Journal of Data Analysis and Information Processing, 2023, 11, 69-80
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Analysis of College Students' Test Scores Based on Two-Component Mixed Generalized Normal Distribution

Luliang Wen[1*], Haiwu Rong[1*], Yanjun Qiu[2]

[1]Foshan University, Foshan, China
[2]Jinan University, Guangzhou, China
Email: *wenluliang@qq.com, *ronghw@foshan.net

How to cite this paper: Wen, L.L., Rong, H.W. and Qiu, Y.J. (2023) Analysis of College Students' Test Scores Based on Two-Component Mixed Generalized Normal Distribution. *Journal of Data Analysis and Information Processing*, **11**, 69-80.
https://doi.org/10.4236/jdaip.2023.111005

Received: December 19, 2022
Accepted: February 5, 2023
Published: February 8, 2023

Abstract

In order to improve the fitting accuracy of college students' test scores, this paper proposes two-component mixed generalized normal distribution, uses maximum likelihood estimation method and Expectation Conditional Maxinnization (ECM) algorithm to estimate parameters and conduct numerical simulation, and performs fitting analysis on the test scores of Linear Algebra and Advanced Mathematics of F University. The empirical results show that the two-component mixed generalized normal distribution is better than the commonly used two-component mixed normal distribution in fitting college students' test data, and has good application value.

Keywords

Two-Component Mixed Generalized Normal Distribution, Two-Component Mixed Normal Distribution, ECM Algorithm, Test Scores

1. Introduction

With regard to the distribution of test scores, the traditional view is to use normal distribution for statistical analysis and inference. However, in reality, many test scores do not conform to the assumption of normal distribution. Li and Zhang (2021) [1] through theoretical reasoning and analysis of test data, suggest reducing or eliminating the requirements for normal distribution of scores in college course tests. In order to find a more accurate distribution to describe college students' test scores, many scholars have carried out extensive research. Yin (2007) [2], Gu and Chi (2010) [3], Zhang and Ma (2021) [4] used the two-component mixed normal distribution to fit the distribution of college students'

test scores. Through numerical simulation and empirical analysis, it is more accurate and reasonable to use the two-component mixed normal distribution to fit the test scores than the normal distribution.

In this paper, the test scores of linear algebra (2417 samples) and advanced mathematics (2035 samples) of the students of relevant majors in F University in the second semester of 2019-2020 academic year are plotted as a distribution histogram. As shown in Figure 1, through observation, it can be found that the two score distributions show double peaks, and the two peaks are respectively located in the score interval [40, 50] and [70, 80], which indicates that the relevant literature has good applicability to select the two component mixed normal distribution for fitting. Wen *et al.* (2022) [5] proposed the mixed generalized normal distribution, and its degenerate distribution includes the mixed normal distribution. Therefore, this paper first introduces the two-component mixed generalized normal distribution into the analysis of college students' test scores, compares the fitting effects of the plan and two-component mixed normal distribution, and tries to find a better fitting distribution of college students' test scores than the two-component mixed normal distribution.

In terms of content arrangement, Section 2 gives the definition of two-component mixed generalized normal distribution; In Section 3, ECM algorithm is proposed to estimate and simulate the parameters of two-component mixed generalized normal distribution and two-component mixed normal distribution; Section 4 compares and analyzes the fitting effects of the two-component mixed generalized normal distribution and the two-component mixed normal distribution by using the higher mathematics and linear algebra test scores; Section 5 is the conclusion.

2. Two-Component Mixed Generalized Normal Distribution

If variable X is subject to two-component generalized normal distribution, its probability density function is:

$$f\left(x\middle|\lambda,\mu_1,\sigma_1,s_1,\mu_2,\sigma_2,s_2\right) = \lambda\left(\frac{s_1}{2\sigma_1\Gamma\left(1/s_1\right)}\right)\exp\left\{-\left|\frac{x-\mu_1}{\sigma_1}\right|^{s_1}\right\}+\left(1-\lambda\right)\left(\frac{s_2}{2\sigma_2\Gamma\left(1/s_2\right)}\right)\exp\left\{-\left|\frac{x-\mu_2}{\sigma_2}\right|^{s_2}\right\}. \quad (1)$$

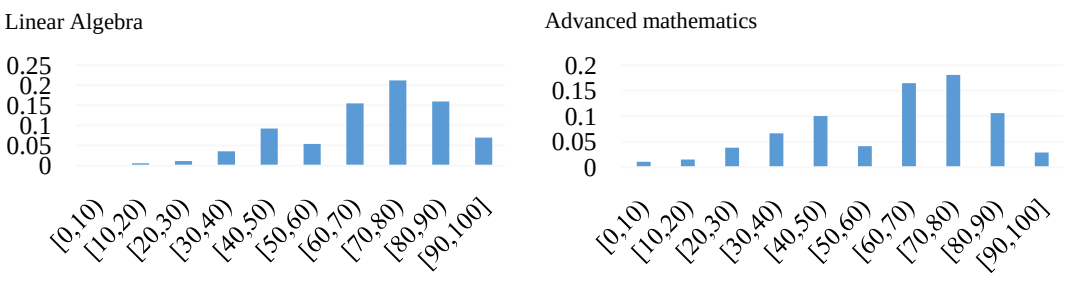


Figure 1. Column chart of the distribution of grades in the 2019-2020-2 linear algebra and advanced mathematics exams of F University.

