Development Model of College English Education Service Talents Based on Nonlinear Random Matrix

Xiaoyan Gao

The English Department, Taiyuan University, Taiyuan 030032, China

Correspondence should be addressed to Xiaoyan Gao; gaoxiaoyan@tyu.edu.cn

Received 4 July 2022; Revised 17 August 2022; Accepted 27 August 2022; Published 14 September 2022

Academic Editor: Ning Cao

Copyright © 2022 Xiaoyan Gao. 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.

The development of artificial intelligence makes people’s life and work easier and more effective, and computer-based online exams and marking not only improve students’ learning efficiency but also reduce the pressure of teachers’ marking work. For objective questions, marking has gone from manual marking to cursor reader marking to computerized character matching, and the correct rate of marking has soared to 100%; for subjective questions, foreign systems such as PEG and E-rater have been used, and domestic systems such as those using English large corpus similarity matching and those based on natural language understanding using intelligent algorithms have been used for marking. Most of these systems are based on some shallow linguistic features such as rules and LSAs for marking, and there is no deep perception of English language sense. Although the current intelligent marking systems have made a lot of achievements, they do not fundamentally solve the problem of the rationality of intelligent marking of subjective questions. In this article, we propose a regularized discriminant analysis algorithm with good estimation of the mean, and a dimensionality reduction algorithm for high-dimensional missing data by using the relevant research results of random matrix theory to address the problems of traditional machine learning methods in high-dimensional data analysis. Although the linear discriminant analysis algorithm performs well in solving many practical problems, it works poorly in dealing with high-dimensional data. The specific analysis is as follows: in terms of age characteristics, the mobile population under the age of 35 has a significant preference for urban consumer comfort, and it increases with age, peaking at the stage of 30–35 years old and then decreasing rapidly. For this reason, a regularized discriminant analysis algorithm based on random matrix theory is proposed. First, a good estimate of the high-dimensional covariance matrix is made by the nonlinear shrinkage method or the eigenvalue interception method, respectively; then, the estimated high-dimensional covariance matrix is used to calculate the discriminant function values and perform the classification. The classification experiments conducted on simulated and real datasets show that the proposed algorithm is not only more widely applicable but also has a high correct classification rate.

1. Introduction

With the advent of the era of big data, massive amounts of data have been collected and stored in all walks of life in society, and the characteristics of these data include large number and dimensionality, high value, and fast growth rate, and these characteristics also pose huge challenges for data analysis [1]. The main reason for the explosive growth of data in recent years is the increasingly low cost of data production and storage. For example, in the field of genetic data analysis, the price of whole genome sequencing has been decreasing dramatically [2]. This is also true in other areas such as social media analytics, biomedical imaging, and retail sales. The data collected in these fields often have dimensions close to or even exceeding the sample size of the data and are called high-dimensional data. How to obtain useful information from these high-dimensional data and make it produce great value in production practice has become an important problem for modern society and poses a serious challenge to traditional machine learning methods. The impact of increasing the number of data dimensions on data analysis is multifaceted. For example, in nonparametric estimation, the high dimensionality of data affects the convergence speed of algorithm estimation; in model
2 Mathematical Problems in Engineering

selection, too many data variables cause degradation of model performance; in regression analysis, the sparsity of high-dimensional data is also one of the difficulties in data prediction [3].

Also in multivariate statistical analysis, it is usually assumed that the dimensionality of the data is fixed and finite, and when the dimensionality of the data approaches or even exceeds the sample size of the data, then the classical theory of multivariate statistical analysis also reflects its own limitations, especially for the estimation of the mean and covariance matrix of high-dimensional data. Mean estimation of data is a fundamental problem in multivariate statistical analysis, and many data analysis methods like diagonal discriminant analysis [4] and Markowitz mean-covariance analysis [5] need to estimate the mean of data. When the data dimension is large, it is difficult for the data obeying a specific distribution to be near the overall mean in the high-dimensional space, and when keeping the data sample constant, the data will gradually move away from the overall mean as the data dimension increases, and it is difficult for the traditional mean estimation methods to accurately estimate the mean of high-dimensional data, which largely causes some data mining algorithms cannot be applied to the analysis of high-dimensional data. The calculation of the covariance matrix or its inverse matrix (precision matrix) of the data is also a key step in many data analysis and mining algorithms. For example, in principal component analysis, the computation of the covariance matrix and the determination of the number of principal components are key aspects of dimensional approximate reduction of the original data [6]; in Bayesian multivariate statistical inference theory, the computation of conditional probabilities under multivariate normal distribution approximation requires consistent estimation of the precision matrix [7]; similarly, in large-scale reinforcement learning methods based on Gaussian process classifiers, the estimation of the covariance matrix and the precision matrix [8]. The dimensionality of high-dimensional data is almost equal to, or even larger than, the sample size, so the classical multivariate statistical analysis methods will no longer be applicable to the estimation of the covariance matrix [9]. At this point, the sample covariance matrix is pathological or singular, and the sample covariance matrix is no longer a good estimate of the overall covariance matrix [10].

