高等数理统计笔记

刘宇扬

January 13, 2020

摘要

这篇笔记是根据上海交通大学统计学专业核心课程高等数理统计上课所讲授的内容以及课后作业整理，主讲教授为刘卫东教授。本笔记分为五个部分：

第一章概述中位数方面的一些统计推断方法：经典的中位数Bahurder展开、分布式系统下中位数的Bahurder展开、分布式系统中基于下降方法的中位数估计、基于median of mean 稳健估计的Bahurder展开以及中位数回归的Bahurder展开。通过这一章的学习我们将不难看出，Bahurder展开在统计学中意义重大，因其直接蕴含着相合性、收敛速度以及中心极限定理等有关统计推断的重要结果。

第二章从统计学的角度（样本与总体）概述机器学习里的下降方法，并从样本推断总体的角度给出相应的理论结果。其内容包括Newton迭代与梯度下降这两种基本的下降方法的理论结果、Online设定下的随机梯度下降以及Communication efficiency（交互有效）的极限理论。最后结合第一章median of mean的想法，给出了稳健Communication efficiency 的理论结果。

第三章重点介绍Lasso作为高维设定下线性模型参数的一种估计方法以及高维设定下线性模型的一种变量选择方法的理论结果。包括证明了基于样本的Lasso估计收敛到总体模型参数的速度、作为变量选择方法的selection consistency。除此之外，概述了一种与Lasso类似的参数估计、变量选择方法：Dantzig selector及其与Lasso估计的关系。最后，结合第二章的communication efficiency的想法，给出了分布式系统下基于communication efficiency的Lasso估计方法，以及对应的收敛速度和selection consistency。

第四章概述现代统计学研究的一大方向：高维矩阵估计的四个分支，稀疏协方差矩阵的估计、稀疏协方差矩阵逆矩阵的估计、类似与变量选择的协方差矩阵逆矩阵支撑的估计以及判别分析。先证明了几个初等但重要的结论，而后介绍几种以这些结论为出发点的经典方法。

最后一部分是笔记的附录，包括如$O_{a.s}$、双指标极限等记号的说明、一些概率论的结论和证明，如正态分布最大次序统计量的极限性质等。

目录

1 中位数 **5**

1.1 样本中位数的Bahurder 展开 5

1.2 分布式系统下总体中位数估计(mean of median)的Bahurder展开 10

1.2.1 分布式系统设定 10

1.2.2 分布式系统下总体中位数的Naive估计 10

1.2.3 分布式系统下总体中位数的Naive估计的Bahurder展开 10

1.3 基于下降方法的中位数估计 16

1.3.1 准备工作 16

1.3.2 基于下降方法的中位数估计 21

1.3.3 基于下降方法的中位数估计在分布式系统下的实现 24

1.4 分布式系统下总体均值稳健估计(median of mean)的Bahurder展开 26

1.4.1 背景简介 26

1.4.2 Median of mean的Bahurder展开 26

1.5 中位数回归中参数估计的Bahurder展开 31

1.5.1 中位数回归简介 31

1.5.2 中位数回归中参数估计的Bahurder展开 32

2 下降方法 **40**

2.1 概述 40

2.1.1 损失函数 40

2.1.2 两种基本的下降法简介 44

2.1.3 参数估计：β^*与$\hat{\beta}$ 45

2.1.4 统计量实现：$\hat{\beta}$与β_t 48

2.2 Online设定下的随机梯度下降 49

2.2.1 背景简介 49

2.2.2 Online设定下的随机梯度下降参数估计问题中β_t与β^* . . . 50
2.3 Communication efficiency . . . 53
2.3.1 背景简介 . . . 53
2.3.2 Communication efficiency的极限性质 . . . 53
2.3.3 基于median of mean的稳健Communication efficiency及其极限性质 . . . 59

3 Lasso **64**
3.1 Lasso简介 . . . 64
3.2 Lasso估计的极限性质 . . . 64
3.2.1 一个关于优化问题的引理 . . . 64
3.2.2 Lasso估计收敛到总体参数的速度 . . . 66
3.3 Lasso作为变量选择方法的selection consistency . . . 69
3.3.1 一个关于Lasso估计的性质 . . . 69
3.3.2 selection consistency . . . 70
3.4 分布式系统下基于communication efficiency的Lasso估计 . . . 75
3.4.1 背景简介 . . . 75
3.4.2 基于communication efficiency的Lasso估计收敛到总体参数的速度 . . . 76
3.4.3 基于communication efficiency的Lasso估计的selection consistency . . . 78
3.5 Lasso与Dantzig selector . . . 83
3.5.1 Dantzig selector简介 . . . 83
3.5.2 Lasso的一种等价形式 . . . 83
3.5.3 Lasso与Dantzig selector . . . 87

4 高维矩阵估计概述 **89**
4.1 稀疏协方差矩阵的估计 . . . 89
4.1.1 传统方法存在的问题以及关于一个基本性质的证明 . . . 89
4.1.2 条状协方差矩阵的估计 . . . 91

4.1.3 稀疏协方差矩阵的估计 92
4.2 稀疏协方差矩阵逆矩阵的估计 94
4.2.1 两个多元统计分析中的结论及其证明 95
4.2.2 限制域最小化估计(CLIME)及其极限性质的证明 97
4.3 协方差矩阵逆矩阵支撑的估计 99
4.3.1 虚假发现率 100
4.3.2 应用多重假设检验估计GGM 100
4.4 高维数据判别分析 102
4.4.1 Fisher线性判别准则 102
4.4.2 基于Dantzig的想法估计线性判别函数 103

A 几个在这篇笔记中反复出现记号说明 **105**

B 一些数学结论与证明 **106**

1 中位数

1.1 样本中位数的Bahurder 展开

Bahurder 展开

$$\hat{m_x} - m_x = \frac{1}{nf(m_x)} \sum_{i=1}^{n} \left[\frac{1}{2} - \mathbf{1}\{X_i \le m_x\} \right] + O_{a.s.}\left(n^{-\frac{3}{4}} logn \right)$$

其中$\{x_i\}$独立同分布，有Lipschitz连续的分布函数F，以及密度函数f

证明 记$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \le x\}$ 为基于样本$\{X_i\}$的经验分布函数

$F(x) = P\{X_1 \le x\}$为总体的分布函数

我们先证明以下结论

$$\sup_{x \in I_n} \left| \Big[F_n(x) - F_n(\theta) \Big] - \Big[F(x) - F(\theta) \Big] \right| = O_{a.s.}(n^{-\frac{3}{4}} logn), \forall \theta \in \mathbf{R} \tag{1}$$

其中，$I_n = [\theta - a_n, \theta + a_n], a_n = O(\frac{logn}{\sqrt{n}})$，这里设$a_n \le C_2 \frac{logn}{\sqrt{n}}$

记$G_n(x) = [F_n(x) - F_n(\theta)] - [F(x) - F(\theta)]$

事实上，$\forall x \in [\theta - a_n, \theta + a_n]$，由Berstein不等式

$$\begin{aligned} P\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} \left(\Big[\mathbf{1}\{X_i \le x\} - \mathbf{1}\{X_i \le \theta\} \Big] - \Big[F(x) - F(\theta) \Big] \right) \right| \ge C_1 n^{-\frac{3}{4}} logn \right\} \\ \le 2exp\left\{ - \frac{\frac{n^2}{2} C_1^2 n^{-\frac{2}{3}} log^2 n}{\sum_{i=1}^{n} Var(\mathbf{1}\{X_i \le x\} - \mathbf{1}\{X_i \le \theta\}) + \frac{1}{3} nMC_1 n^{-\frac{3}{4}} logn} \right\} \end{aligned}$$

其中，

$$
\begin{aligned}
&Var\Big(\mathbf{1}\{X_i \le x\} - \mathbf{1}\{X_i \le \theta\}\Big) \\
&\le E\bigg[\Big(\mathbf{1}\{X_i \le x\} - \mathbf{1}\{X_i \le \theta\}\Big)^2\bigg] \\
&= E\bigg[\Big|\mathbf{1}\{X_i \le x\} - \mathbf{1}\{X_i \le \theta\}\Big|\bigg] \\
&\le E\bigg[\mathbf{1}\{\theta - |x-\theta| < X_i \le \theta + |x-\theta|\}\bigg] \\
&= F(\theta + |x-\theta|) - F(\theta - |x-\theta|) \\
&\le 2C_L|x-\theta| \\
&\le 2C_L a_n \le 2C_L C_2 \frac{logn}{\sqrt{n}}
\end{aligned}
$$

这里，C_L为分布函数F的Lipschitz常数.

M为满足$M \ge |[\mathbf{1}\{X_i \le x\} - \mathbf{1}\{X_i \le \theta\}] - [F(x) - F(\theta)]|, a.e.$的正数。

这里可取$M = 4$,进而，当n充分大使得$\frac{3}{4}C_1 n^{-\frac{1}{4}} \le 1$

$$
\begin{aligned}
&P\bigg\{\Big|\frac{1}{n}\sum_{i=1}^{n}\Big([\mathbf{1}\{X_i \le x\} - \mathbf{1}\{X_i \le \theta\}] - [F(x) - F(\theta)]\Big)\Big| \ge C_1 n^{-\frac{3}{4}} logn\bigg\} \\
&\le 2exp\bigg\{-\frac{\frac{1}{2}C_1^2 logn}{2C_L C_2 + \frac{3}{4}C_1 n^{-\frac{1}{4}}}\bigg\} \\
&\le 2exp\bigg\{-\frac{C_1^2 logn}{2(2C_L C_2 + 1)}\bigg\} \\
&= 2(\frac{1}{n})^{\frac{C_1^2}{C^*}}
\end{aligned}
$$

其中将$2(2C_L C_2 + 1)$记作C^*

$$
\sum_{n=1}^{\infty} P\bigg\{\Big|\frac{1}{n}\sum_{i=1}^{n}\Big([\mathbf{1}\{X_i \le x\} - \mathbf{1}\{X_i \le \theta\}] - [F(x) - F(\theta)]\Big)\Big| \ge C_1 n^{-\frac{3}{4}} logn\bigg\} < +\infty
$$

由$Borel - Cantelli$引理，

$$
P\bigg\{\Big|\frac{1}{n}\sum_{i=1}^{n}\Big([\mathbf{1}\{X_i \le x\} - \mathbf{1}\{X_i \le \theta\}] - [F(x) - F(\theta)]\Big)\Big| \ge C_1 n^{-\frac{3}{4}} logn,\ i.o.\bigg\} = 0
$$

即，

$$\exists N, \forall n, \left| \frac{1}{n} \sum_{i=1}^{n} \Big([\mathbf{1}\{X_i \le x\} - \mathbf{1}\{X_i \le \theta\}] - [F(x) - F(\theta)] \Big) \right| < C_1 n^{-\frac{3}{4}} logn.$$

几乎处处成立

从而，$\forall x \in [\theta - a_n, \theta + a_n], |G_n(x)| = O_{a.s.}(n^{-\frac{3}{4}} logn)$

下面往证 $\sup\limits_{x \in I_n} |[F_n(x) - F_n(\theta)] - [F(x) - F(\theta)]| = O_{a.s.}(n^{-\frac{3}{4}} logn)$

将区间I_n等分为$N = 2a_n n^{\alpha}$个小区间，每个长度均为$n^{-\alpha}$

记这些小区间左端点为$\{x_k\}|_{k=1,2,\ldots,N}$

一方面，

$$\max_{1 \le k \le N} |G_n(x_k)| = O_{a.s.}(n^{-\frac{3}{4}} logn) \tag{2}$$

事实上，

$$\begin{aligned} & P\Big\{ \max_{1 \le k \le N} |G_n(x_k)| \ge C_1 n^{-\frac{3}{4}} logn \Big\} \\ & \le \sum_{k=1}^{N} P\Big\{ |G_n(x_k)| \ge C_1 n^{-\frac{3}{4}} logn \Big\} \\ & = 2N(\frac{1}{n})^{\frac{C_1^2}{C^*}} = 4C_2(\frac{1}{n})^{\frac{C_1^2}{C^*} + \frac{1}{2} - \alpha} logn \end{aligned}$$

对于充分大的C_1，依然可以保证

$$\sum_{n=1}^{\infty} P\Big\{ \max_{1 \le k \le N} |G_n(x_k)| \ge C_1 n^{-\frac{3}{4}} logn \Big\} < +\infty$$

从而，$\max\limits_{1 \le k \le N} |G_n(x_k)| = O_{a.s.}(n^{-\frac{3}{4}} logn)$

另一方面，

$$\max_{1\le k\le N}\sup_{x\in[x_k,x_{k+1}]}|G_n(x)-G_n(x_k)|=O_{a.s.}(n^{-\frac{3}{4}}logn) \tag{3}$$

事实上，不妨取$\alpha>\frac{3}{4}$，当n充分大时

有$C_1n^{-\frac{3}{4}}logn-2C_Ln^{-\alpha}>(C_1-1)n^{-\frac{3}{4}}logn$

$$\begin{aligned}
&P\left\{\sup_{x\in[x_k,x_{k+1}]}|G_n(x)-G_n(x_k)|\ge C_1n^{-\frac{3}{4}}logn\right\}\\
&=P\left\{\sup_{x\in[x_k,x_{k+1}]}\left|\frac{1}{n}\sum_{i=1}^{n}\Big([\mathbf{1}\{X_i\le x\}-\mathbf{1}\{X_i\le x_k\}]-[F(x)-F(x_k)]\Big)\right|\ge C_1n^{-\frac{3}{4}}logn\right\}\\
&\le P\left\{\sup_{x\in[x_k,x_{k+1}]}\left|\frac{1}{n}\sum_{i=1}^{n}[\mathbf{1}\{X_i\le x\}-\mathbf{1}\{X_i\le x_k\}]\right|+\sup_{x\in[x_k,x_{k+1}]}\Big|F(x)-F(x_k)\Big|\ge C_1n^{-\frac{3}{4}}logn\right\}\\
&=P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}[\mathbf{1}\{X_i\le x_{k+1}\}-\mathbf{1}\{X_i\le x_k\}]\right|+\Big|F(x_{k+1})-F(x_k)\Big|\ge C_1n^{-\frac{3}{4}}logn\right\}\\
&\le P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}[\mathbf{1}\{X_i\le x_{k+1}\}-\mathbf{1}\{X_i\le x_k\}]-[F(x_{k+1})-F(x_k)]\right|+2\Big|F(x_{k+1})-F(x_k)\Big|\ge C_1n^{-\frac{3}{4}}logn\right\}\\
&\le P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}[\mathbf{1}\{X_i\le x_{k+1}\}-\mathbf{1}\{X_i\le x_k\}]-[F(x_{k+1})-F(x_k)]\right|+2C_Ln^{-\alpha}\ge C_1n^{-\frac{3}{4}}logn\right\}\\
&\le P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}[\mathbf{1}\{X_i\le x_{k+1}\}-\mathbf{1}\{X_i\le x_k\}]-[F(x_{k+1})-F(x_k)]\right|\ge (C_1-1)n^{-\frac{3}{4}}logn\right\}
\end{aligned}$$

当n充分大时，由Berstein不等式易得上式$\le 2(\frac{1}{n})^{\frac{(C_1-1)^2}{2}}$

从而，$\forall k=1,2,...,N$有$\sup_{x\in[x_k,x_{k+1}]}|G_n(x)-G_n(x_k)|=O_{a.s.}(n^{-\frac{3}{4}}logn)$

与之前同理可得$\max_{1\le k\le N}\sup_{x\in[x_k,x_{k+1}]}|G_n(x)-G_n(x_k)|=O_{a.s.}(n^{-\frac{3}{4}}logn)$

再由

$$\sup_{x\in I_n}|[F_n(x)-F_n(\theta)]-[F(x)-F(\theta)]|\le\max_{1\le k\le N}|G_n(x_k)|+\max_{1\le k\le N}\sup_{x\in[x_k,x_{k+1}]}|G_n(x)-G_n(x_k)|$$

以及(2), (3)可得(1)成立。

接下来我们再证明

$$|V_n - m_x| = O_{a.s.}(\frac{logn}{\sqrt{n}}) \tag{4}$$

其中，$V_n = X_{[k_n]}$为第$[k_n]$次序统计量，$k_n = n(\frac{1}{2} + \frac{logn}{\sqrt{n}})$，$[\cdot]$为高斯函数。

事实上，

$$\begin{aligned}
&P\left\{V_n \geq m_x + C\frac{logn}{\sqrt{n}}\right\} \\
&\leq P\left\{\sum_{i=1}^{n} \mathbf{1}\{X_i \leq m_x + C\frac{logn}{\sqrt{n}}\} \leq k_n\right\} \\
&= P\left\{\frac{1}{n}\sum_{i=1}^{n}[F(m_x + C\frac{logn}{\sqrt{n}}) - \mathbf{1}\{X_i \leq m_x + C\frac{logn}{\sqrt{n}}\}] \geq F(m_x + C\frac{logn}{\sqrt{n}}) - (\frac{1}{2} + \frac{logn}{\sqrt{n}})\right\} \\
&\leq P\left\{\frac{1}{n}\sum_{i=1}^{n}[F(m_x + C\frac{logn}{\sqrt{n}}) - \mathbf{1}\{X_i \leq m_x + C\frac{logn}{\sqrt{n}}\}] \geq C_0\frac{logn}{\sqrt{n}}\right\}
\end{aligned}$$

其中，$C_0 = Cf(m_x) - 2$，下面由Berstein不等式，当n充分大时

并选取C充分大使得$\frac{C_0^2}{4} > 2$

上式$\leq exp\{\frac{-\frac{C_0^2}{2}log^2 n}{1+\frac{2}{3}C_0\frac{logn}{\sqrt{(n)}}}\} \leq (\frac{1}{n})^{\frac{C_0^2}{4}} \leq \frac{1}{n^2}$

从而，$\exists C_1 > 0, \exists N_1, \forall n > N_1, V_n - m_x < C_1\frac{logn}{\sqrt{n}}$，几乎处处成立

同理，$\exists C_2 > 0, \exists N_2, \forall n > N_2, V_n - m_x > -C_2\frac{logn}{\sqrt{n}}$，几乎处处成立

因此$|V_n - m_x| = O_{a.s.}(\frac{logn}{\sqrt{n}})$，即(4)成立。

最后，我们证明样本中位数的Bahurder展开：

由(1)(4)，这里取(1)中$\theta = m_x$可得

$$|G_n(V_n)| \leq \sup_{x \in I_n} |[F_n(x) - F_n(\theta)] - [F(x) - F(\theta)]|$$

几乎处处成立。

即得$|G_n(V_n)| = O_{a.s.}(n^{-\frac{3}{4}}logn)$。

进而，由Taylor公式

$$\begin{aligned}
&\frac{[k_n]}{n} = F_n(V_n) = F_n(m_x) + F(V_n) - F(m_x) + F_n(V_n) - F(V_n) + F(m_x) - F_n(m_x) \\
&= F_n(m_x) + F(V_n) - F(m_x) + O_{a.s.}(n^{-\frac{3}{4}}logn) \\
&= F_n(m_x) + (V_n - m_x)f(m_x) + O_{a.s.}(\frac{(logn)^2}{n}) + O_{a.s.}(n^{-\frac{3}{4}}logn) \\
&\Rightarrow V_n - m_x = \frac{1}{nf(m_x)}\sum_{i=1}^{n}[\frac{[k_n]}{n} - \mathbf{1}\{X_i \le m_x\}] + O_{a.s.}(n^{-\frac{3}{4}}logn)
\end{aligned}$$

1.2 分布式系统下总体中位数估计(mean of median)的Bahurder展开

1.2.1 分布式系统设定

n个样本$X_1, X_2, ..., X_n$分别来自N个子服务器，每个子服务器中有m个样本，满足$n = N \cdot m$。另外存在一个总服务器来汇总处理来自N个子服务器的各类信息，并作出关于总体统计推断。在基于分布式系统下的统计推断，除了要考虑一般统计推断应考虑的问题（如估计的相合性以及收敛速度、假设检验的无偏性以及取值尽可能高的势函数），还需要考虑在n个样本分散分布于N个子服务器的情况下实现统计推断对样本传输的要求（若某估计问题所需统计量的实现必须要求将n个样本汇总到总服务器内，那么在分布式系统下这并不是一个好的推断方法）。

1.2.2 分布式系统下总体中位数的Naive估计

对于每个子服务器，基于m个样本得到样本中位数$\hat{\beta}_i, i = 1, 2, ..., N$，将N 个样本中位数传到总服务器中取平均(mean of median)得到关于总体中位数的估计量$\hat{m_x} = \frac{1}{N}\sum_{i=1}^{N}\hat{\beta}_i$

1.2.3 分布式系统下总体中位数的Naive估计的Bahurder展开

$$\frac{1}{N}\sum_{i=1}^{N}\hat{\beta}_i - m_x = \frac{1}{nf(m_x)}\sum_{i=1}^{n}(\frac{1}{2} - \mathbf{1}\{X_i \le m_x\}) - \frac{f^{'}(m_x)}{8mf^3(m_x)} + O_p(\frac{logm}{\sqrt{n}m^{\frac{1}{4}}} + \frac{(logm)^2}{m^{\frac{5}{4}}})$$

其中，n个样本$X_1, X_2, ..., X_n$独立同分布，其为分布函数$F(x)$，并且有二阶可导且二阶导数有界的密度函数$f(x)$。

另外，一个重要假设是$\exists A > 1, n \leq m^A$.

证明 对于第k组样本，记样本下标集为H_k，由Taylor展开有：

$$F(\hat{\beta_k}) - F(m_x) = f(m_x)(\hat{\beta_k} - m_x) + \frac{1}{2}f^{'}(m_x)(\hat{\beta_k} - m_x)^2 + \frac{1}{6}f^{''}(\xi)(\hat{\beta_k} - m_x)^3$$

从而稍作变形后求和得

$$\frac{1}{N}\sum_{k=1}^{N}\hat{\beta_k} - m_x = \frac{1}{N}\frac{\sum_{k=1}^{N}(F(\hat{\beta_k}) - F(m_x))}{f(m_x)} - \frac{1}{2N}\frac{f^{'}(m_x)}{f(m_x)}\sum_{k=1}^{N}(\hat{\beta_k} - m_x)^2 - \frac{1}{6N}\frac{f^{''}(\xi)}{f(m_x)}\sum_{k=1}^{N}(\hat{\beta_k} - m_x)^3 \quad (5)$$

这里先说明一个证明中用到的极限结论：

由$\exists A > 1, n \leq m^A$可得$N \leq m^{A-1}$

从而有

$$\underset{1\leq k\leq N}{max}|\hat{\beta_k} - m_x - \frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2} - \mathbf{1}\{X_i \leq m_x\})| = O_{a.s.}(m^{-\frac{3}{4}}logm) \quad (6)$$

事实上，由之前所证可得

$$P\{|\hat{\beta_k} - m_x - \frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2} - \mathbf{1}\{X_i \leq m_x\})| > Cm^{-\frac{3}{4}}logm\} \leq m^{-r}$$

这里可以通过调整C使得r任意大

$$P\{\underset{1\leq k\leq N}{max}|\hat{\beta_k} - m_x - \frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2} - \mathbf{1}\{X_i \leq m_x\})| > Cm^{-\frac{3}{4}}logm\} \leq Nm^{-r} \leq m^{-r+A-1}$$

调整C使得上述概率可和，再由Borel-Cantelli引理即得(6)。

同理可得

$$\underset{1\le k\le N}{max}|\frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2}-\mathbf{1}\{X_i\le m_x\})|=O_{a.s.}(m^{-\frac{1}{2}}logm)$$

下面考查(5)中$\frac{1}{6N}\frac{f^{''}(\xi)}{f(m_x)}\sum_{k=1}^{N}(\hat{\beta_k}-m_x)^3$

设$|f^{''}(x)|\le M_2$，并由Cr不等式：$(a+b)^3\le C_3(a^3+b^3)$

$$\begin{aligned}
&|\frac{1}{6N}\frac{f^{''}(\xi)}{f(m_x)}\sum_{k=1}^{N}(\hat{\beta_k}-m_x)^3|\\
&\le\frac{M2}{6N}\frac{1}{f(m_x)}\sum_{k=1}^{N}|\hat{\beta_k}-m_x|^3\\
&\le\frac{M2}{6N}\frac{1}{f(m_x)}\sum_{k=1}^{N}\left[|\frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2}-\mathbf{1}\{X_i\le m_x\})|+|\hat{\beta_k}-m_x-\frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2}-\mathbf{1}\{X_i\le m_x\})|\right]^3\\
&\le\frac{M2}{6N}\frac{C_3}{f(m_x)}\sum_{k=1}^{N}\left[\left|\frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2}-\mathbf{1}\{X_i\le m_x\})\right|^3+\left|\hat{\beta_k}-m_x-\frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2}-\mathbf{1}\{X_i\le m_x\})\right|^3\right]\\
&\le\frac{M2}{6N}\frac{C_3}{f(m_x)}\sum_{k=1}^{N}\left|\frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2}-\mathbf{1}\{X_i\le m_x\})\right|^3\\
&\quad+\frac{M2}{6}\frac{C_3}{f(m_x)}\underset{1\le k\le N}{max}\left|\hat{\beta_k}-m_x-\frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2}-\mathbf{1}\{X_i\le m_x\})\right|^3
\end{aligned}$$

其中对于$\frac{M2}{6N}\frac{C_3}{f(m_x)}\sum_{k=1}^{N}\left|\frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2}-\mathbf{1}\{X_i\le m_x\})\right|^3$

由Rosenthal不等式

$$\begin{aligned}
&E\left[\left|\sum_{i\in H_k}(\frac{1}{2}-\mathbf{1}\{X_i\le m_x\})\right|^3\right]\\
&\le C_4\cdot max\left\{\sum_{i\in H_k}E\left|\frac{1}{2}-\mathbf{1}\{X_i\le m_x\}\right|^3,\left[\sum_{i\in H_k}E\left|\frac{1}{2}-\mathbf{1}\{X_i\le m_x\}\right|^2\right]^{\frac{3}{2}}\right\}\\
&\le(\frac{1}{4})^{\frac{3}{2}}\cdot C_4\cdot m^{\frac{3}{2}}
\end{aligned}$$

从而，由Markov不等式

$$\frac{1}{6N}\frac{f^{''}(\xi)}{f(m_x)}\sum_{k=1}^{N}(\hat{\beta}_k - m_x)^3 = O_p(m^{-\frac{3}{2}}) + O_{a.s.}(m^{-\frac{9}{4}}logm) = O_p(m^{-\frac{3}{2}}) \tag{7}$$

下面考查(5)中$\frac{1}{2N}\frac{f^{'}(m_x)}{f(m_x)}\sum_{k=1}^{N}(\hat{\beta}_k - m_x)^2$

$$\begin{aligned}
(\hat{\beta}_k - m_x)^2 &= \left[\left(\hat{\beta}_k - m_x - \frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2} - \mathbf{1}\{X_i \le m_x\})\right) + \frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2} - \mathbf{1}\{X_i \le m_x\})\right]^2 \\
&= \left[\hat{\beta}_k - m_x - \frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2} - \mathbf{1}\{X_i \le m_x\})\right]^2 \\
&\quad + 2\left[\hat{\beta}_k - m_x - \frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2} - \mathbf{1}\{X_i \le m_x\})\right]\cdot\left[\frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2} - \mathbf{1}\{X_i \le m_x\})\right] \\
&\quad + \left[\frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2} - \mathbf{1}\{X_i \le m_x\})\right]^2
\end{aligned}$$

这里有

$$\begin{aligned}
&\frac{1}{N}\sum_{k=1}^{N}\left[\hat{\beta}_k - m_x - \frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2} - \mathbf{1}\{X_i \le m_x\})\right]^2 \\
&\le \max_{1\le k\le N}\left[\hat{\beta}_k - m_x - \frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2} - \mathbf{1}\{X_i \le m_x\})\right]^2 \\
&= O_{a.s.}(m^{-\frac{3}{2}}log^2 m)
\end{aligned}$$

以及

$$
\begin{aligned}
&\left|\frac{1}{N}\sum_{k=1}^{N}2\left[\hat{\beta}_k - m_x - \frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2}-\mathbf{1}\{X_i\le m_x\})\right]\cdot\left[\frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2}-\mathbf{1}\{X_i\le m_x\})\right]\right| \\
&\le 2\max_{1\le k\le N}\left|\hat{\beta}_k - m_x - \frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2}-\mathbf{1}\{X_i\le m_x\})\right|\cdot\max_{1\le k\le N}\left|\frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2}-\mathbf{1}\{X_i\le m_x\})\right| \\
&= O_{a.s.}(m^{-\frac{5}{4}}log^2 m)
\end{aligned}
$$

另外

$$
\begin{aligned}
&\frac{1}{N}\sum_{k=1}^{N}\left[\frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2}-\mathbf{1}\{X_i\le m_x\})\right]^2 \\
&= E\left[\left(\frac{1}{mf(m_x)}\sum_{i\in H_1}(\frac{1}{2}-\mathbf{1}\{X_i\le m_x\})\right)^2\right] \\
&\quad + \frac{1}{N}\sum_{k=1}^{N}\left(\left(\frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2}-\mathbf{1}\{X_i\le m_x\})\right)^2 - E\left[\left(\frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2}-\mathbf{1}\{X_i\le m_x\})\right)^2\right]\right) \\
&= \frac{1}{4m^2f^2(m_x)} + O_p\left(\sqrt{\frac{Var\left(\left(\frac{1}{mf(m_x)}\sum_{i\in H_1}(\frac{1}{2}-\mathbf{1}\{X_i\le m_x\})\right)^2\right)}{N}}\right)
\end{aligned}
$$

最后一个等号是利用了Markov不等式。

其中$Var\left(\left(\frac{1}{mf(m_x)}\sum_{i\in H_1}(\frac{1}{2}-\mathbf{1}\{X_i\le m_x\})\right)^2\right) = O(\frac{1}{m^2})$

因此$\frac{1}{N}\sum_{k=1}^{N}\left[\frac{1}{mf(m_x)}\sum_{i\in H_k}(\frac{1}{2}-\mathbf{1}\{X_i\le m_x\})\right]^2 = \frac{1}{4m^2f^2(m_x)} + O_p(\frac{1}{\sqrt{N}m})$

即

$$
\frac{1}{2N}\frac{f^{'}(m_x)}{f(m_x)}\sum_{k=1}^{N}(\hat{\beta}_k - m_x)^2 = \frac{f^{'}(m_x)}{8mf^3(m_x)} + O_{a.s.}(m^{-\frac{5}{4}}log^2 m) + O_p(\frac{1}{\sqrt{n}m^{\frac{1}{2}}}) \tag{8}
$$

最后考察(5)中$\frac{1}{N}\frac{\sum_{k=1}^{N}(F(\hat{\beta_k})-F(m_x))}{f(m_x)}$

设$F_k(x)$为基于第k组样本的经验分布函数，并由Markov不等式有

$$\begin{aligned}
&\frac{1}{N}\sum_{k=1}^{N}(F(\hat{\beta_k})-F(m_x))\\
&=\frac{1}{N}\sum_{k=1}^{N}\left(\Big[F(\hat{\beta_k})-F(m_x)\Big]-\Big[F_k(\hat{\beta_k})-F_k(m_x)\Big]\right)+\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}-\mathbf{1}\{X_i\le m_x\}\right)\\
&=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}-\mathbf{1}\{X_i\le m_x\}\right)+O_p\left(E\left[\left(\Big[F(\hat{\beta_k})-F(m_x)\Big]-\Big[F_k(\hat{\beta_k})-F_k(m_x)\Big]\right)^2\right]\right)
\end{aligned}$$

记事件$A=\left[\left|\left(\Big[F(\hat{\beta_k})-F(m_x)\Big]-\Big[F_k(\hat{\beta_k})-F_k(m_x)\Big]\right)^2\right|>Cm^{-\frac{3}{4}}logm\right]$

由之前所证，$P(A)\le O(m^{-r})$，通过调整C可以使得r任意大。

从而，对于充分大的C

$$\begin{aligned}
&E\left[\left(\Big[F(\hat{\beta_k})-F(m_x)\Big]-\Big[F_k(\hat{\beta_k})-F_k(m_x)\Big]\right)^2\right]\\
&=E\left[\left(\Big[F(\hat{\beta_k})-F(m_x)\Big]-\Big[F_k(\hat{\beta_k})-F_k(m_x)\Big]\right)^2\cdot I_A\right]\\
&\quad+E\left[\left(\Big[F(\hat{\beta_k})-F(m_x)\Big]-\Big[F_k(\hat{\beta_k})-F_k(m_x)\Big]\right)^2\cdot I_{A^c}\right]\\
&\le Cm^{-\frac{3}{2}}log^2m+16O(m^{-r})\le(C+1)m^{-\frac{3}{2}}log^2m
\end{aligned}$$

即

$$\frac{1}{N}\frac{\sum_{k=1}^{N}(F(\hat{\beta_k})-F(m_x))}{f(m_x)}=\frac{1}{nf(m_x)}\sum_{i=1}^{n}\left(\frac{1}{2}-\mathbf{1}\{X_i\le m_x\}\right)+O_p(\frac{logm}{\sqrt{n}m^{\frac{1}{4}}})\tag{9}$$

综合(7)(8)(9)即得证。

评注　基于上述结果，易得以下推论：

- 若$\frac{\sqrt{n}}{m}\to 0$，则$\sqrt{n}\Big(\frac{1}{N}\sum_{i=1}^{N}\hat{\beta}_i-m_x\Big)\xrightarrow{D}N(0,\frac{1}{4f^2(m_x)})$

- 若$\frac{\sqrt{n}}{m} \to \alpha$，则$\sqrt{n}\Big(\frac{1}{N}\sum_{i=1}^{N}\hat{\beta}_i - m_x\Big) \xrightarrow{D} N(-\frac{\alpha f^{'}(m_x)}{8f^3(m_x)}, \frac{1}{4f^2(m_x)})$
- 若$\frac{\sqrt{n}}{m} \to \infty$，则$m\Big(\frac{1}{N}\sum_{i=1}^{N}\hat{\beta}_i - m_x\Big) \xrightarrow{P} -\frac{f^{'}(m_x)}{8f^3(m_x)}$

这意味着，分布式系统下总体中位数Naive估计的渐进结果依赖于(n,m)趋向无穷速度的关系。

在下一部分，我们将详细介绍一种其渐进结果不依赖于(n,m)趋向无穷速度关系的估计方法。

1.3 基于下降方法的中位数估计

1.3.1 准备工作

一个常用的极限结论 $\Big\{\{X_{ni}\}|_{i=1,2,\dots,n}\Big\}\Big|_{n=1,2,\dots}$为一个随机变量组列，满足：对于任意固定的$n \in \mathbf{N}^*, \{X_{ni}\}|_{i=1,2,\dots,n}$相互独立，$EX_{ni} = \mu_{ni}, Var(X_{ni}) \le a_n$，并且有$|X_{ni}| \le M$ almost surely.

