DIGITALCOMMONS @WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 15 | Issue 1 Article 20

5-2016

Compound Identification Using Penalized Linear Regression on Metabolomics

Ruiqi Liu
University of Louisville, r0liu009@louisville.edu

Dongfeng Wu
University of Louisville, dongfeng.wu@louisville.edu

Xiang Zhang
University of Louisville, xiang.zhang@louisville.edu

Seongho Kim
Wayne State University, mathan72@gmail.com

Compound Identification Using Penalized Linear Regression on Metabolomics

Cover Page Footnote
This work was supported by NSF grant DMS-1312603 and NIH grant 1RO1GM087735. The Biostatistics Core is supported, in part, by NIH Center Grant P30 CA022453 to the Karmanos Cancer Institute at Wayne State University.

Journal of Modern Applied Statistical Methods
May 2016, Vol. 15, No. 1, 373-388.

ISSN 1538 − 9472

Compound Identification Using Penalized Linear Regression on Metabolomics

Ruiqi Liu
University of Louisville
Louisville, KY

Dongfeng Wu
University of Louisville
Louisville, KY

Xiang Zhang
University of Louisville
Louisville, KY

Seongho Kim
Wayne State University
Detroit, MI

Compound identification is often achieved by matching the experimental mass spectra to the mass spectra stored in a reference library based on mass spectral similarity. Because the number of compounds in the reference library is much larger than the range of mass-to-charge ratio (*m*/*z*) values, so that the data become high dimensional data suffering from singularity. For this reason, penalized linear regressions such as ridge regression and the lasso are used instead of the ordinary least squares regression. Furthermore, two-step approaches using the dot product and Pearson's correlation along with the penalized linear regression are proposed in this study.

Keywords: Compound identification, mass spectral similarity, metabolomics, penalized linear regression

Introduction

One of the critical analyses on GC-MS data is compound identification, and it is often achieved by matching the experimental mass spectra to the mass spectra stored in a reference library based on mass spectral similarity (Stein & Scott, 1994). To improve the accuracy of compound identification, various algorithms measuring mass spectral similarity scores have been developed, such as dot product (Tabb, MacCoss, Wu, Anderson, & Yates, 2003; Beer, Barnea, Ziv, & Admon, 2004; Craig, Cortens, Fenyo, & Beavis, 2006; Frewen, Merrihew, Wu, Noble, &

Ruiqi Liu is a PhD student in the Department of Bioinformatics and Biostatistics. Email at: r0liu009@louisville.edu. Dr. Wu is an Associate Professor in the Department of Bioinformatics and Biostatistics. Email at: dongfeng.wu@louisville.edu. Dr. Zhang is a Professor in the Department of Chemistry. Email at: xiang.zhang@louisville.edu. Dr. Kim is the corresponding author and an Assistant Professor in the Department of Oncology. Email at: kimse@karmanos.org.

MacCoss, 2006), composite similarity (Stein & Scott, 1994), probability-based matching system (Atwater, Stauffer, McLafferty, & Peterson, 1985), Hertz similarity index (Hertz, Hites, & Biemann, 1971), normalized Euclidean distance (L2-norm) (Rasmussen & Isenhour, 1979; Stein & Scott 1994; Julian, Higgs, Gygi, & Hilton, 1998), absolute value distance (L1-norm) (Rasmussen & Isenhour, 1979; Beer et al., 2004), Fourier and wavelet-based composite similarity (Koo, Zhang, & Kim, 2011), and mixture partial and semi-partial correlation measures (Kim et al., 2012).

Because some compounds have mass spectral information that is similar to that of other compounds, an experimental query spectrum of these compounds is often matched to multiple mass spectra in the reference library with high similarity scores, impeding the high confidence compound identification. That is, the mass spectral similarity score of a true positive pair does not always have the top ranked score, and it is instead ranked as the second- or the third-highest similarity score with an ignorable difference from the top-ranked score.

