分类号_____ 密级_____
UDC_____ 编号_____

華中師範大學

硕士学位论文

关于高维协方差矩阵迹的若干估计

学位申请人姓名：周雅诗

申请学位学生类别：全日制硕士

申请学位学科专业：概率论与数理统计

指导教师姓名：李正帮 副教授

硕士学位论文

关于高维协方差矩阵迹的若干估计

论文作者：周 雅 诗

指导教师：李正帮 副教授

学科专业：概率论与数理统计

研究方向：生物统计

华中师范大学数学与统计学学院

2021 年 5 月

硕士学位论文
MASTER'S THESIS

Some estimates of the traces of high-dimensional covariance matrices

A Thesis

Submitted in Partial Fulfillment of the Requirements

For the M.S.Degree in Mathematics

By

Yashi Zhou

Postgraduate Program

School of Mathematics and Statistics

Central China Normal University

Supervisor : **Zhengbang Li**

Academic Title **Professor** **Signature:** 李正邦

Approved

May 2021

华中师范大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的研究成果。除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

作者签名：周雅诗　　日期：2021年5月31日

学位论文版权使用授权书

学位论文作者完全了解华中师范大学有关保留、使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属华中师范大学。学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许学位论文被查阅和借阅：学校可以公布学位论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存、汇编学位论文。(保密的学位论文在解密后遵守此规定) 保密论文注释：本学位论文属于保密，在____年解密后适用本授权书。非保密论文注释：本学位论文不属于保密范围，适用本授权书。

作者签名：周雅诗　　导师签名：李正帮

日期：2021年5月31日　　日期：2021年6月1日

本人已经认真阅读“CALIS高校学位论文全文数据库发布章程”，同意将本人的学位论文提交“CALIS高校学位论文全文数据库”中全文发布，并可按“章程”中的规定享受相关权益。同意论文提交后滞后：□半年；□一年；□二年发布。

作者签名：周雅诗　　导师签名：李正帮

日期：2021年5月31日　　日期：2021年6月1日

摘 要

当今, 随着科学技术的高速发展, 高维数据渐渐出现在我们的视野里, 同时也在越来越多的领域中得以应用, 例如在生物学和金融研究等方面. 我们发现, 研究中的高维数据的一个共同特点就是数据的维数大于样本的大小, 这就是我们所熟知的“大 p, 小 n”, 而这容易出现“维数灾难”现象, 但“维数灾难”现象往往会导致高维协方差矩阵难以被估计. 事实上, 我们在处理高维数据时, 往往需要研究样本的高维协方差矩阵, 这就导致研究高维协方差矩阵的估计方法成为了一个重要的问题. 在本文, 我们将从Bai和Saranadasa [3]、Chen和Qin [4]和Li和Chen [5]中对高维协方差矩阵迹的三种估计方法进行研究, 并提出我们新的估计方法. 在这里, 我们研究的是一个样本的情况, 并且样本中的每个数据是独立同分布的, 我们主要通过理论研究四种估计方法的期望和方差. 除此之外, 我们将通过R语言进行数据模拟. 在协方差矩阵的结构是三种结构的基础上, 当高维数据服从正态分布、拉普拉斯分布时, 模拟高维协方差矩阵迹的期望与方差的真实值与估计值, 并通过模拟结果比较得出高维协方差矩阵迹的较优的估计方法.

关键词：高维数据; 独立同分布; 高维协方差矩阵的迹; 无偏估计; 一致估计.

Abstract

Nowadays, with the rapid development of science and technology, high dimensional data has gradually appeared in our field of vision, and it has also been applied in more and more fields, such as in biology and financial research. We found that a common feature of high-dimensional data in research is that the dimensionality of the data is greater than the size of the sample. This is what we know as "large p, small n" , which is prone to "dimensionality disaster" phenomenon, but "The dimensionality disaster" phenomenon often makes it difficult to estimate the high-dimensional covariance matrix. In fact, when we deal with high-dimensional data, we often need to study the high-dimensional covariance matrix of the sample, which leads to the study of the estimation method of the high-dimensional covariance matrix has become an important issue. In this article, we will study the three estimation methods of high-dimensional covariance matrix traces from Bai and Saranadasa, Chen and Qin and Li and Chen, and propose our new estimation method. Here, we are studying the situation of a sample, and each data in the sample is independent and identically distributed. We mainly study the expectations and variances of the four estimation methods through theory. In addition, we will use the R language to simulate the data. On the basis of the three structures of the covariance matrix, when the high-dimensional data obeys the normal distribution and Laplace distribution, the real value and estimated value of the expectation and variance of the trace of the high-dimensional covariance matrix are simulated, and through the comparison of simulation results, a better estimation method of the high-dimensional covariance matrix trace is obtained.

Keywords: High-dimensional data; Independent and identically distributed; The trace of the high-dimensional covariance matrix; Unbiased estimation; Consistent estimation.

目录

第一章 绪论

1.1 研究背景

随着现代信息技术的快速发展, 人们面对的信息越来越多元化、复杂化, 如何有效收集与分析数据以提取有用的信息和获得知识成为了一项重要的技能. 基于此, 人们对数据处理能力的要求也越来越高. 同时, 当我们在面对复杂化的数据库、金融行业的预测与决策、图像与信号处理等统计问题时, 往往伴随着维数高、结构复杂并且前后息息相关的数据, 因此, 高维数据渐渐被我们的关注, 同时也在各个领域中得以广泛应用, 例如在生物学、医学和金融研究等方面. 当我们在对高维数据进行分析时, 我们经常需要估计协方差矩阵这一个函数, 但在研究中, 我们发现高维数据的一个共同特点就是数据的维度往往高于样本的大小, 这就是我们所熟知的“大 p, 小 n” [6], 而这容易导致“维数灾难”的现象. 当这种情况出现时, 一些传统的检验方法, 例如 t 检验和Hotelling T^2 检验等, 在高维数据的情况下都已经不适用. 这是因为, 在高维数据中, 所有对象在许多方面看起来都是稀疏和不相似的, 在处理数据的过程中容易出现“维度诅咒”的现象, 这就导致无法使用一维和多维经典检验 [7]的数据处理方法. 下面我们来看一个例子:

例: 设样本 $x_1, x_2, \cdots, x_n$ 是 n 个服从正态分布 $N_p(\mu, \Sigma)$ 的 p 维随机变量, 且其均值为

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

记 $\overline{x}_i$ 和 μ_i 分别为 $\overline{x}$ 和 μ 的第 i 个分量, 那么

$$\overline{x}_i - \mu_i \sim N(0, \frac{1}{\sqrt{n}})$$

即

$$\overline{x}_i - \mu_i = O_p(\frac{1}{\sqrt{n}})$$

从而

$$\|\overline{x} - \mu\|^2 = \sum_{i=1}^{p}(\overline{x}_i - \mu_i)^2 = \sum_{i=1}^{p} O_p(\frac{1}{n}) = p \cdot O_p(\frac{1}{n})$$

其中, $O_p(\frac{1}{\sqrt{n}})$ 表示此后所有 n 的指数都大于等于 $1/2$ 的项之和, $\|\cdot\|$ 表示的是 R^P 上的范数. 显而易见, 当样本容量 n 是一个定值时, 样本均值 $\overline{x}$ 作为总体均值 μ 的估计将随着维数 p 的增大而变得越来越差. 因此, 要使在 p 增加的情形下, $\|\overline{x} - \mu\|^2 \to 0$ 仍然成立, 维数 p 增加的速度就要求要慢于或者等于样本大小 n 的

增加速度. 所以, 从上面的结论就可以看出高维数据会给统计推断带来一定的影响, 换句话说, 高维数据不能简单地用已有的一些经典的极限理论来检验. 同时, "维数灾难"这种现象往往会导致高维协方差矩阵难以被估计, 但协方差矩阵在多元统计分析中占有重要的地位, 并且协方差矩阵的估计问题在很多课题的研究中都被涉及且扮演重要的角色, 如在线性判别分析、主成分分析与聚类分析等, 这就使得研究高维协方差矩阵的估计方法变成了一个重要且亟待解决的问题.

当我们在研究协方差矩阵 Σ 的估计问题时 [10], 往往要研究 $tr\Sigma^2$ 的估计问题. 然而, 在许多问题中, 往往研究的都是高维数据的两样本检验问题, 比如: 两独立同分布的随机样本 $X_{11}, X_{12}, \cdots, X_{1N_1}$ 与 $X_{21}, X_{22}, \cdots, X_{2N_2}$, X_{ij} $(i = 1, 2,\ j = 1, 2, \cdots, N_i)$ 是 p 维随机向量, 其均值分别为 $\mu_1 = (\mu_{11}, \mu_{12}, \cdots, \mu_{1p})'$, $\mu_2 = (\mu_{21}, \mu_{22}, \cdots, \mu_{2p})'$, 协方差矩阵为 Σ_1, Σ_2.

我们考虑如下的假设检验问题:

$$H_0 : \mu_1 = \mu_2 \quad vs \quad H_1 : \mu_1 \neq \mu_2$$

著名的Hotelling's T^2 检验定义如下:

$$T^2 = \eta(\overline{X}_1 - \overline{X}_2)' S_n^{-1} (\overline{X}_1 - \overline{X}_2)$$

其中, $\overline{X}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} X_{ij}$ (i=1,2), $S_n = \sum_{i=1}^{2} \sum_{j=1}^{N_i} (X_{ij} - \overline{X}_i)(X_{ij} - \overline{X}_i)'$, $\eta = n \frac{N_1 N_2}{N_1 + N_2}$, 且 $n = N_1 + N_2 - 2$.

当 N_1 和 N_2 大于 p 时 [9], 我们可以较好的利用Hotelling's T^2 检验来解决问题, 但当 N_1 和 N_2 小于 p 时, Hotelling's T^2 检验方法已经不适用. 因此, 在假设检验 H_0 的条件下, Bai和Saranadasa [3]提出了新的检验方法, 并假设:

(a) $x_{ij} = \Gamma z_{ij} + \mu_j, i = 1, 2, \cdots, N_j, j = 1, 2$, Γ 是 $p \times m$ 矩阵, $\Gamma = [\Gamma_1, \cdots, \Gamma_p]'$, $\Gamma\Gamma' = \Sigma$, 正定矩阵 Σ 表示样本的协方差矩阵, z_{ij} 是独立同分布的随机 m 维向量, 满足$E(z_{ij}) = 0, Var(z_{ij}) = I, E(z_{ijk}^4) = 3 + \Delta < \infty$;

(b) $p/n \to c > 0$ 和 $N_1/(N_1 + N_2) \to \kappa \in (0, 1)$;

(c) $\mu'\Sigma\mu = o(\tau tr\Sigma^2)$ 和 $\lambda_{max} = o(\sqrt{tr\Sigma^2})$, $\tau = \frac{N_1 + N_2}{N_1 N_2}$, λ 是 Σ 的特征根.

Bai和Saranadasa [3]提出用 $\|\overline{X}_1 - \overline{X}_2\|^2$ 代替 $(\overline{X}_1 - \overline{X}_2)' S_n^{-1} (\overline{X}_1 - \overline{X}_2)$, 这也就是剔除了 S_n^{-1}, 其中 $\|\cdot\|$ 表示的是 R^P 上的范数, 考虑如下统计量:

$$M_n = (\overline{X}_1 - \overline{X}_2)'(\overline{X}_1 - \overline{X}_2) - \tau tr S_n$$

在假设 H_0 的条件下, $E(M_n) = 0$, 且如果满足条件(a)-(c), Bai和Saranadasa [3]证得了当 $n \to \infty$ 时, 有

$$Z_n = \frac{M_n}{\sqrt{Var(M_n)}} \to N(0, 1) \tag{1.1}$$

如果总体分布是正态分布, 且在假设 H_0 的条件下, 有

$$\sigma_M^2 = Var(M_n) = 2\tau^2(1+\frac{1}{n})tr\Sigma^2$$

如果总体分布不是正态分布, 但满足条件(a)-(c), 有

$$Var(M_n) = \sigma_M^2(1+o(1))$$

这就表明, 如果用 σ_M 代替 Z_n 的分母, (1.1)依然成立. 为了构建检验统计量, Bai和Saranadasa [3]提出了 $tr\Sigma^2$ 的估计方法, 并以此来代替 Z_n 的分母. 在条件(a)-(c)下, 这个改进有较好的功效, 但是对 n 和 p 取值的大小还是存在着限制. 为了解决超高维数据模拟检验问题, Chen和Qin [4]在样本独立的条件下建立了两样本高维数据的渐近正态的统计检验. 由于 $\sum_{j=1}^{N_i} X_{ij}'X_{ij}$, $i=1,2$在检验中起不到作用, 所以Chen和Qin [4]在Bai和Saranadasa [3]的基础上剔除 $\|\overline{X}_1 - \overline{X}_2\|^2$ 上的项 $\sum_{j=1}^{N_i} X_{ij}'X_{ij}$, $i=1,2$, 从而提出以下的统计量:

$$T_n =: \frac{\sum_{i\neq j}^{N_1} X_{1j}'X_{1j}}{N_1(N_1-1)} + \frac{\sum_{i\neq j}^{N_2} X_{2j}'X_{2j}}{N_2(N_2-1)} - 2\frac{\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} X_{1j}'X_{2j}}{N_1N_2}$$

其均值为 $E(T_n) = \|\mu_1-\mu_2\|^2$, 在 H_1 假设的条件下, 且

$$(\mu_1-\mu_2)'\Sigma_i(\mu_1-\mu_2) = o[n^{-1}tr\{(\Sigma_1+\Sigma_2)^2\}],\ \ i=1,2.$$

其方差为 $Var(T_n) = \{\frac{2}{N_1(N_1-1)}tr\Sigma_1^2 + \frac{2}{N_2(N_2-1)}tr\Sigma_2^2 + \frac{4}{N_1N_2}tr(\Sigma_1\Sigma_2)\}\{1+o(1)\}$. 因此, 当 $p\to\infty$ 与 $n\to\infty$ 时,

$$\frac{T_n - \|\mu_1-\mu_2\|^2}{\sqrt{Var(T_n)}} \xrightarrow{d} N(0,1) \tag{1.2}$$

为了基于(1.2)建立检验统计量, $Var(T_n)$ 需要被估计, 即需要估计 $tr\Sigma_i^2$ 与 $tr(\Sigma_i\Sigma_j)$, $(i,j=1,2,i\neq j)$. 因此, $tr\Sigma_i^2$ 与 $tr(\Sigma_i\Sigma_j)$ 的估计方法成为了我们主要的研究问题. 事实上, 有许多学者在这方面也做了很多的探索与研究, 除了我们提到的Bai和Saranadasa [3]、Chen和Qin [4], 其他学者也提出了好几种统计量来估计 $tr\Sigma^2$, 比如Schott [7], Srivastava和Fujikoshi [8], Chen和Li [10]以及Fujikoshi et al. [11]等.

