# IMSmining: A Tool for Imaging Mass Spectrometry Data Biomarker Selection and Classification

**Jingsai Liang, Don Hong, Fengqing (Zoe) Zhang and Jiancheng Zou**

**Abstract** We developed IMSmining, a free software tool combining functions of intuitive visualization of imaging mass spectrometry (IMS) data with advanced analysis algorithms in a single package which is easy to operate. Main functions of IMSmining include data visualization, biomarker selection, and classification using advanced multivariate analysis methods such as elastic net, sparse PCA, as well as wavelets. It can be used to study the correlation and distribution of the IMS data by incorporating the spatial information in the entire image cube and to help finding the distinction of the possible features caused by the biological structure and the potential biomarkers. This software package can be downloaded from http://capone.mtsu.edu/dhong/IMSmining.htm.

**Keywords** IMS data processing · Statistical computing · Wavelet application · Biomarker selection and Classification · Software package

J. Liang · D. Hong (✉) · J. Zou
Computational Science Program, Middle Tennessee State University, Murfreesboro, TN, USA
e-mail: Don.Hong@mtsu.edu

J. Liang
e-mail: JL4Z@mtmail.mtsu.edu

J. Zou
e-mail: zjc@ncut.edu.cn

D. Hong · J. Zou
College of Sciences, North China University of Technology, Beijing, China

F.Z. Zhang
Department of Psychology, Drexel University, Philadelphia, PA, USA
e-mail: fengqingzoezhang@gmail.com



R.N. Mohapatra et al. (eds.), *Mathematics and Computing*,
Springer Proceedings in Mathematics & Statistics 139,
DOI 10.1007/978-81-322-2452-5_11

## 1 Introduction

Mass spectrometry (MS) and imaging mass spectrometry (IMS) are both important techniques in proteomics. IMS is a novel technology that is able to incorporate spatial biochemical information in full molecular range [1]. However, there are still many challenges in data processing due to high dimensionality, huge differences between the number of predictors and the sample size, and the incorporation of both spectral and spatial information. All these challenges pose great difficulties in model selection and data processing.

Several software tools are commonly used for IMS/MS data analysis. Biomap and Tissue View are mainly for data visualization. These software tools lack advanced data analysis functionality such as multivariate analysis methods for biomarker selection and classification. MarkerView and ClinProTools are packages for MS data analysis. Technically, IMS data after using Biomap or Tissue View based on visualization can be exported and then imported to MarkerView or ClinProTools for further data analysis. However, this is not feasible for IMS data processing, especially for those in high resolution. PCA and clustering are most commonly used for IMS data analysis [2]. LDA and multivariate analysis of variance [3] and PCA combined with support vector machine (SVM) [4] were used to process IMS data. However, these methods have their limitations of handling high-dimensional IMS cubes and incorporating spatial information.

It is essential to extract the complex/hidden patterns from the IMS data. Modern statistical methods should be used to complete a series of operations for biomarker selection and classification in potential application to disease and cancer diagnosis.

IMSmining software package is mainly for IMS data visualization, biomarker selection, model validation, and classification. Visualization functions include the spectrum of a single pixel, the average spectrum of an area, and intensity distribution matrix at a fixed $m/z$ value. The analysis functions include not only PCA, SVM, and LDA methods, but also the most recently developed models SPCA [5, 6], Wavelet4IMS [7], EN4IMS (Elastic Net) [8], and WEN (Weighted Elastic Net) [9] using the spatial information. The motivation is to provide a convenient and automatic way to analyze and extract useful information from the high-dimensional and complex IMS data by not only utilizing the spectrum information within individual pixels, but also studying the correlation and distribution using the spatial information.

The remainder of the paper is organized as follows: In Sect. 2, the main algorithms such as EN4IMS, WEN, Wavelet4IMS are briefly introduced. In Sect. 3, we give the detail of the implementation of the software. A summary of the pipeline of this software is given in Sect. 4. Finally, remarks and a brief discussion are presented in Sect. 5.

