应用概率统计 第 36 卷
第 4 期 2020 年 8 月

Chinese Journal of Applied Probability and Statistics
Aug., 2020, Vol. 36, No. 4, pp. 355-364
doi: 10.3969/j.issn.1001-4268.2020.04.003

基于随机矩阵理论的高维数据球形检验 *

袁守成[1,2] 周 杰[1⋆] 沈洁琼[3]

([1]四川大学数学学院, 成都, 610064; [2]普洱学院数学与统计学院, 普洱, 665000)

([3]浙江大学宁波理工学院计算机与数据工程学院, 宁波, 315100)

摘 要: 本文基于随机矩阵理论, 研究了一般总体的高维协方差矩阵的球形检验. 当样本量小于数据维数时, 经典的似然比检验方法在球形检验中已无法使用. 通过引入样本协方差矩阵谱分布的高阶矩, 构造出一个新的检验统计量, 并给出其在零假设下的渐近分布. 模拟实验表明所提出的统计量在控制第一类错误概率的基础上能有效提高检验功效, 对于 Spiked 模型效果尤为显著.

关键词: 高维协方差矩阵; 球形检验; 谱分布; 检验功效

中图分类号: O212.1

英文引用格式: YUAN S C, ZHOU J, SHEN J Q. Sphericity test for high dimensional data based on random matrix theory [J]. Chinese J Appl Probab Statist, 2020, 36(4): 355–364. (in Chinese)

§1. 引 言

协方差矩阵的球形检验是多元统计中一个重要的统计推断问题, 在图像处理、信号处理和遗传基因学等诸多领域有着重要的应用. 在经典的大样本统计理论中, 当数据维数 p 确定而样本量 n 充分大时, 传统的似然比检验方法会取得良好的检验效果. 然而, 随着科技的进步, 数据采集变得更加容易, 需要处理的数据往往是高维数据. 由于这种数据的维数大于样本量, 使得似然比统计量中的样本协方差矩阵不可逆而无法进行正常检验. 甚至, 在高维情形下, 即使保证了样本协方差矩阵可逆, 经典的似然比统计量仍会有较差的检验效果. 从而有必要提出新的检验方法.

设 p 维总体 X 的均值为 $\mathbf{0}$, 协方差矩阵为 Σ_p, $X_1, X_2, \cdots, X_n$ 是来自 X 的一个样本, 其样本协方差矩阵为 $S_n = n^{-1}\sum\limits_{j=1}^{n} X_j X_j'$. 考虑以下球形检验:

$$H_0 : \Sigma_p = \sigma^2 I_p \quad \text{vs.} \quad H_1 : \Sigma_p \neq \sigma^2 I_p, \tag{1}$$

其中 $\sigma > 0$ 是一个未知常量, I_p 是 p 维单位矩阵. 对于检验问题 (1), 当 $p > n$ 时, 经典的似然比检验法[1] 已失效, 从而关于 $p > n$ 的研究一直是统计学家们感兴趣的, 也涌现出很多

*国家自然科学基金项目 (批准号: 61374027、11871357) 和云南省教育厅科学研究基金项目 (批准号: 2018JS516) 资助.

⋆通讯作者, E-mail: jzhou@scu.edu.cn.

本文 2019 年 3 月 21 日收到, 2019 年 7 月 16 日收到修改稿.

研究成果. Ledoit 和 Wolf[2] 在 n 和 p 都趋于无穷且同阶的框架下将球形检验推广为局部最优不变检验. Srivastava[3] 利用协方差矩阵特征值谱分布的前二阶矩 $\mathrm{tr}\Sigma_p^1$ 和 $\mathrm{tr}\Sigma_p^2$ 的无偏估计, 构造了与 Ledoit 和 Wolf 相似的统计量. Fisher 等[4] 采用与文献 [3] 类似的方法, 由协方差矩阵谱分布的前四阶矩的无偏估计, 构造出一个新的检验统计量, 使 Spiked 模型[5] 的检验功效有了一定改进. 以上检验方法都基于正态总体的假设. 对于非正态的情形, Chen 等[6] 利用 U 统计量给出的统计量在基因微阵列分析中具有较好的检验效果. Wang 和 Yao[7] 借助随机矩阵理论把 Ledoit 和 Wolf 给出的检验推广到非正态分布的情形. Tian 等[8] 通过结合文献 [3,4] 中两个统计量构造出一个稳健的检验统计量. 由文献 [4,9], 我们发现在构造球形检验的统计量时, 恰当地引入协方差矩阵谱分布的四阶矩对检验的功效会有一定的提升. 本文在构造检验问题 (1) 的统计量时, 合理地引入了样本协方差矩阵谱分布的四阶矩并利用随机矩阵理论给出其在零假设 H_0 下的渐近分布. 通过数值模拟研究发现, 我们提出的统计量比其它的统计量具有更好的检验功效, 对于 Spiked 模型功效格外显著.

