Modeling and Statistical Analysis for Student Performance

Qin Haocheng 11911317, Liu Zeyang 11911315, Cui Zhengqian 11912505

May 26, 2022

Abstract

In this paper, we conduct comprehensive analysis and evaluations of study performance based on a factor analysis model. In the beginning, 21 representative original evaluation indexes are selected to determine the principal components. On this basis, the weight of each evaluation index is calculated according to the estimation of the factor loading matrix. Then we defined the criteria for doing well on the final and for progress during the semester respectively by computing the weighted average of these indexes where the weights were obtained by the regression coefficients. Based on the criteria, we can directly estimate students' performance at the end of the semester and their progress during the semester. We found in our research that about 25% of the students are judged as bad and retrograding students. To verify the effectiveness of the proposed comprehensive evaluation method, we compared the estimated performance and the actual performance of the students, and these fitted well. Moreover, we did cluster analysis from which we can verify roughly that our factor analysis and modeling are effective enough. At last, we did canonical correlation analysis and found that family factors would also lead to changes in students' individual learning status with a strong correlation.

1 Introduction

1.1 Problem background

It has always been an issue worth discussing to analyze the factors that affect students' performance in school and predict their performance. If we can predict a student's performance at the end of the semester, and if we can identify specific factors, then we can improve the performance of a student who is expected to do poorly at the end of the semester. In other words, we can redistribute teaching resources by predicting students' final performance and progress, so as to maximize the use of teaching resources and achieve the best teaching quality. There are many people who have done related studies before us, and here we will list what they have done and point out the innovations of our study compared to previous studies.

1.2 Literature review

Huang (2013) [1] did a comparative research based on the 2907 data points collected from 323 undergraduate students over four semesters, whose majors contain Mechanical and Aerospace Engineering (MAE), Civil and Environmental Engineering (CEE), Biological Engineering, General Engineering, Pre-engineering, undeclared, or non-engineering majors. The predictor variables of this research contain before-semester grades like cumulative GPA, statics grade, calculus I grade, calculus II grade, and physics grade. Also, they contain during-semester grades like the dynamics mid-exam score. The response variable is the score on the dynamics final comprehensive exam. This research compared the performance of four models based on this data to predict the student's predicted grades. The four models are the multiple linear regression model, the multi-layer perception network model, the radial basis function network model, and the support vector machine model.

This research concludes that the four types of mathematical models do not have a significant difference in terms of the average prediction accuracy while the SVM model generally yields the highest percentage of accurate prediction among the four types of mathematical models. Moreover, for different models and different goals of prediction, this research proposed minimal subsets of variables respectively without loss of prediction accuracy. However, it can be noted that there should be more predictor variables to be considered into the model to ensure the accuracy of the predicted value. Then, Navamani (2015) [2] conducted an analysis of different characteristics of the data obtained from the results of primary school exams in Tamil Nadu (India) showing the relationship between ethnicity, geographic environment, and students' performance.

Bydžovská (2015) [3] did a research mainly based on collaborative filtering methods. It used the data set that comprised 62 courses with 3,423 students and their 42,635 grades and got the conclusion that students can be sufficiently characterized only by their previously passed courses that cover their knowledge of the field of their study if they can find students enrolled in the same courses in the last years that are the most similar to the investigated one. Pero (2014) [4] conducted a research showing that CF methods are still usable on a small and dense data set but compared with the results of Thai-Nghe et al. provided on a large and sparse KDD Cup 2010 data set [5] are much less precise.

Villagrá-Arnedo (2016) [6] did a research to study the use of different sets of data to predict the student performance in a subject and obtained a weekly ranking of each student's probability of belonging to one of these three classification levels: high, medium or low performance based on SVM. Also, it proposed that the use of heterogeneous significant data enriches the final performance of the prediction algorithms.

ML Clustering techniques have been satisfactorily applied in this field. Moreover, Park (2018) [7] enhanced the performance of prediction by considering the computing major-specific course characteristics, such as core courses, course prerequisites, and course levels.

From the literature review above, we can see that the research to study the affluence of background of students on the final grade and their progress during the semester still lacking. Besides, the model based on ML proposed to predict the performance lacks interpretability which leads to the difficulty in improving the performance of students who are expected to perform poorly. So we use factor analysis to decide the crucial factors that will affect the final grade and the progress during the semester. And we will propose comprehensive criteria according to the factors to predict the final level of grade and progress during the semester of students. Our analysis will be based on the data set named "Student Performance". This data set contained two sub-datasets which contain the data of 382 students and their performance in math course and Portuguese language course respectively. Our research will focus on the math performance. The variables for dimension reduction are shown in Fig.13 in the appendix. Moreover, it contains the variables that measure the performance of students. They are shown in Fig.14 in the appendix.

1.3 Structure

Our report will be structured by this: in section 2, we will introduce the methods used in our research; In section 3, experimental design like data procession and idea of the experiment will be stated; In section 4, there will be results and corresponding analysis; At last, we will state our conclusions of this research.

2 Methodology

Before the factor analysis, we apply the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity, which helps us to determine whether it is proper to conduct factor analysis on this data set. For the factor analysis, we use the principal component method to estimate the coefficients in the factor loading matrix. After that, we apply clustering analysis to the factors. Finally, we conduct canonical correlation analyses on some groups of variables.

2.1 Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy

The KMO measure is a statistics comparing the simple correlation coefficients and partial correlation coefficients of a given correlation matrix. The definition is

$$KMO = \frac{\sum\limits_{i<j} r_{ij}^2}{\sum\limits_{i<j} r_{ij}^2 + \sum\limits_{i<j} r_{(p)ij}^2}$$

where r_{ij} is the correlation coefficient between variable X_i and X_j, and $r_{(p)ij}$ is the partial correlation coefficient between X_i and X_j. The latter one is given by the element on the i$^{\text{th}}$ row and j$^{\text{th}}$ column of the matrix $\boldsymbol{Q = DR^{-1}D}$, where $\boldsymbol{R}$ is the correlation matrix of $\boldsymbol{X}$ and $\boldsymbol{D} = [(\text{diag } \boldsymbol{R^{-1}})^{\frac{1}{2}}]^{-1}$. [8]

When the correlations between X_i and X_j's are strong, r_{ij}'s will be relatively large and KMO can be near 1, which is proper for a factor analysis; while when X_i, X_j's are mutually uncorrelated, KMO will be 0, and the factor analysis will not give any help. Based on a subjective criterion, it is reasonable to apply a factor analysis when $KMO > 0.5$.

2.2 Bartlett's Test of Sphericity

The null hypothesis of Bartlett's test of sphericity is that $\boldsymbol{R}$, the correlation matrix of $\boldsymbol{X}$, is an identity matrix, i.e., all of X_i, X_j's are mutually uncorrelated.

To be specific, denote the dimension of $\boldsymbol{X}$ as $n \times p$. The test is

$$H_0 : \boldsymbol{R} = \boldsymbol{I}_p \quad v.s. \quad H_1 : \boldsymbol{R} \neq \boldsymbol{I}_p$$

and the test statistic is

$$\chi^2 = -[n - \frac{1}{6}(2p+5)] \ln|\boldsymbol{R}|$$

where $\chi^2 \sim \chi^2(\frac{1}{2}p(p-1))$ under H_0. [9]

We may conduct a factor analysis when the null hypothesis is rejected. The significance level of this test is usually set to be 0.05.

2.3 Factor Analysis

The main idea of the factor analysis is to express each component in $\boldsymbol{X}$ by a linear combination of a group of common factors, where such factors are usually latent, but the coefficient can be estimated by several methods. When there are correlations between some of the components (this can be tested by computing the KMO measure and applying Bartlett's test of sphericity), we can use a small number of factors to explain a large proportion of correlations, in order to reduce the dimension.

To be specific, let $\boldsymbol{x} = (x_1, x_2, \cdots, x_p)'$ be a p-dimension observable random vector, whose mean is $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_p)'$ and covariance matrix is $\boldsymbol{\Sigma} = (\sigma_{ij})$. Let m be the number of common factors we wish to have (clearly $m < p$ for the usefulness of the factor analysis), and $f_1, f_2, \cdots, f_m$ be the common factors, $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_p$ be the special factors. Then the general model of factor analysis can be expressed as

$$\begin{cases} x_1 = \mu_1 + a_{11}f_1 + a_{12}f_2 + \cdots + a_{1m}f_m + \varepsilon_1 \\ x_2 = \mu_2 + a_{21}f_1 + a_{22}f_2 + \cdots + a_{2m}f_m + \varepsilon_2 \\ \quad \vdots \\ x_p = \mu_p + a_{p1}f_1 + a_{p2}f_2 + \cdots + a_{pm}f_m + \varepsilon_p \end{cases} \tag{2.3.1}$$

or

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{f} + \boldsymbol{\varepsilon}$$

in the form of vectors. Here $\boldsymbol{A} = (a_{ij})$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_p)'$.

