Multivariate Analysis Methods for IMS Data

Biomarker Selection and Classification

A Thesis

Presented to the Faculty of the Department of Mathematical Sciences

Middle Tennessee State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Mathematical Sciences

---

by

Fengqing (Zoe) Zhang

May 2010

APPROVAL

This is to certify that the Graduate Committee of

Fengqing (Zoe) Zhang

met on the

1st day of March, 2010.

The committee read and examined her thesis, supervised her defense of it in an oral examination, and decided to recommend that her study should be submitted to the Graduate Council, in partial fulfillment of the requirements for the degree of Master of Science in Mathematics.

_____

Dr. Don Hong
Chair, Graduate Committee

_____

Dr. Curtis Church

_____

Dr. Lisa B Green

_____

Dr. Don Nelson
Chair,
Department of Mathematical Sciences

Signed on behalf of
the Graduate Council

_____

Dr. Michael Allen
Dean,
School of Graduate Studies

ABSTRACT

Imaging Mass Spectrometry (IMS) has shown great potential and is very promising in proteomics. However, challenges remain in data processing. The main task of this thesis is to find effective and efficient ways for IMS data biomarker selection and classification.

First, we incorporate a spatial penalty term into the elastic net (EN) model for IMS data processing. The EN-based model fully utilizes not only the spectrum information within individual pixels but also the spatial information for the whole IMS image cube. The real data analysis results show that the EN-based model works effectively and efficiently for IMS data processing.

Second, we propose a weighted elastic net (WEN) model combining ion intensity spreading information directly with the elastic net model. Properties including variable selection accuracy of the WEN model are discussed.

Finally, we develop a software package, called IMSmining, including visualization and analysis tools for IMS data processing. The analysis functions include EN-based model, WEN model and other current popular multivariate analysis methods in IMS community. The graphical interface of this software is user friendly.

DEDICATION

This thesis is dedicated to my parents, especially Yangui Zhang, my father, who has taught me, encouraged me and supported me in my life. Thanks for all your patience, love and unconditional support.

ACKNOWLEDGMENTS

A special thanks to my thesis advisor, Dr. Don Hong, who has been very helpful and patient with me through the entire process. I benefit a lot from his academic instruction, deep insights in mathematics and statistics, constant encouragement and valuable suggestions throughout the work on my thesis. Thank you Dr. Hong! To Dr. Anhua Lin, thank you for your helpful discussion and comments. I am also very grateful to the Vanderbilt Mass Spectrometry Research Center, especially Dr. Richard M Caprioli, Dr. Shannon Cornett, and Sara Frappier for valuable discussions and suggestions, as well as providing data sets for this research. And a word of thanks to the Graduate Coordinator Dr. Andrew Worsey and the rest of my committee who have graciously given their time to carefully read through my thesis. Your comments and suggestions have been very helpful.

Contents

List of Tables

List of Figures

1 Introduction

Proteomics is the study of, and the search for, information about proteins. It is much more difficult than genomics primarily due to the highly complex cellular proteomes and the low abundance of many of the proteins, and thus requires more sensitive analytical techniques. The development of 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), greatly speeds up proteomics research. Indeed, the 2002 Nobel Prize in Chemistry recognized MALDI's ability to analyze intact biological macromolecules. MALDI MS has increasingly been used to study therapeutic effects of drugs, disease progression and early cancer detection. Cancer research reported that if ovarian cancer can be detected earlier, more than 90 percent of women with the cancer can live five years or longer. It is essential to extract the information hidden in the noisy mass spectral data, involving operations such as biomarker selection and classification for application of MS techniques in cancer diagnosis. Biomarkers are biological features such as molecules that are indicators of physiologic state, and change during a disease process. Biomarker selection can be employed along with classification techniques to discriminate normal and cancer tissues more reliable and to provide deeper insights into the underlying causal relationships.

MALDI-Imaging is an emerging and very promising new technique for protein analysis from intact biological tissues [6]. It measures a large collection of mass spectra spreading out over an organic tissue section and retains the absolute spatial information of the measurements for analysis and imaging. The current interest in IMS lies in its unique advantage: the ability to correlate anatomical information

provided by histology with the spatially resolved biochemical information provided by the imaging mass spectrometry experiments. Compared with MALDI-MS, IMS, by automatic spotting of matrices on the tissue in an array format, results in comprehensive structural analysis at a higher spatial resolution, saves time, and provides hundreds of identical independent spectra which address the measurement repeatability. It has broad applications in the spatial distribution study of lipids [44], peptides [1], proteins [5] and small molecules with their metabolites [27] in tissue sections to study the biochemical changes related with several diseases, especially multiple forms of cancer ([9], [41], [7]). However, each MALDI imaging data set is multidimensional, with hundreds of pixels covering the tissue section and an entire mass spectrum in which mass-over-charge (m/z) values can range from 2k to 70k Dalton associated to each pixel. In this case, the number of predictors (m/z values) is much larger than the number of observations. 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 for the entire image cube. Another important issue is to distinguish the selected feature m/z values according to the differences caused by biological structure of the tissue or purely by cancer. All these difficulties, compounded together, pose great challenges to IMS data processing and are yet to be well solved.

MALDI imaging software packages such as BioMap as well as many other software tools for IMS do not provide multivariate analysis (MVA) methods for further data analysis. Usually, a common way to visualize IMS data is to generate two-dimensional ion intensity maps for known m/z values of interest [48]. However, a more important application should be the determination of unknown variants for metabolite and protein profiling in both clinical and disease studies. More mature analysis methods

are yet to be implemented in current commercial software, since IMS is a very recent and new technique. 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 in IMS data analysis [34]. The use of PCA in analyzing IMS data has been proposed to identify both spatial and mass trends that can merit further investigation [46]. Also LDA, Multivariate analysis of variance (MANOVA) and clustering methods have been used to analyze IMS data [37]. PCA and clustering are most commonly used for IMS data ([45], [10]). Plas et al.[47] proposed using peak intensity-based PCA to process IMS data. Correlation calculation for ion images, both in and between serial sections, was studied in [33]. PCA and Support Vector Machine (SVM) were also combined to process IMS data in [14]. However, these methods for IMS data processing have different sets of limitations (see section 2.3 for a discussion in detail).

In this thesis, we propose two statistical models for biomarker selection and classification of the high-dimensional and complex IMS data. Our aim is to extract as much useful information as possible from IMS data, by not only utilizing the spectrum information within individual pixels but also studying correlation and distribution using the spatial information. Compared with other currently popular methods, our models work efficiently and effectively for IMS data processing in terms of confirming new biomarkers, producing a more accurate listing of valid peaks by including significant peaks, reducing the number of side peaks, and providing more accurate classification results. A set of biomarkers was obtained with interesting biological explanations. In addition, a software package IMSmining has been developed, which provides data import, export, different ways of data visualization, and effective algorithms for IMS data analysis especially focusing on biomarker selection and classification.

Chapter 2 explains the charateristics of IMS data as well as the experiments and data sets we used for this study. It also states the problem of IMS data biomarker selection and classification, which is the main task of this thesis work. After briefly reviewing other current popular MVA methods for IMS data analysis, the difficulties and challenges of this problem are described.

A newly developed variable selection method, called elastic net (EN)[57], can simultaneously perform automatic variable selection and continuous shrinkage, as well as select groups of correlated variables. Compared to other current commonly used analysis methods, the EN model is much more suitable for IMS data processing. The EN model enjoys a sparsity of representation, which is particularly useful when the number of variables (p) for selecting as predictors is much larger than the number of observations (n), and also encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together.

In Chapter 3, we incorporate a spatial penalty term into the elastic net(EN) model in order to develop a new tool for IMS data biomarker selection and classification [52]. The motivation is to fully utilize not only the spectral information within individual pixels but also the spatial information for the whole IMS image cube. By incorporating the spatial penalty term, this model helps to distinguish the IMS feature peaks caused by biological structure differences from those truly associated with cancer disease. The EN-based model is compared with other current popular analysis methods commonly used in IMS community. The EN-based algorithm is applied to a real IMS data set for biomarker selection. The analysis results showed that this model works efficiently and effectively for IMS data processing in terms of confirming new biomarkers, producing a more precise peak list by including significant peaks and reducing the number of side peaks, and providing more accurate classification results.

A set of biomarkers was obtained with interesting biological explanations.

In the EN-based model, the spatial penalty consideration is applied in the cross validation step. A more general model called weighted elastic net (WEN), which incorporates the spatial penalty directly into the EN model equation, is discussed in Chapter 4, in order to better consider the spatial information for more precise biomarker selection [22]. Theoretical properties of the WEN model, such as variable selection accuracy, are discussed. The WEN algorithm is applied to IMS data sets for predictor selection. The analysis results showed that the WEN method works efficiently and effectively for IMS data processing. A set of biomarkers was obtained with interesting biological explanations.

Chapter 5 concerns the development of IMSmining software. Our motivation is to provide a convenient and automatic way to analyze IMS data. This package provides functions to import and export data, functions to visualize IMS data both in 2 and 3 dimensions ways, functions of EN-based model and WEN model for IMS data analysis, functions of other current popular algorithms in IMS area, serving as comparisons and also function to create the spatial intensity distribution graphics of certain given m/z values. The user interface is very friendly and gives flexible selections to users. For example, users can select the training data sets directly from the image by clicking the area they are interested in. The pseudo codes of EN4IMS algorithm and WEN algorithm are provided in Chapter 5 as well.

2 Biomarker Selection and Classification Problems for IMS Data

2.1 IMS Data and Experiments for This Study

Imaging mass spectrometry (IMS) is currently receiving a significant amount of attention in the mass spectrometric community. It offers the potential for direct examination of biomolecular patterns from cells and tissue. This makes it a seemingly ideal tool for biomedical diagnostics and molecular histology[19]. IMS data sets are image slides associated with mass spectra at pixels and thus, IMS data sets have very high dimensions. 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 indicates the tumor area. The IMS data set consists of an array of pixels covering the tissue section, and behind every pixel is an entire individual mass spectrum, which displays the ion intensities along the mass over charge (m/z) values, just as shown in Figure 1(c). IMS data can be viewed as a three-mode array with two spatial dimensions (x-, y- dimension) and the ion intensity values associated with m/z dimension (z- dimension) as shown in Figure 1(b). Figure 1(d) is the visualization of IMS data represented as a data cube. The x- and y- axes stand for two spatial dimensions and the z- axis represents the mass over charge dimension with the color indicates the ion intensity at m/z points. Five spatial intensity distribution graphs (slides) are shown in Figure 1(d) corresponding to five selected m/z values. Each graph gives an ion intensity distribution image with a false color visualization of the spatial distribution of peak height for a corresponding m/z value. The spatial information provided by IMS helps rapid mapping of protein localization and

the detection of sizeable differences in protein expression ([40], [49]). Conventional images, derived from a specific analyte mass, do not identify the spatially localized correlations between analytes that are latent in IMS data processing. Therefore, it is a great challenge in IMS data processing to use both spectral and spatial information for feature extraction.

The IMS data sets we used contain two serial sections from the mouse brain and are on the GL26 (cell line) glioma study. Figure 2 is an image of GL26 tumor implanted brains. It shows the overlay of the Hemotoxylin and Eosin stains on the actual optical image of the plate. C57 black mice were implanted with a GL26 glioma cell line and tumor growth was allowed to occur for 15 days. The mice brains were excised, flash-frozen, sectioned on a cryostat ($12\mu m$) and thaw-mounted onto gold-coated MALDI targets. Brain tissue was spotted with sinapinic acid for protein images on 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 $300\mu m$ by $300\mu m$. After data acquisition, the data underwent a series of basic preprocessing steps to reduce the experimental variance between spectra through the removal of background noise, normalization of the peak intensity to the total ion current, and peak binning/alignment algorithms 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) [8] to generate the SAM feature list in Table 1 of section 3.3.

The darker area in Figure 2 is the tumor area. The shape of the cancer and non-cancer area in Figure 2 can be compared with the spatial intensity distributions of selected m/z values in Figure 4 of section 3.3, in order to show the correctness of biomarker selection. The two brains, outlined in a rectangle in Figure 2, are the

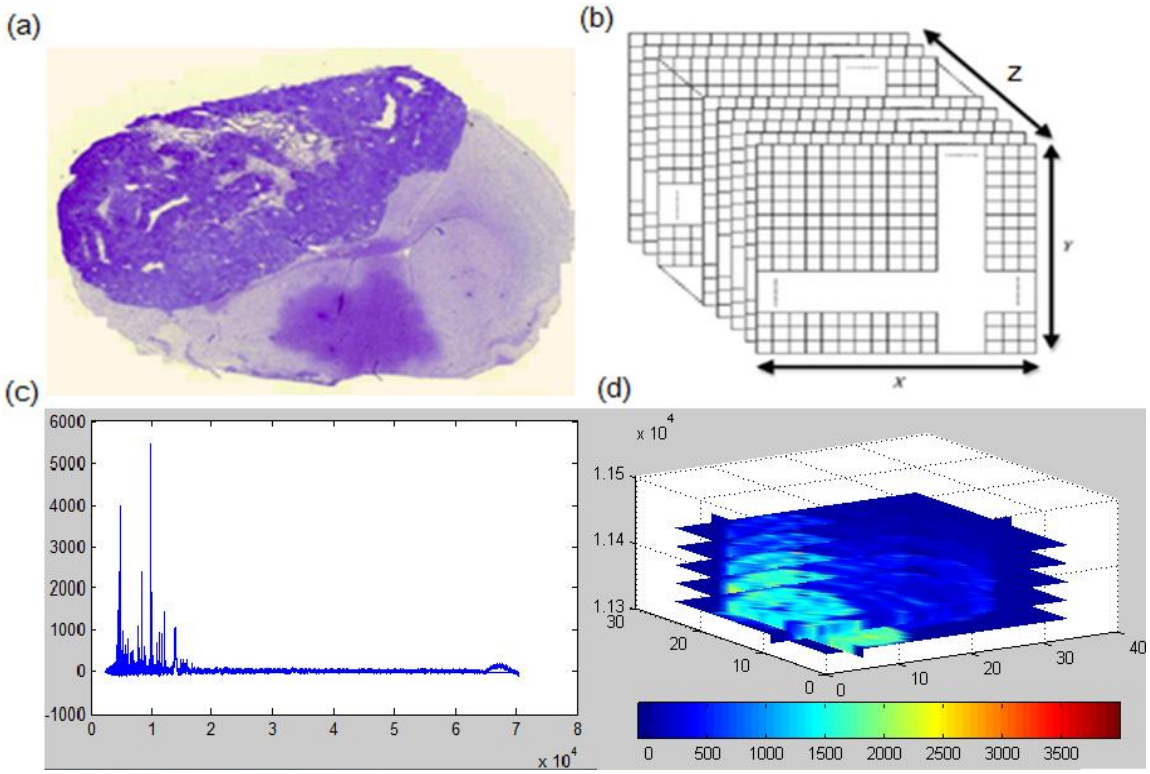

Figure 1: *Mouse brain IMS Data. (a) photomicrograph of a cresyl violet stained mouse brain section, implanted with a GL26 glioma cell line and tumor growth; (b)the three-mode array representation of IMS data set; (c)an entire individual mass spectru m behind one pixel; (d)the data cube visualization of IMS data set.*

sections 1 and 2 respectively. There are 635 pixels for section 1 and 695 pixels for section 2. For each pixel, the m/z values range from $2k$ to $70k$ Dalton with 22195 different m/z values. This is a typical case where $p \gg n$ and poses a great challenge for effective and efficient biomarker selection and classification.

2.2 Biomarker Selection and Classification

The goal in data processing is to effectively and correctly obtain the useful information from the IMS data for bimarkers discovery. Biomarkers are biological features such as molecules that are indicators of physiologic state, and change during a disease process. At the protein level, distinct changes occur during the transformation of a healthy cell into a neoplastic cell, including altered expression, differential protein modification, changes in specific activity, and aberrant localization, all of which may affect cellular function. Identifying and understanding these changes is the underlying theme in cancer proteomics.

Biomarker selection can be considered as a variable selection or feature selection problem in statistics. Variable or feature selection has become the focus of a great deal of application areas for which data sets with tens or hundreds of thousands of variables are available. These areas include for example, text processing of internet documents, gene expression array analysis, combinatorial chemistry, proteomics. The objective of variable selection is three-fold: provide faster and more cost-effective predictors, improve the model prediction performance, and provide a better understanding of the underlying process that generated the data.

