bioRxiv preprint doi: https://doi.org/10.1101/2022.02.05.479235; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The usefulness of sparse k-means in metabolomics data: An example from breast cancer data

Misa Goudo[1], Masahiro Sugimoto[3,4], Satoru Hiwa[2], Tomoyuki Hiroyasu[2,*]

1 Graduate School of Life and Medical Sciences, Doshisha University, Kyoto, Japan
2 Department of Biomedical Sciences and Informatics, Doshisha University, Kyoto, Japan
3 Institute for Medical Science, Tokyo Medical University, Tokyo, Japan
4 Institute for Advanced Biosciences, Keio University, Yamagata, Japan

*** tomo@is.doshisha.ac.jp**

Abstract

In processing metabolomics data, multidimensional quantitative data from thousands of metabolites are often sparse, that is, only a small fraction of metabolites are relevant to the phenotype of interest. Clustering is therefore used to discover subtypes from omics data. Sparse processing, which selects important metabolites from the total omics data, is an effective clustering technique. This study investigated the effectiveness of sparse k-means for metabolomics data. Specifically, sparse k-means was used to cluster blood lipid metabolite data of breast cancer patients in two studies: (1) before and after menopause, and (2) pre- and postoperative chemotherapy. In both cases, sparse k-means showed comparable discrimination accuracy with fewer metabolites than k-means. Furthermore, when the L1 norm values were varied, no significant changes were observed. The mean silhouette coefficients of sparse k-means and k-means were (1) 0.38 ± 0.14 (S.D.) and 0.17 ± 0.01, (2) 0.38 ± 0.07 and 0.17 ± 0.01, indicating that feature selection using sparse k-means can improve clustering results. In addition, metabolite selection using sparse k-means was consistent regardless of the test data or the constrained value of the L1 norm, indicating robustness.

keyword

Sparse, k-means clustering, clustering, metabolomics, multifactor dimensionality reduction

1 Introduction

Recent advances in omics technologies have made it possible to comprehensively quantify molecules in living organisms and generate extensive data, such as proteomes and transcriptomes [1]. By observing changes in hundreds to thousands of molecular-level networks, it has become possible to comprehensively understand the etiology and pathogenesis of various diseases [2], search for diagnostic and predictive markers, and search for therapeutic targets based on patterns of multiple molecules rather than single molecules [3]. Clustering is used to interpret multidimensional quantitative omics data, where highly similar data can be extracted as clusters. It has

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.05.479235; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

been used to search for new subtypes of diseases and classify data using molecular aggregates [4]. Clustering is often used to identify subtypes of a particular disease by integrating methylation and genomic information with the transcriptome [5,6]. Whether in single-omics or multi-omics, clustering is considered one of the most effective methods to consolidate various information inherent in the data.

There are various clustering methods, of which the most widely used are hierarchical clustering and k-means clustering, unsupervised learning methods [7,8]. According to a predefined distance measure, k-means clustering classifies practical information into k clusters [8]. A similar technique is Partitioning Around Medoids (PAM), which introduces dissimilarity (dissimilarity matrix). Other single-layer neural network methods such as Self-Organizing Maps (SOM) [9] have also been developed. These methods classify the similarity of samples and rank the importance of observed variables based on their clustering contribution and use the top-ranked variables for biochemical considerations and markers. However, all of these methods use all observed variables; thus, many variables are involved in extracting information for each cluster, which presents computational challenges. In contrast, sparse k-means, an improved version of k-means [10], allows clustering with fewer observables, eliminating information that has little relevance to the cluster, leaving only important variables for cluster formation.

Metabolomics is a comprehensive and simultaneous molecular profiling method that captures metabolic data. Nuclear magnetic resonance (NMR) and mass spectrometry (MS) are widely used in metabolomics. The instruments used in these techniques are becoming increasingly sensitive, enabling the simultaneous measurement of hundreds to thousands of metabolites. The high sensitive mass spectrometers have enabled the quantification of a wide variety of metabolites, which increased the dimension of the observed information [11,12]. Multivariate analysis, such as principal component analysis (PCA), is a standard method to obtain a complete picture of measured metabolites [13]. Predictive models, such as partial least squares regression (PLS-DA) and two-way orthogonal partial least squares (O2PLS), are widely used to rank molecules that contribute to classification and are correlated with quantitative values [14]. K-means metabolic data analysis has been applied in various fields. For example, in an Irish cohort, triglycerides and glucose in fasting serum samples were clustered using k-means to identify individual dietary habits [15]. In another study, k-means clustered metabolic profiles from liquid chromatography-tandem mass spectrometry (LC-MS/MS) data were used from red wine [16].

