Research Article

A Novel Massive Big Data Analysis of Educational Examination Research Using a Linear Mixed-Effects Model

Jing Zhao and Yiwen Wang

1Beijing Normal University, Business School, Beijing 100875, China
2Beijing Normal University, Faculty of Education, Beijing 100875, China

Correspondence should be addressed to Jing Zhao; 201931030027@mail.bnu.edu.cn

Received 13 May 2021; Accepted 5 July 2021; Published 26 July 2021

Academic Editor: Huihua Chen

Copyright © 2021 Jing Zhao and Yiwen Wang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

To further solve the problems of storage bottlenecks and excessive calculation time when calculating estimators under two different formats of massive longitudinal data, an examination data analysis and evaluation method based on an improved linear mixed-effects model is proposed in this paper. First, a three-step estimation method is proposed to improve the parameters of the linear-effects model, avoiding the complicated iterative steps of maximum likelihood estimation. Second, we perform spectral clustering based on test data on the basis of defining data attributes and basic evaluation rules. Finally, based on cloud technology, a cross-regional, multiuser educational examination big data analysis and evaluation service platform is developed for evaluating the proposed method. Experimental results have shown that the proposed model can not only effectively improve the efficiency of test data acquisition and storage but also reduce the computational burden and the memory usage, solve the problem of insufficient memory, and increase the calculation speed.

1. Introduction

Nowadays, the era of big data affects many industries. Among them, the education industry has also improved its educational examinations through the use of big data technology [1]. In traditional educational examinations, teachers have very little understanding of the role of examinations. Teachers believe that examinations can only provide a simple understanding of students’ knowledge and skills. However, the results measured by traditional educational examinations have many uncertain factors [2]. In terms of teaching examinations, schools can carry out large-scale networked education examinations and online learning. As long as they are from the perspective of big data, students can participate in online learning activities. In the process of students learning through the Internet, teachers can accurately grasp the status of students’ learning ability to find students’ deficiencies and, at the same time, can improve students’ nonintellectual factors, such as emotional intelligence, values, learning motivation, and enthusiasm.

Educational examination big data analysis and evaluation is to mine the original test data of students, feedback a large amount of information and laws hidden in the data relationship, and assist the education department, teachers, and students to analyze the reasons for the formation of test results. The relationship between the test results and the actual ability of students is built to improve the process and methods of education and assessment on this basis. Scientific, accurate, and comprehensive test data analysis and
evaluation are of great significance to students’ self-diagnosis, teachers’ feedback and summation, and educational test policy formulation, which is embodied in [3]:

(1) A rational self-diagnosis helps students to understand the strengths and weaknesses of their own learning and improve their sense of self-efficacy and self-psychology

(2) Teachers can analyze and summarize test results, link teaching, feedback teaching, and reflect on teaching

(3) Educational authorities can adjust education examination policies and improve the forms and methods of examination evaluation through examination evaluation and analysis results

(4) Through multidimensional analysis of student test results, the trust and cooperation between the parents and the school can be strengthened, to provide reference for the joint development of a student’s individual study and life plan

However, the current examination data analysis and evaluation mainly have the following shortcomings [4]:

(1) Simplification of the examination data processing

(2) Fuzzy analysis of test results

(3) Simplification of the test result feedback

However, the current simplification of test results can easily lead to the rigidity of the test question system and loss of the vitality of the test questions, which is not conducive to the control of the education authorities. The overall education and teaching situation formulate appropriate education and teaching policies.

With the rapid development of information and network technology, the speed of data growth and change is amazing, showing a trend of data massive quantification [5]. Traditional linear mixed model estimation methods have encountered a series of challenges under massive data, such as storage bottlenecks, computational efficiency, and other problems [6], so it is necessary to explore new algorithms to improve the previous estimation methods of linear mixed effects models. In order to construct the optimal estimation of the coefficients in the linear mixed effects model, [7] proposed a weighted least squares method based on the covariance structure, but the efficiency of this estimation method is closely related to the covariance structure estimation method of longitudinal data. There are some commonly used covariance structure estimation methods, including maximum likelihood method and limited maximum likelihood method. However, these methods require repeated iterative calculations and optimization steps, which will lead to high calculation time and quantity cost when the amount of data is very large.

To solve such problems, [8] proposed a simple method for estimating variance components when studying random effect variable coefficient models. This method does not require iterative calculation and can adapt to massive data situations. However, directly using the weighted least squares method in matrix form to estimate model coefficients will encounter storage problems, so the divide-and-conquer algorithm has been proposed and has been widely used in the statistical calculation of massive data [9]. The core idea of the divide-and-conquer algorithm is to first decompose a complex problem into several subproblems, find the solutions to these subproblems, and then use a suitable method to combine the results of each subproblem to get the result of the entire problem. The divide-and-conquer algorithm proposed by [10] is the most practical method to solve the problem of massive data analysis. Reference [11] combined the least squares method with the divide-and-conquer algorithm to solve the estimation problem of linear regression models under massive data. The dataset is divided into a series of manageable data blocks. Then, the least square method is run independently on each data block, and finally, the results of each data block are integrated to obtain the final result. However, previous scholars only used the divide-and-conquer algorithm to solve the estimation problem of simple data models under massive data and have not studied the estimation algorithms of more complex and more widely used linear mixed-effect models under massive data.

Based on the weighted least squares estimation method of linear mixed-effects model coefficients in [12] and the estimation method of variance components and divide-and-conquer algorithm in [13], this paper proposes an improved linear mixed-effects model [14]. First, a three-step estimation method for linear mixed model parameters is proposed based on the conventional linear mixed-effects model. In the estimation method, the iterative calculation based on maximum likelihood estimation or limited maximum likelihood estimation is avoided for the requirements of massive data. Then, based on the analysis of two different situations of massive data, the rule algorithm calculates the estimator in steps. Experiments show that the algorithm proposed in this paper can not only reduce memory usage and solve storage problems but also shorten computing time and increase computing speed. Aiming at the single, one-sided, and complex characteristics of the educational examination quality evaluation process, as well as the resulting unscientific and nonobjective examination evaluation problems, based on the improved linear mixed effect model, we have developed a large amount of educational examination big data and the application analysis of the proposed method, including the following.