In order to avoid the aforementioned issue, Kim et al. (2012) developed a novel similarity measure using partial and semi-partial correlations. The partial correlation can be seen as the pure relationship between two random variables after adjusting the effect of other random variables. On the other hand, the semi-partial correlation eliminates the effect of a fraction of other random variables, just adjusting the effect of one random variable from a total of two random variables. When it comes to compound identification, these partial and semi-partial correlations can be applied to calculate the mass spectral similarity score. By removing the effect of other mass spectra over the two mass spectra of interest, the unique relationship between the mass spectra can be extracted. Using partial and semi-partial correlations can obtain high accuracy of compound identification. Indeed, Koo, Kim, and Zhang (2013) recently compared among existing spectral similarity measures in terms of compound identification and concluded that mixture semi-partial correlation measure outperforms others. However, the performance of this method suffers from expensive calculation because the data are ultra-high-dimensional, which propels us to search for an alternative for compound identification.

Another way for compound identification is to use the multiple ordinary linear regression-based methods. In the context of linear regression, the response variable is an experimental mass spectrum (i.e., query) and all the compounds in the reference library are the independent variables. Each regression coefficient reflects the strength of their relationships with the response variable, so we could match the experimental compound with the reference compound which shows the strongest

connection. In particular, the coefficients of the multiple ordinary linear regressions are proportional to the semi-partial correlation coefficient, meaning that both methods will give us the same result if the maximal coefficient is considered only. In other words, the ordinary linear regression is a great alternative to the semi-partial correlation-based compound identification.

However, it is not feasible to apply ordinary linear regression in compound identification for two reasons. First, our data are high-dimensional data. The size of a reference library is much larger than the range of mass-to-charge ratio (m/z) values, and the number of variables becomes much larger than the number of samples so that the ordinary linear regression will suffer from singularity. Second, it is possible that different compounds have identical mass spectra, such as isomers. Note that isomers are compounds with the same molecular formula but different chemical structures. Because of the existence of isomers, several predictors are highly correlated to each other so that their correlation coefficients become almost one. This also causes ordinary linear regression to suffer from singularity.

In order to elude this difficulty, a penalized linear regression is introduced for the compound identification. Penalized linear regression can deal with high-dimensional data, and it is a trade-off between unbiasedness and a smaller estimation variance by putting a penalty constraint on coefficients. Different types of constrains will result in the lasso and ridge regression, which have L1-norm and L2-norm penalties, respectively. To improve the performance of penalized linear regression, two-step approaches are introduced using widely used mass spectral similarity scoring methods, either dot product or Pearson's correlations as the first step, and then penalized linear regression as the second step. Using the NIST mass spectral library, the performance of the proposed penalized linear regression approaches and two-step approaches with the dot product and Pearson's correlation are compared in terms of the accuracy of compound identification.

Methodology

Mass spectrum matching-based compound identification is achieved by matching the experimental mass spectra to the mass spectra stored in a reference library based on mass spectral similarity. In other words, all pairwise similarity scores between an experimental mass spectrum and each of the library mass spectra are first calculated. The compound whose library mass spectrum has the highest mass spectral similarity score is considered as the most probable compound that generated the experimental mass spectrum. Each mass spectrum is composed of

m/*z* values and their intensities. The intensities are used for calculation of the spectral similarity scores.

In this study, the spectral similarity between experimental mass spectrum and each of the reference spectra is calculated. A reference compound is considered as the compound given rise to the experimental spectrum if its reference spectrum has the best similarity with the experimental spectrum. The following methods are applied to calculate the similarity scores between the experimental mass spectrum and each of the reference spectra:

Dot Product

The dot product, which is also known as the cosine correlation (Stein & Scott, 1994), was used to obtain the cosine of the angle between two sequences of intensities, $\mathbf{x} = (x_i)_{i = 1,\ldots, n}$ and $\mathbf{y} = (y_i)_{i = 1,\ldots, n}$. It is defined as

$$S = \mathrm{S}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^{\mathrm{T}}\mathbf{y}}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}, \quad (1)$$

where $\mathbf{x}^{\mathrm{T}}\mathbf{y} = \sum_{i=1}^{n} x_i y_i$ and $\|\mathbf{x}\| = \left(\sum_{i=1}^{n} x_i^2\right)^{1/2}$. We calculate the dot product of mass spectra for each experimental compound and each reference compound, and a greater value of S in (1) indicates a higher chance that the reference compound is the compound that generated the experimental mass spectrum.

