Machine Learning and Artificial Intelligence
A.J. Tallón-Ballesteros and C. Chen (Eds.)

doi:10.3233/FAIA200801

Detection of Specific Components in a PCA Mixture

Jin XIE [a,1], Zian ZHENG[b] and Jian GAO[c]
[a]*School of Mathematics and Statistics, Lanzhou University, China*
[b]*School of Information Science and Engineering, Lanzhou University, China*
[c]*College of Chemistry and Chemical Engineering, Lanzhou University, China*

Abstract. Taking a given mixture as an example, 25,000 samples were selected for the detection of 7 indicators. Firstly, the correlation between each indicator and the test result is analyzed, The T test is used to identify the main indicators that can be used to determine the existence of a specific component. Secondly, three comprehensive indexes are obtained by combining PCA. Determine whether there are specific components in the unknown mixture.

Keywords. T test, PCA, mixture, detection

1. Introduction

Judging the composition structure of mixtures is an important step in the field of physical chemistry. When facing a huge amount of information, the correlation can be analyzed through known samples and a mathematical model can be built. Principal component analysis (PCA) is a kind of multivariate analysis method, which transforms the original correlated variables into some new unrelated variables by means of variable transformation. The method is scientific and effective, and can be widely applied in many fields. For the detection of specific components in the mixture, generally there are chemical methods, professional formula analysis based on microspectrum technology, physical purification method, DAD detection and model analysis algorithm enhance the computer processing capacity and realize the qualitative identification of the mixture detection. Using mathematical modeling theory and method to integrate the detection of specific components of the mixture, applying mathematical knowledge to abstract the calculation problem in the experiment into a mathematical problem and summarized into mathematical models. Through calculation to determine whether the unknown mixture contains specific components, and get the results. Judging the detection of the composition of mixtures is an important step worth discussing in the field of physical chemistry. In the face of a huge amount of information, the correlation of known samples can be analyzed and a mathematical model can be built. The model can be used to judge whether the unknown samples contain specific components, with scientific and effective methods.

[1] Corresponding Author: Xie Jin; E-mail: xijiner0211@163.com

In this paper, 25,000 samples given in the Mathematical Modeling Contest of Lanzhou University in 2020 are taken as an example ,and 7 indicators are known to be carried out on the samples of this mixture (denotable as V1,V2...,V7). The "training data" contained 20,000 samples of the mixture, with the mixture known to contain a specific ingredient, and the "test data" contained 5,000 samples of the mixture, with the mixture unknown to contain a specific ingredient. Through correlation analysis and PCA model, the main indexes for determining the existence of specific components are given.

2. Basic Theory

（1）T Test:

H0 (null hypothesis): attribute A and B are independent of each other, $\overline{x_1} = \overline{x_2}$;
H1 (alternative hypothesis): attribute A is related to B, $\overline{x_1} \neq \overline{x_2}$. To test whether the hypothesis is true, the sample data is processed as follows [1].

Independence test:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{{s_1}^2}{N_1} + \frac{{s_2}^2}{N_2}}}$$

If we assume that the variance of these two samples is the same

$$t = \frac{\overline{x_1} - \overline{x_2}}{s_p\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \qquad s_p = \frac{(N_1-1)S_1^2+(N_2-1)S_2^2}{N_1+N_2-2}$$

Inspection level: $\alpha = 0.05$

（2）PCA

Principal component analysis (PCA) is a multivariate statistical analysis method that transforms multiple indicators into several comprehensive indicators by dimension reduction. The main objective is to explain most of the information in the original data with fewer variables.[2]

First, standardize the indicator data: $x'_{ij} = \frac{x_{ij}-\overline{x_j}}{s_j}, i = 1,2, \ldots m , \overline{x_j}$ is mean, s_j is standard deviation. Secondly, the correlation coefficient matrix and covariance matrix are calculated: $r_{ij} = \frac{\sum_{k=1}^{n}(x_{ki}-\overline{x_i})\cdot(x_{ki}-\overline{x_j})}{\sqrt{\sum_{k=1}^{n}(x_{ki}-\overline{x_i})\cdot\sum_{k=1}^{n}(x_{kj}-\overline{x_j})^2}}, c_{ij} = \frac{\sum_{k=1}^{n}(x_{ki}-\overline{x_i})\cdot(x_{ki}-\overline{x_j})}{m-1}$

After the covariance matrix is obtained, KMO (Kaiser-Meyer-Olkin) and Bartlett ball type test are conducted to determine whether it is suitable for principal component analysis. Principal component analysis is only suitable when the KMO test value is > 0.5 and the Bartlett test value is < 0.05.