Thus, to discover new biochemical mechanisms and establish molecular markers, it is necessary to extract representative metabolites among the metabolites detected. Furthermore, if the discrimination accuracy is similar or improved using only a subset of features compared to all features, these subset metabolites can explain the focused phenotype. Therefore it is essential to extract a subset of metabolites for facilitating the discovery of new biochemical mechanisms and molecular markers.

Among the clustering methods, sparse k-means can identify essential features simultaneously with clustering. Witten and Tibshirani developed an alternative method that identifies features using lasso-type penalties based on conventional k-means [10]. Sparse k-means can also be effective for metabolomics data; however, there are few metabolomics studies that utilize sparse k-means [17]. Therefore, it is necessary to investigate the impact of each clustering process on the accuracy of the data and establish an analytical protocol to evaluate the applicability of sparse k-means in metabolomic data analysis.

This study conducted two comparisons of sparse k-means and k-means using quantitative blood lipid data from breast cancer patients: one for premenopausal and menopausal, and another for preoperative and postoperative chemotherapy.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.05.479235; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2 Materials and Methods

2.1 Mini-review of existing methods

Two unsupervised clustering methods, k-means and sparse k-means, were compared. The k-means clustering method is a well-known algorithm [18]. The number of clusters (k), which is the number of clusters to be extracted, was pre-specified by a data analyst. The clusters were determined by minimizing the within-cluster sum of squares (WCSS). This operation is synonymous with maximizing the between-cluster sum of squares (BCSS). In contrast, sparse k-means is an algorithm for clustering using features selected during adaptation [10], where a weight of ≥0 is assigned to each metabolite. The sparse k-means is a function of the sum of the squares between the clusters multiplied by the weights of the variables, as shown in Equation 1. The clusters are determined by maximizing this sum [10].

$$\begin{gathered} maximize_{C_1 \ldots C_k, w}\{\sum_{j=1}^{p} w_j(\frac{1}{n}\sum_{i=1}^{n}\sum_{i'=1}^{n} d_{i,i',j} - \sum_{k=1}^{K}\frac{1}{n_k}\sum_{i,i' \in C_k} d_{i,i',j})\}; \\ d_{i,i',j} = (X_{i,j} - X_{i',j})^2; \\ subjectto \| w \|_2 \leq 1, \| w \|_1 \leq s, w_j \geq 0 \forall j, \end{gathered} \tag{1}$$

where X is an n (sample) $\times p$ (variable) matrix, i is the number of samples, j is the number of variables, w is the weight of the variable, w_j is the weight of the jth variable, k is the number of clusters, n_k is the number of samples in the kth cluster number, and C_k is the kth cluster. The parameter s is the sum of the weights of the features. If s is large, more features are selected, and if s is small, fewer features are selected.

2.2 Evaluation indicators and data used in this study

2.2.1 Evaluation indicators

Four evaluation metrics were calculated: accuracy, sensitivity, and specificity to validate the data classification and the silhouette coefficient to measure the cohesiveness of the clusters. The first three metrics were calculated from the mixing matrix. Accuracy was calculated by(true positive [TP] + true negative [TN]) / (TP + TN + false positive [FP] + false negative [FN]), sensitivity by TP / (TP + TN), and specificity by TN / (TN + FP). The silhouette coefficient is an index for evaluating clustering performance, based on the idea that clustering results are condensed within clusters and distant in different clusters. The method used to calculate the silhouette coefficient is given by Equation 2.

$$\begin{gathered} a^{(i)} = \frac{1}{|C_{in}-1|}\sum_{x^j \in C_{in}} \| x^i - x^j \|, \\ b^{(i)} = \frac{1}{|C_{near}|}\sum_{x^j \in C_{near}} \| x^i - x^j \|, \\ s^i = \frac{b^i - a^i}{max(b^i, a^i)}. \end{gathered} \tag{2}$$

Here, $a^{(i)}$ is the average distance from sample i to all samples in the same cluster (C_{in}) and $b^{(i)}$ is the average distance from sample i to all samples of the closest cluster (C_{in}). The silhouette coefficient is calculated from -1 to 1, where 1 indicates that the sample is farther away from neighboring clusters and -1 indicates that the sample is far from the center of its cluster, which means that the sample may be in the wrong cluster. A silhouette coefficient of 0 indicates that the sample is on or near the decision boundary. The average silhouette coefficients for all data were used to assess the validity of the clustering [19].

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.05.479235; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.2.2 Data selection

The MTBLS92 dataset from MetaboLights (`https://www.ebi.ac.uk/metabolights/`)was used in this study. This dataset was obtained from a multicenter, randomized, phase Ⅲ trial in which breast cancer patients were randomly assigned to one of the following preoperative neoadjuvant chemotherapies (NACs).

