)}{1 + \gamma m_n(z)}.$$

To learn more about random matrix theory: the book by Terry Tao on random matrices. The book by Anderson, Guionnet and Zeitouni, *An introductin to Random Matrices* ("cupbook.pdf"). The book by Tao is more accessible than the later.