Statistics and Application 统计学与应用, 2024, 13(5), 1677-1689
Published Online October 2024 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2024.135166

Hans汉斯

基于尺度混合偏正态分布的考试成绩分析

周晓东[1]，陈沐沂[2]

[1]上海对外经贸大学统计与信息学院，上海
[2]布里斯托大学社会科学与法律学院，英国 布里斯托

收稿日期：2024年9月14日；录用日期：2024年10月6日；发布日期：2024年10月17日

摘　要

本文针对考试成绩分布多峰、有偏的特点，提出采用有限混合－尺度混合偏正态分布进行统计分析。通过模拟和实证分析，对比了多个潜在混合分布，证实所提方法的有效性。文章进一步采用有限混合－尺度混合偏正态误差回归模型对影响考试成绩的因素进行探讨，并与正态误差回归模型进行对比，证实混合偏正态误差回归模型在对考试成绩评价中的优势。

关键词

尺度混合偏正态分布，ECMC算法，偏态回归模型，考试成绩分析

Statistical Analysis for Exam Scores via Scale Mixture of Skew-Normal Distributions

Xiaodong Zhou[1], Shuyi Chen[2]

[1]School of Statistics and Information, Shanghai University of International Business and Economics, Shanghai
[2]Faculty of Social Science and Law, University of Bristol, Bristol, UK

Received: Sep. 14th, 2024; accepted: Oct. 6th, 2024; published: Oct. 17th, 2024

Abstract

This article proposes the use of finite mixture-scale mixture of skew-normal distributions for statistical analysis of exam scores that exhibit multiple peaks and skewness. Through simulation and empirical studies, multiple potential mixture distributions are compared to demonstrate the effectiveness of the proposed method. Furthermore, a linear regression model with a finite mixture-scale mixture of skew-normal error is used to investigate the factors influencing exam scores, and

文章引用: 周晓东, 陈沐沂. 基于尺度混合偏正态分布的考试成绩分析[J]. 统计学与应用, 2024, 13(5): 1677-1689.
DOI: 10.12677/sa.2024.135166

is compared with a normal error regression model, confirming the advantages of the mixture of skew-normal error regression model in evaluating exam performance.

Keywords

Scale Mixture of Skew-Normal Distribution, ECMC Algorithm, Regression Model with Skew Errors, Analysis of Exam Scores

1. 引言

考试成绩作为衡量学生学习成果和筛选人才的重要依据，一直是教育行业研究的主要对象。传统的考试成绩分析大多是基于正态分布进行的。然而，近年来越来越多的教育工作者在对考试数据进行分析时，发现真实考试成绩与基于正态分布的成绩分析存在较大差异。多位学者的研究亦表明学生的考试成绩不服从正态分布，并分析了考试成绩不服从正态分布的原因，且提出了合理的替代分布模型。马成有(2013) [1]和岳武陵(2007) [2]对考试成绩和正态分布之间的关系进行了分析讨论，得出大学生的考试成绩基本不可能服从正态分布的结论。喻晓莉(2006) [3]研究了学生成绩偏离正态分布的原因。尹向飞(2007) [4]、张军舰和马岱君(2021) [5]等人的研究也指出考试成绩的分布使用正态分布进行分析也是不合适的。因此寻找合适的分布模型对考试成绩进行分析是一个值得探讨的问题。

为寻找合适的分布模型对考试成绩进行分析，众多学者进行了有效探索。张国才(2002) [6]研究认为由于考试成绩分布与多个因素有关，如学生群体大小、学生群体和教师的能动作用、学生的素质和基础、成绩评定标准等，因此积极有效的教学可能导致成绩呈现负偏态分布，并对成绩符合负偏态的合理性进行了分析。李翔等(2011) [7]研究认为在当前我国国情下，正态分布不一定能反映教师和学生的教学和学习水平，而峰值靠右的负偏分布可能是合适的分布，同时提出利用三次 Hermite 样条和 B 样条构造考试成绩标准分布函数的方法。考虑到考试成绩存在多峰分布的特点，尹向飞(2007) [4]、张军舰和马岱君(2021) [5]等建议采用混合正态分布对考试成绩进行分析，通过实证分析，发现相比传统的正态分布，混合正态分布对学生成绩的拟合情况更佳，具有应用上的优势。另外，李金屏(2009) [8]等在考虑了学习时间和学习效率的影响之后给出了一个通用的学生考试成绩分布的数学模型。为了进一步分析影响考试成绩的关键因素，彭长生(2010) [9]基于调查数据采用线性回归模型实证发现学习态度、努力程度、班级学风、家庭背景等对考试成绩有影响。除回归模型外，沈家豪等(2022) [10]采用结构方程模型探索医学硕士研究生课程考试成绩的影响因素。喻铁朔等(2020) [11]通过多种回归模型(广义线性模型、深度学习、梯度提升树、支持向量机)结合各可能影响因素对学生考试成绩进行预测。张莉等(2017) [12]利用支持向量机技术对高考成绩进行预测分析。考虑到考试成绩呈现偏态的特点，Canale 等(2016) [13]采用偏 t 分布拟合大一新生统计学考试成绩，讨论了分布参数的无信息分布，采用贝叶斯方法估计模型参数，并将相应方法应用于多门相关成绩的联合建模。

从现有有关考试成绩建模分析的文献可以发现越来越多的学者关注到考试成绩分布有偏特征，开始尝试用有偏分布进行研究分析，但相关工作还不多见。本文结合考试成绩有偏、多峰的特点，提出采用尺度混合偏态分布对考试成绩进行拟合，重点探讨混合偏正态分布的应用。通过模拟和实证验证所提方