1. 4×epirubicine+cyclophosphamide(EC)→docetaxel(D)
2. 4×EC→4×D/capecitabine(C)
3. 4×EC→4×D→C

Blood samples were collected before NAC (baseline, BL) and in the non-fasting state after NAC at the time of surgery. Serum lipid levels were quantified by liquid chromatography-mass spectrometry (LC) and fatty acids by gas chromatography (GC), and a flame ionization detector (FID) detector. Substances were quantified based on the peak intensity ratio and using internal standards for each pretreatment and were identified by a library search developed internally. The dataset consisted of 253 individuals and 240 metabolites. We refer the reader to the original paper for details of the study design, sample collection, and processing [20]. The number of samples used in this study is listed in Table 1.

Table 1. Number of samples

comparison number	sample contents	samples[1]
1	premenopausal and menopausal	142/111
2	pre- and postoperative chemotherapy	148/105

[1] Any of the three NACs were administered. The data for the premenopausal and menopausal comparisons were from BL and after NAC at the time of surgery were used for the combined data.

2.3 Clustering process

Hilvo et al. showed a difference in the metabolite profiles of patients receiving NAC before and after treatment. They also showed a difference in the menopausal period [20]. This study used two-class clustering to separate pre- and postoperative chemotherapy and premenopausal and menopausal patients based on these results. The data was divided into two datasets: one for class determination and the other for prediction to perform two-class clustering. K-means and sparse k-means clustering were applied to the datasets used for class determination, and the centers of the two clusters were determined. The sum of the weights of the features (s), the option with the highest accuracy when combined with the teacher data, was adopted. The most represented label determined the label of each cluster in each respective cluster. The labels of close centers categorized the dataset for prediction. Finally, the accuracy, sensitivity, specificity, and average of the silhouette coefficients were calculated for performance evaluation. In the premenopausal and menopausal classification, values were assigned as positive and negative, respectively; in the pre- and postoperative chemotherapy clustering, values were assigned as positive and negative, respectively. All statistical analyses used in this study were based on the Welch's t-test ($P < 0.05$). The dataset was split into two datasets by a stratified extraction method to determine 4/5 and predict 1/5 classes. The dataset was split ten times, and the same dataset was used for each method to minimize bias. Missing values in the data were interpolated using half of the mean value of each variable. In addition, the values were log2-transformed and

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.05.479235; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

normalized per sample, and then the normalized values per variable were used. The default number of initial searches for each method was 10, and the number of clustering attempts was set to 100. Sparse k-means requires the sum of the weights of the features (denoted by s in Equation 1). In this study, we did not determine it uniquely but varied it in increments of 0.48 to divide the data into 40 parts between 1.1 and 20. The minimum value of s is 1.1, and at $s = 20$, the results were stable in preliminary trials. Out of ten trials, the variables whose features were nonzero and were ranked according to the most significant coefficient, and the total value of the ten rankings was calculated. For example, ten was calculated if the variable was ranked first ten times.

2.4 Software implementation

K-means and sparse k-means were computed using R (v 4.0.3; R Development Core Team, Vienna, Austria). K-means was computed using the standard R package "stats" (v 4.0.3). Sparse k-means was computed using the R package "sparcl" (v 1.0.4) [21]. The program will be released upon request.

3 Results

3.1 Differences in the selected features depend on the constraints of the weights in sparse k-means

Once the features are selected, the data can be represented in a lower dimension, which is necessary for determining metabolites for practical applications. The sparse k-means algorithm is characterized by features becoming sparse after clustering. However, obtaining a sparse result depends on the hyperparameter setting of the sparse k-means, that is, the sum of the weights of each feature (s). By changing s, the weights attached to each feature will change, also changing the number of features with zero weight. Therefore, in sparse k-means, the feature selected varies greatly depending on s.

Figure 1 shows the relationship between s and the number of metabolites selected using sparse k-means. For high values of s, the selected metabolites included all metabolites (240), indicating that selected features were not sparse. In the two-class clustering dividing premenopausal and menopausal classes, all metabolites were selected when $s \geq 12.73$, indicating that sparsity was not reached. When $s < 12.73$ was sparse, i.e. only some metabolites were selected. Two-class clustering based on pre- or postoperative chemotherapy also yielded the same results as those based on menopausal status. In both clustering methods, features were limited with $s < 12.73$, confirming the sparsity of selected features.

