Concentração de matrizes aleatórias e suas aplicações em estatísticas

September 2024

Sumário

1 Introdução / Objetivos

1) Estudar o capítulo 4 do Livro do Roman Vershynin, "High-Dimensional Probability". Mais precisamente, as seções 4.1, 4.2 e 4.4 (a base necessária para as aplicações) e em seguida as seções 4.5 (aplicação no problema de detecção de comunidades) e 4.7 (aplicação no problema de estimação da matriz de covariância e clusterização).

Link para o livro: `https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.html#`

2) Estudar o artigo Semidefinite Programs on Sparse Random Graphs and their Application to Community Detection

Link para o artigo: `https://web.stanford.edu/~montanar/SDPgraph/sdp.pdf`

2*) Estudar o artigo Clustering in Block Markov Chains

Link para o artigo `https://projecteuclid.org/journals/annals-of-statistics/volume-48/issue-6/Clustering-in-Block-Markov-Chains/10.1214/19-AOS1939.full`

3) Voltar no texto e ir preenchendo lacunas que faltaram

4) A definir

2 Preliminares sobre matrizes (4. 1)

2.1 Decomposição em valores singulares (4. 1. 1)

Def. (Decomposição em valores singulares): Para $A \in \mathbb{R}^{m\times n}$ com $p = \text{posto}(A)$, a decomposição de A em $\sum_{i=1}^{p} s_i u_i v_i^T$, com $v_i \in \mathbb{R}^n$ e $u_i \in \mathbb{R}^m$, em que v_i e u_i são, respectivamente, os autovetores ortonormais de A^TA e AA^T e s_i é a raiz quadrada do autovalor λ_i de A^TA e AA^T.

Teo. 1 (Min-max de Courant-Fisher): Se A é uma matriz simétrica e $\lambda_1 \geq ... \geq \lambda_k \geq ... \geq \lambda_n$ são autovalores de A, então

$$\lambda_k = \max_{\dim(E)=k} \min_{x\in S(E)} \langle Ax, x\rangle$$

em que E é um subespaço k-dimensional de $\mathbb{R}^n$ e $S(E) = \{x \in E : \|x\| = 1\}$.

Exe. 1: Se A é uma matriz invertível que tem decomposição em valores singulares, então

$$A^{-1} = \sum_{i=1}^{n} s_i^{-1} v_i u_i^T$$

Primeiro, notamos que

$$A = \sum_{i=1}^{n} s_i u_i v_i^T = \begin{bmatrix} u_1 & \dots & u_n \end{bmatrix} \begin{bmatrix} s_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & s_n \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = U\Sigma V^T$$

Analogamente,

$$\sum_{i=1}^{n} s_i^{-1} v_i u_i^T = \begin{bmatrix} v_1 & \dots & v_n \end{bmatrix} \begin{bmatrix} s_1^{-1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & s_n^{-1} \end{bmatrix} \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} = V\Sigma^{-1}U^T$$

Mas, uma vez que U e V são matrizes ortogonais, as suas transpostas coincidem

com as suas inversas. Logo,

$$\begin{aligned} A\Big(\sum_{i=1}^{n} s_i^{-1} v_i u_i^T\Big) &= \Big(\sum_{i=1}^{n} s_i u_i v_i^T\Big)\Big(\sum_{i=1}^{n} s_i^{-1} v_i u_i^T\Big) \\ &= U\Sigma V^T V \Sigma^{-1} U^T \\ &= U\Sigma V^{-1} V \Sigma^{-1} U^{-1} \\ &= U\Sigma\Sigma^{-1} U^{-1} \\ &= UU^{-1} \\ &= I \end{aligned}$$

2.2 Norma de operador e valores singulares extremos (4. 1. 2)

Def. (Norma de operador): Se l_2^n denotar o espaço de Hilbert obtido ao munirmos $\mathbb{R}^n$ com a norma euclidiana $\|\cdot\|_2$, então a matriz $A_{m\times n}$ é interpretada como um operador de l_2^n em l_2^m. Logo,

$$\|A\| = \max_{x\in\mathbb{R}^n-\{0\}} \frac{\|Ax\|_2}{\|x\|_2} = \max_{x\in\mathbb{R}^n-\{0\}} \left\|A\Big(\frac{x}{\|x\|_2}\Big)\right\|_2 = \max_{x\in S(\mathbb{R}^n)} \|Ax\|_2$$

Além disso, para $y \in S(\mathbb{R}^m)$, pela desigualdade de Cauchy-Schwarz, $\langle Ax, y\rangle \leq \|Ax\|_2\|y\|_2 = \|Ax\|_2$. Particularmente, se $y^* = \frac{Ax}{\|Ax\|_2}$, então $y^* \in S(\mathbb{R}^m)$ e $\langle Ax, y^*\rangle = \|Ax\|_2$. Assim, Argmax $_{\{y\in S(\mathbb{R}^m)\}}\langle Ax, y\rangle = y^*$. Logo,

$$\begin{aligned} \max_{(x,y)\in S(\mathbb{R}^n)\times S(\mathbb{R}^m)} \langle Ax, y\rangle &= \max_{x\in S(\mathbb{R}^n)} \Big(\max_{y\in S(\mathbb{R}^m)} \langle Ax, y\rangle\Big) \\ &= \max_{x\in S(\mathbb{R}^n)} \Big\langle Ax, \frac{Ax}{\|Ax\|_2}\Big\rangle \\ &= \max_{x\in S(\mathbb{R}^n)} \frac{\langle Ax, Ax\rangle}{\|Ax\|_2} \\ &= \max_{x\in S(\mathbb{R}^n)} \frac{\|Ax\|_2^2}{\|Ax\|_2} \\ &= \max_{x\in S(\mathbb{R}^n)} \|Ax\|_2 \\ &= \|A\| \end{aligned}$$

Também notamos que, se E é um subespaço unidimensional de $\mathbb{R}^n$, então é uma reta. Disso, para $v \in E$ vetor unitário, pela linearidade $Av = -(A(-v))$. Logo, min $_{\{x\in S(E)\}}\|Ax\| = \|Av\|_2 = \|A(-v)\|_2$. Assim, pelo Teo. 1,

$$s_1 = \max_{\dim(E)=1} \min_{x\in S(E)} \|Ax\|_2 = \max_{x\in S(\mathbb{R}^n)} \|Ax\|_2 = \|A\|$$

Percebemos, por fim, que $s_n > 0$ somente se $m > n$. Nesse caso, $p = n$. Além disso, s_n é uma medida da não-degeneração de A. Assim, $s_n = \|A^+\|^{-1}$, em que A^+

é a pseudoinversa de Moore-Penrose de A e $\|A^+\|$ é a norma do operador de A^{-1} restrita à imagem de A.[1]

2.3 Norma de Frobenius (4. 1. 3)

Def. (Norma de Frobenius): Para $A \in \mathbb{R}^{m\times n}$,

$$\|A\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2 \right)^{1/2}$$

Além disso, uma vez que $\langle A, A\rangle = \mathrm{tr}(A^T A)$,

$$\begin{aligned}
\|A\|_F &= \left(\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2 \right)^{1/2} \\
&= \left(\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} A_{ij} \right)^{1/2} \\
&= (\mathrm{tr}(A^T A))^{1/2} \\
&= (\langle A, A\rangle)^{1/2}
\end{aligned}$$

Disso, dado que as matrizes U e V são ortogonais e que $\mathrm{tr}(AB) = \mathrm{tr}(BA)$,

$$\begin{aligned}
\|A\|_F &= (\mathrm{tr}(A^T A))^{1/2} \\
&= (\mathrm{tr}((U\Sigma V^T)^T U\Sigma V^T))^{1/2} \\
&= (\mathrm{tr}(U\Sigma V^T (U\Sigma V^T)^T))^{1/2} \\
&= (\mathrm{tr}(U\Sigma V^T V\Sigma^T U^T))^{1/2} \\
&= (\mathrm{tr}(U\Sigma V^{-1} V\Sigma^T U^{-1}))^{1/2} \\
&= (\mathrm{tr}(U\Sigma\Sigma^T U^{-1}))^{1/2} \\
&= (\mathrm{tr}((U\Sigma)(\Sigma^T U^{-1})))^{1/2} \\
&= (\mathrm{tr}((\Sigma^T U^{-1})(U\Sigma)))^{1/2} \\
&= (\mathrm{tr}(\Sigma^T U^{-1} U\Sigma))^{1/2} \\
&= (\mathrm{tr}(\Sigma^T \Sigma))^{1/2} \\
&= \left(\sum_{i=1}^{p} s_i^2 \right)^{1/2}
\end{aligned}$$

Por fim, se $s = (s_1, ..., s_p)$ é o vetor dos valores singulares de A, então $\|A\|_F = (\sum_{i=1}^{p} s_i^2)^{1/2} = \|s\|_2$ e $\|A\|_2 = s_1 = \max\{s_1, ..., s_p\} = \|s\|_\infty$. Além disso, as normas são equivalentes. De fato, dado que $s_1 \geq ... \geq s_p$,

$$\|A\| = s_1 \leq (s_1^2)^{1/2} \leq \left(\sum_{i=1}^{p} s_i^2 \right)^{1/2} = \|A\|_F$$

[1] Se $A \in \mathbb{R}^{m\times n}$, então a pseudoinversa de Moore-Penrose é $A^+ \in \mathbb{R}^{n\times m}$ tal que $AA^+A = A$, $A^+AA^+ = A^+$, $(AA^+)^T = AA^+$ e $(A^+A)^T = A^+A$.

Por outro lado,

$$\|A\|_F = \left(\sum_{i=1}^{p} s_i^2\right)^{1/2} \leq \left(\sum_{i=1}^{p} s_1^2\right)^{1/2} = \sqrt{p}s_1 = \sqrt{p}\|A\|$$

Exe. 2: Se s é vetor dos valores singulares de A, então

$$s_1 = (s_1^2)^{1/2} \leq \left(\sum_{i=1}^{p} s_i^2\right)^{1/2} = \|s\|_2 = \|A\|_F = \frac{\|A\|_F}{1} = \frac{\|A\|_F}{\sqrt{1}}$$

Agora, para $n = 2, ..., p$,

$$\sqrt{n}s_n = (ns_n^2)^{1/2} \leq \left(\sum_{i=1}^{n} s_i^2\right)^{1/2} \leq \left(\sum_{i=1}^{p} s_i^2\right)^{1/2} = \|s\|_2 = \|A\|_F$$

Portanto, para $s_n \in s$,

$$s_n \leq \frac{1}{\sqrt{n}}\|A\|_F$$

2.4 Aproximação de posto menor (4. 1. 4)

Teo. 2 (Eckart-Young-Mirsky): Dada $A \in \mathbb{R}^{m\times n}$, sejam $k < p$ e $\|\cdot\|$ uma norma unitariamente invariante,[2] A_k é dita melhor aproximação de A de posto k se $\text{Argmin}_{\{\text{posto}(A')\leq k\}}\|A - A'\| = A_k = \sum_{i=1}^{k} s_i u_i v_i^T$.

Exe. 3: Para A_k melhor aproximação de A de posto k,

$$\begin{aligned}
\|A - A_k\|^2 &= (\|A - A_k\|)^2 \\
&= \left(\left\| \sum_{i=1}^{p} s_i u_i v_i^T - \sum_{i=1}^{k} s_i u_i v_i^T \right\|\right)^2 \\
&= \left(\left\| \sum_{i=k+1}^{p} s_i u_i v_i^T \right\|\right)^2 \\
&= (\max\{s_{k+1}, ..., s_p\})^2 \\
&= s_{k+1}^2
\end{aligned}$$

Analogamente,

$$\|A - A_k\|_F^2 = \left(\left\| \sum_{i=k+1}^{p} s_i u_i v_i^T \right\|_F\right)^2 = \left(\left(\sum_{i=k+1}^{p} s_i^2\right)^{1/2}\right)^2 = \sum_{i=k+1}^{p} s_i^2$$

[2] Se $\|\cdot\|$ uma norma em $\mathbb{R}^{m\times n}$, então, para $A \in \mathbb{R}^{m\times n}$, $\|UAV\| = \|A\|$ para todas matrizes unitárias $U \in \mathbb{R}^{m\times m}$ e $V \in \mathbb{R}^{n\times n}$, em que matriz unitária significa que, se $A \in \mathbb{R}^{m\times n}$, então $A^TA = I_n$ e $AA^T = I_m$.

2.5 Isometrias aproximadas (4. 1. 5)

Dada $A \in \mathbb{R}^{m\times n}$, os valores singulares extremos s_1 e s_n são, respectivamente, a menor cota superior e a maior inferior tais que, para $x \in \mathbb{R}^n$, $s_n\|x\|_2 \leq \|Ax\|_2 \leq s_1\|x\|_2$. De fato, para $y \in \mathbb{R}^n - \{0\}$ – no caso em que $y = 0$, o resultado é trivial –, seja $y^* = \frac{y}{\|y\|_2}$. Pela definição, $\|Ay\|_2 = \|y\|_2\|Ay^*\|_2 \leq \|y\|_2 \max_{\{x\in S(\mathbb{R}^n)\}}\|Ax\|_2 = \|y\|_2\|A\|$ Mas, pelo Teo. 1, $s_1 = \|A\|$. Logo, $\|Ay\| \leq \|y\|_2 s_1$. Por outro lado, também pelo Teo. 1, $s_n = \max_{\{\dim(E)=n\}} \min_{\{x\in S(E)\}}\|Ax\|_2 = \min_{\{x\in S(\mathbb{R}^n)\}}\|Ax\|_2$. Já, pela definição, $\|y\|_2\|Ay^*\|_2 \geq \|y\|_2 \min_{\{x\in S(\mathbb{R}^n)\}}\|Ax\|_2$. Disso, $\|Ay\|_2 \geq s_n\|y\|_2$. Disso, dado que a relação também é verdadeira para $x - y \in \mathbb{R}^n$, s_1 e s_n podem ser interpretados como fatores extremos de deformação de distâncias que o operador A causa em $\mathbb{R}^n$.