Some useful information in the data is often not available due to missing, which also further increases the estimation bias of data statistics like mean and covariance, making the analysis of high-dimensional data increasingly difficult. Most of the existing data analysis methods are also mainly used for the analysis of complete data sets, so it is necessary to design an algorithm to solve the problem of analyzing high-dimensional missing data. The data may cause some missing data during the process of data collection and storage. When high-dimensional data contains missing values, most data analysis methods are difficult or unsatisfactory for analyzing high-dimensional missing data. To this end, a principal component analysis algorithm that can be used for dimensionality reduction of high-dimensional missing data is proposed. First, based on the relevant theory of random matrix theory, the covariance matrix estimation of high-dimensional missing data is obtained by using the Lasso estimation of matrix; then, the feature decomposition is performed, the main feature vectors are selected to form the low-dimensional projection matrix, and the projection matrix is used to project the high-dimensional data into the low-dimensional space; finally, the linear discriminant analysis algorithm is combined to classify the high-dimensional missing data. The classification experiments on simulated and real datasets show that the proposed algorithm can accomplish the dimensionality reduction of high-dimensional missing data, and also improve the correct classification rate of linear discriminant analysis algorithm on high-dimensional missing data.

2. Related Work

IEA (Intelligent Essay Assessor) is a system that scores words based on the semantic statistics of the words. In contrast to PEG, IEA is a content-focused scoring system. Its development team claims that IEA can measure both semantic and textual content. This is due to the fact that IEA uses latent semantic analysis (LSA), which is a model for information retrieval that can be used to filter words in the text by the equivalent of mining the text for keywords [11]. In this way, LSA is used on the training and test sets to represent keywords as spatial vectors, and then the semantic similarity between them is compared to measure the semantic “readability” of the text. E-rater is a hybrid scoring system developed by the US. Educational Testing Service, which is used in many large-scale exams, including the TOEFL exam, where essays are graded [12]. TOEFL essay marking generally requires two scorers, with E-rater acting as one scorer and only introducing another scorer for manual marking if it differs significantly from the manually marked score [13]. The calculation of the covariance matrix or its inverse matrix (precision matrix) of the data is also a key step in many data analysis and mining algorithms. For example, in principal component analysis, the computation of the covariance matrix and the determination of the number of principal components are key aspects of dimensional approximate reduction of the original data.

This shows that the E-rater scoring system is relatively well done, because it uses a variety of techniques that incorporate not only the advantages of the PEG system for evaluating language quality but also the IEA system for extracting features for text content quality. E-rater also uses the same linear regression analysis used by PEG to score the above features [14]. Explore the influence of urban comfort on talent mobility under the role of market allocation and government promotion. Traditional Western theory emphasizes the influence of gainful factors such as economic opportunities, migration costs, and migration policies on labor mobility [15]. However, the continuous development of transportation and communication technologies has shrunk the time distance and perceived distance between cities, and local characteristics such as public services, environment, and cultural atmosphere have been gradually incorporated into the spatial.
division of labor mobility, and nonrevenue factors have become important factors in influencing the spatial decision of talent mobility [16].

It was found that the energy sequences of quantum systems can be approximated by the eigenvalues of Hermitian matrices, but the dimensionality of the data matrix is too large to effectively distinguish each energy sequence, and the limiting spectral distribution of all eigenvalues is studied instead [19]. The classical multivariate statistical analysis methods for data analysis are premised on the assumption that the dimensionality of the data is constant and finite, while the sample size gradually tends to infinity, which is no longer applicable to the analysis of large-dimensional matrices. To solve the above problems, the problem of large-dimensional random matrix eigenvalue distribution has been widely discussed and researched, and the famous Semicircle Law and Marchenko-Pastur Law were discovered, which played a fundamental role in the study of random matrix theory. With the development of information and computing science, more and more fields have generated and collected a large amount of high-dimensional data, further promoting the study of random matrix theory, which has also become an important part of modern statistical theory [20]. The research related to random matrix theory is also applicable to the analysis of high-dimensional data, including covariance matrix, regression analysis, and hypothesis testing. This article also further completes the analysis of high-dimensional data based on the research related to random matrix theory, and some knowledge about the random matrix theory covered in this article will be introduced next [21].