§2. 预备知识

定义 1 设 $\mathscr{M}(\mathbb{C})$ 是由 $p\times p$ 复值矩阵构成的集合, 若 $A\in\mathscr{M}(\mathbb{C})$, 其特征值为 $\{\lambda_i\}_{1\leqslant i\leqslant p}$, 则称

$$F^A=\frac{1}{p}\sum_{i=1}^{p}\delta_{\lambda_i}$$

为复值矩阵 A 特征值的经验谱分布, 其中 δ_a 是在 a 点的狄拉克测度.

通常, 经验谱分布 F^A 是在复数域 $\mathbb{C}$ 上的概率测度. 如果 A 是 Hermite 矩阵 (或非负定的 Hermite 矩阵), 那么它的支撑落在 $\mathbb{R}$ (或 $\mathbb{R}^+$) 中.

定义 2 设 $\{A_n\}_{n\geqslant 1}$ 是一列来自 $\mathscr{M}(\mathbb{C})$ 的矩阵, 若它们的经验谱分布 $\{F^{A_n}\}_{n\geqslant 1}$ 弱收敛到一个测度 F, 则称 F 为矩阵列 $\{A_n\}$ 的极限谱分布.

设 H_p 和 F^{S_n} 分别是总体协方差矩阵 Σ_p 和样本协方差矩阵 S_n 的经验谱分布, 并设 F^{S_n} 在 n 和 p 都趋于无穷且同阶时的极限谱分布为 F_c. 定义 Σ_p 和 S_n 经验谱分布的整数阶矩为

$$\alpha_k=\int t^k\mathrm{d}H_p(t)=\frac{1}{p}\mathrm{tr}(\Sigma_p^k),\qquad \widehat{\beta}_k=\int t^k\mathrm{d}F^{S_n}(t)=\frac{1}{p}\mathrm{tr}(S_n^k),$$

其中 $k=0,1,2,\cdots$. 由文献 [7] 可知, α_1, α_2, α_3, α_4 的渐近无偏相合估计量分别可用 $\widehat{\beta}_1$, $\widehat{\beta}_2$, $\widehat{\beta}_3$, $\widehat{\beta}_4$ 的多项式表示为

$$\widehat{\alpha}_1=\widehat{\beta}_1,$$
$$\widehat{\alpha}_2=\frac{n^2}{(n-1)(n+2)}\big(\widehat{\beta}_2-c_n\widehat{\beta}_1^2\big),$$

$$\widehat{\alpha}_3=\frac{n^4}{(n-1)(n-2)(n+2)(n+4)}\big(\widehat{\beta}_3-3c_n\widehat{\beta}_2\beta_1+2c_n^2\widehat{\beta}_1^3\big),$$
$$\widehat{\alpha}_4=\tau\Big(\widehat{\beta}_4-4c_n\widehat{\beta}_3\widehat{\beta}_1-\frac{2n^2+3n-6}{n^2+n+2}c_n\widehat{\beta}_2^2+\frac{10n^2+12n}{n^2+n+2}c_n^2\widehat{\beta}_2\widehat{\beta}_1^2-\frac{5n^2+6n}{n^2+n+2}c_n^3\widehat{\beta}_1^4\Big),$$

其中 $c_n=p/n$, $\tau=n^5(n^2+n+2)/[(n+1)(n+2)(n+4)(n+6)(n-1)(n-2)(n-3)]$.

假设 3　样本量 n 和维数 p 都趋于无穷大, 并且 $c_n=p/n\to c\in(0,\infty)$.

假设 4　随机矩阵 $W_n=(w_{ij})_{1\leqslant i\leqslant p,1\leqslant j\leqslant n}$ 中的随机元素 w_{ij} 独立同分布, 且满足

$$\mathsf{E}(w_{ij})=0,\qquad \mathsf{E}(w_{ij}^2)=1,\qquad \mathsf{E}(w_{ij}^4)=3+\Delta<\infty.$$

并且 X_j, $j=1,2,\cdots,n$ 可表示为 $X_j=\Sigma_p^{1/2}w_{\cdot j}$, 其中 $w_{\cdot j}=(w_{1j},w_{2j},\cdots,w_{pj})'$.

假设 5　$\|\Sigma_p\|_2$ 一致有界, 并且当 $p\to\infty$ 时, H_p 弱收敛到谱分布 H.

