Heliyon 10 (2024) e36871

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

A two-step framework integrating lasso and Relaxed Lasso for resolving multidimensional collinearity in Chinese baijiu aging research

Dongyue An [a], Liangyan Wang [a,**], Jiang He [b], Yuejin Hua [a,*]

[a] MOE Key Laboratory of Biosystems Homeostasis and Protection, Institute of Biophysics, College of Life Sciences, Zhejiang University, Hangzhou, PR China

[b] Sichuan Institute of Atomic Energy, Irradiation Preservation Key Laboratory of Sichuan Province, Chengdu, 610101, PR China

ARTICLE INFO

Keywords:
High dimensional data
Lasso regression
Chinese baijiu
Aging

ABSTRACT

The aging process is crucial for Chinese Baijiu production, significantly enhancing the spirit's flavor, aroma and quality. However, aging involves a complex interplay of numerous compounds, and the extensive duration required for aging leads to a scarcity of samples available for scientific research. These limitations pose a challenge in analyzing high-dimensional data with collinearity, complicating the understanding of the intricate chemical processes at play. In this article, a two-step framework was proposed that integrated Relaxed Lasso regression models with Lasso-selected predictors to address this issue. Baijiu samples subjected to various aging conditions were analyzed using direct GC-MS and HS-GC-MS, and the obtained data was processed by this approach. The results demonstrate significantly superior performance compared to other methods, including PLSR and Gradient Boosting. Analyses were also performed on a previously documented dataset, yielding enhanced results and underscoring the method's advantage in processing high dimensional data with multicollinearity. Moreover, this method proved effective in screening of potential indicative compounds, highlighting its utility in Baijiu aging research.

1. Introduction

Chinese Baijiu (also referred to as Chinese liquor) is a traditional distilled alcoholic beverage of China, and has been widely consumed in China for thousands of years. Baijiu is made from grains and undergoes a co-fermentation process involving multiple microbes. It is then distilled, stored, and skillfully blended. Freshly distilled Chinese Baijiu has undesirable characteristics due to its strong pungent and spicy taste. To eliminate these unpleasant odors and develop a well-balanced, mellow and mature aroma, it is typically stored in a sealed pottery jar for several years. This transformative maturation is commonly referred to as "aging" [1].

Over the past few decades, extensive research has explored various aspects of Chinese Baijiu through comprehensive compound

Abbreviations: Lasso, Least Absolute Shrinkage and Selection Operator; GC-MS, Gas Chromatography-Mass Spectrometer; HS, headspace; PLSR, Partial Least Squares Regression; LOOCV, Leave-One-Out Cross-Validation; CV, Cross-Validation; MSE, Mean Square Error; R^2, coefficient of determination; VIF, Variance Inflation Factor.

* Corresponding author.
** Corresponding author.
E-mail addresses: 12107140@zju.edu.cn (D. An), liangyanwang@zju.edu.cn (L. Wang), 285021736@qq.com (J. He), yjhua@zju.edu.cn (Y. Hua).

https://doi.org/10.1016/j.heliyon.2024.e36871
Received 6 April 2024; Received in revised form 22 August 2024; Accepted 23 August 2024

analyses, including its classification, quality scores, production regions, and vintage characteristics. Aging is among the most complex of them, as it involves a multitude of known and unknown physical and chemical reactions. There are over 2000 chemicals in Baijiu that account for its flavor [2] and change with time. During aging, irritant compounds volatilize, alcohols are oxidized to aldehydes and acids, while alcohols, acids and esters are converted into one another through esterification and hydrolysis reactions [3]. However, natural aging is expensive and time-consuming, alternative methods have been explored to accelerate the aging process of Chinese Baijiu, including irradiation, ultrasonic waves, and varying storage conditions [4,5]. For example, Gamma irradiation can break water and ethanol down into intermediates including H_2O_2, H^+, OH, H, H_2, OH^-, hydrogen and alkane to speed up oxidation, esterification and polymerization reactions [4]. Ultrasonic waves can accelerate Baijiu aging through enhanced mass transfer and diffusion of flavor compounds [5].

The common strategy for investigating Baijiu aging includes two steps: instrumental analysis and statistical analysis. Instrumental techniques for compound analyses include GC-MS, LC-MS (Liquid Chromatography-Mass Spectrometry), E-nose (Electronic nose), 1H NMR (proton Nuclear Magnetic Resonance), and HPLC-NMR (High Performance Liquid Chromatography-Nuclear Magnetic Resonance) [6]. The chemometric tools mainly include PCA (principal component analysis), LDA (linear discriminant analysis), PLSR (partial least squares regression), ANN (artificial neural networks), SVM (support vector machine). For example, Xu et al. applied electronic nose signals, PCA, and PLSR to discriminate Baijiu samples with different ages and to predict their storage time [7]. Chen et al. employed headspace-gas chromatography-ion mobility HS-GC-IMS and PLSR to detect the age of Chinese Baijiu [8].Using time-resolved fluorescence, Zhang et al. developed a relationship between fluorescence lifetimes and storage time of Baijiu [9]. Li et al. used GC and 1H NMR spectroscopy combined with PLSR to identify the storage vintage of Chinese Baijiu samples [3]. Jiang et al. established an ANN model to optimize the irradiation parameters for the desired aging effect [10]. Jia et al. explored the differentiation between natural aging and gamma irradiation maturation, as well as investigating the aging mechanism through UPLC-Orbitrap-MS/MS and PLS-DA [11].

These studies treated the detected compounds as independent variables for unsupervised or supervised learning. However, as detection technologies advance, an increasing number of compounds have been identified in Baijiu, with over 2000 flavor compounds detected to date [2]. This proliferation of detectable compounds presents challenges for data analysis, particularly in the context of supervised learning methods. The complexity of the Baijiu brewing process makes it difficult to provide large-scale observational samples. Furthermore, the analysis is complicated by the vast number of independent variables (detected compounds) and the challenge of multicollinearity within these variables. Multicollinearity, indicative of a dataset where the rank of the matrix X is less than the minimum of n (number of observations) and m (number of variables), reflects non-orthogonality and leads to several adverse effects. High collinearity often results in unstable estimates of regression coefficients in linear models like least squares regression, complicating the assessment of each variable's relative importance. Additionally, collinearity could exacerbate overfitting issues, where the model mistakenly includes random noise as part of significant relationships. Therefore, it is crucial to preprocess the data to reduce collinearity and employ appropriate modeling techniques, such as Stepwise regression, PCR (Massy, 1965), latent root regression (Webster, Gunst, & Mason, 1974), PLSR (Abdi, 2003), Ridge regression, Lasso regression and Elastic Net regression (Hoerl & Kennard, 1970), etc. Despite the availability of various methods, only PCR and PLS have been widely applied in this field.

The Lasso technique has proven particularly useful in situations where the number of predictors (variables) exceeds the number of observations due to its capability of reducing dimensionality. Lasso aims to obtain a subset of predictors that minimizes prediction error for a quantitative response variable by imposing a constraint on the sum of the absolute values of the model parameters [12]. It performs both variable selection and regularization to enhance the prediction accuracy and interpretability of the statistical model. Relaxed Lasso is a variation of the Lasso method that addresses one of Lasso's limitations: the potential for bias in parameter estimation, especially when dealing with highly correlated variables. The Relaxed Lasso procedure first applies the Lasso method to select variables, and then refits the selected variables with a lesser penalty [13]. The versatility and efficiency of these methods make them powerful tools for data scientists across disciplines.

To address the challenges of collinearity and high dimensionality in Baijiu aging studies, we developed a two-step framework employing Lasso for initial variable selection while Relaxed Lasso for further refinement and regression. First, the rationale for integrating Lasso and Relaxed Lasso to enhance model performance and address multicollinearity was discussed. Next, the application of the proposed framework to both artificial and natural aging studies of Chinese Baijiu was detailed, including the fitting process, result evaluations, and comparisons with other methodologies. Moreover, the outcomes of key compound selection were showcased, elucidating their relationship with aging mechanisms. Finally, the differential performance across various datasets was examined, highlighting the potential applicability and limitations of the framework. The results demonstrated that the two-step framework integrating Lasso and Relaxed Lasso can be an effective method for better analysis of the aroma datasets associated with the Chinese Baijiu aging process.

2. Materials and methods

2.1. The two-step framework integrating lasso and Relaxed Lasso

2.1.1. Lasso for feature selection

Let $X = \left(X^1, \ldots, X^p\right)$ be the p-dimensional predictor variable and Y a response variable of interest. Lasso is a regression method that shrink the estimated coefficients of OLS (ordinary least square) regression. The estimated coefficients of OLS are expressed as follows:

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} n^{-1} \sum_{i=1}^{n} \left(Y_i - X_i^T \beta\right)^2 = \left(X^T X\right)^{-1} X^T y$$

If there is multicollinearity in the input values, the matrix attains singularity. Thus, the OLS estimate would be biased and overfit the data. The Lasso (least absolute shrinkage and selection operator) was put forwarded by Tibshirani [14]. Lasso regression can handle the problem by imposing an ℓ_1-penalty on the absolute values of the regression coefficients, which is defined by

$$\widehat{\beta}^{\lambda_1} = \underset{\beta}{arg\,min} n^{-1} \sum_{i=1}^{n} \left(Y_i - X_i^T \beta\right)^2 + \lambda_1 \|\beta\|_1 \tag{1}$$

which can also be written as

$$\widehat{\beta}^{\lambda_1} = \underset{\beta}{arg\,min} n^{-1} \sum_{i=1}^{n} \left(Y_i - X_i^T \beta\right)^2$$

subjected to $\sum_{k\in\{1,\ldots,p\}} |\beta_k| \le t$.

where $\widehat{\beta}$ is the estimated regression coefficients, $\|\beta\|_1 = \sum_{k\in\{1,\ldots,p\}} |\beta_k|$ is the ℓ_1-norm of the vector of coefficients, $\lambda_1 \in [0, \infty)$ is the tuning parameter that controls the strength of penalty in variable selection, n is the number of samples, and p is the number of original features. The constraint makes the sum of the absolute values less than a fixed value and shrinks some of them to be zero. The tuning parameter λ_1 plays key role that controls the strength of penalty. As λ_1 increases, more coefficients are shrunk to zero.