where $\theta=(\lambda,\mu_1,\sigma_1,s_1,\mu_2,\sigma_2,s_2)$, $\Gamma(1/s_1)=\int_0^{\infty}t^{1/s_1-1}e^{-t}dt$, $\Gamma(1/s_2)=\int_0^{\infty}t^{1/s_2-1}e^{-t}dt$, $0<\lambda<1$, $s_1>0$, $s_2>0$, $\sigma_1>0$, $\sigma_2>0$, $-\infty<\mu_1<\infty$, $-\infty<\mu_2<\infty$, $-\infty<x<\infty$. μ_1,μ_2 is called location parameter, σ_1,σ_2 is called scale parameter, and s_1,s_2 is called shape parameter. When $s_1=s_2=2$, it is a mixed normal distribution.

The expectation and variance of two-component mixed generalized normal distribution are:

$$E(X)=\lambda\mu_1+(1-\lambda)\mu_2, \tag{2}$$

$$Var(X)=\lambda\frac{\sigma_1^2\Gamma(3/s_1)}{\Gamma(1/s_1)}+(1-\lambda)\frac{\sigma_2^2\Gamma(3/s_2)}{\Gamma(1/s_2)}+\lambda(1-\lambda)(\mu_1-\mu_2)^2. \tag{3}$$

Given the value of the parameter, the probability density function of the two-component mixed generalized normal distribution and two-component mixed normal distribution can be drawn. It is found from **Figure 2** that it is a bimodal asymmetric graph, in which the thick tail of the control distribution is smaller, and the tail is thicker. By comparing the shapes in **Figure 1** and **Figure 2**, it can be preliminarily judged that it is feasible to use the two-component mixed generalized normal distribution and the two-component mixed normal distribution to fit college students' test scores.

3. Parameter Estimation

Expectation Maximization (EM) algorithm is an effective method to solve mixed distribution parameter estimation. Each iteration is divided into two steps: E-step and M-step (Dempster *et al.*, 1977) [6].

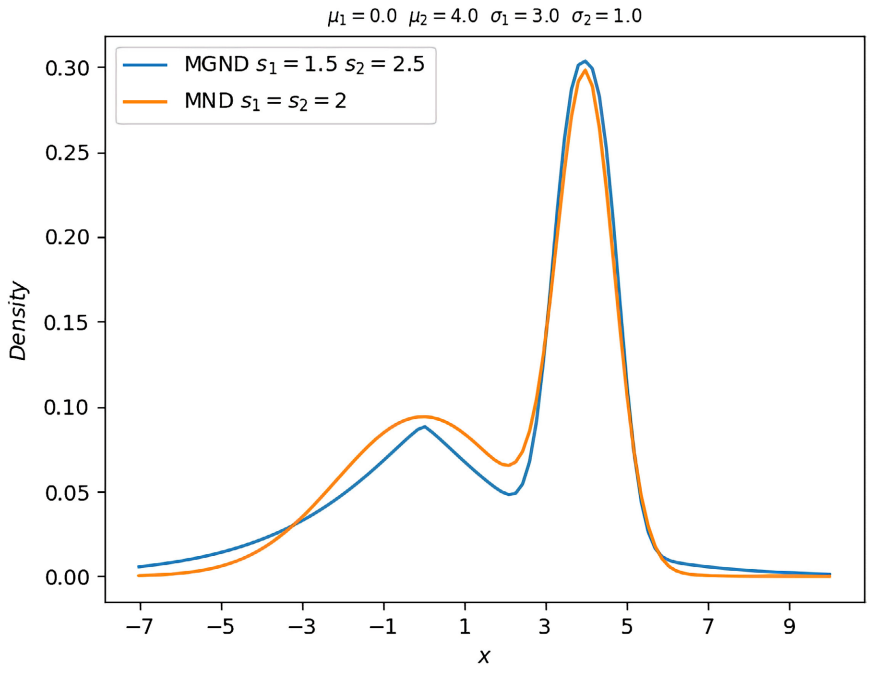


Figure 2. Probability density function diagram of two-component mixed generalized normal distribution (MGND) and two-component mixed normal distribution (MND).