All of the common factors and the special factors are unobservable. The common factors appear in the expression of each of the initial variables $x_i (i = 1, 2, \cdots, p)$, and can be interpreted as the elements which the initial variables have commonly. Each common factor should have some effect on at least two initial variables, i.e., for each common factor, at least two lines of equation 2.3.1 have a coefficient significantly unequal to 0 before this factor; otherwise, it should be considered a special factor. The special factor is kind of like the error term in a linear model, which represents an individual deviation from the model, and is excluded in the estimated result of the model. So, later in the practical analysis, we will only focus on the estimation of $\boldsymbol{A}$, but not consider the special factors.

The most often used model, which is also the one we applied in the following analysis, is the orthogonal factor model. It has five assumptions:

$$\begin{cases} E(\boldsymbol{f}) = \mathbf{0} \\ E(\boldsymbol{\varepsilon}) = \mathbf{0} \\ Var(\boldsymbol{f}) = \boldsymbol{I} \\ Var(\boldsymbol{\varepsilon}) = \boldsymbol{D} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_\mathrm{p}^2) \\ Cov(\boldsymbol{f}, \boldsymbol{\varepsilon}) = E(\boldsymbol{f}\boldsymbol{\varepsilon}') = \mathbf{0} \end{cases} \tag{2.3.2}$$

These assumptions are seemingly strong, but actually, we can always meet them, if the initial variable is proper for factor analysis, and we apply some necessary transformations on $\boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{f}$ and $\boldsymbol{\varepsilon}$. These assumptions can be interpreted as follows.

First, in order to reduce the dimension adequately, the common factors should be mutually uncorrelated, therefore, $Var(\boldsymbol{f})$ should be a diagonal matrix, which is a part of the restrictions in the third line of equation 2.3.2.

Secondly, ε_i should not include any linear information about $f_1, f_2, \cdots, f_m$, i.e., ε_i should be uncorrelated with them. This is a natural state when the common factors do carry enough information, and this forms the fifth line in the equations.

Thirdly, as a continuation of the former interpretation, now $f_1, f_2, \cdots, f_m$ can explain all the covariance (or correlation coefficient equivalently) of $x_1, x_2, \cdots, x_p$, and it is reasonable to assume that $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_p$ are mutually uncorrelated. However, as the variation of the error is hard to be controlled, we may accept that the variances of $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_p$ are different. This leads to the fourth line of 2.3.2.

Finally, the first three lines of the equations are general restrictions on the structure of the model without loss of generality. If $E(\boldsymbol{f}) \neq \mathbf{0}, E(\boldsymbol{\varepsilon}) \neq \mathbf{0}$ or $Var(\boldsymbol{f})$ is diagonal but not an identity matrix, we may transform them by

$$\begin{cases} \boldsymbol{\mu}^\star = \boldsymbol{\mu} + \boldsymbol{A}E(\boldsymbol{f}) + E(\boldsymbol{\varepsilon}) \\ \boldsymbol{\varepsilon}^\star = \boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon}) \\ \boldsymbol{f}^\star = \boldsymbol{Q}^{-1}[\boldsymbol{f} - E(\boldsymbol{f})] \\ \boldsymbol{A}^\star = \boldsymbol{A}\boldsymbol{Q} \end{cases}$$

where $\boldsymbol{Q} = \mathrm{diag}(\sqrt{\mathrm{Var(f_1)}}, \sqrt{\mathrm{Var(f_2)}}, \cdots, \sqrt{\mathrm{Var(f_p)}})$. Then, the model

$$\boldsymbol{x} = \boldsymbol{\mu}^\star + \boldsymbol{A}^\star \boldsymbol{f}^\star + \boldsymbol{\varepsilon}^\star$$

satisfies the assumptions in 2.3.2 again.

From the viewpoint of matrix algebra, an important effect of the factor loading matrix $\boldsymbol{A}$ is that it can help to decompose the covariance matrix $\boldsymbol{\Sigma}$ or the correlation matrix $\boldsymbol{R}$. Notice that

$$\boldsymbol{\Sigma}(\mathrm{or}\boldsymbol{R}) = \mathrm{Var}(\boldsymbol{A}\boldsymbol{f} + \boldsymbol{\varepsilon}) = \mathrm{Var}(\boldsymbol{A}\boldsymbol{f}) + \mathrm{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{A}\mathrm{Var}(\boldsymbol{f})\boldsymbol{A}' + \mathrm{Var}(\varepsilon) = \boldsymbol{A}\boldsymbol{A}' + \boldsymbol{D}$$

where $\boldsymbol{D}$ is the diagonal matrix mentioned in 2.3.2.

Since $\boldsymbol{D}$ is diagonal, the elements in $\boldsymbol{\Sigma}$ or $\boldsymbol{R}$ off the diagonal are determined by $\boldsymbol{A}$ only. This observation is a basis of the computation method we used to estimate $\boldsymbol{A}$. It ca be easily find that if $m = p$ then we may assign $\boldsymbol{A} = \boldsymbol{\Sigma}^{1/2}$ and $\boldsymbol{D} = \mathbf{0}$, but this decomposition provide no assist for dimension reduction. In practical usages, we need to choose a proper m so that it is as small as possible while keep $\boldsymbol{A}\boldsymbol{A}^{\star}$ a not so bad approximately decomposition of Σ or $\boldsymbol{R}$.

We may notice that the factor loading matrix $\boldsymbol{A}$ is not unique for a given $\boldsymbol{X}$ and a fixed number of factors. Let $\boldsymbol{T}$ be an $m \times m$ orthogonal matrix, and let $\boldsymbol{A}^{\star} = \boldsymbol{A}\boldsymbol{T}$, $\boldsymbol{f}^{\star} = \boldsymbol{T}'\boldsymbol{f}$. Then, the initial model

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{f} + \boldsymbol{\varepsilon}$$

can also be expressed as

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{A}^{\star}\boldsymbol{f}^{\star} + \boldsymbol{\varepsilon}$$

Meanwhile, the assumptions in 2.3.2 are also satisfied, and there is still a decomposition

$$\boldsymbol{\Sigma}(\text{or}\boldsymbol{R}) = \boldsymbol{A}^{\star}\boldsymbol{A}^{\star\prime} + \boldsymbol{D}$$

So, we can deem that $\boldsymbol{A}$ and $\boldsymbol{A}^{\star}$ have the same effect of assigning loading to common factors in a sense of explaining the covariance of $\boldsymbol{X}$, and the only difference is that the interpretation of the meaning of each factor is different. Since such orthogonal matrix $\boldsymbol{T}$ can make the factor loading matrix various, we can use such "rotation" method to get a factor loading matrix to make the factors easier to be interpreted.

To get the final $\boldsymbol{A}$, the first step is to compute any of the possible $\boldsymbol{A}$, and the second step is to rotate it. We apply the principal component method to complete the former step.

Since the real covariance or correlation matrix are usually unable to be observed, we usually use the corresponding sample covariance or correlation matrix $\boldsymbol{S}$ or $\hat{\boldsymbol{R}}$ to take the place of them. For the simplicity of notation, we only mention $\boldsymbol{S}$ in the rest of this subsection.

Let the p eigenvalues of $\boldsymbol{S}$ be $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$, and the corresponding orthogonal unit eigenvector be $\hat{\boldsymbol{t}}_1, \hat{\boldsymbol{t}}_2, \cdots, \hat{\boldsymbol{t}}_p$. When the cumulative rate of contribution $\sum_{i=1}^{m} \hat{\lambda}_i / \sum_{i=1}^{p} \hat{\lambda}_i$ is relatively high, $\boldsymbol{S}$ can be approximately decomposed as

$$\begin{aligned}
\boldsymbol{S} &= \hat{\lambda}_1 \hat{\boldsymbol{t}}_1 \hat{\boldsymbol{t}}_1' + \cdots + \hat{\lambda}_m \hat{\boldsymbol{t}}_m \hat{\boldsymbol{t}}_m' + \hat{\lambda}_{m+1} \hat{\boldsymbol{t}}_{m+1} \hat{\boldsymbol{t}}_{m+1}' + \cdots + \hat{\lambda}_p \hat{\boldsymbol{t}}_p \hat{\boldsymbol{t}}_p' \\
&\approx \hat{\lambda}_1 \hat{\boldsymbol{t}}_1 \hat{\boldsymbol{t}}_1' + \cdots + \hat{\lambda}_m \hat{\boldsymbol{t}}_m \hat{\boldsymbol{t}}_m' + \hat{\boldsymbol{D}} \\
&= \hat{\boldsymbol{A}} \hat{\boldsymbol{A}}' + \hat{\boldsymbol{D}}
\end{aligned}$$

where $\hat{\boldsymbol{A}} = (\sqrt{\hat{\lambda}_1}\hat{\boldsymbol{t}}_1, \cdots, \sqrt{\hat{\lambda}_m}\hat{\boldsymbol{t}}_m) = (\hat{a}_{ij})$ is a $p \times m$ matrix.

