上海交通大学硕士学位论文

高维数据的正态性假设检验

硕士研究生：李扬

学　　号：118071910043

导　　师：王成

申请学位：应用统计专业硕士学位

学　　科：应用统计

所在单位：数学科学学院

答辩日期：2020 年 5 月

授予学位单位：上海交通大学

Dissertation Submitted to Shanghai Jiao Tong University
for the Degree of Master

Normality hypothesis testing of high-dimensional data

Candidate：	LiYang
Student ID:	118071910043
Supervisor：	WangCheng
Academic Degree Applied for：	Master of Applied Statistics
Speciality：	Applied Statistics
Affiliation：	School of Mathematical Sciences
Date of Defence：	May, 2020
Degree-Conferring-Institution：	Shanghai Jiao Tong University

高维数据的正态性假设检验

摘 要

正态分布凭借其优良的性质被广泛应用到多种统计理论和方法中，因此检验数据的正态性是非常有必要的。近年来，越来越多的实际问题涉及到高维数据，这也让正态性假设检验问题一直充满活力。

本文从单变量正态性检验着手，总结了四大类检验方法的基本原理、优缺点和适用的数据类型等。本文将高维数据的正态性检验方法划分为基于统计图的检验、多个单变量的检验、合适的降维方法以及直接对高维数据进行检验。本文还归纳了单变量的正态性检验方法推广到高维情况下的常用手段。然后通过计算机模拟比较多种高维正态性检验方法的性能，得到的结论包括：高维正态性检验容易出现第一类错误不可控的情况，提高样本量和维数的比值是解决该问题的有效方法之一；HZ 检验和独立化变换后的 JB 检验效果最好等。最后对两个实际数据集检验其正态性，结果均为拒绝原假设，进一步表明在实际研究中对正态性假设条件进行验证的必要性。

关键词：高维正态分布　假设检验　第一类错误

Normality hypothesis testing of high-dimensional data

ABSTRACT

The normal distribution is widely used in many statistical theories and methods due to its excellent properties, so it is necessary to test whether a dataset follows the normal distribution. In recent years, more and more practical problems use high-dimensional data, which made the normality hypothesis testing problem always full of vitality.

This article starts with the univariate normality tests and summarizes the basic principles, advantages and disadvantages, and applicable data types of the four major test. In this paper, the normality testing methods of high-dimensional data are divided into four categories: statistical chart testing, multiple univariate testing, suitable dimensionality reduction methods, and direct testing of high-dimensional data. In addition, this article summarizes how the univariate normality test is extended to the high-dimensional case. Then the performances of various high -dimensional normality testing methods are compared through computer simulations. The conclusions include: some high-dimensional normality tests can't control the type I error, one of the effective strategies is increasing the ratio of sample size and dimension; the HZ test and the JB test after independent transformation performed best, etc. Finally, this paper tests the normality of the two actual datasets, and the results are both rejecting the null hypothesis, which shows the necessity of verifying the normality hypothesis condition in actual research.

KEY WORDS: High-dimensional normal distribution, hypothesis test, the type I error.

目 录

第一章 绪论

1.1 研究意义

正态分布凭借其优良的性质出现在众多统计学理论和方法中，是统计学科中非常重要的分布。正态分布的优良性质包括：一、分布特征由两个参数唯一确定；二、许多分布的极限分布恰为正态分布，例如t分布，泊松分布，均值分布（中心极限定理）等；三、对于非正态性的数据有许多正态化变换方法，例如对数变换、平方根变换、Box-Cox 变换等。

正态分布是许多分布的基础，从单变量来看，t分布，F分布和卡方分布都是由正态分布衍生而来；从多变量来看，威沙特分布，霍特林分布，威尔克斯分布是由多元正态分布衍生而来。此外，由于正态分布的优良性质，许多经典的统计方法将正态分布作为它们的基本假设，并在此基础上加以研究，例如卡方检验、t检验、F检验、回归分析、区间估计、方差分析、线性判别分析等。近年来，很多研究中也用到了正态假设，比如基于高斯图模型的网络推断[1-7]、高维线性判别分析[8-9]、变量选择后的回归模型推断[10-11]、高维数据的变点分析[12-13]等。

一般来说，许多数据无法严格服从正态分布，但是如果数据本身偏离正态分布的程度较大，尤其是在样本较小的情况下，统计推断的准确性则会付出一定的代价，例如可能使得某些模型的预测结果产生较大的偏差。然而，有的研究中并未讨论数据的分布形态，而直接假定数据服从正态分布并进行下一步的研究。实际上，研究者对正态性假设是否合理并没有充分的把握，从而也影响了研究结果的可信度。这时应该基于实际的观测数据实施正态性检验，因此正态性检验一直以来都是统计学中较为重要的问题。

当数据为单变量时，有许多经典的方法可以做正态性检验，如 Q-Q 图、卡方检验、Shapiro-Wilk 检验。但是，近年来越来越多的实际问题涉及到高维数据，这一现状对高维数据的正态性假设检验方法的性能也提出了越来越高的要求。然而并非所有的单变量正态性检验方法都适用于高维数据，随着数据的维数升高，也会出现许多问题，比如很难通过统计图直观地判断是否服从正态分布，许多检验统计量的计算量也会随着维数的升高而大幅增加，第一类错误不可控等等。因此，对高维数据的正态性检验问题是大数据时代面临的一个新的挑战。

1.2 文献综述

正态分布被应用在许多场景中，而检验样本数据是否来自正态分布的问题已经被多代统计学家广泛研究，学者们通常更喜欢使用检验效果好且容易理解的正态性检验方法[14]。至今为止，学者们从不同的角度提出了多种检验单变量正态性的方法，然而在涉及到高维数据的正态性检验时，则对检验的性能提出了一个新的挑战[15-18]。一般来说，正态性检验方法的研究可以基于以下几个方向进行。

基于统计图的正态性检验，分为两大类，不基于任何假定的统计图和基于正态分布假定的统计图。前者具有统计描述功能，包括直方图、茎叶图和箱线图等；后者则以 P-P 图和 Q-Q 图为代表。通过统计图来判断样本是否来自正态总体具有主观性，后来衍生出许多将统计图所传达的信息量化的方法，尤其是基于 Q-Q 图的量化方法。D'Agostino 和 Stephens[19]在 1987 年提出通过线性回归方法检验统计图中的散点是否分布在直线附近。在此基础上，Coin[20]在 2008 年建议对 Q-Q 图进行多项式回归，并详细探究三次回归系数的变化规律，提出将三次项回归系数作为检验统计量。此外，Richard 等人[21]在 2007 年建议用相关系数量化 Q-Q 图中散点偏离直线的程度，并给出了相应的临界值表。著名的 Shapiro-Wilk 检验[22]（SW 检验）在 1965 年提出，它可以检验多种分布，通过正态概率图的一次回归方程推导出正态假设下的方差估计值，数学表达形式为排序数据的加权平方和，然后计算它与样本方差的比值来衡量正态性。Shapiro 和 Francia[23]在 1972 年简化了权重系数，得到改进的 KS 检验。Royston[24]于 1983 年通过将 SW 检验应用到每一维数据中，并将得到的统计量结合起来，从而把 SW 检验从单变量推广到多变量数据的正态性检验中。1986 年，Fattorini[25]试图寻找最不具有正态性的投影方向，再对其实施 SW 检验，也称 FA 检验。Zhou 和 Shao[26]在 2014 年将这两种方法相结合，同时考虑每一维数据的正态性以及最不具有正态性的投影方向，并将峰度纳入考核正态性的条件。

卡方拟合优度检验是 Pearson[27]于 1900 年提出，在划分子区间后通过计算子区间内的观测频率和理论频率的差异性来衡量数据是否来自正态分布。该方法可以用于检验多种分布，但是检验结果受区间划分影响，针对此问题，Dai[28]于 2009 年提出改进的卡方拟合拟合优度检验。此外，当进行大样本检验时，第一类错误会显著提升，也就是说即使样本只是略微偏离正态分布也容易拒绝正态原假设[29]。

基于偏度和峰度的正态性检验。Mardia[30]在 1970 年定义了高维数据的偏度和峰度并推导出它们的渐近分布，1974 年又提出了改进的渐近分布形式[31]。1975 年，

Browman 和 Shenton[32]证明偏度和峰度系数的 Johnson S_U和S_B变换近似服从标准正态分布。Mardia 和 Kanazawa[33]在 1983 年给出了三阶矩的渐近展开形式，并对 Mardia[30-31]峰度进行 Wilson-Hilferty 变换并推导出其具有正态近似性。Jarque 和 Bera[34]在 1987 年提出一种偏度和峰度联合检验，后称为 JB 检验，它的渐近分布为卡方分布，Deb 等人[34]提供了部分 JB 统计量的临界值。1984 年，Srivastava[35]为高维数据的偏度和峰度提出了另一种定义，并推导出它们的渐近分布。2008 年，Jurgen 等人[37]验证了在某些高维情况下 MSK 统计量的实用性和良好的检验性能，但也表明在某些情况下可能具有较低的检验性能，并且提出将数据进行独立性变换后再适用偏度和峰度统计量，这也为处理高维数据提供了新思路。2019 年，Enomoto 等人[38]对 Mardia[31]峰度进行归一化变换得到 ZNT 检验统计量，推导出它渐近服从标准正态分布并将其用于高维正态性检验，仿真模拟结果表明样本量较大时该统计量与标准正态分布近似，其检验效力强于峰度检验且第一类错误较小。

基于经验分布函数的正态性检验，理论基础为格里汶科定理，一般来说能够检验多种分布。1928 年，Cramer[39]将经验分布函数和原假设分布函数差的平方的积分作为检验统计量，该方法被称为 Cramer-Von Mises 检验（CVM 检验），Pettitt 和 Stephens[40]提供了该检验的临界值。而 Kolmogorov-Smirnov 检验[41]（KS 检验）将经验分布函数和原假设分布函数之差的绝对值的最大值作为统计量。2007 年，Steinskog 等人[42]研究发现当原假设分布函数中的参数为估计值时，无论样本量多大，KS 统计量一定会存在偏差。Anderson-Darling 检验[43]（AD 检验）仍然是刻画两个分布之间的差异性，但对在分布两端的异常值更加灵敏，不过也有可能因为过于灵敏而导致误判。2019 年，Bayoud[44]基于经验分布和理论分布下的公共区域进行正态拟合优度检验，且中间变量可通过 Beta 分布近似。

基于特征函数的正态性检验。1988 年，Baringhaus 和 Henze[45]提出将经验特征函数与原假设分布特征函数之差的平方的积分值作为检验统计量，Henze 和 Zirkler[46]将此方法延伸并被称为 HZ 检验。Thode[47]在比较多种检验方法后，认为 HZ 检验在高维数据正态性检验中性能最佳。

此外，还有将 Friedman-Rafsky 检验[48]（FR 检验）拓展到高维情况的双样本检验的非参数方法[49]。2019 年，Chen 和 Xia[50]指出并验证许多高维正态检验方法当维数增加时无法控制第一类错误，并提出一种双样本检验的非参数方法，证明了它能保证在高维情况下控制第一类错误且有更好的统计检验性能。

1.3 组织架构

本文的主要研究内容包括：针对多种单变量正态性检验方法，探究它们的设计思路、理论原理和优缺点等，再通过计算机仿真比较不同检验方法的性能，进而获得它们的适用数据类型；整理高维正态性检验中常用的方法，以及探究单变量正态性检验方法是如何发展应用到高维数据中的，对部分直接基于高维数据提出的检验方法归纳它们的设计思路、理论原理和优缺点等，并通过计算机模拟比较这些高维正态性检验方法的性能等；采用多种检验方法对实际数据集进行正态假设检验。

本文结构为：第二章介绍了本文涉及到的理论基础以及其他的准备工作；第三章介绍四大类单变量正态性检验方法，并通过计算机模拟对这些检验方法做数值分析，从多个方面比较它们之间的差异性；第四章详细介绍了多种高维数据正态性检验的思路和方法，并通过计算机模拟对这些检验方法做数值分析，再从检验性能等多方面进行总结；第五章采用多种方法检验两个实际数据集的多元正态性；第六章概括本文结论，并展望未来研究工作。

第四章 高维正态性检验

对高维变量进行正态性检验，可以大致分为四个处理方向。一是绘制统计图；二是做多个单变量检验，包括直接对所有边际分布进行单变量正态性检验、对数据进行独立化变换后再对边际分布进行单变量正态性检验；三是将高维数据进行降维，若降维后变成单变量则使用单变量正态性检验方法，否则再对降维后的高维数据进行正态性检验；四是直接检验高维数据的正态性。

4.1 统计图检验

通过绘制统计图探索正态性的优点是直观，但是当维数增加时，根据 2.2.2 节中性质 3，需要检测任一线性组合的正态性，当维数较高时这是非常困难的，可以对充要条件妥协，仅考虑边缘分布的正态性或考虑降维（1 维）后分布的性质，前者只需参考单变量的正态性统计图检验方法，后者可以基于 2.2.2 节中的平方和距离服从卡方分布的性质，进行卡方点图检验，原理与单变量的 Q-Q 图类似。

接下来将详细介绍卡方点图检验。设p维分布样本$x_1,\cdots x_n$的均值向量为$\bar{x}$，样本协方差为S，并计算每个样本的平方和距离

$$d_i^2 = (x_i - \bar{x})'\mathrm{S}^{-1}(x_i - \bar{x}), i = 1,2,\cdots,n,$$

再将样本按d_i^2排序，按从小到大记为$x_{(1)},\cdots x_{(n)}$。卡方点图检验的纵轴为d_i^2，横轴为$\chi^2_{pl(i)}(p)$, 表示自由度为p的卡方分布在$pl(i)$处的分位数，$pl(i)$表示平方和距离为第i大时的点位置，本文中取$pl(i) = \frac{i-0.5}{n}$。若绘制出的散点图越分布在直线附近则越认为样本来自p维正态分布。

