

Sample classification from protein mass  
spectrometry,  
by “peak probability contrasts”

Robert Tibshirani<sup>\*</sup>, Trevor Hastie<sup>†</sup>, Balasubramanian Narasimhan<sup>‡</sup>,  
Scott Soltys<sup>§</sup>, Gongyi Shi<sup>¶</sup>, Albert Koong<sup>||</sup>, Quynh-Thu Le<sup>\*\*</sup>,

June 3, 2004

**Abstract**

**Motivation**

Early cancer detection has always been a major research focus in solid tumor oncology. Early tumor detection can theoretically result in lower stage tumors, more treatable diseases and ultimately higher cure rates with less

---

<sup>\*</sup>Depts. of Health, Research & Policy, and Statistics, Stanford Univ, tibs@stat.stanford.edu

<sup>†</sup>Depts. of Statistics, and Health, Research & Policy, Sequoia Hall, Stanford Univ., CA 94305. hastie@stat.stanford.edu

<sup>‡</sup>Depts. of Statistics, and Health, Research & Policy, Sequoia Hall, Stanford Univ., CA 94305. naras@stat.stanford.edu

<sup>§</sup>Dept. of Radiation Oncology, Stanford Univ., 94305. sgsoltys@stanford.edu

<sup>¶</sup>Dept. of Radiation Oncology, Stanford Univ., 94305. gshi@stanford.edu

<sup>||</sup>Dept. of Radiation Oncology, Stanford Univ., 94305. akoong@stanford.edu

<sup>\*\*</sup>Dept. of Radiation Oncology, Stanford Univ., 94305. qle@stanford.edu

treatment-related morbidities. Protein mass spectrometry is a potentially powerful tool for early cancer detection.

We propose a novel method for sample classification from protein mass spectrometry data. When applied to spectra from both diseased and healthy patients, the “peak probability contrast” technique provides a list of all common peaks among the spectra, their statistical significance, and their relative importance in discriminating between the two groups. We illustrate the method on Matrix-assisted laser desorption and ionization (MALDI) mass spectrometry data from a study of ovarian cancers.

### **Results**

Compared to other statistical approaches for class prediction, the peak probability contrast method performs as well or better than several methods that require the full spectra, rather than just labeled peaks. It is also much more interpretable biologically. The peak probability contrast (PPC) method is a potentially useful tool for sample classification from protein mass spectrometry data.

**Contact:** tibs@stanford.edu

## **1 Introduction**

Early cancer detection has always been a major research focus in solid tumor oncology. Early tumor detection can theoretically result in lower stage tumors, more treatable diseases and ultimately higher cure rates with less treatment-related morbidities. Many screening approaches have therefore been studied in solid cancers. Established screening tools for the early de-

tection of cancer include mammography for breast cancer, colonoscopy for colorectal cancer, PSA test for prostate cancer, and pap smear for cervix cancer (Smith et al. 2003). Imaging techniques such as chest X-ray and spiral computed tomography are also used, but are limited to a tumor size detection limit of 0.5-1.0cm (representing close to  $10^9$  cells) (Swenson et al. 2002).

Several serum markers have been identified through the years but, with a few exceptions such as prostate specific antigen (PSA) for prostate cancers and alpha fetal protein (AFP) for hepatocellular carcinomas, most have failed general integration into general clinical practice (Hansen & Pedersen 1986). Therefore, it is important to identify and to interpret new methods that provide sensitive and reliable diagnostic markers for solid cancers.

Recent advancements in proteomics have yielded novel and promising techniques to aid in biomarker identification (Hanash (2003), Petricoin et al. (2002a)). One such advancement is the development of protein mass spectrometry and the ability to analyze complex samples using this technique. SELDI-TOF (Surface Enhanced Laser Desorption/Ionization Time of Flight) and MALDI (Matrix-assisted laser desorption and ionization) mass spectrometry are the two most popular approaches presently employed for detecting quantitative or qualitative changes in circulating serum or plasma proteins in relation to a pathological state such as the presence of a solid tumor. Both represent high throughput and highly sensitive proteomic approaches that allows protein expression profiling of large sample sets (Hutchens & Yip (1993), Merchant & Weinberger (2000)). Briefly, in SELDI, proteins of interest from biologically complex samples bind selectively to chemically modified affinity surfaces, with non-specifically bound impurities washed away. The retained

sample is complexed with an energy-absorbing molecule, and analyzed by laser desorption/ionization time-of-flight mass spectrometry, producing spectra of mass/charge ratio ( $m/z$ ).

