高维非渐近统计理论

王 成

上海交通大学数学科学学院

chengwang@sjtu.edu.cn

Abstract

高维数据分析是过去三十年统计领域的主要研究方向, 产生了一批新的统计理论结果和统计方法, 并成功应用到各个交叉领域. 高维数据背后的主要理论工具是概率论以及优化, 两本优秀的专著 High-dimensional probability (by Roman Vershynin)和 High-dimensional statistics (by Martin J Wainwright)对相关科学问题以及技术工具进行了系统性的介绍.

本次短期课程主要面向研究生和博士生, 我们采用一种统计专业同学更加熟悉的方式展开整个高维统计的故事. 从多元统计的角度出发, 更加深入的分析相关估计中涉及到的随机向量和随机矩阵的理论性质, 引出高维情形下产生的新的问题以及对应的高维稀疏解决办法. 希望能够通过此次课程对高维统计分析有所了解, 感兴趣同学可以继续深入系统学习相关专著.

感谢国家天元数学西北中心邀请和组织了此次短期课程, 感谢所有参与课程的同学和老师. 此讲义准备匆忙, 难免有一些错误, 欢迎批评指正.

Contents

1 预备知识

给定一组来自于某个参数分布族或者参数模型的样本$X_1, \ldots, X_n$, 经典数理统计研究的科学问题是如何基于样本构造出真实参数θ的一个统计量估计$\hat{\theta}_n$.

1.1 参数的相合估计

依据概率论中随机变量的收敛性, 我们可以定义参数估计的收敛性, 即相合估计:

- 依概率收敛(Convergence in probability):

$$\hat{\theta}_n \xrightarrow{p} \theta \iff \forall \epsilon > 0,\ \mathsf{P}(|\hat{\theta}_n - \theta| > \epsilon) \to 0.$$

- 几乎处处收敛(Almost sure convergence):

$$\begin{aligned}\hat{\theta}_n \xrightarrow{a.s.} \theta &\iff \forall \epsilon > 0,\ \mathsf{P}(\limsup_{n\to\infty}\{|\hat{\theta}_n - \theta| > \epsilon\}) = 0,\\ &\iff \forall \epsilon > 0,\ \mathsf{P}\left(\bigcap_{k=1}^{\infty}\bigcup_{n=k}^{\infty}\{|\hat{\theta}_n - \theta| > \epsilon\}\right) = 0.\end{aligned}$$

- 完全收敛(Complete convergence):

$$\forall \epsilon > 0,\ \sum_{n=1}^{\infty}\mathsf{P}(|\hat{\theta}_n - \theta| > \epsilon) < \infty.$$

统计学中, 技术条件通常假设在随机变量的矩上，即我们可以考虑

Definition 1.1 (r阶矩相合)**.** 对于参数θ的估计$\hat{\theta}_n$

$$\mathbb{E}|\hat{\theta}_n - \theta|^r \to 0.$$

尾部概率$\mathsf{P}(|\hat{\theta}_n - \theta| > \epsilon)$和$r$阶矩$\mathbb{E}|\hat{\theta}_n - \theta|^r$之间可以通过Markov's inequality建立起联系。

Lemma 1.1 (Markov's inequality)**.** 对任意非负随机变量X, $\forall x > 0$,

$$P(X \geq x) \leq \frac{\mathbb{E}X}{x}.$$

Remark 1.1. 对于任意的单调增函数$\phi : \mathbb{R}^+ \to \mathbb{R}^+$,

$$P(X \geq x) \leq P(\phi(X) \geq \phi(x)) \leq \frac{\mathbb{E}\phi(X)}{\phi(x)},$$

即

$$P(X \geq x) \leq \min_{\phi\uparrow} \frac{\mathbb{E}\phi(X)}{\phi(x)}.$$

特别的, 取$\phi(t) = I(t \geq x)$不等式可以取到等号.

1.2 大数定律(Law of large numbers, LLN)

给定独立同分布(independent and identically distributed, i.i.d.)随机变量$X_1, \ldots, X_n$, 考虑样本均值

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

样本均值$\bar{X}_n$是总体均值$\mu \overset{\text{def}}{=} \mathbb{E}(X_1)$的相合估计

- 弱大数定律: $\bar{X}_n \xrightarrow{p} \mu$,
- 强大数定律: $\bar{X}_n \xrightarrow{a.s.} \mu$.

利用Markov's inequality可以得到(弱化)大数定律的证明.

- (弱化的)弱大数定律的证明: 假定$\sigma^2 \overset{\text{def}}{=} \text{var}(X_1) < \infty$, $\forall \epsilon > 0$

$$\mathsf{P}\left(|\bar{X}_n - \mu| \geq \epsilon\right) = \mathsf{P}\left(|\bar{X}_n - \mu|^2 \geq \epsilon^2\right) \leq \frac{\mathbb{E}|\bar{X}_n - \mu|^2}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0.$$

- (弱化的)强大数定律的证明: 假定$a_4 \overset{\text{def}}{=} \mathbb{E}(X_1 - \mu)^4 < \infty$, $\forall \epsilon > 0$

$$\mathsf{P}\left(|\bar{X}_n - \mu| \geq \epsilon\right) \leq \frac{\mathbb{E}|\bar{X}_n - \mu|^4}{\epsilon^4} = \frac{n^4 a_4 + 3n(n-1)\sigma^4}{n^4\epsilon^4} = O(n^{-2}),$$

由Borel–Cantelli lemma, $\bar{X}_n \xrightarrow{a.s.} \mu$.

Question 1.1. 如果$\mathbb{E}(X_1 - \mu)^6 < \infty$, 会如何?

1.3 样本均值

考虑一个多元正态分布的简单样本

$$\mathbf{X}_1, \cdots, \mathbf{X}_n, \ i.i.d. \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

其中$\boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p} \succ 0$分别是总体均值和总体协方差矩阵.

Question 1.2. *对于样本均值:*

$$\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i,$$

如何(定义)理解$\bar{\mathbf{X}}_n$是$\boldsymbol{\mu}$的相合估计?

基于一元参数的相合估计, 我们可以考察每个分量或者任意线性投影

- 对于任意的分量$j \in \{1, \cdots, p\}$: $\bar{\mathbf{X}}_j \xrightarrow{p} \mu_j$.
- 对于任意的单位向量$\mathbf{u} \in \mathbb{R}^p$,

$$\mathbf{u}^\top \bar{\mathbf{X}} \xrightarrow{p} \mathbf{u}^\top \boldsymbol{\mu}.$$

注意, 只要$\boldsymbol{\Sigma}$的特征值有界和$n \to \infty$, 这里的结果对任意维度p都是成立的.

考察每个分量或者投影都只是局部的方式, 从全局的角度我们可以考虑向量的度量.

Definition 1.2 (向量ℓ_q norm)**.** *给定一个向量$\mathbf{x} = (x_1, \ldots, x_p)^\top$, 定义向量的$\ell_q$-norm:*

$$\|\mathbf{x}\|_q = \left(\sum_{j=1}^{p} |x_j|^q\right)^{1/q}, \ q \geq 1.$$

特别的

- ℓ_1-norm: $\|\mathbf{x}\|_1 = \sum\limits_{j=1}^{p} |x_j|$;
- ℓ_2-norm (Frobenius norm): $\|\mathbf{x}\|_2 = \sqrt{\sum\limits_{j=1}^{p} |x_j|^2}$;

- ℓ_∞-norm: $\|\mathbf{x}\|_\infty = \max\limits_{j=1,\dots,p} |x_j|$.

关于ℓ_q-norm相关的不等式包括Hölder's inequality和Minkowski's inequality.

对于多元正态分布的样本均值

$$\bar{\mathbf{X}} \sim N(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}) \iff \sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \sim N(\mathbf{0}, \boldsymbol{\Sigma}).$$

在$\boldsymbol{\Sigma} = \mathbf{I}$的情形下,

- ℓ_1-norm:

$$\mathbb{E}\|\bar{\mathbf{X}} - \boldsymbol{\mu}\|_1 = \frac{1}{\sqrt{n}}\mathbb{E}\sum_{j=1}^{p}|Z_j| = \sqrt{\frac{2}{\pi}}\frac{p}{\sqrt{n}}.$$

- ℓ_2-norm:

$$\begin{aligned}\mathbb{E}\|\bar{\mathbf{X}} - \boldsymbol{\mu}\|_2 =& \frac{1}{\sqrt{n}}\int_0^\infty \sqrt{x}\frac{1}{2^{p/2}\Gamma(p/2)}x^{p/2-1}e^{-x/2}dx \\ =& \frac{\sqrt{2}}{\sqrt{n}}\frac{\Gamma(p/2+1/2)}{\Gamma(p/2)} = \sqrt{2}\frac{\sqrt{p}(1+o(1))}{\sqrt{n}}.\end{aligned}$$

- 任意ℓ_q-norm: 一般的$q > 1$, 即使计算期望也很难得到具体结果, 猜测

$$\mathbb{E}\|\bar{\mathbf{X}} - \boldsymbol{\mu}\|_q = O(1)\frac{p^{1/q}}{\sqrt{n}}.$$

下面考虑$q \to \infty$即最大值

$$\sqrt{n}\|\bar{\mathbf{X}} - \boldsymbol{\mu}\|_\infty \overset{d}{=} \max_{j=1,\dots,p}|Z_j|,$$

其分布函数为

$$\mathsf{P}(\|\bar{\mathbf{X}} - \boldsymbol{\mu}\|_\infty \le x) = \prod_{j=1}^{p}\mathsf{P}(|Z_j| \le \sqrt{n}x) = (\Phi(\sqrt{n}x) - \Phi(-\sqrt{n}x))^p, \forall x \ge 0.$$

Lemma 1.2 (Mills ratio)**.** 对于标准正态分布的分布函数$\Phi(x)$和密度函数$\phi(x)$, $\forall x > 0$,

$$\frac{1}{x} - \frac{1}{x^3} \le \frac{1-\Phi(x)}{\phi(x)} \le \frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5}.$$

由Mills ratio, 记

$$1-\Phi(x)=\frac{c_x}{\sqrt{2\pi}}\frac{1}{x}\exp\{-\frac{x^2}{2}\},\ c_x>0, c_x\to 1,$$

所以

$$\begin{aligned}(\Phi(\sqrt{n}x)-\Phi(-\sqrt{n}x))^p&=\left(1-2(1-\Phi(\sqrt{n}x))\right)^p\\&=\exp\left\{p\log\left(1-2(1-\Phi(\sqrt{n}x))\right)\right\}\\&\approx\exp\left\{-p2\left(1-\Phi(\sqrt{n}x)\right)\right\}\\&=\exp\left\{-\frac{2c_{\sqrt{n}x}}{\sqrt{2\pi}}\frac{p}{\sqrt{n}x}\exp\{-\frac{nx^2}{2}\}\right\}.\end{aligned}$$

取

$$x=\sqrt{\frac{2\log p}{n}},$$

可得

$$\mathsf{P}\left(\|\bar{\mathbf{X}}-\boldsymbol{\mu}\|_\infty>\sqrt{\frac{2\log p}{n}}\right)=1-(\Phi(\sqrt{n}x)-\Phi(-\sqrt{n}x))^p\to 0.$$

Proposition 1.1. 对于$\mathbf{X}_1,\cdots,\mathbf{X}_n,\ i.i.d.\sim N(\boldsymbol{\mu},\mathbf{I})$的样本均值

$$\bar{\mathbf{X}}=\frac{1}{n}\sum_{i=1}^n\mathbf{X}_i,$$

可得

$$P\left(\|\bar{\mathbf{X}}-\boldsymbol{\mu}\|_\infty>\sqrt{\frac{2\log p}{n}}\right)\to 0.$$

进一步，我们考虑最大值的期望。

$$\mathbb{E}\left(\max_{j=1,\dots,p}|Z_j|\right)=\int_0^\infty\left[1-(\Phi(x)-\Phi(-x))^p\right]dx.$$

Proposition 1.2. 对于$Z_1,\cdots,Z_p,\ i.i.d.\sim N(0,1),$

$$\mathbb{E}\left(\max_{j=1,\dots,p}|Z_j|\right)=\sqrt{2\log p}+o(1).$$

对于一般的ℓ_q norm,

$$\mathbb{E}\|\bar{\mathbf{X}}-\boldsymbol{\mu}\|_q\le p^{1/q}\mathbb{E}\|\bar{\mathbf{X}}-\boldsymbol{\mu}\|_\infty\le p^{1/q}\sqrt{\frac{2\log p}{n}}+p^{1/q}\frac{o(1)}{\sqrt{n}}.$$

和我们之前猜想的只相差了一个$\sqrt{2\log p}$项.