法的有效性。进一步，为探讨影响学生考试成绩的因素，提出采用尺度混合偏态误差回归模型对考试成绩进行分析，并与一般回归模型结果进行对比分析，证实所提方法在考试成绩评价中的优势。

2. 有限混合－尺度混合偏正态分布

近年来，有限混合分布被广泛应用于金融、心理、生物医疗等各个领域，相关理论研究成果丰硕[14]。其中，混合正态分布运用较为广泛。但在实际应用中，由于数据分布经常是有偏的，从而正态分布的对称性难以满足，因此混合偏态分布模型常被用来拟合前述的多峰有偏的数据[15] [16]。本文基于 Basso 等(2010) [16]和 Prates 等(2013) [17]的研究，侧重于有限混合尺度混合偏正态分布模型的统计推断和应用。

2.1. 尺度混合偏正态分布

设随机变量 X 具有密度函数：

$$f\left(x;\mu,\sigma^2,\lambda\right)=2/\sigma\,\phi\left(\left(x-\mu\right)/\sigma\right)\Phi\left(\lambda\left(x-\mu\right)/\sigma\right),x\in R, \tag{1}$$

则称随机变量 X 为服从参数 $\boldsymbol{\theta}=\left(\mu,\sigma,\lambda\right)^{\mathrm{T}}$ 的一元偏正态分布，记为 $X\sim SN\left(\mu,\sigma^2,\lambda\right)$，其中，$\mu$ 表示位置参数，σ^2 表示尺度参数，λ 表示偏度参数，ϕ 为标准正态分布的密度函数，Φ 表示标准正态分布的分布函数。当偏度参数 λ 等于 0 时，式(1)所表示的密度函数将退化为正态分布的密度函数，即偏正态分布将退化为正态分布。

设 $Z\sim SN\left(0,\sigma^2,\lambda\right)$。令

$$Y=\mu+U^{-1/2}Z,$$

其中 U 是与 Z 独立且取值为正的随机变量。U 的分布函数和密度函数分别记为 $H\left(u;\nu\right)$ 和 $h\left(u;\nu\right)$，ν 为分布参数。我们称 Y 服从尺度混合偏正态分布，记为 $Y\sim SMSN\left(\mu,\sigma^2,\lambda,\nu\right)$。随机变量 Y 的边缘密度函数为：

$$f\left(y;\mu,\sigma^2,\lambda,\nu\right)=2\int_0^\infty\phi\left(y;\mu,u^{-1}\sigma^2\right)\Phi\left(u^{\frac{1}{2}}\frac{\lambda\left(y-\mu\right)}{\sigma}\right)\mathrm{d}H\left(u;\nu\right).$$

注 1 考虑 U 的不同取值，有：

1. 当 $U=1$ 时，上述尺度混合偏正态分布退化为偏正态分布。

2. 当 U 取不同分布时，尺度混合偏正态分布对应常见的偏 t 分布、偏 Slash 分布、污染偏正态分布等。

若 $U\sim\Gamma\left(\nu/2,\nu/2\right)$，即 U 服从均值为 1 的伽玛分布，则 Y 服从偏 t 分布，记为 $Y\sim ST\left(\mu,\sigma^2,\lambda,\nu\right)$，其密度函数为

$$f\left(y;\mu,\sigma^2,\lambda,\nu\right)=\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu\sigma}}\left(1+\frac{d}{\nu}\right)^{-\frac{\nu+1}{2}}T\left(\sqrt{\frac{\nu+1}{d+\nu}}A;\nu+1\right),y\in R,$$

其中 $d=\left(y-\mu\right)^2/\sigma^2$，$A=\lambda\left(y-\mu\right)/\sigma$，$T\left(\cdot;\nu\right)$ 表示均值为 0，标准差为 1，自由度为 ν 的标准 t 分布函数。当偏度参数 $\lambda=0$ 时，Y 服从 t 分布。

若 $U\sim Beta\left(\nu,1\right)$，其中 $Beta\left(a,b\right)$ 为参数 a,b 的贝塔分布，则 Y 服从偏 Slash 分布，记为 $Y\sim SSL\left(\mu,\sigma^2,\lambda,\nu\right)$，其密度函数为

$$f\left(y;\mu,\sigma^2,\lambda,\nu\right)=2\nu\int_0^1 u^{\nu-1}\phi\left(y;\mu,u^{-1}\sigma^2\right)\Phi\left(u^{1/2}A\right)\mathrm{d}u,\ y\in R.$$

若 U 为离散随机变量，密度函数为 $h\left(u\middle|\mathbf{v}\right)=\nu I\left(u=\gamma\right)+\left(1-\nu\right)I\left(u=1\right)$，$0<\nu<1$，$0<\gamma\le 1$，其中 $\mathbf{v}=\left(\nu,\gamma\right)^{\mathrm{T}}$，$I\left(\cdot\right)$ 为示性函数，则 Y 服从污染偏正态分布，记为 $Y\sim SCN\left(\mu,\sigma^2,\lambda,\nu,\gamma\right)$，其密度函数为

$$f\left(y;\mu,\sigma^2,\lambda,\mathbf{v}\right)=2\left(\nu\phi\left(y;\mu,\gamma^{-1}\sigma^2\right)\Phi\left(\gamma^{1/2}A\right)+\left(1-\nu\right)\phi\left(y;\mu,\sigma^2\right)\Phi\left(A\right)\right).$$

