Research Article

Statistics in Medicine

Received 6 January 2010, Accepted 22 October 2010 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4147

Elastic net-based framework for imaging mass spectrometry data biomarker selection and classification

Fengqing (Zoe) Zhang[a‡] and Don Hong[a,b*†]

Imaging mass spectrometry (IMS) shows great potential for the rapid mapping of protein localization and for detecting of sizeable differences in protein expression. However, data processing remains challenging due to the difficulty of analyzing high dimensionality, the fact that the number of predictors is significantly larger than the number of observations, and the need to consider both spectral and spatial information in order to represent the advantage of IMS technology. Ideally one would like to efficiently analyze all acquired data to find trace features based on both spectral and spatial patterns. Therefore, biomarker selection from IMS data is a problem of global optimization. A recently developed regularization and variable selection method, elastic net (EN), produces a sparse model with admirable prediction accuracy and can be an effective tool for IMS data processing. In this paper, we incorporate a spatial penalty term into the EN model and develop a new tool for IMS data biomarker selection and classification. A comprehensive IMS data processing software package, called EN4IMS, is also presented. The results of applying our method to both simulated and real data show that the EN4IMS algorithm works efficiently and effectively for IMS data processing: producing a more precise listing of selected peaks, helping confirmation of new potential biomarkers discovery, and providing more accurate classification results. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: imaging mass spectrometry; biomarker selection; classification; elastic net

1. Introduction

New advances in mass spectrometry (MS), such as matrix-assisted laser desorption ionization (MALDI) time-of-flight (TOF) MS, surface-enhanced laser desorption/ionization (SELDI) TOF MS, and imaging mass spectrometry (IMS), have greatly enhanced the opportunity for proteomics to be used for investigative studies of molecular interactions in intact tissue with excellent molecular specificity. Particularly, MALDI IMS has emerged as a powerful technique for analyzing the spatial distribution of proteins directly in tissue specimens [1]. Current interest in MALDI IMS lies in the unique advantage of correlating histological information as determined by a pathologist with the spatially resolved biochemical information provided by the IMS experiments. Advances are being made in all areas of MALDI IMS, including in how sample preparation techniques, instrument design and data analysis, a comprehensive structural and molecular analysis can be performed in a high-throughput and reproducible manner. IMS has broad applications for studying the spatial distribution of lipids [2], peptides [3], proteins [4], and small molecules with their metabolites [5] in tissue sections. In such instances, IMS helps to find the biomedical changes related with several diseases, especially multiple forms of cancer [6–8].

However, each MALDI IMS data set is multidimensional and has hundreds of pixels associating a complete mass spectrum with each of them. This contrasts with regular images where for each pixel there is a set of RGB values. Each mass spectrum contains mass-to-charge (m/z) values ranging from

[a] *Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, TN, U.S.A.*
[b] *College of Science, Ningbo University, Zhejiang, People's Republic of China*
[*] *Correspondence to: Don Hong, Department of Mathematical Science, Middle Tennessee State University, Murfreesboro, TN 37132, U.S.A.*
[†] *E-mail: dhong@mtsu.edu*
[‡] *Current address: Department of Statistics, Northwestern University, Evanston, IL, U.S.A.*

2k to 70k Da and ion intensity values that are associated with each pixel. There are hundreds of mass spectra represented in a single MS image. In this case, the number of predictors (m/z values) is much larger than the number of observations (the number of spectra). To fully utilize IMS data, it is desirable to not only identify the peaks of the spectrum within individual pixels, but also to study correlation and distribution using the spatial information of the entire image cube. Another important distinction that should be made is to determine whether the m/z values selected as potential biomarkers are caused by the biological structure of the tissues or by the disease state being investigated. All of these issues compounded together pose great challenges in IMS data processing and statistical analysis.

MALDI IMS software packages such as BioMap as well as many other software tools for IMS do not provide multivariate analysis (MVA) methods. Usually, a common way to visualize IMS data is to generate two-dimensional ion intensity maps for known m/z values of interest [9]. However, a more important application should be the determination of unknown variants for metabolite and protein profiling in both clinical and disease studies. Such mature analysis methods have not yet been implemented in the current commercial software. The IMS community has begun exploring and comparing a few MVA methods, such as principal component analysis (PCA), linear discriminant analysis (LDA), and clustering methods with IMS data analysis [10]. The use of PCA in analyzing IMS data has been proposed to identify both spatial and mass trends that can merit further investigation [11]. Also, LDA, multivariate analysis of variance (MANOVA), and clustering methods have been used to analyze the IMS data [12]. PCA and clustering are most commonly used for IMS data [13, 14]. Plas *et al.* [15] proposed to use peak intensity-based PCA to process IMS data. Correlation calculation for ion images, both in and between serial sections, was studied in [16]. PCA and support vector machine (SVM) were also combined to process IMS data in [17]. However, these methods for IMS data processing are inadequate due to their limited use of spatial information and the advantages of IMS technology (see Section 2.3.3 for a more detailed discussion).

A newly developed variable selection method, called elastic net (EN), can simultaneously perform automatic variable selection and continuous shrinkage [18]. That is, it can continuously shrink the coefficients toward zero as its regularization parameters increase; some coefficients are shrunk to exactly zero if the regularization parameters are sufficiently large [18–20]. The shrinkage often improves the prediction accuracy due to the bias-variance trade-off. Thus, the EN model simultaneously achieves accuracy and sparsity. The achievement of sparsity is particularly useful when the number of variables (p) is much larger than the number of observations (n). In addition, the EN model encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together. Compared with other current commonly used analysis methods, the EN model is much more suitable for IMS data processing. In this paper, we incorporate a spatial penalty term into the EN model in order to develop a new tool for IMS data biomarker selection and classification. Our motivation is to fully utilize not only the spectral information within individual pixels, but also the spatial information for the entire IMS data cube. A software package for comprehensive IMS data processing, called EN4IMS, is developed based on this new model. By incorporating the spatial penalty term, this package helps to distinguish the IMS feature peaks caused by biological structural differences from those truly associated with cancer.

The proposed EN4IMS method has been tested on extensive simulation studies, and the algorithm has also been applied to two real IMS data sets: one with relatively low resolution and other with higher resolution. Both data sets were provided by Vanderbilt University Mass Spectrometry Research Center (VUMSRC). The analysis results of both simulation studies and real data examples show that the EN4IMS algorithm works efficiently and effectively for IMS data processing: producing a more precise listing of feature peaks, help in discovering new potential biomarkers, and providing more accurate classification results. A set of our selected features with interesting biological explanations has been identified following a series of experiments.

2. Methods

2.1. Data interpretation

MALDI IMS offers the potential for direct examination of biomolecular patterns from cells and tissues. This makes it a seemingly ideal tool for biomedical diagnostics and molecular histology [21]. To obtain the MALDI IMS data, thin frozen sections ($10-15\,\mu$m thick) are cut, thaw-mounted on target plates and subsequently an energy absorbing matrix is applied. Areas, typically having a target spot size of about 50 μm in diameter, are ablated with an UV laser, and give rise to ionic molecular species that are recorded according to their mass-to-charge (m/z) values. Thus, a single mass spectrum is acquired from each ablated spot (pixel) in the array. Signal intensities at specific m/z values can be exported from this array to produce a two-dimensional ion intensity map, or ion image, constructed from the specific coordinate location of that signal and its corresponding relative abundance.

For high-resolution images, a matrix is deposited on a homogeneous manner to the surface of the tissue in such a way as to minimize the lateral dispersion of molecules of interest. This can be achieved either by automatically printing arrays of small droplets or by robotically spraying a continuous coating. Each micro spot, or pixel coordinate, is then automatically analyzed by MALDI MS. From the analysis of a single section, images at virtually any molecular weight may be obtained, provided there is sufficient signal intensity to record. The relatively low-resolution IMS data set in our study has 35×24 pixels, whereas the higher resolution IMS data set has 65×44 pixels. The spectrum associated with each pixel has over 30 000 m/z values and corresponding signal intensities. The high dimensionality in IMS data processing is often a very challenging issue.