The ℓ_1-penalty has two effects, model selection and shrinkage estimation. On the one hand, for variables in the selected model, coefficients are shrunken towards zero compared to the least-squares solution, thus improves the overall prediction accuracy by sacrificing a little bias to reduce the variance of the predicted values. On the other hand, a certain set of coefficients is set to zero and hence excluded from the selected model $\mathscr{M}_{\lambda_1} = \left\{1 \le k \le p | \widehat{\beta}_k^{\lambda_1} \ne 0\right\}$.

Compared to other methods such as Ridge regression and Elastic Net regression, which do not shrink coefficients exactly to zero, this property is particularly useful for feature selection, especially in high-dimensional data where the number of predictors exceeds the number of observations. Previous studies have shown that Lasso outperforms other methods in removing the redundant or irrelevant features or features which are strongly correlated in the data without much loss of information [15]. In numerous fields, especially those dealing with high-dimensional data, Lasso has been utilized in feature selection for model fitting to make the model easier to interpret and increase generalization by reducing the variance. For example, it was applied to building predictors for climate variables in environmental modeling, offering interpretations that are climatologically meaningful for the high-dimensional climate datasets [16]. In systems biology research, Lasso was integrated with parallel tempering to reduce model complexity by selecting important variables [17]. For high-dimensional data, it has also been demonstrated that the technique of employing features extracted by Lasso to build the regression model could offer a promising trade-off result in terms of prediction accuracy and model interpretation [15,18].

Lasso was conducted to generate predictor variables from the concentration matrix of Baijiu compounds, which has a large number of potential predictors and a limited sample size.

2.1.2. Relaxed Lasso regression

The Lasso suffers from some drawbacks, such as slow convergence rates for high-dimensional data with sparse structure, over-shrinkage of non-zero coefficients, and selection of noise variables. The over-shrinkage issue of Lasso can introduce bias into the estimation. Addressing the two interrelated effects of Lasso, model selection and shrinkage estimation, may help mitigate this problem to some extent. It is not clear whether it is indeed optimal to control these two effects by a single parameter only. As an example, it might be desirable in some situations to estimate the coefficients of all selected variables without shrinkage. The Relaxed Lasso was proposed by Meinshausen [13] to tackle this challenge. It controls model selection and shrinkage estimation by two separate parameters λ and ϕ, and allows coefficients to be estimated with less bias than Lasso, which is defined by

$$\widehat{\beta}^{\lambda_2,\phi} = \underset{\beta}{arg\,min} n^{-1} \sum_{i=1}^{n} \left(Y_i - X_I^T \left\{\beta \bullet 1_{\mathscr{M}_{\lambda_2}}\right\}\right)^2 + \phi\lambda_2 \|\beta\|_1 \tag{2}$$

where $\mathscr{M}_{\lambda_2}$ denotes the set of predictor variables selected by the ordinary Lasso, $\lambda_2 \in [0, \infty)$ determines which variables are included, as in the ordinary Lasso, $\phi \in [0, 1]$ determines how much the coefficients of the selected variables are shrunken towards zero. The Lasso and Relaxed Lasso estimators are identical when $\phi = 1$. For $\phi < 1$, the shrinkage of coefficients is reduced. The Relaxed Lasso estimator for $\phi = 0$ are defined as the limitation of the above definition. The estimation has less variance than Lasso and can adapt to the sparseness and signal-to-noise ratio of the underlying data by varying the relaxation parameter.

To reduce computational complexity, the Relaxed Lasso follows a two-stage procedure. Firstly, it uses Lars algorithm to compute all ordinary Lasso solutions for a range of λ_2 values. Let $\lambda_2^1 > \ldots > \lambda_2^s = 0$ be the penalty term and $\mathscr{M}_{\lambda_2^1}, \ldots, \mathscr{M}_{\lambda_2^s}$ be the variable sets of results. When $\mathscr{M}_{\lambda_2} = \mathscr{M}_{\lambda_2^k}$, if and only if $\lambda_2 \in \left(\lambda_2^k, \lambda_2^{k-1}\right]$. The second stage computes all Relaxed Lasso solutions for each k, which first compute the direction of Lasso as $f_{(k)} = \left(\widehat{\beta}^{\lambda_2^k} - \widehat{\beta}^{\lambda_2^{k-1}}\right) / \left(\lambda_2^{k-1} - \lambda_2^k\right)$. Let $\widetilde{\beta} = \widetilde{\beta}^{\lambda_2^k} + \lambda_2^k f_{(k)}$. If there is one component l so that $sign(\widetilde{\beta}_l) \ne$

$sign\left(\widetilde{\beta}_l^{\lambda_2^k}\right)$, all Relaxed Lasso solutions for $\lambda_2 \in \Lambda_k$ and $\Phi \in [0,1]$ need to be computed. Otherwise, Relaxed Lasso solutions for $\lambda_2 \in \Lambda_k$ and $\Phi \in [0,1]$ are linear interpolation between $\widehat{\beta}^{\lambda_2^{k-1}}$ and $\widetilde{\beta}$.

Since Relaxed Lasso is a derivative method of Lasso, it also adapts to high dimensional data with collinearity. The application of Relaxed Lasso has been relatively rare so far, but it has shown good performance in feature selection and tumor classification for microarray data [19], which selected feature by reducing the number of irrelevant genes and enhancing the accuracy of gene identification, thereby optimizing the performance of the multi-class support vector machine for tumor classification.

In this study, the Relaxed Lasso was utilized to model the relationship between aging parameters and the selected features by Lasso. A continuum of solutions that minimized the cross-validation error was yielded was conducted by adjusting the parameters λ_2 and ϕ in the Relaxed Lasso model. Ultimately, those variables with nonzero coefficients, which were corroborated by other analytical methods, were pinpointed as the potential indicative aroma compounds of aging parameters amidst a vast pool of potential predictors.

2.1.3. Overall description

The main methodology used in this paper is a two-step framework integrating Lasso and Relaxed Lasso, as demonstrated in Fig. 1: Lasso for feature selection and Relaxed Lasso for further variable selection and regression.

It was notice that, apart from the over shrinkage of coefficients, Lasso's penalty can lead to rapid sparsification of selected variables $\mathscr{M}_\lambda$, potentially leading to the inclusion of noise variables or the exclusion of significant ones. Although Relaxed Lasso mitigates the over-shrinkage of non-zero coefficients caused by large λ, such flaws also arise in Relaxed Lasso due to their shared variable selection strategy. For example, if Lasso selects 10 variables at $\lambda = 0.6$, increasing λ to 0.9 might abruptly reduce the selected variables to 2, bypassing intermediate combinations (3–9 variables) [20]. Such issues detract from the accuracy and efficiency of the estimation procedure in the high-dimensional datasets [15].

By applying Relaxed Lasso after Lasso, we aim to refine the initial variable selection and achieve a better fitting performance. In the first stage of the combined method, a series of Lasso models, each with a distinct regularization factor λ_1 were applied to the original data set. Selected features with non-zero coefficients, as determined by Equation (1), were pooled across the models ($\mathscr{M}_{\lambda_1^1}$,$\mathscr{M}_{\lambda_1^2}$,…). In the second stage, we take each set of the Lasso-selected variables (e.g., $\mathscr{M}_{\lambda_1^k}$) in the feature pool, and perform another round of regression starting from a zero-penalty using Relaxed Lasso. By tuning λ_2, the Relaxed Lasso generated a new set of selected variables ($\mathscr{M}_{\lambda_1^1}$,$\mathscr{M}_{\lambda_1^2}$,…). This allows for the selection of intermediate variable counts (e.g., 4, 6, 8 variables), addressing the limitations of Lasso in variable selection under strong penalties. Furthermore, by tuning ϕ, the advantages of Relaxed Lasso, including better fitting performance and reduced bias in coefficient estimation as calculated by Equation (2), were retained. Cross-validation procedure was performed to estimate the optimal model tuning parameters for both stages: λ_1 of the first Lasso, λ_2 and ϕ of the Relaxed Lasso. This framework could enhance model performance by ensuring a more comprehensive feature pool and mitigating the risks associated with

