MATH888: High-dimensional probability and statistics Fall 2021

Lecture 14 — October 8, 2021

Sebastien Roch, UW-Madison *Scribe: Nayoung Lee*

1 Overview

In the last lecture we discussed tail inequalities for the norm of sub-Gaussian vectors and obtained an error bound for sub-Gaussian vector using the Hanson-Wright inequality.

In this lecture we will discuss mean estimation of i.i.d. sub-Gaussian random vectors and derive an error bound.

2 Application to mean estimation

2.1 Review: Tail inequalities for quadratic forms of sub-Gaussian vectors

We begin by reviewing relevant concepts.

- For a symmetric matrix $\mathbf{\Sigma}$, it has eigenvalue decomposition: $\mathbf{\Sigma} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$.
- $\mathbf{\Sigma} \succeq 0$ denotes that $\mathbf{\Sigma}$ is a positive semi-definite matrix, which means $D_{i,i} \geq 0$ for all i.
- $\|\boldsymbol{B}\|_F^2 = \operatorname{tr}(\boldsymbol{B}^T\boldsymbol{B})$
- $\|\boldsymbol{B}\|_2^2 = \|\boldsymbol{B}^T\boldsymbol{B}\|_2$
- For sub-Gaussian random vector $\boldsymbol{z} \in \mathbb{R}^p$ with $\|\boldsymbol{z}\|_{\psi_2} \leq K$, $\mathbb{E}\boldsymbol{z} = 0$ and $\mathbf{\Sigma} \in \mathbb{R}^{p\times p} \succeq 0$, we have $\forall t \geq 0$:
$$\mathcal{P}(\boldsymbol{z}^T\mathbf{\Sigma}\boldsymbol{z} \geq CK^2\operatorname{tr}(\mathbf{\Sigma}) + t) \leq \exp(-\frac{ct}{K^2\|\mathbf{\Sigma}\|_2}).$$

2.2 Mean estimation of sub-Gaussian random vectors

Lemma 1. *Let $\boldsymbol{z}^{(1)}, \cdots, \boldsymbol{z}^{(n)} \in \mathbb{R}^p$ be independent, mean zero, sub-Gaussian vectors. Then we have*
$$\|\sum_{i=1}^n \alpha_i \boldsymbol{z}^{(i)}\|_{\psi_2}^2 \lesssim \sum_{i=1}^n \alpha_i \|\boldsymbol{z}^{(i)}\|_{\psi_2}^2.$$

Proof. The key idea is to project $\boldsymbol{z}$ on unit vectors.

For any $\boldsymbol{u} \in S^{p-1}$,

$$\begin{aligned}
\|\langle \boldsymbol{u}, \sum_{i=1}^{n} \alpha_i \boldsymbol{z}^{(i)} \rangle\|_{\psi_2}^2 &= \| \sum_{i=1}^{k} \alpha_i \langle \boldsymbol{u}, \boldsymbol{z}^{(i)} \rangle \|_{\psi_2}^2 \\
&\lesssim \sum_{i=1}^{n} \alpha_i^2 \|\langle \boldsymbol{u}, \boldsymbol{z}^{(i)} \rangle\|_{\psi_2}^2 \qquad \text{(by Prop 2.6.1 in [1]).}
\end{aligned}$$

By taking the supremum on the equation above, we have,

$$\begin{aligned}
\sup_{\boldsymbol{u} \in S^{p-1}} \|\langle \boldsymbol{u}, \sum_{i=1}^{n} \alpha_i \boldsymbol{z}^{(i)} \rangle\|_{\psi_2}^2 &= \sup_{\boldsymbol{u}} \| \sum_{i=1}^{n} \alpha_i \langle \boldsymbol{u}, \boldsymbol{z}^{(i)} \rangle \|_{\psi_2}^2 \\
&\lesssim \sup_{\boldsymbol{u}} \sum_{i=1}^{n} \alpha_i^2 \|\langle \boldsymbol{u}, \boldsymbol{z}^{(i)} \rangle\| \\
&\leq \sum \alpha_i^2 \sup_{\boldsymbol{u}} \|\langle \boldsymbol{u}, \boldsymbol{z}^{(i)} \rangle\|_{\psi_2}^2.
\end{aligned}$$

The last inequality uses the fact that the sum of supremum is equal or larger than the superemum of sum. Also note that the second line holds for universal constant.

□

Theorem 2. *Let $\boldsymbol{X}^{(1)}, \cdots, \boldsymbol{X}^{(n)} \in \mathbb{R}^p$ be i.i.d random vectors with $\mathbb{E}(\boldsymbol{X}^{(i)}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{X}^{(i)}) = \mathbb{E}((\boldsymbol{X}^{(i)} - \boldsymbol{\mu})(\boldsymbol{X}^{(i)} - \boldsymbol{\mu})^T) = \boldsymbol{\Sigma} \succeq 0$. Assume further that for all $\boldsymbol{w} \in \mathbb{R}^p$,*

$$\|\langle \boldsymbol{w}, \boldsymbol{X}^{(i)} - \boldsymbol{\mu} \rangle\|_{\psi_2} \leq K \sqrt{\mathbb{E}(\langle \boldsymbol{w}, \boldsymbol{X}^{(i)} - \boldsymbol{\mu} \rangle^2)}$$

holds. Let $\bar{\boldsymbol{X}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}^{(i)}$, the mean of $\boldsymbol{X}^{(1)}, \cdots \boldsymbol{X}^{(n)}$. Then, w.p $\geq 1 - \delta$, we have

$$\|\bar{\boldsymbol{X}} - \boldsymbol{\mu}\|_2^2 \preceq \frac{K^2}{n} \|\boldsymbol{\Sigma}\|_2 \left\{ \frac{\mathrm{tr}\, \boldsymbol{\Sigma}}{\|\boldsymbol{\Sigma}\|_2} + \log \frac{1}{\delta} \right\}.$$