仅通过肉眼判断是否分布在直线附近是非常主观的，可以将单变量 Q-Q 图的量化方法推广到高维上来，包括计算d_i^2与$\chi^2_{pl(i)}(p)$的线性相关系数r_Q、SW 检验等，探索回归方程的三次项系数情况。仅需将单变量方法中的$x_{(i)}$换成$d^2_{(i)}$，$q_{pl(i)}$换成$\chi^2_{pl(i)}(p)$即可。

4.2 多个单变量检验

多个单变量检验指不考虑高维数据联合分布的正态性，而是通过考察边际分布的正态性，以此来推断高维数据是否服从正态分布，单变量检验方法可以参考

第三章内容。当高维数据的每个变量相互独立时，那么“所有的边际分布都服从单变量正态分布”与“总体服从高维正态分布等价”。但当高维数据的变量之间不独立时，“所有的边际分布都服从单变量正态分布”仅是“总体服从高维正态分布等价”的必要不充分条件，因此检验结果会更偏向于接受原假设，即“取伪”概率增加。因此，多个单变量检验可以分为两大类，直接检验边际分布的正态性和将数据变换为独立变量后再检验边际分布的正态性。由 2.2.2 节性质 1 知，当总体服从高维正态分布时，若协方差矩阵为对角阵则必有变量之间相互独立。本节重点介绍后一种，即如何进行数据变换使变量之间相互独立。

设p维分布样本$x_1,\cdots,x_n$服从多元正态分布，均值向量为$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，样本协方差矩阵为$S=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})'$，方差对角阵$V=diag\left(\hat{\sigma}_1^2,\cdots,\hat{\sigma}_p^2\right)$，样本相关性矩阵为$C=V^{-1/2}SV^{-1/2}$，其中$\hat{\sigma}_i^2$表示第$i$维的样本方差。

对矩阵S进行特征分解，Λ_S表示对角元素为特征值的对角矩阵，H_S列向量为对应位置的特征向量，满足$\Lambda_S=H_S'SH_S$且$H_S'H_S=I_p$；同理对矩阵C进行特征分解，存在Λ_C和H_C满足$\Lambda_C=H_C'CH_C$且$H_C'H_C=I_p$。

设原数据为$x_1,\cdots x_n$服从高维正态分布，通过某种变换后转换为$y_1,\cdots,y_n$。设$Y'=(y_1,\cdots,y_n)$，若它的协方差矩阵为对角阵，则称该变换为独立性变换。以下介绍四种独立化变换。

（1）$y_i=S^{-1/2}(x_i-\bar{x}),\ i=1,2,\cdots,n$

（2）$y_i=H_S\Lambda_S^{-1/2}H_S'(x_i-\bar{x}),\ i=1,2,\cdots,n$

（3）$y_i=\Lambda_S^{-1/2}H_S'(x_i-\bar{x}),\ i=1,2,\cdots,n$

（4）$y_i=H_C\Lambda_C^{-1/2}H_C'V^{-1/2}(x_i-\bar{x}),\ i=1,2,\cdots,n$

以上四种变换后的Y都满足均值为零向量，且有$\frac{1}{n}Y'Y=I_p$，则协方差矩阵阵为对角阵。在总体为高维正态分布的假设下，若利用总体数据进行独立性变换得到的Y的变量之间是相互独立的，而使用样本数据变换得到的Y的变量之间则是渐近相互独立的。

4.3 降维

当数据维数过高时，会给正态性检验带来很多困难，而且维数越高则越容易拒绝正态性原假设。解决该问题最直接的方法就是进行降维，但是降维后的数据即使服从正态分布，根据 2.2.2 节中的性质 4，也只是原数据为高维正态分布的

必要不充分条件。但在实际应用中，对降维后的数据进行正态性检验还是很常见的，一般有以下几种降维方法：

（1）每次随机抽取较少的变量进行正态性子检验，重复抽取足够多次，若存在任何一次子检验拒绝正态性假设，则拒绝认为总体服从高维正态分布。

（2）平方和距离服从卡方分布也可以认为是一种降维方法。

（3）使用常用的数据降维的方法。比如主成分分析算法 PCA（Principal Component Analysis），将高维数据映射到低维空间中，同时保证在投影方向上的方差最大，从而尽可能地保留原始数据的信息。

（4）根据 2.2.2 节中的性质 4，高维数据服从高维正态分布的充要条件是是任意线性组合都服从单变量正态分布。如果非正态性最强的几个线性组合都不拒绝正态分布假设，那么不拒绝总体服从高维正态分布的做法是有一定合理性的。FA 检验[25]就是基于此思路提出来的，它结合使用了 SW 检验，具体的统计量为

$$FA = \min_{1\le j\le n} SW\left(z_j' z_1, \cdots, z_j' z_n\right),$$

其中$z_i = S^{-1/2}(x_i - \bar{x})$。

4.4 高维正态性检验

许多高维正态性检验方法是单变量正态性检验的推广，也有部分高维正态性检验是直接基于高维数据提出来的，接下来从四个方向介绍高维正态性检验。

4.4.1 拟合优度检验

拟合优度检验通过比较理论频数和观测频数的差异性判断总体数据是否服从某分布。在进行高维正态性检验时，主要有两个处理思路：

一是直接检验原始数据的正态性，该方法在高维数据上会面临两个问题。一是划分区间时，区间的个数随维数呈幂次递增，从而大幅增加计算量，且区间个数太多对样本量的需求量也会随之增大；二是计算每个区间的理论频率时需要先得到理论概率，这涉及到对分布函数进行多变量积分，计算量较大，若先将数据进行独立化变换则可以将多变量积分问题转换为求单变量积分的乘积问题，而正态分布可以直接通过查表获得积分值，从而大幅降低计算量。

其他需要计算分布函数的检验，比如 KS 检验等，在处理高维数据时，也可以先将数据进行独立化变换，从而简化积分的计算。

二是检验平方和距离是否服从卡方分布，方法可以直接利用单变量的拟合优

度检验或适用于多种分布的检验方法，比如基于经验分布函数的检验方法，包括KS 检验和 AD 检验等。

4.4.2 偏度和峰度检验

将单变量的偏度和峰度检验方法推广到高维主要通过两种方式。

一是直接计算多维偏度和峰度，然后基于它们的统计特性构造渐近分布。Mardia[30-31]通过研究偏度和峰度的均值、方差推导出它们的渐近分布。

设p维分布样本$x_1,\cdots,x_n$服从多元正态分布，均值向量为$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，样本协方差矩阵为$S=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})'$，则样本偏度$\sqrt{b_s}$和样本峰度$b_k$为：

$$\sqrt{b_s}=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[(x_i-\bar{x})'S^{-1}(x_j-\bar{x})\right]^3,$$

$$b_k=\frac{1}{n}\sum_{i=1}^{n}[(x_i-\bar{x})'S^{-1}(x_i-\bar{x})]^2.$$

Mardia[30]于 1970 年推导出$\sqrt{b_s}$和b_k的渐近分布：

$$z_1=\frac{n}{6}\sqrt{b_s}^2\underset{a}{\sim}\chi^2(p(p+1)(p+2)/6)$$

$$z_2=\frac{b_k-\frac{n-1}{n+1}p(p+2)}{\sqrt{\frac{8}{n}p(p+2)}}\underset{a}{\sim}N(0,1),$$

在此基础上构造的偏度峰度联合检验统计量为：

$$M_1=z_1+z_2^2\underset{a}{\sim}\chi^2(p(p+1)(p+2)/6+1)$$

Mardia[31]于 1974 年提出改进的$\sqrt{b_s}$和b_k的渐近分布：

$$z_1^*=\frac{n(p+1)(n+1)(n+3)}{6n[(n+1)(p+1)-6]}\sqrt{b_s}^2\underset{a}{\sim}\chi^2(p(p+1)(p+2)/6),$$

$$z_2^*=\frac{[(n+1)b_k-p(p+2)(n-1)]\sqrt{(n+3)(n+5)}}{\sqrt{8p(p+2)(n-3)(n-p-1)(n-p+1)}}\underset{a}{\sim}N(0,1),$$

在此基础上构造的偏度峰度联合检验统计量为：

$$M_2=z_1^*+(z_2^*)^2\underset{a}{\sim}\chi^2(p(p+1)(p+2)/6+1).$$

二是将数据进行独立化变换后，计算样本偏度和峰度向量，再将单变量检验方法推广到多维。以JB统计量为例：设p维分布样本$x_1,\cdots x_n$通过独立化变换后转

化为变量之间相互独立的样本，记为$y_1,\cdots,y_n$，对每一维求偏度可以得到偏度向量$B_s'=\left(\sqrt{b_{s1}},\cdots,\sqrt{b_{sp}}\right)$，同理可得峰度向量$B_k'=\left(b_{k1},\cdots,b_{kp}\right)$，则对应的$JB$统计量为：

$$JB=n\left(\frac{B_s'B_s}{6}+\frac{(B_k-3l)'(B_k-3l)}{24}\right)\underset{a}{\sim}\chi^2(2p)$$

其中l为元素均是 1 的p维向量。

4.4.3 基于特征函数的检验

基于特征函数的正态性检验方法基本原理是衡量经验特征函数与原假设分布的特征函数的差异性，当差异过大则拒绝原假设。HZ 检验就是一种基于特征函数的正态性检验，它的检验统计量为：

$$HZ_\beta=n\left(4I_E+D_{n,\beta}I_{E^c}\right)$$

其中当E为奇异矩阵时，$I_E=1$，$I_{E^c}=0$；否则$I_E=0$，$I_{E^c}=1$，且

$$\beta=2^{-1/2}\left\{\frac{(2p+1)n}{4}\right\}^{1/(p+4)}$$

$$D_{n,\beta}=n^{-2}\sum_{j=1}^{n}\sum_{k=1}^{n}exp\left(-\frac{\beta^2\parallel Y_j-Y_k\parallel^2}{2}\right)+(1+2\beta^2)^{-\frac{p}{2}}\quad-2(1+\beta^2)^{-p/2}n^{-1}\sum_{j=1}^{n}exp\left[-\frac{\beta^2\parallel Y_j\parallel^2}{2(1+\beta^2)}\right]$$

当HZ_β足够大时，拒绝原假设。但是由于该统计量的直观意义不明显，可解释性不强，当拒绝原假设时，难以获取该数据违反正态性的具体情况。

4.4.4 双样本非参数检验

检验总体的正态性，除了进行单样本检验，还可以尝试双样本检验，第二组样本可以从正态分布随机产生。由于高维正态分布由均值向量μ和协方差矩阵Σ唯一确定，因此当μ和Σ确定后，即可以生成第二组样本，再检验这两组样本是否服从同一分布，从而间接地检验总体的正态性。近年来有许多学者在研究双样本检验，接下来将简要介绍 Chen 和 Xia[50]于 2019 年提出的双样本非参数检验方法。

大多数情况下，μ和Σ是未知的，此时一般用样本$x_1,\cdots,x_n$的均值向量μ_x和样本协方差矩阵Σ_x替代，然后在此基础上生成一组新的样本$y_1,\cdots,y_n\sim N(\mu_x,\Sigma_x)$。值得注意的是检验这两组样本是否服从同一分布并不完全等价于原检验问题。为了

解决该问题，Chen 提出再生成一组独立样本$x_1^*,\cdots,x_n^*\sim N(\mu_x,\Sigma_x)$，并计算它的样本均值向量$\mu_{x^*}$和协方差矩阵$\Sigma_{x^*}$；然后再此基础上生成一组独立样本$y_1^*,\cdots,y_n^*\sim N(\mu_{x^*},\Sigma_{x^*})$。直观来看，若原假设成立，即总体分布服从多元正态分布，那么$\{x_1,\cdots,x_n\}$与$\{y_1,\cdots,y_n\}$之间的关系应该和$\{x_1^*,\cdots,x_n^*\}$与$\{y_1^*,\cdots,y_n^*\}$的关系相似，因此仅需检验这两种关系是否足够相似即可以检验总体分布的正态性。

为了避免 Fridman-Rafsky 双样本检验[48]的缺陷——检验效应（第一类错误）随着维数增加而失去控制（大于显著性水平），Chen[50]在该方法中使用了最近邻信息。具体来说，将$\{x_1,\cdots,x_n\}$与$\{y_1,\cdots,y_n\}$两组样本进行混合，然后针对每个样本寻找它的最近邻点，记YY为$\{y_1,\cdots,y_n\}$中的某一样本的在混合样本下的最近邻点也在落在$\{y_1,\cdots,y_n\}$这一事件，$r(YY)$为该事件发生的概率;相似地可以定义Y^*Y^*和$r(Y^*Y^*)$。最后通过衡量$r(YY)$与$r(Y^*Y^*)$的差异性来检验总体分布的正态性。Chen[50]还证明了YY与Y^*Y^*在一定条件下的渐近等价性以及该检验方法对第一类错误可控。具体的检验步骤为：

Step1：计算样本$x_1,\cdots,x_n$的均值向量μ_x和样本协方差矩阵Σ_x；

Step2：生成样本$y_1,\cdots,y_n\sim N(\mu_x,\Sigma_x)$，并计算$r(YY)$；

Step3：生成样本$x_1^*,\cdots,x_n^*\sim N(\mu_x,\Sigma_x)$，计算均值向量$\mu_{x^*}$和协方差矩阵$\Sigma_{x^*}$，在此基础上生成样本$y_1^*,\cdots,y_n^*\sim N(\mu_{x^*},\Sigma_{x^*})$，并计算$r(Y^*Y^*)$；

Step4：重复 Step3B次，从而获得B个$r(Y^*Y^*)$，表示在原假设下$r(YY)$的经验分布，并计算$r(Y^*Y^*)$的均值$m(r(Y^*Y^*))$；