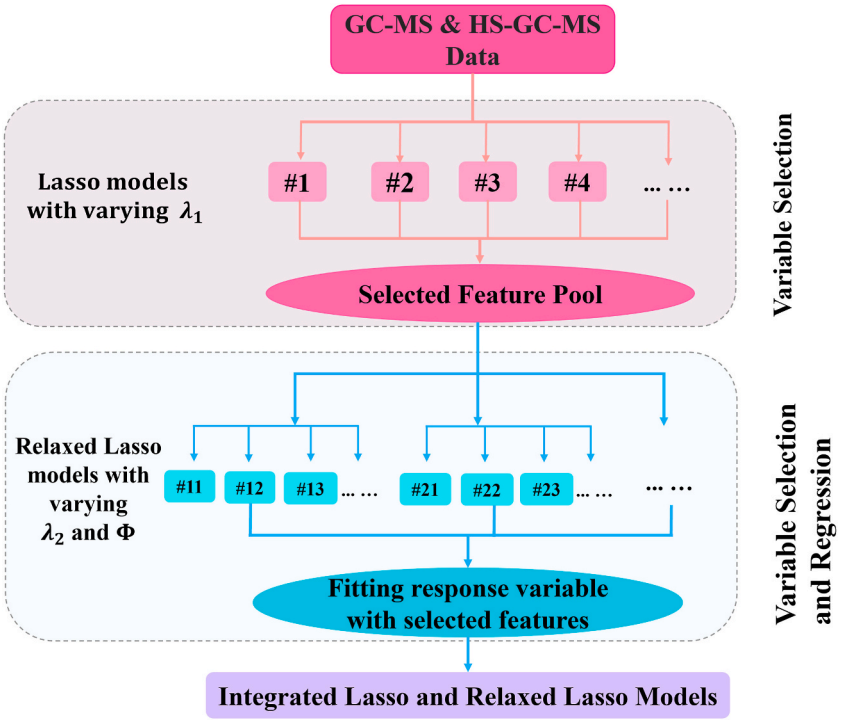

Fig. 1. The Two-Step Scheme integrating Lasso and Relaxed Lasso. The arrows show the data flow.

over shrinkage from a single Lasso penalty.

2.2. Determination of significant variables

The significance of variable selection in multivariate regression techniques has been highlighted by many researchers. In general, variable selection in multivariate regression is based on the principle of either choosing the most contributing variables or eliminating the noncontributing variables [19,21].

The proposed two-step framework integrated Lasso and Relaxed Lasso for variable selection by minimizing the penalized residual sum of squares of the Relaxed Lasso. This process shrank some coefficients of features in the Lasso-selected feature pool to zero, as described by Equations (1) and (2). However, Lasso has certain limitations under specific conditions. When there are highly correlated covariates, it tends to randomly select one or a few predictors and shrink the rest to zero. This can lead to errors when there is high correlation between relevant and unimportant covariates or result in a loss of information when important covariates have a strong dependence structure [20]. On the other hand, PLSR selects variables with high loadings on the principal components, which are linear combinations of the predictors with maximum covariance with the response. It can handle multicollinearity better and take advantage of natural correlations. There are possibly no performance differences between them at realistic quantities of samples [22]. Furthermore, Relaxed Lasso can select variables more flexibly than Lasso by adjusting the tuning parameter. When the tuning parameter is small, Relaxed Lasso can retain more variables than Lasso without compromising the model accuracy. However, this may also include some irrelevant or redundant variables. Therefore, correlation analysis and PLSR were used as supplements to filter out the more relevant and reliable variables in this study [3]. The final potential characteristic markers are those selected by Lasso with coefficients > 0 and either by PLSR with VIP scores > 1 or correlation analysis with p-values < 0.05. This method can select one representative from each group of relevant important variables and eliminate irrelevant ones, yielding credible and comprehensive results.

2.3. The application of integrated lasso and Relaxed Lasso method in studying Chinese baijiu aging

2.3.1. Samples and chemicals

All Chinese Baijiu samples were strong-aroma Baijiu and purchased from Sichuan ** Wine Co., Ltd. The samples were. A series of orthogonal experiments were conducted on 20 artificially aged samples, s-i1, s-i2, …, s-i20, which were freshly distilled spirits immediately following fermentation. These samples first underwent radiation treatment at various doses (3 kGy, 4 kGy, 5 kGy, and 6 kGy), followed by ultrasonic treatment for different durations (0 min, 5 min, 10 min, and15 min), and were then stored at varying temperatures (20 °C, 25 °C, 30 °C, 35 °C). Additionally, five naturally aged samples, s-n1, s-n2, s-n3, s-n4, and s-n5, were stored for different periods (1 year, 2 years, 3 years, 4 years, 5 years) to compare the effects of artificial and natural aging. The irradiation source was 60Co and the absorbed dose rate was 0.71 kGy/min. The ultrasonic treatment was conducted at a frequency of 25k Hz. For specific information, see Table S1. The internal standard, N-amyl acetate ($\geq$ 99.0 %), was obtained from Sigma-Aldrich (Shanghai, China). Water was purified using a Milli-Q water purification system (Millipore, Bedford, MA, USA).

2.3.2. Direct GC-MS analysis

The direct injection method was used, in accordance with the Chinese national standard (GB/T 10345-2007). A mixture of Chinese Baijiu and internal standard solution (N-amyl acetate; final concentration, 200 mg/L) was placed in a vial, and 1 μL was separated and detected by GC-MS. GC-MS analysis was performed on an Agilent 7890A gas chromatography equipped with a HP-5MS column (30 m $\times$ 0.25 mm, 0.25 μm) and an Agilent 5977B mass spectrometer. Helium (flow rate, 0.8 mL/min) was used as the carrier gas. The injector and detector temperatures were both set to 250 °C and the ion source temperature was set to 230 °C. The GC-MS conditions were as follows. Oven temperature was held at 40 °C for 3 min, increased to 180 °C at a rate of 5 °C/min, then further increased to 260 °C at a rate of 8 °C/min, and finally held at this value for 6 min. The mass spectrometer was operated in the electron impact mode at 70 eV over a scan range of 20 $\sim$ 350 *m/z*. Analyses were conducted in triplicate.

2.3.3. HS-GC-MS analysis

This study's static headspace method is validated as a simple, efficient way to quantitatively analyze low-boiling point compounds [23].

A total of 5 mL of each Baijiu sample solution, spiked with internal standard (N-amyl acetate) to a final concentration of 200 mg/L, was transferred to a 20-mL headspace vial. After 20-min heating at a constant temperature of 70 °C in Agilent 7697A static headspace equipment, the upper gas was analyzed by GC-MS. GC-MS conditions were as described above for the direct injection method.

The GC-MS analysis was performed on an Agilent 7890B gas chromatography coupled with an Agilent 7000C mass selective detector, equipped with a HP-5MS column (30 m $\times$ 0.25 mm, 0.25 μm). Helium (flow rate, 0.8 mL/min) was used as the carrier gas in the split mode (30:1). The injector and detector temperatures were both set to 200 °C and the ion source temperature was set to 230 °C. The GC-MS conditions were as follows. Oven temperature was held at 40 °C for 2 min, increased to 200 °C at the rate of 2 °C/min, and held at this value for 5 min. The mass spectrometer was operated in the electron impact mode at 70 eV over a scan range of 20~350 *m/z*. The analyses were conducted in triplicate.

2.3.4. Data pretreatment and statistical analysis

Identification of volatile components was achieved by comparing the mass spectra with the reference spectra of the National

Institute of Standards and Technology (NIST) database. The relative concentrations of volatile components were estimated with the following equation using the internal standard:

$$c_i = c_0 \bullet \frac{s_i}{s_0} \tag{3}$$

where c_i is the relative concentration of an aroma component (mg/L), c_0 is the concentration of the internal standard (N-amyl acetate, 200 mg/L), s_i is the peak area of the analyte, and s_0 is the peak area of the internal standard. All analyses were conducted in triplicate and the average amount of each flavor compound was calculated.

The results of aroma analysis for Baijiu samples were expressed as a heat map, where the aroma compounds were analyzed using HCA (hierarchical cluster analysis). The KNN approach was used to estimate the missing values for compound contents.

Considering the impact of the semi-quantification method on the accuracy of results due to the discrimination effect and the effect of abundance on the significant information of compounds, z-score standardization, defined as the deviation of each point from the mean, divided by the standard deviation, was used to balance signal variances of intensity and abundance. The pairwise correlation coefficients (r) and the variance inflation factor (VIF) between aroma compounds were computed for collinearity diagnostics. The correlation coefficients between aroma compounds and process parameters were calculated, and correlation networks were analyzed.

2.3.5. Model evaluation

To evaluate model performance with our very small dataset, we applied leave-one-out cross-validation (LOOCV). This method involves sequentially removing each sample from the dataset, building a model with the remaining samples, and then predicting the excluded sample. It ensures the training and test sets are disjoint and as independent as possible. Given the dataset of 20 samples, we performed LOOCV by creating 20 folds and conducting 20 cross-validation iterations, fully utilizing each sample, reducing the variability on validation sets. For datasets with a larger sample size, it would be advisable to set aside a separate test set and evaluate the model's performance on it. For methods requiring parameter optimization, LOOCV was also used to determine the optimal parameter configuration. Performance metrics including MSE (Mean Square Error) and the R^2 (coefficient of determination) were employed to assess the models' effectiveness [24]. They are defined as follows:

$$MSE = \frac{\sum_{i=1}^{n} \left(y_{pre} - y_{act}\right)^2}{n} \tag{4}$$

$$R^2 = 1 - \frac{\left(y_{pre} - y_{act}\right)^2}{\left(y_{act} - y_{mean}\right)^2} \tag{5}$$

where n is the number of samples. y_{pre} represents the forecasting value of the response variable, y_{act} is the actual value of the response

