Math. Model. Nat. Phenom.
Vol. 5, No. 3, 2010, pp. 115-133
DOI: 10.1051/mmnp/20105308

Weighted Elastic Net Model for Mass Spectrometry Imaging Processing

D. Hong* **and F. Zhang**

Department of Mathematical Science, Middle Tennessee State University
Murfreesboro, Tennessee, USA

Abstract. In proteomics study, Imaging Mass Spectrometry (IMS) is an emerging and very promising new technique for protein analysis from intact biological tissues. Though it has shown great potential and is very promising for rapid mapping of protein localization and the detection of sizeable differences in protein expression, challenges remain in data processing due to the difficulty of high dimensionality and the fact that the number of input variables in prediction model is significantly larger than the number of observations. To obtain a complete overview of IMS data and find trace features based on both spectral and spatial patterns, one faces a global optimization problem. In this paper, we propose a weighted elastic net (WEN) model based on IMS data processing needs of using both the spectral and spatial information for biomarker selection and classification. Properties including variable selection accuracy of the WEN model are discussed. Experimental IMS data analysis results show that such a model not only reduces the number of side features but also helps new biomarkers discovery.

Key words: biomarker discovery, weighted elastic-net, mass spectrometry imaging, penalized regression, variable selection
AMS subject classification: 62J05, 62J07, 62H35, 62P10

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

*Corresponding author. E-mail: dhong@mtsu.edu

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. MALDI-Imaging, in particular developed from the well-established single-cell detection techniques MALDI-TOF, is an emerging and promising new technique for protein analysis from intact biological tissues [2]. 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. IMS has its unique advantages in discovering biomarkers. A profiling strategy by manually spotting matrix on predefined areas of interest is often biased by design because it requires manual intervention and only a fragmentary analysis of the tissue at low spatial resolution can be obtained. However, IMS, by the automatic spotting of matrix on the tissue in an array format, results in comprehensive structural analysis at a higher spatial resolution and also saves time. Another clear advantage is the imaging strategy is to go further into structure or morphological detail [23].

IMS data have very high dimensions. Each data set generated by IMS has two spatial dimensions and ion intensities along the mass-over-charge (m/z) dimension. As an example, Figure 1(a) shows the stained mouse brain section implanted with a GL26 glioma cell line. The darker region indicates the tumor area. 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). Behind every pixel is an entire individual mass spectrum shown in Figure 1(c) ranging from $2k$ to $70k$ Dalton in our data set. This in itself is a challenge for data processing, but it is further compounded by the low signal intensities found across the image. Figure 1(d) is the visualization of IMS data represented as a data cube. The x-, y- and z-axes are the same as in Figure 1(b). Five spatial distribution graphs are also shown in Figure 1(d) corresponding to five selected m/z values. Each spatial distribution 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. To fully utilize IMS data, it is desirable to not only identify the peaks of the spectrum within individual pixels, but also to preserve the spatial information for the whole images. The combination of spatial and mass resolution results in large and complex data sets that gives a great challenge to the quantitative analysis and interpretation tools. Conventional images, derived from a specific analyte mass, do not identify the spatially localized correlations between analytes that are latent in IMS data processing.

The application of multivariate analysis (MVA) methods has opened new doors for the exploration of IMS data. Most of MVA methods work towards one central task of summarizing the variance patterns within a dataset. The IMS community has begun exploring and comparing these MVA methods but few guidelines have been established for data pre-processing before these MVA methods are applied [9]. A newly developed variable selection method [32], called elastic net (EN), 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 predictors (p) is much larger than the number of observations (n), and also encourages a grouping effect, where strongly correlated