(1) Define test data attributes, including forward attributes and reverse attributes, and establish gradient evaluation rules

(2) On the basis of defining data attributes and basic evaluation rules, perform hierarchical filtering of test data based on spectral clustering to identify basic group clusters

(3) Based on the gradient evaluation rule, the group cluster is corrected twice to avoid the unscientificity of the single score evaluation mechanism

(4) Based on cloud technology, a cross-regional, multiuser educational examination big data analysis and evaluation service platform has been developed
Application results have proved that the platform can effectively improve the efficiency of test data acquisition and storage, and, at the same time, provide effective analysis tools for education supervisors and implementation departments.

2. Improved Linear Mixed-Effects Model

2.1. Linear Mixed-Effects Model and Its Estimation Method

Consider the linear mixed effects model of longitudinal data:

\[ Y_{ij} = X_{ij}'\beta + Z_{ij}'b_i + \epsilon_{ij}, \quad i = 1, 2, \ldots, n; j = 1, 2, \ldots, n. \]  

(1)

Here, the unknown parameter \( \beta \) is a \( p \times 1 \) dimensional fixed effect vector, \( b_i \) is a \( q \times 1 \) dimensional random effect vector, and \( X_{ij} \) and \( Z_{ij} \), respectively, represent the covariates related to it. Assume that the random effects \( b_i \sim N(0, \Sigma) \), and the model error is \( \epsilon_{ij} \sim N(0, \sigma^2) \). Both satisfy the condition of independent and identical distribution, and \( b_i \) and \( \epsilon_{ij} \) are independent of each other.

To avoid iterative calculation, a three-step estimation method is proposed to estimate the variance components and fixed effects of the linear mixed-effects model, respectively. The consistency and asymptotic normality of the estimates can be obtained by using the proof steps similar to those in [15]. Therefore, this paper only considered the estimation algorithm. First, we use the least square method to make a preliminary estimate of the coefficients. Let \( Y = (Y_1^T, Y_2^T, \ldots, Y_n^T)^T \) and \( \epsilon \) are similar defined, and denote \( b = (b_1^T, b_2^T, \ldots, b_n^T)^T \), \( Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{in})^T \), \( Z_i = \text{diag}(Z_1, Z_2, \ldots, Z_n) \). Then, (1) can be reexpressed as follows:

\[ Y = X\beta + Zb + \epsilon. \]  

(2)

The least squares estimation of the model coefficients is directly obtained by (2). Then, a preliminary estimation can also be obtained by the following equation:

\[ \hat{\beta}_0 = (X^TX)^{-1}X^TY. \]  

(3)

Second, the variance component is estimated. We use the principle of [15] to estimate the variance component \( \sigma^2, \Sigma \).

Let \( U_{ij} = Z_{ij}'b_i + \epsilon_{ij} \), then \( Y_{ij} = X_{ij}'\beta + U_{ij} \), and furthermore, \( \bar{U}_{ij} = Y_{ij} - X_{ij}'\hat{\beta} \). Let \( U_i = (U_{i1}, U_{i2}, \ldots, U_{in})^T \), then \( \bar{U}_i = (\bar{U}_{i1}, \bar{U}_{i2}, \ldots, \bar{U}_{in})^T \), and furthermore, \( U_i = Z_i b_i + \epsilon_i \). Therefore, the estimated value of random effect \( \hat{b}_i \) can be obtained by using the least square method \( \hat{b}_i = (Z_i^TZ_i)^{-1}Z_i^T\bar{U}_i \). Furthermore, note that \( \bar{U}_i = Z_i b_i = Z_i(Z_i^TZ_i)^{-1}Z_i^T U_i \equiv H_i U_i \), and then the residual sum of squares of the model \( U_i = Z_i b_i + \epsilon_i \) is \( S_i = (U_i - U_i)^T(U_i - U_i) = U_i^T U_i - U_i^T H_i U_i \), Thus,

\[ S_i = \bar{U}_i^T \bar{U}_i - U_i^T H_i U_i. \]  

(4)

According to the estimation procedure in [15], further considering the influence of the estimated value of the coefficient \( \hat{\beta}_0 \), the estimation of the variance can be constructed similarly:

\[ \hat{\sigma}^2 = \left[ \sum_{i=1}^{n} (n_i - q) - p \right]^{-1} \sum_{i=1}^{n} \hat{S}_i = (N - nq - p)^{-1} \sum_{i=1}^{n} \hat{S}_i, \]  

(5)

where \( N = \sum_{i=1}^{n} n_i \), \( q \) is the dimension of \( b_i \), and \( p \) is the dimension of \( \beta \). Next, the \( \Sigma \) is estimated according to the definition of \( U_i \); \( \hat{b}_i = (Z_i^T Z_i)^{-1}Z_i^T U_i = b_i + (Z_i^T Z_i)^{-1}Z_i^T \epsilon_i \). Furthermore, we have the following equation:

\[ \frac{1}{n} \sum_{i=1}^{n} b_i b_i^T = \frac{1}{n} \sum_{i=1}^{n} \bar{b}_i \bar{b}_i^T + \frac{1}{n} \sum_{i=1}^{n} (Z_i^T Z_i)^{-1} Z_i^T \epsilon_i \epsilon_i^T Z_i (Z_i^T Z_i)^{-1} + \frac{1}{n} \sum_{i=1}^{n} (Z_i^T Z_i)^{-1} Z_i^T \epsilon_i \epsilon_i^T Z_i (Z_i^T Z_i)^{-1}. \]  

(6)

By directly calculating the first and second moments of each item, it can be proved that the order of the last two items in (6) is \( O_p(n^{1/2}) \), so it can be ignored compared with the other items. Thus,

\[ \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \bar{b}_i \bar{b}_i^T - \sigma^2 \frac{1}{n} \sum_{i=1}^{n} (Z_i^T Z_i)^{-1}. \]  

(7)

So, we can obtain

\[ \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \bar{b}_i \bar{b}_i^T - \sigma^2 \frac{1}{n} \sum_{i=1}^{n} (Z_i^T Z_i)^{-1}. \]  

(8)

Finally, calculate the weighted least squares estimate of \( \beta \):

\[ V = \text{diag}(V_1, V_2, \ldots, V_n), \]

\[ V_i = \text{var}(Z_i b_i + \epsilon_i) = Z_i \sum Z_i^T + \sigma^2 I_n, \]

(9)

From (6) and (8), the following is obtained:

\[ \hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} X. \]  

(11)

Or we also can use the calculation of the sample summation equation (12) in [13]:

\[ \hat{\beta} \]
2.2. Improved Linear Mixed-Effects Model for Massive Education Big Data Analysis. Consider two scenarios of massive data:

1. The amount of individual data is larger, but the amount of data in the group is smaller,
2. The amount of individual data is smaller, but the amount of data in the group is larger.