Figure 1(a) shows the photomicrograph of a cresyl violet-stained mouse brain section, implanted with a GL26 glioma cell line and tumor growth. The darker region in the right hemisphere of the brain indicates the presence of the tumor. IMS data can be viewed as a three-mode array (data cube) with two spatial dimensions (x-, y-dimension) and the ion intensity values associated with m/z dimension as shown in Figure 1(b). The IMS data obtained from a serial section of the brain tissue is represented by mass spectra taken from each (x, y) coordinate or pixel. Figure 1(c) displays the mass spectrum at a single pixel. For each fixed m/z value, the false color image based on the ion intensity at each pixel is the ion image. Figure 1(d) depicts the 3D visualization of IMS data by showing five selected ion images. In Figure 1(d), pixel x coordinate, y coordinate, and m/z value are used to form a three-dimensional array, and then the intensity, represented by color, becomes a function value of these three variables. The five selected ion images shown in Figure 1(d) are produced by the MATLAB functions *meshgrid*, which can be used to represent volumetric data, and *slice*, which can be used to slice planes through volumetric data.

The spatial information provided by IMS is very helpful in the visual mapping of protein localization and the detection of sizeable differences in protein expression [22, 23]. Conventional MALDI MS data do not identify the spatially localized correlations between mass analytes that are latent in IMS data processing. Therefore, in IMS data processing it is advantageous to use both spectral and spatial information for feature extraction, even though doing so presents great challenges.

To fully utilize IMS data, it is desirable not only to identify the peaks of the spectrum within individual pixels, but also to study correlation and distribution using the spatial information for the entire image cube. It is also important to distinguish the selected feature m/z values according to the differences caused by biological structures of the tissue or purely by cancer. The combination of spatial information and mass resolution results in large and complex data sets and therefore presents a serious challenge in developing bioinformatics tools for the quantitative analysis and for the biological interpretation of the IMS data.

2.2. *EN model and LAR-EN algorithm*

The usual linear regression model can be described as follows.

Assuming p predictors $\mathbf{x}_1, \ldots, \mathbf{x}_p$, the response $\mathbf{y}$ is predicted by

$$\hat{\mathbf{y}} = \hat{\beta}_0 + \mathbf{x}_1\hat{\beta}_1 + \cdots + \mathbf{x}_p\hat{\beta}_p. \tag{1}$$

Given a data set, a model fitting procedure produces the vector of coefficients $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)^{\mathrm{T}}$. Ordinary least squares (OLS) estimates are obtained by minimizing the residual sum of squares (RSS). Ridge Regression minimizes the RSS subject to a bound on the ℓ_2 norm of the coefficients

$$\hat{\beta} = \operatorname{argmin}_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_2 \|\beta\|_2^2,$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ denotes the predictor matrix, and the penalty term associated with $\|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2$ is also called the ridge penalty term. λ_2 is a non-negative tuning/regularization parameter. The so-called lasso method minimizes the RSS subject to a bound on the ℓ_1 norm of the coefficients

$$\hat{\beta} = \operatorname{argmin}_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_1 \|\beta\|_1,$$

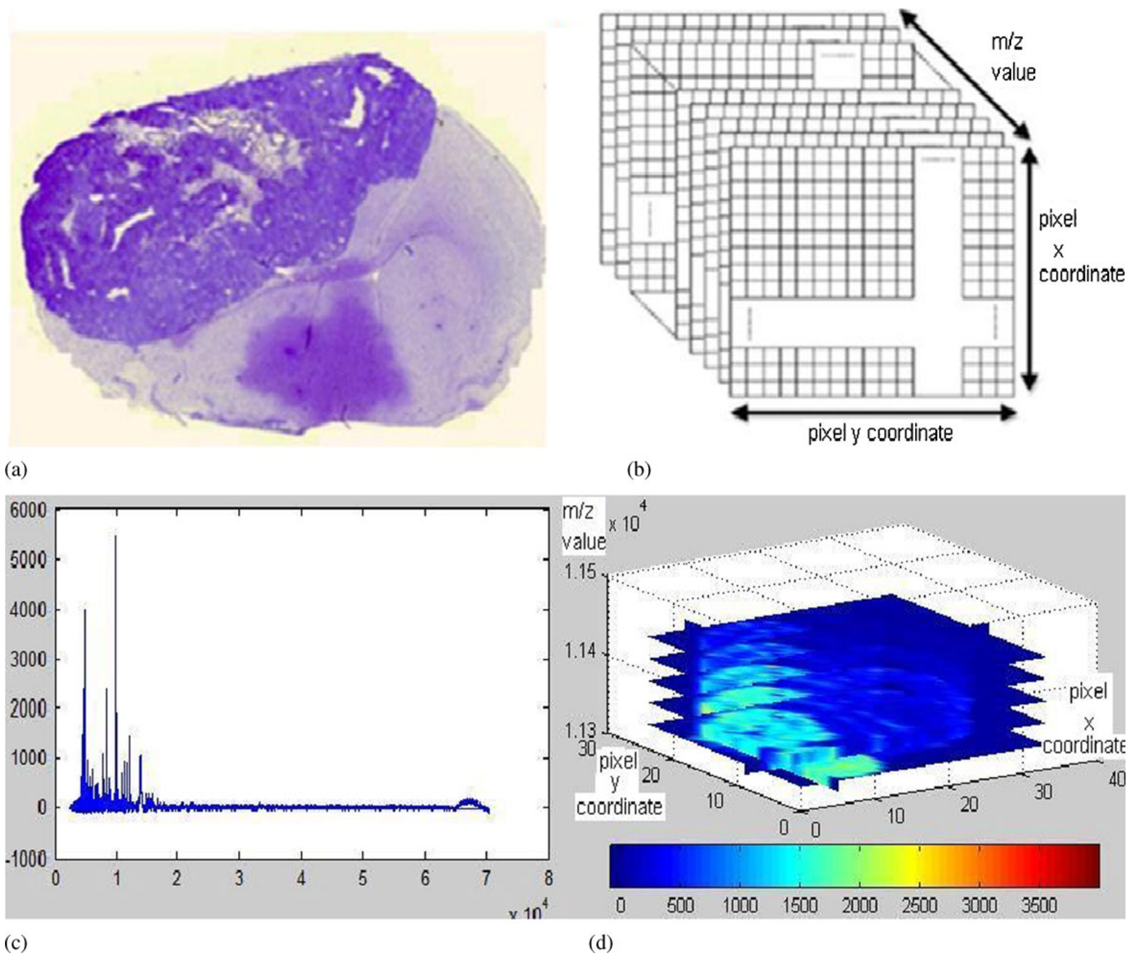

Figure 1. Mouse brain IMS data: (a) photomicrograph of a mouse brain section, implanted with a tumor growth; (b) the three-mode array representation of IMS data set with two spatial dimensions (x, y) and the m/z dimension; (c) an individual mass spectrum behind one pixel; and (d) the data cube visualization of IMS data set by showing five selected ion images.

where the penalty term associated with $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ is usually called the lasso penalty term. λ_1 is a non-negative tuning/regularization parameter. If the number of predictors, p, is greater than the number of observations, n, the lasso selects at most n variables. The number of selected predictors is bounded by the number of observations. In addition, the lasso fails to conduct grouped selection. That is, it tends to select one variable from a group and ignore others. However, the EN [18], a convex combination of the lasso and ridge penalty, typically outperforms them in many situations. The EN method is particularly useful in the case where $p \gg n$. It also encourages a grouping effect where strongly correlated predictors tend to be in or out of the model together.

The naive EN criterion is to minimize the following functional [18]:

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

The ℓ_1 component of the penalty functional generates a sparse model, while the quadratic part removes the limitation on the number of selected variables, encourages a grouping effect, and stabilizes the ℓ_1 regularization path. The non-negative tuning/regularization parameters λ_1 and λ_2 balance the goodness-of-fit and complexity of the model. Zou *et al.* [18] mentioned the naive EN appears to incur a double amount of shrinkage, which does not help to reduce the variances much and introduces an unnecessary extra bias, compared with pure lasso or ridge shrinkage. Therefore, the EN estimates $\hat{\beta}$ are defined by $\hat{\beta}(\text{elastic net}) = (1+\lambda_2)\hat{\beta}(\text{naive elastic net})$, which means that the EN coefficients are rescaled naive EN coefficients. Such a scaling transformation, namely, multiplying the naive EN coefficients by $1+\lambda_2$, removes the $1/(1+\lambda_2)$ shrinkage from ridge regression and overcomes the double shrinkage deficiency of the naive EN (see [18] for details). Finally, the EN estimates $\hat{\beta}$ are given by

$$\hat{\beta} = \text{argmin}_{\beta} \beta^{\mathrm{T}}((\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda_2 \mathbf{I})/(1+\lambda_2))\beta - 2\mathbf{y}^{\mathrm{T}}\mathbf{X}\beta + \lambda_1 \|\beta\|_1.$$

Efron *et al.* [24] proved that, starting from zero, the lasso solution paths grow piecewise linearly in a predictable manner, and a new algorithm called LARS was proposed to solve the entire lasso solution path efficiently by using the same order of computations as a single OLS fitting. For each fixed λ_2, the EN problem is equivalent to a lasso problem on an augmented data set. The algorithm LARS-EN was proposed in [18] to solve the EN efficiently based on the algorithm LARS.