Ridge Regression

Ridge regression is a shrinkage method which imposes a penalty on the size of regression coefficients. The ridge coefficients minimize a penalized residual sum of squares,

$$\beta^{\mathrm{ridge}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \quad (2)$$

where p is the number of variables (e.g., compounds or metabolites), N is the number of observations (e.g., intensities or *m*/*z* values), and $\lambda \geq 0$, which is a complexity parameter and controls the amount of shrinkage. A larger value of λ results in a great amount of shrinkage. The coefficients are shrunk toward zero (and

each other) (Hastie, Tibshirani, & Friedman, 2009). A well-known equivalent method is to solve the following problem, which makes the size constraint on the parameters explicit:

$$\beta^{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 , \tag{3}$$

subject to $\sum_{j=1}^{p} \beta_j^2 \le t$. Note that there is a one-to-one correspondence between the parameters λ and t.

For ridge regression, we can also write the above criterion in matrix form, the ridge regression can be easily solved as

$$\boldsymbol{\beta}^{\text{ridge}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda \mathbf{I} \right)^{-1} \mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{4}$$

where $\mathbf{I}$ is the $p \times p$ identity matrix. In our case, $p \gg N$, so use the singular-value decomposition of $\mathbf{X}$, $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}} = \mathbf{R}\mathbf{V}^{\mathrm{T}}$ to calculate the coefficients, where $\mathbf{V}$ is $p \times N$ with orthonormal columns, $\mathbf{U}$ is $N \times N$ orthogonal, and $\mathbf{D}$ is a diagonal matrix with elements $d_1 \ge d_2 \ge \cdots \ge d_N \ge 0$. The matrix $\mathbf{R}$ is $N \times N$ with rows r_i^{T}. Replacing $\mathbf{X}$ by $\mathbf{R}\mathbf{V}^{\mathrm{T}}$, we have

$$\boldsymbol{\beta}^{\text{ridge}} = \mathbf{V}\left(\mathbf{R}^{\mathrm{T}}\mathbf{R} + \lambda \mathbf{I} \right)^{-1} \mathbf{R}^{\mathrm{T}}\mathbf{y} \ . \tag{5}$$

The Lasso

The lasso (Least Absolute Shrinkage and Selection Operator), which was first proposed by Tibshirani (1996), is a shrinkage method like ridge, but it has subtle and important differences from the ridge regression. The lasso is a penalized least squares procedure that minimizes residual sum of squares (RSS) subject to the non-differentiable constraint expressed in terms of the L1 norm of the coefficients (Kyung, Gill, Ghosh, & Casella, 2010). That is, the lasso estimator is given by

$$\beta^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right| \right\} . \tag{6}$$

This L1 norm constraint makes the solutions nonlinear in the y_i, resulting in no analytical solution different from ridge regression.

Two-Step Approach

To maximize the performance of compound identification and also reduce the data dimensionality, the two-step approaches are proposed by combining the dot product, Pearson's correlation, and penalized linear regression. In this procedure, the first step is made to precede the first match. Then, select a certain amount of the best matches based on the result of the first step and use them to conduct the second step which is penalized linear regression.

Dot product and lasso/ridge regression

In this two-step approach, after calculating the dot product of mass spectra for all experimental mass spectra and reference mass spectra, rank the results of dot product and choose N reference compounds with top N largest dot product values. Then conduct the lasso or ridge regression with only these N reference compounds. The flowchart is shown in Figure 1.

Pearson's correlation and lasso/ridge regression

In this case, after calculating the Pearson's correlation coefficients of an experimental spectrum and all reference spectra, sort the correlation coefficients in descending order and calculate their $(1 - \alpha)\%$ confidence intervals. Then, check if there is overlap between two adjacent intervals from the top compounds and stop at the N^{th} compound, if there is no overlap between the N^{th} interval and $(N + 1)^{\text{th}}$ interval. By doing so, select N reference compounds and then conduct the lasso/ridge regression only with these N reference compounds. Figure 2 shows the flow chart.