MALDI is similar to SELDI except that it does not have the preselection or enrichment steps for certain proteins in the sample mixture by allowing fractionation based on prebinding to different surfaces or chemical coatings. In MALDI, the samples are mixed with a crystal forming matrix, placed on an inert metal target, and subjected to a pulsed laser beam to produce gas phase ions that traverse a field-free flight tube and then are separated by mass/charge ratio. There are theoretical advantages and disadvantages for each of these two approaches; however both have been applied to cancer detection in solid tumors with reported high sensitivity and specificity using a variety of statistical analyses (Petricoin et al. (2002b), Li et al. (2002), Qu et al. (2002), Rai et al. (2002), Adam et al. (2003), Yasui et al. (2003) and Wu et al. (2003)).

In this paper we propose a novel algorithm for pattern classification from protein spectra, and compare it to several other existing techniques. We primarily focus on the comparison of two diagnostic classes (e.g. healthy versus cancer), although our method can be generalized to more than two classes (details are given in the Appendix).

Since there are many possible approaches to this problem, it is important to discuss the desiderata for such a procedure:

1. It should focus on clearly detectable peaks in the spectra, at least for the initial analysis. While there may well be discriminative information in other parts of the spectra, peaks are more likely to represent isolatable

proteins, protein fragments or peptides

2. The method should account for the variation in the  $m/z$  location and heights of the same biological peak in different spectra. The source of this variation may be biological or technical (i.e. due to properties of the mass spectrometer).
3. It should give some measure of discriminatory power for all peaks.
4. If possible, the sample classification rule should use the peak information in a relatively simple way and provide a method for filtering out the less significant peaks.

Point 3 can be important in the following scenario: suppose that other researchers, studying the same disease, find a potentially important peak at a certain  $m/z$  value. You would want to be able to assess the importance of that peak (or a nearby peak) in your data, and hence need an evaluation of all peaks found in your data.

## **2 Methods**

### **2.1 Sample description**

The ovarian cancer dataset was analyzed in Wu et al. (2003), and was provided by the authors. It consists of MALDI-MS spectra generated from a Micromass MALDI-R instrument on pre-treatment serum samples of 89 subjects, consisting of 42 non-cancer controls and 47 ovarian cancer patients.