Thirdly, calculate the eigenvalue and variance contribution rate:

$$G_i = \lambda_i / \sum_{i=1}^{n} \lambda_i \quad , TG_i = \sum_{i=1}^{s} \lambda_i / \sum_{i=1}^{n} \lambda_i$$

Finally, the principal component was extracted and the comprehensive evaluation was carried out, and the comprehensive index F was calculated and evaluated:

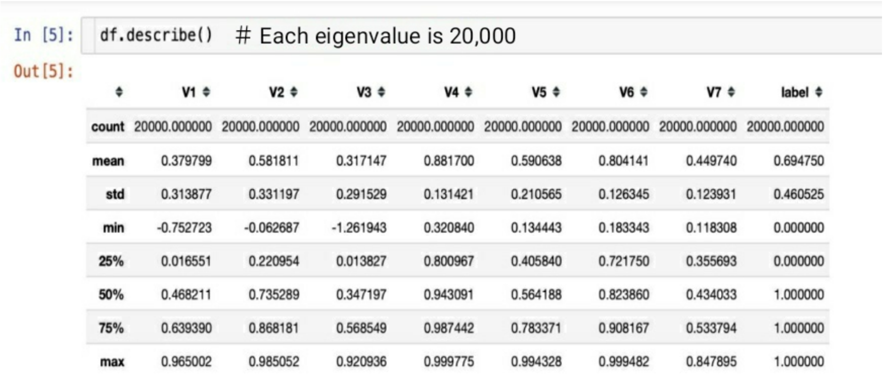
$$F = \sum_{i=1}^{s} \alpha_i F_i \quad \alpha_i = \lambda_i / \sum_{i=1}^{n} \lambda_i$$

3. The Process Of Problem Solving

3.1 Missing Value Detection

First, descriptive analysis of the data was carried out. Visualizing the data through Python lists the number, maximum, minimum, mean, standard deviation, and 4 quantile of each variable, as shown in Table 1.It can be seen from the table that there is no missing value in 20,000 data, so there is no need to fill in the missing value.

Table 1 Data indexing



```
In [5]: df.describe()   # Each eigenvalue is 20,000
Out[5]:
```

	V1	V2	V3	V4	V5	V6	V7	label
count	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000
mean	0.379799	0.581811	0.317147	0.881700	0.590638	0.804141	0.449740	0.694750
std	0.313877	0.331197	0.291529	0.131421	0.210565	0.126345	0.123931	0.460525
min	-0.752723	-0.062687	-1.261943	0.320840	0.134443	0.183343	0.118308	0.000000
25%	0.016551	0.220954	0.013827	0.800967	0.405840	0.721750	0.355693	0.000000
50%	0.468211	0.735289	0.347197	0.943091	0.564188	0.823860	0.434033	1.000000
75%	0.639390	0.868181	0.568549	0.987442	0.783371	0.908167	0.533794	1.000000
max	0.965002	0.985052	0.920936	0.999775	0.994328	0.999482	0.847895	1.000000

3.2 Outlier Handling

Outliers are detected by box diagram, in which outliers are usually defined as values less than Q_L-1.5QR or greater than Q_U+1.5IQR, as shown in Figure 1. It can be seen from the figure that there are some outliers in V3, V4, V6 and V7. A total of 417 outliers were calculated. As the sample size was large enough, the outliers were deleted directly. After processing, see Figure 2.

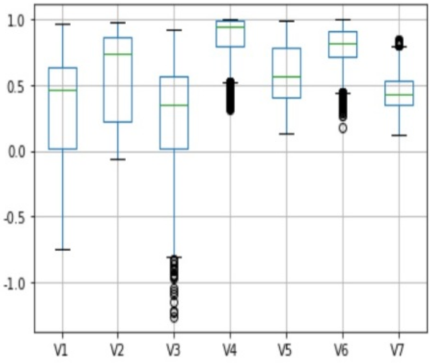


Figure 1 Data outliers

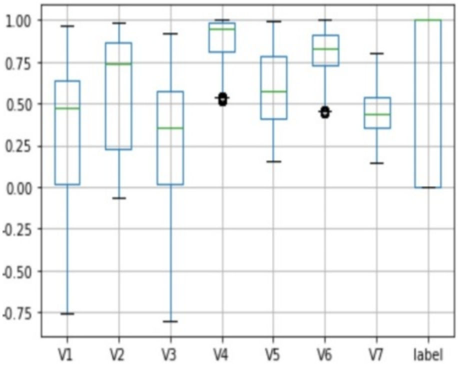
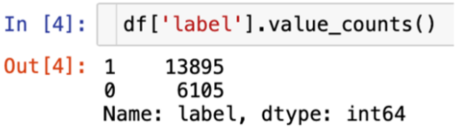