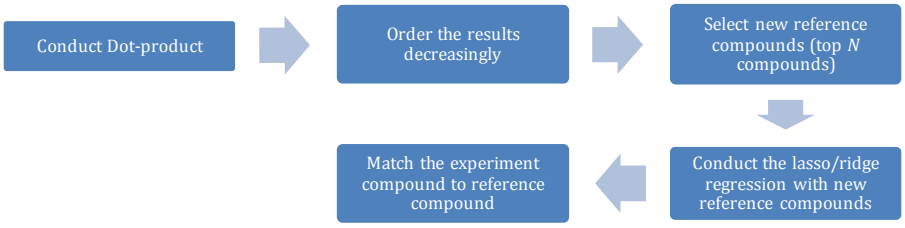

Figure 1. Workflows of the proposed two-step approach using dot product

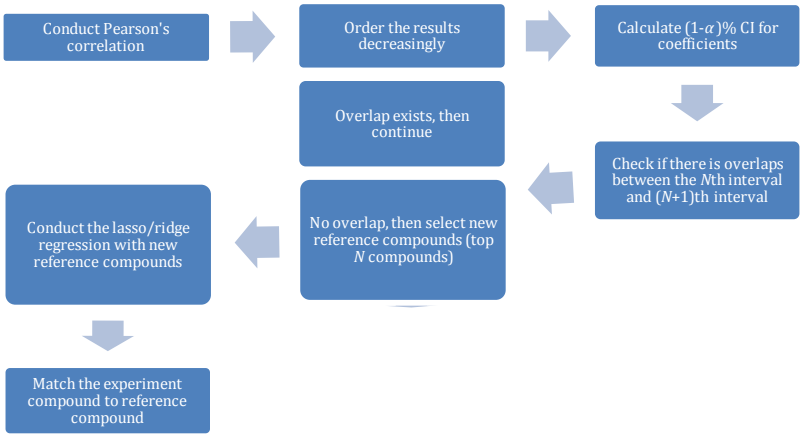

Figure 2. Workflows of the proposed two-step approach using Pearson's correlation along with the lasso/ridge regression

Data

The National Institute of Standards and Technology (NIST) Chemistry WebBook service provides users with chemical and physical information for chemical compounds, including mass spectra generated by electron ionization mass spectrometry (Linstrom & Mallard, 2001). The mass spectra recorded in the NIST main mass spectrometry database and repetitive database were used as the reference mass spectra and experimental mass spectra, respectively. For our reference library, the mass spectra of 2739 compounds were extracted from NIST Chemistry WebBook database. The fragment ion *m*/*z* values ranged from 1 to 1036 with a bin size of 1. The experimental library contains 1530 mass spectra of compounds

extracted from the repetitive database. Because it was assumed the NIST library has the mass spectrum information for all the experimental compounds, all the compounds that were not present in the NIST main library were removed from the repetitive library.

Performance Evaluation

Each compound in the NIST database was assigned to a unique Chemical Abstract Service (CAS) registry number. To evaluate the performance of compound identification of each similarity measure, calculate the identification accuracy. The accuracy is the proportion of the spectra identified correctly in query data. In other words, if a pair of unknown and reference spectra have the same CAS index, we consider this pair as the correct match and if otherwise as the incorrect match. Then by counting all the correct matches, the accuracy of identification can be calculated by

$$\text{accuracy} = \frac{\text{number of spectra matched correctly}}{\text{number of spectra queried}} \tag{7}$$

Software

All the statistical analyses are performed using statistical software R version 2.15.3. The comparison of ridge regression and the lasso is performed by the R package *glmnet*.

Results

The penalized regressions, lasso and ridge regression, were conducted using R package *glmnet* to compare the identification results. In order to find a proper range of the shrinkage factor λ, the shrinkage factor was initially varied widely from 0.0001 to 1000000 and accuracy was calculated for each method. Figure 3(a) shows accuracy along with different shrinkage factor values for these two penalized linear regressions. The accuracy trend for the lasso is very different from that of ridge regression. For larger values of λ, accuracy tends to be a constant for each regression. However, accuracy for the lasso tends to be zero, while the ridge regression levels off at 89.20%. Based on this analysis, the shrinkage factors ranged from 0.10 to 5000 were focused on and then applied the lasso and ridge regression, respectively, to further check the specific trends of each regression.