Exe. 4: Seja $A \in \mathbb{R}^{m\times n}$, com $n \leq m$. Supomos que $A^TA = I_n$, então, ao definirmos $P = AA^T \in \mathbb{R}^{m\times m}$, verificamos que é uma projeção, pois $P^2 = AA^TAA^T = A(A^TA)A^T = AI_nA^T = P$. Além disso, também pela hipótese, $P^T = (AA^T)^T = (A^T)^TA^T = AA^T = P$. Disso, P é uma projeção ortogonal em $\mathbb{R}^m$. Por fim, outra vez, pela hipótese,

$$\begin{aligned}
A^TA &= \begin{bmatrix} a_1^T \\ \vdots \\ a_n^T \end{bmatrix} \begin{bmatrix} a_1 & \dots & a_n \end{bmatrix} \\
&= \begin{bmatrix} \langle a_1, a_1\rangle & \dots & \langle a_1, a_n\rangle \\ \vdots & \ddots & \vdots \\ \langle a_n, a_1\rangle & \dots & \langle a_n, a_n\rangle \end{bmatrix} \\
&= \begin{bmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{bmatrix} \\
&= I_n
\end{aligned}$$

Logo, posto que $\langle a_i, a_j\rangle = \delta_{ij}$, as colunas de A são ortonormais. Assim, para $a_i \in \{a_1, ..., a_n\}$, $Pa_i = AA^Ta_i = Ae_i = a_i$. Logo, P projeta, em $\mathbb{R}^m$, um subespaço ortogonal de dimensão n.

Em seguida, supomos que $P = AA^T$ é uma projeção ortogonal em $\mathbb{R}^m$ de um subespaço de dimensão n. Se $x \in \mathbb{R}^n$, então, para alguns $\alpha_1, ..., \alpha_n$, $x = \sum_{i=1}^n \alpha_i e_i$. Logo, pela ortonormalidade da base canônica,

$$\|x\|_2 = \left\| \sum_{i=1}^n \alpha_i e_i \right\|_2 = \sum_{i=1}^n \|\alpha_i e_i\|_2 = \sum_{i=1}^n |\alpha_i|\|e_i\|_2 = \sum_{i=1}^n |\alpha_i|$$

Disso,

$$Ax = A\left(\sum_{i=1}^{n} \alpha_i e_i\right) = \sum_{i=1}^{n} \alpha_i A e_i = \sum_{i=1}^{n} \alpha_i e_i$$

Mas, $a_1, ..., a_n$ é o conjunto das colunas de A, que é ortonormal. Novamente,

$$\|Ax\|_2 = \left\| \sum_{i=1}^{n} \alpha_i a_i \right\|_2 = \sum_{i=1}^{n} |\alpha_i| \|a_i\|_2 = \sum_{i=1}^{n} |\alpha_i|$$

Portanto, posto que $x \in \mathbb{R}^n$ é arbitrário, $\|Ax\|_2 = \|x\|_2$.

Agora, supomos que A seja tal que, para $x \in \mathbb{R}^n$, $\|Ax\|_2 = \|x\|_2$. Por absurdo, supomos que $s_1 = k > 0$ tal que $k \neq 1$. Disso, para algum $y \in \mathbb{R}^n - \{0\}$, $k\|y\| = \|Ay\|$. Mas isso contradiz a hipótese inicial. Logo, $s_1 = 1$. De modo análogo, concluímos que $s_n = 1$.

Por fim, supomos que $s_1 = 1 = s_n$. Disso, a matriz Σ da decomposição em valores singulares de A é a matriz I_n. Assim, ao também considerarmos a ortogonalidade das matrizes da decomposição em valores singulares,

$$\begin{aligned} A^T A &= (U\Sigma V^T)^T (U\Sigma V^T) \\ &= ((V^T)^T \Sigma^T U^T) U\Sigma V^T \\ &= V I_n^T (U^{-1}U) I_n V^{-1} \\ &= V V^{-1} \\ &= I_m \end{aligned}$$

Lem. 1: Se $A \in \mathbb{R}^{m\times n}$ e $\delta > 0$ tal que $\|A^T A - I_n\| \leq \max\{\delta, \delta^2\}$, então, para $x \in \mathbb{R}^n$ arbitrário, $(1-\delta)\|x\|_2 \leq \|Ax\|_2 \leq (1+\delta)\|x\|_2$. Consequentemente, $1 - \delta < s_n \leq ... \leq s_1 < 1 + \delta$.

Dem.:

Podemos assumir sem perda de generalidade que $||x||_2 = 1$, pois sempre podemos normalizar o vetor. Temos, então, da nossa hipótese,

$$\begin{aligned} \max(\delta, \delta^2) &\geq \|A^T A - I_n\| \\ &\geq |\langle (A^T A - I_n)x, x\rangle| \\ &= |(A^T Ax)^T x - (I_n x)^T x| \\ &= |x^T A^T Ax - x^T x| \\ &= |(Ax)^T Ax - \|x\|_2| \\ &= |\|Ax\|_2^2 - 1| \end{aligned}$$

Agora, provamos que $\max(|z-1|, |z-1|^2) \leq |z^2-1|$, para $z \geq 0$. Se $z < 1$, então $z \geq z^2$. Assim, $|z-1| = 1-z \leq 1-z^2 = |z^2-1|$. Além disso, uma vez que $|z-1| \leq 1$, $|z-1|^2 \leq |z-1|$. Logo, a desigualdade é satisfeita. No caso em que $z > 1$, se $z \leq 2$, então, analogamente, $|z-1| \leq 1$ e, assim, $|z-1| \geq |z-1|^2$. Disso, $|z-1| = z-1 < z^2-1 = |z^2-1|$. Outra vez, a afirmação é correta. Caso contrário, $|z-1| > 1$ e, consequentemente, $|z-1|^2 > |z-1|$. Mas $|z-1|^2 = (z-1)^2 = z^2-2z+1 < z^2-3 < z^2-1 = |z^2-1|$. Portanto, a desigualdade é válida.

Disso, $\max(\delta, \delta^2) \geq |\|Ax\|_2^2 - 1| \geq \max(|\|Ax\|_2 - 1|, |\|Ax\|_2 - 1|^2)$. No caso em que $\max(\delta, \delta^2) = \delta$, $\delta \geq \max(|\|Ax\|_2 - 1|, |\|Ax\|_2 - 1|^2) \geq |\|Ax\|_2 - 1|$. Disso, $\delta \geq \|Ax\|_2 - 1 \geq \delta$, isto é, $1+\delta \geq \|Ax\|_2 \geq 1-\delta$. Analogamente, se $\max(\delta, \delta^2)$, então $\delta^2 \geq |\|Ax\|_2 - 1|^2 \geq 0$. Assim, ao tomarmos a raíz quadrada, chegamos ao primeiro caso. Portanto, uma vez que $\|x\|_2 = 1$, concluímos que $(1+\delta)\|x\|_2 \geq \|Ax\|_2 \geq (1-\delta)\|x\|_2$.

Exe. 5: Primeiro, para $k = 1, ..., n$, $s_k^2 = \lambda_k$, em que λ_k é autovalor de A^TA. Assim, se $\delta \in (0, 1]$, dado que $1-\delta \geq 0$, então $0 < (1-\delta)^2 \leq \lambda_k \leq (1+\delta)^2$. Mas, dado que $\delta \geq \delta^2 > 0$, $(1+\delta)^2 \leq 1+3\delta$ e $(1-\delta)^2 > 1-2\delta$, ou seja, $1-2\delta < \lambda_k \leq 1+3\delta$ e, consequentemente, $-2\delta < \lambda_k - 1 \leq 3\delta$. Logo, $\lambda_k - 1 \in (-2\delta, 3\delta]$. Portanto, $|\lambda_k - 1| \leq 3\delta = 3\max(\delta, \delta^2)$. Agora, se $\delta > 1$, então $0 \leq \lambda_k \leq (1+\delta)^2$. Disso, posto que $\delta^2 > \delta$, $(1+\delta)^2 < 1+3\delta^2$ e, por isso, $0 \leq \lambda_k < 1+3\delta^2$, isto é, $-\delta^2 < -1 \leq \lambda_k - 1 < 3\delta^2$. Logo, $\lambda_k - 1 \in (-\delta^2, 3\delta^2)$. Portanto, $|\lambda_k - 1| < 3\delta^2 = 3\max(\delta, \delta^2)$. Desse modo, uma vez que $|\lambda_1 - 1| = \|A^TA - I_n\|$, $\|A^TA - I_n\| \leq 3\max(\delta, \delta^2)$.

Obs.: Para $Q \in \mathbb{R}^{n\times m}$, pelo Exercício 4, $QQ^T = I_n$ se e somente se Q^TA é uma projeção ortogonal de dimensão n em $\mathbb{R}$, em que Q é dita projeção de $\mathbb{R}^m$ para $\mathbb{R}^n$. Além disso, $A \in \mathbb{R}^{m\times n}$ é uma imersão isométrica de $\mathbb{R}^n$ em $\mathbb{R}^m$ se e somente se A^T é uma projeção de $\mathbb{R}^m$ para $\mathbb{R}^n$. Analogamente, no caso de A ser uma isometria aproximada, A^T é dita uma projeção aproximada.

Exe. 6:

Uma matriz é dita unitária se $U^*U = UU^* = I_n$ em que U^* é o transposto conjugado (no caso real $U^* = U^T$ e matriz unitária é a matriz ortonormal)

Seja $U_{nn} = \begin{bmatrix} | & | & ... & | \\ v_1 & v_2 & ... & v_n \\ | & | & ... & | \end{bmatrix}$ uma matriz ortonormal, ou seja, os vetores colunas são ortogonais entre si.

Escolhendo k colunas quaisquer da matriz e juntando em uma nova matriz (ordenados as colunas para facilitar) temos:

$$V_{nk} = \begin{bmatrix} | & | & ... & | \\ v_1 & v_2 & ... & v_k \\ | & | & ... & | \end{bmatrix} \text{ e } V_{kn}^T = \begin{bmatrix} - & v_1 & - \\ - & v_2 & - \\ ... & ... & ... \\ - & v_k & - \end{bmatrix}$$

Fazendo:

$$V_{kn}^T V_{nk} = \begin{bmatrix} - & v_1 & - \\ - & v_2 & - \\ ... & ... & ... \\ - & v_k & - \end{bmatrix} \cdot \begin{bmatrix} | & | & ... & | \\ v_1 & v_2 & ... & v_k \\ | & | & ... & | \end{bmatrix} = \begin{bmatrix} < v_1, v_1 > & < v_1, v_2 > & ... & < v_1, v_k > \\ < v_2, v_1 > & < v_2, v_2 > & ... & < v_2, v_k > \\ ... & ... & ... & ... \\ < v_k, v_1 > & < v_k, v_2 > & ... & < v_k, v_k > \end{bmatrix}$$

Como os v_i são ortonormais entre si, temos que $< v_i, v_j >= \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$, logo $V_{kn}^T V_{nk} = I_k$, logo V é isometria pela definição.

Agora escolhendo k linhas quaisquer de U temos:

$V_{kn} = \begin{bmatrix} | & | & ... & | \\ v_1' & v_2' & ... & v_n' \\ | & | & ... & | \end{bmatrix}$ Como U é uma matriz unitária, as linhas continuas sendo ortonormais

$P = (V_{kn} V_{nk}^T)^T = V_{kn} V_{nk}^T$ logo é simétrica

$P^2 = PP = V_{kn} V_{nk}^T V_{kn} V_{nk}^T = V_{kn} \cdot I_n \cdot V_{nk}^T = V_{kn} V_{nk}^T$

Seja $u \in Ker(P)$ e $w \in Im(P)$, temos que $< u, w >=< u, Pw >=< P^T u, w >=< Pu, w >=< 0, w >= 0$, logo a imagem é o núcleo são ortogonais entre si.

Decomposição em blocos: $V_{kn} V_{nk}^T = \sum_{i=1}^{n} v_i' v_i'^T$

3 Redes e números de cobertura e de embalagem (4. 2)

Def. (ϵ-rede): Para (X, d) espaço métrico, $K \subset X$ subconjunto e $\epsilon > 0$, subconjunto $\mathcal{N} \subset K$ é dito ϵ-rede de K se, para todo $x \in K$, existe $y \in \mathcal{N}$ tal que $d(x, y) \leq \epsilon$.

Def. (Número de cobertura): A menor cardinalidade possível para que $\mathcal{N}$ seja ϵ-rede de K, notado como $\mathcal{N}(K, d, \epsilon)$.

Obs.: Para (X, d) espaço métrico completo, o subconjunto K é pré-compacto se e somente se, para todo $\epsilon > 0$, $\mathcal{N}(K, d, \epsilon) < \infty$, ou seja, o número de cobertura é uma medida de compacidade de K.