在本文, 我们考虑的是一个高维样本的情况. 我们将对Bai和Saranadasa [3]、Chen和Qin [4]与Li和Chen [5]对 $tr\Sigma^2$ 提出的已有估计方法进行研究, 同时提出我们新的估计方法, 并逐个研究其统计性质. 除了在理论上的研究, 我们也将从数据模拟上来分析比较四种估计方法的优劣.

1.2 本文思路

在本文, 我们将研究一个样本的高维协方差矩阵迹的估计问题, 我们通过综合研究前辈对 $tr\Sigma^2$ 提出的估计方法及其统计性质, 并分析每一种估计方法的有关计算结果和数据模拟结果来找出更优的估计量, 以便我们在研究中找到参考.

接下来, 我们来说一说本文的主要思路安排, 如下:

- 第一节内容主要是叙述本课题的研究背景以及主要思路, 同时对本文应用到的某些定理定义作相关介绍, 为第二节内容的开展做铺垫.

- 第二节内容主要是计算四种估计方法的期望, 并得出其是无偏估计还是一致估计的结论, 同时计算四种估计方法的方差.

- 第三节内容主要是叙述我们的模拟条件以及思路, 同时利用统计软件R语言进行数据模拟, 主要模拟估计方法的期望和方差, 并对得到的模拟结果做分析, 通过直观的数据得出更优的估计量, 让文章更具有说服力.

- 第四节内容主要是本文的总结与展望, 通过前文的分析得出较优的估计方法, 同时对本文作了展望.

1.3 预备知识

在这一节, 我们主要介绍一下在本文中应用到的相关符号和相关定理定义.

定义1.1. *设 $X=(X_1,X_2,\cdots,X_n)'$ 为 n 维随机变量, 称矩阵*

$$C=(c_{ij})_{n\times n}=\begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix}$$

为 n 维随机向量 X 的协方差矩阵, 也记为 $D(X)$, 其中

$$c_{ij}=Cov(X_i,X_j), i,j=1,2,\cdots,n.$$

为 X 的分量 X_i 和 X_j 的协方差 (设它们都存在).

定义1.2. 设 $x=(x_1,\cdots,x_n)'$ 是一个 n 维随机向量，且 $E(x)=\mu$ （实向量）,$Var(x)=\Sigma$ （正定阵或非负定阵），假如它的特征函数为

$$f_x(t)=exp\{it'\mu-\frac{1}{2}t'\Sigma t\}$$

则称 x 为 n 元正态随机向量，其分布称为 n 元正态分布，记为 $x\sim N_n(\mu,\Sigma)$，而矩阵 Σ 的秩 $Rank(\Sigma)=r$ 称为这个分布的秩.

定义1.3. 设随机变量 ε 具有密度函数:

$$f(x)=\frac{1}{2\lambda}e^{-\frac{|x-\mu|}{\lambda}}$$

其中 λ,μ 为常数，且 $\lambda>0$，则称 ε 服从参数为 λ，μ 的拉普拉斯分布.

定义1.4. 设 $(\mathscr{X},\mathscr{B},\mathscr{P})$ 是一个统计结构，$T=T(x)$ 是从可测空间 $(\mathscr{X},\mathscr{B})$ 到 $(\mathscr{T},\mathscr{L})$ 的一个可测映照，假若这个映照 T 不依赖于分布族 $\mathscr{P}$，则称 T 为此结构上的统计量，假如 $\mathscr{P}$ 为参数分布族 $\{P_\theta:\theta\in\Theta\}$，则不依赖于参数 θ 的可测映照 T 称为此结构上的统计量.

定义1.5. 在统计中常把要检验的假设称为原假设，记为 H_0; 在拒绝时所接受的假设称为备择假设，记为 H_1.

双边假设　　$H_0:\mu=\mu_0$　　$H_1:\mu\neq\mu_0$

定义1.6. 定义在 $\mathscr{P}$ 上的一个实值泛函 $g(P)$ 称为参数，而 $(\mathscr{X},\mathscr{B},\mathscr{P})$ 上的用来估计 $g(P)$ 的实值统计量称为 $g(P)$ 的点估计量，简称估计.

定义1.7. 设 $(\mathscr{X},\mathscr{B},\{P_\theta,\theta\in\Theta\})$ 为可控参数统计结构，$g(\theta)$ 是未知参数，$x=(x_1,x_2,\cdots,x_n)$ 是来自该统计结构的一个样本，若用 $\widehat{g}(x)$ 估计 $g(\theta)$，且

$$E_\theta(\widehat{g}(x))=g(\theta),\ \forall\,\theta\in\Theta$$

则称 $\widehat{g}(x)$ 为 $g(\theta)$ 的无偏估计.

定义1.8. 设 $\widehat{g}_n=\widehat{g}(X_1,X_2,\cdots,X_n)$ 为 $g(\theta)$ 的基于样本的 $(X_1,X_2,\cdots,X_n)'$ 的一个估计量，若对任意固定的 $\theta\in X$，都满足: 对于任给的 $\varepsilon>0$，有

$$\lim_{n\to\infty}P_\theta|\widehat{g}_n-g(\theta)|>\varepsilon=0$$

成立，则称 $\widehat{g}_n$ 为 $g(\theta)$ 的一致估计量，上述极限简记为 $\widehat{g}_n\xrightarrow{p}g(\theta)$

定理1.1 (柯西-施瓦茨不等式). 若 $a_1, a_2, \cdots, a_n$ 和 $b_1, b_2, \cdots, b_n$ 是任意实数, 则有

$$(\sum_{i=1}^{n} a_i b_i)^2 \leq (\sum_{i=1}^{n} a_i)^2 (\sum_{i=1}^{n} b_i)^2$$

此外, 如果有某个 $a_i \neq 0$, 则上式中的等号当且仅当存在一个实数 x, 使得对每一个 $i = 1, 2, \cdots, n$ 都有 $a_i x + b_i = 0$ 时成立.

定义1.9. 交叉验证, 是由*Seymour Geisser*提出来的. 它有时候也称为循环估计, 在统计学上是一种将数据样本切割成较小子集的实用方法. 该理论在给定的建模样本中, 拿出大部分样本进行建立模型, 留小部分样本用刚建立的模型进行预报, 并求这小部分样本的预报误差, 同时记录它们的平方加和. 这个过程一直进行, 直到所有的样本都被预报了一次而且仅被预报一次. 我们把每个样本的预报误差平方加和, 称为*PRESS(predicted Error Sum of Squares)*.

第二章 $tr\Sigma^2$ 估计方法的统计性质研究

2.1 条件与符号

在本文中, 我们研究的都是在只有一个样本的情况下. 下面我们介绍相关的已知条件与符号:

令 $x_1, x_2, \cdots, x_n$ 是 n 个独立同分布的 p 维随机向量, 并且每一个 x_i 可以表示为:

$$x_i = \Gamma z_i + \mu, i = 1, 2, \cdots, n.$$

这里, 对于 $m \geq p$, Γ 是 $p \times m$ 矩阵, $\Gamma = [\Gamma_1, \cdots, \Gamma_p]'$, $\Gamma\Gamma' = \Sigma$, 正定矩阵 Σ 表示样本的协方差矩阵, z_i 是独立同分布的随机 m 维向量, 满足$E(z_i) = 0, Var(z_i) = I$, 由 $\Gamma z_i = x_i - \mu$, 可知 $E[(\Gamma z_i)(\Gamma z_i)'] = E[(x_i - \mu)(x_i - \mu)']$, 又因为 $E[(\Gamma z_i)(\Gamma z_i)'] = \Sigma$, 故样本协方差矩阵也可以表示为:

$$\Sigma = E[(x_i - \mu)(x_i - \mu)']$$

并且, 如果 $z_i = (z_{i1}, z_{i2}, \cdots, z_{im})'$, 对一个正数 q 满足 $\sum_{l=1}^{q} \alpha_l \leq 8$ 且 $l_1 \neq l_2 \neq l_q$, 我们假设 $E(z_{ik}^4) = 3 + \Delta < \infty$, Δ 表示常数, 且

$$E(z_{il_1}^{\alpha_1} z_{il_2}^{\alpha_2} \cdots z_{il_q}^{\alpha_q}) = E(z_{il_1}^{\alpha_1}) E(z_{il_2}^{\alpha_2}) \cdots E(z_{il_q}^{\alpha_q})$$

除此之外, 我们定义如下的 κ_{ij} [14], 即

$$\kappa_{ij} = E[z'\Sigma^i z z'\Sigma^j z] - 2tr\Sigma^{i+j} - tr\Sigma^i tr\Sigma^j$$

同时, 我们定义如下的渐近参数 $W1$ 和假设 $W2$:

$$\begin{aligned} W1 &: \ n/p \to c \in (0, 1); \\ W2 &: \ a_i := tr\Sigma^i/p \to a_i^0 \in (0, \infty), i = 1, 2, 3, 4. \end{aligned}$$

以下的计算基于上述这些条件.

2.2 已有估计方法的统计性质

在本节中, 我们将研究 $tr\Sigma^2$ 的已有估计方法及其性质, 即研究Bai和Saranadasa [3]、Chen和Qin [4] 、Li和Chen [5]对 $tr\Sigma^2$ 提出的估计方法, 包括研究其期望与方差, 同时判断出已有的估计方法是一致估计, 还是无偏估计.

2.2.1 Bai和Saranadasa估计方法及其性质

当我们考虑 $tr\Sigma^2$ 的估计时, 我们自然而然会认为 $tr\Sigma^2$ 的估计就是 trS^2, 其中 $S=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(x_i-\overline{x})'$. 然而, 不像在正态假设的情况下, 当 p 固定时, trS^2 既不是无偏的也不是渐近一致的. 因此, 我们经常会用 trS^2 和 $(trS)^2$ 来表示 $tr\Sigma^2$ 的估计. Bai 和Saranadasa [3]提出了有关于用 trS^2 和 $(trS)^2$ 来表示的统计量来估计 $tr\Sigma^2$. 显然, 根据柯西-施瓦茨不等式, $trS^2-\frac{1}{n-1}(trS)^2\geq 0$. 在一个样本的情况下, Bai和Saranadasa [3]提的估计方法如下:

$$\widehat{tr\Sigma_1^2}=\frac{(n-1)^2}{(n+1)(n-2)}\{trS^2-\frac{1}{n-1}(trS)^2\} \tag{2.1}$$

除此之外, 当 z_i 服从 F 分布且 F 分布是正态分布时, Srivastava [9]提出了 $tr\Sigma^2/p(=:a_2)$ 的无偏和一致估计 $\widetilde{a}_2$, 即

$$\widetilde{a}_2=\frac{(n-1)^2}{(n+1)(n-2)}\frac{1}{p}\{trS^2-\frac{1}{n-1}(trS)^2\}$$

随后, Srivastava [9]证明了虽然 F 分布不是正态分布, 但在有很强条件的情况下估计 $\widetilde{a}_2$ 是 a_2 的一致估计, 而不是无偏估计. 对于Bai和Saranadasa [3]提出的估计, 由于当 $n\to\infty$ 时,

$$(n-1)^2/[(n+1)(n-2)]\to 1$$

所以, 我们不妨不考虑统计量(2.1)的系数 $(n-1)^2/[(n+1)(n-2)]$, 从而提出我们的第一种估计, 记为 T_1, 即

$$T_1=trS^2-\frac{1}{n-1}(trS)^2 \tag{2.2}$$

由Srivastava [9], 我们猜想: 当我们不考虑 F 的分布情况时, T_1 是 $tr\Sigma^2$ 的一致估计.

当我们在研究统计量的统计性质时, 我们一般会研究估计量的期望及方差, 因此, 接下来我们主要来讨论第一种估计方法的期望与方差. 对于(2.2), 已知

$$\begin{aligned}\overline{x} &= \frac{1}{n}\sum_{i=1}^{n}x_i\\ S &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(x_i-\overline{x})'\end{aligned}$$

从而得到

$$\begin{aligned}T_1=&\frac{1}{4n^2(n-1)^2}\sum_{i_1,i_2,j_1,j_2}^{\bullet}(x_{i_1}-x_{j_1})'(x_{i_2}-x_{j_2})(x_{i_2}-x_{j_2})'(x_{i_1}-x_{j_1})\\ &-\frac{1}{4n^2(n-1)^3}\sum_{k_1,k_2,l_1,l_2}^{\bullet}(x_{k_1}-x_{l_1})'(x_{k_1}-x_{l_1})(x_{k_2}-x_{l_2})'(x_{k_2}-x_{l_2})\end{aligned} \tag{2.3}$$

其中, $\sum^{\bullet}_{i_1,i_2,j_1,j_2}$ 表示的是 $\{(i_1,i_2,j_1,j_2): i_1 \neq j_1, i_2 \neq j_2\}$ 情况下的指数之和.

不妨令 $T_1 = tr(A1)+tr(B1)$, $tr(A1)$ 代表(2.3)的第一项, $tr(B1)$ 代表(2.3)第二项. 又因为 $\kappa_{11} = E[z'\Sigma z z'\Sigma z] - 2tr\Sigma^2 - (tr\Sigma)^2$, 从而得到以下的引理.