At this time, storage bottlenecks and computational inefficiencies will be encountered, and the previously mentioned estimation methods are no longer applicable. Therefore, in these two scenarios, the divide-and-conquer algorithm is used to adjust the three-step estimation method proposed previously.

2.2.1. The Amount of Individual Data Is Large, and the Amount of Data in the Group Is Small. Let $D = \{ (X_{ij}, Z_{ij}, Y_{ij}) \}_{i=1,j=1}^{n,m}$ be the entire dataset. According to the divide-and-conquer algorithm, the dataset $D$ is divided into $K$ subsets $D_1, D_2, \ldots, D_K$. The number of samples in each subset $m = n/K$, and the data contained in each subset is $X_{kj}, Z_{kj}, Y_{kj}$, where $1 \leq j \leq n_{kj}$, $1 \leq i \leq m$, $1 \leq k \leq K$, and the data contained in each subset cannot exceed the processing capacity of a single machine. First, we have calculated a preliminary estimate. Remember $X_{ki} = (X_{k1i}, X_{k2i}, \ldots, X_{km_i})^T$, $Y_{ki} = (Y_{k1i}, Y_{k2i}, \ldots, Y_{km_i})^T$, and $\epsilon_k$ are similarly defined, and $Z_{ki} = (Z_{k1i}, Z_{k2i}, \ldots, Z_{km_i})^T$, $Z_k = diag(Z_{k1i}, Z_{k2i}, \ldots, Z_{km_i})$, and $b_k = diag(b_{k1i}, b_{k2i}, \ldots, b_{km_i})^T$ are well-defined according to the previously mentioned theory. Then, the data model corresponding to the k-th subset can be written as

$$Y_k = X_k\beta + Z_k\epsilon_k + \epsilon_k.$$  

The least square estimation for each subsystem directly has $\hat{\beta}_{0,k} = (X_k^T X_k)^{-1} X_k^T Y_k$, $1 \leq k \leq K$, according to the divide-and-conquer algorithm, integrate the results of each subset, save each group of $X_k^T X_k, \hat{\beta}_{0,k}$, and finally obtain preliminary estimates:

$$\hat{\beta}_0 = (X^T X)^{-1} X^T Y = \left( \sum_{k=1}^{K} X_k^T X_k \right)^{-1} \left( \sum_{k=1}^{K} X_k^T X_k \hat{\beta}_{0,k} \right).$$  

Second, we use the divide-and-conquer algorithm to estimate the variance component. Remember that $U_{ki} = (U_{k1i}, U_{k2i}, \ldots, U_{km_i})^T$, $U_k = (U_{k1i}, U_{k2i}, \ldots, U_{km_i})^T$ is similar to (4), so we can get

$$\tilde{S}_k = \tilde{U}_k^T \tilde{U}_k - \tilde{U}_k^T H \tilde{U}_k.$$  

Therefore, we can obtain

$$\hat{\sigma}^2 = (N - \eta n - p)^{-1} \sum_{k=1}^{K} \tilde{S}_k,$$  

Finally, a weighted estimate of the model coefficient $\beta$ is calculated. According to the calculation equations (15) and (16) of the variance component, the covariance matrix of the k-th subdata model is estimated as $\hat{V}_k = Z_k^T \hat{\sigma}^2 I_{n \times m}$.

Using the least squares method to estimate that each subset is $\hat{\beta}_k = (X_k^T \tilde{V}_k X_k)^{-1} X_k^T \tilde{V}_k Y_k$, it can be seen that the data $X_k^T \tilde{V}_k^{-1} X_k$, $\hat{\beta}_k$ is retained, and the integrated result of each subset to get:

$$\hat{\beta} = (X^T \tilde{V}^{-1} X)^{-1} X^T \tilde{V}^{-1} Y = \left( \sum_{k=1}^{K} X_k^T \tilde{V}_k^{-1} X_k \right)^{-1} \sum_{k=1}^{K} X_k^T \tilde{V}_k^{-1} X_k \hat{\beta}_k.$$  

2.2.2. The Amount of Individual Data Is Small, and the Amount of Data in the Group Is Large. First, we have applied the divide-and-conquer algorithm to each individual. Let $D_i = \{(X_{ij}, Z_{ij}, Y_{ij})\}_{j=1}^{n_i}$, $i = 1, 2, \ldots, n$, divide the data $D_i$ into $K$ subsets $D_{i1}, D_{i2}, \ldots, D_{ik}$, and the number of samples in each subset is $m_i = n_i/K$. The data contained in each subset are $X_{ikj}, Y_{ikj}$, and $Z_{ikj}$ where $1 \leq j \leq m_i$, $1 \leq k \leq K$, $1 \leq i \leq n$, and the data contained in each subset cannot exceed the processing capacity of a single machine. First, we have calculated a preliminary estimate. Remember $X_{ik} = (X_{ik1}, X_{ik2}, \ldots, X_{ikm_i})^T$, $X_i = (X_{i1}, X_{i2}, \ldots, X_{ik})^T$, $Y_i, Z_i, \epsilon_i$ are similarly defined, and the data model corresponding to the k-th subset of the i-th individual is

$$\hat{\beta}_i = (X_i^T \tilde{V}_k^{-1} X_i)^{-1} X_i^T \tilde{V}_k^{-1} Y_i.$$

