Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 10, No. 3, 2023

Research on Key Substances in the Rating of Strong Aroma Crude Spirits Based on Correlation Algorithms

Yingjie Peng [1,2], Xianguo Tuo [1,2] *, Xuemei Zhu [1,2], Wei Zhuang [1,2], Wenzhu Mu [1,2]

[1] School of Automation & Information Engineering, Sichuan University of Science & Engineering, Yibin 644000, China
[2] Artificial Intelligence Key Laboratory of Sichuan Province, Yibin 644000, China

[*] **Corresponding author**: Xianguo Tuo (Email: gyz_118@163.com)

Abstract: This experiment obtained the substance content of crude spirits at different distillation times using Gas Chromatography-Mass Spectrometry (GC-MS). The relationship between substance content and crude spirits rating was revealed using Spearman's rank correlation coefficient, the Maximal Information Coefficient (MIC), and Principal Component Analysis (PCA). There were 11 substances with a Spearman's coefficient greater than 0.70, 9 substances with a MIC greater than 0.6, and 6 substances in the PCA with an information extraction rate greater than 1.5×105. In combination with these three algorithms, a total of 17 substances were found to be related to the crude spirits grading. These substances are: 1,1-diethoxy-3-methylbutane, ethyl valerate, ethyl hexanoate, 2-methyl-1-butanol, ethyl caproate, ethyl lactate, ethyl nonanoate, butyl lactate, 2-hydroxy-4-methylpentanoic acid ethyl ester, isoamyl lactate, ethyl decanoate, butyric acid, (2,2-diethoxyethyl)-benzene, ethyl laurate, ethyl heptadecanoate, ethyl stearate, ethyl linolenate.

Keywords: GC-MS; Volatile substances; Spearman; MIC; PCA.

1. Introduction

Baijiu is a traditional Chinese spirit made primarily through solid-state fermentation, a unique process that has evolved over more than 2000 years [1]. This spirit typically uses a mixture of grains, most commonly sorghum, along with rice, glutinous rice, wheat, and corn [2]. Compared to other global distilled spirits, baijiu has a more diverse base of ingredients. In addition, the fermentation process of baijiu is distinct from other international distilled spirits. The conversion of starch into sugar and then into alcohol happens simultaneously during the fermentation process, a result of the diverse microbiological composition of the Qu (a type of fermentation starter), which includes yeast, molds, bacteria, and lactic acid bacteria [3]. These various microbes not only break down the ingredients into alcohol, but they also produce a range of volatile compounds such as esters, acids, aldehydes, and alcohols, with a total concentration of about 2-5g/L[4]. The interaction of these micro-constituents contributes to the diverse flavor profiles observed in baijiu.

Recent advances have seen the application of Gas Chromatography-Mass Spectrometry (GC-MS), Gas Chromatography, Liquid Chromatography, Spectroscopy, Electronic Nose, and Nuclear Magnetic Resonance in analyzing the chemical composition of baijiu [5-13]. Notable studies have shown the efficacy of GC-MS in differentiating various types of baijiu based on their volatile compound profiles [14,15], determining the grading of strong-flavor baijiu [16], and identifying and classifying different brands of baijiu [17]. These findings suggest that GC-MS, combined with correlation algorithms, can effectively detect the trace compounds in baijiu and distinguish between different types of spirits based on these constituents.

However, the majority of these studies have focused on the finished product of baijiu, which often has pronounced differences between types. There has been less research on the correlation between the composition and grading of different fractions of crude spirits. Therefore, this study aims to investigate the trace compounds in different distillation periods of baijiu using GC-MS and establish their relationship with the grading. The content of volatile compounds in the crude spirits will be determined using GC-MS, followed by an analysis of how these compounds change during the distillation process. Lastly, Spearman correlation, MIC, and PAC algorithms will be employed to determine the influence of different compounds on the grading of crude spirits, aiming to identify the key compounds that affect the grading.

2. Materials and Data Collection

2.1. Materials

Crude Spirits (17 batches of 202 bottles of crude spirits produced by a well-known liquor factory in Sichuan in May 2022); 2-Ethylbutyric acid (Chromatographically pure, purchased from Macklin Biochemical Co., Ltd. in Shanghai).

