Roman Vershynin - covariance estimation

Buď $p(x)$ hustota na $\mathbb{R}^n$ a nechť X je náhodná veličina s hustotou p. Tedy X je náhodný vektor se složkami

$$X = (x_1, \dots, x_n) \in \mathbb{R}^n.$$

Typická otázka: Jsou $x_1, \dots, x_n$ nezávislé, resp. jaká je covariance matrix $Cov(x_1, \dots, x_n)$?

Označení: $Cov(X) = Cov(x_1, \dots, x_n) = \begin{pmatrix} \mathbb{E} x_1 x_1 & \cdots & \mathbb{E} x_1 x_n \\ \vdots & \ddots & \\ \mathbb{E} x_n x_1 & \cdots & \mathbb{E} x_n x_n \end{pmatrix}$

$$= \mathbb{E} \begin{pmatrix} x_1 x_1 & \cdots & x_1 x_n \\ & \ddots & \\ x_n x_1 & & x_n x_n \end{pmatrix} = \mathbb{E}\left[\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} (x_1, \dots, x_n) \right] = \mathbb{E}[X X^T]$$

Mějme dáno m pozorování $X_1, \dots, X_m$ ... tedy $X_1, \dots, X_m$ jsou kopie X.
$X_i \in \mathbb{R}^n$
NEZÁVISLÉ

Sample covariance matrix:

$$Cov_m(X) = \frac{1}{m} \sum_{i=1}^{m} X_i X_i^T$$

"Population expectation" nahrazeno "sample expectation."

Zákon velkých čísel: $Cov_m(X) \to Cov(X)$ almost surely.

... jak velké je třeba mít m, aby chyba byla malá?!

1, Bez náhodných matic

Budeme předpokládat, že X je sub-gaussovský

- momenty & tail bounds nižší nebo stejné jako pro gaussovské X
- def. např. pomocí $\mathbb{P}(|X| \geq t) \leq 2\exp(-t^2/K_1^2)$, $t \geq 0$

  nebo $\|X\|_p = (\mathbb{E}|X|^p)^{1/p} \leq K_2\sqrt{p}$, $p \geq 1$

  nebo $\mathbb{E}\exp(X^2/K_3^2) \leq 2$

- $\|X\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}\exp(X^2/t^2) \leq 2\}$ ... sub-gaussian norm
- Subgaussian vectors: $\|X\|_{\psi_2} = \sup_{\|x\|_2 \leq 1} \|\langle X, x\rangle\|_{\psi_2} = \sup_{\|x\|_2 \leq 1} \|\sum_{j=1}^n x_j X_j\|_{\psi_2}$

Věta (4.7.1. u R. Vershynin): Nechť X je sub-gaussovský náhodný vektor v $\mathbb{R}^n$. Pak ex. K (závislé na $\|X\|_{\psi_2}$) tak, že pro každé $m \geq 1$ je

$$\mathbb{E}\|\mathrm{Cov}_m(X) - \mathrm{Cov}(X)\| \leq CK^2 \cdot \left(\sqrt{\frac{n}{m}} + \frac{n}{m}\right) \|\mathrm{Cov}(X)\|.$$

"Důkaz": Založeno na odhadu vlastních (resp. singulárních) čísel a "doby před Bernsteinem":

- Nechť A je $m \times n$ matice s řádky A_i, které jsou nezávislé, centrované, sub-gaussovské a izotropické vektory v $\mathbb{R}^n$.

Pak $\sqrt{m} - CK^2(\sqrt{n} + t) \leq s_n(A) \leq s_1(A) \leq \sqrt{m} + CK^2(\sqrt{n} + t)$

s pravd. $\geq 1 - 2\exp(-t^2)$. Zde je $K := \max_i \|A_i\|_{\psi_2}$.

dokonce ... $\|\frac{1}{m}A^TA - I\| \leq K^2\max(\delta, \delta^2)$, $\delta = C\left(\sqrt{\frac{n}{m}} + \frac{t}{\sqrt{m}}\right)$ [Isotropie $\mathbb{E}[XX^T] = \mathrm{Cov}(X) = \mathrm{Id}$

- Je-li Z isotropický a centrovaný, pak $X = \mu + \Sigma^{1/2} Z$ ($\Sigma \geq 0$ fixed) má střední hodnotu μ a $Cov(X) = \Sigma$