引理 2.1. *如果 $x_1, x_2, \cdots, x_n$ 独立同分布, 且 $\Sigma = E[(x_i-\mu)(x_i-\mu)']$, 则*

$$E[tr(A1)] = \frac{n}{n-1}tr\Sigma^2 + \frac{1}{n-1}(tr\Sigma)^2 + \frac{1}{n}\kappa_{11} \tag{2.4}$$

$$E(tr(B1)) = -\frac{2}{(n-1)^2}tr\Sigma^2 - \frac{1}{n-1}(tr\Sigma)^2 - \frac{1}{n(n-1)}\kappa_{11} \tag{2.5}$$

证明: 由 $S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(x_i-\overline{x})'$, 得

$$S = \frac{1}{2n(n-1)}\sum_{i\neq j}^{n}(x_i-x_j)(x_i-x_j)'$$

从而得到

$$S^2 = \frac{1}{4n^2(n-1)^2}\sum^{\bullet}_{i_1,i_2,j_1,j_2}(x_{i_1}-x_{j_1})(x_{i_1}-x_{j_1})'(x_{i_2}-x_{j_2})(x_{i_2}-x_{j_2})'$$

$$(trS)^2 = \frac{1}{4n^2(n-1)^2}\sum^{\bullet}_{i_1,i_2,j_1,j_2}(x_{i_1}-x_{j_1})'(x_{i_1}-x_{j_1})(x_{i_2}-x_{j_2})'(x_{i_2}-x_{j_2})$$

其中, $\sum^{\bullet}_{i_1,i_2,j_1,j_2}$ 表示的是 $\{(i_1,i_2,j_1,j_2): i_1 \neq j_1, i_2 \neq j_2\}$ 情况下的指数之和.
则

$$tr(A1) = trS^2 = \frac{1}{4n^2(n-1)^2}\sum^{\bullet}_{i_1,i_2,j_1,j_2}(x_{i_1}-x_{j_1})'(x_{i_2}-x_{j_2})(x_{i_2}-x_{j_2})'(x_{i_1}-x_{j_1})$$

$$tr(B1) = \frac{1}{n-1}(trS)^2 = \frac{1}{4n^2(n-1)^3}\sum^{\bullet}_{k_1,k_2,l_1,l_2}(x_{k_1}-x_{l_1})'(x_{k_1}-x_{l_1})(x_{k_2}-x_{l_2})'(x_{k_2}-x_{l_2})$$

对于 $tr(A1)$, 因为

$$\begin{aligned}(x_{i_1}-x_{j_1})'(x_{i_2}-x_{j_2}) =& (x_{i_1}-\mu)'(x_{i_2}-\mu) - (x_{i_1}-\mu)'(x_{j_2}-\mu) - (x_{j_1}-\mu)'(x_{i_2}-\mu)\\ &+ (x_{j_1}-\mu)'(x_{j_2}-\mu)\\ (x_{i_2}-x_{j_2})'(x_{i_1}-x_{j_1}) =& (x_{i_2}-\mu)'(x_{i_1}-\mu) - (x_{i_2}-\mu)'(x_{j_1}-\mu) - (x_{j_2}-\mu)'(x_{i_1}-\mu)\\ &+ (x_{j_2}-\mu)'(x_{j_1}-\mu)\end{aligned}$$

所以

$$\begin{aligned}
&\dot{\sum_{i_1,i_2,j_1,j_2}}(x_{i_1}-x_{j_1})'(x_{i_2}-x_{j_2})(x_{i_2}-x_{j_2})'(x_{i_1}-x_{j_1})\\
=&4\dot{\sum_{i_1,i_2,j_1,j_2}}(x_{i_1}-\mu)'(x_{i_2}-\mu)(x_{i_2}-\mu)'(x_{i_1}-\mu)\\
&-8\dot{\sum_{i_1,i_2,j_1,j_2}}(x_{i_1}-\mu)'(x_{i_2}-\mu)(x_{j_2}-\mu)'(x_{i_1}-\mu)\\
&+4\dot{\sum_{i_1,i_2,j_1,j_2}}(x_{i_1}-\mu)'(x_{i_2}-\mu)(x_{j_2}-\mu)'(x_{j_1}-\mu)
\end{aligned}$$

从而

$$\begin{aligned}
tr(A1)=&\frac{1}{n^2(n-1)^2}\dot{\sum_{i_1,i_2,j_1,j_2}}(x_{i_1}-\mu)'(x_{i_2}-\mu)(x_{i_2}-\mu)'(x_{i_1}-\mu)\\
&-\frac{2}{n^2(n-1)^2}\dot{\sum_{i_1,i_2,j_1,j_2}}(x_{i_1}-\mu)'(x_{i_2}-\mu)(x_{j_2}-\mu)'(x_{i_1}-\mu)\\
&+\frac{1}{n^2(n-1)^2}\dot{\sum_{i_1,i_2,j_1,j_2}}(x_{i_1}-\mu)'(x_{i_2}-\mu)(x_{j_2}-\mu)'(x_{j_1}-\mu)\\
=&:tr(A1)_1+tr(A1)_2+tr(A1)_3
\end{aligned}\tag{2.6}$$

对于 $tr(A1)_1$, 当 $i_1=i_2$ 时

$$E[(x_{i_1}-\mu)'(x_{i_2}-\mu)(x_{i_2}-\mu)'(x_{i_1}-\mu)]=E[(x_1-\mu)'(x_1-\mu)]^2$$

当 $i_1\neq i_2$ 时

$$E[(x_{i_1}-\mu)'(x_{i_2}-\mu)(x_{i_2}-\mu)'(x_{i_1}-\mu)]=tr\Sigma^2$$

对于 $tr(A1)_2$, 因为 $i_2\neq j_2$, 所以 $E[(x_{i_1}-\mu)'(x_{i_2}-\mu)(x_{j_2}-\mu)'(x_{i_1}-\mu)]=0$
对于 $tr(A1)_3$, 当 $i_1=i_2$, $j_1=j_2$ 时

$$E[(x_{i_1}-\mu)'(x_{i_2}-\mu)(x_{j_2}-\mu)'(x_{j_1}-\mu)]=(tr\Sigma)^2$$

当 $i_1=j_2$, $j_1=i_2$ 时

$$E[(x_{i_1}-\mu)'(x_{i_2}-\mu)(x_{j_2}-\mu)'(x_{j_1}-\mu)]=tr\Sigma^2$$

故

$$E[tr(A1)_1] = \frac{n-1}{n}tr\Sigma^2+\frac{1}{n}E[(x_1-\mu)'(x_1-\mu)]^2$$

$$E[tr(A1)_2] = 0$$
$$E[tr(A1)_3] = \frac{1}{n(n-1)}tr\Sigma^2 + \frac{1}{n(n-1)}(tr\Sigma)^2$$

又因为 $\kappa_{11} = E[z'\Sigma zz'\Sigma z] - 2tr\Sigma^2 - (tr\Sigma)^2$, 因此, 我们可以得到

$$E[tr(A1)] = \frac{n}{n-1}tr(\Sigma^2) + \frac{1}{n-1}(tr\Sigma)^2 + \frac{1}{n}\kappa_{11}$$

同理, 我们也可以得到 $trB1$ 的期望, 即

$$E[tr(B1)] = -\frac{2}{(n-1)^2}tr(\Sigma^2) - \frac{1}{n-1}(tr\Sigma)^2 - \frac{1}{n(n-1)}\kappa_{11}$$

□

通过观察(2.4)与(2.5), 我们发现 $tr(A1)$ 与 $tr(B1)$ 的期望都与样本均值 μ 无关, 且项 $(tr\Sigma)^2$ 的系数相反, 因此, T_1 的期望也与 $(tr\Sigma)^2$ 无关.

定理2.1. *基于引理(2.1), 已知 $tr(A1)$ 与 $tr(B1)$ 的期望, 则*

$$E(T_1) = \frac{n(n-1)-2}{(n-1)^2}tr\Sigma^2 - \frac{n-2}{n(n-1)}\kappa_{11} \tag{2.7}$$

显然, T_1 是 $tr\Sigma^2$ 的一致估计.

证明: 在引理中, 我们得到了 $tr(A1)$ 与 $tr(B1)$ 的期望, 且由于

$$E(T_1) = E[tr(A1)] + E[tr(B1)]$$

所以, 我们很容易得到 T_1 的期望. 从Himeno和Yamada [17], 我们知道 $\kappa_{11} = o(p^3)$, 且当 $n \to \infty$ 时

$$\frac{n(n-1)-2}{(n-1)^2} \to 1, \quad \frac{n-2}{n(n-1)} \to 0$$

因此

$$E(T_1) \xrightarrow{p} tr\Sigma^2$$

故, T_1 是 $tr\Sigma^2$ 的一致估计. □

通过观察我们发现, T_1 的期望与样本均值 μ 无关, 也就是说无论 μ 是否为 0, 都不影响 T_1 的期望. 显然, 对于 $E(T_1)$, 我们可以得到:

$$E(T_1) = tr\Sigma^2 + o(1)$$

即

$$E(T_1) \xrightarrow{p} tr\Sigma^2$$

从而说明 T_1 是 $tr\Sigma^2$ 的一致估计, 而不是无偏估计. 下面我们来讨论估计 T_1 的方差, 根据(2.6)我们知道了 $tr(A1)$ 的表达式, 因此得到以下推论.

推论2.2. 已知 $x_1, x_2, \cdots, x_n$ 独立同分布, $x_i = \Gamma z_i + \mu$, $i = 1, 2, \cdots, n$, 且 $\Sigma = \Gamma\Gamma' = E[(x_i - \mu)(x_i - \mu)']$, 则

$$\begin{aligned}
E[tr^2(A1)] =& \frac{[(n-1)^4 + 12(n-1)^2 - 32(n-1) + 60](n-2)(n-3)}{n^3(n-1)^3} tr^2\Sigma^2 \\
&+ \frac{[6(n-1)^4 + 60(n-1)^2 - 96(n-1) + 48](n-2)}{n^3(n-1)^3} E[(z_1'\Sigma^2 z_1)^2] \\
&- \frac{[24(n-1)^3 - 48(n-1)^2 + 112(n-1) - 96](n-2)}{n^3(n-1)^3} E[(z_1'\Sigma z_2)^2 z_1'\Sigma^2 z_2] \\
&+ \frac{4(n-1)^2 + 24}{n^3(n-1)} E[(z_1'\Sigma z_1)^2 z_1'\Sigma^2 z_1] \\
&- \frac{24(n-1)^2 + 32}{n^3(n-1)^2} E[(z_1'\Sigma z_1)^2 z_1'\Sigma z_2 z_2'\Sigma z_2] \\
&+ \frac{3(n-1)^4 + 24(n-1)^2 + 8}{n^3(n-1)^3} E[(z_1'\Sigma z_2)^4] \\
&+ \frac{1}{n^3} E[(z_1'\Sigma z_1)^4]
\end{aligned}$$

证明: 我们要求 $E[tr^2(A1)]$, 就要求出每一个 $E[tr^2(A1)_i]$ 与 $E[tr(A1)_i tr(A1)_j]$, 其中 $i, j = 1, 2, 3$ 且 $i \neq j$.

由

$$tr(A1)_1 = \frac{1}{n^2(n-1)^2} \dot{\sum_{i_1, i_2, j_1, j_2}} (x_{i_1} - \mu)'(x_{i_2} - \mu)(x_{i_2} - \mu)'(x_{i_1} - \mu)$$

得

$$\begin{aligned}
E[tr^2(A1)_1] =& \frac{1}{n^4(n-1)^4} E\{[\dot{\sum_{i_{11}, i_{21}, j_{11}, j_{21}}} (x_{i_{11}} - \mu)'(x_{i_{21}} - \mu)(x_{i_{21}} - \mu)'(x_{i_{11}} - \mu)] \cdot \\
& [\dot{\sum_{i_{12}, i_{22}, j_{12}, j_{22}}} (x_{i_{12}} - \mu)'(x_{i_{22}} - \mu)(x_{i_{22}} - \mu)'(x_{i_{12}} - \mu)]\}
\end{aligned}$$

经过对 $i_{11}, i_{21}, j_{11}, j_{21}, i_{12}, i_{22}, j_{12}, j_{22}$ 的分类讨论, 并通过排列组合, 我们得到了 $E[tr^2(A1)_1]$, 即

$$\begin{aligned}
E[tr^2(A1)_1] =& \frac{(n-1)(n-2)(n-3)}{n^3} tr^2\Sigma^2 + \frac{1}{n^3} E[(z_1'\Sigma z_1)^4] \\
&+ \frac{3(n-1)}{n^3} E[(z_1'\Sigma z_2)^4] + \frac{4(n-1)}{n^3} E[(z_1'\Sigma z_1)^2 z_1'\Sigma^2 z_1] \\
&+ \frac{6(n-1)(n-2)}{n^3} E[(z_1'\Sigma^2 z_1)^2]
\end{aligned}$$

同理

$$\begin{aligned}
E[tr^2(A1)_2] &= \frac{8(n-2)(n-3)}{n^3(n-1)}tr^2\Sigma^2+\frac{16}{n^3(n-1)}E[(z_1'\Sigma z_2)^4]\\
&+\frac{16}{n^3(n-1)}E[(z_1'\Sigma z_1)^2z_1'\Sigma^2z_1]\\
&+\frac{32(n-2)}{n^3(n-1)}E[(z_1'\Sigma z_2)^2z_1'\Sigma^2z_2]\\
&+\frac{40(n-2)}{n^3(n-1)}E[(z_1'\Sigma^2z_1)^2]\\
E[tr^2(A1)_3] &= \frac{60(n-2)(n-3)}{n^3(n-1)^3}tr^2\Sigma^2+\frac{8}{n^3(n-1)^3}E[(z_1'\Sigma z_2)^4]\\
&+\frac{96(n-2)}{n^3(n-1)^3}E[(z_1'\Sigma z_2)^2z_1'\Sigma^2z_2]\\
&+\frac{48(n-2)}{n^3(n-1)^3}E[(z_1'\Sigma^2z_1)^2]\\
E[tr(A1)_1tr(A1)_2] &= \frac{6}{n^3}E[(z_1'\Sigma z_1)^2z_1'\Sigma z_2z_2'\Sigma z_2]\\
&+\frac{6(n-2)}{n^3}E[(z_1'\Sigma z_2)^2z_1'\Sigma^2z_2]\\
E[tr(A1)_1tr(A1)_3] &= \frac{2(n-2)(n-3)}{n^3(n-1)}tr^2\Sigma^2+\frac{4}{n^3(n-1)}E[(z_1'\Sigma z_2)^4]\\
&+\frac{4}{n^3(n-1)}E[(z_1'\Sigma z_1)^2z_1'\Sigma^2z_1]\\
&+\frac{8(n-2)}{n^3(n-1)}E[(z_1'\Sigma z_2)^2z_1'\Sigma^2z_2]\\
&+\frac{10(n-2)}{n^3(n-1)}E[(z_1'\Sigma^2z_1)^2]\\
E[tr(A1)_2tr(A1)_3] &= \frac{8(n-2)(n-3)}{n^3(n-1)^2}tr^2\Sigma^2\\
&+\frac{28(n-2)}{n^3(n-1)^2}E[(z_1'\Sigma z_2)^2z_1'\Sigma^2z_2]\\
&+\frac{8}{n^3(n-1)^2}E[(z_1'\Sigma z_1)^2z_1'\Sigma z_2z_2'\Sigma z_2]\\
&+\frac{24(n-2)}{n^3(n-1)^2}E[(z_1'\Sigma^2z_1)^2]
\end{aligned}$$

从而, 我们可以得到 $E[tr^2(A1)]$. □

我们观察 $tr(A1)$ 和 $tr(B1)$, 并且通过验证, 我们可得到:

$$E[tr^2(B1)] = \frac{1}{(n-1)^2}E[tr^2(A1)]$$

$$E[tr(A1)tr(B1)] = \frac{1}{n-1}E[tr^2(A1)]$$

因此

$$\begin{aligned} E(T_1^2) &= E[tr^2(A1)] + E[tr^2(B1)] - 2E[tr(A1)tr(B1)] \\ &= \frac{(n-2)^2}{(n-1)^2}E[tr^2(A1)] \end{aligned}$$

而我们又已知了 $E[tr^2(A1)]$, 并且 $E[(z_1'\Sigma^2 z_1)^2] = \kappa_{22} + 2tr\Sigma^4 + tr^2\Sigma^2$, 所以我们得到了以下推论.