**3. A Model of English Education Talent Development in Universities Based on Random Matrix Theory**

**3.1. Spatial Pattern of Talent Clustering.** Spatial autocorrelation analysis is often used to detect the potential spatial interdependence of geographic data, which decreases with increasing geographic distance. The classical Moran’s I index is used to measure the spatial dependence of talent concentration levels in each city, and the Moran’s I statistical formula is as follows:

\[
I = \frac{\left( n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} x_i x_j \right)}{n \sum_{i=1}^{n} x_i^2},
\]

(1)

where \( x_i \) is the deviation of the attribute of element \( i \) from its mean \( (x_i - x) \), \( w_{i,j} \) are the spatial weights between elements \( i \) and \( j \), and \( S_0 \) is the aggregation of all spatial weights. To solve the above problems, the problem of large-dimensional random matrix eigenvalue distribution has been widely discussed and researched, and the famous Semicircle Law and Marchenko-Pastur Law were discovered, which played a fundamental role in the study of random matrix theory.

The High/Low Clustering (Getis-Ord General G) tool measures the density of high or low values for a specified study area. First, the null hypothesis specifies that there is no spatial clustering of elemental values, and the rejection of the null hypothesis is tested by calculating a statistically significant \( p \)-value. If the null hypothesis is rejected, it indicates that there is spatial clustering of high or low elemental values within the study area, and the discriminating tool is the \( z \)-value returned by the Getis-Ord General G statistic. The General G statistic formula is as follows:

\[
G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} x_i x_j}{n \sum_{i=1}^{n} x_i^2},
\]

(2)

where \( x_i \) and \( x_j \) are the attribute values of elements \( i \) and \( j \), and \( w_{i,j} \) are the spatial weights between elements \( i \) and \( j \).

Spatial Statistics (tool) of Arc GIS10.2 was used to calculate the spatial global Moran’s I statistic of city talent concentration level. As shown in Figure 1, Moran’s I statistic of city talent concentration level is significantly positive, which indicates that high-end human capital has a spatial concentration characteristic at the municipal scale. Based on this, the spatial local high/lower clustering (Getis-Ord General G) (tool) and hotspot analysis (Getis-Ord Gi) (tool) of Arc GIS10.2 are used to explore the spatial pattern of talent agglomeration level.

Starting from the theoretical framework of English education in colleges and universities, the level of talent clustering is influenced by urban comfort and plays a driving role in the economic development of cities. Based on the results of urban comfort clustering, the creative environment comfort factor, public service comfort factor and talent, population flow migration, and the level of urban economic development are taken for correlation verification. Among them, talent and population migration are measured by the proportion of mobile population with college education or above (C_pro) and the proportion of nonlocal household registration in the resident population (M_pro), respectively, and the level of urban economic development is measured by gross domestic product per capita (G_per). Pearson and Spearman correlation analyses were used to test the correlation between urban comfort, population mobility, and economic development level of 253 prefecture-level cities in China, respectively, where Pearson test used the original data and Spearman test used the urban rank order (from the highest to lowest) of the sample cities on each variable instead of the original data.

There are positive correlations between the proportion of college education or above (C_pro) and the proportion of nonlocal household registration in the resident population (M_pro) and urban creative environment comfort (FAC1) and public service comfort (FAC2), among which the correlation between urban public service and the proportion of foreign population is more significant, and urban GDP per capita (G_per) is also significantly and positively correlated with the proportion of foreign population; in addition, there is also a significant positive correlation between English education in both types of colleges and universities and the GDP per capita in cities; with reference to the city rank-order-size rule, Spearman’s test is applied to examine the rank correlation between talent inflow and the city comfort system and economic system in the sample cities, and the difference from the previous results is that the effect of GDP...
per capita on the percentage of talent inflow changes from insignificant to significant, and the effect of public service comfort on the percentage of foreign population changes from significant to insignificant. The reason may be that the proportion of the mobile population with college education or above and the proportion of nonlocal household registration in the resident population do not exactly obey the normal distribution in the field of economics, which leads to the deviation of the two tests. In summary, there is a significant positive correlation between urban talent concentration and urban creative environment and public service level; thus, it can be judged as talent concentration and urban creative environment and public service level, a significant positive correlation between urban talent concentration and urban comfort level and per capita GDP, and a certain rank concentration in the resident population do not exactly obey the normal distribution in the field of economics, which leads to the deviation of the two tests. In summary, there is a significant positive correlation between urban talent concentration and urban creative environment and public service level; thus, it can be judged as talent concentration, urban comfort, and urban economic development. There may be an endogenous interaction between talent concentration, urban comfort, and urban economic development.