1.4 应用：高维稀疏均值估计

对样本$\mathbf{X}_1, \cdots, \mathbf{X}_n$, $i.i.d. \sim N(\boldsymbol{\mu}, \mathbf{I})$, 考虑均值的稀疏估计

$$\begin{aligned}\hat{\boldsymbol{\mu}}(\lambda) =& \arg\min_{\mathbf{x}} \frac{1}{2n}\sum_{i=1}^{n}\|\mathbf{X}_i - \mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1 \\ =& \arg\min_{\mathbf{x}} \frac{1}{2}\|\bar{\mathbf{X}} - \mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1 \\ =& \text{soft}(\bar{\mathbf{X}}, \lambda),\end{aligned}$$

其中soft是soft-thresholding 函数,

$$\text{soft}(x,\lambda) = \text{sign}(x)\max(0, |x| - \lambda) = \text{sign}(x)(|x| - \lambda)_+ = \begin{cases} x - \lambda, & x \geq \lambda \\ 0, & |x| < \lambda \\ x + \lambda, & x \leq -\lambda. \end{cases}$$

设置$\lambda \geq \|\bar{\mathbf{X}} - \boldsymbol{\mu}\|_\infty$, 则

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_\infty \leq \|\hat{\boldsymbol{\mu}} - \bar{\mathbf{X}}\|_\infty + \|\bar{\mathbf{X}} - \boldsymbol{\mu}\|_\infty \leq 2\lambda,$$

且

$$\boldsymbol{\mu}_j = 0 \Rightarrow \hat{\boldsymbol{\mu}}_j = 0.$$

由此可得高维($\log p = o(n)$)情形下的相合估计.

Proposition 1.3 (严格稀疏). 假定总体向量$\boldsymbol{\mu}$是严格稀疏的, 即

$$\|\boldsymbol{\mu}\|_0 = \sum_{j=1}^{p} I(\boldsymbol{\mu}_j \neq 0) \leq s,$$

则

$$P\left(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_\infty \leq 2\sqrt{\frac{2\log p}{n}}\right) \to 1.$$

以及

$$P\left(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_1 \leq 2s\sqrt{\frac{2\log p}{n}}\right) \to 1, \; P\left(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \leq 2\sqrt{\frac{2s\log p}{n}}\right) \to 1.$$

Question 1.3. 尝试控制$\mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_1$.

Remark 1.2 (普适性(Universality)). 上述结果依赖于标准正态分布,

- 上述结果可否推广到一般的协方差矩阵$\boldsymbol{\Sigma}$?
- 上述结果可否推广到更大的分布族?

1.5 样本协方差矩阵

对于样本协方差矩阵(简单起见暂不考虑样本均值$\bar{\mathbf{X}}$)

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})^\top,$$

如何(定义)理解$\widehat{\mathbf{\Sigma}}_n$是$\mathbf{\Sigma}$的相合估计?从分量或者二次型角度,

- 对于任意的分量$j \in \{1, \cdots, p\}$: $\widehat{\mathbf{\Sigma}}_{i,j} \xrightarrow{p} \mathbf{\Sigma}_{i,j}$.
- 对于任意的单位向量$\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$,

$$\mathbf{u}^\top \widehat{\mathbf{\Sigma}} \mathbf{v} \xrightarrow{p} \mathbf{u}^\top \mathbf{\Sigma} \mathbf{v}.$$

Definition 1.3 (矩阵norm)**.** *对于一个矩阵 $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{p\times q}$, 常用的矩阵度量(matrix norm)*

- *the element-wise l_∞ norm* $\|\mathbf{A}\|_\infty = \max_{1\le i\le p, 1\le j\le q} |a_{ij}|$;
- *the spectral norm* $\|\mathbf{A}\| = \sup_{|\mathbf{x}|_2\le 1} |\mathbf{A}\mathbf{x}|_2$;
- *the matrix ℓ_1 norm* $\|\mathbf{A}\|_L = \max\limits_{1\le j\le q} \sum\limits_{i=1}^{p} |a_{ij}|$;
- *the Frobenius norm* $\|\mathbf{A}\|_2 = \sqrt{\sum\limits_{i=1}^{p}\sum\limits_{j=1}^{q} a_{ij}^2}$;
- *the element-wise ℓ_1 norm* $\|\mathbf{A}\|_1 = \sum\limits_{i=1}^{p}\sum\limits_{j=1}^{q} |a_{ij}|$.

Remark 1.3 (ℓ_1 norm)**.** 在高维统计中，矩阵ℓ_1 *norm* $\|\mathbf{A}\|_L$是一个常用到的度量，例如对于对称的$\mathbf{A}$

$$\|\mathbf{A}\mathbf{x}\|_\infty \le \|\mathbf{A}\|_L \|\mathbf{x}\|_\infty,$$

以及*Gershgorin Circle Theorem*

$$\|\mathbf{A}\| \le \|\mathbf{A}\|_L.$$

基于矩阵度量，可以研究样本协方差矩阵$\widehat{\mathbf{\Sigma}}$

$$\mathbb{E}\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2^2 = \frac{(\mathrm{tr}\mathbf{\Sigma})^2 + \mathrm{tr}(\mathbf{\Sigma}^2)}{n}.$$

对于其他矩阵度量计算具体结果都是非常困难的.

Remark 1.4 (普适性(Universality)). 样本协方差矩阵的分布更加复杂, 涉及到矩阵度量

- 一般条件下, 不同的矩阵度量下什么时候样本协方差矩阵是总体协方差矩阵的相合估计?
- 从向量的ℓ_∞结果可否推导得到$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty$?
- 如何得到$\mathbf{\Sigma}$的稀疏估计?
- 如何得到精度矩阵$\mathbf{\Sigma}^{-1}$的稀疏估计?
- 一般统计方法如最小二乘、线性判别分析、二次型判别分析中，如何得到稀疏的相合估计?

1.6 附录: 正态分布最大值的期望

对于$Z_1, \ldots, Z_p, i.i.d \sim N(0,1)$,

$$\begin{aligned}
&\mathbb{E}\left(\max_{j=1,\ldots,p} |Z_j|\right) = \int_0^\infty \left[1 - (\Phi(x) - \Phi(-x))^p\right] dx \\
=&y - \int_0^y (\Phi(x) - \Phi(-x))^p dx + \int_y^\infty \left[1 - (\Phi(x) - \Phi(-x))^p\right] dx
\end{aligned}$$

- $x \geq y$的时候,

$$\begin{aligned}
\int_y^\infty \left[1 - (\Phi(x) - \Phi(-x))^p\right] dx =& \int_y^\infty \left[1 - (1 - 2(1 - \Phi(x)))^p\right] dx \\
\leq& \int_y^\infty 2p(1 - \Phi(x)) dx \\
\leq& 2p \int_y^\infty \phi(x) \frac{c_x}{x} dx \leq 2p \int_y^\infty \phi(x) \frac{2}{y} dx \\
=& \frac{4p}{y}(1 - \Phi(y)) \leq \frac{4p}{y} \frac{2}{y} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{y^2}{2}\},
\end{aligned}$$

所以取

$$y \to \infty, \frac{p}{y^2} \exp\{-\frac{y^2}{2}\} \to 0.$$

- $x \in (0, y)$ 的时候,

$$\begin{aligned}(\Phi(x) - \Phi(-x))^p &= (1 - 2(1 - \Phi(x)))^p \\ &= \exp\left\{p \log\left(1 - 2(1 - \Phi(x))\right)\right\} \\ &\approx \exp\left\{-p2\left(1 - \Phi(x)\right)\right\} \\ &= \exp\left\{-\frac{2c_x}{\sqrt{2\pi}} \frac{p}{x} \exp\{-\frac{x^2}{2}\}\right\},\end{aligned}$$

所以 $\forall \epsilon > 0$

$$\begin{aligned}\int_0^y (\Phi(x) - \Phi(-x))^p dx &= \int_0^{y-\epsilon} (\Phi(x) - \Phi(-x))^p dx + \int_{y-\epsilon}^{y} (\Phi(x) - \Phi(-x))^p dx \\ &\leq (y - \epsilon)(\Phi(y - \epsilon) - \Phi(-y + \epsilon))^p + \epsilon \\ &\leq y \exp\left\{-c\frac{p}{y - \epsilon} \exp\{-\frac{(y - \epsilon)^2}{2}\}\right\} + \epsilon\end{aligned}$$

由此可以取

$$\frac{p}{y \log y} \exp\{-\frac{(y - \epsilon)^2}{2}\} \to \infty,$$

即

$$\frac{p}{y \log y} \exp\{-\frac{y^2}{2}\} \exp\{\epsilon y\} \to \infty,$$

由此, 设置$y = \sqrt{2 \log p}$, $\epsilon = 1/\log(y)$, 可得

$$\mathbb{E}\left(\max_{j=1,\ldots,p} |Z_j|\right) = \sqrt{2 \log p} + o(1).$$

2 集中不等式

在很多统计问题中, 一个重要的目标是得到一个随机变量的尾部概率界不等式, 从而保证一个随机变量与其均值或中位数的接近程度.

2.1 从Markov不等式到Chernoff界

最基本的尾部概率界是*Markov's inequality*: 对任意非负随机变量X, $\forall x > 0$,

$$\mathsf{P}(X \geq x) \leq \frac{\mathbb{E}X}{x}.$$

进一步，对于任意的单调增函数$\phi : \mathbb{R}^+ \to \mathbb{R}^+$,

$$\mathsf{P}(X \geq x) \leq \mathsf{P}(\phi(X) \geq \phi(x)) \leq \frac{\mathbb{E}\phi(X)}{\phi(x)},$$

几种常用的Markov不等式形式

- *Chebyshev*不等式 :

$$\mathsf{P}(|X - \mu| \geq t) \leq \frac{\mathrm{var}(X)}{t^2} \quad \forall t > 0.$$

- 基于高阶矩的Markov不等式

$$\mathsf{P}(|X - \mu| \geq t) \leq \frac{\mathbb{E}\left(|X - \mu|^k\right)}{t^k} \quad \forall t > 0.$$

- Chernoff界: 对随机变量$Y = e^{\lambda(X-\mu)}$运用Markov不等式

$$\log \mathsf{P}\left((X - \mu) \geq t\right) \leq \inf_{\lambda \in [0,b]} \left\{ \log \mathbb{E}\left[e^{\lambda(X-\mu)}\right] - \lambda t \right\}.$$

Remark 2.1. *在最优的k下, 基于高阶矩的Markov不等式不会比Chernoff界中基于矩母函数获得的界更差. 尽管如此, 在实际应用中, 由于矩母函数计算的便利性, Chernoff界仍有着广泛的应用.*