引理 6　设随机矩阵 $\boldsymbol{X}$ 满足假设 3–5, 已知 $f_1,f_2,\cdots,f_k$ 是一列包含 F_c 支撑的开区域上解析的函数, 那么随机向量 $(X_n(f_1),X_n(f_2),\cdots,X_n(f_k))'$ (其中 $X_n(f)=p[F^{S_n}(f)-F_{c_n}(f)]$) 弱收敛到一个高斯随机向量 $(X_{f_1},X_{f_2},\cdots,X_{f_k})'$, 其均值和协方差矩阵为

$$\mathsf{E}(X_{f_i})=I_1(f_i)+I_2(f_i)\cdot\Delta,\qquad \mathsf{Cov}\,(X_{f_i},X_{f_j})=J_1(f_i,f_j)+J_2(f_i,f_j)\cdot\Delta,$$

其中

$$I_1(f_i)=\frac{1}{2\pi i}\oint\frac{c\{\underline{s}(z)/[1+\underline{s}(z)]\}^3f_i(z)}{\{1-c\{\underline{s}(z)/[1+\underline{s}(z)]\}^2\}^2}\mathrm{d}z,$$
$$I_2(f_i)=-\frac{1}{2\pi i}\oint\frac{c\{\underline{s}(z)/[1+\underline{s}(z)]\}^3f_i(z)}{1-c\{\underline{s}(z)/[1+\underline{s}(z)]\}^2}\mathrm{d}z,$$
$$J_1(f_i,f_j)=-\frac{1}{2\pi^2}\oint\oint\frac{f_i(z_1)f_j(z_2)}{[\underline{s}(z_1)-\underline{s}(z_2)]^2}\underline{s}'(z_1)\underline{s}'(z_2)\mathrm{d}z_1\mathrm{d}z_2,$$
$$J_2(f_i,f_j)=-\frac{1}{4\pi^2}\oint f_i(z_1)\frac{\partial}{\partial z_1}\frac{\underline{s}(z_1)}{1+\underline{s}(z_1)}\mathrm{d}z_1\cdot\oint f_j(z_2)\frac{\partial}{\partial z_2}\frac{\underline{s}(z_2)}{1+\underline{s}(z_2)}\mathrm{d}z_2,$$

其中 $1\leqslant i,j\leqslant k$, $\Delta=\mathsf{E}(w_{11}^4-3)$, 复积分的轨道是沿着包含 F_c 支撑的开区间的边界 (在项 J_1 中是不交的).

引理 6 的证明过程详见文献 [10].

在引理 6 中, $\underline{s}(z)$ 与样本协方差矩阵 S_n 谱分布的 Stieltjes 变换[11] $s(z)$ 有如下关系:

$$\underline{s}(z)=-\frac{1-c}{z}+cs(z),\qquad z\in\mathbb{C}^+.$$

一般地, 复积分的轨道是不确定的, 具体应用不是那么容易. 如果把不确定的复积分轨道转换在一个单位圆上, 那么可使计算变得相对简单.

引理 7 引理 6 中的极限参数 I_1, I_2, J_1, J_2 可以表示为

$$I_1(f)=\lim_{r\downarrow 1}\frac{1}{2\pi i}\oint_{|\xi|=1}f(|1+h\xi|^2)\Big(\frac{\xi}{\xi^2-r^{-2}}-\frac{1}{\xi}\Big)\mathrm{d}\xi, \tag{2}$$

$$I_2(f)=\frac{1}{2\pi i}\oint_{|\xi|=1}f(|1+h\xi|^2)\frac{1}{\xi^3}\mathrm{d}\xi, \tag{3}$$

$$J_1(f,g)=\lim_{r\downarrow 1}-\frac{1}{2\pi^2}\oint_{|\xi_1|=1}\oint_{|\xi_2|=1}\frac{f(|1+h\xi_1|^2)g(|1+h\xi_2|^2)}{(\xi_1-r\xi_2)^2}\mathrm{d}\xi_1\mathrm{d}\xi_2, \tag{4}$$

$$J_2(f,g)=-\frac{1}{4\pi^2}\oint_{|\xi_1|=1}\frac{f(|1+h\xi_1|^2)}{\xi_1^2}\mathrm{d}\xi_1\oint_{|\xi_2|=1}\frac{f(|1+h\xi_2|^2)}{\xi_2^2}\mathrm{d}\xi_2, \tag{5}$$

其中 $h=\sqrt{c}$, $r>1$ 并接近于 1.

引理 7 的证明过程详见文献 [12].

§3. 主要结果

对于球形检验问题 (1), 注意到在尺度变换: $x\mapsto ax$ $(a\neq 0)$ 和正交变换: $x\mapsto G(x)$ (G 属于正交变换群) 下是不变的. 为不失一般性, 我们设 $\Sigma_p=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_p)$, $\lambda_i>0$, $i=1,2,\cdots,p$.

