RAMS: Regression Analysis for Cancer Diagnosis based on Proteomic-Profile Mass Spectrometry Data

Han Liu [*]
Statistic Department and Machine Learning Department
Carnegie Mellon University

July 9, 2006

ABSTRACT

A large body of research has been done in using patterns derived from mass spectra from time-of-flight matrix-assisted laser desorption and ionization (MALDI-TOF), or Surface-enhanced laser desorption (SELDI-TOF) mass spectrometer to differentiate patient samples from control cases. Based on the raw spectra, high precision for discriminant analysis has been reached without identification of the underlying proteins responsible. However, the reproducibility and reliability of these procedures are still questionable. One of the main challenges is how to deconvolute the mixture of biologically meaningful process and the artifactually involved noise process. To achieve this goal, a careful analysis and preprocessing of the obtained spectrometry data is needed, the models used for analysis should be justified. More importantly, the whole procedure should be conducted under a unified framework in a reproducible manner. For the sake of its simplicity and interpretability , regression analysis is adopted as our starting point. In this paper, we introduced a newly developed mass spectrometry analysis package named RAMS. In which, different parametric and nonparametric regression techniques are used for data preprocessing. Some simple classification methods are also implemented as modules for sample classifications. We use RAMS to analyze an Ovarian cancer dataset, with 10-fold cross validation, Our analysis achieves a misclassification error rate about 0.35%. Which is comparable with current result of the literature, but our approach is easier.

Keywords: Regression analysis, sample classification, variable selection, mass spectrometry, cancer diagnosis.

[*] Ph.D. candidate in the joint program between Statistics and Machine Learning

1 Introduction

A reliable and precise classification of patient samples is essential for successful cancer diagnosis and treatment. Mass Spectrometry (MS) is a promising technique which is being used increasingly in cancer research. By analyzing the expression profile for thousands of proteins simultaneously, such technique may lead to a more complete understanding of the underlying biomarkers and hence to a finer and more informative diagnosis strategy. The ability to successfully discriminate cancer samples from normal counterparts is by far the most important aspect of this novel approach to early cancer detection. More specifically, with a massively parallel analysis of thousands of proteins in a reproducible manner, comparative proteomic profiling extracted from normal versus cancer tissues enables the possibility to discover important proteins as biomarkers that play a crucial role in disease pathology.

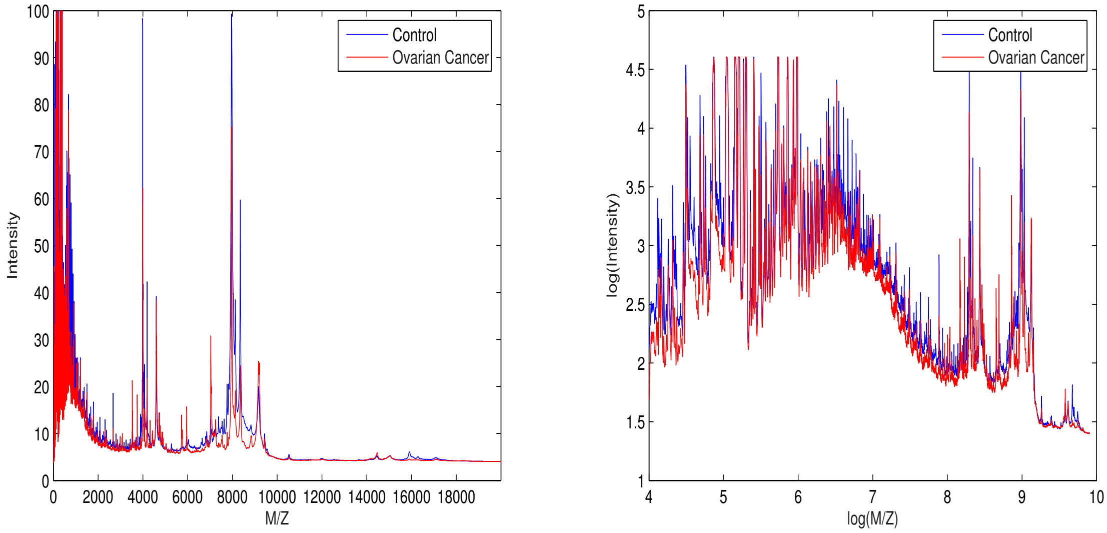
Generally, a complete mass spectrometry analysis includes three stages: (i) data preprocessing, (ii) feature selection, and (iii) sample classification. Data preprocessing is mainly used for noise-reduction and smoothing, so that the products could be used for the downstream analysis; Feature extraction is the most important step for biomarkers identification; While sample classification is mainly used for early cancer diagnosis. In this section, recent progresses in the last three years are briefly summarized according to these three categories. Because of the importance of data preprocessing and feature extraction, we will describe them in more details than the already well-established classification methods. Some problems and challenges of current procedures are also addressed, which motivates the development of a more systematic analytical framework.

1.1 Data Preprocessing

Sample preprocessing includes spectrum calibration, baseline correction, smoothing, peak identification, intensity normalization and peak alignment [1, 2, 3]. It is the most important step for mass spectrometry analysis, all the following analysis crucially depend on the quality of this step, several popular preprocessing techniques are summarized here according to different stages.

Spectra Calibration: Aligning individual spectra is the first step for data preprocessing. Due to some instrument factors, even with the use of internal calibration, the maximum observed intensity for an internal calibrant may not occur exactly at the same m/z site across different spectra. This can be solved by visually align the maximum observed intensity of the internal calibrant. Also, For the collected raw spectra, the distance between each pair of consecutive m/z ratios is not constant. Instead, the increment in m/z values is approximately a linear function of the x-axis values. Wu et al. [4] performed a logarithm transformation of the m/z values to make the scale of the predictor roughly comparable across the range of all m/z values. At the same time, to reduce the dynamic range of the peak intensities, a logarithm transformation on the peaks is also conducted. Figure 1 shows some raw spectra before and after these logarithm-transformations.

Background substraction: During the sample preparation procedure, chemical and electronic noise might produce background fluctuations. This pattern is spectrum specific and tends to dominant the distributions of the peak intensities in the whole spectrum. Therefore, it's important to remove these background noise before further analysis. Wu et al [4] used a "loess" smoother to fit a nonparametric local regression line to estimate the background intensity values,



(a) Before transformation

(b) After transformation

Figure 1: Left: Raw spectra before any transformation Right: the same spectra after the logarithm transformation

and then subtracted the fitted values from the fitted regression mean function. Baggerly et al. [5] proposed a semi-monotonic baseline correction method in their analysis of the SELDI-TOF data. An example of baseline substraction is shown in figure 2 with a windowed piecewise cubic interpolation method. For the MALDI-TOF mass spectrometry data, after background correction, a smoothing procedure is generally needed to smooth over the effect of isotopic envelop presented in the data. Coombes et al. [6] adopted a wavelet based approach for smoothing and de-noising the mass spectrometry data before peak identification. A recent work by Tibshrani et al. [7] used "super-smoother" with a span of 0.002 to achieve this goal.

Peak identification: Some mass spectrometers could provide a list of labelled peaks, but this information is not available in most systems. In these cases, a peak finding procedure is needed to identify peaks in each individual spectrum. The most intuitive and heuristic idea is viewing spectral peaks as local maxima in the mass spectrometry data. For example, Yasui et al. [8] developed a neighbored local intensity method, they look for sites (m/z values) that have a highest peak intensity among the $\pm s$ sites that around it. Also, this candidate peak should have a value higher than the average background at that site across different spectra. Based on these heuristics, several parameters need to be specified, eg. the number of neighborhood points, the threshold for the background intensity values. The validity of these parameter settings depend crucially on the correctness of the underlying noise models. A different approach is trying to identify peaks by plugging into prior biological knowledge, eg. Yu et al. [9] deemed a spectral peak to be biologically meaningful (eg. due to peptide ionization) only if there are more than one isotopic variant of it has presented.

Peak alignment: Due to the existence of chemical noise, co-crystallization, isotopic elements,

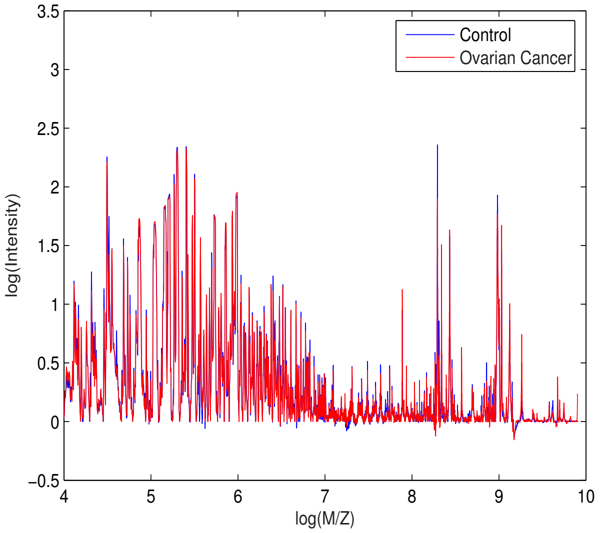


Figure 2: Mass spectra after the logarithm transformation and background substraction

and other unknown factors, the same biological peak may be horizontally shifted in different spectra. This challenge can be addressed by aligning peaks across different spectra. E.g, Yasui et al. [8] believed that the m/z axis shift of peaks is about $\pm0.1\%$ to $\pm0.2\%$ of the m/z value, they simply assume the peak has a width of $\pm0.4\%$ of the m/z value. Then, the m/z values with the largest number of peaks which have overlapped ranges is extracted. Thus, the set of peaks that contribute to this m/z are aligned and removed from the spectra. This process is conducted iteratively until all the peaks are exhaustive. Tibshirani et al. [7] proposed a more intuitive algorithm, called Peak Probability Contrasts (PPC), which can align the peaks by hierarchical clustering. For this, complete linkage hierarchical clustering is applied to the peak positions along the $\log(m/z)$ axis, and the resulting dendrogram is cut at height $\log(0.005)$, all the peaks in the same cluster are aligned together.