2.2 次Gaussian随机变量和Hoeffding界

Example 2.1 (Gaussian分布尾部界)**.** *令$X \sim N\left(\mu, \sigma^2\right)$, 简单的计算可得$X$的矩母函数*

$$\mathbb{E}\left[e^{\lambda X}\right] = e^{\mu\lambda + \frac{\sigma^2\lambda^2}{2}}, \qquad \forall \lambda \in \mathbb{R}.$$

代入最优*Chernoff*界

$$\inf_{\lambda\geq 0}\left\{\log \mathbb{E}\left[e^{\lambda(X-\mu)}\right]-\lambda t\right\}=\inf_{\lambda\geq 0}\left\{\frac{\lambda^2\sigma^2}{2}-\lambda t\right\}=-\frac{t^2}{2\sigma^2},$$

由此可得上偏差不等式*:*

$$P(X\geq \mu+t)\leq e^{-\frac{t^2}{2\sigma^2}}\qquad \forall t\geq 0. \tag{1}$$

对比*Mills ratio,* 这个界是除了多项式修正项之外是最优的*.*

由Gaussian分布启发, 定义一般的sub-Gaussian分布*.*

Definition 2.1 (sub-Gaussian分布)**.** *若存在正实数*σ使得

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right]\leq e^{\sigma^2\lambda^2/2}\qquad \forall\lambda\in\mathbb{R}, \tag{2}$$

那么称这个均值为$\mu=\mathbb{E}[X]$*的随机变量*X*是次*Gaussian*的. 常数*σ*被称作次*Gaussian*参数.*

若X*是参数为*σ*的次*Gaussian*随机变量, 那么它满足上偏差不等式*(1)*. 同样*$-X$*也是参数为*σ*的次*Gaussian*随机变量, 结合可得到对任意次*Gaussian*随机变量*X *的集中不等式*

$$\mathsf{P}\left(|X-\mu|\geq t\right)\leq 2e^{-\frac{t^2}{2x^2}}\qquad \forall t>0.$$

Definition 2.2 (次Gaussian随机变量定义的等价性)**.** *对任意均值为*0*的随机变量*X*,下面的性质是等价的:*

1. *矩母函数: 存在常数*$\sigma\geq 0$*使得*

$$\mathbb{E}\left[e^{\lambda X}\right]\leq e^{\frac{\lambda^2\sigma^2}{2}}\qquad \forall\lambda\in\mathbb{R}.$$

2. *尾部概率: 存在常数*$K_2>0$

$$P(|X|\geq t)\leq 2\exp(-t^2/K_2^2)\quad \forall t\geq 0.$$

3. *各阶矩: 存在常数*$K_3>0$

$$\left(\mathbb{E}|X|^k\right)^{1/k}\leq K_2\sqrt{k},\ \forall k\geq 1.$$

4. X^2矩母函数:存在常数$K_4 > 0$

$$\mathbb{E}\exp(\lambda^2 X^2) \leq \exp(K_4^2\lambda^2), \forall 0 \leq \lambda \leq \frac{1}{K_4}.$$

5. X^2矩母函数:存在常数$K_5 > 0$

$$\mathbb{E}\exp(X^2/K_4^5) \leq 2.$$

非正态的次Gaussian随机变量包括:

- 一个Rademacher随机变量ε是指等概率取$\{-1, 1\}$的随机变量. 它是一个参数为$\sigma = 1$的次Gaussian随机变量.
- 设X是均值为0,支撑集为某个区间$[a, b]$的随机变量. 是一个参数为$b - a$的次Gaussian随机变量, 更复杂的推导可以知道X是参数为$\sigma \leq \frac{b-a}{2}$的次Gaussian随机变量.

类似于正态随机变量在线性运算下还是正态随机变量, 次Gaussian随机变量同样有类似的性质. 例如, X_1和X_2是独立的次Gaussian随机变量,对应的参数分别为σ_1和σ_2, 那么$X_1 + X_2$则是参数 为$\sqrt{\sigma_1^2 + \sigma_2^2}$的次Gaussian随机变量. 对于独立的次Gaussian随机变量的和, 可得到一个重要的结果，也就是*Hoeffding*界 :

Proposition 2.1 (Hoeffding界)**.** 假定次*Gaussian*随机变量$X_i, i = 1, \ldots, n$相互独立, 其对应的均值和次*Gaussian*参数分别为μ_i及σ_i. 那么, 对任意的$t > 0$, 我们有

$$P\left[\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right] \leq \exp\left\{-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right\}.$$

2.3 次指数随机变量和Bernstein界

次Gaussian一般用来研究均值, 进一步如果研究方差就需要更一般的分布. 在正态分布框架中，正态分布的平方对应卡方分布，自由度为2的卡方分布正好是指数分布. 我们接下来研究次指数随机变量, 其对矩母函数的要求相对更加宽松:

Definition 2.3 (次指数分布)**.** 对于一个均值为零的随机变量X, 称之为次指数随机变量, 如果下列任意一条件成立:

1. 尾部概率*:* 存在常数K_1, 使得

$$P(|X| \geq t) \leq 2\exp(-t/K_1),\ \forall t \geq 0.$$

2. 各阶矩*:* 存在常数K_2, 使得

$$\left(\mathbb{E}|X|^k\right)^{1/k} \leq K_2 k,\ \forall k \geq 1.$$

3. 矩母函数*:* 存在常数K_3, 使得

$$\mathbb{E}\exp(\lambda|X|) \leq \exp(K_3\lambda),\ \forall 0 \leq \lambda \leq \frac{1}{K_3}.$$

4. 矩母函数*:* 存在常数K_4,

$$\mathbb{E}\exp(|X|/K_4) \leq 2.$$

5. 矩母函数*:* 存在常数K_5,

$$\mathbb{E}\exp(\lambda X) \leq \exp(K_5^2\lambda^2),\quad \forall|\lambda| \leq \frac{1}{K_5}.$$

按照定义: 所有的次Gaussian随机变量都是次指数的的. 然而, 次指数随机变量并不一定是次Gaussian的, 卡方分布就是一个例子:

Example 2.2 (非次Gaussian的次指数随机变量)**.** 令$Z \sim \mathcal{N}(0,1)$, 考虑随机变量$X = Z^2$. 对$\lambda < \frac{1}{2}$, 我们有

$$\mathbb{E}\left[e^{\lambda(X-1)}\right] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{\lambda\left(z^2-1\right)}e^{-z^2/2}dz = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}}.$$

对$\lambda > \frac{1}{2}$, 矩母函数是发散的, 即说明X不是次*Gaussian*的. 通过简单的计算得知 :

$$\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{2\lambda^2} = e^{4\lambda^2/2}, \qquad \forall|\lambda| < \frac{1}{4}. \tag{3}$$

同次Gaussian性类似, 对于次指数分布, 结合Chernoff界的技巧, 就可以得到次指数随机变量对应的偏差和集中不等式. 当t充分小的时候, 这些界本质上是次Gaussian的(即指数部分为t^2阶); 而当t较大的时候, 这个界的指数部分会与t成线性关系.

Proposition 2.2 (次指数的Chernoff界)**.** 对于均值为零的次指数随机变量

$$\mathbb{E}\exp(\lambda|X|) \le \exp(K^2\lambda^2), \quad \forall|\lambda| \le \frac{1}{K},$$

可得集中不等式:

$$P(|X| \ge t) \le \begin{cases} \exp\left(-\frac{t^2}{4K^2}\right) & 0 \le t \le 2K \\ \exp\left(1-\frac{t}{K}\right) & t > 2K. \end{cases},$$

就像次Gaussian性一样,独立次指数随机变量的和同样保持次指数性, 由此可得Bernstein's inequality.

Theorem 2.1 (Bernstein不等式)**.** 假设$X_1, \ldots, \mathbf{X}_n$是均值为零, 独立同分布的次指数随机变量,

$$\mathbb{E}\exp(\lambda X_i) \le \exp(\sigma_i^2\lambda^2), \quad \forall|\lambda| \le \frac{1}{\sigma_i}.$$

那么, $\forall t \ge 0$

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right| \ge t\right) \le \begin{cases} 2\exp\left(-\frac{n^2t^2}{4\sum_{i=1}^n \sigma_i^2}\right), & 0 \le t \le \frac{2\sum_{i=1}^n \sigma_i^2}{n\max_i \sigma_i} \\ 2\exp\left(\frac{\sum_{i=1}^n \sigma_i^2}{\max_i \sigma_i^2} - \frac{nt}{\max_i \sigma_i}\right), & t > \frac{2\sum_{i=1}^n \sigma_i^2}{n\max_i \sigma_i}. \end{cases}$$

实际使用时候, $t = o(1)$, 即我们使用平方项目, 从而对于次指数分布也可以得到类似于次Gaussian分布的Hoeffding界结果.

Question 2.1. 对于多元正态分布$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$的样本协方差矩阵$\widehat{\boldsymbol{\Sigma}}$, 尝试利用*Bernstein*不等式分析每个元素的界, 即

$$\hat{\Sigma}_{ij} - \Sigma_{ij} = \mathbf{e}_i^\top(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\mathbf{e}_j.$$

3 最大特征值的概率界

3.1 引例: 最大元素的尾部概率

由次Gaussian随机变量和次指数的性质, 我们可以控制独立和随机变量的尾部概率(Hoeffding不等式和Bernstein不等式)。利用这些尾部概率，通过简单求和即可得到随机向量(多元样本均值)和随机矩阵(样本协方差矩阵)的ℓ_∞-norm的刻画.

Example 3.1 (多元样本均值的ℓ_∞-norm)**.** 对于一组独立的多元样本$\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^p$, 考虑总体均值$\boldsymbol{\mu} \overset{def}{=} \mathbb{E}\mathbf{X}_1$的样本均值

$$\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i$$

假设每个分量都是参数为*1*的次*Guassian*随机变量, 即

$$\mathbb{E}\left[e^{\lambda(X_{1j}-\mu_j)}\right] \le e^{\frac{\lambda^2}{2}} \qquad \forall \lambda \in \mathbb{R},\ j = 1, \cdots, p.$$

利用*Hoeffding*界,

$$P\left[|\bar{X}_j - \mu_j| \ge t\right] \le 2\exp\left\{-\frac{nt^2}{2}\right\},\ j = 1, \cdots, p.$$

由此可得

$$P\left[\|\bar{\mathbf{X}} - \boldsymbol{\mu}\|_\infty \ge t\right] \le \sum_{j=1}^{p} P\left[|\bar{X}_j - \mu_j| \ge t\right] \le 2p\exp\left\{-\frac{nt^2}{2}\right\}$$

设置

$$t = \sqrt{c\frac{2\log p}{n}},\ c > 1,$$

可得

$$P\left[\|\bar{\mathbf{X}} - \boldsymbol{\mu}\|_\infty \ge \sqrt{c\frac{2\log p}{n}}\right] \le 2p^{1-c},$$

即

$$\|\bar{\mathbf{X}} - \boldsymbol{\mu}\|_\infty = O_p\left(\sqrt{\frac{2\log p}{n}}\right).$$

注意: 从普适性*(Universality)*的角度, 相比于标准多元正态分布*($\boldsymbol{\Sigma} = \mathbf{I}$)*的结果, 这里把结果推广到了次*Guassian*分布, 同时不需要分量之间的独立性.