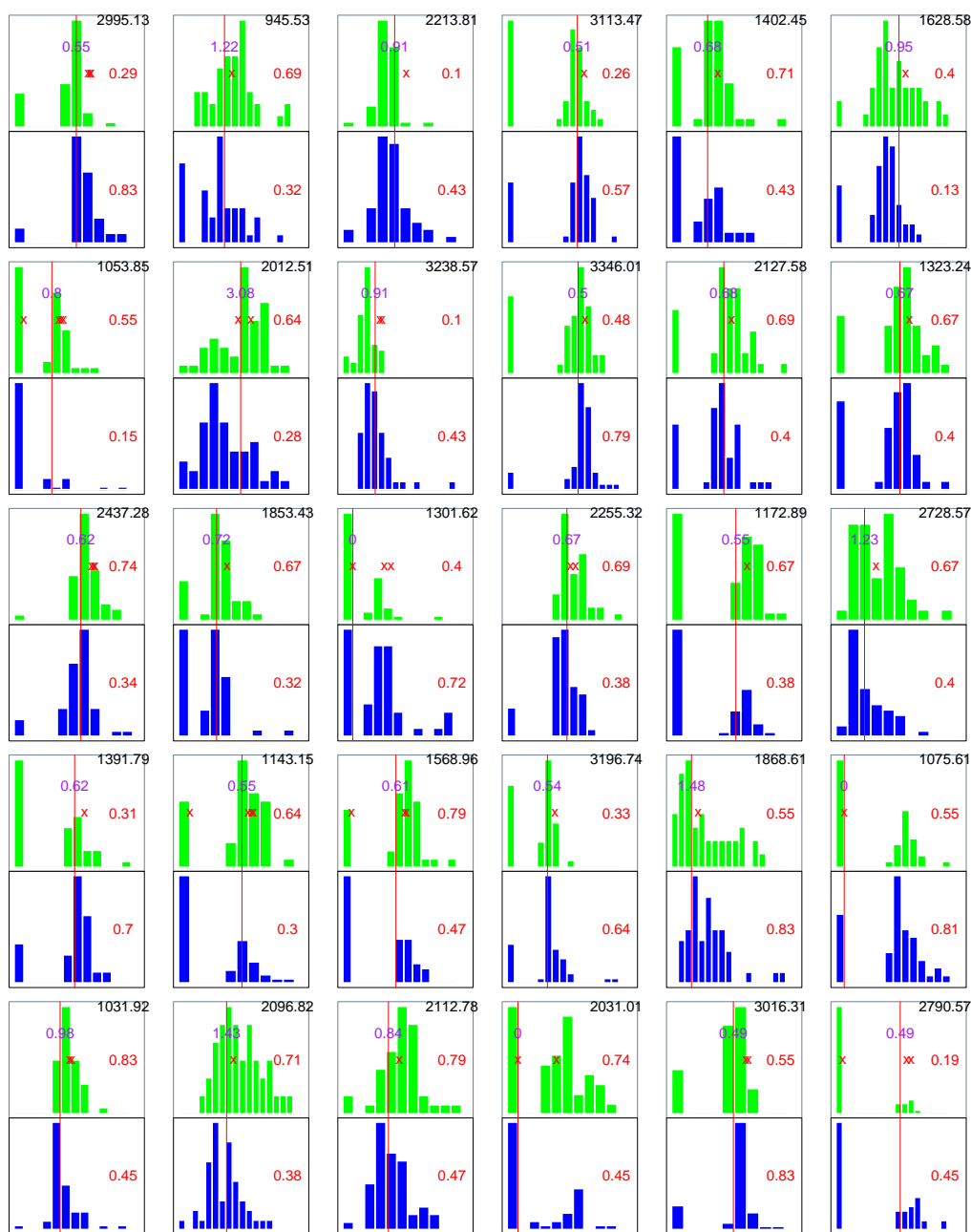


Figure 1: Results of PPC method on the ovarian cancer example. Each panel shows a histogram of peak heights in the training set at one site ( $m/z$  value in black type in top right corner), for healthy patients (green) and cancer patients (blue). Figure 2 gives details of the format. The peaks are ordered from strongest to weakest, as measured by the difference in proportions (red type), starting in the top left corner and moving down the left column. Only the top 30 peaks are shown, out of a total of 192 peak sites.