推论2.3. *已知 $x_1, x_2, \cdots, x_n$ 独立同分布, $x_i = \Gamma z_i + \mu$, $i = 1, 2, \cdots, n$, 且 $\Sigma = \Gamma\Gamma' = E[(x_i - \mu)(x_i - \mu)']$, $E[(z_1'\Sigma^2 z_1)^2] = \kappa_{22} + 2tr\Sigma^4 + tr^2\Sigma^2$, 则*

$$\begin{aligned} E(T_1^2) =& \frac{[(n-1)^4(n+3) + 12(n+2)(n-1)^2 - 32n(n-1) + 60(n-3) + 48](n-2)^3}{n^3(n-1)^5}tr^2\Sigma^2 \\ &+ \frac{[6(n-1)^4 + 60(n-1)^2 - 96(n-1) + 48](n-2)^3}{n^3(n-1)^5}\kappa_{22} \\ &+ \frac{2[6(n-1)^4 + 60(n-1)^2 - 96(n-1) + 48](n-2)^3}{n^3(n-1)^5}tr\Sigma^4 \\ &- \frac{[24(n-1)^3 - 48(n-1)^2 + 112(n-1) - 96](n-2)^3}{n^3(n-1)^5}E[(z_1'\Sigma z_2)^2 z_1'\Sigma^2 z_2] \\ &+ \frac{[4(n-1)^2 + 24](n-2)^2}{n^3(n-1)^3}E[(z_1'\Sigma z_1)^2 z_1'\Sigma^2 z_1] \\ &- \frac{[24(n-1)^2 + 32](n-2)^2}{n^3(n-1)^4}E[(z_1'\Sigma z_1)^2 z_1'\Sigma z_2 z_2'\Sigma z_2] \\ &+ \frac{[3(n-1)^4 + 24(n-1)^2 + 8](n-2)^2}{n^3(n-1)^5}E[(z_1'\Sigma z_2)^4] \\ &+ \frac{(n-2)^2}{n^3(n-1)^2}E[(z_1'\Sigma z_1)^4] \end{aligned}$$

定理2.2. *基于推论(2.3), 且 $tr^2\Sigma^2 = p^2a_2^2$, $tr\Sigma^4 = pa_4$, $n/p \to c \in (0,1)$, 则有*

$$Var(T_1) = (-\frac{8}{n} + \frac{37}{n^2})p^2a_2^2 + (\frac{12}{n} - \frac{72}{n^2})pa_4 + o(p^5) \tag{2.8}$$

证明： 因为 $E(T_1)=\frac{n(n-1)-2}{(n-1)^2}tr\Sigma^2-\frac{n-2}{n-1}\kappa_{11}$, 所以

$$\begin{aligned}(ET_1)^2=&\frac{n^2(n-1)^2-4n(n-1)+4}{(n-1)^4}tr^2\Sigma^2+\frac{(n-2)^2}{(n-1)^2}\kappa_{11}^2\\&-\frac{2[n(n-1)-2](n-2)}{(n-1)^3}(tr\Sigma^2)\kappa_{11}\end{aligned}$$

并且从Himeno和Yamada [17]里, 我们可以知道:

$$\begin{aligned}\kappa_{11}&=o(p^3)\\\kappa_{22}&=o(p^3)\\E[(z_1'\Sigma z_2)^2z_1'\Sigma^2z_2]&=o(p^{5/2})\\E[(z_1'\Sigma z_2)^4]&=o(p^4)\end{aligned}$$

因此

$$\begin{aligned}E[(z_1'\Sigma z_1)^2z_1'\Sigma^2z_1]&=o(p^{9/2})\\\kappa_{11}tr\Sigma^2&=o(p^4)\end{aligned}$$

我们利用柯西-施瓦茨不等式来证明 $E[(z_1'\Sigma z_1)^2z_1'\Sigma^2z_1]=o(p^{9/2})$:

$$\begin{aligned}E[|(z_1'\Sigma z_1)^2z_1'\Sigma^2z_1|]&=\sqrt{E[(z_1'\Sigma z_1)^4]E[z_1'\Sigma^2z_1]^2}\\&\leq\sqrt{[\kappa_{11}+2tr\Sigma^2+(tr\Sigma)^2]^2(\kappa_{22}+2tr\Sigma^4+tr^2\Sigma^2)}\\&=o(p^{9/2})\end{aligned}$$

因为 $tr^2\Sigma^2=p^2a_2^2$, $tr\Sigma^4=pa_4$, $n/p\to c\in(0,1)$, 故估计量 T_1 的方差为:

$$Var(T_1)=(-\frac{8}{n}+\frac{37}{n^2})p^2a_2^2+(\frac{12}{n}-\frac{72}{n^2})pa_4+o(p^5)$$

□

2.2.2 Chen和Qin估计方法及其性质

在这一节, 我们来讨论 $tr\Sigma^2$ 的第二种已有的估计方法. 受剔除 $\|\overline{X}_1-\overline{X}_2\|^2$ 上的项 $\sum_{j=1}^{N_i}X_{ij}'X_{ij}$, $i=1,2$ 的影响, Chen和Qin [4]采用的是剔除样本 $x_1,x_2,\cdots,x_n$ 中任意两个高维数据的方法, 即剔除项 x_i 与项 x_j $(i\neq j)$. 这种剔除方法类似于交叉验证思想. 基于这个思想, Chen和Qin [4]构建了 x_i 和 x_j 样本均值的偏差, 即为了估计 $tr\Sigma^2$ 提出了统计量 $\widehat{tr\Sigma_2^2}$, 不妨记为 T_2, 即

$$T_2=\frac{1}{n(n-1)}tr\{\sum_{i\neq j}^{n}(x_i-\overline{x}_{(i,j)})x_i'(x_j-\overline{x}_{(i,j)})x_j'\}\tag{2.9}$$

其中 $\overline{x}_{(i,j)}$ 表示的是除了 x_i 和 x_j 的样本均值.

由(2.9), 我们可以把 T_2 整理得到下面的式子.

$$\begin{aligned}
T_2 =& \frac{1}{n(n-1)}\sum_{i\neq j}^{n}\{(x_i-\mu)'(x_j-\mu)(x_j-\mu)'(x_i-\mu) \\
&-2(x_i-\mu)'(x_j-\mu)(x_j-\mu)'(\overline{x}_{(i,j)}-\mu) \\
&+2(x_i-\mu)'(x_j-\mu)\mu'(x_i-\mu) \\
&-2(x_i-\mu)'(x_j-\mu)\mu'(\overline{x}_{(i,j)}-\mu) \\
&+2(x_i-\mu)'(\overline{x}_{(i,j)}-\mu)\mu'(\overline{x}_{(i,j)}-\mu) \\
&-2\mu'(x_j-\mu)\mu'(\overline{x}_{(i,j)}-\mu) \\
&+(x_i-\mu)'(\overline{x}_{(i,j)}-\mu)(x_j-\mu)'(\overline{x}_{(i,j)}-\mu) \\
&+\mu'(x_j-\mu)\mu'(x_i-\mu) \\
&+\mu'(\overline{x}_{(i,j)}-\mu)\mu'(\overline{x}_{(i,j)})\} \\
=:& \sum_{k=1}^{10} tr(A2)_k
\end{aligned} \tag{2.10}$$

我们很容易得到$E[tr(A2)_1] = tr\Sigma^2$, $E[tr(A2)_{10}] = \frac{1}{n-2}\mu'\Sigma\mu$, 而当 $k = 2, 3, \cdots, 9$ 时, $E[tr(A2)_k] = 0$. 因此我们得到 T_2 的均值.

定理2.3. *已知 $x_1, x_2, \cdots, x_n$ 独立同分布, 且 $\Sigma = E[(x_i-\mu)(x_i-\mu)']$, 则*

$$E(T_2) = tr\Sigma^2 + \frac{1}{n-2}\mu'\Sigma\mu \tag{2.11}$$

显然, T_2 是 $tr\Sigma^2$ 的一致估计.

证明: 因为$(x_i - \overline{x}_{(i,j)})x_i' = [(x_i-\mu) - (\overline{x}_{(i,j)}-\mu)][(x_i-\mu)' + \mu']$
所以

$$(x_i - \overline{x}_{(i,j)})x_i' = (x_i-\mu)(x_i-\mu)' + (x_i-\mu)\mu' - (\overline{x}_{(i,j)}-\mu)(x_i-\mu)' - (\overline{x}_{(i,j)}-\mu)\mu'$$

同理可得到:

$$(x_j - \overline{x}_{(i,j)})x_j' = (x_j-\mu)(x_j-\mu)' + (x_j-\mu)\mu' - (\overline{x}_{(i,j)}-\mu)(x_j-\mu)' - (\overline{x}_{(i,j)}-\mu)\mu'$$

从而

$$
\begin{aligned}
&(x_i-\overline{x}_{(i,j)})x_i'(x_j-\overline{x}_{(i,j)})x_j' \\
=&(x_i-\mu)(x_i-\mu)'(x_j-\mu)(x_j-\mu)'+(x_i-\mu)(x_i-\mu)'(x_j-\mu)\mu' \\
&-(x_i-\mu)(x_i-\mu)'(\overline{x}_{(i,j)}-\mu)(x_j-\mu)'-(x_i-\mu)(x_i-\mu)'(\overline{x}_{(i,j)}-\mu)\mu' \\
&+(x_i-\mu)\mu'(x_j-\mu)(x_j-\mu)'+(x_i-\mu)\mu'(x_j-\mu)\mu' \\
&-(x_i-\mu)\mu'(\overline{x}_{(i,j)}-\mu)(x_j-\mu)'-(x_i-\mu)\mu'(\overline{x}_{(i,j)}-\mu)\mu' \\
&-(\overline{x}_{(i,j)}-\mu)(x_i-\mu)'(x_j-\mu)(x_j-\mu)'-(\overline{x}_{(i,j)}-\mu)(x_i-\mu)'(x_j-\mu)\mu' \\
&+(\overline{x}_{(i,j)}-\mu)(x_i-\mu)'(\overline{x}_{(i,j)}-\mu)(x_j-\mu)'+(\overline{x}_{(i,j)}-\mu)(x_i-\mu)'(\overline{x}_{(i,j)}-\mu)\mu' \\
&-(\overline{x}_{(i,j)}-\mu)\mu'(x_j-\mu)(x_j-\mu)'-(\overline{x}_{(i,j)}-\mu)\mu'(x_j-\mu)\mu' \\
&+(\overline{x}_{(i,j)}-\mu)\mu'(\overline{x}_{(i,j)}-\mu)(x_j-\mu)'+(\overline{x}_{(i,j)}-\mu)\mu'(\overline{x}_{(i,j)}-\mu)\mu'
\end{aligned}
$$

根据上式, 我们就得到了(2.10), 下面我们对 $tr(A2)_k$, $k=1,2,\cdots,10$, 进行讨论. 其中

$$
\begin{aligned}
E[tr(A2)_1] &= E[\frac{1}{n(n-1)}\sum_{i\neq j}^{n}(x_i-\mu)'(x_j-\mu)(x_j-\mu)'(x_i-\mu)] \\
&= \frac{1}{n(n-1)}\sum_{i\neq j}^{n}trE[(x_j-\mu)(x_j-\mu)'(x_i-\mu)(x_i-\mu)'] \\
&= tr\Sigma^2
\end{aligned}
$$

并且由于 $E[(\overline{x}_{(i,j)}-\mu)(\overline{x}_{(i,j)}-\mu)']=1/(n-2)\Sigma$, 我们可以得到 $E[tr(A2)_{10}]$, 即

$$
\begin{aligned}
E[tr(A2)_{10}] &= E[\frac{1}{n(n-1)}\sum_{i\neq j}^{n}\mu'(\overline{x}_{(i,j)}-\mu)\mu'(\overline{x}_{(i,j)}-\mu)] \\
&= \frac{1}{n(n-1)}\sum_{i\neq j}^{n}E[\mu'(\overline{x}_{(i,j)}-\mu)\mu'(\overline{x}_{(i,j)}-\mu)] \\
&= \frac{1}{n(n-1)}\sum_{i\neq j}^{n}E[\mu'(\overline{x}_{(i,j)}-\mu)(\overline{x}_{(i,j)}-\mu)'\mu] \\
&= \frac{1}{n-2}\mu'\Sigma\mu
\end{aligned}
$$

而对于其他的 $tr(A2)_t$, $t=2,3,\cdots,9$ 的期望都为 0, 那是因为对于每一项 $tr(A2)_t$ $(t=2,3,\cdots,9)$ 都有互不相等的两项, 比如 $x_j-\mu$ 与 $x_i-\mu, i\neq j$. 由于 $x_j-\mu$ 与 $x_i-\mu$, $i\neq j$ 互相独立, 从而我们就可以得到 $E[tr(A2)_t]=0$, $t=2,3,\cdots,9$.