Figure 2 Delete the outliers

3.3 Category Equalization Check

In the classification task, the imbalance of sample categories will have a great impact on the training of the model, so the proportion of positive and negative samples should be checked during the training of the model.



```
In [4]:  df['label'].value_counts()

Out[4]: 1    13895
        0     6105
        Name: label, dtype: int64
```

There were many samples of category 1, and the ratio of 1 to 0 was 2.27:1. There was a certain imbalance in the samples. Therefore, measures should be taken from data sampling, model selection, and algorithm evaluation criteria to reduce the impact of sample imbalance on the model. Instead of random sampling, stratified sampling is used to divide the training set and test set, so that the data of the divided training set and test set have nearly consistent distribution with the overall data in each feature.

4. Model and Solution Method

4.1 T-test Model

First, standardize the data: [3]

$$x = \frac{x-\min(x)}{\max(x)-\min(x)}$$

See Annex 1 for the distribution map, V1 in Figure 3 shows the frequency distribution of indicator V1 for y equals 0 (blue) and 1 (yellow). Considering that the randomized trial is influenced by several random factors that are independent of each other, but it is difficult to determine the leading indicators. Under the condition that the sample size is large enough, the observed values of the random experiment approximately conform to the normal distribution. When one variable is a categorical variable and the other is a continuous variable, we choose to use t test for correlation analysis. The variables significantly correlated in the T test were taken as the main influencing variable.

(1) **Data Visualization**

Using Python to draw the bar chart of each variable, and some of the figures are shown as follows (see Figure 3). It can be seen that it's approximately normally distributed. SPSS was used to conduct normal test respectively, and the data was normalized by its standard deviation.

Let's look at the first line, Sig=significance, if Sig>0.05, the first line shall prevail.

Otherwise, the second row shall prevail. If Sig (bilateral)<0.05, it represents rejection of the null hypothesis and correlation of the variable. Otherwise it's irrelevant. T-test can compare the differences between the two groups of data and combine the analysis results of independent sample test, as shown in the Table 2 and Table 3:

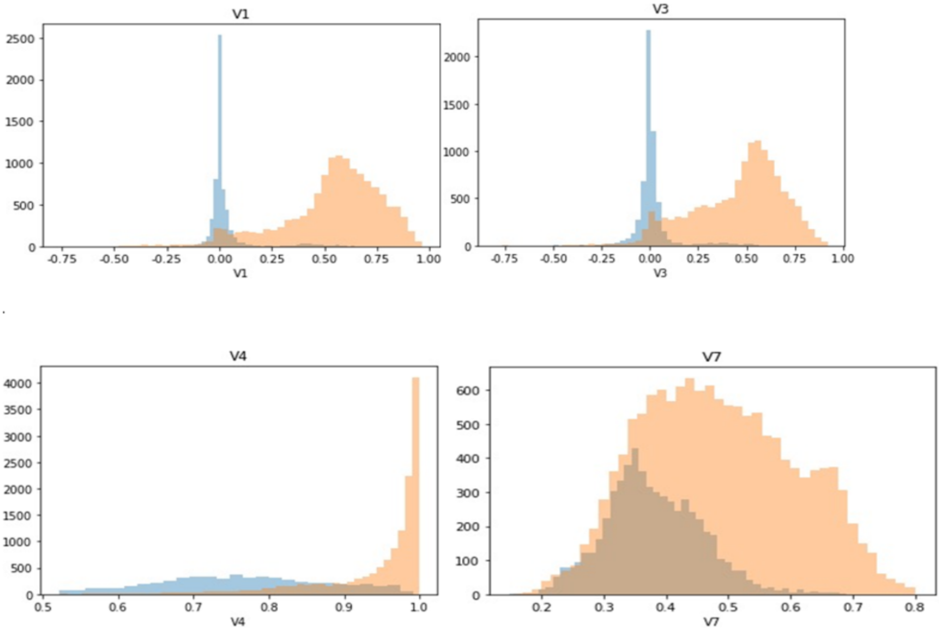
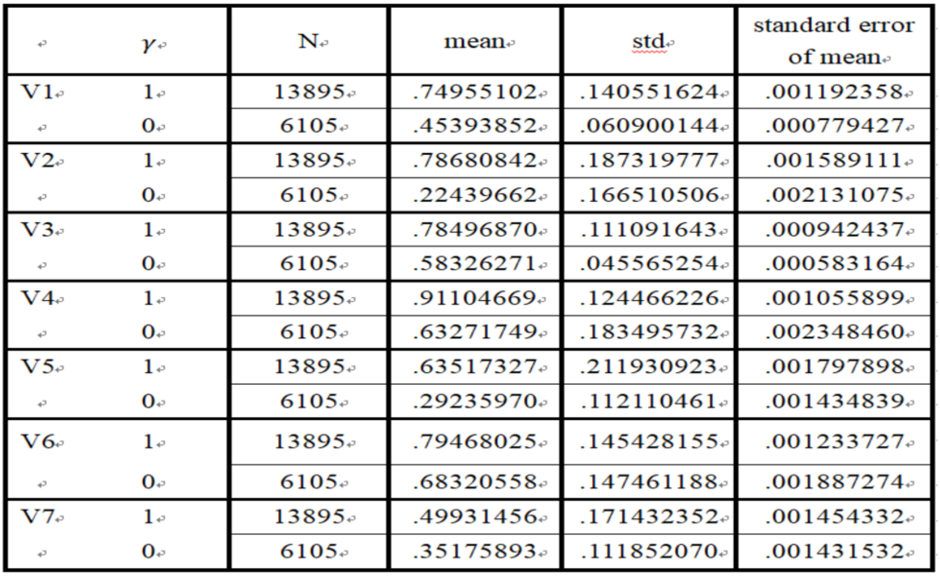


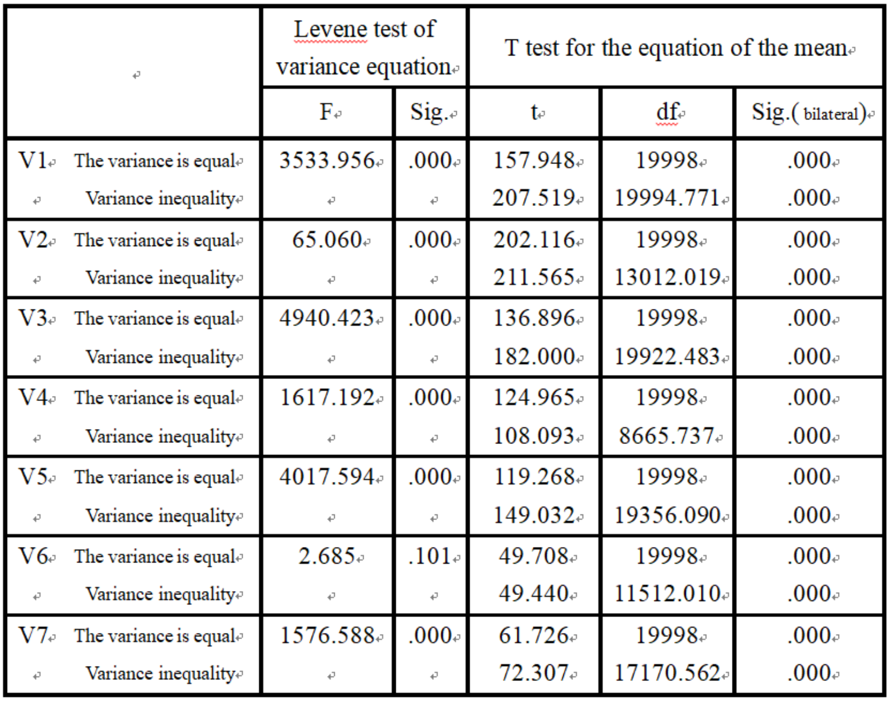
Figure 3 A distribution pattern of variables