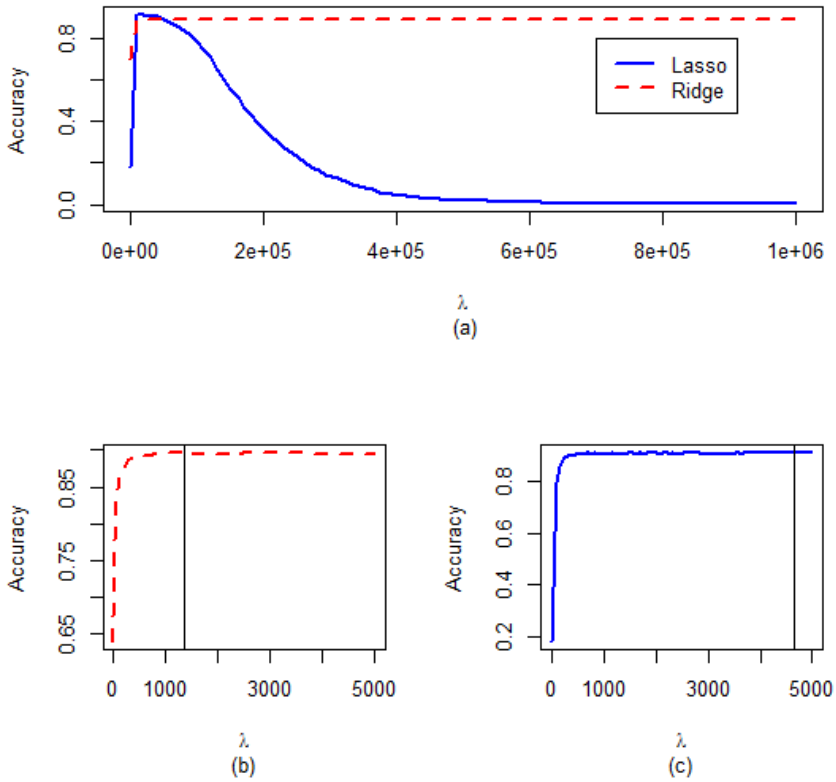

Figure 3. Accuracy vs. shrinkage factor λ. Plot (a) is for the lasso and ridge regression using the wide range of λ. Plots (b) and (c) are for the ridge regression and lasso, respectively, using the smaller range of λ.

The Lasso

After conducting the lasso regression between query data and reference data with 100 different shrinkage factors λ (range from 0.10 to 5000), correct matches and accuracy were calculated. Figure 3(c) displays the change of accuracy corresponding to different shrinkage factor values. After a further check, the best accuracy for the lasso is 91.50% when $\lambda = 4646.47$. This accuracy is higher than the highest accuracy from ridge regression.

Two-Step Approach

Dot product and the lasso/ridge regression

The two-step approach, dot product and the lasso/ridge regression were performed to optimize the performance of compound identification, and to find the relationship between accuracy and different rank levels as well as λ values. A total of 12 different rank levels ranging from 25 to 300 were chosen. For λ, 100 values ranging from 0.10 to 5000 were used, which is the same with the identification using the lasso and ridge regression. Table 1 lists the analysis results. The results for this two-step approach are not so clear to interpret, so a contour plot (Figure 4) is used to show the relationship among accuracy, rank levels, and shrinkage factors for both the lasso and ridge regression.

In Figure 4, the green color indicates relatively low accuracy, while white and pink indicate relatively high accuracy. The highest accuracy, 90.20%, appears at rank level = 25 and λ = 0.10, which is shown as a red point in the left plot of Figure 4. The other four red points in the left plot of Figure 4 also have relatively high accuracy. Comparing with ridge regression only, we can see that this two-step approach performs better than the ridge regression only (accuracy = 90.20% vs. 89.74%). In general, we can also see the following trend: when the shrinkage factor (λ) increases, the corresponding rank needs to be increased in order to achieve better identification accuracy.