Peak normalization: Normalizing the peak intensities across different spectra into the same scale could be quite helpful to the comparison of different groups of spectral peaks in different mass spectra. The naive approach is to normalize each spectrum by a linear transformation that maps the minimum and maximum peak values to 0 and 1 respectively. This approach obviously oversimplified the problem. Considering the possible outliers and influential points, Tibshrani et al. [7] used $10th$ and $90th$ quartiles instead of the minimum and maximum as the range for mapping: all the peaks higher than the $90th$ quartile or lower then the $10th$ quartile are mapped to 1 and 0 respectively.

Most of these data preprocessing methods are very heuristic and *ad hoc*. Many of them come from the past experience and personal knowledge about the data. Some approaches work pretty well on some specific datasets, but not effective on the others. How to develop a systematic, theoretically soundable, framework for mass spectra preprocessing would be an interesting topic. Another crutial point is, these methods generally do not distinguish the underlying noise process and the biological process explicitly. Due to the high sensitivity of the mass spectrometer to the protocol for sample processing , the meaningful biological process is quite prone to be "tainted" by the influential noise patterns. Due to the essential uncertainty nature of the mass spectrometry analysis, it may be useful to treat it from a probabilistic perspective. By viewing each peak as a random variable which subject to some unknown distributions, the data preprocessing problem can be reduced to the a simpler form as how to transform the data to make it satisfy the model assumptions, this is a well-studied topic in the area of regression analysis [10].

1.2 Feature Extraction

After preprocessing, all the peaks from different spectra are aligned and normalized well for comparison. By viewing each m/z site as a feature, generally, we will get hundreds of thousands of features (also called covariates exchangeably in the following discussion) versus a relatively small sample size (about several hundreds). Feature extraction serves three-fold purposes in this context: (i) Trying to discover interesting and interpretable biomarkers for a specific disease, (ii) Reduce the data complexity to make it more tractable to statistical analysis. (iii) Reduce the noise in the data, to make the analyzed results more reliable. The proposed feature extraction approaches could be divided into two divisions according to whether it's *deterministic* or *nondeterministic*. Deterministic approach includes two sample t-test, Principle component analysis, boosting etc. While the nondeterministic methods are mainly represented by genetic algorithm and neural network methods.

Feature extraction Settings: Suppose that, we have n_0 samples from the cancer group (eg. cancer patient) and n_1 from the control group (normal person). There are altogether m covarites (m/z values) in each spectrum. From a probabilistic perspective, assuming that $X_k(G)$ represents the k-th covariate in the group G, $(G = 0, 1)$. 0 for the cancer group and 1 for the control group, $k = 1, ..., m$. The observations of the random variable $X_k(G)$ are

$$X_k(0) = (x_{k,1}(0), x_{k,2}(0), ..., x_{k,n_0}(0))^T$$

for the cancer group, and

$$X_k(1) = (x_{k,1}(1), x_{k,2}(1), ..., x_{k,n_1}(1))^T$$

for the normal group. The task is trying to find a subset of the covariates $\{X_1, ..., X_s\}$, $(s \ll k)$ which could informatively summarize the difference between these two groups. More importantly, in the context of mass spectrometry analysis, we expect them to be biologically interpretable.

Two sample t-test: As in the analysis by Guoan et al. [11], two sample t-test is used to quantify the difference between two groups in the analysis of mass spectrometry data. When we are interested in identifying peak intensities that are differentially expressed in two populations of mass spectrometry samples—– cancer versus normal. According to the formalization of the data above, the parameter of interest is an m–vector of differences in mean expression measures in the two populations, $\psi(k) = \mathrm{E}[X_k(G)|\ G = 0] - \mathrm{E}[X_k(G)|\ G = 1]$, $k = 1, \ldots, m$. To identify peaks with higher mean intensity measures in the abnormal compared to the normal counterparts, one can test the two-sided null hypotheses $H_0(k) = \mathrm{I}(\psi(k) = 0)$ vs. the alternative hypotheses

$H_1(k) = \mathrm{I}(\psi(k) \neq 0)$, using two-sample Welch t-statistics

$$T_k \equiv \frac{\bar{X}_{k,n_0}(0) - \bar{X}_{k,n_1}(1)}{\sqrt{\frac{\sigma^2_{k,n_0}(0)}{n_0} + \frac{\sigma^2_{k,n_1}(1)}{n_1}}}, \tag{1}$$

where n_0, n_1, $\bar{X}_{k,n_0}(0)$, $\bar{X}_{k,n_1}(1)$, and $\sigma^2_{k,n_0}(0)$, $\sigma^2_{k,n_1}(1)$ denote, respectively, the sample size, sample means, and sample variances, for samples with different status (cancer vs. normal). If the null hypotheses are rejected, i.e., the corresponding peak intensities are declared differentially expressed. Two sample t-test may be the most popular statistical method used for feature selection due to its simplicity. However, one potential problem is that t-test is not a very robust statistic, the lack of robustness may impair the feature selection result when a huge amount to features are being screened. Also, the reliability of t-test depends on the underlying data has an approximate normal distribution assumption. When there are only about tens of spectra in each groups, it is not clear that screening on the value of the t-statistic is still effective or not.

Peak probability contrasts (PPC): Peak probability contrasts (PPC) was proposed by Tibshrani et al. [7]. Unlike two sample t-test that takes into account the absolute peak intensity value, PPC cut the peak height at some quantile in such a way as to maximally discriminate between the cancer and normal samples in the training dataset. This approach should be more reliable than the two sample t-test for mass spectrometry analysis. Since it takes into account the fact that the peak intensity values can vary greatly across the m/z range. For more details, assume $q(\alpha, k)$ represents the α quantile of the peaks $x_{k,j}$ at site k, then, given two groups $\mathcal{A}_G$, $G = 0, 1$ of size n_0 and n_1, let $p_{k,\alpha}(G)$ be the proportion of spectra in group G with a peak at site k larger than $q(\alpha, k)$:

$$p_{k,\alpha}(G) = \sum_{j \in \mathcal{A}_G} I[x_{k,j} > q(\alpha, k)]/n_i, \quad i = 0, 1 \tag{2}$$

Fianlly, the optimal $\alpha(k)$ is chosen to maximize $|p_{k,\alpha}(G = 0) - p_{k,\alpha}(G = 1)|$ and set $\hat{p}_k(G) = p_{k,\hat{\alpha(k)}}(G)$. After these calculations, important features could be selected according to the decreasing order of $|\hat{p_k}(G = 0) - \hat{p_k}(G = 1)|$. PPC does not depend on any explicit normality assumptions, it should be more robust than the two sample t-test and is also computationally tractable.

Principle Component Analysis (PCA): Principle component analysis was first applied for mass spectrometry analysis by Lilien et al. [12] in their algorithm package named $Q5$. PCA is an unsupervised technique mainly used for determining orthogonal axes of maximal variance from a dataset. When viewing each m/z site as a feature, the spectra are zero-meaned and an eigendecomposition (EVD) of the covariance matrix is computed. The eigenvector associated with the ith largest eigenvalue lies along the ith principal component. With the corresponding eigenvectors column spanned matrix as a projection matrix, the raw spectra is projected into a reduced-dimension space with little or no information loss. The reduced dimension could be viewed as a linear combination of the original peak intensities. PCA is a deterministic algorithm, when the number of sample is not quite large, it's easy to conduct the EVD decomposition. The biggest challenge for PCA method is how to interpret the reduced-dimensional features. Since this weighted combination mechanism does not have a clear biological interpretation, it is not quite natural to find biologically meaningful bombardiers.

Genetic algorithm: Unlike the above deterministic methods, genetic algorithm is an iterative approach. One of the representative work is done by Petricoin et al. [13]. In their work, each run of the genetic algorithm starts with 1500 logical chromosomes (feature sets) of a size ranging from 5 to 20 index values. The fitness of each feature set is measured by the Euclidean distance. New populations are then produced by preferentially combining pieces of the "most fit" members of the current generation. The process then evolves for 250 generations, with a mutation rate of 0.02% and random crossover locations. All the distinct features in the raw spectra are included in the feature set. There are a whole family of genetic algorithms that can be used for feature selection, these algorithms require multiple iterations to converge. The biggest problem for these algorithms is their *nondeterministicness*, even run on the same dataset, the results may be different due to different initial values. Also, interpretability is a drawback of such kinds of algorithms.