Biomarker selection can be employed along with classification techniques to discriminate normal and cancer tissues more reliable and to provide deeper insights into the underlying causal relationships. Early tumor detection can result in lower stage

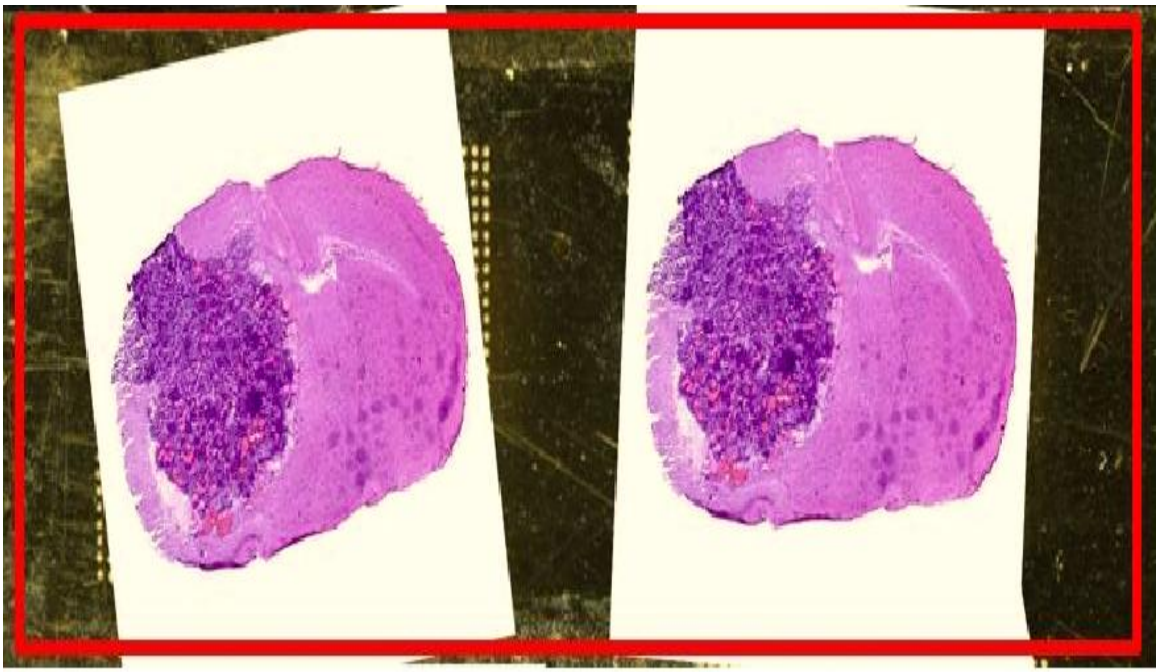

Figure 2: *Slide pictures of a mouse brain with tumor. The left one is section 1 for model training and the right one is section 2 for classification purposes. The shape of the cancer area and non-cancer area can be compared with the spatial intensity distributions of selected biomarkers.*

tumors, more treatable diseases and ultimately higher cure rates with less treatment-related morbidities. For example, if ovarian cancer can be detected earlier, more than 90 percent of women with that cancer can live five years or longer.

Therefore, the main task of this thesis is to find effective and efficient ways to do biomarker selection and classification for high-dimensional and complex IMS data, by making good use of both spectral and spatial information available.

2.3 Popular Statistical Methods and Tools for IMS Data Analysis

MALDI imaging software packages such as BioMap, as well as many other software tools for IMS, do not provide multivariate analysis (MVA) methods for further data analysis. Usually, a common way to visualize IMS data is to generate two-dimensional ion intensity maps for known m/z values of interest [48]. If one already has a known particular m/z value with special biological meaning and plans to find the spatial distribution of a particular molecule, an ion image is very informative. For example, [29] used MALDI-TOF-MS to analyze ovary cancer data and MALDI imaging to validate their biomarkers. However, a more important application should be the ability to determine of unknown variants for metabolite and protein profiling in both clinical and disease studies.

More mature analysis methods are yet to be implemented in current commercial software, since IMS is a very recent and new technique. 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 in IMS data analysis [34]. The use of PCA in IMS data has been proposed to identify

both spatial and mass trends that can merit further investigation [46]. Also LDA, Multivariate analysis of variance (MANOVA) and clustering methods have been used to analyze IMS data [37]. PCA and clustering are most commonly used for IMS data ([45], [10]). Plas et al.[47] proposed using peak intensity-based PCA to process IMS data. Correlation calculation for ion images, both in and between serial sections, was studied in [33]. PCA and Support Vector Machine (SVM) were also combined to process IMS data in [14]. However, these methods for IMS data processing have different sets of limitations.

In general, it is quite challenging to perform classification and biomarker selection accurately with IMS techniques since the m/z dimension is far greater than the sample size and extremely high dimensional. PCA is a general tool for dimension reduction in classifier construction. PCA aims to find an eigenvector $\mathbf{u}$ such that the projected data point $\mathbf{u}^T\mathbf{x}$ has the largest variance while $\|\mathbf{u}\|_2 = 1$. Thus PCA aims to find the eigenvector that maximizes $E[(\mathbf{u}^T\mathbf{x} - \mathbf{u}^T\mu)^2]$.

$$max_{\mathbf{u}} E[(\mathbf{u}^T(\mathbf{x} - \mu))^2] = max_{\mathbf{u}} \mathbf{u}^T E[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T]u = max_{\mathbf{u}} \mathbf{u}^T \mathbf{C}\mathbf{u}.$$

By reordering all discrete spatial positions $I \times J$ in the x- and y-directions, that is "pixels", into one long vector holding $I \cdot J$ elements for PCA on IMS data ([37], [46], [47]), a matrix $\mathbf{D}$ of size $I \cdot J \times k$ is formed, holding all information contained within the original matrix $\mathbf{D}$. It is preferable to apply PCA to all input variables. Although it has very high computation cost, 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 several drawbacks using PCA for analyzing IMS data. PCA is a commonly used dimension reduction technique, which constructs new input variables using linear combinations of all original input variables. However, 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 [31]. Furthermore, although PCA can be helpful in finding m/z values that represent the most significant variance in one section, the variance may be caused by structural difference instead of cancer disease. 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 [33].

K-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. It is similar to the expectation-maximization algorithm for mixtures of Gaussians in that they both attempt to find the centers of natural clusters in the data.

Given N observations $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$, where each $\mathbf{x}_i$ is a D-dimensional real vector, then K-means clustering aims to partition this data set into K clusters, μ_k, $k = 1, \cdots, K$. μ_k can be considered as the mean of all the data points assigned to cluster k. If a data point $\mathbf{x}_n$ is assigned to cluster k, then $r_{nk} = 1$ for $j \neq k$. The objective function is $J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \mu_k\|_2^2$.

One typically uses K-means combined with PCA for IMS data analysis [34]. The inputs of cluster methods are the principal components selected by PCA, since the m/z dimension is too high. Thus, inaccuracy of PCA will introduce inaccuracy in cluster results. This unsupervised classification could be further improved by supervised technique such as LDA and SVM.

Linear discriminant analysis (LDA) aims to maximize the ratio of between-class variance to the within-class variance. A discriminant function can be written as $g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + \mathbf{w}_0$, where $\mathbf{w}$ is the weight vector and $\mathbf{w}_0$ is the threshold weight. For

binary classification, LDA follows this decision rule: decide $\mathbf{w}_1$ if $g(\mathbf{x}) > 0$ and $\mathbf{w}_2$ if $g(\mathbf{x}) < 0$. Thus $\mathbf{x}$ is assigned to $\mathbf{w}_1$ if the inner product $\mathbf{w}^t\mathbf{x}$ exceeds the threshold $\mathbf{w}_0$ and to $\mathbf{w}_2$ otherwise. If $g(\mathbf{x}) = 0$, $\mathbf{x}$ can ordinarily be assigned to either class.

LDA was combined with PCA for IMS data analysis ([34],[37]). 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. Thus, the authors of [34] and [37] proposed to use PCA as a dimension reduction first and thus, the inaccuracy of PCA will result in inaccuracy in clustering and classification. In addition, LDA implicitly assumes that the mean is the discriminating factor (not variance) and the data are normally distributed. Such assumptions may be violated, which limits the application of LDA.

Linear SVM combined with PCA was also used for IMS data analysis in [14]. Under linear SVM binary classification, a labeled set of features $\{\mathbf{x}_i, \mathbf{y}_i\}$ is constructed for all k features in the training data set. The binary output is recorded as $\mathbf{y}_i = \{-1, 1\}$. SVM aims to construct a separating hyper-plane which maximizes the margin between the two data sets. The margin is defined as the shortest distance from the separating hyper-plane to the closest data point. Intuitively the larger the margin is, the lower the generalization error of the classifier will be. SVM can be formulated as the following optimization problem in a dual form

Maximize

$$\sum_{i=1}^{n} \alpha_i - 1/2 \sum_{i,j} \alpha_i \alpha_j c_i c_j \mathbf{X}_i^T \mathbf{X}_j$$

Subject to (for any $i = 1, \cdots, n$) $\alpha_i \geq 0$ and $\sum_{i=1}^{n} \alpha_i c_i = 0$

The above functional can be easily extended to the nonlinear case by using a nonlinear kernel $k(\mathbf{X}, \mathbf{X}^T)$ instead of the inner product $\mathbf{X}_i^T \mathbf{X}_j$. To use linear SVM, one usually uses PCA for dimension reduction. Thus, SVM suffers from PCA inaccuracy

similar to the methods discussed each. In addition, SVM is for classification but not for 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 [57]. Consequently, this method is not effective for biomarker selection.

It is reasonable and more effective to use the learning process for biomarker selection, and to fully utilize the spatial information provided by IMS data. Cancer disease may affect some functional proteins, and thus peptides related to these proteins should be in or out of the model together. Therefore, it is desirable to take this grouping effect into consideration. After combining with the spatial penalty term, our EN-based model and WEN model not only inherit good properties from elastic net, but also can effectively select cancer related features instead of those related with structure difference, and thus theoretically outperform these algorithms discussed above. Experimental results also show the advantages of our models compared with the methods discussed each, which should be referred to Section 3.3 and Section 4.4.

2.4 Challenges and Difficulties

The complexity and high dimensionality of IMS data pose great challenges and difficulties for information extraction and data analysis. One experiment has several mouse brain samples. For one mouse brain sample, there are several data section series. And for each section, there are 35×24 pixels for the relatively low resolution data sets, and there are 65×44 pixels for the higher resolution data sets. As shown in Figure 1, behind each pixel there is an entire spectrum with a large m/z range from $2k$ to $70k$ Dalton. This large Dalton range includes 22195 different m/z values for our data. All these together form a really huge data set and, this requires researchers to

pay special attention to dimension reduction, computational cost of algorithms and variable selection. Furthermore, the number of predictors (m/z values) is much larger than the number of observations. This itself is a great challenge for data processing, but is further compounded by the low signal intensities found across the image for feature selection. 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 for the entire image cube. It is also important to distinguish the selected feature m/z values according to the differences caused by biological structure 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 great challenge in developing of new methods and tools for quantitative analysis and biological interpretation of the IMS data.

In this thesis work, we propose the EN-based model and Weighted Elastic Net model for effective and efficient biomarker selection and classification by utilizing both spectrum and spatial information of IMS data. Additionally, we have developed a software package, IMSmining, for IMS data visualization and quantitative analysis with a user friendly interface.

3 EN-Based Model for IMS Data Analysis

3.1 Motivation and Background

3.1.1 The Elastic Net

The usual linear regression model can be described as follows: Assuming p predictors $\mathbf{X}_1, \cdots, \mathbf{X}_p$, 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 gives the vector of coefficients $\hat{\beta} = (\hat{\beta}_0, \cdots, \hat{\beta}_p)^T$. Ordinary least squares (OLS) estimates are obtained by minimizing the residual sum of squares (RSS). Ridge Regression minimizes RSS subject to a bound on the ℓ_2 norm of the coefficients

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

where the ℓ_2 norm penalty term $\|\beta\|_2^2 = \sum_j^p \beta_j^2$ is also called the ridge penalty term. The Lasso minimizes RSS subject to a bound on the ℓ_1 norm of the coefficients

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

where the ℓ_1 norm penalty term $|\beta|_1 = \sum_j^p |\beta_j|$ is usually called the lasso penalty term.

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. This means that the lasso fails to conduct grouped selection. That is, it tends to select one variable from a group and ignores the others. However elastic net [57], a convex combination of the lasso and ridge penalty term, usually

outperforms them in many situations. The EN method is particularly useful when $p \gg n$. The group effect is like a stretchable fishing net that retains "all the big fish" [57].

The naive elastic net criterion is to minimize the function:

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

The ℓ_1 part of the penalty term generates a sparse model, while the quadratic part removes the limitation on the number of selected variables, encourages the grouping effect, and stabilizes the ℓ_1 regularization path. The non-negative tuning parameters λ_1 and λ_2 balance the goodness-of-fit and complexity of the model. By using the scaling transformation β(elastic net) $= (1+\lambda_2)\beta$(naive elastic net), the double shrinkage deficiency of the naive elastic net can be solved. Finally the elastic net estimates β can be given as

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

3.1.2 Computation: the LARS-EN Algorithm

Least angle regression (LARS)([11]),which LARS-EN derives from, can be viewed as a version of the stagewise method that uses mathematical formulas to reduce the computation cost. The general idea of LARS is as follows:

LARS starts with all coefficients equal to zero, and find the predictor most correlated with the response, say $\mathbf{x}_{j_1}$. It takes the largest step possible in the direction of this predictor until some other predictor, say $\mathbf{x}_{j_2}$, has as much correlation with the current residual. At this point, LARS proceeds in a direction equiangular between the two predictors until a third variable $\mathbf{x}_{j_3}$ earns its way into the "most correlated"

set. LARS then proceeds equiangularly between $\mathbf{x}_{j_1}$, $\mathbf{x}_{j_2}$ and $\mathbf{x}_{j_3}$, that is, along the "least angle direction", until a fourth variable enters, and so on.

Efron et al.[11] proved that, starting from zero, the lasso solution paths grow piecewise linearly in a predictable manner. The new algorithm 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 elastic net problem is equivalent to a lasso problem on the augmented data set. Thus an efficient algorithm LARS-EN was proposed to solve the elastic net, much like the LARS algorithm does for the lasso. Details can be found in [57].

3.2 EN-Based Model for IMS Data

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 [52]. The motivation is to fully utilize not only the spectral information within individual pixels but also the spatial information for the whole IMS data cube.

3.2.1 Spatial Penalty Consideration

The importance of fully utilizing spatial information provided by IMS techniques has been emphasized in recent literature (see [34], [33], [37] for example). Thus, it is very desirable to develop an algorithm of biomarker selection and classification by combining the spectral information within individual pixels with the spatial information for entire images. Another important issue is to distinguish the selected m/z values according to the differences caused by biological structure of the tissue or by disease. The true cancer related biomarkers can effectively describe the cancer and can be

used for cancer diagnosis. Those feature peaks in terms of m/z values in IMS show differences in cancer and non-cancer areas but depend on tissue structure and thus are not the true features for cancer study. Therefore, it is critical to find only cancer related biomarkers independent of tissue structure in IMS data processing with consideration of spatial information.

In IMS data analysis, if a feature m/z value in the MS spectrum is truly related to a cancer disease, then it is reasonable to expect that the ion intensity values at this m/z from different pixel locations in a cancer area are approximately the same. Therefore, the standard deviation of the intensities at the m/z should be small. In comparison, if the feature peak selected by the statistical model based on differentiation mainly caused by the tissue structure, then the ion intensities at the m/z point vary significantly from pixel to pixel. Therefore, the standard deviation of intensities at such an m/z point should be relatively large. Thus, it is proper to associate standard deviations at all selected predictors to the optimal model selection in order to enforce penalty on predictors caused by structure differences. Hence, in our work we incorporate such a spatial penalty into the EN model to develop EN4IMS algorithm for IMS data analysis.

3.2.2 EN-based Model for IMS

Based on the biological considerations discussed in Section 3.2.1, we incorporate the spatial standard deviation for the spatial penalty consideration into the EN model cross validation (CV) step. More precisely, the algorithm chooses the optimal model by determining the tuning parameter λ_1 in order to minimize the functional (3) in

the ten-fold cross validation:

$$\|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2^2 + \frac{\lambda_3}{M}\sum_{j=1}^{M}\sqrt{\frac{\sum_{i=1}^{N}(x_{ij} - \mu_j)^2}{N-1}}, \tag{3}$$

where N is the number of all cancer pixels, M represents the number of all m/z values selected by EN-based model. x_{ij} is the intensity for a fixed m/z value j in the pixel i, and μ_j is the mean intensity over all these cancer pixels for a fixed m/z value j. This penalty term is incorporated into the cross validation step.

