International Journal of Mathematics and
Computer Science, 5(2010), no. 2, 123-140

Mathematical Tools and Statistical Techniques for Proteomic Data Mining

Don Hong[1,2], Shi Yin Qin[3], and Fengqing (Zoe) Zhang[4]

[1] Department of Mathematical Sciences
Program of Computational Sciences
Middle Tennessee State University
Murfreesboro, TN, TN 37132, USA

[2] College of Science
Ningbo University
Ningbo, Zhejiang, China

[3] School of Automation Science and Electrical Engineering
Beihang University
Beijing, China

[4] Department of Statistics
Northwestern University
Evanston, IL, 60208, USA

e-mail: dhong@mtsu.edu, qsy@buaa.edu.cn, FengqingZhang2015@u.northwestern.edu

(Received June 30, 2010, Revised Aug. 15, 2010, Accepted Nov. 1, 2010)

Abstract

Proteomics is the study of and the search for information about proteins. The development of mass spectrometry (MS) such as matrix-assisted laser desorption ionization (MALDI) time-of-flight (TOF) MS and imaging mass spectrometry (IMS), greatly speeds up proteomics studies. At the same time, the MS and IMS applications in medical science give rise to many challenges in mathematics and statistics regarding to the MS and IMS data analysis including data preprocessing, classification, and biomarker discovery. In this paper, we give a review of recent development of mathematical techniques and statistical tools for MS and IMS based proteomic data mining including wavelet based MS data preprocessing and multivariate statistical methods for IMS data classification and biomarker discovery.

Key words and phrases: biomarker discovery, classification, data analysis, elastic net, mass spectrometry, image MS, proteomics, wavelet denoising

AMS subject classification: 62-07, 62P10, 65T50, 92C55

ISSN 1814-0424

1. Introduction

The widespread adoption of matrix-assisted laser desorption ionization (MALDI) time-of-flight (TOF) MS for protein identification in proteomics studies is driven by the sensitivity, unlimited mass range capabilities, versatility and intact protein analysis (see [1], [11], [30] for example). One promising area of MS based proteomics is that the information hidden in the noisy mass spectral data can help people to detect cancer even in early stage. MS has already been widely used to find disease related proteomic patterns in complex mixtures of proteins. These new techniques have made proteomics possible, especially when involving large molecules. Indeed, the Nobel prize in chemistry in 2002 recognized MALDI's ability to analyze intact biological macromolecules. MALDI IMS has emerged as a powerful technique for analyzing the spatial distribution of proteins directly in tissue specimens. IMS as a platform has shown great potential and is very promising for rapid mapping of protein localization and the detection of sizeable differences in protein expression (see [11], [28], [31], and references therein). However, the complexity and high dimensionality of the MS and IMS data pose great challenges for data processing.

Usually, a raw MS spectrum consists of three components: true peaks, baseline, and noise. Disentangling these three components is a complex task. Concerning to IMS, the data processing becomes even more difficult. IMS data has two spatial dimensions ($x-$ and $y-$ dimensions) and the mass-over-charge (m/z) dimension. Each MALDI IMS data set is multidimensional and has hundreds of pixels. Each pixel is associated with a complete mass spectrum. This contrasts with regular images where for each pixel there is a set of RGB values. Each mass spectrum contains mass-to-charge (m/z) values ranging from $2k$ to $70k$ Daltons and ion intensity values which are associated with each pixel. There are hundreds of mass spectra represented in a single MS image. 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 distinction that should be made is determining if the m/z values selected as potential biomarkers are caused by the biological structure of the tissues or by the disease state being investigated.

Generally, the mathematical processing of MS signals can be roughly divided into two steps. First, in the preprocessing step, we attempt to recover true signals from the raw data, as accurately as possible. This step includes calibration, denoising, baseline correction, data alignment cross samples, and peak selection. The second type processing is to involve operations such as dimension reduction, feature selection, clustering, and pattern recognition for classification. Preprocessing is of great importance and can improve the performance of classifiers to separate cancer and non-cancer samples. It has been studied that ineffective or inadequate algorithms in preprocessing will introduce substantial biases and thus prevent to extract valuable biomarkers from raw data.

Wavelet is an important method in signal processing and has very broad applications in image processing and statistical data analysis. The characteristics of wavelet analysis such as localized representation, orthogonal decomposition and multi-resolution analysis (MRA)

guarantee that wavelets are very suitable for MS data analysis ([6], [7], [9], [12], [13], [18], [19] for example). Wavelet could reveal more information than other conventional methods. Combing the zoom-in and pan-out properties, wavelets, as building blocks of models, are well localized in both time and frequency scale. The MRA enables us to analyze the signal in different frequency bands and thus enables us to observe any transient in time domain as well as in frequency domain. Furthermore, wavelet is very useful in analyzing data with gradual frequency changes. In addition, wavelets select widths of time slices according to the local frequency in the signal. This adaptive property of wavelets certainly can help us to determine the location of peak differences of MS protein expressions between cancer and non-cancer samples.