1.3 Sample Classification

After data preprocessing and feature extraction, almost every discriminant methods in the machine learning community can be applied directly for mass spectrometry analysis. Markey et. al [14] proposed to use classification and regression tree (CART) model for spectra classification. The basis of their algorithm is the recursive partitioning of the data into more homogenous subsets. The advantage of CART is it's good interpretability, with the binary decision tree representations learned from their algorithm, a series of if-then rules could be extracted. However, when the number of features are significantly larger than the number of samples, a tree model is prone to overfitting, it must be pruned to avoid this. Another approach is proposed by Lilien et al. [12], in which Fisher Discriminant Analysis (FDA) is used for mass spectrometry classification, with a supervised learning manner. They try to project the sample points with the extracted features onto a hyperplane which maximizes the between-class variation and within-class variation. The disadvantage of this approach is that the number of features should be less than the number of samples. In the context of mass spectra classification, this generally means a very strict pre-screening procedure, which is not always easy to be obtained without sacrificing the interpretability. In the work of Wu et al. [4], Support Vector Machine (SVM) and Random Forest (RF) are used for sample classification. These two methods use the whole spectra as input features and could conduct feature extraction simultaneous with sample classification. The problem for these method is that there are too many parameters needs to setup, even though it's possible to tune these parameters automatically, it's not easy to produce reproducible results based on these iterative procedures. A recent work by Tibshrani et al. utilized Nearest Shrunken Centroids [15] for sample classification, by comparing the extracted feature vectors in Euclidian distance (or other metrics), their classifier is essentially a nonparametric nearest neighbor classifier. The performance of Nearest shrunken Centroids was compared with those of SVM and RF under an external cross validation [16]. They found that the classification performance is comparable, which is quite promising, since their Nearest Shrunken Centroids method only used a selected feature sets, but not the full spectra.

In conclusion, the main task for the mass spectrometry analysis is trying to discriminate cancer samples from the normal samples simply based on proteomic profile patterns. For this purpose, we need to carefully process the raw data to reduce the effects of systematic noise. Also, important features should be extracted for biomarker discovery and sample classification. In the next section, we will summarize the dataset we use and report our results, a regression framework for our analysis. An integrated regression framework is proposed for data preprocessing, a R package named “RAMS” (Regression Analysis for Mass Spectrometry) is also developed and made available online. Further experimental design and details will be in section 3. Further discussions are presented in section 4.

2 Dataset and Results

Ovarian Cancer Dataset: The dataset was obtained from Yale’s biostatics group, published in Wu et al. [17]. The spectra were obtained from ovarian cancer and control serum samples from the National Ovarian Cancer Early Detection Program at Northwestern University Hospital. These samples were then automated desalted and were conducted a MALDI-TOF mass spectrometry analysis on a Micromass MALDI-L/R instrument. There are two modes for the MALDI-L/R mass spectrometer: linear and reflectron modes. Two sets of data could be automatically acquired in a positive ion detection model, with 700-3500 Da for the reflectron model and 3450-28000 Da for the linear. There are altogether 170 spectra samples, consisting of 93 patients with ovarian cancer and 77 non-cancer controls. For our analysis, we are only interested in the high resolution reflecton data, these spectra are measured at 94,780 sites, spaced 0.019Da apart. A visualization of one of the raw spectra is shown as in figure 3

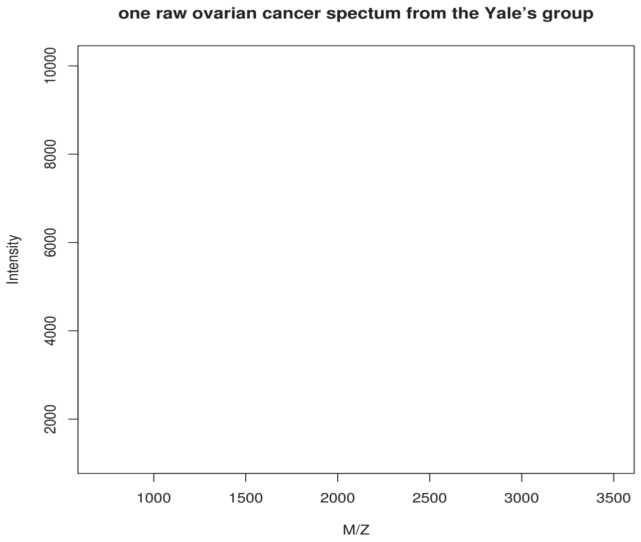


Figure 3: Visualization of one raw spectrum for our dataset

RAMS framework: RAMS (Regression Analysis for Mass Spectrometry) is an R package developed for MALDI-TOF or SELDI-TOF mass spectra analysis. Figure 4 is its flowchart. From which, we see that the whole analysis could be divided into three components. Regression analysis (both parametric and nonparametric approaches) is intensively used for data preprocessing. Parametric methods include the simple linear regression, regression through origin and robust regression. Nonparametric regression mainly includes Kernel regression and cubic splines; Model selection techniques, like two sample t-test, Lasso or forward stepwise selection are adapted for feature extraction; Different classification models are also included as modules, including Naive Bayes (NB), Nearest Neighbor Classifier (KNN), Logistic Regression (LR), Generalized Additive Models (GAM), Support Vector Machines (SVM), Trees, and Linear/Quadratic Discriminant Analysis (LDA/QDA). As shown in figure 4, all these procedures are under an external cross-validation framework, we have two versions of the cross validation: 10-fold cross validation and subset randomly splitting cross validation. For each step, different algorithm modules could be selected. For example, when conducting background substraction, either a loess estimator or a local linear regressor could be used. For the classification step, we may use either Logistic Regression or Generalized Additive Models. All of our analysis is under this RAMS framework.

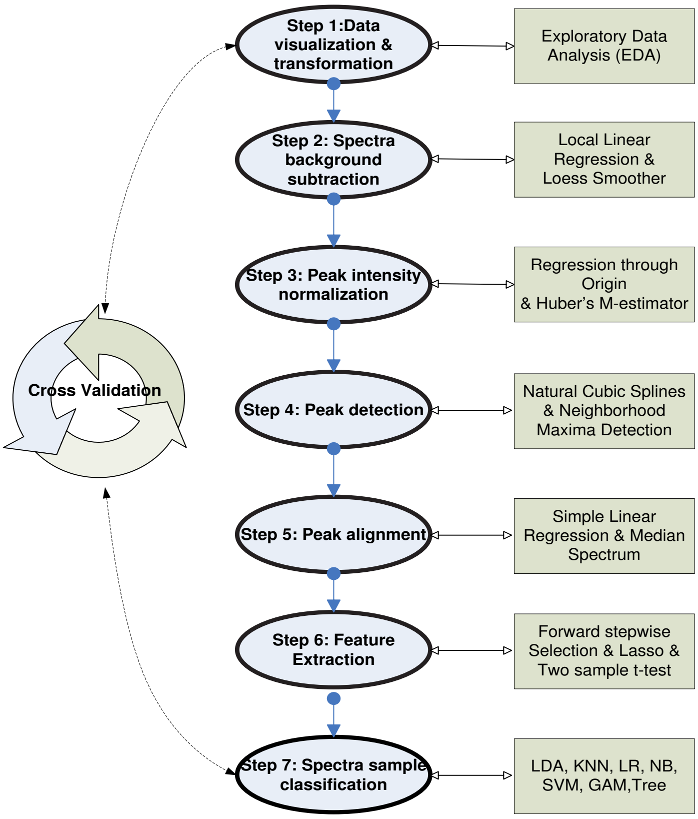

Step 1:Data visualization & transformation
Exploratory Data Analysis (EDA)
Step 2: Spectra background subtraction
Local Linear Regression & Loess Smoother
Step 3: Peak intensity normalization
Regression through Origin & Huber's M-estimator
Cross Validation
Step 4: Peak detection
Natural Cubic Splines & Neighborhood Maxima Detection
Step 5: Peak alignment
Simple Linear Regression & Median Spectrum
Step 6: Feature Extraction
Forward stepwise Selection & Lasso & Two sample t-test
Step 7: Spectra sample classification
LDA, KNN, LR, NB, SVM, GAM,Tree

Figure 4: Flowchart of the RAMS analysis

Results: Using RAMS, we compared the classification performance of different classifiers. The classification performance with and without some preprocessing steps (e.g. peak alignment) are measured and showed that these preprocessing procedures are really crucial. By conducting some goodness-of-fit test, we validate the consistency between the modelling assumption and the observed data. Under the assumption that the misclassification error is approximately distributed as Normal, we showed that the classification performance based on proteomic-profile spectra data could lead to a statistically significant improvement than random guessing, which demonstrates the prominence of the mass spectrometry approach for early cancer detection. The performance of RAMS is fairly comparable with the best published results on this dataset, but our approach is simpler and much easier to be understood by both statisticians and biologists. By a careful examination of the extracted biomarkers, we also evaluated and verified the internal reproducibility of RAMS.

3 Experimental Results and Analysis

In this section, individual steps and results will be described in detail, for each step showed in figure 4, we briefly illustrate the methods that RAMS adapted and some intuitive justifications about its rationale.