The ten-fold CV method divides the IMS data into ten equal batches and estimates parameters ten times, leaving one batch out each time. The testing error for each omitted batch is computed using the estimates derived from the remaining batches. The algorithm runs some number large enough as the total steps in order to ensure sufficient accuracy. The sum squared residuals and the sum of spatial standard deviation of selected ion intensities are measured at every step. Finally, the optimal model position is chosen as the one that can minimize both the residual error and spatial standard deviation of selected ion intensities.

The tuning parameter λ_3 controls a weight on the spatial penalty term for balancing the contribution of the residual error and spatial penalty term for the IMS data. Due to the complexity of the determination of λ_3 together with other parameters λ_1 and λ_2, in this algorithm, we select λ_3 based on experimental experience of the IMS data study. For those peaks or m/z values with a smaller standard deviation, the penalty effect on them will be smaller than the ones with a larger standard deviation. For the purpose of taking advantage of the LARS algorithm, we first fix λ_3 and then use two dimensional cross validation to select the parameters of λ_1 and λ_2.

The parameter λ_1 can be associated with k, the number of steps in cross validation algorithm. In the LARS algorithm, the lasso is described as a forward stagewise

additive fitting procedure and shown to be (almost) identical to $\varepsilon - L_2$ boosting [11]. From this point, we see that the number of steps k of algorithm LARS can be considered as a tuning parameter λ_1 for the lasso. Therefore, two tuning parameters in the EN model are k and λ_2. In the two dimensional cross validation, typically the tuning parameter λ_2 is picked as a relatively small grid, say $(0, 0.01, 0.1, 1, 10, 100)$. We first fix λ_2. Algorithm EN4IMS produces all possible estimates of the vector β for the IMS data. However, we just want a single optimal $\hat{\beta}$; thus, some rules for selection are needed. A modified cross validation step is used to choose the other tuning parameter ($\lambda_1 or\ k$). The optimal step k is chosen as the one that can minimize both the residual error and spatial standard deviation of selected ion intensities. Then comparing the error of each optimal model at each fixed λ_2, we choose the λ_2 as the one that can minimize the error.

The EN-based model proposed here is for pixel-level classification. When entering the data, cancer pixels and non-cancer pixels are selected from the mouse brain IMS data sets to be as symmetric as possible with the consideration of structure 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 allowed to use them without dimension reduction. The early stopping feature of the LARS-EN algorithm saves computation cost and time [57]. In the case where $p \gg n$, if the algorithm ends in m steps, then it only requires $O(m^3 + pm^2)$ operations. We include all m/z values as predictors in (1), and y takes negative one for a non-cancer pixel and one otherwise.

By incorporating the spatial penalty term, true biomarkers are able to be distinguished from features selected through structure difference, which is an urgent need in current IMS data processing. It also considers the group effect of these m/z val-

ues, where strongly correlated predictors tend to be in or out of the model together. Furthermore, the optimal model provides us with selected variables (m/z list) serving as potential biomarkers and can do 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 cross validation step. A more general model that includes the spatial penalty directly into the EN model equation (2) is discussed in Chapter 4, to better consider the spatial information for more precise biomarker selection.

3.3 Evaluation of the EN-Based Model

In this section, we apply the EN-based model to a set of mouse brain IMS data generated from the Vanderbilt Mass Spectrometry Research Center (VUMSRC). Details of the data sets can be found in Section 2.1. The analysis includes a comparison of results obtained by applying the EN method, PCA, LDA, SVM and the EN-based model to the IMS data set, as well as the results obtained by using the commercial software SAM. From our results, the EN4IMS algorithm produces a more concise listing of peaks in the sense of including all significant features but a smaller number of side peaks. Our algorithm confirms a new biomarker which is not included in the SAM list and also provides better classification results, compared with other current popular analysis methods in IMS community.

Table 1 shows peak lists based on the mice brain IMS data sets by running the SAM, PCA, EN, and EN4IMS algorithms. 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) [8] to generate the SAM feature list. In comparison, EN4IMS processes IMS data using only basic prepro-

cessing steps with no peak binning beforehand and saves significant amount of time for data processing. For the $GL26$ IMS data set, the EN4IMS algorithm generates a list of m/z values which matches significant features obtained by using SAM and provides a much more condensed list by removing side peaks.

Comparing the m/z list generated by the regular EN algorithm, the newly developed EN based algorithm that incorporates the spatial penalty term produces an even more concise list by including all significant features and has a smaller number of side peaks. Side peaks are fake peaks caused by noise, and are not main peaks with biological meaning. For instance, the m/z list, obtained by using the EN4IMS algorithm, does not include the side peak ($m/z = 6794$) which is misidentified by the EN method. Figure 3 shows the partial length of the cancer mean spectrum and non-cancer mean spectrum, from which it is clear to see that peak ($m/z = 6794$) is a side peak from a biological point of view. It is time consuming to reduce side peaks by checking the list manually. In addition, EN4IMS identifies tumor signal peak ($m/z = 14788$) which is not on the SAM list. Therefore, EN4IMS helps in confirming new biomarkers discovered by biological experiments. The biological interpretation of this peak is explained below.

As we discussed in section 2.3, 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 [33]. EN4IMS found more important biomarkers compared to the PCA method ([37], [46], [47]). The peaks (m/z: 6700, 8380, 10952, 14788) described below are on the EN4IMS list but not on the PCA list in Table 1.

Noticing that MS peaks need to run a so-called binning algorithm for cross samples alignment in MS data processing [21], peaks within 5 Dalton shift of an m/z value

($ge5000$), are usually considered to be the same peaks. Therefore, these four peaks correspond to the values (m/z: 6702, 8384, 10949, 14786) in the EN4IMS list. This fact has also been confirmed by biochemists at VUMSRC by checking them manually. The corresponding spatial intensity distributions of these four peaks are shown in Figure 4. They have been proved to be important biomarkers in this cancer research.

In fact, protein identification 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, O2 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],[30]). The Warburg effect is described as the dependency of tumors on glycolysis rather than 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 identified signals include calgizzarin ($m/z = 10952$) and an acetylated form of Histone $H2A$ ($m/z = 14788$). 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, differentiation and accelerated metabolism in cancer cells, although its detailed function is not yet known ([39], [42], [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 ([50], [16]). Histone modifications promote or

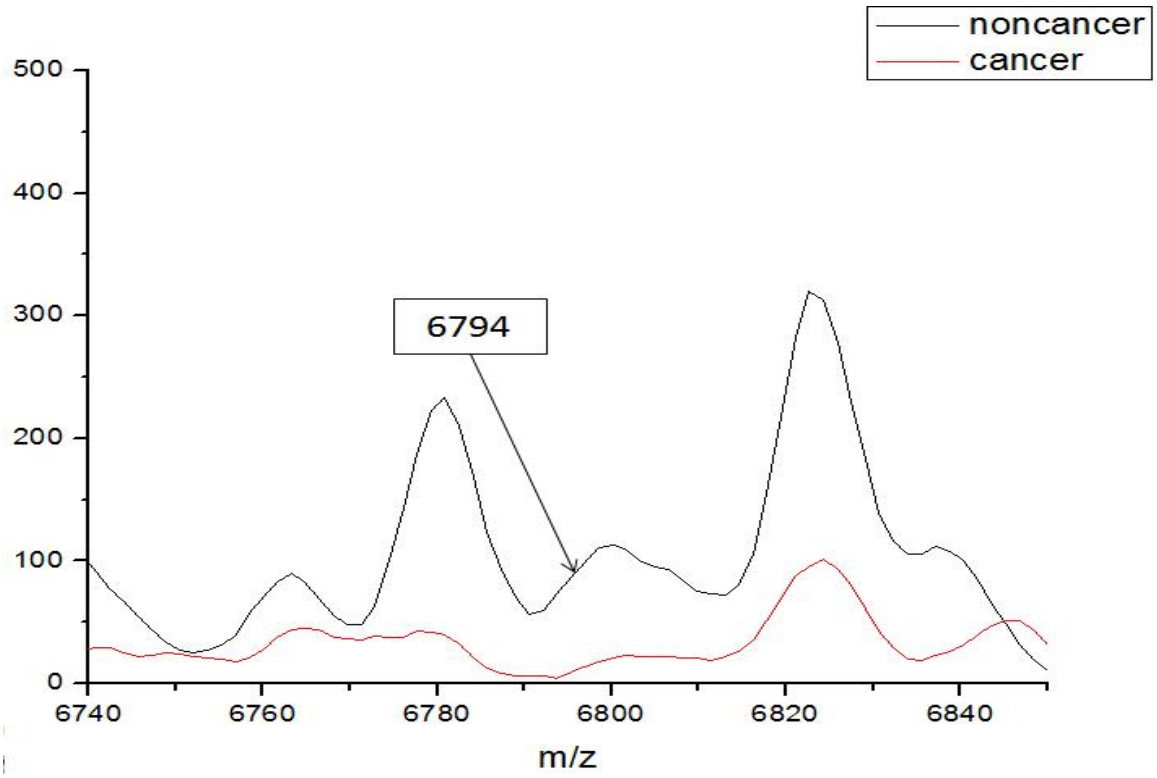

Figure 3: *Side peak(m/z=6794). The peak at m/z=6794 is a fake peak caused by noise, and is not a main peak with biological meaning.*

Table 1: Listings of selected peaks in terms of m/z values using the SAM, PCA, EN and EN4IMS algorithms

SAM list	PCA list	EN list	EN4IMS list
2791 3434 8337	4934 8567	4476 13562	4664
3010 3764 8366	4936 10257	4664 14327	4667
3056 4011 8380	4937 10259	4670 14336	4670
3734 4076 8395	4938 10261	4812 14343	4812
3800 4271 8492	4939 10263	4884 14781	5446
3920 4538 8672	4960 14969	5425 14786	5753
4206 4566 8945	4962 14971	5429 14805	5754
4341 4665 8982	4963 14974	5446	5756
4605 4676 9327	4964 14976	5753	5757
4734 4899 9343	4966 14979	5754	6165
4767 5106 9531	5439 14981	5756	6702
4921 5120 9602	5441 14983	6165	6706
4936 5428 9619	5442 14986	6702	7799
4964 5444 10238	5444 15603	6706	8019
4981 5707 10267	5445 15606	6794	8024
5001 5753 10466	5446 15608	7799	8384
5024 6166 10662	5448 15611	8019	8386
5170 6186 12434	5449 15613	8024	9344
6225 6251 13560	5451 15616	8028	10172
7706 6310 14525	6571 15618	8384	10261
8420 6574	6572 15620	8386	10263
8603 6700	6574 15623	8495	10265
8709 6719	6575 15625	8524	10267
8747 6780	6577 16780	9344	10282
9062 7099	7749 16782	9553	10366
9736 7118	7751 16785	10172	10374
9956 7297	7752 16787	10261	10825
10167 7315	7792	10263	10949
10952 7338	7794	10267	13562
11388 7357	7795	10282	14336
11640 7751	7797	10366	14343
12203 7776	8560	10374	14781
14865 7795	8562	10811	14786
14927 8025	8564	10825	14805
14978 8107	8566	10949	

prevent 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 increase and decrease trend of selected m/z list, 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 m/z values with a decreasing trend in the tumor area.

Interestingly, when $p \gg n$, linear classifiers often perform better than non-linear ones in many applications [18], even though non-linear methods are known to be more flexible. This fact is related to the asymptotic results[17]: when $p \gg n$, under mild assumptions for data distribution, the pairwise distances between any two points are approximately identical to each other so the data points form an n-simplex. Linear classifiers then become natural choices to discriminate two simplices [53].

In Table 2, we compare the results of our algorithm with other current popular methods in the IMS data processing area, including PCA+LDA ([34], [37]) and PCA+SVM [14]. These algorithms are applied to IMS data section one to learn the optimal model and then are used to classify IMS data section two as shown in Figure 2. The EN4IMS algorithm shows the best classification results and also has an internal variable selection facility.

Figure 4 shows images of spatial intensity distribution of four selected biomarkers ($m/z:\ 6702, 8384, 10949, 14786$) described above. It is clear to see the distribution differences between cancer and non-cancer area and the shape similarity between

Table 2: Comparison of EN4IMS classification result with the results by using other current popular methods in the IMS data processing area.

Methods	Accuracy	Sensitivity	Specificity
PCA+LDA	78.64%	100%	57.27%
PCA+SVM	71.82%	84.56%	59.09%
EN4IMS	99.09%	100%	98.18%

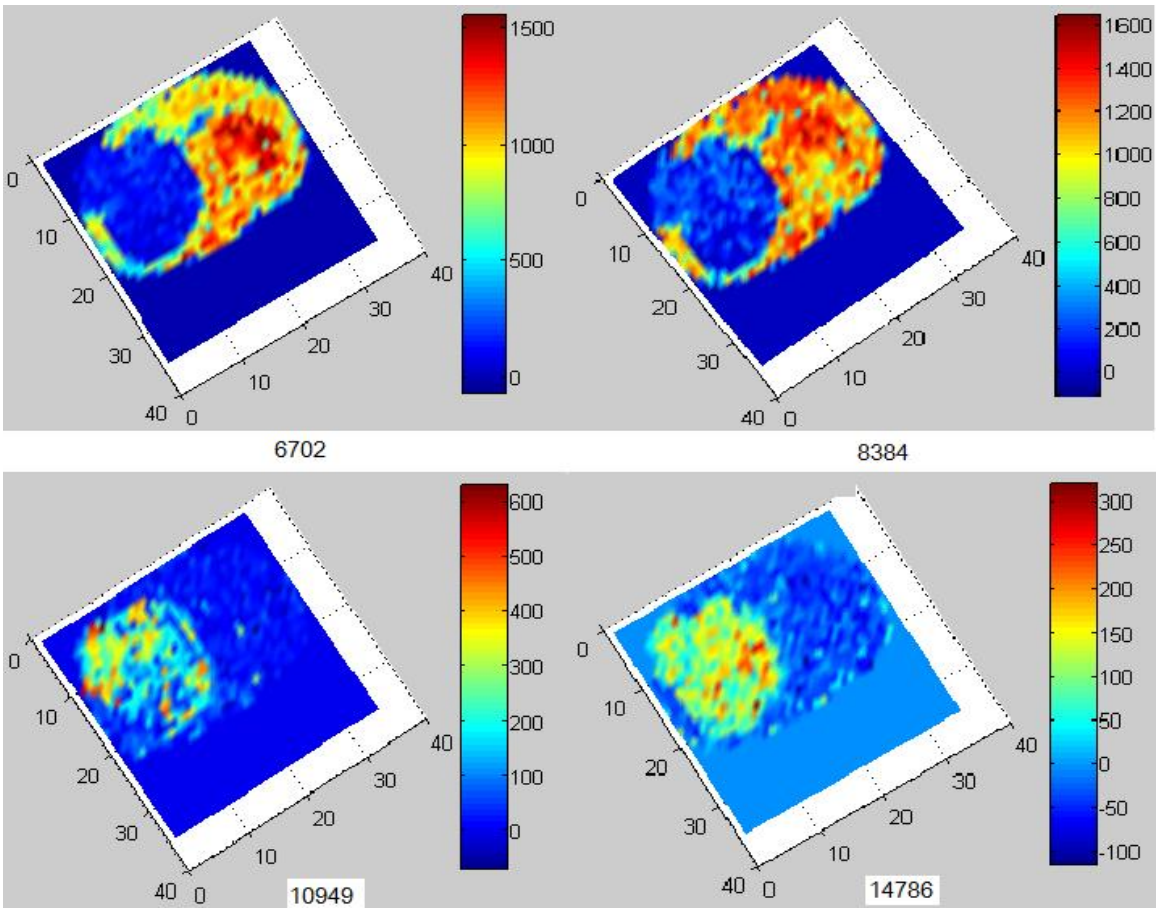

Figure 4: *Spatial intensity distribution graphs of 4 selected important biomarkers. The x- and y- dimensions are the spatial dimensions which correspond to Figure 2. The intensity of the selected m/z value is represented by the color.*

these images and the original tissue section picture shown in Figure 2.

3.4 Summary

In this chapter, we incorporated a spatial penalty term into the EN model, which is a very recently developed regularization and variable selection method, and developed a new tool for IMS data biomarker selection and classification [52]. The motivation is to fully utilize not only the spectrum information within individual pixels but also the spatial information for the whole IMS image cube.