In this paper, we present an overview of wavelet applications in the MALDI MS and IMS proteomics data processing and multivariate statistical tools for IMS data biomarker discovery. We explore the characteristics of MALDI MS and IMS data as well as wavelet application in this area, both theoretically and practically. The application includes not only in preprocessing steps but also in the feature selection. We provide some guidelines to algorithm development and parameter selection in different data processing stages of both the conventional MALDI MS data and IMS data. After analyzing the MS and IMS data from the view of wavelet transform, we suggest integrating all MS and IMS data processing steps by using wavelet transform. Biomarker selection from IMS data is a problem of global optimization. A recently developed regularization and variable selection method, elastic-net (EN), produces a sparse model with admirable prediction accuracy and can be an effective tool for IMS data processing. Very recently, we have incorporated a spatial penalty term into the EN model and developed a new tool for IMS data biomarker selection and classification ([20], [35]). The remainder of the paper is organized as follows. The characteristics of MS and IMS based proteomics data and wavelet applications in this area are discussed in the next section. In Section 3, the preprocessing steps and wavelet applications are presented in detail. Wavelet-based procedure for feature selection and classification algorithms are also discussed this section. Newly developed elastic net based biomarker discovery tools used for IMS data are discussed in Section 4.

2. Characteristics of MALDI MS and IMS Data and Wavelet Applications

In this section, we describe in detail the characteristics of MALDI TOF MS and IMS data and apply wavelets in the data (pre)processing.

2.1. Mathematical Model for MALDI MS and IMS Proteomics Data

A commonly used model ([6], [9], [18], [25]) for MS data analysis is that each raw spectrum data can be represented in three parts: $f(t) = B(t) + N * S(t) + \epsilon(t)$, where $f(t)$ is the observed signal. $B(t)$ stands for baseline, a systematic artifact commonly seen in mass

spectrometry data. $S(t)$ is the true signal, which consists of a sum of possibly overlapping peaks, each corresponding to a particular biological molecule such as a protein or a peptide. N is the normalization factor, a constant multiplicative factor to adjust for spectrum specific variability. $\epsilon(t)$ stands for the noise function. In general, the preprocessing steps include calibration, denoising, baseline correction, normalization, peak alignment, and peak detection and quantification. The difficulty in processing MS data stems from the mixture of true peaks, baseline and noise. Separating these three components from each other is very complex.

Concerning the MADLI IMS data, many of the characteristics in mass spectrum preprocessing stage are the same as those for the MS data. However, the main difference is that the IMS data has two spatial dimensions (both x- and y- dimensions) plus a mass-over-charge ($m/$z) dimension. The combination of spatial and mass resolution results in large and complex data sets, which gives a great challenge to the quantitative analysis and interpretation tools. Figure 1 shows a mouse brain IMS data set, and the darker region in Figure 1(a) indicates the presence of the tumor. The grid in Figure 1(a) forms a matrix of points of the sample surface. Individual mass spectra are acquired for every point (pixel) of the sample surface and stored digitally. Behind each pixel, it is an entire mass spectrum with a very large range of m/z values. Figure 1(b) displays three mass spectra corresponding to three different pixels of the mouse brain tissue section. Specially designed software enables the election of an analyte signal (m/z value) from the mass range and plots the intensity of the signal for each individual point in a matrix. If the signal intensity is plotted by a color scale, the matrix can be represented by an ion image of the analyte distribution, which is shown in Figure 1(c) and (d). Ion images can show the spatial spread of a particular peak's intensity over the tissue and the mass spectral peak represented by the amount of a particular ion that was measured. From Figure 1(c), we can see that the m/z value 5442.704 is differentially expressed between the cancer region and the normal region, which maybe due to the latent biological function of this m/z value and its effect to cancer growth. Figure 1(c) and Figure 1(d) have clearly different intensity distributions over the mouse brain. If one has already known a particular m/z value having biological meaning and plans to know the spatial distribution of a particular molecule, ion image is very informative. However, a more important application should be the determination of unknown variants for metabolite and protein profiling in both clinical and disease studies. Noticing that we have huge number of ion images per data set, it is necessary to have statistical models to do biomaker selection instead of doing visually checking. In addition, these conventional images, derived from a specific analyte mass do not identify the spatially localized correlations between analytes that are latent in IMS data processing. Although it is difficult to make full utilization of both spatial information and spectrum information in IMS data, it is very necessary. For IMS data processing, multi-scale representation and global analysis of spatial and protein information of the biological samples will be the key for data processing. Therefore, the multivariate data analysis methods can be applied in IMS data for identifying both spatial and mass trends and merit further investigation.