由 Cauchy-Schwartz 不等式可知,

$$\Big(\sum_{i=1}^{p}\lambda_i\times 1\Big)^2\leqslant p\times\sum_{i=1}^{p}\lambda_i^2, \tag{6}$$

当且仅当 $\lambda_1=\lambda_2=\cdots=\lambda_p$ 时等号成立. (6) 式两边同时平方取倒数, 再乘以 $p^{-1}\sum\limits_{i=1}^{p}\lambda_i^4$, 整理为

$$\Big(\sum_{i=1}^{p}\lambda_i^4/p\Big)\Big/\Big(\sum_{i=1}^{p}\lambda_i/p\Big)^4\geqslant\Big(\sum_{i=1}^{p}\lambda_i^4/p\Big)\Big/\Big(\sum_{i=1}^{p}\lambda_i^2/p\Big)^2.$$

记

$$\phi\triangleq\Big(\sum_{i=1}^{p}\lambda_i^4/p\Big)\Big/\Big(\sum_{i=1}^{p}\lambda_i/p\Big)^4-\Big(\sum_{i=1}^{p}\lambda_i^4/p\Big)\Big/\Big(\sum_{i=1}^{p}\lambda_i^2/p\Big)^2=\frac{\alpha_4}{\alpha_1^4}-\frac{\alpha_4}{\alpha_2^2}\geqslant 0, \tag{7}$$

可知 $\phi=0$ 当且仅当零假设 H_0 成立. 从而, 构造检验统计量

$$\gamma=\frac{\widehat{\alpha}_4}{\widehat{\alpha}_1^4}-\frac{\widehat{\alpha}_4}{\widehat{\alpha}_2^2}.$$

其中 $\widehat{\alpha}_1$, $\widehat{\alpha}_2$, $\widehat{\alpha}_4$ 是 α_1, α_2, α_4 的相合估计. 从 (7) 可以看出, 零假设 $\Sigma=\sigma^2I_p$ 中 σ 的取值大小对检验问题 (1) 不产生影响, 故不妨设 $\sigma=1$.

定理 8　在假设 3–5 下, 当零假设 H_0 成立时, 有

$$T=\frac{n\gamma-\widetilde{\mu}}{\sqrt{\widetilde{\Gamma}}}\xrightarrow{D}\mathrm{N}(0,1),$$

其中 $\widetilde{\mu}=2\Delta$, $\widetilde{\Gamma}=16+16\Delta/c$.

证明:　当 $\Sigma_p=I_p$ 时, 由引理 6 可得,

$$p\begin{pmatrix}\widehat{\beta}_1-1\\ \widehat{\beta}_2-(1+c_n)\\ \widehat{\beta}_3-(1+3c_n+c_n^2)\\ \widehat{\beta}_4-(1+6c_n+6c_n^2+c_n^3)\end{pmatrix}\xrightarrow{D}\mathrm{N}_4(\boldsymbol{m},\boldsymbol{V}),$$

其中 $\boldsymbol{m}$ 和 $\boldsymbol{V}$ 分别可由 (2)、(3) 和 (4)、(5) 得到. 经计算

$$\boldsymbol{m}=(1+\Delta)(0,c,3c+3c^2,6c+17c^2+6c^3)',$$

正定矩阵 $\boldsymbol{V}=(v_{ij})_{4\times4}$, 其元素为

$$\begin{aligned}
&v_{11}=2c+c\Delta,\qquad v_{12}=4c(c+1)+2c(c+1)\Delta,\\
&v_{13}=6c(c^2+3c+1)+3c(c^2+3c+1)\Delta,\\
&v_{14}=8c(c^3+6c^2+6c+1)+4c(c^3+6c^2+6c+1)\Delta,\\
&v_{22}=4c(2c^2+5c+2)+4c(c^2+2c+1)\Delta,\\
&v_{23}=12c(c^3+5c^2+5c+1)+6c(c^3+4c^2+4c+1)\Delta,\\
&v_{24}=8c(2c^4+17c^3+32c^2+17c+1)+8c(c^4+7c^3+12c^2+7c+1)\Delta,\\
&v_{33}=6c(3c^4+24c^3+46c^2+24c+3)+9c(c^4+6c^3+11c^2+6c+1)\Delta,\\
&v_{34}=24c(c^5+12c^4+37c^3+37c^2+12c+1)+12c(c^5+9c^4+25c^3+25c^2+9c+1)\Delta,\\
&v_{44}=8c(4c^6+66c^5+300c^4+485c^3+300c^2+66c+4)\\
&\qquad+16c(c^6+12c^5+48c^4+74c^3+48c^2+12c+1)\Delta.
\end{aligned}$$