The EN-based model inherits good properties from EN which produces a sparse model with admirable prediction accuracy. By incorporating the spatial penalty term, the EN-based model helps to distinguish the IMS feature peaks caused by biological structure differences from those truly associated with diseases. The EN-based model was applied to real IMS data sets and compared with other current popular analysis methods commonly used in IMS community. The results show that the EN-based model works effectively and efficiently for IMS data processing in terms of confirming new biomarkers, producing a more precise peak list by including significant peaks and reducing the number of side peaks, and providing more accurate classification results.

4 Weighted Elastic Net Model for IMS Data Analysis

4.1 Background and Needs for Model Design

Two fundamental criteria for evaluating the quality of a model in statistical modeling are high prediction accuracy and discovering relevant predictive variables. In the practice of statistical modeling, variable selection is especially important; it is often desirable to have an accurate predictive model with a sparse representation since modern data sets are usually high dimensional with a large number of predictors. One would like to have a simple model to enlighten the relationship between the response and covariates and also to predict future data as accurate as possible.

Let us consider a multiple linear regression model with n observations. Suppose that $\mathbf{x}_j = (x_{1j}, \cdots, x_{nj})^T$, $j = 1, \cdots, p$ are the linear independent predictors and $\mathbf{y} = (y_1, \cdots, y_n)^T$ is the response vector. $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_p]$ denotes the predictor matrix. If the data are centered, then the linear regression model can be expressed as

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \tag{1}$$

where $\beta = (\beta_1, \cdots, \beta_p)^T$ and the noise term $\epsilon \sim N(0, \sigma^2 \mathbf{I}_n)$. A model fitting procedure produces the vector of coefficients $\beta = (\beta_1, \ldots, \beta_p)^T$.

Ordinary least squares (OLS) estimates are obtained by minimizing the residual sum of squares (RSS). It is well known that OLS does poorly in both prediction and variable selection. Penalized methods have been proposed to improve OLS, starting with Ridge regression [20], followed by Bridge regression [13], the Garotte [3], the Lasso [43], LARS [11], and very recently the elastic net [57]. The Dantzig selector

method was proposed in [4] by using sparse approximation and compressive sensing. It was designed for linear regression models where p is large but the vector of coefficients is sparse, its ℓ_1-minimization produces coefficient estimates that are exactly 0 in a similar fashion to the Lasso [25] and hence can be used as a variable selection tool.

Penalization methods achieve feature selection and classifier construction simultaneously by computing $\hat{\beta}$, estimate of β that minimizes a penalized objective function. By properly tuned penalties, estimated β can have components exactly equal to zero and thus achieve the sparsity needed. Therefore, feature selection is achieved in the sense that only variables with nonzero coefficients will be used in the classification model. Specifically, here we define $\hat{\beta}$ as

$$\hat{\beta} = \text{argmin}_{\beta}\{\frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\, \text{pen}(\beta)\} \tag{2}$$

Where λ is a tuning parameter. The penalty $\text{pen}(\beta)$ in (2) controls the complexity of the model. Here $\text{pen}(\beta)$ could be the ridge penalty, Lasso penalty, elastic net penalty and any other appropriate penalty function. The tuning parameter $\lambda > 0$ balances the goodness-of-fit and complexity of the model. As $\lambda \to 0$, the model has better goodness-of-fit. However, this may cause classifiers to be too complex with unsatisfactory prediction and thus less interpretable. As $\lambda \to \infty$, the classifier is the simplest one with no input variable used for classification [31]. With the proper tuning parameter λ, the classifier can have satisfactory prediction accuracy and is interpretable. When only training data are available, tenfold cross validation (CV) is a popular method to estimate the tuning parameter λ, the prediction error and comparing different models ([18], chapter 7). Work is still needed to investigate and compare model selection methods including C_p, Akaike information criterion (AIC), Bayesian Information Criterion (BIC), CV and empirical Bayes.

For the linear regression model (1), one would like to recover the sparse parameter $\beta \in \mathbb{R}^p$. Assume $S = \text{supp}(\beta^*) = \{j : \beta_j^* \neq 0\}$, the support set of β^*, and let $s = |S|$. The set S sometimes is called the active index set. We also denote $S^c = \{1, \cdots, p\} \setminus S$ and correspondingly the vectors (matrices) β_S and β_{S^c} ($\mathbf{X}_S$ and $\mathbf{X}_{S^c}$) defined on S and S^c, respectively.

The importance of the oracle property of the learning model is emphasized in [12]. This ensures the model has good statistical properties, that the model can correctly select the nonzero coefficients with probability converging to one and that the estimators of the nonzero coefficients are asymptotically normal with the same mean and covariance that they would have if the zero coefficients were known in advance. We call the estimating procedure δ an oracle procedure if $\hat{\beta}(\delta)$ (asymptotically) has the following oracle properties:

(1) Identifies the right subset model, $\{j : \hat{\beta}_j \neq 0\} = S$,

(2) Has the optimal estimate rate, $\sqrt{n}(\hat{\beta}(\delta)_S - \beta_S^*) \rightarrow_d \mathbb{N}(0, \mathbf{C})$, where $\mathbf{C}$ is the covariance matrix knowing the true subset model.

Usually, we call property-(1) consistency in variable selection and property-(2) asymptotic normality.

The ridge penalty is defined as

$$\text{pen}(\beta) = \sum_{j=1}^{p} \beta_j^2. \tag{3}$$

Ridge Regression minimizes RSS subject to a bound on the ℓ_2 norm of the coefficients. It projects $\mathbf{y}$ onto the singular values of $\mathbf{X}$ and then shrinks the coefficients of the low-variance components more than the high-variance components. Although it is continuous shrinkage, ridge regression always keeps all the predictors in the model and thus does not have sparsity presentation for input data. Subset selection in

contrast produces a sparse model, but it is a discrete process - variables are either retained or discarded. Thus, it often exhibits high variance and does not reduce the prediction error of the full model [18].

Lasso is a regularization technique for simultaneous estimation and variable selection [43]. The Lasso penalty is defined as

$$\text{pen}(\beta) = \sum_{j=1}^{P} |\beta_j|. \tag{4}$$

Lasso minimizes RSS subject to a bound on the ℓ_1 norm of the coefficients. Due to the nature of the ℓ_1 penalty, Lasso does both continuous shrinkage and automatic variable selection simultaneously. Generally, Lasso is not variable selection consistent in the sense that the whole Lasso path may not contain the true model. Recent research results ([54], [56], [35], [51]) have been focused on the model selection consistency of the Lasso. The condition for the Lasso's model selection consistency by using a so-called the Irrepresentable Condition (IC) was studied in [54] for the classical case when p and s (the number of nonzero coefficients associated with the predictors in the model) are fixed.

For a given estimator $\hat{\beta}$, one would like to have $\text{supp}(\hat{\beta}) = \text{supp}(\beta^*)$ with high probability. More precisely, we want

$$\text{sgn}(\hat{\beta}) = \text{sgn}(\beta^*), \text{ with high probability.}$$

This question was recently considered in [55] for the adaptive Lasso model. In the following, we would like to address this problem on the weighted elastic net model.

We assume that

$$\frac{1}{n}\mathbf{X}^T\mathbf{X} \to \mathbf{C}, \tag{5}$$

where $\mathbf{C}$ is a positive definite matrix. Without loss of generality, let

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{1\,1} & \mathbf{C}_{1\,2} \\ \mathbf{C}_{2\,1} & \mathbf{C}_{2\,2} \end{bmatrix}, \tag{6}$$

where $\mathbf{C}_{1\,1} = \mathbf{X}_S^T \mathbf{X}_S$ is an $s \times s$ matrix corresponding to the covariance matrix on the active index set S.

A so-called irrepresentable condition (IC) states that there exists a positive constant $\eta > 0$ such that

$$\|\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathrm{sgn}(\beta_S)\|_\infty \leq 1 - \eta \tag{7}$$

where the inequality holds element-wise. IC is necessary and sufficient for the Lasso's model selection consistency [54].

To improve the Lasso model, the adaptive Lasso was proposed in [56] by using a weighted ℓ_1 penalty.

$$\mathrm{pen}(\beta) = \sum_{j=1}^{p} \hat{\omega}_j |\beta_j|, \tag{8}$$

where $\hat{\omega}_j = 1/|\hat{\beta}_j|^\gamma$ for an initial estimator $\hat{\beta}$ and a power $\gamma > 0$. By adding such weights to the coefficients, adaptive Lasso enjoys the oracle properties for linear models with $n \gg p$. For the case where $p \gg n$, Lasso can still be variable selection consistent under certain orthogonality conditions [24]. More general situations for Lasso based models to be consistent were recently studied in [55].

If the number of predictors, p, is greater than the sample size, n, Lasso selects at most n variables. Therefore, the number of selected features is bounded by the number of samples. Furthermore, Lasso fails to conduct grouped selection. That is, it tends to select one variable from a group and ignores the others. However elastic net [57], a convex combination of the lasso and ridge penalty, usually outperforms

them in many situations. The EN penalty term with coefficients is defined as

$$\text{pen}(\beta) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2. \tag{9}$$

The EN method is particularly useful when $p \gg n$. The group effect is like a stretchable fishing net that retains "all the big fish" [57]. In high-dimensional data analysis, the number of variables can greatly exceed the number of observations, and strong correlations often exist among subsets of variables. This is the case for IMS data and thus we choose to develop a statistical model based on the elastic net for IMS data processing.

A necessary and sufficient condition for the elastic net to be variable selection consistent in the classical settings when p and s are fixed is given in [51]. Corresponding to the IC condition, the Elastic Irrepresentable Condition (EIC) is defined as

EIC: There exists λ_1, λ_2 and a positive constant $\eta > 0$ such that

$$\|\mathbf{C}_{21}(\mathbf{C}_{11} + \frac{\lambda_2}{n}\mathbf{I})^{-1}(\text{sgn}(\beta_S) + \frac{2\lambda_2}{\lambda_1}\beta_S)\|_\infty \leq 1 - \eta. \tag{10}$$

EIC is necessary and sufficient for the EN model selection consistency [51]. The model selection consistency of EN model for $p \gg n$ case and the relationship of IC and EIC was discussed in [26]. IC implies EIC, but EIC does not imply IC. In order to achieve the oracle property, the following adaptive elastic net by combining adaptive ℓ_1 penalty and ridge penalty was proposed [56].

$$\hat{\beta} = (1 + \frac{\lambda_2}{n})\{\text{argmin}_\beta \|\mathbf{y} - X\beta\|_2^2 + \lambda_2\|\beta\|_2^2 + \lambda_1^* \Sigma_{j=1}^p \hat{\omega}_j |\beta_j|\} \tag{11}$$

where $\hat{\omega}_j = 1/|\beta_j|^\gamma$ for $\gamma > 0$.

The so-called Bridge penalty is defined as

$$\text{pen}(\beta) = \sum_{j=1}^{p} |\beta_j|^\gamma, \gamma > 0. \tag{12}$$

The ℓ_1 Lasso penalty is a special case of the bridge penalty where $\gamma = 1$. Also, the ℓ_2 ridge penalty is a special case where $\gamma = 2$. When $0 < \gamma \leq 1$, some components of the estimator minimizing (2) can be exactly zero if λ is sufficiently large [28]. For linear models with $n \gg p$ and $\gamma < 1$, bridge penalty is consistent in variable selection. For the high dimension case where $n \ll p$ and $\gamma < 1$, the bridge can still be consistent if the features associated with the phenotype and those not associated with the phenotype are only weakly correlated [24].

In applications, it is very common that $n \ll p$ because of the time and cost constraints in collecting samples. The EN model will be an ideal choice for feature selection. However, the elastic net model forces the coefficients to be equally penalized in the penalty terms. We can certainly assign different weights to different coefficients. This makes a great deal of sense in biomarker selection from IMS data sets.

In the next section, we propose a so-called weighted EN model to meet the needs in IMS data processing by considering both the spectral and spatial information of the data sets. Compared to the adaptive EN model (11), the WEN methods choose standard deviations as the weight coefficients associated with the estimators for practical applications. We study the variable selection accuracy for the WEN model in the next section as well. The model provides a data driven method and is easy to implement. The results of applying our algorithm to real data collected from biological experiments are satisfactory.

4.2 Weighted Elastic Net Model

In IMS data analysis, if a biomarker in terms of an m/z value in the MS spectrum is truly related to a cancer disease, then it is reasonable to expect that the ion intensity values at this m/z from different pixel locations in a cancer area are approximate

the same. Therefore, the standard deviation of the intensities at the m/z should be small. In comparison, if the biomarker selected by the statistical model based on differentiation mainly caused by the tissue structure, then the ion intensities at the m/z point vary significantly from pixel to pixel. Therefore, the standard deviation of intensities at such an m/z point should be relatively large. Thus, it is proper to associate standard deviations at each predictor to the coefficient in the model to enforce penalty on predictors caused by structure differences. In Chapter 3, the standard deviations of ion intensity at each m/z point have been combined with elastic net model in the tenfold cross validation (CV) step to select the tuning parameter step k. To better consider the spatial information for more precise biomarker selection, we propose the following weighted elastic net (WEN) model [22]:

$$\text{argmin}_\beta \frac{1}{2}\|\mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j\beta_j\|_2^2 + n\lambda_1 \sum_{j=1}^{p} |w_j\beta_j| + \frac{n}{2}\lambda_2 \sum_{j=1}^{p} |w_j\beta_j|^2, \tag{13}$$

where $w_j > 0, j = 1, \cdots, p$ are weighted penalty coefficients. Let $\mathbf{W} = \text{diag}[w_1, \cdots, w_p]$. Then the WEN model can be rewritten as

$$\text{argmin}_\beta \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + n\lambda_1\|\mathbf{W}\beta\|_1 + \frac{n}{2}\lambda_2\|\mathbf{W}\beta\|_2^2. \tag{14}$$

The Weighted elastic net model (13) puts the weights associated with ion intensity spreading information directly into the elastic net model and thus enforces a larger penalty on the coefficients of predictors caused by differences in structure. This model inherits good properties from the EN model including sparse representation, ability to deal with $p \gg n$ problem and group effect. Additionally, compared with EN model, it is more suitable for IMS data analysis since it makes good use of the spatial information and thus it helps to distinguish the selected feature m/z values according to the differences caused by biological structure of the tissue or purely by cancer.

Recall the WEN model (13), with a scaled coefficient difference, we can rewrite it as:

$$f(\lambda_1, \lambda_2, \omega, \beta) = \|\mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \beta_j\|_2^2 + \lambda_1 \sum_{j=1}^{p} w_j |\beta_j| + \lambda_2 \sum_{j=1}^{p} |w_j \beta_j|^2, \tag{15}$$

where $w_j > 0$, $j = 1, \cdots, p$ are weighted penalty coefficients.

Let $\mathbf{y}^*_{(n+p)} = \binom{\mathbf{y}}{o}$, $\mathbf{X}^*_{(n+p)\times p} = (1+\lambda_2)^{-1/2}\binom{\mathbf{X}}{\sqrt{\lambda_2}\mathbf{W}}$, $\gamma = \frac{\lambda_1}{\sqrt{1+\lambda_2}}$, and $\beta^* = \sqrt{1+\lambda_2}\beta$. Then

$$\begin{aligned} f(\lambda_1, \lambda_2, \mathbf{W}, \beta) &= \|\begin{pmatrix}\mathbf{y}\\ 0\end{pmatrix} - \begin{pmatrix}\mathbf{X}\\ \sqrt{\lambda_2}\mathbf{W}\end{pmatrix} \frac{1}{\sqrt{1+\lambda_2}} \sqrt{1+\lambda_2}\,\beta\|_2^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}} w_j \sqrt{1+\lambda_2}|\beta_j| \\ &= \|\mathbf{y}^* - \sum \mathbf{x}_j^* \beta_j^*\|_2^2 + \gamma \Sigma_j w_j |\beta_j^*| \\ &= \|\mathbf{y}^* - \sum \frac{\mathbf{x}_j^*}{w_j} \beta_j^* w_j\|_2^2 + \gamma \Sigma w_j |\beta_j^*| \qquad (15.1) \\ &= g(\gamma, \mathbf{W}, \beta) \end{aligned}$$

Define $\beta_j^{**} = w_j \beta_j^*$ and $\mathbf{x}_j^{**} = \frac{\mathbf{x}_j^*}{w_j}$. Then,

$$g(\gamma, \mathbf{W}, \beta) = \|\mathbf{y}^* - \sum \mathbf{x}_j^{**} \beta_j^{**}\|_2^2 + \gamma \Sigma |\beta_j^{**}|. \tag{15.2}$$

From above formulas (15), (15.1) and (15.2), it is exciting to see that the minimizing problem in the WEN model (15) can be transformed into an equivalent weighted Lasso-type optimization problem (15.1) on augmented data and, further it is also equivalent to a Lasso-type optimization problem (15.2). This implies that WEN also enjoys the computational advantage of the Lasso. This leads us to develop an algorithm for the WEN method based on the algorithm LARS [11]. The pseudo code of WEN algorithm is discussed in section 5.3.2.