3.1 Step 1: Data transformation

Notice that there are several order of magnitudes of peak intensities, we take the logarithm of the intensity values, one of the log-transformed spectra was shown in figure 5

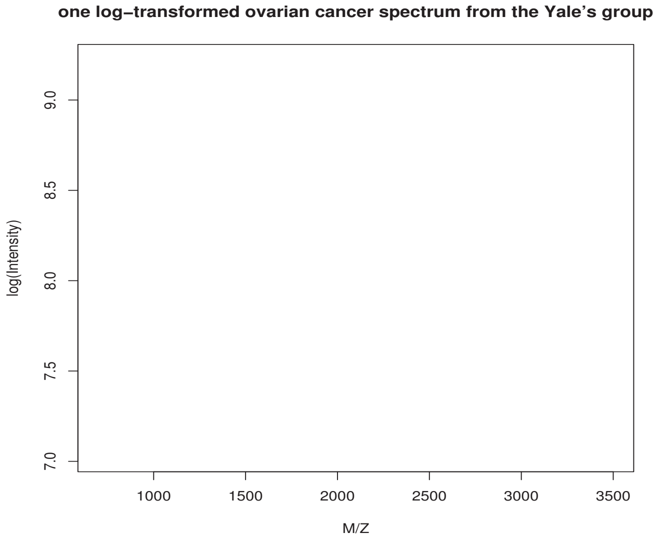


Figure 5: Visualization of one logarithm-transformed spectrum

From figure 5, we see that the peak intensity range are from 7∼10, while that of the raw spectra are ranged from 1000∼10000. Also, followed from Tibshrani et al. [7], we log-transformed the m/z axis. As pointed by Tibshrani et al., the peak widths were about 0.5% of the corresponding m/z value, and this relationship is approximately linear. Therefore, the logarithmic transformation of m/z values could make the peak widths approximately constant across the whole range and will greatly ease the downstream peak alignment algorithms. The visualization of the spectrum looks the same as in figure 5, the only difference is that the x-axis is now in log-scale. I figure 6, we showed the density and Q-Q plot of one particular peak intensity at the 7th site, this further justifies why a logarithmic-transformation is suitable.

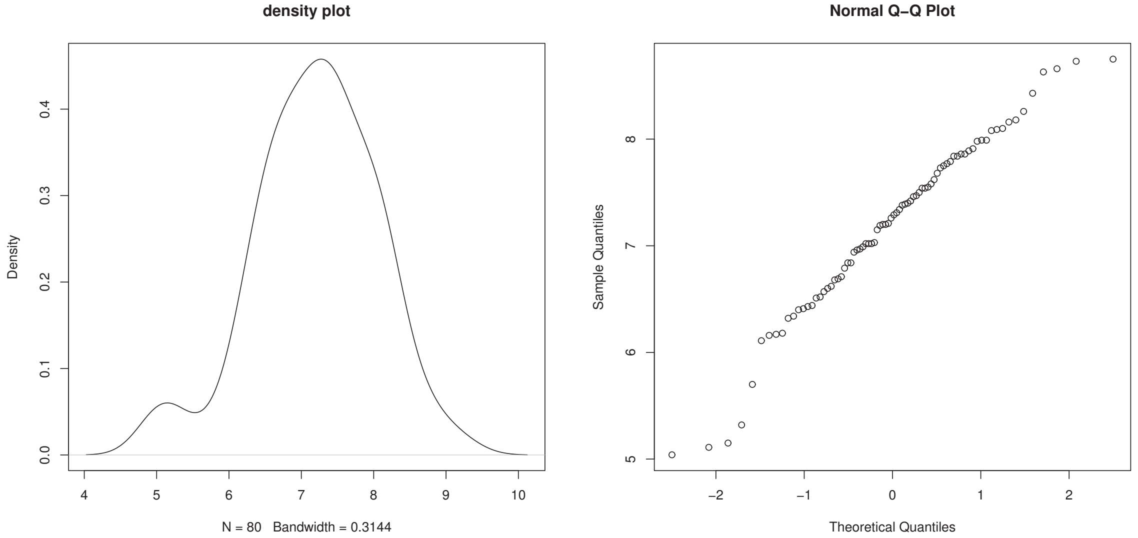


Figure 6: The density and Q-Q plot for a particular peak intensity

3.2 Step 2: Baseline Subtraction

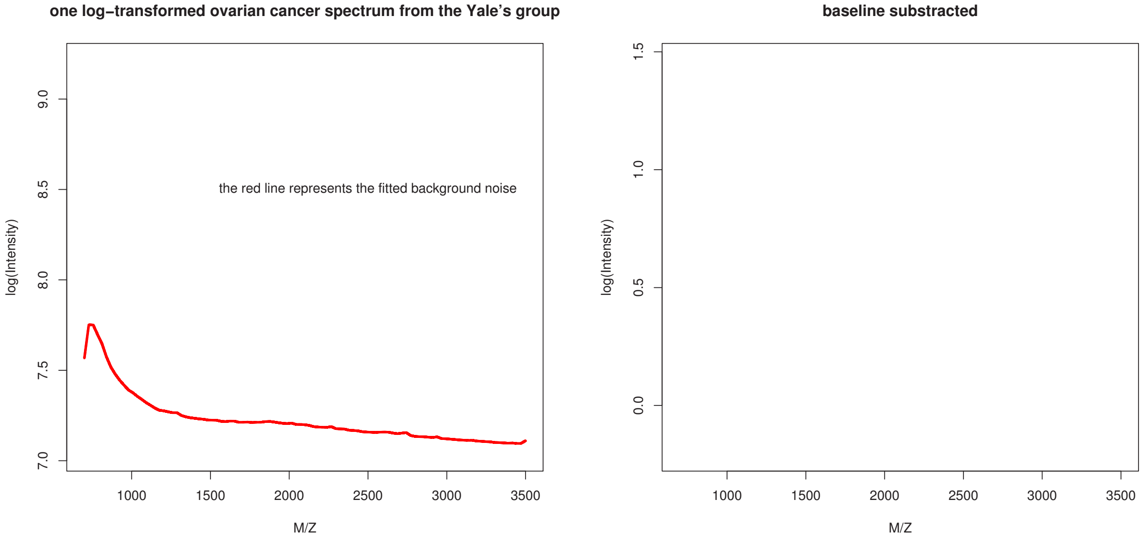
For the purpose of baseline subtraction, the basic approach is to fit the spectra-specific background first and subtract the fitted curve from the raw spectra. This is essentially a problem of nonparametric curve fitting. Both local linear smoother and "Loess" smoother could be applied for this purpose. The local linear smoother is a kind of kernel regressor which minimizes the locally weighted sums of squares

$$RSS_l = \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)(Y_i - a_0 - a_1(X_i - x))^2 \tag{3}$$

where $K(\cdot)$ is a selected kernel, which does not effects the fitted results much. Y_i is the ith peak intensity. X_i is the corresponding m/z value at the ith site. Rewrite it in the form as $(Y - X_x a)^T W_x (Y - X_x a)$, the weighted least squared estimator for a is

$$\hat{a}(x) = (X_x^T W_x X_x)^{-1} X_x^T W_x Y \tag{4}$$

In particular, $\hat{r}_n(x) = \hat{a}_0(x)$ is the inner product of the first element of the estimated $\hat{a}$. The optimal bandwidth h could be selected by minimizing the generalized cross-validation (GCV) score. The *Loess* smoother, originally proposed by Cleveland [18], is a method that is (somewhat) more descriptively known as locally weighted polynomial regression. At each point in the data set a low-degree polynomial is fit to a subset of the data, with covariate values near the point whose response is being estimated. A parameter named "span" represents the proportion of data we want to use to fit a particular point. Thus, it could be used to control the model complexity. From our experience, there is not much difference between these two types of local smoothers. With 94,780 samples and only 1 covariate, both of them could be fitted very well. The first plot of figure 7 illustrates the fitted baseline for the selected spectrum.



(a) The red line represents the fitted background

(b) Background subtracted spectrum by local linear regression

Figure 7: The fitted baseline for the selected spectrum and the baseline subtracted spectrum

Figure 8 showed the estimated 95% confidence band for the mean of the estimated baseline and the estimated variance at each point. To estimate the variance, we assume that

$$Y_i = r(x_i) + \sigma(x_i)\epsilon_i \tag{5}$$

Let $Z_i = \log(Y_i - r(x_i))^2$ and $\delta_i = \log\epsilon_i^2)$. then $Z_i = \log(\sigma^2(x_i)) + \delta_i$. In more detail, we first estimate $r(x)$ with local linear smoother to get an estimate $\hat{r_n}(x)$. Then, simply regress Z_i's on the x_i's with local linear smoother again to achieve an estimate $\hat{q}(x)$ of $\log\sigma^2(x)$ and $\sigma^2\hat{(}x)$ should be $e^{\hat{q}(x)}$. To estimate the confidence band, assume we use a linear smoother, that is, $\hat{r}_(x) = \sum_{i=1}^{n} l_i(x) r_(x_i)$. The following formula could be used [19]

$$I(x) = (\hat{r_n}(x) - c\hat{\sigma}||l(x)||, \hat{r_n}(x) + c\hat{\sigma}||l(x)||) \tag{6}$$

where, the constant c is chosen as $2\,(1-\Phi(c)) + \frac{\kappa_0}{\pi}e^{-\frac{c^2}{2}} = \alpha$, while $\kappa_0 = \int_a^b ||T'(x)||dx$, $T'(x) = (T_1'(x), ..., T_n'(x))$, and $T_i(x) = l_i(x)/||l(x)||$. From figure 8, we see that the confidence band is