Example 3.2 (样本协方差矩阵的ℓ_∞-norm)**.** 对于一组独立的多元样本$\mathbf{X}_1,\ldots,\mathbf{X}_n\in\mathbb{R}^p$, 考虑总体协方差矩阵$\mathbf{\Sigma}\overset{def}{=}cov(\mathbf{X}_1)$的样本协方差矩阵

$$\widehat{\mathbf{\Sigma}}=\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i-\boldsymbol{\mu})(\mathbf{X}_i-\boldsymbol{\mu})^\top.$$

假设每个分量都是参数为1的次*Guassian*随机变量, 即

$$\mathbb{E}\left[e^{\lambda(X_{1j}-\mu_j)}\right]\le e^{\frac{\lambda^2}{2}}\qquad\forall\lambda\in\mathbb{R},\ j=1,\cdots,p.$$

利用*Bernstein*界,

$$P\left[|\hat{\Sigma}_{ij}-\Sigma_{ij}|\ge t\right]\le 2\exp\left(-\frac{nt^2}{8}\right),\quad\forall t\in(0,1).$$

由此可得

$$P\left[\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|_\infty\ge t\right]\le\sum_{i,j=1}^{p}P\left[|\hat{\Sigma}_{ij}-\Sigma_{ij}|\ge t\right]\le 2p^2\exp\left(-\frac{nt^2}{8}\right),\quad\forall t\in(0,1).$$

设置

$$t=\sqrt{c\frac{8\log p}{n}}\in(0,1),\ c>2$$

可得

$$P\left[\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|_\infty\ge\sqrt{c\frac{8\log p}{n}}\right]\le 2p^{2-c},$$

即$\log p=o(n)$时

$$\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|_\infty=O_p\left(\sqrt{\frac{\log p}{n}}\right).$$

Remark 3.1. 对于带有样本均值的协方差矩阵

$$\mathbf{S}_n=\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i-\bar{\mathbf{X}})(\mathbf{X}_i-\bar{\mathbf{X}})^\top=\widehat{\mathbf{\Sigma}}+(\bar{\mathbf{X}}-\boldsymbol{\mu})(\bar{\mathbf{X}}-\boldsymbol{\mu})^\top,$$

由此, $\log p=o(n)$时

$$\|\mathbf{S}_n-\mathbf{\Sigma}\|_\infty\le\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|_\infty+\|\bar{\mathbf{X}}-\boldsymbol{\mu}\|_\infty^2=O_p\left(\sqrt{\frac{\log p}{n}}\right)+O_p\left(\frac{\log p}{n}\right)=O_p\left(\sqrt{\frac{\log p}{n}}\right).$$

Example 3.3 (ℓ_∞-norm的期望)**.** 上述通过对尾部概率求和的方式不能用来得到期望的结果, 例如

$$\begin{aligned}\mathbb{E}\left[\|\bar{\mathbf{X}}-\boldsymbol{\mu}\|_\infty\right] =& \int_0^\infty P\left[\|\bar{\mathbf{X}}-\boldsymbol{\mu}\|_\infty \geq t\right] dt \leq \int_0^\infty 2p\exp\left\{-\frac{nt^2}{2}\right\} dt \\ =&\sqrt{2\pi}\frac{p}{\sqrt{n}} \quad (\approx \mathbb{E}\left[\|\bar{\mathbf{X}}-\boldsymbol{\mu}\|_1\right]).\end{aligned}$$

这里可以利用更加高级的*Chernoff*界的技巧, 由*Jessen*不等式

$$\begin{aligned}\exp\left(t\mathbb{E}\left[\|\bar{\mathbf{X}}-\boldsymbol{\mu}\|_\infty\right]\right) \leq& \mathbb{E}\exp\left[t\|\bar{\mathbf{X}}-\boldsymbol{\mu}\|_\infty\right] = \mathbb{E}\max_j \exp\left[t|\bar{X}_j-\mu_j|\right] \\ \leq& \sum_{j=1}^p \mathbb{E}\exp\left[t|\bar{X}_j-\mu_j|\right] \leq 2p\exp\left(\frac{t^2}{2n}\right), \forall t>0,\end{aligned}$$

注意这里要控制的是一个*sub-Gaussian*随机变量的绝对值,所以多了一个常数项*2*. 由此

$$\mathbb{E}\left[\|\bar{\mathbf{X}}-\boldsymbol{\mu}\|_\infty\right] \leq \min_{t>0}\frac{\log(2p)+\frac{t^2}{2n}}{t} = \sqrt{\frac{2\log p+2\log 2}{n}}.$$

特别的, 对于正态分布，这里的结果之前的也是吻合的.

3.2 最大特征值表示

上述求最大值需要用到和控制一系列随机变量，这些变量由一个有限的指标集来标记, 例如

$$\|\bar{\mathbf{X}}-\boldsymbol{\mu}\|_\infty = \max_{j=1,\ldots,p}|\mathbf{e}_j^\top(\bar{\mathbf{X}}-\boldsymbol{\mu})|, \quad \|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_\infty = \max_{i,j=1,\ldots,p}|\mathbf{e}_i^\top(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})\mathbf{e}_j|,$$

这里$\mathbf{e}_i \in \mathbb{R}^p$表示第$i$个元素为1, 其余为0的向量. 相关结果可以通过集合元素个数乘以尾部概率来实现控制, 最终结果依赖于集合基数的的对数, 即$\log|\mathcal{S}|$.

如果指标集包含了无穷个元素, 例如样本协方差矩阵的最大特征值,

$$\|\widehat{\boldsymbol{\Sigma}}\| = \sup_{\mathbf{u}\in\mathcal{S}^{p-1}}\|\widehat{\boldsymbol{\Sigma}}\mathbf{u}\|_2 = \sup_{\mathbf{u}\in\mathcal{S}^{p-1}}\mathbf{u}^\top\widehat{\boldsymbol{\Sigma}}\mathbf{u},$$

这里的指标集为Euclidean球面$\mathcal{S}^{p-1}$上的向量, 是一个无穷集合. 上述计算最大元素的方式就不能直接使用.

有限集的大小可以通过它的基数来度量, 对含有无限多元素的集合度量其"大小"则需要非常谨慎. 度量熵的概念为解决这个问题提供了一种方法, 这个概念可以追溯到Kolmogorov, Tikhomirov及其他俄罗斯学派成员的开创性工作. 尽管是以度量空间的包装和覆盖的形式在非随机的意义下被定义的, 度量熵对于理解随机过程的性质是非常重要的. 这里简单起见我们主要关注单位球面的"大小".

3.3 $\epsilon-$nets

Definition 3.1 ($\epsilon-$net)**.** 考虑一个度量空间(X,d), $\forall\epsilon>0$, X的一个子集X_ϵ称为X的$\epsilon-net$,

$$\forall x\in X,\exists y\in X_\epsilon,\text{使得 } d(x,y)\le\epsilon.$$

X_ϵ的最小基数$N(X,\epsilon)$称为X的覆盖数*(covering number)*

$\epsilon-$net相当于在无穷个元素的集合中选取有限个点来代表所有元素, 例如

- 对于$[0,1]$区间, 选取欧式距离$d(x,y)=|x-y|$, 则$N([0,1],1/n)\le n$;
- 对于单位正方形$[0,1]^2$, 选取欧式距离$d(\mathbf{x},\mathbf{y})=\|\mathbf{x}-\mathbf{y}\|_2$, 则$N\left([0,1],\sqrt{2}/n\right)\le n^2$;

Lemma 3.1 (单位球面的覆盖数)**.** 对于p维欧式空间的单位球面$\mathcal{S}^{p-1}$, 考虑欧式距离下的$\epsilon-net$, 我们有

$$N(\mathcal{S}^{p-1},\epsilon)\le\left(1+\frac{2}{\epsilon}\right)^p.$$

利用单位球面的$\epsilon-$net, 我们可以通过有限个指标集上的度量来控制最大特征值

Proposition 3.1. 对于对称矩阵$\mathbf{A}\in\mathbb{R}^{p\times p}$, 记$\mathcal{S}_\epsilon$为单位球面$\mathcal{S}^{p-1}$的$\epsilon-net$, 其中$\epsilon\in[0,1/2)$, 则

$$\sup_{\mathbf{x}\in\mathcal{S}_\epsilon}|\mathbf{x}^\top\mathbf{A}\mathbf{x}|\le\|\mathbf{A}\|=\sup_{\mathbf{x}\in\mathcal{S}^{p-1}}|\mathbf{x}^\top\mathbf{A}\mathbf{x}|\le(1-2\epsilon)^{-1}\sup_{\mathbf{x}\in\mathcal{S}_\epsilon}|\mathbf{x}^\top\mathbf{A}\mathbf{x}|$$

Proof. 记矩阵$\mathbf{A}$绝对值最大的特征值对应特征向量为$\mathbf{x}\in\mathcal{S}^{p-1}$, 即

$$\|\mathbf{A}\|=|\mathbf{x}^\top\mathbf{A}\mathbf{x}|.$$

由$\epsilon-$net的定义, $\exists y\in\mathcal{S}_\epsilon$使得 $\|\mathbf{x}-\mathbf{y}\|_2\le\epsilon$. 利用三角不等式

$$\|A\|-\mathbf{y}^\top\mathbf{A}\mathbf{y}\le|\mathbf{x}^\top\mathbf{A}\mathbf{x}-\mathbf{y}^\top\mathbf{A}\mathbf{y}|=|\mathbf{x}^\top\mathbf{A}(\mathbf{x}-\mathbf{y})+\mathbf{y}^\top\mathbf{A}(\mathbf{x}-\mathbf{y})|$$
$$\le\|\mathbf{A}\mathbf{x}\|_2\|\mathbf{x}-\mathbf{y}\|_2+\|\mathbf{A}\mathbf{y}\|_2\|\mathbf{x}-\mathbf{y}\|_2\le2\epsilon\|\mathbf{A}\|,$$

即

$$\|\mathbf{A}\|\le(1-2\epsilon)^{-1}\mathbf{y}^\top\mathbf{A}\mathbf{y}\le(1-2\epsilon)^{-1}\sup_{\mathbf{y}\in\mathcal{S}_\epsilon}|\mathbf{y}^\top\mathbf{A}\mathbf{y}|.$$

□

3.4 最大特征值的尾部概率

下面我们控制样本协方差矩阵和总体协方差矩阵的谱范数, 由Proposition 3.1,

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| \leq (1-2\epsilon)^{-1} \sup_{\mathbf{x}\in\mathcal{S}_\epsilon} |\mathbf{x}^\top(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{x}|,$$

而

$$\mathbf{x}^\top(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{x} = \frac{1}{n}\sum_{i=1}^n \left[\left(\mathbf{x}^\top(\mathbf{X}_i - \boldsymbol{\mu})\right)^2 - \mathbb{E}\left(\mathbf{x}^\top(\mathbf{X}_i - \boldsymbol{\mu})\right)^2\right].$$

为了控制上述独立和随机变量的尾部概率, 我们可以假定$\mathbf{x}^\top(\mathbf{X}_i - \boldsymbol{\mu})$满足次Gaussian性质, 即向量型次Gaussian分布.

Definition 3.2 (向量次Gaussian分布)**.** $\forall \mathbf{x} \in \mathcal{S}^{p-1}$, 定义$\mathbf{X} \in \mathbb{R}^{p-1}$满足

$$\mathbb{E}\left[e^{\lambda \mathbf{x}^\top(\mathbf{X} - \mathbb{E}\mathbf{X})}\right] \leq e^{\frac{\lambda^2\sigma^2}{2}} \qquad \forall \lambda \in \mathbb{R},$$

称为参数为σ^2的向量次*Gaussian*分布.