Table 2 The set of statistics



	γ	N	mean	std	standard error of mean
V1	1	13895	.74955102	.140551624	.001192358
	0	6105	.45393852	.060900144	.000779427
V2	1	13895	.78680842	.187319777	.001589111
	0	6105	.22439662	.166510506	.002131075
V3	1	13895	.78496870	.111091643	.000942437
	0	6105	.58326271	.045565254	.000583164
V4	1	13895	.91104669	.124466226	.001055899
	0	6105	.63271749	.183495732	.002348460
V5	1	13895	.63517327	.211930923	.001797898
	0	6105	.29235970	.112110461	.001434839
V6	1	13895	.79468025	.145428155	.001233727
	0	6105	.68320558	.147461188	.001887274
V7	1	13895	.49931456	.171432352	.001454332
	0	6105	.35175893	.111852070	.001431532

Table 3 Independent sample test



		Levene test of variance equation		T test for the equation of the mean		
		F	Sig.	t	df	Sig.(bilateral)
V1	The variance is equal	3533.956	.000	157.948	19998	.000
	Variance inequality			207.519	19994.771	.000
V2	The variance is equal	65.060	.000	202.116	19998	.000
	Variance inequality			211.565	13012.019	.000
V3	The variance is equal	4940.423	.000	136.896	19998	.000
	Variance inequality			182.000	19922.483	.000
V4	The variance is equal	1617.192	.000	124.965	19998	.000
	Variance inequality			108.093	8665.737	.000
V5	The variance is equal	4017.594	.000	119.268	19998	.000
	Variance inequality			149.032	19356.090	.000
V6	The variance is equal	2.685	.101	49.708	19998	.000
	Variance inequality			49.440	11512.010	.000
V7	The variance is equal	1576.588	.000	61.726	19998	.000
	Variance inequality			72.307	17170.562	.000

(2) **Analysis Conclusion**

Because the Sig values of the seven analysis items in the figure were all less than 0.05, indicating a significant difference, and the seven variables V1--V7 were all correlated. According to the comparison of the mean absolute values of the group statistics, the larger the difference was, the more correlated it was. If the absolute value is greater than 0.2, it indicates that there is a strong correlation between the two which is the main indicator. Thus, v1-V5 is the main indicator, while V6 and V7 are the secondary indicators.

(3) **Test of Conclusion**

Python's Pandas library is used to calculate the correlation coefficients for each indicator (V1-V7) and category (whether there is a specific indicator) and draw the correlation coefficient matrix, as shown in Figure 4.

It can be seen that the correlation coefficient of V1-V5 and label (whether containing a specific indicator) is above 0.6, which can confirm the conclusion obtained by T test that V1-V5 has a strong correlation with a specific component and is the main indicator, while the correlation of V6 and V7 is slightly weak.

4.2 PCA

Since the data has been standardized during data preprocessing, NumPY library directly calculates the standardized covariance matrix, and the visualization result of the covariance matrix is shown in Figure 5. KMO(Kaiser-Meyer-Olkin) and Bartlett ball tests were then performed on the covariance matrix. The KMO test value is 1.948 and the Bartlett test value is 1.07×10^{-12} by using the custom function. So there is a correlation between the data and principal component.