2.2. Instruments

Agilent 7890B–G7000D Triple Quadrupole Gas Chromatography Mass Spectrometry System (Agilent, USA); Agilent 123-7032 DB-Wax gas chromatography column [30m×0.32mm×0.25μm] (Agilent, USA).

2.3. Sample Collection Method

A total of 17 batches of 202 bottles of crude spirits were collected, each batch from a different fermentation pit. Each batch of crude spirits was divided into head, middle, and tail fractions. Each batch varied based on actual field conditions, with samples collected according to alcohol content, distillation time, and field tasting conditions. The head and tail fractions had significant and unstable changes in quality, so the head and tail sections were primarily collected for analysis and classification of liquor grades. After collection, five national-level evaluators graded the liquor based on color, aroma, taste, and style, and divided the liquors into final grades (Grade 1: 40 bottles, Grade 2: 90 bottles, Grade 3: 72 bottles). Each sample was labeled with the format "Grade-

Position within grade", where the batch number refers to batches 1-17, grade refers to grades 1-3, and the position within the grade refers to the order of collection within that grade, with higher numbers indicating later collection times.

2.4. Physical and Chemical Data Detection Method

GC-MS detection conditions: Automatic GC injection, chromatographic column selected was Agilent DB-WAX (30 × 320 × 0.25 μm), with a FID detector, and the liner chosen was Agilent 5062-3587 (900 μL). The temperature program was: held at 60 ℃ for five minutes, then increased at 10 ℃ per minute to 250 ℃, and held for 2 minutes. High purity helium was used as the carrier gas, with a flow rate of: 2.25 mL/min, non-split, total flow of: 34.5 mL/min, vaporizer temperature of 250 ℃; 1μl injection volume; MS interface temperature of 280 ℃; EI (electron ionization, EI) ion source of 70 eV ionization; ion source temperature of 230 ℃; quadrupole temperature of 150 ℃; full scan mode; scan range of 30~540 m.

Quantitative method: 2-Ethylbutyric acid was used as an internal standard, referring to the national standard GB/T 10345—2007 "White Wine Analysis Method", using the internal standard method, the content of each flavor component was calculated based on peak area.

Material analysis software: All data was analyzed using Excel for basic data analysis, and then the relationship between material content and grade was determined using Python library references and custom MATLAB programming.

2.5. Substance Distribution in Samples

A total of 89 substances were detected in the samples via Gas Chromatography-Mass Spectrometry (GC-MS). The data varies slightly for each fermentation pit and each segment of the crude spirits, but the overall trends are generally similar. Here, we selected the top 35 substances with the highest detection rates in all samples for analysis. The names and concentration information of these substances are as as shown in Table 1.