3.2 Comparison of sparse k-means and k-means classification accuracy

This study used accuracy, sensitivity, and specificity as indices for class classification. In addition, average silhouette coefficients were used as indicators of cluster cohesion. The classification accuracy was determined based on the distance between the centers of the clusters. Figure 2 shows the evaluation indices for the sparse k-means and k-means classification, resulting in a higher accuracy than the clusters' data. In the two-class clustering separating premenopausal and menopausal classes, the mean silhouette coefficients of sparse k-means and k-means were 0.38 ± 0.14 (S.D.) and 0.17 ± 0.01, respectively, indicating a significant difference. No significant differences were found for the other indices. Similarly, in the two-class clustering separating pre- and postoperative chemotherapy classes, the mean silhouette coefficients of sparse k-means

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.05.479235; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

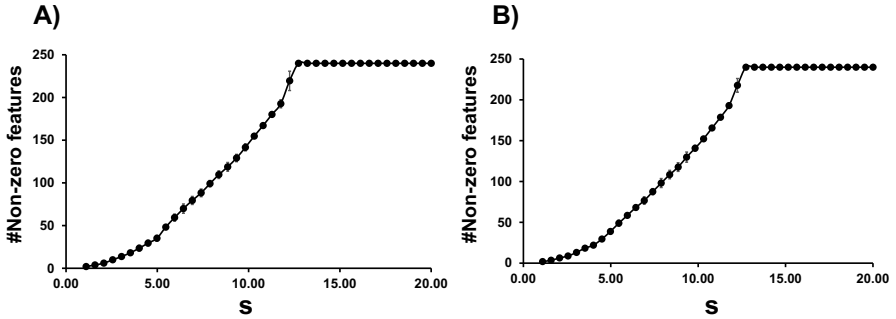

Figure 1. Relationship between the sum of the weights of the features (s) and the number of variables with nonzero weights for two-class clustering. Results for (**A**) premenopausal and menopausal, and (**B**) pre- and postoperative chemotherapy. The x-axis shows s, and the y-axis shows the number of variables with nonzero weights. Error bars indicate standard deviations.

and k-means were 0.38 ± 0.07 and 0.17 ± 0.01, respectively, indicating a significant difference. Otherwise, no significant differences were observed. Figure 3 shows the accuracy of the created clusters compared with the learning data. The labels of the learning data were assigned to the center of the clusters. There was a significant difference between the two classifications separating premenopausal and menopausal classes, with a sparse k-means of 0.62 ± 0.01 and k-means of 0.60 ± 0.01. In the two-class clustering separating the pre- and postoperative chemotherapy classes, the sparse k-means and k-means were 0.57 ± 0.02 and 0.56 ± 0.02, respectively, and no significant difference was obtained. A variable (row) × sample (column) matrix was created for each TP, FP, FN, and TN result obtained from the mixing matrix. Figure 4 shows the matrices for k-means and sparse k-means clustering. For each comparison, the sparse k-means selected metabolites that characterized the cluster better than k-means because the heat map colors differed between the labels. This suggests that metabolites selected by sparse k-means capture the characteristics of the clusters.

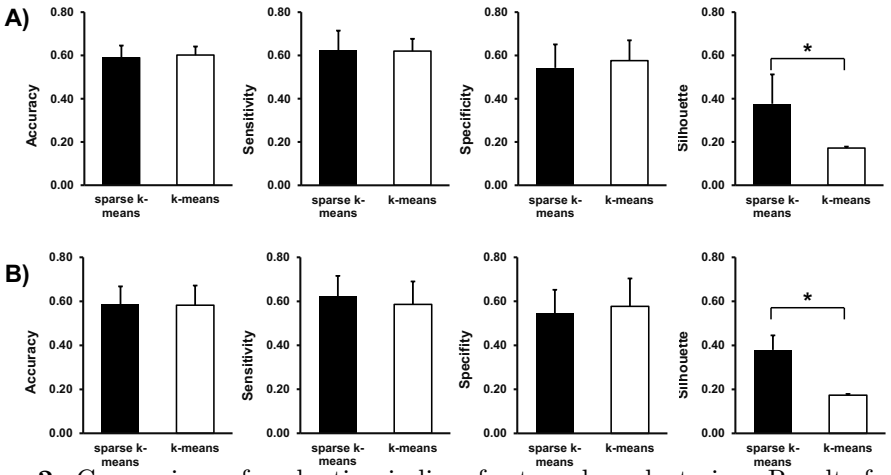

Figure 2. Comparison of evaluation indices for two-class clustering. Results for (**A**) premenopausal and menopausal, and (**B**) pre- and postoperative chemotherapy. From left to right, both panels show the accuracy, sensitivity, specificity, and silhouette coefficient. Error bars indicate standard deviation obtained from 10 trials. $*P < 0.05$ (Welch's t-test).

