高维协方差矩阵的检验综述及其推广

指导老师：王国长　　答辩人：丁文亮

1 绪论

1.1 研究背景和目的

1.2 国内外研究现状综述

（仅展示单个总体协方差矩阵的假设检验问题，两个或两个以上总体的假设检验综述见论文）

1.3 研究思路与框架（篇幅所限，详见论文）

1.1 研究背景和目的

随着科技和计算机的发展，在线收集数据得以实现，使得收集的数据在形式上、结构上和数量上发生了革命性的变化。

为了适应数据发展的要求，统计研究也应呈现相应的变化。传统的统计研究方法关注的是$p < n$的情形，当今的研究更多的是关注p与n相当（高维，$p/n \to c$）和$p \gg c$（超高维）的情形。

1.1 研究背景和目的

- 改进传统检验统计量
- 推广高维协方差矩阵假设检验类型
- 寻找高维其他条件下检验统计量构建的灵感

1.2 国内外研究现状综述（单个总体协方差矩阵的假设检验问题）

- 传统检验统计量（其中S指样本协方差矩阵）：
  - Nagao(1973)提出的V统计量：
$$\frac{1}{p}tr(S-I_p)^2 .$$
  - 似然比检验统计量（Anderson(2004)）：
$$n(trS-\log|S|-p) .$$

1.2 国内外研究现状综述（单个总体协方差矩阵的假设检验问题）

随着p增大，且$p/n \to c \in (0,+\infty)$，在原假设下似然比检验统计量对应的卡方检验不再适用

- Lediot & Wolf(2002)改进V统计量：

$$\frac{1}{p}tr(S-I_p)^2-\frac{p}{n-1}\left[\frac{1}{p}\cdot tr(S)\right]^2 .$$

1.2 国内外研究现状综述（单个总体协方差矩阵的假设检验问题）

放松正态假定，在总体分布峰度为3且$p/n \to c \in (0,1)$等前提下

- Bai et al.(2009)利用随机矩阵理论提出统计量：

$$\frac{1}{p}tr(B) - \frac{1}{p}\log|B| - 1 ,$$

其中$B = \frac{n-1}{n}S$.

- Jiang et al.(2012)利用Selberg不等式将其扩展到$p < n$且$p/n \to c \in (0,1]$的情形。

1.2 国内外研究现状综述（单个总体协方差矩阵的假设检验问题）

- Wang et al.(2013)在$p/n \to c \in (0,+\infty)$前提下提出检验统计量

$$\frac{1}{p}tr(S)-\frac{1}{p}\log|S|-1\ ,$$

相比于Bai et al.(2009)的检验统计量，该统计量更适用于均值向量未知且没有正态分布假定的假设检验问题。

1.2 国内外研究现状综述（单个总体协方差矩阵的假设检验问题）

在$p > n$的情形，样本协方差矩阵是奇异的，由于上述检验统计量包含$\log|S|$项不再适用。

- Chen et al.(2010)在$tr(\Sigma^4) = o(tr^2(\Sigma^2))$假设下，提出统计量

$$\frac{1}{p}T_1 - \frac{1}{p}T_2 - 1 \ ,$$

其中T_1和T_2分别为$tr(\Sigma)$和$tr(\Sigma^2)$的无偏估计。

- Ahmad & Rosen(2015)改进了上述的方法，用$tr(\Sigma)$和$tr(\Sigma^2)$的一致无偏估计代替上述的T_1和T_2 .

1.2 国内外研究现状综述（单个总体协方差矩阵的假设检验问题）

- 由于传统的似然比检验统计量是在固定p且$n \to \infty$情形所提出， Jiang & Yang(2013), Jiang & Qi(2015)和Chen & Jiang(2018)在正态假定和$p/n \to c \in (0,1]$情形，对多类假设检验问题上提出了似然比检验统计量的改进方法。