Table 1. Substance and content of the Crude Baijiu

Index	Compound	Range (mg/L)	Difference (mg/L)	Average (mg/L)
1	1,1-Diethoxy-3-methylbutane	55.63~0.81	54.82	19.87
2	Ethyl valerate	92.00~0.51	91.49	16.79
3	Ethyl hexanoate	1487.01~63.72	1423.29	447.76
4	2-Methylbutanol	14.67~1.23	13.44	5.63
5	Isopentanol	95.62~7.31	88.31	38.95
6	Hexyl acetate	16.41~1.02	15.39	6.50
7	Propyl hexanoate	32.71~1.01	31.70	6.78
8	Ethyl heptanoate	99.17~3.44	95.73	31.19
9	Ethyl lactate	595.05~0.51	594.54	154.66
10	Ethyl formate	81.09~21.06	60.03	35.80
11	Butyl hexanoate	59.28~5.05	54.23	18.13
12	Ethyl octanoate	92.13~7.02	85.11	35.53
13	Isopentyl hexanoate	16.43~1.42	15.01	5.27
14	Acetic acid	124.59~34.56	90.03	55.61
15	Pentyl acetate	9.71~0.70	9.01	2.60
16	Ethyl nonanoate	2.32~0.64	1.68	1.25
17	Butyl lactate	10.76~1.01	9.75	3.16
18	2-Hydroxy-4-methyl-pentanoic acid ethyl ester	46.14~0.67	45.47	13.39
19	Isopentyl lactate	7.56~1.21	6.35	3.67
20	Hexyl hexanoate	78.32~1.65	76.67	19.32
21	Ethyl decanoate	3.20~0.61	2.59	1.32
22	Butyric acid	58.01~1.71	56.30	23.05
23	(2,2-Diethoxyethyl)-benzene	4.62~0.53	4.09	2.27
24	Valeric acid	22.01~0.52	21.49	4.19
25	Ethyl dodecanoate	16.40~0.21	16.19	0.90
26	Hexanoic acid	127.64~15.01	112.63	54.54
27	Ethyl tetradecanoate	34.05~0.42	33.63	5.33
28	Ethyl pentadecanoate	10.40~0.27	10.13	2.67
29	Ethyl hexadecanoate	458.81~1.37	457.44	60.09
30	9-Hexadecenoic acid ethyl ester	43.09~0.38	42.71	10.92
31	Ethyl heptadecanoate	5.05~0.23	4.82	1.19
32	Ethyl octadecanoate	38.15~0.32	37.83	6.24
33	Elaidic acid ethyl ester	261.78~4.16	257.62	34.73
34	Linoleic acid ethyl ester	366.60~5.73	360.87	51.70
35	Linolenic acid ethyl ester	52.66~2.01	50.65	12.03

During the distillation process, substances with low boiling points and high volatility distill out first, while substances

with high boiling points and low volatility distill out later. Figure 2 shows a diagram of the changes in the total content of substances during the distillation process, drawn from 10 randomly selected samples out of 27 groups of liquor samples. The horizontal axis represents samples collected at different times (named by the "grade - grade internal position" rule). It can be seen from the figure that as the crude spirits grade gradually increases, the total substance content decreases gradually in the early stage and tends to stabilize in the later stage.

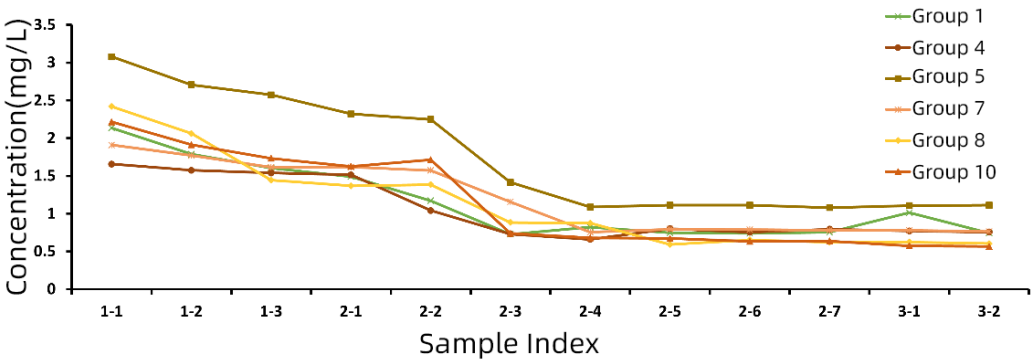

Figure 1. Changes of total content in distillation process

Further combining the information in Table 1 and the content of each substance, it is found that the content of most esters is negatively correlated with the distillation stage. As the distillation progresses, their content in the sample decreases continuously. The content of the ester substances numbered 2, 3, 5, 6, 7, 8, 11, 12, 21, 27, 29, 30, 31, 32, 33, 34, and 35 in the table gradually decreases as the distillation process progresses. A few esters do not follow this pattern. The content of the esters numbered 9, 15, 17, 18, 19, 20, 23, 24, 25, and 28 in the table gradually increases as the distillation process progresses. The content of acids that are distilled out is almost all increasing, and the acids numbered 14, 22, 24, and 26 in the table gradually increase their content as the distillation process progresses. It can be found that the outflow of substances is closely related to the grading of the crude spirits, so further analysis is needed on the correlation between the content of substances and the grade of the crude spirits.