## 2 Algorithm Content

### 2.1 EN4IMS

Let us consider the multiple linear regression model with $n$ observations. Suppose that $x_j = (x_{1j}, \ldots, x_{nj})^T$, $j = 1, \ldots, p$ are linear independent predictors and $y = (y_1, \ldots, y_n)^T$ is the response vector. If we use $X = [x_1, \ldots, x_p]$ represent the predictor matrix, the linear regression model can be expressed as

$$y = X\beta + \varepsilon \tag{1}$$

where $\beta = (\beta_1, \ldots, \beta_p)^T$ and the noise term $\varepsilon \sim N(0, \sigma^2 I_n)$. The naive EN criterion is to minimize the following function [10]:

$$L(\lambda_1, \lambda_2, \beta) = \|y - X\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2. \tag{2}$$

There are totally two penalty parts in Eq. 2. The $\ell_1$ term enforces the model to generate sparse solution and the quadratic term can achieve the group effect. Zou et al. [10] mentioned that the naive EN has some weakness that will result in double amount of shrinkage. Therefore, the EN algorithm modified the naive elastic net as

$$\hat{\beta}_1 = (1 + \lambda_2)\hat{\beta}_0. \tag{3}$$

where $\beta_1$ is named elastic net and $\beta_0$ is the naive elastic net. Also, the EN estimates $\hat{\beta}$ is given in [10] by

$$\hat{\beta} = \arg\min_{\beta} \beta^T((X^T X + \lambda_2 I)/(1 + \lambda_2))\beta - 2y^T X\beta + \lambda_1\|\beta\|_1. \tag{4}$$

In the IMSmining software, we apply EN4IMS based on the above EN algorithm to estimate the biomarkers. EN4IMS algorithm incorporates a spatial penalty term into the EN model. IMS information provides huge spatial information located in each individual pixel. One important fact is that pixels in different locations of the same disease should have similar ion intensity values, which means the standard deviation of the intensities at the true biomarkers should be small. Conversely, the standard deviation would be very large among the complex tissue structure like bones.

So in EN4IMS, we use a parameter $\tau$ to balance two items together. One is the RSS of the linear model and another is the average of spatial standard deviations of the selected ion intensities. In detail, we use tenfold CV to minimizing the following formula:

$$(1 - \tau)\|y - \hat{y}\|_2^2 + \frac{\tau}{M}\sum_{j=1}^{M}\sqrt{\frac{\sum_{i=1}^{N}(x_{ij} - \mu_j)^2}{N - 1}}, \quad 0 < \tau < 1. \tag{5}$$

where $N$ is the number of all cancer pixel, $x_{ij}$ is the intensity of a fixed $j$th $m/z$ value at pixel $i$, $\mu_j$ is the average intensity of all cancer pixels at this fixed $j$th $m/z$ value, and M is the cardinality of active set as defined in [8].

### 2.2 WEN

In order to consider more precise biomarker selection, Hong and Zhang [9] proposed the following model named WEN:

$$\arg\min_{\beta} \frac{1}{2}\|y - \sum_{j=1}^{p} x_j\beta_j\|_2^2 + n\lambda_1 \sum_{j=1}^{p} \omega_j|\beta_j| + \frac{n}{2}\lambda_2 \sum_{j=1}^{p} |\omega_j\beta_j|^2. \tag{6}$$

where $\omega_j > 0$, $j = 1, \ldots, p$ are weighted penalty coefficients. In [9], the LARS-WEN algorithm is provided to solve the above WEN model. Experiments show that WEN not only reduces the number of side features but also helps new biomarkers discovery.

### 2.3 Wavelet4IMS

To meet challenges in IMS data processing, an effective and efficient algorithm for IMS data biomarker selection and classification using methods of multiresolution analysis are proposed. In [7], the authors proposed Wavelet4IMS algorithm. In addition to apply wavelet transform for IMS data denoising, measurement for the similarity of wavelet coefficients is introduced, and the idea of wavelet pyramid method for image matching is applied for biomarker selection and the Naive Bayes classifier is used for classification in the wavelet coefficient space. Performance of the algorithm is evaluated with real data and the results of our experiments show that the multiresolution method has higher accuracy in classification.

## 3 Software Description

IMSmining allows users to visualize IMS data, to discover biomarkers, and to perform a pixel level classification for different IMS data sections. This software package is designed to give users a maximum level of convenience together with high flexibility.

### 3.1 Interface

Figure 1 shows the interface of the software based on MATLAB GUI. The first menu is for the data-type options. We can import the data from .mat file or .txt folder or export the biomarker. The next menu contains seven algorithmic options: EN4IMS, WEN, PCA+SVM/LDA, SPCA+SVM/LDA, and Wavelet4IMS. We can also use "view menu" to view the spectrum of a single pixel or the average spectrum of selected area. Toolbar icons can be used to zoom in, zoom out, drag, or rotate the data cube. There are also four figure windows including training, spectrum, testing, and result. We can use the mouse to drag the squares to select the cancer and noncancer area for training and testing.