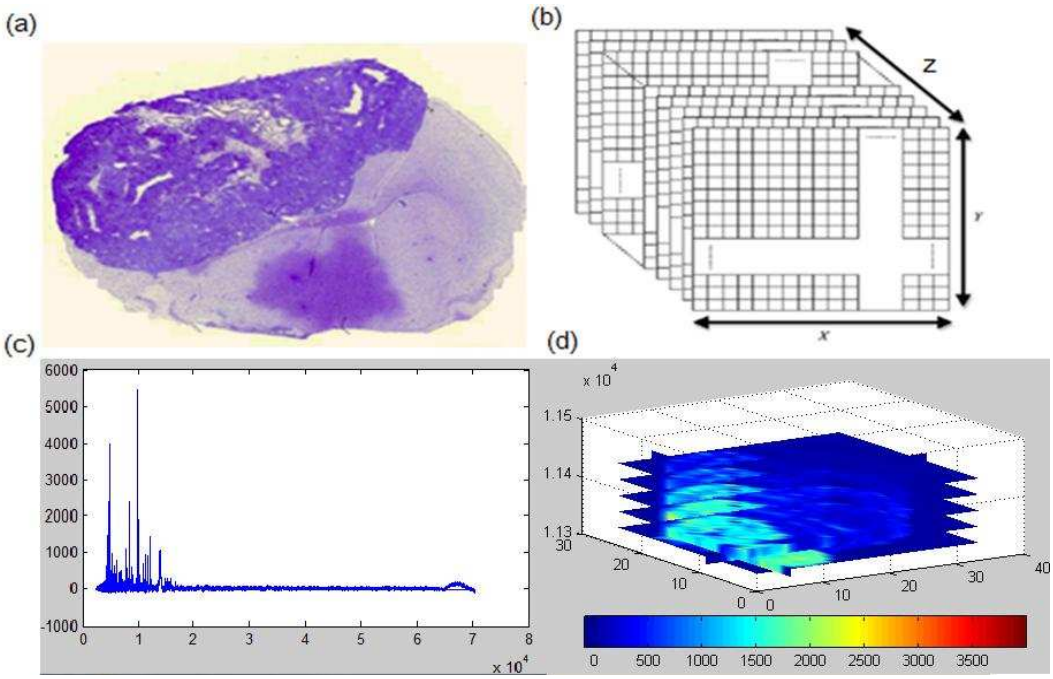

Figure 1: Mouse brain IMS Data.

predictors tend to be in or out of the model together. In this paper, we construct a weighted elastic net (WEN) model for predicting variable selection based on the consideration of both spectral and spatial information of IMS data and develop a new tool for IMS data feature selection and classification. The WEN model fully utilizes not only the spectral information within individual pixels but also the spatial information for the whole images. Properties of WEN model such as the variable selection accuracy are discussed. The WEN algorithm is applied to an IMS data set for predictor selection. The analysis results showed that the WEN method works efficiently and effectively for IMS data processing. A set of biomarkers has been identified with interesting biological explanations.

The remainder of the paper is organized as follows: In Section 2, penalized feature selection models such as Lasso, bridge and ridge regression, elastic net, and adaptive Lasso models are briefly reviewed together with the conditions for variable selection consistency. The weighted elastic net model is then introduced and theoretically studied on its variable selection accuracy in Section 3. In Section 4, we present the WEN algorithm and its application to an IMS data set for predictors selection.

2. Sparse Representation Models

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 accuracy 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

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \tag{2.1}$$

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 $\beta = (\beta_0, \ldots, \beta_p)^T$.

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 [12], followed by Bridge regression [7], the Garotte [1], the Lasso [25], LARS [5], and very recently the elastic net [32]. 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 [15] 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

$$\hat{\beta} = \mathrm{argmin}_\beta\{\frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\,\mathrm{pen}(\beta)\} \tag{2.2}$$

The penalty $\mathrm{pen}(\beta)$ in (2.2) controls the complexity of the model. Here $\mathrm{pen}(\beta)$ 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 [19]. With 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 ([11], 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 (2.1), 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 [6]. 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) the consistency in variable selection and property-(2) the asymptotic normality.

The Ridge penalty is defined as

$$\text{pen}(\beta) = \sum_{j=1}^{p} \beta_j^2. \tag{2.3}$$

Ridge Regression minimizes RSS subject to a bound on the ℓ_2 norm of the coefficients. It projects $\mathbf{y}$ onto these 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 the sparse representation 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 [11].