Def. (ϵ-separabilidade): Para (X, d) espaço métrico, $K \subset X$ subconjunto e $\epsilon > 0$, subconjunto $\mathcal{N} \subset X$ é dito ϵ-separado em K se, para todo $x, y \in \mathcal{N}$, $d(x, y) > \epsilon$.

Def. (Número de embalagem): A maior cardinalidade possível para que, dado subconjunto $K \subset X$, $\mathcal{N}$ seja ϵ-separado em K, notado como $\mathcal{P}(K, d, \epsilon)$.

Exe. 7:

a) Sejam $(X, \|\cdot\|)$ um espaço normado, d a métrica induzida pela norma $\|\cdot\|$ e $K \subset X$ subconjunto. Logo, pela definição, o número de embalagem é

$$\begin{aligned}\mathcal{P}(K, d, \epsilon) &= \max\{|\mathcal{N}| : \mathcal{N} \subset K \text{ é } \epsilon\text{-separado em } K\} \\ &= \max\{|\mathcal{N}| : \mathcal{N} \subset K \text{ tal que, se } x, y \in \mathcal{N}, \text{ então } d(x, y) > \epsilon\} \\ &= \max\{|\mathcal{N}| : \mathcal{N} \subset K \text{ tal que, se } x, y \in \mathcal{N}, \text{ então } \|x - y\| > \epsilon\}\end{aligned}$$

Seja $\mathcal{N}_0 = \text{Argmax}\{|\mathcal{N}| : \mathcal{N} \subset K \text{ tal que, se } x, y \in \mathcal{N}, \text{ então } \|x - y\| > \epsilon\}$. Disso, por absurdo, supomos que existam $x_0, y_0 \in \mathcal{N}_0$ distintos tais que, para algum $z_0 \in K$, $z_0 \in \mathcal{B}(x_0, \epsilon/2) \cap \mathcal{B}(y_0, \epsilon/2)$. Logo, $\|x_0 - z_0\| + \|y_0 - z_0\| \leq \epsilon$. Mas, pela desigualdade triangular, $\|x_0 - y_0\| \leq \epsilon$, uma contradição com o fato de que $\mathcal{N}_0$ é ϵ-separado. Portanto, $\mathcal{N}_0$ é o conjunto maximal de centros de bolas fechadas de raio $\epsilon/2$.

b) Sejam X um conjunto finito qualquer e d a métrica discreta. Logo, para $x, y \in X$ distintos, $d(x, y) = 1$. Disso, $\mathcal{P}(X, d, 1) = 1$. Por outro lado, o maior número de bolas fechadas disjuntas centradas nos pontos de X e de raio $1/2$ é $|X|$.

Lem. 2: Se $\mathcal{N}$ é ϵ-separado maximal em K, então $\mathcal{N}$ é ϵ-rede em K.

Dem.:

Seja $x \in K$. Se $x \in \mathcal{N}$, então, trivialmente, existe $x_0 \in \mathcal{N}$ tal que $d(x,x) = d(x,x_0) \leq \epsilon$. Caso contrário, por absurdo, supomos que não exista $x_0 \in \mathcal{N}$ tal que $d(x,x_0) \leq \epsilon$. Logo, $\mathcal{N}\prime = \mathcal{N} \cup \{x\}$ é ϵ-separado em K, uma contradição à maximalidade de $\mathcal{N}$. Portanto, $\mathcal{N}$ é, realmente, ϵ-rede de K.

Obs.: o último lema indica um algoritmo para a construção de uma ϵ-rede em um conjunto K compacto: escolhemos $x_1 \in X$ qualquer; em seguida, escolhemos $x_2 \in X$ tal que $d(x_1,x_2) > \epsilon$; e, a partir daí, escolhemos $x_n \in X$ tal que, para $x_i = x_1, ..., x_{n-1} \in X$, $d(x_i,x_n) > \epsilon$. Dado que K é compacto, existe uma subcobertura finita dele.

Lem. 3: Se $K \subset X$ subconjunto e $\epsilon > 0$, então $\mathcal{P}(K,d,2\epsilon) \leq \mathcal{N}(K,d,\epsilon) \leq \mathcal{P}(K,d,\epsilon)$.

Dem.:

Pelo Lema 2, se $\mathcal{N}_0$ é ϵ-separado maximal em K, então $\mathcal{N}_0$ é ϵ-rede em K. Disso, $\mathcal{N}(K,d,\epsilon) = \min\{|\mathcal{N}| : \mathcal{N} \subset K \text{ é } \epsilon\text{-rede em } K\} \leq |\mathcal{N}_0| = \max\{|\mathcal{N}| : \mathcal{N} \subset K \text{ é } \epsilon\text{-separado em } K\} = \mathcal{P}(K,d,\epsilon)$.

Por outro lado, sejam $\mathcal{N}_1$ 2ϵ-separado em K e $\mathcal{N}_2$ ϵ-rede em K. Logo, para cada $x_i \in \mathcal{N}_1$, existe $y_j \in \mathcal{N}_2$ tal que $x_i \in \mathcal{B}(y_j,\epsilon)$. Além disso, se $x_i, x_j \in \mathcal{N}_1$ distintos, então $d(x_i,x_j) > 2\epsilon$, ou seja, para cada $y_i \in \mathcal{N}_2$, existe, no máximo, um $x_j \in \mathcal{N}_1$ tal que $x_j \in \mathcal{B}(y_i,2\epsilon)$. Logo, pelo princípio das casas dos pombos, $|\mathcal{N}_1| \leq |\mathcal{N}_2|$. Porém, dado que $\mathcal{N}_1$ e $\mathcal{N}_2$ são arbitrários, $\mathcal{P}(K,d,2\epsilon) \leq \mathcal{N}(K,d,\epsilon)$.

Exe. 8: Sejam (X,d) espaço métrico, $K \subset X$ subconjunto e $\epsilon > 0$. Supomos que $\mathcal{N}_0$ é ϵ-rede minimal em K. Disso, se $x \in \mathcal{N}_0$, então $x \in K$ e, consequentemente, $x \in X$. Portanto, tomamos a ϵ-rede externa em K como $\mathcal{N}^{\text{ext}} = \mathcal{N}_0$. Logo, $\mathcal{N}^{\text{ext}}(K,d,\epsilon) = \min\{|\mathcal{N}| : \mathcal{N} \subset X \text{ é } \epsilon\text{-rede em } K\} \leq |\mathcal{N}^{\text{ext}}| = |\mathcal{N}_0| = \min\{|\mathcal{N}| : \mathcal{N} \subset K \text{ é } \epsilon\text{-rede em } K\} = \mathcal{N}(K,d,\epsilon)$

Agora, supomos que $\mathcal{N}_0^{\text{ext}}$ seja $(\epsilon/2)$-rede externa minimal em K. Logo, para cada $x_i \in \mathcal{N}_0^{\text{ext}}$, existe $y_i \in K$ tal que $y_i \in \mathcal{B}(x_i,\epsilon/2) - \cup_{x_j \in \mathcal{N}_0^{\text{ext}} - \{i\}} \mathcal{B}(x_j,\epsilon/2)$ – se isso fosse falso, então $\mathcal{N}_0^{\text{ext}}$ não seria minimal, já que poderíamos retirar x_i e ainda cobrir K. Porém, $\mathcal{B}(x_i,\epsilon/2) \subset \mathcal{B}(y_i,\epsilon)$, em que $y_i \in K$. Disso, consideramos $\mathcal{N}_0 = \{y_1, ..., y_n\}$ que é uma ϵ-rede em K, pois contém a $(\epsilon/2)$-rede em K. Portanto, $\mathcal{N}(K,d,\epsilon) = \min\{|\mathcal{N}| : \mathcal{N} \subset K \text{ é uma } \epsilon\text{-rede em } K\} \leq |\mathcal{N}_0| = |\mathcal{N}_0^{\text{ext}}| = \min\{|\mathcal{N}| : \mathcal{N} \subset X \text{ é uma } (\epsilon/2)\text{-rede em } K\} = \mathcal{N}^{\text{ext}}(K,d,\epsilon/2)$.

Exe. 9: Sejam $K = [0,1]$, $L = [0,1/3] \cup [2/3,1]$ e d euclidiana. Obviamente, $L \subset K$. Além disso, $\mathcal{N}_K = \{1/2\}$ é uma ϵ-rede em K, pois, se $x \in K$, então $d(x,1/2) \leq 1/2$. Logo, $\mathcal{N}(K,d,1/2) = 1$. Porém, se $\mathcal{N}_L$ é uma ϵ-rede em L, então existem $x_1 \in [0,1/3]$ e $x_2 \in [2/3,1]$ em $\mathcal{N}_L$. Disso, $\mathcal{N}(L,d,1/2) = 2$. Portanto, $L \subset K$ não implica $\mathcal{N}(L,d,\epsilon) \leq \mathcal{N}(K,d,\epsilon)$.

Pela cadeia de desigualdades da última questão, $\mathcal{N}(L,d,\epsilon) \leq \mathcal{N}^{\text{ext}}(L,d,\epsilon/2) = \mathcal{N}(X,d,\epsilon/2)$. Analogamente, $\mathcal{N}(X,d,\epsilon/2) = \mathcal{N}^{\text{ext}}(L,d,\epsilon/2) \leq \mathcal{N}(K,d,\epsilon/2)$. Portanto, $\mathcal{N}(L,d,\epsilon) \leq \mathcal{N}(K,d,\epsilon/2)$.

3.1 Número de cobertura e volume (4. 2. 1)

Def. (Soma de Minkowski): Sejam $A, B \subset \mathbb{R}^n$ subconjuntos, definimos $A+B = \{a+b : a \in A, b \in B\}$

Pro. 1: Se $K \subset \mathbb{R}^n$ subconjunto e $\epsilon > 0$, então $|K|/|\epsilon B_2^n| \leq \mathcal{N}(K,\epsilon) \leq \mathcal{P}(K,\epsilon) \leq |K + (\epsilon/2)B_2^n|/|(\epsilon/2)B_2^n|$, em que $|\cdot|$ é a norma euclidiana de $\mathbb{R}^n$ e B_2^n a bola euclidiana unitária em $\mathbb{R}^n$.

Dem.:

Se $n = \mathcal{N}(K,\epsilon)$, então K pode ser coberto por n bolas de raio ϵ, ou seja, $|K| \leq n|\epsilon B_2^n|$. Portanto, $|K|/|\epsilon B_2^n| \leq \mathcal{N}(K,\epsilon)$.

A desigualdade intermediária é a do Lema 3.

Se $n = \mathcal{P}(K,\epsilon)$, então podemos, pelo Exercício 7, ter n bolas disjuntas de raio $\epsilon/2$ e centro em K. Essas bolas podem não estar totalmente contidas em K, mas o estão em $K + (\epsilon/2)B_2^n$, na qual a soma de Minkowski. Disso, $n|(\epsilon/2)B_2^n| \leq |K + (\epsilon/2)B_2^n|$. Portanto, $\mathcal{P}(K,\epsilon) \leq |K + (\epsilon/2)B_2^n|/|(\epsilon/2)B_2^n|$.

Cor. 1: O número de cobertura de bola euclidiana unitária B_2^n satisfaz, para $\epsilon > 0$, $\epsilon^{-n} \leq \mathcal{N}(B_2^n,\epsilon) \leq ((2/\epsilon)+1)^n$.

Dem.:

Dado que $|\epsilon B_2^n| = \epsilon^n |B_2^n|$, pela primeira desigualdade da cadeia da Proposição 1, $\epsilon^{-n} = |B_2^n|/|\epsilon B_2^n| \leq \mathcal{N}(\mathcal{B}_2^n,\epsilon)$.

Novamente, pela Proposição 1, $((2/\epsilon)+1)^n = ((1+(\epsilon/2))/(\epsilon/2))^n = ((1+$

$(\epsilon/2))^n)/((\epsilon/2)^n) = |(1 + (\epsilon/2))B_2^n|/|(\epsilon/2)B_2^n| \geq \mathcal{N}(B_2^n, \epsilon)$.

Def. (Espaço métrico de Hamming): O cubo de Hamming é $\{0,1\}^n$ e a distância de Hamming é $d_H(x,y) = |\{i : x_i \neq y_i\}|$, em que $x, y \in \{0,1\}^n$. Disso, o espaço métrico de Hamming é $(\{0,1\}^n, d_H)$.

Exe. 10: Sejam d_H a métrica de Hamming e $x, y, z \in \{0,1\}^n$

(i) Se $x = y$, então, para todo $i = 1, ..., n$, $x_i = y_i$. Disso, $d_H(x,y) = |\{i : x_i \neq y_i\}| = 0$. Reciprocamente, seja $d_H(x,y) = 0$. Por absurdo, supomos que exista $i = 1, ..., n$ tal que $x_i \neq y_i$. Logo, $d_H(x,y) = |\{i : x_i \neq y_i\}| = 1 > 0$, uma contradição à hipótese inicial. Desse modo, $x = y$. Portanto, $x = y$ se e somente se $d_H(x,y) = 0$.

(ii) Se $x \neq y$, então existe $i = 1, ..., n$ tal que $x_i \neq y_i$. Portanto, $d_H(x,y) = |\{i : x_i \neq y_i\}| = 1 > 0$.

(iii) Dado que $\{i : x_i \neq y_i\} = \{j : y_j \neq x_j\}$, $d_H(x,y) = |\{i : x_i \neq y_i\}| = |\{j : y_j \neq x_j\}| = d_H(y,x)$.