3. Methods

The content of a substance at a particular moment in the crude spirits distillation process is closely related to the progression of the distillation, implying that both the amount and proportion of substances can impact the grade of the crude spirit. Thus, it is vital to accurately quantify the correlation between substances and the grade of the crude spirit. In this study, we employed the Spearman rank correlation coefficient, MIC, PCA algorithms to investigate the degree of influence of substances on the grade of the crude spirit. Finally, the key volatile substances impacting the grade of the crude spirit were obtained by forming the union of substances selected by the three methods.

3.1. Principle of Spearman Coefficient Algorithm

During the collection of the crude spirits, the outflow time of the crude spirits of different grades is not the same, and the content of substances in the crude spirits of different grades varies significantly. In this situation, the order of substance content size can better reflect the variation of substances with distillation time. The Spearman rank correlation coefficient replaces the numbers themselves with the order of data size, which can determine the correlation of non-normally distributed data and discrete data. The treatment of "grade" in this algorithm coincides with the grade of the crude spirit and the size of the substance content. The calculation formula is as follows:

$$r_s = \rho_{\mathrm{R}(X),\mathrm{R}(Y)} = \frac{\mathrm{cov}(\mathrm{R}(X),\mathrm{R}(Y))}{\sigma_{\mathrm{R}(X)}\sigma_{\mathrm{R}(Y)}} \quad (1)$$

Where: Ω_i is the Pearson product-moment correlation coefficient, $\mathrm{cov}(\mathrm{R}(X),\mathrm{R}(Y))$ is the covariance of the rank variables, and $\sigma_{\mathrm{R}(X)}\sigma_{\mathrm{R}(Y)}$ is the standard deviation of the rank variables.

According to the formula (1), the larger the absolute value of the Spearman rank coefficient, the more correlated the substance is with the grade of the crude spirit. A positive Spearman rank correlation coefficient indicates a positive impact, while a negative Spearman rank correlation coefficient indicates a reverse impact.

3.2. Principle of Maximum Mutual Information Coefficient Algorithm

The maximum mutual information coefficient algorithm is proposed by Reshef and others to measure the correlation strength between two variables. It is an improvement of the mutual information score and is often used to measure the degree of association between two variables X and Y. MIC ranges from 0 to 1, calculates the joint probability density through the scatter plot divided by the grid, and finally obtains the mutual information value, which can capture linear and nonlinear associations. Let and represent the grade of the crude spirit and the substance content, respectively, then the joint distribution of the grade and content is , and the marginal distributions are , respectively. The mutual information is the relative entropy of the joint distribution and the marginal distribution, calculated as follows:

$$\mathrm{MIC}(\mathrm{X};Y) = \max_{m\times n<B} \frac{\sum_{x_i\in x}\sum_{y_i\in y} f(x_i,y_i)\log_2\frac{f(x_i,y_i)}{f(x_i)f(y_i)}}{\log_2\min(m,n)} \quad (2)$$

Where: m, n represent the desired divisions of the x, y direction grid, f(xi, yi) is the joint probability density function, is the upper limit of the number of m×n grids, is a function related to the sample scale n, B=n0.6.

3.3. Principle of PCA Algorithm

PCA is a statistical analysis method that transforms a set of potentially correlated variables into a set of linearly independent variables through orthogonal transformation. These new linearly independent variables are called principal components. This method can reveal the inherent structure of the data and reduce high-dimensional data to low-dimensional data while retaining the original features [36]. In PCA, the principal components are arranged in the direction from large to small of the original data variance. The first principal component retains the most original information, the second principal component retains slightly less original information, and so on. The load factor is the contribution rate of each data to the principal component. The larger the load factor, the greater the contribution to the principal component. The information extraction rate is a measure of the extraction rate of the original variable by the principal component. The higher the information extraction rate, the stronger the correlation between the original variable and the dependent variable.

$$\Omega_i = \sum_{j=1}^{m} \frac{\lambda_i \mu_{ij}^2}{\sigma_i^2} = \sum_{j=1}^{m} \rho_{ij}^2 \tag{3}$$

Where:

$$\rho_{ij} = \rho(x_i, F_i) = \frac{\mu_{ij}\sqrt{\lambda_j}}{\sigma_i} \tag{4}$$

Where: Ω_i represents the extraction rate of the original variable xi by the first m principal components, λ_i is the eigenvalue corresponding to the j-th principal component, μ_{ij} is the load factor of the i-th original variable and the j-th principal component, σ_i is the variance of the i-th original variable. ρ_{ij} is the correlation coefficient of the i-th original variable with the j-th principal component?