2.2. 有限混合－尺度混合偏正态分布

令 $\boldsymbol{\theta}_j=\left(\mu_j,\sigma_j^2,\lambda_j,\mathbf{v}_j^{\mathrm{T}}\right)^{\mathrm{T}}$，$\boldsymbol{\Theta}=\left(\pi_1,\pi_2,\cdots,\pi_m,\boldsymbol{\theta}_1^{\mathrm{T}},\cdots,\boldsymbol{\theta}_m^{\mathrm{T}}\right)^{\mathrm{T}}$。若随机变量 Y 具有密度函数：

$$f\left(y;\boldsymbol{\Theta}\right)=\sum_{j=1}^{m}\pi_j f\left(y;\boldsymbol{\theta}_j\right),\tag{2}$$

其中 $f\left(y;\boldsymbol{\theta}_j\right)$ 为尺度混合偏正态分布 $SMSN\left(\boldsymbol{\theta}_j\right)$ 的密度函数，$\pi_1,\pi_2,\cdots,\pi_m$ 为各成分分布的混合比例，满足 $0\le\pi_j\le 1$ 且 $\sum_{j=1}^{m}\pi_j=1$，则称 Y 服从由 m 个尺度混合偏正态成分分布组成的有限混合尺度混合偏正态分布，记为 $FMSMSN\left(\boldsymbol{\Theta}\right)$。为计算方便，下文只考虑 $\mathbf{v}_1=\mathbf{v}_2=\cdots=\mathbf{v}_m=\mathbf{v}$ 的情况。

2.3. 参数估计

假设 $y_1,y_2,\cdots,y_n$ 为来自有限混合尺度混合偏正态分布 $FMSMSN\left(\boldsymbol{\Theta}\right)$ 的一组简单随机样本。由 $FMSMSN$ 分布的定义可知，样本 $y_1,y_2,\cdots,y_n$ 应属于 m 个成分分布中的某一类，但无法确切知道属于哪一类，因此引入隐变量 Z_{ij}，其取值为 0 或 1。若样本 y_i 来自第 j 个成分分布则 $Z_{ij}=1$，否则 $Z_{ij}=0$。另外，Z_{ij} 满足以下条件：

$$P\left(Z_{ij}=1\right)=\pi_j,\sum_{j=1}^{m}Z_{ij}=1,\ y_i\middle|Z_{ij}=1\sim SMSN\left(\boldsymbol{\theta}_j\right).$$

令 $\mathbf{Z}_i=\left(Z_{i1},\cdots,Z_{im}\right)^{\mathrm{T}},i=1,\cdots,n$，则 $\mathbf{Z}_i,i=1,\cdots,n$ 相互独立，且服从多项式分布 $M\left(1;\pi_1,\cdots,\pi_m\right)$。为获得模型参数的估计量，Basso 等[16]给出了样本 y_i 分层表示形式：

$$y_i\middle|u_i,t_i,Z_{ij}=1\sim N\left(\mu_j+\Delta_j t_i,u_i^{-1}\Gamma_j\right),$$

$$T_i\middle|u_i,Z_{ij}=1\sim HN\left(0,u_i^{-1}\right),$$

$$U_i\middle|Z_{ij}=1\sim H\left(u_i;\nu\right),$$

$$Z_{ij}\sim M\left(1;\pi_1,\cdots,\pi_m\right),i=1,\cdots,n,\ j=1,\cdots,m,$$

其中 $\Gamma_j=\left(1-\delta_j^2\right)\sigma_j^2$，$\Delta_j=\sigma_j\delta_j$，$\delta_j=\lambda_j\Big/\sqrt{1+\lambda_j^2}$，$HN\left(0,\cdot\right)$ 为区间 $\left(0,\infty\right)$ 上的半正态分布。

基于上述分层表示形式，把潜变量 Z_{ij}，U_i 的值看成是缺失数据，Basso 等[16]给出了求解模型参数 $\boldsymbol{\Theta}$ 极大似然估计的 ECME 算法。Prates 等[17]给出了 ECME 算法实现的 R 包 mixsmsn。对于有限混合偏正态分布以及有限混合偏 t 分布，Fruithwirth-Schnater 和 Pyne [15]给出了获得模型参数 $\boldsymbol{\Theta}$ 的贝叶斯方法。本文我们采用 R 包 mixsmsn 计算模型参数 $\boldsymbol{\Theta}$ 的极大似然估计。

2.4. 有限混合－尺度混合偏正态分布的数字特征

若 $Y\sim FMSMSN\left(\boldsymbol{\Theta}\right)$，密度函数如式(2)所示，则由 Basso 等[16]易得

$$E(Y)=\sum_{j=1}^{m}\pi_j\left(\mu_j+\sqrt{\frac{2}{\pi}}\kappa_1\Delta_j\right),$$

$$D(Y)=\sum_{j=1}^{m}\pi_j\left[\sigma_j^2\left(\kappa_2-\frac{2}{\pi}\kappa_1^2\delta_j^2\right)+\left(\mu_j+\sqrt{\frac{2}{\pi}}\kappa_1\Delta_j\right)^2\right]-\left[E(Y)\right]^2,$$

其中 $\delta_j=\lambda_j\Big/\sqrt{1+\lambda_j^2}$，$\kappa_m=E\left(U^{-\frac{m}{2}}\right)$，$\Delta_j=\sigma_j\delta_j$。令 y_p 为分布 $FMSMSN(\boldsymbol{\Theta})$ 的 p 分位数，则 y_p 满足

$$F(y_p;\boldsymbol{\Theta})=\int_{-\infty}^{y_p}f(y;\boldsymbol{\Theta})\mathrm{d}y=\sum_{j=1}^{m}\pi_j\int_{-\infty}^{y_p}f(y;\boldsymbol{\theta}_j)\mathrm{d}y=p.$$