假定$\mathbf{X}_1, \ldots, \mathbf{X}_n$满足参数为1的向量次Gaussian分布, 则

$$\mathsf{P}\left[\left|\mathbf{x}^\top(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{x}\right| \geq t\right] \leq 2\exp\left(-\frac{nt^2}{8}\right), \quad \forall t \in (0,1),$$

结合$\epsilon-$net的覆盖数, 可得

$$\begin{aligned}
\mathsf{P}\left[\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| \geq t\right] \leq & \mathsf{P}\left[(1-2\epsilon)^{-1} \sup_{\mathbf{x}\in\mathcal{S}_\epsilon} |\mathbf{x}^\top(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{x}| \geq t\right] \\
= & \mathsf{P}\left[\sup_{\mathbf{x}\in\mathcal{S}_\epsilon} |\mathbf{x}^\top(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{x}| \geq (1-2\epsilon)t\right] \\
\leq & \sum_{\mathbf{x}\in\mathcal{S}_\epsilon} \mathsf{P}\left[\left|\mathbf{x}^\top(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{x}\right| \geq t\right] \\
\leq & 2\left(1 + \frac{2}{\epsilon}\right)^p \exp\left(-\frac{n(1-2\epsilon)^2t^2}{8}\right), \quad \forall t \in (0,1),
\end{aligned}$$

从而

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| = O_p\left(\sqrt{\frac{p}{n}}\right), \ \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2 \leq \sqrt{p}\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| = O_p\left(\frac{p}{\sqrt{n}}\right).$$

3.5 应用: PCA的相合性

在主成分分析(Principal Component Analysis, PCA)方法中, 通过样本协方差矩阵$\widehat{\mathbf{\Sigma}}$的特征向量来估计总体协方差矩阵$\mathbf{\Sigma}$的的特征向量. 由Davis–Kahan Theorem (Yu et al., 2015), 真实的降维空间与估计出来的降维空间之间的夹角距离可以通过$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|$来控制. 所以在一定条件下$p/n \to 0$, 可得到PCA方法的相合性.

对于其他涉及到样本协方差矩阵(或其逆矩阵)和样本均值的统计方法, 相合性分析可以通过谱范数或者欧式距离来实现, 即在$p = o(n)$或者$p = o(\sqrt{n})$的条件下具有相合性.

3.6 延伸: 随机矩阵与最大特征值

最大特征值的随机性质是数学、物理、统计、信号处理等多个领域关注的问题. 随机矩阵领域关于样本协方差矩阵最大特征值有很多更加精细的结果, 例如Tracy–Widom distribution, Bai-Yin law等. 特别的，对于最大特征值的期望, 统计物理中可以利用Replica Trick计算出更加精细的结果. 感兴趣读者可以参阅相关材料文献.

4 均值和协方差矩阵的稀疏估计

考虑一个多元样本$\mathbf{X}_1, \cdots, \mathbf{X}_n \in \mathbb{R}^p$, 本节主要关注高维情形下总体均值和总体协方差矩阵的相合估计

$$\boldsymbol{\mu} \stackrel{\text{def}}{=} \mathbb{E}\left[\mathbf{X}_1\right], \ \boldsymbol{\Sigma} = \operatorname{cov}(\mathbf{X}_1) = \mathbb{E}(\mathbf{X}_1 - \boldsymbol{\mu})(\mathbf{X}_1 - \boldsymbol{\mu})^\top.$$

基于样本，我们可以得到样本均值

$$\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i,$$

和样本协方差矩阵

$$\mathbf{S}_n = \frac{1}{n}\sum_{i=1}^n \left(\mathbf{X}_i - \bar{\mathbf{X}}\right)\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)^\top.$$

4.1 $\ell_\infty-$norm下的相合估计

高维情形下, 尽管有维数带来的误差积累等问题, 在$\ell_\infty-$norm下样本均值和样本协方差矩阵是相合估计.

Proposition 4.1. 对于独立同分布样本$\mathbf{X}_1, \cdots, \mathbf{X}_n \in \mathbb{R}^p$, 假定每个分量是次*Guassian*随机变量

$$\mathbb{E}\left[e^{\lambda(X_{1j}-\mu_j)}\right] \le e^{\frac{\lambda^2\sigma^2}{2}}, \qquad \forall \lambda \in \mathbb{R}, \ j = 1, \cdots, p,$$

则存在仅依赖于σ^2的常数c_1, c_2使得当$\log(p) = o(n)$时候

$$P\left(\|\bar{\mathbf{X}} - \boldsymbol{\mu}\|_\infty \le c_1\sqrt{\frac{\log p}{n}}\right) \to 1, \ P\left(\|\mathbf{S}_n - \boldsymbol{\Sigma}\|_\infty \le c_2\sqrt{\frac{\log p}{n}}\right) \to 1.$$

4.2 稀疏均值估计

对于$\boldsymbol{\mu}$, 考虑稀疏估计

$$\hat{\boldsymbol{\mu}}(\lambda) = \operatorname{soft}(\bar{\mathbf{X}}, \lambda).$$

Proposition 4.2 (严格稀疏均值估计)**.** 假定总体向量$\boldsymbol{\mu}$是严格稀疏的, 即

$$\|\boldsymbol{\mu}\|_0 = \sum_{j=1}^p I(\boldsymbol{\mu}_j \neq 0) \le s,$$

则设置$\lambda = c_1\sqrt{\frac{\log p}{n}}$可得

$$\mathbb{P}\left(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_\infty \le 2c_1\sqrt{\frac{\log p}{n}}\right) \to 1.$$

以及

$$\mathbb{P}\left(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_1 \le 2c_1 s\sqrt{\frac{\log p}{n}}\right) \to 1,\ \mathbb{P}\left(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \le 2c_1\sqrt{\frac{s\log p}{n}}\right) \to 1.$$

Proof. 在事件

$$\|\bar{\mathbf{X}} - \boldsymbol{\mu}\|_\infty \le c_1\sqrt{\frac{\log p}{n}}$$

成立时,

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_\infty \le \|\bar{\mathbf{X}} - \boldsymbol{\mu}\|_\infty + \|\bar{\mathbf{X}} - \boldsymbol{\mu}\|_\infty \le 2c_1\sqrt{\frac{\log p}{n}}.$$

注意到

$$\mu_j = 0 \Rightarrow |\bar{X}_j| \le c_1\sqrt{\frac{\log p}{n}} \Rightarrow \hat{\mu}_j = \mathrm{soft}\left(\bar{X}_j, c_1\sqrt{\frac{\log p}{n}}\right) = 0,$$

因此

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_1 = \sum_{j:\mu_j\neq 0}|\hat{\mu}_j - \mu_j| + \sum_{j:\mu_j = 0}|\hat{\mu}_j - \mu_j| = \sum_{j:\mu_j\neq 0}|\hat{\mu}_j - \mu_j| \le s\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_\infty,$$

以及

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2^2 \le \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_1\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_\infty \le s\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_\infty^2.$$

□

Proposition 4.3 (渐近稀疏均值估计)**.** 假定总体向量$\boldsymbol{\mu}$是渐近稀疏的, 即存在$q \in [0, 1)$

$$\|\boldsymbol{\mu}\|_q^q = \sum_{j=1}^p |\boldsymbol{\mu}_j|^q \le s_q,$$

则设置$\lambda = 2c_1\sqrt{\frac{\log p}{n}}$可得

$$\mathbb{P}\left(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_\infty \le 3c_1\sqrt{\frac{\log p}{n}}\right) \to 1.$$

以及

$$\mathbb{P}\left(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_1 \le 3c_1 s_q \left(\sqrt{\frac{\log p}{n}}\right)^{1-q}\right) \to 1,\ \mathbb{P}\left(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \le 3c_1\sqrt{s_q}\left(\sqrt{\frac{\log p}{n}}\right)^{1-q/2}\right) \to 1.$$

Proof. 在事件

$$\|\bar{\mathbf{X}} - \boldsymbol{\mu}\|_\infty \le c_1\sqrt{\frac{\log p}{n}}$$

发生的条件下,

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_\infty \le \|\bar{\mathbf{X}} - \boldsymbol{\mu}\|_\infty + \|\bar{\mathbf{X}} - \boldsymbol{\mu}\|_\infty \le 3c_1\sqrt{\frac{\log p}{n}},$$

以及

$$|\mu_j| \le c_1\sqrt{\frac{\log p}{n}} \Rightarrow |\bar{X}_j| \le 2c_1\sqrt{\frac{\log p}{n}} \Rightarrow \hat{\mu}_j = 0.$$

由此可得

$$\begin{aligned}
\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_1 = & \sum_{j:|\mu_j| \le c_1\sqrt{\frac{\log p}{n}}} |\hat{\mu}_j - \mu_j| + \sum_{j:|\mu_j| > c_1\sqrt{\frac{\log p}{n}}} |\hat{\mu}_j - \mu_j| \\
\le & \sum_{j:|\mu_j| \le c_1\sqrt{\frac{\log p}{n}}} |\mu_j| + \sum_{j:|\mu_j| > c_1\sqrt{\frac{\log p}{n}}} 3c_1\sqrt{\frac{\log p}{n}} \\
\le & \sum_{j:|\mu_j| \le c_1\sqrt{\frac{\log p}{n}}} |\mu_j|^q \left(c_1\sqrt{\frac{\log p}{n}}\right)^{1-q} + \sum_{j:|\mu_j| > c_1\sqrt{\frac{\log p}{n}}} 3|\mu_j|^q \left(c_1\sqrt{\frac{\log p}{n}}\right)^{1-q} \\
\le & 3c_1 s_q \left(\sqrt{\frac{\log p}{n}}\right)^{1-q}
\end{aligned}$$

以及

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2^2 \le \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_1 \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_\infty \le 3^2 c_1^2 s_q \left(\sqrt{\frac{\log p}{n}}\right)^{2-q}.$$

□

Remark 4.1 (capped-ℓ_1 稀疏). 另外一种严格稀疏的弱化形式是*Capped-ℓ_1 sparsity(Zhang and Zhang, 2012),*

$$\sum_{j=1}^{p} \min\left(1, \frac{|\mu_j|}{c_1\sqrt{\frac{\log p}{n}}}\right) \leq s,$$

此时上述两项也是可以控制的

$$\begin{aligned}
\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_1 \leq & \sum_{j:|\mu_j| \leq c_1\sqrt{\frac{\log p}{n}}} |\mu_j| + \sum_{j:|\mu_j| > c_1\sqrt{\frac{\log p}{n}}} 3c_1\sqrt{\frac{\log p}{n}} \\
= & \sum_{j:|\mu_j| \leq c_1\sqrt{\frac{\log p}{n}}} \min\left(1, \frac{|\mu_j|}{c_1\sqrt{\frac{\log p}{n}}}\right) c_1\sqrt{\frac{\log p}{n}} + 3c_1\sqrt{\frac{\log p}{n}} \sum_{j:|\mu_j| > c_1\sqrt{\frac{\log p}{n}}} \min\left(1, \frac{|\mu_j|}{c_1\sqrt{\frac{\log p}{n}}}\right) \\
\leq & 3c_1 s\sqrt{\frac{\log p}{n}},
\end{aligned}$$

从而可以得到严格稀疏类似的结果.

4.3 稀疏协方差估计

对于总体协方差矩阵$\boldsymbol{\Sigma}$的稀疏估计, 可以对样本协方差矩阵$\mathbf{S}_n$进行截断, 即考虑估计

$$\widehat{\boldsymbol{\Sigma}}(\lambda) = \text{soft}(\mathbf{S}_n, \lambda).$$

对于矩阵的每一行(或者每一列)进行类似于稀疏均值估计的理论分析, 可以得到

$$\max_j \|\widehat{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\|_\infty = \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty, \ \max_j \|\widehat{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\|_1 = \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{L_1}, \max_j \|\widehat{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\|_2$$

的相关结果. 结合Gershgorin Circle Theorem, 可以得到谱范数的界. 综上, 对于严格稀疏和渐近稀疏我们有下述结果.