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.05.479235; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

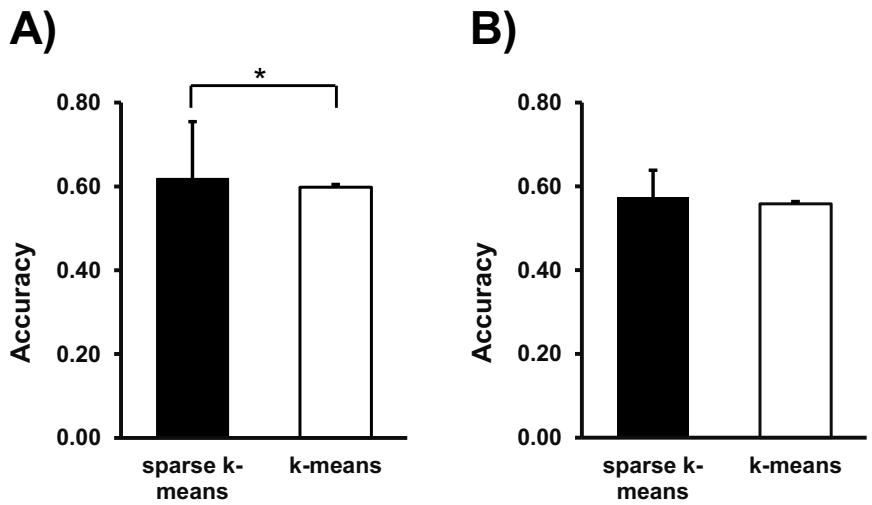

Figure 3. Accuracy of cluster labeling for (**A**) premenopausal and menopausal, and (**B**) pre- and postoperative chemotherapy two-class clustering. Error bars indicate standard deviation obtained from 10 trials. $*P < 0.05$ (Welch's t-test).

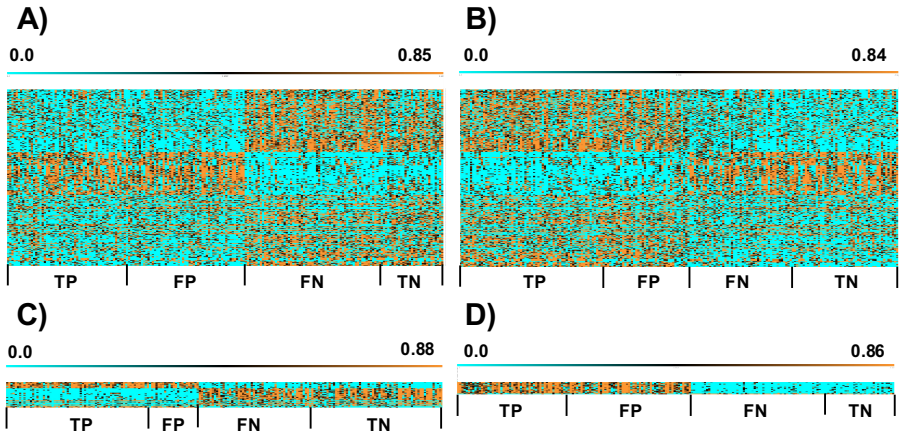

Figure 4. Matrices for (**A**, **C**) premenopausal and menopausal, and (**B**, **D**) pre- and postoperative chemotherapy two-class clustering. In (**A**) and (**B**), k-means produced a matrix of 240 (metabolites) $\times$ 202 (samples); in (**C**), sparse k-means ($s = 4.49$) produced a matrix of 35 (metabolites) $\times$ 202 (samples); in (**D**), sparse k-means ($s = 3.52$) produced a matrix of 18 (metabolites) $\times$ 202 (samples).All panels show, from left to right, True positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN), where premenopausal and preoperative chemotherapy are considered positive and menopausal and postoperative are considered negative. The scale value indicates the value of the metabolic profiles.

3.3 selected metabolites

Table 2 and 3 show the list of metabolites selected by sparse k-means for comparison 1 (premenopausal and menopausal) and comparison 2 (pre- and postoperative chemotherapy), respectively. In Table 2, unidentified_42 (unidentified substance) was selected nine times out of 10, triglycerides (TG; 14:0/16:0/18:1) were selected seven times, and the total ranking value was the smallest among all metabolite profiles. In Table 3, the median number of variables was 33.5, with a minimum of 6 and a