4. Correlation Analysis of Substance Content and Crude Spirits Grade

Figure 1 shows the heat distribution map of the 11 substances with Spearman rank correlation coefficient greater than 0.70 in Table 1. These 11 substances are numbered 1, 2, 6, 9, 17, 18, 19, 23, 31, 32, 35, corresponding to the names of the substances: 1,1-diethoxy-3-methylbutane, pentanoic acid ethyl ester, hexanoic acid ethyl ester, lactic acid ethyl ester, butyric acid lactate, 2-hydroxy-4-methyl-pentanoic acid ethyl ester, lactic acid isoamyl ester, (2,2-diethoxyethyl)-benzene, heptadecanoic acid ethyl ester, octadecanoic acid ethyl ester, linoleic acid ethyl ester. Among them, the contents of substances 1, 2, 6, 31, 32, 35 decreased with the increase of distillation time, while the contents of substances 9, 17, 18, 19, 23 increased with the increase of distillation time. Besides, the content of these substances is not only closely related to the grade of the original liquor, but also highly correlated with each other, indicating that although the content of substances in the original liquor differs greatly, there is a relatively stable proportion between the contents of each substance.

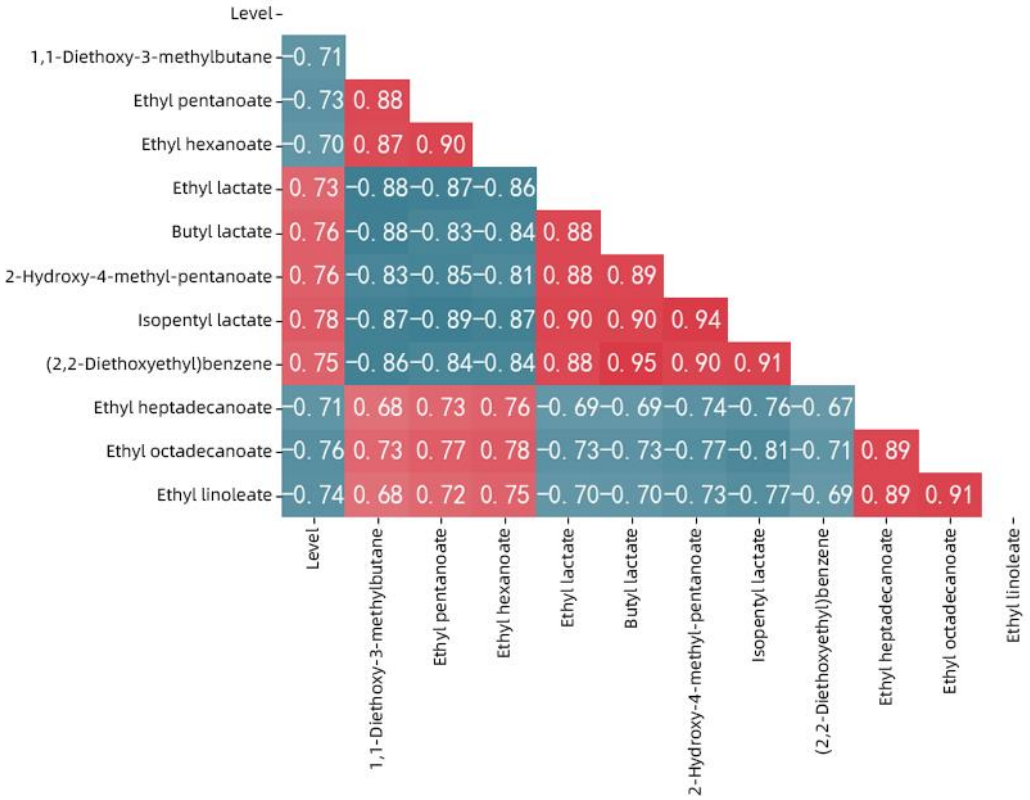

Figure 2. Spearman grade correlation coefficient heat map

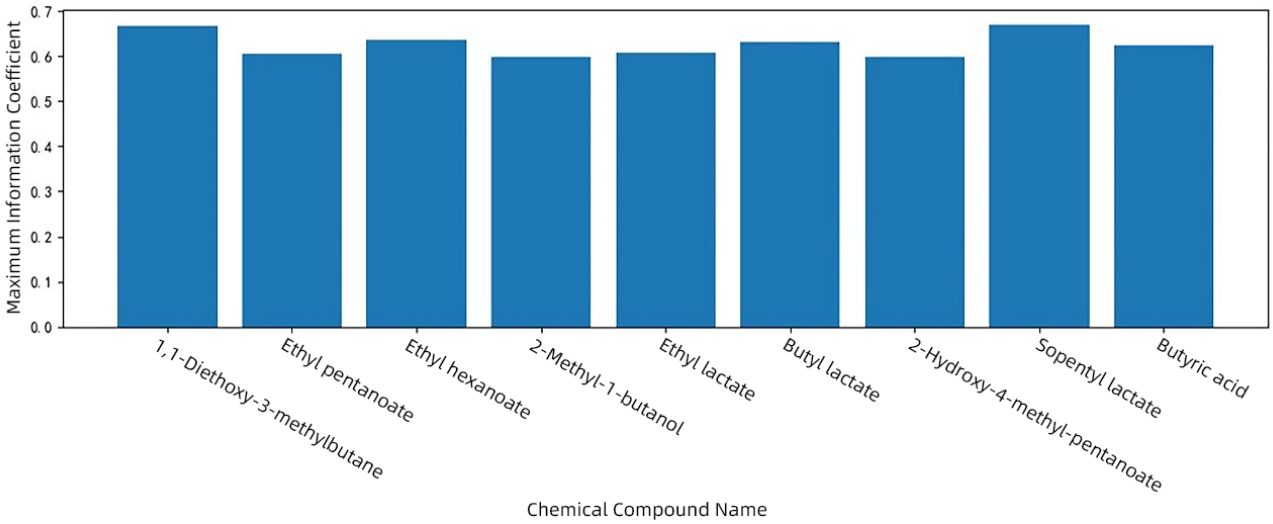

Figure 3. MIC coefficient of substances

Figure 2 shows the coefficient distribution map of 9 substances with MIC coefficient greater than 0.60 in Table 1. These 9 substances are numbered 1, 2, 3, 4, 9, 17, 18, 19, 22, corresponding to the names of the substances: 1,1-diethoxy-3-methylbutane, pentanoic acid ethyl ester, hexanoic acid ethyl ester, 2-methylbutanol, lactic acid ethyl ester, butyric acid lactate, 2-hydroxy-4-methyl-pentanoic acid ethyl ester, lactic acid isoamyl ester, butyric acid. According to the previous research, the contents of substances 1, 2, 3, 4 increased in the first, second, and third grade liquor samples, while the contents of substances 9, 17, 18, 19, 22 decreased in the first, second, and third grade liquor samples.

