Advances in Applied Mathematics 应用数学进展, 2023, 12(1), 73-80
Published Online January 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.121010

Hans汉斯

基于贝叶斯准则的偏态分布拟合效果分析

徐　鹏

南京邮电大学理学院，江苏　南京

收稿日期：2022年12月15日；录用日期：2023年1月8日；发布日期：2023年1月17日

摘　要

正态分布、*t*分布、双指数分布等概率统计分布已经广泛地应用于社会的各个领域，但是由于实际数据的复杂性，上述分布对数据的拟合并不能很好地表现数据的特征，因此偏态分布逐渐引起了人们的注意。为了进一步拓展偏态分布的应用范围，借助R软件，将偏态分布应用到了台风灾害造成的经济损失数据中，并与常用的分布进行对比分析，利用贝叶斯信息准则评价分布的优劣性。通过研究发现，偏*t*分布对台风灾害造成的经济损失数据表现出了较好的拟合效果，能够更好的体现数据特征，这无论是对台风灾害的研究还是偏*t*分布的研究都具有一定的意义。

关键词

偏*t*分布，偏正态分布，贝叶斯准则，极大似然估计

Analysis of Fitting Effect of Skewed Distribution Based on Bayesian Information Criterion

Peng Xu

College of Science, Nanjing University of Posts and Telecommunications, Nanjing Jiangsu

Received: Dec. 15th, 2022; accepted: Jan. 8th, 2023; published: Jan. 17th, 2023

Abstract

Probability statistical distributions such as normal distribution, *t*-distribution and double exponential distribution have been widely used in various fields of society. However, due to the com-

文章引用: 徐鹏. 基于贝叶斯准则的偏态分布拟合效果分析[J]. 应用数学进展, 2023, 12(1): 73-80.
DOI: 10.12677/aam.2023.121010

plexity of actual data, the data fitting of the above distribution can not well represent the characteristics of the data, so the skewed distribution has gradually attracted people's attention. In order to further expand the application scope of the skewed distribution, R software is used to apply the skewed distribution to the economic loss data caused by typhoon disasters, and the comparison analysis is made with the commonly used distribution, the advantages and disadvantages of the distribution are evaluated by the Bayesian Information Criterion. Through the study, it is found that the skew-*t* distribution has a better fitting effect on the economic loss data caused by typhoon disasters, which can better reflect the characteristics of the data, which has certain significance for both the study of typhoon disasters and the study of skew-*t* distribution.

Keywords

Skew-*t* Distribution, Skew Normal Distribution, Bayes Information Criterion, Maximum Likelihood Estimation

1. 引言

随着计算机技术的发展，相关软件的出现对数据分析提供了更多便捷的方式。R 是一套完整的数据处理、计算和制图软件系统[1]，其可编程的特点以及包含丰富的函数和数据集程序包，更有助于使用者灵活机动的进行数据分析。

偏正态分布和偏 t 分布具有优美的数学形式和易于处理的特点。偏正态分布是正态分布的扩展，偏正态分布可以显示数据的偏态特征[2] [3]，偏 t 分布思想是由 t 分布逐渐发展到广义 t 分布[4]，最后发展到偏 t 分布[5]。近年来，偏态分布的应用越来越广泛，如地震风险建模研究[6]、保险赔偿金评估[7]、鲜烟叶成熟度判定[8]、预测大豆产量[9]等，但是偏态分布在台风灾害损失数据拟合分析方面仍然是空白。

因此，本文借助于 R 编程，分析偏态分布对台风灾害造成的经济损失数据拟合效果，探究相较与常用的分布，如正态分布、威尔分布、伽马分布等，利用偏态分布对台风数据进行拟合时是否更具有优势。这里选择了偏正态分布和偏 t 分布两种比较经典的偏态分布进行研究，采用贝叶斯信息准则(bayes information criterion, BIC)评价拟合效果。首先使用 R 软件产生仿真模拟数据，针对仿真数据，采用极大似然估计生成密度函数曲线对偏态分布的拟合效果进行初步分析，继而利用 BIC 值对拟合效果进行更准确的分析，从而验证整体过程的可行性，在此基础上，探究偏态数据对台风造成的经济损失数据的拟合效果。

2. 仿真数据的拟合效果分析

由于偏正态分布和偏 t 分布是在正态分布和 t 分布基础上拓展而来，因此选用样本量 $N = 200$、均值为 0、方差为 1 的正态分布作为仿真数据，非正态数据选用自由度为 5 的卡方分布作为仿真数据，研究偏正态分布、偏 t 分布对数据的拟合效果，并与正态分布和 t 分布的拟合效果进行对比。