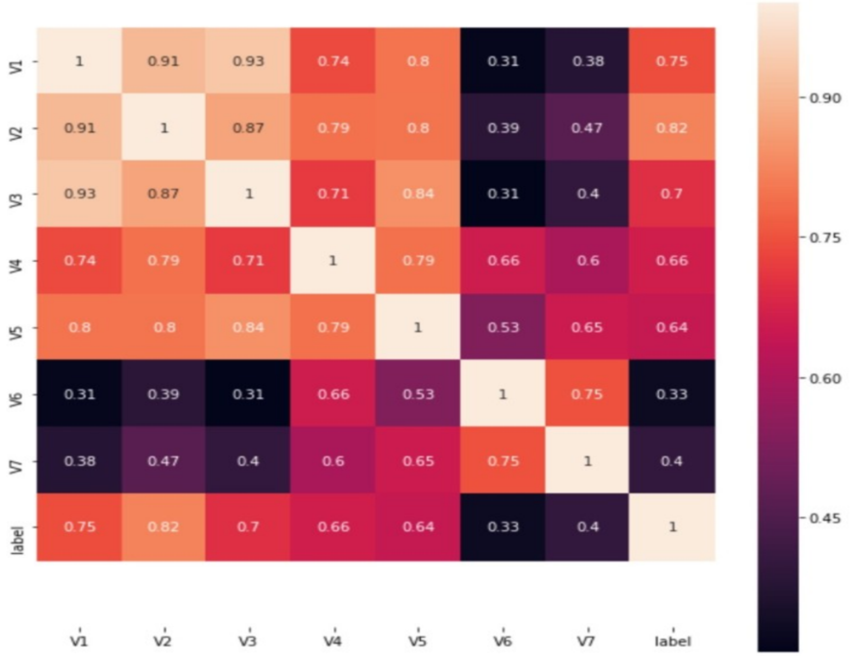


Figure 4 Correlation coefficient matrix

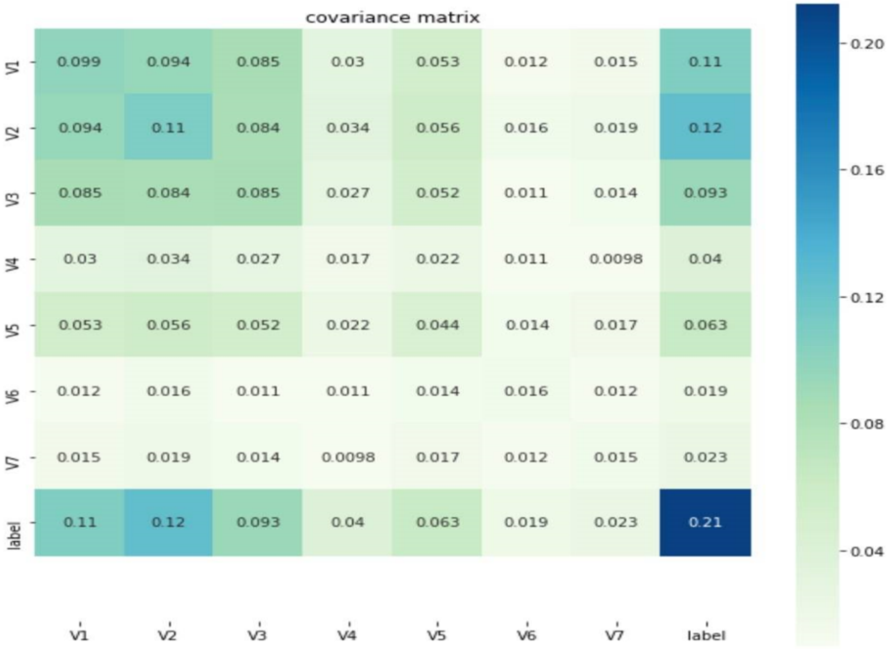


Figure 5 Covariance matrix

Step 3 and step 4 complete by calling the Sklearn library's Decomposition module. Figure 7 shows the edge information of each new variable that the principal component can create. It can be seen that the new variable f1 contains 70.6% of the original 7 variables, f2 contains 17.6% of the original 7 variables, f3 contains 4.6%, and so on.

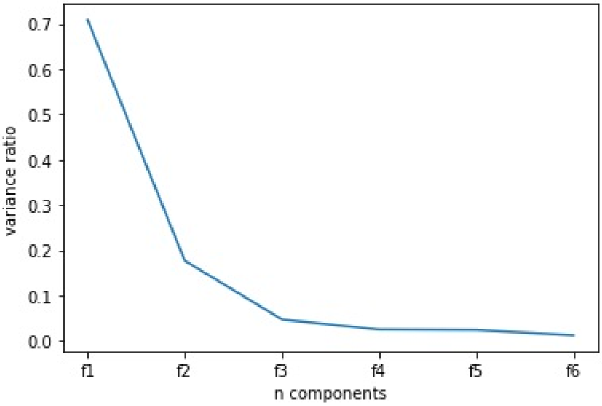


Figure 6 The edge information of the new variable in PCA

For the selection of the number of new variables, refer to the "Elbow rule", which is to find the obvious inflection point in the graph, it is easy to find the elbow position in figure 6 is at f3, that is, the use of f1, f2, and f3 variables can retain most of the information of the data set (93.2%).