3.2. Nonlinear Shrinkage Estimation. In order to solve the problem of estimating the overall covariance matrix, various estimation methods have been proposed by domestic and foreign scholars. Among them, Rotation Invariant Estimation (RIE) has been widely and successfully used because it can estimate the overall covariance matrix well without any prior knowledge of the covariance structure.

In the rotationally invariant estimation method, the spectral decomposition of the overall covariance matrix is first assumed to be of the form:

\[ Q = \sum_{i=1}^{n} \sum_{j=1}^{n} t_{i,j} v_i v_j^T, \]

where \( v_i \) and \( t_i \) are the eigenvalues of \( T \) and the corresponding eigenvectors, respectively. Similarly, the sample covariance matrix \( S \) can be decomposed as:

\[ S = \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_{i,j} u_i u_j^T, \]

where \( \lambda_{i,j} \) are the eigenvalues and the corresponding eigenvectors of \( S \), respectively. Rotation invariant estimation is the expectation to find the estimate of the overall covariance matrix \( S \) from the sample covariance matrix \( S \) in a rotation invariant manner. That is, there exists any \( p \times p \) orthogonal matrix \( Q \) such that \( Q \) satisfies:

\[ Q(S)Q^T(S) = \sum_{i=1}^{n} \sum_{j=1}^{n} u_i u_j^T. \] (5)

Kernel estimates of the probability density function and its Hilbert transform can be obtained using the kernel method:

\[ f_k(x) = \frac{1}{n} \sum_{i=1}^{p} k\left(\frac{x - \lambda_i}{B_{nk}}\right), \quad \forall x \in R. \] (6)

Based on the above estimates, the contraction equation for nonzero eigenvalues can be further derived when \( p > n \):

\[ \xi_k = \frac{1}{n^2 \lambda_1} \sum_{i=1}^{p} k\left(\frac{x - \lambda_i}{B_{nk}}\right), \quad \forall x \in R. \] (7)

And for \( (p > n) 0 \) eigenvalues, the following nonlinear contraction function can be established:

\[ SF(o) = \sum_{i=p+1}^{n} k\left(\frac{x - \lambda_i}{B_{nk}}\right). \] (8)

The eigenvalue interception method is a high-dimensional covariance matrix estimation method that directly uses the Marchenko–Pastur law in random matrix theory to adjust the sample eigenvalues, and has also been applied in the fields of gas identification and financial analysis. The Marchenko–Pastur law can well describe the distribution of eigenvalues of a sample. The basic idea is that all eigenvalues of the sample covariance matrix \( S \) greater than or equal to \( q \) are considered as useful information, while eigenvalues less than \( + q \) are random noise in accordance with the Marchenko–Pastur law. In order to be able to infer the overall covariance matrix from the sample covariance.
matrix $S$, the redundant information contained in the sample covariance matrix $S$ needs to be removed. First, the sample covariance matrix $S$ is spectrally decomposed to find $v_i$ and $u_i$, and the eigenvector $u_i$ is kept constant, and then the overall covariance matrix is estimated using the following equation.

The implementation of the linear discriminant analysis algorithm requires estimating the prior, mean, and covariance matrix for each class using the training data set trainX. However, in the high-dimensional case, the estimated covariance matrices are usually pathological or even singular. To address the problem of estimating high-dimensional covariance matrices in LDA, this article applies two covariance matrix estimation methods, nonlinear shrinkage, and eigenvalue interpolation, to LDA algorithms, respectively, using relevant studies in random matrix theory, and obtains a discriminant algorithm applied to high-dimensional data classification. The algorithm design is shown in Figure 2.

To ensure the stability of the impulse response and variance decomposition, the variables need to be tested for smoothness, and considering the use of short panel data in this article, LLC, IPS, and HT tests are selected. As shown in Figure 3, except for $S_{pro}$, which fails the LLC test, all variables pass the LLC test at the 1% level and the IPS test at the 1% level. Pro and $G_{rat}$ all pass the smoothness test and cointegration test and are suitable for P-VAR analysis. The GMM estimation of the four variables $sev$, $C_{pro}$, $G_{rat}$, and $S_{pro}$ was performed, and the optimal lag order of the model estimation was determined to be order 1 based on the AIC and SC information criteria.