(iv) Para $i = 1, ..., n$, ou $x_i = y_i$, ou $x_i \neq y_i$. Neste caso, dado que $x_i, y_i, z_i \in \{0,1\}$, $z_i = x_i$ ou $z_i = y_i$, ou seja, $z_i \neq y_i$ ou $z_i \neq x_i$. Desse modo, $|\{i : x_i \neq y_i\}| = 1 \leq 1 + 0 = |\{i : x_i \neq z_i\}| + |\{i : z_i \neq x_i\}|$. Naquela situação, ou $x_i = y_i = z_i$, ou $x_i = y_i \neq z_i$. Na primeira situação, $|\{i : x_i \neq y_i\}| = 0 \leq 0 + 0 = |\{i : x_i \neq z_i\}| + |\{i : z_i \neq x_i\}|$ e, na outra, $|\{i : x_i \neq y_i\}| = 0 \leq 1 + 1 = |\{i : x_i \neq z_i\}| + |\{i : z_i \neq x_i\}|$. Disso, ao considerarmos que $\sum_{i=1}^n |\{i : x_i \neq y_i\}| = |\cup_{i \in \{1,...,n\}} \{i : x_i \neq y_i\}|$, $d_H(x,y) = |\{i : x_i \neq y_i\}| \leq |\{i : x_i \neq z_i\}| + |\{i : z_i \neq y_i\}| = d_H(x,z) + d_H(z,y)$. Portanto, a desigualdade triangular é válida.

Exe. 11:

Provar:

$$\frac{2^n}{\sum_{k=0}^{m} \binom{n}{k}} \leq \mathcal{N}(K, d_H, m) \leq \mathcal{P}(K, d_H, m) \leq \frac{2^n}{\sum_{k=0}^{\lfloor m/2 \rfloor} \binom{n}{k}}$$

$k = \{0,1\}^n$, logo k tem 2^n vetores ao todo

▪ A desigualdade $\mathcal{N}(k, d_H, m) \leq \mathcal{P}(k, d_H, m)$ vem dos lemas 4.2.6 e 4.2.8

▪ A desigualdade da esquerda (direto da proposição 4.2.12):

$\mathcal{N}(K, d_H, m)$ é o número mínimo de bolas de raio m que cobrem k. Para um vetor estar na bola ele pode diferir em no máximo m elementos do centro da bola. Temos $\binom{n}{0}$ que diferem em 0 elementos, $\binom{n}{1}$ que diferem em 1 elemento, ... , $\binom{n}{i}$ que diferem em i elementos, logo temos $\sum_{k=0}^{m} \binom{n}{k}$ elementos dentro de cada bola.

Temos ao todo $\sum_{k=0}^{m} \binom{n}{k} \cdot \mathcal{N}(K, d_H, m)$ elementos cobertos, como cobrimos todo k então $2^n \leq \sum_{k=0}^{m} \binom{n}{k} \cdot \mathcal{N}(K, d_H, m)$

▪ Desigualdade da direita:

$\mathcal{P}(K, d_H, m)$ é a cardinalidade do maior subconjunto m-separável de K. Podemos fazer bolas de raio $\lfloor m/2 \rfloor$ que essas bolas vão ser disjuntas. Temos:

$\sum_{k=0}^{m} \binom{n}{k} \cdot \mathcal{P}(K, d_H, m)$ elementos cobertos pelas bolas. Como os centros das bolas estão a uma distância de pelo menos $m+1$ e o raio é de $\lfloor m/2 \rfloor$, não cobrimos os 2^n pontos por completo, logo $\sum_{k=0}^{m} \binom{n}{k} \cdot \mathcal{P}(K, d_H, m) \leq 2^n$

4 Cotas superiores para matrizes aleatórias sub-gaussianas (4.4)

4.1 Cálculo da norma de uma rede (4. 4. 1)

Lem. 4: Se $A \in \mathbb{R}^{m\times n}$ e $\epsilon \in (0,1)$, então, para $\mathcal{N}$ ϵ-rede da esfera $S(\mathbb{R}^n)$, $\sup_{x\in\mathcal{N}}\|Ax\|_2 \leq \|A\| \leq (1-\epsilon)^{-1}\sup_{x\in\mathcal{N}}\|Ax\|_2$.

Dem.:

Dado que $\mathcal{N} \subset S(\mathbb{R}^n)$, trivialmente, se $x \in \mathcal{N}$, então $\|Ax\|_2 \leq \|A\|$. Portanto, $\sup_{x\in\mathcal{N}}\|Ax\|_2 \leq \|A\|$.

Agora, uma vez que $S(\mathbb{R}^n)$ é compacta e $\|\cdot\|$ é contínua, existe $x \in S(\mathbb{R}^n)$ tal que $\|Ax\|_2 = \sup_{x\in S(\mathbb{R}^n)} = \|A\|$. Além disso, se $\mathcal{N}$ é uma ϵ-rede, então existe $y \in \mathcal{N}$ tal que $\|x-y\|_2 \leq \epsilon$. Disso, pela desigualdade da norma do operador, $\|Ax-Ay\|_2 = \|A(x-y)\|_2 \leq \|A\|\|x-y\|_2 \leq \epsilon\|A\|$. Logo, pela desigualdade triangular, $(1-\epsilon)\|A\| = \|A\| - \epsilon\|A\| \leq \|Ax\| - \|Ax-Ay\|_2 \leq \|Ay\|_2 \leq \sup_{z\in\mathcal{N}}\|Az\|_2$. Portanto, ao multiplicarmos por $(1-\epsilon)^{-1}$, $\|A\| \leq (1-\epsilon)\sup_{x\in\mathcal{N}}\|Ax\|_2$.

Exe. 12: Para $x \in \mathbb{R}^n$, $\epsilon \in (0,1]$ e $\mathcal{N}$ ϵ-rede de $S(\mathbb{R}^n)$, se $x = 0$, então $\|x\|_2 = 0$ e, para todo $y \in \mathbb{R}^n$, $\langle x,y\rangle = 0$. Portanto, $\sup_{y\in\mathcal{N}}\langle x,y\rangle \leq \|x\|_2$. Caso contrário, definimos $x^* = x/\|x\|_2$. Disso, dado que $\sup_{y\in S(\mathbb{R}^n)}\langle x^*,y\rangle = 1$ em que $\operatorname{Argmax}\sup_{y\in S(\mathbb{R}^n)}\langle x^*,y\rangle = x^*$, se $y \in \mathcal{N}$, então $\langle x^*,y\rangle \leq 1$, ou seja, $\langle x,y\rangle \leq \|x\|_2$. Portanto, $\sup_{y\in\mathcal{N}}\langle x,y\rangle \leq \|x\|_2$.

Analogamente, no caso trivial em que $x = 0$, a desigualdade $\|x\|_2 = 0 \leq 0 = (1-\epsilon)^{-1}\|x\|_2$ é válida. Para além disso, se $x \neq 0$, consideramos, outra vez, $x^* = x/\|x\|$ e a desigualdade $\sup_{y\in S(\mathbb{R}^n)}\langle x^*,y\rangle \leq 1$. Mas, por absurdo, supomos que $\sup_{y\in\mathcal{N}}\langle x^*,y\rangle < 1-\epsilon$. Logo, uma vez que $x^*, y \in S(\mathbb{R}^n)$, $d(x^*,y) = \|x^*-y\|_2 = \sqrt{\langle x^*-y, x^*-y\rangle} = \sqrt{\langle x^*,x^*\rangle - 2\langle x^*,y\rangle + \langle y,y\rangle} = \sqrt{2(1-\langle x^*,y\rangle)} > \sqrt{2\epsilon} > \epsilon$, uma contradição ao fato de que $\mathcal{N}$ é uma ϵ-rede em $S(\mathbb{R}^n)$. Disso, $\sup_{y\in S(\mathbb{R}^n)}\langle x^*,y\rangle \geq 1-\epsilon$. Portanto, $(1-\epsilon)^{-1}\sup_{y\in S(\mathbb{R}^n)}\langle x,y\rangle \geq \|x\|_2$.

Exe. 13: Sejam $A \in \mathbb{R}^{m\times n}$ e $\epsilon \in [0,1/2)$.

a) Para $\mathcal{N}$ ϵ-rede em $S(\mathbb{R}^n)$ e $\mathcal{M}$ ϵ-rede em $S(\mathbb{R}^m)$,

b)

Exe. 14:

4.2 As normas para matrizes aleatórias sub-gaussianas (4. 4. 2)

Teo. 3:

Seja A uma matriz aleatória m por n cujas entradas A_{ij} sejam variáveis aleatórias sub-gaussianas[3] independentes de média zero. Então, para cada $t > 0$ temos:

$$||A|| \leq C \cdot K \cdot (\sqrt{m} + \sqrt{n} + t)$$

com probabilidade pelo menos $1 - 2 \cdot exp(-t^2)$. Onde $k = \max_{i,j} ||A_{ij}||_{\Psi_2}$

Dem.:

Precisamos verificar $< Ax, y >$ para todos os vetores x, y da esfera unitária. Para isso vamos usar uma estratégia: Passo 1 (Aproximação): Discretizar a esfera usando uma ϵ-rede, assim vamos poder olhar só para os pontos da ϵ-rede ao invés de olhar para toda a esfera. Passo 2 (Concentração): Estabeleceremos um controle sobre $< Ax, y >$ para todos os x, y da ϵ-rede. Passo 3 (União): Aplicar a união de probabilidades para estender o controle obtido nos pontos discretos da rede para toda a esfera. (C e c são constantes positivas)

Passo 1 (Aproximação): Escolha $\epsilon = \frac{1}{4}$. Usando o corolário 4.2.13, podemos achar uma ϵ-rede $\mathcal{N}$ da esfera S^{n-1} e uma uma ϵ-rede $\mathcal{M}$ da esfera S^{m-1} com cardinalidades:

$$\mathcal{N} \leq 9^n \text{ e } \mathcal{M} \leq 9^m$$

Pelo exercício 4.4.3 a norma de operador de A pode ser limitada usando redes:

$$||A|| \leq 2 \max_{x \in \mathcal{N}, y \in \mathcal{M}} < Ax, y >$$

Passo 2 (Concentração): Fixe $x \in \mathcal{N}$ e $y \in \mathcal{M}$. Então:
$< Ax, y >= y^T Ax = \sum_{j=1}^{m} y_i (Ax)_i$; $(Ax)_i = \sum_{j=1}^{n} A_{ij} x_j \rightarrow < Ax, y >= \sum_{i=1}^{m} \sum_{j=1}^{n} y_i A_{ij} x_j$

É uma soma de variáveis aleatórias sub-gaussianas independentes. Pela proposição 2.6.1 (Falar alguma coisa sobre essa proposição) temos que a soma é sub-gaussiana e que:

$$|| < Ax, y > ||_{\Phi_2}^2 \leq C \sum_{i=1}^{m} \sum_{j=1}^{n} ||y_i A_{ij} x_j||_{\Phi_2}^2 \leq CK^2 \sum_{i=1}^{m} \sum_{j=1}^{n} y_i^2 x_j^2 = CK^2 \left(\sum_{i=1}^{m} y_i^2 \right) \left(\sum_{j=1}^{n} x_j^2 \right)$$

Como x e y são unitários então: $|| < Ax, y > ||_{\Phi_2}^2 \leq CK^2$

Usando (2.14) (Falar sobre o 2.14) podemos reafirmar isso como o limite da cauda:

[3] A cauda da distribuição decai pelo menos tão rápido quanto uma gaussiana. É sub-gaussiana se $\exists \alpha > 0 \; t.q. \mathbb{P}(|X - \mu| > t) \leq 2 \cdot exp\left(\frac{-t^2}{2\alpha^2}\right)$

$$\mathbb{P}\{< Ax, y >\geq u\} \leq 2 \cdot exp\left(\frac{-cu^2}{k^2}\right),\ u \geq 0$$

Passo 3 (União): Agora, desfixamos x e y usando uniões . Suponha que o evento $\max_{x\in\mathcal{N},y\in\mathcal{M}} < Ax, y >\geq u$ ocorra. Então existe $x \in \mathcal{N}$ e $y \in \mathcal{M}$ tais que $< Ax, y >\geq u$. Então:

$$\mathbb{P}\{\max_{x\in\mathcal{N},y\in\mathcal{M}} < Ax, y >\geq u\} \leq \sum_{x\in\mathcal{N},y\in\mathcal{M}} \mathbb{P}\{< Ax, y >\geq u\}$$

Usando que $\mathcal{N} \leq 9^n$ e $\mathcal{M} \leq 9^m$ e $\mathbb{P}\{< Ax, y >\geq u\} \leq 2 \cdot exp\left(\frac{-cu^2}{k^2}\right),\ u \geq 0$, temos que:

$\sum_{x\in\mathcal{N},y\in\mathcal{M}} \mathbb{P}\{< Ax, y >\geq u\} \leq \mathcal{N} \cdot \mathcal{M} \cdot 2 \cdot exp\left(\frac{-cu^2}{k^2}\right) \leq 9^{n+m} \cdot 2 \cdot exp\left(\frac{-cu^2}{k^2}\right)$

Escolha:

$$u = CK(\sqrt{n} + \sqrt{m} + t)$$

Estão $u^2 \geq C^2K^2(n + m + t^2)$ e se a constante C é escolhida suficientemente grande, então $\frac{cu^2}{k^2}$ também será, digamos $\frac{cu^2}{k^2} \geq 3(m + n) + t^2$, então:

$$9^{n+m} \cdot 2 \cdot exp\left(\frac{-cu^2}{k^2}\right) \leq 9^{n+m} \cdot 2 \cdot exp\left(-3(m + n) - t^2\right) \leq 2 \cdot exp(-t^2)$$

lembrando que $||A|| \leq 2 \max_{x\in\mathcal{N},y\in\mathcal{M}} < Ax, y >$, logo:

$$2 \cdot exp(-t^2) \geq \mathbb{P}\{2 \max_{x\in\mathcal{N},y\in\mathcal{M}} < Ax, y >\geq 2u\} \geq \mathbb{P}\{||A|| \geq 2u\}$$

$\mathbb{P}\{||A|| \geq 2u\} \leq 2 \cdot exp(-t^2)$, o 2 pode ser incorporado na constante C. Basicamente estamos falando que a probabilidade da norma de A se afastar de u decai exponencialmente.