E-step: According to the observed data and the estimated initial values of the current parameters, first calculate the log likelihood function $\log f\left(\theta\middle|X,Z\right)$ of the complete data, and then calculate the conditional expectation about the potential data Z:

$$Q\left(\theta\middle|\theta^{(t)},X\right)=E_Z\left[\log f\left(\theta\middle|X,Z\right)\middle|\theta^{(t)},X\right]=\int\log f\left(\theta\middle|X,Z\right)f\left(Z\middle|\theta^{(t)},X\right)\mathrm{d}Z$$

M-step: Maximize $Q\left(\theta\middle|\theta^{(t)},X\right)$, solve $\theta^{(t+1)}$, make $Q\left(\theta^{(t+1)}\middle|\theta^{(t)},X\right)=\max_{\theta}Q\left(\theta\middle|\theta^{(t)},X\right)$, an iteration is completed $\theta^{(t)}\to\theta^{(t+1)}$, and repeated until $\left|Q\left(\theta^{(t+1)}\middle|\theta^{(t)},X\right)-Q\left(\theta^{(t)}\middle|\theta^{(t)},X\right)\right|$ is sufficiently small. This is the basic principle of EM algorithm.

Meng and Rubin (1993) [7] proposed a special EM algorithm called ECM or GEM algorithm. It decomposes the M-step in the original EM algorithm into the next K^{th} conditional maximization: in the $i+1$ iteration, remember that $\theta^{(i)}=\left(\theta_1^{(i)},\theta_2^{(i)},\cdots,\theta_k^{(i)}\right)$, after obtaining $Q\left(\theta\middle|\theta^{(t)},Y\right)$, first, under the condition of $\theta_1^{(i)},\theta_2^{(i)},\cdots,\theta_k^{(i)}$ keeping unchanged, $Q\left(\theta\middle|\theta^{(i)},Y\right)$ seek to maximize $\theta_1^{(i+1)}$, and then under the conditions of $\theta_1=\theta_1^{(i+1)}$, $\theta_j=\theta_j^{(i)}$, $j=3,\cdots,k$, $Q\left(\theta\middle|\theta^{(i)},Y\right)$ seek to maximize $\theta_2^{(i+1)}$. Continue like this. After the K^{th} condition maximized, we can get $\theta^{(i+1)}$ and complete an iteration.

Chen *et al.* (2015) [8] used iterative Newton Raphson algorithm to solve the parameter estimation problem of generalized linear mixed model (GLMM). In this section, ECM algorithm is mainly used for parameter estimation and numerical simulation of two-component mixed generalized normal distribution.

3.1. ECM Algorithm

If the random sample obeys two-component mixed generalized normal distribution, its logarithmic likelihood function is:

$$\begin{aligned}\log L\left(\theta\right)=\sum_{j=1}^{n}\log\Bigg\{&\lambda\left(\frac{s_1}{2\sigma_1\Gamma\left(1/s_1\right)}\right)\exp\left\{-\left|\frac{x_j-\mu_1}{\sigma_1}\right|^{s_1}\right\}\\&+\left(1-\lambda\right)\left(\frac{s_2}{2\sigma_2\Gamma\left(1/s_2\right)}\right)\exp\left\{-\left|\frac{x_j-\mu_2}{\sigma_2}\right|^{s_2}\right\}\Bigg\}\end{aligned}\tag{4}$$

with reference to Wen *et al.* (2022) [4], the maximum likelihood estimation of the two-component mixed generalized normal distribution ECM algorithm in the case of complete data is given below. Set the sample $X_1,X_2,\cdots,X_n$ with the capacity of n from the two-component mixed generalized normal distribution, $x_1,x_2,\cdots,x_n$ are the sample observation values:

$$f_i\left(x_i\middle|\lambda,\mu_1,\mu_2,\sigma_1,\sigma_2,s_1,s_2\right)=f_i\left(x_i\middle|\theta\right)=\lambda f_{1i}+\left(1-\lambda\right)f_{2i}\tag{5}$$

where $f_{1i}=f_{1i}\left(x_i\middle|\mu_1,\sigma_1,s_1\right)=\dfrac{s_1}{2\sigma_1\Gamma\left(1/s_1\right)}\exp\left\{-\left|\dfrac{x_i-\mu_1}{\sigma_1}\right|^{s_1}\right\}$,

$$f_{2i} = f_{2i}\left(x_i \middle| \mu_2, \sigma_2, s_2\right) = \frac{s_2}{2\sigma_2\Gamma\left(1/s_2\right)} \exp\left\{-\left|\frac{x_i-\mu_2}{\sigma_2}\right|^{s_2}\right\}.$$

The indicator function I_i is introduced. Suppose it follows the two-point distribution:

$$P\left(I_i=1\right)=\lambda,\quad P\left(I_i=0\right)=1-\lambda,$$

since the X_i of generalized normal distribution population from f_{1i} or f_{2i} is unknown, the joint distribution of X_i and I_i is: $g\left(x_i, I_i, \theta\right) = \left(\lambda f_{1i}\right)^{I_i}\left[\left(1-\lambda\right) f_{2i}\right]^{1-I_i}$, and in a given case X_i, the conditional distribution of I_i is:

$$P\left(I_i=1\middle|x_i,\theta\right)=\frac{\alpha f_{1i}}{f_i},\quad P\left(I_i=0\middle|x_i,\theta\right)=\frac{\left(1-\alpha\right) f_{2i}}{f_i}.$$