Lasso is a regularization technique for simultaneous estimation and variable selection [25]. The Lasso penalty is defined as

$$\text{pen}(\beta) = \sum_{j=1}^{P} |\beta_j|. \tag{2.4}$$

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 ([29], [31], [22], [26]) 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 [29] 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 [30]for the adaptive Lasso model. In the following, we would like to address this problem on weighted elastic net model.

We assume that

$$\frac{1}{n}\mathbf{X}^T\mathbf{X} \to \mathbf{C}, \tag{2.5}$$

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

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{1\,1} & \mathbf{C}_{1\,2} \\ \mathbf{C}_{2\,1} & \mathbf{C}_{2\,2} \end{bmatrix}, \tag{2.6}$$

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

$$\|\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathrm{sgn}(\beta_S)\|_\infty \leq 1 - \eta \tag{2.7}$$

where the inequality holds element-wise. IC is necessary and sufficient for the Lasso's model selection consistency [29].

To improve the Lasso model, the adaptive Lasso was proposed in [31] by using a weighted ℓ_1 penalty.

$$\mathrm{pen}(\beta) = \sum_{j=1}^{p} \hat{\omega_j}|\beta_j|, \tag{2.8}$$

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 [14]. More general situations for Lasso based models to be consistent were recently studied in [30].

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. In addition, Lasso fails to conduct grouped selection. That is, it tends to select one variable from a group and ignores the others. However elastic net [32], a convex combination of the lasso and ridge penalty, usually outperforms them in many situations. The EN penalty term with coefficients is defined as

$$\mathrm{pen}(\beta) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2. \tag{2.9}$$

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" [32]. 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 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 [26]. 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

$$\|\mathbf{C}_{21}(\mathbf{C}_{11} + \frac{\lambda_2}{n}I)^{-1}(\text{sgn}(\beta_S) + \frac{2\lambda_2}{\lambda_1}\beta_S)\|_\infty \leq 1 - \eta. \tag{2.10}$$

EIC is necessary and sufficient for the EN model selection consistency [26]. The model selection consistency of the EN model for $p \gg n$ case and the relationship of IC and EIC was discussed in [16]. 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 [31].

$$\hat{\beta} = (1 + \frac{\lambda_2}{n})\{\text{argmin}_\beta\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_2\|\beta\|_2^2 + \lambda_1^*\Sigma_{j=1}^p\hat{\omega}_j|\beta_j|\} \tag{2.11}$$

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

The so-called Bridge penalty is defined as

$$\text{pen}(\beta) = \sum_{j=1}^p |\beta_j|^\gamma, \gamma > 0. \tag{2.12}$$

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 (2.2) can be exactly zero if λ is sufficiently large [17]. 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 [14].

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 the biomarker selection from the 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 (2.11), 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.

3. 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 a very recent work [27], 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:

$$\text{argmin}_\beta \frac{1}{2}\|\mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j\beta_j\|_2^2 + n\lambda_1 \sum_{j=1}^{p} w_j|\beta_j| + \frac{n}{2}\lambda_2 \sum_{j=1}^{p} |w_j\beta_j|^2, \tag{3.1}$$

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

$$\text{argmin}_\beta \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + n\lambda_1\|\mathbf{W}\beta\|_1 + \frac{n}{2}\lambda_2\|\mathbf{W}\beta\|_2^2. \tag{3.2}$$

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 (3.1), 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 (3.1), 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

$$\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\hat{\beta}) - n\lambda_2 w_j^2\hat{\beta}_j = \lambda_1 n w_j \text{sgn}(\beta_j^*), \quad \text{if } \hat{\beta}_j \neq 0; \tag{3.3}$$

$$|\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\hat{\beta})| \leq \lambda_1 n w_j, \text{ otherwise}, \tag{3.4}$$

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

$$Z = \{\mathbf{z} \in \mathbb{R}^p; z_j = b_j \text{ for } \hat{\beta}_j \neq 0, \text{ and } |z_j| \leq w_j, \text{otherwise}\}. \tag{3.5}$$