记$U_n = \frac{1}{n}\sum_{i=1}^{n}\mu_{ni}$,

(1) 若$\exists C_1, \exists N, \forall n > N, a_n \ge C_1\frac{logn}{n}$，则$\frac{1}{n}\sum_{i=1}^{n}X_{ni} = O_{a.s.}\Big(|U_n| + \sqrt{\frac{a_n logn}{n}}\Big)$

(2) 若$\exists C_2, \exists N, \forall n > N, a_n \le C_2\frac{logn}{n}$，则$\frac{1}{n}\sum_{i=1}^{n}X_{ni} = O_{a.s.}\Big(|U_n| + \frac{logn}{n}\Big)$

证明 由Berstein不等式:

$$
\begin{aligned}
&P\Big\{\Big|\frac{1}{n}\sum_{i=1}^{n}(X_{ni}-\mu_{ni})\Big| \ge C\sqrt{\frac{a_n logn}{n}}\Big\} \\
&\le 2exp\Big\{-\frac{\frac{n^2}{2}\cdot C^2\frac{a_n logn}{n}}{na_n + \frac{n}{3}MC\sqrt{\frac{a_n logn}{n}}}\Big\} \\
&\le 2exp\{-\frac{C\cdot logn}{2(\frac{1}{C}+\frac{1}{3}M\sqrt{\frac{1}{C_1}})}\}
\end{aligned}
$$

调整C使得上述概率关于n可和，并利用Borel-Cantelli引理即得

$$\left|\frac{1}{n}\sum_{i=1}^{n}(X_{ni}-\mu_{ni})\right| = O_{a.s.}\left(\sqrt{\frac{a_n logn}{n}}\right)$$

再由绝对值三角不等式：

$$\left|\frac{1}{n}\sum_{i=1}^{n}X_{ni}\right| \leq \left|\frac{1}{n}\sum_{i=1}^{n}(X_{ni}-\mu_{ni})\right| + \left|U_n\right|$$

即得

$$\frac{1}{n}\sum_{i=1}^{n}X_{ni} = O_{a.s.}\left(|U_n| + \sqrt{\frac{a_n logn}{n}}\right)$$

从而(1)得证，而(2)的证明是类似的。

利用这个结论不难证明:

若$\exists C_1, \exists N, \forall n > N, a_n \geq C_1\frac{logn}{n}$,则$\forall\theta \in \mathbf{R}, \forall x \in [\theta - a_n, \theta + a_n]$

$$\frac{1}{n}\sum_{i=1}^{n}\left(\left[\mathbf{1}\{X_i \leq x\} - \mathbf{1}\{X_i \leq \theta\}\right] - E\left[\mathbf{1}\{X_i \leq x\} - \mathbf{1}\{X_i \leq \theta\}\right]\right) = O_{a.s.}\left(\sqrt{\frac{a_n logn}{n}}\right)$$

事实上，由之前所证$Var\left(\mathbf{1}\{X_i \leq x\} - \mathbf{1}\{X_i \leq \theta\}\right) \leq C \cdot a_n$，之后利用上述结论即可。

对于上面的假设，更进一步地，可以证明：

$$\sup_{x\in[\theta-a_n,\theta+a_n]}\left|\frac{1}{n}\sum_{i=1}^{n}\left(\left[\mathbf{1}\{X_i \leq x\} - \mathbf{1}\{X_i \leq \theta\}\right] - E\left[\mathbf{1}\{X_i \leq x\} - \mathbf{1}\{X_i \leq \theta\}\right]\right)\right| = O_{a.s.}\left(\sqrt{\frac{a_n logn}{n}}\right)$$

非参数核密度估计 $\{X_i\}|_{i=1,2,\ldots,n}$为一组来自同一分布的样本，假设该分布具有有界的密度函数，考虑其密度函数的估计量:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{X_i - x}{h})$$

其中，K为核函数，一般有如下假定

- $K(u) = K(-u)$
- $K(u)$有界且Lipschitz连续
- $\int K(u)du = 1$
- $0 < \int u^2 K(u)du < \infty$
- $0 < \int K^2(u)du < \infty$

事实上，由K的Lipschitz连续性可得核密度估计作为x的函数也是Lipschitz连续的。

若存在正整数l满足：

$$\int u^i K(u)du = 0, i = 1, 2, ..., l$$
$$\int u^{l+1} K(u)du > 0$$

则称K为l阶核函数。

下面我们考查核密度估计作为密度函数估计收敛到总体密度函数的速度：

$$\left|\hat{f}(x) - f(x)\right| \leq \left|\hat{f}(x) - E[\hat{f}(x)]\right| + \left|E[\hat{f}(x)] - f(x)\right|$$

其中$\left|\hat{f}(x) - E[\hat{f}(x)]\right| = O_p\left(\sqrt{Var(\hat{f}(x))}\right)$，

由假定：总体密度函数有界，$|f| \leq C_1$

核函数K满足：$0 < \int K(u)du \leq C_2$，从而有：

$$
\begin{aligned}
&Var(\hat{f}(x)) = Var\left(\frac{1}{nh}\sum_{i=1}^{n}K(\frac{X_i - x}{h})\right) \\
&= \frac{1}{nh^2}Var\left(K(\frac{X_1 - x}{h})\right) \leq \frac{1}{nh^2}E\left[K^2(\frac{X_1 - x}{h})\right] \\
&= \frac{1}{nh^2}\int_{-\infty}^{+\infty}K^2(\frac{y-x}{h})f(y)dy \\
&\leq \frac{C_1}{nh^2}\int_{-\infty}^{+\infty}K^2(\frac{y-x}{h})dy \\
&\leq \frac{C_1}{nh}\int_{-\infty}^{+\infty}K^2(y)dy \\
&\leq \frac{C_1C_2}{nh}
\end{aligned}
$$

因此，$\left|\hat{f}(x) - E[\hat{f}(x)]\right| = O_p\left(\sqrt{\frac{1}{nh}}\right)$

而对于$\left|E[\hat{f}(x)] - f(x)\right|$，

若假定K为l阶核函数，满足$0 < \int u^{l+1}K(u)du \leq C_3$，

并假定总体密度函数$f(x)$有$(l+1)$阶有界导数，设$|f^{(l+1)}(x)| \leq C_4$，由Taylor 展开可得：

$$
\begin{aligned}
&\left|\hat{f}(x) - E[\hat{f}(x)]\right| \\
&= \left|\int_{-\infty}^{+\infty}K(t)\Big[f(x+ht) - f(x)\Big]\right| \\
&= \left|\int_{-\infty}^{+\infty}K(t)\Big[\sum_{j=1}^{l}f^{(j)}(x)\frac{h^jt^j}{j!} + f^{(l+1)}(\xi)\frac{h^{l+1}t^{l+1}}{(l+1)!}\Big]\right| \\
&= \left|\int_{-\infty}^{+\infty}K(t)\cdot f^{(l+1)}(\xi)\frac{h^{l+1}t^{l+1}}{(l+1)!}\right| \\
&\leq \frac{C_3C_4}{(l+1)!}h^{l+1}
\end{aligned}
$$

因此，$\left|E[\hat{f}(x)] - f(x)\right| = O(h^{l+1})$。

综上，对于l阶核函数$K(u)$，加上一些对总体密度函数一般的假设可得：

$$\left|\hat{f}(x)-f(x)\right|=O_p\left(\sqrt{\frac{1}{nh}}+h^{l+1}\right)$$

此时取$h=Cn^{-\frac{1}{2l+3}}$，可得$\forall x\in\mathbf{R},\left|\hat{f}(x)-f(x)\right|=O_p\left(n^{-\frac{2}{2l+3}}\right)$。

事实上，利用上一个准备工作，我们有关于$\left|\hat{f}(x)-E[\hat{f}(x)]\right|$ 几乎处处的极限结果：

若$\exists C_5>0,h>C_5\cdot\frac{logn}{n}$，则$\forall x\in\mathbf{R},\left|\hat{f}(x)-E[\hat{f}(x)]\right|=O_{a.s.}(\sqrt{\frac{logn}{nh}})$。

这里是因为:

$$\begin{aligned}&\left|\hat{f}(x)-E[\hat{f}(x)]\right|\\&=\left|\frac{1}{nh}\sum_{i=1}^{n}\left(K(\frac{X_i-x}{h})-E\big[K(\frac{X_i-x}{h})\big]\right)\right|\end{aligned}$$

再由之前所证：

$$Var\left(K(\frac{X_1-x}{h})\right)\leq C_1C_2\cdot h\leq C_1C_2C_5\cdot\frac{logn}{n}$$

以及上一个准备工作的结论即得证。

进一步地，对于$\{x_1,x_2,...,x_m\}$，其中$\exists A>0,m\leq n^A$，有:

$$\max_{1\leq k\leq m}\left|\hat{f}(x_k)-E[\hat{f}(x_k)]\right|=O_{a.s.}\left(\sqrt{\frac{logn}{nh}}\right)$$

更进一步地，对于任意有限区间I:

$$\sup_{x\in I}\left|\hat{f}(x)-E\left[\hat{f}(x)\right]\right|=O_{a.s.}\left(\sqrt{\frac{logn}{nh}}\right)$$

事实上，设区间I的长度为L，取充分大的α，将区间I等分为$N = L \cdot n^{\alpha}$ 个小区间$I_1, I_2, ..., I_N$，记这些小区间的左端点为$\{x_k\}|_{k=1,2,...,N}$，另外x_{N+1} 为区间I 的右端点。

$$
\begin{aligned}
&\sup_{x\in I}\left|\hat{f}(x) - E\left[\hat{f}(x)\right]\right| \\
&= \max_{1\le k\le N}\left[\sup_{x\in I_k}\left|\hat{f}(x) - E\left[\hat{f}(x)\right]\right|\right] \\
&\le \max_{1\le k\le N}\left[\left|\hat{f}(x_k) - E\left[\hat{f}(x_k)\right]\right| + \sup_{x\in I_k}\left|\hat{f}(x) - \hat{f}(x_k)\right| + \sup_{x\in I_k}\left|E\left[\hat{f}(x)\right] - E\left[\hat{f}(x_k)\right]\right|\right] \\
&= O_{a.s.}\left(\sqrt{\frac{logn}{nh}}\right)
\end{aligned}
$$

最后一步用到了核密度估计作为x的函数是Lipschitz连续的这一结论。

1.3.2 基于下降方法的中位数估计

假定已有对总体中位数的估计$\hat{\theta}_0$，记$|\hat{\theta}_0 - m_x| = O_{a.s.}(a_n)$。由样本中位数的Bahurder展开，可知样本中位数$\hat{m}_x$收敛速度为$|\hat{m}_x - m_x| = O_{a.s.}(\sqrt{\frac{logn}{n}})$，因此，不妨假设$\exists C_1, a_n \ge C_1\sqrt{\frac{logn}{n}}$。

下面我们证明：

假设总体有绝对连续的分布函数F，其密度函数f在$\mathbf{R}$上Lipschitz连续，可导且导函数有界，并且在总体中位数上取值大于0，

对于1 阶核函数K，满足之前核函数的几个基本假设，对于适当的带宽h，

基于已有估计$\hat{\theta}_0$的*boosting*估计量$\hat{\theta}_1$:

$$
\hat{\theta}_1 = \hat{\theta}_0 - \frac{1}{n\hat{f}(\hat{\theta}_0)}\sum_{i=1}^{n}\left[\mathbf{1}\{X_i \le \hat{\theta}_0\} - \frac{1}{2}\right]
$$

满足：

$$\hat{\theta}_1 - m_x = \frac{1}{nf(m_x)} \sum_{i=1}^{n} \left[\frac{1}{2} - \mathbf{1}\{X_i \leq m_x\} \right] + O_{a.s.}\left(a_n^2 + a_n \left(\frac{logn}{n} \right)^{\frac{2}{5}} + \sqrt{a_n \frac{logn}{n}} \right)$$

这里*boosting*估计的构造想法会在下一章节具体阐述，下面先证明*boosting*估计的上述性质：

证明 先考虑$\hat{f}(\hat{\theta}_0)$收敛到$f(m_x)$的速度：

由于$|\hat{\theta}_0 - m_x| = O_{a.s.}(a_n)$，因此$\exists C_2, \exists N, \forall n > N, |\hat{\theta}_0 - m_x| \leq C_2 \cdot a_n \quad a.s.$

$$\begin{aligned}
&\left| \hat{f}(\hat{\theta}_0) - f(m_x) \right| \\
&\leq \sup_{x \in [m_x - C_2 a_n, m_x + C_2 a_n]} \left| \hat{f}(x) - E\left[\hat{f}(x) \right] \right| + \sup_{x \in [m_x - C_2 a_n, m_x + C_2 a_n]} \left| E\left[\hat{f}(x) \right] - f(x) \right| + \left| f(\hat{\theta}_0) - f(m_x) \right| \quad a.s. \\
&= O_{a.s.}\left(\sqrt{\frac{logn}{nh}} + h^2 + a_n \right)
\end{aligned}$$

这里要求带宽h满足：$\exists C, h \geq C \cdot \frac{logn}{n}$。

根据假设$f(m_x) > 0$，对于充分大的n有$\hat{f}(\hat{\theta}_0) > \frac{f(m_x)}{2} > 0 \quad a.s.$

而这保证了*boosting*统计量在数学定义上几乎处处有意义。

由于

$$\begin{aligned}\hat{\theta}_1 - m_x &= \hat{\theta}_0 - m_x - \frac{1}{\hat{f}(\hat{\theta}_0)}\left[F(\hat{\theta}_0) - F(m_x)\right] \\ &+ \frac{1}{n\hat{f}(\hat{\theta}_0)}\sum_{i=1}^{n}\left[\frac{1}{2} - \mathbf{1}\{X_i \le m_x\}\right] \\ &- \frac{1}{\hat{f}(\hat{\theta}_0)}\left(\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{1}\{X_i \le \hat{\theta}_0\} - \mathbf{1}\{X_i \le m_x\}\right] - \left[F(\hat{\theta}_0) - F(m_x)\right]\right)\end{aligned} \tag{10}$$

考虑(10)中的$\frac{1}{\hat{f}(\hat{\theta}_0)}\left(\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{1}\{X_i \le \hat{\theta}_0\} - \mathbf{1}\{X_i \le m_x\}\right] - \left[F(\hat{\theta}_0) - F(m_x)\right]\right)$,

对于充分大的n有$\frac{1}{\hat{f}(\hat{\theta}_0)} \le \frac{2}{f(m_x)}$并结合准备工作中的极限结果可得：

$$\frac{1}{\hat{f}(\hat{\theta}_0)}\left(\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{1}\{X_i \le \hat{\theta}_0\} - \mathbf{1}\{X_i \le m_x\}\right] - \left[F(\hat{\theta}_0) - F(m_x)\right]\right) = O_{a.s.}\left(\sqrt{\frac{a_n logn}{n}}\right) \tag{11}$$

考虑(10)中的$\frac{1}{n\hat{f}(\hat{\theta}_0)}\sum_{i=1}^{n}\left[\frac{1}{2} - \mathbf{1}\{X_i \le m_x\}\right]$,

$$\begin{aligned}&\frac{1}{n\hat{f}(\hat{\theta}_0)}\sum_{i=1}^{n}\left[\frac{1}{2} - \mathbf{1}\{X_i \le m_x\}\right] \\ &= \frac{1}{n\hat{f}(m_x)}\sum_{i=1}^{n}\left[\frac{1}{2} - \mathbf{1}\{X_i \le m_x\}\right] + \left(\frac{1}{n\hat{f}(\hat{\theta}_0)} - \frac{1}{n\hat{f}(m_x)}\right)\sum_{i=1}^{n}\left[\frac{1}{2} - \mathbf{1}\{X_i \le m_x\}\right]\end{aligned}$$

结合$\left|\hat{f}(m_x) - \hat{f}(\hat{\theta}_0)\right| = O_{a.s.}\left(\sqrt{\frac{logn}{nh}} + h^2 + a_n\right)$

以及$\left|\sum_{i=1}^{n}\left[\frac{1}{2} - \mathbf{1}\{X_i \le m_x\}\right]\right| = O_{a.s.}\left(\sqrt{\frac{logn}{n}}\right)$

可得：

$$\frac{1}{n\hat{f}(\hat{\theta}_0)}\sum_{i=1}^{n}\left[\frac{1}{2} - \mathbf{1}\{X_i \le m_x\}\right] = \frac{1}{n\hat{f}(m_x)}\sum_{i=1}^{n}\left[\frac{1}{2} - \mathbf{1}\{X_i \le m_x\}\right] + O_{a.s.}\left(\sqrt{\frac{1}{h}}\frac{logn}{n} + h^2\sqrt{\frac{logn}{n}} + a_n\sqrt{\frac{logn}{n}}\right) \tag{12}$$

考虑(10)中的$\hat{\theta}_0 - m_x - \frac{1}{\hat{f}(\hat{\theta}_0)}\Big[F(\hat{\theta}_0) - F(m_x)\Big]$,

由Taylor展开：

$$\begin{aligned}
&\hat{\theta}_0 - m_x - \frac{1}{\hat{f}(\hat{\theta}_0)}\Big[F(\hat{\theta}_0) - F(m_x)\Big] \\
&= (1 - \frac{f(m_x)}{\hat{f}(\hat{\theta}_0)})(\hat{\theta}_0 - m_x) + \frac{f^{'}(\xi)}{\hat{f}(\hat{\theta}_0)}(\hat{\theta}_0 - m_x)^2 \\
&= O_{a.s.}\Big(a_n\sqrt{\frac{logn}{nh}} + a_n h^2 + a_n^2\Big)
\end{aligned} \tag{13}$$

结合(11)(12)(13)可得：

$$\hat{\theta}_1 - m_x = \frac{1}{n\hat{f}(m_x)}\sum_{i=1}^{n}\Big[\frac{1}{2} - \mathbf{1}\{X_i \le m_x\}\Big] + O_{a.s.}\Big(a_n\sqrt{\frac{logn}{nh}} + a_n h^2 + a_n^2 + \sqrt{\frac{a_n logn}{n}}\Big)$$

取$h = \Big(\frac{logn}{n}\Big)^{\frac{1}{5}}$即得所证。

1.3.3 基于下降方法的中位数估计在分布式系统下的实现

基于下降方法的中位数*boosting*估计中：

$$\hat{\theta}_1 = \hat{\theta}_0 - \frac{1}{n\hat{f}(\hat{\theta}_0)}\sum_{i=1}^{n}\Big[\mathbf{1}\{X_i \le \hat{\theta}_0\} - \frac{1}{2}\Big]$$

$\hat{f}(\hat{\theta}_0)$与$\sum_{i=1}^{n}\Big[\mathbf{1}\{X_i \le \hat{\theta}_0\} - \frac{1}{2}\Big]$都具有线性结构。由于在分布式系统中实现具有线性结构的统计量对数据的传输量的要求很低，所以具有线性结构的统计量十分契合分布式系统。

更直观地，上述观点不难从如下实现算法看出：

- 基于第1个子服务器中的数据计算样本中位数$\hat{\theta}_0$，此时根据Bahurder 展开有$\left|\hat{\theta}_0 - m_x\right| = O_{a.s.}\left(\sqrt{\frac{logm}{m}}\right)$
- 将$\hat{\theta}_0$广播到所有子服务器中，基于第k个服务器的样本计算：

  $< 1 >$ $\quad \sum_{i\in H_k} K\left(\frac{X_i - \hat{\theta}_0}{h}\right)$

  $< 2 >$ $\quad \sum_{i\in H_k}\left[\mathbf{1}\{X_i \le \hat{\theta}_0\} - \frac{1}{2}\right]$

  $k = 1, 2, ..., n$，并将它们传入总服务器中。
- 在总服务器中计算

  $< 1 >$ $\quad \hat{f}(\hat{\theta}_0) = \frac{1}{nh}\sum_{k=1}^{N}\sum_{i\in H_k} K\left(\frac{X_i - \hat{\theta}_0}{h}\right)$

  $< 2 >$ $\quad \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{1}\{X_i \le \hat{\theta}_0\} - \frac{1}{2}\right] = \frac{1}{n}\sum_{k=1}^{N}\sum_{i\in H_k}\left[\mathbf{1}\{X_i \le \hat{\theta}_0\} - \frac{1}{2}\right]$

  从而得到基于$\hat{\theta}_0$的*boosting*统计量$\hat{\theta}_1$，并将其广播到各个子服务器中进行下一次*boosting*估计。

循环这个迭代若干次后的*boosting*估计量$\hat{\theta}$即有收敛速度$\left|\hat{\theta} - m_x\right| = O_{a.s.}\left(\sqrt{\frac{logn}{n}}\right)$。

事实上，根据:

$$\hat{\theta}_1 - m_x = \frac{1}{nf(m_x)}\sum_{i=1}^{n}\left[\frac{1}{2} - \mathbf{1}\{X_i \le m_x\}\right] + O_{a.s.}\left(a_n^2 + a_n\left(\frac{logn}{n}\right)^{\frac{2}{5}} + \sqrt{a_n\frac{logn}{n}}\right)$$

对于第k次*boosting*估计量$\hat{\theta}_k$，记$|\hat{\theta}_k - m_x| = O_{a.s.}(t_n^{(k)})$，可知：

(1) 若$\exists C_1, \exists N, \forall n > N, t_n^{(k)} > C_1(\frac{logn}{n})^{\frac{1}{4}}$，则对于第k+1次*boosting*估计量$\hat{\theta}_{k+1}$有：$|\hat{\theta}_{k+1} - m_x| = O_{a.s.}(a_n^2)$

注意到$\left|\hat{\theta}_0 - m_x\right| = O_{a.s.}\left(\sqrt{\frac{logm}{m}}\right)$以及$\exists A > 1, n < m^A$可知

$\exists L, \exists C_2, C_3, \exists N, \forall n > N, C_2(\frac{logn}{n})^{\frac{1}{2}} \le t_n^{(L)} \le C_3(\frac{logn}{n})^{\frac{1}{4}}$

(2) 对于第L+1次*boosting*估计量$\hat{\theta}_{L+1}$有：$|\hat{\theta}_{L+1} - m_x| = O_{a.s.}\left(\sqrt{\frac{logn}{n}}\right)$

即少量几次迭代后即可达到收敛速度$O_{a.s.}\left(\sqrt{\frac{logn}{n}}\right)$。

1.4 分布式系统下总体均值稳健估计(median of mean)的Bahurder展开

1.4.1 背景简介

如之前的分布式系统设定：n个样本$X_1, X_2, ..., X_n$独立同分布，有均值μ以及方差σ^2，分别来自N个子服务器，每个子服务器中有m个样本，满足$n = N \cdot m$，对于第i 组样本记样本下标集为H_i。

我们希望在分布式系统下做出关于总体均值μ的统计推断，一个直接且契合分布式系统特点的估计即

$$U_n = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{m}\sum_{k\in H_i} X_k\right)$$

而上述估计在统计学的稳健意义下并不是一个好的估计，直观地看，如果一个子服务器出现了计算失误，导致传递给总服务器的样本均值相当高，会大幅度地影响到U_n。

因此，我们对每个子服务器传递的样本均值取中位数(median of mean)作为总体均值μ 的估计，即

$$T_n = \underset{1\le i\le N}{med}\left(\frac{1}{m}\sum_{k\in H_i} X_k\right)$$

直观地看，如果一个子服务器出现了计算失误，由于中位数的稳健性，并不会对T_n造成太大影响。

下面假定N个子服务器中有αN个子服务器计算出现问题，我们考察median of mean 作为样本均值μ估计的极限性质。

1.4.2 Median of mean的Bahurder展开

若n个样本$X_1, X_2, ..., X_n$独立同分布，有均值μ,方差σ^2以及有限三阶矩，分别来自N个子服务器，并且设αN个子服务器计算其样本均值时出现问题，其中α 可随着子服务器数目N的变化而改变，并且$\exists 1 > r > 0, \alpha \le r$ 以及$\lim_{N\to\infty}\alpha = 0$。

则有：

$$\sqrt{n}(T_n-\mu)=\sqrt{2\pi}\sigma\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[\frac{1}{2}-\mathbf{1}\{Y_i\leq 0\}\right]+O_{a.s.}\left(\sqrt{\frac{N}{m}}+\frac{logN}{N^{\frac{1}{4}}}+\frac{\sqrt{logN}}{m^{\frac{1}{4}}}+\sqrt{\alpha logN}+\alpha^2\sqrt{N}\right)$$

其中，$Y_i=\frac{1}{\sqrt{m}}\sum_{k\in H_i}(X_k-\mu), i=1,2,...,N.$

证明　先证明

$$\underset{1\leq i\leq N}{med}\{Y_i\}=\sqrt{m}(T_n-\mu)=O_{a.s.}\left(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\right) \tag{14}$$

容易证明$Y_i, i=1,2,...,N$独立同分布，记其分布函数为F_Y。

由Berry-Esseen不等式知 $\underset{-\infty<x<+\infty}{sup}\left|F_Y(x)-\Phi(\frac{x}{\sigma})\right|=O(\frac{1}{\sqrt{m}})$，其中$\Phi$为标准正态分布的分布函数，并且这里$O$对所有满足0均值，且有相同二阶矩和三阶矩的随机变量是一致的。

另外，不妨设为后αN个子服务器计算样本均值时出现问题，并记$N_1=(1-\alpha)N$。

那么，

$$
\begin{aligned}
&P\Big\{\underset{1\le i\le N}{med}\{Y_i\} \ge C\big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\big)\Big\} \\
&= P\Big\{\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\Big\{Y_i \le C\big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\big)\Big\} \le \frac{1}{2}\Big\} \\
&\le P\Big\{\frac{1}{N}\sum_{i=1}^{N_1}\mathbf{1}\Big\{Y_i \le C\big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\big)\Big\} \le \frac{1}{2}\Big\} \\
&= P\Big\{\frac{1}{N_1}\sum_{i=1}^{N_1}\mathbf{1}\Big\{Y_i \le C\big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\big)\Big\} \le \frac{1}{2(1-\alpha)}\Big\} \\
&= P\Big\{\frac{1}{N_1}\sum_{i=1}^{N_1}\Big[P\Big\{Y_i \le C\big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\big)\Big\} - \mathbf{1}\Big\{Y_i \le C\big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\big)\Big\}\Big] \ge \\
&\quad P\Big\{Y_1 \le C\big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\big)\Big\} - \frac{1}{2(1-\alpha)}\Big\}
\end{aligned}
$$

考察最后的$P\big\{Y_1 \le C\big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\big)\big\} - \frac{1}{2(1-\alpha)}$：

由Berry-Esseen不等式、Lagrange中值定理以及正态函数密度函数ϕ在$[0,+\infty)$ 上单调减可得:

对于充分大的N以及充分大的C有：

$$
\begin{aligned}
&P\Big\{Y_1 \le C\big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\big)\Big\} - \frac{1}{2(1-\alpha)} \\
&\ge \Phi\Big(C\big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\big)/\sigma\Big) - \frac{1}{2(1-\alpha)} - \frac{C_1}{\sqrt{m}} \\
&= \frac{1}{\sigma}\phi(\xi)\cdot C\big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\big) - \frac{\alpha}{2(1-\alpha)} - \frac{C_1}{\sqrt{m}} \\
&\ge \frac{1}{\sigma}\phi(\frac{1}{\sigma})\cdot C\big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\big) - \frac{\alpha}{2(1-\alpha)} - \frac{C_1}{\sqrt{m}} \\
&\ge \frac{1}{2\sigma}\phi(\frac{1}{\sigma})\cdot C\frac{logN}{\sqrt{N}}
\end{aligned}
$$

因此，记$C_2 = \frac{1}{2\sigma}\phi(\frac{1}{\sigma})\cdot C$，并结合上式以及Berstein 不等式可得：

对于充分大的N：

$$
\begin{aligned}
&P\Big\{\underset{1\le i\le N}{med}\{Y_i\} \ge C\big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\big)\Big\} \\
&\le P\Bigg\{\frac{1}{N_1}\sum_{i=1}^{N_1}\Big[P\Big\{Y_i \le C\big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\big)\Big\} - \mathbf{1}\Big\{Y_i \le C\big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\big)\Big\}\Big] \ge C_2\frac{logN}{\sqrt{N}}\Bigg\} \\
&\le exp\Bigg\{-\frac{(N_1\cdot C_2\frac{logN}{\sqrt{N}})^2}{N_1+\frac{2N_1\cdot C_2\frac{logN}{\sqrt{N}}}{3}}\Bigg\} \\
&\le N^{-C_2}
\end{aligned}
$$

调整C_2使得上述概率关于N可和，再由Borel-Cantelli引理，即得：

$\exists C, \exists N^*, \forall N > N^*, \underset{1\le i\le N}{med}\{Y_i\} \le C\big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\big)$.