2.3. EN4IMS model and its algorithm

In this section, we incorporate a spatial penalty term into the EN model in order to develop a new tool for IMS data biomarker selection and classification. Our motivation is to fully utilize not only the spectral information within individual pixels, but also the spatial information for the entire IMS data cube. A software package for comprehensive IMS data processing, called EN4IMS, is developed based on this new model.

2.3.1. Spatial penalty term. The importance of fully utilizing spatial information provided by IMS techniques has been emphasized in the recent literature (see [10, 12, 16] for example). Thus, we want to develop an algorithm of biomarker selection and classification that combines the spectral information within individual pixels with the spatial information for the entire IMS data set. It is also important to distinguish the selected m/z values according to the differences caused by biological structures of the tissue or by disease. The true cancer biomarkers can be effectively related to the cancer tissues and can potentially be used for cancer diagnosis. Therefore, it is critical to consider the spatial information in order to find the cancer-related features independent of tissue structures in IMS data processing.

In IMS data analysis, if a selected m/z value is truly associated with the disease being studied, then it is reasonable to expect that the ion intensity values for this m/z in different pixel locations in a diseased area are approximately the same. Therefore, the standard deviation of the intensities at the m/z value should be small. Conversely, if the m/z value selected by the statistical model is based on a differentiation mainly caused by the tissue structure, then the ion intensities at the m/z value would vary significantly from pixel to pixel in various tissue regions and, consequently, the standard deviation of intensities at that m/z value would be relatively large. Thus, it is appropriate to associate standard deviations at all selected predictors to the optimal model selection in order to impose the penalty functional on predictors caused by structure differences. In our work we incorporate such a spatial penalty into the EN model to develop the EN4IMS algorithm for IMS data analysis.

2.3.2. EN4IMS model. In the EN model, there are two tuning parameters λ_1, λ_2. It is shown in [18] that the parameter λ_1 is associated with the number of steps (k) in the LARS-EN algorithm and k can be used as the second tuning parameter besides λ_2. Therefore, two tuning parameters in the EN model can be considered as k and λ_2. There are well-established methods for choosing such tuning parameters [20]. In the two-dimensional cross-validation (CV) suggested in [18], λ_2 is typically chosen as a relatively small grid, say (0, 0.01, 0.1, 1, 10, 100). For a fixed λ_2, algorithm LARS-EN produces all possible EN estimates $\hat{\beta}$. The other tuning parameter k is selected by 10-fold CV. Some rules can be applied to select an optimal solution $\hat{\beta}$. The optimal step corresponding to the optimal solution can be chosen so that it minimizes the prediction error, that is the RSS in the CV step of the LARS-EN algorithm.

Based on the biological considerations discussed in Section 2.3.1, we incorporate the spatial standard deviation for the spatial penalty consideration into the EN model CV step. The 10-fold CV method divides the IMS data set into ten equally sized batches and estimates parameters ten different times by leaving one batch out each time. The testing error for each omitted batch is computed using the estimates derived from the remaining batches. The RSS and the average of spatial standard deviations of selected ion intensities are measured at every step. Finally, the optimal model is chosen as the one that can minimize the convex combination of the RSS and the average of spatial standard deviations of selected ion intensities.

We now present a detailed description for the EN4IMS algorithm. The pseudo code is given in Appendix A.

First, we apply the EN algorithm to find the entire solution path as described in [18]. Then, we select the optimal step k_{opt} in the 10-fold CV by minimizing

$$(1-\tau)\|\mathbf{y}-\hat{\mathbf{y}}_{S_k}\|_2^2+\frac{\tau}{M}\sum_{j=1}^{M}\sqrt{\frac{\sum_{i=1}^{N}(x_{ij}-\mu_j)^2}{N-1}}, \quad 0<\tau<1, \tag{3}$$

where N is the number of all cancer pixels, M is the cardinality of the active set S_k determined by the EN4IMS model in the step k (see Appendix A). x_{ij} is the intensity for a fixed jth m/z value in the pixel i and μ_j is the mean intensity over all these cancer pixels for a fixed jth m/z value. Comparing the error of each optimal model for each fixed λ_2, we choose the value of λ_2 that minimizes the error. This incorporates a penalty into the modified CV step.

The tuning parameter τ in (3) weighs the spatial penalty for balancing the contribution of the RSS against the spatial penalty term for the IMS data. For those peaks, or m/z values, with smaller standard deviations, the penalty effect on them will be smaller. In order to incorporate the LARS-EN algorithm into the EN4IMS package, we first fix τ and then use two-dimensional CV to select the parameters k, λ_2. We select τ based on experimental experience in the IMS data analysis. The effects of various choices of τ on the number of selected predictors and the number of selected structure-related predictors are discussed at the end of Section 3.1.

The EN4IMS model proposed here is for pixel-level classification and potential biomarker discovery. When entering the data, spectrum pixels from a cancer area and a noncancer area are purposely selected from the mouse brain IMS data sets to be as symmetric as possible relative to biological structural similarity. A master peak list of m/z values for all these pixels is generated. Although the number of m/z values is significantly larger than the sample size, the EN model is able to use them without dimension reduction. The early stop feature of the LARS-EN algorithm saves computation cost and time [18]. In the case where $p \gg n$, if the algorithm ends in q steps, then it only requires $O(q^3+pq^2)$ operations. We include all m/z values as predictors in (1), and $y_i=-1$ for the ith pixel in a noncancer area, else $y_i=1$.

The pseudo code of EN4IMS has been implemented by using MATLAB. Applications for a non-MATLAB environment are also provided in the software package. The package includes functions to visualize intensity distributions of IMS data both in two and three dimensions, functions of the EN model for real MS data analysis, the EN4IMS model for real IMS data analysis, other current popular algorithms used in IMS data analysis such as PCA, LDA, and SVM, as well as provides the functionality to create the ion images of certain given m/z values. By incorporating the spatial penalty term, this package is able to distinguish true biomarkers from features selected through structural difference, which is a crucial need in current IMS data processing. It also considers the grouping effect of these m/z values, where strongly correlated predictors tend to be in or out of the model together. Furthermore, the optimal model provides us with selected variables (a listing of m/z values) serving as potential biomarkers and can perform classification for unknown IMS data sets. The application results of biomarker selection and classification are shown in Section 3.3.

In the EN4IMS algorithm, the spatial penalty consideration is applied in the CV step. A more general model that incorporates a spatial penalty term directly into the EN model Equation (2) has been discussed in [25].

2.3.3. Comparison of the EN4IMS model with other algorithms. Data analysis generally has two components: preprocessing of data, followed by statistical analysis. Various algorithms are used for all of the spectra preprocessing steps including baseline subtraction, peak alignment, normalization, and peak picking [26, 27]. The use of MVA methods has opened new doors for the analysis of IMS data, such as PCA [11, 13–15], LDA [10, 12], and SVM [17]. To compare these methods, guidelines are needed for data preprocessing before they can be applied [28]. The requirements of data preprocessing for different software packages on IMS data processing are quite different. In comparison, the EN4IMS package can take data that have been minimally preprocessed even without peak binning/alignment beforehand.

In general, it is challenging to perform classification and biomarker selection accurately with IMS techniques since the m/z dimension is extremely high, and far greater than the sample size. PCA is a general tool for dimension reduction in classifier construction. By reordering all discrete spatial positions in the x- and y-directions, that is, 'pixels', into one long vector holding $I \cdot J$ elements for PCA on IMS data [11, 12, 15], a matrix D of size $(I \cdot J)\times K$ is formed, holding all the original information.

Because of the issue of global correlation, the use of PCA for the whole m/z range is typically preferred. Although it might seem to be computationally expensive, by transposing a matrix of dimension, say $a \times b$ with $a \ll b$, it is possible to reduce the computation cost from $O(b^3)$ to $O(a^3)$. However, there are still several drawbacks to using PCA for IMS data analysis.

PCA is a commonly used dimension reduction technique, which constructs new input variables using linear combinations of all original input variables. Since all input variables are used in construction of the super variables and hence classification, the biomedical implications of the classifiers are usually not obvious [29]. Furthermore, although PCA can be helpful in finding m/z values that represent the most significant variance, the variance may be caused by structural difference instead of cancer. PCA typically highlights the variations due to anatomical features first, and the features of interest are hidden in the later principal components described by just a small amount of variance [16].