Then, conditions (3.3) and (3.4) are equivalent to saying that there exists a subgradient vector $\mathbf{g} \in Z$ such that its components $g_j, j = 1, \cdots, p$, satisfy

$$-\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\hat{\beta}) + n\lambda_2 w_j^2\hat{\beta}_j + n\lambda_1 g_j = 0. \tag{3.6}$$

Substituting $\mathbf{y} = \mathbf{X}\beta^* + \epsilon$ in (3.6), 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:

$$\mathbf{C}(\hat{\beta} - \beta^*) - \frac{1}{n}\mathbf{x}_j^T\epsilon + \lambda_2 w_j^2\hat{\beta}_j + \lambda_1 g_j = 0. \tag{3.7}$$

Then, we see that for given $\mathbf{X}$, β^*, and $\lambda_1 > 0, \lambda_2 > 0$, $\mathrm{sgn}(\hat{\beta}) = \mathrm{sgn}(\beta^*)$ 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 (3.7) holds and
(ii) $\mathrm{sgn}(\hat{\beta_S}) = \mathrm{sgn}(\beta_S^*)$ 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 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*

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

if and only if the following conditions hold:

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

and

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

Proof. Recall that $\mathbf{y} = \mathbf{X}\beta^* + \epsilon$, $\mathbf{W} = \mathrm{diag}[w_1, \cdots, w_p]$, and $\mathbf{b} = \mathbf{W}_S\mathrm{sgn}(\beta_S^*)$. Substituting $\hat{\beta}_{S^c} = \beta_{S^c}^* = 0$ and $\mathbf{g}_S = \mathbf{b}$ in (3.7), we obtain

$$\mathbf{C}_{21}(\hat{\beta_S} - \beta^*) - \frac{\mathbf{X}_{S_c}^T\epsilon}{n} = -\lambda_1\mathbf{g}_{S^c}, \tag{3.8}$$

$$\mathbf{C}_{11}(\hat{\beta_S} - \beta^*) - \frac{\mathbf{X}_S^T\epsilon}{n} + \lambda_2\mathbf{W}^2\hat{\beta_S} = -\lambda_1\mathbf{g}_S = -\lambda_1\mathbf{b}, \tag{3.9}$$

and also

$$\mathrm{sgn}(\hat{\beta_S}) = \mathrm{sgn}(\beta^*) \text{ and } \hat{\beta_{S^c}} = \beta_{S^c}^* = 0. \tag{3.10}$$

From (3.8) and (3.9), solving for $\hat{\beta_S}$ and $\mathbf{g}_{S^c}$, we obtain

$$\hat{\beta_S} = (\mathbf{C}_{11} + \lambda_2\mathbf{W}^2)^{-1}(\mathbf{C}_{11}\beta_S^* + \frac{\mathbf{X}_S^T\epsilon}{n} - \lambda_1\mathbf{b}), \tag{3.11}$$

and

$$-\lambda_1\mathbf{g}_{S^c} = \mathbf{C}_{21}[(\mathbf{C}_{11} + \lambda_2\mathbf{W}^2)^{-1}(\mathbf{C}_{11}\beta_S^* + \frac{\mathbf{X}_S^T\epsilon}{n} - \lambda_1\mathbf{b}) - \beta_S^*] - \frac{\mathbf{X}_{S_c}^T\epsilon}{n}. \tag{3.12}$$

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

$$|\mathbf{x}_j^T\mathbf{X}_S[(\mathbf{C}_{11} + \lambda_2\mathbf{W}^2)^{-1}(\mathbf{C}_{11}\beta_S^* + \frac{\mathbf{X}_S^T\epsilon}{n} - \lambda_1\mathbf{b}) - \beta_S^*] - \frac{\mathbf{X}_j^T\epsilon}{n}| = |-\lambda_1\mathbf{g}_j| \leq \lambda_1 w_j, \tag{3.13}$$

and

$$\text{sgn}(\hat{\beta}_S) = \text{sgn}((\mathbf{C}_{11} + \lambda_2 \mathbf{W}^2)^{-1}(\mathbf{C}_{11}\beta_S^* + \frac{\mathbf{X}_S^T \epsilon}{n} - \lambda_1 \mathbf{b})) = \text{sgn}(\beta_S^*). \tag{3.14}$$

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

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

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

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

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 (3.12) and (3.11) are satisfied. This completes the proof of the lemma.