同理可得：

$\exists C, \exists N^*, \forall N > N^*, \underset{1\le i\le N}{med}\{Y_i\} \ge -C\big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\big)$.

从而(14)得证。

下面证明Median of mean的Bahurder展开：

记$\hat{F}_Y$为基于$Y_1, ..., Y_N$的经验分布函数，由之前的准备工作以及Berry-Esseen不等式可得：

$$
\begin{aligned}
&\frac{1}{2} = \hat{F}_Y\Big(\underset{1\le i\le N}{med}\{Y_i\}\Big) \\
&= \Big[F_Y\Big(\underset{1\le i\le N}{med}\{Y_i\}\Big) - F_Y(0)\Big] + \Big[\hat{F}_Y\Big(\underset{1\le i\le N}{med}\{Y_i\}\Big) - \hat{F}_Y(0)\Big] - \Big[F_Y\Big(\underset{1\le i\le N}{med}\{Y_i\}\Big) - F_Y(0)\Big] + \hat{F}_Y(0) \\
&= \Big[\Phi\Big(\underset{1\le i\le N}{med}\{Y_i\}/\sigma\Big) - \Phi(0)\Big] + \hat{F}_Y(0) + O_{a.s.}\Bigg(\frac{1}{\sqrt{m}} + \sqrt{\frac{logN}{N}\Big(\frac{logN}{\sqrt{N}} + \frac{1}{\sqrt{m}} + \alpha\Big)}\Bigg) \\
&= \frac{\phi(0)}{\sigma}\cdot \underset{1\le i\le N}{med}\{Y_i\} + \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{Y_i\le 0\} + O_{a.s.}\Bigg(\frac{1}{\sqrt{m}} + \sqrt{\frac{logN}{N}\Big(\frac{logN}{\sqrt{N}} + \frac{1}{\sqrt{m}} + \alpha\Big)} + \Big(\frac{logN}{\sqrt{N}} + \frac{1}{\sqrt{m}} + \alpha\Big)^2\Bigg) \\
&= \frac{\phi(0)}{\sigma}\cdot \underset{1\le i\le N}{med}\{Y_i\} + \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{Y_i\le 0\} + O_{a.s.}\Bigg(\frac{1}{\sqrt{m}} + \frac{logN}{N^{\frac{3}{4}}} + \frac{\sqrt{logN}}{N^{\frac{1}{2}}m^{\frac{1}{4}}} + \sqrt{\frac{\alpha logN}{N}} + \alpha^2\Bigg)
\end{aligned}
$$

移项后两边乘以$\sqrt{N}$整理即得:

$$
\sqrt{n}(T_n-\mu) = \sqrt{2\pi}\sigma\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\Big[\frac{1}{2} - \mathbf{1}\{Y_i\le 0\}\Big] + O_{a.s.}\Bigg(\sqrt{\frac{N}{m}} + \frac{logN}{N^{\frac{1}{4}}} + \frac{\sqrt{logN}}{m^{\frac{1}{4}}} + \sqrt{\alpha logN} + \alpha^2\sqrt{N}\Bigg)
$$

评注 **基**于上述结果并结合Berry-Esseen不等式以及Slusky定理可知：

若$\frac{n}{m^2}\to 0(N,m,n\to\infty)$ 以及$\alpha^2\sqrt{N}\to 0(N,m,n\to\infty)$

则$\sqrt{n}\frac{T_n-\mu}{\sigma}\xrightarrow{D} N(0,\frac{\pi}{2})$.

事实上，

- $\frac{n}{m^2}\to 0(N,m,n\to\infty)$的一个充分条件即$\exists 1<r<2, n<m^r$，而在之前的证明中我们也有过类似的假定。
- $\alpha^2\sqrt{N}\to 0(N,m,n\to\infty)$ 等价于$\alpha N = o(N^{\frac{3}{4}})$，意味着计算失误子服务器的数目是总子服务器数目0.75次幂的无穷小量。特别地，若$\exists r<\frac{3}{4}, \exists C>0$，当总样本量、子服务器数量以及每个子服务器中的样本量都较大时，若能确信出现失误的子服务器数目不会超过$C\cdot N^r$ 时，可以基于median of mean 来对总体均值μ 作出一些统计推断。
- 当没有子服务器计算失误时，利用各个子服务器的样本均值求平均得U_n有中心极限定理$\sqrt{n}\frac{U_n-\mu}{\sigma}\xrightarrow{D} N(0,1)$，其渐进方差1小于median of mean的渐进方差$\frac{\pi}{2}$。

1.5 中位数回归中参数估计的Bahurder展开

1.5.1 中位数回归简介

统计学中的回归问题，从总体的角度看，即通过建立适当的模型，用随机向量$\mathbf{X}$ 预测随机变量Y，对于一般的线性回归模型：

$$Y = \mathbf{X}^T\beta^* + e$$

其中，e为零均值、有限二阶矩的随机变量，并且与$\mathbf{X}$独立，β^*为总体中的未知参数。

然而，在很多的金融、经济学问题中，有限二阶矩这一假设有时并不成立，因此我们建立中位数回归模型：

$$Y = \mathbf{X}^T\beta^* + e$$

其中，e满足$P\{e \le 0\} = \frac{1}{2}$，并且与$\mathbf{X}$独立，$\beta^*$ 为总体中的未知参数。

另外，中位数回归模型的总体参数β^*有如下性质：

$$\beta^* = \underset{\beta}{argmin} E\Big[\big|Y - \mathbf{X}^T\beta^*\big|\Big]$$

事实上，注意到$E\Big[\big|Y - \mathbf{X}^T\beta^*\big|\Big]$为关于$\beta$的凸函数并且$e$与$\mathbf{X}$独立，从而由:

$$\begin{aligned}
&\frac{\partial}{\partial\beta} E\Big[\big|Y - \mathbf{X}^T\beta^*\big|\Big]\Big|_{\beta=\beta^*} \\
&= 2E\Big[\mathbf{X}\big(\mathbf{1}\{e \le 0\} - \frac{1}{2}\big)\Big] \\
&= 2E\Big[\mathbf{X}\Big] \cdot E\Big[\big(\mathbf{1}\{e \le 0\} - \frac{1}{2}\big)\Big] = 0
\end{aligned}$$

可知结论成立。

基于上述关于β^*的性质，基于样本$(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), ..., (\mathbf{X}_n, Y_n)$，对$\beta^*$一个很自然的参数估计即为：

$$\hat{\beta} = \underset{\beta}{argmin} \frac{1}{n} \sum_{i=1}^{n} \left| Y_i - \mathbf{X}_i^T \beta^* \right|$$

下面一节，我们假定$\mathbf{X}$的维数p是固定的，证明关于上述参数估计的Bahurder 展开。

1.5.2 中位数回归中参数估计的Bahurder展开

对于上述的中位数回归模型。若假定固定维数p维随机向量$\mathbf{X}$有正定协方差矩阵Σ，$\exists M > 0, |\mathbf{X}|_2 \leq M$几乎处处成立，$e$有分布函数$F$以及密度函数$f$，$f$可导且导函数有界，则有：

$$\hat{\beta} - \beta^* = \Sigma^{-1} \frac{\sum_{i=1}^{n} \mathbf{X}_i \left[\frac{1}{2} - \mathbf{1}\{e_i \leq 0\} \right]}{n f(0)} + O_{a.s.}\left(n^{-\frac{3}{4}} (logn)^{\frac{3}{4}} \right) \tag{15}$$

证明 先证明$\left| \hat{\beta} - \beta^* \right|_2 = O_{a.s.}\left(\sqrt{\frac{logn}{n}} \right)$

即证$\exists C > 0, \exists N, \forall n > N, \left| \hat{\beta} - \beta^* \right|_2 \leq C\sqrt{\frac{logn}{n}}$

几乎处处成立。

由凸优化的性质，往证：

$\exists C > 0, \exists N, \forall n > N, \forall \beta : \left| \beta - \beta^* \right|_2 = C\sqrt{\frac{logn}{n}}$，有$\sum_{i=1}^{n} \left| Y_i - \mathbf{X}_i^T \beta^* \right| \leq \sum_{i=1}^{n} \left| Y_i - \mathbf{X}_i^T \beta \right|$几乎处处成立。

亦即$\exists C > 0, \exists N, \forall n > N, \forall \beta : \left| \beta - \beta^* \right|_2 = C\sqrt{\frac{logn}{n}}$，有$\sum_{i=1}^{n} \left| e_i \right| - \sum_{i=1}^{n} \left| e_i - \mathbf{X}_i^T \left(\beta - \beta^* \right) \right| < 0$几乎成立。

考虑$\sum_{i=1}^n E\left[\left|e_i\right| - \sum_{i=1}^n \left|e_i - \mathbf{X}_i^T\left(\beta - \beta^*\right)\right|\right]$:

由恒等式$|y| - |y-z| = \int_0^z 1 - 2\cdot\mathbf{1}\{y \le x\}dx$以及$e$与$\mathbf{X}$的独立性可得:

$$
\begin{aligned}
&\sum_{i=1}^n E\left[\left|e_i\right| - \left|e_i - \mathbf{X}_i^T\left(\beta - \beta^*\right)\right|\right] \\
&= \sum_{i=1}^n E\left[\int_0^{\mathbf{X}_i^T(\beta-\beta^*)} 1 - 2F(x)dx\right] \\
&= 2\sum_{i=1}^n E\left[\int_0^{\mathbf{X}_i^T(\beta-\beta^*)} -f(0)x + \frac{f^{'}(\xi)x^2}{2}dx\right] \\
&= 2\sum_{i=1}^n E\left[-\frac{f(0)}{2}(\beta-\beta^*)^T\mathbf{X}_i\mathbf{X}_i^T(\beta-\beta^*) + \int_0^{\mathbf{X}_i^T(\beta-\beta^*)} \frac{f^{'}(\xi)x^2}{2}dx\right] \\
&= -nf(0)(\beta^*-\beta)^T\Sigma(\beta^*-\beta) + \sum_{i=1}^n E\left[\int_0^{\mathbf{X}_i^T(\beta-\beta^*)} f^{'}(\xi)x^2dx\right] \\
&\le -nf(0)(\beta^*-\beta)^T\Sigma(\beta^*-\beta) + \sum_{i=1}^n E\left[\int_{-|\mathbf{X}_i^T(\beta-\beta^*)|}^{|\mathbf{X}_i^T(\beta-\beta^*)|} f^{'}(\xi)x^2dx\right] \\
&\le -nf(0)(\beta^*-\beta)^T\Sigma(\beta^*-\beta) + \sum_{i=1}^n E\left[\int_{-|\mathbf{X}_i|_2|\beta-\beta^*|_2}^{|\mathbf{X}_i|_2|\beta-\beta^*|_2} f^{'}(\xi)x^2dx\right] \\
&\le -nf(0)(\beta^*-\beta)^T\Sigma(\beta^*-\beta) + \frac{n}{3}C_{f'}E[|\mathbf{X}_i|_2^3]\cdot|\beta-\beta^*|_2^3 \\
&= -nf(0)(\beta^*-\beta)^T\Sigma(\beta^*-\beta) + O\left(\frac{(logn)^{\frac{3}{2}}}{\sqrt{n}}\right)
\end{aligned}
$$

从而，当n充分大时有:

$$
\sum_{i=1}^n E\left[\left|e_i\right| - \left|e_i - \mathbf{X}_i^T\left(\beta - \beta^*\right)\right|\right] \le -\frac{3}{4}C^2\cdot\lambda_{min}(\Sigma)\cdot f(0)\cdot logn
$$

下面考虑$var\left[\left|e_i\right| - \left|e_i - \mathbf{X}_i^T\left(\beta - \beta^*\right)\right|\right]$

$$var\left(\left|e_i\right|-\left|e_i-\mathbf{X}_i^T\left(\beta-\beta^*\right)\right|\right)=E\left[var\left(\left|e_i\right|-\left|e_i-\mathbf{X}_i^T\left(\beta-\beta^*\right)\right|\Big|\mathbf{X}_i\right)\right]+var\left(E\left[\left|e_i\right|-\left|e_i-\mathbf{X}_i^T\left(\beta-\beta^*\right)\right|\Big|\mathbf{X}_i\right]\right) \quad (16)$$

考虑(16)式中的第一项，

$$\begin{aligned}
&E\left[var\left(\left|e_i\right|-\left|e_i-\mathbf{X}_i^T\left(\beta-\beta^*\right)\right|\Big|\mathbf{X}_i\right)\right]\\
&=E\left[var\Big(\int_0^{\mathbf{X}_i^T(\beta-\beta^*)}1-2\cdot\mathbf{1}\{e_i\le x\}dx\Big|\mathbf{X}_i\Big)\right]\\
&=E\left[4E\left[\Big(\int_0^{\mathbf{X}_i^T(\beta-\beta^*)}F(x)-2\cdot\mathbf{1}\{e_i\le x\}dx\Big)^2\Big|\mathbf{X}_i\right]\right]\\
&=4E\left[\int_0^{\mathbf{X}_i^T(\beta-\beta^*)}\int_0^{\mathbf{X}_i^T(\beta-\beta^*)}F\big(min(u,v)\big)-F(u)F(x)dudv\right]\\
&=(\beta^*-\beta)^T\Sigma(\beta^*-\beta)\cdot(1+o(1))
\end{aligned}$$

最后两个等号分别利用了Fubini定理以及F的Taylor展开以及f的非负有界性。

考虑(16)式中的第二项，由之前的计算以及C_r不等式可得：

$$\begin{aligned}
&var\left(E\left[\left|e_i\right|-\left|e_i-\mathbf{X}_i^T\left(\beta-\beta^*\right)\right|\Big|\mathbf{X}_i\right]\right)\\
&var\Big(-f(0)(\beta-\beta^*)^T\mathbf{X}_i\mathbf{X}_i^T(\beta-\beta^*)+\int_0^{\mathbf{X}_i^T(\beta-\beta^*)}f^{'}(\xi)x^2dx\Big)\\
&\le E\left[\Big(-f(0)(\beta-\beta^*)^T\mathbf{X}_i\mathbf{X}_i^T(\beta-\beta^*)+\int_0^{\mathbf{X}_i^T(\beta-\beta^*)}f^{'}(\xi)x^2dx\Big)^2\right]\\
&\le 2E\left[\Big(-f(0)(\beta-\beta^*)^T\mathbf{X}_i\mathbf{X}_i^T(\beta-\beta^*)\Big)^2\right]+2E\left[\Big(\int_0^{\mathbf{X}_i^T(\beta-\beta^*)}f^{'}(\xi)x^2dx\Big)^2\right]\\
&=O\Big(\frac{(logn)^2}{n^2}\Big)
\end{aligned}$$

最后一个等号用到了$|\mathbf{X}|_2$有界。

因此(16)式中，对于充分大的n有：

$$var\left(\left|e_i\right| - \left|e_i - \mathbf{X}_i^T\left(\beta - \beta^*\right)\right|\right) \leq 2 \cdot \lambda_{max} \cdot C^2 \cdot \frac{logn}{n}$$

注意到$\left|\left|e_i\right| - \left|e_i - \mathbf{X}_i^T\left(\beta - \beta^*\right)\right|\right| \leq \left|\mathbf{X}_i^T\left(\beta - \beta^*\right)\right| \leq \left|\mathbf{X}_i^T\right|_2 \cdot \left|\beta - \beta^*\right|_2$

即$\left|\left|e_i\right| - \left|e_i - \mathbf{X}_i^T\left(\beta - \beta^*\right)\right|\right|$有界，记其一个上界为M。

由Berstein不等式：

$$\begin{aligned}
&P\left\{\left|\sum_{i=1}^{n}\left[\left|e_i\right| - \left|e_i - \mathbf{X}_i^T\left(\beta - \beta^*\right)\right|\right] - \sum_{i=1}^{n} E\left[\left|e_i\right| - \left|e_i - \mathbf{X}_i^T\left(\beta - \beta^*\right)\right|\right]\right| \geq 0.5C^2 \cdot \lambda_{min}(\Sigma) \cdot f(0) \cdot logn\right\} \\
&\leq 2exp\left\{-\frac{\frac{n^2}{2} \cdot \frac{C^4 \cdot (f(0))^2 \cdot \lambda_{min}^2(\Sigma) \cdot (logn)^2}{4n^2}}{2n \cdot \lambda_{max}(\Sigma) \cdot C^2 \frac{logn}{n} + \frac{n}{6} M \frac{C^2 \cdot \lambda_{min}(\Sigma) \cdot f(0) \cdot logn}{n}}\right\} \\
&= 2exp\left\{-\frac{C^2 \cdot (f(0))^2 \cdot \lambda_{min}^2(\Sigma) \cdot logn}{8\left(2\lambda_{max}(\Sigma) + \frac{M}{6} \cdot f(0) \cdot \lambda_{min}(\Sigma)\right)}\right\}
\end{aligned}$$

对于充分大的C，使得:

$$P\left\{\left|\sum_{i=1}^{n}\left[\left|e_i\right| - \left|e_i - \mathbf{X}_i^T\left(\beta - \beta^*\right)\right|\right] - \sum_{i=1}^{n} E\left[\left|e_i\right| - \left|e_i - \mathbf{X}_i^T\left(\beta - \beta^*\right)\right|\right]\right| \geq 0.5C^2 \cdot \lambda_{min}(\Sigma) \cdot f(0) \cdot logn\right\}$$

可和，再有Borel - Cantelli引理即得，$\exists N, \forall n > N,$

$$\left|\sum_{i=1}^{n}\left[\left|e_i\right| - \left|e_i - \mathbf{X}_i^T\left(\beta - \beta^*\right)\right|\right] - \sum_{i=1}^{n} E\left[\left|e_i\right| - \left|e_i - \mathbf{X}_i^T\left(\beta - \beta^*\right)\right|\right]\right| \leq 0.5C^2 \cdot \lambda_{min}(\Sigma) \cdot f(0) \cdot logn$$

几乎处处成立。

结合$\sum_{i=1}^{n} E\left[\left|e_i\right| - \left|e_i - \mathbf{X}_i^T\left(\beta - \beta^*\right)\right|\right] \leq -\frac{3}{4}C^2 \cdot \lambda_{min}(\Sigma) \cdot f(0) \cdot logn$,

可知$\left|\hat{\beta} - \beta^*\right|_2 = O_{a.s.}\left(\sqrt{\frac{logn}{n}}\right)$得证。

即$\exists C_1 > 0, \exists N, \forall n > N, \left|\hat{\beta} - \beta^*\right|_2 \leq C_1\sqrt{\frac{logn}{n}}$

几乎处处成立。

再由：

$$
\begin{aligned}
&\frac{\sum_{i=1}^{n}\left(\frac{1}{2}\mathbf{X}_i\right)}{n} \\
&= \frac{\sum_{i=1}^{n}\left(F(0)\cdot\mathbf{X}_i\right)}{n} - \frac{\sum_{i=1}^{n}\mathbf{X}_i\cdot\mathbf{1}\{e_i \leq 0\}}{n} + \frac{\sum_{i=1}^{n}\mathbf{X}_i\cdot\mathbf{1}\{e_i \leq 0\}}{n} \\
&+ \frac{\sum_{i=1}^{n}\left(F(0)\cdot\mathbf{X}_i\right)}{n} - \frac{\sum_{i=1}^{n}\left(F(0)\cdot\mathbf{X}_i\right)}{n} + \frac{\sum_{i=1}^{n}\left(F(\mathbf{X}_i^T(\hat{\beta}-\beta^*))\cdot\mathbf{X}_i\right)}{n} - \frac{\sum_{i=1}^{n}\left(F(\mathbf{X}_i^T(\hat{\beta}-\beta^*))\cdot\mathbf{X}_i\right)}{n} \\
&+ \frac{\sum_{i=1}^{n}\mathbf{X}_i\cdot\mathbf{1}\{e_i \leq \mathbf{X}_i^T(\hat{\beta}-\beta^*)\}}{n} - \frac{\sum_{i=1}^{n}\mathbf{X}_i\cdot\mathbf{1}\{e_i \leq \mathbf{X}_i^T(\hat{\beta}-\beta^*)\}}{n}
\end{aligned}
$$

可得：

$$
\begin{aligned}
&\frac{\sum_{i=1}^{n}\mathbf{X}_i\cdot\left(\frac{1}{2}-\mathbf{1}\{e_i\le 0\}\right)}{n}\\
&=\frac{\sum_{i=1}^{n}\left(\left[F(\mathbf{X}_i^T(\hat{\beta}-\beta^*))-F(0)\right]\cdot\mathbf{X}_i\right)}{n}\\
&+\frac{\sum_{i=1}^{n}\left(\left[\mathbf{1}\{e_i\le\mathbf{X}_i^T(\hat{\beta}-\beta^*)\}-\mathbf{1}\{e_i\le 0\}\right]-\left[F(\mathbf{X}_i^T(\hat{\beta}-\beta^*))-F(0)\right]\right)\cdot\mathbf{X}_i}{n}\\
&+\frac{\sum_{i=1}^{n}\left(\left[F(0)-\mathbf{1}\{e_i\le\mathbf{X}_i^T(\hat{\beta}-\beta^*)\}\right]\cdot\mathbf{X}_i\right)}{n}
\end{aligned}
\tag{17}
$$

对于(17)式中的第一项，结合$|\mathbf{X}|_2$的有界性并由Taylor展开可得：

$$
\begin{aligned}
&\frac{\sum_{i=1}^{n}\left(\left[F(\mathbf{X}_i^T(\hat{\beta}-\beta^*))-F(0)\right]\cdot\mathbf{X}_i\right)}{n}\\
&=f(0)\cdot\frac{\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T}{n}\cdot(\hat{\beta}-\beta^*)+O_{a.s.}\left(\frac{logn}{n}\right)\\
&=f(0)\cdot\Sigma\cdot(\hat{\beta}-\beta^*)+\left(\frac{\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T}{n}-\Sigma\right)\cdot(\hat{\beta}-\beta^*)+O_{a.s.}\left(\frac{logn}{n}\right)\\
&=f(0)\cdot\Sigma\cdot(\hat{\beta}-\beta^*)+O_{a.s.}\left(\frac{logn}{n}\right)
\end{aligned}
$$

最后一步是结合随机矩阵理论中$\lambda_{max}\left(\hat{\Sigma}_{p\times p}-\Sigma\right)=O_{a.s.}\left(\sqrt{\frac{p+logp}{n}}\right)$的结论。

对于(17)式中的第二项，由于：

$$
\begin{aligned}
&var\Big(\mathbf{1}\{e_i \le \mathbf{X}_i^T(\hat{\beta}-\beta^*)\} - \mathbf{1}\{e_i \le 0\}\Big) \\
&\le E\Big[\Big|\mathbf{1}\{e_i \le \mathbf{X}_i^T(\hat{\beta}-\beta^*)\} - \mathbf{1}\{e_i \le 0\}\Big|^2\Big] \\
&\le E\Big[\mathbf{1}\Big\{-\big|\mathbf{X}_i^T(\hat{\beta}-\beta^*)\big| \le e_i \le \big|\mathbf{X}_i^T(\hat{\beta}-\beta^*)\big|\Big\}^2\Big] \\
&\le E\Big[\mathbf{1}\Big\{-\big|\mathbf{X}_i^T\big|_2\big|\hat{\beta}-\beta^*\big|_2 \le e_i \le \big|\mathbf{X}_i^T\big|_2\big|\hat{\beta}-\beta^*\big|_2\Big\}^2\Big] \\
&\le E\Big[\mathbf{1}\Big\{-\big|\mathbf{X}_i^T\big|_2\big|\hat{\beta}-\beta^*\big|_2 \le e_i \le \big|\mathbf{X}_i^T\big|_2\big|\hat{\beta}-\beta^*\big|_2\Big\}^2\Big] \\
&\le 2F(MC_1\sqrt{\frac{logn}{n}}) \\
&= O\Big(\sqrt{\frac{logn}{n}}\Big)
\end{aligned}
$$

结合1.3中的准备工作可知：

$$
\frac{\sum_{i=1}^n \Big(\Big[\mathbf{1}\{e_i \le \mathbf{X}_i^T(\hat{\beta}-\beta^*)\} - \mathbf{1}\{e_i \le 0\}\Big] - \Big[F(\mathbf{X}_i^T(\hat{\beta}-\beta^*)) - F(0)\Big]\Big)\cdot \mathbf{X}_i}{n} = O_{a.s.}\Big(n^{-\frac{3}{4}}(logn)^{-\frac{3}{4}}\Big)
$$

对于(17)式中的第三项：

由$\hat{\beta} = \underset{\beta}{argmin}\sum_{i=1}^n |Y_i - \mathbf{X}_i^T\beta|$，

可知$\exists a \in [-1,1]$：

$$
\sum_{i=1}^n -\mathbf{X}_i \cdot \mathbf{1}\{Y_i - \mathbf{X}_i^T\hat{\beta} > 0\} + \sum_{i=1}^n \mathbf{X}_i \cdot \mathbf{1}\{Y_i - \mathbf{X}_i^T\hat{\beta} < 0\} + \sum_{i=1}^n (-a)\mathbf{X}_i \cdot \mathbf{1}\{Y_i - \mathbf{X}_i^T\hat{\beta} = 0\} = 0
$$

从而：

$$\Big|\sum_{i=1}^{n}\mathbf{X}_i\cdot\mathbf{1}\{Y_i-\mathbf{X}_i^T\hat{\beta}\le 0\}-\frac{1}{2}\Big|_2\le\Big|\frac{a+1}{2}\Big|\sum_{i=1}^{n}\Big|\mathbf{X}_i\Big|_2\cdot\mathbf{1}\{Y_i-\mathbf{X}_i^T\hat{\beta}=0\}$$

由于e作为p维连续型随机变量，所以在几乎处处意义下有$\sum_{i=1}^{n}\Big|\mathbf{X}_i\Big|_2\cdot\mathbf{1}\{Y_i-\mathbf{X}_i^T\hat{\beta}=0\}$ 至多有p项非零，

事实上，若存在$i_1,i_2,...,i_{p+1}$使得，

$$e_{i_j}=\mathbf{X}_i^T(\hat{\beta}-\beta^*),j=1,2,...,p+1.$$

由$rank\left(\left[\mathbf{X}_{i_1},\mathbf{X}_{i_2},...,\mathbf{X}_{i_p}\right]\right)\le p$

从而$\exists\mu_1,\mu_2,...,\mu_{p+1}$不全为零，使得：$\sum_{j=1}^{p+1}\mu_j\mathbf{X}_{i_j}=0$，

即$\sum_{j=1}^{p+1}\mu_j e_{i_j}=0$，

而由e的连续性可知，$P\Big\{\sum_{j=1}^{p+1}\mu_j e_{i_j}=0\Big\}=0$。

即在几乎处处意义下有$\sum_{i=1}^{n}\Big|\mathbf{X}_i\Big|_2\cdot\mathbf{1}\{Y_i-\mathbf{X}_i^T\hat{\beta}=0\}$至多有$p$项非零。

从而$\dfrac{\sum_{i=1}^{n}\left(\left[F(0)-\mathbf{1}\{e_i\le\mathbf{X}_i^T(\hat{\beta}-\beta^*)\}\right]\cdot\mathbf{X}_i\right)}{n}=O_{a.s.}\left(\frac{1}{n}\right)$

综合上述结果，即得中位数回归中参数估计的Bahurder展开。

2 下降方法

2.1 概述

2.1.1 损失函数

统计学中最基本同时最重要的一组概念为总体与样本。本节我们分别从总体与样本的角度来分析机器学习中最基本的概念：损失。

2.1.1.1 总体角度 在总体意义下，考虑用随机向量$(X_1, X_2, ..., X_p)^T$来刻画（统计学专业术语为预测）随机变量Y，也就是统计学的回归问题。

事实上，每种预测方法唯一对应着一个$\mathbf{R}^p \to \mathbf{R}$的函数$g$。

对于一个$\mathbf{R}^2 \to \mathbf{R}$的函数$f$，我们以$E\left[f\Big(g(X_1, X_2, ..., X_p), Y\Big)\right]$来度量$g$作为一种预测方法的优劣程度。

加上若干正则条件（这些条件类似于泛函分析定义距离的几个条件），即称f为一个损失函数。

下面我们考虑两个特殊的损失：平方损失与分位数损失，在总体角度下的一些性质。

- **平方损失**
  损失函数$f(x, y) = (x - y)^2$在机器学习中被称为平方损失。

  我们考虑在平方损失意义下用随机向量$\mathbf{X} = (X_1, X_2, ..., X_p)^T$对随机变量$Y$的最优预测：

$$g^* = \underset{g}{argmax} E\left[\Big(g(X_1, X_2, ..., X_p) - Y\Big)^2\right]$$

  事实上，由

$$E\left[\Big(g(X_1, X_2, ..., X_p) - Y\Big)^2\right] = E\left[\Big(g(X_1, X_2, ..., X_p) - E[Y|X_1, X_2, ..., X_p]\Big)^2\right]$$
$$+ E\left[\Big(E[Y|X_1, X_2, ..., X_p] - Y\Big)^2\right] + 2E\left[\Big(E[Y|X_1, X_2, ..., X_p] - Y\Big) \cdot \Big(g(X_1, X_2, ..., X_p) - E[Y|X_1, X_2, ..., X_p]\Big)\right]$$

  利用双期望公式不难证明交叉项为0，因此：在平方损失意义下用随机向量$(X_1, X_2, ..., X_p)^T$对随机变量Y的最优预测即$E[Y|X_1, X_2, ..., X_p]$.