对于仿真生成的数据，其描述性统计表如表 1 所示。根据描述性统计表，利用极大似然估计得到正态分布、偏正态分布、t 分布、偏 t 分布的估计值进而绘制仿真数据真实值以及各分布估计值的密度函数曲线，不同分布极大似然估计值如表 2 所示。

Table 1. Descriptive statistics of sample data generated
表 1. 生成的仿真数据的描述性统计表

	偏正态分布数据	卡方分布数据
平均值	0.118	4.979
标准偏差	1.049	3.398
中位数	0.110	4.317
最小值	−3.364	0.516
最大值	2.971	19.266
偏度	/	1.570
峰度	/	3.000

Table 2. Parameter estimation results of sample data generated under four kinds of distribution
表 2. 生成的仿真数据下四种分布的参数估计结果

分布函数	统计量	正态分布数据估计值	卡方分布数据估计值
正态分布	均值	0.117	4.979
	标准差	1.046	3.390
	偏度	0.000	0.000
偏正态分布	均值	0.118	5.001
	标准差	1.045	3.505
	偏度	0.971	7.465
偏 t 分布	均值	0.120	4.877
	标准差	1.046	3.505
	偏度	0.983	5.697
t 分布	均值	0.121	4.234
	标准差	1.046	4.031
	偏度	0.000	0.000

两种仿真数据真实值以及估计值的密度函数曲线如图 1、图 2 所示。由密度函数曲线可以直观反映出，针对正态仿真数据，四种分布的极大似然估计值密度函数曲线基本相似，针对卡方仿真数据，t 分布的极大似然估计值的密度函数曲线与真实值的曲线差异最大，偏 t 分布的极大似然估计值的密度函数曲线与真实值的曲线最接近。

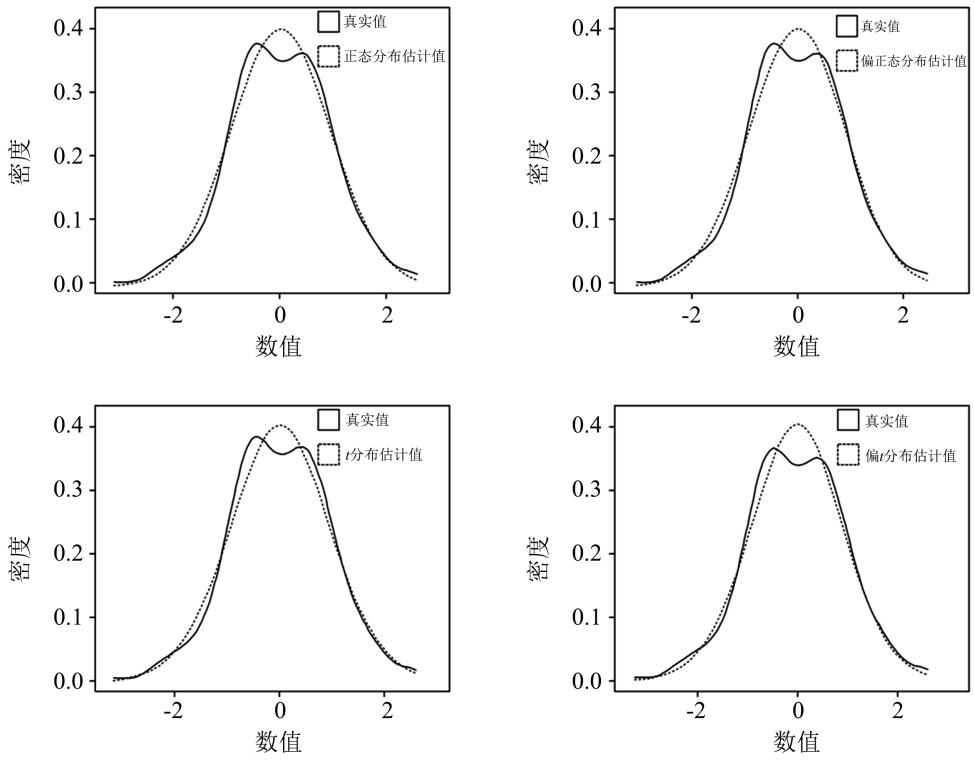

Figure 1. Density function curves of different distributions to normally distributed data
图 1. 不同分布对正态分布数据的密度函数曲线