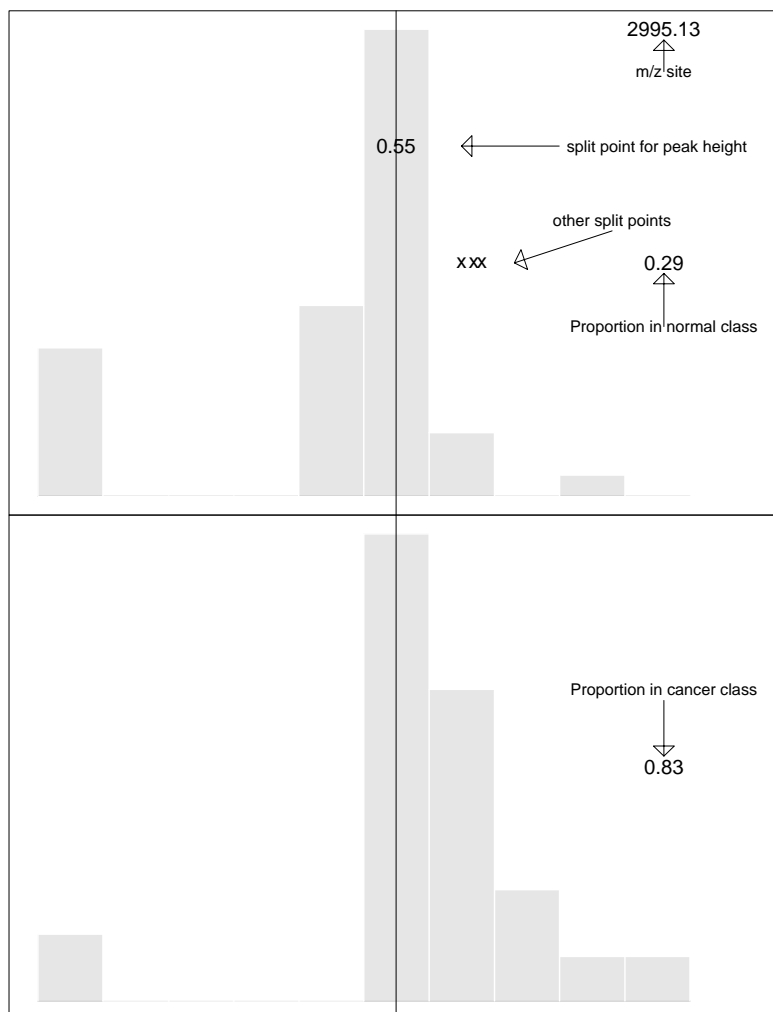


Figure 2: Exploded view of the top left panel of figure 1, with a legend detailing the format. The vertical line shows the estimated optimal height split point. The proportions of samples in each class having peaks higher than the split point are indicated. The “x”s indicate the horizontal positions of split points that achieve a difference in proportion within 10% of the best at that site.

The MS spectra are measured at 91,360 sites, spaced 0.019 Da apart and extending from 800 to 3500 Da. Following Wu et al. (2003) we log-transformed the intensities and then did a baseline subtraction using a “loess” smoother with span of 1000/91360. Finally, we normalized each spectrum by a linear transformation that mapped the 10th and 90th percentiles to 0 and 1 respectively.

A flowchart of the peak probability contrast (PPC) procedure is shown in Figure 3. We now describe the individual steps in detail.

## 2.2 (a) Peak extraction

We begin with the raw MALDI spectra. In some systems, the spectrometry software provides a list of labelled peaks. These were not available for our data, so, we developed a simple peak-finding procedure based on the ideas of Yasui et al. (2003). It looks for sites ( $m/z$  values) whose intensity is higher than that at the  $\pm s$  sites surrounding it, and higher than the estimated average background at that site; here we used the value  $s = 100$ .

First we smoothed the raw spectra, as illustrated in Figure 6. For this we used a “supersmoother” with a span of .002. This step would normally only be carried out for MALDI data, and not SELDI data. It has the effect of smoothing over the isotopic envelop that is present in MALDI data, which is helpful for the purposes of finding peak locations. However after determining the peak locations that are important for sample classification, one should examine the raw spectra to determine the actual width and location of the primary peak in each envelop. Alternatively, one could apply a de-isotoping method to extract the primary peak from each envelop and eliminate the

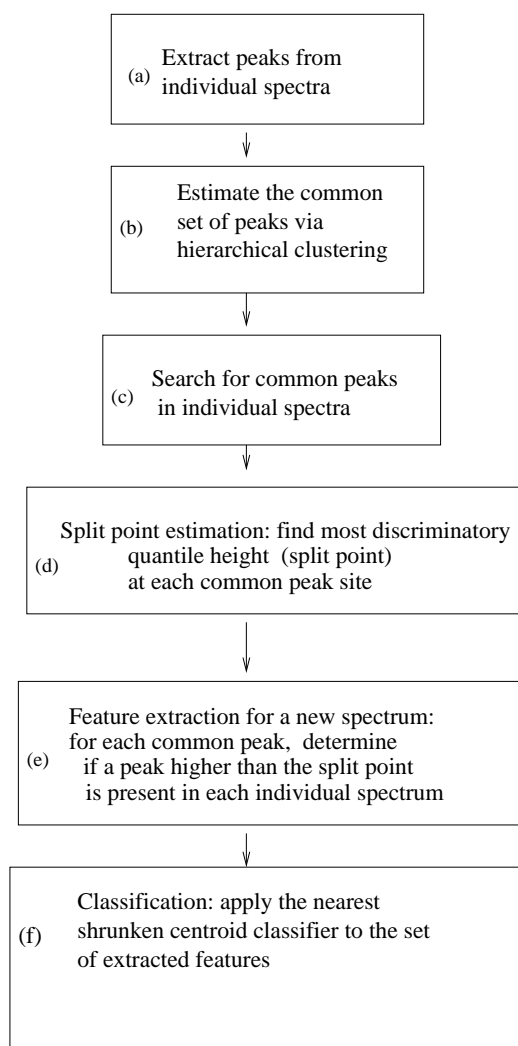


Figure 3: *Flow chart of PPC analysis.*

secondary ones. The output of a de-isotoping peak finder can be fed directly into our procedure.