For each dataset of each individual, direct least squares estimation is $\hat{\beta}_{0,k} = (X_k^T X_k)^{-1} X_k^T Y_{ik}$, $1 \leq i \leq n$, $1 \leq k \leq K$. According to the divide-and-conquer algorithm, we have integrated the results of each individual and each subset, saved each group of $X_k^T X_k, \hat{\beta}_{0,k}$, and finally obtained a preliminary estimate:

$$\hat{\beta}_k = (X^T X)^{-1} X^T Y = \left( \sum_{i=1}^{n} \left( \sum_{k=1}^{K} X_{ik}^T X_{ik} \right) \right)^{-1} \left( \sum_{i=1}^{n} \left( \sum_{k=1}^{K} X_{ik}^T X_{ik} \hat{\beta}_{0,k} \right) \right).$$  

Second, use the divide-and-conquer algorithm to estimate the variance component. Remember the definition of $U_k = (U_{k1i}, U_{k2i}, \ldots, U_{km_i})^T$ is similar to (4), so we can get
3. The Application Architecture of the Improved Model in Massive Education Examination

Big Data

3.1. Data Attributes. The continuous score attribute of the test data is defined by combining the test data mining theme, characteristics of the tested students, gradient evaluation rules, and secondary calibration target [16]. Take the objective question type as an example: a multiple-choice question has four answers, one of which is the correct answer. However, from the perspective of the proposition, the other three options are not added casually without basis, and there are some wrong options that are “close” to the correct answer. The reason why students choose the wrong option is not that they do not know anything about this knowledge, but just because they are inadequate or careless and are deceived by the confusing options. In the traditional examination evaluation process, this feature is not accurate for the reflection.

On the contrary, there are individual options that do not have any connection with the knowledge point, and selecting this option can basically judge that the students have no knowledge of this point. Obviously, these two types of subjects are different, so examination analysis and evaluation should treat them differently and conduct targeted guidance and treatment. In addition, for different test objectives, the ability to examine students should not be solely based on test scores. Some assessment objectives hope to get students’ comprehensive ability assessment, including practical ability, specialty, interest, and so on. Therefore, in the proposition of the test questions, for different question types, possible different answers, and different test objectives, the positive and negative multidimensional attributes of the test data are defined, and the multidimensional attribute model of the test data is established [17]. The data model is shown in Figure 1, except the score attribute. The fact attribute is set at the same time to reflect the factual corresponding to each individual subset, the data to be calculated and saved are \( \{ \bar{U}_i^T \bar{U}_i, \bar{U}_i^T H \bar{U}_i, Z^T \bar{U}_i \} \). Finally, a weighted estimate of the model coefficient \( \beta \) is calculated.

According to the calculation equations (21) ~ (22) of the variance component, the covariance matrix of the \( k \)-th subdata model of the \( i \)-th individual is estimated as \( \bar{V}_{ik} = Z_{ik} \sum Z_{ik}^T + \sigma^2 I_m \). The weighting is applied to each subset of each individual, and the least squares estimates are

\[
\beta_{ik} = \left( X_{ik}^T \bar{V}_{ik}^{-1} X_{ik} \right)^{-1} X_{ik}^T \bar{V}_{ik}^{-1} Y_{ik}.
\]

It can be seen that the data is retained for each subset of each individual:

\[
\beta = \left( X^T \bar{V}^{-1} X \right)^{-1} X^T \bar{V}^{-1} Y = \left[ \sum_{i=1}^{n} \left( \sum_{k=1}^{K} X_{ik}^T \bar{V}_{ik}^{-1} X_{ik} \right) \right]^{-1} \sum_{i=1}^{n} \left( \sum_{k=1}^{K} X_{ik}^T \bar{V}_{ik}^{-1} X_{ik} \beta_{ik} \right).
\]
subject groups, such as “careless mistakes,” “nothing at all,” and “random selection,” and provide targeted guidance and treatment to them. To achieve this goal, this paper first performs hierarchical filtering based on spectral clustering on the test data [19] and then performs secondary correction on each data cluster that is hierarchically filtered.

3.2. Hierarchical Filtering Based on Spectral Clustering. Clustering is the process of distinguishing and classifying things according to certain rules and requirements. In this process, there is no prior knowledge about classification, and only the similarity between things is used as the criterion for classification. Compared with traditional clustering algorithms, spectral clustering has the advantage of being able to cluster on sample spaces of arbitrary shapes and converging to the global optimal solution. The spectral clustering algorithm can be briefly described as follows: given a dataset \( X = \{x_1, x_2, \ldots, x_n\}, x_i \in \mathbb{R}^p \), according to the dataset \( X \), a weighted graph \( G = (V, E) \) is established, where \( V = \{v_i, i = 1, 2, \ldots, n\} \) is the set of vertices, and \( E = \{e_{ij}\} \) is the edge connecting the vertices \( (v_i, v_j) \). Each node \( v_i \) in the graph \( G \) is related to \( x_i \) in the dataset \( X \). A similarity criterion is used to construct the similarity matrix \( W \) between the vertices of the graph \( G, W \in \mathbb{R}^{n \times n} \).

For two objects \( x_i \) and \( x_j \), the similarity \( s \) is defined \( s(x_i, x_j) = \exp(-\|x_i - x_j\|^2/2\sigma^2) \). The algorithm inputs the similarity matrix \( W \) of the graph \( G \) and the number of clusters \( k \) and outputs the clustering results of \( x_1, x_2, \ldots, x_n \). The hierarchical filtering algorithm based on spectral clustering adds the distance matrix of the answer data object based on the characteristics of the test answer data object based on the spectral clustering and obtains a spectral clustering algorithm suitable for answering data analysis [20]. Assuming that the questionnaire data object \( x \) of the study has \( p \) test points, expressed as \( (x_1, x_2, \ldots, x_p) \), the distance \( d(u, v) \) between the two questionnaire data objects \( u \) and \( v \) can be defined as follows: \( d(u, v) = \sum_{i=1}^{n} |u_i - u_j| \). Among them, \( |u_i - u_j| = 1, u_i \neq u_j \) or \( |u_i - u_j| = 0, u_i = u_j \); \( i = 1, 2, \ldots, p \). According to the distance between the questionnaire data objects, the similarity \( s(u, v) \) between the questionnaire data objects \( u \) and \( v \) of two subjects can be defined as follows: \( S(u, v) = \exp(-d(u, v)/2\sigma^2) \), normally, \( \sigma = 1 \). The hierarchical filtering and analysis algorithm of spectral clustering algorithm for answering data includes the following steps.