E-Step: seek expectations.

$$\begin{aligned} Q\left(\theta,\theta^{(m-1)}\right) = &\sum_{i=1}^{n} z_{1i}^{(m-1)}\left(\log\frac{\lambda s_1}{2\sigma_1\Gamma\left(1/s_1\right)} - \left|\frac{x_i-\mu_1}{\sigma_1}\right|^{s_1}\right) \\ &+\sum_{i=1}^{n} z_{2i}^{(m-1)}\left(\log\frac{\left(1-\lambda\right)s_2}{2\sigma_2\Gamma\left(1/s_2\right)} - \left|\frac{x_i-\mu_2}{\sigma_2}\right|^{s_2}\right). \end{aligned} \tag{6}$$

where $z_{1i}^{m-1} = \dfrac{\lambda^{(m-1)} f_{1i}^{(m-1)}}{f_i^{(m-1)}}$, $z_{2i}^{m-1} = \dfrac{\left(1-\lambda^{(m-1)}\right) f_{2i}^{(m-1)}}{f_i^{(m-1)}}$.

CM-Step: maximizing conditions.

$\Psi\left(\dfrac{1}{\upsilon}\right)$ represents the digamma function and $\Psi'\left(\dfrac{1}{\upsilon}\right)$ represents the trigamma function. The iterative formula for the seven parameters of the two-component mixed generalized normal distribution is derived as follows:

$$\begin{cases} \lambda^{(m)} = \dfrac{\sum\limits_{i=1}^{n} z_{1i}^{(m-1)}}{\sum\limits_{j=1}^{2}\sum\limits_{i=1}^{n} z_{ji}^{(m-1)}} = \dfrac{\sum\limits_{i=1}^{n} z_{1i}^{(m-1)}}{\sum\limits_{i=1}^{n}\left(z_{1i}^{(m-1)}+z_{2i}^{(m-1)}\right)} \\ \mu_j^{(m)} = \mu_j^{(m-1)} + \dfrac{\sum\limits_{x_i\ge\mu_j^{(m-1)}} z_{ji}^{(m-1)}\left(x_i-\mu_j^{(m-1)}\right)^{s_j^{(m-1)}-1} - \sum\limits_{x_i<\mu_j^{(m-1)}} z_{ji}^{(m-1)}\left(\mu_j^{(m-1)}-x_i\right)^{s_j^{(m-1)}-1}}{\left(\sigma_j^{(m-1)}\right)^{\upsilon_j^{(m-1)}-2}\left(s_j^{(m-1)}-1\right)\sum\limits_{i=1}^{n} z_{ji}^{(m-1)}\left|\dfrac{x_i-\mu_j^{(m-1)}}{\sigma_j^{(m-1)}}\right|^{s_j^{(m-1)}-2}}, \quad j=1,2 \\ \sigma_j^{(m)} = \left[\dfrac{s_j^{(m-1)}\cdot\sum\limits_{i=1}^{n} z_{ij}^{(m-1)}\left|x_i-\mu_j^{(m)}\right|^{s_j^{(m-1)}}}{\sum\limits_{i=1}^{n} z_{ij}^{(m-1)}}\right]^{\frac{1}{s_j^{(m-1)}}}, \quad j=1,2 \\ s_j^{(m)} = s_j^{(m-1)} - \dfrac{\sum\limits_{i=1}^{n} z_{ij}^{(m-1)}A_1 - \sum\limits_{i=1}^{n} z_{ij}^{(m-1)}A_2}{\sum\limits_{i=1}^{n} z_{ij}^{(m-1)}A_3 - \sum\limits_{i=1}^{n} z_{ij}^{(m-1)}A_4}, \quad j=1,2 \end{cases} \tag{7}$$

where

$$A_1 = \frac{1}{s_j^{(m-1)}}\left(\frac{1}{s_j^{(m-1)}}\Psi\left(\frac{1}{s_j^{(m-1)}}\right)+1\right),$$

$$A_2 = \left|\frac{x_i-\mu_j^{(m)}}{\sigma_j^{(m)}}\right|^{s_j^{(m-1)}} \log\left|\frac{x_i-\mu_j^{(m)}}{\sigma_j^{(m)}}\right|,$$

$$A_3 = -\frac{1}{s_j^{2(m-1)}}\left[1+\frac{2}{s_j^{(m-1)}}\Psi\left(\frac{1}{s_j^{(m-1)}}\right)+\frac{1}{s_j^{2(m-1)}}\Psi'\left(\frac{1}{s_j^{(m-1)}}\right)\right],$$

$$A_4 = \left|\frac{x_i-\mu_j^{(m)}}{\sigma_j^{(m)}}\right|^{s_j^{(m-1)}} \left(\log\left|\frac{x_i-\mu_j^{(m)}}{\sigma_j^{(m)}}\right|\right)^2.$$

On the basis of known observation data, numerical iterative method can be used to solve the above equations, because the transcendental equation is involved, the solution process is difficult. Two propositions of consistency and asymptotic normality for maximum likelihood estimation of two-component mixed generalized normal distribution are given below.