在正态假设下，即$(Y, X_1, X_2, ..., X_p)$的分布为$N(\mu, \Sigma)$，由多元正态的性质有：

1 $E[Y|X_1, X_2, ..., X_p] = \mu_Y + \Sigma_{YX} \cdot \Sigma_{XX}^{-1} \cdot \mu_X$

2 $cov(\mathbf{X}, Y - E[Y|X_1, X_2, ..., X_p]) = \mathbf{0}$，在正态假设下即独立。

自然地，在非正态假设下，我们可以有如下建模（假设）：

1 $E[Y|X_1, X_2, ..., X_p] = \mathbf{X}^T\beta^*$

2 $Y = \mathbf{X}^T\beta^* + \varepsilon$，其中$\varepsilon$与$\mathbf{X}$独立。

即统计学中线性回归的设定。

由模型假定不难得到ε满足$E[\varepsilon] = 0$。

在这种建模（假设）下，$\underset{g}{max}E\left[\Big(g(X_1, X_2, ..., X_p) - Y\Big)^2\right] = \underset{\beta\in\mathbf{R}^p}{max}E\left[\Big(\mathbf{X}^T\beta - Y\Big)^2\right]$

注意到$E\left[\Big(\mathbf{X}^T\beta - Y\Big)^2\right]$是关于$\beta$的凸函数，可由$\frac{\partial E[(\mathbf{X}^T\beta - Y)^2]}{\partial\beta} = 2E[\mathbf{X}(\mathbf{X}^T\beta - Y)] = 0$ 解得：

$\underset{\beta\in\mathbf{R}^p}{argmax}E\left[\Big(\mathbf{X}^T\beta - Y\Big)^2\right] = (E[\mathbf{X}\mathbf{X}^T])^{-1}E[\mathbf{X}Y].$

另外，由$\frac{\partial E[(\mathbf{X}^T\beta - Y)^2]}{\partial\beta}\Big|_{\beta=\beta^*} = 2E[\mathbf{X}\varepsilon] = 2E[\mathbf{X}] \cdot E[\varepsilon] = 0.$

可知模型中β^*满足$\beta^* = \underset{\beta\in\mathbf{R}^p}{argmax}E\left[\Big(\mathbf{X}^T\beta - Y\Big)^2\right].$

- **分位数损失**

记$f_q(x) = x \cdot \Big(\mathbf{1}\{x > 0\} - (1 - q)\Big)$，称$f(x, y) = f_q(y - x)$ 为q- 分位数损失。

我们考虑在q-分位数损失意义下用随机向量$(X_1, X_2, ..., X_p)^T$对随机变量Y的最优预测：

$$g^* = \underset{g}{argmax}E\left[f_q\Big(Y - g(X_1, X_2, ..., X_p)\Big)\right]$$

记$Q_q(Y|\mathbf{X})$为给定$(X_1, X_2, ..., X_p)^T$下Y的条件q-分位数，从而

$$
\begin{aligned}
&E\Big[f_q\Big(Y-g(\mathbf{X})\Big)\Big]\\
&=E\Big[\Big(Y-g(\mathbf{X})\Big)\cdot\Big(\mathbf{1}\{Y-g(\mathbf{X})>0\}-(1-q)\Big)\Big]\\
&=E\Big[\Big(Y-Q_q(Y|\mathbf{X})\Big)\cdot\Big(\mathbf{1}\{Y-Q_q(Y|\mathbf{X})>0\}-(1-q)\Big)\Big]\\
&\quad+E\Big[\Big(Y-g(\mathbf{X})\Big)\cdot\Big(\mathbf{1}\{Y-g(\mathbf{X})>0\}-\mathbf{1}\{Y-Q_q(Y|\mathbf{X})>0\}\Big)\Big]\\
&\quad+E\Big[\Big(Q_q(Y|\mathbf{X})-g(\mathbf{X})\Big)\cdot\Big(\mathbf{1}\{Y-Q_q(Y|\mathbf{X})>0\}-(1-q)\Big)\Big]
\end{aligned}
$$

其中对于第二项$E\Big[\Big(Y-g(\mathbf{X})\Big)\cdot\Big(\mathbf{1}\{Y-g(\mathbf{X})>0\}-\mathbf{1}\{Y-Q_q(Y|\mathbf{X})>0\}\Big)\Big]$

$$
\begin{aligned}
&E\Big[\Big(Y-g(\mathbf{X})\Big)\cdot\Big(\mathbf{1}\{Y-g(\mathbf{X})>0\}-\mathbf{1}\{Y-Q_q(Y|\mathbf{X})>0\}\Big)\Big]\\
&=E\Big[\Big(Y-g(\mathbf{X})\Big)\cdot\Big(\mathbf{1}\{Q_q(Y|\mathbf{X})\geq Y>g(\mathbf{X})\}\Big)+\Big(g(\mathbf{X})-Y\Big)\cdot\Big(\mathbf{1}\{Q_q(Y|\mathbf{X})<Y\leq g(\mathbf{X})\}\Big)\Big]\\
&\geq 0
\end{aligned}
$$

而第三项利用双期望公式不难证明为0。

因此在q-分位数损失意义下用随机向量$(X_1,X_2,...,X_p)^T$对随机变量Y的最优预测即为$Q_q(Y|\mathbf{X})$。

做如下建模（假定）：

1 $Q_q(Y|\mathbf{X})=\mathbf{X}^T\beta^*$

2 $Y=\mathbf{X}^T\beta^*+\varepsilon$，其中$\varepsilon$与$\mathbf{X}$独立。

即统计学中分位数回归的设定。

由模型假定不难得到ε满足$P\Big\{\varepsilon\leq 0\Big\}=q$。

与之前中位数回归中的证明类似，可知模型中的β^*满足$\beta^*=\underset{\beta\in\mathbf{R}^p}{argmax}E\Big[f_q\Big(Y-\mathbf{X}^T\beta\Big)\Big]$

2.1.1.2 样本角度

- **平方损失**

  在上面关于总体方面的讨论中求解$\beta^* = \underset{\beta \in \mathbf{R}^p}{argmax} E\left[\left(\mathbf{X}^T\beta - Y\right)^2\right]$问题，

  但是对于总体未知、仅有样本$(Y_i, \mathbf{X}_i) = (Y_i, X_{i1}, ..., X_{ip}), i = 1, 2, ..., n$的情况，上述问题是无法求解的。

  那么很自然地，我们求解上述问题的样本版本：$\hat{\beta} = \underset{\beta \in \mathbf{R}^p}{argmax} \frac{1}{n} \sum_{i=1}^n \left(\mathbf{X}_i^T\beta - Y_i\right)^2$.

  即基于样本的最小二乘问题。

- **分位数损失**

  同理，在总体未知，仅有样本的情况下，我们求解$\beta^* = \underset{\beta \in \mathbf{R}^p}{argmax} E\left[f_q\left(Y - \mathbf{X}^T\beta\right)\right]$问题的样本版本:

  $\hat{\beta} = \underset{\beta \in \mathbf{R}^p}{argmax} \frac{1}{n} \sum_{i=1}^n f_q\left(Y_i - \mathbf{X}_i^T\beta\right)$

  若取$q = \frac{1}{2}$即基于样本的最小一乘问题。

- **支持向量机**

  考虑总体方面的分类问题，即用随机向量$\mathbf{X} = (X_1, X_2, ..., X_p)^T$对支撑为$\{-1, 1\}$的随机变量$Y$进行预测。

  考虑损失函数$f(x, y) = (1 - xy)_+ = max\{1 - xy, 0\}$,

  此时求解样本版本:$\hat{\beta} = \underset{\beta \in \mathbf{R}^p}{argmax} \frac{1}{n} \sum_{i=1}^n \left(1 - Y_i\mathbf{X}_i^T\beta\right)_+$即SVM支持向量机问题。

- **极大似然诱导出的损失函数**

  继续考虑总体方面的分类问题，即用随机向量$\mathbf{X} = (X_1, X_2, ..., X_p)^T$ 对支撑为$\{-1, 1\}$的随机变量Y进行预测。

  做如下建模：

  $$P\{Y = 1|\mathbf{X}\} = \frac{1}{1 + e^{\mathbf{X}^T\beta^*}}$$

  根据统计学中的极大似然思想，基于样本我们有β^*的极大似然估计$\hat{\beta} = \underset{\beta \in \mathbf{R}^p}{argmax} \Pi_{i=1}^n \frac{1}{1+e^{Y_i\mathbf{X}_i^T\beta}}$.

  取损失函数$f(x, y) = log(1 + e^{xy})$.

  则上述问题等价于$\hat{\beta} = \underset{\beta \in \mathbf{R}^p}{argmin} \sum_{i=1}^n f(\mathbf{X}_i^T\beta, Y_i)$

  这里的损失函数构造想法实际上是源于统计学的极大似然理论。

2.1.2 两种基本的下降法简介

考虑一般优化问题的总体版本与样本版本：

$$\beta^* = \underset{\beta \in \mathbf{R}^p}{argmin} E\Big[f(\mathbf{X};\beta)\Big]$$
$$\hat{\beta} = \underset{\beta \in \mathbf{R}^p}{argmin} \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{X}_i;\beta)$$

其中,β^*即统计学中总体中的未知参数，$\hat{\beta}$即基于样本对未知参数β^* 的估计，而如何实现$\hat{\beta}$这一统计量便需要一些计算数学的工具。

考虑解$\hat{\beta} = \underset{\beta \in \mathbf{R}^p}{argmin} \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{X}_i;\beta)$,

可以利用两种基本的下降方法求解：

- **Newton迭代**

$$\beta_{t+1} = \beta_t - \eta\Big(H_n\Big|_{\beta=\beta_t}\Big)^{-1}\frac{1}{n}\sum_{i=1}^{n}\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta_t}$$

其中，β_t为上一次迭代得到的结果，$\partial\cdot$代表对β求偏导的算子，η为步长，H_n为Hessian矩阵即：

$$H_n = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 f(\mathbf{X}_i;\beta)}{\partial\beta\partial\beta^T} \in \mathbf{R}^{p\times p}$$

- **梯度下降**

$$\beta_{t+1} = \beta_t - \eta\frac{1}{n}\sum_{i=1}^{n}\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta_t}$$

符号含义同上。

事实上，1.3中基于下降方法的中位数估计正是利用了上面的思想构造boosting估计量。

在下面两节中，我们重点阐述：

- $\beta^*, \hat{\beta}$的关系，即参数估计问题（统计学）
- 下降方法中的$\hat{\beta}, \beta_t$的关系，即统计量实现问题（优化）

2.1.3 参数估计：β^*与$\hat{\beta}$

若$f(\mathbf{X};\beta)$是关于β的凸函数，关于β的梯度$\partial f(\mathbf{X};\beta)$在$\beta=\beta^*$处的每个分量作为随机变量方差有限，

并且$H_n = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 f(\mathbf{X}_i;\beta)}{\partial\beta\partial\beta^T}$的特征值有在几乎处处意义下关于样本$\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ 以及β 一致的非负下界λ_L。

则$|\beta^* - \hat{\beta}|_2 = O_{a.s.}(\frac{logn}{\sqrt{n}})$.

证明 等价于证明$\exists N, \exists C, \forall n > N, |\beta^* - \hat{\beta}|_2 \leq C\frac{logn}{\sqrt{n}}$,几乎处处成立。

由f的凸性，等价于证明:

$$\exists N, \exists C, \forall n > N, \forall \beta : |\beta^* - \beta|_2 = C\frac{logn}{\sqrt{n}}$$
$$\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{X}_i;\beta) > \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{X}_i;\beta^*)$$

几乎处处成立。

对于$\forall \beta : |\beta^* - \beta|_2 = C\frac{logn}{\sqrt{n}}$，由Taylor 展开可得：

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{X}_i;\beta) &= \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{X}_i;\beta^*) \\ &+ \frac{1}{n}\sum_{i=1}^{n}(\beta-\beta^*)^T\cdot\partial f(\mathbf{X}_i;\beta^*) \\ &+ \frac{1}{2n}\sum_{i=1}^{n}(\beta-\beta^*)^T\cdot\partial^2 f(\mathbf{X}_i;\tilde{\beta}_i)\cdot(\beta-\beta^*)\end{aligned}$$

对于$\frac{1}{n}\sum_{i=1}^{n}(\beta-\beta^*)^T\cdot\partial f(\mathbf{X}_i;\beta^*)$:

注意到$\partial f(\mathbf{X}_i;\beta^*)$独立同分布，每个分量期望为0、方差有限，由之前的准备工作可知

$$\Big|\frac{1}{n}\sum_{i=1}^{n}\partial f(\mathbf{X}_i;\beta^*)\Big|_2 = O_{a.s.}\Big(\frac{logn}{\sqrt{n}}\Big)$$

即$\exists C_1,\exists N_1,\forall n>N_1,|\frac{1}{n}\sum_{i=1}^{n}\partial f(\mathbf{X}_i;\beta^*)|_2 \le C_1\frac{logn}{\sqrt{n}}$,几乎处处成立。

并由Cauchy不等式，当$n>N_1$时有：

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}(\beta-\beta^*)^T\cdot\partial f(\mathbf{X}_i;\beta^*) \\ &\ge -\Big|\frac{1}{n}\sum_{i=1}^{n}\partial f(\mathbf{X}_i;\beta^*)\Big|_2\cdot\Big|\beta-\beta^*\Big|_2 \\ &\ge -C\cdot C_1\frac{(logn)^2}{n}\end{aligned}$$

几乎处处成立。

对于$\frac{1}{2n}\sum_{i=1}^{n}(\beta-\beta^*)^T\cdot\partial^2 f(\mathbf{X}_i;\tilde{\beta}_i)\cdot(\beta-\beta^*)$:

$$\begin{aligned}
&\frac{1}{2n}\sum_{i=1}^{n}(\beta-\beta^*)^T\cdot\partial^2 f(\mathbf{X}_i;\tilde{\beta}_i)\cdot(\beta-\beta^*)\\
&=(\beta-\beta^*)^T\cdot\frac{1}{2n}\sum_{i=1}^{n}\partial^2 f(\mathbf{X}_i;\tilde{\beta}_i)\cdot(\beta-\beta^*)\\
&\geq\frac{\lambda_L}{2}\cdot\frac{C^2(logn)^2}{n}
\end{aligned}$$

几乎处处成立。

取C充分大使得：$\frac{\lambda_L}{2}\cdot\frac{C^2(logn)^2}{n}-C\cdot C_1\frac{(logn)^2}{n}>0$

此时有$\frac{1}{n}\sum_{i=1}^{n}f(\mathbf{X}_i;\beta)>\frac{1}{n}\sum_{i=1}^{n}f(\mathbf{X}_i;\beta^*)$,几乎处处成立。

得证。

评注 上述结论的条件之一：Hessian 矩阵H_n的特征值有在几乎处处意义下关于样本$\mathbf{X}_1,\mathbf{X}_2,...,\mathbf{X}_n$ 以及β 一致的非负下界λ_L。

在统计学中有很多模型可以得到满足，下面考虑之前的线性回归模型：

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}f(\mathbf{X}_i;\beta)=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\mathbf{X}_i^T\beta)^2\\
&\frac{\partial^2 f(\mathbf{X};\beta)}{\partial\beta\partial\beta^T}=\frac{2}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T
\end{aligned}$$

当$p<n$时结合随机矩阵理论，容易证明条件满足。

2.1.4 统计量实现：$\hat{\beta}$与β_t

下面我们考虑统计量实现，即当迭代次数t足够大时，算法第t次迭代解β_t与优化问题$\underset{\beta\in\mathbf{R}^p}{argmin}\frac{1}{n}\sum_{i=1}^{n}f(\mathbf{X}_i;\beta)$的解$\hat{\beta}$的关系。

对于Newton迭代，由Taylor展开可得：

$$\begin{aligned}
&\beta_{t+1}-\hat{\beta}\\
&=\beta_t-\hat{\beta}-\eta\Big(H_n\Big|_{\beta=\beta_t}\Big)^{-1}\frac{1}{n}\sum_{i=1}^{n}\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta_t}\\
&=\beta_t-\hat{\beta}-\eta\Big(H_n\Big|_{\beta=\beta_t}\Big)^{-1}\cdot\Big[\frac{1}{n}\sum_{i=1}^{n}\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta_t}-\frac{1}{n}\sum_{i=1}^{n}\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\hat{\beta}}\Big]\\
&=\Big[I-\eta\Big(H_n\Big|_{\beta=\beta_t}\Big)^{-1}\cdot H_n\Big|_{\beta=\tilde{\beta}}\Big]\cdot\Big(\beta_t-\hat{\beta}\Big)
\end{aligned}$$

从而，$\left|\beta_{t+1}-\hat{\beta}\right|_2\le\left|\Big[I-\eta\Big(H_n\Big|_{\beta=\beta_t}\Big)^{-1}\cdot H_n\Big|_{\beta=\tilde{\beta}}\Big]\right|\cdot\left|\beta_t-\hat{\beta}\right|_2$

其中，$\left|\Big[I-\eta\Big(H_n\Big|_{\beta=\beta_t}\Big)^{-1}\cdot H_n\Big|_{\beta=\tilde{\beta}}\Big]\right|$为矩阵在$l_2-norm$下的算子模，即特征值的绝对值中最大的。

因此，粗略地看，只需要调整η使得$\left|\Big[I-\eta\Big(H_n\Big|_{\beta=\beta_t}\Big)^{-1}\cdot H_n\Big|_{\beta=\tilde{\beta}}\Big]\right|\le r<1$ 即可。

由之前所证，$|\beta^*-\hat{\beta}|_2=O_{a.s.}(\frac{logn}{\sqrt{n}})$，

因此要$|\beta_t-\hat{\beta}|_2$的阶数也是$\frac{logn}{\sqrt{n}}$，只需迭代$t=O(logn)$ 即可，此时$|\beta^*-\beta_t|_2=O_{a.s.}(\frac{logn}{\sqrt{n}})$.

对于梯度下降，由Taylor展开同理可得：

$$
\begin{aligned}
&\beta_{t+1} - \hat{\beta} \\
&= \beta_t - \hat{\beta} - \eta \frac{1}{n} \sum_{i=1}^{n} \partial f(\mathbf{X}_i; \beta)\Big|_{\beta=\beta_t} \\
&= \beta_t - \hat{\beta} - \eta \cdot \Big[\frac{1}{n} \sum_{i=1}^{n} \partial f(\mathbf{X}_i; \beta)\Big|_{\beta=\beta_t} - \frac{1}{n} \sum_{i=1}^{n} \partial f(\mathbf{X}_i; \beta)\Big|_{\beta=\hat{\beta}}\Big] \\
&= \Big[I - \eta \cdot H_n\Big|_{\beta=\tilde{\beta}}\Big] \cdot \Big(\beta_t - \hat{\beta}\Big)
\end{aligned}
$$

只需要调整η使得$\left|\left[I - \eta \cdot H_n\Big|_{\beta=\tilde{\beta}}\right]\right| \leq r < 1$即可。

通过比较Newton迭代与梯度下降，我们发现，梯度下降对η的选择更加敏感。事实上，在一定条件下，Newton迭代中的η取$\eta = 1$即可。

2.2 Online设定下的随机梯度下降

2.2.1 背景简介

随机梯度下降通过：

$$
\beta_{t+1} = \beta_t = \eta_t \cdot \partial f(\mathbf{X}_{i_t}; \beta)\Big|_{\beta=\beta_t}
$$

来求解$\hat{\beta} = \underset{\beta \in \mathbf{R}^p}{argmin} \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{X}_i; \beta)$.

不同于梯度下降需要计算所有样本下对应的梯度$\partial f(\mathbf{X}_1; \beta)\Big|_{\beta=\beta_t}, ..., \partial f(\mathbf{X}_n; \beta)\Big|_{\beta=\beta_t}$，

随机梯度下降只需要计算随机一个样本$\mathbf{X}_{i_t}$对应的梯度。

而相应地，其步长η依赖于迭代次数t，随着迭代次数的增加而减少。

下面我们重点考虑一种特殊的情况：

假设样本获取过程是一个数据流，也就是我们随着时间的推移依次逐个地获取样本$\mathbf{X}_1, ..., \mathbf{X}_n$，即所谓Online设定。

在Online设定下，梯度下降算法即：

$$\beta_{t+1} = \beta_t = \eta_t \cdot \partial f(\mathbf{X}_t; \beta)\Big|_{\beta=\beta_t}$$

而重要的是，若假定样本之间的相互独立性，那么在这种设定下，显然β_t与$\mathbf{X}_t$是独立的。

2.2.2 Online设定下的随机梯度下降参数估计问题中β_t与β^*

下面我们考虑算法第t次迭代解β_t与总体参数β^*的关系：

这里，β^*满足：

$$\beta^* = \underset{\beta \in \mathbf{R}^p}{argmin} E\Big[f(\mathbf{X}; \beta)\Big]$$

记$g(\beta) = E\Big[\partial f(\mathbf{X}; \beta)\Big]$，对于$\exists C_1, \forall \beta \in \mathbf{R}^p, |g(\beta)|_2^2 \leq C_1$.

以及$\frac{\partial g}{\partial \beta^T}\Big|_{\beta=\beta^*}$半正定，并记其最小特征值为$\lambda_m$，

除此之外，还假设g各分量的混合三阶导数对$\forall \beta \in \mathbf{R}^p$一致有界。

并且Online设定下梯度下降方法中步长η_t选取为$\frac{\theta}{t+1}$，此时当迭代次数t充分大时，$\exists C_2, E[|\beta_t - \beta^*|_2^2] \leq \frac{C_2}{t+1}$.

证明　由:

$$\beta_{t+1} - \beta^* = \beta_t - \beta^* - \eta_t \cdot \partial f(\mathbf{X}_t; \beta)\Big|_{\beta=\beta_t}$$

可得：

$$E\left[\left|\beta_{t+1}-\beta^*\right|_2^2\right]=E\left[\left|\beta_t-\beta^*\right|_2^2\right]-2\eta_t\cdot E\left[(\beta_t-\beta^*)^T\cdot\partial f(\mathbf{X}_t;\beta)\big|_{\beta=\beta_t}\right]+\eta_t^2\cdot E\left[\left|\partial f(\mathbf{X}_t;\beta)\big|_{\beta=\beta_t}\right|_2^2\right] \tag{18}$$

对于(18)式中第二项：$\eta_t\cdot E\left[(\beta_t-\beta^*)^T\cdot\partial f(\mathbf{X}_t;\beta)\big|_{\beta=\beta_t}\right]$,

注意到β_t只依赖于$\mathbf{X}_1,\mathbf{X}_2,...,\mathbf{X}_{t-1}$，从而与$\mathbf{X}_t$独立，并由双期望公式可得：

对于充分大的t：

$$\begin{aligned}
&\eta_t\cdot E\left[(\beta_t-\beta^*)^T\cdot\partial f(\mathbf{X}_t;\beta)\big|_{\beta=\beta_t}\right]\\
&=\eta_t\cdot E\left[E\left[(\beta_t-\beta^*)^T\cdot\partial f(\mathbf{X}_t;\beta)\big|_{\beta=\beta_t}\Big|\beta_t\right]\right]\\
&=\eta_t\cdot E\left[(\beta_t-\beta^*)^T\cdot E\left[\partial f(\mathbf{X}_t;\beta)|\beta_t\right]\Big|_{\beta=\beta_t}\right]\\
&=\eta_t\cdot E\left[(\beta_t-\beta^*)^T\cdot E\left[\partial f(\mathbf{X}_t;\beta)\right]\Big|_{\beta=\beta_t}\right]\\
&=\eta_t\cdot E\left[(\beta_t-\beta^*)^T\cdot E\left[\partial f(\mathbf{X};\beta)\right]\Big|_{\beta=\beta_t}\right]\\
&=\eta_t\cdot E\left[(\beta_t-\beta^*)^T\cdot g(\beta)\Big|_{\beta=\beta_t}\right]\\
&=\eta_t\cdot E\left[(\beta_t-\beta^*)^T\cdot\left(g(\beta)\Big|_{\beta=\beta_t}-g(\beta)\Big|_{\beta=\beta^*}\right)\right]\\
&=\eta_t\cdot E\left[(\beta_t-\beta^*)^T\cdot\frac{\partial g}{\partial\beta^T}\Big|_{\beta=\beta^*}\cdot(\beta_t-\beta^*)\right]+\eta_t\cdot O\left(E\left[\left|\beta_t-\beta^*\right|_2^3\right]\right)\\
&\geq\eta_t\frac{\lambda_m}{2}E\left[\left|\beta_t-\beta^*\right|_2^2\right]
\end{aligned}$$

对于(18)式中第三项：$\eta_t^2 \cdot E\left[\left|\partial f(\mathbf{X}_t;\beta)\big|_{\beta=\beta_t}\right|_2^2\right]$,

由双期望公式：

$$
\begin{aligned}
&\eta_t^2 \cdot E\left[\left|\partial f(\mathbf{X}_t;\beta)\big|_{\beta=\beta_t}\right|_2^2\right] \\
&= \eta_t^2 \cdot E\left[E\left[\left|\partial f(\mathbf{X}_t;\beta)\big|_{\beta=\beta_t}\right|_2^2\middle|\beta_t\right]\right] \\
&= \eta_t^2 \cdot E\left[E\left[\left|\partial f(\mathbf{X}_t;\beta)\right|_2^2\middle|\beta_t\right]\Big|_{\beta=\beta_t}\right] \\
&= \eta_t^2 \cdot E\left[E\left[\left|\partial f(\mathbf{X}_t;\beta)\right|_2^2\right]\Big|_{\beta=\beta_t}\right] \\
&= \eta_t^2 \cdot E\left[E\left[\left|\partial f(\mathbf{X};\beta)\right|_2^2\right]\Big|_{\beta=\beta_t}\right] \\
&= \eta_t^2 \cdot E\Big[g(\beta_t)\Big] \\
&\leq C_1 \cdot \eta_t^2
\end{aligned}
$$

从而，

$$
\begin{aligned}
&E\left[\left|\beta_{t+1}-\beta^*\right|_2^2\right] \\
&= E\left[\left|\beta_t-\beta^*\right|_2^2\right] - 2\eta_t \cdot E\left[(\beta_t-\beta^*)^T \cdot \partial f(\mathbf{X}_t;\beta)\big|_{\beta=\beta_t}\right] + \eta_t^2 \cdot E\left[\left|\partial f(\mathbf{X}_t;\beta)\big|_{\beta=\beta_t}\right|_2^2\right] \\
&\leq (1-\lambda_m \cdot \eta_t) E\left[\left|\beta_t-\beta^*\right|_2^2\right] + C_1 \cdot \eta_t^2
\end{aligned}
$$

再由数学归纳法，即得所证结论。

2.3 Communication efficiency

2.3.1 背景简介

对于分布式系统设定，n个样本$\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$分别来自N个子服务器，每个子服务器中有m个样本，满足$n = N \cdot m$并且$\exists A > 1, n \leq m^A$。另外存在一个总服务器来汇总处理来自N个子服务器的各类信息，并作出关于总体统计推断。

仍然考虑一般优化问题的总体版本:

$$\beta^* = \underset{\beta \in \mathbf{R}^p}{argmin} E\Big[f(\mathbf{X}; \beta)\Big]$$

中β^*的估计问题

对于已有初步估计$\hat{\beta}_0$，设$|\hat{\beta}_0 - \beta^*| = O_{a.s.}(a_n)$，这里不妨设$\exists C_c, C_c\sqrt{\frac{logn}{n}} \leq a_n \leq 1,$

我们考虑基于$\hat{\beta}_0$的boosting估计量：

$$\hat{\beta}_1 = \underset{\beta \in \mathbf{R}^p}{argmin}\left\{ \frac{1}{m} \sum_{i \in H_1} f(\mathbf{X}_i; \beta) - \beta^T \Big[\frac{1}{m} \sum_{i \in H_1} \partial f(\mathbf{X}_i; \beta)\big|_{\beta = \hat{\beta}_0} - \frac{1}{n} \sum_{i=1}^{n} \partial f(\mathbf{X}_i; \beta)\big|_{\beta = \hat{\beta}_0}\Big] \right\}$$

事实上，由$\sum_{i=1}^{n} \partial f(\mathbf{X}_i; \beta)\big|_{\beta = \hat{\beta}_0} = \sum_{k=1}^{N} \sum_{i \in H_k} \partial f(\mathbf{X}_i; \beta)\big|_{\beta = \hat{\beta}_0}$,

可知，上述boosting估计量在分布式系统中很容易实现：只需要每个子服务器计算基于各自样本所有梯度后取加和，将这个p维的加和梯度向量传输到总服务器中再加和，并将经过两次求和的p 维向量传输给第一个子服务器，在第一个子服务器上解一个优化问题。而这个过程即本节标题中的"communication"。

在下一小节中，我们重点论证本节标题中的"efficiency"。

2.3.2 Communication efficiency的极限性质

若总体上，$f(\mathbf{X}; \beta)$满足：

- $|\partial f(\mathbf{X};\beta)|_2^2$的期望对$\forall\beta\in\mathbf{R}^p$ 一致有界.
- $f(\mathbf{X};\beta)$关于β在l_2-norm下Lipschitz连续，即$|\partial f(\mathbf{X};\beta_1)-\partial f(\mathbf{X};\beta_2)|_2\le C_{\mathbf{X}}|\beta_1-\beta_2|_2$，其中随机变量$C_{\mathbf{X}}$存在有限二阶矩。

另外，$H_m=\frac{1}{m}\sum_{i\in H_1}\frac{\partial^2 f(\mathbf{X}_i;\beta)}{\partial\beta\partial\beta^T}$的特征值有在几乎处处意义下关于样本$\mathbf{X}_1,\mathbf{X}_2,...,\mathbf{X}_n$ 以及β 一致的非负下界λ_L，

则对于基于$\hat\beta_0$的boosting估计量$\hat\beta_1$满足：

$$\left|\hat\beta_1-\beta^*\right|_2=O_{a.s.}\left(\sqrt{\frac{logn}{n}}+a_n\cdot\sqrt{\frac{logm}{m}}\right)$$

证明 记$h(\beta)=\frac{1}{m}\sum_{i\in H_1}f(\mathbf{X}_i;\beta)-\beta^T\left[\frac{1}{m}\sum_{i\in H_1}\partial f(\mathbf{X}_i;\beta)\big|_{\beta=\hat\beta_0}-\frac{1}{n}\sum_{i=1}^n\partial f(\mathbf{X}_i;\beta)\big|_{\beta=\hat\beta_0}\right]$.