### 3.2 Data Visualization

IMSmining provides different methods of visualization for IMS data. Users can see intensity distribution images of different $m/z$ values by clicking on different $m/z$

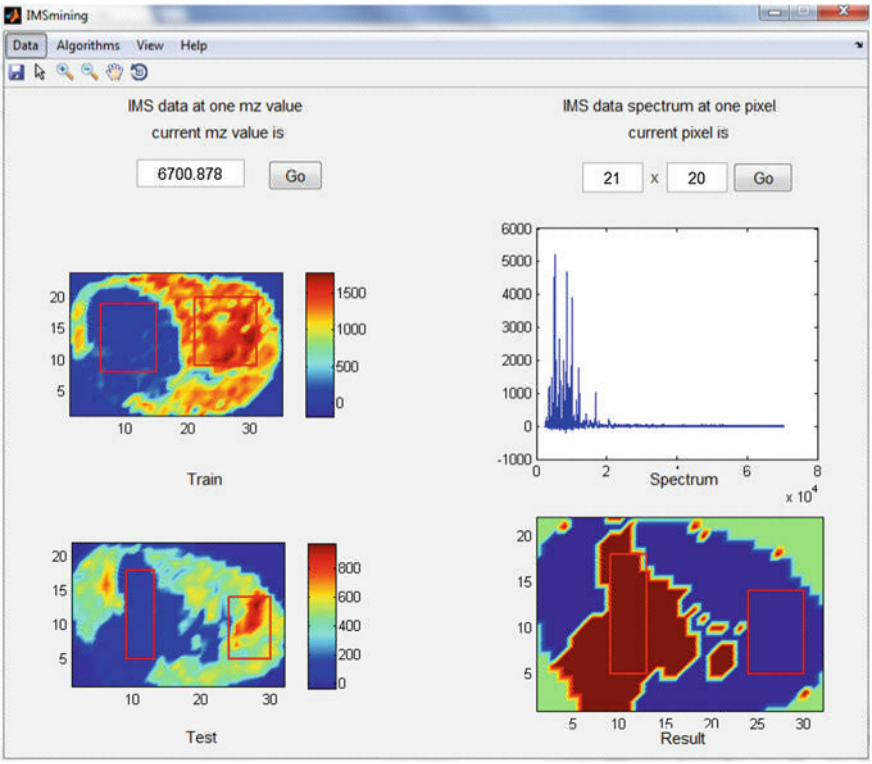


**Fig. 1** Interface of GUI

values on the spectrum image. Users can also see spectrum of different pixels just by clicking on different pixel positions on the distribution images. Users can enlarge the spectrum to see whether the $m/z$ value is corresponding to a true peak. The interactive responses between the intensity images (Left Upper Window) and the spectra (Right Upper Window) are extremely convenient and provide a better understanding of the spatial distribution information for a selected $m/z$ peak. Furthermore, users can directly select an area of pixels from the left upper window to see the mean spectrum of these selected pixels.

### *3.3 Biomarker Selection*

IMSmining provides a series of algorithms, which include very recently developed EN4IMS, WEN models, and Wavelet4IMS for IMS data analysis, and other methods such as PCA, SPCA, and SVM popularly used in IMS community. Here, $m/z$ values selected by the model are considered as potential biomarkers.

In EN4IMS algorithm, a spatial penalty term is incorporated into the cross validation step of the EN model [10] for IMS data processing [8]. The WEN model associates the weighted coefficients of EN model with ion intensity spreading information, and thus provides a systematic consideration for the spatial information of the IMS data for biomarker selection and classification. Both models inherit good properties from the EN method which produces a sparse model with admirable prediction accuracy. By taking the spatial information into consideration, these two models help distinguish the IMS feature peaks caused by biological structure differences from those truly associated with diseases. In Wavelet4IMS algorithm, IMSmining transforms each mass spectrometry to wavelet space and select biomarkers based on multiresolution analysis.

### *3.4 Classification*

IMSmining provides model validation and classifies testing samples. Users can select the training data region directly from the training data figure. After analyzing the training data sets to create the predictive model, validation of models can be done on the selected cancer and noncancer square area of the testing data sets. To enhance the chance of finding the best model, the tuning parameter $\lambda$ of EN4IMS and WEN algorithms can be changed accordingly by users. As a result, we can obtain the classification rates of the selected testing area. Besides implementing EN4IMS or WEN algorithm, IMSmining has one method named Wavelet4IMS which uses feature vectors selected from wavelet domain combining with a naive Bayes classifier for classification. IMSmining can also use PCA or SPCA to reduce the dimension of the data and then continue to use SVM or LDA for classification.