This procedure can also be done for sample correlation matrix if the units of variables are different or the variances of the variables have large difference.

As for the rotation, we want $\boldsymbol{A}$ to have one element with an absolute value near 1 (not possible to be larger than 1) while other elements near 0 on each row. A popular approach is the varimax rotation method, and we use it as well.

2.4 Clustering Analysis

The main target of clustering analysis is to divide the objects into several classes by certain criteria. The number of classes and the condition to put an object into a class may not be decided before we do the analysis, so such analysis is powerful when we are researching samples that we do not know a good criterion for classifying.

Generally speaking, we need to choose a definition of the distance of the samples before we start. Then, samples are assigned to the same class if they are near. This procedure will be repeated several times and the classes are formed gradually step by step. The two main kinds of clustering methods are systematic clustering and dynamic clustering. The most important difference is that the latter permits a sample to be assigned to another class after it has been assigned to one, while systematic clustering does not allow that. We conduct a dynamic clustering analysis using the k means method. When applying this procedure we need to decide the number of classes k in advance.

The detailed procedure of this method is:

1. Choose k samples as initial condensation points, or divide all samples into k initial classes, and use the k centers of gravity (means) as initial condensation points. This step needs to be done subjectively.

2. Classify the samples one by one. At each time a sample is assigned to the class whose condensation point is the nearest to it. Then the condensation points are updated. One term of such operation ends when every sample is classified once.

3. Repeat the whole term of procedure in 2 until in one term all of the samples keep staying in the classes they were in in the last term.

When using k means, the distance of samples is usually the Euclidean distance.

2.5 Canonical Correlation Analysis

Canonical correlation analysis is used to explore the linear correlation between two groups of variables.

Let $\boldsymbol{x} = (x_1, x_2, \cdots, x_p)'$ and $\boldsymbol{y} = (y_1, y_2, \cdots, y_q)$ be two random vectors, and

$$Var\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix}$$

where $Var(\boldsymbol{x}) = \boldsymbol{\Sigma}_{11}$ and $Var(\boldsymbol{y}) = \boldsymbol{\Sigma}_{22}$ are positive definite.

Then, let $\boldsymbol{a} = (a_1, a_2, \cdots, a_p)'$ and $\boldsymbol{b} = (b_1, b_2, \cdots, b_q)'$ be nonzero constant vectors, and "compress" $\boldsymbol{x}$ and $\boldsymbol{y}$ into two single variables by $\boldsymbol{a}$ and $\boldsymbol{b}$, i.e., let $u = \boldsymbol{a'x}$ and $v = \boldsymbol{b'y}$.

The purport of canonical correlation analysis is to find proper u and v to maximize $Corr(u, v)$, since we want to find linear combinations of component of $\boldsymbol{x}$ and $\boldsymbol{y}$ respectively in the way where the linear relationship is furthest expressed by u and v.

Now, begin with

$$\begin{aligned}\mathrm{Cov}(u, v) &= \mathrm{Cov}\left(\boldsymbol{a'x}, \boldsymbol{b'y}\right) = \boldsymbol{a}'\,\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})\boldsymbol{b} = \boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{b}\\ Var(u) &= Var\left(\boldsymbol{a'x}\right) = \boldsymbol{a}'Var(\boldsymbol{x})\boldsymbol{a} = \boldsymbol{a}'\boldsymbol{\Sigma}_{11}\boldsymbol{a}\\ Var(v) &= Var\left(\boldsymbol{b'y}\right) = \boldsymbol{b}'Var(\boldsymbol{y})\boldsymbol{b} = \boldsymbol{b}'\boldsymbol{\Sigma}_{22}\boldsymbol{b}\end{aligned}$$

we can get

$$\rho(u, v) = \frac{\boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{b}}{\sqrt{\boldsymbol{a}'\boldsymbol{\Sigma}_{11}\boldsymbol{a}}\sqrt{\boldsymbol{b}'\boldsymbol{\Sigma}_{22}\boldsymbol{b}}}$$

Since $\rho(u, v)$ will be the same if we simply multiple one or two of them by constants, in order to get an unique solution, we add a normalizing restriction on u and v, i.e., $Var(u) = 1$ and $Var(v) = 1$, or $\boldsymbol{a}'\boldsymbol{\Sigma}_{11}\boldsymbol{a} = 1$ and $\boldsymbol{b}'\boldsymbol{\Sigma}_{22}\boldsymbol{b} = 1$ equivalently.

Now the computational question is clear: we need to compute $\boldsymbol{a}$ and $\boldsymbol{b}$ which maximize $\rho(u, v) = \boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{b}$ under restrictions $\boldsymbol{a}'\boldsymbol{\Sigma}_{11}\boldsymbol{a} = 1$ and $\boldsymbol{b}'\boldsymbol{\Sigma}_{22}\boldsymbol{b} = 1$. Before show the explicit expressions of the solutions, we denote them as $\boldsymbol{a} = \boldsymbol{a}_1$ and $\boldsymbol{b} = \boldsymbol{b}_1$; and we say $u_1 = \boldsymbol{a}_1'\boldsymbol{x}$, $v_1 = \boldsymbol{b}_1'\boldsymbol{y}$ are the first pair of canonical variables. Similarly, we call $\boldsymbol{a}_1, \boldsymbol{b}_1$ as the first pair of canonical coefficient vectors, and $\rho_1 = \rho(u_1, v_1)$ as the first canonical correlation coefficient.

After computing u_1 and v_1 above, we may want to give another pair of canonical variables to capture the most correlations between $\boldsymbol{x}$ and $\boldsymbol{y}$ which are not captured by the first pair, in a sense like the second principal component after the first one in the computation of principal components. A natural restriction on the second pair of canonical variables, $u_2 = \boldsymbol{a'x}$ and $v_2 = \boldsymbol{b'y}$, apart from the standardizing restrictions $\boldsymbol{a}'\boldsymbol{\Sigma}_{11}\boldsymbol{a} = 1$ and $\boldsymbol{b}'\boldsymbol{\Sigma}_{22}\boldsymbol{b} = 1$, is that they should not include any information about correlation that was included in u_1 and v_1. That is,

$$\begin{aligned}\rho\left(u_2, u_1\right) &= \rho\left(\boldsymbol{a'x}, \boldsymbol{a}_1'\boldsymbol{x}\right) = \mathrm{Cov}\left(\boldsymbol{a'x}, \boldsymbol{a}_1'\boldsymbol{x}\right) = \boldsymbol{a}'\boldsymbol{\Sigma}_{11}\boldsymbol{a}_1 = 0\\ \rho\left(v_2, v_1\right) &= \rho\left(\boldsymbol{b'y}, \boldsymbol{b}_1'\boldsymbol{y}\right) = \mathrm{Cov}\left(\boldsymbol{b'y}, \boldsymbol{b}_1'\boldsymbol{y}\right) = \boldsymbol{b}'\boldsymbol{\Sigma}_{22}\boldsymbol{b}_1 = 0\end{aligned}$$

so the target is to maximize $\rho\left(u_2, v_2\right) = \rho\left(\boldsymbol{a'x}, \boldsymbol{b'y}\right) = \boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{b}$ under the two restrictions above.

It can be shown (the details are omitted here) that, denoting $m = \text{rank}(\mathbf{\Sigma_{12}})$, we can compute m pairs of canonical variables step by step. To be general, the $i^{th}(1 < i \leq m)$ canonical variables $u_i = \boldsymbol{a'x}$ and $v_i = \boldsymbol{b'y}$ are computed by maximize

$$\rho(u_i, v_i) = \rho\left(\boldsymbol{a'x}, \boldsymbol{b'y}\right) = \boldsymbol{a'}\mathbf{\Sigma}_{12}\boldsymbol{b}$$

under restrictions

$$\boldsymbol{a'}\mathbf{\Sigma}_{11}\boldsymbol{a} = 1, \quad \boldsymbol{b'}\mathbf{\Sigma}_{22}\boldsymbol{b} = 1$$
$$\boldsymbol{a'}\mathbf{\Sigma}_{11}\boldsymbol{a}_k = 0, \quad \boldsymbol{b'}\mathbf{\Sigma}_{22}\boldsymbol{b}_k = 0$$

where $k = 1, 2, \cdots, i-1$.