Se $\mathbb{P}\{||A|| \geq 2u\} \leq 2 \cdot exp(-t^2) \rightarrow \mathbb{P}\{||A|| \leq 2u\} \leq 1 - 2 \cdot exp(-t^2)$

Exe. 15:

Exe. 16:

Cor. 2:

Seja A uma matriz aleatória n por n simétrica cujas entradas A_{ij} na diagonal e acima são variáveis aleatórias sub-gaussianas independentes de média zero. Então, para qualquer $t > 0$, temos:

$$||A|| \leq CK(\sqrt{n} + t)$$

com probabilidade pelo menos $1 - 4 \cdot exp(-t^2)$. Aqui $K = \max_{i,j} ||A_{ij}||_{\psi_2}$
Dem.:

Decomponha A em uma matriz triangular superior A^+ e uma matriz triangular inferior A^-. Não importa qual das duas matrizes vai ficar com a diagonal, mas vamos coloca-la na A^+ para sermos específicos. Então:

$$A = A^+ + A^-$$

Aplicando o teorema 3 em cada parte (A^+ e A^-) separadamente. Temos que:

$$||A^+|| \leq CK(\sqrt{n} + t) \text{ e } ||A^-|| \leq CK(\sqrt{n} + t)$$

Analogamente a demonstração do teorema 3:

Escolhemos $u = CK(\sqrt{n} + t)$
$\mathbb{P}\{||A^+|| \geq 2u\} \leq 2 \cdot exp(-t^2)$ e $\mathbb{P}\{||A^-|| \geq 2u\} \leq 2 \cdot exp(-t^2)$
$\mathbb{P}\{||A^+|| + ||A^-|| \geq 2u\} \leq \mathbb{P}\{||A^+|| \geq 2u\} + \mathbb{P}\{||A^-|| \geq 2u\} \leq 4 \cdot exp(-t^2)$
Pela desigualdade triangular temos que $||A|| \leq ||A^+|| + ||A^-||$, logo:
$\mathbb{P}\{||A|| \geq 2u\} \leq \mathbb{P}\{||A^+|| + ||A^-|| \geq 2u\} \leq 4 \cdot exp(-t^2)$

Então: $\mathbb{P}\{||A|| \leq 2u\} \geq 1 - 4 \cdot exp(-t^2)$

5 Aplicação: Detecção de comunidades (4. 5)

Comunidades são aglomerados de vértices bem conectados. Vamos ver como achar as comunidades de forma precisa e eficiente

5.1 Modelo de bloco estocástico (4. 5. 1.)

Vamos tentar resolver o problema de detecção de comunidades para uma rede com duas comunidades.

Def. (Modelo de bloco estocástico)

Divida de n vértices em dois conjuntos (comunidades)[4] com $\frac{n}{2}$ vértices cada. Construa um Grafo aleatório G conectando cada par de vértices de forma independente com probabilidade **p** se eles forem da mesma comunidade e **q** se forem de comunidades diferentes.

Vamos assumir que $p > q$, nesse caso arestas tendem a acontecer entre vértices da mesma comunidade. No caso de $p = q$ temos o modelo **Erdös - Rényi**.

5.2 Matriz de Adjacência Esperada (4. 5. 2)

Será conveniente identificar um grafo G com sua matriz de adjacências[5] A. Para um grafo aleatório $G(n, p, q)$ a matriz de adjacências A vai ser uma matriz aleatória.

Podemos dividir A em uma parte determinística e uma parte aleatória:

$$A = D + R$$

Onde D é a esperança de A (D é o "sinal"e R o "ruído")

Podemos computar a auto estrutura da matriz D. As entradas A_{ij} são Bernoullis com parâmetros **p** ou **q** (Depende se os vértices i e j são da mesma comunidade ou não) então as entradas de D são **p** ou **q**.

Exemplo: Se agruparmos os vértices que pertencem a mesma comunidade juntos então para $n = 4$ a matriz D vai ser:

$$D = \mathbb{E}[A] = \begin{bmatrix} p & p & q & q \\ p & p & q & q \\ q & q & p & p \\ q & q & p & p \end{bmatrix}$$

De modo geral $D = \begin{bmatrix} \mathbf{1}p & \mathbf{1}q \\ \mathbf{1}q & \mathbf{1}p \end{bmatrix}$ em que $\mathbf{1}$ é uma matriz quadrada $\frac{n}{2}$ de 1's

[4] o modelo de bloco estocástico também abrange modelos com mais de duas comunidades diferentes e com tamanhos diferentes

[5] matriz de adjacências de um grafo com n vértices vai ser uma matriz simétrica n por n com 1 na entrada A_{ij} se i e j são ligados e 0 caso contrario

Exe. 17: Mostraremos que a matriz D tem posto 2 e que os autovalores λ_i não 0 e seus correspondentes autovetores são:

$$\lambda_1 = \left(\frac{p+q}{2}\right) n, \quad , \lambda_2 = \left(\frac{p-q}{2}\right) n, \qquad u_1 = \begin{bmatrix} 1 \\ ... \\ 1 \\ 1 \\ ... \\ 1 \end{bmatrix}, \qquad u_2 = \begin{bmatrix} 1 \\ ... \\ 1 \\ -1 \\ ... \\ -1 \end{bmatrix}$$

(Os vetores são n por 1 e em u_2 metade das entradas são 1 e a outra metade -1)

Que a matriz D tem posto 2 é obvio pois as primeiras $\frac{n}{2}$ colunas são iguais umas as outras e as últimas $\frac{n}{2}$ colunas também e se $p \neq q$ então as primeiras $\frac{n}{2}$ colunas são diferentes das outras, logo temos 2 vetores LI.

$$Du_1 = \begin{bmatrix} \mathbf{1}p & \mathbf{1}q \\ \mathbf{1}q & \mathbf{1}p \end{bmatrix} \begin{bmatrix} 1 \\ ... \\ 1 \\ 1 \\ ... \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{n}{2}p + \frac{n}{2}q \\ ... \\ \frac{n}{2}p + \frac{n}{2}q \\ \frac{n}{2}q + \frac{n}{2}p \\ ... \\ \frac{n}{2}q + \frac{n}{2}p \end{bmatrix} = \left(\frac{p+q}{2}\right) n \begin{bmatrix} 1 \\ ... \\ 1 \\ 1 \\ ... \\ 1 \end{bmatrix} = \lambda_1 u_1$$

$$Du_2 = \begin{bmatrix} \mathbf{1}p & \mathbf{1}q \\ \mathbf{1}q & \mathbf{1}p \end{bmatrix} \begin{bmatrix} 1 \\ ... \\ 1 \\ -1 \\ ... \\ -1 \end{bmatrix} = \begin{bmatrix} \frac{n}{2}p - \frac{n}{2}q \\ ... \\ \frac{n}{2}p - \frac{n}{2}q \\ \frac{n}{2}q - \frac{n}{2}p \\ ... \\ \frac{n}{2}q - \frac{n}{2}p \end{bmatrix} = \left(\frac{p-q}{2}\right) n \begin{bmatrix} 1 \\ ... \\ 1 \\ -1 \\ ... \\ -1 \end{bmatrix} = \lambda_2 u_2$$

■

O autovetor u_2 contém toda a informação sobre a estrutura das comunidades (uma ficou com os valores 1 e a outra -1). Se soubéssemos u_2 poderíamos identificar as comunidades olhando para o valor de cada entrada. Mas não temos acesso a $D = \mathbb{E}[A]$ então não temos acesso a u_2, somente sabemos $A = D + R$, uma versão com ruído de D.

Pela norma do operador sabemos que $||D|| = \sigma_1$ mas como D é simétrica $\sigma_1 = |\lambda_1|$, logo $||D|| = \lambda_1 \sim n$. Enquanto que o ruído pode ser limitado usando o corolário 2:

$||R|| \leq C\sqrt{n}$ com probabilidade de pelo menos $1 - 4e^{-n}$

Então para $n >> 1$, o ruído R é muito menor que o sinal D, ou seja, A está próximo de D e vamos poder usar A no lugar de D para extrair informações sobre as comunidades. Vamos justificar isso usando a Teoria de perturbação clássica para matrizes.

5.3 Teoria da perturbação (4. 5. 3)

A teoria da perturbação vai descrever como os autovalores e autovetores da matriz mudam sobre perturbações na matriz.

Teo. 4 (Desigualdade de Weyl):

Para quaisquer matrizes simétricas S e T com mesma dimensão, temos:

$\max_i |\lambda_i(S) - \lambda_i(T)| \leq ||S - T||$. Então a norma do operador determina a estabilidade do espectro.

Exe. 18: Deduza a desigualdade de Weyl usando a caracterização de Courant-Fisher min-max de autovalores:

$$\lambda_i(S) = \max_{Dim(E)=i} \min_{x \in S(E)} < Sx, x >, \qquad \lambda_i(T) = \max_{Dim(E)=i} \min_{x \in S(E)} < Tx, x >$$

$$S = T + (S - T) \text{ e } T = S + (T - S)$$

Para qualquer subespaço E de dimensão i e $x \in E$ unitário:

$< Sx, x >=< Tx, x > + < (S - T)x, x >$

Agora:

$| < (S - T)x, x > | \leq ||(S - T)x|| \cdot ||x||$ por cauchy-schwarz e

$||(S - T)x|| \leq ||S - T|| \cdot ||x||$ pela norma do operador

Então: $| < (S - T)x, x > | \leq ||S - T|| \cdot ||x||^2$ e como $||x|| = 1$, temos:

$| < (S - T)x, x > | \leq ||S - T||$

Então temos:

$< Sx, x > \leq < Tx, x > +||S - T||$ e pra qualquer subespaço E vale:

$\min_{x \in S(E)} < Sx, x > \leq \min_{x \in S(E)} (< Tx, x > +||S - T||) = \min_{x \in S(E)} (< Tx, x >) + ||S - T||$

Tirando o máximo sobre todos os subespaços de dimensão i temos:

$\lambda_i(S) \leq \lambda_i(T) + ||S - T||$

Fazendo a mesma coisa mas trocando S com T temos:

$\lambda_i(T) \leq \lambda_i(S) + ||S - T||$, logo;

$|\lambda_i(S) - \lambda_i(T) \leq ||S - T|||$ e como isso vale para todo i então podemos tirar o máximo:

$$\max_i |\lambda_i(S) - \lambda_i(T) \leq ||S - T|||$$

■

Vale um resultado parecido para os autovetores. Temos que ter cuidado para rastrear o mesmo autovetor antes e depois da perturbação pois se os autovalores $\lambda_i(S)$ e $\lambda_{i+1}(S)$ estiverem muito próximos eles podem acabar trocando de ordem e vamos comparar autovetores diferentes. Para prevenir isso vamos assumir que os

autovalores de S são bem separados

Teo. 5 (Davis - Kahan): Sejam S e T matrizes simétricas com mesmas dimensões. Fixe i e assuma que o i-ésimo maior autovetor de S está bem separado do resto do espectro, ou seja:

$$\min_{j:j\neq i} |\lambda_i(S) - \lambda_j(S)| = \delta > 0$$

Então o ângulo entre os autovetores de S e T correspondentes ao i-ésimo autovetor mais largo (como um número entre 0 e 2π) satisfaz:

$sin\angle(v_i(S), v_i(T)) \leq \frac{2||S-T||}{\delta}$

A conclusão do teorema implica que os autovetores unitários são perto um do outro até um sinal, ou seja:

$\exists\theta \in \{-1, 1\} : ||v_i(S) - \theta v_i(T)|| \leq \frac{2^{\frac{3}{2}}||S-T||}{\delta}$

Tentar fazer a demonstração no apêndice e explicar a conclusão

5.4 Clusterização Espectral (4.5.4)

Aplicar o teorema de Davis - Kahan para $S = D$ e $T = A = D + R$ e para o segundo maior autovetor (u_2).

Precisamos checar se λ_2 é bem separado do resto do espectro de D, isto é, 0 e λ_1. A distância é:

$\delta = min(\lambda_2, \lambda_1 - \lambda_2) = \min\left(\frac{p-q}{2}, q\right) n =: \mu n$

Aplicando $||R|| \leq C\sqrt{n}$ com probabilidade pelo menos $1 - 4e^{-n}$ à:

$\exists\theta \in \{-1, 1\} : ||v_i(S) - \theta v_i(T)|| \leq \frac{2^{\frac{3}{2}}||S-T||}{\delta}$

Podemos limitar a distância entre o autovetor **unitário** de D e A:

$\exists\theta \in \{-1, 1\} : ||v_2(D) - \theta v_2(A)|| \leq \frac{C\sqrt{n}}{\mu n} = \frac{C}{\mu\sqrt{n}}$ com probabilidade pelo menos $1 - 4e^{-n}$.