We estimated that in the smoothed spectra, that peak widths were approximately 0.5% of the corresponding  $m/z$  value. Hence we log-transformed the  $m/z$  values so that the peak widths were approximately constant over the entire range. This produced a roughly constant peak width of 0.005. This same approach and peak width has been used by other authors, e.g Yasui et al. (2003). Visual examination of the individual spectra showed that this peak width was fairly reasonable for these data. By log-transforming the data, the peak widths are approximately constant across the  $m/z$  range and this facilitates application of a clustering procedure, described next. In general, the peak width is an important adjustable parameter in our procedure. The data analyst should try varying it, and examine the results both visually and in terms of the cross-validated misclassification (described later). In the ovarian cancer example in this paper, smaller widths such as 0.25% resulted in slightly higher error rates.

In some cases, two peaks within 0.5% are found in the same spectra, and these are combined. Note that any peak finding method can be used to provide peaks for the PPC procedure. The one we have used is crude, and a more refined peak-finder could yield improved classification results.

### **2.3 (b) Peak alignment via clustering**

To align peaks from the set of spectra, we applied complete linkage hierarchical clustering to the collection of all 14,067 peaks from the individual spectra. The clustering here is somewhat novel: it is one dimensional, using

the distance along the  $\log m/z$  axis. This is depicted in Figure 4.

The idea is that tight clusters should represent the same biological peak that has been horizontally shifted in different spectra. We then extract the centroid (mean position) of each cluster, to represent the “consensus” position for that peak across all spectra,

Since this clustering can be performed very quickly, a special routine was written for this purpose. Cutting off the dendrogram at height 0.005 produced 192 clusters with corresponding cluster centers taken as the midpoints between the ranges of the cluster. Since complete linkage was used, we are guaranteed that every peak in the cluster is at most 0.005 from any other peak in that same cluster.

## 2.4 (c) Search for common peaks in individual spectra

Given the list of common peaks from clustering in step (b), we go back to the individual spectra and record whether each spectrum exhibits each of these common peaks. A peak in the individual spectra is deemed to one of the common peaks if its center lies within  $\log(.005)$  of estimated center position of the common peak. If it is present, the height of the individual peak in the spectrum is also recorded.

## 2.5 (d) Split point estimation for each peak

From the previous steps, we have spectrum peak heights  $y_{ij}$ , for observations  $j = 1, 2, \dots, n$  and sites  $i = 1, 2, \dots, m$ . These are the centroids from a hierarchical clustering of all of individual spectra peaks. If there is no peak at site  $i$ , we take  $y_{ij} = 0$ . In this step we cut the peak height at some quantile,