Proof. Let $\boldsymbol{z}^{(i)} = \sqrt{\Sigma^\dagger}(\boldsymbol{X}^{(i)} - \boldsymbol{\mu})$

Recall, the eigenvalue decomposition $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$. The pseudo inverse of $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}^\dagger = \boldsymbol{U}\boldsymbol{D}^\dagger\boldsymbol{U}^T$, where $D_{ii}^\dagger$ is $1/D_{i,i}$ for $D_{i,i} > 0$ and 0 otherwise. Also, $\sqrt{\boldsymbol{\Sigma}^\dagger} = \boldsymbol{U}\sqrt{\boldsymbol{D}^\dagger}\boldsymbol{U}^T$ and $\sqrt{\boldsymbol{\Sigma}} = \boldsymbol{U}\sqrt{\boldsymbol{D}}\boldsymbol{U}$, where $\sqrt{\boldsymbol{D}^\dagger}$ and $\sqrt{\boldsymbol{D}}$ takes a square root of diagonal entries of $\boldsymbol{D}$ and $\boldsymbol{D}^\dagger$. respectively.

$\boldsymbol{X}^{(i)} = \sqrt{\boldsymbol{\Sigma}}\boldsymbol{z}^{(i)} + \boldsymbol{\mu}$ with probability 1. Therefore, $\|\bar{\boldsymbol{X}} - \boldsymbol{\mu}\|_2^2 = \bar{\boldsymbol{z}}^T \boldsymbol{\Sigma} \bar{\boldsymbol{z}}$, where $\bar{\boldsymbol{z}}$ is the average of $\boldsymbol{z}^{(i)}$.

Now it remains to bound $\|\boldsymbol{z}^{(i)}\|_{\psi_2}$. For $\boldsymbol{w} \in S^{p-1}$,

$$\begin{aligned}
\|\langle \boldsymbol{w}, \boldsymbol{z}^{(i)} \rangle\|_{\psi_2} &= \|\langle \boldsymbol{w}, \sqrt{\boldsymbol{\Sigma}^\dagger}(\boldsymbol{X}^{(i)} - \boldsymbol{\mu}) \rangle\|_{\psi_2} \\
&= \|\langle \sqrt{\boldsymbol{\Sigma}^\dagger}\boldsymbol{w}, (\boldsymbol{X}^{(i)} - \boldsymbol{\mu}) \rangle\|_{\psi_2} \qquad (\sqrt{\boldsymbol{\Sigma}^\dagger} : \text{symmetric}) \\
&\leq K \sqrt{\mathbb{E}(\langle \sqrt{\boldsymbol{\Sigma}^\dagger}\boldsymbol{w}, \boldsymbol{X}^{(i)} - \boldsymbol{\mu} \rangle^2)} \\
&\leq \cdots \\
&\leq K
\end{aligned}$$

The rest of the proof uses Lemma to show $\|\bar{\boldsymbol{z}}\|_{\psi_2}^2 \lesssim \frac{K^2}{n}$. Then plug into to $\mathcal{P}(\boldsymbol{z}^T\boldsymbol{\Sigma}\boldsymbol{z} \geq CK^2 \operatorname{tr}(\boldsymbol{\Sigma}))+t$

$\square$

Remark:

- Notice the resemblance of Theorem 2. to the mean estimation of sub-Gaussian random vectors discussed on the last lecture. The intuition is that similar to the fact that sub-Gaussian is closely related to σ^2, we now project $\boldsymbol{X}^{(i)} - \boldsymbol{\mu}$ to $\boldsymbol{w}$ and observe the standard deviation.
- $\frac{\operatorname{tr}\boldsymbol{\Sigma}}{\|\boldsymbol{\Sigma}\|_2}$ can be interpreted as a *stable rank*. The numerator $\operatorname{tr}\boldsymbol{\Sigma}$ is the sum of eigenvalues of $\boldsymbol{\Sigma}$ and the denominator $\|\boldsymbol{\Sigma}\|_2$ is the largest eigenvalue. If the eigenvalues of $\boldsymbol{\Sigma}$ are the same, the stable rank is equal to p whereas if many eigenvalues of $\boldsymbol{\Sigma}$ are close to zero, stable rank is small.

Next time, we will be looking at application to linear regression.

References

[1] R. Vershynin, *High-Dimensional Probability: An introduction with Applications in Data Science*, Cambridge University Press, 2008.