给定模型参数 $\boldsymbol{\theta}_j$ 以及成分比例 π_j 的值，解上述隐函数方程可得 FMSMSN 分布的 p 分位数 y_p。但上述隐函数方程由于涉及到多个成分分布的分布函数，因此求解需要借助数值解法一种比较容易实现的方法是可以通过模拟的方法获取 FMSMSN 分布的分位数，即利用 R 包 mixsmsn 中的 rmix 函数生成足够多的服从 $FMSMSN(\boldsymbol{\Theta})$ 分布的样本，利用样本的 p 分位数近似 $FMSMSN(\boldsymbol{\Theta})$ 分布的 p 分位数。

2.5. 有限混合－尺度混合偏正态误差回归模型

除对学生成绩本身的拟合外，为了分析可能影响学生成绩的因素，考虑采用有限混合-尺度混合偏正态误差线性回归模型对学生考试成绩及其影响因素进行分析。

有限混合-尺度混合偏正态误差线性回归模型表示为

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon,$$

其中 $x_1,x_2,\cdots,x_p$ 为解释变量，$\beta_0,\beta_1,\cdots,\beta_p$ 为回归系数，误差 $\varepsilon\sim FMSMSN(\boldsymbol{\Theta})$，$\boldsymbol{\Theta}=\left(\pi_1,\pi_2,\cdots,\pi_m,\boldsymbol{\theta}_1^{\mathrm{T}},\cdots,\boldsymbol{\theta}_m^{\mathrm{T}}\right)^{\mathrm{T}}$，$\boldsymbol{\theta}_j=\left(\mu_j,\sigma_j^2,\lambda_j,\boldsymbol{\nu}_j^{\mathrm{T}}\right)^{\mathrm{T}}$。

对于上述模型中回归系数 β_i 以及 $\boldsymbol{\Theta}$ 同样可以采用 ECME 算法获得参数的极大似然估计[18]。具体可以采用 R 包 FMsmsnReg 实现。

3. 数值模拟

尽管上述 FMSMSN 分布涵盖了多个可能的有限混合分布，如有限混合正态分布、有限混合偏正态分布，有限混合偏 t 分布，有限混合偏 Slash 分布等，但考虑到考试成绩有偏、多峰的特点以及模型可接受性等因素，本文重点考察有限混合偏正态分布 $FMSN(\theta)$ 在对考试成绩分析中的应用。首先通过一个数值例子研究在真实数据来自混合偏正态分布时，采用 ECME 算法的估计效果。

采用模型参数估计量的相对绝对偏差(RB)和相对二次均方误差(RMSE)来度量估计效果的有效性，定义如下：

$$\mathrm{RB}\left(\hat{\theta}\right)=\frac{1}{Q}\sum_{i=1}^{Q}\frac{\left|\hat{\theta}_i-\theta_0\right|}{\left|\theta_0\right|},\ \mathrm{RMSE}\left(\hat{\theta}\right)=\frac{1}{Q}\sum_{i=1}^{Q}\frac{\left(\hat{\theta}_i-\theta_0\right)^2}{\theta_0^2},$$

其中 Q 为总的模拟次数，$\hat{\theta}_i$ 为模型参数 θ 在第 i 次模拟的估计值，而 θ_0 为参数 θ 的真实值。考虑只含有两个成分分布的有限混合偏正态分布：

$$FMSN\left(\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{\lambda}\right)=\pi_1SN\left(\mu_1,\sigma_1^2,\lambda_1\right)+\pi_2SN\left(\mu_2,\sigma_2^2,\lambda_2\right),\pi_1+\pi_2=1, \tag{3}$$

其中 $\boldsymbol{\mu}=\left(\mu_1,\mu_2\right)^{\mathrm{T}}$，$\boldsymbol{\sigma}^2=\left(\sigma_1^2,\sigma_2^2\right)^{\mathrm{T}}$，$\boldsymbol{\lambda}=\left(\lambda_1,\lambda_2\ \right)^{\mathrm{T}}$。令式(3)中的参数真值为 $\mu_1^{(0)}=30$，$\mu_2^{(0)}=80$，$\sigma_1^{2(0)}=25$，

$\sigma_2^{2(0)}=100$，$\lambda_1^{(0)}=1$，$\lambda_2^{(0)}=-1$。另外考虑成分分布比例为 $\pi_1^{(0)}=0.1$，$\pi_2^{(0)}=0.9$ 以及 $\pi_1^{(0)}=0.5$，$\pi_2^{(0)}=0.5$ 两种情况。样本量分别取为 200、500、1000，重复模拟 1000 次，具体计算结果如下表 1 和表 2 所示。

Table 1. Parameter estimation based on the mixture of skew-normal distribution ($\pi_1 = 0.1$, $\pi_2 = 0.9$)
表 1. 基于混合偏正态分布模型的参数估计($\pi_1 = 0.1$, $\pi_2 = 0.9$)