The Hausman test was conducted for its fixed benefits and random effects, and the hypothesis of individual randomness was rejected, so the model of individual solidity was set. Due to the variability of each region, i.e., the existence of heteroskedasticity in the cross-section, White Cross section was chosen for the regression. Considering the regional variability, regression analysis was also conducted with East-West and Northeast regions as samples, and the regression results are shown in Figure 4. As can be seen from the table, the significance test is passed at the national level, and for every 1 unit increase in the development of high-tech industries, the concentration of scientific and technological talents increases by 0.4906 units. The eastern region has a more significant effect than the other regions in terms of agglomeration, with every 1 unit increase in high-tech industry in the eastern region contributing to a 0.3493 unit increase in the agglomeration of scientific and technological talent, and passing the 10% significance test. The Western region has the smallest effect, but does not pass the significance test. It is noteworthy that wages have a dampening effect on the concentration of scientific and technological talent, both at the national and regional levels. The strongest inhibitory effect is found in the Northeast region, where every 1 unit increase in wages inhibits technological talent clustering by 0.3168 units. The Western region has the weakest inhibitory effect. The brain drain in the Northeast has been very serious in recent years, and relying on wages alone to attract talent is not reliable.

### 3.3. Measurement Model Design.

The $q$ values in the experiments are set to 0.1, 0.3, 0.6, and 0.8, and the number of variables $p$ is 300. Since the difference between the mean of the optimal shrinkage estimate and the estimate of the sample mean on the actual mean is small when the data samples are large, it basically has no effect on the regularized discriminant classification algorithm, and the next simulation experiments focus on the high-dimensional data with $p_n$. The number of samples for all training data is set to 180, and another 1200 test data are generated. Based on the significance of the threshold regression results, this article establishes sequential logistic regression models with the mobile population in cities of categories II, III, and IV as the samples, the creative environmental comfort level of the inflowing cities as the explanatory variables, and the age, education level, number of children, mobility time, monthly income, and working hours per week of the mobile population as the explanatory variables, and the results of the likelihood ratio test (Model Fitting Information) is less than 0.001, the model is statistically significant, and the Test of Parallel Lines is less than 0.001, the hypothesis of equal coefficients of independent variables is accepted.

The classification results of the proposed RDAIMV algorithm and other comparison algorithms are shown in Figure 5, and RDAIMV1 and RDAIMV2 in the table correspond to RMRDA1 and RMRDA2 with improved mean estimation, respectively. The average correct classification rate of each algorithm in the table is averaged from the results of 50 experiments, where these are the classification results corresponding to data of type (a) and (b) of the mean value, respectively. From the comparison of the correct classification rates of the algorithms, the algorithm with the mean value of type (b) data has better classification results than the algorithm with the mean value of type (a) data. The algorithm RDAIMV consistently outperforms the RMRDA algorithm on simulated data and in most cases also outperforms the other compared algorithms. The correct classification rate of the RDAIMV algorithm is slightly lower than that of the smDLDA algorithm only when the sample correlation is low. In addition, as the sample correlation increases, the classification correctness of the RDAIMV2 algorithm is significantly higher than that of the other comparison algorithms.

In addition, the classification performance of the RDAIMV1 algorithm is better than that of the RDAIMV2 algorithm in most cases. The class accuracy of the RDAIMV algorithm is relatively high on the microarray dataset compared with other compared algorithms, while the classification accuracy of the RDAIMV algorithm is relatively low on the Mfeat handwritten character data, but still maintains a high classification accuracy. The effect of the number of samples $n$ in the Mfeat handwritten character dataset and the number of variables $p$ in the microarray dataset on the classification results is also verified. It was verified that for the Mfeat handwritten character dataset, the average correct classification rate of the RDAIMV algorithm increases gradually with the increase of sample $n$. The correct classification rate of the RDAIMV algorithm increases compared with that of the RMRDA algorithm, but the
improvement is not significant. For the microarray dataset, the RDAIMV algorithm always maintains a higher classification correct rate as the data dimension $p$ increases, which significantly improves the classification performance of the RMRDA algorithm, and the RDAIMV algorithm also outperforms the other compared algorithms in most cases.