而本文针对高维协方差矩阵等于单位矩阵、对角矩阵和非负定矩阵（3种情形），提出新检验统计量，主要和Chen & Jiang(2018)的检验统计量进行对比。

2 单个高维总体协方差矩阵的检验

2.1 单个高维总体协方差矩阵等于给定非负定矩阵的检验

2.2 单个高维总体协方差矩阵等于给定单位矩阵的检验

2.3 单个高维总体协方差矩阵等于给定对角矩阵的检验

2.1 单个高维总体协方差矩阵等于给定非负定矩阵的检验($H_0: \Sigma = \Sigma_1\ v.s.\ H_1: \Sigma \neq \Sigma_1$)

Σ_1为给定$p \times p$非负定矩阵。则由奇异值分解$\Sigma_1 = U\Sigma_0 U'$，其中Σ_0为对角矩阵，U为正交矩阵.

方法1：对X左乘U^{-1}，得到$U^{-1} \cdot X \sim N_p(U^{-1}\mu, U^{-1} \cdot \Sigma \cdot U^{-1^T})$，进而转化为检验$U^{-1} \cdot X \sim N_p\left(U^{-1}\mu, U^{-1} \cdot U\Sigma_0 U' \cdot U^{-1^T}\right) = N_p(U^{-1}\mu, \Sigma_0)$的问题。

方法2：对X左乘${\Sigma_1}^{-\frac{1}{2}}$，得到${\Sigma_1}^{-\frac{1}{2}} \cdot X \sim N_p({\Sigma_1}^{-\frac{1}{2}}\mu, {\Sigma_1}^{-\frac{1}{2}} \cdot \Sigma \cdot {\Sigma_1}^{-\frac{1}{2}^T})$，进而转化为检验${\Sigma_1}^{-\frac{1}{2}} \cdot X \sim N_p\left(\mu{\Sigma_1}^{-\frac{1}{2}}, I_p\right)$的问题。

2.2 单个高维总体协方差矩阵等于给定单位矩阵的检验（$H_0: \Sigma = I_p \ v.s. H_1: \Sigma \neq I_p$）

假设$X = (X_1, X_2, \cdots, X_n)$是独立同分布于$N_p(\mu, \Sigma)$的实值随机向量，传统似然比检验统计量为：

$$\Lambda^* = \left(\frac{e}{m}\right)^{\frac{mp}{2}} e^{-\frac{1}{2}tr(A)} |A|^{\frac{m}{2}} .$$

使用**方法2**，则转化为单位矩阵的假设检验，**Chen & Jiang(2018) Corollary 1**提出，在一定条件下，当原假设$H_0: \Sigma = I_p$成立时，有：

$$\frac{\log \Lambda^* - \mu_{n,0}}{n \cdot \sigma_{n,0}} \xrightarrow{D} N(0,1) .$$

$$\mu_{n,0} = -\frac{1}{4}(n-1)\left[2p + (2n-2p-3)\log\left(1 - \frac{p}{n-1}\right)\right],$$

$${\sigma_{n,0}}^2 = -\frac{1}{2}\left[\frac{p}{n-1} + \log\left(1 - \frac{p}{n-1}\right)\right] .$$

对于方法2，回顾$\Sigma_1^{-\frac{1}{2}} \cdot X \sim N_p\left(\mu\Sigma_1^{-\frac{1}{2}}, I_p\right)$，我们构建的检验统计量$\frac{\log \Lambda^* - \mu_{n,0}}{n \cdot \sigma_{n,0}}$中，$\log \Lambda^*$由样本离差阵$A = \left(\Sigma_1^{-\frac{1}{2}} \cdot X\right)' (I_{n\times n} - \frac{1}{n} J_{n\times n})(\Sigma_1^{-\frac{1}{2}} \cdot X)$的特征值决定，$\mu_{n,0}$和$\sigma_{n,0}$由原假设$H_0: \Sigma = I_p$对应的$I_p$的特征值决定。

当Σ_1中含有较大的特征值时，因为$(\Sigma_1^{-\frac{1}{2}} \cdot X)$接近奇异阵，导致样本离差阵$A$会接近奇异阵（不满秩），这对检验统计量中$\log \Lambda^*$的构建产生大的影响（另一种说法，$\Lambda^*$中含有$|A|^{\frac{m}{2}}$），对数函数在区间$(0, 1]$要比$(1, +\infty)$陡峭，从而方法2检验效果不稳定。