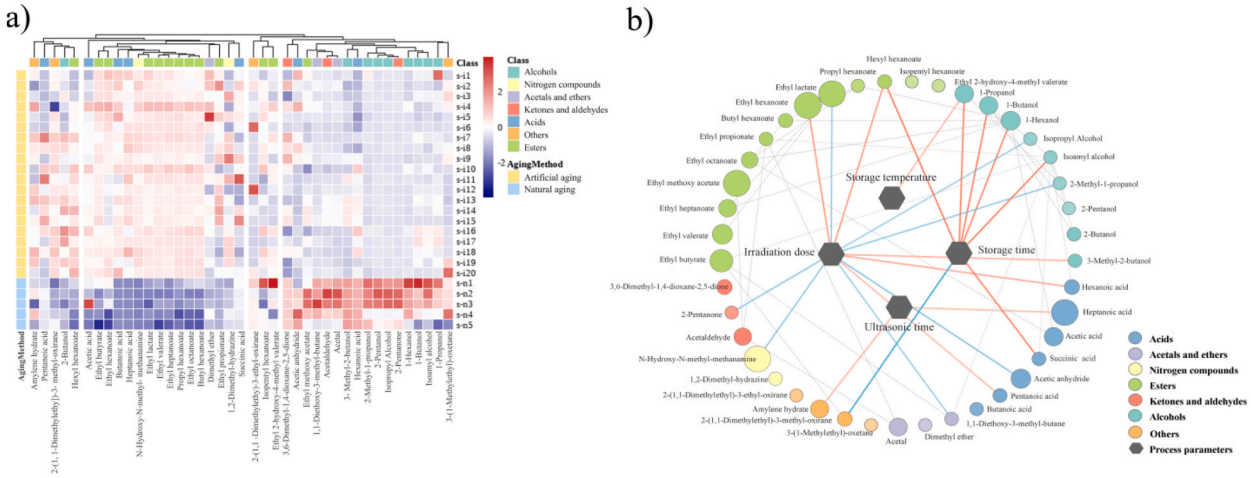

Fig. 2. Cluster analysis and correlation analysis of detected compounds a) Heat map and HCA clustering results of 41 compounds in artificially and naturally aged Baijiu samples. b) Correlation network between process parameters and aroma compounds ($p < 0.05$) and between aroma compounds ($|\rho| > 0.8$) of Baijiu samples. (The gray nodes represent the process variables of Baijiu including storage time as the variable of natural aging, and irradiation dose, ultrasonic time and storage temperature; the outer nodes represent aroma compounds; different colors represent different aroma classes. The size of nodes is proportional to the amount of compound and the thickness of lines is proportional to the value of Pearson's correlation. Solid lines represent correlations between process parameters and aroma compounds and dotted lines represent correlations between aroma compounds. Red solid lines represent positive correlations, blue solid lines represent negative correlations). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

variable and y_{mean} is the mean value of the response variables. In general, the smaller the MSE of the model, the closer R^2 is to one, the better the model's accuracy and predictive ability. The performance measured by them of our method was compared to other multicollinearity-solving methods to prove that it's effective and feasible.

3. Results and discussion

3.1. Changes in compounds during natural aging and artificial aging

Volatile components were identified and then quantified by Equation (3) to depict natural aging and artificial aging. Fig. 2 a) shows the standardized, scaled (by column), and clustered results, where the color at the top indicates the category of compound. For more details, refer to Table S2, supplementary data. The relative amounts of various compounds in different samples can be compared from the heatmap. There were obvious differences between artificially aged Baijiu samples and naturally aged Baijiu samples. In naturally aged Baijiu samples, these compounds showed a regular change pattern as the storage time increased. The contents of compounds in artificially aged samples were also related to the parameter values (irradiation dose, storage temperature and ultrasonic time). We observed that compounds of the same class tended to be closer to each other, indicating their similar change pattern. In the naturally

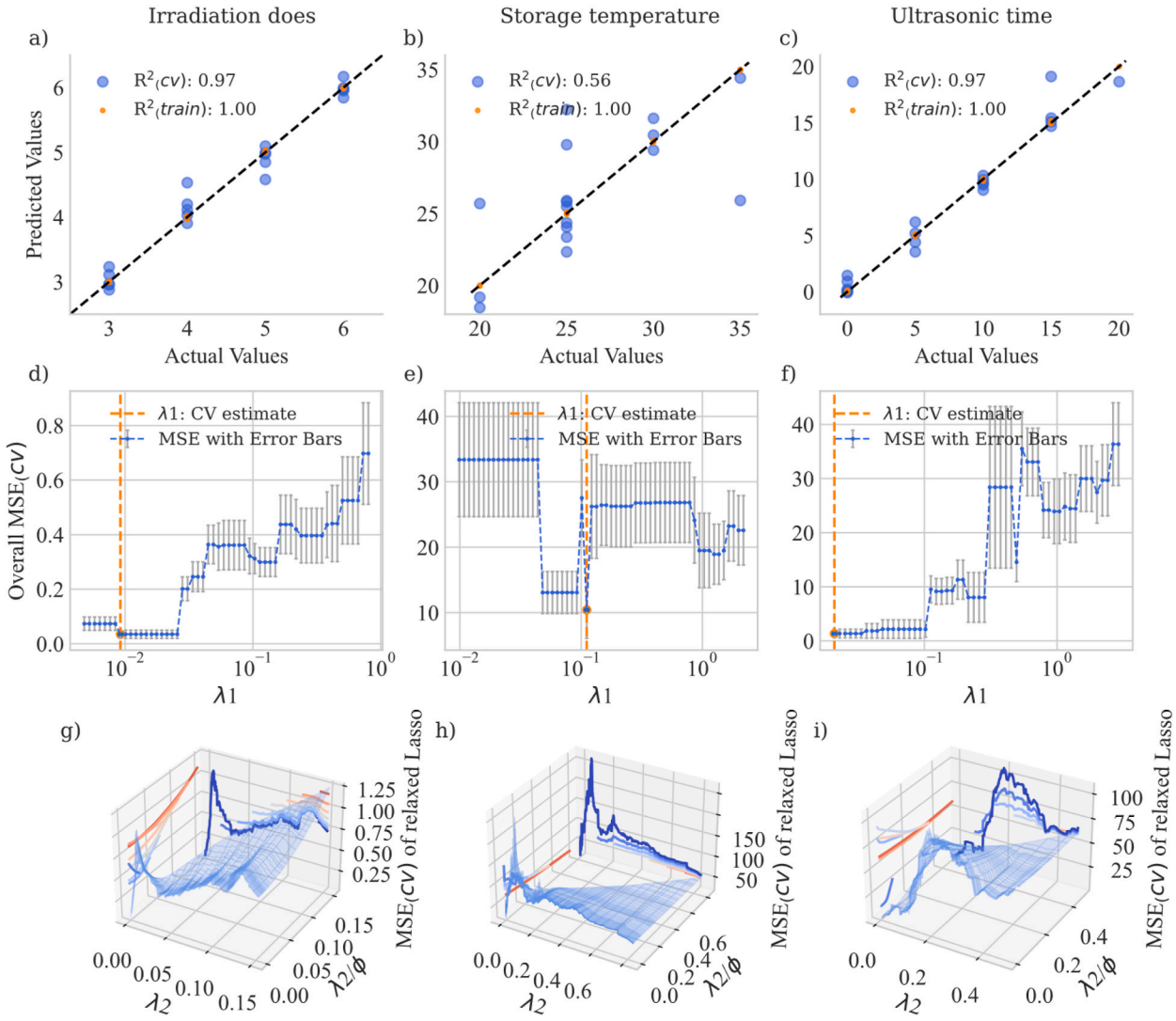

Fig. 3. Model optimization process. The results of predicting a) irradiation dose, b) storage temperature and c) ultrasonic time by integrated Lasso and Relaxed Lasso models on the training set and validation set. The effects of variable shrinkage and selection of the first Lasso on the overall cross-validation MSE can be seen in e) irradiation dose, f) storage temperature and g) ultrasonic time. The tuning effects of λ_2 and ϕ of Relaxed Lasso models on the overall cross-validation MSE with λ_1 fixed to optimal values were shown in g) irradiation dose, h) storage temperature and i) ultrasonic time.

aged Baijiu samples, ketones, aldehydes, acids and esters all increased with time due to oxidation and esterification reactions, which has been shown previously [25]. The contents of most acids, esters, aldehydes and ketones in artificially aged Baijiu samples were higher than those in naturally aged ones, whereas the contents of alcohols were lower. These results are consistent with the earlier finding that irradiation can accelerate the physical and chemical reactions of natural aging and shorten the aging time.

The correlation network was obtained by calculating Pearson correlation coefficients. As shown in Fig. 2 b), in the strongly correlated node pairs, there were four esters, five alcohols, one acid and one nitrogen compound positively correlated and one nitrogen compound negatively correlated with storage time. In terms of irradiation dose, compounds positively correlated included hexyl hexanoate, hexanoic acid and methylal, whereas compounds negatively correlated were ethyl lactate, ethyl hexanoate, amylene hydrate, isopropyl alcohol, acetic anhydride acetal and 2-pentanone. Ethyl octanoate, propyl pyruvate and succinic anhydride were strongly correlated with ultrasonic time. Only 1-propanol had a significant correlation with storage temperature, which might be because the irradiation dose was too strong to be affected by temperature. Therefore, the main factors influencing aroma compounds were irradiation dose and ultrasonic time. Based on our results and previous studies, we can infer that artificial aging can accelerate physical and chemical reactions, shortening the aging time, but specific effects still need further exploration.