4. Numerical Experiments and Results Analysis

In terms of the shock response effect, English education service employees produce a significant positive impulse response to a unit shock that peaks in period 1 and then tend to converge, as shown in Figure 6. This indicates that the economic vitality of cities is the main source of new jobs in consumer service industries; the talent concentration level ($C_{pro}$) produces a significant negative impulse response to GDP growth rate ($G_{rat}$) a unit shock that peaks in period 1 and then converges, which may be due to the fact that the industrial structure of cities with faster economic growth rate in the first size gradient often needs to be optimized, and the slowdown of economic growth rate or the slowdown of economic growth may be a signal of the initial completion of industrial transformation of the city, and at the same time attracts a large number of talents to gather; the positive impulse response of economic growth to the number of employees in the consumer service industry (sev) and the proportion of output value of tertiary industry ($S_{pro}$) indicates that the service-oriented economy is an important component of the comprehensive economic strength of the city and is the endogenous driving force of the city’s economic growth; the proportion of output value of tertiary industry ($S_{pro}$) has a positive effect on the talent. The positive impulse response of one unit shock of tertiary industry output value ($S_{pro}$) to talent concentration level ($C_{pro}$) peaks in period 2-3 and then converges, indicating that talent resources are an important influencing factor in the transformation and development of urban industrial structure, and cities in transition should focus on the improvement of human capital quality to drive the optimization of urban resource allocation and industrial structure adjustment.

This article intends to explore the nonlinear effects of wages, housing prices, and various types of comfort on the educational structure of the inflowing population in different types of cities, and the following requirements should be met when choosing the threshold variables: ① better reflect the gradient change of city scale; ② better reflect the overall consumption demand and payment level of city residents; and ③ better reflect the regional differences in the level of economic development of China’s cities. Based on the above considerations, the threshold model is constructed by combining the actual situation of China’s urban economic development: the average wage of employed workers (Income) is selected as the threshold variable $q_1$, which reflects the regional differences in the wages of employed workers in China’s cities and influences the regional flow of senior human capital to a certain extent; the Housing Price-to-Income Ratio (H2I) is selected as the threshold variable $q_2$. To demonstrate the classification performance of the regularized discriminant analysis algorithm with improved mean estimation for high-dimensional data, several real high-dimensional data are experimentally validated separately. In this article, the ACCR and its SD for three real data with different classification algorithms are given, and the results of each experiment are obtained by averaging the results of 50 experiments. According to the experimental results, it can be seen that the mean-improved RDAIMV algorithm performs better than the RMRDA algorithm in the classification of real data.

Ratio (H2I) is selected as the threshold variable $q_2$, which reflects the gradient change of urban development scale in China to a certain extent, and this phenomenon has been
empirically proved in the studies of many geographers in China, while the larger the scale of a city, the more economic advantage it has compared with other cities in the vicinity and the more obvious effect it has on the agglomeration of various factors such as human capital; the number of people with college education or above among the mobile population in each city (C_pro) is selected as the explanatory variable in equation (1); the city house price-to-income ratio (H2I), wage level (income), and other comfort variables after clustering are selected as the explanatory variables in equation (1), as shown in Figure 7.

Based on the threshold estimation results, as shown in Figure 8, the 253 sample cities are divided into thresholds, and the divided cities are defined into four types: Class I, Class II, Class III, and Class IV, and their spatial distribution and comfort characteristics are investigated. Cities in Class IV are mainly provincial capitals and subprovincial cities in terms of administrative level, which have higher income level and housing cost, and their residents enjoy better creative environment and urban public services. Cities in Class III are widely distributed in southeast coastal areas, which have a large inflow of labor force, which increases housing pressure on the one hand, and promotes urban economic construction and consumption, public services, and other functions on the other hand. Cities in Class II are scattered. These cities are scattered in the northern and southwestern regions of China, and their economic development and infrastructure construction are comparable to those of Class III cities, and their housing prices are low, but their geographical location is far from coastal and economically developed provinces, which limits their development space. The creative environment and public service construction of cities are poor, and they do not have advantages in the competition of labor force and especially senior human capital.

The coefficient of Income is significantly positive, indicating that income is still the main factor influencing the inflow of talents in mega and mega cities; the employment environment factor mainly reflects the degree of industrialization and modernization of cities and the ability to provide employment opportunities for manual workers, and its coefficient is significantly negative, indicating that in mega and mega cities, the degree of industrialization has become a negative comfort for cities: on the one hand, the concentration of a large number of industries, on the other hand, the employment of senior human capital is often not provided by industrial enterprises, which reduces the development space of tertiary industry in disguise; the
Figure 6: Impact of English education services practitioners on unit shock.