Finally, we show the expression of the m pairs of solutions $\boldsymbol{a}_i, \boldsymbol{b}_i$ (the detailed proof is omitted). We can show that $\text{rank}(\mathbf{\Sigma}_{22}^{-1/2}\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1/2}) = \text{m}$, and $\mathbf{\Sigma}_{22}^{-1/2}\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1/2}(\geq 0)$ has m nonzero eigenvectors $\rho_1^2 \geq \rho_2^2 \geq \cdots \geq \rho_m^2 > 0$. Let $\rho_i = \sqrt{\rho_i^2},\ i = 1, 2, \cdots, m$. Denote the m orthogonal unit eigenvectors corresponding to $\rho_1^2, \rho_2^2, \cdots, \rho_m^2$ of $\mathbf{\Sigma}_{22}^{-1/2}\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1/2}$ as $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_m$, and let $\boldsymbol{\alpha}_i = \frac{1}{\rho_i}\mathbf{\Sigma}_{11}^{-1/2}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1/2}\boldsymbol{\beta}_i$ for $i = 1, 2, \cdots, m$. Then, for such i's, the canonical variables are given by $\boldsymbol{a}_i = \mathbf{\Sigma}_{11}^{-1/2}\boldsymbol{\alpha}_i$ and $\boldsymbol{b}_i = \mathbf{\Sigma}_{22}^{-1/2}\boldsymbol{\beta}_i$.

The above procedure compute the canonical variables and canonical correlation coefficients from the population covariance matrix $\mathbf{\Sigma}$, but similarly we can also compute them from the population correlation matrix $\boldsymbol{R}$. It can be shown that although the canonical variables from $\boldsymbol{R}$ are centralized, the m canonical correlation coefficient are all the same with those from $\mathbf{\Sigma}$.

When the population covariance or correlation matrix are unobservable, we can replace them with sample covariance matrix $\boldsymbol{S}$ or the sample correlation matrix $\hat{\boldsymbol{R}}$ respectively, and follow the steps given above to apply a canonical correlation analysis.

3 Experimental Design

3.1 Data Procession

First of all, as the factor analysis is not particularly suitable for categorical variables, in order to deal with discrete variables more reasonably, we can regard them as a kind of explanation of the degree. For example, with the variable "paid", we can think that 0 represents no extra effort at all, and 1 represents the degree of incubating extra effort very deep. Numbers between 1 and 0 represent degrees in between, and although we can't observe such numbers, they are conceptually well defined.

Thus, in a sense, we can treat the original data set as a continuous variable. Of course, this also means that we must eliminate variables that do not fit well with the concept of "degree" : **Mjob**, **Fjob**, **reason**, **Guardian**, these are typical categorical variables, and there are more than two categories.

And since we only want to know the impact of student-related factors on students' performance, we should remove course differences and school differences. We only select the math score data set of one school here.

Of course, the scale of each variable is different. Therefore, we also need to scale them between 0 and 1, that is, transform each variable X_i as follows:

$$X_i \longleftarrow \frac{X_i - Min\{X_i\}}{Max\{X_i\} - Min\{X_i\}} \tag{1}$$

Of course, this is only a preliminary data processing, and then we will choose the appropriate variables in other ways

3.2 Experiment Idea

Since we have test scores at different stages, naturally, we can also define the progress of each student.

$$Advance = \frac{G2 - G1}{G1} + \frac{G3 - G2}{G2} \tag{2}$$

Please note that this is just my personal definition. Intuitively, this can represent the size of progress "in a sense". And, for the convenience of calculation, of course, we should delete all missing values.

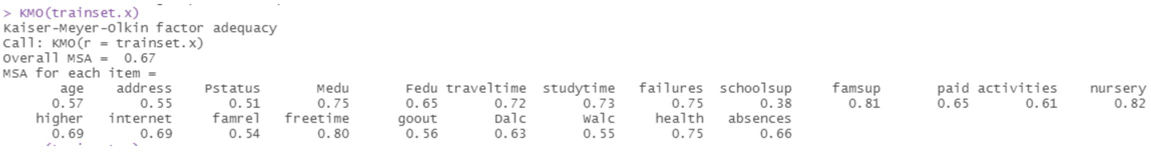
Our experimental idea is to analyze the results of factor analysis, but in order to achieve this, we first need to verify that it has an enough large KMO measure and passes Bartlett's test. We plan to delete the variables that perform poorly in the evaluation of KMO measure to enhance the explanatory degree of the model. Then we use the principal component method to conduct factor analysis and select the first several variables with a total explanatory degree greater than 70% (this is reasonable. It is generally believed that it is difficult to have an explanatory degree greater than 60% for sociological problems [10]). In order to better explain these variables, we need to rotate the factors. Finally, we can use these factors to model students' achievement and progress respectively, and the modeling coefficient can be approximately obtained by linear regression, because it can at least make our index positively correlated with the response variable.

In addition, we can make cluster analysis on the indicators after dimensionality reduction to see if we can significantly find out several different types of students; through our analysis of factors, maybe we can summarize several groups of variables associated with different aspects, and conduct Canonical Correlation Analysis (CCA) between variable groups to see whether the variable groups will affect each other.

4 Empirical Analysis

4.1 KMO test and Bartlett's Test of Sphericity

We first use R for KMO test and the output is shown in figure1.



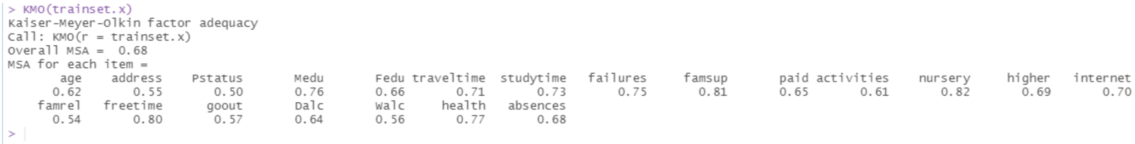
```
> KMO(trainset.x)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = trainset.x)
Overall MSA =  0.67
MSA for each item = 
       age    address    Pstatus       Medu       Fedu traveltime  studytime   failures  schoolsup     famsup       paid activities    nursery 
      0.57       0.55       0.51       0.75       0.65       0.72       0.73       0.75       0.38       0.81       0.65       0.61       0.82 
    higher   internet     famrel   freetime      goout       Dalc       Walc     health   absences 
      0.69       0.69       0.54       0.80       0.56       0.63       0.55       0.75       0.66 
```

Figure 1: Primary KMO Test

From the results, we have basically met the conditions for factor analysis. However, when we look at each variable separately, we find that the MSA of **schoolsup** is very low, which indicates that the effect of factor analysis will become worse after adding this variable, so we need to delete it.

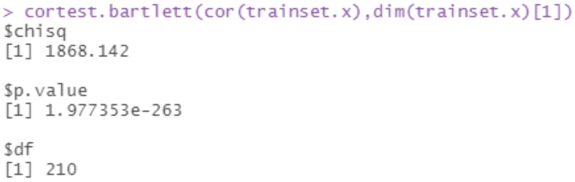
After deletion, it is tested and found that the effect is much better. KMO > 0.5, which is suitable for factor analysis. And also passed the Bartlett's test for p >> 0.05, so we finally selected these 21 variables for factor analysis.

More specificly, they are: **age** (student's age), **address** (whether student lives in urban), **Pstatus** (whether student's parents live together), **Medu/Fedu** (Mother/Father's education degree), **traveltime** (home-school travel time), **studytime** (time spent on study per week), **failures** (class failures), **famsup** (family education support), **paid** (whether have extra paid classes), **activities** (whether take extra-curricular activities), **nursery** (whether went to nursery school), **higher** (whether want higher education), **internet** (whether can use internet), **famrel** (family relationship degree), **freetime** (free time after school), **goout** (out play times), **Dalc/Walc** (workday/weekday alcohol consumption), **health** (health status), **absences** (absent times).



```
> KMO(trainset.x)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = trainset.x)
Overall MSA =  0.68
MSA for each item = 
       age    address    Pstatus       Medu       Fedu traveltime  studytime   failures     famsup       paid activities    nursery     higher   internet 
      0.62       0.55       0.50       0.76       0.66       0.71       0.73       0.75       0.81       0.65       0.61       0.82       0.69       0.70 
    famrel   freetime      goout       Dalc       Walc     health   absences 
      0.54       0.80       0.57       0.64       0.56       0.77       0.68 
> 
```

Figure 2: KMO Test

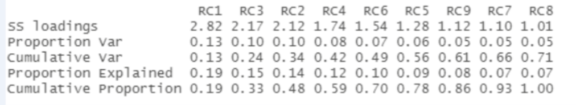


```
> cortest.bartlett(cor(trainset.x),dim(trainset.x)[1])
$chisq
[1] 1868.142

$p.value
[1] 1.977353e-263

$df
[1] 210
```