Proposition 1: (Consistency) For the two-component mixed generalized normal distribution, given arbitrarily θ_0, its maximum likelihood estimates $\hat{\theta}$ are continuous in the interval, and then $\hat{\theta}$ converge to in probability θ_0, that is $\hat{\theta}\xrightarrow{p}\theta_0$.

It is proved that the parameters $\theta=(\alpha,\mu_1,\sigma_1,s_1,\mu_2,\sigma_2,s_2)$ of the two-component mixed generalized normal distribution are assumed to be open sets:

$$\Theta=(0,1)\times(-\infty,+\infty)\times(0,+\infty)\times(0,+\infty)\times(-\infty,+\infty)\times(0,+\infty)\times(0,+\infty)$$

Let $\theta_0=(\alpha_0,\mu_{01},\sigma_{01},s_{01},\mu_{02},\sigma_{02},s_{02})$ denote the true parameter value. Assume that for any $\theta_0=(\alpha_0,\mu_{01},\sigma_{01},s_{01},\mu_{02},\sigma_{02},s_{02})$, there is a compact set $\theta\subset\Theta$ that satisfies:

1) $\theta_0\in\theta$,
2) $\forall\xi\neq\xi_0,\xi\in\theta$, $f(x_i|\theta)\neq f(x_i|\theta_0)$,
3) $\forall\xi\in\theta$, $\log f(x_i|\xi)$ is a continuous function,
4) $E\left[\sup_\xi\left|f(x_i|\xi)\right|\right]<\infty$.

According to the relevant theorem content of Newey and McFadden (1994) [9], it can be proved that the two-component mixed generalized normal distribution satisfies the above four conditions.

With reference to some bounded conditions given by Redner and Walker (1984) [10], such as the set value range with s_1,s_2, it can be proved that the maximum likelihood estimator of the two-component mixed generalized normal distribution satisfies the asymptotic normality.

Proposition 2: (Asymptotic normality) If $\upsilon_1>1,\upsilon_2>1$, $I(\theta_0)$ is an information matrix, then the maximum likelihood estimate $\hat{\theta}$ of θ_0 satisfies the asymptotic normality, that is

$$\sqrt{T}\left(\hat{\theta}-\theta_0\right)\xrightarrow{d}N\left(0,I^{-1}(\theta_0)\right).$$

To prove proposition 2, refer to Redner and Walker (1984) [10] and McLachlan and Peel (2004) [11] for the derivation process. It should be noted that the information matrix expression of the two component mixed generalized normal distribution is more complex. In the process of seeking expectations, it is a difficult problem how to effectively simplify an analytic expression.

$$I(\theta_0) \equiv E\left[\frac{\partial \ln f}{\partial \theta_i} \cdot \frac{\partial \ln f}{\partial \theta_j}\right] = -E\left[\frac{\partial^2 \ln f}{\partial \theta_i \partial \theta_j}\right].$$

3.2. Numerical Simulation

Before the analysis of real data, numerical simulation experiments are conducted. In this section, we will evaluate the performance of ECM algorithm for parameter estimation of two-component mixed generalized normal distribution. First, the formula of skewness and mean square error is given:

$$\text{Bias}\left(\hat{\theta}\right) = \left|\frac{1}{N}\sum_{j=1}^{N}\hat{\theta}_j - \theta\right|, \ \text{MSE}\left(\hat{\theta}\right) = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{\theta}_j - \theta\right)^2 \tag{8}$$

where θ represents the real parameter value and $\hat{\theta}_j$ represents the θ estimated value for the j time. Skewness and mean square error are measures that reflect the difference between the estimator and the estimated. The smaller the skewness and mean square error, the better the estimation effect.

Based on the iterative formula given in Section 3.1, and referring to the rounding sampling method proposed by Tadikamalla (1980) [12] and Chiodi (1995) [13], random numbers of two-component mixed generalized normal distribution are generated. Since the sample size of higher mathematics and linear algebra examination scores selected in this paper is 2035 and 2417 respectively, in order to better simulate real data, we choose to generate 2000 and 2500 random numbers for each simulation.

For λ, we select the initial value $\lambda \in (0,1)$; for s_1, s_2, we select the initial values by $s_1 \in [1,4]$, $s_2 \in [1,4]$; for μ_1, μ_2 and σ_1, σ_2, we use the initial values by $\mu_1 \in [1,4]$, $\mu_2 \in [1,4]$, $\sigma_1 \in [1,3]$, $\sigma_2 \in [1,3]$. The convergence criterion is set as $\left|\hat{\theta}^{(m+1)} - \hat{\theta}^{(m)}\right| < 10^{-4}$, the numerical simulation experiment is carried out for 30 times, and the average value is calculated for analysis. Programming calculation with Python software.