K-means is combined with PCA for IMS data analysis in [10]. The inputs of cluster methods are the principal components selected by PCA, since the m/z dimension is too high. Based on our experience, these inaccuracies in PCA do, in fact, introduce inaccuracies in cluster results. This unsupervised classification could be further improved by using a supervised technique such as LDA or SVM.

LDA aims to maximize the ratio of between-class variance to within-class variance. LDA was combined with PCA for IMS data analysis [10, 12]. To use LDA, the number of pixels in the groups being analyzed should be larger than the number of data points in spectra, which is usually not the case with IMS data. Because of that [10, 12] propose using PCA as a technique for dimension reduction first. In addition, LDA implicitly assumes that the mean is the discriminating factor (not variance) and that the data are normally distributed. Such assumptions limit the application of LDA.

Linear SVM combined with PCA was also used for IMS data analysis in [17]. To use linear SVM, one usually uses PCA for dimension reduction. Thus, SVM suffers from the same problems as these in PCA noted above. Furthermore, SVM is designed for classification, not feature selection. SVM itself cannot select features automatically and uses either univariate ranking or recursive feature elimination to reduce the number of features in the final model [18]. Consequently, this method is not effective for biomarker selection.

By incorporating the spatial penalty term, our proposed EN4IMS model can effectively select cancer-related features instead of structure-related features. It encourages a grouping effect as well. Cancer may affect some functional proteins, and thus peptides related to these proteins should be in or out of the model together. In summary, the EN4IMS model for IMS data processing shows many advantages and can outperform the algorithms that it has been compared against. In Section 3, simulation data and real IMS data sets are used to further evaluate the EN4IMS model.

3. Results

In this section, extensive simulation studies are conducted to evaluate the performance of the proposed EN4IMS algorithm. Further, the EN4IMS algorithm has been applied to two real IMS data sets of mouse brain cancer generated by the Vanderbilt Mass Spectrometry Research Center. The analysis includes a comparison of results of the EN4IMS algorithm, EN, PCA, LDA, and SVM, as well as the results obtained by using the commercial software SAM. The results show that the EN4IMS algorithm, compared with the EN algorithm, produces a more concise listing of peaks in the sense of including all significant features, but a smaller number of shoulder/noise peaks. The EN4IMS algorithm confirms new biomarkers, and also provides better classification results comparing with other currently popular analysis methods in the IMS community.

3.1. Simulation

For simulation purposes, we first generate an IMS data set based on cancer mean spectrum and noncancer mean spectrum from a real data set consisting of 104 cancer spectra (pixels) and 105 noncancer spectra (pixels). From these mean spectra of this real IMS data set, 502 peak features were selected. The m/z values with intensity difference between cancer and noncancer mean spectra larger than 350 are defined as DE-features, that is, the features are differentially expressed between these two groups. In this data set, the number of DE-features is 46. We divide these DE-feature peaks randomly into predefined ‘true biomarkers’ and predefined structure-related feature peaks.

The simulation was conducted based on two cases of the number ℓ of predefined ‘true biomarkers’ and two scenarios of the intensities’ standard deviation σ_2 from these predefined structure-related peaks.

For the first case, 16 'true biomarkers' are randomly selected from 46 DE-feature peaks ($\ell=16$). The remaining 30 peaks are defined as structure-related peaks. For the second case, 18 'true biomarkers' are randomly selected from 46 DE-feature peaks ($\ell=18$). The remaining 28 peaks are then defined as structure-related peaks. The set of biomarkers in the first case only shares two peaks with the set of biomarkers in the second case. This setting of the simulation helps to reduce biases.

We assume that intensities of true biomarkers have smaller variation in cancer areas, whereas intensities of structure-related feature peaks have larger variation in cancer areas. Given that, and based on experimental observations of the data sets, we took the standard deviation σ_1 for these predefined true biomarkers in the cancer group as $\frac{1}{2}(0.5\% \times$ corresponding mean intensity$)+\frac{1}{2}c_1$ for $c_1=10$; the standard deviation σ_2 for these predefined structure-related feature peaks in the cancer group is selected in two scenarios: $\frac{1}{2}(4\% \times$ corresponding mean intensity$)+\frac{1}{2}c_2$ for $c_2=40$ and $c_2=50$, respectively; the standard deviations σ_3 and σ_4 both as $\frac{1}{2}(1.5\% \times$ corresponding mean intensity$)+\frac{1}{2}c_3$ with $c_3=20$ for other 456 peak features in the cancer group and for these 502 peak features in the noncancer group, respectively. In summary, we have the following four corresponding settings for simulation: $(\ell, c_2)=(16, 50), (16, 40), (18, 50), (18, 40)$.

Cancer and noncancer intensity data are generated from a normal distribution with predefined standard deviation σ_1, σ_2, σ_3, σ_4 for those 502 peak features and corresponding means based on the cancer and noncancer mean spectra. The sample sizes for both cancer and noncancer data are chosen to be 50. The EN and EN4IMS are implemented using the same parameters: $STOP=-40$ (see Appendix A) and $\lambda_2=0.1$. The parameter for spatial penalty τ in the EN4IMS algorithm is set to be $\frac{1}{4}$ when $c_2=50$ and $\frac{1}{3}$ when $c_2=40$. The discussion of the selection of τ can be found at the end of this section.

Based on these settings, the simulations were run 50 times for each of the three algorithms (EN4IMS, EN, PCA) in each of the four settings. Simulation results are shown in Figures 2 and 3, respectively.

Figure 2 shows the number of selected predictors using EN4IMS, EN, and PCA methods in four different settings. Since the $STOP=-40$ for EN4IMS and EN, the threshold for PCA is chosen to make PCA algorithm selecting predictors to be around 40. Both EN4IMS and EN select all the 'true biomarkers' but PCA misses at least one-third of the 'true biomarkers'. In settings of both $(\ell, c_2)=(16, 50)$ and (16, 40), PCA selects around half of the 'true biomarkers' and it selects around two-third of the 'true biomarkers' in settings of $(\ell, c_2)=(18, 50)$ and (18, 40). Furthermore, the EN4IMS generally selects predictors less than the other two methods. If the number of selected predictors by PCA is further reduced by using a larger threshold, this will cause PCA to miss even more 'true biomarkers'. However, PCA has a smaller variation in the number of selected predictors.

Figure 3 shows the number of selected structure-related predictors of EN4IM, EN, and PCA methods in four different settings. With the same predefined 'true biomarkers' ($\ell=16$), both EN and EN4IMS

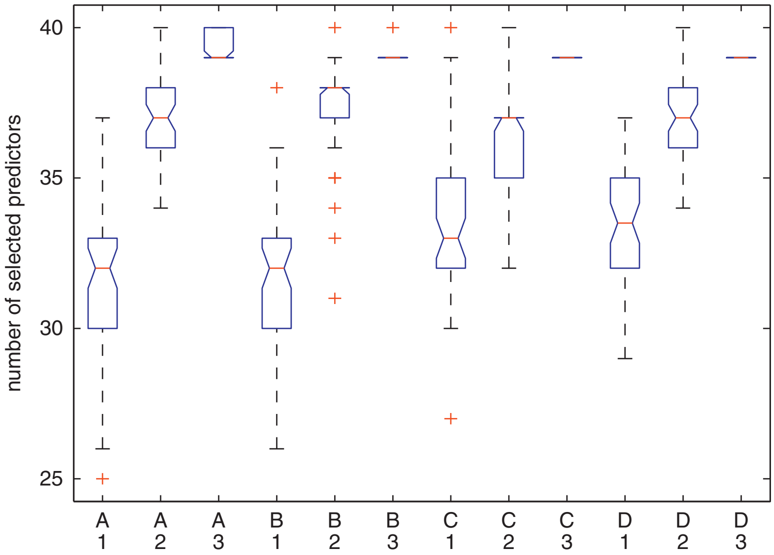

Figure 2. Box-plot of the number of selected predictors [numbers 1, 2, and 3 stand for three algorithms EN4IMS, EN and PCA; letters A, B, C, and D stand for four settings $(\ell, c_2)=(16, 50)$, (16, 40), (18, 50), and (18, 40)].