Step5：计算双边p值，即

$$p=E\left[I\left(\left|r(Y^*Y^*)-m\left(r(Y^*Y^*)\right)\right|\geq\left|r(YY)-m\left(r(Y^*Y^*)\right)\right|\right)\right]$$

其中$I(\cdot)$为示性函数。

Step6：对给定的显著性水平α，若$p\leq\alpha$则拒绝原假设。

应该注意，求最近邻点的时间复杂度是$O(n^2)$，总的时间复杂度为$O(Bn^2)$，当样本量较大时，计算量也较大。

4.5 数值分析

为了探究以上检验方法在不同样本量下，以及对不同的样本分布的检验效应和功效，接下来将进行蒙特卡洛模拟数值分析，选定样本量分别取 50 和 200，数据维数取 2 和 5，显著性水平为 5%。用于检验功效的分布包括：偏态分布$lognormal(0,0.2)^p$、轻尾分布$\beta(2,2)^p$、重尾分布$t(5)^p$、双重分布$exp(1)^5$, 以及高

维混合正态分布$MVNMIX(a,b,c,d)$。形如$(\cdot)^p$表示每一维都是$(\cdot)$分布的p维数据。$MVNMIX(a,b,c,d)$表示占比为a的$N_p(\mathbf{0},\Sigma_1)$与占比为$1-a$$N_p(b\mathbf{1},\Sigma_2)$的混合正态分布，其中$\Sigma_1=(1-c)I_p+c\mathbf{1}\mathbf{1}'$，$\Sigma_2=(1-d)I_p+d\mathbf{1}\mathbf{1}'$，$I_p$表示$p$阶单位阵，$\mathbf{1}$表示元素全为 1 的$p$维向量。部分检验统计量的临界值可以通过查表获得，另一部分检验统计量的临界值通过计算机模拟获得，详细步骤见 2.5 节，取 M=100000，所得临界值的具体数值见表 4-1。

表 4-1 部分检验统计量的临界值
Table 4-1 Critical value of partial test statistics

统计量分位数	$n=50$		$n=200$	
	$p=2$	$p=5$	$p=2$	$p=5$
$q_\alpha(r_Q)$	0.9688	0.9691	0.9826	0.9884
$q_\alpha(FA)$	0.9480	0.9273	0.9848	0.9809
$q_{1-\alpha}(KS)$	0.1355	0.1142	0.0724	0.0632
$q_{1-\alpha}(HZ)$	0.9385	0.9568	0.6391	1.0099

接下来对p维正态分布以及其他 7 种分布进行高维正态性检验，用到的检验统计量为相关系数、FA统计量、χ^2统计量（基于平方和距离）、KS统计量（预先进行独立化变换）、M_1统计量、JB统计量（预先进行独立化变换）和HZ统计量。检验过程为每个分布随机生成n个样本，重复 1000 次并记录这 1000 次中拒绝正态性原假设的次数。当样本的真实分布为正态分布时，拒绝原假设的概率即为该检验方法的效应，当样本的真实分布是其他分布时，拒绝原假设的概率即为该检验方法的功效。效应越低，功效越高意味着检验性能更好。若某个检验的效应远远高于显著性水平 0.05，即第一类错误失去控制，此时即使功效很高也只认为这是一种虚假的拒绝，认为该检验是不可靠的。表 4-2 对应样本量为 200 的 2 维数据的检验结果，表 4-3 对应样本量为 50 的 2 维数据检验结果，表 4-4 对应样本量为 200 的 5 维数据的检验结果，表 4-3 对应样本量为 50 的 5 维数据检验结果，表中平均功效是指某分布基于 7 种非正态分布的拒绝次数的均值，功效排名是指按照平均功效从大到小排序，名次越小意味着拒绝概率越高，也就是功效越高，“\”是指第一类错误失去控制，功效失去可信度，不参与排序。

表 4-2 样本量为 200 下的 2 维正态检验结果

Table 4-2 Two-dimensional normality test results with the sample size of 200

分布名称	r_Q	FA	χ^2	KS	M_1	JB	HZ
$N(0,1)^2$	39	48	74	51	36	36	425
$lognormal(0,0.2)^2$	446	730	96	660	793	955	976
$\beta(2,2)^2$	872	516	635	495	246	950	988
$t(5)^2$	815	811	671	677	917	964	971
$exp(1)^2$	999	1000	1000	1000	1000	1000	1000
$MVNMIX(0.8,5,0,0)$	120	803	486	988	1000	779	1000
$MVNMIX(0.8,0,0.9,-0.1)$	876	610	652	717	961	616	999
$MVNMIX(0.8,5,0.9,-0.1)$	996	1000	1000	1000	1000	1000	1000
平均功效	732	781.4	648.6	791	845.3	894.9	990.6
功效排名	5	4	6	3	2	1	\

观察表 4-2，第一行是指对正态分布样本下进行正态性检验时拒绝的次数，大部分检验的拒绝次数都与 50 接近，恰好对应 5%的显著性水平，然而 HZ 检验的拒绝次数超过了 400 次，也就是说 HZ 检验在样本量为 200 的 2 维数据检验中的第一类错误不受控制，虽然 HZ 检验对来自其他分布的样本拒绝次数很高，但这是一种虚假的拒绝，可以认为此处的 HZ 检验失效了。

功效排名第一的是将样本进行独立化变换后的 JB 检验，它在对偏离正态性不是很严重的混合正态分布的检验中表现相对较差，而另一种偏度峰度联合检验M_1却能有效的拒绝它们，然而它在轻尾分布上表现较差，与单变量情况相一致。而 KS 检验在$lognormal(0,0.2)^2$等多个独立二维变量的样本数据上表现较差，拒绝率普遍不到 70%，除了$exp(1)^2$这种几乎每个检验都能有效拒绝的严重偏离正态性的分布。FA 检验效果与 KS 检验类似。相关系数检验在每个分布上的表现不算特别突出，可能是因为平方和距离损失了原来的样本信息导致，但拒绝次数比较均匀，这意味着该检验对样本的真实分布没有过多要求。拟合优度检验表现最差，除了所有检验都能正确拒绝的两个分布外，拒绝率均不超过 70%，该检验也是基于平方和距离，因此可以进一步认证将高维数据转化为平方和距离来检验正态性会丢失精确度。

表 4-3 样本量为 50 下的 2 维正态检验结果

Table 4-3 Two-dimensional normality test results with the sample size of 50

分布名称	r_Q	FA	χ^2	KS	M_1	JB	HZ
$N(0,1)^2$	57	45	52	46	150	41	30
$lognormal(0,0.2)^2$	170	251	63	193	479	290	194
$\beta(2,2)^2$	245	69	163	100	31	0	95
$t(5)^2$	330	355	167	247	623	527	281
$exp(1)^2$	781	960	569	981	1000	999	999
$MVNMIX(0.8,5,0,0)$	79	634	226	221	887	126	1000
$MVNMIX(0.8,0,0.9,-0.1)$	354	335	121	257	675	195	423
$MVNMIX(0.8,5,0.9,-0.1)$	638	964	616	949	998	992	998
平均功效	371	509.7	275	421.1	670.4	447	570
功效排名	5	2	6	4	\	3	1

表 4-3 相较于表 4-2，样本量从 200 降至 50，平均功效大幅降低的检验有相关系数检验，拟合优度检验、JB 检验，也就是说这些检验对样本量变化比较敏感，若想获得较好的检验性能则需要较大的样本量。从这些检验统计量本身也容易理解，只有较多的样本量才能更加准确的衡量两组数据的线性相关关系；拟合优度检验的原理是比较观测频数和理论频数的差距，当样本量较小时观测频数不稳定性较大，则很难提供符合自身分布特点的观测频数，进而影响检验效果；JB 检验的渐近分布是卡方分布，这也就意味当样本量较大时该性质才能体现出来。此外，HZ 检验在样本量为 50 时的第一类错误控制住 5%以内，且相对其他检验的平均功效最高，而偏度峰度联合统计量M_1的第一类错误失去控制。

表 4-4 样本量为 200 下的 5 维正态检验结果

Table 4-4 Five-dimensional normality test results with the sample size of 200

分布名称	r_Q	FA	χ^2	KS	M_1	JB	HZ
$N(0,1)^5$	50	40	53	41	628	55	49
$lognormal(0,0.2)^5$	449	544	165	458	1000	1000	751
$\beta(2,2)^5$	667	63	718	352	352	1000	742
$t(5)^5$	924	779	920	620	1000	999	906
$exp(1)^5$	1000	999	1000	1000	1000	1000	1000
$MVNMIX(0.8,5,0,0)$	61	582	118	676	1000	126	1000
$MVNMIX(0.8,0,0.9,-0.1)$	1000	990	1000	1000	1000	1000	1000
$MVNMIX(0.8,5,0.9,-0.1)$	1000	1000	1000	1000	1000	1000	998
平均功效	728.7	708.1	703	729.4	907.4	875	913.9
功效排名	4	5	6	3	\	2	1

表 4-4 相较于表 4-2，数据维数从 2 变为 4，除了偏度峰度联合统计量M_1的第一类错误仍然失去控制外，其他检验的第一类错误都更加趋近 50，且平均功效都在一定程度上下降。功效排名第一的是 HZ 检验，除了在$lognormal(0,0.2)^2$等两个分布上功效较低外，在其他分布上全部以超过 90%的概率拒绝原假设。其次是 JB 检验，仅在一个具有偏态性质的混合正态分布上检验功效较低外，在其余分布上的功效都接近 100%。而 KS 等其余检验均在$lognormal(0,0.2)^2$等多个独立二维变量的样本数据上表现较差。值得注意的是 HZ 检验，没有其他检验能以较高的概率拒绝来自$MVNMIX(0.8,5,0,0)$，这表明该分布的非正态性并不强，从而可以知道 HZ 检验是一个相对敏感的检验。

表 4-5 样本量为 50 下的 5 维正态检验结果

Table 4-5 Five-dimensional normality test results with the sample size of 50

分布名称	r_Q	FA	χ^2	KS	M_1	JB	HZ
$N(0,1)^5$	47	59	35	52	990	30	45
$lognormal(0,0.2)^5$	103	131	53	126	999	535	198
$\beta(2,2)^5$	197	46	146	770	995	1000	54
$t(5)^5$	287	299	235	155	1000	790	382
$exp(1)^5$	667	808	735	774	1000	1000	1000
$MVNMIX(0.8,5,0,0)$	49	295	72	913	999	45	992
$MVNMIX(0.8,0,0.9,-0.1)$	862	695	885	998	1000	842	964
$MVNMIX(0.8,5,0.9,-0.1)$	996	940	994	828	1000	997	999
平均功效	451.6	459.1	445.7	652	999	744.1	655.6
功效排名	5	4	6	3	\	1	2

表 4-5 相较于表 4-3，数据维数从 2 变为 4，大部分检验的平均功效都得到了提升。表 4-5 相较于表 4-2，样本量从 200 变为 50，大部分检验的平均功效都相对下降。可见检验的功效会随着样本量和维数的增大而增大，其实无论是增加样本量还是维数都在某种意义上增加了样本数据的信息量，故而提升了正确拒绝非正态分布样本的概率，相应的，错误拒绝正态分布样本的概率也会减少，因此效应也应该随着样本量和维数的增加而被逐渐控制住。表 4-5 中的偏度峰度联合统计量M_1的第一类错误仍然失去控制外，其余检验的第一类错误都得到了较好的控制。纵观表 4-2 至表 4-5，只有在样本量为 200 的 2 维数据的检验中M_1的第一类错误没有失去控制，这种情况下的样本量与维数的比值最大，尝试更改这个比值，并对应地计算M_1检验的第一类错误概率，即效应水平。表 4-6 刻画了M_1的效应水

r_Q的取值范围为[0,1]，r_Q越大表示线性关系越强，因此r_Q可以作为检验统计量，当$r_Q \le (r_Q)_\alpha$时应拒绝正态性原假设。

（2）SW 检验

SW 检验由 Shapiro 和 Wilk[22]在 1965 年提出，也称 W 检验，它的基本思想是对排列后的样本数据和正态分布分位数拟合线性回归方差，在此基础上得到方差的 BLUE 估计（The Best Linear unbiased estimate，最佳线性无偏估计），再计算和样本方差的比值来衡量总体的正态性。该比值称为 SW 统计量，它的取值范围为[0-1]，越接近 1 则正态假设越成立，当比值小于临界值时应该拒绝原假设。

在正态性原假设下，$x_{(i)}$与$q_{pl(i)}$会满足以下线性回归方程：

$$x_{(i)} = \mu + \sigma q_{pl(i)}.$$

记$w_i = E(x_{(i)})$，V为$x_{(1)}, x_{(2)}, \cdots, x_{(n)}$的协方差矩阵。通过广义最小二乘估计，可以得到$\sigma$的 BLUE 估计：

$$\hat{\sigma}_{BLUE} = a'x,$$

其中$a = (w'V^{-1})/(w'V^{-1}V^{-1}w)^{1/2}$，且满足$a'a = 1$。从数学表达式来看，$\hat{\sigma}_{BLUE}$也可以看成排序样本的加权平均和。由于$a_i = -a_{n-i+1}$，因此$\hat{\sigma}_{BLUE}$的另一计算表达式为：

$$\hat{\sigma}_{BLUE} = \sum_{i=1}^{[n/2]} a_{n-i+1}\left(x_{(n-i+1)} - x_{(i)}\right),$$

其中[·]为向下取整符号。将$\hat{\sigma}_{BLUE}$的平方与样本方差相除，得到 SW 统计量：

$$SW = \left(\sum_{i=1}^{[n/2]} a_{n-i+1}\left(x_{(n-i+1)} - x_{(i)}\right)\right)^2 \Big/ \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