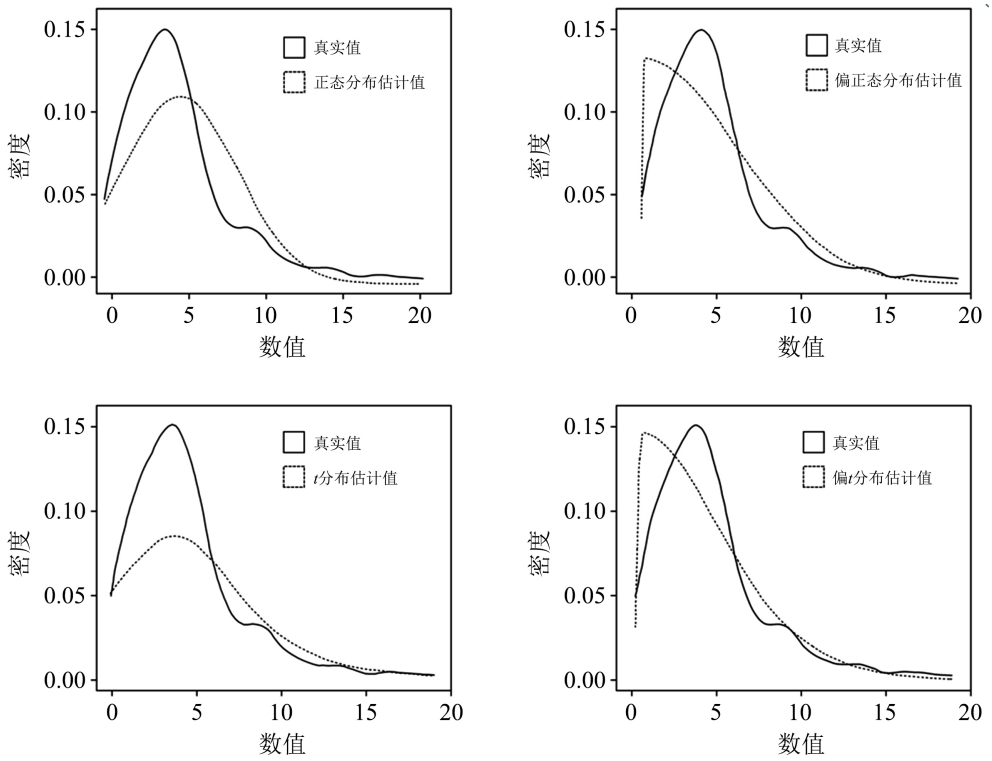

Figure 2. Density function curves of chi-square data with different distributions
图 2. 不同分布对卡方分布数据的密度函数曲线

从上述对密度函数曲线的分析可以发现，对于卡方分布的仿真数据可以看出较明显的区别，但是对于正态分布的仿真数据四种分布估计值的密度函数曲线是相似的，因此采用贝叶斯信息准则来对分布的拟合效果进行更准确的判定就十分必要。表 3 给出了四种分布在仿真数据下的 BIC 值，同时借助常用来判断拟合效果的相关系数 R^2 来进行判定结果的对比，这里 BIC 值越小说明拟合效果越好，R^2 的值越接近 1 说明拟合效果越好。

从上述对密度函数曲线的分析可以发现，对于卡方分布的仿真数据可以看出较明显的区别，但是对于正态分布的仿真数据四种分布估计值的密度函数曲线是相似的，因此采用贝叶斯信息准则来对分布的拟合效果进行更准确的判定就十分必要。表 3 给出了四种分布在仿真数据下的 BIC 值，同时借助常用来判断拟合效果的相关系数 R^2 来进行判定结果的对比，这里 BIC 值越小说明拟合效果越好，R^2 的值越接近 1 说明拟合效果越好。

Table 3. BIC values and correlation coefficients R^2 of different models under sample data generated by simulation
表 3. 生成的仿真数据下不同模型的 BIC 值以及相关系数 R^2 值

分布	正态分布数据		卡方分布数据	
	BIC	相关系数 R^2	BIC	相关系数 R^2
正态分布	**−262.828**	**0.986**	693.766	0.860
偏正态分布	−205.066	0.978	434.094	0.959
偏 t 分布	−219.534	0.983	**301.040**	**0.981**
t 分布	−1.425	0.954	532.885	0.771

根据 BIC 值可以看出，在仿真生成的正态分布数据下，正态分布模型的 BIC 值最小，因此相较于其他分布，正态分布模型的拟合效果最好。在仿真生成的卡方分布数据下，偏 t 分布模型的 BIC 值最小，因此相较于其他分布，偏 t 分布模型的拟合效果最好，这是由偏 t 分布的定义中涵盖了卡方分布导致的。同时，相关系数 R^2 的结果也与 BIC 值的判定结果一致。