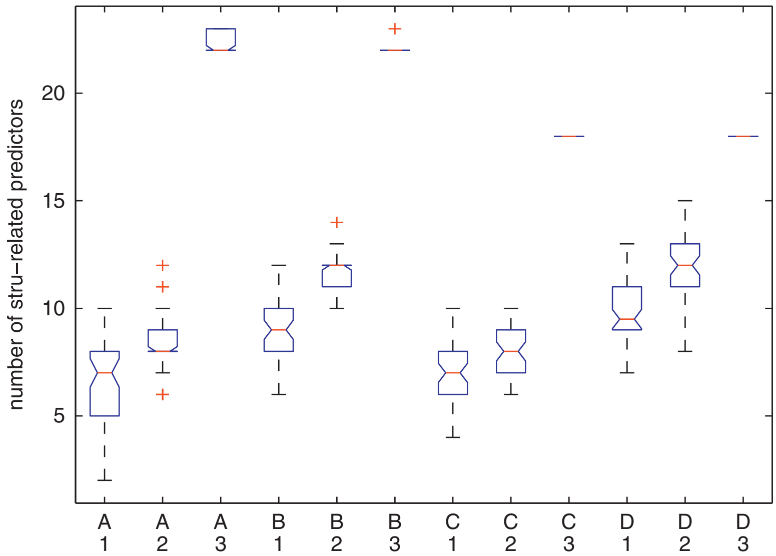

Figure 3. Box-plot of the number of selected structure-related predictors [numbers 1, 2, and 3 stand for, respectively, three algorithms EN4IMS, EN, and PCA; letters A, B, C, and D stand for, respectively, four settings $(\ell, c_2) = (16, 50)$, (16, 40), (18, 50), and (18, 40)].

select more structure-related predictors in the setting of $c_2 = 40$ than in the setting of $c_2 = 50$. Similarly, with $\ell = 18$, both EN and EN4IMS select more structure-related predictors in the setting of $c_2 = 40$ than in the setting of $c_2 = 50$. However, the EN4IMS generally keeps selecting a more concise list of predictors compared with EN and PCA in all four settings. In addition, there are almost no differences for the number of selected structure-related predictors by PCA when c_2 is changed from 40 to 50 for a given ℓ value. Similarly, in Figure 2, the number of selected predictors by PCA varies little when c_2 is changed from 40 to 50 for a given ℓ value. PCA is not sensitive to a variation of intensities in a subset of the data, though PCA operation can be thought of as revealing the internal structure of the data in a way which best explains variance in the data.

The parameter τ controls a weight on the spatial penalty term by balancing the contribution of the RSS and spatial penalty term. In the CV step, the optimal step is chosen so as to minimize the convex combination of the RSS and the average of spatial standard deviations of selected ion intensities. Therefore, the larger value of τ gives more weight to the spatial penalty.

Extensive simulation studies on various choices of τ were also conducted. Simulation settings were $\ell = 18$ with $c_2 = 40$. The EN4IMS algorithm was run 50 times for each of the four τ values. The box-plots in Figure 4 illustrate the effect of various choices of τ on the number of selected predictors and on the number of selected structure-related predictors. We can see that the number of selected predictors and the number of selected structure-related predictors decrease as τ increases.

In real IMS data analysis, we can calculate standard deviations for each m/z value from the sample data. Then based on the obtained standard deviations and the level of penalty on the spatial variations, the value τ can be chosen subjectively. In simulation results illustrated in Figures 2 and 3, we have chosen τ to be $\frac{1}{4}$ when $c_2 = 50$ and $\frac{1}{3}$ when $c_2 = 40$ in the EN4IMS algorithm. Also, we use $\tau = \frac{1}{4}$ in the real data analysis in Section 3.3.

3.2. Experiments

The EN4IMS algorithm has been applied to two real IMS data sets. Both data set-1 and set-2 are on a GL26 glioma study. Figure 5 shows the stained mouse brain sections corresponding to data set-1 (left) and set-2 (right), respectively. The darker areas in this figure indicate the presence of the tumor as confirmed by a trained pathologist. Data set-1 has relatively higher resolution (more pixels/spectra) than data set-2.

In corresponding experiments, C57 black mice were implanted with a GL26 glioma cell line, and tumor growth was allowed to occur for 15 days. The mouse brains were excised, flash-frozen, sectioned on a cryostat (12 µm), and thaw-mounted onto gold-coated MALDI targets. The brain tissue was then

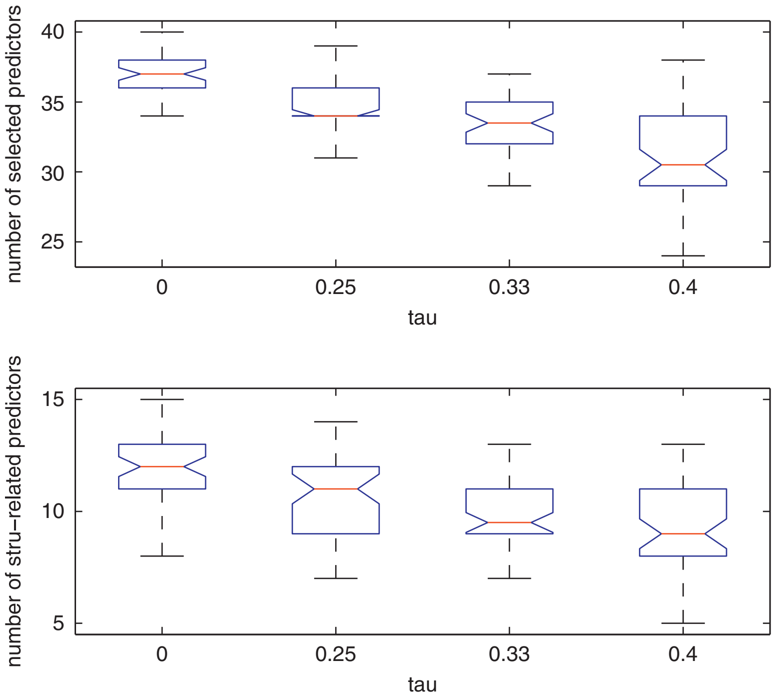

Figure 4. Relationship of the number of selected predictors and the number of selected structure-related predictors with τ.

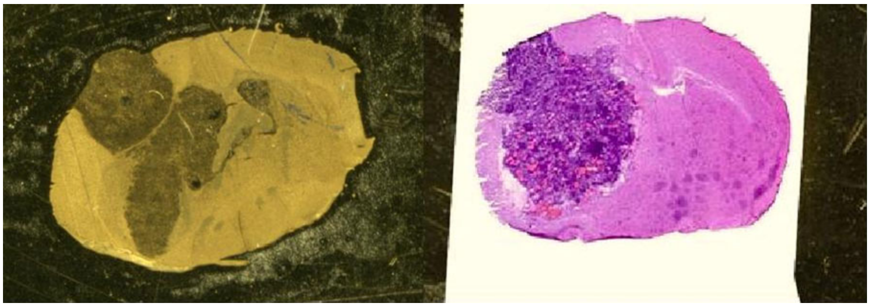

Figure 5. Slide pictures of mouse brains with tumor. The left one is from the data set-1 and the right one is from the data set-2. The shape of the cancer area and noncancer area can be used to compare with the ion images of selected biomarkers.

spotted with sinapinic acid for protein images using an acoustic reagent multispotter (Labcyte). Protein images were acquired for each of the brain sections using a MALDI-TOF-IMS (Bruker) at a resolution of 300 µm by 300 µm. After data acquisition, the data underwent a series of basic preprocessing steps to reduce the experimental variance between spectra by removing background noise, normalizing the peak intensity to the total ion current, and binning peaks if needed.

Various algorithms were employed for all of the spectra processing steps as a part of the PROTS Data program from BioDesex before applying significance analysis of microarrays (SAM) [30] to generate the SAM feature list in Appendix B. In comparison, the EN4IMS package can take data that have been minimally preprocessed even without peak binning/alignment beforehand. This saves a significant amount of time spent on data processing.

3.3. Results and biological interpretation

Listings of selected peaks obtained by processing the IMS data set-2 using the EN4IMS, SAM, EN, and PCA algorithms are included in Appendix B. For the IMS data set-2, the EN4IMS algorithm generates a list of m/z values that matches significant features obtained by using the SAM and provides a much more condensed list by removing noise/shoulder peaks. In addition, the EN4IMS identifies the tumor signal peak ($m/z = 14\,788$), which is not on the SAM list. Therefore, the EN4IMS algorithm helps in confirming peaks found by other algorithms and finds interesting regions in the spectrum missed by other algorithms that are potential new biomarkers. The biological interpretation of the biomarker ($m/z = 14\,788$) is discussed later in the paper.

Compared with the m/z list generated by the EN algorithm, the newly developed EN4IMS algorithm that incorporates the spatial penalty term produces an even more concise list by including all significant features and has a smaller number of noise/shoulder peaks. These noise/shoulder peaks are most likely fake ones caused mainly by noise.