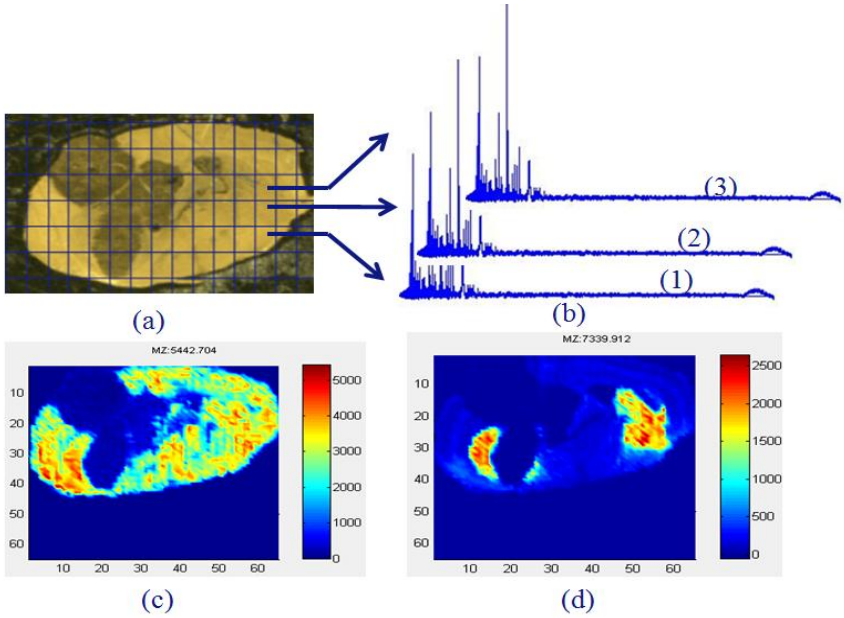

Figure 1: Mouse brain IMS Data. (a) photomicrograph of a mouse brain tissue section, implanted with a GL26 glioma cell line and tumor growth; (b) three mass spectra from three different pixels; (c) the ion image of $m/z = 5442.704$; and (d) the ion image of $m/z = 7339.912$.

2.2. Wavelets for MALDI MS and IMS Data

To extract the true signal from MS/IMS data, we need to remove the noise and the incoherent signal from the observed data. Wavelet analysis serves as an efficient mathematical tool that can be utilized to extract or encode the feature signal. Wavelet theory represents data signals by breaking them down into many interrelated component pieces, similar to the pieces of a jig-saw puzzle. When pieces are scaled and translated by wavelets, the breaking down process is called a wavelet decomposition or wavelet transform. A discrete wavelet transform (DWT) decomposes a signal into several vectors of wavelet coefficients. Different coefficient vectors contain information about the signal function at different scales. Coefficients at coarse scale capture gross and global features of the signal while coefficients at fine scale contain detailed information. Applying wavelet transform to MALDI-TOF MS data, the protein expression difference can be measured at different resolution scales based on a molecular weight-scale analysis. It may reveal more information than other conventional methods. Wavelets, as building blocks of models, are well localized in both time and scale (frequency). In wavelet analysis, a function is approximated by a weighted sum over the scaled and translated mother wavelets. Each weighted wavelet acts as a building block, and when all the blocks are summed together, an approximation is found. Wavelet is very useful in analyzing data with gradual frequency changes. Signals with rapid local changes (signals with discontinuities, cusps, sharp spikes, etc) can be precisely represented with just a few wavelet coefficients.

The wavelet approximation to a signal function f is built up over multiple scales and many localized positions. The fundamental concept involved in multi-resolution analysis (MRA) is to find the average features and the details of the signal via scalar products with scaling signals and wavelets. The MRA enables us to analyze the signal in different frequency bands and thus enables us to observe any transient in time domain as well as in frequency domain. The high frequency band output is viewed as the wavelet transform coefficients for a fine scale and the low frequency band output is decimated by a factor of 2. This low frequency band is then split into a high and low band again. The band splitting and decimation process continues and produces an octave band representation of the signal. The high pass filter output wavelet coefficients represent the signal's characteristics and energy at a particular scale. The output of the final low pass filter is the residual namely the most coarse signal.

Since true signal $S(x)$, the baseline $B(x)$ and the machine noise $\epsilon(x)$ have different time-frequency attributes, it is then possible to separate them in wavelet coefficients. In the wavelet representation, the noise $\epsilon(x)$ is concentrated in the fine scale wavelet coefficients and the incoherent signal can be approximated by the projection onto the coarse space. In contrast to Fourier transforms, wavelets select widths of time slices according to the local frequency in the signal. This adaptive property of wavelets certainly can help us to determine the location and intensity (peak) difference(s) of MALDI-TOF MS protein expressions between cancerous and normal tissues in term of molecular weights. Most peak signals can be represented by a small number of wavelet coefficients while white noise is distributed equally over all wavelet coefficients. However the separation of noise and peaks is not so straightforward since they both have fast changing parts. A variety of threshold strategies can be used to remove the machine noise from the data including feedback strategies from MS data information[7]. But inappropriate thresholding may cause peak attenuation. Baseline correction is an important step in MS data preprocessing. Through wavelet transforms, baseline $B(x)$ can be considered as a coarse approximation and a component with small coefficients in a wavelet space[12]. As for the peak selection step, peaks of MS data can be viewed as the singularities of the MS output signal. Singularities of a signal can be represented by the modulus maxima of their wavelet transforms (see [15], [27], [34], for example). Concerning the 3D IMS data, wavelet is also suitable for data preprocessing as well as feature selection.

3. Wavelet Applications for MALDI MS Data Preprocessing

In this section, we discuss in detail the wavelet applications for MALDI MS data preprocessing with an emphasis on denoising, baseline correction, and feature (peak) selection. Figure 2 illustrates the effects of the preprocessing steps visually. Recall that each raw spectra data can be represented in three parts: $f(t) = B(t) + N * S(t) + \epsilon(t)$. Figure 2 displays the MALDI MS data preprocessing framework in terms of denoising, baseline correction and peak selection.