Proposition 4.4 (严格稀疏协方差估计). 假定总体协方差$\boldsymbol{\Sigma}$是严格稀疏的, 即

$$\max_{i=1,\ldots,p} \sum_{j=1}^{p} I(\Sigma_{ij} \neq 0) \leq s,$$

则设置$\lambda = c_2\sqrt{\frac{\log p}{n}}$可得

$$\mathbb{P}\left(\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty \leq 2c_2\sqrt{\frac{\log p}{n}}\right) \to 1.$$

以及

$$\mathbb{P}\left(\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| \leq 2c_2 s\sqrt{\frac{\log p}{n}}\right) \to 1,\ \mathbb{P}\left(\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2/p \leq 2c_1\sqrt{\frac{s\log p}{n}}\right) \to 1.$$

Proposition 4.5 (渐近稀疏协方差估计)**.** 假定总体协方差$\mathbf{\Sigma}$是渐近稀疏的, 即存在$q \in [0,1)$

$$\max_{i=1,\ldots,p}\sum_{j=1}^{p}|\Sigma_{ij}|^q \leq s_q,$$

则设置$\lambda = 2c_2\sqrt{\frac{\log p}{n}}$可得

$$\mathbb{P}\left(\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty \leq 3c_2\sqrt{\frac{\log p}{n}}\right) \to 1.$$

以及

$$\mathbb{P}\left(\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| \leq 3c_2 s_q\left(\sqrt{\frac{\log p}{n}}\right)^{1-q}\right) \to 1,\ \mathbb{P}\left(\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2/p \leq 3c_2\sqrt{s_q}\left(\sqrt{\frac{\log p}{n}}\right)^{1-q/2}\right) \to 1.$$

4.4 相关文献

Bickel and Levina (2008)最早考虑了对样本协方差矩阵进行截断来得到稀疏估计, 并且证明了所得估计在高维($\log(p) = o(n)$)情形下是谱范数相合的. 进一步的, Rothman et al. (2009)考虑了一般形式的截断, 包括了Hard-thresholding, soft-thresholding以及SCAD惩罚下的截断等等. Cai and Liu (2011a)考虑了对每一行采用不同的截断参数从而提升估计的精度。

统计方法角度, Shao et al. (2011)和Li and Shao (2015)的两个工作把截断的稀疏均值估计和稀疏协方差矩阵估计代入了线性判别分析和二次型判别分析中, 从理论上分析了所得分类器可以达到Bayes错判率.

5 高维稀疏线性回归

本节我们考虑回归问题, 样本

$$(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_n, Y_n), i.i.d. \sim (\mathbf{X}, Y),$$

其中$\mathbf{X} \in \mathbb{R}^p$为解释变量, $Y \in \mathbb{R}$为响应变量. 从总体角度, 最优线性投影

$$\begin{aligned}\boldsymbol{\beta}^* &= \arg\min_{\beta \in \mathbb{R}^p} \mathbb{E}\left[(Y - \mathbb{E}Y) - \boldsymbol{\beta}^\top (\mathbf{X} - \mathbb{E}\mathbf{X})\right]^2 \\ &= \arg\min_{\beta \in \mathbb{R}^p} \left[\boldsymbol{\beta}^\top \mathbb{E}(\mathbf{X} - \mathbb{E}\mathbf{X})(\mathbf{X} - \mathbb{E}\mathbf{X})^\top \boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \mathbb{E}(\mathbf{X} - \mathbb{E}\mathbf{X})(Y - \mathbb{E}Y)\right] = \boldsymbol{\Sigma}^{-1}\mathbf{a},\end{aligned}$$

其中

$$\boldsymbol{\Sigma} = \mathbb{E}(\mathbf{X} - \mathbb{E}\mathbf{X})(\mathbf{X} - \mathbb{E}\mathbf{X})^\top = \mathrm{cov}(\mathbf{X}),\ \mathbf{a} = \mathbb{E}(\mathbf{X} - \mathbb{E}\mathbf{X})(Y - \mathbb{E}Y) = \mathrm{cov}(\mathbf{X}, Y).$$

基于样本$(\mathbf{X}_i, Y_i)$, 可以得到样本协方差矩阵和样本协方差向量

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top, \widehat{\mathbf{a}} = \frac{1}{n}\sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(Y_i - \bar{Y})$$

其中$\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i$和$\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$分别为解释变量和响应变量的样本均值. 由此平方损失可以表述为

$$\frac{1}{2n}\sum_{i=1}^n \left[(Y_i - \bar{Y}) - \boldsymbol{\beta}^\top (\mathbf{X}_i - \bar{\mathbf{X}})\right]^2 = \frac{1}{2}\boldsymbol{\beta}^\top \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \boldsymbol{\beta}^\top \widehat{\mathbf{a}} + \frac{1}{2n}\sum_{i=1}^n (Y_i - \bar{Y})^2.$$

在经典统计中(p固定, n趋于无穷), 这时候样本协方差矩阵$\widehat{\boldsymbol{\Sigma}}$是可逆的, 可得到经典最小二乘估计

$$\begin{aligned}&\arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2n}\sum_{i=1}^n \left[(Y_i - \bar{Y}) - \boldsymbol{\beta}^\top (\mathbf{X}_i - \bar{\mathbf{X}})\right]^2 \\ &= \arg\min_{\beta \in \mathbb{R}^p} \left(\frac{1}{2}\boldsymbol{\beta}^\top \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \boldsymbol{\beta}^\top \widehat{\mathbf{a}}\right) = \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathbf{a}}.\end{aligned}$$

对于高维情形(例如$p \gg n$), 因为样本协方差矩阵不可逆从而最小二乘估计不再适用, 方法上需要引入新的稀疏线性回归方法. 从理论角度, 一定条件下ℓ_∞-norm下样本协方差估计还是相合的, 从这种无穷范数下的相合性能否推导出稀疏线性回归方法的相合性, 也是高维稀疏回归的关键科学问题.

5.1 LASSO

Tibshirani (1996)在最小二乘法的基础上引入ℓ_1惩罚, 提出了LASSO方法(Least Absolute Shrinkage and Selection Operator),

$$\min_{\beta\in\mathbb{R}^p}\frac{1}{2n}\sum_{i=1}^{n}\left[(Y_i-\bar{Y})-\boldsymbol{\beta}^\top(\mathbf{X}_i-\bar{\mathbf{X}})\right]^2+\lambda\|\boldsymbol{\beta}\|_1,$$

等价的可表示为

$$\widehat{\boldsymbol{\beta}}_{\text{lasso}}(\lambda)=\arg\min_{\beta\in\mathbb{R}^p}\left(\frac{1}{2}\boldsymbol{\beta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}-\boldsymbol{\beta}^\top\widehat{\mathbf{a}}+\lambda\|\boldsymbol{\beta}\|_1\right),$$

这里$\lambda>0$是一个调节参数.

假定真实参数$\boldsymbol{\beta}^*$的支撑集$\mathcal{S}=\{i:\beta_i^*\neq 0\}$元素不是特别多, 可以得到严格稀疏下的LASSO相合性.

Theorem 5.1 (LASSO相合性)**.** 假设

$$\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_\infty\le c\sqrt{\frac{\log p}{n}},\ \|\widehat{\mathbf{a}}-\mathbf{a}\|_\infty\le c\sqrt{\frac{\log p}{n}},\ \|\boldsymbol{\beta}\|_0\|\boldsymbol{\Sigma}_{\mathcal{S},\mathcal{S}}^{-1}\|_L\sqrt{\frac{\log p}{n}}\to 0,$$

以及*Irrepresentable Condition*

$$\|\boldsymbol{\Sigma}_{\mathcal{S}^c,\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S},\mathcal{S}}^{-1}\|_L<1-\alpha,\ \alpha>0.$$

设置

$$\lambda=\frac{3c}{\alpha}\sqrt{\frac{\log p}{n}}(1+\|\boldsymbol{\beta}^*\|_1),$$

可得

- 支撑集相合性*:* $\widehat{\boldsymbol{\beta}}_{lasso}(\mathcal{S}^c)=\mathbf{0}$.
- 无穷范数相合性*:*

$$\|\widehat{\boldsymbol{\beta}}_{lasso}-\boldsymbol{\beta}^*\|_\infty\le 2c(1+\frac{3}{\alpha})\|\boldsymbol{\Sigma}_{\mathcal{S},\mathcal{S}}^{-1}\|_L(1+\|\boldsymbol{\beta}^*\|_1)\sqrt{\frac{\log p}{n}}.$$

从理论结果可以看出通过设置合适的调节参数, LASSO方法可以有效的去除噪音变量, 估计效果几乎等价于在已知支撑集的条件下的最小二乘法, 两者制相差了一个$\sqrt{\log p}$项.

5.2 Dantzig Selector

对于LASSO估计, 考虑KKT(Karush–Kuhn–Tucker)条件,

$$\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\beta}}_{\text{lasso}} - \widehat{\mathbf{a}} + \lambda \cdot \text{sign}\left(\widehat{\boldsymbol{\beta}}_{\text{lasso}}\right) = \mathbf{0},$$

即LASSO估计满足性质

$$\|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \widehat{\mathbf{a}}\|_\infty \leq \lambda.$$

Candes and Tao (2007)提出Dantzig Selector(下文简写DS)方法

$$\widehat{\boldsymbol{\beta}}_{\text{ds}}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\arg\min} \|\boldsymbol{\beta}\|_1, \text{ subject to } \|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \widehat{\mathbf{a}}\|_\infty \leq \lambda.$$

理论上, 对于同样的调节参数λ, 一定有$\|\widehat{\boldsymbol{\beta}}_{\text{ds}}\|_1 \leq \|\widehat{\boldsymbol{\beta}}_{\text{lasso}}\|_1$, 即在$\ell_1$范数下, DS估计比LASSO估计更加稀疏. 理论分析过程中, 这种更加稀疏的特性也会大大简化DS的证明过程, 例如这里我们可以考虑更为宽松的渐近稀疏情形.

Theorem 5.2 (Dantzig Selector)**.** 当

$$\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty \leq c\sqrt{\frac{\log p}{n}}, \ \|\widehat{\mathbf{a}} - \mathbf{a}\|_\infty \leq c\sqrt{\frac{\log p}{n}}$$

成立时, 设置

$$\lambda = c\sqrt{\frac{\log p}{n}}(1 + \|\boldsymbol{\beta}^*\|_1),$$

可得

$$\|\widehat{\boldsymbol{\beta}}_{ds} - \boldsymbol{\beta}\|_\infty \leq 2c\|\boldsymbol{\Sigma}^{-1}\|_L(1 + \|\boldsymbol{\beta}^*\|_1)\sqrt{\frac{\log p}{n}}.$$

进一步假定$\boldsymbol{\beta}^*$是渐近稀疏的,

$$\|\widehat{\boldsymbol{\beta}}_{ds} - \boldsymbol{\beta}\|_1 \leq 4\|\boldsymbol{\beta}^*\|_q^q c^{1-q}\|\boldsymbol{\Sigma}^{-1}\|_L^{1-q}(1 + \|\boldsymbol{\beta}^*\|_1)^{1-q}\left(\sqrt{\frac{\log p}{n}}\right)^{1-q},$$

$$\|\widehat{\boldsymbol{\beta}}_{ds} - \boldsymbol{\beta}\|_2 \leq 3\sqrt{\|\boldsymbol{\beta}^*\|_q^q} c^{1-q/2}\|\boldsymbol{\Sigma}^{-1}\|_L^{1-q/2}(1 + \|\boldsymbol{\beta}^*\|_1)^{1-q/2}\left(\sqrt{\frac{\log p}{n}}\right)^{1-q/2}.$$

5.3 相关文献

高维线性回归是整个高维数据分析的最核心问题, 除了LASSO和Dantzig Selector还有其他一些方法, 这里不再展开. 关于LASSO的证明, Zou (2006); Zhao and Yu (2006); Wainwright (2009)等最早取得了理论上的突破. 除了这里的不可表示条件, 分析LASSO支撑集相合性还有一些其他的条件, Van De Geer and Bühlmann (2009)对这些条件进行了详细的比较. 进一步的，Zhang and Zhang (2012)利用基本不等式研究了渐近稀疏情形.