即$\hat\beta_1=\underset{\beta\in\mathbf{R}^p}{argmin}\{h(\beta)\}$.

由凸优化的性质，往证$\exists C>0,\forall\beta:\left|\beta-\beta^*\right|_2=C\left(\sqrt{\frac{logn}{n}}+a_n\cdot\sqrt{\frac{logm}{m}}\right),h(\beta)>h(\beta^*)$，几乎处处成立。

由Taylor展开：

$$\begin{aligned}
&h(\beta)-h(\beta^*)\\
&=\left(\beta-\beta^*\right)^T\cdot\Big[\frac{1}{m}\sum_{i\in H_1}\partial f(\mathbf{X}_i;\beta)|_{\beta=\beta^*}-\frac{1}{m}\sum_{i\in H_1}\partial f(\mathbf{X}_i;\beta)|_{\beta=\hat\beta_0}+\frac{1}{n}\sum_{i=1}^n\partial f(\mathbf{X}_i;\beta)|_{\beta=\hat\beta_0}\Big]\\
&\quad+\frac{1}{2}\left(\beta-\beta^*\right)^T\cdot H_m|_{\beta=\tilde\beta}\cdot\left(\beta-\beta^*\right)\\
&=\left(\beta-\beta^*\right)^T\cdot\frac{1}{m}\sum_{i\in H_1}\Big(\big[\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta^*}-\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\hat\beta_0}\big]-\big[E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\beta^*}-E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\hat\beta_0}\big]\Big)\\
&\quad+\left(\beta-\beta^*\right)^T\cdot\frac{1}{n}\sum_{i=1}^n\Big(\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\hat\beta_0}-E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\hat\beta_0}\Big)\\
&\quad+\frac{1}{2}\left(\beta-\beta^*\right)^T\cdot H_m\Big|_{\beta=\tilde\beta}\cdot\left(\beta-\beta^*\right)
\end{aligned}\tag{19}$$

对于(19)式中$\Big|\frac{1}{m}\sum_{i\in H_1}\Big([\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta^*}-\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\hat{\beta}_0}]-[E[\partial f(\mathbf{X}_i;\beta)]\Big|_{\beta=\beta^*}-E[\partial f(\mathbf{X}_i;\beta)]\Big|_{\beta=\hat{\beta}_0}]\Big)\Big|_2$.

由$|\hat{\beta}_0-\beta^*|=O_{a.s.}(a_n)$可知，$\exists C_1,\exists N,\forall n>N,|\beta^*-\hat{\beta}_0|\le C_1\cdot a_n$，几乎处处成立。因此，对于$n>N$：

$$\begin{aligned}&\Big|\frac{1}{m}\sum_{i\in H_1}\Big([\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta^*}-\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\hat{\beta}_0}]-[E[\partial f(\mathbf{X}_i;\beta)]\Big|_{\beta=\beta^*}-E[\partial f(\mathbf{X}_i;\beta)]\Big|_{\beta=\hat{\beta}_0}]\Big)\Big|_2\\&\le\sup_{\beta_C:|\beta_C-\beta^*|_2\le C_1\cdot a_n}\Big|\frac{1}{m}\sum_{i\in H_1}\Big([\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta^*}-\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta_C}]-[E[\partial f(\mathbf{X}_i;\beta)]\Big|_{\beta=\beta^*}-E[\partial f(\mathbf{X}_i;\beta)]\Big|_{\beta=\beta_C}]\Big)\Big|_2\end{aligned}$$

由:

$$\begin{aligned}&E\left[\Big|\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta^*}-\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta_C}\Big|_2^2\right]\\&\le E\left[C_{\mathbf{X}}^2\right]\cdot\Big|\beta^*-\beta_C\Big|_2^2\\&\le E\left[C_{\mathbf{X}}^2\right]\cdot C_1\cdot a_n\end{aligned}$$

可知$\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta^*}-\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta_C}$每一维的方差皆小于等于$E\left[C_{\mathbf{X}}^2\right]\cdot C_1^2\cdot a_n^2$.

这里由于考虑的问题中p固定，则由1.3的准备工作，可得：$\forall\beta_C:|\beta_C-\beta^*|_2\le C_1\cdot a_n,\exists C_2$,对于充分大的$n$有：

$$P\left\{\Big|\frac{1}{m}\sum_{i\in H_1}\Big([\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta^*}-\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta_C}]-[E[\partial f(\mathbf{X}_i;\beta)]\Big|_{\beta=\beta^*}-E[\partial f(\mathbf{X}_i;\beta)]\Big|_{\beta=\beta_C}]\Big)\Big|_2>C_2\cdot a_n\sqrt{\frac{logm}{m}}\right\}\le m^{-A_1}$$

这里可以通过调整C_2使得A_1任意，结合$n\le m^A$，上述概率小于$n^{-\frac{A_1}{A}}$。

将圆形区域$\Big\{\beta_C\in\mathbf{R}^p:|\beta_C-\beta^*|_2\le C_1\cdot a_n\Big\}$的外接矩形邻域$\Big\{\beta_C\in\mathbf{R}^p:|\beta_C-\beta^*|_1\le\sqrt{p}C_1\cdot a_n\Big\}$，划分为$\sqrt{p}C_1\cdot a_n\cdot n^r$若干矩形区域$I_1,I_2,...,I_{\tilde{n}}$，对应的中心点$\beta_1,\beta_2,...,\beta_{\tilde{n}}$，每个区域的对角线长度为$O(n^{-r})$。

此时，与之前类似，容易证明:

$$\underset{1\le j\le \tilde{n}}{max}\Big|\frac{1}{m}\sum_{i\in H_1}\Big(\big[\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta^*}-\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta_j}\big]-\big[E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\beta^*}-E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\beta_j}\big]\Big)\Big|_2$$

大于$C_2\cdot a_n\sqrt{\frac{logm}{m}}$的概率小于$n^{-\frac{A_1}{A}+r+1}$。

而对于充分大的n，对于$\forall 1\le j\le\tilde{n},\forall\beta_C\in I_j$：

$$\begin{aligned}
&\Big|\frac{1}{m}\sum_{i\in H_1}\Big(\big[\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta_j}-\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta_C}\big]-\big[E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\beta_j}-E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\beta_C}\big]\Big)\Big|_2\\
&\le\frac{1}{m}\sum_{i\in H_1}\Big(\Big|\big[\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta_j}-\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta_C}\big]\Big|_2+\Big|\big[E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\beta_j}-E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\beta_C}\big]\Big|_2\Big)\\
&\le\frac{1}{m}\sum_{i\in H_1}C_{\mathbf{X}_i}\Big|\beta_C-\beta_j\Big|_2+E\Big[C_{\mathbf{X}}\Big]\cdot\Big|\beta_C-\beta_j\Big|_2\\
&\le\frac{1}{m}\sum_{i\in H_1}C_{\mathbf{X}_i}n^{-r+1}+E\Big[C_{\mathbf{X}}\Big]\cdot n^{-r+1}
\end{aligned}$$

由随机变量$C_{\mathbf{X}_1},C_{\mathbf{X}_2},...,C_{\mathbf{X}_m}$独立同分布，存在有限二阶矩。

由强大数律，可得$\lim_{m\to\infty}\frac{1}{m}\sum_{i\in H_1}C_{\mathbf{X}_i}=E\Big[C_{\mathbf{X}}\Big]$，

对β_C一致地几乎处处成立（即$\lim_{m\to\infty}\frac{1}{m}\sum_{i\in H_1}C_{\mathbf{X}_i}=E\Big[C_{\mathbf{X}}\Big]$不成立对应的零测集不随$\beta_C$改变）。

因此对于充分大的m：

$$\begin{aligned}
&\underset{\beta_C\in I_j}{sup}\Big|\frac{1}{m}\sum_{i\in H_1}\Big(\big[\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta_j}-\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta_C}\big]-\big[E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\beta_j}-E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\beta_C}\big]\Big)\Big|_2\\
&\le 3E\Big[C_{\mathbf{X}}\Big]\cdot n^{-r+1}
\end{aligned}$$

几乎处处成立。

从而：

$$\Big|\frac{1}{m}\sum_{i\in H_1}\Big(\big[\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta^*}-\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\hat{\beta}_0}\big]-\big[E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\beta^*}-E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\hat{\beta}_0}\big]\Big)\Big|_2=O_{a.s.}\Big(a_n\sqrt{\frac{logm}{m}}\Big)$$

对于(19)式中$\frac{1}{n}\sum_{i=1}^{n}\Big(\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\hat{\beta}_0}-E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\hat{\beta}_0}\Big)$

由：

$$\begin{aligned}&\Big|\frac{1}{n}\sum_{i=1}^{n}\Big(\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\hat{\beta}_0}-E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\hat{\beta}_0}\Big)\Big|_2\\&\leq\sup_{\beta_C:|\beta_C-\beta^*|_2\leq C_1\cdot a_n}\Big|\frac{1}{n}\sum_{i=1}^{n}\Big(\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\hat{\beta}_0}-E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\hat{\beta}_0}\Big)\Big|_2\end{aligned}$$

由条件中：$\partial f(\mathbf{X};\beta)$的二阶矩对$\forall\beta\in\mathbf{R}^p$ 一致有界。以及与上面相同的技巧，可以证明：

$$\Big|\frac{1}{n}\sum_{i=1}^{n}\Big(\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\hat{\beta}_0}-E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\hat{\beta}_0}\Big)\Big|_2=O_{a.s.}\Big(\sqrt{\frac{logn}{n}}\Big)$$

从而：

$$\begin{aligned}&\Big|\frac{1}{m}\sum_{i\in H_1}\Big(\big[\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta^*}-\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\hat{\beta}_0}\big]-\big[E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\beta^*}-E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\hat{\beta}_0}\big]\Big)\\&\quad+\frac{1}{n}\sum_{i=1}^{n}\Big(\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\hat{\beta}_0}-E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\hat{\beta}_0}\Big)\Big|_2\\&=O_{a.s.}\Big(a_n\sqrt{\frac{logm}{m}}+\sqrt{\frac{logn}{n}}\Big)\end{aligned}$$

也就是$\exists C_3, \exists N, \forall n > N$，有:

$$\begin{aligned}
&\Big|\frac{1}{m}\sum_{i\in H_1}\Big(\big[\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta^*} - \partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\hat{\beta}_0}\big] - \big[E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\beta^*} - E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\hat{\beta}_0}\big]\Big)\\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\Big(\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\hat{\beta}_0} - E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\hat{\beta}_0}\Big)\Big|_2\\
&\le C_3\bigg(a_n\sqrt{\frac{logm}{m}} + \sqrt{\frac{logn}{n}}\bigg)
\end{aligned}$$

几乎处处成立。

在(19)式中，由Cauchy不等式：

$$\begin{aligned}
&= \Big(\beta-\beta^*\Big)^T\cdot\frac{1}{m}\sum_{i\in H_1}\Big(\big[\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta^*} - \partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\hat{\beta}_0}\big] - \big[E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\beta^*} - E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\hat{\beta}_0}\big]\Big)\\
&\quad + \Big(\beta-\beta^*\Big)^T\cdot\frac{1}{n}\sum_{i=1}^{n}\Big(\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\hat{\beta}_0} - E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\hat{\beta}_0}\Big)\\
&\ge -C\cdot C_3\cdot\bigg(\sqrt{\frac{logn}{n}} + a_n\cdot\sqrt{\frac{logm}{m}}\bigg)^2
\end{aligned}$$

几乎处处成立。

再由(19)式中最后一项：

$$\begin{aligned}
&\frac{1}{2}\Big(\beta-\beta^*\Big)^T\cdot H_m\Big|_{\beta=\tilde{\beta}}\cdot\Big(\beta-\beta^*\Big)\\
&\ge C^2\frac{\lambda_L}{2}\bigg(\sqrt{\frac{logn}{n}} + a_n\cdot\sqrt{\frac{logm}{m}}\bigg)^2
\end{aligned}$$

取$C > \frac{2C_3}{\lambda_L}$，即可得到：$\forall \beta : \left|\beta - \beta^*\right|_2 = C\left(\sqrt{\frac{logn}{n}} + a_n \cdot \sqrt{\frac{logm}{m}}\right), h(\beta) > h(\beta^*)$，几乎处处成立。

得证。

2.3.3 基于median of mean的稳健Communication efficiency及其极限性质

分布式系统基本设定承接2.3.2，若假设N个子服务器中，有αN个子服务器出现问题，其中α满足$\exists 1 > r > 0, \alpha \leq r$以及$\lim_{N\to\infty} \alpha = 0$，计算平均梯度$L_k(\beta) = \frac{1}{m}\sum_{i\in H_k} \partial f(\mathbf{X}_i;\beta)$时出现失误，但对于$\mathbf{R}^p$上的两个不同向量$\beta_1, \beta_2$，平均梯度差$L_k(\beta_1) - L_k(\beta_2)$不发生改变。并且已知第一台子服务器没有出现失误。在这种情况下，我们基于median of mean 的想法来构造boosting 估计量：

$$\hat{\beta}_1 = \underset{\beta\in\mathbf{R}^p}{argmin}\left\{\frac{1}{m}\sum_{i\in H_1} f(\mathbf{X}_i;\beta) - \beta^T\left[\frac{1}{m}\sum_{i\in H_1}\partial f(\mathbf{X}_i;\beta)\big|_{\beta=\hat{\beta}_0} - \underset{1\leq k\leq N}{med}\frac{1}{m}\sum_{i\in H_k}\partial f(\mathbf{X}_i;\beta)\big|_{\beta=\hat{\beta}_0}\right]\right\}$$

这里对一个p维向量取中位数即逐维取中位数。

在与2.3.2相同的假设下，我们有：

$$\left|\hat{\beta}_1 - \beta^*\right|_2 = O_{a.s.}\left(a_n\sqrt{\frac{logm}{m}} + \frac{logN}{\sqrt{n}} + \frac{1}{m} + \frac{\alpha}{\sqrt{m}}\right)$$

证明 不妨假定前$N_1 = (1-\alpha)N$个子服务器不出现问题。

下面的证明中我们引入一个记号：若x为$\mathbf{R}^p$上的p维向量，$x_{(i)}$表示其第i个分量。

由2.3.2的证明可知，往证：

$$\left|\frac{1}{m}\Big(\sum_{i\in H_1}\partial f(\mathbf{X}_i;\beta)\big|_{\beta=\beta^*}-\sum_{i\in H_1}\partial f(\mathbf{X}_i;\beta)\big|_{\beta=\hat{\beta}_0}\Big)+\underset{1\le k\le N}{med}\frac{1}{m}\sum_{i\in H_k}\partial f(\mathbf{X}_i;\beta)\big|_{\beta=\hat{\beta}_0}\right|_2=O_{a.s.}\left(a_n\sqrt{\frac{logm}{m}}+\frac{logN}{\sqrt{n}}+\frac{1}{m}+\frac{\alpha}{\sqrt{m}}\right)$$

再由$|\hat{\beta}_0-\beta^*|=O_{a.s.}(a_n)$可知，$\exists C_1,\exists N,\forall n>N,|\beta^*-\hat{\beta}_0|\le C_1\cdot a_n$，几乎处处成立。从而：

$$\begin{aligned}
&\left|\frac{1}{m}\Big(\sum_{i\in H_1}\partial f(\mathbf{X}_i;\beta)\big|_{\beta=\beta^*}-\sum_{i\in H_1}\partial f(\mathbf{X}_i;\beta)\big|_{\beta=\hat{\beta}_0}\Big)+\underset{1\le k\le N}{med}\frac{1}{m}\sum_{i\in H_k}\partial f(\mathbf{X}_i;\beta)\big|_{\beta=\hat{\beta}_0}\right|_2\\
&\le\left|\frac{1}{m}\sum_{i\in H_1}\Big(\big[\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta^*}-\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\hat{\beta}_0}\big]-\big[E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\beta^*}-E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\hat{\beta}_0}\big]\Big)\right|_2\\
&\quad+\left|\underset{1\le k\le N}{med}\frac{1}{m}\sum_{i\in H_k}\Big(\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\hat{\beta}_0}-E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\hat{\beta}_0}\Big)\right|_2\\
&\le\sup_{\beta_C:|\beta_C-\beta^*|_2\le C_1\cdot a_n}\left|\frac{1}{m}\sum_{i\in H_1}\Big(\big[\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta^*}-\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta_C}\big]-\big[E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\beta^*}-E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\beta_C}\big]\Big)\right|_2\\
&\quad+\sup_{\beta_C:|\beta_C-\beta^*|_2\le C_1\cdot a_n}\left|\underset{1\le k\le N}{med}\frac{1}{m}\sum_{i\in H_k}\Big(\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta_C}-E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\beta_C}\Big)\right|_2
\end{aligned}$$

几乎处处成立。

由2.3.2的证明，往证：

$$\sup_{\beta_C:|\beta_C-\beta^*|_2\le C_1\cdot a_n}\left|\underset{1\le k\le N}{med}\frac{1}{m}\sum_{i\in H_k}\Big(\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta_C}-E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\beta_C}\Big)\right|_2=O_{a.s.}\left(\frac{logN}{\sqrt{n}}+\frac{1}{m}+\frac{\alpha}{\sqrt{m}}\right)$$

首先，对于固定的$\beta_C\in\{\beta_C:|\beta_C-\beta^*|_2\le C_1\cdot a_n\}$，

记：$Y_k(\beta_C)=\left(\frac{1}{\sqrt{m}}\sum_{i\in H_k}\Big(\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta_C}-E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\beta_C}\Big)\right)_{(1)}$，

这里，由假设中$|\partial f(\mathbf{X};\beta)|_2^2$的期望对$\forall\beta\in\mathbf{R}^p$ 一致有界，设一个上界为M，

可知$var\Big(\big(\partial f(\mathbf{X}_i;\beta)\big|_{\beta=\beta_C}\big)_{(1)}\Big) \le E\Big[\big(\partial f(\mathbf{X}_i;\beta)\big)^2_{(1)}\Big]\Big|_{\beta=\beta_C} \le E\Big[\big|\partial f(\mathbf{X}_i;\beta)\big|_2^2\Big]\Big|_{\beta=\beta_C} \le M.$

从而，由Berry-Esseen不等式以及标准正态密度函数的单调性可得：

对于充分大的N以及m：

$$
\begin{aligned}
&P\Big\{\underset{1\le k\le N}{med} Y_k(\beta_C) \ge C\Big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\Big)\Big\}\\
&= P\Big\{\sum_{k=1}^{N}\mathbf{1}\Big\{Y_k(\beta_C)\le C\Big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\Big)\Big\}\le\frac{N}{2}\Big\}\\
&\le P\Big\{\frac{1}{N_1}\sum_{k=1}^{N_1}\mathbf{1}\Big\{Y_k(\beta_C)\le C\Big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\Big)\Big\}\le\frac{1}{2(1-\alpha)}\Big\}\\
&\le P\Big\{\frac{1}{N_1}\sum_{k=1}^{N_1}\Big[P\Big(Y_k(\beta_C)\le C\Big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\Big)\Big)-\mathbf{1}\Big\{Y_k(\beta_C)\le C\Big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\Big)\Big\}\Big]\\
&\quad\ge P\Big(Y_1(\beta_C)\le C\Big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\Big)\Big)-\frac{1}{2(1-\alpha)}\Big\}\\
&\le P\Big\{\frac{1}{N_1}\sum_{k=1}^{N_1}\Big[P\Big(Y_k(\beta_C)\le C\Big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\Big)\Big)-\mathbf{1}\Big\{Y_k(\beta_C)\le C\Big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\Big)\Big\}\Big]\\
&\quad\ge \Phi\Big(\frac{C\Big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\Big)}{\sqrt{var\Big(\big(\partial f(\mathbf{X}_i;\beta)\big|_{\beta=\beta_C}\big)_{(1)}\Big)}}\Big)-\frac{1}{2}-\frac{\alpha}{2(1-\alpha)}-C_2\frac{1}{\sqrt{m}}\Big\}\\
&\le P\Big\{\frac{1}{N_1}\sum_{k=1}^{N_1}\Big[P\Big(Y_k(\beta_C)\le C\Big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\Big)\Big)-\mathbf{1}\Big\{Y_k(\beta_C)\le C\Big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\Big)\Big\}\Big]\\
&\quad\ge \Phi\Big(\frac{C\Big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\Big)}{\sqrt{M}}\Big)-\frac{1}{2}-\frac{\alpha}{2(1-\alpha)}-C_2\frac{1}{\sqrt{m}}\Big\}\\
&\le P\Big\{\frac{1}{N_1}\sum_{k=1}^{N_1}\Big[P\Big(Y_k(\beta_C)\le C\Big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\Big)\Big)-\mathbf{1}\Big\{Y_k(\beta_C)\le C\Big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\Big)\Big\}\Big]\\
&\quad\ge \phi(1)\frac{C\Big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\Big)}{\sqrt{M}}-\frac{\alpha}{2(1-\alpha)}-C_2\frac{1}{\sqrt{m}}\Big\}\\
&\le P\Big\{\frac{1}{N_1}\sum_{k=1}^{N_1}\Big[P\Big(Y_k(\beta_C)\le C\Big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\Big)\Big)-\mathbf{1}\Big\{Y_k(\beta_C)\le C\Big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\Big)\Big\}\Big]\ge C\cdot\frac{\sqrt{1-\alpha}\phi(1)}{\sqrt{M}}\frac{logN_1}{\sqrt{N_1}}\Big\}
\end{aligned}
$$

再由Berstein不等式，可得上述概率：

$$
\begin{aligned}
&P\Big\{\underset{1\le k\le N}{med}\, Y_k(\beta_C) \ge C\Big(\frac{logN}{\sqrt{N}}+\frac{1}{\sqrt{m}}+\alpha\Big)\Big\}\\
&\le N_1^{-A_C} \le (1-r)^{-A}\cdot N^{-A_C} \le (1-r)^{-A}\cdot m^{-A_C(A-1)}
\end{aligned}
$$

其中，可以调整C使得A_C任意大。

再由Borel-Cantelli引理，以及维数p为固定的假设，对于固定的$\beta_C \in \big\{\beta_C : |\beta_C-\beta^*|_2 \le C_1\cdot a_n\big\}$：

$$
\Big|\underset{1\le k\le N}{med}\frac{1}{m}\sum_{i\in H_k}\Big(\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta_C} - E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\beta_C}\Big)\Big|_2 = O_{a.s.}\Big(\frac{logN}{\sqrt{n}}+\frac{1}{m}+\frac{\alpha}{\sqrt{m}}\Big)
$$

将球形区域$\big\{\beta_C : |\beta_C-\beta^*|_2 \le C_1\cdot a_n\big\}$扩张为矩形区域$\big\{\beta_C : |\beta_C-\beta^*|_1 \le C_1\cdot\sqrt{p}\cdot a_n\big\}$，

并将矩形区域划分为$C_1\cdot\sqrt{p}\cdot a_n\cdot\Big(\frac{logN}{\sqrt{n}}+\frac{1}{m}+\frac{\alpha}{\sqrt{m}}\Big)^{-2}$个小的矩形区域$I_1, I_2, ..., I_{\tilde{n}}$，记其中心为$\beta_1, \beta_2, ..., \beta_{\tilde{n}}$。

从而每个小区域的对角线长度为$2\Big(\frac{logN}{\sqrt{n}}+\frac{1}{m}+\frac{\alpha}{\sqrt{m}}\Big)^2$。

显然$\exists R>0, \tilde{n}=o(n^R)$，从而：

$$
\underset{1\le j\le\tilde{n}}{max}\Big|\underset{1\le k\le N}{med}\frac{1}{m}\sum_{i\in H_k}\Big(\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta_j} - E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\beta_j}\Big)\Big|_2 = O_{a.s.}\Big(\frac{logN}{\sqrt{n}}+\frac{1}{m}+\frac{\alpha}{\sqrt{m}}\Big) \tag{20}
$$

对于$\forall\beta_C\in I_1$，

这里由于对于$\mathbf{R}^p$上的两个不同向量β_1,β_2，平均梯度差$L_k(\beta_1)-L_k(\beta_2)$不发生改变。从而有：

$$
\begin{aligned}
&|\sqrt{\frac{1}{m}}\Big[Y_k(\beta_C)-Y_k(\beta_1)\Big]| \\
&=\frac{1}{m}\sum_{i\in H_1}\Big|\Big(\partial f(\mathbf{X}_i;\beta)\big|_{\beta=\beta_C}\Big)_{(1)}-\Big(\partial f(\mathbf{X}_i;\beta)\big|_{\beta=\beta_1}\Big)_{(1)}\Big|+\Big|E\Big[\Big(\partial f(\mathbf{X}_i;\beta)\big|_{\beta=\beta_C}\Big)_{(1)}\Big]-E\Big[\Big(\partial f(\mathbf{X}_i;\beta)\big|_{\beta=\beta_1}\Big)_{(1)}\Big]\Big| \\
&\le\frac{1}{m}\sum_{i\in H_1}\Big|\partial f(\mathbf{X}_i;\beta)\big|_{\beta=\beta_C}-\partial f(\mathbf{X}_i;\beta)\big|_{\beta=\beta_1}\Big|_2+\Big|E\Big[\partial f(\mathbf{X}_i;\beta)\big|_{\beta=\beta_C}\Big]-E\Big[\partial f(\mathbf{X}_i;\beta)\big|_{\beta=\beta_1}\Big]\Big|_2 \\
&\le\Big(\frac{1}{m}\sum_{i\in H_1}C_{\mathbf{X}_i}+E[C_{\mathbf{X}}]\Big)\cdot\Big|\beta_C-\beta_1\Big|_2
\end{aligned}
$$

由强大数律，$\lim_{m\to\infty}\frac{1}{m}\sum_{i\in H_1}C_{\mathbf{X}_i}=E[C_{\mathbf{X}}]$，几乎处处成立。

从而，当m充分大（不依赖于β_C以及小矩形区域）时有：$|\sqrt{\frac{1}{m}}\Big[Y_k(\beta_C)-Y_k(\beta_1)\Big]|\le 3E[C_{\mathbf{X}}]\cdot\Big|\beta_C-\beta_1\Big|_2$，对$\beta_C\in I_1$ 以及不同小矩形区域一致几乎处处成立。

进而容易证明：

$$
\begin{aligned}
&\Big|\Big|\underset{1\le k\le N}{med}\sqrt{\frac{1}{m}}Y_k(\beta_C)\Big|-\Big|\underset{1\le k\le N}{med}\sqrt{\frac{1}{m}}Y_k(\beta_1)\Big|\Big| \\
&\le\Big|\underset{1\le k\le N}{med}\sqrt{\frac{1}{m}}Y_k(\beta_C)-\underset{1\le k\le N}{med}\sqrt{\frac{1}{m}}Y_k(\beta_1)\Big| \\
&\le 3E[C_{\mathbf{X}}]\cdot\Big|\beta_C-\beta_1\Big|_2 \\
&\le 3E[C_{\mathbf{X}}]\cdot\left(\frac{logN}{\sqrt{n}}+\frac{1}{m}+\frac{\alpha}{\sqrt{m}}\right)^2
\end{aligned}
$$

对$\beta_C\in I_1$ 以及不同小矩形区域一致几乎处处成立。

从而可得：

$$
\underset{1\le j\le\tilde{n}}{max}\ \underset{\beta_C\in I_j}{sup}\left|\underset{1\le k\le N}{med}\frac{1}{m}\sum_{i\in H_k}\Big(\partial f(\mathbf{X}_i;\beta)\Big|_{\beta=\beta_C}-E\big[\partial f(\mathbf{X}_i;\beta)\big]\Big|_{\beta=\beta_C}\Big)\right|_2=O_{a.s.}\left(\frac{logN}{\sqrt{n}}+\frac{1}{m}+\frac{\alpha}{\sqrt{m}}\right)\qquad(21)
$$

结合(20)与(21)式得证。

3 Lasso

3.1 Lasso简介

之前研究的随机向量$\mathbf{X}$中的维数p皆被假设为固定的，即不随样本量n而改变。在现代统计学研究中，常常会在总体方面：$Y = \mathbf{X}^T\beta^* + e$，假设$p$随着样本量的增大而增大，并且总体参数$\beta^*$除其中$s$维之外皆为0，但这$s$ 维非零分量在p维向量中的具体位置未知，即稀疏性假设。

在这种设定下，原先的参数估计方法$\hat{\beta} = \underset{\beta\in\mathbf{R}^p}{argmin}\frac{1}{2n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i^T\beta)^2$，不再是一个好的估计。一方面，其实现这个参数估计时涉及到广义逆，导致参数估计不唯一；另一方面，这种参数估计方法并没有体现稀疏性这一假设。

因此，基于$l_1 - norm$惩罚的Lasso估计应运而生：$\hat{\beta} = \underset{\beta\in\mathbf{R}^p}{argmin}\frac{1}{2n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i^T\beta)^2 + \lambda|\beta|_1$ ，其中λ被称为$tuning\ parameter$，在统计量实现（也就是解上述优化）之前给定。

这里向量的l_1范数为向量各元素绝对值求和。

下一小节，我们考察这种参数估计收敛到总体参数的速度。

3.2 Lasso估计的极限性质

3.2.1 一个关于优化问题的引理

我们先证明一个关于优化问题的引理：

记向量或是矩阵的∞范数为其中元素绝对值的最大值。

考虑二次优化问题$\hat{\beta} = \underset{\beta\in\mathbf{R}^p}{argmin}\frac{1}{2}\beta^T A\beta - b^T\beta + \lambda|\beta|_1$，其中$p$阶方阵$A$半正定。

若$\tilde{\beta}\in\mathbf{R}^p$满足：

- $|A\tilde{\beta}-b|_\infty \le \frac{1}{2}\lambda$
- 令$C_1=4, \exists C_2$满足：$\underset{|\delta|_1 \le C_1\sqrt{s}|\delta|_2}{min} \frac{\delta^T A\delta}{|\delta|_2^2} \ge C_2$

其中，s为$\tilde{\beta}$非零元的个数,并将其下标集也记为s。则：$\exists C, |\hat{\beta}-\tilde{\beta}|_2 \le C\sqrt{s}\lambda$.