Therefore, the first three features are considered as potential major components, and their relationship with the original V1-V7 is as follows:

F1 = -0.40043*V1 -0.41*V2 -0.40339*V3 -0.4074*V4 -0.41684*V5 -0.27661*V6 -0.30428*V7

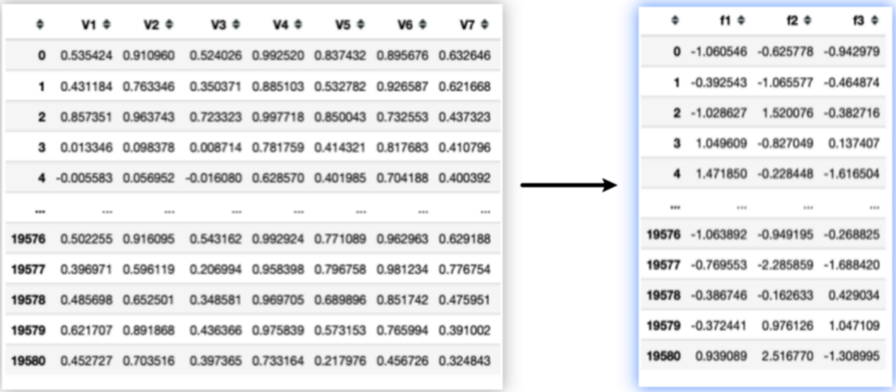
F2 = 0.34374*V1 + 0.24579*V2 + 0.32363*V3 -0.07159*V4 + 0.01153*V5 -0.6345* V6 -0.55574*V7

F3 = 0.04547*V1 + 0.08565*V2 -0.13065*V3 + 0.49405*V4 -0.3449*V5+0.47352* V6 -0.6215*V7

As can be seen from the principal component formula, the indexes closely related to each principal component are V1, V2, V3 and V4.V5, V6 and V7 are secondary indicators. This result conforms to the correlation test obtained in Figure 4.

5. Result

In order to verify whether the three new variables F1, F2 and F3 are potential principal components, we used F1, F2 and F3 as features to train the SVM model, predict whether there are specific components, and compare the accuracy obtained with the SVM model trained by using V1-V7 as features.



	V1	V2	V3	V4	V5	V6	V7
0	0.535424	0.910960	0.524026	0.992520	0.837432	0.895676	0.632646
1	0.431184	0.763346	0.350371	0.885103	0.532782	0.926587	0.621668
2	0.857351	0.963743	0.723323	0.997718	0.850043	0.732553	0.437323
3	0.013346	0.098378	0.008714	0.781759	0.414321	0.817683	0.410796
4	-0.005583	0.056952	-0.016080	0.628570	0.401985	0.704188	0.400392
...	...	...	...	...	...	...	...
19576	0.502255	0.916095	0.543162	0.992924	0.771089	0.962963	0.629188
19577	0.396971	0.596119	0.206994	0.958398	0.796758	0.981234	0.776754
19578	0.485698	0.652501	0.348581	0.969705	0.689896	0.851742	0.475951
19579	0.621707	0.891868	0.436366	0.975839	0.573153	0.765994	0.391002
19580	0.452727	0.703516	0.397365	0.733164	0.217976	0.456726	0.324843

	f1	f2	f3
0	-1.060546	-0.625778	-0.942979
1	-0.392543	-1.065577	-0.464874
2	-1.028627	1.520076	-0.382716
3	1.049609	-0.827049	0.137407
4	1.471850	-0.228448	-1.616504
...	...	...	...
19576	-1.063892	-0.949195	-0.268825
19577	-0.769553	-2.285859	-1.688420
19578	-0.386746	-0.162633	0.429034
19579	-0.372441	0.976126	1.047109
19580	0.939089	2.516770	-1.308995

Figure 7 Feature data transforms the data of the principal component

According to the operation results, the SVM model trained with V1-V7 as the feature has an accuracy rate of 94% on the test set, and the accuracy rate of using F1-F3 as the feature is 92.6%. It can be known that the three new features can well determine whether there is a specific component, so F1, F2 and F3 are considered as the potential main components.

6. Conclusion

Since T test depends on a large sample size, a good approximation can be obtained in this paper, and T test almost uses all the data information so it's best to find the difference. In this paper, correlation analysis, PCA and other methods are comprehensively used to find the main indicators for the detection of specific components in mixtures. In the implementation process, make full use of data to illustrate problems, intuitive and visual, from the actual characteristics of data to choose the modeling strategy, scientific and objective.

References

[1] Fisher Box, Joan. Guinness, Gosset, Fisher, and Small Samples. Statistical Science. 1987, 2 (1): 45–52.

[2] Si Shou-Kui, SUN Xi-jing. Python Mathematical Experiment and Modeling [M]. Beijing: Science Press,2020.04:335.

[3] Si Shoukui, SUN Zhaoliang. Mathematical Modeling Algorithm and Application (2nd edition) [M]. Beijing: National Defense Industry Press, 2019, 07:223.