关于Dantzig Selector的理论性质, Candes and Tao (2007)原始论文研究的是严格稀疏情形, Cai et al. (2011)在研究精度矩阵估计时候考虑了Dantzig Selector估计的渐近稀疏情形, 这里介绍的正是后者的证明思路.

5.4 附录1: LASSO的理论证明

Proof. 这里采用的是Wainwright (2009)的证明手法. 首先考虑考虑支撑集$\mathcal{S}$上的LASSO估计

$$\begin{aligned}\widehat{\mathbf{b}} &= \underset{\mathbf{b}\in\mathbb{R}^q,\ \mathbf{b}_{\mathcal{S}^c}=0}{\arg\min}\ \frac{1}{2}\mathbf{b}^\top\widehat{\mathbf{\Sigma}}\mathbf{b} - \mathbf{b}^\top\widehat{\mathbf{a}} + \lambda\|\mathbf{b}\|_1 \\ &= \underset{\mathbf{b}\in\mathbb{R}^q,\ \mathbf{b}_{\mathcal{S}^c}=0}{\arg\min}\ \frac{1}{2}\mathbf{b}_{\mathcal{S}}^\top\widehat{\mathbf{\Sigma}}_{\mathcal{S},\mathcal{S}}\mathbf{b}_{\mathcal{S}} - \mathbf{b}_{\mathcal{S}}^\top\widehat{\mathbf{a}}_{\mathcal{S}} + \lambda\|\mathbf{b}_{\mathcal{S}}\|_1.\end{aligned}$$

由KKT条件可得,

$$\widehat{\mathbf{\Sigma}}_{\mathcal{S},\mathcal{S}}\widehat{\mathbf{b}}_{\mathcal{S}} - \widehat{\mathbf{a}}_{\mathcal{S}} + \lambda\cdot\mathrm{sign}(\widehat{\mathbf{b}}_{\mathcal{S}}) = 0.$$

由定义$\boldsymbol{\beta}^* = \mathbf{\Sigma}^{-1}\mathbf{a}$,

$$\begin{pmatrix}\mathbf{a}_{\mathcal{S}}\\ \mathbf{a}_{\mathcal{S}^c}\end{pmatrix} = \begin{pmatrix}\mathbf{\Sigma}_{\mathcal{S},\mathcal{S}} & \mathbf{\Sigma}_{\mathcal{S},\mathcal{S}^c}\\ \mathbf{\Sigma}_{\mathcal{S}^c,\mathcal{S}} & \mathbf{\Sigma}_{\mathcal{S}^c,\mathcal{S}^c}\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta}^*_{\mathcal{S}}\\ \mathbf{0}\end{pmatrix} = \begin{pmatrix}\mathbf{\Sigma}_{\mathcal{S},\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}\\ \mathbf{\Sigma}_{\mathcal{S}^c,\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}\end{pmatrix},$$

所以

$$\begin{aligned}\widehat{\mathbf{b}}_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}} =& \mathbf{\Sigma}^{-1}_{\mathcal{S},\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S},\mathcal{S}}(\widehat{\mathbf{b}}_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}})\\ =& \mathbf{\Sigma}^{-1}_{\mathcal{S},\mathcal{S}}\left((\mathbf{\Sigma}_{\mathcal{S},\mathcal{S}} - \widehat{\mathbf{\Sigma}}_{\mathcal{S},\mathcal{S}})(\widehat{\mathbf{b}}_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}}) + (\widehat{\mathbf{\Sigma}}_{\mathcal{S},\mathcal{S}}\widehat{\mathbf{b}}_{\mathcal{S}} - \widehat{\mathbf{a}}_{\mathcal{S}}) + (\widehat{\mathbf{\Sigma}}_{\mathcal{S},\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}} - \widehat{\mathbf{a}}_{\mathcal{S}})\right).\end{aligned}$$

由三角不等式, 可得

$$\|\widehat{\mathbf{b}}_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}}\|_\infty \le \|\mathbf{\Sigma}^{-1}_{\mathcal{S},\mathcal{S}}\|_L\left(|\mathcal{S}|\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty\|\widehat{\mathbf{b}}_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}}\|_\infty + \lambda + \alpha\lambda/3\right),$$

整理可得

$$\|\widehat{\mathbf{b}}_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}}\|_\infty \le (1+\frac{\alpha}{3})\frac{\|\mathbf{\Sigma}^{-1}_{\mathcal{S},\mathcal{S}}\|_L\lambda}{1 - \|\mathbf{\Sigma}^{-1}_{\mathcal{S},\mathcal{S}}\|_L|\mathcal{S}|\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty}.$$

接下来, 验证上述带约束估计满足原始LASSO问题的KKT条件

$$\|(\widehat{\mathbf{\Sigma}}\widehat{\mathbf{b}}-\widehat{\mathbf{a}})_{\mathcal{S}}\|_\infty \le \lambda, \text{和 } \|(\widehat{\mathbf{\Sigma}}\widehat{\mathbf{b}}-\widehat{\mathbf{a}})_{\mathcal{S}^c}\|_\infty < \lambda,$$

即

$$\|\widehat{\mathbf{\Sigma}}_{\mathcal{S},\mathcal{S}}\widehat{\mathbf{b}}_{\mathcal{S}}-\widehat{\mathbf{a}}_{\mathcal{S}}\|_\infty \le \lambda,\ \|\widehat{\mathbf{\Sigma}}_{\mathcal{S}^c,\mathcal{S}}\widehat{\mathbf{b}}_{\mathcal{S}}-\widehat{\mathbf{a}}_{\mathcal{S}^c}\|_\infty < \lambda.$$

第一项就是上述带约束LASSO问题的KKT条件. 这里考虑第二项,

$$\begin{aligned}
&\widehat{\mathbf{\Sigma}}_{\mathcal{S}^c,\mathcal{S}}\widehat{\mathbf{b}}_{\mathcal{S}}-\widehat{\mathbf{a}}_{\mathcal{S}^c}\\
=&(\widehat{\mathbf{\Sigma}}_{\mathcal{S}^c,\mathcal{S}}-\mathbf{\Sigma}_{\mathcal{S}^c,\mathcal{S}})(\widehat{\mathbf{b}}_{\mathcal{S}}-\boldsymbol{\beta}^*_{\mathcal{S}})+\mathbf{\Sigma}_{\mathcal{S}^c,\mathcal{S}}\mathbf{\Sigma}^{-1}_{\mathcal{S},\mathcal{S}}\{\mathbf{\Sigma}_{\mathcal{S},\mathcal{S}}(\widehat{\mathbf{b}}_{\mathcal{S}}-\boldsymbol{\beta}^*_{\mathcal{S}})\}+\widehat{\mathbf{\Sigma}}_{\mathcal{S}^c,\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}-\widehat{\mathbf{a}}_{\mathcal{S}^c}\\
=&(\widehat{\mathbf{\Sigma}}_{\mathcal{S}^c,\mathcal{S}}-\mathbf{\Sigma}_{\mathcal{S}^c,\mathcal{S}})(\widehat{\mathbf{b}}_{\mathcal{S}}-\boldsymbol{\beta}^*_{\mathcal{S}})+\widehat{\mathbf{\Sigma}}_{\mathcal{S}^c,\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}-\widehat{\mathbf{a}}_{\mathcal{S}^c}\\
&+\mathbf{\Sigma}_{\mathcal{S}^c,\mathcal{S}}\mathbf{\Sigma}^{-1}_{\mathcal{S},\mathcal{S}}\{(\mathbf{\Sigma}_{\mathcal{S},\mathcal{S}}-\widehat{\mathbf{\Sigma}}_{\mathcal{S},\mathcal{S}})(\widehat{\mathbf{b}}_{\mathcal{S}}-\boldsymbol{\beta}^*_{\mathcal{S}})+\widehat{\mathbf{\Sigma}}_{\mathcal{S},\mathcal{S}}(\widehat{\mathbf{b}}_{\mathcal{S}}-\boldsymbol{\beta}^*_{\mathcal{S}})\},
\end{aligned}$$

利用三角不等式

$$\begin{aligned}
\|\widehat{\mathbf{\Sigma}}_{\mathcal{S}^c,\mathcal{S}}\widehat{\mathbf{b}}_{\mathcal{S}}-\widehat{\mathbf{a}}_{\mathcal{S}^c}\|_\infty \le& s\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|_\infty\|\widehat{\mathbf{b}}_{\mathcal{S}}-\boldsymbol{\beta}^*_{\mathcal{S}}\|_\infty+\frac{\alpha}{3}\lambda\\
&+(1-\alpha)\left(s\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|_\infty\|\widehat{\mathbf{b}}_{\mathcal{S}}-\boldsymbol{\beta}^*_{\mathcal{S}}\|_\infty+\lambda+\frac{\alpha}{3}\lambda\right)\\
=&\lambda-\frac{\alpha(1+\alpha)}{3}\lambda+(2-\alpha)s\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|_\infty\|\widehat{\mathbf{b}}_{\mathcal{S}}-\boldsymbol{\beta}^*_{\mathcal{S}}\|_\infty\\
\le&\lambda-\frac{\alpha(1+\alpha)}{3}\lambda+(2-\alpha)sc\sqrt{\frac{\log p}{n}}\frac{2\|\mathbf{\Sigma}^{-1}_{\mathcal{S},\mathcal{S}}\|_L}{1-c\|\mathbf{\Sigma}^{-1}_{\mathcal{S},\mathcal{S}}\|_L s\sqrt{\frac{\log p}{n}}}\frac{3+\alpha}{3}\lambda.
\end{aligned}$$

注意

$$s\|\mathbf{\Sigma}^{-1}_{\mathcal{S},\mathcal{S}}\|_L\sqrt{\frac{\log p}{n}}\to 0,$$

所以

$$\|\widehat{\mathbf{\Sigma}}_{\mathcal{S}^c,\mathcal{S}}\widehat{\mathbf{b}}_{\mathcal{S}}-\widehat{\mathbf{a}}_{\mathcal{S}^c}\|_\infty < \lambda.$$

□

5.5 附录2: 渐近稀疏下的LASSO的理论证明

5.6 附录3: DS的理论证明

对于Dantzig Selector估计

$$\widehat{\boldsymbol{\beta}}_{\mathrm{ds}}(\lambda) = \underset{\beta\in\mathbb{R}^p}{\arg\min} \|\boldsymbol{\beta}\|_1, \text{ subject to } \|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \widehat{\mathbf{a}}\|_\infty \le \lambda.$$

设置

$$\lambda \ge \|(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\|_\infty \|\boldsymbol{\beta}^*\|_1 + \|\mathbf{a} - \widehat{\mathbf{a}}\|_\infty \ge \|(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\beta}^* + \mathbf{a} - \widehat{\mathbf{a}}\|_\infty = \|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^* - \widehat{\mathbf{a}}\|_\infty,$$

即真实参数满足约束条件, 从而可得

$$\|\widehat{\boldsymbol{\beta}}_{\mathrm{ds}}\|_1 \le \|\boldsymbol{\beta}^*\|_1.$$

由此,

$$\begin{aligned}
\|\widehat{\boldsymbol{\beta}}_{\mathrm{ds}} - \boldsymbol{\beta}^*\|_\infty =& \|\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}}_{\mathrm{ds}} - \boldsymbol{\beta}^*)\|_\infty \le \|\boldsymbol{\Sigma}^{-1}\|_L \|\boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}}_{\mathrm{ds}} - \boldsymbol{\beta}^*)\|_\infty \\
=& \|\boldsymbol{\Sigma}^{-1}\|_L \|(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}})\widehat{\boldsymbol{\beta}}_{\mathrm{ds}} + \widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\beta}}_{\mathrm{ds}} - \widehat{\mathbf{a}} + \widehat{\mathbf{a}} - \mathbf{a}\|_\infty \\
\le& \|\boldsymbol{\Sigma}^{-1}\|_L \left(\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty \|\widehat{\boldsymbol{\beta}}_{\mathrm{ds}}\|_1 + \|\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\beta}}_{\mathrm{ds}} - \widehat{\mathbf{a}}\|_\infty + \|\widehat{\mathbf{a}} - \mathbf{a}\|_\infty \right) \\
\le& 2\|\boldsymbol{\Sigma}^{-1}\|_L \lambda.
\end{aligned}$$