Table 1 shows that the skewness of the seven parameters is within 0.25 and the mean square error is within 7% when estimating the parameters of the two-component mixed generalized normal distribution; Table 2 shows that when estimating the five parameters of the two-component mixed normal distribution, the skewness of the parameters is within 0.09 and the mean square error is within 4%. Under the same sample size, it is found that the fewer parameters to be estimated, the better the estimation effect. In general, when the sample size is 2000 and 2500, the parameter estimates of the two distributions have reached convergence. Through the analysis of skewness and mean square error, the parameter estimates are also relatively ideal.

Table 1. Parameter estimation results of two-components mixed generalized normal distribution ($e-0x=10^{-x}$).

True value	$\lambda=0.65$	$\mu_1=1.5$	$\mu_2=3.5$	$\sigma_1=1.2$	$\sigma_2=2.6$	$s_1=3.2$	$s_2=1.5$	sample size
Est.	0.6418	1.4948	3.4443	1.2394	2.8493	3.0563	1.5973	
Bias	0.0082	0.0052	0.0557	0.0394	0.2493	0.1437	0.0973	**2000**
MSE	6.7e−05	2.7e−05	0.0031	0.0016	0.0621	0.0207	0.0095	
Est.	0.6513	1.5053	3.4871	1.2714	2.8015	3.1837	1.5982	
Bias	0.0013	0.0053	0.0129	0.0714	0.2015	0.0163	0.0982	**2500**
MSE	1.6e−05	2.8e−05	0.0002	0.0051	0.0406	0.0003	0.0096	

Table 2. Parameter estimation results of two-components mixed normal distribution ($s_1=s_2=2$).

True value	$\lambda=0.65$	$\mu_1=1.5$	$\mu_2=3.5$	$\sigma_1=1.2$	$\sigma_2=2.6$	$s_1=2$	$s_2=2$	sample size
Est.	0.6804	1.5046	3.6270	1.2811	2.5637	2.0	2.0	
Bias	0.0304	0.0046	0.1270	0.0811	0.0363	0.0	0.0	**2000**
MSE	0.0009	2.2e−05	0.0161	0.0066	0.0013	0.0	0.0	
Est.	0.6922	1.5108	3.703	1.3008	2.5080	2.0	2.0	
Bias	0.0422	0.0108	0.2031	0.1007	0.0920	0.0	0.0	**2500**
MSE	0.0018	0.0001	0.0413	0.0102	0.0085	0.0	0.0	

4. Real Data Analysis

4.1. Descriptive Statistics

We choose the linear algebra (2417 samples) and advanced mathematics (2035 samples) test scores of students of relevant majors in F University in the second semester of 2019-2020 academic year, and the column chart of the distribution is shown in Figure 1. In order to facilitate data analysis, the scores of linear algebra test (abbreviated as XXDS) and advanced mathematics test (abbreviated as GDSX) are normalized. To generate new data for descriptive statistics, the scores of each examinee are set as x_i, $y_i=\frac{x_i}{100}\in[0,1]$.

It can be seen from Table 3 that linear algebra (XXDS) and advanced mathematics (GDSX) have common features, such as a small difference between their mean values; The skewness coefficients are all less than 0, showing the characteristics of left bias, and the kurtosis are all less than 3; At the 5% significance level, the results of J-B statistics are far greater than 5.99, and the assumption of normal distribution is rejected.

4.2. Fitting Evaluation

AIC and BIC criteria are generally used to evaluate the model fitting effect:

$$\begin{cases}\mathrm{AIC}=2\varphi-2\log L\left(\hat{\theta}|x\right),\\ \mathrm{BIC}=\varphi\log(n)-2\log L\left(\hat{\theta}|x\right),\end{cases} \tag{9}$$

Table 3. Descriptive statistics of linear algebra (XXDS) and advanced mathematics (GDSX).

	sample size	Mean	Std.	Skewness	Kurtosis	J-B value	P-value
XXDS	2417	0.685	0.175	−0.684	0.087	189.483	0
GDSX	2035	0.610	0.205	−0.682	0.179	160.465	0

Table 4. Estimates of two-component mixed generalized normal distribution (MGND) and two-component mixed normal distribution (MND) parameters using the ECM algorithm for XXDS and GDSX test score data.