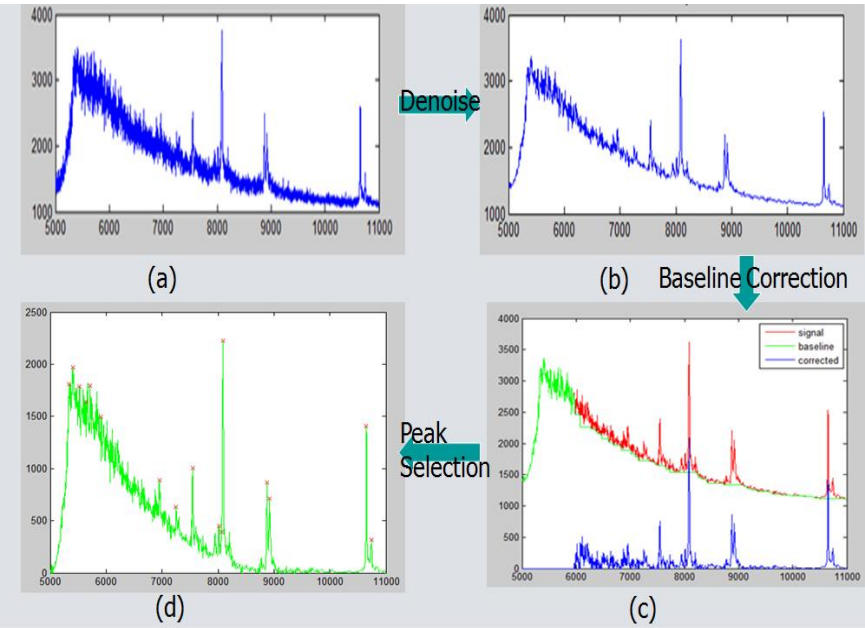

Figure 2: MALDI MS data preprocessing framework. (a) a raw mass spectrum; (b) the denoised mass spectrum; (c) the mass spectrum after baseline correction; and (d) the mass spectrum marked with selected peaks.

3.1. Denoising

Yasui et al.[33] and Coombes et al.[9] independently proposed the application of wavelet transformation in proteomics. Recent applications of wavelets for MS data processing can be found, for example in [6], [7], [12], [18], and [22].

The general wavelet denoising procedure is as follows:

1. Apply wavelet transform to the noisy signal to produce the noisy wavelet coefficients to the level which we can properly distinguish the peaks.

2. Select an appropriate threshold limit at each level and a threshold method (hard or soft thresholding) to best remove the noises.

3. Inverse wavelet transform of the thresholded wavelet coefficients to obtain a denoised signal.

In denoising step, the most important issues are the selection of a suitable wavelet, the decomposition levels, and the coefficients of the denoising threshold. Selecting the mother wavelet is of great importance for wavelet transform. There is no strict rule for selection. However, the analysis becomes more precise if the wavelet shape is adapted to the signal. DWT is sufficient for exact reconstruction, and the discrete forms are necessary for most computer implementations. Du at al.[13] proposed an improved DWT smoothing algorithm which utilizes the cross-level DWT coefficients information during smoothing. It estimates the noise distribution based on the first DWT decomposition level, and then infers the thresh-

old at other levels. In order to reduce the peak attenuation in smoothing, the related DWT coefficients of the detected peak-related DWT coefficient are also used for reconstruction. Coombes et al[9] proposed to use undecimated discrete wavelet transform (UDWT). They suggested when using the DWT for denoising, it tends to create significant artifacts near the ends of the signal. With the DWT, there is usually a trade-off between the smoothness of the denoised signal and its squared-error performance ([8], [23]). While DWT denoising effect will change drastically if the starting position of the signal is shifted, UDWT is shift-invariant. Comparing with DWT, it is reported that UDWT gives better visual and qualitative denoising, with a small added cost in computational complexity[9].

For the choice of a decomposition level, the maximum level to apply the wavelet transform depends on how many data points are contained in a data set, since there is a down-sampling by 2 operation from one level to the next one. One factor that affects the noise removal results is the signal-to-noise ratio (SNR) in the original signal. Fewer levels of wavelet transform are needed to remove most of the noise if the signals have higher SNR. The signals with lower SNR should be decomposed by relatively more levels of wavelet transform.

For the denoising threshold, different kinds of methods have been tried to extract and preserve the desired signals as much and accurately as possible. The commonly used universal threshold is computed as $\lambda = \sigma\sqrt{2\log N}$, where σ stands for the estimation of the variation of the coefficients on the standard deviation scale. N represents the number of data points (wavelet coefficients). However, local threshold performs better than universal threshold when applied to MS data. The noise distribution of MS data over the m/z value is heterogeneous[13]. The most important issue in a denoising procedure is the wavelet coefficient threshold selection. In [9], UDWT is used to denoise the spectra and Daubechies wavelet of degree 8 is applied. The denoising procedure starts by transforming observed signal from time domain to the wavelet domain, then computes the median absolute deviation (MAD) of the wavelet coefficients, sets coefficients to zero with a hard thresholding (some threshold expressed as a multiple of 0.67 MAD) and finally transforms the signal back to the time domain. However, the denoising parameters are usually chosen by experience. Chen et al [7] introduced feed-back concepts to the MS data denoising in order to target the optimal parameters setup as objectively as possible. It is reported in the paper that the elevated baseline and the height of such baseline can be also associated with the proportion of falsely detected peaks. Thus an adaptive threshold selection algorithm is suggested by utilizing the proportion of the baseline as an index to adjust the thresholds. There are many other useful and practical principles for adaptive threshold selection such as Stein's unbiased risk estimate (SURE), minimal mean squared error, generalized cross validation and etc.. In addition, Du et al.[13] proposed to utilize the cross-level DWT coefficients information for denoising. It estimates the noise distribution based on the first DWT decomposition level, and then infers the threshold at other levels. There is an approximate linear relationship of the noise component distribution at different levels. It is known that wavelet transform modulus maxima of signal and noise have different transmission properties across different scales. The method of utilizing wavelet coefficients relativity to tell desired signal from noise is called spatially selective noise filtration. This method can provide more steady denoising