参数	n	RB	RMSE	参数	n	RB	MSE
	200	0.7620	0.5281		200	0.3517	0.1208
μ_1	500	0.7530	0.5242	μ_2	500	0.3600	0.1283
	1000	0.7105	0.4637		1000	0.3490	0.1212
	200	0.6243	0.3103		200	0.2627	0.0533
σ_1	500	0.5619	0.2834	σ_2	500	0.2544	0.0594
	1000	0.5198	0.2550		1000	0.2479	0.0595
	200	2.0643	0.3643		200	2.7684	6.5559
λ_1	500	0.9020	0.0186	λ_2	500	1.8685	3.3453
	1000	0.7055	0.0155		1000	1.7597	3.0459
	200	4.0451	15.7200		200	0.4495	0.1941
π_1	500	4.0654	16.1225	π_2	500	0.4517	0.1990
	1000	3.8915	14.8802		1000	0.4324	0.1837

Table 2. Parameter estimation based on the mixture of skew-normal distribution ($\pi_1 = 0.5$, $\pi_2 = 0.5$)
表 2. 基于混合偏模型参正态分布数估计结果($\pi_1 = 0.5$, $\pi_2 = 0.5$)

参数	n	RB	RMSE	参数	n	RB	MSE
	200	0.7600	0.5576		200	0.3391	0.1099
μ_1	500	0.8015	0.63.2	μ_2	500	0.3621	0.1283
	1000	0.7304	0.5263		1000	0.3449	0.1182
	200	0.5507	0.2666		200	0.2819	0.5392
σ_1	500	0.5414	0.2711	σ_2	500	0.2870	0.6973
	1000	0.4909	0.2272		1000	0.2659	0.6608
	200	0.9601	0.4515		200	1.5322	0.1759
λ_1	500	0.7297	0.3813	λ_2	500	1.5311	0.2102
	1000	0.5312	0.2179		1000	1.4942	0.2175
	200	0.0560	0.0000		200	0.0560	0.0000
π_1	500	0.0355	0.0000	π_2	500	0.0355	0.0000
	1000	0.0248	0.0000		1000	0.0248	0.0000

同时为考察在模型为有限混合－尺度混合偏正态分布误差回归模型时 ECMC 算法估计的效果。我们假定多元线性回归模型为

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, i = 1,2,\cdots,n,$$

其中 $\beta_0 = 20$，$\beta_1 = 3$，$\beta_2 = 0.2$，$x_1 \sim B(1,0.5)$，$x_2 \sim N\left(70,10^2\right)$，$\varepsilon \sim FMSN\left(\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{\lambda}\right)$，$\mu_1^{(0)} = -40$，$\mu_2^{(0)} = 4$，$\sigma_1^{2(0)} = 25$，$\sigma_2^{2(0)} = 100$，$\lambda_1^{(0)} = 1$，$\lambda_2^{(0)} = -1$。同样考虑两种情况 $\pi_1^{(0)} = 0.1$，$\pi_2^{(0)} = 0.9$ 以及 $\pi_1^{(0)} = 0.5$，$\pi_2^{(0)} = 0.5$。样本量分别取为 200、500、1000，重复模拟 1000 次，具体计算结果如下表 3 和表 4 所示。

从模拟和计算结果可以得出以下几个结论：

(1) 表 1 和表 2 结果表明混合偏正态分布的总体参数估计具有相合性，即随着样本量的增加，模型参数的估计值趋于真值。并且随着样本量的增加，估计的均方误差(RMSE)也越来越小。相对于参数 μ_j 和 σ_j，参数 λ_j 和 π_j 估计的 RB 和 RMSE 值偏大一些，尤其是当两总体的比例相差很大的时候。

(2) 表 3 和表 4 结果表明总体而言对于有限混合－尺度混合偏正态分布误差回归模型除个别参数外随着样本量的增加，参数估计量的 RB 和 RMSE 值呈现减少趋势。相对于其他参数，参数 λ_j 和 σ_j 估计的 RB 和 RMSE 值偏大一些。

Table 3. Parameter estimation based on the regression model with a mixture of skew-normal error distribution ($\pi_1 = 0.1$, $\pi_2 = 0.9$)
表 3. 基于混合偏正态分布误差回归模型的参数估计($\pi_1 = 0.1$, $\pi_2 = 0.9$)

参数	n	RB	RMSE	参数	n	RB	MSE
	200	0.2717	0.0738		200	0.1739	0.0053
μ_1	500	0.2346	0.0550	μ_2	500	0.1178	0.0057
	1000	0.2213	0.0049		1000	0.0986	0.0070
	200	14.0620	197.74		200	8.8502	78.326
σ_1	500	11.0126	121.28	σ_2	500	8.3177	69.185
	1000	8.8281	77.934		1000	8.2534	68.119
	200	1.4985	0.9390		200	1.7260	2.4884
λ_1	500	1.0663	0.8640	λ_2	500	1.8357	3.2623
	1000	0.8408	0.6630		1000	1.9113	3.6483
	200	0.3009	0.0527		200	0.0334	0.0000
π_1	500	0.1501	0.0095	π_2	500	0.0167	0.0000
	1000	0.0872	0.0029		1000	0.0097	0.0000
	200	0.3268	0.0682		200	0.3289	0.0003
β_0	500	0.2805	0.0707	β_1	500	0.1865	0.0002
	1000	0.2648	0.0684		1000	0.1223	0.0001
	200	0.3353	0.0000				
β_2	500	0.2129	0.0000				
	1000	0.1495	0.0000				

Table 4. Parameter estimation based on the regression model with a mixture of skew-normal error distribution ($\pi_1 = 0.5$, $\pi_2 = 0.5$)
表 4. 基于混合偏正态分布误差回归模型的参数估计($\pi_1 = 0.5$, $\pi_2 = 0.5$)