（3）多项式回归

基于图 3-2，多种非正态分布的 Q-Q 图均具有 S 曲线型或反 S 曲线型特征，Daniele Coin[20]提出建立多项式回归模型：

$$z_{(i)} = \beta_1 q_{pl(i)} + \beta_3 q_{pl(i)}^3 + \varepsilon_i.$$

若三次项系数β_3显著不等于 0，则意味着标准化 Q-Q 图呈现 S 曲线型或反 S 曲线型。因此可以将$\hat{\beta}_3^2$作为检验统计量，在显著性水平α下，若$\hat{\beta}_3^2 \ge q(\beta_3^2)_{1-\alpha}$，则拒绝原假设，否则没有充分理由拒绝原假设。Daniele Coin[20]研究发现$q(\beta_3^2)_{1-\alpha}$是关于样本量n的函数。

该方法虽然存在一定的合理性，不过由于它是基于 Q-Q 图建立的回归模型，

而 Q-Q 图判断正态性的机理是散点是否分布在直线附近，衡量三次项系数是否为 0 仅仅是对某一类曲线性的量化，显然不如相关系数统计量衡量线性相关程度来的直接。

3.2 拟合优度检验

拟合优度检验的可以分为两大类：吻合度检验和独立性检验。正态性检验属于吻合度检验。首先将连续变量划分成若干个子区间并统计各个子区间的观测频数，同时根据正态分布原假设计算该区间的理论频数，若观察频数与理论频数差异较大，则认为样本数据与正态分布原假设不吻合，从而拒绝原假设。

若总体总共分为k个区间，那么检验统计量为

$$\chi^2 = \sum_{i}^{k} \frac{(f_i - np_i)^2}{np_i} \sim \chi^2(k-1),$$

其中f_i表示区间i的观测频数，p_i表示在原假设分布下区间i的理论概率，np_i则表示区间i的理论频数，该统计量服从自由度为$k-1$的卡方分布。若分布中含有r个待估参数，则需要先估计这些参数，在此基础上计算理论概率的估计值$\hat{p}_i$，最后得到卡方统计量，它服从自由度为$k-r-1$的卡方分布。

$$\chi^2 = \sum_{i}^{k} \frac{(f_i - n\hat{p}_i)^2}{n\hat{p}_i} \sim \chi^2(k-r-1).$$

当原假设为真时，观测频数与理论频数差异应该较小，得到的χ^2也较小；反之，当原假设不成立时，观测频数与理论频数差异应该较大，得到的χ^2也较大。因此，在显著性水平α下，拒绝域为$\chi^2 \geq \chi^2_{1-\alpha}(k-3)$。拟合优度检验的效果会受到区间划分的影响，一般来说划分的区间越多，检验效果则会更好[52]。

3.3 基于偏度和峰度的检验

早在 1930 年就有许多学者[53-55]研究正态分布的矩，当时对矩的研究阶数已经高达七阶，不过后来普遍认为矩在大于四阶后的正态性检验功效反而较低，本节中主要介绍数据标准化后的三阶矩和四阶矩，分别叫偏度和峰度。正态分布的偏度为 0，峰度为 3，当某分布的偏度和峰度分别与这两个值有显著性差异时，则拒绝正态性原假设。基于偏度系数和峰度系数的正态性检验分为两大类，一是偏度或峰度变量单独进行检验，二是偏度峰度联合检验。

3.3.1 偏度系数检验和峰度系数检验

本文 2.4 节中发现部分非正态分布的偏度为 0 或峰度为 3，因此单独使用偏度系数或者峰度系数检验正态性是不一定准确的，除非事先获得总体分布的先验信息：当进行偏度系数检验时，总体应该仅在偏度方向上偏离正态，而在峰度方向上与正态分布相同或相近；当进行峰度系数检验时，总体应该仅在峰度方向上偏离正态，而在偏度方向上与正态分布相同或相近，即满足对称性。

（1）偏度系数检验

偏度系数检验用到的检验统计量为$\sqrt{b_s}=\frac{m_3}{m_2^{3/2}}$，$\sqrt{b_s}>0$表示数据右偏，$\sqrt{b_s}<0$表示数据左偏。在大多数情况下，该检验是双边的，即备择假设为偏度不等于 0；在已知数据的偏斜方向的前提下，该检验是单边的，即备择假设是偏度大于 0 或者偏度小于 0。D'Agostino[56]证明偏度系数的 Johnson S_U变换近似服从标准正态分布，具体的变换过程为：

$$Y=\sqrt{b_s}\left\{\frac{(n+1)(n+3)}{6(n+2)}\right\}^{1/2},$$

$$\beta_2=\frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)},$$

$$W^2=\sqrt{2(\beta_2-1)}-1,$$

$$\delta=\frac{1}{\sqrt{\log W}},$$

$$\alpha=\sqrt{2/(W^2-1)},$$

$$X\left(\sqrt{b_s}\right)=\delta\log\left(Y/\alpha+\sqrt{(Y/\alpha)^2+1}\right),$$

其中要求样本容量$n\geq 8$。则有$X\left(\sqrt{b_s}\right)\underset{a}{\sim}N(0,1)$，其中“$\underset{a}{\sim}$”表示渐近服从某分布。通过 Johnson S_U变换，仅需查询正态分布的分位数表即可得到检验临界值。

（2）峰度系数检验

峰度系数检验用到的检验统计量为$b_k=\frac{m_4}{m_2^2}$，当$\sqrt{b_k}>3$表示数据具有长尾分布，当$b_k<3$表示数据具有短尾分布。在大多数情况下，该检验是双边的，即备择假设为峰度不等于 3；在已知数据分布的长短尾前提下，该检验是单边的即备择假设是峰度大于 3 或者峰度小于 3。Browman 和 Shenton[32]推导出偏度和峰度系数的 Johnson S_U和S_B变换后近似服从标准正态分布，具体变换过程如下：

当$n\geq 25$时，

$$X_S\left(\sqrt{b_s}\right) = \delta_1 \sinh^{-1}\left(\sqrt{b_s}/\lambda_1\right),$$
$$X_S(b_k) = \gamma_2 + \delta_2 \sinh^{-1}\{(b_k - \xi)/\lambda_2\};$$

当$n < 25$时，

$$X_S(b_k) = \gamma_2 + \delta_2 \log\left(\frac{b_2 - \xi}{\xi + \lambda_2 - b_2}\right).$$

式中涉及的参数由$\sqrt{b_s}$和b_k的均值、方差、偏度、峰度确定，但是$X_S(b_k)$的计算比较复杂，因为相关的参数需要进行迭代才能得到。$X_s\left(\sqrt{b_s}\right)$和$X_s(b_k)$都近似服从$N(0,1)$，因此通过 Johnson S_U和S_B变换，仅需查询正态分布的分位数表即可获得检验临界值。

3.3.2 偏度峰度联合检验

单独使用偏度系数或峰度系数检验正态性都需要有先验条件，当数据的偏度、峰度的具体情况难以获取或不满足先验条件时，则应考虑使用偏度峰度联合检验。

Browman 和 Shenton[32]曾考虑以下联合检验变量：

$$\frac{\left(\sqrt{b_s}\right)^2}{\sigma_1^2} + \frac{(b_k - 3)^2}{\sigma_2^2} \underset{a}{\sim} \chi^2(2) \tag{3-1}$$

其中$\sigma_1^2 = 6/n$为$\sqrt{b_1}$的渐近方差，$\sigma_2^2 = 24/n$为b_2的渐近方差。然而该联合变量是难以在现实情况中发挥强大的检验功效的，原因有三：一是只有在样本量很大时才渐近服从卡方分布；二是$\sqrt{b_1}$和b_2并不独立，Jurgen[37]等人通过绘制基于正态分布生成的样本的偏度和峰度散点图，进一步阐释了这两个统计量的相依性；三是b_2趋于正态分布的速度非常慢[32]。Browman 和 Shenton[32]基于$X_s\left(\sqrt{b_s}\right)$和$X_s(b_k)$提出另一种联合检验变量，由于$X_s\left(\sqrt{b_s}\right)$和$X_s(b_k)$并不容易计算，因此该检验统计量的计算量也较大。具体形式为：

$$Y_S^2 = X_S^2\left(\sqrt{b_s}\right) + X_S^2(b_k) \underset{a}{\sim} \chi^2(2).$$

Jarque 和 Bera[34]提出当被检验的真实分布是 Pearson 分布族时，将式（3-1）作为检验变量的检验效果很好，这极大地促进了该联合检验变量的普及，且得名为JB统计量。在正态性假设前提下，JB统计量为 0，当JB统计量大于某临界值时，应该拒绝原假设。JB统计量计算简单，但当样本量较小时，检验效果较差。

$$JB = n\left(\frac{\left(\sqrt{b_s}\right)^2}{6} + \frac{(b_k-3)^2}{24}\right) \underset{a}{\sim} \chi^2(2)$$

3.4 基于经验分布函数的检验

基于经验分布函数的检验是指通过比较经验分布函数和原假设分布函数之间的差距来判别样本是否来自原假设分布。该类方法的理论基础是格里汶科定理，即当样本量很大时，经验分布函数是总体分布函数的良好近似。一般来说此类方法适用于多种分布的检验，比较常用的检验方法包括 CVM 检验[39]、AD 检验[41]、KS 检验[43]等。由于它们的检验原理比较相似，接下来仅详细介绍 KS 检验。

KS 检验由 Kolmogorov[41]在 1933 年提出，它是一种基于经验分布函数的检验方法，使用经验分布函数和原假设分布函数之差的绝对值的最大值来衡量两种分布函数是否近似，若差异小则没有充分的理由拒绝原假设，若差异太大则拒绝原假设。统计量KS的具体形式为：

$$KS = \max_{1\le i\le n}\left\{\left|\Phi(x_i) - \frac{i-1}{n}\right|, \left|\frac{i}{n} - \Phi(x_i)\right|\right\},$$

其中$\Phi(x)$表示标准正态分布函数。在 KS 检验中，为了保证经验分布函数与总体的分布函数相近 ，样本量不能太小。此外，如果每个样本都与正态分布都存在差异，但是最大差异并没有超过临界值，这种情况下检验结果的准确性会受到影响。

3.5 数值分析

为了探究以上检验方法在不同样本量下，以及对不同的样本分布的检验效应和功效，接下来将进行蒙特卡洛模拟数值分析，选定样本量分别取 20、200，用于检验功效的分布详见表 1-1，显著性水平为 5%。部分检验统计量的临界值可以通过查表获得，另一部分检验统计量的临界值通过计算机模拟获得，详细步骤见 2.5 节，取 M=100000，所得临界值的具体数值见表 3-1，“\”表示不需要的分位数。

接下来对正态分布以及其他 25 种分布进行正态性检验，用到的检验统计量为相关系数、SW统计量、三次项系数β_3、χ^2统计量、偏度$\sqrt{b_s}$、偏度b_k、JB统计量和KS统计量。检验过程为每个分布随机生成n个样本，重复 1000 次并记录这 1000 次中拒绝原假设的次数。当样本的真实分布为正态分布时，拒绝原假设的概率即是第一类错误的概率，也为该检验方法的效应；当样本的真实分布是其他分布时，

表 3-1 部分检验统计量的临界值
Table 3-1 Critical value of partial test statistics

统计量分位数	$n=20$			$n=200$		
	$\alpha=0.025$	$\alpha=0.05$	$\alpha=0.975$	$\alpha=0.025$	$\alpha=0.05$	$\alpha=0.975$
$q_\alpha(r_Q)$	\	0.95	\	\	0.99	\
$q_\alpha(\beta_3)$	-0.14	\	0.14	-0.03	\	0.03
$q_\alpha(\sqrt{b_s})$	-0.94	\	0.94	-0.34	\	0.33
$q_\alpha(b_k)$	1.74	\	4.65	2.44	\	3.74

布时，拒绝原假设的概率即为该检验方法的功效。效应越低，功效越高意味着检验性能更好。若某个检验的效应远远高于显著性水平 0.05，即第一类错误失去控制，此时即使功效很高也只认为这是一种虚假的拒绝，进而认为该检验是不可靠的。表 3-2 对应样本量为 200 的检验结果，表 3-3 对应样本量为 20 的检验结果，表中平均功效是指某检验方法基于 25 种非正态分布的拒绝次数的均值，功效排名是指按照平均功效从大到小排序，名次越小意味着拒绝概率越高，也是功效越高，“\”是指第一类错误失去控制，功效失去可信度，不参与排序。

观察表 3-2，第一行是指对正态分布样本下进行正态性检验时拒绝的次数，大部分检验的拒绝次数都与 50 接近，恰好对应 5%的显著性水平。唯一远远超出此标准的是三次项系数β_3，不仅对正态样本拒绝次数过多，也对来自其他分布的样本拒绝次数较少，尤其是对于偏态分布，拒绝次数几乎全部为 0，它在重尾分布以及部分非正态性很强的双峰分布上表现相对较佳。由此可以推出，虽然许多非正态分布的 Q-Q 图呈现 S 型，却并不一定恰好是三次曲线型；此外 Q-Q 分布的检验原理是比较两组数据的线性关系，β_3统计量的原理却是间接地寻找某一类曲线的显著性来证明它的非线性性，虽然在理论上有一定的联系，但仍会损失部分数据信息，从而导致效果不佳。

功效最高的检验是 KS 检验，平均功效超过 83%且检验效应 2%也为最低，可以认为KS 检验是当样本量为200 是正态检验性能最佳的检验方法。即使是$\beta(3,3)$、$SCN(0.8,3)$、$LCN(0.5,1)$、$LCN(0.5,3)$和$LCN(0.5,5)$等其他检验的功效欠佳的分布，观察表 1-1 可以发现它们的偏度和峰度来看都与正态分布较为接近，但 KS 检验仍然展现了良好的检验性能，这说明 KS 检验是对非正态性敏感的。从直观上解释，其他检验统计量大多基于整个样本计算，因此样本的个体性差异容易被忽略，而 KS 检验是找到最远离正态分布的样本计算统计量，因此这个“最不正态”的样本被单独保留，因此对非正态性较为敏感，然而可能因为过于灵敏导致容易拒绝含