由于 $\widehat{\alpha}_1$, $\widehat{\alpha}_2$, $\widehat{\alpha}_3$, $\widehat{\alpha}_4$ 是 $\widehat{\beta}_1$, $\widehat{\beta}_2$, $\widehat{\beta}_3$, $\widehat{\beta}_4$ 的多项式函数, 利用 Delta 方法[13], 可得

$$n(\widehat{\alpha}_1-1,\widehat{\alpha}_2-1,\widehat{\alpha}_4-1)'\xrightarrow{D}\mathrm{N}_3(\mu,\Gamma),$$

其中

$$\begin{aligned}
&\mu=\Delta\cdot(0,1,c+6)',\\
&\Gamma=\frac{1}{c}\begin{pmatrix}2+\Delta & 4+2\Delta & 8+4\Delta\\ 4+2\Delta & 4(c+2)+4\Delta & 8(3c+2)+8\Delta\\ 8+4\Delta & 8(3c+2)+8\Delta & 8(c^3+12c^2+18c+4)+16\Delta\end{pmatrix}.
\end{aligned}$$

设 $\boldsymbol{t}=(x,y,w)'$, 定义函数

$$G(\boldsymbol{t})=\frac{w}{x^4}-\frac{w}{y^2}.$$

显然, $G(\boldsymbol{t})$ 在点 $\boldsymbol{t}_0=(1,1,1)'$ 处连续可微. 利用 Delta 方法知

$$n[\gamma-G(\boldsymbol{t}_0)]\xrightarrow{D}\mathrm{N}(\widetilde{\mu},\widetilde{\Gamma}),$$

其中 $\widetilde{\mu}=J'(\boldsymbol{t}_0)\mu$, $\widetilde{\Gamma}=J'(\boldsymbol{t}_0)\Gamma J(\boldsymbol{t}_0)/c^2$, 且

$$J(\boldsymbol{t}_0)=\frac{\partial G(\boldsymbol{t})}{\partial\boldsymbol{t}}\Big|_{\boldsymbol{t}=\boldsymbol{t}_0}=(-4,2,0)'.$$

经过简单计算, $G(\boldsymbol{t}_0)=0$, $\widetilde{\mu}=2\Delta$, $\widetilde{\Gamma}=16+16\Delta/c$, 即

$$\frac{n\gamma-\widetilde{\mu}}{\sqrt{\widetilde{\Gamma}}}\xrightarrow{D}\mathrm{N}(0,1).$$

定理 8 得证. □

当 $\mathsf{E}(w_{11}=3)$ 时, $\Delta=0$, 由定理 8 可知, $\widetilde{\mu}=0$, $\widetilde{\Gamma}=16$, 从而可得如下定理.

定理 9 在假设 3–5 下, 如果原假设 H_0 成立且 $\mathsf{E}(w_{11}^4)=3$, 那么

$$\frac{n}{4}\gamma\xrightarrow{D}\mathrm{N}(0,1).$$

定理 9 包含了总体是正态分布的情形.

§4. 数值模拟

这一节主要通过数值模拟阐明我们提出的检验统计量在高斯分布和非高斯分布两种不同场景下所得结论的正确性和在第一类错误控制上的表现; 并且通过两个不同的模型比较出所得统计量与其它统计量检验功效的大小.

记 Srivastava[3] 给出的统计量为 T_{s}, Fisher 等[4] 给出的统计量为 T_{f}, Wang 和 Yao[7] 给出的统计量为 T_{c}, Tian 等[8] 给出的统计量为 T_{m}. 我们采用蒙特卡洛法进行数值模拟, 在相应的设定下独立重复模拟 2 000 次. 首先, 考虑以下两种情形:

情形 1 $w_{\cdot j}\sim\mathrm{N}(0,I_p)$ 且 $\Sigma=I_p$.

情形 2 $w_{\cdot j}=(w_{1j},w_{2j},\cdots,w_{pj})'$, w_{ij} 互相独立且服从标准化的 Gamma 分布$\Gamma(4,2)-2$, $i=1,2,\cdots,p$ 且 $\Sigma_p=I_p$.