当Σ_1中含有较大的特征值时，化成单位矩阵的检验有如下缺陷：

- 检验不等价
- 数值不稳定

为了克服这种检验缺陷，我们提出一种新的检验统计量，将假设检验推广到“对角矩阵”的情形。

2.3 单个高维总体协方差矩阵等于给定对角矩阵的检验（$H_0: \Sigma = \Sigma_0\ v.s.\ H_1: \Sigma \neq \Sigma_0$，$\Sigma_0$对角元为$0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_p < \infty$）

定理3：假设$p = p_n$，对任意$n \geq 3$，满足$n > p + 3$，且$\lim_{n\to\infty} \frac{p}{n} \to c \in (0,1]$。$\Lambda_m^* = \Lambda^*$，当$n \to \infty$，在原假设$H_0: \Sigma = \Sigma_0$成立时，有：

$$\frac{\log \Lambda_m^* - \mu_m}{m \cdot \sigma_m} \xrightarrow{D} N(0,1)$$

$$\mu_m = \frac{1}{4} m \left[(2m - 2p - 1) \cdot {r_m}^2 + 2 \sum_{i=1}^{p} \log \lambda_i - 2 \sum_{i=1}^{p} \lambda_i \right],$$

$${\sigma_m}^2 = \frac{1}{2} \left[{r_m}^2 - \frac{2}{m} \sum_{i=1}^{p} \left(\lambda_i - \frac{{\lambda_i}^2}{2} \right) \right], r_m = \left[-\log\left(1 - \frac{p}{m}\right) \right]^{\frac{1}{2}}$$

对于方法1，回顾$U^{-1}\cdot X\sim N_p(U^{-1}\mu,\Sigma_0)$，当$\Sigma_1$中含有较大的特征值时，化成对角矩阵的检验：我们构建的检验统计量$\frac{\log\Lambda_m^*-\mu_m}{m\cdot\sigma_m}$中，$\log\Lambda_m^*$由样本离差阵$A=(U^{-1}\cdot X)'(I_{n\times n}-\frac{1}{n}J_{n\times n})(U^{-1}\cdot X)$的特征值决定，$\mu_m$和$\sigma_m$由原假设$H_0$: $\Sigma=\Sigma_0$对应的Σ_0的特征值决定。

这个方法不会使样本离差阵A接近奇异阵（不满秩），保证了检验数值的稳定。

3 数值模拟

3.1 极限分布的模拟（略）

3.2 检验显著性水平和功效模拟（略）

（单个高维总体协方差矩阵等于给定单位矩阵）

3.3 检验显著性水平和功效模拟（略）

（单个高维总体协方差矩阵等于给定对角矩阵）

3.4 检验功效变化的数值模拟

（单个高维总体协方差矩阵等于给定对角矩阵）

3.3 检验显著性水平和功效模拟
（单个高维总体协方差矩阵等于给定对角矩阵）

选取样本量$n = 100$和变量个数$p = 60$，对于第k个数值实验（其中编号$k = 1,\cdots,60$，模拟次数都为10,000次），对于检验显著性水平的模拟，生成均值向量为零向量，协方差矩阵为$\Sigma_0 = diag\{1,\cdots,1000\}$（其中第$k$个数值实验$\Sigma_0$对角线中1000的个数为$k$）的$N_p(\mu,\Sigma_0)$随机数，检验水平设置在$\alpha = 0.05$；对于检验功效的模拟，生成均值向量为零向量，协方差矩阵为$\Sigma^{**} = diag\{1,\cdots,1200\}$（其中第$k$个数值实验$\Sigma_1$对角线中1200的个数为$k$）$N_p(\mu,\Sigma^{**})$随机数。

- T_1 为本文提出的定理3的检验统计量
- T_2 为Chen & Jiang(2018) Theorem 1的检验统计量
- T_3 表示使用方法2（化成单位矩阵检验问题）的检验统计量