Table 1. Top 5 best accuracies and corresponding shrinkage factors for the dot product and the lasso/ridge regression

Method	Rank	Shrinkage factor (λ)	Number of query	Number of correct matches	Accuracy
Dot Product and Ridge	25	0.10	1530	1380	90.20%
	100	202.12	1530	1380	90.20%
	100	303.12	1530	1380	90.20%
	250	505.14	1530	1380	90.20%
	275	555.64	1530	1380	90.20%
Dot Product and Lasso	200	3838.41	1530	1395	91.18%
	300	1363.71	1530	1395	91.18%
	300	1414.21	1530	1395	91.18%
	300	1464.72	1530	1395	91.18%
	300	1515.22	1530	1395	91.18%

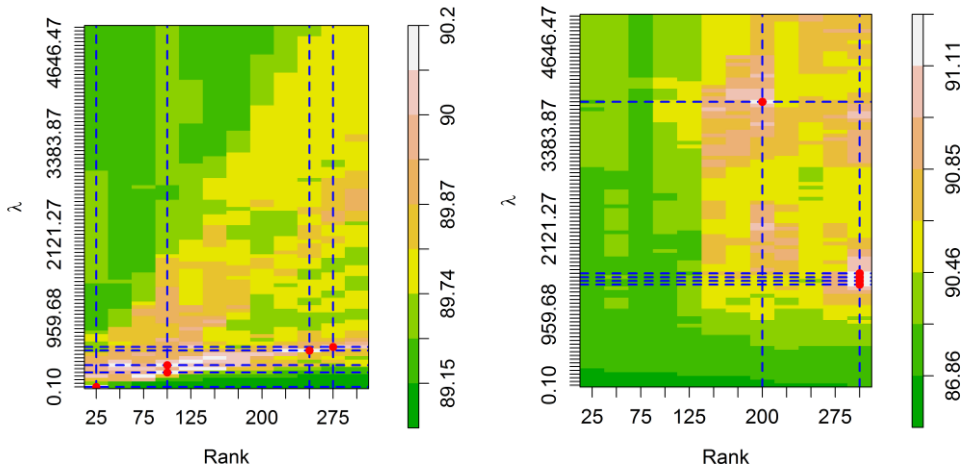

Figure 4. Accuracy of two-step approach using dot product and ridge (left) and the lasso regression (right)

The right plot of Figure 4 displays the relationship among accuracy, rank levels, and λ values for the two-step approache using the dot product and the lasso regression. The highest accuracy 91.18% appears at rank level = 200 and $\lambda = 3838.41$, which are shown as a red point in the plot. Comparing to the identification using the lasso only, this two-step approach has no improvement in accuracy, which is different from the two-step approach using ridge regression.

Pearson's correlation and the lasso/ridge regression

For the Pearson's correlation and penalized linear regression two-step approach, we intend to find the relationship among accuracy, different confidence levels, and λ values. The α levels of 0.01, 0.025, 0.05, and 0.1 were chosen, along with 100 shrinkage factor (λ) values ranging from 0.10 to 5000. The top 5 highest accuracies and corresponding shrinkage factors are shown in Table 2.

The best accuracies for this two-step approach using the lasso and ridge all appear at $\alpha = 0.1$, which are 89.41% (ridge regression) and 77.91% (the lasso). However, in this two-step approach, the lasso regression does not seem as good as the ridge regression. The contour plots are shown in Figure 5.

The relationship of accuracy, α levels, and λ values in this two-step approach seems much clearer. In the left plot of Figure 5, when the shrinkage factor (λ) is greater than a certain value (around 300), it does not influence the accuracy so much.

The red points, which indicate the best accuracies, all appear at α=0.1, making a red vertical line.