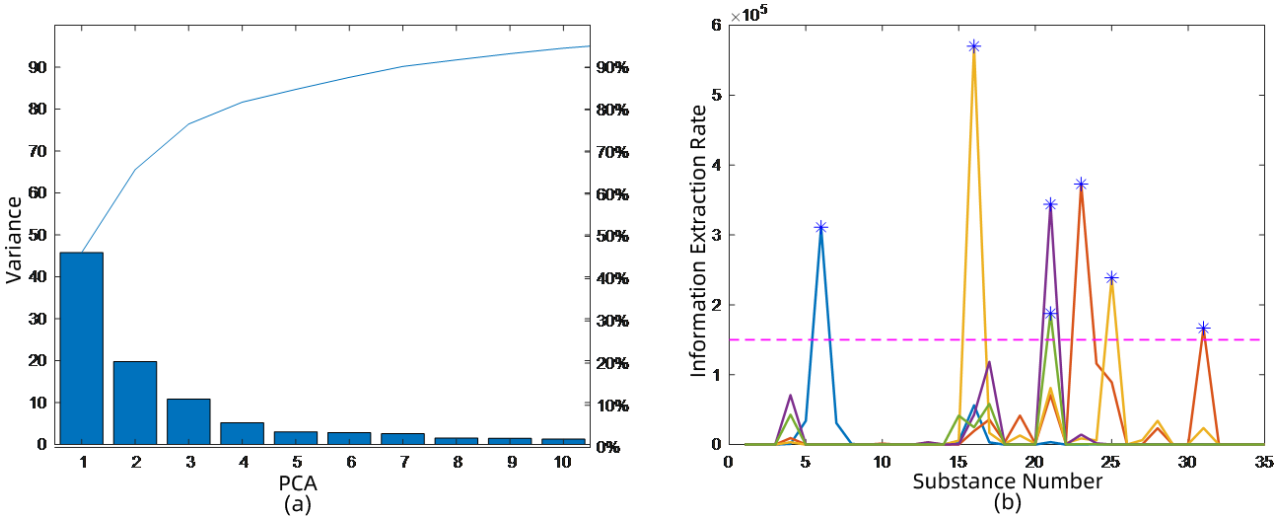

Figure 4. Correlation analysis of PCA grade

Figure 3(a) shows the contribution and cumulative contribution of the top ten PCs after dimension reduction of the crude spirits data. In order to avoid the information extraction rate of each PC being submerged in the total of all PC' information extraction rates, the information extraction rate formula is used to measure the impact of substance content on the crude spirits grade. The information extraction rates of the top five PCs were calculated separately, and shows in Figure 3(b), taking 1.5×105 as the threshold, and extracting highly correlated substances.

Using the PCA algorithm, six substances related to the grading were selected from Table 1, with serial numbers: 6, 16, 21, 23, 25, 31. Their names are: Hexanoic acid ethyl ester, Nonanoic acid ethyl ester, Decanoic acid ethyl ester, (2,2-Diethoxyethyl)-benzene, Dodecanoic acid ethyl ester, Heptadecanoic acid ethyl ester. In combination with previous research, it was found that substances 6, 16, 31 appear in the early stage of distillation, and a spirit sample that does not contain these three substances is likely to be a third segment spirit.

Taking the union of the substances obtained from the three methods, finally, 17 key volatile substances that affect the grade of crude spirits were selected. Their names are: 1,1-Diethoxy-3-methylbutane, Pentanoic acid ethyl ester, Hexanoic acid ethyl ester, 2-Methylbutanol, Hexanoic acid ethyl ester, Lactic acid ethyl ester, Nonanoic acid ethyl ester, Butyl lactate, 2-Hydroxy-4-methyl-pentanoic acid ethyl ester, Isopentyl lactate, Decanoic acid ethyl ester, Butyric acid, (2,2-Diethoxyethyl)-benzene, Dodecanoic acid ethyl ester, Heptadecanoic acid ethyl ester, Stearic acid ethyl ester, Linolenic acid ethyl ester.

5. Conclusion

In this experiment, 85 trace substances were detected from 202 samples of crude spirits using GC-MS, and the relationship between the content of 35 trace substances and the grading of crude spirits was studied using Spearman's coefficient, MIC, and PAC. These three algorithms were used to select 17 substances related to the grading of crude spirits from the samples. The substances are: 1,1-Diethoxy-3-

methylbutane, Pentanoic acid ethyl ester, Hexanoic acid ethyl ester, 2-Methylbutanol, Hexanoic acid ethyl ester, Lactic acid ethyl ester, Nonanoic acid ethyl ester, Butyl lactate, 2-Hydroxy-4-methyl-pentanoic acid ethyl ester, Isopentyl lactate, Decanoic acid ethyl ester, Butyric acid, (2,2-Diethoxyethyl)-benzene, Dodecanoic acid ethyl ester, Heptadecanoic acid ethyl ester, Stearic acid ethyl ester, Linolenic acid ethyl ester.

During the entire distillation process, these 17 substances show a steady upward or downward trend. The content and proportion of substances at different distillation times vary, and only when the proportions of various substances are moderate in the middle of the distillation will the crude spirits present a full-bodied taste.