Mas os autovetores $u_i(D)$ tem norma $\sqrt{n}$. Então multiplicamos ambos os lados por $\sqrt{n}$, obtendo nessa normalização que:

$$||u_2(D) - \theta u_2(A)||_2 \leq \frac{C}{\mu}$$

Segue que a maioria dos coeficientes de $\theta u_2(A)$ e $u_2(D)$ devem ser iguais. De fato, sabemos que:

$$\sum_{j=1}^{n} |u_2(D)_j - \theta u_2(A)_j|^2 \leq \frac{C}{\mu^2}$$

E sabemos que os coeficientes de $u_2(D)$ são todos ± 1. Então todo coeficiente j para os quais os sinais $\theta u_e(A)_j$ e $u_2(D)_j$ discordam contribuem com pelo menos 1 para a soma. Então o número de discordâncias de sinais deve ser limitado por $\frac{C}{\mu^2}$.

Resumindo, podemos usar o vetor $u_2(A)$ para estimar precisamente o $u_2(D)$ e usar os sinais das entradas de $u_2(A)$ para detectar as comunidades.

Algoritmo Clusterização Espectral:

Input: Grafo G

Output: Partição de vértices de G em 2 comunidades

1. Compute a matriz de Adjacências A do Grafo

2. Compute $u_2(A)$ (segundo maior autovetor)

3. Particione os vértices em duas comunidades de acordo com os sinais das entradas de $u_2(A)$

Teo. 6 (Clusterização Espectral para o modelo de bloco estocástico)

Seja $G \sim G(n, p, q)$ com $p > q$ e $\min(q, p - q) = \mu > 0$. Então, com probabilidade de pelo menos $1 - 4e^{-n}$, o algorítimo de clusterização espectral identifica as comunidades de G corretamente com até $\frac{C}{\mu^2}$ classificações incorretas

Resumindo: O algoritmo de clusterização espectral classifica corretamente todos exceto um constante número de vértices, desde que o grafo aleatório seja denso o suficiente ($q \geq const$) e que a probabilidade de ligações dentro e entre comunidades estejam bem separadas ($p - q \geq conts$)

6 Aplicação: Estimação de covariância e clusterização (4.7)

Suponha que queremos analisar dados de alta dimensão representados por pontos $X_1, X_2, ..., X_m$ amostrados de uma distribuição desconhecida de $\mathbb{R}^n$. Uma das ferramentas mais básicas para a análise exploratória dos dados é a análise de componentes principais (PCA).

Como não temos acesso a toda distribuição mas só uma amostra finita $\{X_1, X_2, ..., X_m\}$, só podemos esperar conseguir calcular a matriz de covariância da distribuição subjacente aproximada. Se pudermos fazer isso, o teorema de Davis-Kahan nos permitiria estimar os componentes principais da distribuição subjacente, que são os autovetores da matriz de covariância.

Então, como podemos estimar a matriz de covariância a partir dos dados? Seja X o vetor aleatório extraído da distribuição (desconhecida). Assuma por simplicidade que X tenha média 0 então podemos denotar a matriz de covariância por $\Sigma = \mathbb{E}[XX^T]$ (Não vamos depender de média 0 mas facilita Σ ser matriz de segundo momento de X)

Para estimar Σ podemos usar a matriz de covariância amostral Σ_m, que é computada da amostra como:

$$\Sigma_m = \frac{1}{m}\sum_{i=1}^{m} X_i X_i^T$$

Como X_i e X tem a mesma distribuição, nossa estimativa é não viesada, isto é, $\mathbb{E}[\Sigma_m] = \Sigma$. Então a lei dos grandes números aplicada a cada entrada de Σ nos da que $\Sigma_m \to \Sigma$ quase certamente a medida que o tamanho da amostra m cresce para infinito. Isso leva a pergunta: quão grande o tamanho da amostra precisa ser para que $\Sigma_m \approx \Sigma$ com alta probabilidade? Vamos mostrar que $m \approx n$ é suficiente.

Teo. 7 (Estimação de covariância)

Seja X um vetor aleatório sub gaussiano em $\mathbb{R}^n$. Mais precisamente, assuma que exista $K \geq 1$ tal que:

$$|| < X, x > ||_{\Psi_2} \leq K || < X, x > ||_{L_2} \text{ para qualquer } x \in \mathbb{R}^n$$

Então para qualquer inteiro positivo m, temos:

$$\mathbb{E}||\Sigma_m - \Sigma|| \leq CK^2 \left(\sqrt{\frac{n}{m}} + \frac{n}{m}\right) ||\Sigma||$$

Antes da Prova:

Detalhes da norma

$||X||_{L^p} = \mathbb{E}[|X|^p]^{\frac{1}{p}}$

$|| < X, x > ||^2_{L^2} = \mathbb{E}[< X, x >^2]$, mas $< X, x >^2 = (X^T x)^2 = x^T X X^T x$, logo:

$\mathbb{E}[< X, x >^2] = \mathbb{E}[x^T X X^T x] = x^T \mathbb{E}[XX^T] x =< \mathbb{E}[XX^T]x, x >$, então caso $\mathbb{E}[XX^T] = I$ vamos ter $|| < X, x > ||_{L^2} = ||x||_2$

Def. (Norma sub gaussiana de variável aleatória sub gaussiana)

Se X é variável aleatória sub gaussiana então $||X||_{\Psi_2} = \inf \left\{ t > 0 : \mathbb{E}\left[exp\left(\frac{X^2}{t^2} \right) \right] \right\}$

Def. (Norma sub gaussiana de vetor aleatório sub gaussiano)

Se X é vetor aleatório sub gaussiano de $\mathbb{R}^n$ $||X||_{\Psi_2} = \sup_{x \in S^{n-1}} || < X, x > ||_{\Psi_2}$

Def. (Vetor aleatório isotrópico)

Um vetor aleatório X em $\mathbb{R}^n$ é chamado isotrópico se $\Sigma(X) = \mathbb{E}[XX^T] = I_n$.

Prova: Primeiro vamos trazer os vetores $X, X_1, ..., X_n$ para posição isotrópica. Existem vetores aleatórios isotrópicos independentes $Z, Z_1, ..., Z_n$ tais que:

$$X = \Sigma^{\frac{1}{2}} Z \text{ e } X_i = \Sigma^{\frac{1}{2}} Z_i$$

(Isso é checado no exercício 3.2.2)

A partir de $|| < X, x > ||_{\Psi_2} \leq K || < X, x > ||_{L_2}$ e $X = \Sigma^{\frac{1}{2}} Z$ temos:

$|| < \Sigma^{\frac{1}{2}} Z, x > ||_{\Psi_2} \leq K || < \Sigma^{\frac{1}{2}} Z, x > ||_{L_2}$ e como $\Sigma^{\frac{1}{2}}$ é simétrico podemos passar para o outro lado, obtendo $|| < Z, \Sigma^{\frac{1}{2}} x > ||_{\Psi_2} \leq K || < Z, \Sigma^{\frac{1}{2}} x > ||_{L_2}$. Se $y = \Sigma^{\frac{1}{2}} x$ então $|| < Z, y > ||_{\Psi_2} \leq K || < Z, y > ||_{L_2} = K||y||_2$, pois $\mathbb{E}[ZZ^T] = I$.

Tomando o $\sup_{y \in S^{n-1}}$ e pela definição de vetores aleatórios sub-gaussianos temos: $||Z||_{\Psi_2} \leq K$ e analogamente $||Z_i||_{\Psi_2} \leq K$.

Então:

$||\Sigma_m - \Sigma|| = ||\Sigma^{\frac{1}{2}} R_m \Sigma^{\frac{1}{2}} \leq ||R_m|| \cdot ||\Sigma||$ onde $R_m := \frac{1}{m} \sum_{i=1}^{m} Z_i Z_i^T - I_n$

Considere a matriz aleatória $A_{m,n}$ cujas linhas são Z_i^T. Então:

$\frac{1}{m} A^T A - I_n = \frac{1}{m} \sum_{i=1}^{m} Z_i Z_i^T - I_n = R_m$

Podemos aplicar o teorema 4.6.1 (Mais precisamente o exercício 4.6.2) para A e obtemos:

$\mathbb{E}(||R_m||) \leq CK^2 \left(\sqrt{\frac{n}{m}} + \frac{n}{m} \right)$

Agora, voltando para: $||\Sigma_m - \Sigma|| \leq ||R_m|| \cdot ||\Sigma||$, tirando a esperança dos dois lados obtemos:

$$\mathbb{E}||\Sigma - \Sigma_m|| \leq CK^2 \left(\sqrt{\frac{n}{m}} + \frac{n}{m} \right) ||\Sigma||$$

■

Observação: O teorema implica que para qualquer $\epsilon \in (0,1)$, temos garantido de ter uma estimação de covariância com um bom erro, $\mathbb{E}||\Sigma_m - \Sigma|| \leq \epsilon||\Sigma||$ se tivermos o tamanho da amostra $m \approx \epsilon^{-2}n$

Em outras palavras, a matriz de covariância pode ser estimada precisamente pela matriz de covariância amostral se o tamanho da amostra m for proporcional a dimensão n.

Exe. 19: Cheque que para qualquer $u \leq 0$, temos:

$$||\Sigma - \Sigma_m|| \leq CK^2 \left(\sqrt{\frac{n+u}{m}} + \frac{n+u}{m} \right) ||\Sigma||$$

com probabilidade pelo menos $1 - 2 \cdot exp(-u)$

$||\Sigma_m - \Sigma|| = ||\Sigma^{\frac{1}{2}} R_m \Sigma^{\frac{1}{2}} \leq ||R_m|| \cdot ||\Sigma||$ onde $R_m := \frac{1}{m} \sum_{i=1}^{m} Z_i Z_i^T - I_n$

Considere a matriz aleatória $A_{m,n}$ cujas linhas são Z_i^T. Então:
$\frac{1}{m} A^T A - I_n = \frac{1}{m} \sum_{i=1}^{m} Z_i Z_i^T - I_n = R_m$

Podemos aplicar o teorema 4.6.1 para A e obtemos:
$||R_m|| = \left|\left| \frac{1}{m} A^T A - I_n \right|\right| \leq K^2 \max(\delta, \delta^2)$, com probabilidade pelo menos $1 - 2 \cdot exp(-t^2)$ onde $\delta = C \left(\sqrt{\frac{n}{m}} + \frac{t}{\sqrt{m}} \right)$

Botando $t = \sqrt{u}$ temos que $\delta = C \left(\sqrt{\frac{n}{m}} + \sqrt{\frac{u}{m}} \right)$ e $\delta^2 = C^2 \left(\frac{n+u}{m} + \frac{2\sqrt{nu}}{m} \right)$

Usando $\sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)}$, $\left(\sqrt{\frac{n}{m}} + \sqrt{\frac{u}{m}} \right) = \sqrt{2} \left(\sqrt{\frac{n+u}{m}} \right)$

Usando $2\sqrt{ab} \leq a + b$, $\left(\frac{n+u}{m} + \frac{2\sqrt{nu}}{m} \right) \leq 2 \left(\frac{n+u}{m} \right)$

$\max(\delta, \delta^2) \leq \delta + \delta^2 \leq C\sqrt{2} \left(\sqrt{\frac{n+u}{m}} \right) + C^2 2 \left(\frac{n+u}{m} \right)$, fazendo $C' = \max(C\sqrt{2}, C^2 2)$, temos:

$||R_m|| \leq K^2 \max(\delta, \delta^2) \leq K^2 C' \left(\sqrt{\frac{n+u}{m}} + \frac{n+u}{m} \right)$, substituindo:

$||\Sigma_m - \Sigma|| \leq K^2 C' \left(\sqrt{\frac{n+u}{m}} + \frac{n+u}{m} \right) \cdot ||\Sigma||$

■

6.1 Aplicação: Clusterização de conjunto de pontos

Vamos ilustrar o teorema com uma aplicação de clustering. Como na aplicação anterior, vamos tentar identificar clusters nos dados. Mas a natureza dos dados vai ser diferente, ao invez de redes, vamos trabalhar com conjuntos de pontos em $\mathbb{R}^n$. O objetivo geral vai ser particionar o conjunto de pontos dado em alguns clusters. O

que exatamente constitui um cluster não está bem definido na ciência de dados. Mas o senso comum sugere que pontos no mesmo cluster tendam a estar mais próximos uns dos outros do que pontos de clusters diferentes.

Assim como fizemos para redes, vamos criar um modelo probabilístico básico de conjuntos de pontos em $\mathbb{R}^n$ com duas comunidades e estudar o problema de clusterização para esse modelo

Def. (Modelo de mistura Gaussiana)

Gere m pontos aleatórios em $\mathbb{R}^n$ da seguinte forma: Jogue uma moeda honesta; se cair cara amostre um ponto de $N(\mu, I_n)$ e se cair coroa, de $N(-\mu, I_n)$. Essa distribuição de pontos é chamada Modelo de mistura gaussiano com médias μ e $-\mu$.