表 3-2 样本量为 200 下的正态检验结果
Table 3-2 Normality test results with the sample size of 200

类型	分布名称	r_Q	SW	β_3	χ^2	$\sqrt{b_s}$	b_k	JB	KS
正态分布	$N(0,1)$	56	45	381	61	51	51	46	2
左偏分布	$\beta(2,1)$	1000	1000	0	999	979	602	1000	1000
	$Weibull(10)$	898	914	0	365	935	267	845	1000
右偏分布	$\beta(1,2)$	1000	1000	0	999	976	610	1000	1000
	$lognormal(0,0.2)$	845	844	0	279	899	317	795	1000
	$Weibull(2)$	993	991	42	521	956	183	929	1000
轻尾分布	$\beta(3,3)$	310	448	0	122	4	786	155	1000
	$\beta(2,2)$	861	931	0	245	2	993	625	1000
重尾分布	$SCN(0.8,3)$	362	337	733	224	101	248	264	1000
	$t(5)$	848	791	447	311	478	859	852	1
	$SCN(0.2,7)$	1000	1000	764	1000	738	1000	1000	43
	$t(2)$	1000	1000	681	991	890	1000	1000	28
	$Cauchy(0,1)$	1000	1000	928	1000	987	1000	1000	751
双重分布	$exp(1)$	1000	1000	342	1000	1000	990	1000	1000
	$Gamma(1,1)$	1000	1000	359	1000	1000	986	1000	1000
	$\chi^2(1)$	1000	1000	489	1000	1000	1000	1000	1000
	$lognormal(0,1)$	1000	1000	619	1000	1000	1000	1000	1000
对称双峰分布	$LCN(0.5,1)$	34	38	423	51	39	68	25	991
	$LCN(0.5,3)$	974	985	624	848	5	996	877	1000
	$LCN(0.5,5)$	1000	1000	731	1000	2	1000	1000	1000
左偏双峰分布	$LCN(0.8,1)$	55	62	399	51	69	48	54	1000
	$LCN(0.8,3)$	998	997	562	931	994	65	992	1000
	$LCN(0.8,5)$	1000	1000	677	1000	1000	143	1000	1000
右偏双峰分布	$LCN(0.2,1)$	55	68	398	48	69	46	50	160
	$LCN(0.2,3)$	999	997	530	942	997	46	983	968
	$LCN(0.2,5)$	1000	1000	663	1000	1000	149	1000	988
	平均功效	809.33	816.12	416.4	677.15	644.86	576.17	777.84	837.21
	功效排名	3	2	\	5	6	7	4	1

有异常值的但同时是来源于正态分布的样本。值得注意的是 KS 检验在多个对称重尾分布上检验效果非常差，考虑到重尾分布具有密度曲线陡峭且尾部很长的特征，而此处 KS 检验主要是衡量经验分布函数和正态分布函数之差的绝对值的最大值，结合这些分布的对称性，在对称重尾分布两端会存在较多的差的绝对值并不大的样本点，因而 KS 统计量确实的值也不会很大，因此在使用 KS 检验时应该同时使用另一种对峰度敏感的检验方法。

位列第二的是 SW 检验，除了一些非正态性不强的分布外，SW 检验基本都呈现出较好的检验功效。位列第三的是相关系数，检验性能的表现与 KS 检验较为相似，而比同样是基于 Q-Q 图的三次项系数β_3检验效果好很多，其主要原因是因为相关系数衡量两组数据的线性相关性与 Q-Q 图检验原理一致。

其次是 JB 检验，一种偏度峰度联合检验，在轻尾分布和一些非正态性不强的分布是检验功效不佳，这一性质与JB统计量的定义相对应。而偏度统计量位列第六，对于偏态分布检验效果很好，但对不具备偏态性质的分布则难以发现它的非正态性；与之相对的是峰度统计量，在大部分轻尾和重尾分布上均表现良好，但对于一些峰度接近 3 的双峰分布则效果很差。

拟合优度检验位列第五，它的检验效应相对较高，且对部分分布的检验功效较低，没有特别明显的规律可循。由于拟合优度检验的效果受到区间划分的影响，本文对不同的分布均采取了相同的划分区间方式，效果好坏不一是正常的，可以通过细化划分区间来改善该检验的性能。

表 3-3 与表 3-2 的区别是样本量从 200 降至 20，从平均功效来看每个检验的性能都随着样本量的减少下降，不变的是三次项系数β_3的效应仍远高于显著性水平且在其他分布上的功效较低，而平均功效前三位保持不变。KS 检验仍然是表现最佳的检验方法，不过在右偏双峰分布上的功效下降较多。相关系数检验和 SW 检验在大部分严重偏离正态性的分布上保持了较好的功效水平，但在其他分布上的检验效果也下降较为明显。从总体来看，JB 检验的功效排名下降最多，可以认为它是受样本量减少影响最大的检验方法，由于 JB 统计量的渐近卡方性本来就是基于大样本的，因此，该现象是合理的。

3.6 本章小结

本章主要四个大方向介绍了单变量的正态性检验方法。一是统计图检验，包括用于统计描述的统计图和基于正态性假设下的统计图；二是拟合优度检验；三是基于偏度和峰度的检验，包括单独适用偏度和峰度的检验以及偏度峰度联合检验；四是基于经验分布函数的检验，主要介绍了 KS 检验。

最后通过计算机模拟比较各检验方法的性能并总结出与它们相适应的数据集，得到的结论有检验功效一般会随着样本的增加而提高；检验效果最佳的是 KS 检验、SW 检验和相关系数检验；KS 检验最为敏感，能够检测出非正态性较弱的其他分布、但它在对称重尾分布上检验功效很低，可以考虑结合使用一种峰度敏

表 3-3 样本量为 20 下的正态检验结果
Table 3-3 Normality test results with the sample size of 20

类型	分布名称	r_Q	SW	β_3	χ^2	$\sqrt{b_s}$	b_k	JB	KS
正态分布	$N(0,1)$	59	57	407	48	44	45	29	1
左偏分布	$\beta(2,1)$	208	283	0	172	106	125	35	1000
	$Weibull(10)$	143	151	0	77	173	100	69	1000
右偏分布	$\beta(1,2)$	221	323	0	181	103	113	30	1000
	$lognormal(0,0.2)$	126	140	0	66	149	98	74	1000
	$Weibull(2)$	147	155	87	85	162	79	73	1000
轻尾分布	$\beta(3,3)$	20	38	0	51	13	60	2	1000
	$\beta(2,2)$	37	49	0	58	9	100	1	1000
重尾分布	$SCN(0.8,3)$	104	75	759	69	103	76	61	78
	$t(5)$	194	183	525	85	206	175	177	1
	$SCN(0.2,7)$	869	856	753	693	665	770	750	5
	$t(2)$	584	536	651	334	498	526	486	3
	$Cauchy(0,1)$	874	857	856	785	791	832	820	8
双重分布	$exp(1)$	799	834	383	655	677	361	496	1000
	$Gamma(1,1)$	798	835	401	652	700	372	476	1000
	$\chi^2(1)$	978	983	526	789	897	513	696	1000
	$lognormal(0,1)$	920	929	626	833	868	567	723	1000
对称双峰分布	$LCN(0.5,1)$	40	41	485	59	38	47	16	397
	$LCN(0.5,3)$	59	113	643	104	13	211	2	953
	$LCN(0.5,5)$	647	775	773	546	20	733	7	972
左偏双峰分布	$LCN(0.8,1)$	44	63	429	51	56	47	27	811
	$LCN(0.8,3)$	231	258	598	155	199	88	97	1000
	$LCN(0.8,5)$	841	847	739	601	568	244	297	999
右偏双峰分布	$LCN(0.2,1)$	47	58	461	42	49	52	32	90
	$LCN(0.2,3)$	254	258	600	126	204	85	83	260
	$LCN(0.2,5)$	851	869	715	568	567	207	304	288
	平均功效	401.4	420.4	440.4	313.5	313.4	263.2	233.4	674.6
	功效排名	3	2	\	4	5	6	7	1

感的检验；JB 检验在轻尾分布上检验功效较低，且对样本量敏感，不适用于小样本的检验；单独使用偏度和峰度检验需要先验信息且适用范围狭窄；通过三次曲线量化 Q-Q 曲线不如直接计算线性相关系数的效果好。

表 3-4 汇总了以上理论和数值模拟结果的信息。

表 3-4 单变量正态性检验方法比较

Table 3-4 Comparison of Univariate Normality Test Methods

方法	检验原理	优点	缺点
统计图	统计描述是否与正态分布相符、基于正态性假设的散点图	易于理解，简洁直观	受主观判断影响，茎叶图受区间划分影响
r_Q	排序样本与正态分位数是否线性相关	易于理解，计算简便	小样本时准确率较低，容易受受异常值影响
SW	正态假设下方差的估计值与样本方差之比是否显著小于 1	适用于小样本，准确率高	敏感性不高，难以检测出非正态性不强的分布
β_3	排序样本与正态分位数的三次回归系数是否显著不为 0	易于理解，比较适用于重尾分布	对许多分布检验功效不佳
χ^2	观测频数和理论频数的是否存在较大差距	适用于多种分布	检验结果受区间划分影响
$\sqrt{b_s}$	正态分布的偏度为 0	适用于非对称分布	需要先验信息，适用范围窄
b_k	正态分布的峰度为 3	适用于对称非正态分布	需要先验信息、适用范围窄
JB	结合偏度和峰度	同时考虑了偏度和峰度，适用范围较广	小样本时不准确，轻尾分布检验功效不佳
KS	经验分布函数与原假设分布函数的最大差值	适用于多种分布，敏感性高	样本量不能太小、对异常值敏感、重尾分布检验效果差

第二章 准备工作

2.1 总体与样本

一般情况下，如果总体的容量较大，不便分析其数据特征，我们可以随机抽取一定数量的样本，再通过样本的数据特征来对总体的数据特征进行估计。因此，在本文的正态性拟合检验中，一般都是通过样本呈现出的特征是否与正态分布相符合来判断总体是否满足正态性原假设。样本能够代表总体做统计推断的理论基础是格里汶科定理。

首先介绍经验分布函数。设$X_1, X_2, \cdots, X_n$是来自总体X的一个随机样本，将其观测值$x_1, x_2, \cdots, x_n$按从小到大的顺序排列为$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$，则总体$X$的经验分布函数为

$$F_n(x) = \begin{cases} 0, & x < x_{(1)} \\ \dfrac{k}{n}, & x_{(k)} \le x < x_{(k+1)} \\ 1, & x = x_{(n)} \end{cases}.$$

格里汶科定理：设总体X的真实分布函数为$F(x)$，经验分布函数为$F_n(x)$，则

$$P\left\{\lim_{n\to\infty} \sup_{-\infty<x<+\infty} |F_n(x) - F(x)| = 0\right\} = 1.$$

这表明当n足够大时，$F_n(x)$与$F(x)$之差的绝对值足够小这一事件发生的概率为 1，也就是说此时经验分布函数近似地等于真实分布函数。变量的各种数字特征都会随分布的确定而唯一确定，这也正是能够用样本推断总体分布及其数字特征的理论依据。

实际上，对于$F_n(x)$与$F(x)$之差的绝对值取值的概率有更加精确的概率公式，Dvoretzky–Kiefer–Wolfowitz 不等式[51]：

$$P\left\{\sup_{-\infty<x<+\infty} |F_n(x) - F(x)| > \sqrt{\frac{\log n}{n}}\right\} \le \frac{c}{n^2}) \qquad (1\text{-}1)$$

其中c是与$F(x)$独立的常量。

2.2 正态分布

正态分布，于 1733 年由德国的数学家和天文学家 Abraham de Moivre 在研究

二项分布时提出。1809 年，德国数学家和天文学家 Carl Friedrich Gauss 首次将正态分布应用到天文观测误差的研究中，因此正态分布也叫“高斯分布”。这项研究促使 Gauss 总结出观测误差定律，并提出最小二乘法理论。法国数学家和物理学家 Pierre-Simon Laplace 在 Gauss 的工作基础上推广了中心极限定理，他认为误差可以看成许多小误差的叠加，根据中心极限理论，误差服从高斯分布。正态分布的另一个著名的早期应用是在 1859 年，英国数学家和物理学家 James Clerk Maxwell 提出分子速度分布定律，后来被推广为麦克斯韦-玻尔兹曼分布定律。

2.2.1 单变量正态分布

单变量正态分布的密度函数曲线（以下简称正态曲线）中间高，两边低，左右对称，形似钟型。正态分布可由两个参数唯一确定——均值和方差。均值恰好是正态曲线的对称轴和峰所在的位置，因此均值也称位置参数；方差决定了正态分布的离散程度，方差越大，数据越分散，方差越小，数据越集中，因此方差也称形状参数。

若随机变量X服从正态分布，且它的均值为μ，方差为σ^2，则$X\sim N(\mu,\sigma^2)$，它的密度函数为

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

标准正态分布的均值为 0，方差为 1，记为$N(0,1)$。

偏度（Skewness）是描述分布的偏斜方向和偏斜程度的统计量，可以刻画分布的不对称性。正态分布的偏度为 0，是对称分布。对称分布意味着分布的一半是另一半的镜像，若某分布的一条尾巴比另一条尾巴长，则该分布是偏斜的，它的偏度不为 0，此时称该分布为偏态分布。若偏度小于 0，则称该分布为左偏分布（负偏分布），此时均值位于峰的左侧，分布中小于均值的数据比大于均值的数据少，也就是会存在部分数据非常小，从而密度曲线的左尾比右尾长；若偏度大于 0，则称该分布为右偏分布（正偏分布），此时均值位于峰的右侧，分布中小于均值的数据比大于均值的数据多，也就是会存在部分数据非常大，从而密度曲线的右尾比左尾长。在正态分布中，平均数和中位数的取值相同，而在偏态分布中，一般来说，均值和中位数的取值不同：左偏分布的平均数在中位数的左侧，右偏分布的平均数在中位数的右侧，如图 1-1 所示。