参数	n	RB	RMSE	参数	n	RB	MSE
	200	1.0323	1.0657		200	4.2473	1.7045
μ_1	500	1.0629	1.1298	μ_2	500	4.4090	2.5259
	1000	1.0597	1.1230		1000	4.4517	2.4460
	200	11.7818	138.81		200	5.0273	25.273
σ_1	500	12.2169	149.25	σ_2	500	4.6847	21.946
	1000	12.5434	157.34		1000	4.8138	23.172
	200	0.6480	0.1189		200	1.8300	3.0715
λ_1	500	0.3505	0.0247	λ_2	500	1.8646	3.3966
	1000	0.2155	0.0023		1000	1.9213	3.6163
	200	0.0572	0.0003		200	0.0572	0.0003
π_1	500	0.0375	0.0001	π_2	500	0.0375	0.0001
	1000	0.0263	0.0000		1000	0.0263	0.0000
	200	0.9729	0.9466		200	0.2141	0.0000
β_0	500	0.9720	0.9448	β_1	500	0.1399	0.0000
	1000	0.9705	0.9419		1000	0.0962	0.0000
	200	0.3454	0.0000				
β_2	500	0.2625	0.0000				
	1000	0.2009	0.0000				

4. 实证分析

4.1. 基于混合偏正态分布的考试成绩分析

运用某高校 599 名大学生《统计学》考试成绩进行实证分析。表 5 为考试成绩的描述性统计结果。图 1 为考试成绩的 QQ 图。

Table 5. Descriptive statistics of exam scores
表 5. 考试成绩的描述性统计

均值	标准差	中位数	最小值	最大值	偏度系数	峰度系数
72.03	16.97	76	8	97	−1.73	2.86

对数据进行正态性检验，考虑 Anderson-Darling、Cramer-vonMises、Lilliefors、pearson 卡方以及 Shapiro-Francia 正态性检验方法，结果均显示 p 值小于 0.005，拒绝原假设，表明所使用的成绩数据是非正态的。

由表 5、图 1 以及正态性检验的结果，可知学生考试成绩的分布具有明显的偏斜，不符合正态分布。以下我们采用有限混合－尺度混合偏正态分布(FMSN)进行拟合。为确定最优成分数，表 6 计算了不同信息准则下不同成分组数对应模型的信息准则值。

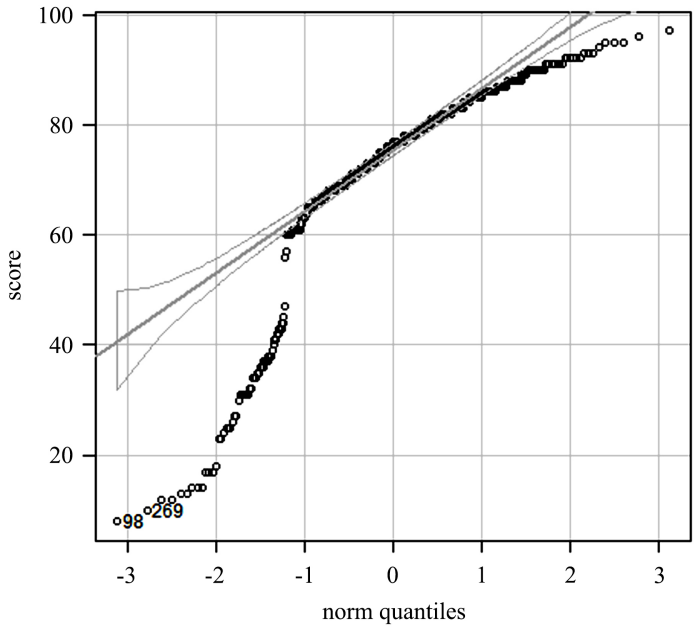

Figure 1. QQ plot for exam scores
图 1. 考试成绩 QQ 图

Table 6. Values of different information criteria for cases with different number of components
表 6. 不同信息准则不同成分组数下信息准则值

信息准则	$m = 1$	$m = 2$	$m = 3$	$m = 4$
AIC	4524.877	4369.032	**4362.306**	4362.372
BIC	4537.856	**4399.315**	4409.894	4427.264
EDC	4533.063	**4388.132**	4392.321	4403.301
ICL	4537.856	**4399.892**	4560.78	4573.571

由表 6，结合图 1，我们选用成分数为 2 的混合偏正态模型(4)，并使用 ECMC 算法求解模型参数的极大似然估计，结果如表 7 所示。

$$Y \sim \pi_1 SN\left(\mu_1, \sigma_1^2, \lambda_1\right) + \pi_2 SN\left(\mu_2, \sigma_2^2, \lambda_2\right) \tag{4}$$

Table 7. Parameter estimation for the mixture of skew-normal distribution
表 7. 混合偏正态分布参数估计结果

$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$
0.1104	0.8896	43.3539	82.8538	269.3054	100.0410	−4.9657	−0.9954

估计的密度曲线如图 2 所示。作为对比，图 2 还给出了核密度估计曲线。结果表明构成混合偏正态的两个偏态分布均有偏，其中第一个成分的偏度($\hat{\lambda}_1 = -4.9657$)要小于第二成分的偏度($\hat{\lambda}_2 = -0.9954$)。两成分的混合偏正态分布密度与核密度估计结果比较接近。表 8 给出了基于估计分布的均值，以及 20%，40%，60%，80%分位数，同时表中也给出了实际考试分数的平均分，以及 20%，40%，60%，80%分位数。结果表明基于混合偏正态分布的估计结果与实际结果比较接近。