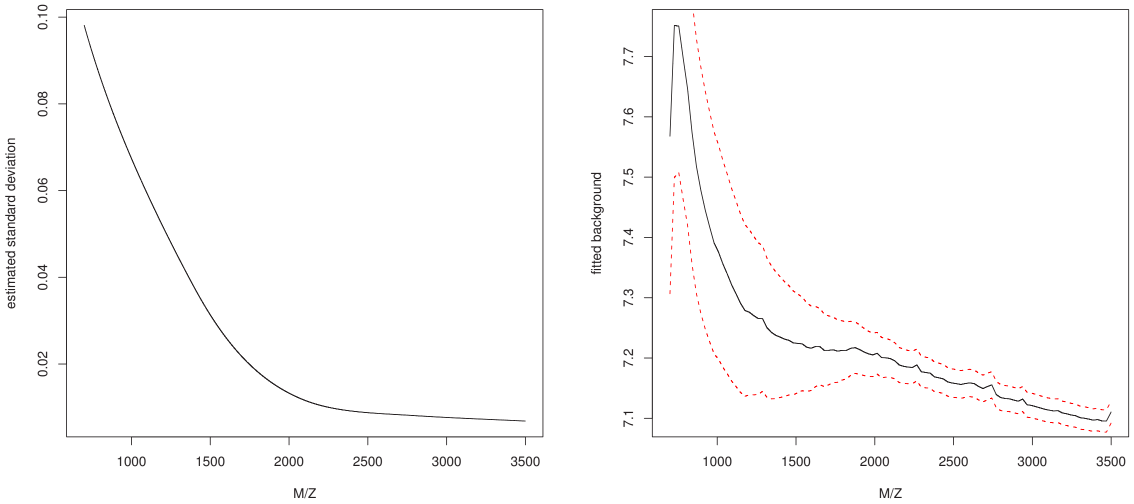


Figure 8: The estimated variance and confidence band for the fitted baseline on each m/z point

very loose when m/z is small and it becomes tighter with the increase of m/z. Also, the second plot of figure 7 showed the background subtracted version of the same spectrum in figure 5. It's obvious that some peak intensities will become negative by this approach, but this will not affect their discriminative power.

3.3 Step 3: Spectra Normalization

Robust regression is used by RAMS for spectra normalization. First, we calculate the median of the m/z values and the peak intensities at each point to form a "median spectrum". The intensities of every spectrum are then normalized by calculating a normalization factor with this median spectrum. This is essentially a problem of regression through origin. Huber's M-estimator is adopted. For which, we choose the factor β to minimize

$$RSS_M = \sum_{i=1}^{n} \rho \left(\frac{Y_i - x_i \cdot \beta}{s} \right)^2 \tag{7}$$

where s is defined as $\text{median}_i|Y_i - \text{median}_j Y_j/0.6745$, while ρ is the Huber function, (c is chosen to give a 95% efficiency at the Normal)

$$\rho(x) = \begin{cases} x^2 & |x| \leq c \\ c \times (2|x| - c) & |x| > c \end{cases} \tag{8}$$

For each spectrum, we could calculate a normalization factor with respect to the median spectrum and their Pearson's correlation coefficient. One thing to note is that the calculated normalization factor could be used as a metric for spectra quality assessment. Due to the effects of chemical noise or electronic fluctuation, some spectra may not be valid at all. From the figure 9, we see that, among 80 spectra in the training set (40 from the cancer group, 40 from the control group), the spectra # 2 and # 47 have a very negative normalization factor and the Pearson's correlation coefficient is less than 0, if a threshold 0.2 is setup, these two spectra could be removed from the current training set as outliers. From figure 9, even though the normalization factor plot and the Pearson's correlation plot looks very similar, they convey different information. The normalization factor plot represents the "magnitude" of a spectrum corresponding to the median spectrum, while the Pearson's correlation represents the "linear relationship" between them.

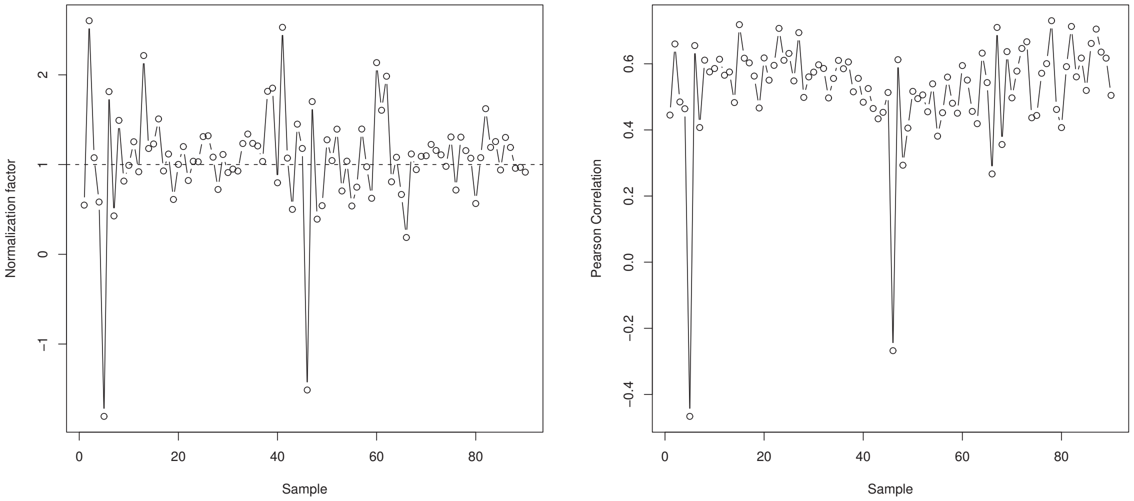


Figure 9: Normalization factor and the Pearson correlation coefficient for 80 spectra in the training dataset

3.4 Step 4: Peak Detection

Since we are only interested in the biological meaningful peaks as biomarkers, it's a better strategy to detect them before the feature selection step. For this, a natural cubic spline is used to further smooth the data. Then, the peaks are deemed as local maxima in a bandwidth s neighborhood. For the cubic spline, by choosing $B_1, ..., B_N$ as the power basis, we need to find the coefficients β to minimize

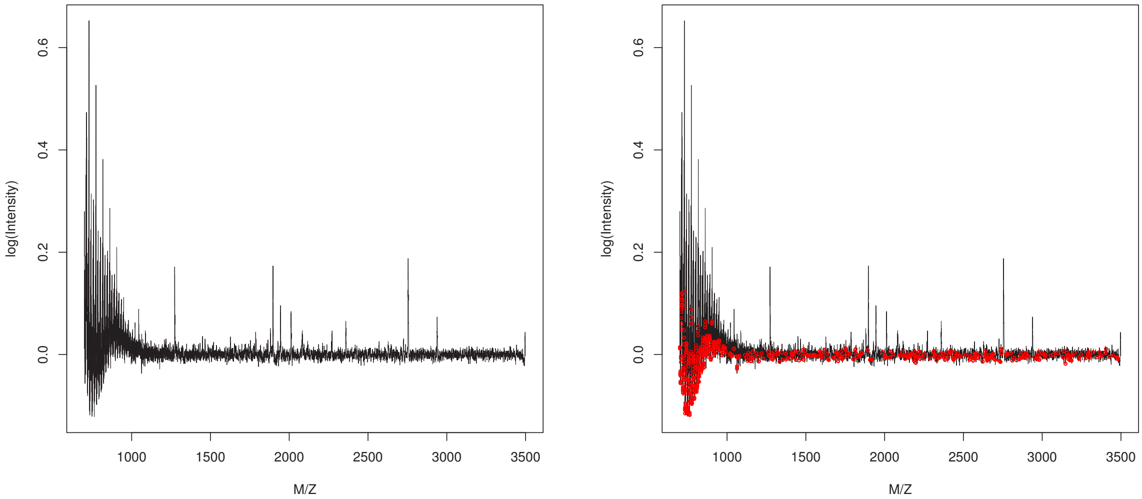
$$(Y - B\beta)^T(Y - B\beta) + \lambda\beta^T \int B_j{}''(x)B_j{}''(x)dx\beta \tag{9}$$

the value of β that minimizes it is

$$\hat{\beta} = (B^T B + \lambda \int B_j{}''(x)B_j{}''(x)dx)^{-1}B^T Y \tag{10}$$

Also, we deem a point as a peak only if there are at least s successive points in the neighborhood show a progressive increase and decrease in the background corrected, spline smoothed spectra.

The tuning parameter s could be controlled by cross-validation. However, for the purpose of our analysis, we want to keep as more information as possible, we choose $s = 5$. The first plot of figure 10 shows the cubic spline smoothed version of the same spectrum as in figure 5.



(a) Natural cubic spline smoothed spectrum

(b) Detected peaks (red point) from the spline smoothed spectrum with $s = 20$

Figure 10: The spline smoothed spectrum and the detected peaks

The second plot of figure 10 illustrates the detected peaks by the local maxima neighborhood algorithm with $s = 20$, from which, we see that even though many intensities are obvious local maxima, they can not be deemed as biologically meaningful peaks because their neighborhood do not show a progressive ascending and descending trend.