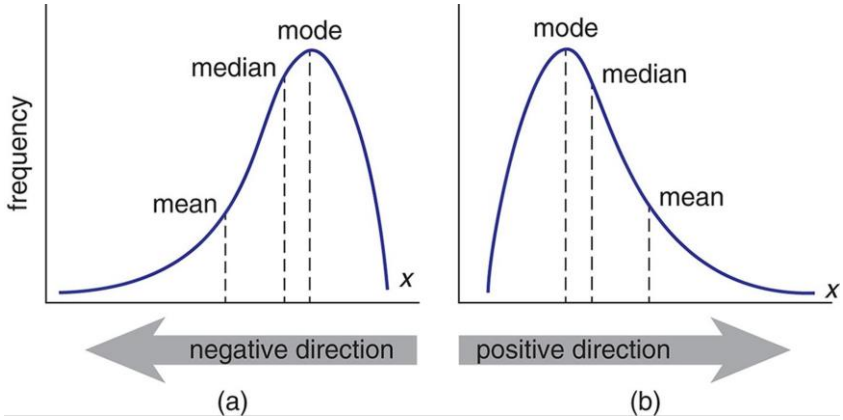


图 1-1 偏度分布的峰、平均数、中位数位置图

Fig.1-1 The position of peak, average, median position for skewness distribution

峰度（Kurtosis）是描述分布形态陡缓程度的统计量，可以直观反映峰的尖度。正态分布的峰度为 3。计算其他分布的峰度其实正是和正态分布相比较，若峰度大于 3，则存在较多的极端值，分布的尾部就越长/越厚，在相同方差下，其余数据则更加集中分布在均值附近，从而峰的形状比较尖，此处命名为“重尾分布”；若峰度小于 3，则存在较少的极端值，分布的尾部就越短/越轻，在相同方差下，集中分布在均值附近的数据则会较少，从而峰的形状比较平，此处命名为“轻尾分布”。

正态分布有四个常用的统计量，它们都与矩相关。均值是一阶原点矩，方差是二阶中心矩，偏度和峰度分别是数据标准化后的三阶原点矩和四阶原点矩，也可以是两个中心矩的比值，偏度等于$\frac{m_3}{m_2^{3/2}}$，峰度等于$\frac{m_4}{m_2^2}$，其中m_i表示i阶中心矩：

$$m_i = E\left[(X - E[X])^i\right].$$

2.2.2 高维正态分布

高维正态分布是单变量正态分布到在高维数据上的推广，它仍由两个参数唯一确定——均值向量和协方差矩阵。标准高维正态分布具有零向量均值和单位阵协方差。设p维随机变量X服从高维正态分布，均值向量为μ，协方差矩阵为Σ，则记为$X \sim N(\mu, \Sigma)$，它的密度函数为：

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right].$$

接下来介绍高维正态分布的常用性质：

性质 1：若$X = (X_1, \cdots, X_p)' \sim N_p(\mu, \Sigma)$，$\Sigma$是对角阵，则$X_1, \cdots, X_p$相互独立。

性质 2：设$X=\left(X_1,\cdots,X_p\right)'\sim N_p(\mu,\Sigma)$，设$A$为$s*p$阶矩阵，$d$为$s$维常数向量，则$AX+d\sim N_s(A\mu+d,A\Sigma A')$。

推论 1：对正态分布X作标准化变换$Z_i=\Sigma^{-1/2}(Z_i-\mu)$后仍服从正态分布，且$Z=\left(Z_1,\cdots,Z_p\right)'\sim N_p\left(0_p,I_p\right)$，其中$0_p$表示$p$维零向量，$I_p$表示$p$阶单位矩阵，则$Z$服从$p$维标准正态分布。由性质 1 知，$Z$的各变量之间相互独立。

推论 2：从p维正态分布中选取任意$k(k\le p)$维变量，则它们的联合分布为k维正态分布，特殊的，当$k=1$时，边际分布服从单变量正态分布。

性质 3：设$X=\left(X_1,\cdots,X_p\right)'\sim N_p(\mu,\Sigma)$，$Z=\left(Z_1,\cdots,Z_p\right)'$且$Z_i=\Sigma^{-1/2}(X_i-\mu)$，则有$d_i^2=Z_i'Z_i=(X_i-\mu)'\Sigma^{-1}(X_i-\mu)\sim\chi_p^2$。

d_i^2是第i个数据点到标准p维正态分布的马氏距离的平方，这里简称为“平方和距离”。若两个数据点的平方和距离相同，则在某种意义上认为它们在该分布中处于相同的位置。

性质 4：设X为p维随机向量，则X服从p元正态分布的充分必要条件是对任意p维实向量a，$\xi=a'X$是单变量正态随机向量。

实际上，证明任意线性组合是一元正态变量是比较困难的。在对多维数据进行正态性检验的时候，若仅尝试证明每一维度或者较少维度上的正态性，会导致拒绝原假设的概率相对较低，但可以减少检验的工作量，在实际应用中是可行的。

2.3 效应和功效

评判一个正态检验方法的性能的好坏可以从两个方面进行——效应和功效。效应是指错误拒绝原假设的概率，即当真实分布是正态分布，而检验的结果为拒绝原假设这一事件的概率，也是第一类错误（拒真）的概率；功效是指正确拒绝原假设的概率，即当真实分布不是正态分布，检验的结果为拒绝原假设的概率，等于 1-第二类错误发生（取伪）的概率。

在进行效应分析时，如果效应小于显著性水平α，也就意味着该检验不容易拒绝原假设，我们称该检验是宽容的；反之如果效应大于显著性水平α，就意味着该检验容易拒绝原假设，那么当真实分布不是正态分布时，正确拒绝原假设的概率一般也会较大，此时即使得到较高的功效也不能认为是可靠的结果，这种现象也称第一类错误不可控。一般认为效应越小且功效越大的检验的性能更优。

2.4 选择用于功效分析的分布

在进行功效分析时，数据来自非正态总体分布，且它的变量维数和样本量是可以调节的。一般来说正态检验方法在不同的维数下功效不同，但对其中的每一维来说，它们的密度曲线可能和正态分布类似，具有中间高两边低的特点，但不再是具有对称性或者更加平缓/陡峭，也可能是密度曲线包括不止一个峰等。具体来说可以分为六大类：左偏分布、右偏分布、轻尾分布、重尾分布、双重分布、双峰分布。其中左偏分布是指具有单峰形态，偏度小于 0，峰度与正态分布的峰度 3 相差不大的分布；右偏分布与左偏分布相对，区别在于它的偏度大于 0；轻尾分布是指具有单峰形态，峰度小于 3，且具有对称性或近似对称性，即偏度约等于 0 的分布；重尾分布与轻尾分布相对，区别在于它的峰度大于 3； 双重分布是指具有单峰形态，偏度和峰度均与正态分布差异较大，具有双重差异性的分布；双峰分布是指具有两个峰的形态的分布。

接下来对以下常见分布进行归类，将它们分到这六大类分布中的一类，并在后续做功效分析时作为该大类分布的代表，用于研究某正态检验方法在该大类分布下的性能情况。本文用到的常见分布包括：指数分布（exp分布）、t分布、$Gamma$分布、β分布、对数正态分布（Log Normal 分布）、Weibull 分布、χ^2分布、$Cauchy$分布、混合正态分布。其中混合正态分布包括两大类：$LCN(p,a)$，表示由占比为$1-p$的标准正态分布$N(0,1)$和占比为p的$N(a,1)$的混合正态分布；$SCN(p,b)$，表示由占比为 $1-p$的标准正态分布$N(0,1)$和占比为p的$N(0,b)$的混合正态分布。

最后挑选出 25 种分布见，详情见表 1-1。偏度和峰度通过计算机模拟计算得到：每次生成 100000 个服从某分布的数据并计算它的偏度和峰度；重复操作 1000 次，计算对应的平均值，并将平均值作为该分布的偏度和峰度的估计，表中保留两位小数。观察发现， $lognormal(0,1)$的峰度估计值超过 80，$Cauchy(0,1)$的峰度估计值超过接近 5000，可以认为它们属于与正态分布差异甚远的极端分布，直观来看，该类分布的检验功效应该会相对更好。此外，多个分布的偏度估计值为 0 或峰度估计值为 3，这也表明在没有任何先验信息的情况，若只用偏度或者峰度来衡量正态性很有可能得到错误的检验结果。

2.5 蒙特卡罗方法计算临界值

蒙特卡罗（Monte Carlo）方法，又称随机抽样或统计试验方法，是基于概率论和统计理论的一种计算方法。该方法通过抓住某事件的数量关系并建立合适的

表 1-1 用于功效分析的分布
Table 1-1 Distributions for power analysis

类型	分布名称	偏度估计	峰度估计
左偏分布	$\beta(2,1)$	-0.57	2.40
	$Weibull(10)$	-0.64	3.57
右偏分布	$\beta(1,2)$	0.57	2.40
	$lognormal(0,0.2)$	0.61	3.68
	$Weibull(2)$	0.63	3.24
轻尾分布	$\beta(3,3)$	0.00	2.33
	$\beta(2,2)$	0.00	2.14
	$SCN(0.8,3)$	0.00	3.55
重尾分布	$t(5)$	0.00	8.67
	$SCN(0.2,7)$	0.00	12.69
	$t(2)$	0.00	953
双重分布	$Cauchy(0,1)$	0.61	4839.60
	$exp(1)$	2.00	8.94
	$Gamma(1,1)$	2.00	8.99
	$\chi^2(1)$	2.82	14.84
	$lognormal(0,1)$	5.86	85.97
对称双峰分布	$LCN(0.5,1)$	0.00	2.91
	$LCN(0.5,3)$	0.00	2.04
	$LCN(0.5,5)$	0.00	1.52
左偏双峰分布	$LCN(0.8,1)$	0.00	3.00
	$LCN(0.8,3)$	-0.07	3.09
	$LCN(0.8,5)$	-1.07	3.17
右偏双份分布	$LCN(0.2,1)$	0.08	2.99
	$LCN(0.2,3)$	0.68	3.08
	$LCN(0.2,5)$	1.07	3.17

概率模型，利用数学方法在计算机上进行模拟抽样，从而获得问题的近似解，一般来说，求解的问题可以为估计某事件发生的概率或者随机变量的期望值等，例如2.4节中对25种分布偏度和峰度的估计就是应用蒙特卡罗方法求期望的一个应用。在正态性检验中，原假设H_0为总体服从正态分布，如果检验方法是通过计算某统计量的取值是否落入拒绝域中来判断是否拒绝原假设，那么在确定显著性水平α的条件下，应该首先确定临界值（critical value）。显著性水平是指当原假设正确而检验结果却是拒绝原假设，即第一类错误发生的概率；临界值是构成拒绝域的重要参数，可以是检验统计量的$1-\alpha$分位数或其他位置的分位数。

当检验统计量是常见分布且它的分位数容易得到时，则可以直接确定临界值；

否则可以通过蒙特卡洛模拟来获得临界值。由 Dvoretzky–Kiefer–Wolfowitz 不等式知，若通过计算机模拟生成样本量为 1000 的模拟样本，根据公式（1-1），则经验分布函数与总体分布函数绝对值差距大于 0.0548 的概率小于10^{-6}数量级；当模拟样本量为 10000 时，经验分布函数与总体分布函数绝对值差距大于 0.02 的概率小于10^{-8}数量级。因此，当设定较大的模拟样本量时，通过蒙特卡洛模拟来获得临界值是可取的。具体步骤为：

Step1：基于原假设分布生成 N 个随机样本；

Step2：计算该组样本对应的检验统计量；

Step3：重复 Step1 和 Step2 共 M 次，则可以获得 M 个检验统计量；

Step4：在给定的显著性水平α下，计算这 M 个统计量的相关分位数，即临界值。

2.6 本章小结

本章介绍了样本与总体的关系，阐述为何能够通过样本检验总体的分布，还介绍了的正态分布的定义与相关性质，接下来阐述了如何衡量检验性能并简要介绍了效应和功效的概念，再接着讨论进行功效分析需要用到哪些分布，最后介绍如何通过蒙特卡罗摸模拟获得检验的临界值。

第三章 单变量正态性检验

统计分析一般分为描述统计和推断统计两大类，对应到正态性检验方法中则是统计图和统计指标。利用统计图进行正态性检验非常直观，但同时有一定的主观性；而通过统计指标检验正态性则需要根据特定的统计性质设计合适的统计量，是客观的。单变量正态性检验发展至今已超过一百年，相关的检验方法非常多，本文将从以下四个大方向各挑选一至两个检验方法详细阐述，包括统计图检验、拟合优度检验、基于偏度和峰度的检验、基于经验分布函数的检验。

3.1 统计图检验

通过绘制统计图来检验正态性是直观的但也是主观的。用于正态性检验的统计图分为两大类，第一类统计图不基于任何假定，主要是对数据进行统计描述，包括直方图、茎叶图和箱线图等，第二类统计图的原理是基于正态性假设的，通过比较样本分布的某些统计量是否与正态分布近似来判断总体是否服从正态分布，例如 PP 图和 QQ 图。