results but requires higher computational cost. In addition, Li et al [24] proposed Bayesian wavelet shrinkage and thresholding estimators which outperform the classical data adaptive wavelet thresholding estimators in terms of mean squared error with finite samples. It is proposed to use block threshold strategy in [18] because the high frequency components decrease as the mass weight increases. Block threshold is to threshold the wavelet coefficients in groups (blocks) rather than individually to increase estimation accuracy by utilizing information about neighboring coefficients.

Two rules are generally used for thresholding the wavelet coefficients (soft/hard thresholding). Hard thresholding sets zeros for all wavelet coefficients whose absolute value is less than the specified threshold limit. Generally hard thresholding provides an improved signal to noise ratio but the reconstructed signal may have additional oscillation. Soft thresholding only reduces these wavelet coefficients less than the specified threshold limit instead of setting them to be zeros. Soft thresholding can preserve the smoothness but not as good as hard thresholding in the sense of mean squared error. However the selection of hard or soft thresholding should refer to the principles for thresholding selection. With an appropriate threshold, noise can be removed without biasing the signal too much, since the wavelet coefficients greater than the particular threshold level still remain unaltered. However, if the threshold is too large, then the signal will be altered. If the threshold is too small, then the level of denoising is not enough. This causes a crucial problem of peak attenuation. An ideal transform can project signal to a domain where the signal energy is concentrated in a small number of coefficients. On the other hand, if the noise is evenly distributed across this domain, this domain will be a good place to do denoising, due to the fact that the SNR is significantly increased in some important coefficients, or we can say the signal is highlighted in this domain while the noise is not. However, the signal peaks also have fast changing components just like noise and will be attenuated by wavelet transform as well. All these algorithms discussed above aim to extract noise from desired signal as accurately as possible.

3.2. Baseline Correction

Another important issue that needs to be addressed in MS data preprocessing is the baseline correction. Baselines of different spectra can have large variation. Usually baseline is viewed as a very low frequency component of the observed signal. The region of the spectrum below 950 m/z is typically dominated by noise from the matrix molecules and may also contain extensive areas of saturation where the number of ions hitting the detector exceeds its ability to count them. Denoising also plays a critical role in baseline correction. Without it, the extremes of the noise (on the low end) will tend to drive the estimated baseline below the actual baseline, and the baseline-corrected spectra will tend to drift upward to the right [2].

Baselines are corrected by fitting a monotone local minimum curve to the denoised spectra in [9]. However, wavelet transform can also serve as an effective tool in baseline correction. The discrete wavelet transform decomposes the MS signal into an approximate component and several detail components. It has been mentioned in [12] that the approximation component at a certain higher level of DWT is a good estimation of the smoothly decaying baseline.

But later, they point out that the baseline removal does not perform well when large peak regions exist in the spectrum. However this does not mean baseline correction in wavelet domain is not useful. If the spectrum has large peak regions, the approximate component at a certain higher level of DWT still contains information of peaks. Thus the method to remove the approximate component as baseline will, of course, fail in this situation. It will badly affect the peak selection. If we perform baseline correction by fitting a monotone local minimum curve in the approximate component, it would be more effective and reasonable. In addition, the empirical mode decomposition (EMD) introduced by Huang et al [21] could also serve as a useful method in MS data baseline correction. EMD is adaptive and therefore highly efficient method for analyzing nonlinear and non-stationary data.

3.3. Feature Peak Selection

One of the critical problems in the MS data analysis is to select meaningful feature peaks. Therefore, the final goal of the MS data preprocessing is to identify the locations and intensities of important peaks which can be used for biomarker discovery. Actually, the peak selection procedure can be considered as two parts: peak detection that is to find the m/z values of peaks and peak quantification that is to quantify the intensities of peaks. Peak selection procedure can reduce data dimension significantly. Most of the current methods use the height of the local maximum to quantify the peak within estimated boundaries. Peaks are selected in [9] using all local maxima of a spectrum after denoising, baseline correction, and normalization. The height of the peak is used to quantify peaks. For this purpose, a local maximum is defined as a point where the intensities change from increasing to decreasing (allowing for flat plateaus when the tops of peaks are more than one clock tick in width). The signal-to-noise ratio (SNR) of a peak is estimated as the height above baseline divided by a median-smoothed version of the wavelet-defined noise. Also the local maximum points need to be larger than certain intensity threshold and SNR threshold in order to be considered as peaks. However, point measurement may be subject to high variation form various resources. Also, the height may not be a good index of the total amount of ions for a specific feature.