情形 1 对应正态总体, $\Delta=0$; 情形 2 对应非正态总体, $\Delta=1.5$. 对于这两种情形, 分别取样本量 $n=200$, 数据维数 $p=400$, 它们的正态 QQ 图如下所示:

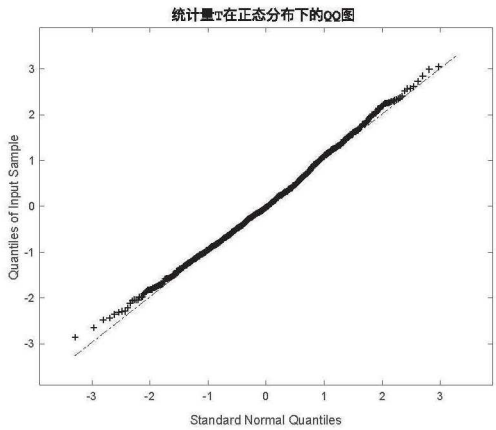

图 1　统计量 T 在情形 1 下的 QQ 图

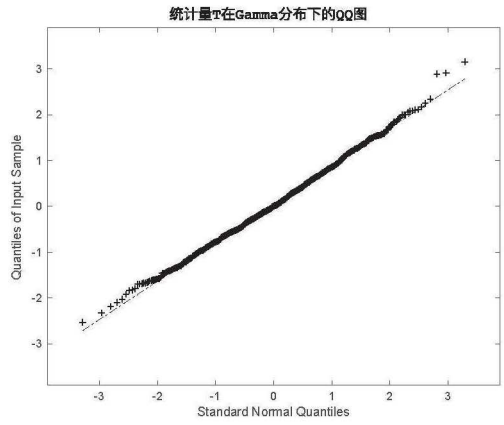

图 2　统计量 T 在情形 2 下的 QQ 图

从图 1 和图 2 可以看出在原假设 H_0 下, 不论研究总体服从正态分布还是 Gamma 分布, 给出的检验统计量都具有渐近正态性, 与所给定理相符合.

取检验的显著性水平 $\alpha = 0.05$, 样本量 n 分别取为 20, 40, 60, 80, 100, 120, $c = p/n$ 取为 1, 2, 5. 对于情形 1 和情形 2, 这些统计量在第一类错误上的控制如表 1 和表 2 所示.

表 1　对于情形 1 统计量 $T_{\rm s}$、$T_{\rm f}$ 与 T 的在第一类错误控制上的比较

$p = cn$	$c = 1$			$c = 2$			$c = 5$		
	$T_{\rm s}$	$T_{\rm f}$	T	$T_{\rm s}$	$T_{\rm f}$	T	$T_{\rm s}$	$T_{\rm f}$	T
$n = 20$	0.0635	0.0300	0.0645	0.0595	0.0380	0.0640	0.0530	0.0425	0.0595
$n = 40$	0.0595	0.0475	0.0575	0.0420	0.0430	0.0565	0.0475	0.0535	0.0520
$n = 60$	0.0480	0.0575	0.0530	0.0560	0.0555	0.0570	0.0520	0.0485	0.0490
$n = 80$	0.0530	0.0525	0.0545	0.0505	0.0390	0.0510	0.0475	0.0500	0.0525
$n = 100$	0.0505	0.0535	0.0520	0.0465	0.0500	0.0515	0.0510	0.0480	0.0515
$n = 120$	0.0545	0.0505	0.0535	0.0490	0.0540	0.0480	0.0495	0.0455	0.0480

表 2　对于情形 2 统计量 $T_{\rm c}$、$T_{\rm m}$ 与 T 的在第一类错误控制上的比较

$p = cn$	$c = 1$			$c = 2$			$c = 5$		
	$T_{\rm c}$	$T_{\rm m}$	T	$T_{\rm c}$	$T_{\rm m}$	T	$T_{\rm c}$	$T_{\rm m}$	T
$n = 20$	0.0705	0.1165	0.0705	0.0725	0.0985	0.0740	0.0805	0.1050	0.0745
$n = 40$	0.0760	0.1060	0.0735	0.0595	0.0850	0.0690	0.0655	0.0985	0.0650
$n = 60$	0.0680	0.0945	0.0630	0.0600	0.0805	0.0640	0.0570	0.0845	0.0560
$n = 80$	0.0590	0.0870	0.0495	0.0535	0.0755	0.0585	0.0590	0.0740	0.0520
$n = 100$	0.0550	0.0755	0.0480	0.0570	0.0645	0.0560	0.0595	0.0680	0.0495
$n = 120$	0.0530	0.0645	0.0465	0.0565	0.0660	0.0505	0.0560	0.0620	0.0485

由表 1 和表 2 可以清晰地看出: 在情形 1 中, 随着 n 和 p 的不断增大, 统计量 $T_{\rm s}$、$T_{\rm f}$ 与 T 的第一类错误概率越来越接近 α; 在情形 2 中, 当 n 和 p 的不断增大时, 统计量 $T_{\rm s}$ 与T 的第一类错误概率同样越来越接近 α, 比统计量 $T_{\rm m}$ 的控制效果更理想.