3.1.1 统计描述图

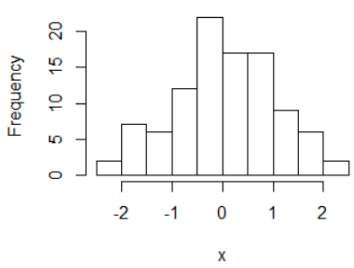
直方图由 Karl Pearson 提出，是一种条形图，横轴一般表示数据类型或数据取值范围，纵轴一般表示频率。对于连续分布，首先应将数据按取值范围分段，再统计每个区间内数据出现的次数即频率，通常要保证每个区间的频率不小于 5。它的检验原理是：当划分的区间长度越短，区间数越多，则条形图的包络线就会更加接近样本数据分布的密度曲线，此时只需将它与正态曲线相比较，就可以直观的呈现样本数据的正态逼近效果。直方图的优点是易于理解，简便直观。缺点是不能量化样本数据服从正态分布的程度，检验结果带有主观性；且区间数是可变的，不同的区间划分会产生不同的包络线从而可能导致不同的检验结果。

茎叶图由英国统计学家 Arthur Bowley 提出，具体画法是将数据按位进行排列比较，茎一般为数据取值较少的位，叶则为数据取值较多的位。利用茎叶图检验正态性的原理与优缺点都与直方图类似，且同时能呈现数据的原始取值信息。

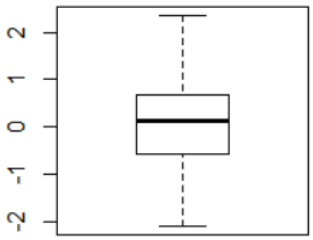
箱线图由美国统计学家 John Tukey 提出，因为形似箱子得名，一般包括五个统计量：最小值，下四分位数、中位数、上四分位数、最大值。若下四分位数和

上四分位数关于中位数对称，则可近似认为样本数据对称，四分位数所处的位置高低则体现了样本数据的分布情况，因此也需要与正态分布的箱线图进行比较才能粗略判断总体是否具有正态性。

图 3-1 绘制了 100 个正态分布样本的直方图、茎叶图和箱线图。可以看到直方图在横轴方向上形似正态曲线，而茎叶图从纵轴方向上形似正态曲线，箱线图基本呈现对称形式。




```
-2 | 100
-1 | 987766
-1 | 3321100
-0 | 987766666655
-0 | 444443322221111100000
 0 | 1111222223333344
 0 | 5556666677777889
 1 | 000111234
 1 | 5556788
 2 | 034
```



（a）直方图　　（b）茎叶图　　（c）箱线图

图 3-1 正态分布样本的直方图、茎叶图和箱线图

Fig.3-1 Histogram, stem-leaf plot and box plot of samples with normal distribution

3.1.2 基于正态性假设的统计图

基于正态性假设的统计图是指刻画样本数据与正态分布的某统计量之间关系的统计图。常用的基于正态性假设的统计图包括 P-P 图和 Q-Q 图。

P-P 图，P 指累积概率，它是以样本数据的累积概率为纵轴，以某特定分布，此处为正态分布的累积概率为横轴绘制散点图。若散点近似地分布在$y = x$对角直线附近，则认为样本数据来自正态分布总体。若散点不满足该条件，但满足其他分布规律，则意味着通过某种数据变换，能够使得变换后的数据的 P-P 图散点近似分布在对角直线附近，也就是说变换后的样本服从正态分布。

Q-Q 图，Q 指分位数，它以从小到大排序的数据为纵轴，以某特定分布对应点位置的分位数为横轴绘制散点图。若该特定分布为正态分布，Q-Q 图也可称为正态概率图。若要判断样本数据是否服从正态分布，仅需观察散点是否近似地分布在某条直线附近，且该直线的截距和斜率恰好表示均值和标准差。同样的，若散点虽然不分布在直线附近但有规律可循，也可以通过某种数据变换方法使得变换后的样本服从正态分布。

点位置的取法在不同文献中存在差异，本文采取的点位置是$pl(i) = \frac{i-0.5}{n}$，对

应的正态分位数记为$q_{pl(i)}$。

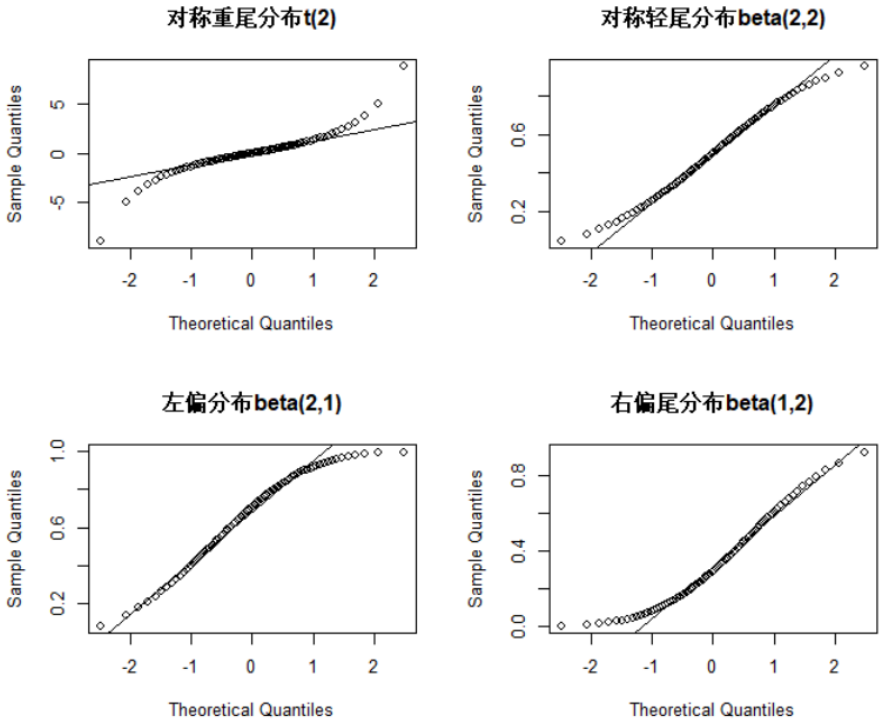


图 3-2 四种分布的 Q-Q 图

Fig.3-2 Q-Q plot of the four distributions

图 3-2 表明，无论是重尾分布、轻尾分布、左偏分布、右偏分布，它们的 Q-Q 图都不分布在某条直线附近而是呈现 S 曲线型或反 S 曲线型，因此可以计算排序样本与正态分布分位数的线性相关系数、S 曲线弯曲程度或基于线性回归方法来量化 Q-Q 图中散点偏离直线的程度，从而以更客观的方法来度量样本是否来自正态总体。

设排序后的样本值为$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$，计算标准化样本值$z_{(i)} = \frac{x_{(i)} - \hat{\mu}}{\hat{\sigma}}$，其中$\hat{\mu}$为样本均值，$\hat{\sigma}$为样本标准差，对应点位置的正态分位数为$q_{pl(i)}$。

（1）相关系数

相关系数可以衡量变量之间的线性相关程度，因此可以计算排序样本和正态分布分位数之间的相关系数来量化 Q-Q 图中散点分布在某条直线附近的程度[19]，计算公式为：

$$r_Q = \frac{\sigma_{xq}}{\sigma_x \sigma_x} = \frac{\sum_{i=1}^{n}(q_{pl(i)} - \bar{q})}{\sqrt{\sum_{i=1}^{n}(x_{(i)} - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(q_{pl(i)} - \bar{q})^2}}.$$

致　谢

本论文是在王成老师的悉心指导下完成的，从确定论文题目到初步设计论文大纲，再到完成本篇论文，非常感谢王老师提供的相关最新文献以及对我的疑惑的解答。一年多前，在多元统计课程上第一次见到王老师，就觉得他学识渊博、看事通透还诙谐幽默。之后很有幸成为了王老师的学生，在此期间王老师给予我们许多学习上和生活上的关怀，还为我们搭起和师兄师姐师弟师妹们们之间的桥梁，帮助我们更好的适应学校生活。此外还要感谢蔡涵玥和李也同学在我读研究生期间的陪伴，在她们的帮助下，我总能更快从逆境走出来。最后，真心的感谢在校期间所有给予我帮助和温暖的老师和同学们。

上海交通大学
学位论文原创性声明

本人郑重声明：所呈交的学位论文《高维数据的正态性假设检验》，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：李扬

日期：2020 年 5 月 19 日

上海交通大学
学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校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，允许论文被查阅和借阅。本人授权上海交通大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密□，在___年解密后适用本授权书。

本学位论文属于

不保密☑。

（请在以上方框内打“√”）

学位论文作者签名：李扬　　指导教师签名：

日期：2020 年 5 月 19 日　　日期：2020 年 5 月 19 日

平随n/p的变化，可以看出效应水平基本随着n/p 的变化而减小，但并非呈现线性关系，若要达到同一效应，越小的p对n/p的要求越低。

表 4-6 M_1检验效应随n/p的变化表

Table 4-6 The effect of M_1 test changes with n/p

n	50	50	200	200	500	400	1000
p	5	2	5	2	5	2	5
n/p	10	25	40	100	100	200	200
效应(%)	99	15	62.8	3.6	15.2	1	1

此外，对于 Chen[50]提出的两样本非参数检验，本文也进行了模拟，尝试多种样本量（50、100、200)和维数（5、10）的 6 种组合后，没有一种情况能得到较高的功效。这可能是因为仅就$x_1,\cdots,x_n$的均值向量μ_x和样本协方差矩阵Σ_x生成一次样本$y_1,\cdots,y_n \sim N(\mu_x,\Sigma_x)$，除非在非常大的样本量下，$r(YY)$的方差很大，是一个非常不稳定的值。因此该问题适合大样本的正态性检验，然而该检验方法的时间复杂度是为$O(Bn^2)$，所以样本量变大也会增加许多计算工作。对于样本量较小的数据，可以通过计算机模拟得到n个$r(YY)$，再计算它的均值，这样可以将为$r(YY)$的方差缩小n倍，从而增加小样本检验的稳定性；或者增加B的次数，因为B个$r(Y^*Y^*)$代表在原假设下$r(YY)$的经验分布，而只有较大的样本量B下得到的经验分布才能更好的接近方差较大的$r(YY)$的真实分布。

4.6 本章小结

本章重点介绍了多种高维正态性检验方法，并总结了四个处理方向。一是绘制统计图，着重介绍了卡方点图检验，它的量化方法与 Q-Q 图类似；二是做多个单变量检验，分为直接对所有边际分布进行单变量正态性检验以及对数据进行独立化变换后再对边际分布进行单变量正态性检验，重点介绍了多种独立化变换方法；三是将高维数据进行降维，比如抽取较少变量进行多次检验、选取非正态性最强的方向进行检验等；四是直接检验高维数据的正态性，包括单变量正态性检验到高维情况的推广，以及其他几种直接针对高维变量提出的检验方法。

单变量正态性检验推广到高维情况的常用手段包括：直接对检验统计量赋予高维情况的定义，推导它的性质或渐近分布从而获得相应的检验统计量；利用高

维正态分布的性质得到相关统计量，并对该统计量进行检验；通过例如独立性变换将高维样本拆解为彼此独立的向量，从而便于使用单变量正态性检验方法。

最后通过计算机模拟进行数值分析，比较多种检验方法的性能和效果。最后得到以下结论：高维正态性检验比较容易出现第一类错误不可控的情况，提高样本量和维数的比值是解决该问题的有效方法之一；HZ 检验和独立化变换后的 JB 检验性能最好；HZ 检验十分敏感，可以检验出非正态性较弱的其他分布，但在某些情况下会发生第一类错误不可控的现象；独立化变换后的 JB 检验对样本量敏感，不适用于样本量较少的情况；M_1检验对轻尾分布的检验效果较差，当样本量和维数比值较小时容易出现第一类错误不可控的情况；基于平方和距离的相关系数检验和拟合优度检验的检验性能不太高，可能是因为平方和距离丢失了部分样本信息，且它们都对样本量敏感，不适用于样本量较少的情况；检验的功效和效应一般会随着样本量和维数的增加就而向好的方向发展；两样本非参数检验适用于大样本情况，如果要将其应用到小样本可以通过计算机模拟出中间变量的均值，从而增加小样本检验的稳定性。

第五章 实际数据集分析

分类是一个重要的统计问题，无论在传统的低维情况下还是近年来发展起来的高维情况下都有广泛的研究。在一定的维数下，当样本量趋于无穷大时，费希尔线性判别分析已显示出良好的性能并具有一定的最优性[57]。即使在高维情况下，样本协方差阵不可逆时，费希尔线性判别分析仍得到了广泛的应用[9, 58-59]。在所有这些研究中，获得线性判别规则并进行随后的误分类率分析，数据的正态性是一个关键假设。本章中将采用七种检验方法对两个常用于分类的实际数据集进行正态假设检验。

5.1 Iris 数据集

Iris 数据集是 Fisher 提供的多变量数据集[60]，包括三类鸢尾花的四个特征变量（皮瓣长度，萼片宽度，花瓣长度和花瓣宽度），其中每类样本数均为 50。该数据集可以直接在 R 软件中获取。Lonney[61]曾检验鸢尾花的四个变量的每一个单变量的正态性，发现除了花瓣宽度以外的变量都不拒绝正态性原假设。现在采用七种检验方法检验该数据集的多元正态性，并计算 p 值，详情见表 5-1。观察发现，除了相关系数检验的 p 值略大为 0.0744 外，其他检验方法的 p 值均小于 0.01，因此拒绝 Iris 数据集的多元正态性原假设，这一结果正与 Lonney[61]的结论相吻合。