因此, 我们就可以得到 T_2 的期望, 即

$$E(T_2) = tr\Sigma^2 + \frac{1}{n-2}\mu'\Sigma\mu$$

显而易见, 当 $n \to \infty$ 时, $\mu'\Sigma\mu$ 的系数 $1/(n-2) \to 0$, 此时我们得到

$$E(T_2) = tr\Sigma^2 + o(1)$$

即

$$E(T_2) \xrightarrow{p} tr\Sigma^2$$

从而说明 T_2 是 $tr\Sigma^2$ 的一致估计. □

显然, 当样本均值 $\mu = 0$ 时, $E(T_2) = tr\Sigma^2$, 此时 T_2 是 $tr\Sigma^2$ 的无偏估计, 而当样本均值 $\mu \neq 0$ 时, T_2 是 $tr\Sigma^2$ 的一致估计. 接下来我们讨论第二种估计 T_2 的方差, 由于 $\overline{x} = 1/n\sum_{i=1}^{n} x_i$, 并且由于当 $i \neq j$ 时, $\overline{x}_{(i,j)} - \mu$ 与 $x_i - \mu$, $x_j - \mu$ 独立同分布, 所以我们可以得到:

$$\begin{aligned} E[(\overline{x}_{(i,j)} - \mu)(\overline{x}_{(i,j)} - \mu)'] &= 1/(n-2)\Sigma \\ E[(\overline{x}_{(i,j)} - \mu)(\overline{x}_{(i,k)} - \mu)'] &= (n-3)/(n-2)^2\Sigma \\ E[(\overline{x}_{(i,j)} - \mu)(\overline{x}_{(k,l)} - \mu)'] &= (n-4)/(n-2)^2\Sigma \end{aligned}$$

其中 i, j, k, l 两两不相等.

我们知道$T_2 = \sum_{k=1}^{10} tr(A2)_k$, 我们为了要计算$E(T_2^2)$, 就要得出每一个 $E[(tr^2(A2)_i)]$ 及 $E[tr(A2)_i tr(A2)_j]$, 其中 $i \neq j$ 且 $i, j = 1, 2, \cdots, 10$, 因此, 我们便可以得到 $\sum_{i=1}^{10} E[tr^2(A2)_i]$ 和 $\sum_{i \neq j} E[tr(A2)_i tr(A2)_j]$. 从而, 我们可以得到以下推论.

推论2.4. *已知 $x_1, x_2, \cdots, x_n$ 独立同分布, $x_i = \Gamma z_i + \mu$, $i = 1, 2, \cdots, n$, 且 $\Sigma = \Gamma\Gamma' = E[(x_i - \mu)(x_i - \mu)']$, $E[(z_1'\Sigma^2 z_1)^2] = \kappa_{22} + 2tr\Sigma^4 + tr^2\Sigma^2$, 则*

$$\begin{aligned} E(T_2^2) =& \frac{(n+1)(n-2)^3(n-3) + 8(n-2)(n-3)^2 + 8(n-2)^2 + 18(n+3) + 10}{n(n-1)(n-2)^3} tr^2\Sigma^2 \\ &+ \frac{4(n-2)^4 + 8(n-2)^2 - 8(n-2) + 2}{n(n-1)(n-2)^3}\kappa_{22} \\ &+ \frac{8(n-2)^4 + 16(n-2)^2 - 16(n-2) + 2}{n(n-1)(n-2)^3} tr\Sigma^4 + \frac{2}{n(n-1)} E[(z_1'\Sigma z_2)^4] \\ &- \frac{8(n-2)^3 - 20(n-2)^2 + 16(n-2) - 4}{n(n-1)(n-2)^3} E[(z_1'\Sigma z_2)^2 z_1'\Sigma^2 z_2] \\ &+ \frac{4(n-1)(n-2) + 4}{n(n-1)(n-2)^2}\mu'\Sigma E[(z_1'\Sigma z_1)^2]\mu \end{aligned}$$

$$
\begin{aligned}
&+\frac{4(n-2)(n-3)+4}{n(n-1)(n-2)^2}\mu'\Sigma^3E[(z_1z_2')^3]\mu \\
&+\frac{2(n-1)(n-2)^3+28(n-2)^2-24(n-3)+4}{n(n-1)(n-2)^3}\mu'\Sigma^3\mu \\
&+\frac{(n-4)^2+1}{n(n-1)(n-2)^2}(\mu'\Sigma\mu)^2+\frac{1}{n(n-2)^2}E[\mu'\Gamma z_1]^4 \\
&+\frac{8}{n(n-1)(n-2)}E[(z_1'\Sigma z_1)^2z_2'\Sigma z_2z_2'\Gamma']\mu \\
&+\frac{n(n-3)(n-5)-3}{n(n-1)(n-2)^3}\mu'E[\Sigma^2\Gamma z_1z_1'\Sigma z_1] \\
&+\frac{4(n-3)(2n-3)}{n(n-1)(n-2)^2}(\mu^3)'E[\Sigma\Gamma z_1z_1'\Sigma z_1]
\end{aligned}
$$

证明： 要计算 $E(T_2)^2$, 必须求出每一个 $E[tr^2(A2)_i]$ 及 $E[tr(A2)_itr(A2)_j]$, $i\neq j$, $i,j=1,2,\cdots,10$. 计算方法同推论2.2, 它们的计算结果如下：

$$
\begin{aligned}
E[tr^2(A2)_1] &= \frac{(n-2)(n-3)}{n(n-1)}tr^2\Sigma^2+\frac{4(n-2)}{n(n-1)}E[(z_1'\Sigma^2z_1)^2]+\frac{2}{n(n-1)}E[(z_1'\Sigma z_2)^4] \\
E[tr^2(A2)_2] &= \frac{8(n-3)}{n(n-1)(n-2)}tr^2\Sigma^2+\frac{8}{n(n-1)(n-2)}E[(z_1'\Sigma^2z_1)^2] \\
&\quad+\frac{16}{n(n-1)(n-2)}E[(z_1'\Sigma z_2)^2z_1'\Sigma^2z_2] \\
E[tr^2(A2)_3] &= \frac{4}{n(n-1)}\mu'\Sigma E[(z_1'\Sigma z_1)^2]\mu+\frac{4}{n(n-1)}\mu'E[(\Sigma z_1z_2')^3]\mu+\frac{4(n-2)}{n(n-1)}\mu'\Sigma^3\mu \\
E[tr^2(A2)_4] &= \frac{24}{n(n-1)(n-2)}\mu'\Sigma^3\mu \\
E[tr^2(A2)_5] &= \frac{4}{n(n-1)}\mu'\Sigma E[(z_1'\Sigma z_1)^2]\mu+\frac{4}{n(n-1)}\mu'E[(\Sigma z_1z_2')^3]\mu+\frac{4(n-2)}{n(n-1)}\mu'\Sigma^3\mu \\
E[tr^2(A2)_6] &= \frac{4}{n(n-1)(n-2)^2}\mu'\Sigma E[(z_1'\Sigma z_1)^2]\mu+\frac{4}{n(n-1)(n-2)^2}\mu'E[(\Sigma z_1z_2')^3]\mu \\
&\quad+\frac{4(n-2)+28(n-3)^2}{n(n-1)(n-2)^3}\mu'\Sigma^3\mu \\
E[tr^2(A2)_7] &= \frac{8(n-3)+4}{n(n-1)(n-2)}(\mu'\Sigma\mu)^2 \\
E[tr^2(A2)_8] &= \frac{26(n-3)}{n(n-1)(n-2)^3}tr^2\Sigma^2+\frac{2}{n(n-1)(n-2)^3}E[(z_i'\Sigma^2z_1)^2] \\
&\quad+\frac{4}{n(n-1)(n-2)^3}E[(z_1'\Sigma z_2)^2z_1'\Sigma^2z_2] \\
E[tr^2(A2)_9] &= \frac{2}{n(n-1)}(\mu'\Sigma\mu)^2
\end{aligned}
$$

$$E[tr^2(A2)_{10}] = \frac{1}{n(n-2)^2}E[\mu'\Gamma z_1]^4 + \frac{3(n-2)(n-3)-(n-3)+4}{n(n-1)(n-2)^2}(\mu'\Sigma\mu)^2$$

而

$$\begin{aligned}\sum_{i\neq j}^{10} E[tr(A2)_i tr(A2)_j] = &-\frac{4(n-3)}{n(n-1)(n-2)^2}tr^2\Sigma^2 - \frac{4}{n(n-1)(n-2)^2}E[(z_1'\Sigma^2 z_1)^2] \\ &-\frac{4(n-2)^2-2(n-2)+8}{n(n-1)(n-2)^2}E[(z_1'\Sigma z_2)^2 z_1'\Sigma^2 z_2] \\ &+\frac{-4(n-2)^2(n-4)+20(n-3)+8}{n(n-1)(n-2)^3}\mu' E[\Sigma^2\Gamma z_1 z_1'\Sigma z_1] \\ &+\frac{4}{n(n-1)(n-2)}E[(z_1'\Sigma z_1)^2 z_2'\Sigma z_2 z_2'\Gamma']\mu \\ &+\frac{-3(n-1)(n-2)(n-3)+(n-16)(n-3)}{n(n-1)(n-2)^2}\mu'\Sigma^3\mu \\ &-\frac{2}{n(n-2)}\mu'\Sigma^3 E[(z_1'z_2)^3]\mu \\ &-\frac{2(n-3)}{n(n-1)(n-2)}\mu'\Sigma E[(z_1'\Sigma z_1)^2]\mu \\ &-\frac{4(n-2)+2(n-3)}{n(n-1)(n-2)}(\mu'\Sigma\mu)^2\end{aligned}$$

故得到 $E(T_2^2)$. □

定理2.4. 基于推论*(2.4)*, 且 $tr^2\Sigma^2 = p^2a_2^2$, $tr\Sigma^4 = pa_4$, $n/p \to c \in (0,1)$, 则我们有

$$\begin{aligned}Var(T_2) =&(-\frac{1}{n}+\frac{8}{n^2})p^2a_2^2 + (\frac{8}{n}-\frac{64}{n^2})pa_4 + \frac{2}{n}\mu'\Sigma^3\mu \\ &+\frac{1}{n^2}(\mu'\Sigma\mu)^2 + o(p^5)\end{aligned} \tag{2.12}$$

证明： 已知 $E(T_2) = tr\Sigma^2 + \frac{1}{n-2}\mu'\Sigma\mu$, 所以

$$(ET_2)^2 = tr^2\Sigma^2 + \frac{2}{n-2}(\mu'\Sigma\mu)tr\Sigma^2 + \frac{1}{(n-2)^2}(\mu'\Sigma\mu)^2$$

因为 $Var(T_2) = E(T_2^2) - (ET_2)^2$, 且推论(2.4)中我们已经计算了 $E(T_2^2)$.
经过计算, 得

$$\begin{aligned}\mu'\Sigma E[(z_1'\Sigma z_1)^2]\mu &= o(p) \\ \mu'\Sigma^3 E[(z_1'z_2)^3]\mu &= o(p^3)\end{aligned}$$

$$\begin{aligned}\mu' E[\Sigma^2 \Gamma z_1 z_1' \Sigma z_1] &= o(p^2) \\ (\mu^3)' E[\Sigma \Gamma z_1 z_1' \Sigma z_1] &= o(p)\end{aligned}$$

因为 $tr^2\Sigma^2 = p^2 a_2^2$, $tr\Sigma^4 = pa_4$, $n/p \to c \in (0,1)$, 故估计量 T_2 的方差为:

$$\begin{aligned}Var(T_2) =& (-\frac{1}{n} + \frac{8}{n^2})p^2 a_2^2 + (\frac{8}{n} - \frac{64}{n^2})pa_4 + \frac{2}{n}\mu'\Sigma^3\mu \\ &+ \frac{1}{n^2}(\mu'\Sigma\mu)^2 + o(4)\end{aligned}$$

□

2.2.3 Li和Chen估计方法及其性质

在两样本的情况下, Li和Chen [5]在假设 $H_0 : \Sigma_1 = \Sigma_2$ vs $H_1 : \Sigma_1 \neq \Sigma_2$ 的基础上, 构建了检验统计量来估计 $tr(\Sigma_1 - \Sigma_2)^2$. 由于 $tr(\Sigma_1 - \Sigma_2)^2 = tr\Sigma_1^2 + tr\Sigma_2^2 - 2tr(\Sigma_1\Sigma_2)$, 因此需要估计上式的每一项, 但是在高维的情况下, trS_{nh}^2 是 $tr\Sigma_h^2$ 的较差的估计, 其中 trS_{nh}^2 表示的是第 h $(h = 1, 2)$ 个样本的样本协方差. 为了得到 $tr\Sigma_h^2$ 的无偏估计, 同时也为了方便分析后续的渐近估计, Li和Chen [5]在 trS_{nh}^2 中提取了效率更高的项, 他们考虑了形式为 $\frac{1}{n_h(n_h-1)}\sum_{i\neq j}(x_{hi}' x_{hj})^2$ 的 U 统计量; 当 $\mu_h = 0$, 这个统计量是无偏的. 然而, 为了考虑 $\mu_h \neq 0$ 的情况, Glasser [12] [13] 使用了减去其他两个 U 统计量的方法. 基于此, Li 和Chen [5] 提出新的估计方法.

Li和Chen [5]提出的第三种统计量 $\widehat{tr\Sigma_3^2}$ 是 $tr\Sigma^2$ 的无偏估计, 在一个样本的情况下, 我们把这个估计记为 T_3. 具体分析过程如下:

$$\begin{aligned}T_3 =& \frac{1}{n(n-1)}\sum_{i\neq j}^{n}(x_i' x_j)^2 - \frac{2}{n(n-1)(n-2)}\sum_{i,j,k}^{\star} x_i' x_j x_j' x_k \\ &+ \frac{1}{n(n-1)(n-2)(n-3)}\sum_{i,j,k,l}^{\star} x_i' x_j x_k' x_l \\ =& : tr(A3) + tr(B3) + tr(C3)\end{aligned} \tag{2.13}$$

其中, $\sum^{\star}$ 表示的是互不相同的指数之和. 例如, $\sum_{i,j,k}^{\star}$ 表示的是在 $\{(i,j,k) : i \neq j, j \neq k, k \neq i\}$ 情况下的指数之和.

接下来, 我们研究第三种估计方法的统计性质, 我们主要研究其期望与方差, 而要得出 T_3 的期望, 首先要知道 $tr(A3)$, $tr(B3)$, $tr(C3)$ 的期望, 我们得到以下的定理.

定理2.5. 如果 $x_1, x_2, \cdots, x_n$ 独立同分布, 且 $\Sigma = E[(x_i - \mu)(x_i - \mu)']$, 则

$$\begin{aligned} E[tr(A3)] &= tr\Sigma^2 + 2\mu'\Sigma\mu + (\mu'\mu)^2 && (2.14)\\ E[tr(B3)] &= -2\mu'\Sigma\mu - 2(\mu'\mu)^2 && (2.15)\\ E[tr(C3)] &= (\mu'\mu)^2 && (2.16)\\ E(T_3) &= tr\Sigma^2 && (2.17) \end{aligned}$$

显然, T_3 是 $tr\Sigma^2$ 的无偏估计.