3.2. Model building and optimization

The integrated Lasso and Relaxed Lasso models were built with compounds as input and artificial aging parameters as output, and optimization was performed by tuning model parameters. The optimization process consists of two stages: variable subset generation and Relaxed Lasso regression. Fig. 3 a)-c) show the results of the Relaxed Lasso model with predictive variables selected by Lasso and evaluated by cross-validation. In the first stage, ordinary Lasso was used to generate predictors. As the tuning parameter λ_1 increases, some coefficients become zero and are eliminated, resulting in a pool of predictive variables for the next stage. In the second stage, the Relaxed Lasso models was refitted using the selected variables and their performance were measured using leave-one-out cross-validation. The minimum MSE_{cv} of each Relaxed Lasso model, as determined by Equation (4), was plotted against λ_1 in Fig. 3 d)- f). It can be seen that the MSE_{cv} has a U-shaped relationship with λ_1 for irradiation dose and storage temperature, but a positive relationship for ultrasonic time. This suggests that coefficient shrinkage can reduce overfitting and instability by removing redundant predictors, but excessive sparsity can also impair the model performance.

In the second part, we tuned the regularization parameter, $\lambda_2 \in [0,\infty)$, and the relaxation parameter, $\phi \in [0,1]$, to obtain sparse and consistent solutions of the Relaxed Lasso model. For each set of predictors selected by λ_1, Relaxed Lasso models were established with different values of λ_2 and ϕ, and their predictive performance was evaluated by cross-validation. The exact solutions of the Relaxed Lasso estimator were computed by the Lars-algorithm [26]. Fig. 3 g), h) and i) show how MSE_{cv} of the Relaxed Lasso models varies with λ_2 and ϕ. For all datasets, the minimum MSE_{cv} was achieved when both λ_2 and ϕ were small, which indicates that the variable selection by coefficient shrinkage is similar to that of the first Lasso stage. This suggests that the variables selected in the first stage are indeed significant, while ordinary Lasso tends to over-shrink the coefficients for our datasets, and the relaxation parameter ϕ helps to balance variance and bias for a lower prediction loss [13].

It can be seen that the increasement of λ_2 first enlarged and then reduced MSE_{cv}, which might be due to the trade-off between bias and variance. Growth of λ_2 raised bias and lowered variance when it was small, while achieved variable selection and refitting when it was large, which could improve the accuracy. On the other hand, varying ϕ changed the penalty term from 0 to λ_2 but didn't change the predictors. Irradiation dose had significant correlation with most of the compounds, making it more prone to overfit than the other two factors. For irradiation dose models, increasing ϕ had opposite effects on the MSEcv depending on λ_2. When λ_2 was small, it increased the MSE_{cv}, which might be because it reduced the penalty necessary to avoid overfitting. When λ_2 was large, it decreased the MSE_{cv}, which may be due to fewer predictors and less penalty. For storage temperature and ultrasonic time, increase of ϕ would result in reduction of MSE_{cv} when λ_2 was extremely small, while enlarged MSE_{cv} when λ_2 was larger. A possible explanation is that the number of predictors did not cause overfitting and might even be underfitting and did not need much penalty, while larger λ_2 removed some redundant noise variables and showed more accurate prediction. These patterns depend on multiple factors, such as overfitting or underfitting of the predictors, and the optimal relaxation for coefficient estimation and predictor selection.

In summary, Table 1 and Fig. 3 show the optimization results and model performance, measured by MSE, R^2 and VIF. For

Table 1
Model performance and optimization results.

Responding Variable	Number of predictors	MSE		R^2		λ_1	λ_2	ϕ	VIF
		Training	Validation	Training	Validation				
Irradiation dose	17	1.705E-04	3.470E-02	9.999E-01	9.722E-01	1.033E-01	2.193E-05	3.139E-01	9.908E+00
Storage temperature	18	1.838E-03	1.042E+01	9.999E-01	5.600E-01	2.181E+00	1.254E-04	6.455E-02	2.399E+01
Ultrasonic time	16	6.228E-03	1.365E+00	9.998E-01	9.650E-01	2.100E-02	3.901E-03	3.645E-02	6.245E+00

Notes: MSE: mean square error, R^2: coefficient of multiple determination, λ_1: the tuning parameters of the ordinary Lasso model, λ_2 and ϕ: the tuning parameters of the Relaxed Lasso model, VIF: variance inflation factor.

irradiation dose prediction, the model involving 17 variables selected by $\lambda_1 = 1.033e-1$ and $\lambda_2 = 2.193e-5$, and Relaxed Lasso by $\phi = 3.139e-1$ performed best. Its R^2_{cv} was 0.9722. For ultrasonic time, the model with optimal performance involving 18 predictors was achieved when $\lambda_1 = 2.1e-2$, $\lambda_2 = 3.901e-3$ and $\phi = 3.645e-2$. Its R^2_{cv} was as high as 0.965. For storage temperature, the best model used the predictors selected by $\lambda_1 = 2.181$, $\lambda_2 = 1.254e-4$, and Relaxed Lasso by $\phi = 6.455e-2$. Its R^2_{cv} was 0.56. The Relaxed Lasso performed better on irradiation dose and ultrasonic time than on storage temperature, suggesting that they are the dominant factors affecting the volatile compounds of Baijiu samples. The weak performance on storage temperature was consistent with the correlation analysis, which showed that storage temperature has little impact on the aroma composition of Baijiu in the aging process, and is overshadowed by the other factors in the dataset.

The dataset had severe collinearity, as shown by the high average VIF values ($\overline{VIF} = 6.24E+16$). However, the features selected by our framework had much lower VIFs than the original ones. The average VIFs of irradiation dose and ultrasonic time models were below 10, and only the less important factor, storage temperature, had an average VIF above 20.

3.3. Model comparison

Multicollinearity, where predictor variables are highly correlated, can make the least squares method less effective. And various strategies have been devised for linear regression to mitigate this issue [27]. In this study, the proposed two-step framework integrating Lasso and Relaxed Lasso was compared with six commonly used methods to evaluate its effectiveness.

The other six methods each have their advantages and disadvantages. PLSR is a commonly used technique to handle multicollinearity and reduce dimensionality by constructing orthogonal latent variable that maximize the covariance with the response variables. In this study, A previously suggested scheme was employed to sequentially eliminate variables based on their significance for model refinement [3]. Ridge regression is a shrinkage regression method that adds a small positive constant to the diagonal elements of the predictor matrix (ℓ_2-norm), making it nonsingular and invertible. Lasso, Relaxed Lasso and the combination of them are also shrinkage regression techniques with ℓ_1-norm penalty that were described above. Lasso is preferred when the solution have sparse features because ℓ_1 regularization promotes sparsity, but its covariate selection is arbitrarily when the dataset is highly collinear [28]. Elastic-net regression combines ridge regression and Lasso regression, overcoming the limitations of both. It adds a penalty term to the sum of squared residuals that consists of both the ℓ_1-norm and the L2-norm of the regression coefficients and can handle multicollinearity by encouraging sparse solutions and grouping correlated predictors [29]. Gradient Boosting, a nonlinear regression method, was also included in the comparison. It is a type of boosting algorithm that uses least squares as the loss function and builds an ensemble of weak learners, usually regression trees. Gradient Boosting regression can handle multicollinearity to some extent, as it does not rely on the inversion of the predictor matrix or the estimation of individual coefficients [30].

The six aforementioned methods, along with the integrated Lasso and Relaxed Lasso approach proposed in this paper, were employed to perform regression analysis. The models used Baijiu aging factors (irradiation dose, storage temperature, and ultrasonic time) as response variables and Baijiu compound concentrations as predictor variables. Due to the limited sample size, LOOCV was utilized for model optimization and evaluation. Specifically, LOOCV involves calculating the predicted values for each observation when it is used as the validation set. The model's performance on the validation set was assessed using the MSE and R^2 between the predicted and actual values according to Equations (4) and (5). The results of multivariate regression comparative analyses were reported in Table 2.

Regardless of the modeling approaches, irradiation dose models always had the highest R^2, storage time models had the lowest R^2, and ultrasonic time models had intermediate R^2. These results support evidence from the above correlation analysis and model optimization results, indicating that irradiation dose was the dominant factor affecting the contents of aroma compounds, followed by ultrasonic time, while the tiny contribution of ultrasonic time was unexposed in the dataset.

For all the three response variables, the two-step framework integrating Lasso and Relaxed Lasso outperformed the other methods, with R^2 values of 0.9722, 0.5560 and 0.9650, indicating its excellent fitting and generalization ability on our dataset. Apart from that method, Ridge regression models performed best, followed by Relaxed Lasso and Gradient Boosting regression in the case of irradiation dose prediction. For storage temperature prediction, Gradient Boosting regression performed better than the other methods. For ultrasonic time prediction, Relaxed Lasso and Gradient Boosting worked well, whereas PLSR, Lasso and Elastic-Net regression models

Table 2
Comparison of regression method performances.

Regression Method	R^2_{cv}			MSE_{cv}		
	Irradiation dose	Storage temperature	Ultrasonic time	Irradiation dose	Storage temperature	Ultrasonic time
PLS	0.4353	−0.0893	0.2113	0.7059	25.8028	30.5625
Ridge	0.6578	−0.1153	−0.0188	0.4278	26.4183	39.4779
Lasso	0.4002	−0.1080	0.2681	0.7497	26.2465	28.3624
Elastic-Net	0.4002	−0.1580	0.2373	0.7498	27.4308	29.5557
Relaxed Lasso	0.5857	−0.1451	0.4783	0.5178	27.1239	20.2161
Integrated Lasso and Relaxed Lasso	0.9722	0.5560	0.9650	0.0347	10.4232	1.3555
Gradient Boosting	0.5024	0.3788	0.5102	0.6220	14.7142	18.9805

Notes: R^2_{cv}: coefficient of multiple determination on the validation set (LOOCV), MSE_{cv}: mean square error on the validation set (LOOCV).

were inferior. Ridge regression was more suitable for predicting the dominant parameter, irradiation dose, because it shrank all coefficients without eliminating any, capturing the complex relationships between irradiation dose and many significant predictors. On the other hand, Lasso regression was better suited for predicting the other two parameters because it performed feature selection, which helped to reduce model complexity and avoid overfitting when there were fewer relevant predictors.