As we know, high amplitudes do not always guarantee real peaks: some sources of noise can result in high amplitude spikes. Conversely, low amplitude peaks can still be real. Measuring a small region or bounded neighborhood around that peak would be more robust and informative. It is pointed in [12] that the estimated peak strength is proportional to the area under the curve (AUC) of the peak in simple situations. Therefore, they suggested to study AUC estimation in spectra with multiple overlapping peaks for possibility to improve peak detection. A new peak selection algorithm for MS data analysis is proposed in [23] by using asymmetric Lorentzian and Sech2 functions to fit peak shape. A Bayesian wavelet-based functional mixed model is used to represent mass spectra as functions in [29]. This flexible framework in modeling nonparametric fixed and random effect functions enables it to simultaneously model the effects of multiple factors. From the model output, they identify spectral regions that are differentially expressed across experimental conditions, while

controlling the Bayesian FDR, in a way that takes both statistical and clinical significance into account.

All these peak selection algorithms discussed above require denoising, baseline correction, and normalization beforehand. A peak detection algorithm called MassSpecWavelet, by applying CWT-based pattern matching and wavelet transform modulus maxima, is introduced in [12]. This method can be directly applied to the raw data and requires no baseline removal or peak smoothing preprocessing steps before this peak detection. As mentioned earlier, peaks of MS data can be viewed as the singularities of the MS output signal. Singularities of a signal can be represented by the modulus maxima of their wavelet transforms ([27], [34], [15]). For the CWT, a symmetric Mexican Hat wavelet is used in [12]. The Mexican Hat wavelet is proportional to the second derivative of the Gaussian probability density function. The symmetric property of Mexican Hat wavelet can help to remove the baseline component after continuous wavelet transform. In the peak identification process, instead of using Lipschitz exponent to check the singularity, the SNR is used. However, as we know Lipschitz exponent is a very reliable and popular way to measure the singularity of signal. Thus by taking advantage of Lipschitz exponent, this peak identification algorithm may be further improved.

The peak quantification algorithms can be measured in terms of reproducibility while the evaluation of peak detection algorithms can base on sensitivity and false discovery rate. Guerra et al. [26] use ANOVA and F-tests to measure the reproducibility of peak quantification. Their results show for peak quantification, among five peak selection algorithms, MassSpecWavelet has the best performance. This wavelet-based direct peak detection method shows its advantage with sensitivities above 0.95 with a FDR of 0.1 in their experiment results. It is possible, as well as quite meaningful, to integrate all MS data preprocessing steps by wavelet transform. Another wavelet-based peak selection algorithm for MS data analysis, which is available in an open source framework OpenMS, can be found in [23]. This algorithm is a three-step technique including determining the positions of putative peaks in the wavelet-transformed signal, fitting an analytically given peak function to the data in that region, and optionally improving the resulting fit by using nonlinear optimization.

Algorithms above are mainly about peak selection in individual spectrum. In real application, however, the positions of peaks in each spectrum around the same m/z value may be slightly different from each other. Thus a processing step that determines which peaks found in individual spectrum should be identified as representing the same biochemical substance across spectra is necessary. Coombes et al. [9] started selecting the set of peaks with $S/N > 10$ first, then coalesced two peaks if they differed in location by at most 7 clock ticks or if they differed in relative mass by at most 0.3%. These parameters were determined empirically by visually checking the spectra. Then they considered the peaks with $2 < S/N < 10$, and added these to the list if they fell within the same distance limits of a previously identified peak. A new algorithm called project spectrum binning (PSB) for the cross sample peak alignment was introduced by Hong et al in [19]. Averaging is a fundamental principle underlying many statistical methods. Peak detection on average spectrum is a

direct and simple way for spectrum alignment. Using mean spectrum for peak extraction and quantification can be found in [25].

4. Classification and Biomarker Discovery

After preprocessing procedures, the MS data is ready for biological feature extraction or called biomarker discovery associated with certain diseases. Machine learning or pattern recognition methods can be applied to extract features from wavelet coefficients of the MS data after wavelet transform. In this section, we emphasize on image mass spectrometry data classification and biomarker discovery using multivariate statistical analysis. For IMS data, while heat map plots of individual ion intensities make pretty pictures, viewing the data one ion at a time is tedious and relies on the expertise and interpretation of the operator. There has been good progress in applying multivariate statistical methods such as PCA, LDA and SVM to IMS data analysis. However, these methods for IMS data processing are inadequate due to their limited use of spatial information and the advantages of IMS technology [35].

MALDI-Imaging is an emerging and very promising new technique for protein analysis from intact biological tissues [11]. 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 more time, and provides hundreds of identical independent spectra which address the measurements repeatability. 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 greatly 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.

The application of MVA methods has opened new doors for the exploration of IMS data. Very recently, two statistical models are presented in [20] and [35], respectively, for biomarker selection and classification of the high dimensional and complex IMS data. The 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, these models work efficiently and effectively for IMS data processing in terms of confirming new biomarkers, producing more precise peak list by including significant peaks and reducing the

number of side peaks, and providing more accurate classification results.

4.1. EN4IMS Model for IMS Data Processing

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. Ordinary least squares (OLS) estimates are obtained by minimizing the residual sum square (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 [17], followed by Bridge regression [16], the Garotte [4], the Lasso [32], LARS [14], and very recently the elastic net [36]. The Dantzig selector method was proposed in [5] by using sparse approximation and compressive sensing.