证明: 由于

$$x_i'x_j = (x_i - \mu)'(x_j - \mu) + (x_i - \mu)'\mu + \mu'(x_j - \mu) + \mu'\mu$$

所以

$$\begin{aligned} \sum_{i\neq j}^{n}(x_i'x_j)^2 =& \sum_{i\neq j}^{n}(x_i - \mu)'(x_j - \mu)(x_i - \mu)'(x_j - \mu) + 2\sum_{i\neq j}^{n}(x_i - \mu)'\mu(x_i - \mu)'\mu \\ &+ 4\sum_{i\neq j}^{n}(x_i - \mu)'(x_j - \mu)(x_i - \mu)'\mu + 4\sum_{i\neq j}^{n}(x_i - \mu)'(x_j - \mu)\mu'\mu \\ &+ 4\sum_{i\neq j}^{n}(x_i - \mu)'\mu\mu'\mu + \sum_{i\neq j}^{n}(\mu'\mu)^2 \end{aligned}$$

从而可得

$$\begin{aligned} tr(A3) =& \frac{1}{n(n-1)}\sum_{i\neq j}^{n}(x_i - \mu)'(x_j - \mu)(x_i - \mu)'(x_j - \mu) \\ &+ \frac{2}{n(n-1)}\sum_{i\neq j}^{n}(x_i - \mu)'\mu(x_i - \mu)'\mu \\ &+ \frac{4}{n(n-1)}\sum_{i\neq j}^{n}(x_i - \mu)'(x_j - \mu)(x_i - \mu)'\mu \\ &+ \frac{4}{n(n-1)}\sum_{i\neq j}^{n}(x_i - \mu)'(x_j - \mu)\mu'\mu \\ &+ \frac{4}{n(n-1)}\sum_{i\neq j}^{n}(x_i - \mu)'\mu\mu'\mu \end{aligned}$$

$$+\frac{1}{n(n-1)}\sum_{i\neq j}^{n}(\mu'\mu)^2$$
$$=:\sum_{i=1}^{6}tr(A3)_i$$

根据引理2.1的证明过程, 我们知道 $E[tr(A3)_1]=tr\Sigma^2$, 而对于 $tr(A3)_2$, 由于

$$(x_i-\mu)'\mu(x_i-\mu)'\mu=\mu'(x_i-\mu)(x_i-\mu)'\mu$$

并且 $E(\mu'(x_i-\mu)(x_i-\mu)'\mu)=\mu'\Sigma\mu$, 因此,

$$E[tr(A3)_2]=2\mu'\Sigma\mu$$

显然, $E[tr(A3)_6]=(\mu'\mu)^2$, $E[tr(A3)_i]=0, i=3,4,5$, 这是因为 $tr(A3)_3$, $tr(A3)_4$, $tr(A3)_5$ 里都含有 i,j 不相等的两项或是单一的项. 从而, 我们可以得到 $tr(A3)$ 的期望, 即

$$E[tr(A3)]=tr\Sigma^2+2\mu'\Sigma\mu+(\mu'\mu)^2$$

同理

$$E[tr(B3)]=-2\mu'\Sigma\mu$$
$$E[tr(C3)]=-(\mu'\mu)^2$$

故, 我们得到 T_3 的期望, 即

$$E(T_3)=tr\Sigma^2$$

显然, T_3 是 $tr\Sigma^2$ 的无偏估计. □

接着, 我们来讨论第三种估计量 T_3 的方差. 由(2.13), 我们得到以下推论.

推论2.5. 已知 $x_1,x_2,\cdots,x_n$ 独立同分布, $x_i=\Gamma z_i+\mu$, $i=1,2,\cdots,n$, 且 $\Sigma=\Gamma\Gamma'=E[(x_i-\mu)(x_i-\mu)']$, $E[(z_1'\Sigma^2z_1)^2]=\kappa_{22}+2tr\Sigma^4+tr^2\Sigma^2$, 则

$$\begin{aligned}E[tr^2(A3)]=&\frac{(n-2)(n+1)}{n(n-1)}tr^2\Sigma^2+\frac{4(n-2)}{n(n-1)}\kappa_{22}\\&+\frac{8(n-2)}{n(n-1)}tr\Sigma^4+\frac{2}{n(n-1)}E[(z_1'\Sigma z_2)^4]\\&+\frac{4}{n}E[\mu'\Gamma z_1]^4+\frac{2(3n+14)(n-1)+32}{n(n-1)}(\mu'\Sigma\mu)^2\\&+\frac{4(n-2)(n-3)}{n(n-1)}\mu'\Sigma^3\mu+\frac{8(n+1)}{n(n-1)}\mu'\Sigma E[(z_1'\Sigma z_1)^2]\mu\\&+\frac{32}{n(n-1)}\mu'\Sigma^3E[(z_1z_2')^3]\mu+\frac{4(n+4)}{n}(\mu'\Sigma^{\frac{1}{3}}\mu)^3\end{aligned}$$

$$
\begin{aligned}
&+(\mu'\mu)^4+\frac{16(n-2)}{n(n-1)}\mu'E[\Sigma^2\Gamma z_1z_1'\Sigma z_1]\\
&+\frac{16}{n(n-1)}E[(z_1'\Sigma z_1)^2z_2'\Sigma z_2z_2'\Gamma']\mu\\
&+\frac{16(2n+1)}{n(n-1)}(\mu^3)'E[\Sigma\Gamma z_1z_1'\Sigma z_1]
\end{aligned}
$$

$$
\begin{aligned}
E[tr^2(B3)]=&\frac{8}{n(n-1)}tr^2\Sigma^2+\frac{8}{n(n-1)(n-2)}\kappa_{22}+\frac{16}{n(n-1)(n-2)}tr\Sigma^4\\
&+\frac{16}{n(n-1)(n-2)}E[(z_1'\Sigma z_2)^2z_1'\Sigma^2z_2]\\
&+\frac{16(n-2)^2+96}{n(n-1)(n-2)}\mu'\Sigma^3\mu+\frac{16}{n(n-1)}\mu'\Sigma E[(z_1'\Sigma z_1)^2]\mu\\
&+\frac{16}{n(n-1)}\mu'\Sigma^3E[(z_1z_2')^3]\mu+\frac{4(n+15)(n-1)+200}{n(n-1)}(\mu'\Sigma\mu)^2\\
&+\frac{4}{n}E[\mu'\Gamma z_1]^4+\frac{8(n+8)}{n}(\mu'\Sigma^{\frac{1}{3}}\mu)^3\\
&+4(\mu'\mu)^4+\frac{32(n+1)}{n(n-1)(n-2)}\mu'E[\Sigma^2\Gamma z_1z_1'\Sigma z_1]\\
&+\frac{80}{n(n-1)}\mu'\Sigma^3\mu+\frac{16(n+9)}{n(n+1)}(\mu^3)'E[\Sigma\Gamma z_1z_1'\Sigma z_1]\\
&+\frac{32}{n}(\mu^5)'E(\Gamma z_1z_1'\Sigma z_1)
\end{aligned}
$$

$$
\begin{aligned}
E[tr^2(C3)]=&\frac{24}{n(n-1)(n-2)(n-3)}tr^2\Sigma^2+\frac{96}{n(n-1)(n-2)}\mu'\Sigma^3\mu\\
&+\frac{72}{n(n-1)}(\mu'\Sigma\mu)^2+\frac{16}{n}(\mu'\Sigma^{\frac{1}{3}}\mu)^3+(\mu'\mu)^4
\end{aligned}
$$

$$
\begin{aligned}
E[tr(A3)tr(B3)]=&-\frac{4}{n(n-1)}E[(z_1'\Sigma z_2)^2z_1'\Sigma^2z_2]-\frac{8}{n(n-1)}E[(z_1'\Sigma z_1)^2z_2'\Sigma z_2z_2'\Gamma']\mu\\
&-\frac{8}{n-1}\mu'E[\Sigma^2\Gamma z_1z_1'\Sigma z_1]-\frac{36}{n(n-1)}\mu'\Sigma^3E[(z_1z_2')^3]\mu\\
&-\frac{4(n+3)}{n(n-1)}\mu'\Sigma E[(z_1'\Sigma z_1)^2]\mu-\frac{32n+80}{n(n-1)}(\mu^3)'E[\Sigma\Gamma z_1z_1'\Sigma z_1]\\
&-\frac{4}{n}E[\mu'\Gamma z_1]^4-\frac{24}{n}(\mu^5)'E(\Gamma z_1z_1'\Sigma z_1)\\
&-\frac{2(n+6)}{n}\mu'\Sigma^3\mu-\frac{2(3n+22)(n-1)+80}{n(n-1)}(\mu'\Sigma\mu)^2\\
&-\frac{6n+32}{n}(\mu'\Sigma^{\frac{1}{3}}\mu)^3-2(\mu'\mu)^4
\end{aligned}
$$

$$\begin{aligned}E[tr(A3)tr(C3)] =&\frac{12}{n(n-1)}\mu'\Sigma^3E[(z_1z_2')^3]\mu+\frac{8(n+5)}{n(n-1)}(\mu^3)'E[\Sigma\Gamma z_1z_1'\Sigma z_1]\\&+\frac{(n+16)(n-1)+48}{n(n-1)}(\mu'\Sigma\mu)^2+\frac{8}{n}(\mu^5)'E(\Gamma z_1z_1'\Sigma z_1)\\&+\frac{2(n+8)}{n}(\mu'\Sigma^{\frac{1}{3}}\mu)^3+(\mu'\mu)^4\end{aligned}$$

$$\begin{aligned}E[tr(B3)tr(C3)] =&-\frac{48}{n(n-1)(n-2)}\mu'E[\Sigma^2\Gamma z_1z_1'\Sigma z_1]-\frac{48}{n(n-1)}(\mu^3)'E[\Sigma\Gamma z_1z_1'\Sigma z_1]\\&-\frac{8}{n}(\mu^5)'E(\Gamma z_1z_1'\Sigma z_1)-\frac{24(n+2)}{n(n-1)(n-2)}\mu'\Sigma^3\mu\\&-\frac{16(n-1)+120}{n(n-1)}(\mu'\Sigma\mu)^2-\frac{2n+32}{n}(\mu'\Sigma^{\frac{1}{3}}\mu)^3-2(\mu'\mu)^4\end{aligned}$$

由推论(2.5), 我们容易得到 $E(T_3^2)$, 又因为 $E(T_3)=tr\Sigma^2$, 因此, 我们得到以下定理.

定理2.6. 基于推论*(2.5)*, 且 $tr^2\Sigma^2=p^2a_2^2$, $tr\Sigma^4=pa_4$, $n/p\to c\in(0,1)$, 我们有

$$Var(T_3)=(-\frac{4}{n}+\frac{34}{n^2})p^2a_2^2+(\frac{8}{n}-\frac{32}{n^2})pa_4+o(p^4) \tag{2.18}$$

证明： 根据Himeno和Yamada [17], 我们可得

$$\begin{aligned}\kappa_{22} &= o(p^2)\\ E[(z_1'\Sigma z_2)^2z_1'\Sigma^2z_2] &= o(p^{5/2})\\ E[(z_1'\Sigma z_2)^4] &= o(p^4)\end{aligned}$$

且 $[E(T_3)]^2=tr^2\Sigma^2$, $tr^2\Sigma^2=p^2a_2^2$, $tr\Sigma^4=pa_4$, $n/p\to c\in(0,1)$, 又 $Var(T_3)=E(T_3^2)-[E(T_3)]^2$, 故 T_3 的方差为:

$$Var(T_3)=(-\frac{4}{n}+\frac{34}{n^2})p^2a_2^2+(\frac{8}{n}-\frac{32}{n^2})pa_4+o(p^4)$$

□

2.3 新估计方法及统计性质

2.3.1 U统计量

在这一节的开始, 我们先来介绍U统计量.

U 统计量是由 W.Hoeffding 提出的一种非参数统计量, 是样本均值的推广, 其作用是能够将样本中有关总体的信息聚集起来, 便于对总体进行研究.

定义2.6. 对分布族 $\mathscr{F}$ 的参数 θ, 如果存在样本量为 n 的样本 $X_1, X_2, \cdots, X_m$ 的统计量 $h(X_1, X_2, \cdots, X_m)$, 使得

$$E_F h(X_1, X_2, \cdots, X_m) = \theta$$

则称参数 θ 对分布族 $\mathscr{F}$ 是 m 可估的, $h(X_1, X_2, \cdots, X_m)$ 称为 θ 的核.
一般的, 对 $(1, 2, \cdots, m)$ 的任一个排列 $(i_1, i_2, \cdots, i_m)$, 都有

$$h(X_1, X_2, \cdots, X_m) = h(X_{i_1}, X_{i_2}, \cdots, X_{i_m})$$

则称为对称核.

定义2.7. 设独立随机变量 $X_1, X_2, \cdots, X_n$ 是总体 $F(x) \in \mathscr{F}$ 的样本, n 可估参数 θ 有对称核 $h(X_1, X_2, \cdots, X_m)$. 则参数θ 的 U 统计量构造如下:

$$U_n = \frac{1}{C_n^m} \sum_{1 \le i_1 \le \cdots \le i_m \le n} h(X_1, X_2, \cdots, X_m)$$

我们都知道, U 统计量是解决非参数统计结构中检验问题的一种有效方法, 因此, U 统计量检验方法在检验问题中被广泛应用, 例如用于解决两总体分布均值相等的检验问题及两总体方差差异的检验问题等. 本节利用 U 统计量检验法构造非参数统计量来估计 $tr\Sigma^2$, 与Bai 和Saranadasa [3]、Chen和Qin [4] 、Li和Chen [5]提出的 $tr\Sigma^2$ 的三种估计方法相比, 适用于更广泛的场合, 也就是这种检验方法不仅仅适用于已知总体分布且含有未知参数的统计结构, 而且也适用于总体结构未知的非参数结构.