The EN4IMS algorithm found more important biomarkers than the PCA method [11, 12, 15]. The peaks ($m/z = 6700, 8380, 10\,952, 14\,788$) described below are on the EN4IMS list but not on the PCA list (Appendix B). Notice also that a so-called binning algorithm needs to be run usually on MS peaks in order to obtain cross samples alignment in MS data processing [31]. Peaks within a five-dalton shift of a central m/z value ($\geqslant 5000$) are usually considered to be the same peaks. Therefore, these four biomarkers correspond to the m/z values of 6702, 8384, 10949, and 14786 in the EN4IMS list.

The feature peaks ($m/z = 6700$, 8380, 10 952, 14 788) have been proved to be important biomarkers in cancer research [32–36]. Furthermore, a series of experiments were performed to correctly identify the proteins of our selected features of interest. First, tissues were homogenized in Tissue Protein Extraction Reagent (TPER; Pierce, Rockford, IL) and supplemented with protease inhibitors. For each extract, a Vydac C8 polymeric reversed-phase column (3.2 × 150 mm) fractionated 300 g of protein solution (96 min linear gradient from 2 to 90 plate). During separation, a liquid handling robot moves the transfer capillary sequentially into each of the 96 wells at 1 min intervals. To identify wells containing proteins of interest in an automated process, 0.2 L was removed from each well, mixed with SA matrix, and analyzed by MALDI MS (Bruker Autoflex). Fractions containing m/z values of interest were run on a gel (10–20tricine) and bands of interest were excised and digested with trypsin gold (Promega; Madison, WI). In the solution trypsin digest was also performed on fractions that contained m/z values of interest. Either an LTQ or an LCQ (Thermo Scientific; Waltham, MA) was used to analyze digested proteins. The peptides were separated on a packed capillary column, 75 m × 10.5 cm, with C18 resin (Monitor C18, 5 m; Column Engineering, Ontario, CA), using a linear gradient (5MS/MS spectra were initially analyzed by searching the mouse International Protein Index database using Sequest software). ProteinProphet software was then used to determine the probability that a protein had been correctly identified based on the available peptide sequence evidence.

Figure 6 shows ion images of six m/z values, including four cancer biomarkers ($m/z = 6700, 8380, 10\,952, 14\,788$) mentioned above together with two non-DE feature peaks ($m/z = 7500, 4000$) for comparison. Non-DE features are defined as the m/z values where intensity is not differentially expressed between cancer and noncancer groups. For the four images in the right two columns, the intensity differences between cancer and noncancer areas are very clear. For the two ion images of biomarkers with m/z values of 6700 and 8380, respectively, intensities are much lower in one area than intensities outside that area. In the other two ion images of biomarkers with m/z values of 10 952 and 14 788, respectively, intensities are much higher in one area than intensities outside that area. Compared to the mouse brain tissue section in Figure 5 (right), these four ion images show the shape of the tumor area very well. By contrast, the images in the left column of the non-DE features ($m/z = 7500, 4000$) do not depict differences between cancer and noncancer regions at all.

Protein identification experiments provided identities of important biomarker peaks, including cytochrome c oxidase copper chaperone ($m/z = 6700$) and cytochrome C oxidase subunit 6c ($m/z = 8380$), which are involved in the electron transport chain. The electron transport chain removes electrons from the donor, NADH, and passes them to a terminal electron acceptor, O_2, via a series of redox reactions. Several recent studies have linked impaired mitochondrial function as well as impaired respiration to the growth, division, and expansion of tumor cells; this is known as the Warburg effect [32, 33]. The Warburg effect is described as the dependency of tumors on glycolysis rather than on oxidative phosphorylation for ATP even in the presence of oxygen. This explains why the cytochrome c oxidase copper chaperone and the cytochrome c oxidase subunit 6c have decreased signal intensities in the tumor areas of the brain.

Additional experiments identified signals including calgizzarin ($m/z = 10\,952$) and an acetylated form of Histone $H2A$ ($m/z = 14\,788$). These signal intensities were found to be increased in the tumor areas of the brain. Calgizzarin, a calcium binding protein, has been implemented in the processes of proliferation,

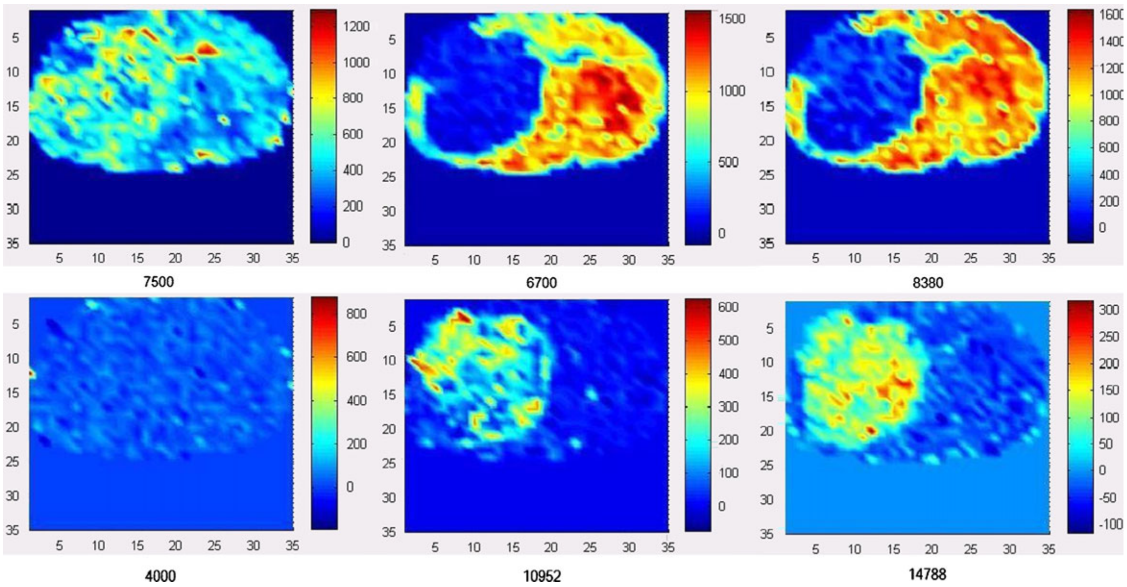

Figure 6. Ion images of six m/z values including four important biomarkers (right two columns) and two non-DE features (the left column). The shape of cancer area can be compared with Figure 5 (right) based on data set-2.

differentiation, and accelerated metabolism in cancer cells, although its detailed function is not yet known [34, 37, 38]. Histones tail modifications, such as acetylation, methylation, phosphorylation, and ubiquitination, along with DNA methylation, are the most studied epigenetic events related to cancer progression [35, 36]. Histone modifications promote or prevent the binding of proteins and protein complexes that drive particular regions of the genome into active transcription or repression.

By examining the details of the intensity increasing or decreasing trends, we found that most m/z values in the EN4IMS list have a decreasing trend in the tumor area. By plotting the difference of mean spectrum of normal data and mean spectrum of tumor data, we can see the whole data set is negatively associated overall. Since the EN4IMS algorithm is based on a linear regression model, if the data set is negatively associated overall, then it is likely to only pick up negative differentiations.

Interestingly, when $p \gg n$, linear classifiers often perform better than non-linear ones in many applications [20], even though non-linear methods are known to be more flexible. This fact is related to the asymptotic results [39]: when $p \gg n$, under mild assumptions for data distribution, the pairwise distances between any two points are approximately identical to each other so that the data points form an n-simplex. Linear classifiers then become natural choices to discriminate between two simplices [40].

Both IMS technology and IMS data processing are new fields, therefore it is interesting to incorporate new methods into IMS data analysis. In Table I, we compare classification results of the EN4IMS algorithm with those of other current popular methods used in the IMS community. Here, PCA+LDA, PCA+SVM algorithms are implemented according to [10] and [17]. In data set-2, there are two mouse brain tissue sections. One is used for model training and the other section is used for model testing. 110 pixels are selected from the cancer area to be used as the training cancer data set, and 110 pixels are selected from the normal area to be used as the training noncancer data set. Similarly, 110 cancer pixels and 110 noncancer pixels are selected from the second mouse brain tissue section as test data. Results are shown in the first three rows of Table I. In addition, we use data set-1 from a different mouse brain for testing. In this test, 99 cancer pixels and 96 noncancer pixels are selected. Results are shown in the last three rows of Table I. In both cases, the EN4IMS algorithm shows the best classification results.