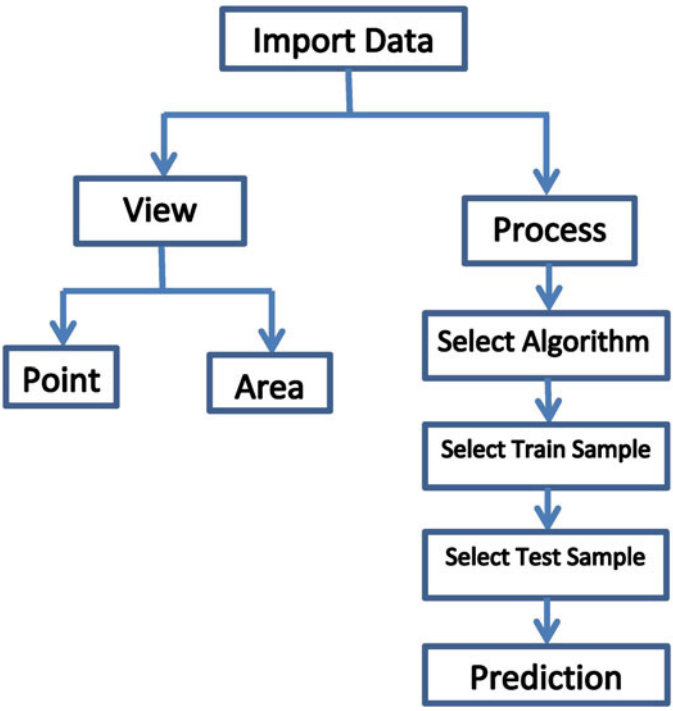


**Fig. 2** Pipeline of GUI

## 4 Pipeline

Figure 2 shows the pipeline of IMSmining. After importing the data, we can either view the image of the data or process the data-based variety of algorithms. If you only want to view the image, you have two choices: point or area. Then you can import a single pixel or just simply click on the data image. Or you can drag the mouse to select an area to calculate the major statistical value of this specifical area. In another branch, you have three steps to complete the model prediction: algorithm selecting, training image selecting, and testing image selecting. You can stop the algorithm at each step and start over in another algorithm. And after you select the images, you need to use the mouse to drag both of the cancer and noncancer area. After the calculation, IMSmining will show the comparative cancer and noncancer result.

## 5 Discussion

We developed a software package called IMSmining based on algorithms of EN4IMS, WEN, SPCA, and Wavelet4IMS. We have applied this software tool to real IMS data [8, 9]. Compared with other current popular methods, the models of EN4IMS, WEN, and Wavelet4IMS work more efficiently and effectively for IMS data processing in terms of confirming new biomarkers, producing a more accurate feature list including significant peaks, and providing more accurate classification results.

**Acknowledgments** The authors would like to thank Shannon Cornett, Sara Frappier, and Richard M. Caprioli from the VUMSRC for valuable discussions and providing IMS data sets for the study. DH is grateful for the support from the program of Beijing Overseas High Caliber Talents.

## References

1. Trede, D., Kobarg, J. H., Oetjen, J., Thiele, H., Maass, P., Alexandrov, T.: On the importance of mathematical methods for analysis of maldi imaging mass spectrometry data. J Integr Bioinform **9**(1), 189 (2012)
2. de Plas, R.V., Ojeda, F., Dewil, M., Bosch, L.V.D., Moor, B.D., Waelkens, E.: Prospective exploration of biochemical tissue composition via imaging mass spectrometry guided by principal component analysis. In: Pacific Symposium on Biocomputing, World Scientific, pp. 458–469 (2007)
3. Muir, E.R., Ndiour, I., Le Goasduff, N.A., Moffitt, R., Liu, Y., Sullards, M.C., Merrill, A., Chen, Y., Wang, M.: Multivariate analysis of imaging mass spectrometry data. Bioinform. Bioeng. 472–479 (2007)
4. Gerhard, M., Deininger, S.-O., Schleif, F.: Statistical classification and visualization of maldi-imaging data. Comput. Based Med Syst 403–405 (2007)
5. Zou, H.T., Tibshirani, H.R.: Sparse principal component analysis. J. Comput. Graph. Stat. **15**(2), 265–286 (2006)
6. Wang, Y., Wu, Q.: Sparse PCA by iterative elimination algorithm. Adv. Comput. Math. **36**(1), 137–151 (2012)
7. Xiong, L., Hong, D.: Multi-resolution analysis method for ims data biomarker selection and classification. British J. Math. Comp. Sci. **5**(1), 64–80 (2015)
8. Zhang, F., Hong, D.: Elastic net based framework for imaging mass spectrometry data biomarker selection and classification. Stat. Med. **30**, 753–768 (2010)
9. Hong, D., Zhang, F.: Weighted elastic net model for mass spectrometry imaging processing. Math. Model. Nat. Phenom. **5**(3), 115–133 (2010)
10. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. B **67**, 301–320 (2005)