Relaxed Lasso outperformed Lasso because it avoided over-shrinkage of the coefficients and made them closer to the orthogonal projection on the variables. Moreover, the combination of Lasso and Relaxed Lasso achieved higher prediction accuracy than Relaxed Lasso alone, likely due to the increased diversity of variable combinations selected. In the two-step framework integrating Lasso and Relaxed Lasso, each set of variables selected by the first Lasso was subjected to a second Relaxed Lasso with λ_2 increased from 0, generating new subsets of variables that could compensate for the excessive sparsity caused by large λ_1 in the Lasso regression.

3.4. Determination of potential artificial aging markers

As described above, significant overlapping compounds selected by Relaxed Lasso and either PLSR or correlation analysis were considered characteristic markers of the artificial aging process. The results were represented in Fig. 4. The importance scores of compounds, computed by different algorithms, were described in Fig. 4 a)-c), with details provided in Table S3, supplementary data. The count of variables selected by each method and the number of shared variables identified across different methods were shown in Fig. 4 d)-f).

Regarding irradiation dose, 17, 10 and 12 compounds were screened out by the integrated Lasso and Relaxed Lasso, PLSR and correlation analysis, respectively. Eight selected important compounds by the integrated Lasso and Relaxed Lasso models were confirmed by other methods. Seven of them were determined as indicators of aging by irradiation with different doses according to their practical significance: isopropyl alcohol, 3-methyl-2-butanol, hexanoic acid, heptanoic acid, 1,1-diethoxyethane (acetal), 1,1-diethoxy-3-methyl-butane and hexyl hexanoate. Isopropyl alcohol and 3-methyl-2-butanol are fusel oils, which confer a bitter and spicy sensation to Baijiu [31]. Hexanoic acid has the aroma characteristics of fat, cheese and pit aroma, while heptanoic acid is mainly responsible for the fruity and floral aroma. Organic acids have a substantial effect on the balance of flavor and can effectively reduce the pungent and bitter taste, increase the sweetness of Baijiu, which are evaluated to monitor the production of Baijiu and ensure

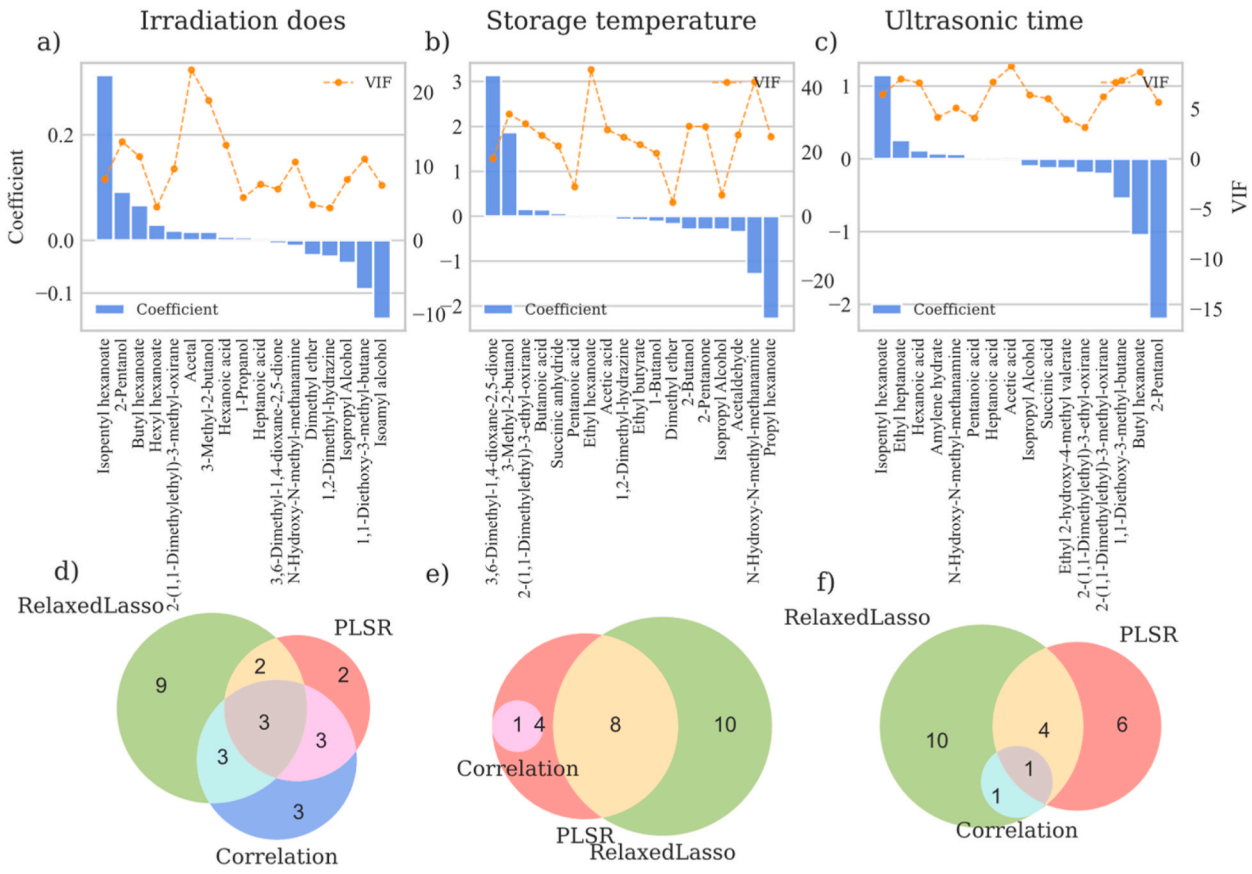

Fig. 4. Variable selection results. Coefficients and VIF of variables selected by Lasso -Relaxed Lasso models for predicting a) irradiation dose, b) storage temperature and c) ultrasonic time. Venn plots for the overlapping compounds selected by the integrated Lasso and Relaxed Lasso framework, PLSR and correlation analysis corresponding to d) irradiation dose, e) storage temperature and f) ultrasonic time.

product quality [31]. It has been reported that the content of organic acids increased through natural aging, and good Baijiu quality often corresponds to high organic acid content in a certain proportion range [32]. 1,1-Diethoxyethane and 1,1-diethoxy-3-methyl-butane are main acetals, recognized as important indicators during the maturity of Baijiu and essential contributors to the pleasant flavor [33]. They can alleviate the pungency of alcohol content and enhance the soft aroma of Baijiu. 1,1-Diethoxy-3-methyl-butane was detected as one of the most important aroma compounds in Wuliangye Baijiu [34].

The comparison between natural aging and gamma irradiation treatment illustrates their effects on the components of Baijiu during aging, with irradiation treatment notably accelerating this process. Hexanoic and heptanoic acids reached equilibrium rapidly in the naturally aged samples, while their concentrations increased with higher irradiation doses in the irradiated samples. Specifically, heptanoic acid levels were significantly elevated under irradiation (6 kGy), reaching much higher levels compared to natural aging. Hexanoic acid approached levels comparable to natural aging only at higher irradiation doses. The content of hexyl hexanoate exhibited an ascending trend over storage time or irradiation dose, with a more pronounced rise following irradiation treatment. Both aging methods enhanced the content of acetal, but the effect was more pronounced at a high dose. Similarly, 1,1-diethoxy-3-methyl-butane and 1,1-diethoxyethane contents rose with time in natural aging and with irradiation dose, but only a high dose matched the effect of 1–2 years of natural aging. These phenomena are likely driven by the mechanism whereby irradiation treatment, particularly at higher doses, generates free radicals. The free radicals can accelerate the oxidation of alcohols to aldehydes and acids, the esterification of alcohols and acids to esters, and the condensation of alcohols and aldehydes to acetals, thus facilitating the aging process and improve the flavor quality. This is consistent with the previous research findings [11,35].

Regarding ultrasonic time, 16, 10 and 2 compounds were detected by the integrated Lasso and Relaxed Lasso method, PLSR and correlation analysis, respectively. Six compounds occurred more than once, four of which were identified as significant markers of ultrasonic time, including hexanoic acid, heptanoic acid, acetic acid and succinic acid. The contribution of the first two acids to the overall aroma of Baijiu has been emphasized in the previous paragraph. Acetic acid, with the highest Osme value in strong-aroma Baijiu [36], is an important organic acid produced by oxidation, which contributed to vinegar-like. Succinic acid is an non-volatile acid that reconciles the taste of Baijiu [37]. With other factors fixed, these acid compounds rose with the extension of ultrasonic time. Moreover, extending ultrasonic exposure reduces alcohol compounds. This could be attributed to the cavitation effect of ultrasound. The ultrasound could produce cavitation bubbles in liquid system. The large amount of energy released during the cracking of these small bubbles is able to change properties of substances in the system, such as the cracking of water molecules to generate free radicals, which accelerates the oxidation, esterification reactions [38]. It also changes the degree of association of hydrogen bonds in the solution molecules, thereby reducing the number of irritating free molecules, making the white wine more soft and mellow.