4.3 Variable Selection Accuracy of the WEN Model

4.3.1 Lemma for Sign Consistency Condition

The WEN model (13) we proposed is different from the newly developed adaptive elastic net model([58]), because the weight coefficients are generated based on biological considerations of the spread information of the intensities of the m/z values on the cancer area, instead of theoretical considerations. In such a practical model setting, it is quite interesting to test whether the model possesses the oracle property or at least the sign consistency under certain assumptions. These properties would guarantee the reasonableness of the variable selection of the model.

Let us first consider the variable selection accuracy of the WEN model. Results of applying the algorithm will be given in the next section.

Let $\hat{\beta}$ and β^* denote the estimator and the true parameter vector in the linear regression model (13), respectively. We would like to first study necessary and sufficient conditions for $\text{sgn}(\hat{\beta}) = \text{sgn}(\beta^*)$.

To find a solution in nonlinear programming of the optimization problem (13), we first check its Kauush-Kuhn-Tucker (KKT) conditions, a generalization of the method of Lagrange multipliers to inequality constraints. We found that the KKT conditions of the WEN model are equivalent to

$$\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\hat{\beta}) - n\lambda_2 w_j^2 \hat{\beta}_j = \lambda_1 n\, w_j \text{sgn}(\beta_j^*), \quad \text{if } \hat{\beta}_j \neq 0; \tag{16}$$

$$|\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\hat{\beta})| \leq \lambda_1 n\, w_j, \text{ otherwise}, \tag{17}$$

for any $j = 1, \cdots, p$.

Let $b_j = w_j \text{sgn}(\beta_j^*)$ and $\mathbf{b} = \mathbf{W}_S \text{sgn}(\beta_S^*)$. Define the set

$$Z = \{\mathbf{z} \in \mathbb{R}^p; z_j = b_j \text{ for } \hat{\beta}_j \neq 0, \text{ and } |z_j| \leq w_j, \text{otherwise}\}. \tag{18}$$

Then, conditions (16) and (17) are equivalent to say that there exists a subgradient vector $\mathbf{g} \in Z$ such that its components $g_j, j = 1, \cdots, p$, satisfy

$$-\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\hat{\beta}) + n\lambda_2 w_j^2 \hat{\beta}_j + n\lambda_1 g_j = 0. \tag{19}$$

Substituting $\mathbf{y} = \mathbf{X}\beta^* + \epsilon$ in (19), we obtain

$$\mathbf{x}_j^T\mathbf{X}(\hat{\beta} - \beta^*) - \mathbf{x}_j^T\epsilon + n\lambda_2 w_j^2 \hat{\beta}_j + n\lambda_1 g_j = 0.$$

Equivalently, we have:

$$\mathbf{C}(\hat{\beta} - \beta^*) - \frac{1}{n}\mathbf{x}_j^T\epsilon + \lambda_2 w_j^2 \hat{\beta}_j + \lambda_1 g_j = 0. \tag{20}$$

Then, we see that for given $\mathbf{X}$, β^*, and $\lambda_1 > 0, \lambda_2 > 0$, $\text{sgn}(\hat{\beta}) = \text{sgn}(\beta^*)$ holds if and only if

(i) there exists a point $\hat{\beta} \in \mathbb{R}^p$ and a subgradient $\mathbf{g} \in Z$ such that (20) holds,

(ii) $\text{sgn}(\hat{\beta}_S) = \text{sgn}(\beta_S^*)$ and $\hat{\beta}_{S^c} = \beta_{S^c}^* = 0$ implies that $\mathbf{g}_S = \mathbf{b}$ and $|g_j| \leq w_j$ for $j \in S^c$.

Lemma 4.1 *Assume that the weight coefficients $w_j > 0$ for $j = 1, \cdots, p$ and $\mathbf{C}_{11}$ is invertible. Then there is a solution $\hat{\beta}$ for the weighted elastic net such that*

$$\text{sgn}(\hat{\beta}) = \text{sgn}(\beta^*)$$

if and only if the following conditions hold:

$$|\mathbf{x}_j^T\mathbf{X}_S[(\mathbf{C}_{11} + \lambda_2\mathbf{W}^2)^{-1}(\mathbf{C}_{11}\beta_S^* + \frac{\mathbf{X}_S^T\epsilon}{n} - \lambda_1\mathbf{b}) - \beta_S^*] - \frac{\mathbf{x}_j^T\epsilon}{n}| \leq \lambda_1 w_j, \text{ for } j \in S^c, \ (*)$$

and

$$\text{sgn}((\mathbf{C}_{11} + \lambda_2\mathbf{W}^2)^{-1}(\mathbf{C}_{11}\beta_S^* + \frac{\mathbf{X}_S^T\epsilon}{n} - \lambda_1\mathbf{b})) = \text{sgn}(\beta_S^*). \tag{**}$$

Proof. Recall that $\mathbf{y} = \mathbf{X}\beta^* + \epsilon$, $\mathbf{W} = \text{diag}[w_1, \cdots, w_p]$, and $\mathbf{b} = \mathbf{W}_S \text{sgn}(\beta^*_S)$. Substituting $\hat{\beta}_{S^c} = \beta^*_{S^c} = 0$ and $\mathbf{g}_S = \mathbf{b}$ in (20), we obtain

$$\mathbf{C}_{21}(\hat{\beta}_S - \beta^*) - \frac{\mathbf{X}^T_{S_c}\epsilon}{n} = -\lambda_1 \mathbf{g}_{S^c}, \tag{21}$$

$$\mathbf{C}_{11}(\hat{\beta}_S - \beta^*) - \frac{\mathbf{X}^T_S\epsilon}{n} + \lambda_2 \mathbf{W}^2 \hat{\beta}_S = -\lambda_1 \mathbf{g}_S = -\lambda_1 \mathbf{b}, \tag{22}$$

and also

$$\text{sgn}(\hat{\beta}_S) = \text{sgn}(\beta^*) \text{ and } \hat{\beta_{S^c}} = \beta^*_{S^c} = 0. \tag{23}$$

From (21) and (22), solving for $\hat{\beta}_S$ and $\mathbf{g}_{S^c}$, we obtain

$$\hat{\beta}_S = (\mathbf{C}_{11} + \lambda_2 \mathbf{W}^2)^{-1}(\mathbf{C}_{11}\beta^*_S + \frac{\mathbf{X}^T_S\epsilon}{n} - \lambda_1 \mathbf{b}), \tag{24}$$

and

$$-\lambda_1 \mathbf{g}_{S^c} = \mathbf{C}_{21}[(\mathbf{C}_{11} + \lambda_2 \mathbf{W}^2)^{-1}(\mathbf{C}_{11}\beta^*_S + \frac{\mathbf{X}^T_S\epsilon}{n} - \lambda_1 \mathbf{b}) - \beta^*_S] - \frac{\mathbf{X}^T_{S_c}\epsilon}{n}. \tag{25}$$

Therefore, for $j \in S^c$,

$$|\mathbf{x}_j^T \mathbf{X}_S[(\mathbf{C}_{11} + \lambda_2 \mathbf{W}^2)^{-1}(\mathbf{C}_{11}\beta^*_S + \frac{\mathbf{X}^T_S\epsilon}{n} - \lambda_1 \mathbf{b}) - \beta^*_S] - \frac{\mathbf{X}^T_j\epsilon}{n}| = |-\lambda_1 \mathbf{g}_j| \leq \lambda_1 w_j, \tag{26}$$

and

$$\text{sgn}(\hat{\beta}_S) = \text{sgn}((\mathbf{C}_{11} + \lambda_2 \mathbf{W}^2)^{-1}(\mathbf{C}_{11}\beta^*_S + \frac{\mathbf{X}^T_S\epsilon}{n} - \lambda_1 \mathbf{b})) = \text{sgn}(\beta^*_S). \tag{27}$$

This proves the lemma in one direction. To prove the reverse direction, we assume the conditions (*) and (**) in the lemma hold for some $\lambda_1 > 0$ and $\lambda_2 > 0$, and thus we can construct an estimator $\hat{\beta} \in \mathbb{R}^p$ by letting $\hat{\beta}_{S^c} = \beta^*_{S^c} = 0$ and

$$\hat{\beta}_S = [(\mathbf{C}_{11} + \lambda_2 \mathbf{W}^2)^{-1}(\mathbf{C}_{11}\beta^*_S + \frac{\mathbf{X}^T_S\epsilon}{n} - \lambda_1 \mathbf{b})]$$

which guarantees $\text{sgn}(\hat{\beta}_S) = \text{sgn}(\beta^*_S)$ by the condition (**). We can also construct $\mathbf{g}$ by letting $\mathbf{g}_S = \mathbf{b}$ and

$$g_{S^c} = \frac{-1}{\lambda_1}\{\mathbf{C}_{21}[(\mathbf{C}_{11} + \lambda_2 \mathbf{W}^2)^{-1}(\mathbf{C}_{11}\beta^*_S + \frac{\mathbf{X}^T_S\epsilon}{n} - \lambda_1 \mathbf{b}) - \beta^*_S] - \frac{\mathbf{X}^T_{S_c}\epsilon}{n}\}$$

which guarantees that $|g_j| \leq w_j$ for $j \in S^c$ due to the condition (*). Therefore, there exists a parameter vector $\hat{\beta} \in \mathbb{R}^p$ and a subgradient $\mathbf{g} \in Z$ such that $\text{sgn}(\hat{\beta}) = \text{sgn}(\beta^*)$ and equations (25) and (24) are satisfied. This completes the proof of the lemma.

4.3.2 Main Theorem for WEN Model Selection Accuracy

To state and prove the main theorem of this section, we follow the notations defined in [55].

Let $\mathbf{e}_j = \mathbb{R}^s$ be the vector with one in the jth position and zero elsewhere. Then $\|\mathbf{e}_j\|_2 = 1$. We define probability event sets $\mathcal{E}(U)$ and $\mathcal{E}(V)$ relevant to the conditions of (*) and (**) in Lemma 4.1 as follows.

For $j \in S^c$,

$$V_j = \mathbf{x}_j^T \mathbf{X}_S[\beta_S^* - (\mathbf{C}_{11} + \lambda_2 \mathbf{W}^2)^{-1}(\mathbf{C}_{11}\beta_S^* - \lambda_1 \mathbf{b})] + \mathbf{x}_j^T(\mathbf{I}_{n\times n} - \mathbf{X}_S(\mathbf{C}_{11} + \lambda_2 \mathbf{W}^2)^{-1}\mathbf{X}_S^T)\frac{\epsilon}{n}.$$

Then the condition (*) in Lemma 4.1 holds if and only if it is true for the event

$$\mathcal{E}(V) = \{V_j; j \in S^c, |V_j| \leq \lambda_1 w_j\}. \tag{28}$$

Since

$$\hat{\beta_S} = (\mathbf{C}_{11} + \lambda_2 \mathbf{W}^2)^{-1}(\mathbf{C}_{11}\beta_S^* + \frac{\mathbf{X}_S^T \epsilon}{n} - \lambda_1 \mathbf{b})$$

for $j \in S$, we define

$$U_j = \mathbf{e}_j^T(\mathbf{C}_{11} + \lambda_2 \mathbf{W}^2)^{-1}(\frac{\mathbf{X}_S^T \epsilon}{n} - \lambda_2 \mathbf{W}^2 \beta_S^* - \lambda_1 \mathbf{b}).$$

Therefore, we have that the condition (**) in Lemma 4.1 holds if the following event is true:

$$\mathcal{E}(U) = \{U_j; j \in S, \max_{i \in S} |U_j| \leq \beta_{\min}\}, \tag{29}$$

where $\beta_{\min} = \min_{j \in S}|\beta_j|$.

For a symmetric matrix $\mathbf{A}$, $\Lambda_{\min}(\mathbf{A})$ denotes the smallest eigenvalue of $\mathbf{A}$. We assume there exists some constant λ_0 such that

$$\Lambda_{\min}(\mathbf{C}) \geq \lambda_0 > 0.$$

Furthermore, we assume that the ℓ_2-norm of each column of the predictor matrix $\mathbf{X}$ is bounded above by $c_0\sqrt{n}$ for some constant $c_0 > 0$.

Define a probability event set

$$\mathcal{T} = \{\mathbf{X}^T\epsilon; \|\frac{\mathbf{X}^T\epsilon}{n}\|_\infty \leq c_0\sigma\sqrt{\frac{6\log p}{n}}\},$$

here we let $c_0 = \max_{j\in S^c}\|\mathbf{X}_j\|_2/\sqrt{n}$. From known results of inequalities between matrix norms and the assumption that $\Lambda_{\min}(\mathbf{C}_{11}) \geq \lambda_0 > 0$, we know that

$$\|(\mathbf{C}_{11}+\lambda_2\mathbf{W}^2)^{-1}\|_\infty \leq \sqrt{s}\|(\mathbf{C}_{11}+\lambda_2\mathbf{W}^2)^{-1}\|_2 = \frac{\sqrt{s}}{\Lambda_{\min}(\mathbf{C}_{11}+\lambda_2\mathbf{W}^2)} \leq \frac{\sqrt{s}}{\lambda_0}$$

since $\Lambda_{\min}(\mathbf{C}_{11}+\lambda_2\mathbf{W}^2) \geq \Lambda_{\min}(\mathbf{C}_{11}) \geq \lambda_0$. Therefore, by using the triangle inequality, we obtain

$$\max_{j\in S}|U_j| \leq \|(\mathbf{C}_{11}+\lambda_2\mathbf{W}^2)^{-1}\|_\infty\|\frac{\mathbf{X}^T\epsilon}{n}\|_\infty + \|(\mathbf{C}_{11}+\lambda_2\mathbf{W}^2)^{-1}\|_\infty\|\lambda_2\mathbf{W}^2\beta_S^* + \lambda_1\mathbf{b}\|_\infty.$$

Let $\beta^*_{\max} = \max|\beta_j^*|$, $w_{\max}(S) = \max_{j\in S} w_j$. We have that

$$\begin{aligned}\|\lambda_2\mathbf{W}^2\beta_S^* + \lambda_1\mathbf{b}\|_\infty &= \max_{j\in S}(\lambda_2 w_j^2|\beta_j^*| + \lambda_1 w_j) \leq \lambda_2 w_j^2\beta^*_{\max} + \lambda_1 w_j \\ &\leq \lambda_2 w^2_{\max}(S)\beta^*_{\max} + \lambda_1 w_{\max}(S).\end{aligned}$$

From the assumption that

$$\beta_{\min} > \max\{\frac{4c_0\sigma}{\lambda_0}\sqrt{\frac{6s\log p}{n}}, \frac{2(\lambda_2 w^2_{\max}(S)\beta^*_{\max} + \lambda_1 w_{\max}(S))\sqrt{s}}{\lambda_0}\},$$

we have that

$$\frac{\sqrt{s}}{\lambda_0}(c_0\sigma\sqrt{\frac{24\log p}{n}} + \lambda_2 w^2_{\max}(S)\beta^*_{\max} + \lambda_1 w_{\max}(S)) < \beta_{\min}.$$

Thus, we obtain

$$\max_{j\in S}|U_j| \le \|(C_{11}+\lambda_2\mathbf{W}^2)^{-1}\|_\infty\|\frac{X^T\epsilon}{n}\|_\infty + \|(C_{11}+\lambda_2\mathbf{W}^2)^{-1}\|_\infty\|\lambda_2\mathbf{W}^2\beta_S^*+\lambda_1\mathbf{b}\|_\infty$$

$$\le \frac{\sqrt{s}}{\lambda_0}(c_0\sigma\sqrt{\frac{24\log p}{n}}+\lambda_2 w_{\max}^2(S)\beta_{\max}^*+\lambda_1 w_{\max}(S)) < \beta_{\min}.$$

Therefore, we have shown that $j\in\mathcal{T}$ implies $j\in\mathcal{E}(U)$. Hence $P[\mathcal{E}(U)^c]\le P[\mathcal{T}^c]\le 1/p^2$ according to Lemma 9.1 in [55]. Thus, the event $\mathcal{E}(U)$ in (29) holds on the set $\mathcal{T}$.