2.3.2 新估计方法及其性质

基于 U 统计量的检验方法, 我们构建了如下的非参数统计量来估计 $tr\Sigma^2$. 具体过程如下:
由于

$$(x_1-x_2)(x_1-x_2)' = (x_1-\mu)(x_1-\mu)'-(x_1-\mu)(x_2-\mu)'-(x_2-\mu)(x_1-\mu)'+(x_2-\mu)(x_2-\mu)'$$

$$(x_3-x_4)(x_3-x_4)' = (x_3-\mu)(x_3-\mu)'-(x_3-\mu)(x_4-\mu)'-(x_4-\mu)(x_3-\mu)'+(x_4-\mu)(x_4-\mu)'$$

且 $(x_i - \mu)$ 与 $(x_j - \mu)$ 独立同分布, 其中 $i \neq j$, 所以

$$E[(x_i - \mu)(x_j - \mu)'] = 0$$

而我们已知 $E[(x_i-\mu)(x_i-\mu)']=\Sigma$, 从而我们可以得到:

$$E[(x_1-x_2)(x_1-x_2)'(x_3-x_4)(x_3-x_4)']=4\Sigma^2$$

因此

$$E\{[(x_1-x_2)'(x_3-x_4)(x_3-x_4)'(x_1-x_2)]/4\}=tr\Sigma^2$$

由此可知 $[(x_1-x_2)'(x_3-x_4)'(x_3-x_4)'(x_1-x_2)']/4$ 是 $tr\Sigma^2$ 的无偏估计. 根据 U 统计量检验法, 当样本容量为 n 时, 我们可以提出 $tr\Sigma^2$ 的第四种非参数估计 $\widehat{tr\Sigma_4^2}$, 记为 T_4, 即:

$$T_4=\frac{1}{4n(n-1)(n-2)(n-3)}\sum_{i,j,k,l}^{*}(x_i-x_j)'(x_k-x_l)(x_k-x_l)'(x_i-x_j) \quad (2.19)$$

其中, $\sum_{i,j,k,l}^{*}$ 表示的是在 $\{(i,j,k,l): i\neq j, i\neq k, i\neq l, j\neq k, j\neq l, k\neq l\}$ 情况下的指数之和.

我们猜想: T_4 是 $tr\Sigma^2$ 的无偏估计, 即 $E(T_4)=tr\Sigma^2$, 具体讨论过程如下:
由(2.19), 并且

$$(x_i-x_j)'(x_k-x_l)=(x_i-\mu)'(x_k-\mu)-(x_i-\mu)'(x_l-\mu)-(x_j-\mu)'(x_k-\mu)+(x_j-\mu)'(x_l-\mu)$$

$$(x_k-x_l)'(x_i-x_j)=(x_k-\mu)'(x_i-\mu)-(x_k-\mu)'(x_j-\mu)-(x_l-\mu)'(x_i-\mu)+(x_l-\mu)'(x_j-\mu)$$

把上面两式相乘, 经过整理可得:

$$\begin{aligned}T_4=&\frac{1}{n(n-1)(n-2)(n-3)}\sum_{i,j,k,l}^{*}\{(x_i-\mu)'(x_k-\mu)(x_k-\mu)'(x_i-\mu)\\&-2(x_i-\mu)'(x_k-\mu)(x_k-\mu)'(x_j-\mu)\\&+(x_i-\mu)'(x_k-\mu)(x_l-\mu)'(x_j-\mu)\}\\=&: tr(A4)_1+tr(A4)_2+tr(A4)_3\end{aligned} \quad (2.20)$$

其中, $\sum_{i,j,k,l}^{*}$ 表示的是在 $\{(i,j,k): i\neq j, i\neq k, i\neq l, j\neq k, j\neq l, k\neq l\}$ 情况下的指数之和.

定理2.7. 如果 $x_1,x_2,\cdots,x_n$ 独立同分布, 且 $\Sigma=E[(x_i-\mu)(x_i-\mu)']$, 则

$$E(T_4)=tr\Sigma^2$$

显然, T_4 是 $tr\Sigma^2$ 的无偏估计.

证明： 由(2.20)知, $T_4 = tr(A4)_1 + tr(A4)_2 + tr(A4)_3$, 即

$$E(T_4) = E[tr(A4)_1] + E[tr(A4)_2] + E[tr(A4)_3]$$

由 $E[(x_i-\mu)'(x_k-\mu)(x_k-\mu)'(x_i-\mu)] = tr\Sigma^2$, 则

$$E[tr(A4)_1] = tr\Sigma^2$$

而 $E[tr(A4)_i] = 0$, $i = 2,3$, 原因如下: $tr(A4)_i$ 中含有项 $(x_i-\mu)$ 与 $(x_j-\mu)$, $i \neq j$, 并且这两项是独立的, 故 $E(T_4) = tr\Sigma^2$. □

接下来, 我们继续讨论 T_4 的方差. 由于 $Var(T_4) = E(T_4^2) - E^2(T_4)$, 所以我们要计算 $E(T_4^2)$, 从而就必须要得到每一个 $E[tr^2(A4)_i]$ 和 $E[tr(A4)_i tr(A4)_j]$, 其中 $i \neq j$ 且 $i,j = 1,2,3$. 我们便得到了以下的推论.

推论2.8. *已知 $x_1, x_2, \cdots, x_n$ 独立同分布, $x_i = \Gamma z_i + \mu$, $i = 1,2,\cdots,n$, 且 $\Sigma = \Gamma\Gamma' = E[(x_i-\mu)(x_i-\mu)']$, $E[(z_1'\Sigma^2 z_1)^2] = \kappa_{22} + 2tr\Sigma^4 + tr^2\Sigma^2$, 则*

$$\begin{aligned} E(T_4^2) =& \frac{(n+1)(n-2)^2(n-3) + 8(n-2)(n-3) + 24}{n(n-1)(n-2)(n-3)} tr^2\Sigma^2 + \frac{4(n-2)^2+8}{n(n-1)(n-2)}\kappa_{22} \\ &+ \frac{8(n-2)^2+16}{n(n-1)(n-2)} tr\Sigma^4 - \frac{8(n-2)-16}{n(n-1)(n-2)} E[(z_1'\Sigma z_2)^2 z_1'\Sigma^2 z_2] \\ &+ \frac{2}{n(n-1)} E[(z_1'\Sigma z_2)^4] \end{aligned}$$

证明： 我们知道 $T_4 =: tr(A4)_1 + tr(A4)_2 + tr(A4)_3$, 因此要求 $E(T_4^2)$, 就要求出每一个 $E[(trA4_i)^2]$ 及 $E[tr(A4)_i tr(A4)_j], i \neq j$, $i,j = 1,2,3$. 计算方法同推论2.2, 它们的计算结果如下:

$$\begin{aligned} E[tr^2(A4)_1] &= \frac{(n-2)(n-3)}{n(n-1)} tr^2\Sigma^2 + \frac{4(n-2)}{n(n-1)} E[(z_1'\Sigma^2 z_1)^2] \\ &\quad \frac{2}{n(n-1)} E[(z_1'\Sigma z_2)^4] \\ E[tr^2(A4)_2] &= \frac{8(n-3)}{n(n-1)(n-2)} tr^2\Sigma^2 + \frac{8}{n(n-1)(n-2)} E[(z_1'\Sigma^2 z_1)^2] \\ &\quad \frac{16}{n(n-1)(n-2)} E[(z_1'\Sigma z_2)^2 z_1'\Sigma^2 z_2] \\ E[tr^2(A4)_3] &= \frac{24}{n(n-1)(n-2)(n-3)} tr^2\Sigma^2 \\ E[tr(A4)_1 tr(A4)_2] &= -\frac{4}{n(n-1)} E[(z_1'\Sigma z_2)^2 z_1'\Sigma^2 z_2] \\ E[tr(A4)_1 tr(A4)_3] &= 0 \quad E[tr(A4)_2 tr(A4)_3] = 0 \end{aligned}$$

因此, 我们可以得到 $E(T_4^2)$. □

定理2.8. 基于推论*(2.8)*, 且 $tr^2\Sigma^2 = p^2a_2^2$, $tr\Sigma^4 = pa_4$, $n/p \to c \in (0,1)$, 我们有

$$Var(T_4) = (-\frac{4}{n} + \frac{34}{n^2})p^2a_2^2 + (\frac{8}{n} - \frac{32}{n^2})pa_4 + o(p^4) \tag{2.21}$$

证明过程同定理2.6, 这里将不再赘述. 显然, 通过理论研究, 我们发现已有的第三种估计方法与我们新提的估计方法具有相同的统计性质, 但两者提出的角度不同, 因此, 适用范围也不同.

2.4 小结

在本章, 通过计算, 我们知道了已有的估计量 T_1 和 T_2 是 $tr\Sigma^2$ 的一致估计, 而不是无偏估计. 这也就是说, 只有当样本容量 n 无限大的时候, 其期望才无限接近 $tr\Sigma^2$, 但我们观察发现, 如果样本均值 $\mu = 0$ 时, T_2 是 $tr\Sigma^2$ 的无偏估计. 而已有估计量 T_3 与我们提的新估计方法 T_4 是 $tr\Sigma^2$ 的无偏估计, 同时, 我们也计算了每种估计方法的方差. 从估计方法的统计性质上来看, 我们发现 T_3 和 T_4 有相同的期望与方差. 但由于估计方法的方差与样本的大小 n、维度 p 有关, 所以我们不能从理论上比较四种估计方法之间的优劣. 因此, 在下一章, 我们将利用数据模拟来检验估计量的期望与方差, 进而来比较估计量的好坏.

第三章 模拟与分析

由于我们不能直接通过理论计算结果来比较四种估计方法的优劣, 因此在本章, 我们将利用R语言对四种统计方法的相关理论进行数据模拟, 同时对模拟结果进行分析.

3.1 模拟

在这里, 我们生成服从两种分布的高维数据, 即服从正态分布与服从拉普拉斯分布. 我们把正态分布表示为 $\boldsymbol{K_1}$, 把拉普拉斯分布表示为 $\boldsymbol{K_2}$.

$\boldsymbol{K_1}$: 从 p 维正态分布中生成高维数据 $x_i = (x_{i_1}, x_{i_2}, \cdots, x_{ip})'$ $(i = 1, 2, \cdots, n)$, 其均值 $\mu = (\mu_1, \mu_2, \cdots, \mu_p)'$, 协方差矩阵为 Σ.

$\boldsymbol{K_2}$: 从 p 维拉普拉斯分布中生成高维数据 $x_i = (x_{i_1}, x_{i_2}, \cdots, x_{ip})'$ $(i = 1, 2, \cdots, n)$, 其均值 $\mu = (\mu_1, \mu_2, \cdots, \mu_p)'$, 协方差矩阵为 Σ.

为了使模拟结果更具有真实性、可靠性, 在这里我们采用三种协方差结构. 令 $D = (d_{q,s})_{q,s=1,\cdots,p}$ 为对角矩阵, 其对角元素 $d_{s,s}$ $(s = 1, 2, \cdots, p)$ 产生于均匀分布(1,3). 三种协方差结构分别设置如下:

结构1: 协方差矩阵为 $\Sigma = D^{1/2}AD^{1/2}$, 其中 $A = (\rho^{|q-s|})_{q,s=1,2,\cdots,p}$, 且 $\rho = 0, 0.1, 0.2, 0.3, 0.4, 0.5$.

结构2: 协方差矩阵为 $\Sigma = D^{1/2}A^{-1}D^{1/2}$, 其中 $A = (\rho^{|q-s|})_{q,s=1,2,\cdots,p}$, 且 $\rho = 0, 0.1, 0.2, 0.3, 0.4, 0.5$.

结构3: 协方差矩阵为 $\Sigma = D^{1/2}BD^{1/2}$, 其中 $B = (b_{ij})_{p\times p}$, 当 $i = j$ 时, $b_{ij} = 1$, 当 $i \neq j$ 时, $b_{ij} = \rho$ 且 $\rho = 0, 0.1, 0.2, 0.3, 0.4, 0.5$.

表 1: 基于结构1 在 K_1 与 K_2 下的期望

	p	$E(T)$	$E(T_1)$	$E(T_2)$	$E(T_3)$	$E(T_4)$
K_1	0	188.3454	196.3091	191.5582	191.5623	191.5623
	0.1	191.8426	196.3463	191.8302	191.8342	191.8342
	0.2	159.1758	159.4027	155.6889	155.6855	155.6855
	0.3	174.4933	181.1948	177.0281	177.0294	177.0294
	0.4	198.8783	202.7860	198.4206	198.4261	198.4261
	0.5	236.6446	243.9269	237.8413	237.8397	237.8397
K_2	0	163.4720	181.3004	164.5780	164.4173	164.4173
	0.1	166.3390	186.4919	169.5606	169.3188	169.3188
	0.2	167.0464	182.7660	166.9325	166.9370	166.9370
	0.3	183.3734	199.9532	183.3884	183.3920	183.3920
	0.4	209.4579	224.6430	208.6270	208.6227	208.6227
	0.5	250.0204	271.9682	253.4164	253.4100	253.4100

表 2: 基于结构1 在 K_1 与 K_2 下的方差

	p	$D(T)$	$D(T_1)$	$D(T_2)$	$D(T_3)$	$D(T_4)$
K_1	0	0.0000	319.4782	307.8459	308.0372	308.0372
	0.1	0.0000	513.0026	496.5229	496.9610	496.9610
	0.2	0.0000	268.6716	262.1309	262.2875	262.2875
	0.3	0.0000	420.6025	400.8274	401.2131	401.2131
	0.4	0.0000	486.3927	473.3746	473.5191	473.5191
	0.5	0.0000	1004.8830	1007.6111	1008.0386	1008.0386
K_2	0	0.0000	657.6824	507.4437	507.6931	507.6931
	0.1	0.0000	793.9154	592.1449	591.3403	591.3403
	0.2	0.0000	743.9255	532.8849	533.2918	533.2918
	0.3	0.0000	1015.3853	806.6079	807.6481	807.6481
	0.4	0.0000	1261.5797	1080.1788	1080.2585	1080.2585
	0.5	0.0000	3180.2965	2552.7996	2553.7355	2553.7355

表 3: 基于结构2 在 K_1 与 K_2 下的期望

	p	$E(T)$	$E(T_1)$	$E(T_2)$	$E(T_3)$	$E(T_4)$
K_1	0	162.9765	166.0961	162.2150	162.2151	162.2151
	0.1	172.4787	173.0836	168.5189	168.5151	168.5151
	0.2	228.2031	234.8378	229.1424	229.1438	229.1438
	0.3	296.5023	308.6035	301.0794	301.0925	301.0925
	0.4	419.9276	437.2013	426.3323	426.3039	426.3039
	0.5	647.2373	676.2380	660.7274	660.7179	660.7179
K_2	0	169.7440	186.9992	170.8136	170.6471	170.6471
	0.1	179.6232	196.7655	178.7912	178.5269	178.5269
	0.2	166.9503	185.4008	168.7685	168.7715	168.7715
	0.3	216.6083	234.6489	216.4482	216.4529	216.4529
	0.4	306.3350	326.6258	301.5379	301.5382	301.5382
	0.5	471.5576	509.5991	472.6856	472.6632	472.6632