- Vyjádříme $X_i = \Sigma^{1/2} Z_i = [Cov(X)]^{1/2} Z_i$, Z_j ... též isotrop., $\|Z_i\|_{\psi_2} \leq K$

$$\& \quad \|Cov_m(X) - Cov(X)\| = \left\| \frac{1}{m} \sum_{i=1}^{m} X_i X_i^T - Cov(X) \right\|$$

$$= \left\| \frac{1}{m} \sum_{i=1}^{m} [Cov(X)]^{1/2} Z_i Z_i^T [Cov(X)]^{1/2} - Cov(X) \right\|$$

$$= \left\| Cov(X)^{1/2} \cdot \left(\frac{1}{m} \sum_{i=1}^{m} Z_i \cdot Z_i^T \right) \cdot Cov(X)^{1/2} - Cov(X)^{1/2} \cdot Id \cdot Cov(X)^{1/2} \right\|$$

$$= \| Cov(X)^{1/2} \underbrace{\left[\frac{1}{m} \sum_{i=1}^{m} Z_i Z_i^T - Id \right]}_{R_m} Cov(X)^{1/2} \| \leq \|Cov(X)\| \cdot \|R_m\|$$

... A ... $m \times n$ mat. matice s řádky Z_i^T ... $\frac{1}{m} A^T A - I = R_m$

& $\|R_m\| \leq K^2 \max(\delta, \delta^2)$, $\delta = C\left(\sqrt{\frac{n}{m}} + \frac{t}{\sqrt{m}}\right)$ ... & integrace

• Využitím Bernsteinovy nerovnosti

Opět $X \in \mathbb{R}^n$ je náh. vektor, $Cov(X) = \mathbb{E} XX^T$

Věta (5.6.1. u R. Vershynin)

Nechť X je náhodný vektor v $\mathbb{R}^n$, $n \geq 2$. Předpokládejme, že $\|X\|_2 \leq K (\mathbb{E}\|X\|_2^2)^{1/2}$. Pak pro $m \geq 1$ platí

$$\mathbb{E}\,\|Cov_m(X) - Cov(X)\| \leq C \left(\sqrt{\frac{K^2 n \log n}{m}} + \frac{K^2 n \log n}{m} \right) \|Cov(X)\|.$$

"Důkaz": • $\mathbb{E}\|X\|_2^2 = \mathbb{E} \sum_{i=1}^n X_i^2 = \sum_{i=1}^n \mathbb{E}[X_i^2] = tr\, Cov(X)$.

$$\Rightarrow \|X\|_2^2 \leq K^2 \cdot tr(Cov(X))$$

$$\mathbb{E}\,\|Cov_m(X) - Cov(X)\| = \mathbb{E}\,\Big\| \frac{1}{m} \sum_{i=1}^m X_i X_i^T - Cov(X) \Big\|$$

$$= \mathbb{E}\,\Big\| \frac{1}{m} \underbrace{\sum_{i=1}^m \left(X_i X_i^T - Cov(X) \right)}_{\sum_{i=1}^m Z_i} \Big\| = \frac{1}{m}\, \mathbb{E}\,\Big\| \sum_{i=1}^m Z_i \Big\| \leq \frac{1}{m} \left(\sigma \sqrt{\log n} + M \log n \right)$$

kde • $\sigma^2 = \Big\| \sum_{i=1}^m \mathbb{E} \left(X_i X_i^T - Cov(X) \right)^2 \Big\| = m \,\| \mathbb{E} (XX^T - Cov(X))^2 \|$

• $M \geq \|XX^T - Cov(X)\|$ a.s.

"Stačí" vyjádřit σ^2 a M.

• M: $\|XX^T - Cov(X)\| \le \|\underbrace{XX^T}_{\text{Rank-1 matrix}}\| + \|Cov(X)\| \le \|X\|_2^2 + \underbrace{\|Cov(X)\|}_{\le tr(Cov(X))}$

$\le K^2 tr(Cov(X)) + tr Cov(X) \le 2K^2 tr(Cov(X)) \quad (K \ge 1)$

$\le 2K^2 \cdot n \|Cov(X)\|$

• G^2: $E(XX^T - Cov(X))^2 = E\{XX^T XX^T - \underbrace{XX^T}_{Cov(X)} Cov(X) - Cov(X)\underbrace{XX^T}_{Cov(X)} + Cov(X)^2\}$

$= E(XX^T)^2 - Cov(X)^2 \le E(XX^T)^2$

• $(XX^T)^2 = X \underbrace{X^T X}_{\|X\|_2^2} X^T = \|X\|_2^2 \cdot XX^T \le K^2 \cdot tr(Cov(X)) \cdot XX^T$

• $E(XX^T)^2 \le K^2 tr(Cov(X)) \cdot Cov(X)$; $\|\cdot\| \le K^2 \cdot n \cdot \|Cov(X)\|^2$

$\Rightarrow E\|Cov_m(X) - Cov(X)\| \le C \cdot \frac{1}{m}\left(\sqrt{m K^2 \cdot \|Cov(X)\|^2 \cdot n \log n} + 2K^2 \cdot n \|Cov(X)\| \cdot \log n\right)$

$= C \cdot \|Cov(X)\| \cdot \left\{\sqrt{\frac{n}{m} \cdot K^2 \cdot \log n} + \frac{n}{m} \cdot K^2 \cdot \log n\right\}$ ∎