Figure 3: Bartlett's Test

4.2 Factor Analysis

We firstly conduct Factor Analysis based on Principle Components. We found that for the first 9^{th} principles, the cumulative variance go beyond 70%, so we can just take these 9.



```
                       RC1  RC3  RC2  RC4  RC6  RC5  RC9  RC7  RC8
SS loadings           2.82 2.17 2.12 1.74 1.54 1.28 1.12 1.10 1.01
Proportion Var        0.13 0.10 0.10 0.08 0.07 0.06 0.05 0.05 0.05
Cumulative Var        0.13 0.24 0.34 0.42 0.49 0.56 0.61 0.66 0.71
Proportion Explained  0.19 0.15 0.14 0.12 0.10 0.09 0.08 0.07 0.07
Cumulative Proportion 0.19 0.33 0.48 0.59 0.70 0.78 0.86 0.93 1.00
```

Figure 4: Cumulative Variance

By rotation procedure, we get the result shown in figure.5. Then we explain the meaning of each dimensionally reduced variable according to the size of the coefficient (For convinient, we masked the value below 0.4. The full result is attached as Appendix B. And for meaning of each variable, recheck figure.13):

RC1 ⇐ **Study Devotion Factor**: It is significantly positively correlated with"studytime" and "paid", and negatively correlated with "freetime". We believe that the more devoted and focused a person is in learning, naturally he will spend more time on learning and less time on playing. So we view RC1 as **Study Devotion Factor**

RC2 ⇐ **(-)Self-control Factor**: As it is significantly positively related to "goout", "Dalc", and "Walc" and negatively related to "health". Obviously, these are related to the degree of self indulgence. Due to the lack of self-management ability, one then often go out to play, indulge in alcohol and ignore health management. So we view RC2 as **(-)Self-control Factor**

Note that in practical use, what we take is the negative direction of this factor, that's why we add (-).

RC3 ⇐ **Family Education Background Factor**: As it is significantly positively related to "Medu" and "Fedu" and negatively related to "traveltime" and "failures". The first two factors are directly related to family education background. And according to our personal experience, families with high education level tend to have more strict tutoring and higher income level, so the latter two factors can also be explained. So we view RC3 as **Family Education Background Factor**

RC4 ⇐ **Family Harmony Factor**: As it is significantly positively related to "Pstatus" and "famrel". Obviously, this is determined by the harmony of family relations. So we view RC4 as **Family Harmony Factor**

RC5 ⇐ **Traffic Convenience Factor**: As it is significantly positively related to "traveltime" and "activities" and negatively related to "address". The more convenient the transportation, the shorter the natural commuting time and the easier it is to participate in extracurricular activities. Of course, this is also closely related to living in the city. So we view RC5 as **Traffic Convenience Factor**

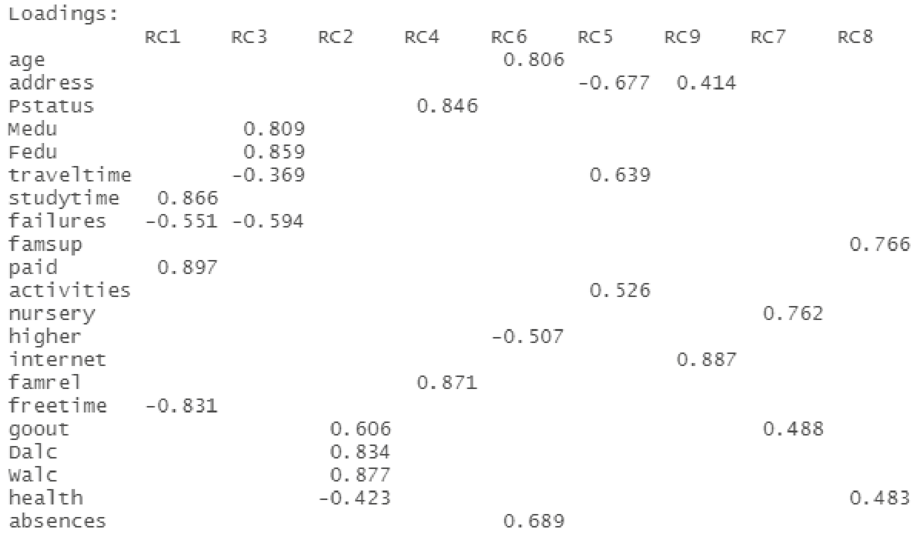
RC6 ⇐ **(-)Ambition Factor**: As it is significantly positively related to "age" and "absences" and negatively related to "higher". Lack of ambition can result in no pursuit of a higher degree and frequent absence. This is positively related to age, of course, because it is possible to repeat grades. So we view RC6 as **Ambition Factor**

Note that in practical use, what we take is the negative direction of this factor, that's why we add (-).

RC7 ⇐ **Competence Education Factor**: As it is significantly positively related to "nursery" and "goout". This is not an obvious explanation, but we can think that the more families attach importance to competence education, the more willing they are to let students receive preschool education and go out to play. So we view RC7 as **Competence Education Factor**

RC8 ⇐ **Family Care Factor**: As it is significantly positively related to "famsup" and "health". The more the family cares about the students, naturally the more it will provide them with family education and pay more attention to their health. So we view RC8 as **Family Care Factor**

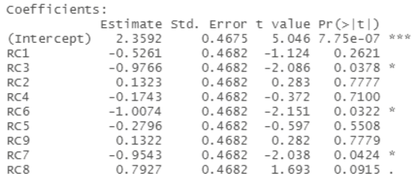
RC9 ⇐ **Modernization Factor**: As it is significantly positively related to "Internet" and relatively related to "adress", the mode modernized a family is, the more it will live in an urban and allow students to use internet. So we view RC9 as **Modernization Factor**



```
Loadings:
           RC1    RC3    RC2    RC4    RC6    RC5    RC9    RC7    RC8
age                                     0.806
address                                       -0.677  0.414
Pstatus                          0.846
Medu               0.809
Fedu               0.859
traveltime        -0.369                       0.639
studytime   0.866
failures   -0.551 -0.594
famsup                                                              0.766
paid        0.897
activities                                     0.526
nursery                                                      0.762
higher                                 -0.507
internet                                              0.887
famrel                           0.871
freetime   -0.831
goout                     0.606                              0.488
Dalc                      0.834
Walc                      0.877
health                   -0.423                                     0.483
absences                                0.689
```

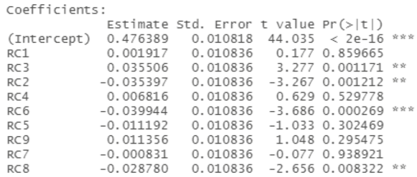
Figure 5: Rotated and Masked Result

4.3 Modeling by Linear Regression



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.3592     0.4675   5.046 7.75e-07 ***
RC1          -0.5261     0.4682  -1.124   0.2621    
RC3          -0.9766     0.4682  -2.086   0.0378 *  
RC2           0.1323     0.4682   0.283   0.7777    
RC4          -0.1743     0.4682  -0.372   0.7100    
RC6          -1.0074     0.4682  -2.151   0.0322 *  
RC5          -0.2796     0.4682  -0.597   0.5508    
RC9           0.1322     0.4682   0.282   0.7779    
RC7          -0.9543     0.4682  -2.038   0.0424 *  
RC8           0.7927     0.4682   1.693   0.0915 .  
```

Figure 6: LR for Grade



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.476389   0.010818  44.035  < 2e-16 ***
RC1          0.001917   0.010836   0.177 0.859665    
RC3          0.035506   0.010836   3.277 0.001171 ** 
RC2         -0.035397   0.010836  -3.267 0.001212 ** 
RC4          0.006816   0.010836   0.629 0.529778    
RC6         -0.039944   0.010836  -3.686 0.000269 ***
RC5         -0.011192   0.010836  -1.033 0.302469    
RC9          0.011356   0.010836   1.048 0.295475    
RC7         -0.000831   0.010836  -0.077 0.938921    
RC8         -0.028780   0.010836  -2.656 0.008322 ** 
```

Figure 7: LR for Progress

We call such LR coefficients as $Coef_{Grade}$ and $Coef_{Progress}$, and reorder the factors as the LR model does as $Factors$ (Actually that is Factors = Vector[Study Devotion, Family Education Background, (-)Self-control, Family Harmony, (-)Ambition, Traffic Convenience, Modernization, Competence Education, Family Care]), From such coefficients, our final model is:

$$
\begin{aligned}
Index_{Grade} &= {Coef_{Grade}}^{T} Factors \\
Judge_{Grade} &= \begin{cases} *Bad\ Student & Index < -0.1 \\ Normal\ Student & -0.1 < Index < 0.1 \\ Good\ Student & Index > 0.1 \end{cases} \\
Index_{Progress} &= {Coef_{Progress}}^{T} Factors \\
Judge_{Progress} &= \begin{cases} *Retrogressive\ Student & Index < -3 \\ Stable\ Student & -3 < Index < 3 \\ Progressive\ Student & Index > 3 \end{cases}
\end{aligned}
$$

Notice that we choose the threshold (±)0.1 and (±)3, This is no accident. Our modeling criterion is simple and effective, so we try some thresholds symmetrically in steps of (±)0.1, and finally find that the effect is the best when we take (±)0.1 and (±)3 respectively. We can see this from the box plots below.