证明 **先**证明：$|\hat{\beta}-\tilde{\beta}|_1 \le 4\sqrt{s}|\hat{\beta}-\tilde{\beta}|_2$.

由：

$$\frac{1}{2}\hat{\beta}^T A\hat{\beta}-\hat{\beta}^T b+\lambda|\hat{\beta}|_1 \le \frac{1}{2}\tilde{\beta}^T A\tilde{\beta}-\tilde{\beta}^T b+\lambda|\tilde{\beta}|_1$$

可知：

$$\begin{aligned}
&\Big(\frac{1}{2}\hat{\beta}^T A\hat{\beta}-\hat{\beta}^T b\Big)-\Big(\frac{1}{2}\tilde{\beta}^T A\tilde{\beta}-\tilde{\beta}^T b\Big)\\
&\le \lambda\cdot\Big(|\tilde{\beta}|_1-|\hat{\beta}|_1\Big)\\
&= \lambda\cdot\Big(|\tilde{\beta}_s|_1-|\hat{\beta}_s|_1-|\hat{\beta}_{s^c}|_1\Big)\\
&\le \lambda\cdot\Big(|\tilde{\beta}_s-\hat{\beta}_s|_1-|\tilde{\beta}_{s^c}-\hat{\beta}_{s^c}|_1\Big)
\end{aligned}$$

另一方面：

$$\begin{aligned}
&\Big(\frac{1}{2}\hat{\beta}^T A\hat{\beta}-\hat{\beta}^T b\Big)-\Big(\frac{1}{2}\tilde{\beta}^T A\tilde{\beta}-\tilde{\beta}^T b\Big)\\
&\ge (A\tilde{\beta}-b)^T(\hat{\beta}-\tilde{\beta})\\
&\ge -|A\tilde{\beta}-b|_\infty\cdot|\hat{\beta}-\tilde{\beta}|_1\\
&\ge -\frac{\lambda}{2}|\hat{\beta}-\tilde{\beta}|_1\\
&= -\frac{\lambda}{2}\Big(|\hat{\beta}_s-\tilde{\beta}_s|_1+|\hat{\beta}_{s^c}-\tilde{\beta}_{s^c}|_1\Big)
\end{aligned}$$

从而：$|\hat{\beta}_{s^C} - \tilde{\beta}_{s^C}|_1 \leq 3|\hat{\beta}_s - \tilde{\beta}_s|_1$.

再由Cauchy不等式：$|\hat{\beta} - \tilde{\beta}|_1 \leq 4|\hat{\beta}_s - \tilde{\beta}_s|_1 \leq 4\sqrt{s}|\hat{\beta}_s - \tilde{\beta}_s|_2 \leq 4\sqrt{s}|\hat{\beta} - \tilde{\beta}|_2$.

这时，取$C_1 = 4$，从而$\exists C_2$，使得$(\hat{\beta} - \tilde{\beta})^T \cdot A \cdot (\hat{\beta} - \tilde{\beta}) \geq C_2 \cdot |\hat{\beta} - \tilde{\beta}|_2^2$.

而由$|A\hat{\beta} - b|_\infty \leq \lambda$以及$|A\tilde{\beta} - b|_\infty \leq \frac{\lambda}{2}$可得：$|A(\hat{\beta} - \tilde{\beta})|_\infty \leq \frac{3\lambda}{2}$.

从而$(\hat{\beta} - \tilde{\beta})^T \cdot A \cdot (\hat{\beta} - \tilde{\beta}) \leq \frac{3\lambda}{2}|\hat{\beta} - \tilde{\beta}|_1 \leq 6\lambda\sqrt{s}|\hat{\beta} - \tilde{\beta}|_2$.

综上即得：$|\hat{\beta} - \tilde{\beta}|_2 \leq \frac{6\sqrt{s}}{C_2}\lambda$.

3.2.2 Lasso估计收敛到总体参数的速度

在总体方面$Y = \mathbf{X}^T\beta^* + e$，设误差$e$的标准差为$\sigma$，

若假定标准化后的$\mathbf{X}$的总体协方差矩阵Σ的最小特征值大于0，$\exists M, |\mathbf{X}|_2 \leq M, |e| \leq M$，几乎处处成立（事实上证明中利用指数不等式替代Berstein不等式可以放宽这里的假设），

维数p与样本量n满足：$\lim_{(n,p)\to\infty} \frac{logp}{n} = 0$，并且$\mathbf{X}$与误差$e$独立。

则$\lambda = 4\sigma\sqrt{\frac{logp}{n}}$，在这种$\lambda$选取下的Lasso估计满足：$\exists C > 0, \lim_{(n,p)\to\infty} P\left\{\left|\hat{\beta} - \beta^*\right|_2 \leq C\sqrt{\frac{slogp}{n}}\right\} = 1$.

证明 不妨假定样本$\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$已经中心化，Lasso估计即：

$$\underset{\beta\in\mathbf{R}^p}{argmin}\frac{1}{2}\beta^T\Big(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^T\Big)\beta - \Big(\frac{1}{n}\sum_{i=1}^n Y_i\mathbf{X}_i\Big)^T\beta + \lambda|\beta|_1$$

即一个3.2.1研究的二次优化问题，这里$A = \hat{\Sigma}$，即基于样本$\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$的样本协方差矩阵。

这里由总体方面的假设：$Y = \mathbf{X}^T\beta^* + e,$

一方面有：

$$\Big|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T\beta^* - \frac{1}{n}\sum_{i=1}^{n}Y_i\mathbf{X}_i\Big|_\infty = \Big|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i e_i\Big|_\infty = \underset{1\le j\le p}{max}\Big|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i)_{(j)}e_i\Big| = O_p\Big(\sqrt{\frac{logp}{n}}\Big)$$

事实上，由关于总体方面的一些假设并由Berstein不等式可知：当n, p充分大时

$$\begin{aligned}
&P\Big\{\underset{1\le j\le p}{max}\Big|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i)_{(j)}e_i\Big| > 2\sigma\sqrt{\frac{logp}{n}}\Big\}\\
&\le p\cdot 2exp\Big\{-\frac{2n^2\sigma^2\frac{logp}{n}}{n\sigma^2 + \frac{2\sigma}{3}M^2\sqrt{\frac{logp}{n}}}\Big\}\\
&\le 2p^{-\frac{1}{3}}
\end{aligned}$$

即：

$$\forall\varepsilon, \exists L_1, \forall n > L_1, \forall p > L_1, P\Big\{\Big|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i)e_i\Big|_\infty > 2\sigma\sqrt{\frac{logp}{n}}\Big\} < \varepsilon \tag{22}$$

另一方面，对于$C_1 = 4, \forall\delta\in\mathbf{R}^p : |\delta|_1 \le C_1\cdot\sqrt{s}|\delta|_2$有：

$\frac{\delta^T\hat{\Sigma}\delta}{|\delta|_2^2} = \frac{\delta^T(\hat{\Sigma}-\Sigma)\delta}{|\delta|_2^2} + \frac{\delta^T\Sigma\delta}{|\delta|_2^2},$

其中，$\frac{\delta^T\Sigma\delta}{|\delta|_2^2} \ge \lambda_{min}(\Sigma).$

以及，$\frac{\delta^T(\hat{\Sigma}-\Sigma)\delta}{|\delta|_2^2} \ge -\frac{|\delta^T(\hat{\Sigma}-\Sigma)\delta|}{|\delta|_2^2} \ge -\frac{|\delta|_1|(\hat{\Sigma}-\Sigma)\delta|_\infty}{|\delta|_2^2} \ge -\frac{|\delta|_1^2|\hat{\Sigma}-\Sigma|_\infty}{|\delta|_2^2} \ge -C_1^2\cdot s|\hat{\Sigma}-\Sigma|_\infty.$

而

$$|\hat{\Sigma}-\Sigma|_\infty = \underset{1\le j\le k\le p}{max}\Big|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i)_{(j)}(\mathbf{X}_i)_{(k)} - E\Big[\mathbf{X}_{(j)}\mathbf{X}_{(k)}\Big]\Big| = O_p\Big(\sqrt{\frac{logp}{n}}\Big)$$

事实上，由关于总体方面的一些假设并由Berstein不等式可知：当n,p充分大时，

$$\begin{aligned}
&P\Big\{\underset{1\le j\le k\le p}{max}\Big|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i)_{(j)}(\mathbf{X}_i)_{(k)} - E\Big[\mathbf{X}_{(j)}\mathbf{X}_{(k)}\Big]\Big| > 2M^2\sqrt{\frac{logp}{n}}\Big\}\\
&\le p^2\cdot 2exp\Big\{-\frac{8n^2M^4\frac{logp}{n}}{nM^4+\frac{4n}{3}M^4\sqrt{\frac{logp}{n}}}\Big\}\\
&\le 2p^{-2}
\end{aligned}$$

即：

$$\forall\varepsilon,\exists L_2,\forall n>L_2,\forall p>L_2, P\Big\{\underset{1\le j\le k\le p}{max}\Big|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i)_{(j)}(\mathbf{X}_i)_{(k)} - E\Big[\mathbf{X}_{(j)}\mathbf{X}_{(k)}\Big]\Big| > 2M^2\sqrt{\frac{logp}{n}}\Big\} < \varepsilon \tag{23}$$

由(22)(23)以及$\lim_{(n,p)\to\infty}\frac{logp}{n}=0$可知，

$$\begin{aligned}
&\forall\varepsilon,\exists L,\forall n>L,\forall p>L,\\
&P\Big\{\underset{1\le j\le p}{max}\Big|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i)_{(j)}e_i\Big| > 2\sigma\sqrt{\frac{logp}{n}}\Big\} < \varepsilon\\
&P\Big\{\underset{1\le j\le k\le p}{max}\Big|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i)_{(j)}(\mathbf{X}_i)_{(k)} - E\Big[\mathbf{X}_{(j)}\mathbf{X}_{(k)}\Big]\Big| > 2M^2\sqrt{\frac{logp}{n}}\Big\} < \varepsilon\\
&C_1^2\cdot s\cdot M^2\sqrt{\frac{logp}{n}} \le \frac{\lambda_{min}(\Sigma)}{2}
\end{aligned}$$

即$P\Big\{\Big|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T\beta^* - \frac{1}{n}\sum_{i=1}^{n}Y_i\mathbf{X}_i\Big|_\infty \le 2\sigma\sqrt{\frac{logp}{n}} \quad AND \quad \underset{|\delta|_1\le C_1\sqrt{s}|\delta|_2}{min}\frac{\delta^T\hat{\Sigma}\delta}{|\delta|_2^2} \ge \frac{\lambda_{min}(\Sigma)}{2}\Big\} \ge 1-2\varepsilon$

由3.2.1的引理可得：

$P\Big\{\Big|\hat{\beta}-\beta^*\Big|_2 \le \frac{48\sigma}{\lambda_{min}(\Sigma)}\sqrt{\frac{slogp}{n}}\Big\} \ge 1-2\varepsilon.$

即:

$$\exists C, \lim_{(n,p)\to\infty} P\Big\{\Big|\hat{\beta}-\beta^*\Big|_2 \le C\sqrt{\frac{slogp}{n}}\Big\} = 1$$

3.3 Lasso作为变量选择方法的selection consistency

3.3.1 一个关于Lasso估计的性质

在证明Lasso作为变量选择方法具有selection consistency之前，我们先证明一个关于Lasso 估计的一个性质。

事实上，Lasso估计：$\hat{\beta} = \underset{\beta\in\mathbf{R}^p}{argmin}\frac{1}{2n}\sum_{i=1}^n(Y_i-\mathbf{X}_i^T\beta)^2+\lambda|\beta|_1$ ，从严格意义上，并不是一个估计。

这是因为优化问题$\underset{\beta\in\mathbf{R}^p}{argmin}\frac{1}{2n}\sum_{i=1}^n(Y_i-\mathbf{X}_i^T\beta)^2+\lambda|\beta|_1$不一定只有一个解，但关于这个优化所有的解有一个共同点：

若β_1,β_2为上述优化问题的两个不同解，则$\forall i=1,2,...,n$有$\mathbf{X}_i^T\beta_1=\mathbf{X}_i^T\beta_2$.

证明 若存在$j\in\{1,2,...,n\}$使得$\mathbf{X}_j^T\beta_1\neq\mathbf{X}_j^T\beta_2$.

对于$\beta_3=\frac{1}{2}(\beta_1+\beta_2)$,

则由$f(x)=(Y_j-x)^2$的严格凸性以及$g(x)=|x|$的凸性知：

$$\frac{1}{2n}\sum_{i=1}^n(Y_i-\mathbf{X}_i^T\beta_3)^2+\lambda|\beta_3|_1 < \frac{1}{2}\Big(\frac{1}{2n}\sum_{i=1}^n(Y_i-\mathbf{X}_i^T\beta_1)^2+\lambda|\beta_1|_1\Big)+\frac{1}{2}\Big(\frac{1}{2n}\sum_{i=1}^n(Y_i-\mathbf{X}_i^T\beta_2)^2+\lambda|\beta_2|_1\Big)$$

这与β_1, β_2为优化问题的解矛盾。

3.3.2 selection consistency

对于总体方面$Y = \mathbf{X}^T\beta^* + e$，记误差$e$的标准差为$\sigma$，$\mathbf{X}$的协方差矩阵为$\Sigma$，且$p$维向量$\beta^*$有$s$维非零分量，并将其下标集也记为$s$，并且这$s$维的样本协方差$\hat{\Sigma}_{s\times s}$ 可逆。

记矩阵A的l_1范数A_{l_1}为$\left||A|\cdot\mathbf{1}\right|_\infty$，这里$|A|$表示矩阵$A$逐个元取绝对值后的矩阵。

假定$\exists M, |\mathbf{X}|_2 \le M, |e| \le M$，几乎处处成立（事实上这里与3.2 同理可以放宽这里的假设），

若：

- $\left|\Sigma^{-1}_{s\times s}\right|_{l_1} \le C_1.$
- $\min\limits_{i\in s}|\beta_i^*| \ge 6C_2\sigma\sqrt{\frac{logp}{n}}$，其中$C_2 > C_1.$
- $\left|\Sigma_{s^c\times s}\cdot\Sigma^{-1}_{s\times s}\right|_{l_1} < \frac{1}{3}.$

则取$\lambda = 4\sigma\sqrt{\frac{logp}{n}}$，在这种$\lambda$选取下的任意Lasso 估计的非零下标集$\hat{s}$满足:$\lim_{(n,p)\to\infty} P\{\hat{s} = s\} = 1.$

证明 往证$\lim_{(n,p)\to\infty} P\{s \subset \hat{s}\} = 1$以及$\lim_{(n,p)\to\infty} P\{\hat{s} \subset s\} = 1$，

先证明$\lim_{(n,p)\to\infty} P\{\hat{s} \subset s\} = 1$，

考虑下面一个优化问题：

$$\beta^o = \underset{\beta_{s^c}=0}{argmin}\frac{1}{2n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i\beta)^2 + \lambda|\beta|_1$$

由凸优化理论：

$$-\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i(Y_i-\mathbf{X}_i^T\beta^o)\right)_s+\lambda\cdot\tilde{z}_s=0$$

从而，

$$\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T(\beta^o-\beta^*)\right)_s=\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_ie_i\right)_s-\lambda\tilde{z}_s$$

即：

$$\hat{\Sigma}_{s\times s}\cdot(\beta^o-\beta^*)_s=\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_ie_i\right)_s-\lambda\tilde{z}_s$$

从而，

$$\begin{aligned}
&\frac{1}{\lambda}\cdot\left|\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T(\beta^o-\beta^*)\right)_{s^c}\right|_\infty\\
&=\frac{1}{\lambda}\cdot\left|\left(\hat{\Sigma}(\beta^o-\beta^*)\right)_{s^c}\right|_\infty\\
&=\frac{1}{\lambda}\cdot\left|\hat{\Sigma}_{s^c\times s}(\beta^o-\beta^*)_s\right|_\infty\\
&=\frac{1}{\lambda}\cdot\left|\hat{\Sigma}_{s^c\times s}\cdot\hat{\Sigma}_{s\times s}^{-1}\cdot\left((\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_ie_i)_s-\lambda\tilde{z}_s\right)\right|_\infty\\
&\le\left|\hat{\Sigma}_{s^c\times s}\cdot\hat{\Sigma}_{s\times s}^{-1}\right|_{l_1}\cdot\left(\frac{\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_ie_i\right|_\infty}{\lambda}+1\right)
\end{aligned}$$

所以可得，

$$
\begin{aligned}
&\left|\left(\frac{1}{n\lambda}\sum_{i=1}^{n}\mathbf{X}_i(Y_i-\mathbf{X}_i^T\beta^o)\right)_{s^c}\right|_\infty \\
&=\left|\left(\frac{1}{n\lambda}\sum_{i=1}^{n}\mathbf{X}_i e_i-\frac{1}{n\lambda}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T(\beta^o-\beta^*)\right)_{s^c}\right|_\infty \\
&\le\left|\frac{1}{n\lambda}\sum_{i=1}^{n}\mathbf{X}_i e_i\right|_\infty+\left|\left(\frac{1}{n\lambda}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T(\beta^o-\beta^*)\right)_{s^c}\right|_\infty \\
&\le\left|\frac{1}{n\lambda}\sum_{i=1}^{n}\mathbf{X}_i e_i\right|_\infty+\left|\hat{\Sigma}_{s^c\times s}\cdot\hat{\Sigma}_{s\times s}^{-1}\right|_{l_1}\cdot\left(\left|\frac{1}{n\lambda}\sum_{i=1}^{n}\mathbf{X}_i e_i\right|_\infty+1\right)
\end{aligned}
$$

对于任意$\varepsilon>0$，

结合随机矩阵理论以及假设$\left|\Sigma_{s^c\times s}\cdot\Sigma_{s\times s}^{-1}\right|_{l_1}<\frac{1}{3}$，当样本量$n$充分大时，

$$
P\left\{\left|\hat{\Sigma}_{s^c\times s}\cdot\hat{\Sigma}_{s\times s}^{-1}\right|_{l_1}<\frac{1}{3}\right\}>1-\varepsilon
$$

再结合(22)式，对于充分大的L，当$n>L$并且$p>L$时，

$$
P\left\{\left|\frac{1}{n\lambda}\sum_{i=1}^{n}(\mathbf{X}_i)e_i\right|_\infty>\frac{1}{2}\right\}<\varepsilon
$$

因此：

$$
P\left\{\left|\left(\frac{1}{n\lambda}\sum_{i=1}^{n}\mathbf{X}_i(Y_i-\mathbf{X}_i^T\beta^o)\right)_{s^c}\right|_\infty<1\right\}>1-2\varepsilon
$$

而当$\left|\left(\frac{1}{n\lambda}\sum_{i=1}^{n}\mathbf{X}_i(Y_i-\mathbf{X}_i^T\beta^o)\right)_{s^c}\right|_\infty<1$成立时，存在$z^o$满足：$z_{s^c}^o\in(-1,1)$以及$z_s^o=\tilde{z}_s$，使得：

$$-\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i(Y_i-\mathbf{X}_i^T\beta^o)+\lambda z^o=0$$

根据凸优化理论，此时β^o即为Lasso的一个解。而对于任意Lasso的解$\hat{\beta}$：

$$-\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i(Y_i-\mathbf{X}_i^T\hat{\beta})+\lambda\cdot z=0$$

根据3.3.1的结论，$z=z^o$。所以$z_{s^c}\in(-1,1)$，即$s^c\subset\hat{s}^c$。

由此：

$$\forall\varepsilon,\exists L,\forall n>L,\forall p>L,P\left\{\hat{s}\subset s\right\}>1-2\varepsilon$$

这样就证明了：

$$\lim_{(n,p)\to\infty}P\{\hat{s}\subset s\}=1$$

再证$\lim_{(n,p)\to\infty}P\{s\subset\hat{s}\}=1$：

若假定$\hat{s}\subset s$，

对于满足上述设定的任意Lasso估计$\hat{\beta}$，由凸优化理论：

$$-\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i(Y_i-\mathbf{X}_i^T\hat{\beta})+\lambda\cdot z=0$$

这里，$z \in \frac{\partial|\beta|}{\partial\beta}\Big|_{\beta=\hat{\beta}}$.

从而，$\hat{\beta}_s - \beta_s^* = \hat{\Sigma}_{s\times s}^{-1} \cdot \Big[(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i e_i)_s - \lambda z_s\Big]$.

进而，$\Big|\hat{\beta}_s - \beta_s^*\Big|_\infty \le \Big|\hat{\Sigma}_{s\times s}^{-1}\Big|_{l_1} \cdot \Big[\lambda + \Big|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i)e_i\Big|_\infty\Big]$.

对于上述$\varepsilon > 0$，

根据随机矩阵理论并结合条件$\Big|\Sigma_{s\times s}^{-1}\Big|_{l_1} \le C_1$，当样本量$n$ 充分大时，

$$P\left\{\Big|\hat{\Sigma}_{s\times s}^{-1}\Big|_{l_1} \ge \frac{C_1 + C_2}{2}\right\} < \varepsilon$$

再结合(22)式，对于充分大的L，当$n > L$并且$p > L$时，

$$P\left\{\lambda + \Big|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i)e_i\Big|_\infty > 6\sigma\sqrt{\frac{logp}{n}}\right\} < \varepsilon$$

以及第一部分所证，

$$P\Big\{\hat{s} \subset s\Big\} > 1 - 2\varepsilon$$

从而：

$$P\left\{\Big|\hat{\beta}_s - \beta_s^*\Big|_\infty \le \frac{C_1 + C_2}{2} \cdot 6\sigma\sqrt{\frac{logp}{n}}\right\} > 1 - 4\varepsilon$$

再结合条件$\underset{i\in s}{min}|\beta_i^*| \geq 6C_2\sigma\sqrt{\frac{logp}{n}}$，并且$\underset{i\in s}{min}|\hat{\beta}_i| \geq \underset{i\in s}{min}|\beta_i^*| - \left|\hat{\beta}_s - \beta_s^*\right|_\infty$，可知：

$$\forall\varepsilon > 0, \exists L > 0, \forall n > L, \forall p > L, P\left\{\underset{i\in s}{min}|\hat{\beta}_i| > 0\right\} > 1 - 4\varepsilon$$

这样就证明了：

$$\lim_{(n,p)\to\infty} P\{s \subset \hat{s}\} = 1$$

证毕。

3.4 分布式系统下基于communication efficiency的Lasso估计

3.4.1 背景简介

在本节，假若我们已有关于总体方面$Y = \mathbf{X}^T\beta^* + e$中参数$\beta^*$ 的一个初步估计$\hat{\beta}_0$，

并假设：

$$\lim_{(n,p)\to\infty} P\left\{\frac{|\hat{\beta}_0 - \beta^*|_1}{4\sqrt{s}} \leq |\hat{\beta}_0 - \beta^*|_2 \leq a_n\right\} = 1$$

其中s为总体参数β^*中非零元的个数。事实上，由3.2.1的证明不难得知这里的假设是合理的。

其他关于总体方面的假设与之前相同：误差e的标准差为σ，维数p与样本量n 满足：$\lim_{(n,p)\to\infty}\frac{logp}{n} = 0$，并且$\mathbf{X}$与误差$e$ 独立。

$\mathbf{X}$已经过标准化，总体协方差矩阵Σ的最小特征值大于0，$\exists M, |\mathbf{X}|_2 \leq M, |e| \leq M$，几乎处处成立（与之前同理可以放宽这里的假设）。

关于分布式系统，n个样本$(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), ..., (\mathbf{X}_n, Y_n)$ 分别来自N个子服务器，记各自样本相对所有n个样本的下标集为$H_1, H_2, ..., H_N$，每个子服务器中有m个样本，满足$n = N \cdot m$并且$\exists A > 1, n \leq m^A$。另外存在一个总服务器来汇总处理来自N个子服务器的各类信息，并作出关于总体统计推断。

下面，我们考虑在第一台子服务器上，基于初步估计$\hat{\beta}_0$的boosting估计量：

$$\hat{\beta}_1 = \underset{\beta \in \mathbf{R}^p}{argmin} \frac{1}{2m} \sum_{i \in H_1} (Y_i - \mathbf{X}_i^T \beta)^2 - \beta^T \cdot \Big[\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i (Y_i - \mathbf{X}_i^T \hat{\beta}_0) - \frac{1}{m} \sum_{i \in H_1} \mathbf{X}_i (Y_i - \mathbf{X}_i^T \hat{\beta}_0) \Big] + \lambda |\beta|_1$$

在下面的两小节里，我们考虑boosting估计量$\hat{\beta}_1$收敛到总体参数的速度以及selection consistency.

3.4.2 基于communication efficiency的Lasso估计收敛到总体参数的速度

若取$\lambda = 4\sigma \sqrt{\frac{logp}{n}} + 32M^2 \sqrt{\frac{slogp}{m}} \cdot a_n$，则在这种$\lambda$选取下，存在常数$C > 0$，使得：

$$\lim_{(n,p) \to \infty} P \Bigg\{ \frac{|\hat{\beta}_1 - \beta^*|_1}{4\sqrt{s}} \leq |\hat{\beta}_1 - \beta^*|_2 \leq C \cdot \left(\sigma \sqrt{\frac{slogp}{n}} + 8a_n \cdot M^2 \cdot \sqrt{\frac{s^2 logp}{m}} \right) \Bigg\} = 1$$

证明　由：

$$\hat{\beta}_1 = \underset{\beta \in \mathbf{R}^p}{argmin} \frac{1}{2m} \sum_{i \in H_1} (Y_i - \mathbf{X}_i^T \beta)^2 - \beta^T \cdot \Big[\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i (Y_i - \mathbf{X}_i^T \hat{\beta}_0) - \frac{1}{m} \sum_{i \in H_1} \mathbf{X}_i (Y_i - \mathbf{X}_i^T \hat{\beta}_0) \Big] + \lambda |\beta|_1$$

可得：

$$\hat{\beta}_1 = \underset{\beta \in \mathbf{R}^p}{argmin} \frac{1}{2} \beta^T \cdot \Big(\frac{1}{m} \sum_{i \in H_1} \mathbf{X}_i \mathbf{X}_i^T \Big) \cdot \beta - \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i Y_i - \Big(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^T - \frac{1}{m} \sum_{i \in H_1} \mathbf{X}_i \mathbf{X}_i^T \Big) \hat{\beta}_0 \right)^T \beta$$

即仍然是一个二次优化问题。

一方面：

$$\begin{aligned}
&\left|\Big(\frac{1}{m}\sum_{i\in H_1}\mathbf{X}_i\mathbf{X}_i^T\Big)\cdot\beta^*-\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_iY_i+\Big(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T-\frac{1}{m}\sum_{i\in H_1}\mathbf{X}_i\mathbf{X}_i^T\Big)\hat{\beta}_0\right|_\infty\\
&=\left|\Big(\frac{1}{m}\sum_{i\in H_1}\mathbf{X}_i\mathbf{X}_i^T-\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T\Big)\cdot(\beta^*-\hat{\beta}_0)-\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_ie_i\right|_\infty\\
&\le\left|\frac{1}{m}\sum_{i\in H_1}\mathbf{X}_i\mathbf{X}_i^T-\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T\right|_\infty\cdot\left|\beta^*-\hat{\beta}_0\right|_1+\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_ie_i\right|_\infty\\
&\le\left(\left|\frac{1}{m}\sum_{i\in H_1}\mathbf{X}_i\mathbf{X}_i^T-\Sigma\right|_\infty+\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T-\Sigma\right|_\infty\right)\cdot\left|\beta^*-\hat{\beta}_0\right|_1+\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_ie_i\right|_\infty
\end{aligned}$$

对于任意$\varepsilon>0$，

由3.2.2中(22)式以及(23)式可得，存在充分大的L_1，$\forall n>L_1,\forall p>L_1$，

$$P\left\{\left|\Big(\frac{1}{m}\sum_{i\in H_1}\mathbf{X}_i\mathbf{X}_i^T\Big)\cdot\beta^*-\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_iY_i+\Big(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T-\frac{1}{m}\sum_{i\in H_1}\mathbf{X}_i\mathbf{X}_i^T\Big)\hat{\beta}_0\right|_\infty<\frac{1}{2}\lambda\right\}>1-3\varepsilon$$

另一方面，3.2.2已证明了：存在充分大的L_2，$\forall n>L_2,\forall p>L_2$，

$$P\Big\{\min_{|\delta|_1\le C_1\sqrt{s}|\delta|_2}\frac{\delta^T\hat{\Sigma}\delta}{|\delta|_2^2}\ge\frac{\lambda_{min}(\Sigma)}{2}\Big\}\ge1-\varepsilon$$

由引理3.2.1可得：

$$P\left\{\frac{|\hat{\beta}_1-\beta^*|_1}{4\sqrt{s}}\le|\hat{\beta}_1-\beta^*|_2\le\frac{48}{\lambda_{min}(\Sigma)}\cdot\left(\sigma\sqrt{\frac{slogp}{n}}+8a_n\cdot M^2\cdot\sqrt{\frac{s^2logp}{m}}\right)\right\}>1-4\varepsilon$$