To state and prove the main theorem of this section, we follow the notations defined in [30]. Let $\mathbf{e}_j = \mathbb{R}^s$ be the vector with one in the jth position and zero elsewhere. Then $\|\mathbf{e}_j\| = 1$. We define probability event sets $\mathcal{E}(U)$ and $\mathcal{E}(V)$ relevant to the conditions of (*) and (**) in Lemma 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 1 holds if and only if it is true for the event

$$\mathcal{E}(V) = \{V_j; j \in S^c, |V_j| \leq \lambda_1 w_j\}. \tag{3.15}$$

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 1 holds if the following event is true:

$$\mathcal{E}(U) = \{U_j; j \in S, \max_{i\in S} |U_j| \leq \beta_{\min}\}, \tag{3.16}$$

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 on inequalities on 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

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

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

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

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

Therefore, we have shown that $j \in \mathcal{T}$ implies $j \in \mathcal{E}(U)$. Hence $P[\mathcal{E}(U)^c] \leq P[\mathcal{T}^c] \leq 1/p^2$ according to Lemma 9.1 in [30]. Thus, the event $\mathcal{E}(U)$ in (3.16) holds on the set $\mathcal{T}$.

V_j in (3.15) is a function of ϵ, thus, a random variable. Its expected value

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

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 $\|\mathbf{I}-\mathbf{P}\|_2 \le 1$. Then, by using the tail bound for a Gaussian random variable, the probability value

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

with

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

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

We then obtain

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

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

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

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 this paper.

Theorem 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*

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

Assume

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

where $\beta_{\max}^* = \max |\beta_i^*|$*. Then for the* $\hat{\beta}$ *in (3.1), the probability*

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

Proof. According to Lemma 1, $sgn(\hat{\beta}) = \operatorname{sgn}(\beta^*)$ if and only if conditions (*) and (**) hold. On the other hand, under the assumptions of this theorem, the condition (*) in Lemma 1 holds if the event $\mathcal{E}(V)$ is true and the condition (*) holds if the event $\mathcal{E}(U)$ is true. Therefore,

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

This completes the proof.

4. Algorithm and Experimental Results

In this section, we develop an algorithm for the WEN method based on the algorithm LARS [5]. It turns out that the minimizing problem in WEN model (4.1) could be transformed into an equivalent weighted Lasso-type optimization problem (4.2) on augmented data and then it can be even further equivalent to a Lasso-type optimization problem (4.3). This fact implies that WEN also enjoys the computational advantage of the Lasso. Experimental IMS data analysis results show that WEN model not only reduces the number of side features but also helps new biomarkers discovery.

4.1. WEN Algorithm

Recall the WEN model (3.1), with a scaled coefficient difference, we can rewrite it as:

$$f(\lambda_1, \lambda_2, \omega, \beta) = \|\mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \beta_j\|_2^2 + \lambda_1 \sum_{j=1}^{p} w_j |\beta_j| + \lambda_2 \sum_{j=1}^{p} |w_j \beta_j|^2, \tag{4.1}$$

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

$$\begin{aligned} f(\lambda_1, \lambda_2, \omega, \beta) &= \|\begin{pmatrix}\mathbf{y}\\ 0\end{pmatrix} - \begin{pmatrix}\mathbf{X}\\ \sqrt{\lambda_2}\mathbf{W}\end{pmatrix} \frac{1}{\sqrt{1+\lambda_2}} \sqrt{1+\lambda_2}\,\beta\|_2^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}} w_j \sqrt{1+\lambda_2}|\beta_j| \\ &= \|\mathbf{y}^* - \sum \mathbf{x}_j^* \beta_j^*\|_2^2 + \gamma \Sigma_j w_j |\beta_j^*| \\ &= \|\mathbf{y}^* - \sum \frac{\mathbf{x}_j^*}{w_j} \beta_j^* w_j\|_2^2 + \gamma \Sigma w_j |\beta_j^*| \\ &= g(\gamma, \mathbf{W}, \beta) \end{aligned} \tag{4.2}$$