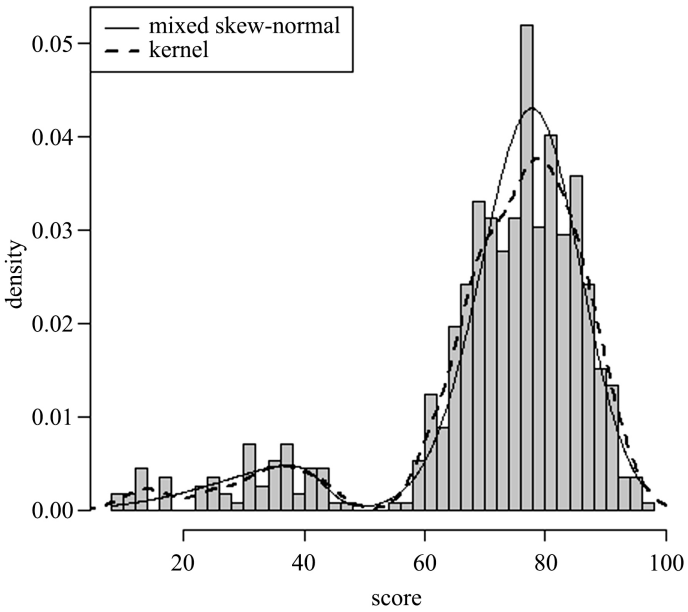

Figure 2. Fitting of exam scores to a mixture of skew-normal distribution
图 2. 考试成绩混合偏正态分布拟合情况

Table 8. Mean and quantile values under different distributions
表 8. 不同分布模型下均值以及分位数值

模型	均值	20%分位数	40%分位数	60%分位数	80%分位数
实际数据	72.034	66	73	78	84
FMSN	**72.034**	66.404	73.608	78.441	83.481
FMN	71.988	66.610	73.435	78.233	83.391
FMT	71.852	66.444	**73.374**	**78.160**	83.371
FMST	71.977	**66.380**	73.555	78.400	83.441
FMSSL	72.009	66.483	73.622	78.423	83.482
NORM	70.203	57.003	66.867	75.109	**84.134**

除两成分的混合偏正态分布外，我们考虑了两成分的其他混合分布，如混合正态分布(FMN)、混合 t 分布(FMT)、混合偏 t 分布(FMST)，并与混合偏正态分布(FMSN)进行比较，各信息准则的值见表 9。表 8 也给出了各估计分布下均值以及分位数的估计值。表 8 同时还给出了单个正态分布(NORM)模型下平均分以及各分位数的值。表 9 给出的结果表明 FMSN 和 FMN 模型要优于其他模型。表 8 的结果表明相对

其他混合分布模型，单一成分的正态分布模型效果最差。整体 FMSN 模型要略好于 FMN 模型，尤其是尾部分位数。

Table 9. Values of information criteria under different distributions
表 9. 不同分布模型下各信息准则的值

模型	AIC	BIC	EDC	ICL
FMSN	**4369.032**	4399.315	**4388.132**	4399.892
FMN	4375.345	**4396.975**	4388.988	**4398.444**
FMT	4379.042	4404.999	4395.414	4406.633
FMST	4372.516	4407.125	4394.345	4407.852
FMSSL	4371.039	4405.648	4392.868	4406.226

4.2. 基于有限混合－尺度混合偏正态分布误差的线性回归模型

本节考虑影响学生成绩可能的因素。限于数据的可获取性，我们考虑两个影响因素：学生性别(sex)和上一学期《概率论》成绩(prob)，建立如下有限混合偏正态分布误差的线性回归模型(SNLM)：

$$y_i = \beta_0 + \beta_1 sex_i + \beta_2 prob_i + \varepsilon_i, \varepsilon_i \sim \sum_{j=1}^{2} \pi_j SN\left(\mu_j, \sigma_j^2, \lambda_j\right).$$

除误差服从有限混合偏正态分布的线性回归模型(FSNLM)外，我们还考虑了误差服从有限混合偏 t 分布的线性回归模型(FSTLM)和有限混合偏 slash 分布的线性回归模型(FSSLLM)进行了对比分析。计算结果见表 10。

Table 10. Estimation of linear regression model with finite mixture skew error distribution
表 10. 基于有限混合偏态分布的线性回归模型的估计

	FSNLM		FSTLM		FSSLLM		LM	
	估计值	标准误差	估计值	标准误差	估计值	标准误差	估计值	标准误差
β_0	24.534	0.009	24.388	0.828	22.124	0.836	21.580	6.995
β_1	3.240	0.001	3.428	0.048	3.246	0.045	6.583	1.431
β_2	0.499	0.029	0.504	0.015	0.520	0.037	0.510	0.080
μ_1	−37.623	17.000	−40.57	9.059	−35.175	12.162	-	-
μ_2	4.749	16.400	4.963	5.265	4.396	9.457	-	-
σ_1^2	180.315	0.764	162.96	4.203	264.801	3.989	-	-
σ_2^2	83.014	125.000	67.812	258.913	97.031	545.036		
λ_1	−0.685	2.290	−0.704	27.064	1.551	35.168	-	-
λ_2	0.625	0.769	0.313	3.115	−1.175	3.400	-	-
π_1	0.112	0.088	0.109	1.128	0.111	0.593	-	-
π_2	0.888	0.088	0.801	1.128	0.889	0.593	-	-