基于$|\widehat{\boldsymbol{\beta}}|_1 \le |\boldsymbol{\beta}|_1$以及$t_n = |\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_\infty$比较小, 可以控制$\ell_1$-norm. 记集合$J = \{j : |\beta_j| < t_n\}$, 则

$$\sum_{j\in J} |\hat{\beta}_j| + \sum_{j\in J^c} |\hat{\beta}_j| \le \sum_{j\in J} |\beta_j| + \sum_{j\in J^c} |\beta_j|,$$

由此可得

$$\sum_{j\in J} |\hat{\beta}_j| \le \sum_{j\in J} |\beta_j| + \sum_{j\in J^c} \left(|\beta_j| - |\hat{\beta}_j| \right) \le \sum_{j\in J} |\beta_j| + \sum_{j\in J^c} |\hat{\beta}_j - \beta_j|.$$

因此

$$\begin{aligned}
|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}|_1 =& \sum_{j\in J^c}|\hat{\beta}_j-\beta_j|+\sum_{j\in J}|\hat{\beta}_j-\beta_j| \\
\leq& 2\sum_{j\in J^c}|\hat{\beta}_j-\beta_j|+2\sum_{j\in J}|\beta_j| \\
\leq& 2\sum_{j:|\beta_j|\geq t_n} t_n+2\sum_{j:|\beta_j|<t_n}|\beta_j| \\
\leq& 2\sum_{j:|\beta_j|\geq t_n}|\beta_j|^q|t_n|^{1-q}+2\sum_{j:|\beta_j|<t_n}|\beta_j|^q|t_n|^{1-q} \\
=& 2|t_n|^{1-q}\sum_{j=1}^{p}|\beta_j|^q.
\end{aligned}$$

由此

$$|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}|_2^2 \leq |\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}|_\infty|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}|_1 \leq 12t_n^{2-q}\sum_{i=1}^{p}|\beta_i|^q.$$

6 其他高维稀疏估计

在高维稀疏线性回归问题中, 估计的参数形式为:

$$\mathbf{b}^* = \mathbf{A}^{-1}\mathbf{a}$$

其中$\mathbf{A} \in \mathbb{R}^{p\times p}$是一个正定矩阵, $\mathbf{a} \in \mathbb{R}^p$是一个向量. 从损失函数角度,

$$\mathbf{b}^* = \arg\min_{\mathbf{b}\in\mathbb{R}^p} \left(\frac{1}{2}\mathbf{b}^\top \mathbf{A}\mathbf{b} - \mathbf{b}^\top \mathbf{a}\right).$$

基于统计样本, 我们可以构造出$\mathbf{A}$和$\mathbf{a}$在无穷范数下的相合估计, 即

$$\|\widehat{\mathbf{A}} - \mathbf{A}\|_\infty = O_p(\sqrt{\frac{\log p}{n}}),\ \|\widehat{\mathbf{a}} - \mathbf{a}\|_\infty = O_p(\sqrt{\frac{\log p}{n}}).$$

由此可得LASSO估计

$$\widehat{\mathbf{b}}_{\text{lasso}} = \arg\min_{\mathbf{b}\in\mathbb{R}^p} \left(\frac{1}{2}\mathbf{b}^\top \widehat{\mathbf{A}}\mathbf{b} - \mathbf{b}^\top \widehat{\mathbf{a}} + \lambda\|\mathbf{b}\|_1\right),$$

和Dantzig Selector估计

$$\widehat{\mathbf{b}}_{\text{ds}} = \arg\min_{\mathbf{b}\in\mathbb{R}^p} \|\mathbf{b}\|_1,\ \text{subject to } \|\widehat{\mathbf{A}}\mathbf{b} - \widehat{\mathbf{a}}\|_\infty \le \lambda,$$

其中$\lambda > 0$是调节参数. 第五章的LASSO和Dantzig Selector估计的理论性质分析适用于这里的一般情形, 本节我们在这个基础上考虑一下其他高维稀疏多元统计分析问题.

6.1 线性判别分析

在线性判别分析(Linear Discriminant Analysis, LDA)问题中, 考虑两个总体$N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$和$N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, 要估计的参数为:

$$\boldsymbol{\beta}^* = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),$$

形式上取$\mathbf{A} = \boldsymbol{\Sigma}$, $\mathbf{a} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$即可转化为高维稀疏回归问题, 基于样本均值和样本协方差矩阵可以构造出无穷范数下的估计. 特别的, LASSO型的稀疏LDA方法可参见Mai et al. (2012), Dantzig Selector型估计可参考Cai and Liu (2011b).

6.2 精度矩阵估计

总体协方差矩阵的逆矩阵称为精度矩阵$\mathbf{\Omega} \overset{\text{def}}{=} \mathbf{\Sigma}^{-1}$. 在多元正态分布下, 精度矩阵刻画了变量之间的条件协方差(Gaussian graphical model). 高维统计中通过惩罚似然函数可以得到稀疏精度矩阵的估计方法, 即graphical LASSO. 利用回归, 我们也可以构造出另外形式的稀疏精度矩阵估计.

考虑精度矩阵的每一列,

$$\mathbf{\Omega}_j = \mathbf{\Sigma}^{-1}\mathbf{e}_j,$$

所以形式上取$\mathbf{A} = \mathbf{\Sigma}, \mathbf{a} = \mathbf{e}_j$即可构造出每一列的稀疏估计. 特别的, LASSO形式的估计参见Liu and Luo (2015), Dantzig Selector形式估计参见Cai et al. (2011). 考虑到精度矩阵估计要求对称性和正定性, 实际方法中还需要一些对称化和正定化处理技巧, 完整的证明过程参考相应文献.

6.3 二次型判别分析

在二次型判别分析(Quadratic Discriminant Analysis)问题中, 考虑两个总体$N(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1)$和$N(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2)$, 最优的分类准则为

$$(\mathbf{X}-\boldsymbol{\mu}_1)^\top \left(\mathbf{\Sigma}_2^{-1} - \mathbf{\Sigma}_1^{-1}\right)(\mathbf{X}-\boldsymbol{\mu}_1) - 2\left(\mathbf{X} - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\right)^\top \mathbf{\Sigma}_2^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1) + \log\frac{|\mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_1|} \geq 0,$$

线性项系数完全等价于线性回归系数, 这里的难点主要是估计一个$p \times p$的矩阵$\mathbf{\Sigma}_2^{-1} - \mathbf{\Sigma}_1^{-1}$.

从矩阵拉直运算的角度,

$$\begin{aligned}\text{vec}\left(\mathbf{\Sigma}_2^{-1} - \mathbf{\Sigma}_1^{-1}\right) =& \text{vec}\left(\mathbf{\Sigma}_2^{-1}(\mathbf{\Sigma}_2 - \mathbf{\Sigma}_1)\mathbf{\Sigma}_1^{-1}\right)\\ =& \left(\mathbf{\Sigma}_1^{-1} \otimes \mathbf{\Sigma}_2^{-1}\right)\text{vec}(\mathbf{\Sigma}_2 - \mathbf{\Sigma}_1)\\ =& \left(\mathbf{\Sigma}_1 \otimes \mathbf{\Sigma}_2\right)^{-1}\text{vec}(\mathbf{\Sigma}_2 - \mathbf{\Sigma}_1),\end{aligned}$$

这里$\otimes$表示Kronecker乘积.

所以形式上可设置

$$\mathbf{A} = \mathbf{\Sigma}_1 \otimes \mathbf{\Sigma}_2,\ \mathbf{a} = \text{vec}(\mathbf{\Sigma}_2 - \mathbf{\Sigma}_1),$$

即考虑一个p^2维度的回归问题. 给定样本协方差矩阵$\widehat{\mathbf{\Sigma}}_1$和$\widehat{\mathbf{\Sigma}}_2$, LASSO形式的估计为

$$\arg\min_{\mathbf{b}\in\mathbb{R}^{p^2}} \left(\frac{1}{2}\mathbf{b}^\top(\widehat{\mathbf{\Sigma}}_1 \otimes \widehat{\mathbf{\Sigma}}_2)\mathbf{b} - \mathbf{b}^\top \mathrm{vec}(\widehat{\mathbf{\Sigma}}_2 - \widehat{\mathbf{\Sigma}}_1) + \lambda\|\mathbf{b}\|_1 \right).$$

如果记$b = \mathrm{vec}(\mathbf{B})$, $\mathbf{B} \in \mathbb{R}^{p\times p}$, 上述LASSO形式的估计可以采用矩阵形式表述为

$$\arg\min_{\mathbf{B}\in\mathbb{R}^{p\times p}} \left[\frac{1}{2}\mathrm{tr}\left(\mathbf{B}^\top\widehat{\mathbf{\Sigma}}_2\mathbf{B}\widehat{\mathbf{\Sigma}}_1\right) - \mathbf{B}^\top(\widehat{\mathbf{\Sigma}}_2 - \widehat{\mathbf{\Sigma}}_1) + \lambda\|\mathbf{B}\|_1 \right].$$

该方法可参看Jiang et al. (2018). 类似的, Dantzig Selector形式可参看Zhao et al. (2014).

更加一般的, 可以考虑估计系数矩阵

$$\mathbf{\Sigma}_1^{-1}\mathbf{M}\mathbf{\Sigma}_2^{-1},$$

这一类型的参数在矩阵型数据的判别分析, 二次型回归分析以及充分降维等领域中会出现, 也有很多文献基于类似的LASSO或者Dantzig Selector形式来估计参数.

6.4 总结

至此, 我们从高维角度对多元统计分析相关课题进行了延伸和拓展. 除此之外, 高维统计方法还包括稀疏主成分分析、基于Nuclear-norm的低秩矩阵估计等等其他形式的稀疏估计, 这里不再展开.

References

P. J. Bickel and E. Levina. Covariance regularization by thresholding. *Annals of Statistics*, 36(6):2577–2604, 2008.

T. Cai and W. Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494): 672–684, 2011a.

T. Cai, W. Liu, and X. Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

T. T. Cai and W. Liu. A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106 (496):1566–1577, 2011b.

E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35(1):2313–2351, 2007.

B. Jiang, X. Wang, and C. Leng. A direct approach for sparse quadratic discriminant analysis. *Journal of Machine Learning Research*, 19(31):1–37, 2018.

Q. Li and J. Shao. Sparse quadratic discriminant analysis for high dimensional data. *Statistica Sinica*, pages 457–473, 2015.

W. Liu and X. Luo. Fast and adaptive sparse precision matrix estimation in high dimensions. *Journal of Multivariate Analysis*, 135:153–162, 2015.

Q. Mai, H. Zou, and M. Yuan. A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99(1):29–42, 2012.

A. J. Rothman, E. Levina, and J. Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104 (485):177–186, 2009.

J. Shao, Y. Wang, X. Deng, and S. Wang. Sparse linear discriminant analysis by thresholding for high dimensional data. *Annals of Statistics*, 39(2): 1241–1265, 2011.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

S. A. Van De Geer and P. Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

C.-H. Zhang and T. Zhang. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4): 576–593, 2012.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

S. D. Zhao, T. T. Cai, and H. Li. Direct estimation of differential networks. *Biometrika*, 101(2):253–268, 2014.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.