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

$$g(\gamma, \mathbf{W}, \beta) = |\mathbf{y}^* - \sum \mathbf{x}_j^{**} \beta_j^{**}|^2 + \gamma \Sigma |\beta_j^{**}|. \tag{4.3}$$

For a fixed λ_2, the weighted EN optimization problem is equivalent to a weighted lasso problem on an augmented data set and, further, it can be 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 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 ([11], chapter 7). In our algorithm, the other tuning parameter λ_1 or say step k is selected by tenfold CV. The pseudo code for LARS-WEN is listed 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.

4.2. Experimental Results

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 analysis includes a comparison of results by applying the EN method, WEN method, s well as the results obtained using the commercial software SAM and other popular methods and software programs used in mass spectrometry community.

The IMS data set is on $GL26$ glioma study. $C57$ 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 μm) 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 300μm by 300μm. After data acquisition, the data underwent a series of basic preprocessing steps to reduce the experimental variance between spectra through the removal of background, 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) [3] to generate the SAM feature list in Table 1. In comparison, WEN processes IMS data underwent only basic preprocessing steps with no peak binning beforehand and saves significant amount of time for data processing.

The WEN model proposed here is for pixel-level classification. For the data entering, cancer pixels and non-cancer pixels are selected from the mouse brain IMS data sets 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 [32]. We include all

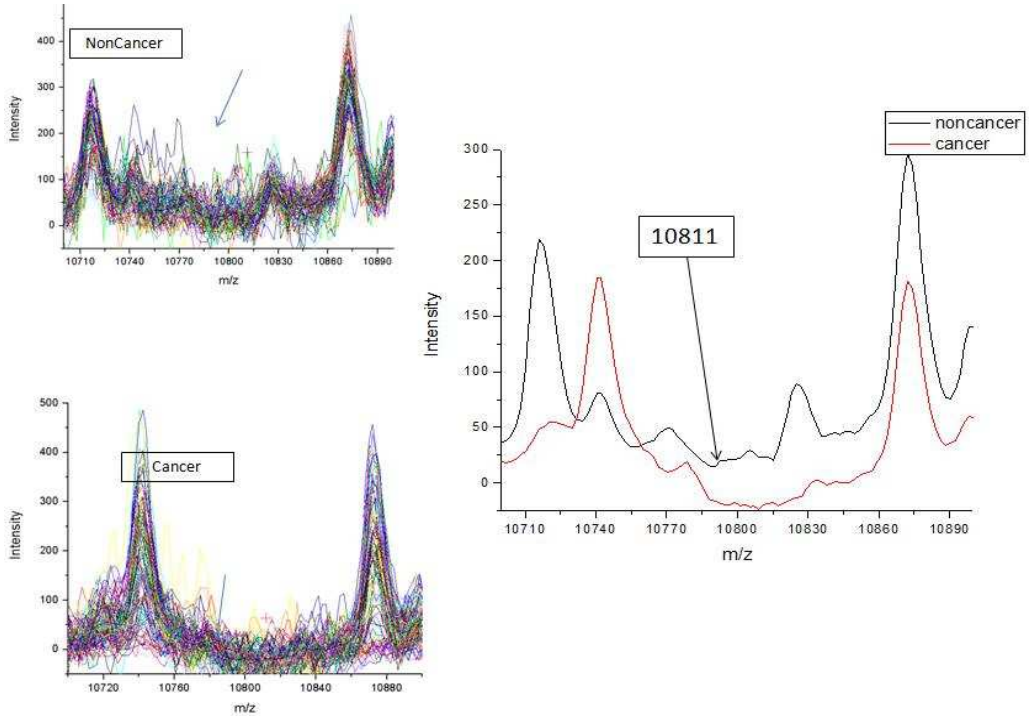

Figure 2: Side peak ($m/z = 10811$). The peak at $m/z = 10811$ is a fake peak caused by noise.

peak m/z values before binning as predicator in (3.1), and $\mathbf{y}$ takes negative one for a noncancer pixel and one otherwise.