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.05.479235; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

maximum of 89 in 10 trials; TG (14:0/16:0/18:0), TG (16:0/16:0/18:0), phosphatidylcholine (PC), and sphingomyelin (SM) were included in the majority of trials. All other molecules were selected at least twice, with the exception of ceramide, (Cer; d18:1/17:0), TG (16:0/18:1/20:1), TG (18:0/18:1/18:1), PC (34:1e) | PE (phosphatidylethanolamine (37:1e), and a few unidentified substances.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.05.479235; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 2. Variables selected in comparison 1 (top 15)

metabolite	number of times selected	total ranking value
unidentified_42	9	46
TG(14:0/16:0/18:1)	7	24
TG(48:2)	6	175
TG(16:0/18:1/16:0)	6	146
TG(14:0/16:0/18:0)	6	134
TG(16:0/16:0/18:0)	6	159
TG(46:1)	6	192
unidentified_18	6	258
TG(14:0/16:0/16:0) | TG(16:0/18:0/12:0)	6	267
SM(d18:1/22:1)	6	281
unidentified_47	6	278
TG(14:0/18:1/18:1) | TG(16:0/16:1/18:1)	6	273
unidentified_53	6	456
TG(49:1)	5	190
TG(16:0/18:0/18:1)	5	188

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.05.479235; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 3. Variables selected in comparison 2 (top 15)

metabolite	number of times selected	total ranking value
TG(14:0/16:0/18:1)	10	28
TG(14:0/16:0/18:0)	10	38
unidentified_42	10	41
TG(16:0/16:0/18:0)	10	67
unidentified_47	10	91
TG(14:0/16:0/16:0) | TG(16:0/18:0/12:0)	9	66
TG(46:1)	8	63
TG(16:0/18:1/16:0)	8	90
TG(48:2)	8	108
TG(16:0/18:0/18:1)	8	138
unidentified_18	8	176
TG(47:1)	8	179
Custom adduct of 796,742 m/z - TG(14:0/16:0/16:0)+TG(16:0/18:0/12:0)	7	111
unidentified_71	7	125
TG(44:1)	7	146

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.05.479235; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4 Discussion

4.1 Clustering Accuracy

This study used sparse k-means and k-means to investigate classification accuracy and cluster cohesion by performing two-class clustering using two comparisons, one for premenopausal and menopausal and the other for pre- and postoperative chemotherapy. There was no significant difference in classification accuracy between sparse k-means and k-means, but sparse k-means showed a significant difference in cluster cohesion due to variable selection. Sparse k-means was more in line with the teacher data in the premenopausal and menopausal comparison than k-means. This result suggests that sparse k-means can perform clustering and feature selection, providing more information than k-means. Originally, it was necessary to measure the gap statistics to determine s [10, 17, 22]. In this study, this was not considered. In addition, we did not compare sparse k-means with other methods, such as PLS-DA. Future studies may investigate the effects of different clustering methods in metabolite selection. Gal et al. compared the metabolic signatures of breast cancer obtained using five unsupervised learning methods, including sparse k-means [17]. They determined the s in sparse k-means using the gap statistic, whereas, in this paper, we compared the differences in metabolites with changes in s. We also examined the accuracy of the classification based on its proximity to the center of the cluster. The k-means method uses all the features for clustering, so all the observed metabolites can be divided into similar groups. On the other hand, the sparse k-means method is feature-selective to extract specific metabolites within similar groups.

4.2 Metabolites selected by sparse k-means

Although the number of variables selected by sparse k-means varied among trials, most variables were selected in multiple trials. In particular, only triglycerides were selected in the smallest number of trials in the pre- and postoperative comparisons. In breast cancer patients treated with NAC, the blood lipid profile remains unchanged, but changes in thiobarbituric acid reactive substances (TBARS) and superoxide dismutase (SOD) indicate that oxidative stress in the body is significantly altered [23], resulting in a change in TG levels. Epirubicin, an anthracycline anticancer drug, was included in the NAC. It has been reported that some SMs may contribute to the activation efficiency of anthracycline uptake [24]. PC has been shown to protect against peripheral neurotoxicity, one of the side effects caused by docetaxel [25]. Changes in the blood profile with NAC treatment are also speculated. In breast cancer subtypes, triple-negative tumors contain more specific PCs [26], suggesting that the baseline may differ in each subtype rather than uniformly changing in all cases. In contrast, there are reports that blood PC in breast cancer patients decreases during anticancer drug treatment [27]. PC itself can be used as a marker for predicting cervical cancer susceptibility to NAC [28]. Therefore, changes in the blood profile may vary from individual to individual.