1. Read into the exam database record. The set of answer data objects is represented as \( X = \{x_1, x_2, \ldots, x_p\}, x_i \in \mathbb{R}^p \). Among them, \( x_i \) is the answer information of the \( i \)-th subject, and \( P \) is the number of questions. We have calculated the distance matrix \( A \) of the subject’s answer sheet data object.
(2) Calculate the similarity matrix \( W \) of the subject’s answer data object. The order of the matrix \( W \) is the same as the order of the matrix \( A \).

(3) Calculate the extended adjacency matrix \( L \). According to the similarity matrix \( W \), a diagonal matrix is established, denoted as \( D = (d_{ij})_{n \times n} \) where \( d_{ij} = \sum_{k=1}^{n} w_{ij} \), and the matrix is defined as \( L = D^{-1/2} W D^{-1/2} \).

(4) Calculate the eigenvalues and eigenvectors of the matrix \( L \); select the first \( k \) eigenvectors \( u_1, u_2, \ldots, u_k \); construct a matrix \( U \in \mathbb{R}^{n \times k} \) with the eigenvectors \( u_1, u_2, \ldots, u_k \) as the columns, and compare the matrix each line of \( U \) is unitized.

(5) Let \( y_i \in \mathbb{R}^k \) be the vector corresponding to the \( i \)-th row of the matrix \( U \). Using the \( k \)-means algorithm, cluster the vector \( y_i \) into \( C_1, C_2, \ldots, C_k \).

(6) Establish a mapping \( x_i \in \mathbb{R}^k \rightarrow y_i \in \mathbb{R}^k \), \( i = 1, 2, \ldots, n \). According to the correspondence between \( x_i \) and \( y_i \), the clustering results of the row vector \( y_i \) of the matrix \( U \) in step (5) are used to determine the results of the points \( x_1, x_2, \ldots, x_n \) in the clustering.

### 3.3. Quadratic Correction Algorithm Based on Improved Linear Effect Model

The hierarchical filtering based on spectral clustering is the first classification process for the subjects. In the classification process, only the test score information of the subjects is used. However, in some specific selection processes, only scores used to measure the ability of the examinee appeared to be insufficiently scientific and cannot truly reflect the overall ability of the examinee and the special group caused by special circumstances during the examination process. Therefore, on the basis of the first hierarchical filtering of test scores, the clustering results are corrected twice [21]. Take a single-choice objective multiple-choice question as an example. Suppose a test question has four answers A, B, C, and D. Among them, A is the correct answer with a score of \( x \); B is a strongly related answer with a score of \( k_1 \); C is a weakly related answer with a score of \( k_2 \);\( \epsilon \) is an irrelevant answer with a score of \( k_3 \>; \epsilon \); \( k_3 \) \( \epsilon \) is the value interval is \([0, 1]\). Suppose the number of question exams is \( N \), and the number of questions answered by a certain student is \( M \), where \( M \leq N \). The number of wrong questions with weights \( k_1, k_2, k_3 \) in the wrong answers are \( M_1, M_2, M_3 \), respectively, and \( M_1 + M_2 + M_3 = M \). Then, the final correction score \( X \) of the participant is recorded as follows:

\[
X = \left( \frac{M_1 k_1}{M} + \frac{M_2 k_2}{M} + \frac{M_3 k_3}{M} + (N - M) \right) \times x. \quad (26)
\]

Similarly, for multiple-choice questions, suppose there are six options A, B, C, D, E, F for a question, and the number of correct answers for the question with an incorrect answer is \( N^* \), \( 1 \leq N^* \leq 6 \). The number of correct answer options selected by the participant is \( M^* \), \( 0 \leq M^* \leq N^* \). The number of wrong answer options selected by the participant is \( K^* \), \( 0 \leq K^* \leq 6 - N^* \). Then, the participant gives the right answer. The final score \( Y \) of this multiple choice question is defined as follows:

\[
Y = \left( \frac{M^*}{N^*} - \frac{K^* + 6 - N^*}{6} \right) \times x^*. \quad (27)
\]

Based on the previously mentioned description, a new correction and measurement can be made on the actual ability of the participant based on the classification of the basic test scores. The second correction algorithm is described as follows.

(1) Enter the total number of single-choice and multiple-choice questions as \( N, N^* \), and the starting and ending question numbers of the single-choice and multiple-choice questions, each with a score of \( x \) and \( x^* \) for each single-choice and multiple-choice questions, and enter the multiple-choice questions answer score standard table, with each record in the table being a triple \((M^*, N^*, K^*)\).

(2) Scan the answer sheet.

(3) For single-choice questions, calculate the single-choice question correction score \( X \) according to (26).

(4) For multiple-choice questions, calculate the multiple-choice question correction score \( Y \) according to (27).

(5) Calculate the total correction score \( Y \). After hierarchical filtering and secondary correction, each subject has two scores: real score and corrected score. Each subject is in its own cluster. This will analyze the test examinee and the ability of subject and teacher to provide a more accurate basis for judgment.

### 4. Simulation Experiment and Result Analysis

#### 4.1. Example Analysis of Improved Linear Mixed-Effects Model

In order to compare the estimation algorithm in this paper with the maximum likelihood method [12], verify the advantages of the algorithm in this paper in terms of calculation time, and test the feasibility of the algorithm at the same time, the Matlab software is used for numerical simulation and experiment. The experiment was performed on a hardware platform of Intel i7 9700K CPU with 16 GB memories machine, running on the Windows 10 operating system. Suppose the linear mixed-effects model of longitudinal data is as follows:

\[
Y_{ij} = X_{ij}^T \beta + Z_{ij}^T \gamma + \epsilon_{ij}, \quad i = 1, 2, \ldots, n; j = 1, 2, \ldots, n. \quad (28)
\]

In the equation, we consider two sample scenarios under massive data.