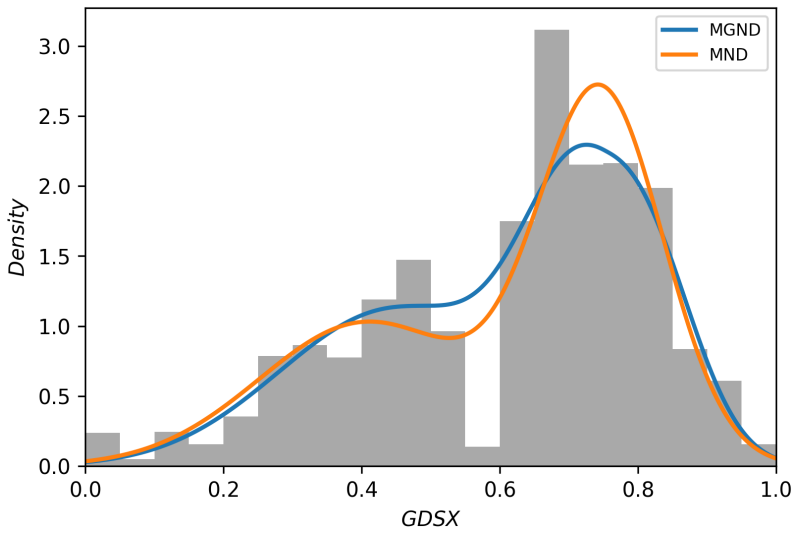
Parameter	MND		MGND		Data Sources
	Est.	S.E.	Est.	S.E.	
λ	0.2674	3.4e−05	0.5166	0.0213	XXDS
	0.4030	3.7e−06	0.4918	0.0533	GDSX
μ_1	0.4493	0.0008	0.5195	0.0057	XXDS
	0.4088	2.4e−06	0.4593	0.0299	GDSX
μ_2	0.7710	3.7e−06	0.7925	0.0005	XXDS
	0.7464	4e−07	0.7541	0.0056	GDSX
σ_1	0.1573	2.2e−05	0.1626	0.0030	XXDS
	0.2210	2.1e−06	0.2448	0.0503	GDSX
σ_2	0.1380	2.6e−06	0.1617	0.0006	XXDS
	0.1285	2.4e−07	0.1442	0.0024	GDSX
s_1	2.0	--	1.2350	0.0069	XXDS
	2.0	--	2.0822	0.1034	GDSX
s_2	2.0	--	5.7705	0.2971	XXDS
	2.0	--	2.4348	0.3449	GDSX
$L\left(\hat{\theta}\middle\|x\right)$	−297.2428		−226.0337		XXDS
	−312.0936		−307.7131		GDSX
AIC	604.4856		466.0673		XXDS
	634.1871		629.4261		GDSX
BIC	645.0176		506.5993		XXDS
	673.5149		668.7539		GDSX

where φ represents the number of parameters and the sample size. The smaller the AIC and BIC values, the better the fitting effect. To achieve more accurate results, the parameter estimation is repeated for 30 times to get the final result.

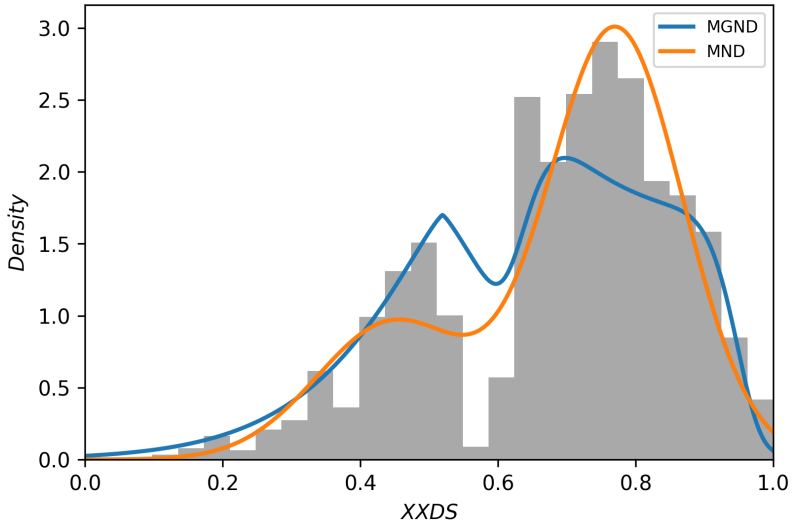
Table 4 shows the parameter estimation and AIC and BIC calculation results of the two-component mixed normal distribution and two-component generalized normal distribution using ECM algorithm. By analyzing the calculation results of AIC and BIC, it can be seen that the fitting effect of the two-component mixed generalized normal distribution is obviously better than the two-component mixed normal distribution for the test results of linear algebra (XXDS), and for

the test results of advanced mathematics (GDSX), the fitting effect of the two-component mixed generalized normal distribution is slightly better than the two-component mixed normal distribution, and there is little difference between the two distributions.

Through the analysis of the parameter estimation s_1, s_2 results, we can see that for the linear algebra test scores, the estimation s_1 results are significantly less than 2, and the s_2 results are significantly greater than 2. For the estimation of the higher mathematics scores, we find that the s_1, s_2 results are all near 2, which also explains the reason why the AIC and BIC results of the two-component mixed generalized normal distribution and the two-component mixed normal distribution are very close.