3.5 Step 5: Peak Alignment

Due to the instrumental measurement error and some other unknown factors, the same peak across different spectra has a nonlinear shift. We applied simple linear regression to make a coarse alignment. Then, a refined alignment is done by a restamping/binning based on the median spectrum. For each spectrum, a global mass shift δ is calculated as

$$\delta = \bar{Y} - \bar{X} \tag{11}$$

where $\bar{Y}$ is the mean intensity of the median spectrum, while $\bar{X}$ is the mean intensity of this spectrum.

Since we have already log-transformed the m/z values, the peak width now should be stationary with respective to $\log(m/z)$. We binned the $\log(m/z)$ of the median spectrum. When calibrating different spectra, only the maximum peak in each bin is retained. The bin number m could be selected by cross-validation. For our analysis, $m = 10,000$ is a fixed value.

3.6 Step 6: Feature Selection

Feature selection is a crucial step for Biomarker discovery. Since the data is in high dimensions, we use a combination of two sample t-test and Lasso/forward stepwise selection. The standard setting for the two sample t-test is defined as in formula 1. Having calculated the T_k, the p-value is the tail probablity of the t-distribution with a degree of freedom $n_1 + n_2 - 2$. A coarse variable selection for RAMS is done by choosing the smallest 100-200 p-values calculated from the two sample t-test.

Based on the coarsely selected variables from two sample t-test, Lasso is then used to conduct a finer selection. For Lasso, we want to find the β to minimize

$$RSS_{L_1} = \left(\sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda ||\beta||_1 \right) \tag{12}$$

where $\lambda > 0$ and $||\beta||_1$ is the L_1 norm of the vector β. The constant λ can be chosen by the cross validation score. This L_1 norm regularization could lead to a sparsity solution. Besides Lasso, we also tried forward stepwise selection with Bayesian Information Criterion (BIC) for variable selection. Even though its property, the performance is close to Lasso for this dataset. Figure 11 shows the coefficients vs. degrees of freedom plot for Lasso when the number of features is 30 and illustrates the corresponding Mallow's Cp score for feature selection. We see that, after a coarse selection by the two sample t-test, if we apply Lasso on the remained 30 features, the Mallow's Cp score is minimized when 14 variables are selected. As pointed by Tibshrani et al. [7], two sample t-test may not be the best criterion for feature selection. RAMS only adopted two-sample t-test as a coarse selection step, while a fine feature extraction step is based on Lasso.

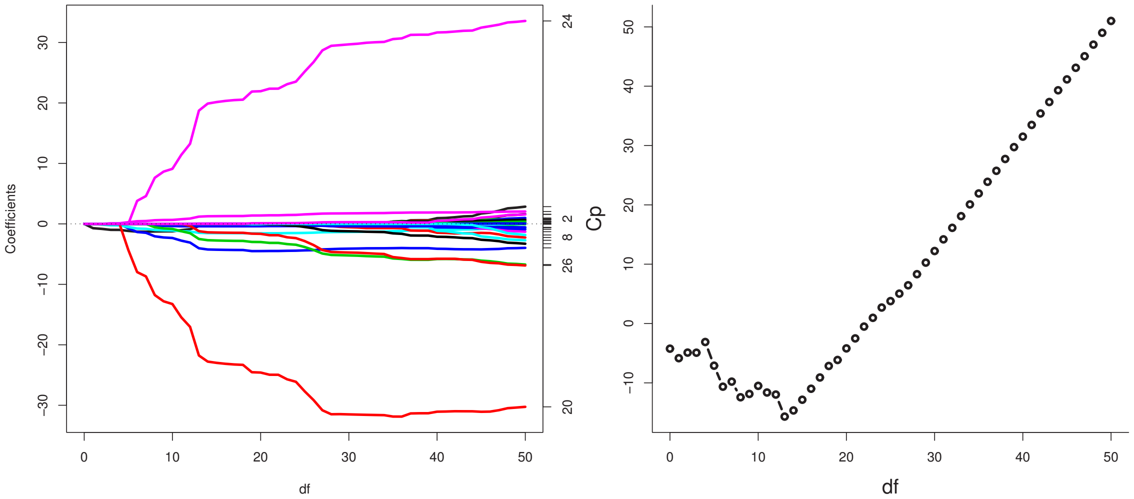


Figure 11: Left: The Lasso plot for feature selection Right: The Mallow's Cp score vs. the degree of freedom

Some more sophisticated methods, such as PPC as discussed before, could also be plugged in as a module into the whole framework.

3.7 Step 7: Sample Classification

For the chosen ovarian cancer dataset, there are only two classes, cancer ($G = 1$) or control ($G = 0$). Therefore, it's a binary classification problem. We compared and evaluated 7 methods under the RAMS framework, including: Naive Bayes (NB), Generalized Additive Model (GAM), Linear and Logistic Regression (LR), Linear/Quadratic Discriminant Analysis (LDA/QDA), Classification and Regression Trees (CART), and the Support Vector Machine (SVM), these classifiers are briefly summarized here

Bayes' rule for classification Most of the classification methods are based on Bayes' Rule: assume that $r(x) = \mathrm{E}(Y|X = x) = \mathrm{P}(Y = 1|X = x)$ denotes the regression function, Bayes' rule h^* satisfies

$$h^*(x) = \begin{cases} 1 & \text{if } r(x) > \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

The set $\mathcal{D}(h) = \{x : r(x) = 1/2\}$ is called the decision boundary. If we know the true function, a classifier satisfies Bayes' rule should be optimal [19]

Linear and Logistic Regression: The approach of Linear and Logistic Regression for sample classification is to estimate the regression function $\hat{r}(x)$ first, then, the Bayes' rule is used

$$\hat{h}^*(x) = \begin{cases} 1 & \text{if } \hat{r}(x) > \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

For the linear regression model, the regression function is estimated as $r(x) = \beta_0 + \sum_{j=1}^{d} \beta_j X_j + \epsilon$. For the Logistic Regression

$$r(x) = \mathrm{P}(Y = 1|X = x) = \frac{\exp\{\beta_0 + \sum_j \beta_j x_j\}}{1 + \exp\{\beta_0 + \sum_j \beta_j x_j\}} \tag{15}$$

Linear/Quadratic Discriminant Analysis (LDA/QDA) For Quadratic Discriminant Analysis (LDA), each class is modeled by a multivariate Gaussian distribution, for $k = 0, 1$

$$f_k(x) = \frac{1}{(2\pi)^d |\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right\}$$

Thus, $X|Y = k \sim \mathcal{N}(\mu_k, \Sigma_k)$, then the Bayes' rule is

$$\hat{h}^*(x) = \begin{cases} 1 & \text{if } r_1^2 < r_0^2 + 2\log\left(\frac{\pi_1}{\pi_0} + \log(\frac{\Sigma_0}{\Sigma_1})\right) \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

where $r_k = (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k), \quad (k = 0), 1$ is the *Manalahobis* distance, if we assume that different classes have a equal variance-covariance matrix, QDA is reduced to LDA.

Naive Bayes and Generalized Additive Model: Naive Bayes works by estimating the underlying density, rewrite the Bayes' rule as

$$h^*(x) = \begin{cases} 1 & \text{if } \frac{f_1(x)}{f_0(x)} > \frac{1-\pi}{\pi} \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

Assuming that X_1,..., X_d are independent, that is $f_k(x_1,...,x_d) = \prod_{j=1}^d f_{kj}(x_j), \quad (k = 0, 1)$. When using some one-dimensional density estimators and multiply them, we have $\hat{f}_k(x_1,...,x_d) = \prod_{j=1}^d \hat{f}_{kj}(x_j), \quad (k = 0, 1)$. If π is estimated as $\hat{\pi} = \frac{1}{n}\sum_i Y_i$, the Bayes' rule for the Naive Bayes classifier is

$$\hat{h}^*(x) = \begin{cases} 1 & \text{if } \frac{\hat{f}_1(x)}{\hat{f}_0(x)} > \frac{1-\hat{\pi}}{\hat{\pi}} \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

The Generalized Linear Model has the form

$$\text{logit}\left(\frac{\mathrm{P}(Y=1|X)}{\mathrm{P}(Y=0|X)}\right) = \beta_0 + \sum_{j=1}^d g_j(X_j) \tag{19}$$

Naive Bayes and Generalized Additive models are closely related, the discussion could be found in [19]

Nearest Neighbors K Nearest Neighbor classifiers (KNN) finds the K objects (or neighbors) in the training data that are closest to it, it follows a so-called k-nearest neighbor rule

$$h^*(x) = \begin{cases} 1 & \text{if } \frac{\sum_{i=1}^n w_i(x) I(Y_i=1)}{\sum_{i=1}^n w_i(x) I(Y_i=0)} > 1 \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

where $w_i = 1$ if X_i is one of the k nearest neighbors of x, $w_i(x) = 0$ otherwise, often, Euclidean distance $||X_i - X_j||$ is used to evaluate the neighborhood relationship.