Figure 7: Mobility of English-speaking service professionals in universities.

Figure 8: Regional division of English services in universities based on threshold values.
coefficient of environmental quality factor is significantly positive, which indicates that mega and mega cities, as the main inflow of talents, are often densely populated and heavily polluted, although some cities in China have shown a reverse urbanization. Although some of the cities in China have shown a tendency to reverse urbanization, senior human capital will continue to gather in the central cities due to various considerations, and the density of senior human capital and the degree of environmental pollution in these cities have not reached the inflection point; meanwhile, the creative environment (FAC1) and the level of public services (FAC2) in Class IV cities have obvious advantages and strong heterogeneity, as shown in Figure 9, but the regression results are not significant. After reaching a certain threshold of city size, these cities have formed their own central business districts (CBDs) and other "elite spaces" and "comfort spaces" that meet the needs of urban elites for consumption and entertainment, medical education, and living services, and no longer rely on the overall city creative environment and living service facilities.

The regression results showed that three variables, such as age, education level, and working hours per week, had significant effects on the choice of the comfort of the creative environment of the mobile population. The specific analysis is as follows: in terms of age characteristics, the mobile population under the age of 35 has a significant preference for urban consumer comfort, and it increases with age, peaking at the stage of 30–35 years old and then decreasing rapidly. This is mainly because although young mobile individuals have a strong desire to consume and a strong desire to explore novelties, they are still at the stage of accumulating purchasing power and will give more consideration to the price level of the city of employment; as they get older, mobile individuals gradually accumulate more wealth and have better economic conditions to pursue consumption for a better life, while in middle age (after 35), their consumption desire may shift to some more rigid needs (mortgage expenses, children’s education, and support).

5. Conclusion

There is an obvious spatial correlation between the degree of urban talent concentration and the level of urban comfort, and highly educated talents prefer economically developed areas and regional central cities in making mobility decisions. There is an obvious endogenous interaction between urban consumption environment, talent concentration, and economic growth. Economic growth and the development of consumer service industries have a coupling effect, and there is an inertial growth mechanism for the development of service industries, talent concentration, and economic growth dynamics in cities; cities with a higher proportion of tertiary industry output have a positive and long-term impulse response to the level of talent concentration in cities, and talent resources are an important influencing factor for the transformation and development of urban industrial structure; talent concentration has a significant impact on new consumer service industries in cities, which accelerates the transformation of urban economic structure. In this article, we propose a regularized discriminant analysis algorithm with good estimation of the mean and a dimensionality reduction algorithm for high-dimensional missing data to address the problems of traditional machine learning methods in the analysis of high-dimensional data, using the relevant research results of random matrix theory. Although the linear discriminant analysis algorithm performs well in solving many practical problems, it works poorly in dealing with high-dimensional data. The reason is that when the data dimension \( p \) is close to or larger than the number of samples \( n \), the sample covariance matrix is no longer a good estimate of the true covariance matrix, resulting in large deviations in the linear discriminant function values. For this reason, a regularized discriminant analysis algorithm based on random matrix theory is proposed. First, a good estimate of the high-dimensional covariance matrix is made by the nonlinear shrinkage method or the eigenvalue interception method, respectively; then, the estimated high-dimensional covariance matrix is used to calculate the discriminant function values and perform the classification. The classification experiments conducted on simulated and real datasets show that the proposed algorithm is not only more widely applicable but also has a high correct classification rate [17, 18].

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest.
Acknowledgments

This work was supported by The English Department, Taiyuan University.

References

[1] E. Aucejo and J. James, “The path to college education: the role of math and verbal skills,” Journal of Political Economy, vol. 129, no. 10, pp. 2905–2946, 2021.

[2] M. Chen, H. A. Alterazi, and N. M. Lavin, “Task design and analysis of English reading teaching from the perspective of Activity Theory,” Applied Mathematics and Nonlinear Sciences, vol. 6, no. 2, pp. 261–268, 2021.

[3] Z. Luo, M. Zhang, and L. Jiang, “Innovation model of college education based on nonlinear random matrix organization and management[II],” Mathematical Problems in Engineering, p. 2022, 2022.

[4] C. De Cat, “Socioeconomic status as a proxy for input quality in bilingual children?” Applied Psycholinguistics, vol. 42, no. 2, pp. 301–324, 2021.