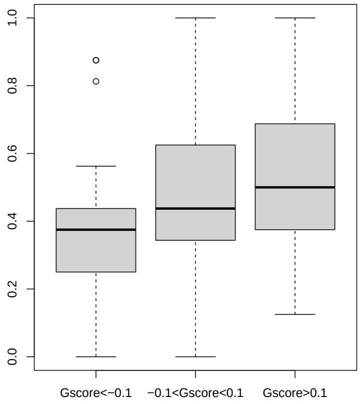


Figure 8: verification for Grade

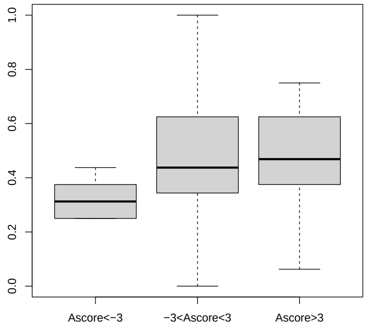


Figure 9: Verification for Progress

It can be seen from the figure that under the threshold given by us, there are differences between groups, especially for bad students and retrograding students. This shows that our model can well judge whether a student may be a bad student or a retrograding student according to his life state. For students with better performance, although there are differences, the difference is not obvious and can only be used as a reference.

Please note that the data set we use here is very subjective, which will weaken the generalization and interpretability of the model. When actually using this method, we should first get a more objective value of variables, Such as for “health”, we can use more strict standards, such as grading according to the number of diseases one has had before.

4.4 Cluster Analysis

We now do KNN clustering for the reduced factors by FA.

From figure.10, we can significantly indicate the students with bad and retrograding performance (the red points which lie mainly in the area with advance less than 0 and final score less than 0.6), which matches the result of our factor analysis and modeling that we can well judge whether a student may be a bad student or a retrograding student according to his life state.

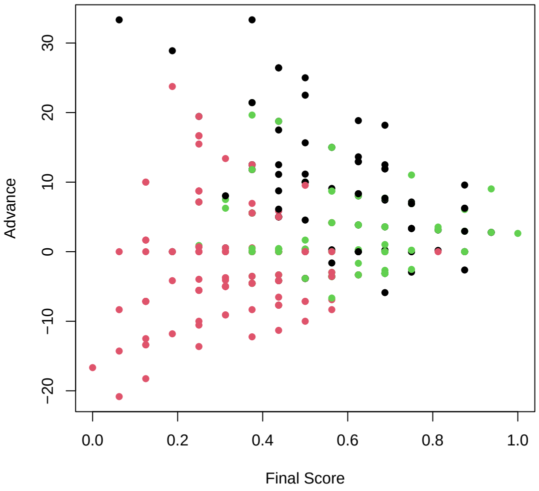


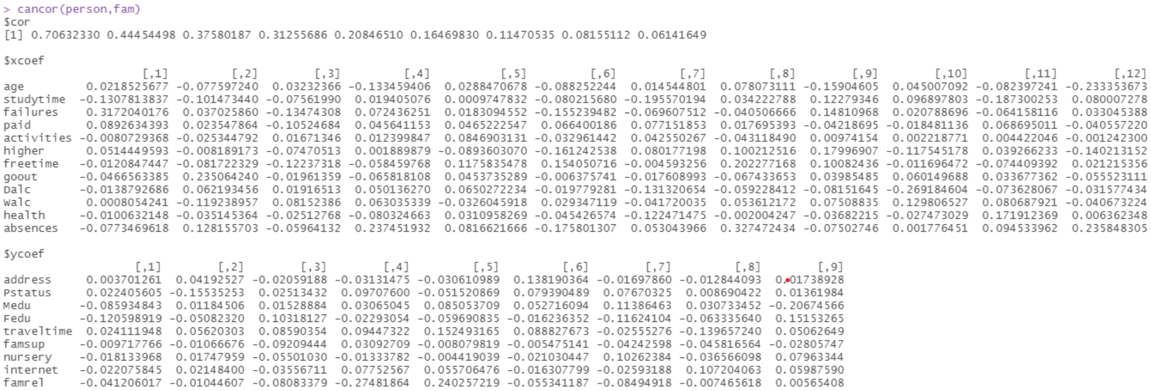
Figure 10: Cluster Analysis

4.5 Canonical Correlation Analysis

Through factor analysis, we found that some factors are related to students themselves, and some factors are related to families. We take these factors out separately, and put the variables most closely related to them (i.e., the variables not covered in the figure 5) into two groups, namely:

Family Variables Group: {**address**, **Pstatus**, **Medu**, **Fedu**, **traveltime**, **famsup**, **nursery**, **famrel**, **internet**}

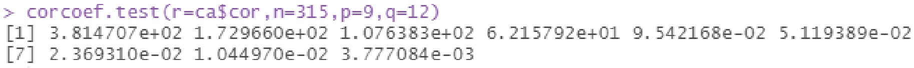
Student Status Variables Group: {**age**, **studytime**, **failures**, **paid**, **activities**, **higher**, **freetime**, **goout**, **Dalc**, **Walc**, **health**, **absences**}



```
> cancor(person,fam)
$cor
[1] 0.70632330 0.44454498 0.37580187 0.31255686 0.20846510 0.16469830 0.11470535 0.08155112 0.06141649

$xcoef
                    [,1]         [,2]        [,3]         [,4]          [,5]         [,6]         [,7]         [,8]        [,9]        [,10]        [,11]        [,12]
age         0.0218525677 -0.077597240  0.03232366 -0.133459406  0.0288470678 -0.088252244  0.014544801  0.078073111 -0.15904605  0.045007092 -0.082397241 -0.233353673
studytime  -0.1307813837 -0.101473440 -0.07561990  0.019405076  0.0009747832 -0.080215680 -0.195570194  0.034222788  0.12279346  0.096897803 -0.187300253  0.080007278
failures    0.3172040176  0.037025860 -0.13474308  0.072436251  0.0183094552 -0.155239482 -0.069607512 -0.040506666  0.14810968  0.020788696 -0.064158116  0.033045388
paid        0.0892634393  0.023547864 -0.10524684  0.045641153  0.0465222547  0.066400186  0.077151853  0.017695393 -0.04218695 -0.018481136  0.068695011 -0.040557220
activities -0.0080729368 -0.025344792  0.01671346  0.012399847  0.0846903131 -0.032961442  0.042550267 -0.043118490  0.00974154  0.002218771  0.004422046 -0.001242300
higher      0.0514449593 -0.008189173 -0.07470513  0.001889879 -0.0893603070 -0.161242538  0.080177198  0.100212516  0.17996907 -0.117545178  0.039266233 -0.140213152
freetime   -0.0120847447 -0.081722329 -0.12237318 -0.058459768  0.1175835478  0.154050716 -0.004593256  0.202277168  0.10082436 -0.011696472 -0.074409392  0.021215356
goout      -0.0466563385  0.235064240 -0.01961359 -0.065818108  0.0453735289 -0.006375741 -0.017608993 -0.067433653  0.03985485  0.060149688  0.033677362 -0.055523111
Dalc       -0.0138792686  0.062193456  0.01916513  0.050136270  0.0650272234 -0.019779281 -0.131320654 -0.059228412 -0.08151645 -0.269184604 -0.073628067 -0.031577434
Walc        0.0008054241 -0.119238957  0.08152386  0.063035339 -0.0326045918  0.029347119 -0.041720035  0.053612172  0.07508835  0.129806527  0.080687921 -0.040673224
health     -0.0100632148 -0.035145364 -0.02512768 -0.080324663  0.0310958269 -0.045426574 -0.122471475 -0.002004247 -0.03682215 -0.027473029  0.171912369  0.006362348
absences   -0.0773469618  0.128155703 -0.05964132  0.237451932  0.0816621666 -0.175801307  0.053043966  0.327472434 -0.07502746  0.001776451  0.094533962  0.235848305

$ycoef
                   [,1]        [,2]        [,3]        [,4]         [,5]         [,6]        [,7]         [,8]        [,9]
address     0.003701261  0.04192527 -0.02059188 -0.03131475 -0.030610989  0.138190364 -0.01697860 -0.012844093  0.01738928
Pstatus     0.022405605 -0.15535253  0.02513432  0.09707600 -0.051520869  0.079390489  0.07670325  0.008690422  0.01361984
Medu       -0.085934843  0.01184506  0.01528884  0.03065045  0.085053709  0.052716094  0.11386463  0.030733452 -0.20674566
Fedu       -0.120598919 -0.05082320  0.10318127 -0.02293054 -0.059690835 -0.016236352 -0.11624104 -0.063335640  0.15153265
traveltime  0.024111948  0.05620303  0.08590354  0.09447322  0.152493165  0.088827673 -0.02555276 -0.139657240  0.05062649
famsup     -0.009717766 -0.01066676 -0.09209444  0.03092709 -0.008079819 -0.005475141 -0.04242598 -0.045816564 -0.02805747
nursery    -0.018133968  0.01747959 -0.05501030 -0.01333782 -0.004419039 -0.021030447  0.10262384 -0.036566098  0.07963344
internet   -0.022075845  0.02148400 -0.03556711  0.07752567  0.055706476 -0.016307799 -0.02593188  0.107204063  0.05987590
famrel     -0.041206017 -0.01044607 -0.08083379 -0.27481864  0.240257219 -0.055341187 -0.08494918 -0.007465618  0.00565408
```

Figure 11: Canonical Correlation Analysis

We then test for the significance [11]:



```
> corcoef.test(r=ca$cor,n=315,p=9,q=12)
[1] 3.814707e+02 1.729660e+02 1.076383e+02 6.215792e+01 9.542168e-02 5.119389e-02
[7] 2.369310e-02 1.044970e-02 3.777084e-03
```

Figure 12: Canonical Correlation Analysis

Thus we can take the first 3 pairs of canonical variables for analysis for they are significant.