续表

Loglik	**−2179.592**	−2181.767	−2181.444	−2339.313
AIC	**4379.192**	4383.533	4382.889	4686.627
BIC	**4422.453**	4426.795	4426.15	4703.931
EDC	**4406.478**	4410.819	4410.175	-
ICL	19048.17	17702.45	**15698.22**	-

由表 10 可知，对于两个解释变量，学生性别(sex)、上一学期《概率论》成绩(prob)的 p 值均小于 0.01。因此，两个因素影响都是显著的。

依据信息准则结果，由 Loglik 值、AIC、BIC、EDC 准则值可知，有限混合偏正态分布的线性回归模型的拟合效果最好，在 ICL 准则下有限混合偏 slash 分布的线性回归模型的拟合效果较好。

为进行比较，我们还考虑了误差项服从正态分布，即 $\varepsilon_i \sim N\left(0,\sigma^2\right)$ 的线性回归模型

$$y_{ij} = \beta_0 + \beta_1 sex_i + \beta_2 prob_i + \varepsilon_i, \varepsilon_i \sim N\left(0,\sigma^2\right).$$

模型估计的结果见表 10 的最后两列。由 Loglik 值、AIC 和 BIC 信息准则值可知，所建立的误差项为有限混合偏态分布的线性回归模型比普通的线性回归方程拟合效果好，其中有限混合偏正态分布的线性回归模型的拟合效果最好。

5. 结论

本文针对学生考试成绩有偏、多峰的特点，提出采用有限混合－尺度混合偏正态分布进行分析，利用 ECMC 算法求解模型参数的极大似然估计，数值模拟结果表明该方法是有效可行的。实证结果表明有限混合偏正态分布或有限混合正态分布在对考试成绩进行分析时较正态分布具有较好的拟合效果。同时在对影响考试成绩的因素进行研究时，基于有限混合－尺度混合偏正态误差的线性回归模型也较正态误差线性回归模型拟合效果要好。

基金项目

上海高校市级重点课程建设项目“贝叶斯统计”(2024)；国家自然科学基金项目“多因子试验具有最小支撑点的最优回归设计”(11971318)。

参考文献

[1] 马成有. 对学生成绩“正态分布”现象的看法[J]. 现代交际, 2013(2): 256.

[2] 岳武陵. 对考试结果要用正态分布评价的思考[J]. 科技咨询导报, 2007(22): 226-227.

[3] 喻晓莉. 学生成绩偏离正态分布的原因分析[J]. 重庆科技学院学报, 2006(S1): 106.

[4] 尹向飞. 基于混合正态分布的大学生考试成绩分布的拟合[J]. 统计与决策, 2007(8): 133-135.

[5] 张军舰, 马岱君. 考试成绩的混合正态分布分析[J]. 数理统计与管理, 2021, 40(5): 815-821.

[6] 张国才. 学生学习成绩负偏态分布的合理性[J]. 江苏高教, 2002(2): 74-76.

[7] 李翔, 冯珉, 丁澍, 缪柏其. 考试成绩分布函数特点研究[J]. 中国科学技术大学学报, 2011, 41(6): 531-534.

[8] 李金屏, 黄艺美, 刘蓓等. 一个通用的学生考试成绩分布的数学模型研究[J]. 数学的实践与认识, 2009, 39(11): 88-97.

[9] 彭长生. 大学课程考试成绩影响因素的实证分析——一个本科阶段《计量经济学》的教学案例[J]. 安庆师范学

院学报(社会科学版), 2010, 29(12): 75-78.

[10] 沈家豪, 关颖, 欧春泉, 等. 基于结构方程模型的医学硕士研究生课程考试成绩影响因素分析[J]. 中国卫生统计, 2022, 39(5): 695-698.

[11] 喻铁朔, 李霞, 甘琤. 基于学生成绩回归预测的多模型适用性对比研究[J]. 中国教育信息化, 2020(17): 23-28.

[12] 张莉, 卢星凝, 陆从林, 王邦军, 李凡长. 支持向量机在高考成绩预测分析中的应用[J]. 中国科学技术大学学报, 2017, 47(1): 1-9.

[13] Canale, A., Pagui, E.C.K. and Scarpa, B. (2016) Bayesian Modeling of University First-Year Students' Grades after Placement Test. *Journal of Applied Statistics*, **43**, 3015-3029. https://doi.org/10.1080/02664763.2016.1157144

[14] Mclachlan, G. and Peel, D. (2000) Finite Mixture Models. Wiley. https://doi.org/10.1002/0471721182

[15] Fruithwirth-Schnater, S. and Pyne, S. (2010) Bayesian Inference for Finite Mixtures of Univariate and Multivariate Skew-Normal and Skew-t Distributions. *Biostatistics*, **11**, 317-336. https://doi.org/10.1093/biostatistics/kxp062

[16] Basso, R.M., Lachos, V.H., Cabral, C.R.B. and Ghosh, P. (2010) Robust Mixture Modeling Based on Scale Mixtures of Skew-Normal Distributions. *Computational Statistics and Data Analysis*, **54**, 2926-2941. https://doi.org/10.1016/j.csda.2009.09.031

[17] Prates, M.O., Cabral, C.R.B. and Lachos, V.H. (2013) Mixsmsn: Fitting Finite Mixture of Scale Mixture of Skew-Normal Distributions. *Journal of Statistical Software*, **54**, 1-20. https://doi.org/10.18637/jss.v054.i12

[18] Lachos, V.H., Ghosh, P. and Arellano-Valle, R.B. (2010) Likelihood Based Inference for Skew-Normal Independent Linear Mixed Models. *Statistica Sinica*, **20**, 303-322.