其次, 我们通过以下两个模型:

模型 1 $\Sigma_p = I_p + \sqrt{c}\cdot \mathrm{diag}(\underbrace{1.2,\cdots,1.2}_{n/10},\underbrace{0,\cdots,0}_{p-n/10})$;

模型 2 $\Sigma_p = I_p + \sqrt{c}\cdot \mathrm{diag}(\underbrace{0.5,\cdots,0.5}_{p/2},\underbrace{0,\cdots,0}_{p/2})$,

来考察统计量 T 的检验功效. 同样取样本量 $n = 20, 40, 60, 80, 100, 120$, $c = p/n$, c 取 1, 2, 5, 显著性水平设定为 0.05. 它们的功效比较如表 3 和 4 所示.

表 3 当总体分布为正态分布时统计量 $T_{\rm s}$、$T_{\rm f}$ 与 T 的经验功效比较

$p=cn$	$c=1$			$c=2$			$c=5$		
				模型 1					
	$T_{\rm s}$	$T_{\rm f}$	T	$T_{\rm s}$	$T_{\rm f}$	T	$T_{\rm s}$	$T_{\rm f}$	T
$n=20$	0.2640	0.2025	0.2855	0.3160	0.2565	0.3365	0.3480	0.2810	0.3745
$n=40$	0.6015	0.4970	0.6400	0.6360	0.6150	0.7225	0.6915	0.6685	0.8000
$n=60$	0.8675	0.7530	0.8875	0.9105	0.8685	0.9460	0.9410	0.9095	0.9795
$n=80$	0.9750	0.9215	0.9785	0.9930	0.9700	0.9975	0.9975	0.9840	1
$n=100$	0.9930	0.9830	0.9965	0.9985	0.9960	0.9995	1	0.9980	1
$n=120$	1	0.9995	1	1	1	1	1	1	1
				模型 2					
	$T_{\rm s}$	$T_{\rm f}$	T	$T_{\rm s}$	$T_{\rm f}$	T	$T_{\rm s}$	$T_{\rm f}$	T
$n=20$	0.2780	0.1420	0.2450	0.4590	0.1305	0.4455	0.7180	0.1390	0.6890
$n=40$	0.6305	0.2185	0.5735	0.9205	0.2630	0.9050	0.9975	0.2620	0.9545
$n=60$	0.9120	0.4340	0.8960	0.9965	0.4725	0.9925	1	0.4585	0.9995
$n=80$	0.9955	0.6130	0.9930	1	0.6565	1	1	0.6350	1
$n=100$	0.9990	0.7700	0.9995	1	0.8350	1	1	0.97770	1
$n=120$	1	0.8970	1	1	0.9330	1	1	0.8735	1

由表 3 和表 4 可以发现, 当样本容量 n 不断增大时, 检验功效都逐渐接近 1. 在模型 1 (主对角元素非 1 的个数较少), 即 Spiked 结构模型中, 不论总体服从正态分布还是 Gamma 分布, 我们提出的检验统计量 T 要比其它的检验统计量功效更大, 从而检验效果也更好. 在模型 2 (主对角元素非 1 的个数较多) 中, 我们提出的统计量 T 的检验功效明显高于 $T_{\rm f}$, 略小于 $T_{\rm s}$.

表 4　当总体分布为 Gamma 分布时统计量 $T_{\rm c}$、$T_{\rm m}$ 与 T 的经验功效比较

$p=cn$	$c=1$			$c=2$			$c=5$		
				模型 1					
	$T_{\rm c}$	$T_{\rm m}$	T	$T_{\rm c}$	$T_{\rm m}$	T	$T_{\rm c}$	$T_{\rm m}$	T
$n=20$	0.2100	0.1615	0.2255	0.2810	0.2020	0.2840	0.2805	0.1935	0.3085
$n=40$	0.4885	0.3210	0.4985	0.6035	0.4785	0.6595	0.6355	0.5475	0.7115
$n=60$	0.7785	0.6575	0.7900	0.8730	0.7785	0.9080	0.8940	0.8610	0.9380
$n=80$	0.9380	0.8590	0.9465	0.9750	0.9590	0.9885	0.9880	0.9785	0.9980
$n=100$	0.9880	0.9630	0.9930	0.9990	0.9950	0.9990	1	0.9990	1
$n=120$	1	0.9990	1	1	1	1	1	1	1
				模型 2					
	$T_{\rm c}$	$T_{\rm m}$	T	$T_{\rm c}$	$T_{\rm m}$	T	$T_{\rm c}$	$T_{\rm m}$	T
$n=20$	0.2390	0.1170	0.2140	0.4130	0.2245	0.3655	0.6890	0.4070	0.4215
$n=40$	0.5655	0.2820	0.5060	0.9105	0.6895	0.8540	0.9960	0.9555	0.9320
$n=60$	0.8855	0.6505	0.8450	0.9960	0.9695	0.9875	1	1	0.9990
$n=80$	0.9870	0.9160	0.9745	1	1	1	1	1	1
$n=100$	0.9990	0.9880	0.9990	1	1	1	1	1	1
$n=120$	1	1	1	1	1	1	1	1	1