得证。

评注 由假设中$\exists A > 1, n \leq m^A$可知：在每次迭代中，适当选取$\lambda$，即可在$2A$左右次迭代后达到$\sqrt{\frac{slogp}{n}}$的收敛速度。

3.4.3 基于communication efficiency的Lasso估计的selection consistency

若：

- $\left|\Sigma_{s\times s}^{-1}\right|_{l_1} \leq C_1.$
- $\min\limits_{i\in s}|\beta_i^*| \geq C_2 \cdot (10\sigma\sqrt{\frac{logp}{n}} + 100M^2 \cdot \sqrt{\frac{slogp}{m}}a_n)$，其中$C_2 > C_1.$
- $\left|\Sigma_{s^c\times s} \cdot \Sigma_{s\times s}^{-1}\right|_{l_1} < \frac{1}{3}.$

取$\lambda = 8\sigma\sqrt{\frac{logp}{n}} + 64M^2\sqrt{\frac{slogp}{m}} \cdot a_n$，则在这种$\lambda$选取下的任意boosting估计有：

$$\lim_{(n,p)\to\infty} P\left\{\hat{s} = s\right\} = 1$$

证明 这里证明的思路与Lasso估计的selection consistency思路基本一致。

记$\hat{\Sigma}^m = \frac{1}{m}\sum_{i\in H_1}\mathbf{X}_i\mathbf{X}_i^T$ 以及$\hat{\Sigma}^n = \frac{1}{n}\sum_{i=1}^n\mathbf{X}_i\mathbf{X}_i^T.$

先证明：$\lim_{(n,p)\to\infty} P\left\{\hat{s} \subset s\right\} = 1.$

先考虑下面这个优化问题：

$$\beta^o = \underset{\beta_{s^c}=0}{argmin} \frac{1}{2m} \sum_{i \in H_1} (Y_i - \mathbf{X}_i^T \beta)^2 - \beta^T \cdot \Big[\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i (Y_i - \mathbf{X}_i^T \hat{\beta}_0) - \frac{1}{m} \sum_{i \in H_1} \mathbf{X}_i (Y_i - \mathbf{X}_i^T \hat{\beta}_0) \Big] + \lambda |\beta|_1$$

由凸优化理论可知：

$$\Big(- \frac{1}{m} \sum_{i \in H_1} \mathbf{X}_i (Y_i - \mathbf{X}_i^T \beta^o) \Big)_s - \Big(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i (Y_i - \mathbf{X}_i^T \hat{\beta}_0) - \frac{1}{m} \sum_{i \in H_1} \mathbf{X}_i (Y_i - \mathbf{X}_i^T \hat{\beta}_0) \Big)_s + \lambda \cdot \tilde{z}_s = 0$$

稍作化简即得：

$$(\beta^o - \beta^*)_s = (\hat{\Sigma}^m_{s \times s})^{-1} \cdot \Big(\Big(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i e_i \Big)_s + \Big((\hat{\Sigma}^n - \hat{\Sigma}^m) \cdot (\beta^* - \hat{\beta}_0) \Big)_s - \lambda \cdot \tilde{z}_s \Big)$$

从而，

$$\begin{aligned}
&\frac{1}{\lambda} \Big| \Big(\hat{\Sigma}^m (\beta^o - \beta^*) \Big)_{s^c} \Big|_\infty \\
&= \frac{1}{\lambda} \Big| \hat{\Sigma}^m_{s^c \times s} (\beta^o - \beta^*)_s \Big|_\infty \\
&= \frac{1}{\lambda} \Big| \hat{\Sigma}^m_{s^c \times s} (\hat{\Sigma}^m_{s \times s})^{-1} \cdot \Big(\Big(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i e_i \Big)_s + \Big((\hat{\Sigma}^n - \hat{\Sigma}^m) \cdot (\beta^* - \hat{\beta}_0) \Big)_s - \lambda \cdot \tilde{z}_s \Big) \Big|_\infty \\
&\leq \Big| \hat{\Sigma}^m_{s^c \times s} (\hat{\Sigma}^m_{s \times s})^{-1} \Big|_{l_1} \cdot \Big(\frac{\Big| \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i e_i \Big|_\infty}{\lambda} + \frac{\Big| (\hat{\Sigma}^n - \hat{\Sigma}^m) \cdot (\beta^* - \hat{\beta}_0) \Big|_\infty}{\lambda} + 1 \Big)
\end{aligned}$$

进而，

$$\frac{1}{\lambda}\left|\left(-\frac{1}{m}\sum_{i\in H_1}\mathbf{X}_i(Y_i-\mathbf{X}_i^T\beta^o)\right)_{s^c}-\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i(Y_i-\mathbf{X}_i^T\hat{\beta}_0)-\frac{1}{m}\sum_{i\in H_1}\mathbf{X}_i(Y_i-\mathbf{X}_i^T\hat{\beta}_0)\right)_{s^c}\right|_\infty$$
$$=\frac{1}{\lambda}\left|\left(\hat{\Sigma}^m(\beta^o-\beta^*)\right)_{s^c}+\left((\hat{\Sigma}^m-\hat{\Sigma}^n)\cdot(\beta^*-\hat{\beta}_0)\right)_{s^c}-\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_ie_i\right)_{s^c}\right|_\infty$$
$$\leq\left|\hat{\Sigma}^m_{s^c\times s}(\hat{\Sigma}^m_{s\times s})^{-1}\right|_{l_1}\cdot\left(\frac{\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_ie_i\right|_\infty}{\lambda}+\frac{\left|(\hat{\Sigma}^n-\hat{\Sigma}^m)\cdot(\beta^*-\hat{\beta}_0)\right|_\infty}{\lambda}+1\right)+\frac{\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_ie_i\right|_\infty}{\lambda}+\frac{\left|(\hat{\Sigma}^n-\hat{\Sigma}^m)\cdot(\beta^*-\hat{\beta}_0)\right|_\infty}{\lambda}$$

对于任意$\varepsilon>0$，

由(22)式与3.4.2的证明，并结合条件中$\left|\Sigma_{s^c\times s}\cdot\Sigma^{-1}_{s\times s}\right|_{l_1}<\frac{1}{3}$，可知存在充分大的$L$，$\forall n>L,\forall p>L$，

$$P\left\{\frac{1}{\lambda}\left|\left(-\frac{1}{m}\sum_{i\in H_1}\mathbf{X}_i(Y_i-\mathbf{X}_i^T\beta^o)\right)_{s^c}-\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i(Y_i-\mathbf{X}_i^T\hat{\beta}_0)-\frac{1}{m}\sum_{i\in H_1}\mathbf{X}_i(Y_i-\mathbf{X}_i^T\hat{\beta}_0)\right)_{s^c}\right|_\infty<1\right\}>1-\varepsilon$$

而当$\frac{1}{\lambda}\left|\left(-\frac{1}{m}\sum_{i\in H_1}\mathbf{X}_i(Y_i-\mathbf{X}_i^T\beta^o)\right)_{s^c}-\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i(Y_i-\mathbf{X}_i^T\hat{\beta}_0)-\frac{1}{m}\sum_{i\in H_1}\mathbf{X}_i(Y_i-\mathbf{X}_i^T\hat{\beta}_0)\right)_{s^c}\right|_\infty<1$成立时，

存在z^o满足：$z^o_{s^c}\in(-1,1)$以及$z^o_s=\tilde{z}_s$，使得：

$$\left(-\frac{1}{m}\sum_{i\in H_1}\mathbf{X}_i(Y_i-\mathbf{X}_i^T\beta^o)\right)-\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i(Y_i-\mathbf{X}_i^T\hat{\beta}_0)-\frac{1}{m}\sum_{i\in H_1}\mathbf{X}_i(Y_i-\mathbf{X}_i^T\hat{\beta}_0)\right)+\lambda\cdot z^o=0$$

根据凸优化理论，此时β^o即为boosting估计对应优化问题的一个解。而对于其任意的解$\hat{\beta}$：

$$\left(-\frac{1}{m}\sum_{i\in H_1}\mathbf{X}_i(Y_i-\mathbf{X}_i^T\hat{\beta})\right)-\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i(Y_i-\mathbf{X}_i^T\hat{\beta}_0)-\frac{1}{m}\sum_{i\in H_1}\mathbf{X}_i(Y_i-\mathbf{X}_i^T\hat{\beta}_0)\right)+\lambda\cdot z=0$$

根据3.3.1的结论，$z = z^o$。所以$z_{s^c} \in (-1, 1)$，即$s^c \subset \hat{s}^c$。

由此：

$$\forall \varepsilon, \exists L, \forall n > L, \forall p > L, P\Big\{\hat{s} \subset s\Big\} > 1 - \varepsilon$$

这样就证明了：

$$\lim_{(n,p)\to\infty} P\{\hat{s} \subset s\} = 1$$

再证$\lim_{(n,p)\to\infty} P\{s \subset \hat{s}\} = 1$：

若假定$\hat{s} \subset s$，

对于boosting估计对应的优化问题的任意解$\hat{\beta}$，由凸优化理论：

$$\Big(-\frac{1}{m}\sum_{i\in H_1}\mathbf{X}_i(Y_i - \mathbf{X}_i^T\hat{\beta})\Big) - \Big(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i(Y_i - \mathbf{X}_i^T\hat{\beta}_0) - \frac{1}{m}\sum_{i\in H_1}\mathbf{X}_i(Y_i - \mathbf{X}_i^T\hat{\beta}_0)\Big) + \lambda \cdot z = 0$$

由于：

$$\hat{\beta}_s - \beta_s^* = (\hat{\Sigma}^m_{s\times s})^{-1} \cdot \Big[(\hat{\Sigma}^n - \hat{\Sigma}^m)\cdot(\beta^* - \hat{\beta}_0) + \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i e_i - \lambda z\Big]_s$$

从而，

$$\left|\hat{\beta}_s - \beta_s^*\right|_\infty \leq \left|(\hat{\Sigma}^m_{s\times s})^{-1}\right|_{l_1} \cdot \left(\left|(\hat{\Sigma}^n - \hat{\Sigma}^m)\cdot(\beta^* - \hat{\beta}_0)\right|_\infty + \left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i e_i\right|_\infty + \lambda\right)$$

对于上述ε，

根据假设条件以及之前的结论，存在充分大的L，$\forall n > L, \forall p > L$，

由第一部分所证，

$$P\left\{\hat{s} \subset s\right\} > 1 - \varepsilon$$

可知：

$$P\left\{\left|\hat{\beta}_s - \beta_s^*\right|_\infty \leq \frac{C_1 + C_2}{2} \cdot \left(10\sigma\sqrt{\frac{logp}{n}} + 100M^2 \cdot \sqrt{\frac{slogp}{m}}a_n\right)\right\} > 1 - 2\varepsilon$$

再结合条件$\underset{i\in s}{min}|\beta_i^*| \geq C_2 \cdot \left(10\sigma\sqrt{\frac{logp}{n}} + 100M^2 \cdot \sqrt{\frac{slogp}{m}}a_n\right)$，并且$\underset{i\in s}{min}|\hat{\beta}_i| \geq \underset{i\in s}{min}|\beta_i^*| - \left|\hat{\beta}_s - \beta_s^*\right|_\infty$，可知：

$$\forall \varepsilon > 0, \exists L > 0, \forall n > L, \forall p > L, P\left\{\underset{i\in s}{min}|\hat{\beta}_i| > 0\right\} > 1 - 2\varepsilon$$

这样就证明了：

$$\lim_{(n,p)\to\infty} P\{s \subset \hat{s}\} = 1$$

证毕。

3.5 Lasso与Dantzig selector

3.5.1 Dantzig selector简介

在总体方面，依然考虑与上面相同的线性模型。

对于样本$(\mathbf{X}_i, Y_i)_{i=1,2,...,n}$，记数据矩阵为$\mathbf{X}$，也就是$\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n]^T$，并记$\mathbf{Y} = [Y_1, Y_2, ..., Y_n]^T$。

关于线性模型中的参数β^*的Dantzig 估计$\hat{\beta}_D$为下面优化问题的解：

$$\min_{\beta \in \mathbf{R}^p} |\beta|_1 \quad s.t. \ |\mathbf{X}^T(\mathbf{Y} - X\beta)|_\infty \leq \lambda_D$$

其中，λ_D为*tunning parameter*。

事实上，与Lasso类似，由于这里的优化问题的解并不唯一，因此这不是一个严格的估计量定义。

由于上述优化问题的解是稀疏的，因此我们可以将其作为一种变量选择方法，称为*Dantzig selector*。

3.5.2 Lasso的一种等价形式

在这一小节，对于Lasso估计，我们给出一种形式上与Dantzig估计类似等价形式。

对于样本$(\mathbf{X}_i, Y_i)_{i=1,2,...,n}$，

任意*tunning parameter* λ_L，相应的Lasso估计解集S_L：

$$S_L = \left\{ \hat{\beta} \in \mathbf{R}^p \ \middle| \ \frac{1}{2}|\mathbf{Y} - \mathbf{X}\beta|_2^2 + \lambda_L|\beta|_1 \geq \frac{1}{2}|\mathbf{Y} - \mathbf{X}\hat{\beta}|_2^2 + \lambda_L|\hat{\beta}|_1, \ \forall \beta \in \mathbf{R}^p \right\}$$

存在一个*tunning parameter* λ_l，与之相应的解集S_l：

$$S_l = \left\{ \hat{\beta} \ : \ |\mathbf{Y} - \mathbf{X}\hat{\beta}|_2^2 \leq \lambda_l \ \middle| \ |\beta|_1 \geq |\hat{\beta}|_1, \ \forall \beta \ : \ |\mathbf{Y} - \mathbf{X}\beta|_2^2 \leq \lambda_l \right\}$$

使得：$S_l = S_L$。

证明 任取S_L中的一元素$\hat{\beta}_0$，取$\lambda_l = |\mathbf{Y} - \mathbf{X}\hat{\beta}_0|_2^2$，由3.3.1的结论可知，$\lambda_l$与$\hat{\beta}_0$ 的选取无关。

此时，$\forall \beta \ : \ |\mathbf{Y} - \mathbf{X}\beta|_2^2 \leq |\mathbf{Y} - \mathbf{X}\hat{\beta}_0|_2^2$：

$$\begin{aligned}
|\beta|_1 &= \frac{\lambda_L |\beta|_1}{\lambda_L} \\
&\geq \frac{\frac{1}{2}|\mathbf{Y} - \mathbf{X}\beta|_2^2 + \lambda_L |\beta|_1 - \frac{1}{2}|\mathbf{Y} - \mathbf{X}\hat{\beta}_0|_2^2}{\lambda_L} \\
&\geq \frac{\frac{1}{2}|\mathbf{Y} - \mathbf{X}\hat{\beta}_0|_2^2 + \lambda_L |\hat{\beta}_0|_1 - \frac{1}{2}|\mathbf{Y} - \mathbf{X}\hat{\beta}_0|_2^2}{\lambda_L} \\
&\geq |\hat{\beta}_0|_1
\end{aligned}$$

这便证明了，对于上述选取的*tunning parameter* λ_l，有$S_L \subset S_l$。

另一方面，对于$\forall \hat{\beta} \in S_l$，有：

$$\begin{aligned}
&|\mathbf{Y} - \mathbf{X}\hat{\beta}|_2^2 \leq |\mathbf{Y} - \mathbf{X}\hat{\beta}_0|_2^2 \\
&|\hat{\beta}|_1 \leq |\hat{\beta}_0|_1
\end{aligned}$$

从而有：$\frac{1}{2}|\mathbf{Y} - \mathbf{X}\hat{\beta}_0|_2^2 + \lambda_L |\hat{\beta}_0|_1 \geq \frac{1}{2}|\mathbf{Y} - \mathbf{X}\hat{\beta}|_2^2 + \lambda_L |\hat{\beta}|_1$.

所以有$S_l \subset S_L$。

评注

- 在上述的结论中，λ_l的选取是依赖于样本$(\mathbf{X}_i, Y_i)_{i=1,2,\ldots,n}$的。

  由于在实际数据处理中，*tunning parameter*的选取本就依赖于样本，因此这是合理的。

- 对于基于样本的对参数的任意一个最小二乘估计β_{LS}，利用广义逆的性质可知：

  任意*tunning parameter* λ_L，相应的Lasso估计解集S_L：

  存在一个*tunning parameter* λ_c，与之相应的解集S_c：

$$S_c = \left\{ \hat{\beta} \ : \ |\mathbf{X}(\hat{\beta} - \hat{\beta}_{LS})|_2^2 \leq \lambda_c \ \middle| \ |\beta|_1 \geq |\hat{\beta}|_1, \ \forall \beta \ : \ |\mathbf{X}(\beta - \hat{\beta}_{LS})|_2^2 \leq \lambda_c \right\}$$

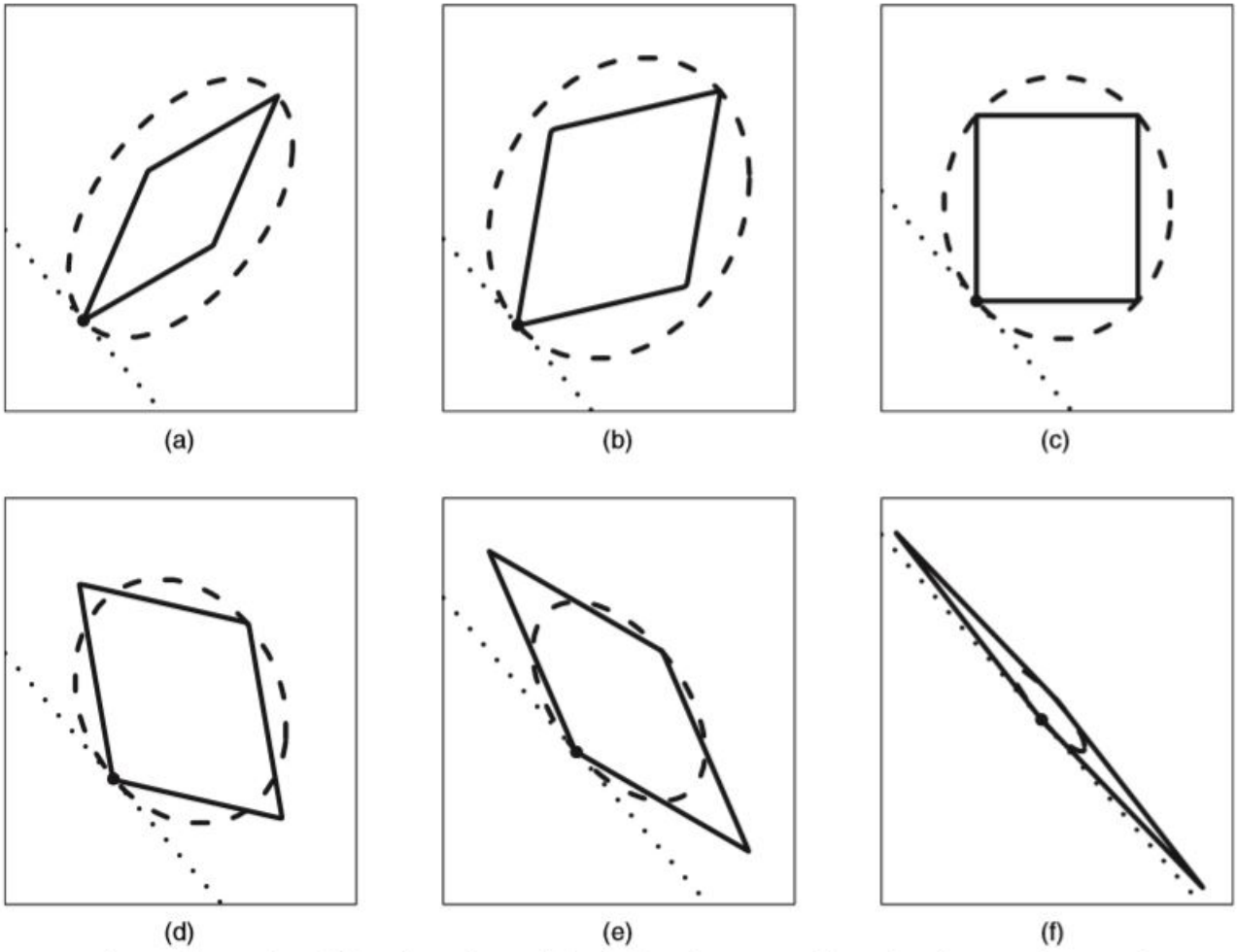
  使得：$S_c = S_L$。

  事实上，Dantzig估计的解集S_D可以写成：

$$S_D = \left\{ \hat{\beta} \ : \ |\mathbf{X}^T\mathbf{X}(\hat{\beta} - \hat{\beta}_{LS})|_\infty \leq \lambda_D \ \middle| \ |\beta|_1 \geq |\hat{\beta}|_1, \ \forall \beta \ : \ |\mathbf{X}^T\mathbf{X}(\beta - \hat{\beta}_{LS})|_\infty \leq \lambda_D \right\}$$

这里基于数据矩阵的不同情况（两两相关系数从小变大），更直观地反映S_c与S_D的关系。

此处图片出自论文DASSO: connections between the Dantzig selector and lasso(Gareth M.James, Peter Radchenko and Jinchi Lv, 2009)。



Lasso (– – –) and Dantzig selector (◊) solutions in a $p = 2$ dimensional space (·····, L_1-norm that is being minimized): (a) $\rho = -0.5$; (b) $\rho = -0.2$; (c) $\rho = 0$; (d) $\rho = 0.2$; (e) $\rho = 0.5$; (f) $\rho = 0.9$

3.5.3 Lasso与Dantzig selector

下面我们严格证明Lasso估计与Dantzig selector估计的关系。

设$\hat{\beta}$为Lasso估计$\underset{\beta\in\mathbf{R}^p}{argmin}\frac{1}{2}|\mathbf{Y}-\mathbf{X}\beta|_2^2+\lambda|\beta|_1$的一个解，记$\hat{\beta}$非零分量对应的下标集为$I_L$，并记$p$ 维向量$\mathbf{X}^T(\mathbf{Y}-\mathbf{X}\beta)$中等于$\lambda$的分量对应的下标集为$I_1$，$p$ 维向量$\mathbf{X}^T(\mathbf{Y}-\mathbf{X}\beta)$中等于$-\lambda$的分量对应的下标集为$I_2$，$\mathbf{X}_{I_1}$为$X$ 关于下标集I_1的列切片组成的矩阵，$\mathbf{X}_{I_2}$为$-X$ 关于下标集I_2的列切片组成的矩阵，$\mathbf{X}=[\mathbf{X}_{I_1},\mathbf{X}_{I_2}]$.

若$\mathbf{X}^T\mathbf{X}_I$列满秩，

则$\hat{\beta}$为下列优化问题的解：

$$\underset{\beta\in\mathbf{R}^p}{min}|\beta|_1 \quad s.t.\ |\mathbf{X}^T(\mathbf{Y}-\mathbf{X}\beta)|_\infty\le\lambda$$

当且仅当存在一个各分量非负的$|I|$维向量$\mathbf{u}$

以及一个p维向量$z\in\left\{(v_1,v_2,...,v_p)^T\middle|v_j=sgn(\hat{\beta}_j)\cdot\mathbf{1}\{j\in I_L\}+\theta\cdot\mathbf{1}\{j\notin I_L\},\ \theta\in[-1,1]\right\}$,

使得$z-\mathbf{X}^T\mathbf{X}_I\mathbf{u}=0$.

证明 因为$\hat{\beta}$为Lasso估计$\underset{\beta\in\mathbf{R}^p}{argmin}\frac{1}{2}|\mathbf{Y}-\mathbf{X}\beta|_2^2+\lambda|\beta|_1$的一个解，由凸优化理论可得：

$$\exists\tilde{z}\ :\ |\tilde{z}|_\infty\le1,\ -\mathbf{X}^T(\mathbf{Y}-\mathbf{X}\hat{\beta})+\lambda\tilde{z}=0$$

由此可得$|\mathbf{X}^T(\mathbf{Y}-\mathbf{X}\hat{\beta})|_\infty\le\lambda$.

即$\hat{\beta}$是优化问题：

$$\underset{\beta\in\mathbf{R}^p}{min}|\beta|_1 \quad s.t.\ |\mathbf{X}^T(\mathbf{Y}-\mathbf{X}\beta)|_\infty\le\lambda$$

的一个可行解。

而$\hat{\beta}$为下列优化问题的解：

$$\min_{\beta \in \mathbf{R}^p} |\beta|_1 \quad s.t.\ |\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\beta)|_\infty \leq \lambda$$

若记$\mathbf{X} = [x_1, x_2, ..., x_p]$，则上述优化问题等价于：

$$\begin{aligned} &\min_{\beta \in \mathbf{R}^p} |\beta|_1 \\ &s.t. \quad x_j^T(\mathbf{Y} - \mathbf{X}\beta) \leq \lambda \\ &\quad -x_j^T(\mathbf{Y} - \mathbf{X}\beta) \leq \lambda \end{aligned}$$

由KKT条件的充分必要性即得结论成立。

4 高维矩阵估计概述

4.1 稀疏协方差矩阵的估计

4.1.1 传统方法存在的问题以及关于一个基本性质的证明

传统方法的一个基本性质 总体上的随机向量$\mathbf{X}$，有$E\mathbf{X}=\mu,\ Cov(\mathbf{X})=\Sigma.$

考虑我们有独立同分布的样本$\mathbf{X}_1,\mathbf{X}_2,...,\mathbf{X}_n.$

在经典统计学中，我们以样本协方差矩阵来估计Σ.

$$\hat{\Sigma}=\frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{X}_k-\bar{\mathbf{X}})(\mathbf{X}_k-\bar{\mathbf{X}})^T$$

有统计学的经典理论，样本协方差矩阵具有无偏性，以及相对矩阵中的每个元素而言具有相合性。

更进一步地，记$\hat{\Sigma}=(\hat{\sigma}_{ij})_{p\times p}$，

若:

$$\lim_{(n,p)\to\infty}\frac{logp}{n}=0$$

则有：

$$\underset{1\le i,j\le p}{max}|\hat{\sigma}_{ij}-\sigma_{ij}|/\sqrt{\sigma_{ii}\sigma_{jj}}=O_p\left(\sqrt{\frac{logp}{n}}\right)$$

这里，O_p是相对双指标(n,p)的。

下面我们在$\exists M>0,\ |\mathbf{X}_i|\le M,\ i=1,2,...,p,\ a.s.$下证明上述结论。

证明：

对于任意(i,j) ： $1 \le i,j \le p$，将$\mathbf{X}_i, \mathbf{X}_j$对应的样本记为$x_1, x_2, ..., x_n$以及$y_1, y_2, ..., y_n$。

不妨设$E[\mathbf{X}_i] = E[\mathbf{X}_j] = 0,\ Var[\mathbf{X}_i] = Var[\mathbf{X}_j] = 1$。

此时，由Berstein不等式，对于充分大的n与p：

$$
\begin{aligned}
&P\left\{\left|\frac{1}{n-1}\sum_{k=1}^{n}\Big(x_k-\bar{x}\Big)\Big(y_k-\bar{y}\Big)\right| > C\sqrt{\frac{logp}{n}}\right\} \\
&= P\left\{\left|\frac{1}{n-1}\sum_{k=1}^{n}\Big(x_k y_k-\sigma_{xy}\Big) - \frac{n}{n-1}\Big(\bar{x}\bar{y}-\frac{\sigma_{xy}}{n}\Big)\right| > C\sqrt{\frac{logp}{n}}\right\} \\
&\le P\left\{\left|\frac{1}{n-1}\sum_{k=1}^{n}\Big(x_k y_k-\sigma_{xy}\Big)\right| > \frac{C}{2}\sqrt{\frac{logp}{n}}\right\} + P\left\{\left|\frac{n}{n-1}\Big(\bar{x}\bar{y}-\frac{\sigma_{xy}}{n}\Big)\right| > \frac{C}{2}\sqrt{\frac{logp}{n}}\right\} \\
&\le P\left\{\left|\frac{1}{n}\sum_{k=1}^{n}\Big(x_k y_k-\sigma_{xy}\Big)\right| > \frac{C}{4}\sqrt{\frac{logp}{n}}\right\} + P\left\{\left|\bar{x}\bar{y}-\frac{\sigma_{xy}}{n}\right| > \frac{C}{4}\sqrt{\frac{logp}{n}}\right\} \\
&\le P\left\{\left|\frac{1}{n}\sum_{k=1}^{n}\Big(x_k y_k-\sigma_{xy}\Big)\right| > \frac{C}{4}\sqrt{\frac{logp}{n}}\right\} + P\left\{\left|\bar{x}\bar{y}\right| > \frac{C}{8}\sqrt{\frac{logp}{n}}\right\} \\
&\le P\left\{\left|\frac{1}{n}\sum_{k=1}^{n}\Big(x_k y_k-\sigma_{xy}\Big)\right| > \frac{C}{4}\sqrt{\frac{logp}{n}}\right\} + P\left\{\left|\bar{x}\right| > \frac{C}{8M}\sqrt{\frac{logp}{n}}\right\} \\
&\le 2exp\Big\{-\frac{C^2 logp}{32M^4 + \frac{16}{3}M^2 C\sqrt{\frac{logp}{n}}}\Big\} + 2exp\Big\{-\frac{C^2 logp}{128M^4 + \frac{16}{3}M^2 C\sqrt{\frac{logp}{n}}}\Big\} \\
&\le p^{-A}
\end{aligned}
$$

最后一个不等号利用了条件$\lim_{(n,p)\to\infty}\frac{logp}{n} = 0$，另外，$A$可以通过调整C达到任意大。

从而，对于充分大的n与p：

$$
P\left\{\max_{1\le i,j\le p}|\hat{\sigma}_{ij}-\sigma_{ij}|/\sqrt{\sigma_{ii}\sigma_{jj}} > C\sqrt{\frac{logp}{n}}\right\} \le p^{-A+2}
$$

得证。

存在的问题

- 当$p > n$时，$\hat{\Sigma}$不可逆，而总体参数Σ是可逆的。
- 当$\lim_{(n,p)\to\infty} \frac{p}{n} = r,\ r > 0$时，$\hat{\Sigma}$在以算子模作为度量下不具有相合性。

4.1.2 条状协方差矩阵的估计

在总体方面，考虑这样一族协方差矩阵$\mathcal{B}_\alpha = \left\{\Sigma : \max_i \sum_{j:j-i|\geq k} |\sigma_{ij}| \leq C_0 k^{-\alpha},\ \alpha > 0\right\}$.