4. Conclusion

The major result of this paper shows that incorporating a modern statistical regularization and variable selection method, called the EN model, into the IMS data processing procedure by adding a spatial penalty term results in a better utilization of advantages of the IMS technology. An algorithm package, called EN4IMS, has been developed, which allows us to process IMS data that have been minimally preprocessed even without peak binning beforehand, obtain an accurate classification rate, and increase

Table I. Classification results comparison of the EN4IMS algorithm with other methods.

Methods	Accuracy (per cent)	Sensitivity (per cent)	Specificity (per cent)
PCA+LDA	78.64	100	57.27
PCA+SVM	71.82	84.56	59.09
EN4IMS	99.09	100	98.18
PCA+LDA	82.05	98.99	64.58
PCA+SVM	70.77	100	40.63
EN4IMS	100	100	100

reproducibility. The EN4IMS algorithm provides a framework that can deal with MALDI-TOF IMS data and helps in confirming new potential biomarkers. In summary, detection and quantification of IMS biomedical data features is a key step to biomarker discovery. The EN theory and the EN4IMS algorithm development ensure that the analysis results possess the desired qualities of precision, efficiency, robustness, and reproducibility. MATLAB scripts used to implement the methods described in this paper, along with supplementary data sets, can be found at `http://www.mtsu.edu/~dhong/EN4IMS.htm`.

Appendix A

In the following pseudo code, $\mathbf{X}$ is the input variable matrix and $\mathbf{y}$ is the response variable vector. The value of $\mathbf{y}_j$ is positive one when the pixel is in cancer region or negative one when the pixel is in noncancer area. λ_1, λ_2 are the regularization parameters. The EN4IMS algorithm inherits the early stop feature of LARS-EN [18] by including the input value STOP which has the following functions:

1. If STOP is negative, its absolute value is the desired number of predict variables selected for the model.
2. If STOP is positive, it corresponds to an upper bound on the ℓ_1 norm of the beta coefficients.
3. If STOP is zero, the pseudo code as below allows the generation of the entire solution path.

Algorithm (EN4IMS)

```
1. Input predictor matrix $\mathbf{X}$ of covariate vectors $\mathbf{x}_j$, the response vector $\mathbf{y}$. Set estimate coefficient $\hat{\beta}=0$, step $k=0$.
2. Define correlations $\hat{\mathbf{C}}_k=\mathbf{X}^{\mathrm{T}}(\mathbf{y}-\hat{\mathbf{y}}_{S_k})$, where estimate $\hat{\mathbf{y}}_{S_k}=\mathbf{X}_{S_k}\hat{\beta}_{S_k}, \mathbf{X}_{S_k}=(\ldots s_j\mathbf{x}_j\ldots), s_j=\mathrm{sgn}\{\hat{c}_{jk}\}$ for $j\in S_k$. Active set $S_k$ is the set of indices corresponding to covariates with the greatest absolute correlation, $S_k=\{j:|\hat{c}_{jk}|=C_M\}$. The greatest absolute correlation $C_M=\max_j\{|\hat{c}_{jk}|\}$.
   While $(S_k^c\neq\emptyset)$ Do
   (a) $\mathbf{G}_{S_k}=\mathbf{X}_{S_k}^{\mathrm{T}}\mathbf{X}_{S_k}$, $A_{S_k}=(\mathbf{1}_{S_k}^{\mathrm{T}}\mathbf{G}_{S_k}^{-1}\mathbf{1}_{S_k})^{-1/2}$
   (b) Calculate equiangular vector
   $\mathbf{u}_1=\mathbf{X}_{S_k}\mathbf{\Omega}_{S_k}d_2$
   $\mathbf{u}_2=d_1d_2\mathbf{\Omega}_{S_k}$
   where $\mathbf{\Omega}_{S_k}=A_{S_k}\mathbf{G}_{S_k}^{-1}\mathbf{1}_{S_k}$, $d_1=\sqrt{\lambda_2}, d_2=\frac{1}{\sqrt{1+\lambda_2}}$.
   (c) Calculate the inner product vector which represents the correlation between each variable and equiangular vector
   $\mathbf{a}=(\mathbf{X}^{\mathrm{T}}\mathbf{u}_1+\mathbf{u}_2d_1)d_2$
   (d) Update current algorithm estimate
   $\hat{\mathbf{y}}_{S_{k+1}}=\hat{\mathbf{y}}_{S_k}+\hat{\gamma}\mathbf{u}_1$
   where $\hat{\gamma}=\min^{+}_{j\in S_k^c}\left\{\frac{C_M-\hat{c}_{jk}}{A_{S_k}-a_j}, \frac{C_M+\hat{c}_{jk}}{A_{S_k}+a_j}\right\}$
   (e) Update the active set $S_k$
   if $\tilde{\gamma}<\hat{\gamma}$, $S_{k+1}=S_k-\{\tilde{j}\}$
   else $S_{k+1}=S_k+\{\tilde{j}\}$
   where $\tilde{\gamma}=\min_{\gamma_j>0}\{\gamma_j\}, \gamma_j=-\hat{\beta}_j/(s_j\mathbf{\Omega}_{S_kj})$ for $j\in S_k$
   (f) $k=k+1$
   End Do
3. Find step $k_{\mathrm{opt}}$ to select the optimal model by using 10-fold cross-validation to minimize the following functional $(1-\tau)\times\|\mathbf{y}-\hat{\mathbf{y}}_{S_k}\|_2^2+\frac{\tau}{M}\sum_{j=1}^{M}\sqrt{\sum_{i=1}^{N}(x_{ij}-\mu_j)^2/N-1}$ for $j\in S_k$
```

Appendix B

Listings of selected peaks in terms of m/z values using the EN4IMS, SAM, EN, and PCA algorithms.

EN4IMS list	SAM list	EN list	PCA list
4664	2791 3434 8337	4476 13 562	4934 8567
4667	3010 3764 8366	4664 14 327	4936 10 257
4670	3056 4011 8380	4670 14 336	4937 10 259
4812	3734 4076 8395	4812 14 343	4938 10 261
5446	3800 4271 8492	4884 14 781	4939 10 263
5753	3920 4538 8672	5425 14 786	4960 14 969
5754	4206 4566 8945	5429 14 805	4962 14 971
5756	4341 4665 8982	5446	4963 14 974
5757	4605 4676 9327	5753	4964 14 976
6165	4734 4899 9343	5754	4966 14 979
6702	4767 5106 9531	5756	5439 14 981
6706	4921 5120 9602	6165	5441 14 983
7799	4936 5428 9619	6702	5442 14 986
8019	4964 5444 10 238	6706	5444 15 603
8024	4981 5707 10 267	6794	5445 15 606
8384	5001 5753 10 466	7799	5446 15 608
8386	5024 6166 10 662	8019	5448 15 611
9344	5170 6186 12 434	8024	5449 15 613
10 172	6225 6251 13 560	8028	5451 15 616
10 261	7706 6310 14 525	8384	6571 15 618
10 263	8420 6574	8386	6572 15 620
10 265	8603 6700	8495	6574 15 623
10 267	8709 6719	8524	6575 15 625
10 282	8747 6780	9344	6577 16 780
10 366	9062 7099	9553	7749 16 782
10 374	9736 7118	10 172	7751 16 785
10 825	9956 7297	10 261	7752 16 787
10 949	10 167 7315	10 263	7792
13 562	10 952 7338	10 267	7794
14 336	11 388 7357	10 282	7795
14 343	11 640 7751	10 366	7797
14 781	12 203 7776	10 374	8560
14 786	14 865 7795	10 811	8562
14 805	14 927 8025	10 825	8564
	14 978 8107	10 949	8566

Acknowledgements

The authors are very grateful to anonymous referees for their constructive comments. They thank Shannon Cornett, Sara Frappier, and Richard M. Caprioli from the VUMSRC for valuable discussions and providing IMS data sets for the study, and also Andrew Worsey for many valuable suggestions, which substantially improved the manuscript. This research was partially supported by Middle Tennessee State University Faculty Research Grant #2-21519 and Yue-kong Pao Endowment Fund at Ningbo University, China.