**Scenario (1):** the number of individuals is large, but the amount of data in the group is small. Two samples are considered as follows:

(1) \( n = 10000, n_i = 10 \), and the total sample size is 100, 000.
Scenario (2): there are a limited number of individuals, but the amount of data in the group is large. The sample is \( n = 10 \) and \( n_j = 100000 \). In addition, the parameter coefficients are \( \beta = (1, 2, -3, 1, -2)^T \); \( X_{ij} = (X_{ij1}, X_{ij2}, X_{ij3}, X_{ij4}, X_{ij5})^T \), where \( X_{ij1} \sim N(0, 1), X_{ij2} \sim N(0, 2), X_{ij3} \sim N(0, 1), X_{ij4} \sim N(0, 2), X_{ij5} \sim N(0, 1), \) and \( X_{ij1}, X_{ij2}, X_{ij3}, X_{ij4}, X_{ij5} \) are independent of each other. \( Z_{ij} = (Z_{ij1}, Z_{ij2})^T \), \( Z_{ij1} \) and \( Z_{ij2} \) are all independently and identically distributed in \( N(0, 2) \); \( b_j = (b_{j1}, b_{j2})^T \), \( b_{j1} \) and \( b_{j2} \) are all independently and identically distributed in \( N(0, 1) \); \( \varepsilon_i \sim N(0, \sigma^2) \), where \( \sigma^2 = 2 \), and \( b_j \) and \( \varepsilon_i \) are independent of each other. The number of iterative simulations is 50 times.

4.1.1. Simulation Results and Analysis of Scenario (1). In the two sample cases of scenario (1), the maximum likelihood method [12] and the algorithm in this paper are used to estimate the model coefficients and variance components, respectively, and the calculation time and estimation accuracy used by the two methods are compared.

(1) Estimate Accuracy and Calculation Efficiency. The mean square error (MSE) is used to measure the deviation between the simulated estimated value and the true value, and the estimation accuracy is analyzed. A smaller MSE value causes the higher estimation accuracy. The calculation equation of the mean square error is as follows:

\[
Y_{\text{MSE}} = \frac{1}{m} \sum_{k=1}^{m} (x_{\text{obs},k} - x_{\text{real}})^2.
\]  

In (29), \( m \) is the number of simulations, \( x_{\text{real}} \) is the true value of the parameter, and \( x_{\text{obs},k} \) is the estimated value of the parameter obtained from the \( k \)-th simulation. The calculation time and the mean square error values under the two sample sizes in scenario (1) are obtained by simulation as shown in Table 1.

It can be seen from Table 1 that the MSE values of the model parameters estimated by the maximum likelihood method and the algorithm in this paper are both small, indicating that the parameter estimation effect is better and the estimation accuracy is higher. In addition, compared with the maximum likelihood method, the proposed algorithm can increase the operating efficiency by four times while almost not reducing the estimation accuracy, indicating that the algorithm can greatly reduce the calculation time and increase the calculation speed.

To study the relationship between the calculation time and the number of subdataset blocks and illustrate the necessity of using the divide-and-conquer algorithm for block calculation, the data under the two sample sizes are divided into different dataset block numbers, and the estimated parameters are recorded. The relationship between the calculation time of sample (1) and sample (2) and the number of dataset blocks is shown in Figure 2. The data generated in the figure is the time used by the sample summation method.

It can be seen from Figure 2 that when \( K = 1 \), the weighted least square equation (11) is directly used to estimate the model coefficients. At this time, it will exceed the memory and cannot be calculated. Under samples (1) and (2), that is, when \( K \) takes 10000 and 100000, respectively, use the sample sum equation (12) to estimate the model coefficients. \( K \) takes other values in the table, according to the algorithm calculated in this article. When \( K \) slowly increases, the time used gradually decreases. Taking an appropriate value of \( K \), such as 2000 and 20000 in the samples (1) and (2), respectively, the calculation time is minimized compared with the sample sum method. The running speed can be increased by more than 20%. This is because when \( K \) increases, the number of samples allocated to each dataset block is reduced at the same time, so the amount of calculation time is also reduced. At the same time, in the case of limited and fixed stand-alone processing capacity, a certain level of data is suitable. The best calculation speed can be achieved under the number of single machines. In practice, the amount of data is larger, usually reaching millions or even tens of millions, and the effect of the algorithm in this paper will be better at this time. In summary, the algorithm in this paper can reduce memory overhead and calculation time and increase calculation speed.

4.1.2. Simulation Results and Analysis of Scenario (2). In scenario (2), the relationship between the calculation time used by the algorithm in this paper and the number of dataset blocks is shown in Figure 3. The data generated in the figure is the time used by the sample summation method. When \( K = 1 \), use the sample summation equation (12) to calculate the coefficient. When \( K \) takes other values calculated by this algorithm, it can be seen that the change trend of time with the number of data blocks is similar to scenario (1), and when \( K = 200, t = 1.38s \), the calculation time reaches the minimum. At the same time, the maximum likelihood method is used to calculate the model parameters of the scenario (2). The time used is 2269.71 s, which proves the efficiency of the algorithm in this paper again.

Select the data of the year-end gross domestic product (GDP), industrial output, and total retail sales of consumer goods. For the selected data, we fixed asset investment and urban residents’ savings at the year-end balance of 285 prefecture-level across the country from 2006 to 2010. Take the regional gross product \( (Y) \) as the response variable and the gross industrial output value \( (X_1) \) and the total retail sales of consumer goods \( (X_2) \) as covariates of fixed effects. Considering the fixed asset investment \( (Z_1) \) and the year-end balance of urban and rural residents’ savings \( (Z_2) \), such factors will also affect the regional GDP. Therefore, these two factors are taken as part of the random effect, and a linear mixed effect model is established:

\[
Y_{ij} = X_{ij}^T \beta + Z_{ij}^T b_i + \varepsilon_{ij}, \quad i = 1, 2, \ldots, 285; j = 1, 2, \ldots, 5.
\]
Bringing the data into (30) shows that the data structure at this time conforms to the scenario (1), so the estimation algorithm in the scenario (1) is used to estimate the parameters. The specific results are shown in Table 2.