References

[1] HONG J, ZHAO D, SUN B. Research Progress on the Profile of Trace Components in Baijiu[J/OL]. Food Reviews International, 2021: 1-27.

DOI:10.1080/87559129.2021.1936001.

[2] LIU H, SUN B. Effect of Fermentation Processing on the Flavor of Baijiu[J/OL]. Journal of Agricultural and Food Chemistry, 2018, 66(22): 5425-5432.

DOI:10.1021/acs.jafc.8b00692.

[3] ZHU Y, TRAMPER J. Koji – where East meets West in fermentation[J/OL]. Biotechnology Advances, 2013, 31(8): 1448-1457.

DOI:10.1016/j.biotechadv.2013.07.001.

[4] FAN W, XU Y, QIAN M. Current Practice and Future Trends of Aroma and Flavor Research in Chinese Baijiu[M/OL]//GUTHRIE B, BEAUCHAMP J D, BUETTNER A, et al. ACS Symposium Series: Vol. 1321. Washington, DC: American Chemical Society, 2019: 145-175[2023-02-25].

https://pubs.acs.org/doi/abs/10.1021/bk-2019-1321.ch012.

[5] LIU LL, YANG H, JING X, et al. Analysis of volatile constituents of vintage wine based on GC-MS and GC-IMS [J/OL]. Science and Technology of Food Industry: 1-16[2022-08-22].

DOI: 10.13386/ j.issn.ssn1002-0306.2022040054.

[6] HU J, MA YF, LUO JX, et al. Simultaneous determination of 57 flavor compounds in Liquor by gas chromatography [J]. China Brewing,202,41(05):206-211.

[7] XIONG YF, MA Z, PENG YS, et al. Research progress on chromatographic analysis of flavor components of chinese liquor [J]. China brewing,2019,38(11):1-5.

[8] DENG B, SHEN CH, DING HL, et al. Application progress of infrared spectrum analysis technology in liquor industry [J]. China brewing,2020,39(09):13-17.

[9] TANG JD ZHAO YM, RAN GY, et al. Research progress of spectral technology in quality control of liquor [J/OL]. Science and Technology of Food Industry :1-16[2022-08-23].

DOI: 10.13386/ J.ISSN 1002-0306.2022050161.

[10] LI JJ, SUN ZH, MENG QH. Liquor recognition based on hand-held electronic nose [J]. Journal of food and fermentation industry, 2019, (24) : 218-222. The DOI: 10.13995 / j.carol carroll nki. 11-1802 / ts. 021867.

[11] ZHANG Y, XIA Al. Discriminant analysis of liquor brand based on low field NMR [J]. China brewing,2021,40(10):207-209.

[12] GUO YX, CHENG W, CHEN XJ, et al. Modern instrument analysis technique in the study of liquor flavor group [J]. Journal of food safety and quality testing, 2022, 13 (16) : 5218-5226. The DOI: 10.19812 / j.carol carroll nki jfsq11-5956 / ts. 2022.16.021.

[13] ZHAO DR, SHI DM, SUN J Y, et al. Characterization of key aroma compounds in gujinggong chinese baijiu by gas chromatography--olfactometry, quantitative measurements, and sensory evaluation [J].Food Research International, 2018, 105: 616—627.

[14] ZHANG JJ, YU Q, HU SW, et al. Analysis on the difference of flavor components between light elegance and luzhou-flavor old liquor [J]. China brewing,2021,40(10):157-162.

[15] FU X, NIE QY, ZHANG Y, et al. Analysis of volatile constituents of typical luzhou-flavor liquor by gc-ims [J]. China brewing,2021,40(11):178-183.

[16] LI XF, ZHANG L, LI FF, et al. Quality evaluation of luzhou-flavor liquor based on GC-QTOF MS [J]. Food industry science and technology, 2019, 40 (15) : 235-241. The DOI: 10.13386 / j.i ssn1002-0306.2019.15.039.

[17] QIAN Y, HU X, SUN Y, et al. Study on classification of luzhou-flavor liquor based on fingerprint and stoichiometry [J]. China brewing,2021,40(06):152-156.