3.3 检验显著性水平和功效模拟
（单个高维总体协方差矩阵等于给定对角矩阵）

表3-2 三种方法的检验统计量的检验显著性水平和检验功效对比

编号k	检验显著性水平（在H_0下）			检验功效（在H_1下）		
	$T_1\ size$	$T_2\ size$	$T_3\ size$	$T_1\ power$	$T_2\ power$	$T_3\ power$
1	0.0499	0.0471	0.0494	0.3117	0.3104	0.0545
2	0.0486	0.0481	0.0497	0.4880	0.4969	0.0520
3	0.0510	0.0478	0.0540	0.6416	0.6405	0.0521
4	0.0504	0.0498	0.0530	0.7489	0.7521	0.0549
5	0.0518	0.0425	0.0521	0.8366	0.8365	0.0564
6	0.0472	0.0476	0.0584	0.8961	0.8878	0.0535
7	0.0481	0.0475	0.0539	0.9346	0.9359	0.0604
8	0.0504	0.0455	0.0526	0.9565	0.9571	0.0607
9	0.0526	0.0503	0.0540	0.9725	0.9715	0.0581
10	0.0466	0.0475	0.0539	0.9814	0.9835	0.0608
11	0.0501	0.0468	0.0527	0.9881	0.9900	0.0641
12	0.0479	0.0519	0.0541	0.9941	0.9932	0.0666
13	0.0529	0.0511	0.0518	0.9960	0.9962	0.0616
14	0.0517	0.0464	0.0541	0.9972	0.9976	0.0629
15	0.0492	0.0488	0.0505	0.9986	0.9987	0.0713
16	0.0471	0.0478	0.0533	0.9988	0.9989	0.0689
17	0.0507	0.0468	0.0496	0.9993	0.9992	0.0678
18	0.0490	0.0524	0.0533	0.9998	0.9995	0.0796
19	0.0503	0.0508	0.0492	0.9997	0.9999	0.0753
20	0.0521	0.0464	0.0530	0.9998	0.9998	0.0748

- 对$T_1\ size$、$T_2\ size$和$T_3\ size$分别对应的60个值取均值，分别得到0.04983（0.0020）、0.0486（0.0024）和0.0530（0.0021）。因而$\frac{\log \Lambda_m^* - \mu_m}{m \cdot \sigma_m}$（标为$T_1$，定理3）的显著性水平是更接近理论显著性水平的，而且也更稳定。
- 当样本协方差矩阵含有较高的特征值时，方法2($T_3\ size$)的检验功效很低，而且随着所含较高特征值的个数的增加，检验功效的波动也很大，而$\frac{\log \Lambda_m^* - \mu_m}{m \cdot \sigma_m}$（标为$T_1$，定理3）和$\frac{\log \Lambda_n^* - \mu_n}{n \cdot \sigma_n}$（标为$T_2$，定理4）的检验功效都很高，且随着所含较高特征值个数的增加而迅速增加。

4 结论和展望

本文对现有的高维总体协方差矩阵假设检验的方法进行综述与回顾，并针对单个高维总体协方差矩阵等于单位矩阵、对角矩阵和对称正定矩阵的假设检验问题，分别提出检验方法。结果表明，该检验方法比传统的似然比检验有更好的检验显著性水平和检验功效。

并且对于协方差矩阵等于特定对称正定矩阵的检验问题，在样本协方差矩阵含有较大特征值时，提出了更稳健的检验流程。

4 结论和展望

- 对多种假设检验问题，本文的似然比检验框架下得到的检验统计量所含成分，对于在超高维等情形提出浓缩信息的统计量有启发作用。（样本协方差矩阵特征值的和、特征值的平方和）
- 本文假定样本服从多元正态分布假定，并且有$\lim_{n\to\infty}\frac{p}{n}\to c\in(0,1]$的限制，进一步研究可考虑放松正态假定，或改为更易验证的接近正态假定的条件；或改进检验统计量使之适用于$p>n$的情形。
  - ……

从论文到答辩，从综述到理论推导，对前辈学人的等身著作和读书治学精神愧不能表达万一——倘若有若干真实处，则是读者自己的学识之真实。