V_j in (28) is a function of ϵ, thus, a random variable. Its expected value $E[V_j]$

$$\begin{aligned}
\mu_j &= \mathbf{x}_j^T\mathbf{X}_S(\mathbf{C}_{11}+\lambda_2\mathbf{W}^2)^{-1}\lambda_1\mathbf{b}+\mathbf{x}_j^T\mathbf{X}_S[\beta_S^*-(\mathbf{C}_{11}+\lambda_2\mathbf{W}^2)^{-1}\mathbf{C}_{11}\beta_S^*]\\
&= \mathbf{x}_j^T\mathbf{X}_S(\mathbf{C}_{11}+\lambda_2\mathbf{W}^2)^{-1}\lambda_1\mathbf{b}+\mathbf{x}_j^T\mathbf{X}_S(\mathbf{C}_{11}+\lambda_2\mathbf{W}^2)^{-1}[(\mathbf{C}_{11}+\lambda_2\mathbf{W}^2)\beta_S^*-\mathbf{C}_{11}\beta^*]\\
&= \mathbf{x}_j^T\mathbf{X}_S(\mathbf{C}_{11}+\lambda_2\mathbf{W}^2)^{-1}[\lambda_1\mathbf{b}+\lambda_2\mathbf{W}^2\beta_S^*]\\
&= \lambda_1\mathbf{x}_j^T\mathbf{X}_S(\mathbf{C}_{11}+\lambda_2\mathbf{W}^2)^{-1}[\mathbf{b}+\frac{\lambda_2\mathbf{W}^2}{\lambda_1}\beta_S^*].
\end{aligned}$$

Assume for any $j\in S^c$, there exists $\eta\in(0,1)$, such that

$$|\mathbf{x}_j^T\mathbf{X}_S(\mathbf{C}_{11}+\lambda_2\mathbf{W}^2)^{-1}[\mathbf{b}+\frac{\lambda_2\mathbf{W}^2}{\lambda_1}\beta_S^*]| \le w_j(1-\eta).$$

Then, $|\mu_j|\le\lambda_1 w_j(1-\eta)$.

Define

$$\tilde{V}_j = \mathbf{x}_j^T(\mathbf{I}_{n\times n}-\mathbf{X}_S(\mathbf{C}_{11}+\lambda_2\mathbf{W}^2)^{-1}\mathbf{X}_S^T)\frac{\epsilon}{n}, j\in S^c,$$

which is a zero-mean Gaussian random variable with variance

$$\mathrm{Var}(\tilde{V}_j) = \frac{\sigma}{n^2}\mathbf{x}_j^T[(\mathbf{I}_{n\times n}-\mathbf{P})(\mathbf{I}_{n\times n}-\mathbf{P})^T]\mathbf{x}_j \le \frac{\sigma^2}{n^2}\|\mathbf{x}_j\|_2^2 \le \frac{\sigma^2 c_0^2}{n},$$

where $\mathbf{P}=\mathbf{X}_S(\mathbf{C}_{11}+\lambda_2\mathbf{W}^2)^{-1}\mathbf{X}_S^T$.

By using singular value decomposition, we can show that $\|\mathbf{I}-\mathbf{P}\|_2 \leq 1$. Then, by using the tail bound for a Gaussian random variable, the probability value

$$\text{Prob}[|\tilde{V}_j| \geq t] \leq \frac{\sqrt{Var(\tilde{V}_j)}}{t}\exp(\frac{-t^2}{2Var(\tilde{V}_j)}) \leq \frac{\sigma c_0}{\sqrt{n}t}\exp(\frac{-nt^2}{2\sigma^2 c_0^2})$$

with

$$t = \frac{\eta\lambda_1 w_{\min}(S^c)}{2} \geq 2c_0\sigma\sqrt{\frac{2\log(p-s)}{n}}.$$

where $w_{\min}(S^c) = \min_{j\in S^c} w_j$.

We then obtain

$$\text{Prob}\left[\max_{j\in S^c}|\tilde{V}_j| \geq \frac{\eta\lambda_1 w_{\min}(S^c)}{2}\right] \leq \frac{1}{2(p-s)^3\sqrt{2\log(p-s)}}.$$

Thus with probability at least $1-\frac{1}{2(p-s)^3}$, we have for $\forall j \in S^c$,

$$|V_j| \leq |\mu_j| + |\tilde{V}_j| \leq \lambda_1 w_j(1-\eta) + \frac{\eta\lambda_1 w_{\min}(S^c)}{2} \leq \lambda_1 w_j(1-\eta/2) < \lambda_1\omega_j.$$

Therefore, the probability of the event $\mathcal{E}(V)^c$ is at most $\frac{1}{2(p-s)^3}$ and thus less than $\frac{1}{p^2}$ for $s < p$.

Now we are ready to prove the following main result of WEN model.

Theorem 4.2 *For $0 < \eta < 1$, if the predictor matrix $\mathbf{X}$ satisfies*

$$\forall j \in S^c, |\mathbf{x}_j^T\mathbf{X}_S(\mathbf{C}_{11}+\lambda_2\mathbf{W}^2)^{-1}[\mathbf{b}+\frac{\lambda_2\mathbf{W}^2}{\lambda_1}\beta_S^*]| \leq w_j(1-\eta)$$

and

$$\Lambda_{\min}(\mathbf{C}_{11}) \geq \lambda_0 > 0.$$

Where λ_0 is a constant value. Let $c_0 = \max_{j\in S^c}\|\mathbf{x}_j\|_2/\sqrt{n}$. Suppose $w_j > 0$ for $j = 1,\cdots,p$, $w_{\min}(S^c) = \min_{j\in S^c} w_j$, $w_{\max}(S) = \max_{j\in S} w_j$, and λ_1 is chosen such that

$$\lambda_1 w_{\min}(S^c) \geq \frac{4c_0\sigma}{\eta}\sqrt{\frac{2\log(p-s)}{n}}.$$

Assume

$$\beta_{\min} > \max\{\frac{4c_0\sigma}{\lambda_0}\sqrt{\frac{6s\log p}{n}}, \frac{2(\lambda_2 w^2_{\max}(S)\beta^*_{\max} + \lambda_1 w_{\max}(S))\sqrt{s}}{\lambda_0}\},$$

where $\beta^*_{\max} = \max|\beta^*_i|$. *Then for the* $\hat{\beta}$ *in (13), the probability*

$$P[\text{sgn}(\hat{\beta}) = \text{sgn}(\beta^*)] \geq 1 - \frac{2}{p^2}.$$

Proof. According to Lemma 4.1, $sgn(\hat{\beta}) = \text{sgn}(\beta^*)$ if and only if conditions (*) and (**) hold. On the other hand, under the assumptions of this theorem, the condition (*) in Lemma 4.1 holds if the event $\mathcal{E}(V)$ is true and the condition (*) holds if the event $\mathcal{E}(U)$ is true. Therefore,

$$\text{Prob}[\text{sgn}(\hat{\beta}) = \text{sgn}(\beta^*)] \geq 1 - \text{Prob}[\mathcal{E}(U)^c \cup \mathcal{E}(V)^c] \geq 1 - \frac{2}{p^2}.$$

This completes the proof.

4.4 Experimental Results for the WEN Model

In the following, we apply the WEN model to analyze a set of mouse brain IMS data generated from the Vanderbilt Mass Spectrometry Research Center, the same one as we used for the EN4IMS algorithm. The analysis includes a comparison of results by applying the EN method, WEN method, as well as the results obtained using the commercial software SAM, and other popular methods and software programs used in mass spectrometry community.

The WEN model proposed here is for pixel-level classification. When entering the data, cancer pixels and non-cancer pixels are selected from the mouse brain IMS data sets to be as symmetric as possible with the consideration of structure 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 WEN-model is able to use them with no need to reduce dimensions. The early stopping feature of the LARS-type algorithm saves computation cost and time [57].

Comparing the m/z list generated by the regular EN algorithm, the newly developed WEN algorithm that incorporates the spatial penalty term produces an even more concise list by including all significant features with a smaller number of side peaks. For instance, the side peak ($m/z = 10811$) shown in Figure 5 has been removed from the EN list by using the WEN algorithm. In addition, around eighty percent of m/z values in the WEN list are also in the SAM list.

By examining the details of the intensity increase and decrease trends of selected m/z list, similar to the EN based model, we found that most m/z values in the WEN list have a decreasing trend in the tumor region. Again, we remark that when $p \gg n$, linear classifiers often perform better than non-linear ones in many applications [18], even though non-linear methods are known to be more flexible. This fact is related to the asymptotic results in [17]: when $p \gg n$, under mild assumptions for data distribution, the pairwise distances between any two points are approximately identical to each other so the data points form an n-simplex. Therefore, linear classifiers then become natural choices to discriminate two simplices [53].

The WEN peak list include important biomarkers such as Cytochrome c oxidase copper chaperone ($m/z = 6700$), NADH dehydrogenase ($m/z = 7799$) and Cytochrome C oxidase subunit 6c ($m/z = 8380$), which are involved in the electron transport chain. See Section 3.3 for details of the relationship between these proteins and tumor growth.

Table 3 is for comparison of the classification results of algorithms by using PCA+LDA ([34], [37]), PCA+SVM [14] with WEN. These algorithms are applied

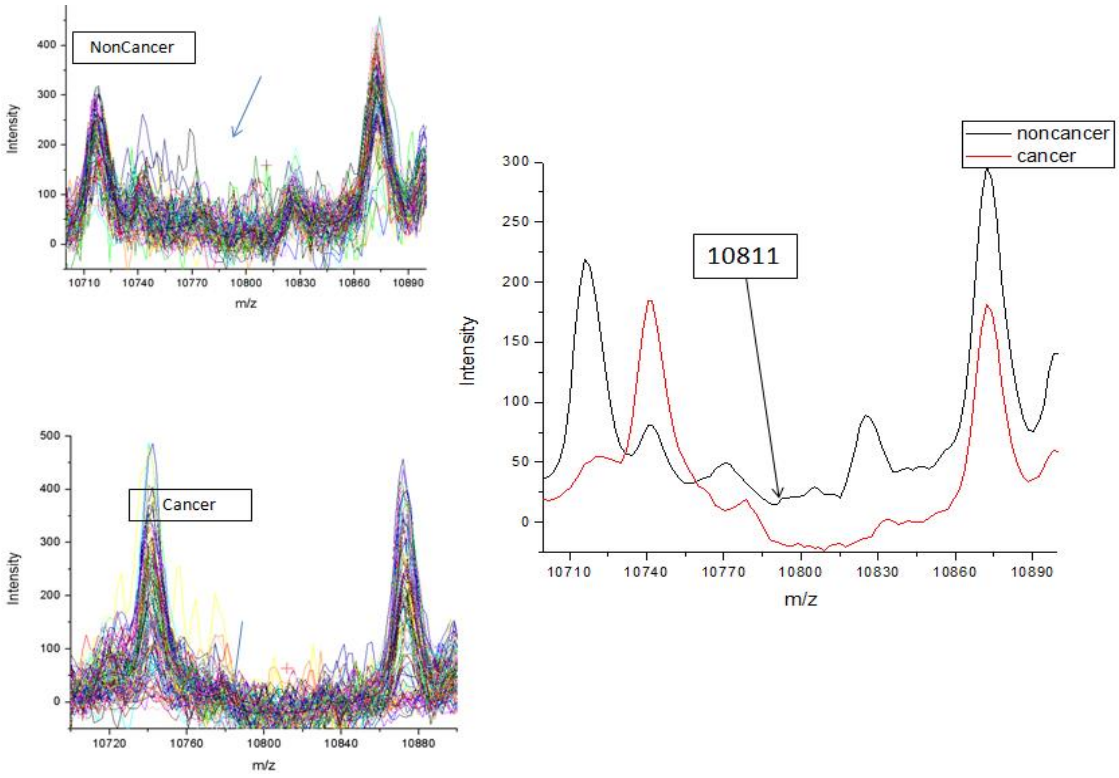

Figure 5: *Side peak ($m/z = 10811$). The peak at $m/z = 10811$ is a fake peak caused by noise, and is not a main peak with biological meaning.*

Table 3: Comparison of WEN classification result with the results by using other current popular methods in the IMS data processing area.

Methods	Accuracy	Sensitivity	Specificity
PCA+LDA	78.64%	100%	57.27%
PCA+SVM	71.82%	84.56%	59.09%
WEN	99.55%	100%	99.09%

to section 1 IMS data to determine the optimal model and then are used to classify section 2 IMS data. The WEN algorithm shows the best classification results and also has an internal feature selection facility.

Compared with the EN-based model, the WEN model directly incorporates the spatial penalty as weighted coefficients into the EN model instead of in the cross validation step. Thus the WEN model is a systematic consideration for the spatial information. Therefore, we expect the WEN model to be more reliable. It provides a platform for a theoretical study of the model as well. Based on the consideration of computation cost, the EN-based model is better than the WEN model.

4.5 Summary and Discussion

In this chapter, we proposed a weighted elastic net model [22] based on IMS data processing needs of using both spectral and spatial information for biomarker selection and classification. The WEN model associates the weight coefficients with ion intensity spreading information directly into the elastic net model instead of in the cross validation step, and thus provides a systematic consideration for the spatial information. This model inherits good properties from the EN model including sparse

representation, ability to deal with problem where $p \gg n$ and grouping effect. In addition, by taking spatial information into account, this model can distinguish the selected feature m/z values according to the differences caused by biological structure of the tissue or purely by cancer. Properties including variable selection accuracy of the WEN model are discussed. The WEN algorithm is applied to IMS data sets for predictor selection, and results show that the WEN method works effectively and efficiently for IMS data processing.

Since the WEN model is based on linear regression, it would be interesting to consider piecewise linear spline regression classifiers for IMS data analysis. However, due to the nonlinearity and the mixed ℓ_1 and ℓ_2 constrains, we expect that such a study is non-trivial, and beyond the scope of this research.

5 Software Development

5.1 Overview of Software Tool IMSmining

The significant advantages of IMS make it a very promising tool for proteomics study. It measures a large collection of mass spectra spreading out over an organic tissue section and retains the absolute spatial information of the measurements for analysis and imaging. It offers the potential of direct examination of biomolecular patterns from cells and tissue. However, IMS data is of a high dimension and complex, and this poses great challenges for data analysis. Furthermore, IMS is a recently new technique and thus mature quantitative analysis methods are not yet implemented into current available software for IMS data. The information extraction and data mining would heavily depend on the quantitative software development.

In order to make better use of the promising IMS technique, we think it is necessary and important to develop software to make the data analysis more convenient and automatic for IMS researchers. Suitable methods for data visualization and effective algorithms for biomarker selection and classification, are our main focus here. We would like to give users more freedom and convenience in visualizing IMS data, to input the m/z values and tuning parameters, to select individual pixels in an area of interest, and to export data sets they need for further analysis.

The software tool IMSmining is a collection of functions that extend the capability of the MATLAB numeric computing environment. It is intended for the visualization and quantitative analysis of IMS data produced by MALDI/SELDI TOF imaging mass spectrometry instruments in experiments. The purpose of this software is to discover biomarkers by learning process and perform a pixel level classification for different IMS data sections, with a special advantage in case where $p \gg n$.

This software provides functions to visualize IMS data in different ways including spectrum figures, 2D and 3D ion intensity distribution graphs of certain input m/z values, original optical image of the plate, and 3D data cube, functions of EN4IMS algorithm and WEN algorithm we proposed for IMS data analysis, functions of other current popular MVA algorithms in IMS area including PCA, LDA, and SVM, and graph operations to rotate and zoom in these figures.

The graphical user interface (GUI) of this software is very friendly and convenient. Users can select an area of pixels from the ion intensity distribution graph as the training data sets, choose the test data sets they want to use from folders, and export data sets which they think are useful for later usage. Also users can input the m/z range they want to use for the 3D data cube, can input the tuning parameter lambda2 for EN based algorithms and can input the m/z value they want to be displayed in the 3D ion intensity distribution graph.

Another convenience is the mutual response between the ion intensity distribution graph and the spectrum figure. When users click on the ion intensity distribution graph, the spectrum figure will show the corresponding spectrum of the pixel which users click on. When users click on the spectrum figure, the ion intensity distribution graph will show the corresponding ion intensity distribution of the m/z value which users click on. We believe this software will provide lots of help and convenience for users to analyze IMS data.

This software package is developed using MATLAB but also provides application for a non-MATLAB environment. To run this software, users need to install MCRInstaller.exe first and then "IMSmining.exe" is ready for use. MCRInstaller.exe is the MATLAB Compiler Runtime, which enables users to run "IMSmining.exe" without MATLAB installed on their computers. The current version of this software only

supports a Microsoft Windows platform. Some algorithms (depending on the size of data) require strong computational effort, so it is recommended to have at least 1 GB of RAM and 2 Gz of processor speed.