Trees: Trees are classification methods that partition the covariate space $\mathcal{X}$ into disjoint piectes and classify them by majority vote. To construct a tree, suppose that $y \in \mathcal{Y} = \{0, 1\}$. For the jth covariate X_j, $j = 1,..,d$, a split point t is chosen so that it divides the real line into $A_i = (-\infty, t]$ and $A_2 = (t, \infty)$. Let $\hat{p}_s(k)$ be the proportion of observations in A_s, such that $Y_i = k$, for $k = 0, 1$, $s = 1, 2$:

$$\hat{p}_s(k) = \frac{\sum_{i=1}^n I(Y_i = k, X_j i \in A_s)}{\sum_{i=1}^n I(X_j i \in A_s)} \tag{21}$$

The impurity of the split t is defined as $I(t) = \sum_{k=1}^2 \gamma_k$, where $\gamma_s = 1 - \sum_{k=0}^1 \hat{p}_s(k)^2$, this particular measure is named Gini index, The whole construction procedure is greedy, each time, we look for the j, so that spliting X_j could lead to the greatest decrease of the impurity. Of course, other criteria, such as BIC, AIC, or least sqaured error could be used alternatively.

Support Vector Machine: Support vector machine is basically a linear classify which trying to find a *maximum margin hyperplane*, for $j = 1,...,d$, define

$$\hat{\alpha_j} = \sum_{i=1}^d \hat{\alpha_i} Y_i X_j(i)$$

where $X_j(i)$ is the value of the covariate X_j for the ith data point, and $\hat{\alpha} = (\hat{\alpha_1},...,\hat{\alpha_n})$ is the vector that maximizes

$$\sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i=1}^n \sum_{k=1}^n \alpha_i \alpha_k Y_i Y_k \langle X_i, X_k \rangle \tag{22}$$

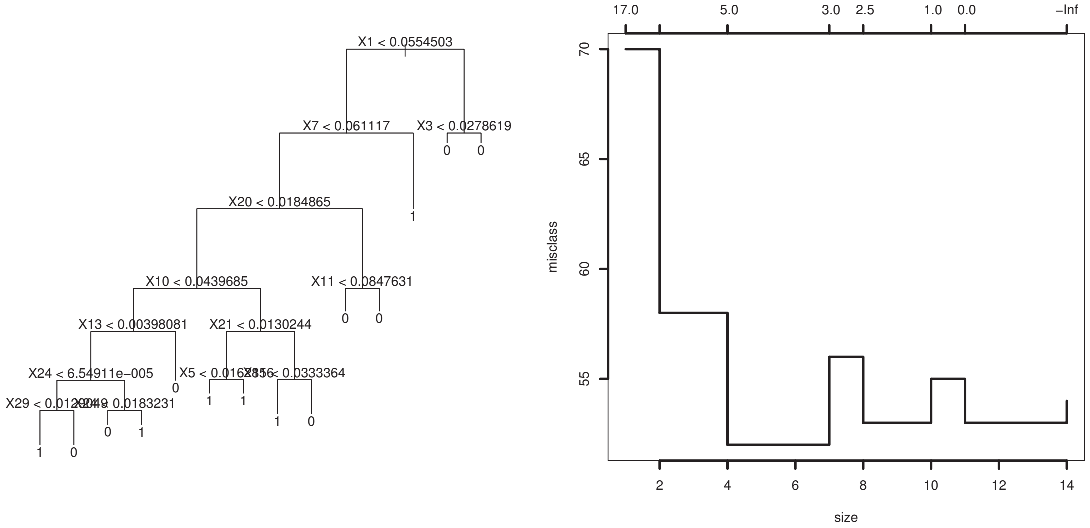


Figure 12: Left: A fitted full tree with 30 features Right: Leave-one-out cross validation score for the tree

subject to $\alpha_i \geq 0$ and $\sum_i \alpha_i Y_i = 0$. This problem could be solved by quadratic programming package. The final hyperplane could be written as

$$\hat{H}(x) = \hat{\alpha_0} + \sum_{i=1}^{n} \hat{\alpha_i} Y_i \langle x, X_i \rangle \tag{23}$$

3.8 Classification Performance Evaludation

To evaluate the performance of different classifiers, since Naive Bayes and Generalized Additive Model are mathematically equivalent [19], their performance are expected to be similar. We only consider Generalized Additive Model in the following analysis. Some classifiers, like Trees, are always trying to overfit the data, we need to prune it to control the model complexity. The first plot of figure 12 shows a fitted tree with 30 features. The second plot of figure 12 shows the leave-one-out cross validation plot with misclassification error as criterion. We see that only 4 nodes are needed. For the purpose of evaluation, simply divide the raw dataset into training and testing set is not enough. As shown in figure 13, in which we randomly split the raw dataset into two parts, 80 spectra for training and the remaining 90 for testing, and applied Logistic regression on it. The figures illustrates the relationship between the misclassification error and the number of features included for training the classifier, the line is quite bumpy, which is a sign that randomly split the data once may not be enough. For our analysis, 10-fold cross validation are used. ie. 10 percent data are left out as the validation set, the classifier are trained on the remaining 90 percent data and the prediction performance are evaluated on the validation set. After 10 iterations, all the data in the raw dataset have been utilized. The misclassification error on 10 parts will be averaged to get a 10-fold CV score. Table 1 shows the classification results for 8 classifiers with 10-fold cross validation (with 20 and 30 features respectively).

Table 1: Misclassification rates for eight classification methods applied to the Ovarian cancer dataset, with and without peak alginment.

	20	features	30	features
	Peak Aligned	Not Aligned	Peak Aligned	Not Aligned
Linear Discriminant Analysis	0.5176471	0.5588235	0.4705882	0.5823529
Quadratic Discriminant Analysis	0.4764706	0.5176471	0.4705882	0.5623592
Linear Regression	0.5176471	0.5764706	0.4705882	0.5823529
Logistic Regression	0.4941176	0.5764706	0.5588235	0.5470588
1-Nearest Neighbour Classifier	0.3588235	0.4352941	0.40	0.3947059
Generalized Additive Model	0.4941176	0.5588235	0.5588235	0.5529412
Support Vector Machine	0.3588235	0.4352941	0.40	0.4176471
Trees	0.5411765	0.5235294	0.4941176	0.5235294

For the practical purpose of cancer diagnosis, we do not want the number of biomarkers to be too huge, generally speaking, 10-30 biomarkers should be fine. Therefore, for the results we reported here, we only considered two cases: 20 features kept and 30 features. Also, to show that the peak alignment does really help, we compared the mis-classification rate with and without the peak alignment procedure. From this table, we see that Support Vector Machine and 1 Nearest Neighbor perform the best. When there are 20 features included, the misclassification rate is 0.3528 for both of them. From the analysis of Wu et al. [4], both SVM and LDA performed as the best method on another dataset. In our analysis, SVM is still also the best, which is consistent with their's results. However, LDA performs worese than SVM and 1-NN. One reason might be that we use a different dataset here, another reason may due to the difference between the peak alignment and feature selection procedures by RAMS with their approaches. The classification performance of different classifiers with the peak alignment as a preprocessing procedure are uniformly better than those without the peak alignment. This shows that the step for peak alignment does really helpful.

Figure 14 shows the boxplots for these classifiers when number of features are 20 and 30. The last boxplot is a dummy method as our baseline, it simply randomly guess 1 or 0 according to a Binomial distribution with parameter 0.5. From these two boxplots, it's obvious that SVM and 1-NN's performances are better than the remaining methods. If we assume that the underlying distribution for the CV score is Normal, a pairwise two sample t-test [20] is condcuted to test whether the improvement of classification performance is statistically significant or not.When feature number = 20, the 10-fold cross validation result from SVM are compared with that from the baseline method, the T-statistic is -3.4245 with 9 degrees of freedom. The obtained p-value = 0.007574, with level $\alpha = 0.05$, which is a strong evidence against the null hypothesis that these two groups have the same mean. Therefore, we conclude that this decreasement of the misclassification rate is statistically significant.

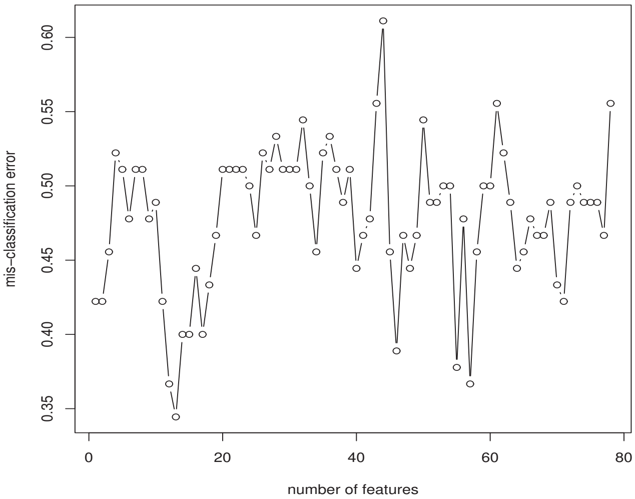


Figure 13: Misclassification rate of the Logistic Regression for different number of features