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 2 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, we found that most m/z values in the WEN list have a decreasing trend in the tumor region. 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 WEN 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 region.

Interestingly, when $p \gg n$, linear classifiers often performs better than non-linear ones in many applications [11], even though non-linear methods are known to be more flexible. This fact is related to the asymptotic results in [10]: 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 [28].

In fact, protein identification provided identities of important biomarker peaks, including 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. 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 [20]. 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.

The Table 1 shows the comparison results of the classification algorithms using principal component analysis (PCA) with linear discriminant analysis (LDA) ([21], [24]), PCA with support vector machine (SVM) [8], and WEN. These algorithms are applied to section 1 IMS data to learn 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.

Table 1.

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%

Acknowledgments

The authors are grateful to valuable discussions with Anhua Lin and collaborators at the Mass Spectrometry Research Center at Vanderbilt University for IMS data study that motivated this research. The authors are also very grateful to the referee's valuable comments and suggestions, which helped to improve the writing of this paper. This work is partially supported by MTSU FRCAC grant #2-21519.

References

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

[2] P. Chaurand, M.E. Sanders, R.A. Jensen, R.M. Caprioli. *Profiling and imaging proteins in tissue sections by MS*. Anal. Chem., 76 (2004), 86A-93A.

[3] 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/]

[4] E. Candes, T. Tao. *The dantzig selector: statistical estimation when p is much larger than n.* Annals of Statistics, 35 (2007), 2313(C2351.

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

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

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

[8] M. Gerhard, S.O. Deininger, F.M. Schleif. *Statistical Classification and visualization of MALDI imaging data*. CBMS'07 2007; 0-7695-2905-4/07.

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

[10] P. Hall, J.S. Marron, A. Neeman. *Geometric representation of high dimension low sample size data*. J. R. Statist. Soc. B, 67 (2005), 427(C444.

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

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

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

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

[15] 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(C142.

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

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

[18] S. Matoba, J.G. Kang, W.D. Patino, A. Wragg, M. Boehm, O. Gavrilova, P.J. Hurley, F. Bunz, P.M. Hwang. *P53 regulates mitochondrial respiration*. Science, 312 (2006), 1650-1653.

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

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

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

[22] N. Meinshausen, B. Yu. *Lasso-type recovery of sparse representations for high-dimensional data*. Annals of Statistics, 37 (2009), no. 1, 246-270.

[23] H. Meistermann, J.L. Norris, H.R. Aerni, D.S. Cornett, A. Friedlein, A.R. Erskine, A. Augustin, M.C. De Vera Mudry, S. Ruepp, L. Suter, H. Langen, R.M. Caprioli, A. Ducret. *Biomarker discovery by imaging mass spectrometry: transthyretin is a biomarker for gentamicin-induced nephrotoxicity in rat*. Mol Cell Proteomics, 5 (2006), 1876-1886.

[24] E.R. Muir, I.J. Ndiour, N.A. Le Goasduff, R.A. Moffitt, Y. Liu, M.C. Sullards, A.H. Merrill, Y. Chen, M.D. Wang. *Multivariate analysis of imaging mass spectrometry data*. BIBE 2007 proceedings of the 7th IEEE international conference 472-479.

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

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

[27] F. Zhang, D. Hong, S. Frappier, D.S. Cornett, R.M. Caprioli. *Elastic Net Based Framework for Imaging Mass Spectrometry Data Biomarker Selection and Classification*. Manuscript, 2009.

[28] H. Zhang, J. Ahn, X. Lin, C. Park. *Gene selection using support vector machines with non-convex penalty*. Bioinformatics, 22 (2006), 88-95.

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

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

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

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

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