事实上，这类协方差矩阵在时间序列研究中十分常见。

Banding estimator Bickel与Levina根据截断的思想提出了关于这类协方差矩阵的估计：

$$\hat{\mathbf{\Sigma}}_{band} = \left(\hat{\sigma}_{ij} I\left\{|i-j| \leq k_n\right\}\right)_{p\times p}$$

下面我们依然在$\exists M > 0,\ |\mathbf{X}_i| \leq M,\ i = 1, 2, ..., p,\ a.s.$ 下证明：

若：

$$\lim_{(n,p)\to\infty} \frac{logp}{n} = 0$$

则：

$$|\hat{\mathbf{\Sigma}}_{band} - \Sigma|_{l_1} = O_p\left(k_n\sqrt{\frac{logp}{n}} + k_n^{-\alpha}\right)$$

证明：

由于

$$\left|\hat{\mathbf{\Sigma}}_{band} - \mathbf{\Sigma}\right|_{l_1} \leq \left|\hat{\mathbf{\Sigma}}_{band} - \left(\sigma_{ij} I\left\{|i-j| \leq k_n\right\}\right)_{p\times p}\right|_{l_1} + \left|\left(\sigma_{ij} I\left\{|i-j| > k_n\right\}\right)_{p\times p}\right|_{l}$$

可知

$$P\left\{\underset{1\le i,j\le p}{max}|\hat{\sigma}_{ij}-\sigma_{ij}|/\sqrt{\sigma_{ii}\sigma_{jj}}\le C\sqrt{\frac{logp}{n}}\right\}\le P\left\{\left|\hat{\mathbf{\Sigma}}_{band}-\mathbf{\Sigma}\right|_{l_1}\le max\{C_0,M^2C\}\cdot\left(k_n\sqrt{\frac{logp}{n}}+k_n^{-\alpha}\right)\right\}$$

得证。

此时若取$k_n=\left(\frac{n}{\log p}\right)^{\frac{1}{2+2\alpha}}$便可得到：

$$\left\|\hat{\mathbf{\Sigma}}_{band}-\mathbf{\Sigma}\right\|_{l_1}=O_p\left((\log p/n)^{\alpha/(2+2\alpha)}\right)$$

4.1.3 稀疏协方差矩阵的估计

在总体方面，考虑这样一族协方差矩阵$\mathcal{S}(s_p)=\left\{\mathbf{\Sigma}\succ 0:\max_i\sigma_{ii}\le K,\quad \max_i\sum_{j=1}^{p}I\left\{\sigma_{ij}\neq 0\right\}\le s_p\right\}$.

Thresholding estimator 根据之前所证的结论：

$$\max_{1\le i,j\le p}|\hat{\sigma}_{ij}-\sigma_{ij}|=O_p\left(\sqrt{\frac{\log p}{n}}\right)$$

El Karoui(2008)与Bickel and Levina(2008)根据截断想法提出thresholding estimator:

$$\hat{\mathbf{\Sigma}}(\lambda_n)=\left(\hat{\sigma}_{ij}I\left\{|\hat{\sigma}_{ij}|\ge\lambda_n\right\}\right)_{p\times p},\quad \lambda_n=C\sqrt{\log p/n}$$

该统计量具有以下性质：

若：

$$\lim_{(n,p)\to\infty}\frac{logp}{n}=0$$

则：

$$\forall \varepsilon,\ \exists C,\ \exists A, \forall n > A,\ \forall p > A,\ P\left\{ \left| \hat{\mathbf{\Sigma}}\left(\lambda_n\right) - \mathbf{\Sigma} \right|_{l_1} > 2Cs_p\sqrt{\frac{logp}{n}} \right\} < \varepsilon$$

其中$\lambda_n = C\sqrt{\log p/n}$依赖于$\varepsilon,\ p,\ n.$

而这对于作为*tunning parameter*的λ_n而言，这是合理的。

证明：

记总体协方差矩阵Σ的第j列非零元素的行标集为S_j。

由

$$\left| \hat{\mathbf{\Sigma}}\left(\lambda_n\right) - \mathbf{\Sigma} \right|_{l_1} \leq \max_j \sum_{i=1}^{p} \left| \hat{\sigma}_{ij} I\left\{ |\hat{\sigma}_{ij}| \geq \lambda_n \right\} - \sigma_{ij} \right| \leq \max_j \left[\sum_{i \in S_j} \left| \hat{\sigma}_{ij} I\left\{ |\hat{\sigma}_{ij}| \geq \lambda_n \right\} - \sigma_{ij} \right| + \sum_{i \in S_j^c} |\hat{\sigma}_{ij}| I\left\{ |\hat{\sigma}_{ij}| \geq \lambda_n \right\} \right]$$

以及

$$\begin{aligned}
&\max_j \sum_{i \in S_j} \left| \hat{\sigma}_{ij} I\left\{ |\hat{\sigma}_{ij}| \geq \lambda_n \right\} - \sigma_{ij} \right| \\
&\leq \max_j \sum_{i \in S_j} |\hat{\sigma}_{ij} - \sigma_{ij}| + \max_j \sum_{i \in S_j} |\hat{\sigma}_{ij}| I\left\{ |\hat{\sigma}_{ij}| < \lambda_n \right\} \\
&\leq \max_j \sum_{i \in S_j} |\hat{\sigma}_{ij} - \sigma_{ij}| + Cs_p\sqrt{\frac{\log p}{n}}
\end{aligned}$$

可知

$$P\left\{ \max_{1 \leq i,j \leq p} |\hat{\sigma}_{ij} - \sigma_{ij}| < C\sqrt{\frac{logp}{n}} \right\} \leq P\left\{ \left| \hat{\mathbf{\Sigma}}\left(\lambda_n\right) - \mathbf{\Sigma} \right|_{l_1} \leq 2Cs_p\sqrt{\frac{logp}{n}} \right\}$$

得证。

自适应thresholding estimator 自适应thresholding estimator仍然基于与之前相同的截断思想。

但考虑到对于不同下标的估计量$\hat{\sigma}_{ij}$，其作为随机变量有着不同的方差，设$\hat{\sigma}_{ij}$的渐进方差为θ_{ij}。若各个(i,j)对应的渐进方差差异过大，则对于基于截断的thresholding estimator，截断门限λ_n应随着(i,j)而改变。

基于这一想法，Cai与Liu在2011年提出自适应thresholding estimator：

$$\hat{\mathbf{\Sigma}}^{\star}(\delta) = \left(\hat{\sigma}_{ij} I\left\{|\hat{\sigma}_{ij}| \geq \lambda_{ij}\right\}\right)_{p\times p}$$

其中

$$\lambda_{ij} = \delta\sqrt{\frac{\hat{\theta}_{ij}\log p}{n}}, \delta > 0$$

$$\hat{\theta}_{ij} = \frac{1}{n}\sum_{k=1}^{n}\left[\left(X_{ki} - \bar{X}^i\right)\left(X_{kj} - \bar{X}^j\right) - \hat{\sigma}_{ij}\right]^2, \quad \bar{X}^i = n^{-1}\sum_{k=1}^{n} X_{ki}$$

这里可以证明$\hat{\theta}_{ij}$是对θ_ij的一个相合估计。

并且可以证明，在适当条件下：

$$\lim_{(n,p)\to\infty} P\left\{\bigcap_{1\leq i,j\leq p}\left[|\hat{\sigma}_{ij} - \sigma_{ij}| \leq 2\sqrt{\frac{\hat{\theta}_{ij}\log p}{n}}\right]\right\} = 1$$

4.2 稀疏协方差矩阵逆矩阵的估计

高维数据统计推断中，判别分析、对均值的假设检验以及高斯图模型的估计都涉及到协方差矩阵逆矩阵（精度矩阵）的估计。

这里高斯图模型是指对于由p个随机变量组成的顶点集合$\{X_1, X_2, ..., X_p\}$，定义X_i与X_j有边相连，当且仅当$Cov\Big(X_i, X_j \mid X_k,\ k \neq i, j\Big) \neq 0$。而在下面一个多元统计分析中的结论中，我们可以看出估计高斯图模型等价于估计逆矩阵。

在这一小节，我们先证明两个多元统计中重要的结论，而后重点介绍Cai与Liu提出的CLIME 方法。

4.2.1 两个多元统计分析中的结论及其证明

下面两个结论只涉及到总体层面上的性质，但这两个基本的性质是很多统计推断方法的出发点，因此我们单独列出并证明之。

设$\mathbf{X} = [X_1, X_2, ..., X_p]^T \sim N(0, \Sigma)$，并记$\Sigma^{-1} = \Omega = (w_{ij})_{p\times p}$.

逆矩阵与条件相关性

$$Cov\Big(X_i, X_j\Big|X_k, k \neq i, j\Big) = 0 \Leftrightarrow w_{ij} = 0,\ \forall i \neq j.$$

证明：

由对称性，等价于证明$Cov\Big(X_1, X_2\Big|X_k, k \neq 1, 2\Big) = 0 \Leftrightarrow w_{12} = 0$

记$\mathbf{X}_1 = [X_1, X_2]^T,\ \mathbf{X}_2 = [X_3, X_4, ..., X_p]^T$，并记$[\mathbf{X}_1, \mathbf{X}_2]^T$的协方差矩阵为

$$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

由多元统计学的结论：

$$\mathbf{X}_1\Big|\mathbf{X}_2 \sim N(0, \Sigma_{1\cdot 2})$$

其中$\Sigma_{1\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$，而根据分块矩阵求逆的结果，$\Sigma_{1\cdot 2}$恰好为$\Omega$的左上角$2\times 2$ 的子块。

因此得证。

逆矩阵与线性模型

$$X_k - \sum_{i \neq k} \beta_{ki} X_i \perp [X_1, ..., X_{k-1}, X_{k+1}, ..., X_p]^T \Leftrightarrow \beta_{ki} = -\frac{w_{ki}}{w_{kk}},\ i \neq k$$

证明： 由对称性，只证明$k = 1$的情况。

将协方差矩阵Σ写成分块矩阵的形式：

$$\begin{bmatrix} \Sigma_{11}^0 & \Sigma_{12}^0 \\ \Sigma_{21}^0 & \Sigma_{22}^0 \end{bmatrix}$$

这里，Σ_{11}^0表示Σ的左上角第一个元素。

此时$X_1 - \sum_{i \neq 1} \beta_{1i} X_i = [1, -\beta_{12}, ..., -\beta_{1p}]^T \mathbf{X}$.

以及$[X_2, X_3, ..., X_p]^T = A\mathbf{X}$.

其中，

$$A = \begin{bmatrix} 0 & 0 \\ 0 & I_{p-1} \end{bmatrix}$$

注意到$[X_1 - \sum_{i \neq 1} \beta_{1i} X_i, X_2, X_3, ..., X_p]^T = B\mathbf{X}$.

其中，

$$B = \begin{bmatrix} 1 & [-\beta_{12}, ..., -\beta_{1p}] \\ 0 & I_{p-1} \end{bmatrix}$$

为满秩矩阵。因此$X_1 - \sum_{i\neq 1} \beta_{1i} X_i \perp [X_2, X_3, ..., X_p]^T$ 等价于：

$$[0, I_{p-1}] \ \Sigma \ [1, -\beta_{12}, ..., -\beta_{1p}]^T = 0$$

进一步等价于

$$[\beta_{12}, ..., \beta_{1p}]^T = (\Sigma^0_{22})^{-1} \Sigma^0_{21}$$

而注意到，当$\beta_{1i} = -\frac{w_{1i}}{w_{11}}$时，

$$[1, -\beta_{12}, ..., -\beta_{1p}] = \frac{1}{w_{11}} \cdot (1, 0, ..., 0) \ \Sigma^{-1}$$

此时显然有:

$$[0, I_{p-1}] \ \Sigma \ [1, -\beta_{12}, ..., -\beta_{1p}]^T = 0$$

因此结论得证。

事实上，若记$\varepsilon_k = X_k - \sum_{i\neq k} \beta_{ki} X_i$.

由上述证明中，

$$[1, -\beta_{12}, ..., -\beta_{1p}] = \frac{1}{w_{11}} \cdot (1, 0, ..., 0) \ \Sigma^{-1}$$

可以直接得到$Cov(\varepsilon_i, \varepsilon_j) = 0 \ \Leftrightarrow w_{ij} = 0$.

4.2.2 限制域最小化估计(CLIME)及其极限性质的证明

基于Dantzig selector的想法，记$\{\hat{\Omega}_1\}$为下述优化问题的解：

$$\min |\Omega|_1 \ \ s.t.\ |\Sigma\Omega - I|_\infty \le \lambda_n, \quad \Omega \in R^{p\times p} \tag{24}$$

这里，$|\cdot|_1$指矩阵中各个元素绝对值之和，$|\cdot|_\infty$指矩阵中各元素绝对值的最大值，λ_n 为$tunning\ parameter$.

事实上，上述优化问题等价于：

$$\min |\omega_j|_1 \ s.t.\ |\Sigma\omega_j - e_j|_\infty \le \lambda_n,\ \omega \in R^p,\ 1 \le j \le p. \tag{25}$$

记$\{\hat{\Omega}_1\} = (\hat{\omega}_{ij}^1)$.

此时，CLIME估计定义为：

$$\hat{\Omega} = (\hat{\omega}_{ij}), \quad \hat{\omega}_{ij} = \hat{\omega}_{ji} = \hat{\omega}_{ij}^1 I\left\{\left|\hat{\omega}_{ij}^1\right| \le \left|\hat{\omega}_{ji}^1\right|\right\} + \hat{\omega}_{ji}^1 I\left\{\left|\hat{\omega}_{ij}^1\right| > \left|\hat{\omega}_{ji}^1\right|\right\}$$

下面我们证明CLIME估计的极限性质：

$$\forall\varepsilon,\ \exists C,\ \exists A,\ \forall n > A,\ \forall p > A,\ P\left\{\left|\hat{\Omega}_1 - \Omega\right|_{l_1} \ > \ 4Cs_p \cdot |\Omega|_{l_1} \cdot \sqrt{\frac{logp}{n}}\right\} < \varepsilon$$

其中，$\hat{\Omega}_1$是指（24）式中λ_n取为$C\sqrt{\frac{logp}{n}}$时的优化解。

也就是λ_n依赖于$\varepsilon,\ p,\ n$，而这对于作为$tunning\ parameter$ 的λ_n而言，这是合理的。

证明：

对于$\lambda_n = C\sqrt{\frac{logp}{n}}$，

假若$|\Sigma_n\Omega - I|_\infty \le \lambda_n$，

由（25）式，$\left|\hat{\Omega}_{1\cdot j}\right|_1 \le |\Omega_{\cdot j}|_1$，其中$A_{\cdot j}$表示矩阵A的第$j$列形成的列向量。

记S_j为$\Omega_{\cdot j}$的支撑，S为Ω的支撑，$h_j = \hat{\Omega}_{1\cdot j} - \Omega_{\cdot j}$.

进而

$$\left|\Omega_{\cdot j}\right|_1 \geq \left|\hat{\Omega}_{1\cdot j} - \Omega_{\cdot j} + \Omega_{\cdot j}\right|_1 = \left|(h_j)_{S_j^c}\right|_1 + \left|(h_j)_{S_j} + (\Omega_{\cdot j})_{S_j}\right|_1 \geq \left|(h_j)_{S_j^c}\right|_1 - \left|(h_j)_{S_j}\right|_1 + \left|(\Omega_{\cdot j})_{S_j}\right|_1$$

即得$\left|(h_j)_{S_j^c}\right|_1 \leq \left|(h_j)_{S_j}\right|_1$以及$\left|\left(\hat{\Omega}_1 - \Omega\right)_{S^c}\right|_{l_1} \leq \left|\left(\hat{\Omega}_1 - \Omega\right)_S\right|_{l_1}$.

所以，$\left|\hat{\Omega}_1 - \Omega\right|_{l_1} \leq 2\left|\left(\hat{\Omega}_1 - \Omega\right)_S\right|_{l_1} \leq 2s_p \cdot \left|\hat{\Omega}_1 - \Omega\right|_\infty$.

而，$\left|\hat{\Omega}_1 - \Omega\right|_\infty = \left|(\Omega\hat{\Sigma} - I)\hat{\Omega}_1 - \Omega\left(\hat{\Sigma}\hat{\Omega}_1 - I\right)\right|_\infty \leq 2\lambda_n\left|\Omega\right|_{l_1}$.

综上即有：

$$P\left\{\left|\Sigma_n\Omega - I\right|_\infty \leq C\sqrt{\frac{logp}{n}}\right\} \leq P\left\{\left|\hat{\Omega}_1 - \Omega\right|_{l_1} \leq 4Cs_p \cdot |\Omega|_{l_1} \cdot \sqrt{\frac{logp}{n}}\right\}$$

而由之前所证

$$\forall\varepsilon,\ \exists C,\ \exists A,\ \forall n > A,\ \forall p > A,\ P\left\{\left|\Sigma_n\Omega - I\right|_\infty > C\sqrt{\frac{logp}{n}}\right\} < \varepsilon$$

这也就完成了证明。

4.3 协方差矩阵逆矩阵支撑的估计

事实上，任意一种具有稀疏性的估计协方差逆矩阵的方法都可以用作估计协方差逆矩阵的支撑。

但是，一般的具有稀疏性的估计协方差逆矩阵的方法不方便直接地针对总体协方差矩阵逆矩阵支撑作出统计推断，我们在本小节介绍基于多重假设检验的方法来进行关于协方差矩阵逆矩阵支撑的统计推断。

我们先介绍多重假设检验中的重要概念：虚假发现率

4.3.1 虚假发现率

考虑同时做m个假设检验：$H_1, H_2, ..., H_m$，对应的$p-value$分别为$p_1, p_2, ..., p_m$.

若当$p_i \leq t$时拒绝H_i，则定义此时虚假发现率（FDR）为：

$$E\left[\frac{\sum_{i:\ H_i} \mathbf{1}\{p_i \leq t\}}{\sum_{i=1}^{m} \mathbf{1}\{p_i \leq t\} \ \vee 1}\right]$$

事实上，当$m=1$时，上式即为第一类错误的发生概率。

将FDR的一次样本实现称为FDP，在假设检验中，我们以下面的方式选取t：

$$\hat{t}_0 = sup\left\{t \in [0,1] \ : \ EDP \leq \alpha\right\}$$

由于假设检验问题中，原假设一般都是成立。因此也可以用下面的方式选取t：

$$\hat{t}_0 = sup\left\{t \in [0,1] \ : \ \frac{mt}{\sum_{i=1}^{m} \mathbf{1}\{p_i \leq t\} \ \vee 1} \leq \alpha\right\}$$

4.3.2 应用多重假设检验估计GGM

对于多元正态总体，

由4.2.1的结论可知如下模型是成立的：

$$X_{k,i} - \bar{X}_i = \left(\mathbf{X}_{k,-i} - \overline{\mathbf{X}}_{-i}\right)\boldsymbol{\beta}_i + \varepsilon_{k,i},\ 1 \leq k \leq n. \tag{26}$$

其中$\boldsymbol{\beta}_{j,i}=-\omega_{ij}/\omega_{ii},\boldsymbol{\beta}_i=(\beta_{1,i},\ldots,\beta_{p-1,i})^T$以及$\varepsilon_{k,i}$独立于$\mathbf{X}_{k,-i}$.

再根据4.2.1的结论$Cov(\varepsilon_i,\varepsilon_j)=0\ \Leftrightarrow w_{ij}=0$，我们可以根据残差来构造统计量进行假设检验。

上述理论都是在正态下成立，下面我们不假定总体正态分布，依然按照上面的想法构造统计量。

对于模型(26)，我们利用Lasso或是Dantzig selector来进行参数估计$\hat{\beta}_i$，

这里，我们记：

$$\max_{1\leq i\leq p}\left|\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i\right|_1=O_{\mathbf{P}}\left(a_{n1}\right)$$

以及

$$\min\left\{\lambda_{\max}^{1/2}(\boldsymbol{\Sigma})\max_{1\leq i\leq p}\left|\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i\right|_2,\ \max_{1\leq i\leq p}\sqrt{\left(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i\right)'\hat{\boldsymbol{\Sigma}}_{-i,-i}\left(\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}-\boldsymbol{\beta}_{\boldsymbol{i}}\right)}\right\}=O_{\mathbf{P}}\left(a_{n2}\right)$$

由此得到残差序列$\hat{\varepsilon}_{ki}=X_{k,i}-\bar{X}_i-\left(\mathbf{X}_{k,-i}-\overline{\mathbf{X}}_i\right)\hat{\boldsymbol{\beta}}_i$.

基于残差的样本协方差定义如下的检验统计量：

$$T_{ij1}:=\frac{1}{n}\left(\sum_{k=1}^{n}\hat{\varepsilon}_{ki}\hat{\varepsilon}_{kj}+\sum_{k=1}^{n}\hat{\varepsilon}_{ki}^2\hat{\beta}_{i,j}+\sum_{k=1}^{n}\hat{\varepsilon}_{kj}^2\hat{\beta}_{j-1,i}\right)$$

记

$$\begin{aligned}&b_{nij}=\omega_{ii}\hat{\sigma}_{ii,\varepsilon}+\omega_{jj}\hat{\sigma}_{jj,\varepsilon}-1\\&\left(\hat{\sigma}_{ij,\varepsilon}\right)_{1\leq i,j\leq p}=\tfrac{1}{n}\textstyle\sum_{k=1}^{n}\left(\varepsilon_k-\bar{\varepsilon}\right)\left(\varepsilon_k-\bar{\varepsilon}\right)'\\&\varepsilon_k=\left(\varepsilon_{k1},\ldots,\varepsilon_{kp}\right)',\bar{\varepsilon}=\tfrac{1}{n}\textstyle\sum_{k=1}^{n}\varepsilon_k\end{aligned}$$

那么若

$$\lim_{(n,p)\to\infty} \frac{logp}{n} = 0$$
$$\lim_{(n,p)\to\infty} a_{n1}\sqrt{logp} = 0$$
$$\lim_{(n,p)\to\infty} a_{n2} \cdot n^{-\frac{1}{4}} = 0$$

则

$$\sqrt{\frac{n}{\hat{r}_{ii}\hat{r}_{jj}}}\left(T_{ij1} + b_{nij}\frac{\omega_{ij}}{\omega_{ii}\omega_{jj}}\right) \xrightarrow{d} N\left(0, 1 + \frac{\omega_{ij}^2}{\omega_{ii}\omega_{jj}}\right)$$

这里是指关于(n,p)双指标的依分布收敛。

那么我们取检验统计量：

$$\hat{T}_{ij} = \sqrt{\frac{n}{\hat{r}_{ii}\hat{r}_{jj}}}\,|T_{ij1}|$$

以及拒绝规则：若$\hat{T}_{ij} > t$，则拒绝H_{0ij}.

最后我们依照4.3.1中FDP的定义，并根据极限分布确定t即可。

A 几个在这篇笔记中反复出现记号说明

1. $O_{a.s.}$ 首先阐述数学分析中，对于数列$\{a_n\}$与$\{b_n\}$，$a_n = O(b_n)$的含义是：

$$\exists C,\ \exists N,\ \forall n > N,\ |\frac{a_n}{b_n}| \leq C$$

对于随机变量序列$\{X_n\}$与$\{Y_n\}$，$X_n = O_{a.s.}(Y_n)$的含义是：

$$\exists C,\ P\left\{\exists N,\ \forall n > N,\ |\frac{X_n}{Y_n}| \leq C\right\} = 1$$

事实上，在本篇笔记中$O_{a.s.}$往往利用Borel-Cantelli引理得到。

2. 双指标极限 在高维统计中，研究往往涉及到关于二元组(n,p)的极限结果，在本篇笔记中符号$\lim_{(n,p)\to\infty}$反复出现，因此在这里给出严格定义。

对于双指标数列$\{a_{mn}\}$，

$$\lim_{(m,n)\to\infty} a_{mn} = a$$

定义为：

$$\forall \varepsilon > 0,\ \exists A \in \mathbf{N}^*,\ \forall m > A, \forall n > A,\ |a_{mn} - a| < \varepsilon.$$

- Cauchy收敛定理表明，数列$\{c_n\}$极限存在等价于$\lim_{(m,n)\to\infty} |c_n - c_m| = 0$
- 若$\lim_{(m,n)\to\infty} a_{mn} = a$，则对于任意单增数列$p_n$ 有$\lim_{n\to\infty} a_{p_n,n} = a$
- 与之同理，我们可以定义双指标O_p，双指标依分布收敛等

B 一些数学结论与证明

1.正态分布尾概率估计 对于任意$x>0$，

$$\frac{1}{\sqrt{2\pi}}\Big(\frac{1}{x}-\frac{1}{x^3}\Big)e^{-x^2/2}\leq\frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-t^2/2}dt\leq\frac{1}{x\sqrt{2\pi}}e^{-x^2/2}$$

更进一步地：

$$\frac{1}{\sqrt{2\pi}}\frac{x}{1+x^2}e^{-x^2/2}\leq\frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-t^2/2}dt$$

事实上，只需要对x求导即可得证。

2. 正态分布最大次序统计量 若$Y_i,i=1,2,...,n$服从标准正态分布，则：

$$\lim_{n\to\infty}P\Big\{\max_{1\leq i\leq n}|Y_i|^2-2logn+loglogn\leq x\Big\}=exp\Big(-\frac{e^{-\frac{x}{2}}}{\sqrt{\pi}}\Big)$$

证明 先证明一个引理：

若数列$\{a_n\}$满足$\lim_{n\to\infty}a_n=a$，

则有$\lim_{n\to\infty}(1-\frac{a_n}{n})^n=e^{-a}$.

事实上，

$$
\begin{aligned}
&(1-\frac{a_n}{n})^n \\
&=exp\Big(nlog(1-\frac{a_n}{n})\Big) \\
&=exp\Big(n\cdot(-\frac{a_n}{n}+\xi\frac{a_n^2}{n^2})\Big)
\end{aligned}
$$

对于充分大的n，$\xi\in[\frac{1}{2},2]$，从而引理得证。

由：

$$
\begin{aligned}
&P\Big\{\max_{1\le i\le n}|Y_i|^2-2logn+loglogn\le x\Big\} \\
&=\prod_{i=1}^{n}P\Big\{|Y_i|\le\sqrt{x+2logn-loglogn}\Big\} \\
&=\Big[1-2P\Big\{Y_1\ge\sqrt{x+2logn-loglogn}\Big\}\Big]^n
\end{aligned}
$$

其中，

$$
\begin{aligned}
&\frac{\sqrt{x+2logn-loglogn}}{\sqrt{2\pi}(1+x+2logn-loglogn)}e^{-\frac{1}{2}(x+2logn-loglogn)} \\
&\le P\Big\{Y_1\ge\sqrt{x+2logn-loglogn}\Big\} \\
&\le\frac{1}{\sqrt{2\pi}\sqrt{x+2logn-loglogn}}e^{-\frac{1}{2}(x+2logn-loglogn)}
\end{aligned}
$$

因此有：

$$
\lim_{n\to\infty}nP\Big\{Y_1\ge\sqrt{x+2logn-loglogn}\Big\}=\frac{1}{2\sqrt{\pi}}e^{-\frac{1}{2}x}
$$

由前面的引理即得所证。

评注：由此可以直接得出正态分布的最大次序统计量$Y_{(n)} = O_p(\sqrt{logn})$.

3. 矩阵的两种范数 若对于向量x，$|x|_2$表示其欧几里得模，$|x|_1$表示其各分量绝对值之和，$|x|_\infty$表示各分量绝对值的最大值。

那么对于$q \times p$维矩阵A：

- 算子模（谱模） $|A| : \sup_{x \in \mathbf{R}^p:\ |x|_2=1} |Ax|_2$
- l_1 模 $|A|_{l_1} = \big||A| \cdot 1\big|_\infty$，这里$|A|$表示矩阵A各分量取绝对值后的矩阵。

这里可以证明：$|Ax|_\infty \le |A|_{l_1}|x|_\infty$。

进一步地，若A为p维实对称矩阵，则显然有$|A|$等于A的特征值的绝对值中的最大者。

设这个特征值为λ，则由$A\xi = \lambda\xi$，两边取无穷范数即得$|A||\xi|_\infty = |A\xi|_\infty \le |A|_{l_1}|\xi|_\infty$，从而$|A| \le |A|_{l_1}$.

4. Berstein不等式 若随机变量$X_1, X_2, ..., X_n$独立，零均值

并且$\exists M,\ |X_i| \le M,\ i = 1, 2, ..., n,\ a.s.$

则

$$\forall t > 0,\ P\left\{\frac{1}{n}\sum_{i=1}^{n} X_i > t\right\} \le exp\left(-\frac{\frac{n^2}{2}t^2}{\sum_{i=1}^{n} E[X_i^2] + \frac{n}{3}Mt}\right)$$

5. 绝对值与示性函数不等式

$$\Big|\mathbf{1}\{X \le 0\} - \mathbf{1}\{X \le A\}\Big| \le 1\Big\{ -|A| < X \le |A|\Big\}$$

对于任意$A \in \mathbf{R},\ X \in \mathbf{R}$成立。

6. Borel - Cantelli引理

$$\sum_{n=1}^{\infty} P\{A_n\} < \infty \Rightarrow P\{A_n,\ i.o.\} = 0$$

7. Markov不等式

$$P\{|X| > C\} \le \frac{E|X|}{C}$$

对于任意随机变量X以及任意$C > 0$成立。

事实上，由这个不等式可以直接地得到$X = O_p\Big(\big(E|X|^r\big)^{\frac{1}{r}}\Big),\ \forall r > 0.$

8. Rosenthal不等式

$$E|S_n|^t \le C_t \cdot max\Big\{\sum_{i=1}^{n} E|\xi_i|^t,\ (\sum_{i=1}^{n} E\xi_i^2)^{\frac{t}{2}}\Big\}$$

其中，对于任意$t > 2$，

若随机变量$\xi_1, \xi_2, ..., \xi_n$独立零均值，并且t阶矩有限，并记$S_n = \sum_{i=1}^{n} \xi_i$

则上述不等式成立。

9. Lipschitz连续　若$\exists C > 0,\ \forall x, y \in I \subset \mathbf{R},\ |f(x) - f(y)| \le C|x - y|$，

则称f在集合I上Lipschitz连续，并将C称作f在I上的Lipschitz常数。

事实上，区间上的Lipschitz连续蕴含着一致连续和绝对连续。

10. 绝对值与示性函数积分不等式

$$|y| - |y - z| = \int_0^z 1 - 2 \cdot \mathbf{1}\{y \leq x\} dy$$

对于任意$y, z \in \mathbf{R}$成立。