通过以上验证，可以确定整体过程是可行的并且可靠的。

3. 台风损失数据的拟合分析

选用我国东南沿海台风多发的广东省从 1983 年到 2019 年 36 年间台风造成的经济损失数据作为研究对象，数据来自于《中国统计年鉴》，具体数据见表 4。

Table 4. Data of economic losses caused by typhoons in Guangdong Province from 1983 to 2019
表 4. 1983~2019 年广东省台风造成经济损失的数据

年份	财产损失/亿元	年份	财产损失/亿元
1983	5.000	2002	8.080
1985	1.000	2003	12.000

Continued

1986	3.200	2004	13.590
1987	4.100	2005	18.890
1988	6.760	2006	19.620
1989	22.600	2007	23.700
1990	2.500	2008	8.480
1991	1.630	2009	15.490
1992	10.000	2010	38.620
1993	3.000	2011	175.70
1994	11.130	2012	32.910
1995	0.710	2013	2.090
1996	0.787	2014	7.529
1997	0.981	2015	12.106
1998	1.390	2016	6.062
1999	4.700	2017	0.600
2000	5.300	2018	8.250
2001	23.600	2019	28.580

根据表 4 数据，表 5 给出了数据的描述性统计表。根据描述性统计表，利用极大似然估计给得到偏正态分布、偏 t 分布的估计值进而绘制仿真数据真实值以及各分布估计值的密度函数曲线，表 6 给出了偏正态分布和偏 t 分布的极大似然估计值，图 3 给出了两种偏态分布和其他常用分布估计值的密度函数曲线和真实值曲线。

Table 5. Descriptive statistics of economic losses caused by typhoon
表 5. 台风造成的经济损失数据的描述性统计表

	台风经济损失/亿元
平均值	17.55
标准偏差	29.10
最小值	0.284
最大值	175.7
偏度	3.941
峰度	18.098

Table 6. Parameter estimation results of skew normal distribution and skew-*t* distribution
表 6. 偏正态分布、偏 *t* 分布的参数估计结果

分布函数	参数	估计值
偏正态分布	均值	−1.1995
	标准差	33.8388
	偏度	25.8880
偏 *t* 分布	均值	18.0052
	标准差	0.9998
	自由度	1.1047
	偏度	1.3566

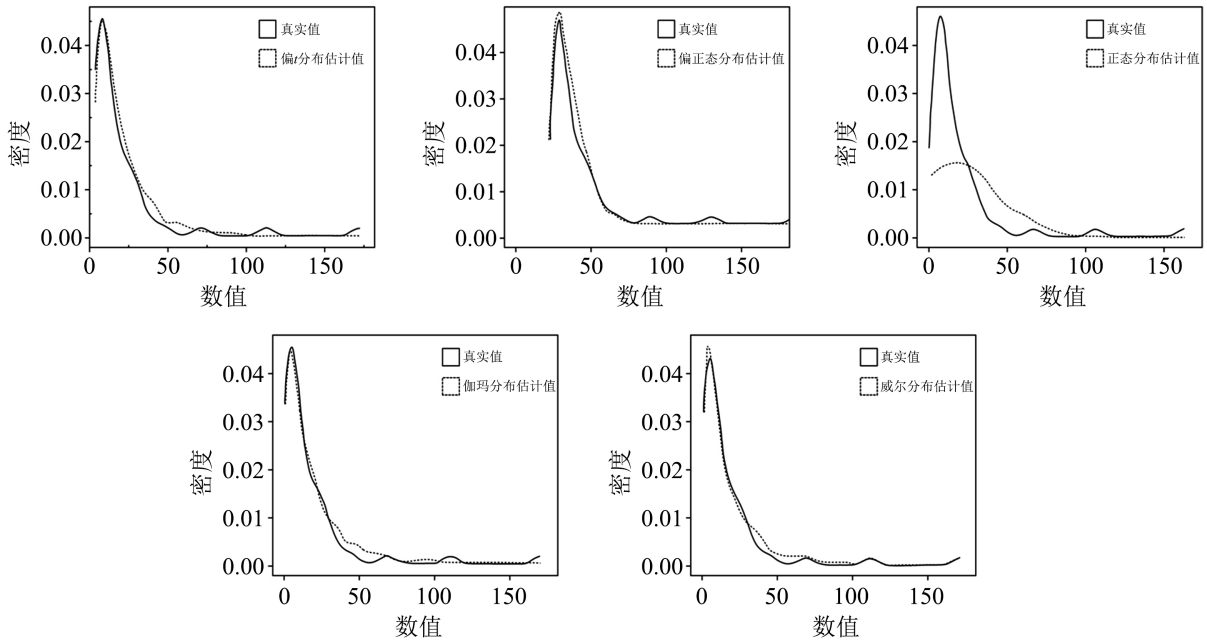

Figure 3. Probability density curves of typhoon loss data with different distributions
图 3. 不同分布对台风损失数据的密度函数曲线