Table 2. Top 5 best accuracies and corresponding shrinkage factors for Pearson's correlation and the lasso/ridge regression

Method	α	Shrinkage factor (λ)	Number of query	Number of correct matches	Accuracy
Dot Product and Ridge	0.1	101.11	1530	1368	89.41%
	0.1	353.63	1530	1368	89.41%
	0.1	404.13	1530	1368	89.41%
	0.1	454.64	1530	1368	89.41%
	0.1	505.14	1530	1368	89.41%
Dot Product and Lasso	0.1	0.10	1530	1192	77.91%
	0.1	50.60	1530	1192	77.91%
	0.1	101.11	1530	1192	77.91%
	0.1	151.61	1530	1192	77.91%
	0.1	202.12	1530	1192	77.91%

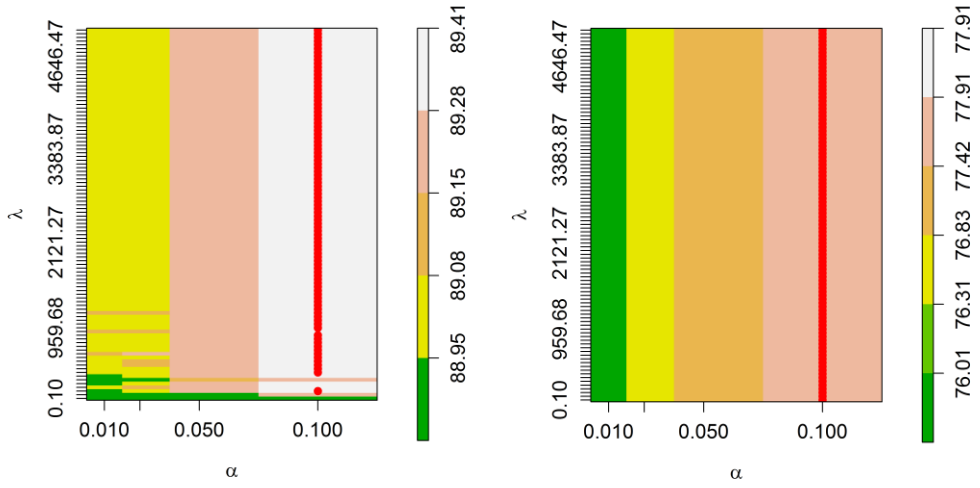

Figure 5. Accuracy of two-step approach using Pearson's correlation and ridge (left) and the lasso regression (right)

The relationship among accuracy, α levels, and λ values in Pearson's correlation and the lasso two-step approach is similar to that when ridge regression is used, as can be seen in the right plot of Figure 5. As in the two-step approach

using ridge regression, the red points all appear at α=0.1, which make a red vertical line. The selection of λ value does not influence the accuracy, although it is clear that a greater α level results in higher accuracy.

Table 3. Compound identification methods and their performance.

Method	Lambda	Rank (Alpha)	Accuracy (%)
Dot Product	--	--	89.54
Pearson's Correlation	--	--	89.54
Ridge	1363.71	--	89.74
Lasso	4646.47	--	91.50
Dot Product and Ridge	0.10	25.0	90.20
Pearson's Correlation and Ridge	353.63~858.67	0.1	89.41
Dot Product and Lasso	3838.41	200.0	91.18
	1363.71~1515.22	300.0	
Pearson's Correlation and Lasso	0.10~960.00	0.1	77.91

The Best Performance

The performance of four compound identification methods involving penalized linear regression were tested. In addition, previously widely used methods were included. Table 3 shows these new methods and their best performance (accuracy), including the corresponding shrinkage factor (λ) value, rank selection (for dot product and the lasso/ridge regression two-step approach), and alpha selection (for Pearson's correlation and the lasso/ridge regression two-step approach). The performance of the dot product and Pearson's correlation in compound identification are also listed. Overall, the lasso only performs the best among other approaches (accuracy = 91.50%, line 4 in Table 3).

Conclusion

New approaches for compound identification were proposed using penalized linear regressions, and further two-step approaches are introduced. In particular, an alternative to the semi-partial correlation-based approach using multiple linear regressions was pursued.

From the results using a small data set, it can be seen that the lasso achieves the highest accuracy of compound identification, which is 91.50% with λ of 4646.5,

resulting in 1% greater accuracy than that of the dot product. Nevertheless, the accuracy for the lasso is highly related to the selection of shrinkage factor λ, so we have to tune up the shrinkage factor, such as using cross-validation, when using the lasso for compound identification. This additional work will result in a longer calculation time. Although ridge regression shows a worse accuracy than the lasso, its property that accuracy becomes constant after a certain λ value makes the ridge regression a better choice in terms of computational expense. In addition, the two-step approach using the dot product and the lasso has accuracy 91.18 %, which is similar to that of the lasso only. Because the dot product reduces the size of library, the following lasso regression becomes much inexpensive that the lasso regression only in terms of computational time. In this regard, this method could be a best alternative to the lasso regression only to achieve a higher accuracy.