(a)



(b)

Figure 3. Histograms of fitting linear algebra (XXDS) and advanced mathematics (GDSX) test scores using two-component mixed generalized normal distribution (MGND) and two-component mixed normal distribution (MND).

Figure 3 is the histogram of the two-component mixed normal distribution and the two-component generalized normal distribution. It can be found that the two-component generalized mixed normal distribution better fits the peak on the left, and the two-component mixed normal distribution better fits the peak on the right. In general, the fitting effect of the two-component generalized mixed normal distribution is better than the two-component mixed normal distribution, which is consistent with the conclusion drawn from the analysis of AIC and BIC calculation results.

5. Conclusion

In this paper, a two-component mixed generalized normal distribution is proposed, and the results of linear algebra and higher mathematics examinations are fitted and analyzed. The following conclusions are drawn: 1) When studying the maximum likelihood estimation of two-component mixed generalized normal distribution, ECM algorithm is proposed to estimate parameters, which is verified to be an effective method by numerical simulation. 2) Through the comparative analysis of the fitting effect of two-component mixed generalized normal distribution and two-component mixed normal distribution on college students' math test scores, the empirical results show that the overall fitting effect of two-component mixed generalized normal distribution is better than that of two-component mixed normal distribution, especially in characterizing low or high scoring groups, it avoids the problem of too much or too little tail fitting of two-component mixed normal distribution, It is the innovation of research methods of Zhang and Ma (2021) [4] and other scholars. 3) For the "bimodal distribution" of college students' test scores, there are mainly two types of students, one is the students who fail the test, and the other is the students who pass the test. Huang *et al.* (2019) [14] gave a statistical analysis of the influencing factors for students who failed in the exam. The two-component mixed generalized normal distribution proposed in this paper has a good application value for accurately analyzing the test scores of different types of college students and optimizing the teaching methods of different types of students.

Acknowledgements

We thank the reviewers for their valuable suggestions, which helped us to improve the manuscript. This research is supported by the 13th Five-Year Plan of Philosophy and Social Sciences of Guangdong Province (No. GD20XYJ19).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Li, L. and Zhang, W.H. (2021) Analysis on Normal Distribution of College Course

Examination Results. *China Exam*, **4**, 86-93.

[2] Yin, X.F. (2007) Fitting the Distribution of College Students' Test Scores Based on Mixed Normal Distribution. *Statistics and Decision Making*, **8**, 133-135.

[3] Gu, C.C. and Chi, Z.Y. (2010) Research on the Distribution Law of Students' Achievements. *Journal of Anyang Institute of Technology*, **9**, 88-90.

[4] Zhang, J.J. and Ma, D.J. (2021) Analysis of Mixed Normal Distribution of Test Results. *Mathematical Statistics and Management*, **40**, 815-821.

[5] Wen, L.L., Qiu, Y.J., Wang, M.H., Yin, J.L. and Chen, P.Y. (2022) Numerical Characteristics and Parameter Estimation of Finite Mixed Generalized Normal Distribution. *Communications in Statistics—Simulation and Computation*, **51**, 3596-3620. https://doi.org/10.1080/03610918.2020.1720733

[6] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*: *Series B* (*Methodological*), **39**, 1-22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

[7] Meng, X.L. and Rubin, D.B. (1993) Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika*, **80**, 267-278. https://doi.org/10.1093/biomet/80.2.267

[8] Chen, Y., Fei, Y. and Pan, J.X. (2015) Statistical Inference in Generalized Linear Mixed Models by Joint Modelling Mean and Covariance of Non-Normal Random Effects. *Open Journal of Statistics*, **5**, 568-584. https://doi.org/10.4236/ojs.2015.56059

[9] Newey, W.K. and McFadden, D. (1994) Large Sample Estimation and Hypothesis Testing. *Handbook of Econometrics*, **4**, 2111-2245. https://doi.org/10.1016/S1573-4412(05)80005-4

[10] Redner, R.A. and Walker, H.F. (1984) Mixture Densities, Maximum Likelihood and the EM Algorithm. *SIAM Review*, **26**, 195-239. https://doi.org/10.1137/1026034

[11] McLachlan, G. J. and Peel, D. (2004) Finite Mixture Models. John Wiley & Sons, New York.

[12] Tadikamalla, P. (1980) Random Sampling from the Exponential Power Distribution. *Journal of the American Statistical Association*, **75**, 683-686. https://doi.org/10.1080/01621459.1980.10477533

[13] Chiodi, M. (1995) Generation of Pseudo-Random Variates from a Normal Distribution of Order P. *Italian Journal of Applied Statistics*, **7**, 401-416.

[14] Huang, G.J., Ou, S.D. and Li, Q. (2019) Estimation of Failure Rate of College Mathematics Examination and Statistical Analysis of Its Influencing Factors. *Journal of Guangxi University*: *Natural Science Edition*, **44**, 1835-1841.