And through CCA, we found that the correlation coefficient between the first pair of canonical variables was more than 70%, indicating that family factors would also lead to changes in students' individual learning status, and the correlation was quite strong.

5 Conclusion

5.1 Research Results

In this research, we mainly analyze the data set with 382 students that contains 30 attributes of each student and 3 variables that indicate the math course performance of each student.

By factor analysis, 21 variables of this data set can be summarized into 9 factors that are Study Devotion Factor, Self-control Factor, Family Education Background Factor, Family Harmony Factor, Traffic Convenience Factor, Ambition Factor, Competence Education Factor, Family Care Factor and Modernization Factor with a degree of explanation go beyond 70%. Also, This result can be verified well by the actual performance and cluster analysis.

At last, if we divide the 9 factors into two groups representing family variables and student status variables respectively, by canonical correlation analysis we can conclude that family factors would also lead to changes in students' individual learning status.

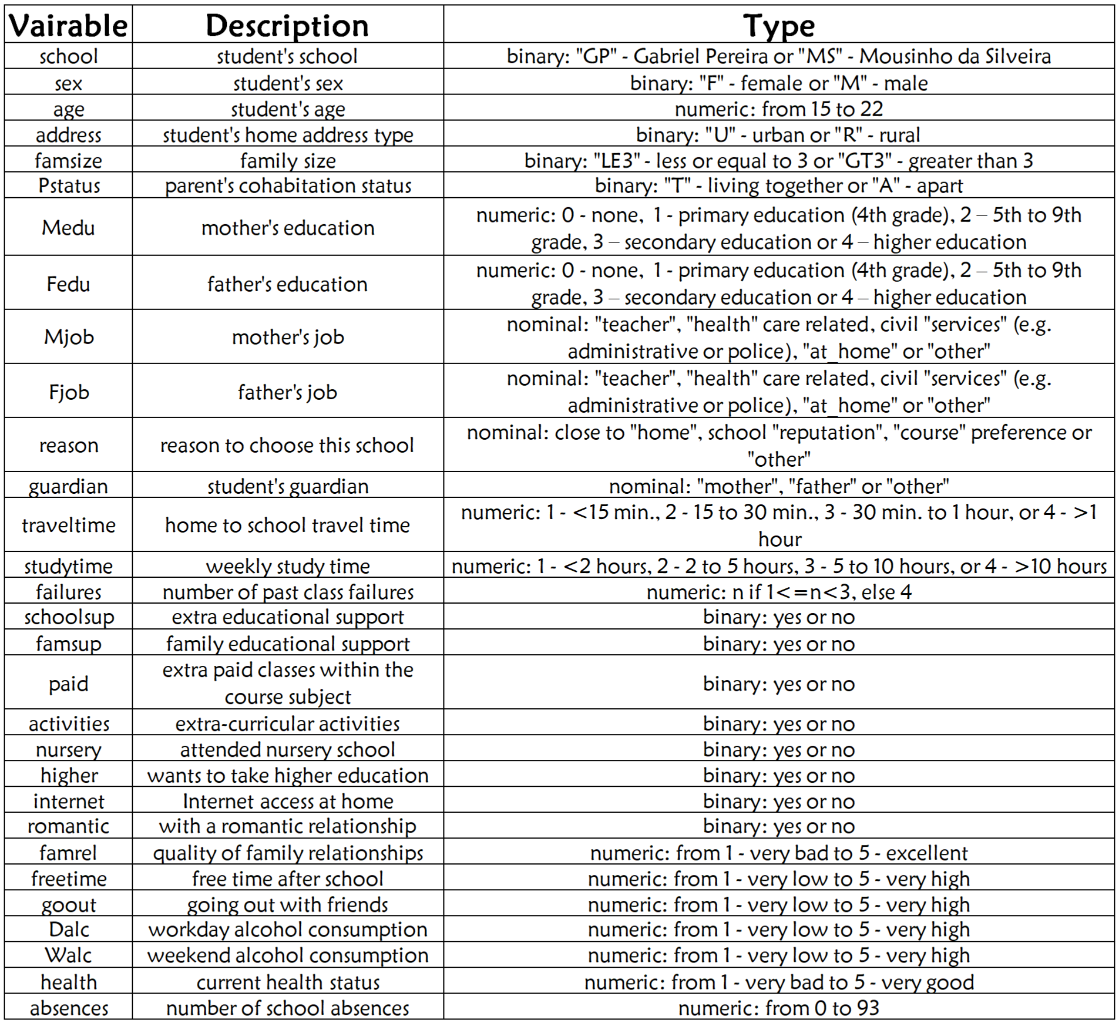
Therefore, we educators can predict which students are likely to have academic problems according to their usual performance and family conditions. On the other hand, it also inspires us that family has a great influence on students' study, which has been proved by many researches not only ours [12]. Parents should pay attention to providing better learning conditions for their children and pay more attention to students' learning status. Of course, students can also use our methods to test and adjust themselves.

5.2 Further Thinking

From research above, we can see that there is still progress in the accuracy of criteria we have proposed to predict the students performance since our evaluation can be only accurate when it comes to judging whether a student will do badly and the criteria perform poorly in distinguishing that whether a student will do well. Besides, there are still more attributes need to be considered such as friendship, class they have taken or their previous performance in school.

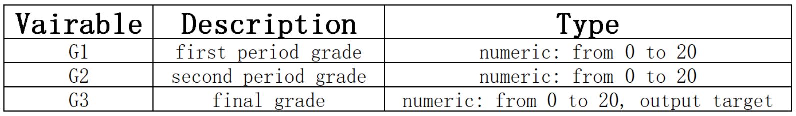
On the other hand, our model learning is actually affected a lot by categorical variables [13], and if we can replace some of these variables with continuous variables, or at least integer variables, obviously the analysis results will be better. This is certainly feasible. For example, for the variable "freetime", we can track the specific time of a student, at least specify to the hour. This will certainly make data collection more difficult, but it would be an effective and meaningful direction of improvement.

A Appendix



Vairable	Description	Type
school	student's school	binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira
sex	student's sex	binary: "F" - female or "M" - male
age	student's age	numeric: from 15 to 22
address	student's home address type	binary: "U" - urban or "R" - rural
famsize	family size	binary: "LE3" - less or equal to 3 or "GT3" - greater than 3
Pstatus	parent's cohabitation status	binary: "T" - living together or "A" - apart
Medu	mother's education	numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education
Fedu	father's education	numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education
Mjob	mother's job	nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other"
Fjob	father's job	nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other"
reason	reason to choose this school	nominal: close to "home", school "reputation", "course" preference or "other"
guardian	student's guardian	nominal: "mother", "father" or "other"
traveltime	home to school travel time	numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour
studytime	weekly study time	numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours
failures	number of past class failures	numeric: n if 1<=n<3, else 4
schoolsup	extra educational support	binary: yes or no
famsup	family educational support	binary: yes or no
paid	extra paid classes within the course subject	binary: yes or no
activities	extra-curricular activities	binary: yes or no
nursery	attended nursery school	binary: yes or no
higher	wants to take higher education	binary: yes or no
internet	Internet access at home	binary: yes or no
romantic	with a romantic relationship	binary: yes or no
famrel	quality of family relationships	numeric: from 1 - very bad to 5 - excellent
freetime	free time after school	numeric: from 1 - very low to 5 - very high
goout	going out with friends	numeric: from 1 - very low to 5 - very high
Dalc	workday alcohol consumption	numeric: from 1 - very low to 5 - very high
Walc	weekend alcohol consumption	numeric: from 1 - very low to 5 - very high
health	current health status	numeric: from 1 - very bad to 5 - very good
absences	number of school absences	numeric: from 0 to 93