References

1. Chaurand P, Schwartz SA, Caprioli RM. Profiling and imaging proteins in tissue sections by MS. *Analytical Chemistry* 2004; **76**(5):86–93.
2. Touboul D, Kollmer F, Niehuis E, Brunelle A, Laprevote O. Improvement of biological time-of-flight secondary ion massspectrometry imaging with bismuth cluster ion source. *Journal of the American Society for Mass Spectrometry* 2005; **16**:1608–1618.
3. Altelaar AF, Taban IM, McDonnell LA, Verhaert PD, de Lange RP, Adan RA, Mooi WJ, Heeren RA, Piersma SR. High-resolution MALDI imaging mass spectrometry allows localization of peptide distributions at cellular length scales in pituitary tissue sections. *International Journal of Mass Spectrometry* 2007; **260**:203–211.
4. Chaurand P, Rahman MA, Hunt T, Mobley JA, Gu G, Latham JC, Caprioli RM, Kasper S. Monitoring mouse prostate development by profiling and imaging mass spectrometry. *Molecular and Cellular Proteomics* 2008; **7**:411–423.
5. Khatib-Shahidi S, Andersson M, Herman JL, Gillespie TA, Caprioli RM. Direct molecular analysis of whole-body animal tissue sections by imaging MALDI mass spectrometry. *Analytical Chemistry* 2006; **78**:6448–6456.

6. Cornett DS, Reyzer ML, Chaurand P, Caprioli RM. MALDI imaging mass spectrometry: molecular snapshots of biochemical systems. *Nature Methods* 2007; **4**:828–833.
7. Stoeckli M, Chaurand P, Hallahan DE, Caprioli RM. Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues. *Nature Medicine* 2001; **7**:493–496.
8. Chaurand P, Schwartz SA, Caprioli RM. Assessing protein patterns in disease using imaging mass spectrometry. *Journal of Proteome Research* 2004; **3**:245–252.
9. Wickes BT, Yongmin K, Castner DG. Denoising and multivariate analysis of time-of-flight SIMS images. *Surface and Interface Analysis* 2003; **35**:640–648.
10. McCombie G, Staab D, Stoeckli M, Knochenmuss R. Spatial and Spectral correlation in MALDI mass spectrometry images by clustering and multivariate analysis. *Analytical Chemistry* 2005; **77**:6118–6124.
11. Van de Plas R, Ojeda F, Dewil M, Van Den Bosch L, De Moor B, Waelkens E. Prospective exploration of biochemical tissue composition via imaging mass spectrometry guided by principal component analysis. *Pacific Symposium on Biocomputing*, Maui, vol. 12, 2007; 458–469.
12. Muir ER, Ndiour IJ, Le Goasduff NA, Moffitt RA, Liu Y, Sullards MC, Merrill AH, Chen Y, Wang MD. Multivariate analysis of imaging mass spectrometry data. *BIBE 2007 Proceedings of the Seventh IEEE International Conference*, Boston, MA, 2007; 472–479.
13. Trim PJ, Atkinson SJ, Princivalle AP, Marshall PS, West A. Matrix-assisted laser desorption/ionisation mass spectrometry imaging of lipids in rat brain tissue with integrated unsupervised and supervised multivariant statistical analysis. *Rapid Communications in Mass Spectrometry* 2008; **22**:1503–1509.
14. Deininger SO, Ebert MP, Futterer A, Gerhard M, Rocken C. MALDI imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers. *Journal of Proteomic Research* 2008; **7**(12):5230–5236.
15. Van de Plas R, De Moor B, Waelkens E. Imaging mass spectrometry based exploration of biochemical tissue composition using peak intensity weighted PCA. *Life Science Systems and Applications Workshop*, *LISA*. *IEEE/NIH*, Bethesda, MD, 2007; 209–212.
16. McDonnell LA, Van Remoortere A, Van Zeijl RJ, Deelder AM. Mass spectrometry image correlation: quantifying colocalization. *Journal of Proteomic Research* 2008; **7**:3619–3627.
17. Gerhard M, Deininger SO, Schleif FM. Statistical classification and visualization of MALDI imaging data. *CBMS'07*, Maribor, Slovenia, 2007; 0-7695-2905-4/07.
18. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, *Series B* 2005; **67**:301–320.
19. Zou H, Hastie T, Tibshirani R. On the 'Degrees of Freedom' of the Lasso. *Annals of Statistics* 2007; **35**:2173–2192.
20. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. *Data Mining*, *Inference and Prediction*. Springer: New York, 2001.
21. Heeren RMA, Smith DF, Stauber J, Kkrer-Kaletas B, MacAleese L. Imaging mass spectrometry: hype or hope? *Journal of the American Society for Mass Spectrometry* 2009; **20**(6):1006–1014.
22. Schwartz SA, Weil RJ, Johnson MD, Toms SA, Caprioli RM. Protein profiling in brain tumors using mass spectrometry: feasibility of a new technique for the analysis of protein expression. *Clinical Cancer Research* 2004; **10**:981–987.
23. Yanagisawa K, Shyr Y, Xu BJ, Massion PP, Larsen PH, White BC, Roberts JR, Edgerton M, Gonzalez A, Nadaf S, Moore JH, Caprioli RM, Carbone DP. Proteomic patterns of tumour subsets in non-small-cell lung cancer. *Lancet* 2003; **362**(9382):433–439.
24. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Annals of Statistics* 2004; **32**:407–499.
25. Hong D, Zhang F. Weighted elastic net model for mass spectrometry imaging processing. *Mathematical Modelling Natural Phenomena* 2010; **5**(3):115–133.
26. Norris JL, Cornett DS, Mobley JA, Andersson M, Seeley EH, Chaurand P, Caprioli RM. Processing MALDI mass spectra to improve mass spectral direct tissue analysis. *International Journal of Mass Spectrometry* 2007; **260**:212–221.
27. Hong D, Shyr Y. Mathematical framework and wavelets applications in proteomics for cancer study. In *Handbook of Cancer Models with Applications to Cancer Screening*, *Cancer Treatment and Risk Assessment*, Tan WY, Hannin L (eds). World Scientific Publication: Singapore, 2008; 471–499.
28. Wagner MS, Graham DJ, Castner DG. Simplifying the interpretation of ToF-SIMS spectra and images using careful application of multivariate analysis. *Applied Surface Science* (*SIMS XV International Conference on Secondary Ion Mass Spectrometry*, *U.K.*) 2006; **252**:6575–6581.
29. Ma S, Huang J. Penalized feature selection and classification in bioinformatics. *Brief Bioinformatics* 2008; **9**(5):392–403.
30. Chu G, Narasimhan B, Tibshirani R, Tusher VG. SAM Version 1. 12: user's guide and technical document, 2001. Available from: http://www-stat.stanford.edu/ tibs/SAM/.
31. Hong D, Li HM, Li M, Shyr Y. Wavelets and projecting spectrum binning for proteomic data processing. In *Quantitative Medical Data Analysis Using Math Tools and Statistical Techniques*, Hong D, Shyr Y (eds). World Scientific Publications, LLC: Singapore, 2007; 159–178.
32. Mayevsky A. Mitochondrial function and energy metabolism in cancer cells: past overview and future perspectives. *Mitochondrion* 2009; **9**:165–179.
33. Matoba S, Kang JG, Patino WD, Wragg A, Boehm M, Gavrilova O, Hurley PJ, Bunz F, Hwang PM. P53 regulates mitochondrial respiration. *Science* 2006; **312**:1650–1653.
34. Rehman I, Azzouzi AR, Catto JWF, Ahmad M, Deloulme JC, Cross SS, Feeley K, Eaton CL, Meuth M, Hamdy FC. Calgizzarin (S100A11) immunostaining pattern is altered in prostate cancer suggesting a role in tumourigenesis. *European Urology Supplements* 2003; **2**:68–68.
35. Yoo CB, Jones PA. Epigenetic therapy of cancer: past, present and future. *Nature Reviews Drug Discovery* 2006; **5**:37–50.
36. Hadnagy A, Beaulieu R, Balicki D. Histone tail modifications and noncanonical functions of histones: perspectives in cancer epigenetics. *Molecular Cancer Therapeutics* 2008; **7**:740.

37. Tanaka M, Adzuma K, Iwami M, Yoshimoto K, Monden Y, Itakura M. Human calgizzarin; one colorectal cancer related gene selected by a large scale random cDNA sequencing and Northern blot analysis. *Cancer Letters* 1995; **89**:195–200.
38. Ohuchida K, Mizumoto K, Ogura Y, Ishikawa N, Nagai E, Yamaguchi K, Tanaka M. Quantitative assessment of telomerase activity and human telomerase reverse transcriptase messenger RNA levelsin pancreatic juice samples for the diagnosis of pancreatic cancer. *Clinical Cancer Research* 2006; **12**:5417–5422.
39. Hall P, Marron JS, Neeman A. Geometric representation of high dimension low sample size data. *Journal of the Royal Statistical Society*, *Series B* 2005; **67**:427–444.
40. Zhang HH, Ahn J, Lin X, Park C. Gene selection using support vector machines with non-convex penalty. *Bioinformatics* 2006; **22**:88–95.