The impact of storage temperature on the aging process was assessed across uniform natural aging durations and irradiation levels, despite its relatively minor effect. Seven components were considered as key aroma compounds related to the storage time. Among them, acidic compounds, which contributed to sweet and cheesy odor, specifically acetic acid, pentanoic acid, and butanoic acid, escalated with both time and temperature. Acetaldehyde is a metabolite in alcoholic fermentation by yeast, and can be formed by the oxidation during the ageing period. It is a volatile chemical with low boiling point with an unpleasant odors that can evaporate and be eliminated during maturation [39]. Acetaldehyde exhibited a peak during aging, yet slightly declined with temperature increase, which might be due to the oxidation of ethanol at early stage, while decreased slowly because of evaporation and oxidation into acetic acid by dissolved oxygen in later period [3]. 3,6-Dimethyl-1,4-dioxane-2,5-dione and 2-(1,1-dimethylethyl)-3-ethyl-oxirane are plastic materials that might be associated with the packaging of Baijiu and could be detrimental to human health at elevated levels. The analysis revealed that they were stable during natural aging, but increased with increasing temperature. These findings indicate that elevated storage temperatures facilitate the volatilization of low boiling point substances and oxidation reactions, which can help alleviate the pungent and spicy flavor of Baijiu and enhance the taste quality. However, high temperatures might also trigger the release of harmful substances from packaging, compromising the quality and safety of Baijiu.

Our findings demonstrate that different artificial aging methods have distinct yet similar effects on the volatile components of Baijiu. Generally, irradiation, high temperature and ultrasound can enhance organic acids, reduce irritating odors, and soften the wine, giving Baijiu a soft and elegant taste. However, the potential risks of harmful by-products should also be evaluated.

3.5. Application to another natural aging dataset

To evaluate the widespread applicability of our proposed analytical framework, the two-step scheme integrating Lasso and Relaxed Lasso was applied to a dataset from a study on Chinese Baijiu aging, published in existing literature [40]. Table S2, supplementary data shows details about the data. This dataset included measurements of 30 compounds across 36 Baijiu samples with varying aging times (0–30 years), aligning perfectly with our study's focus on multidimensional data characterized by strong inter-variable correlations and a relatively small sample size compared to the number of variables. The concentrations of these 30 compounds were used as input, with storage time as the dependent response variable, and the data were fitted according to the procedures described in our methodology, with adjustments made due to the dataset's relatively large size.

To explore the framework's applicability across different sample sizes, subsets of 7, 12, 18, 24, and 30 samples were extracted from the initial dataset randomly. The sub-samples were then split into training, validation, and test sets, comprising approximately 70 %, 15 %, and 15 % of the total samples, respectively. We performed 5-fold cross-validation on the training and validation sets to optimize model parameters and evaluate validation set performance. Subsequently, the test set was used to assess each model's performance on unseen samples.

Furthermore, multiple regression methods, including Ridge regression, Elastic-Net regression, Lasso regression, Relaxed Lasso regression, PLS regression Gradient Boosting, and the integrated Lasso and Relaxed Lasso, were compared to ensure a comprehensive

evaluation of our framework. These comparisons offered valuable insights into the capabilities and applicability of our approach across different modeling techniques and sample sizes. The evaluation metrics for the validation and test sets are presented in Table 3.

The results shown in Table 3 indicate that the proposed framework, which combines Lasso and Relaxed Lasso, performed well on both the validation and test sets across different sample sizes, especially suitable for small sample sizes. When the sample size was 30, this method achieved R^2 values of 0.9851 and 0.9822 on the validation and test sets, respectively. With a small sample size of 7, the R^2 values reached 0.9937 and 0.9637. Compared to other methods, our approach demonstrated significant advantages with smaller sample sizes (7, 12, 18, 24). As the sample size increased (27, 30), the performance gap between the methods narrowed, with our method performing on par with Relaxed Lasso and surpassing other methods. This suggests that with larger datasets, the standalone Relaxed Lasso method is capable of effectively selecting the potentially relevant variables.

The improved performance of the integrated framework can be attributed to its effective feature selection and capability to handle multicollinearity among variables, which reduces overfitting and enhances model generalization, crucial given the strong inter-variable correlations in the dataset. These factors make it a good choice for high-dimensional datasets with strong inter-variable correlations and relatively small sample sizes.

4. Conclusions

In this paper, a framework was proposed based on Lasso and Relaxed Lasso regression, and applied to fitting multiple aging parameters of Baijiu using compound content matrices. This framework showed superior accuracy over traditional methods, making it suitable for addressing high-dimensionality, multicollinearity, and limited sample size issues in Baijiu aging research. The three artificial aging factors were fitted separately, proving its potential in analyzing complex experimental data involving multiple aging factors. The higher prediction accuracy for irradiation dose and ultrasonic time over storage temperature indicates their dominant effects. The application across datasets of varying sample sizes demonstrated its effectiveness.

Additionally, the two-step framework integrating Lasso and Relaxed Lasso could identify chemically characteristic compounds and complements other variable selection methods. Key compounds were pinpointed for various factors: isopropyl alcohol, 3-methyl-2-butanol, hexanoic acid, and others for irradiation dose; 3-methyl-2-butanol, acetaldehyde, and others for storage temperature; hexanoic acid, heptanoic acid, and others for ultrasonic time; and ethyl hexanoate for storage time. These findings underscore that artificial aging techniques like irradiation and ultrasound accelerate aging, potentially reducing pungent flavors and enhancing overall taste quality of Baijiu. Irradiation generates free radicals that speed up oxidation, esterification, and acetal formation, while ultrasound induces cavitation bubbles that alter substance properties, promoting oxidation and esterification. Consequently, both methods increase the content of acids and esters. Moreover, high temperatures facilitate chemical reactions and the volatilization of low-boiling-point substances such as acetaldehyde, but they may also release harmful compounds from packaging materials.

In future research, increasing the sample size will allow for a more comprehensive evaluation. Each aging method can be individually designed for experiments to explore the aging mechanisms in greater depth. This study can serve as a foundational reference

Table 3
Comparison of regression method performances in another dataset.

Size	Evaluation Metrics	Regression Method						
		Ridge	Elastic-Net	Lasso	Relaxed Lasso	PLS	Gradient Boosting	Integrated Lasso and Relaxed Lasso
7	R^2_{cv}	0.9893	0.9929	0.9767	0.9167	0.8717	0.2380	0.9937
	MSE_{cv}	1.1097	0.7367	2.4237	8.6682	13.3382	79.2442	0.6539
	R^2_{test}	0.8548	0.9550	0.9439	0.9505	0.7503	0.3832	0.9637
	MSE_{test}	7.1139	2.2055	2.7466	2.4231	12.2362	30.2209	1.7783
12	R^2_{cv}	0.9793	0.9599	0.9362	0.9385	0.9150	0.8259	0.9691
	MSE_{cv}	1.6874	3.2700	5.2025	5.0169	6.9389	14.2048	2.5218
	R^2_{test}	0.8211	0.9418	0.9498	0.8144	0.4761	−0.0774	0.9659
	MSE_{test}	2.8632	0.9311	0.8025	2.9691	8.3819	17.2383	0.5452
18	R^2_{cv}	0.9779	0.9123	0.9123	0.9695	0.9463	0.9268	0.9913
	MSE_{cv}	2.0688	8.1979	8.1979	2.8513	5.0233	6.8410	0.8162
	R^2_{test}	0.6039	0.6684	0.6684	0.9769	0.8810	0.9532	0.9888
	MSE_{test}	16.9001	14.1465	14.1465	0.9860	5.0783	1.9966	0.4799
24	R^2_{cv}	0.9557	0.8933	0.8722	0.9774	0.9805	0.9735	0.9847
	MSE_{cv}	3.4742	8.3590	10.0160	1.7688	1.5293	2.0791	1.1982
	R^2_{test}	0.8934	0.8980	0.5953	0.8339	0.8896	0.8863	0.9049
	MSE_{test}	15.8132	15.1314	60.0102	24.6295	16.3657	16.8556	14.1060
27	R^2_{cv}	0.9900	0.9745	0.9733	0.9911	0.9767	0.9870	0.9911
	MSE_{cv}	1.0713	2.7380	2.8612	0.9579	2.5181	1.3905	0.9579
	R^2_{test}	0.4804	0.5675	0.5433	0.9589	0.9477	0.8897	0.9589
	MSE_{test}	27.6638	23.0260	24.3145	2.1863	2.3351	5.8748	2.1863
30	R^2_{cv}	0.9804	0.9385	0.9391	0.9851	0.9753	0.9769	0.9851
	MSE_{cv}	1.9338	6.0660	6.0072	1.4720	2.4370	2.2731	1.4720
	R^2_{test}	0.9419	0.7744	0.8719	0.9822	0.9516	0.9777	0.9822
	MSE_{test}	3.2558	12.6364	7.1734	0.9959	2.7115	1.2469	0.9959

Notes: R^2_{cv}: coefficient of multiple determination on the validation set (5-fold cross-validation), MSE_{cv}: mean square error on the validation set (5-fold cross-validation), R^2_{test}: coefficient of multiple determination on the test set, MSE_{test}: mean square error on the validation set.

for future research aimed at optimizing baijiu aging technology, thereby saving labor. Additionally, the developed methodology has the potential to be extended to other similar fields, such as wine and beer, thus playing a more significant role. Furthermore, while the integrated Lasso and Relaxed Lasso framework is advantageous for high-dimensional data with collinearity, offering improved feature selection, interpretability, and robustness, it has limitations, including computational complexity for large datasets, risk of incorrect variable selection with highly correlated covariates, and potential information loss with dependent important covariates. Careful attention is needed for its assumptions and computational demands, especially with large or highly nonlinear datasets.