Equivalentemente, podemos considerar um vetor aleatório $X = \theta\mu + g$ onde θ é uma variável aleatória Bernoulli, $g \in N(0, I_n)$, e θ e g são independentes. Pegue uma amostra $X_1, ..., X_m$ de vetores aleatórios independentes que são distribuídos identicamente a X. Então a amostra será distribuída de acordo com o modelo de mistura gaussiana.

Suponha que temos uma amostra de m pontos distribuídos de acordo com o modelo de mistura gaussiana. Nosso objetivo é identificar qual ponto pertence a qual cluster. Para isso, podemos usar uma variante do algoritmo de clusterização espectral que introduzimos para redes.

Para ver por que um espectral tem chance de funcionar aqui, note que a distribuição de X não é isotrópica mas sim esticado na direção de μ. Assim, podemos calcular aproximadamente μ calculando o primeiro componente principal dos dados. Após isso, podemos projetar os pontos projetar os pontos na linha gerada por μ e assim classificá-los apenas olhando de que lado da origem as projeções estão. Isso leva ao seguinte algoritmo:

Algoritmo Clusterização Espectral:

Input: pontos $X_1, ..., X_m$ de $\mathbb{R}^n$

Output: Partição dos pontos em 2 clusters

1. Compute a matriz de covariância amostral $\Sigma_m = \frac{1}{m}\sum_{i=1}^{m} X_i X_i^T$

2. Compute o autovetor $v = v_1(\Sigma_m)$ correspondente ao maior autovalor de Σ_m

3. Particione os vértices em duas comunidades de acordo com os sinais do produto interno de v com os dados. (Para ser específico, se $< v, X_i >> 0$ ponha o ponto X_i na primeira comunidade, caso contrário na segunda)

Teo. 8 (Garantias da Clusterização espectral do modelo de mistura gaussiano)

Seja $X_1, ..., X_m$ pontos de $\mathbb{R}^n$ amostrados do modelo de mistura gaussiano como acima, i.e. existem duas comunidades com médias $-\mu$ e μ.

Seja $\epsilon > 0$ tal que $||\mu||_2 \geq C\sqrt{log\left(\frac{1}{\epsilon}\right)}$. Suponha que o tamanho da amostra

satisfaz

$$m \geq \left(\frac{n}{||\mu||_2} \right)^c$$

onde $c > 0$ é uma constante absoluta apropriada.

Então com probabilidade de pelo menos $1 - 4e^{-n}$, o algoritmo de clusterização espectral acima identifica as comunidades corretamente com até ϵm vértices classificados erradamente.

Exe. 20: Prove o Teorema 8 para o algoritmo de clusterização espectral aplicado ao modelo de mistura gaussiana.

a) Compute a matriz de covariância

Agora começamos o estudo do artigo CLUSTERING IN BLOCK MARKOV CHAINS

---

7 Introdução para clusterização BLOCK MARKOV CHAINS

7.1 Notação

Para quaisquer dois conjuntos $A, B \subseteq V \triangleq \{1, ..., n\}$, definimos sua **diferença simétrica** por $A \Delta B = (A \setminus B) \cup (B \setminus A)$. Para quaisquer dois números $a, b \in \mathbb{R}$, introduzimos a notação abreviada $a \wedge b = \min\{a, b\}$ e $a \vee b = \max\{a, b\}$.

Para qualquer matriz $m \times n$, $A \in \mathbb{R}^{m \times n}$, indicamos suas linhas por $A_{r,\cdot}$ para $r = 1, ..., m$ e suas colunas por $A_{\cdot,c}$ para $c = 1, ..., n$. Também introduzimos a notação abreviada $A_{A,B} = \sum_{x \in A} \sum_{y \in B} A_{x,y}$ para todos os subconjuntos $A, B \subseteq V$.

Definimos o **simplex de probabilidade** de dimensão $n-1$ por $\Delta^{n-1} = \{x \in [0,1]^n : \|x\|_1 = 1\}$, bem como o conjunto de **matrizes estocásticas** (linhas somam 1) por $\mathcal{S}_{n \times n} = \{(x_{r,c}) \in [0,1]^{n \times n} : \sum_{c=1}^{n} x_{r,c} = 1 \text{ para } r = 1, ..., n\}$ similarmente.

Em nossas análises assintóticas, escrevemos $f(n) \sim g(n)$ se $\lim_{n \to \infty} f(n)/g(n) = 1$, $f(n) = o(g(n))$ se $\lim_{n \to \infty} f(n)/g(n) = 0$ e $f(n) = O(g(n))$ se $\limsup_{n \to \infty} f(n)/g(n) < \infty$.

Sempre que $\{X_n\}_{n=1}^{\infty}$ é uma sequência de variáveis aleatórias de valor real e $\{a_n\}_{n=1}^{\infty}$ uma sequência determinística, escrevemos

$$\begin{aligned} X_n = o_{\mathbb{P}}(a_n) &\iff P\left(\left|\frac{X_n}{a_n}\right| \geq \delta\right) \to 0 \quad \forall \delta > 0 \\ &\iff \forall \varepsilon, \delta > 0, \exists N_{\varepsilon,\delta} : P\left(\left|\frac{X_n}{a_n}\right| \geq \delta\right) \leq \varepsilon \quad \forall n > N_{\varepsilon,\delta} \end{aligned}$$

e

$$X_n = O_{\mathbb{P}}(a_n) \iff \forall \varepsilon > 0, \exists \delta_\varepsilon, N_\varepsilon : P\left(\left|\frac{X_n}{a_n}\right| \geq \delta_\varepsilon\right) \leq \varepsilon \quad \forall n > N_\varepsilon.$$

Similarmente, $X_n = \Omega_P(a_n)$ denota $\forall \varepsilon > 0, \exists \delta_\varepsilon, N_\varepsilon : P[|X_n/a_n| \leq \delta_\varepsilon] \leq \varepsilon \quad \forall n > N_\varepsilon$, e $X_n \asymp_{\mathbb{P}} (a_n)$ significa $\forall \varepsilon > 0, \exists \delta_\varepsilon^-, \delta_\varepsilon^+, N_\varepsilon : P[\delta_\varepsilon^- \leq |X_n/a_n| \leq \delta_\varepsilon^+] \geq 1 - \varepsilon \quad \forall n > N_\varepsilon$.

7.2 Cadeias de Markov em Blocos (BMCs)

Assumimos que temos n estados $V = \{1, ..., n\}$, cada um dos quais está associado a um de K clusters, ou seja, o conjunto de estados é particionado tal que $V = \bigcup_{k=1}^{K} V_k$ com $V_k \cap V_l = \emptyset$ para todo $k \neq l$. Seja $\sigma(.)$ um função que associa o estado $v \in V$ ao seu cluster. Também assumimos que existem constantes $\alpha \in \Delta^{K-1}$ tais que $\lim_{n \to \infty} |V_k|/(n\alpha_k) = 1$, ou seja, α_i representa a proporção de estados no cluster i.

Para quaisquer $\alpha \in \Delta^{K-1}$ e $p \in \mathcal{S}_{K\times K}$ (o conjunto de matrizes estocásticas $K \times K$), definimos a BMC $\{X_t\}_{t\geq 0}$ da seguinte forma. Sua matriz de transição $P \in \mathcal{S}_{n\times n}$ será definida como

$$P_{x,y} = \frac{p_{\sigma(x),\sigma(y)}}{|V_{\sigma(y)}| - \mathbf{1}_{[\sigma(x)=\sigma(y)]}} \cdot \mathbf{1}_{[x\neq y]} \quad \text{para todo } x, y \in V, \tag{1}$$

Aqui $p_{\sigma(x),\sigma(y)}$ é a entrada da matriz $p \in \mathcal{S}_{K\times K}$ que representa as probabilidades de transição do cluster de x para o cluster de y. $|V_{\sigma(y)}|$ é a quantidade de elementos do cluster de y e devemos subtrair 1 caso x também seja do mesmo cluster pois um estado não vai para si mesmo. $\mathbf{1}_{[x\neq y]}$ é pelo menos motivo de que um estado não vai para si mesmo.

Note que esta cadeia de Markov não é necessariamente reversível. Adicionalmente, assumimos que K, α, p são fixos, e que estudamos o regime assintótico $n \to \infty$. Assumimos que o menor cluster tem um tamanho que cresce linearmente com n: $\alpha_{\min} \triangleq \min_k \alpha_k > 0$.

Finalmente, como estamos interessados no agrupamento (clustering) dos estados, assumiremos que

$$\exists\, \eta > 1 \text{ tal que } \max_{a,b,c}\{p_{b,a}/p_{c,a}, p_{a,b}/p_{a,c}\} \leq \eta,$$

o que garante um nível mínimo de separabilidade dos parâmetros e também que é sempre possível ir de um cluster para o outro.

7.3 Comportamento de Equilíbrio

Assumimos que a matriz estocástica p é tal que a distribuição de equilíbrio de $\{X_t\}_{t\geq 0}$ existe, e a denotaremos por Π_x para $x \in V$. Por simetria,

$$\Pi_x = \Pi_y \triangleq \bar{\Pi}_k \quad \text{para quaisquer dois estados } x, y \in V_k \text{ para todo } k = 1, ..., K.$$

Considere a quantidade escalonada

$$\pi_k \triangleq \lim_{n\to\infty} \sum_{x\in V_k} \Pi_x = \lim_{n\to\infty} |V_k|\bar{\Pi}_k \quad \text{para } k = 1, ..., K.$$

7.3.1 Proposição

A quantidade π resolve $\pi^T p = \pi^T$, e é, portanto, a distribuição de equilíbrio de uma cadeia de Markov com matriz de transição p e espaço de estados $\{1, ..., K\}$.

7.3.2 Prova

Primeiro, provamos que π é uma distribuição de probabilidade. Isso decorre de (i) definição de π, (ii) simetria de todos os estados no mesmo cluster, e (iii) porque

Π é uma distribuição de probabilidade:

$$\begin{aligned}\sum_{k=1}^{K}\pi_k &\overset{\text{(i)}}{=} \sum_{k=1}^{K}\lim_{n\to\infty}|V_k|\bar{\Pi}_k \\ &\overset{\text{(ii)}}{=} \lim_{n\to\infty}\sum_{k=1}^{K}\sum_{x\in V_k}\Pi_x = \lim_{n\to\infty}\sum_{x\in V}\Pi_x \\ &\overset{\text{(iii)}}{=} 1.\end{aligned}$$

Em seguida, mostramos que as equações de balanço são satisfeitas. Para $k = 1, \ldots, K$, segue da simetria de quaisquer dois estados $x, z \in V_k$ que $\Pi_x = \Pi_z = \bar{\Pi}_k$. Portanto, para qualquer $y \in V_l$, pelo (iv) balanço global (A chance de se estar em y é o somatório das chances de se estar em x multiplicado pela chances de ir de x para y):

$$\begin{aligned}\Pi_y = \bar{\Pi}_l &\overset{\text{(iv)}}{=} \sum_{k=1}^{K}\sum_{x\in V_k}\Pi_x P_{x,y} \\ &= \sum_{k=1}^{K}\sum_{x\in V_k}\bar{\Pi}_k\frac{p_{k,l}}{|V_l| - \mathbf{1}_{[k=l]}}\cdot\mathbf{1}_{[x\neq y]} \\ &= \sum_{k=1}^{K}\bar{\Pi}_k(|V_k| - \mathbf{1}_{[k=l]})\frac{p_{k,l}}{|V_l| - \mathbf{1}_{[k=l]}}.\end{aligned}$$

Tomando o limite $n \to \infty$, e notando que $|V_j| \to \infty$ para todo j e que $\lim_{n\to\infty}|V_k|\bar{\Pi}_k = \pi_k$, descobrimos que

$$\pi_l = \sum_{k=1}^{K}\pi_k p_{k,l} \quad \text{para todo } l = 1, \ldots, K.$$

Isso é equivalente a $\pi = \pi p$ e completa a prova.

■

Com isso podemos concluir que o comportamento de longo prazo do BMC com n estados pode ser descrito por uma cadeia de Markov muito mais simples, com apenas K estados (os clusters).

7.4 Tempo de mistura

A próxima proposição fornece um limite para o **tempo de mistura** $t_{\text{mix}} \in (0, \infty)$, que é definido por $d(t) \triangleq \sup_{x\in V}\{d_{\text{TV}}(P^t_{x,\cdot}, \mu)\}$ e $t_{\text{mix}}(\varepsilon) \triangleq \min\{t \geq 0 : d(t) \leq \varepsilon\}$, onde

$$d_{\text{TV}}(\mu, \nu) \triangleq \frac{1}{2}\sum_{x\in V}|\mu_x - \nu_x|.$$

7.4.1 Proposição

Para qualquer BMC com $n \geq 4/\alpha_{\min}$, $t_{\text{mix}}(\varepsilon) \leq -c_{\text{mix}} \ln \varepsilon$, onde $c_{\text{mix}} = -1/\ln(1 - 1/(2\eta))$.

7.4.2 Prova

Escrever prova

A Proposição acima implica que os tempos de mistura são suficientemente curtos para que nossos resultados se mantenham, independentemente de assumirmos que a cadeia de Markov está inicialmente em equilíbrio. Mostraremos mais a frente que o importante é que a cadeia atinja a estacionariedade dentro de T passos (o comprimento da trajetória observada) e, consequentemente, T precisa ser escolhido suficientemente grande em relação a n para garantir que isso ocorra. Portanto, assumimos por simplicidade que a cadeia é iniciada a partir do equilíbrio. Isso elimina a necessidade de rastrear termos de correção de ordem superior.