The newly developed variable selection method, elastic net (EN), can simultaneously perform automatic variable selection and continuous shrinkage [36]. That is, it can continuously shrink the coefficients toward zero as its regularization parameters increase; some coefficients are shrunk to exactly zero if the regularization parameters are sufficiently large. The shrinkage often improves the prediction accuracy due to the bias-variance trade-off. Thus, the EN model simultaneously achieves accuracy and sparsity. The achievement of sparsity is particularly useful when the number of variables (p) is much larger than the number of observations (n). In addition, the EN model encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together. Compared with other current commonly used analysis methods, the EN model is much more suitable for IMS data processing. In [35], a spatial penalty term is incorporated into the EN model in order to develop a new tool for IMS data biomarker selection and classification. The motivation of this new model EN4IMS is to fully utilize not only the spectral information within individual pixels, but also the spatial information for the entire IMS data cube. A software package for comprehensive IMS data processing, called IMSmining, is developed based on this new model. By incorporating the spatial penalty term, this package helps to distinguish the IMS feature peaks caused by biological structural differences from those truly associated with cancer.

The EN4IMS method has been tested on extensive simulation studies, and the algorithm has also been applied to real IMS data sets provided by Vanderbilt University Mass Spectrometry Research Center (VUMSRC). The analysis results of both simulation studies and real data examples show that the EN4IMS algorithm works efficiently and effectively for IMS data processing: producing a more precise listing of feature peaks, helping to discover new potential biomarkers, and providing more accurate classification results.

4.2. Weighted Elastic Net Model

In order to better consider the spatial information for more precise biomarker selection, a more general model called weighted elastic net (WEN), which incorporates the spatial penalty directly into the EN model equation, is developed in [20]. Theoretical properties of the WEN model such as the variable selection accuracy are discussed there as well.

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 its standard deviation at each predictor with the corresponding coefficient in the model to enforce penalty on predictors caused by structure differences.

To better consider the spatial information for more precise biomarker selection, we propose the following weighted elastic net (WEN) model [20]:

$$\mathrm{argmin}_\beta \frac{1}{2}\|\mathbf{y}-\sum_{\mathbf{j=1}}^{\mathbf{p}}\mathbf{x_j}\beta_\mathbf{j}\|_\mathbf{2}^\mathbf{2}+\mathbf{n}\lambda_\mathbf{1}\sum_{\mathbf{j=1}}^{\mathbf{p}}\mathbf{w_j}|\beta_\mathbf{j}|+\frac{\mathbf{n}}{\mathbf{2}}\lambda_\mathbf{2}\sum_{\mathbf{j=1}}^{\mathbf{p}}|\mathbf{w_j}\beta_\mathbf{j}|^\mathbf{2}, \tag{4.1}$$

where $w_j > 0$, $j = 1, \cdots, p$ are weighted penalty coefficients. Let $\mathbf{W} = \mathrm{diag}[\mathrm{w}_1, \cdots, \mathrm{w}_\mathrm{p}]$. Then the WEN model can be rewritten as

$$\mathrm{argmin}_\beta \frac{1}{2}\|\mathbf{y}-\mathbf{X}\beta\|_\mathbf{2}^\mathbf{2}+\mathbf{n}\lambda_\mathbf{1}\|\mathbf{W}\beta\|_\mathbf{1}+\frac{\mathbf{n}}{\mathbf{2}}\lambda_\mathbf{2}\|\mathbf{W}\beta\|_\mathbf{2}^\mathbf{2}. \tag{4.2}$$

The weighted elastic net model (4.1) puts the weights associated with ion intensity spreading information directly into the elastic net model and thus enforces larger penalty on the coefficients of predictors caused by structure differences. This model inherits good properties from the EN model including sparse representation, ability to deal with $p \gg n$ problem and grouping effect. In addition, compared with the 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.

By an algebraic simplification, we can see that WEN also enjoys the computational advantage of the Lasso. Thus, an algorithm for the WEN method based on the algorithm LARS [14] can be developed [20]. The WEN algorithm is applied to IMS data sets for predictor selection and classification, and results show that the WEN method works effectively and efficiently for IMS data processing.

The WEN algorithm together with EN4IMS plus many other functions are integrated into a software package called IMSmining. Classification results of using the EN4IMS and WEN models are compared with those of other current popular methods used in the IMS

community. In a real data set of two mouse brain tissue sections, one is used for model training and the other section is used for model testing. 110 pixels are selected from the cancer area to be used as the training cancer data set, and 110 pixels are selected from the normal area to be used as the training noncancer data set. Similarly, 110 cancer pixels and 110 noncancer pixels are selected from the second mouse brain tissue section as test data. Classification rates show that the EN4IMS and WEN models outperform the other methods.

Since both the EN4IMS and the WEN models are 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 at all. It would be also very interesting to incorporate wavelet transform of IMS data into the study of classification and biomarker discovery.

Acknowledgements

The authors are grateful to an anonymous referee for constructive comments. This research was partially supported by Middle Tennessee State University Faculty Research Grant #2-21519 and Yue-Kong Pao Endowment Fund at Ningbo University, China..

References

[1] R. Aebersold and M. Mann M, Mass spectrometry-based proteomics, Nature, 422 (2003), 198-207.