图 3 密度函数曲线直观表现出了五种分布对台风损失数据的拟合效果，正态分布模型拟合效果最差，偏 *t* 分布、偏正态分布、伽玛分布和威尔分布四种分布对于台风损失数据的峰值以及尾部特征各有优劣，因而利用贝叶斯信息准则进一步评价拟合效果。表 7 为不同分布模型对台风损失数据的 BIC 值和相关系数 R^2 值。

从 BIC 值可更加准确判定几种分布模型对台风损失数据的拟合效果。根据表 7 中 BIC 值结果，偏 *t* 分布模型对台风损失数据的拟合更具有优势，相较于除偏 *t* 分布以外 BIC 值最小的威尔分布，偏 t 分布的 BIC 值也具有 22.26%的优势。同时，相关系数 R2 也表现出一致的结果，即偏 *t* 分布模型对台风造成的经济损失数据具有较好的拟合效果。

Table 7. BIC values and R^2 values obtained by fitting the typhoon loss data with five distribution fits
表 7. 五种分布模型对台风损失数据的 BIC 值以及相关系数 R^2 值

分布	BIC 值	相关系数 R^2
偏正态分布	308.0779	0.5034
偏 *t* 分布	**184.8136**	**0.9657**
正态分布	312.6345	0.5621
伽玛分布	383.2735	0.9364
威布尔分布	237.7294	0.9087

4. 总结

本文借助 R 软件中的函数以及数据集程序包，通过编程验证了贝叶斯信息准则在评价分布模型拟合效果的准确性，进而研究了偏态分布在对我国广东省台风造成经济损失数据的拟合效果，并同常用分布模型对比，发现针对我国台风损失数据，偏 *t* 分布具有最佳的拟合效果。这一结论拓展了偏态分布的应用范围，为分析台风灾害造成的经济损失提供了一个新的手段。同时，应该注意到的是，不同的数据所具备的特征不尽相同，因此在选用模型拟合数据时要根据具体情况具体分析。

基金项目

此研究由国家自然科学基金项目(31971029)资助。

参考文献

[1] 吴剑, 钱进. R 软件在工科概率论与数理统计教学中的应用[J]. 考试周刊, 2019(29): 29.

[2] Azzalini, A. (1985) A Class of Distributions That Includes the Normal Ones. *Scandinavian Journal of Statistics*, **12**, 171-178.

[3] Gupta, A.K., Gonzalez-Farias, G. and Dominguez-Molina, J.A. (2004) A Multivariate Skew Normal Distribution. *Journal of Multivariate Analysis*, **89**, 181-190. https://doi.org/10.1016/S0047-259X(03)00131-3

[4] McDonald, J.B. and Newey, W.K. (1988) Skewed Adaptive Estimation of Regression Models via the Generalized *t* Distribution. *Econometric Theory*, **4**, 428-457. https://doi.org/10.1017/S0266466600013384

[5] Theodossiou, P. (1998) Financial Data and the Skewed Generalized *t* Distribution. *Management Science*, **44**, 1650-1661. https://doi.org/10.1287/mnsc.44.12.1650

[6] 郝军章, 翟嘉. 基于有偏分布的我国地震风险测度与保费厘定[J]. 数学的实践与认识, 2020, 50(23): 57-68.

[7] 王明高, 孟生旺. 基于尺度混合偏正态分布的稳健未决赔款准备金评估方法[J]. 数理统计与管理, 2021, 40(4): 634-642.

[8] 沈平, 童德文, 陈郑盟, 等. 基于叶色偏态分布模式的鲜烟叶成熟度判定[J]. 烟草科技, 2021, 54(8): 26-35.

[9] 张佩, 陈郑盟, 马顺登, 尹帝, 江海东. 用冠层叶色偏态分布模式 RGB 模型预测大豆产量[J]. 农业工程学报, 2021, 37(9): 120-126.