8 Apêndice

8.1 Cadeias de Markov

Considere um processo estocástico de tempo discreto, X_n, $n = 0, 1, 2, \ldots$, em que X_n assume valores no conjunto finito $S = \{1, \ldots, N\}$. Chamamos os valores possíveis de X_n de **estados** do sistema. Para descrever as probabilidades de tal processo, precisamos fornecer os valores de $P\{X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n\}$, para todo n e toda sequência finita de estados $(i_0, \ldots, i_n)$.

Equivalentemente, poderíamos fornecer a distribuição de probabilidade inicial $\phi(i) = \mathbb{P}\{X_0 = i\}$ e as "probabilidades de transição",

$$q_n(i_n|i_0, \ldots, i_{n-1}) = \mathbb{P}\{X_n = i_n|X_0 = i_0, \ldots, X_{n-1} = i_{n-1}\}$$

pois então

$$\mathbb{P}\{X_0 = i_0, \ldots, X_n = i_n\} = \phi(i_0)q_1(i_1|i_0)q_2(i_2|i_0, i_1)\cdots q_n(i_n|i_0, \ldots, i_{n-1}) \qquad (2)$$

8.1.1 Propriedade de Markov

O futuro e o passado são condicionalmente independentes dado o presente. Ou seja, para fazer previsões do comportamento de um sistema no futuro, é suficiente considerar apenas o estado presente do sistema e não o histórico passado.

$$\mathbb{P}\{X_{n+1} = j|X_0 = i_0, \ldots, X_{n-1} = i_{n-1}, X_n = i\} = P\{X_{n+1} = j|X_n = i\}$$

8.1.2 Homogeneidade no tempo

Uma **cadeia de Markov homogênea no tempo** é um processo tal que

$$P\{X_n = i_n|X_0 = i_0, \ldots, X_{n-1} = i_{n-1}\} = p(i_{n-1}, i_n),$$

para alguma função $p : S \times S \to [0, 1]$. Geralmente quando dizemos **Cadeia de Markov**, queremos dizer cadeia de Markov homogênea no tempo.

8.1.3 Matriz de Transição

Para fornecer as probabilidades de uma cadeia de Markov, precisamos fornecer uma distribuição de probabilidade inicial $\phi(i) = \mathbb{P}\{X_0 = i\}$ e as probabilidades de transição $p(i, j)$, pois então, por (2),

$$P\{X_0 = i_0, \ldots, X_n = i_n\} = \phi(i_0)p(i_0, i_1)p(i_1, i_2)\cdots p(i_{n-1}, i_n) \qquad (3)$$

A **matriz de transição** P para a cadeia de Markov é a matriz $N \times N$ cuja entrada (i, j), P_{ij}, é $p(i, j)$. A matriz P é uma **matriz estocástica**, i.e.,

$$0 \leq P_{ij} \leq 1, \quad 1 \leq i, j \leq N, \tag{4}$$

$$\sum_{j=1}^{N} P_{ij} = 1, \quad 1 \leq i \leq N. \tag{5}$$

Qualquer matriz que satisfaça (4) e (5) pode ser a matriz de transição para uma cadeia de Markov.

8.1.4 Distribuição estacionária

Suponha que π é um vetor de probabilidade limite, isto é, para algum vetor de probabilidade inicial v,

$$\pi = \lim_{n \to \infty} vP^n.$$

Então

$$\pi = \lim_{n \to \infty} vP^{n+1} = \left(\lim_{n \to \infty} vP^n \right) P = \pi P.$$

Chamamos um vetor de probabilidade π de uma **distribuição de probabilidade invariante** para P se

$$\pi = \pi P.$$

Tal π também pode ser chamado de distribuição de probabilidade **estacionária** ou de **equilíbrio**. Note que um vetor de probabilidade invariante é um **autovetor à esquerda** de P com **autovalor 1**.

Escrever sobre periodicidade e redutibilidade para falar quando π existe

8.1.5 Reversível e Simétrica

Uma cadeia de Markov de tempo discreto com matriz de transição P é dita **reversível** em relação a π se

$$\pi(x)P(x, y) = \pi(y)P(y, x),$$

para todos $x, y \in S$, e **simétrica** se $P(x, y) = P(y, x)$.

8.2 Teorema de Perron—Frobenius

Escrever o teorema de Perron-Frobenius

8.3 Lema de Jonson Lindenstrauss

Pode ser útil para redução de dimensionalidade

8.4 Modelo Erdös - Rényi

O modelo de bloco estocástico é na verdade uma extensão do modelo Erdös - Rényi:

$G(n, p)$: Dados n vértices, conectamos cada par de vértices distintos com probabilidade **p** de forma independente.

O grau esperado de cada vértice é $(n-1)p =: d$ e se $d \gtrsim log(n)$ (grafo denso, ou seja, o número de arestas é próximo ao número máximo de arestas) então o grafo é regular (grafo onde cada vértice tem o mesmo número de adjacências) com alta probabilidade.

8.5 Algoritmo de K-médias

K-médias é um algoritmo de clusterização que tem como objetivo minizar a variância dentro de cada cluster

Algoritmo:

1. Aleatoriamente atribua cada observação a um dos número de 1 até k. Isso vai servir de atribuição inicial de clusters para as observações.

2. Itere até as atribuições de cluster pararem de mudar:

(a) Para cada um dos clusters K, calcule o centróide do cluster.

(b) Atribua cada observação ao cluster cujo centróide está mais próximo.

8.6 Análise de componentes principais (PCA)

O autovetor u_1 correspondente ao maior autovalor s_1 define a primeira direção principal. Isto é, a direção em que a distribuição é mais estendida e explica a maior parte da variabilidade dos dados. O próximo autovalor u_2 (correspondente ao segundo maior autovalor s_2) define a próxima direção principal que melhor explica a variação restante nos dados e assim por diante.

A análise de componentes principais (PCA) computa os primeiros componentes principais e projeta os dados de $\mathbb{R}^n$ no subespaço E gerado por eles (os componentes principais). Isso reduz consideravelmente a dimensão dos dados e simplifica a análise.

Por exemplo, se somente 2 ou 3 autovalores forem grandes e sejam considerados informativos e os outros considerados ruídos então a projeção vai permitir a visualização dos dados.

8.7 Convergências

Colocar convergência em probabilidade, em distribuição e quase certamente e suas relações aqui

8.8 Desigualdades úteis

Botar as desigualdades de jensen, markov, triangular,... que vamos usar alguma hora

8.9 SVD

8.9.1 Ideia da SVD

Uma discussão extra sobre decomposição em valores singulares (SVD)

Queremos diagonalizar uma matriz $A : \mathbb{R}^n \to \mathbb{R}^m$ mas não podemos usar o teorema espectral diretamente pois $m \neq n$. Só temos uma base de vetores ortonormais que é levada em outra base de vetores ortogonais. $v_j \perp v_i \Rightarrow Av_j \perp Av_i$

$Av_i = ||Av_i|| \dfrac{Av_i}{||Av_i||} = \sigma_i u_i$, $||u_i|| = 1$, onde o σ é o valor singular

▪ Quando $n > m$, alguns vetores vão ser amassados (Vetores do núcleo), e eles tem os valores singulares 0. Mas eles nem vão importar pois vão ser amassados mesmo. Quando $m > n$ alguns vetores vão surgir que não estavam na história antes.

▪ Essas Bases v e u são únicas (exceto nas identidades), e qualquer matriz A tem uma Decomposição em valores singulares.

8.9.2 Existência de v's e u's

Seja a matriz $A : \mathbb{R}^n \to \mathbb{R}^m$, vamos supor que temos uma coleção de v_i que satisfazem $Av_i = \sigma_i u_i$, $i = 1, 2, ..., k \leq m \; ou \; n$ com $\sigma_i > 0$, também supomos:

$$\langle u_i, u_j \rangle = \langle v_i, v_j \rangle = \delta_{ij} = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases} \text{ (Os } u\text{'s e } v\text{'s são ortonormais)}$$

$A = \sum_{i=1}^{k} \sigma_i u_i v_i^T$ (SVD em forma de blocos independentes)

$Av_i = \sigma_1 u_1 v_1^T v_i + \sigma_2 u_2 v_2^T v_i + ... + \sigma_k u_k v_k^T v_i = \sigma_i u_i \to Av_i = \sigma_i u_i$, $i \leq k$

Se $k < n$: $Av_{k+1} = ... = Av_n = 0$ (ou seja quem é maior do que k está no núcleo de A e tem o valor singular 0)

▪ Também podemos escrever a SVD em forma de matriz:

$$U_{mk} = \begin{pmatrix} | & & | \\ u_1 & ... & u_k \\ | & & | \end{pmatrix}, V_{nk} = \begin{pmatrix} | & & | \\ v_1 & ... & v_k \\ | & & | \end{pmatrix}, \Sigma_{kk} = \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & \sigma_k \end{pmatrix}$$

$$\mathbf{A = U\Sigma V^T}$$

$j \leq k$: $V^T v_j = e_j \to \Sigma e_j = \sigma_j e_j \to U\sigma_j e_j = \sigma_j u_j \Rightarrow Av_j = \sigma_j u_j$
$j > k$: $V^T v_j = 0 \Rightarrow Av_j = 0$

. $V \rightarrow$ Vetores singulares à direita.

. $U \rightarrow$ Vetores singulares à esquerda.

. $\Sigma \rightarrow$ Valores singulares.

. U e V são ortogonal, ou seja, $U^TU = V^TV = I$. E Σ é Diagonal.

Obs: Por conveniência nos ordenamos os σ em uma ordem não crescente, ou seja $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_k$

Como não temos nada de errado até o momento, só precisamos encontrar quem de fato são os nossos u's, v's e σ's.

8.9.3 Encontrando os v's e u's

Até agora temos que $A = \sum_{i=1}^{k} \sigma_i u_i v_i^T = U\Sigma V^T$, logo $A^T = \sum_{i=1}^{k} \sigma_i v_i u_i^T = V\Sigma U^T$

Transpor a matriz A nos faz trocar os u's de lugar com os v's, então se temos uma decomposição singular para A, também temos uma decomposição singular para A^T.

$A^T : \mathbb{R}^m \rightarrow \mathbb{R}^n$ e $A^T u_j = \sigma_j v_j u_j^T u_j = \sigma_j v_j$, $j \leq k$

. Vamos olhar as matrizes simétricas AA^T e A^TA:

$AA^T u_j = \sigma_j A v_j = \sigma_j^2 u_j$, ou seja, os u_j são autovetores da matriz AA^T e os σ_j são as raizes quadradas dos autovalores.

$A^TAv_j = \sigma_j A^T u_j = \sigma_j^2 v_j$, ou seja, os v_j são autovetores da matriz A^TA e os σ_j são as raizes quadradas dos autovalores.

Obs:

$\langle AA^T u, u \rangle = \langle A^T u, A^T u \rangle \geq 0 \ \forall u$ (positiva-definida, ou seja, $\lambda > 0 \ \forall \lambda$) e mesma coisa para A^TA.

Como A^TA é positiva definida e simétrica, então pelo teorema espectral sempre vamos ter uma base ortonormal tal que $A^TAv_j = \lambda_j v_j$, com $\lambda_j \geq 0$

Igualmente para AA^T sempre vamos ter uma base ortonormal tal que $AA^T u_j = \lambda_j u_j$, com $\lambda_j \geq 0$

Então nossos u's v's e σ's sempre vão estar definidos.

Obs2:

Um caso particular é quando a matriz A é uma matriz simétrica. Nesse caso, os valores singulares vão ser os valores absolutos dos autovalores de A e os vetores singulares a esquerda e a direita serão ambos iguais aos autovetores de A.

8.10 Provas omitidas no livro

Teoremas e suas provas que os livros e artigos usados deixaram omitidos

8.10.1 Teorema min-max de Courant-Fisher

O Teorema Min-Max de Courant-Fischer é um resultado fundamental na álgebra linear, que oferece uma forma de caracterizar os autovalores de uma matriz simétrica real. Ele estabelece uma relação entre os autovalores de uma matriz e os subespaços vetoriais em termos de valores máximos e mínimos.

Definição: Quociente de Rayleigh

$$R_A(x) = \frac{x^T A x}{x^T x}$$

Onde A é uma matriz simétrica e x é um vetor de dimensão compatível
Se x é um autovetor de A temos que $R_A(x) = \frac{x^T A x}{x^T x} = \frac{x^T \lambda x}{x^T x} = \lambda \frac{x^T x}{x^T x} = \lambda$

O Teorema de Courant-Fischer fala que os vetores x que maximizam $R_A(x)$ são justamente os autovetores do maior autovalor de A. Na verdade, ele fornece uma caracterização de todos os autovalores de uma matriz simétrica.

Teorema (Courant-Fischer)
Seja A uma matriz simétrica com autovalores $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$, então:

$\lambda_k = max_{S \subseteq \mathbb{R}^n}\ min_{x \in S}\ R_A(x)$, onde $dim(S) = k$ e $x \neq 0$

$\lambda_k = min_{T \subseteq \mathbb{R}^n}\ max_{x \in T}\ R_A(x)$, onde $dim(T) = n - k + 1$ e $x \neq 0$
(A maximização e minimização são feitas sobre os espaços S e T de $\mathbb{R}^n$)

Prova:

8.10.2 Teorema de Eckart-Young-Mirsky

8.10.3 Teorema de Davis - Kahan