Figure 13: Raw variable list

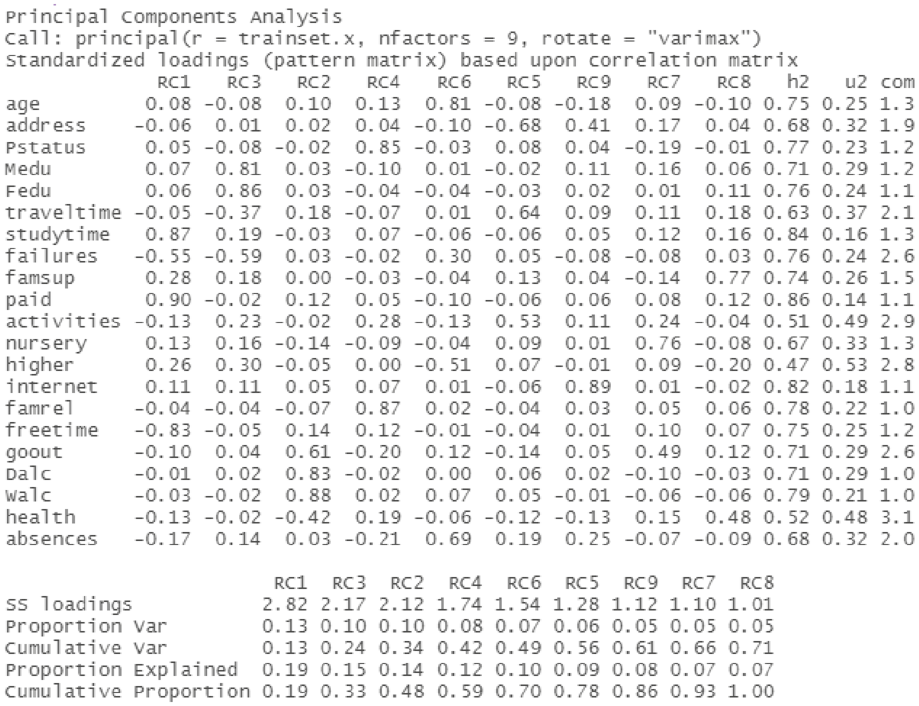


Vairable	Description	Type
G1	first period grade	numeric: from 0 to 20
G2	second period grade	numeric: from 0 to 20
G3	final grade	numeric: from 0 to 20, output target

Figure 14: Raw performance variable list

B Appendix

The Full FA Result



```
Principal Components Analysis
Call: principal(r = trainset.x, nfactors = 9, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
             RC1   RC3   RC2   RC4   RC6   RC5   RC9   RC7   RC8   h2   u2 com
age         0.08 -0.08  0.10  0.13  0.81 -0.08 -0.18  0.09 -0.10 0.75 0.25 1.3
address    -0.06  0.01  0.02  0.04 -0.10 -0.68  0.41  0.17  0.04 0.68 0.32 1.9
Pstatus     0.05 -0.08 -0.02  0.85 -0.03  0.08  0.04 -0.19 -0.01 0.77 0.23 1.2
Medu        0.07  0.81  0.03 -0.10  0.01 -0.02  0.11  0.16  0.06 0.71 0.29 1.2
Fedu        0.06  0.86  0.03 -0.04 -0.04 -0.03  0.02  0.01  0.11 0.76 0.24 1.1
traveltime -0.05 -0.37  0.18 -0.07  0.01  0.64  0.09  0.11  0.18 0.63 0.37 2.1
studytime   0.87  0.19 -0.03  0.07 -0.06 -0.06  0.05  0.12  0.16 0.84 0.16 1.3
failures   -0.55 -0.59  0.03 -0.02  0.30  0.05 -0.08 -0.08  0.03 0.76 0.24 2.6
famsup      0.28  0.18  0.00 -0.03 -0.04  0.13  0.04 -0.14  0.77 0.74 0.26 1.5
paid        0.90 -0.02  0.12  0.05 -0.10 -0.06  0.06  0.08  0.12 0.86 0.14 1.1
activities -0.13  0.23 -0.02  0.28 -0.13  0.53  0.11  0.24 -0.04 0.51 0.49 2.9
nursery     0.13  0.16 -0.14 -0.09 -0.04  0.09  0.01  0.76 -0.08 0.67 0.33 1.3
higher      0.26  0.30 -0.05  0.00 -0.51  0.07 -0.01  0.09 -0.20 0.47 0.53 2.8
internet    0.11  0.11  0.05  0.07  0.01 -0.06  0.89  0.01 -0.02 0.82 0.18 1.1
famrel     -0.04 -0.04 -0.07  0.87  0.02 -0.04  0.03  0.05  0.06 0.78 0.22 1.0
freetime   -0.83 -0.05  0.14  0.12 -0.01 -0.04  0.01  0.10  0.07 0.75 0.25 1.2
goout      -0.10  0.04  0.61 -0.20  0.12 -0.14  0.05  0.49  0.12 0.71 0.29 2.6
Dalc       -0.01  0.02  0.83 -0.02  0.00  0.06  0.02 -0.10 -0.03 0.71 0.29 1.0
Walc       -0.03 -0.02  0.88  0.02  0.07  0.05 -0.01 -0.06 -0.06 0.79 0.21 1.0
health     -0.13 -0.02 -0.42  0.19 -0.06 -0.12 -0.13  0.15  0.48 0.52 0.48 3.1
absences   -0.17  0.14  0.03 -0.21  0.69  0.19  0.25 -0.07 -0.09 0.68 0.32 2.0

                       RC1  RC3  RC2  RC4  RC6  RC5  RC9  RC7  RC8
SS loadings           2.82 2.17 2.12 1.74 1.54 1.28 1.12 1.10 1.01
Proportion Var        0.13 0.10 0.10 0.08 0.07 0.06 0.05 0.05 0.05
Cumulative Var        0.13 0.24 0.34 0.42 0.49 0.56 0.61 0.66 0.71
Proportion Explained  0.19 0.15 0.14 0.12 0.10 0.09 0.08 0.07 0.07
Cumulative Proportion 0.19 0.33 0.48 0.59 0.70 0.78 0.86 0.93 1.00
```

Figure 15: Full FA Result

C Appendix

Dataset: https://www.kaggle.com/datasets/larsen0966/student-performance-data-set
Codes: https://github.com/ShimizuYukii/Project_Codes-2022Spring-

References

[1] S. Huang and N. Fang, "Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models," *Computers & Education*, vol. 61, pp. 133–145, 2013.

[2] J. M. A. Navamani and A. Kannammal, "Predicting performance of schools by applying data mining techniques on public examination results," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 9, no. 4, pp. 262–271, 2015.

[3] H. Bydžovská, "Student performance prediction using collaborative filtering methods," in *International Conference on Artificial Intelligence in Education*, pp. 550–553, Springer, 2015.

[4] Š. Pero and T. Horváth, "Comparison of collaborative-filtering techniques for small-scale student performance prediction task," in *Innovations and Advances in Computing, Informatics, Systems Sciences, Networking and Engineering*, pp. 111–116, Springer, 2015.

[5] N. Thai-Nghe, L. Drumond, A. Krohn-Grimberghe, and L. Schmidt-Thieme, "Recommender system for predicting student performance," *Procedia Computer Science*, vol. 1, no. 2, pp. 2811–2819, 2010.

[6] C.-J. Villagrá-Arnedo, F. J. Gallego-Durán, P. Compañ, F. Llorens Largo, R. Molina-Carmona, *et al.*, "Predicting academic performance from behavioural and learning data," 2016.

[7] Y. Park, "Predicting personalized student performance in computing-related majors via collaborative filtering," in *Proceedings of the 19th Annual SIG Conference on Information Technology Education*, pp. 151–151, 2018.

[8] B. K. Nkansah, "On the kaiser-meier-olkin's measure of sampling adequacy," *Mathematical Theory and Modeling*, vol. 8, no. 7, 2018.

[9] M. S. Bartlett, "A note on the multiplying factors for various $\chi 2$ approximations," *Journal of the royal statistical society series b-methodological*, vol. 16, pp. 296–298, 1954.

[10] R. B. Cattell, "Factor analysis: an introduction and manual for the psychologist and social scientist.," 1952.

[11] Tiaaaaa, "Canonical correlation coefficient significance test by r." `https://blog.csdn.net/tiaaaaa/article/details/58137522`, 2019.

[12] L. Wößmann, "Educational production in east asia: The impact of family background and schooling policies on student performance," *German Economic Review*, vol. 6, no. 3, pp. 331–353, 2005.

[13] R. J. Mislevy, "Recent developments in the factor analysis of categorical variables," *Journal of educational statistics*, vol. 11, no. 1, pp. 3–31, 1986.