5 Conclusions

In metabolomics data analysis, k-means and hierarchical clustering are used for discriminating multiple groups. However, it is difficult to examine metabolic characteristics in a vast number of metabolites. Therefore, clustering with a few metabolites and feature selection is essential for analysis. Although sparse k-means, which simultaneously performs feature selection and clustering, is one method to solve

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.05.479235; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

this problem, it has not been widely adopted for metabolomics analysis. This study investigates the effect of constraint values on the weights of sparse k-means when applied to metabolomics data and examines its usefulness for future analysis. In this study, two breast cancer metabolomics data classifications were conducted using k-means and sparse k-means. No statistically significant difference in classification accuracy was found between sparse k-means and k-means clustering. The sparse k-means method selected variables that explained the clusters by setting appropriate weight constraint values. In addition, the selected metabolites were suitable for discriminating the given problems. Because feature selection can be made simultaneously with clustering, the sparse k-means method is useful for analyzing metabolomics data.

References

1. Bensu Karahalil. Overview of Systems Biology and Omics Technologies. *Current Medicinal Chemistry*, 23(37):4221–4230.

2. Edwin K. Silverman, Harald H. H. W. Schmidt, Eleni Anastasiadou, Lucia Altucci, Marco Angelini, Lina Badimon, Jean-Luc Balligand, Giuditta Benincasa, Giovambattista Capasso, Federica Conte, Antonella Di Costanzo, Lorenzo Farina, Giulia Fiscon, Laurent Gatto, Michele Gentili, Joseph Loscalzo, Cinzia Marchese, Claudio Napoli, Paola Paci, Manuela Petti, John Quackenbush, Paolo Tieri, Davide Viggiano, Gemma Vilahur, Kimberly Glass, and Jan Baumbach. Molecular networks in Network Medicine: Development and applications. *Wiley Interdisciplinary Reviews. Systems Biology and Medicine*, 12(6):e1489, 2020.

3. Nikos Koundouros and George Poulogiannis. Reprogramming of fatty acid metabolism in cancer. *British Journal of Cancer*, 122(1):4–22, 2020.

4. T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lønning, and A. L. Børresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*, 98(19):10869–10874, 2001.

5. SriGanesh Jammula, Annalise C. Katz-Summercorn, Xiaodun Li, Constanza Linossi, Elizabeth Smyth, Sarah Killcoyne, Daniele Biasci, Vinod V. Subash, Sujath Abbas, Adrienn Blasko, Ginny Devonshire, Amber Grantham, Filip Wronowski, Maria O'Donovan, Nicola Grehan, Matthew D. Eldridge, Simon Tavaré, Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) consortium, and Rebecca C. Fitzgerald. Identification of Subtypes of Barrett's Esophagus and Esophageal Adenocarcinoma Based on DNA Methylation Profiles and Integration of Transcriptome and Genome Data. *Gastroenterology*, 158(6):1682–1697.e1, 2020.

6. Prabhakar Chalise, Devin C. Koestler, Milan Bimali, Qing Yu, and Brooke L. Fridley. Integrative clustering methods for high-dimensional molecular data. *Translational Cancer Research*, 3(3):202–216, 2014.

7. Patrick K. Kimes, Yufeng Liu, David Neil Hayes, and James Stephen Marron. Statistical significance for hierarchical clustering. *Biometrics*, 73(3):811–821, 2017.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.05.479235; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

8. J. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 5.1:281–298, 1967.

9. Benjamin H Ott. A convergence criterion for self-organizing maps. *Dissertations and Master's Theses (Campus Access)*, pages 1–70, 2012.

10. Daniela M. Witten and Robert Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726, 2010.

11. John L. Markley, Rafael Brüschweiler, Arthur S. Edison, Hamid R. Eghbalnia, Robert Powers, Daniel Raftery, and David S. Wishart. The future of NMR-based metabolomics. *Current Opinion in Biotechnology*, 43:34–40, 2017.

12. Minnie Jacob, Andreas L. Lopata, Majed Dasouki, and Anas M. Abdel Rahman. Metabolomics toward personalized medicine. *Mass Spectrometry Reviews*, 38(3):221–238, 2019.

13. Bradley Worley and Robert Powers. Multivariate Analysis in Metabolomics. *Current Metabolomics*, 1(1):92–107, 2013.

14. Tusharkanti Ghosh, Weiming Zhang, Debashis Ghosh, and Katerina Kechris. Predictive Modeling for Metabolomics Data. *Methods in molecular biology (Clifton, N.J.)*, 2104:313–336, 2020.

15. Anna Riedl, Elaine Hillesheim, Nina Wawro, Christa Meisinger, Annette Peters, Michael Roden, Florian Kronenberg, Christian Herder, Wolfgang Rathmann, Henry Völzke, Martin Reincke, Wolfgang Koenig, Henri Wallaschofski, Hannelore Daniel, Hans Hauner, Lorraine Brennan, and Jakob Linseisen. Evaluation of the Metabotype Concept Identified in an Irish Population in the German KORA Cohort Study. *Molecular Nutrition & Food Research*, 64(8):1900918, 2020.