It can be seen from Table 2 that the coefficients of the gross industrial output value and the total retail sales of consumer goods are both positive, indicating that these two factors positively affect the regional gross product and the model built is effective. Therefore, to increase the regional GDP, we must strive to increase the level of industrial production, increase the total retail sales of consumer goods, increase people's income, improve the consumption environment, stabilize the price level, and solve medical, health, and employment issues.

4.2. Empirical Analysis of Massive Education Examination Big Data

4.2.1. Architecture of Mass Education Examination Big Data System. In order to verify the efficiency and accuracy of the hierarchical filtering and secondary correction strategy for test big data analysis, a cloud computing-based education test big data analysis and evaluation service platform was developed. The overall construction of the platform includes a centre, two-level applications, and three layers of subsystems. The overall architecture of this system is shown in Figure 4.
The platform mainly includes three layers: infrastructure layer, public support base layer, and application service layer. The functional modules are shown in Figure 5. The infrastructure layer mainly includes cloud infrastructure and system software. Its main function is to complete the distributed cloud storage of educational examination data. The hadoop distributed file system (HDFS) system based on cloud technology is mainly used for the organization and management of the file system; the public basic support layer is the evaluation analysis system, and the main components of the data warehouse include the data warehouse part [22] and secondary data processing tools. The data warehouse part mainly includes the realization of basic functions of data collection, data exchange, data mining, and evaluation engine. In units of schools and subjects, the public support application part includes the test question bank, data cleaning, test paper generation, historical information analysis, subject analysis engine, and so on. The application service layer is the user-oriented service module of the platform, which mainly includes basic examination data query, online marking comprehensive service system, evaluation analysis and management subsystem, and test question bank application system. The educational examination data analysis and evaluation cloud platform built by this project directly interacts with users through these services, and such platform is therefore the most user-oriented part.

4.2.2. Analysis of Massive Education Examination Big Data Results. The system can perform statistical analysis for a certain course in a certain class or perform statistical analysis on a certain course in a certain area. In addition, it can also perform statistical analysis on all students within a certain range for a certain topic [23]. Taking a second-grade political and English test in a certain area as an example, the original scores and the second-corrected scores of the subjects are calculated, respectively. The statistical results are shown in Figures 6 and 7.

It can be seen from Figures 6 and 7 that the subjects are basically divided into three categories: the first category has a score of 35 to 65, the second category has a score of 65 to 90, and the third category has a score of 90 or more. At the same time, it can be seen that, after the second correction, the scores of all students have improved compared with the original scores, but the increase in the first and third categories is significantly smaller than the second category. The reason for the small increase in the first category is that this group usually has a poor academic performance. Therefore, during the examination process, the wrong questions are basically of the type that is not understood, so the correction range is small. The reason for the small increase in the third category is that the performance of this group is already very high, and the room for correction has been severely compressed. The second type of improvement is relatively large because this group of students is in the middle class, and there is a large room for improvement. Teachers should pay special attention and guidance to the problems that lead to mistakes.

4.2.3. Massive Education Examination Big Data Analysis Report. The cloud computing-based education examination big data analysis and evaluation service system currently serves 13 provinces and more than 100 cities across the country. The examination data acquisition, storage, and evaluation analysis business accounts for 40% of the national academic examination and analysis evaluation business market. The system supports multiple types of report formats and provides relevant analysis reports for different groups of people through various parameters. There are three reports mainly included as follows:

1) Student Report. The student user report mainly shows the performance of the students in the exam. It analyzes the students from the overall situation of the students and the individual subjects, points out their strengths and weaknesses, and gives targeted improvement opinions [24]. An example of a student report is shown in Figure 8(a).

2) Institutional Report. The institutions here cover the school and the school’s competent department. The institutional user report mainly displays the overall examination conditions of the institution’s students in the examination. The institution is statistically analyzed from the overall situation of the students and the individual subjects of the students. There are various indicators that are used to study the
Institutional reports are divided into test overview report, institutional subject report, subquestion answering report, institutional knowledge comparison, and student knowledge comparison. An example of an agency report is shown in Figure 8(b).

**Teacher Report.** The teacher’s report is mainly based on the establishment of a one-to-many multidimensional mapping relationship of knowledge points, knowledge sections, cognitive levels, and ability fields [13]. The score of each dimension item is the main statistical indicator. As a sample group of

![Figure 5: The functional architecture of the massive education big data cloud platform.](image)

![Figure 6: Comparison of political classes before and after the linear mixed model proofreading.](image)
Figure 7: Comparison of English class before and after the linear mixed model proofreading.

Figure 8: Continued.
different candidates, grades or classes point out the weak links in teaching and play a role in guidance, inspiration, and feedback for teachers’ follow-up teaching. An example of the analysis report is shown in Figure 8(c).

5. Conclusions

With the rapid development of the big data era nowadays, many applications derived from big data have also been used in educational examinations. This paper proposes a three-step estimation algorithm, so that complex statistical models can also be applied to massive data. Numerical simulation shows that the algorithm can greatly reduce the calculation time, improve the calculation efficiency, and solve the problem of insufficient computer memory under massive data. On applying the improved linear mixed-effects model to the massive educational examination big data, this paper carries out the corresponding platform architecture and experiments: (1) define the test data attributes; (2) based on the definition of data attributes and basic evaluation rules, the group cluster data is corrected twice based on the gradient evaluation rules to avoid single score evaluation; (3) the cloud-based services technology research and development is a cross-regional, multiuser educational examination big data analysis and evaluation cloud platform. Future works will concentrate on how to reduce the relationships between different massive big data on educational examination and pay more attention to improving the analyzed performance and efficiency.

Data Availability

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Conflicts of Interest

The authors declared that they have no conflicts of interest regarding this work.