[2] K.A. Baggerly, J.S. Morris, J. Wang, D. Gold, L.C. Xiao, and K.R. Coombes, A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples, Proteomics, 3 (2003), 1667-1672.

[3] K.A. Baggerly, J.S. Morris, and K.R. Coombes, Reproducibility of SELDI- TOF protein patterns in serum: comparing datasets from different experiments, Bioinformatics, 20 (2004), 777-785.

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

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

[6] S. Chen, D. Hong, and S. Yu, Wavelet-based procedures for proteomic mass spectrometry data processing, Computational Statistics & Data Analysis, 52 (2007), 211-220.

[7] S. Chen, M. Li, D. Billheimer, D. Hong, B. Xu, and Y. Shyr, A Novel Comprehensive Wave-form MS Data Processing Method, Bioinformatics, 25 (2009), 808-814.

[8] R. Coifman and D. Donoho, Translation invariant de-noising, In: Wavelets and Statistics, pp. 125-150, New York. Springer-Verlag, 1995.

[9] K.R. Coombes, S. Tsavachidis, J.S. Morris, K.A. Baggerly, M.-C. Hung, and H.M. Kuerer, Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform, Proteomics, 5 (2007), 4107-4117.

[10] Coombes, K.R., et al. Preprocessing mass spectrometry data, In: Fundamentals of Data Mining in Genomics and Proteomics, pp. 79-99, Kluwer, Boston, 2007.

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

[12] Pan Du, Warren A. Kibbe, and Simon M. Lin, Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching, Bioinformatics, 22 (2006), 2059-2065.

[13] P. Du, S.M. Lin, W.A. Kibbe, and H.H. Wang, Application of wavelet transform to the MS-based proteomics data preprocessing, Bioinformatics and Bioengineering, BIBE. Proceedings of the 7th IEEE International Conference, pp. 680 - 686, Boston, MA, 2007.

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

[15] A. Faghfouri and W. Kinsner, Local and global analysis of multifractal singularity spectrum through wavelets, IEEE CCECE/CCGEI, pp.2163-2169, Saskatoon, 2005.

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

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

[18] D. Hong and Y. Shyr, Mathematical framework and wavelets applications in proteomics for cancer study. In: Handbook of Cancer Models with Applications to Cancer Screening, Cancer Treatment and Risk Assessment (W.Y. Tan and L. Hannin Eds.), pp. 471-499, World Scientific, Singapore, 2008.

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

[20] D. Hong and F. Zhang, Weighted Elastic Net Model for Mass Spectrometry Imaging processing, Math. Model. Nat. Phenom., 5(2010), 115-133.

[21] N.E. Huang, S. Zheng, S.R. Long, M.C. Wu4, H.H. Shih, Q. Zheng, N.-C. Yen7, C.C. Tung, and H.H. Liu, The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis, Proc. R. Soc. Lond. A., 454 (1998), 903-995.

[22] D. Kwon, M. Vannucci, J.J. Song, J. Jeong, and R.M. Pfeiffer, A novel wavelet-based thresholding method for the preprocessing of mass spectrometry data that accounts for heterogeneous noise, Proteomics, 8 (2008), 30193029.

[23] E. Lange, C. Gropl, K. Reinert, High-accuracy peak picking of proteomics data using wavelet techniques, Biocomputing, 11 (2006), 243-254.

[24] X.L. Li, J. Li, and X. Yao, A wavelet-based data preprocessing analysis approach in mass spectrometry, Computers in Biology and Medicine 37 (2007), 509-516.

[25] J.S. Morris, K.R. Coombes, J. Koomen, K.A. Baggerly, and R. Kobayashi, Feature extraction and quantification for mass spectrometry in biomedical application using the mean spectrum, Bioinformatics, 21 (2005), 1764-1775.

[26] R. Guerra, M. Vannucci, Y. Li, C.C. Lau, T.K. Man, and A.C. Marcelo, Comparison of algorithms for preprocessing of SELDI-TOF mass spectrometry data, Bioinformatics, 24 (2008), 2129-2136.

[27] S. Mallat and W.L. Hwang, Singularity detection and processing with wavelets. IEEE transactions on information theory, 38 (1992), 617-643.

[28] H. Meistermann, J.L. Norris, H.R. Aerni, and et al., Biomarker discovery by imaging mass spectrometry, Molecular & Cellular Proteomics, 5 (2006), 1876-1886.

[29] J.S. Morris, P.J. Brown, R.C. Herrick, K.A. Baggerly, and K.R. Coobes, Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models, Biometrics, 64 (2008), 479-489.

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

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

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

[33] Y. Yasui, M. Pepe, M.L. Thompson, B.L. Adam, G.L. Wright, Y. Qu Jr., J.D. Potter, M. Winget, M. Thornquist, and Z. Feng, A data analytic strategy for protein biomarker discovery: profiling of high dimensional proteomic data for cancer detection, Biostatistics, 4 (2003), 449-463.

[34] C.L. Tu, W.L. Hwang, and J. Ho, Analysis of singularities from modulus maxima of complex wavelets, IEEE transactions on information theory, 51 (2005), 1049-1062.

[35] F. Zhang and D. Hong, Elastic Net Based Framework for Imaging Mass Spectrometry Data Biomarker Selection and Classification, Stat. in Medicine, in press.

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