5.2 Software Function Description

The GUI of the software package IMSmining is shown in Figure 6. The software allows users to visualize IMS data, to discover biomarkers by learning process and to perform a pixel level classification for different IMS data sections, with a special advantage in $p \gg n$ cases. Main functions of the software will be discussed in this section.

From the menu File, users can click the submenu Import Data and select the folder they put their data sets. Figure 7 shows the GUI after data is entered. The upper left figure shows the ion intensity distribution graph of certain m/z value and the upper right figure shows the spectrum of certain pixel. The interactive response between the ion intensity distribution graph and the spectrum figure gives great convenience for users.

In Figure 7, the m/z value is 14935 and the pixel is the one with $x = 10$ and $y = 22$. When users click on the upper left figure, the upper right figure will show the corresponding spectrum of the pixel which users click on. When users click on the upper right figure, the upper left figure will show the corresponding ion intensity distribution of the m/z value which users click on. The software can automatically match the (x, y) position of the upper left figure with the original data, namely all these text files where all spectrum information (m/z values and intensities) is stored.

From the submenu Original Picture under the menu Display, users can select the original optical image, which is shown in the lower left figure. Combined together,

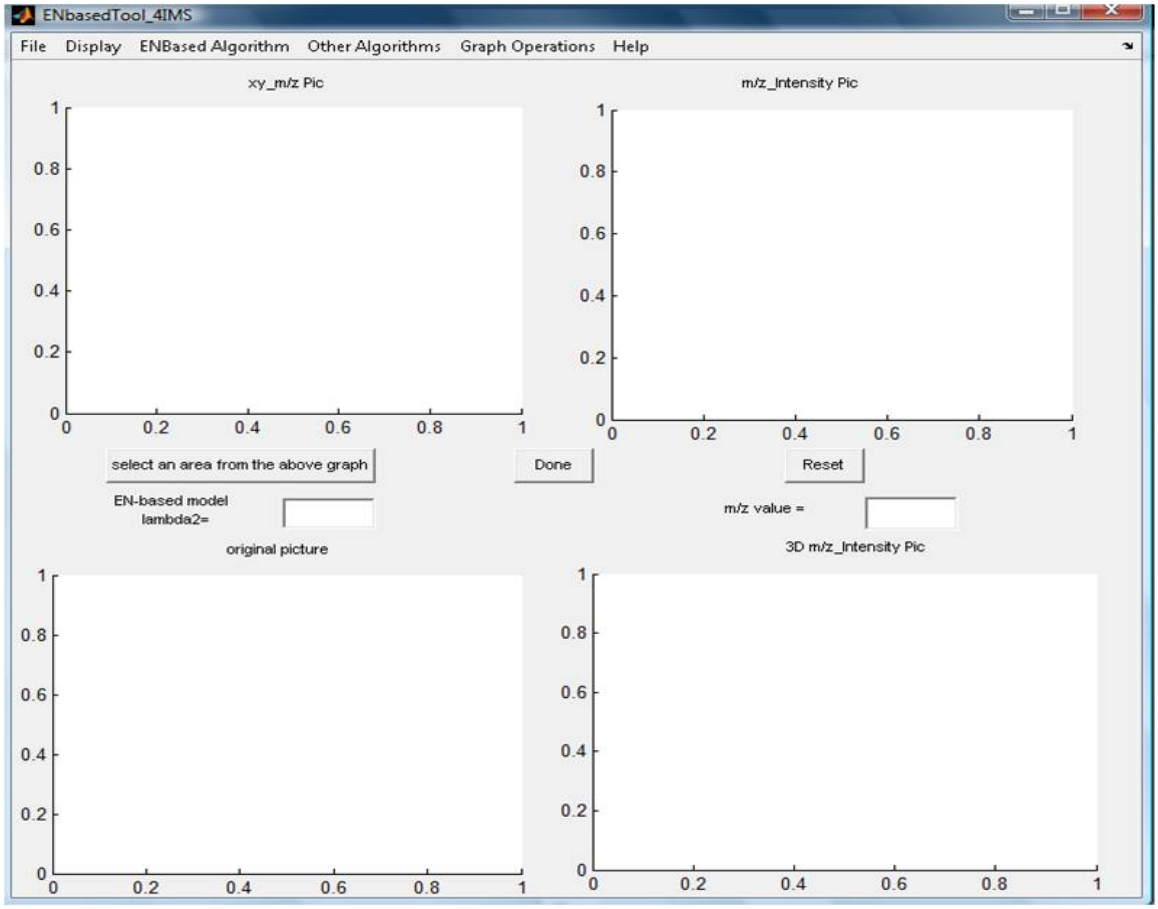

Figure 6: *Graphical user interface of the software package IMSmining*

these three figures can give users rich information about the IMS data and aid in biomarker selection.

(1) Users can directly compare the shape of the upper left figure with the shape of the lower left figure, which helps to check biological meaning of certain m/z value.

(2) Users can see ion intensity distribution graphs of different m/z values just by clicking on different m/z values on the upper right figure. It is very convenient, fast and easy to compare different ion intensity distribution graphs.

(3) Users can see spectrum of different pixels just by clicking on different positions on the upper left figure. Users can use zoom on to enlarge the spectrum which is convenient to check whether the m/z value is a side peak or not.

(4) Users can select an area of pixels and see the mean spectrum of these selected pixels on the upper right figure. The mean spectrum can be saved for later usage. Besides, Users can use these selected pixels as the training data sets for classification and can export the data sets they need for later analysis.

(5) Users can input certain m/z value and generate the 3D ion intensity distribution graph at the lower right figure. This figure is 3D instead of 2D and rotatable, which can provide additional information. This design of combining these figures, provides as much information as possible for users.

From the submenu of Data Cube under the menu Display, users can visualize IMS data in a 3D data cube, which is shown in the upper right figure of Figure 8. It is similar to Figure 1 (b) and (d) in section 2.1. The software allows users to input maximum and minimum values for m/z on the z axis of the 3D data cube. In addition users can do graph operations to 3D data cube figure such as rotation, enlargement and etc.

The above functions are mainly for data visualization and help users to check

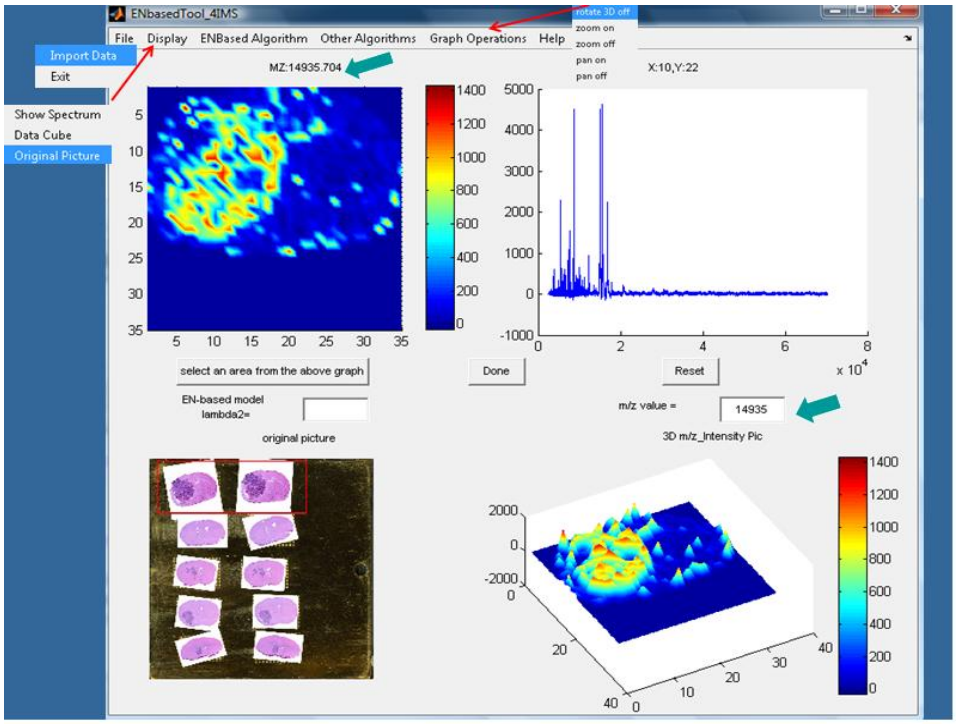

Figure 7: *GUI after data is entered. The upper left figure shows the ion intensity distribution graph, which has interactive response with the upper right figure of the spectrum. The lower left figure is the original optical image of the mouse brain section. The lower right figure is the 3D representation of ion intensity distribution.*

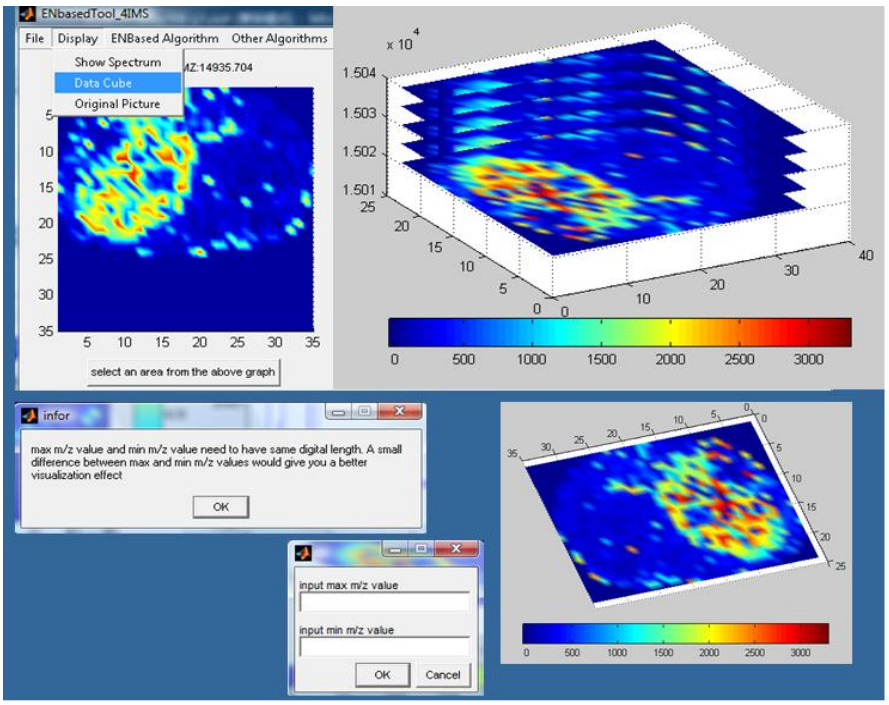

Figure 8: *GUI after 3D data cube function is applied. Several dialogs and figures show the details of this process*

their selected biomarkers. Figure 9 shows the functions mainly related with analysis algorithms for biomarker selection and classification. The algorithms include EN4IMS and WEN we proposed, and other current popular MVA methods such as PCA, LDA and SVM. By clicking on the "select an area from the above graph" button, users can select training cancer data set and training noncancer data set directly from the ion intensity distribution graph, which is shown in the upper left figure of Figure 6. Users can also export the data set of selected pixels if they want, and select the folders where they put their test data sets.

For EN4IMS and WEN algorithms, users can input the tuning parameter lambda2. After running these algorithms, the selected m/z list and classification rate will be available and can be saved. If users are satisfied with this trained model, they can continue to use this trained model to classify other unknown data sets. All these functions together greatly help users to analyze their IMS data.

The computation cost of the algorithm is one consideration here since IMS data is of a very high dimension. For example, the global PCA algorithm has very high computation cost and seems impossible to be applied to all 22195 m/z values (all the input variables). However, by transposing a matrix of dimension, say $a \times b$ with $a \ll b$, the computation cost can be reduced from $O(b^3)$ to $O(a^3)$, which makes it possible to use PCA. Details are given in [2]. We implement the LDA algorithm as discussed in [34] and [37] and the SVM algorithm as discussed in [14]. Since these algorithms all use PCA for dimension reduction before applying LDA and SVM, the computation cost would be acceptable here. The computational aspect of the EN4IMS and WEN algorithms is discussed in the following section.

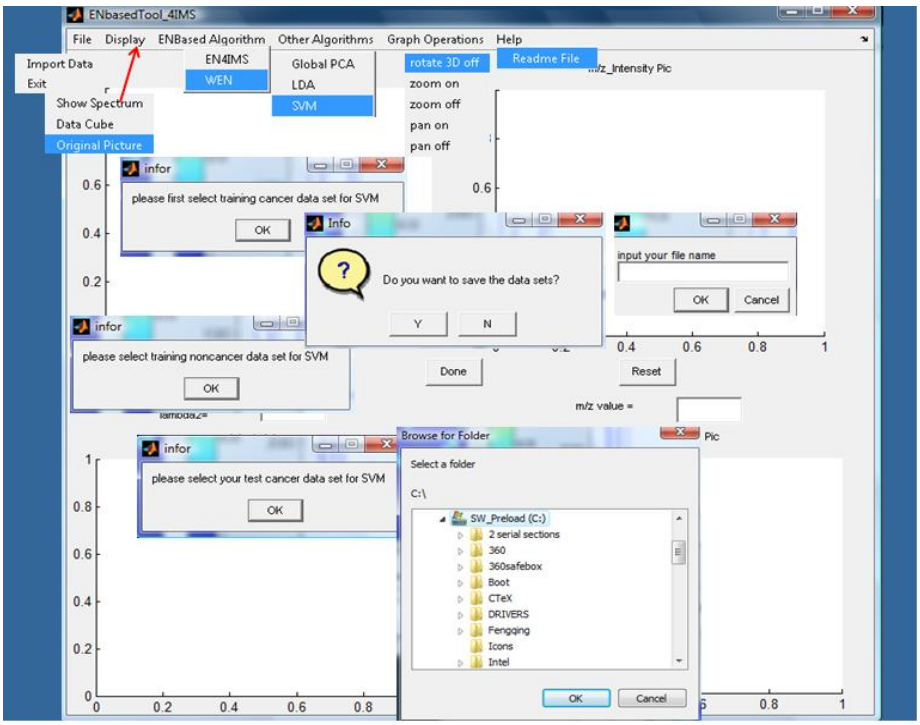

Figure 9: *GUI for SVM classifier function is applied. Several dialogs show the details of this process.*

5.3 Computation: Pseudo Codes

5.3.1 The EN4IMS Algorithm

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 non-cancer area. Lambdas are the regularization parameters. The subindex set A denotes the set of all selected variables as defined in the pseudo code. The EN4IMS algorithm inherits the early stop feature of LARS-EN 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 $\hat{\beta} = 0, k = 0$.

2. Let $\hat{\mathbf{C}} = \mathbf{X}^T(\mathbf{y} - \hat{\mu}_S), C_M = \max_j\{|\hat{c}_j|\}, S = \{j : |\hat{c}_j| = C_M\}, s_j = \text{sgn}\{\hat{c}_j\}$ for $j \in S, \mathbf{X}_S = (\ldots s_j\mathbf{x}_j \ldots)_{j\in S}, \hat{\mu}_S = \mathbf{X}_S\hat{\beta}_S, d_1 = \sqrt{\lambda_2}, d_2 = \frac{1}{\sqrt{1+\lambda_2}}$.

While $(S^c \neq \varnothing)$ **Do**

(a) $\mathbf{G}_S = \mathbf{X}_S^T\mathbf{X}_S$, $A_S = (\mathbf{1}_S^T\mathbf{G}_S^{-1}\mathbf{1}_S)^{-1/2}$

(b) Calculate equiangular vector

$\mathbf{u}_1 = \mathbf{X}_S\mathbf{\Omega}_S d_2$

$\mathbf{u}_2 = \mathbf{\Omega}_S d_1 d_2$

where $\mathbf{\Omega}_S = A_S \mathbf{G}_S^{-1} \mathbf{1}_S$

(c) Calculate the inner product vector

$\mathbf{a} = (\mathbf{X}^T \mathbf{u}_1 + \mathbf{u}_2 d_1) d_2$

(d) Update current algorithm estimate

$\hat{\mu}_S = \hat{\mu}_S + \hat{\gamma} \mathbf{u}_1$

where $\hat{\gamma} = \min_{j \in S^c}^{+} \{ \frac{C_M - \hat{c}_j}{A_S - a_j}, \frac{C_M + \hat{c}_j}{A_S + a_j} \}$

(e) Update the support (active) set S

if $\tilde{\gamma} < \hat{\gamma}, S = S - \{\tilde{j}\}$

else $S = S + \{\tilde{j}\}$

where $\tilde{\gamma} = \min_{\gamma_j > 0} \{\gamma_j\}, \gamma_j = -\hat{\beta}_j / \hat{d}_j, \hat{d}_j = s_j \mathbf{\Omega}_{Sj}$

(f) $k = k + 1$

End Do

3. Find step k_{opt} to select the optimal model by using ten-fold cross validation to minimize the following functional

$\|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2^2 + \frac{\lambda_3}{M} \sum_{j=1}^{M} \sqrt{\frac{\sum_{i=1}^{N} (x_{ij} - \mu_j)^2}{N-1}} \}$

5.3.2 The LARS-WEN Algorithm

For a fixed λ_2, the weighted EN optimization problem is equivalent to a weighted lasso problem on an augmented data set and could be further transformed into a lasso problem. We therefore develop algorithm LARS-WEN based on the LARS algorithm to create the entire solution path. In the WEN model, there are two tuning parameters λ_1, λ_2. Typically, the tuning parameter, λ_2, is picked as a relatively small grid, say (0, 0.01, 0.1, 1, 10, 100). For each λ_2 , algorithm LARS-WEN produces

all possible WEN estimates of the vector β for the IMS data. We just want a single optimal β^*; thus, some rules for selecting among the possibilities are needed. When only training data are available, tenfold cross validation is a popular method for estimating the prediction error and comparing different models ([18], chapter 7). In our algorithm, the other tuning parameter λ_1 is selected by tenfold CV. The pseudo code for LARS-WEN is listed as below.