Furthermore, the same data used here were applied to the mixture semi-partial correlation approach with the mixture weight of 0.7 and the rank of 100 (Kim et al. 2012), resulting in a slightly better performance than that of the lasso only with 92.9% of identification accuracy. Although the two-step approach using Pearson's correlation and the lasso/ridge regression has no improvement in identification accuracy, it shows that the shrinkage factor selection has no effect upon the accuracy of compound identification, which means that there should be no concern about the selection of shrinkage factors.

References

Atwater, B. L., Stauffer, D. B., McLafferty, F. W., & Peterson, D. W. (1985). Reliability ranking and scaling improvements to the probability based matching system for unknown mass spectra. *Analytical Chemistry, 57*(4), 899-903. doi: 10.1021/ac00281a028

Beer, I., Barnea, E., Ziv, T., & Admon, A. (2004). Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics, 4*(4), 950-960. doi: 10.1002/pmic.200300652

Craig, R., Cortens, J. C., Fenyo, D., & Beavis, R. C. (2006). Using annotated peptide mass spectrum libraries for protein identification. *Journal of Proteome Research, 5*(8), 1843-1849. doi: 10.1021/pr0602085

Frewen, B. E., Merrihew, G. E., Wu, C. C., Noble, W. S., & MacCoss, M. J. (2006). Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Analytical Chemistry, 78*(16), 5678-5684. doi: 10.1021/ac060279n

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer series in statistics. New York, NY: Springer-Verlag.

Hertz, H. S., Hites, R. A., & Biemann, K. (1971). Identification of mass spectra by computer-searching a file of known spectra. *Analytical Chemistry, 43*(6), 681-691. doi: 10.1021/ac60301a009

Julian, R. K., Higgs, R. E., Gygi, J. D., & Hilton, M. D. (1998). A method for quantitatively differentiating crude natural extracts using high-performance liquid chromatography-electrospray mass spectrometry. *Analytical Chemistry, 70*(15), 3249-3254. doi: 10.1021/ac971055v

Kim, S., Koo, I., Jeong, J., Wu, S., Shi, X., & Zhang, X. (2012). Compound identification using partial and semipartial correlations for gas chromatography-mass spectrometry data. Analytical Chemistry, 84(15), 6477-6487. doi: 10.1021/ac301350n

Koo, I., Kim, S., & Zhang, X. (2013). Comparative analysis of mass spectral matching-based compound identification in gas chromatography-mass spectrometry. *Journal of Chromatography A, 1298*, 132-138. doi: 10.1016/j.chroma.2013.05.021

Koo, I., Zhang, X., & Kim, S. (2011). Wavelet- and Fourier-transform-based spectrum similarity approaches to compound identification in gas

chromatography/mass spectrometry. *Analytical Chemistry, 83*(14), 5631-5638. doi: 10.1021/ac200740w

Kyung, M., Gill, J., Ghosh, M., & Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis, 5*(2), 369-411. Available from http://projecteuclid.org/euclid.ba/1340218343

Linstrom, P. J., & Mallard, W. G. (Eds.). (2015). *NIST chemistry webbook: NIST standard reference database number 69*. Retrieved from http://webbook.nist.gov/chemistry/

Rasmussen, G. T., & Isenhour, T. L. (1979). The evaluation of mass spectral search algorithms. *Journal of Chemical Information and Modeling, 19*(3), 179-186. doi: 10.1021/ci60019a014

Stein, S. E., & Scott, D. R. (1994). Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society Mass Spectrometry, 5*(9), 859-866. doi: 10.1016/1044-0305(94)87009-8

Tabb, D. L., MacCoss, M. J., Wu, C. C., Anderson, S. D., & Yates, J. R., III. (2003). Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Analytical Chemistry, 75*(10), 2470-2477. doi: 10.1021/ac026424o

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological), 58*(1), 267-288. Available from http://www.jstor.org/stable/2346178