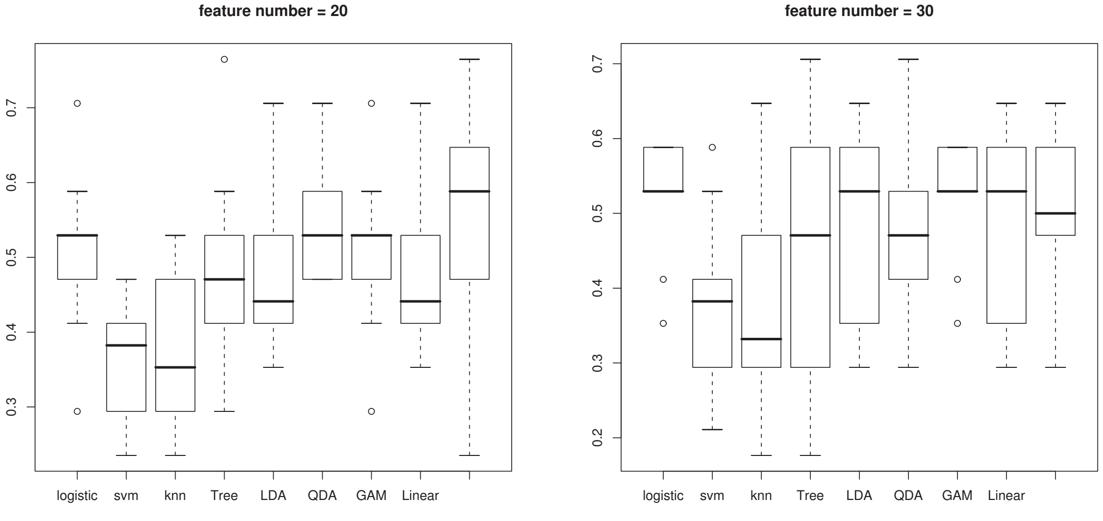


Figure 14: The boxplot for different classifiers for comparison, the last one is random guessing as a baseline

3.9 Step 8: Reported Biomarkers and internal reproducibility

With number of features = 20, there were 20 potential biomarkers detected, they were reported in table 2 according to the ranking:

Table 2: Reported potential biomarkers by 10-fold cross validation, the m/z and Intensity are values from the median spectrum

order	M/Z (Da)	Intensity
1	828.520	0.024475165
2	835.950	-0.032946414
3	880.700	-0.002384472
4	847.170	0.001361237
5	828.580	0.007853288
6	835.890	-0.039581336
7	768.270	0.035737468
8	825.710	-0.001525047
9	704.460	0.062437628
10	836.010	-0.030958591
11	871.330	0.024213808
12	884.360	0.039047785
13	880.660	-0.001213880
14	825.650	0.006621555
15	871.375	0.036258617
16	812.380	0.061968575
17	726.570	0.045650803
18	710.700	0.054321409
19	871.420	0.051560495
20	844.840	-0.031171868

This table was generated by 10-fold cross validation. For each fold, on the 90 percent of the data, RAMS select the most important 20 features. After an iteration of all the 10 folds, all the features are ranked by their show up frequency across these 10 folds. In table, some intensities are negative, this is resulted from the baseline correction of the median spectrum. We also noticed that almost all of the biomarkers are in the mass range 700 - 900. However, from figure 8, we found that the estimated variance for the estimated baseline curve is quite large in this area, there we do not have a high confidence about our estimation. Generally, a mass spectrometer can not achieve a stable model in the low m/z area. Therefore, whether the discriminative power of these potential biomarkers comes from the underlying biological process or simply comes from some artifactual noise is a meaningful topic which needs further investigation. Checked with 10 fold cross validation, we found that these biomarkers are internally reproducible within this dataset. Whether it's reproducible across different datasets or serum samples is still not clear.

4 Conclusion and Discussion

Cancer classification based on proteomic-profile data is a hard task due to the convolution of biological signals and artifacts noise. A series of preprocessing steps are needed to make the analysis valid: baseline correction, spectra normalization, peak detection and alignment, feature selection, data transformation could all affect the final classification performance. From a statistician's perspective, we proposed and developed a R package RAMS which intensively uses modern regression and classification methods as its module for cancer diagnosis based on MALDI-TOF or SELDI-TOF mass spectrometry data. Nonparametric kernel regression is used for baseline correction. Robust regression is used for spectra normalization. Natural cubic splines is used as an initial smoothing step for peak identification. Some modern methods for nonparametric variance estimation and confidence band construction are also included. Lasso regression, forward stepwise selection, and two sample t-test are mainly used for feature selection. Different classification methods, like nearest neighbor classifiers, naive Bayes, support vector machine, generalized additive model, linear/quadratic discriminant analysis, Trees are also included in RAMS for sample classification. Using RAMS, we analyze an Ovarian cancer dataset and achieved a classification result which is comparable with the other groups. For the purpose of biomarker discovery, instead of using the whole raw spectra, RAMS selected a subset of more simple and tractable features and try to classify on it. For some classifiers, like SVM, which may deal with the original high dimensional raw spectra directly based on some kernel tricks and the classification performance may achieve a little bit better performance. However, it's not natural for biomarker discovery, which is a very important goal for MALDI-TOF or SELDI-TOF mass spectrometry analysis.

Discussion - Classification with noise

The concept so-called "classification with noise" was first introduced into the area of proteomic-profile mass spectrometry analysis by Baggerly et al. [5] for a reanalysis of the SELDI-TOF datasets published by Petricoin et al. [13]. By some simple iterative method, they found multiple feature sets that could perfectly classify the spectra samples. They believed that this low misclassification rate is due to some artificial factors. e.g., a shift between the chip types or a change of electronic voltage, dominates the underlying patterns of the raw data. A perfect classifier trained with the pattern of these artifactual noise can not generalize well. Which makes the discovered biomarkers invalid. Our analysis did not deal with this issue explictly. How to generalize the RAMS framework to handle this problem is an interesting topic for the future investigation.

Acknowledgment

The author to this paper would like to thank Elly Kaiser for the help on efficient and large scale R programming.

References

[1] M. Wagner, D. Naik, and A. Pothen. Protocols for disease classification from mass spectrometry data. *Proteomics*, 3:1692–1698, 2003.

[2] K. R. Coombes, H. A. Fritsche, C. Clarke Jr, J. Chen, K. A. Baggerly, J. S. Morris, L. Xiao, M. Hung, and H. M. Kuerer. Quality control and peak finding fror proteomics data collected from nipple asirate fluid by surface-enchanced laser desorption and ionization. *Clinical Chemistry*, 49(10):1615–1623, 2003.

[3] Y. Yasui, D. McLerran, B. L. Adam, M. Winget, M. Thornquist, and Z.D. Feng. An automated peak identification/calibration procedure for high-dimensional protein measures from mass spectrometers. *Journal of Biomedicines and Biotechnology*, 4:242–248, 2003.

[4] B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, and H. Zhao. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 19(13):1636–1643, 2003.

[5] K. A. Baggerly, J. S. Morris, S. R. Edmonson, and K. R. Coombes. Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer. *J. Natl. Cancer Inst*, 97:307–309, 2005.

[6] K.R. Coombes, S. Tsavachidis, J.S. Morris, K.A. Baggerly, M. Hung, and H.M. Kuerer. Improved peak detection and quantification of mass spectrometry data acquiered from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Technical Report, The University of Texas M.D. Anderson Cancer Center*, 2004.

[7] R. Tibshirani, T. Hastie, B. Narasimhan, S. Soltys, G. Hi, A. Koong, and Q.T. Le. Sample classification from protein mass spectrometry by peak probability contrasts. *Bioinformatics*, 20(1):3034–3044, 2004.

[8] Y. Yasui, M. Pepe, M.L. Thompson, B. 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(3):449–463, 2003.

[9] W. Yu, B. Wu, N. Lin, K. Stone, K. Williams, and H. Zhao. Detecting and aligning peaks in mass spectrometry data with applications to maldi. *Applied Bioinformatics*, pages 626–633, 2004.

[10] S. Weisberg. *Applied linear regression.* Wiley series in Probability and Statistics, 2005. (Third Edition).

[11] C. Guoan, G. G. Tarek, H. Chiang-Ching, A. S. Dafydd, G. T. Kerby, M.G.T. Jereny, L. R. K. Sharon, E. M. David, J. G. Thomas, and D. I. Mark. Proteomic analysis of lung adenocarcinoma: identification of a highly expressed set of proteins in tumors. *Clinical Cancer Research*, 8:2298–2305, 2002.

[12] R.H. Lilien, H. Farid, and B. R. Donald. Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. *Journal of computational biology*, 10(6):925–946, 2003.

[13] E.F. Petricoin, A.M. Ardekani, B.A. Hitt, P.J. Levine, V. Fusaro, S. M. Steinberg, G.B. Mills, C. Simone, D. A. Fishman, E. Kohn, and L. A. Liotta. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, 359:572–577, 2002.

[14] M.K. Markey, G.D. Tourassi, and C.E. Floyd Jr. Decision tree classification of proteins identified by mass spectrometry of blood serum samples from people with and without lung cancer. *Proteomics*, 3:1678–1679, 2003.

[15] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Statistical Sciences*, 18(1):104–117, 2003.

[16] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: Data mining, inference, and prediction.* Springer series in statistics, 2001.

[17] B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, and H. Zhao. Ovarian cancer classification based on mass spectrometry analysis of sera. *Statistical Sciences*, 18(1):104–117, 2003.

[18] W.S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836, 1979.

[19] L. Wasserman. *All of Statistics : A Concise Course in Statistical Inference.* Springer Texts in Statistics, 2004.

[20] W.N.Venables and B.D. Ripley. *Modern Applied Statistics with S.* Springer: Statistics and Computing, 2002.

[21] K.A.Baggerly, J.S.Morris, and K.R.Coombes. Reproducibility of seldi-tof protein patternsin serum: comparing data sets from different experiments. *Bioinformatics*, 20(5):777–786, 2004.