Algorithm (LARS-WEN)

1. Input predictor matrix $\mathbf{X}$ of covariate vectors $\mathbf{x}_j$, the response vector $\mathbf{y}$ and weight coefficients w_j. Set $\hat{\beta} = 0, k = 0$ and $\mathbf{x}_j = \mathbf{x}_j / w_j$.

2. Let $\hat{\mathbf{C}} = \mathbf{X}^T(\mathbf{y} - \hat{\mu}_S), C_M = \max_j\{|\hat{c}_j|\}, S = \{j : |\hat{c}_j| = C_M\}, s_j = \text{sgn}\{\hat{c}_j\}$ for $j \in S, \mathbf{X}_S = (\ldots s_j\mathbf{x}_j \ldots)_{j\in S}, \hat{\mu}_S = \mathbf{X}_S\hat{\beta}_S, d_1 = \sqrt{\lambda_2}, d_2 = \frac{1}{\sqrt{1+\lambda_2}}$, and $\mathbf{W} = \text{diag}[w_1, \cdots, w_p]$.

While $(S^c \neq \varnothing)$ **Do**

(a) $\mathbf{G}_S = \mathbf{X}_S^T\mathbf{X}_S$, $A_S = (\mathbf{1}_S^T\mathbf{G}_S^{-1}\mathbf{1}_S)^{-1/2}$

(b) Calculate equiangular vector

$\mathbf{u}_1 = \mathbf{X}_S\mathbf{\Omega}_S d_2$

$\mathbf{u}_2 = \mathbf{W}_S\mathbf{\Omega}_S d_1 d_2$

where $\mathbf{\Omega}_S = A_S\mathbf{G}_S^{-1}\mathbf{1}_S$

(c) Calculate the inner product vector

$\mathbf{a} = (\mathbf{X}^T\mathbf{u}_1 + \mathbf{W}^T\mathbf{u}_2 d_1)d_2$

(d) Update current algorithm estimate

$\hat{\mu}_S = \hat{\mu}_S + \hat{\gamma}\mathbf{u}_1$

where $\hat{\gamma} = \min^+_{j\in S^c}\{\frac{C_M - \hat{c}_j}{A_S - a_j}, \frac{C_M + \hat{c}_j}{A_S + a_j}\}$

(e) Update the support (active) set S

if $\tilde{\gamma} < \hat{\gamma}, S = S - \{\tilde{j}\}$

else $S = S + \{\tilde{j}\}$

where $\tilde{\gamma} = \min_{\gamma_j > 0}\{\gamma_j\}, \gamma_j = -\hat{\beta}_j / \hat{d}_j, \hat{d}_j = s_j \mathbf{\Omega}_{Sj}$

(f) $k = k + 1$

End Do

3. Output $\hat{\beta}_j = \hat{\beta}_j / w_j$. Find step k_{opt} to select the optimal model by using ten-fold cross validation.

5.4 Summary

In this chapter, we described the functions of a self-developed software package called IMSmining. This package aims to provide different ways for data visualization, effective algorithms for biomarker selection and classification, which derive from the EN-based model, the WEN model and multivariate analysis methods discussed in Chapter 3 and Chapter 4. The graphical user interface is very friendly and provides convenient, fast and easy ways for data visualization, comparison and analysis. One advantage of the software is the interactive response between the ion intensity distribution graph and the spectrum figure. Another great convenience is that users can select an area of interest directly from the ion intensity distribution graph and use selected data sets for model training, biomarker selection and classification. We believe this software will be extremely helpful to researchers in analyzing IMS data where complexity and high dimensionality pose great challenges for information extraction and data mining.

BIBLIOGRAPHY

[1] A.F. Altelaar, I.M. Taban, L.A. Mcdonnell,et al., *High-resolution MALDI imaging mass spectrometry allows localization of peptide distributions at cellular length scales in pituitary tissue sections*, Int. J. Mass Spectrom, 260 (2007), pp. 203-211.

[2] CM. Bishop, Pattern recognition and machine learning, Springer, 2007.

[3] L. Breiman, *Better subset regression using the nonnegative garrote*, Technometrics, 37 (1995), pp. 373-384.

[4] E. Candes and T. Tao, *The dantzig selector: statistical estimation when p is much larger than n*, Annals of Statistics, 35 (2007), pp. 2313-2351.

[5] P. Chaurand, M.A. Rahman, T. Hunt, et al., *Monitoring mouse prostate development by profiling and imaging mass spectrometry*, Mol. Cell. Proteomics, 7 (2008), pp. 411-423.

[6] P. Chaurand, S.A. Schwartz, R.M. Caprioli, *Profiling and imaging proteins in tissue sections by MS*, Anal. Chem., 76 (2004), 5, pp. 86A-93A.

[7] P. Chaurand, S.A. Schwartz, R.M. Caprioli, *Assessing protein patterns in disease using imaging mass spectrometry*, J. Proteome Res., 3 (2004), pp. 245-252.

[8] G. Chu, B. Narasimhan, R. Tibshirani, V.G. Tusher, *SAM Version 1.12: user's guide and technical document*, [http://www-stat.stanford.edu/ tibs/SAM/]

[9] D.S. Cornett, M.L. Reyzer, P. Chaurand, R.M. Caprioli, *MALDI imaging mass spectrometry: molecular snapshots of biochemical systems*, Nat. Methods, 4 (2007), pp. 828-833.

[10] S.O. Deininger, M.P. Ebert, A. Futterer, M. Gerhard, C. Rocken, *MALDI Imaging Combined with Hierarchical Clustering as a New Tool for the Interpretation of Complex Human Cancers*, J. Proteome Res., 7 (2008), 12, pp. 5230-5236.

[11] B. Efron, T. Hastie, R. Tibshirani, *Least angle regression*, Annals of Statistics, 32 (2004), pp. 407-499.

[12] J. Fan, R. Li, *Variable selection via nonconcave penalized Likelihood and Its Oracle Properties*, Journal of the American Statistical Association, 96 (2001), pp. 1348-1360.

[13] I. Frank, J. Friedman, *A statistical view of some chemometrics regression tools*, Technometrics, 35 (1993), pp. 109-148.

[14] M. Gerhard, S.O. Deininger, F.M. Schleif, *Statistical Classification and visualization of MALDI imaging data*, CBMS'07 2007, pp. 403-405.

[15] D.J. Graham, M.S. Wagner, D.G. Castner, *Information from complexity: challenges of TOF-SIMS data interpretation*, Applied surface science, 252 (2006), pp. 6860-6868.

[16] A. Hadnagy, R. Beaulieu, D. Balicki, *Histone tail modifications and noncanonical functions of histones : perspectives in cancer epigenetics*, Molecular Cancer Therapeutics, 7 (2008), pp. 740-748.

[17] P. Hall et al., *Geometric representation of high dimension low sample size data*, J. R. Statist. Soc., B, 67 (2005), pp. 427-444.

[18] T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning; Data mining, inference and prediction*, Springer, New York, 2001.

[19] R.M. Heeren, D.F. Smith, J. Stauber, B. Kkrer-Kaletas, L. MacAleese, *Imaging Mass Spectrometry: Hype or Hope?*, J. American Society for Mass Spectrometry, 20 (2009), 6, pp. 1006-1014.

[20] A. E. Hoerl, R. W. Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics, 12 (1970), pp. 55-67.

[21] D. Hong, H.M. Li, M. Li, Y. Shyr, *Wavelets and Projecting Spectrum Binning for Proteomic Data Processing, In: Quantitative Medical Data Analysis Using Math Tools and Statistical Techniques*, World Scientific Publications, LLC., Singapore. 2007, pp. 159-178.

[22] D. Hong and F. Zhang, *Weighted Elastic Net Model for Mass Spectrometry Imaging processing*, Math. Model. Nat. Phenom., 2010, to appear.

[23] J. Huang, J. Horowitz, S. Ma, *Asymptotic properties of bridge estimators in sparse high-dimensional regression models*, Annals Statatistics, 36 (2008), pp. 587-613.

[24] J. Huang, S. Ma, C. Zhang, *Adaptive Lasso for sparse high dimensional regression models*, Stat Sin, 18 (2008), pp. 1603-1618.

[25] G.M. James, P. Radchenko, and J. Lv, *DASSO: connections between the Dantzig selector and lasso*, J. R. Statist. Soc., B, 71 (2009), pp. 127-142.

[26] J. Jia, B. Yu, *On model selection consistency of the elastic net when* $p \gg n$, Tech. Report, 756, Statistics, UC Berkeley, 2008.

[27] S. Khatib-Shahidi, M. Andersson, J.L. Herman, T.A. Gillespie, R.M. Caprioli,

Direct molecular analysis of whole-body animal tissue sections by imaging MALDI mass spectrometry, Anal. Chem., 78 (2006), pp. 6448-6456.

[28] K. Knight, W. Fu, *Asymptotics for Lasso-type estimators*, Annals Statistics, 28 (2000), pp. 1356-1378.

[29] R. Lemaire et al., *Specific MALDI imaging and profiling for biomarker hunting and validation: fragment of the 11s proteasome activator complex, reg alpha fragment, is a new potent ovary cancer biomarker*, J. Proteome Res., 6 (2007), 11, pp. 4127-4134.

[30] S. Matoba, J.G. Kang, et al., *P53 regulates mitochondrial respiration*, Science, 312 (2006), pp. 1650-1653.

[31] S. Ma, J. Huang, *Penalized feature selection and classification in bioinformatics*, Brief in Bioinform., 9 (2008), pp. 392-403.

[32] A. Mayevsky, *Mitochondrial function and energy metabolism in cancer cells: Past overview and future perspectives*, Mitochondrion, 9 (2009), pp. 165-179.

[33] L.A. McDonnell, A.V. Remoortere, J.M. Rene, et al., *Mass spectrometry image correlation: quantifying colocalization*, J. Proteome Res., 7 (2008), pp. 3619-3627.

[34] G. McCombie, D. Staab, M. Stoeckli, R. Knochenmuss, *Spatial and Spectral correlation in MALDI mass spectrometry images by clustering and multivariate analysis*, Anal. Chem., 77 (2005), pp. 6118-6124.

[35] N. Meinshausen, B. Yu, *Lasso-type recovery of sparse representations for high-dimensional data*, Annals of Statistics, 37 (2009), 1, pp. 246-270.

[36] H. Meistermann, J.L. Norris, et al., *Biomarker discovery by imaging mass spectrometry: transthyretin is a biomarker for gentamicin-induced nephrotoxicity in rat*, Mol Cell Proteomics, 5 (2006), pp. 1876-1886.

[37] E.R. Muir, I.J. Ndiour, et al., *Multivariate analysis of imaging mass spectrometry data*, BIBE 2007 proceedings of the 7th IEEE international conference, pp. 472-479.

[38] K. Ohuchida, K. Mizumoto, Y. Ogura, et al., *Quantitative assessment of telomerase activity and human telomerase reverse transcriptase messenger RNA levelsin pancreatic juice samples for the diagnosis of pancreatic cancer*, Clin Cancer Res, 12 (2006), pp. 5417-5422.

[39] I. Rehman, A. Azzouzi, J. Catto, et al., *Calgizzarin (S100A11) immunostaining pattern is altered in prostate cancer suggesting a role in tumourigenesis*, European Urology Supplements, 2 (2003), pp. 68-68.

[40] S.A. Schwartz, R.J. Weil, M.D. Johnson, S.A. Toms, R.M. Caprioli, *Protein profiling in brain tumors using mass spectrometry: feasibility of a new technique for the analysis of protein expression*, Clin Cancer Res., 10 (2004), pp. 981-987.

[41] M. Stoeckli, P. Chaurand, D.E. Hallahan, R.M. Caprioli, *Imaging mass spectrometry: a new technology for the analysis of of protein expression in mammalian tissues*, Nat. Med., 7 (2001), pp. 493-496.

[42] M. Tanaka, K. Adzuma, M. Iwami, K. Yoshimoto, Y. Monden, M. Itakura , *Human calgizzarin; one colorectal cancer related gene selected by a large scale random cDNA sequencing and Northern blot analysis*, Cancer Letters, 89 (1995), pp. 195-200.

[43] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. R. Statist. Soc., Series B., 58 (1996), 1, pp. 267-288.

[44] D. Touboul, F. Kollmer, E. Niehuis, A. Brunelle, O. Laprevote, *Improvement of biological time-of-flight secondary ion massspectrometry imaging with bismuth cluster ion source*, J. Am. Soc. Mass Spectrom., 16 (2005), pp. 1608-1618.

[45] P.J. Trim, S.J. Atkinson, A.P. Princivalle, P.S. Marshall, A. West, *Matrix-assisted laser desorption/ionisation mass spectrometry imaging of lipids in rat brain tissue with integrated unsupervised and supervised multivariant statistical analysis*, Rapid Commun. Mass Spectrom., 22 (2008), pp. 1503-1509.

[46] R. Van de Plas, F. Ojeda, M. Dewil, L. Van Den Bosch, B. De Moor, E. Waelkens, *Prospective exploration of biochemical tissue composition via imaging mass spectrometry guided by principal component analysis*, Pacific Sympossium on Biocomputing, 12 (2007), pp. 458-469.

[47] R. Van de Plas, B. De Moor, E. Waelkens, *Imaging mass spectrometry based exploration of biochemical tissue composition using peak intensity weighted PCA*, Life Science Systems and Applications Workshop, LISA. IEEE/NIH. 2007, pp. 209-212.

[48] B.T. Wickes, K. Yongmin, D.G. Castner, *Denoising and multivariate analysis of time-of-flight SIMS images*, Surf. Interface Anal., 35 (2003), pp. 640-648.

[49] K. Yanagisawa, Y. Shyr, B.J. Xu, et al., *Proteomic patterns of tumour subsets in non-small-cell lung cancer*, Lancet, 362 (2003), 9382, pp. 433-439.

[50] C.B. Yoo, P.A. Jones, *Epigenetic therapy of cancer: past, present and future*, Nat. Rev. Drug Discov., 5 (2006), pp. 37-50.

[51] M. Yuan, Y. Lin, *On the nonnegative garrote estimator*, J. R. Statist. Soc., B., 69 (2007), pp. 143-161.

[52] F. Zhang, D. Hong, et al., *Elastic Net Based Framework for Imaging Mass Spectrometry Data Biomarker Selection and Classification*, manuscript.

[53] H. Zhang et al., *Gene selection using support vector machines with non-convex penalty*, Bioinformatics, 22 (2006), pp. 88-95.

[54] P. Zhao, B. Yu, *On model selection consistency of lasso*, The Journal of Machine Learning Research, 7 (2006), pp. 2541-2563.

[55] S. Zhou, S. Geer, P. Buhlmann, *Adaptive lasso for high dimensional regression and gaussian graphical modeling*, manuscript, 2009.

[56] H. Zou, *The adaptive lasso and its oracle properties*, Journal of the American Statistical Association, 101 (2006), pp. 1418-1429.

[57] H. Zou, T. Hastie, *Regularization and variable selection via the elastic net*, J. R. Statist. Soc., B., 67(2005), 2, pp. 301-320.

[58] H. Zou, H. Zhang, *On the adaptive elastic-net with a diverging number of parameters*, Annals of statistics, 37 (2009), pp. 1733-1751.