[5] G. Sun, S. Bin, M. Jiang et al., “Research on public opinion propagation model in social network based on blockchain,” Computers, Materials & Continua, vol. 60, no. 3, pp. 1015–1027, 2019.

[6] F. Farnia and E. Geva, “Late-emerging developmental language disorders in English-speaking monolinguals and English-language learners: a longitudinal perspective,” Journal of Learning Disabilities, vol. 52, no. 6, pp. 468–479, 2019.

[7] M. E. Foster, J. L. Anthony, T. A. Zucker, and L. Branum-Martin, “Prediction of English and Spanish kindergarten mathematics from English and Spanish cognitive and linguistic abilities in Hispanic dual language learners,” Early Childhood Research Quarterly, vol. 46, pp. 213–227, 2019.

[8] L. Gabriele, F. Bertacchini, A. Tavernise, L. Vaca-Cárdenas, P. Pantano, and E. Bilotta, “Lesson planning by computational thinking skills in Italian pre-service teachers,” Informatics in Education, vol. 18, no. 1, pp. 69–104, 2019.

[9] S. Bin and G. Sun, “Optimal energy resources allocation method of wireless sensor networks for intelligent railway systems,” Sensors, vol. 20, no. 2, p. 482, 2020.

[10] P. Guerrero-Rosada, C. Weiland, M. McCormick et al., “Null relations between CLASS scores and gains in children’s language, math, and executive function skills: a replication and extension study,” Early Childhood Research Quarterly, vol. 54, pp. 1–12, 2021.

[11] M. Hanly, B. Edwards, S. Goldfeld et al., “School starting age and child development in a state-wide, population-level cohort of children in their first year of school in New South Wales, Australia,” Early Childhood Research Quarterly, vol. 48, pp. 325–340, 2019.

[12] Y.-C. Lin and K. A. Magnuson, “Classroom quality and children’s academic skills in child care centers: understanding the role of teacher qualifications,” Early Childhood Research Quarterly, vol. 42, pp. 215–227, 2018.

[13] E. N. Muratov, J. Bajorath, R. P. Sheridan et al., “QSAR without borders,” Chemical Society Reviews, vol. 49, no. 11, pp. 3525–3564, 2020.

[14] Y. Qian, “Exploration of machine algorithms based on deep learning model and feature extraction,” Mathematical Biosciences and Engineering, vol. 18, no. 6, pp. 7602–7618, 2021.

[15] R. Raggio, V. Esposito, M. Di Mario, S. Masiello, M. Viscovo, and R. D. Cramer, “Teaching and learning computational drug design: student investigations of 3D quantitative structure–activity relationships through web applications,” Journal of Chemical Education, vol. 97, no. 7, pp. 1922–1930, 2020.

[16] S. S. R. Rizvi, M. A. Khan, S. Abbas, M. Asadullah, N. Anwer, and A. Fatima, “Deep extreme learning machine-based optical character recognition system for nastalique Urdu-like script languages,” The Computer Journal, vol. 65, no. 2, pp. 331–344, 2022.

[17] M. H. Sibley, S. J. Coxe, T. F. Page et al., “The ADHD teen integrative data analysis longitudinal (TIDL) dataset: background, methodology, and aims,” Journal of Clinical Child and Adolescent Psychology: The Official Journal for the Society of Clinical Child and Adolescent Psychology, American Psychological Association, Division 53, vol. 20, no. 1, pp. 1–14, 2020.

[18] P. Troncoso and N. Humphrey, “Playing the long game: a multivariate multilevel non-linear growth curve model of long-term effects in a randomized trial of the Good Behavior Game,” Journal of School Psychology, vol. 88, pp. 68–84, 2021.

[19] J. A. Washington, L. Branum-Martin, C. Sun, and R. Lee-James, “The impact of dialect density on the growth of language and reading in African American children,” Language, Speech, and Hearing Services in Schools, vol. 49, no. 2, pp. 232–247, 2018.

[20] W. Yong, R. Kharabsheh, and F. Bester, “Quality model construction and continuous improvement mechanism design in OBE education,” Applied Mathematics and Nonlinear Sciences, vol. 6, no. 2, pp. 339–346, 2021.

[21] G. Ziegler, M. Moutoussis, T. U. Hauser et al., “Childhood socio-economic disadvantage predicts reduced myelin growth across adolescence and young adulthood,” Human Brain Mapping, vol. 41, no. 12, pp. 3392–3402, 2020.