References

[1] D. Fu, X. Xia, M. Duan et al., “Mapping nighttime PM2.5 from VIIRS DNB using a linear mixed-effect model,” Atmospheric Environment, vol. 178, pp. 214–222, 2018.
[2] E. Moreira-Marconi, M. C. Moura-Fernandes, P. Lopes-Souza et al., “Evaluation of the temperature of posterior lower limbs skin during the whole body vibration measured by infrared thermography: cross-sectional study analysis using linear mixed effect model,” PLoS One, vol. 14, no. 3, Article ID e0212512, 2019.
[3] S. Sotoudeheian and M. Arhami, “Using linear mixed effect model to estimate ground-level PM2.5: case study for Tehran,” Iranian Journal of Health and Environment, vol. 10, no. 2, pp. 213–224, 2017.
[4] M. Huda, A. Maselelo, P. Atmotiyoso et al., “Big data emerging technology: insights into innovative environment for online learning resources,” International Journal of Emerging Technologies in Learning (iJET), vol. 13, no. 1, pp. 23–36, 2018.
[5] M. Tsubak and Y. Maeda, “A research of analysing the effectiveness of speaking-pen on English learning in
consideration of individual differences using a linear mixed-effect model,” *Education and Information Technologies*, vol. 25, no. 2, pp. 731–753, 2020.

[6] S. Roy and S. N. Singh, “Emerging trends in applications of big data in educational data mining and learning analytics,” in *Proceedings of 2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence*, pp. 193–198, IEEE, Noida, India, January 2017.

[7] R. Swathi, N. P. Kumar, L. KiranKranth et al., “Systematic approach on big data analytics in education systems,” in *Proceedings of 2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 420–423, IEEE, Madurai, India, June 2017.

[8] O. Yuki, K. Yousuke, S. Yul-Wan et al., “A linear mixed-effect model analysis of the effect of schizotypal personality traits on confidence in reality monitoring,” *Tohoku Psychologica Folia*, vol. 77, pp. 83–90, 2018.

[9] S. Chahine, K. Kulasegaram, S. Wright et al., “A call to investigate the relationship between education and health outcomes using big data,” *Academic Medicine*, vol. 93, no. 6, pp. 829–832, 2018.

[10] A. K. Nikoloulopoulos, “A vine copula mixed effect model for trivariate meta-analysis of diagnostic test accuracy studies accounting for disease prevalence,” *Statistical Methods in Medical Research*, vol. 26, no. 5, pp. 2270–2286, 2017.

[11] J. Deng, “Research on effectiveness of colleges ideological and political education in the age of big data,” *Journal of Physics: Conference Series*, IOP Publishing, vol. 1744, no. 3, Article ID 032058, 2021.

[12] T. Chen and R. Wang, “Inference for variance components in linear mixed-effect models with flexible random effect and error distributions,” *Statistical Methods in Medical Research*, vol. 29, no. 12, pp. 3586–3604, 2020.

[13] K. Maruo, Y. Yamaguchi, H. Noma, and M. Goshio, “Interpretable inference on the mixed effect model with the Box-Cox transformation,” *Statistics in Medicine*, vol. 36, no. 15, pp. 2420–2434, 2017.

[14] I. Gasmí, B. Aljoumani, J. A. Sánchez-Espigares et al., “A linear mixed effect (LME) model for soil nutrients and soil salinity changes based on two localized irrigation techniques (drip irrigation and buried diffuser),” *Arabian Journal of Geosciences*, vol. 12, no. 11, pp. 1–12, 2019.

[15] S. Guo and L. Lin, “Research on voluntary intelligent reporting system of college entrance examination based on big data technology,” in *Proceedings of International Conference on Advanced Hybrid Information Processing*, pp. 98–111, Springer, Cham, Switzerland, September 2020.

[16] C. Ngufor, H. Van Houten, B. S. Caffo, N. D. Shah, and R. G. McCoy, “Mixed Effect Machine Learning: a framework for predicting longitudinal change in hemoglobin A1c,” *Journal of Biomedical Informatics*, vol. 89, pp. 56–67, 2019.

[17] Z. G. Dessie, T. Zewotir, H. Mwambi, and D North, “Modelling of viral load dynamics and CD4 cell count progression in an antiretroviral naive cohort: using a joint linear mixed and multistate Markov model,” *BMC Infectious Diseases*, vol. 20, pp. 246–314, 2020.

[18] C. Zhenyu, “The application of big data in higher vocational education based on Holland vocational interest theory,” in *Proceedings of 2017 International Conference on Industrial Informatics—Computing Technology, Intelligent Technology, Industrial Information Integration (ICIIIC)*, pp. 37–40, IEEE, Hubei, China, December 2017.

[19] D. Narinc and N. Oksuz Narinc, “Comparison of resampling and bayesian approaches in variance component estimation of a hierarchical univariate mixed effect model,” *Selcuk Journal of Agricultural and Food Sciences*, vol. 34, no. 1, pp. 91–98, 2020.

[20] N. M. Nielsen, W. A. C. Smink, and J.-P. Fox, “Small and negative correlations among clustered observations: limitations of the linear mixed effects model,” *Behaviormetrika*, vol. 48, no. 1, pp. 51–77, 2021.

[21] P. Liu, P. Chen, Y. Yuan et al., “A teaching assistant system for big data analysis,” *Journal of Physics: Conference Series*, IOP Publishing, vol. 1678, no. 1, Article ID 012090, 2020.

[22] N. W. Sobe, “Problematizing comparison in a post-exploration age: big data, educational knowledge, and the art of criss-crossing,” *Comparative Education Review*, vol. 62, no. 3, pp. 325–343, 2018.

[23] I. Bakbergenuy and E. Kulinskaya, “Meta-analysis of binary outcomes via generalized linear mixed models: a simulation study,” *BMC Medical Research Methodology*, vol. 18, no. 1, 2018.

[24] D. Appelbaum, A. Kogan, and M. A. Vasarhelyi, “Big data and analytics in the modern audit engagement: research needs,” *AUDITING: A Journal of Practice & Theory*, vol. 36, no. 4, pp. 1–27, 2017.