表 5-1 Iris 数据集的正态性检验结果
Table 5-1 The normality test results of Iris dataset

检验方法	r_Q	FA	χ^2	KS	M_1	JB	HZ
p-value	0.0744	0	0	0.0084	0	0	0

5.2 Leukemia 数据集

Leukemia 数据集是由 Gloub 提供的白血病数据集[62]，共有 72 个样本，它们均来自急性白血病患者，此处急性白血病包括两类，急性淋巴细胞白血病（ALL）和急性骨髓性白血病（AML），每个样本由 7129 个基因描述。该数据集的下载网址为 http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi。正确区分 ALL 和 AML 对于治疗成功性非常重要。该数据集曾被 Fan 等人[58]用 FAIR 和 NSC 方法以及 Cai

等人[9]用 LPD 规则进行分析，从而用于区分两种白血病。然而，他们在使用他们提出的方法之前均未检验数据的正态性（高维线性判别分析的一个基本假设）。如果数据不具有正态性，那么错分率会受到很大影响且降低了结论的可信度。

该包含 7129 个基因的数据集拒绝正态性原假设的必要条件是任意基因子集的数据不满足正态性。因此，在本文中将随机选取 10 个基因进行七种正态性检验（同 5.1 节中的七种方法），重复 2000 次，若任意一次检验拒绝原假设则拒绝原假设，七种方法的检验结果均为拒绝。重复以上步骤 100 次，七种方法都拒绝了 100 次，因此没有理由认为该数据集服从正态性原假设，从而不能将该数据直接用于高维线性判别分析来区分 ALL 和 AML，例如 Fan 等人[58]和 Cai 等人[9]的研究。

5.3 本章小结

本章检验两个实际数据集是否服从正态原假设。第一个为 Iris 鸢尾花数据集，除了相关系数检验方法的 p 值略大外，基本可以认定拒绝正态原假设，与前人研究结果相吻合；第二个为 Leukemia 白血病数据集，对变量采样降低维数后再进行正态假设检验，重复多次检验结果均为拒绝原假设，并得出该数据集不能直接用于高维线性判别分析的结论。这也进一步说明在实际研究中对正态性假设条件进行验证的必要性以及重要性。

第六章 总结与展望

6.1 主要工作

本文主要研究高维正态性假设检验问题，从单变量的正态性假设检验入手，总结概括了统计图及其量化方法、拟合优度检验、基于偏度和峰度的检验以及基于经验分布函数的检验这四大类检验方法的思路和原理，并通过计算机模拟比较多种检验方法在不同类型样本下的检验性能，从而总结出各检验方法的优缺点以及适用的数据类型。

接下来概括对高维数据进行正态性假设检验的四个可行的处理方向：一是绘制统计图，着重介绍 Q-Q 图及其量化方法在高维情况下的推广；二是多个单变量检验，基于多元正态分布的充要条件引入了多种独立化变换，增强了使用多个单变量检验高维正态性的合理性；三是引入合适的降维方法，介绍了除传统降维方法外的其他可行降维方法，包括引入非正态性最强的投影方向的概念；四是直接检验高维数据的正态性，包括高维变量的偏度和峰度检验、HZ 检验等等。此外，还总结了单变量正态性检验推广到高维情况下的常用手段。然后通过计算机模拟进行数值分析，得到多种高维正态性检验方法的性能。总结出的结论包括：高维正态性检验比较容易出现第一类错误不可控的情况，提高样本量和维数的比值是解决该问题的有效方法之一；HZ 检验和独立化变换后的 JB 检验效果最好；基于平方和距离的相关系数检验和拟合优度检验的性能一般；检验的功效一般会随着样本量和维数的增加而增加，且效应也能够得到控制。

最后对 Iris 鸢尾花数据集和 Leukemia 白血病数据集检验是否服从正态原假设，结果均为拒绝原假设，不能直接使用该数据进行线性判别分析。这也进一步说明在实际研究中若涉及正态性假设条件，则进行正态假设检验是必要的。

6.2 后续研究工作

（1）在本文研究中发现 KS 检验除了对称重尾分布外的检验效果都比较好，除了可以尝试将 KS 检验与其他峰度敏感的检验同时使用，还可以考虑将两者进行有机结合，从而简化检验流程。

（2）由于第一类错误不可控在高维正态性检验中比较常见，除了提高样本量和维

数的比值还可以探索如何从检验本身控制第一类错误。

（3）Chen[50]提出的两样本非参数正态性检验适用于大样本数据，可以尝试通过计算机模拟得到$r(YY)$的估计值，从而减小它的方差，使得小样本检验更加稳定。

参 考 文 献

[1] Whittaker J. Graphical Models in Applied Multivariate Statistics [M]. Chichester: John Wiley and Sons, 1990.

[2] Ma S, Gong Q, Bohnert H J. An Arabidopsis gene network based on the graphical Gaussian model[J]. Genome Research, 2007, 17(11): 1614-1625.

[3] Yuan Ming, Lin Yi. Model selection and estimation in the Gaussian graphical model[J]. Biometrika, 2007(1):1.

[4] Jerome, Friedman, Trevor, et al. Sparse inverse covariance estmation with the graphical lasso.[J]. Biostatistics, 2008.

[5] Jianqing Fan, Yang Feng and Yichao Wu. Network Exploration via the Adaptive LASSO and SCAD Penalties[J]. Annals of Applied Statistics, 3(2):521-541.

[6] Yuan M. High Dimensional Inverse Covariance Matrix Estimation via Linear Programming[J]. Journal of Machine Learning Research, 2010, 11(12):2261-2286.

[7] Yin X, Tianxi C, Tony C T. Testing differential networks with applications to the detection of gene-gene interactions[J]. Biometrika(2):2.

[8] Fan J, Fan Y. High Dimensional Classification Using Features Annealed Independence Rules[J]. Annals of Statistics, 2008, 36(6):2637.

[9] Cai T, Liu W. A Direct Estimation Approach to Sparse Linear Discriminant Analysis[J]. Journal of the American Statistical Association, 2011, 106(496):1566-1577.

[10] Sun, Dennis, L, et al. EXACT POST-SELECTION INFERENCE, WITH APPLICATION TO THE LASSO[J]. Annals of Statistics An Official Journal of the Institute of Mathematical Statistics, 2016.

[11] Taylor J, Tibshirani R. Post-selection inference for-penalized likelihood models[J]. Canadian Journal of Statistics, 2017.

[12] Xie Y, Siegmund D. Sequential multi-sensor change-point detection[J]. Annals of Statistics, 2013, 41(2):670-692.

[13] Liu K, Zhang R, Mei Y. Scalable SUM-Shrinkage Schemes for Distributed Monitoring Large-Scale Data Streams[J]. Statistica Sinica, 2016, 29(1).

[14] D'Agostino R B, Belanger A, D'Agostino R B. A Suggestion for Using Powerful and Informative Tests of Normality[J]. The American Statistician, 1990, 44(4):316-321.

[15] Gnanadesikan R. Methods for Statistical Data Analysis of Multivariate Observations[J]. John Wiley and Sons, London, 1977.

[16]Cox D R, Small N J H. Testing multivariate normality[J]. Biometrika, 1978, 65:263-272.

[17]Cox D R , Wermuth N . Tests of Linearity, Multivariate Normality and the Adequacy of Linear Scores[J]. Journal of the Royal Statistical Society, 1994, 43(2):347-355.

[18]Looney S W . How to Use Tests for Univariate Normality to Assess Multivariate Normality[J]. American Statistician, 1995, 49(1):64-70.

[19]King T, D'Agostino R B, Stephens M A. Goodness-of-Fit Techniques[J]. Technometrics, 1987, 29(4):493.

[20]Coin D. A goodness-of-fit test for normality based on polynomial regression[J]. Computational Statistics & Data Analysis, 2008, 52(4):2185-2198.

[21]Richard J, Dean W. Applied Multivariate Statistical Analysis[M]. USA: Pearson,2007.

[22]Shapiro S S, Wilk M B. An analysis of variance test for normality[J]. Biometrika, 1965, 52(3):591-599.

[23]Shapiro S S, Francia R S. An approximate analysis of variance test for normality. Journal of the American Statistical Association, 1972, 67 (337):215-6.

[24]Royston J. Some Techniques for Assessing Multivariate Normality Based on the Shapiro- Wilk W[J]. Journal of the Royal Statistical Society, 1983.

[25]Fattorini L. Remarks on the use of the Shapiro-Wilk statistic for testing multivariate normality[J]. Stata, 1986, 46(2):209-217.

[26]Zhou M, Shao Y. A powerful test for multivariate normality[J]. Journal of applied stats, 2014, 41(1-2):351-363.

[27]Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonable supposed to have arisen from random sampling. Philosophical Magazine, 1990, 50:157.

[28]Dai J J, Yang A J. A Modified Chi-Squared Goodness-of-Fit Test[J]. Journal of Mathematical Research & Exposition, 2009, 29(1):113-123.

[29]Gerald K．Statistics for Management and Economics[M]．USA：South Western, 2019, 617-621.

[30]Mardia K V. Measures of Multivariate Skewness and Kurtosis with Applications[J]. Biometrika, 1970, 57(3):519-530.

[31]Mardia K V. Applications of Some Measures of Multivariate Skewness and Kurtosis in Testing Normality and Robustness Studies[J]. Sankhyā: The Indian Journal of Statistics, Series B (1960-2002), 1974, 36(2):115-128.

[32]Bowman K O, Shenton L R. Omnibus test contours for departures from normality based on b1 and b2[J]. Biometrika, 1975, 62(2).

[33]Mardia K V, Kanazawa M. The null distribution of multivariate kurtosis[J]. Communications in Statistics - Simulation and Computation, 1983, 12(5):569-76.

[34]Jarque C M, Bera A K. A test for normality of observations and regression residuals[J]. International Statal Review, 1987, 55(2):163-72.

[35]Deb P , Sefton M . The distribution of a Lagrange multiplier test of normality[J]. Economics Letters, 1996, 51(2):123-130.

[36]Srivastava M S. A measure of skewness and kurtosis and a graphical method for assessing multivariate normality[J]. Statistics & Probability Letters, 1984, 2(5):263-7.

[37]Jurgen A, Doornik, Hansen H. An Omnibus Test for Univariate and Multivariate Normality[J]. Oxford Bulletin of Economics & Statistics, 2008, 70:927-939.

[38]Enomoto R, Hanusz Z, Hara A, et al. Multivariate normality test using normalizing transformation for Mardia's multivariate kurtosis[J]. Communication in Statistics Simulation & Computation, 2019:1-15.

[39]Cramer H. On the composition of elementary errors[J]. Skand Aktuarietidskr, 1928, 11:141-80.

[40]Pettitt A N, Stephens M A. Modified Cramer-von Mises Statistics for Censored Data[J]. Biometrika, 1976, 63(2):291-298.

[41]Kolmogorov A. Sulla determinazione empirica di una legge di distribuzione[J]. Giornale dell'Istituto Italiano Attuari, 1933, 4: 83-91.

[42]Steinskog D J, TjØstheim B, KvamstØ G. A Cautionary Note on the Use of the Kolmogorov–Smirnov Test for Normality[J]. Monthly Weather Review, 2007, 135(3):1151-1157.

[43]Anderson T W, Darling D A. Asymptotic Theory of Certain 'Goodness of Fit' Criteria Based on Stochastic Processes[J]. The Annals of Mathematical Statistics, 1952, 23(2):193-212.

[44]Bayoud H A. Tests of normality: new test and comparative study[J]. Communications in Statistics - Simulation and Computation,2019.

[45]Baringhaus L, Henze N. A consistent test for multivariate normality based on the empirical characteristic function[J]. Metrika, 1988, 35(1):339-348.

[46]Henze N, Zirkler B. A class of invariant consistent tests for multivariate normality[J]. Communications in Statistics Theory & Methods, 1990, 19(10):3595-3617.

[47]Thode H C. Testing for Normality[M]. Marcel Dekker, Inc., New York, 2002.

[48]Friedman J H, Rafsky L C. Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-Sample Tests[J]. Annals of Statistics, 1979, 7(4):697-717.

[49]Smith S P, Jain A K. A test to determine the multivariate normality of a data set[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1988, 10(5):0-761.

[50]Chen H, Xia Y. A Nonparametric Normality Test for High-dimensional Data[J]. 2019.

[51]Shao J. Mathematical Statistics, 2nd ed[M]. Springer-Verlag, New York, 2003.

[52]Stephens M A. EDF Statistics for Goodness of Fit and Some Comparisons[J]. Journal of the American Statistical Association, 1974, 69(347):730-737.

[53]Pearson E S. A Further Development of Tests for Normality[J]. Biometrika, 1930, 22(1-2):239-249.

[54]Fisher R A. The Moments of the Distribution for Normal Samples of Measures of Departure from Normality[J]. Proceedings of the Royal Society A Mathematical Physical & Engineering Sciences, 1930, 130(812):16-28.

[55]Lawley C T, Lawley D N. The Derivation of the Fifth and Sixth Moments of the Distribution of b2in Samples from a Normal Population[J]. Biometrika, 1940, 31(3-4):238-248.

[56]D'Agostino R B. Transformation to Normality of the Null Distribution of g1[J]. Biometrika, 1970, 57(3):679-681.

[57]Anderson T W. An introduction to multivariate statistical analysis[M]. Wiley-Intersceince, 3rd ed, New York, 2003.

[58]Fan J, Fan Y. High-dimensional classification using features annealed independence rules[J]. Ann. Statist, 2008.

[59]Zou H. A direct approach to sparse discriminant analysis in ultra-high dimensions[J]. Biometrika, 2012, 99(1):p.29-42.

[60]Fisher R A. The Use of Multiple Measurements in Taxonomic Problems[J]. Annals of Human Genetics, 1936, 7(7):179-188.

[61]Looney S W. How to Use Tests for Univariate Normality to Assess Multivariate Normality[J]. American Statistician, 1995, 49(1):64-70.

[62]Golub T. Molecular classification of cancer : class discovery and class prediction by gene expression monitoring[J]. Science, 1999, 286.