Data availability statement

Data included in article/supp. material/referenced in article.

Funding

This work was financially supported by Sichuan Science and Technology Program, China [grant numbers 2020YFSY0006]

CRediT authorship contribution statement

Dongyue An: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Liangyan Wang:** Writing – review & editing, Supervision, Methodology, Investigation. **Jiang He:** Methodology, Investigation. **Yuejin Hua:** Supervision, Resources, Project administration, Investigation, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found at file: supplementary data.

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e36871.

References

[1] W. Fan, M.C. Qian, Headspace solid phase microextraction and gas Chromatography–Olfactometry dilution analysis of young and aged Chinese "yanghe daqu" liquors, J. Agric. Food Chem. 53 (20) (2005) 7931–7938.
[2] J. Hong, W. Tian, D. Zhao, Research progress of trace components in sesame-aroma type of baijiu, Food Res. Int. 137 (2020) 109695.
[3] Y. Li, S. Fan, A. Li, G. Liu, W. Lu, B. Yang, F. Wang, X. Zhang, X. Gao, Z. Lu, N. Su, G. Wang, Y. Liu, X. Ji, P. Xin, G. Li, D. Wang, F. Lu, Q. Zhong, Vintage analysis of Chinese Baijiu by GC and H-1 NMR combined with multivariable analysis, Food Chem 360 (2021) 129937.
[4] J.M. Wetherill, Gamma irradiation of food, Can. J. Public Health 56 (12) (1965) 521–524.
[5] A.C. Chang, Study of ultrasonic wave treatments for accelerating the aging process in a rice alcoholic beverage, Food Chem. 92 (2) (2005) 337–342.
[6] W. Jia, Z.B. Fan, A. Du, Y.L. Li, R. Zhang, Q.Y. Shi, L. Shi, X.G. Chu, Recent advances in Baijiu analysis by chromatography based technology-A review, Food Chem 324 (2020) 126899.
[7] M.L. Xu, S.M. Zhu, Y. Yu, Quality assessment of Chinese liquor with different ages and prediction analysis based on gas chromatography and electronic nose, Sci. Rep. 7 (1) (2017) 6541.
[8] S. Chen, J.L. Lu, M.C. Qian, H.K. He, A.J. Li, J. Zhang, X.M. Shen, J.J. Gao, Y. Xu, Untargeted headspace-gas chromatography-ion mobility spectrometry in combination with chemometrics for detecting the age of Chinese liquor (baijiu), Foods 10 (11) (2021) 2888.
[9] Y. Zhang, J. Gu, C. Ma, Y. Wu, L. Li, C. Zhu, H. Gao, Z. Yang, Y. Shang, C. Wang, G. Chen, Flavor classification and year prediction of Chinese Baijiu by time-resolved fluorescence, Appl. Opt. 60 (19) (2021) 5480–5487.
[10] J.C. Jiang, L.F. Zhang, X.D. Zhang, S. Liang, D.Q. Ji, Research of 60Co-γ irradiation effect on the quality of white wine based on BP neural network, Industry Control and Applications 34 (10) (2015) 16–19.
[11] W. Jia, Y. Li, A. Du, Z. Fan, R. Zhang, L. Shi, C. Luo, K. Feng, J. Chang, X. Chu, Foodomics analysis of natural aging and gamma irradiation maturation in Chinese distilled Baijiu by UPLC-Orbitrap-MS/MS, Food Chem 315 (2020) 126308.
[12] R. Tibshirani, Regression shrinkage and selection via the lasso, J. Roy. Stat. Soc. b-methodological 58 (1996) 267–288.
[13] N. Meinshausen, Relaxed lasso, Comput. Stat. Data Anal. 52 (1) (2007) 374–393.
[14] R. Tibshirani, Regression shrinkage and selection via the lasso, J. Roy. Stat. Soc. B 58 (1) (1996) 267–288.
[15] R. Muthukrishnan, R. Rohini, LASSO: a feature selection technique in predictive modeling for machine learning, IEEE (2016) 18–20.
[16] S. Chatterjee, K. Steinhaeuser, A. Banerjee, S. Chatterjee, A. Ganguly, Sparse group lasso: consistency and climate applications, Sdm (2012) 47–58.
[17] S. Gupta, R.E.C. Lee, J.R. Faeder, Parallel Tempering with Lasso for model reduction in systems biology, PLoS Comput. Biol. 16 (3) (2020) e1007669.
[18] C. Cui, D. Wang, High dimensional data regression using Lasso model and neural networks with random weights, Inf. Sci. 372 (2016) 505–517.
[19] C. Kang, Y. Huo, L. Xin, B. Tian, B. Yu, Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine, J. Theor. Biol. 463 (2019) 77–91.

[20] L. Freijeiro-González, M. Febrero-Bande, W. González-Manteiga, A critical review of LASSO and its derivatives for variable selection under dependence among covariates, Int. Stat. Rev. 90 (1) (2022) 118–145.
[21] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Variables selection methods in near-infrared spectroscopy, Anal. Chim. Acta 667 (1) (2010) 14–32.
[22] M. Farrés, S. Platikanov, S. Tsakovski, R. Tauler, Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation, J. Chemometr. 29 (10) (2015) 528–536.
[23] C. Kong, Z. Wang, Determination of volatile compounds in cabernet sauvignon wine by static headspace extraction coupled with GC-MS, Liquor-making Sci. Technol. (2012) 112–115, 09.
[24] M. Basalekou, C. Pappas, Y. Kotseridis, P.A. Tarantilis, E. Kontaxakis, S. Kallithraka, Red wine age estimation by the alteration of its color parameters: fourier transform infrared spectroscopy as a tool to monitor wine maturation time, Journal of Analytical Methods in Chemistry 2017 (2017) 5767613.
[25] Y.H. Deng, A. Xiong, K. Zhao, Y.R. Hu, B.S. Kuang, X. Xiong, Z.L. Yang, Y.G. Yu, Q. Zheng, Mechanisms of the regulation of ester balance between oxidation and esterification in aged Baijiu, Sci. Rep. 10 (1) (2020) 17169.
[26] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Ann. Stat. 32 (2) (2004) 407–499.
[27] R. Fildes, D.A. Belsley, Conditioning diagnostics: collinearity and weak data in regression, J. Oper. Res. Soc. 44 (1) (1993) 88–89.
[28] M. Prakash, J.K. Sarin, L. Rieppo, I.O. Afara, J. Toyras, Optimal regression method for near-infrared spectroscopic evaluation of articular cartilage, Appl. Spectrosc. 71 (10) (2017) 2253–2262.
[29] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. Roy. Stat. Soc. B 67 (2) (2005) 301–320.
[30] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. 29 (5) (2001) 1189–1232.
[31] M. Xu, Y. Yu, H. Ramaswamy, S. Zhu, Characterization of Chinese liquor aroma components during aging process and liquor age discrimination using gas chromatography combined with multivariable statistics, Sci. Rep. 7 (1) (2017) 39671.
[32] J. Xu, M. Chen, X. Liu, Changes in main trace components of different alcoholic crude Chinese spirits during storage, Int. J. Food Eng. 14 (4) (2018) 20170083.
[33] F.Q. Fqs, Wang X. Safety, X. Song, L. Zhu, X. Geng, F. Zheng, Q. Zhao, X. Sun, D. Zhao, S. Feng, M. Zhao, B. Sun, Unraveling the acetals as ageing markers of Chinese Highland Qingke Baijiu using comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry combined with metabolomics approach, Food Quality and Safety 5 (2021) 1–8.
[34] W. Fan, M.C. Qian, Characterization of aroma compounds of Chinese “Wuliangye” and “jiannanchun” liquors by aroma extract dilution analysis, J. Agric. Food Chem. 54 (7) (2006) 2695–2704.
[35] J. He, Q. Chen, X. Jia, Y. Wang, M. Huang, G.X. Wang, H. Chen, P. Gao, The effects of gamma irradiation and natural aging on the composition of Nongxiangxing baijiu, J. Food Process. Preserv. 46 (12) (2022) e17146.
[36] J. Zheng, Z. He, K. Yang, Z. Liu, D. Zhao, M.C. Qian, Volatile analysis of Wuliangye baijiu by LiChrolut EN SPE fractionation coupled with comprehensive GC×GC-TOFMS, Molecules 27 (4) (2022) 1318.
[37] Y. Fan, B.-Z. Han, Functional Microorganisms Associated with Baijiu Fermentation, Springer Nature, Singapore, 2023, pp. 503–545.
[38] W.E.I. Qunshu, Y. Yong, C. Yu, F. Kun, L.I.U. Peihua, O.U. Zhifeng, L.I. Jianbin, Accelerating aging of rice-flavor Baijiu by ultrasonic wave, China Brew. 36 (10) (2017) 66–70.
[39] L. Zhu, X. Wang, X. Song, F. Zheng, H. Li, F. Chen, Y. Zhang, F. Zhang, Evolution of the key odorants and aroma profiles in traditional Laowuzeng baijiu during its one-year ageing, Food Chem 310 (2020) 125898.
[40] Y. Ma, S. Zhang, M. Li, H. Qiao, Mathematical modeling for identification of fen-flavor liquor aging time, Food Sci. (N. Y.) 33 (10) (2012) 184–189.