§5. 结　　语

面对高维数据的球形检验, 经典的似然比方法已无法正常使用或是检验效果不好, 需要提出新的检验方法. 我们基于随机矩阵理论, 通过引入样本协方差矩阵谱分布的高阶矩, 合理地构造出一个新的检验统计量, 并求出其在零假设下的渐近分布. 模拟实验表明, 我们给出的统计量具有良好的渐近正态性, 并且在较好地控制第一类错误概率的基础上, 不论用于检验的高维数据服从正态分布还是其它分布, 与其它检验统计量比较功效会有一定的提高, 尤其对于 Spiked 模型 (协方差矩阵主对角元素非 1 个数较少) 功效更加显著, 检验效果更好.

参 考 文 献

[1] ANDERSON T W. *An Introduction to Multivariate Statistical Analysis* [M]. 3rd ed. New York: Wiley, 2003.

[2] LEDOIT O, WOLF M. Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size [J]. *Ann Statist*, 2002, **30(4)**: 1081–1102.

[3] SRIVASTAVA M S. Some tests concerning the covariance matrix in high dimensional data [J]. *J Japan Statist Soc*, 2005, **35(2)**: 251–272.

[4] FISHER T J, SUN X Q, GALLAGHER C M. A new test for sphericity of the covariance matrix for high dimensional data [J]. *J Multivariate Anal*, 2010, **101(10)**: 2554–2570.

[5] JOHNSTONE I M. On the distribution of the largest eigenvalue in principal components analysis [J]. *Ann Statist*, 2001, **29(2)**: 295–327.

[6] CHEN S X, ZHANG L X, ZHONG P S. Tests for high-dimensional covariance matrices [J]. *J Amer Statist Assoc*, 2010, **105(490)**: 810–819.

[7] WANG Q W, YAO J F. On the sphericity test with large-dimensional observations [J]. *Electron J Stat*, 2013, **7**: 2164–2192.

[8] TIAN X T, LU Y T, LI W M. A robust test for sphericity of high-dimensional covariance matrices [J]. *J Multivariate Anal*, 2015, **141**: 217–227.

[9] FISHER T J. On testing for an identity covariance matrix when the dimensionality equals or exceeds the sample size [J]. *J Statist Plann Inference*, 2012, **142(1)**: 312–326.

[10] PAN G M, ZHOU W. Central limit theorem for signal-to-interference ratio of reduced rank linear receiver [J]. *Ann Appl Probab*, 2008, **18(3)**: 1232–1270.

[11] BAI Z D, SILVERSTEIN J W. *Spectral Analysis of Large Dimensional Random Matrices* [M]. 2nd ed. New York: Springer-Verlag, 2009.

[12] 王勤文. 大维 Spike 模型下的极限谱性质 [D]. 杭州: 浙江大学, 2015.

[13] BILODEAU M, BRENNER D. *Theory of Multivariate Statistics* [M]. New York: Springer-Verlag, 1999.

Sphericity Test for High Dimensional Data Based on Random Matrix Theory

YUAN Shoucheng[1,2] ZHOU Jie[1] SHEN Jieqiong[3]

([1] *College of Mathematics, Sichuan University, Chendu, 610064, China*)

([2] *College of Mathematics and Statistical Science, Puer University, Puer, 665000, China*)

([3] *School of Computer and Data Engineering, Zhejiang University Ningbo Institute of Technology, Ningbo, 315100, China*)

Abstract: In this article we study test of sphericity for high-dimensional covariance matrix in the general population based on random matrix theory. When the sample size is less than data dimension, the classical likelihood ratio test has poor performance for test of sphericity. Thus, we propose a new statistic for test of sphericity by using the higher moments of spectral distribution of the sample covariance matrix, and derive the asymptotic distribution of the statistic under the null hypothesis. Simulation results show that the proposed statistics can effectively improve the power of the test of sphericity for high dimensional data, and have especially significant effects for Spiked model, on the basis of controlling the type-one error probability.

Keywords: high-dimensional covariance matrix; sphericity test; spectral distribution; test power

2010 Mathematics Subject Classification: 62N03