16. Xin-Ke Zhang, Yi-Bin Lan, Yue Huang, Xu Zhao, and Chang-Qing Duan. Targeted metabolomics of anthocyanin derivatives during prolonged wine aging: Evolution, color contribution and aging prediction. *Food Chemistry*, 339:127795, 2021.

17. Jocelyn Gal, Caroline Bailleux, David Chardin, Thierry Pourcher, Julia Gilhodes, Lun Jing, Jean-Marie Guigonis, Jean-Marc Ferrero, Gerard Milano, Baharia Mograbi, Patrick Brest, Yann Chateau, Olivier Humbert, and Emmanuel Chamorey. Comparison of unsupervised machine-learning methods to identify metabolomic signatures in patients with localized breast cancer. *Computational and Structural Biotechnology Journal*, 18:1509–1524, 2020.

18. Amit Saxena, Mukesh Prasad, Akshansh Gupta, Neha Bharill, Om Prakash Patel, Aruna Tiwari, Meng Joo Er, Weiping Ding, and Chin-Teng Lin. A review of clustering techniques and developments. *Neurocomputing*, 267:664–681, 2017.

19. Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

20. Mika Hilvo, Stephan Gade, Tuulia Hyötyläinen, Valentina Nekljudova, Tuulikki Seppänen-Laakso, Marko Sysi-Aho, Michael Untch, Jens Huober, Gunter von Minckwitz, Carsten Denkert, Matej Orešič, and Sibylle Loibl. Monounsaturated fatty acids in serum triacylglycerols are associated with response to neoadjuvant

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.05.479235; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

chemotherapy in breast cancer patients. *International Journal of Cancer*, 134(7):1725–1733, 2014.

21. Daniela M. Witten, Robert Tibshirani, and Maintainer Daniela Witten. Package 'sparcl'. 2013.

22. Yujia Li, Xiangrui Zeng, Chien-Wei Lin, and George Tseng. Simultaneous Estimation of Number of Clusters and Feature Sparsity in Clustering High-Dimensional Data. *arXiv:1909.01930 [stat]*, 2019.

23. Luana Almeida Gonzaga, Thais Reis Silva de Paulo, Juliana Viezel, Laís Manata Vanzella, Ismael Forte Freitas Jr., and Luiz Carlos Marques Vanderlei. Changes in Cardiac Autonomic Modulation in Women with Breast Cancer Using Aromatase Inhibitors and the Relation with Biochemical Variables. *Arquivos Brasileiros de Cardiologia*, 112(5):555–563, 2019.

24. R J Veldman, S Zerp, W J van Blitterswijk, and M Verheij. N-hexanoyl-sphingomyelin potentiates in vitro doxorubicin cytotoxicity by enhancing its cellular influx. *British Journal of Cancer*, 90(4):917–925, 2004.

25. Sung Tae Kim, Eun Jung Kyung, Jung Sook Suh, Ho Sung Lee, Jun Ho Lee, Soo In Chae, Eon Sub Park, Yoon Hee Chung, Jinhyung Bae, Tae Jin Lee, Won Mo Lee, Uy Dong Sohn, and Ji Hoon Jeong. Phosphatidylcholine attenuated docetaxel-induced peripheral neurotoxicity in rats. *Drug and Chemical Toxicology*, 41(4):476–485, 2018.

26. Yuko Hosokawa, Noritaka Masaki, Shiro Takei, Makoto Horikawa, Shoko Matsushita, Eiji Sugiyama, Hiroyuki Ogura, Norihiko Shiiya, and Mitsutoshi Setou. Recurrent triple-negative breast cancer (TNBC) tissues contain a higher amount of phosphatidylcholine (32:1) than non-recurrent TNBC tissues. *PloS One*, 12(8):e0183724, 2017.

27. Xin Li, John Lim, Anand Kolatkar, Lisa Welter, Kathryn Waitman, Eun-Sil Shelley Hwang, Janice M. Lu, Peter Kuhn, and Jorge J. Nieva. Serum phosphatidylcholine is lower among breast cancer patients on systemic chemotherapy. *Journal of Clinical Oncology*, 35:e12571–e12571, 2017.

28. Ming-Zhu Yin, Shu Tan, Xia Li, Yan Hou, Guosheng Cao, Kang Li, Junping Kou, and Ge Lou. Identification of phosphatidylcholine and lysophosphatidylcholine as novel biomarkers for cervical cancers in a prospective cohort study. *Tumour Biology: The Journal of the International Society for Oncodevelopmental Biology and Medicine*, 37(4):5485–5492, 2016.