表 4: 基于结构2 在 K_1 与 K_2 下的方差

	p	$D(T)$	$D(T_1)$	$D(T_2)$	$D(T_3)$	$D(T_4)$
K_1	0	0.0000	328.5584	326.2697	326.6089	326.6089
	0.1	0.0000	221.1478	207.6808	208.0015	208.0015
	0.2	0.0000	457.4163	458.5383	459.0675	459.0675
	0.3	0.0000	1130.1510	1088.6448	1089.7659	1089.7659
	0.4	0.0000	2247.9642	2157.4491	2157.9777	2157.9777
	0.5	0.0000	5379.5349	5320.9799	5326.4933	5326.4933
K_2	0	0.0000	745.3373	619.4639	619.4163	619.4163
	0.1	0.0000	985.1974	713.6914	713.9236	713.9236
	0.2	0.0000	840.2055	662.5534	663.2430	663.2430
	0.3	0.0000	1168.6464	959.6907	960.1231	960.1231
	0.4	0.0000	3501.7116	2686.9188	2688.8023	2688.8023
	0.5	0.0000	5463.9442	4634.0885	4633.4764	4633.4764

表 5: 基于结构3 在 K_1 与 K_2 下的期望

	p	$E(T)$	$E(T_1)$	$E(T_2)$	$E(T_3)$	$E(T_4)$
K_1	0	191.5687	192.7112	187.7602	187.7505	187.7505
	0.1	261.4761	262.9958	257.3171	257.3120	257.3120
	0.2	471.1983	454.1981	443.7640	443.7294	443.7294
	0.3	820.7353	768.5767	748.8621	748.8280	748.8280
	0.4	1310.087	1284.466	1253.279	1253.003	1253.003
	0.5	1939.254	2141.192	2081.151	2081.459	2081.459
K_2	0	168.7960	182.9723	166.9323	166.7880	166.7880
	0.1	229.3443	236.4740	218.9563	218.7309	218.7309
	0.2	410.9891	456.3157	433.2631	433.0178	433.0178
	0.3	713.7305	736.5839	713.1035	712.8590	712.8590
	0.4	1137.568	1219.217	1178.907	1178.546	1178.546
	0.5	1682.503	1731.255	1677.005	1676.393	1676.393

表 6: 基于结构3 在 K_1 与 K_2 下的方差

	p	$D(T)$	$D(T_1)$	$D(T_2)$	$D(T_3)$	$D(T_4)$
K_1	0	0.0000	329.9559	314.0926	314.4618	314.4618
	0.1	0.0000	1881.3817	1856.7320	1856.9716	1856.9716
	0.2	0.0000	18794.2821	18034.4549	18056.5492	18056.5492
	0.3	0.0000	96719.9178	92001.4740	92106.1252	92106.1252
	0.4	0.0000	306062.425	288727.175	288940.093	288940.093
	0.5	0.0000	1391804.130	1296079.881	1297493.690	1297493.690
K_2	0	0.0000	673.1079	496.8279	497.2092	497.2092
	0.1	0.0000	2735.9507	2421.2754	2420.6182	2420.6182
	0.2	0.0000	26988.0098	24175.7707	24194.4151	24194.4151
	0.3	0.0000	94518.3208	89930.6324	90043.0635	90043.0635
	0.4	0.0000	323148.620	302484.705	302327.917	302327.917
	0.5	0.0000	525105.510	497258.179	497676.846	497676.846

3.2 模拟结果与分析

在模拟中, 我们取样本容量 $n = 40$, 数据维数 $p = 40$, 并且重复循环次数 $b = 100$. 模拟结果得到表1至表6, 分别是协方差矩阵的结构基于结构1、结构2与结构3的 $tr\Sigma^2$ 期望与方差的真实值 $E(T)$, $D(T)$ 和估计值 $E(T_i)$, $D(T_i)$, $i = 1, 2, 3, 4$.

通过观察6个表, 无论协方差结构基于哪种结构, 且无论高维数据服从哪种分布, 我们发现Li和Chen提出的第三种估计方法与我们提出来的估计方法的期望与方差都是相等的, 这就很好的从数据模拟上验证了我们理论部分中已有的第三种估计方法与我们提出来的新估计方法有相同的统计性质.

通过观察表1与表2发现, 在结构1的基础上, 且数据服从正态分布的情况下, 当 $\rho = 0.2$ 时, Bai 和Saranadasa [3]提的第一种估计方法的期望与真实值较接近, 但此时方差却不是最小的; 当 $\rho = 0.5$ 时, 第一种估计方法的方差比较小. 而在其他条件下, Chen和Qin [4]、Li和Chen [5]与我们提的新估计方法的期望值和真实值比较贴合, 与真实值的误差都较好的控制在5以内, 且这三种估计方法的期望值也相差不大, 方差也比Bai 和Saranadasa [3]提的估计方法小. 所以, 从总体上来说, 在协方差结构是结构1的基础上, Chen和Qin [4]、Li和Chen [5]与我们提的新估计方法三者的期望相差甚小, 但我们发现Chen和Qin [4]提的第二种估计量的方差一直比第三、第四种估计量小, 从而说明第二种估计量的稳健性更好.

通过观察表3与表4发现, 在结构2的基础上, 且数据服从正态分布的情况下, 当 $\rho = 0.1$ 时, Bai 和Saranadasa [3]提的第一种估计方法的期望与真实值较接近; 而在其他条件下, Bai和Saranadasa [3]估计方法的期望与真实值相差最大, 且其方差也最大, 所以其估计效果并不理想, 而其他三种估计方法的期望值和真实值比较贴合, 但与真实值的误差控制在13以内. 显然, 在协方差结构是结构2时, 总体期望与真实值的误差、总体方差较结构1大, 因此, 其估计效果没有结构1稳健.

通过观察表5与表6发现, 在结构3及正态分布的基础上, Bai和Saranadasa [3]估计方法的期望与真实值较接近, 但随着 ρ 的增大, 其误差也越来越大. 在拉普拉斯分布下, 虽然 ρ 的某些取值会使Bai和Saranadasa [3]估计方法的期望与真实值比较接近, 但在一些取值上, 其误差会发生突变, 这就说明其稳健性是非常差的.

最后, 无论基于哪种结构、哪种分布, Bai和Saranadasa [3]估计方法的估计效果最不理想, Chen和Qin [4]、Li 和Chen [5]提的估计方法与我们提的新估计方法的期望与方差相差不大, 但Chen和Qin [4]提的估计方法在方差上却是最小的, 说明这种估计方法是较稳健的, 但从统计方法的构建上, 我们提出的新估计方法是基于U 统计量的非参数方法, 其适用范围更广, 且从数据模拟上看, 其模拟结果也较好.

第四章　总结与展望

在本文中, 我们考虑的是一个样本的情况, 并基于高维数据是独立同分布的, 我们综合研究了 $tr\Sigma^2$ 已有的三种估计方法, 并提出了我们新的估计方法. 通过对四种估计方法性质的研究, 我们知道了Bai和Saranadasa [3]与Chen和Qin [4]提的已有估计方法是 $tr\Sigma^2$ 的一致估计, 而Li和Chen [5]提的估计方法与我们提的新估计方法是 $tr\Sigma^2$ 的无偏估计. 同时, 我们也研究了四种估计方法的方差, 但我们发现从理论上不能直接比较估计方法的优劣, 因此, 我们利用R语言统计软件进行数据模拟, 并得到了如下的结论:

在性质上, 从不同的模型中我们发现Li和Chen [5]提出的第三种估计方法与我们提出来的估计方法的性质一致, 同时, 在四种估计方法的期望和方差与真实值的比较上, 我们发现Bai和Saranadasa [3]提的估计方法模拟结果变化较大, 即这种估计方法用于估计 $tr\Sigma^2$ 偏差较大, 说明其不是很好的估计方法; 而从总体上来说, 无论在哪种模型的基础上其他三种估计方法区别不大, 但在大多数情况下, Chen和Qin [4]提出的第二种估计方法的期望与真实值更为接近, 并且其方差也较小. 但从统计方法的构建上, 我们提出的新估计方法是基于U统计量的非参数方法, 其适用范围更广, 且从数据模拟上看, 其模拟结果也较好.

然而, 我们正处在信息不断更新的时代, 这也意味着我们面对的高维数据将是越来越复杂的, 这就需要对高维数据进行更深入的研究. 在今后, 在我们对 $tr\Sigma^2$ 提出了新的估计方法之后, 我们就可以对研究背景中的假设检验问题提出更好的估计办法, 同时, 在此基础上, 我们也可以对 $tr\Sigma^2$ 估计方法进行推广, 比如, 我们可以研究 $tr(\Sigma_1\Sigma_2)$ 及 $tr(\Sigma^4)$ 的估计方法, 其中 Σ_i $(i=1,2)$ 表示的是第 i 个样本的协方差矩阵, 我相信这一定会为我们处理高维数据带来便利.

参考文献

[1] 宋鹏, 刘程程, 胡永宏. 稳健高维协方差矩阵估计及其投资组合应用——基于中心正则化算法[J]. 统计研究, 2020(7):312-320.

[2] 李铭, 赵强. 一种估计高维协方差矩阵的新方法[J]. 山东师范大学学报(自然科学版), 2019, 034(001):43-47.

[3] Bai Z, Saranadasa H. Effect of high dimension: By an example of a two sample problem[J]. Statist.Sinica, 1996, 6(2):311-329.

[4] Chen S X, Qin Y L. A two-sample test for high-dimensional data with applications to gene-set testing[J]. Annals of Statistics, 2010, 38(2):808-835.

[5] Li J, Chen S X. Two sample tests for high-dimensional covariance matrices[J]. The Annals of Statistics, 2012 40(2):908-940.

[6] Kosorok M R, Ma S. Marginal asymptotics for the "large P, small N" paradigm: With applications to microarray data[J]. The Annals of Statistics, 2005, 35(4):124-133.

[7] Schott J R. Some high-dimensional tests for a one-way MANOVA[J]. Journal of Multivariate Analysis, 2007, 98(9):1825-1839.

[8] Srivastava M S, Fujikoshi. Multivariate analysis of variance with fewer observations than the dimension[J]. Multivariate Anal, 2006, 97(9):1927-1940.

[9] Srivastava M S . Some Tests Concerning the Covariance Matrix in High Dimensional Data[J]. Journal of the Japan Statistical Society, 2005, 35(2):251–272.

[10] Chen S X, Zhang L X, Zhong P S. Tests for high-dimensional covariance matrices[J]. Am.Statist.Assoc, 2010, 105:810-819.

[11] Fujikoshi Y, Himeno T, Wakaki H. Asymptotic results of a high dimensional MANOVA test and power comparison when the dimension is large compared to the sample size[J]. Japan Statist.Soc, 2004, 34(1):19-26.

[12] Glasser G J. An unbiased estimator for powers of the arithmetic mean[J]. Roy.Statist.Soc.Ser, 1961, B23:154-159.

[13] Glasser G J. Estimators for the product of arithmetic means[J]. Roy.Statist.Soc.Ser. 1962, B24:180-184.

[14] Magnus J R, Neudecker H. The commutation matrix: Some properties and applications[J]. Ann.Statist, 1979, 17(7):381-894.

[15] Huang J Z, Liu N, Pourahmadi M, Liu L. Covariance matrix selection and estimation via penalised normal likelihood[J]. Biometrika, 2006, 93(1):85-98.

[16] Bickel P L, Levina E. Regularized estimation of large covariance matrices[J]. Ann.Statist, 2008, 36(1):199-227.

[17] Himeno T, Yamada T. Estimations for some functions of covariance matrix in high dimension under non-normality and its applications[J]. Journal of Multivariate Analysis, 2014, 130:27-44.

[18] Fan J, Peng H, Huang T. Semilinear High-Dimensional Model for Normalization of Microarray Data[J]. Publications of the American Statistical Association, 2005, 100(471):781-796.

[19] Ledoit O, Wolf M. Some Hypothesis Tests for the Covariance Matrix When the Dimension is Large Compared to the Sample Size[J]. The Annals of Statistics, 2001, 30(575):733-751.

[20] Portnoy S. On the central limit theorem in R p when p$\rightarrow\infty$[J]. probability theory related fields, 1986, 73(4):571-583.

[21] Storey J D, Taylor J E, D Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach[J]. Journal of the Royal Statistical Society, 2010, 66(4):534-547.

[22] Yin Y Q, Bai Z D, Krishnaiah P R. On the limit of the largest eigenvalue of the large dimensional sample covariance matrix[J]. Probability Theory Related Fields, 1988, 78(4):509-521.

致　谢

时光飞逝, 我在华中师范大学三年的学习生活也即将画上句号. 在这三年里, 我得到了很多老师、朋友和同学的关心与帮助, 让我不断成长. 在硕士论文即将完成之际, 我内心感慨颇多, 在此我要向所有在研究生期间给予我支持、鼓励和帮助的人表示我最真挚的感谢.

感谢我的导师李正帮老师. 在论文选题之后, 对论文如何构思、如何撰写, 李正帮老师都给予我耐心的指导和热情的帮助, 使我的毕业论文得以顺利的完成. 其次, 李正帮老师的严谨的治学态度和乐观的生活态度也深深地影响了我, 使我受益终生. 同时, 我还要感谢在硕士期间所有授予我学业的老师, 正是你们的悉心教导, 我才收获了知识, 才更有信心完成我的毕业论文.

感谢我同窗的小伙伴们, 在我撰写毕业论文的过程中能够给予我及时的建议与帮助, 帮助我解决一个个疑惑与难点. 除此之外, 我还要感谢我的同学朋友们, 千里迢迢来到陌生的城市上学, 你们给予了我太多的关怀和欢乐, 让我感受到了校园里的温暖.

感谢我的家人, 感谢父母对我的无私付出, 你们的默默支持让我能够安心地完成学业. 养育之恩, 无以回报, 你们的健康快乐就是我最大的期盼.

毕业在即, 在今后的工作和生活中, 我将牢牢谨记大家对我的教导, 继续努力. 再一次感谢所有给予我帮助的人, 祝愿你们身体健康, 万事顺意.

周雅诗

2021 年 5 月