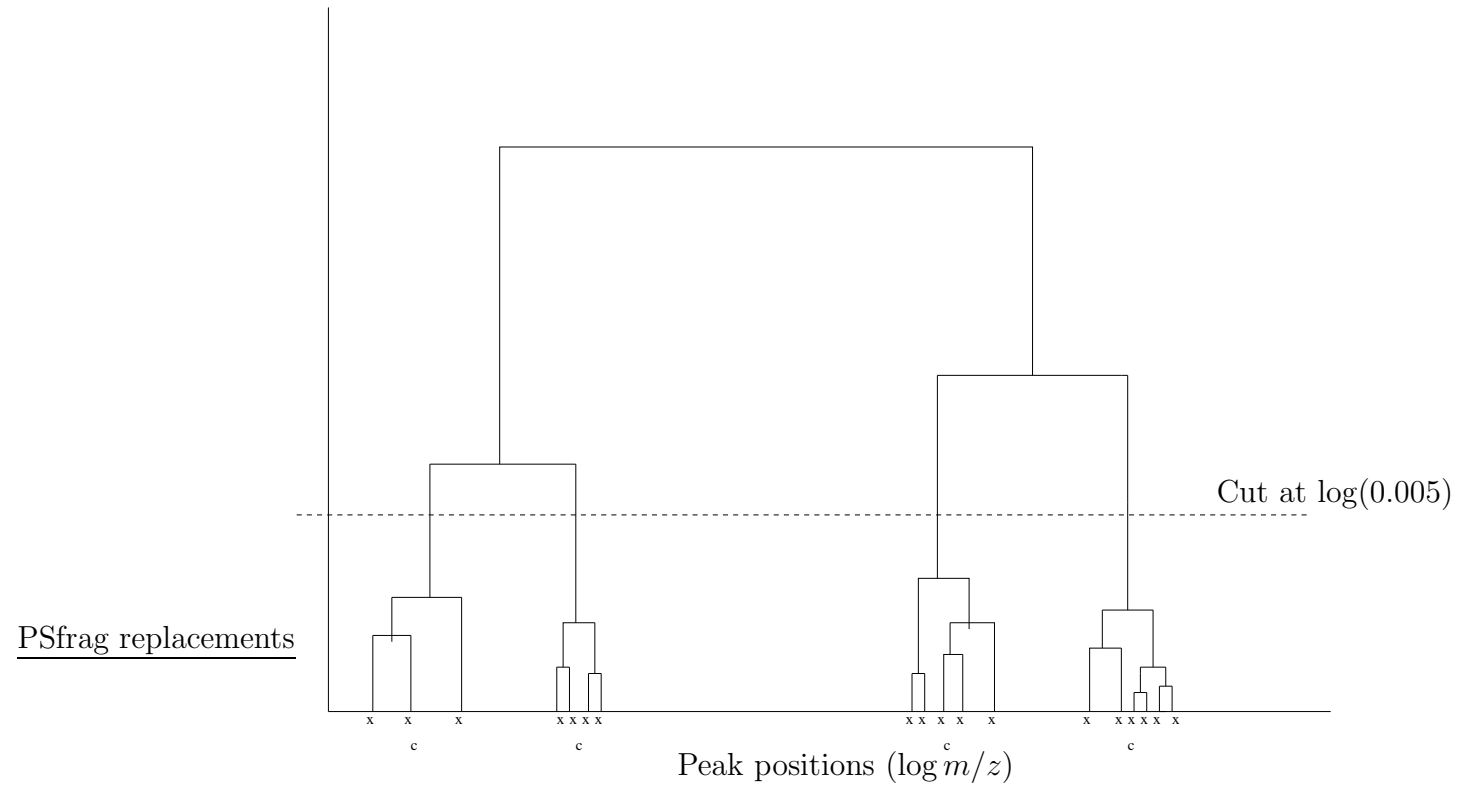


Figure 4: *Illustration of hierarchical clustering for peak alignment and clustering. The points marked “x” represent the positions of extracted peaks from the individual spectra. Complete linkage hierarchical clustering is applied to the peak positions along the  $\log(m/z)$  axis, and the resulting dendrogram (clustering tree) is cut at height  $\log(0.005)$ . In this simple illustrative example, this process produced four clusters with associated centroids indicated by a “c”.*

in such a way as to maximally discriminate between the healthy and normal samples in the training set. Basing the splits on the quantiles of all heights at a peak position, rather than absolute peak heights, is important: it accounts for the fact that peaks heights can vary greatly across the  $m/z$  range. Here are the details:

- Let  $q(\alpha, i)$  be the  $\alpha$  quantile of the peaks  $y_{ij}$  at site  $i$ .
- Given two groups  $G_1, G_2$  of size  $n_1, n_2$ , let  $p_{ik}(\alpha)$  be the proportion of spectra in group  $k$  with a peak at site  $i$  larger than  $q(\alpha, i)$ :

$$p_{ik}(\alpha) = \sum_{j \in G_k} I[y_{ij} > q(\alpha, i)] / n_k, \quad k = 1, 2$$

Where  $I[\cdot]$  is the indicator function, equally one if the event is true and zero otherwise.

- Choose  $\hat{\alpha}(i)$  to maximize  $|p_{i2}(\alpha) - p_{i1}(\alpha)|$  and set  $\hat{p}_{ik} = p_{ik}(\hat{\alpha}(i))$ .

This process produced the cutpoints (red vertical lines) and class probabilities  $\hat{p}_{ik}$  shown in Figure 1. The panels in the Figure are arranged in decreasing strength, that is, decreasing value of  $|\hat{p}_{i2} - \hat{p}_{i1}|$ . These histograms are informative in themselves. For some sites (e.g at 1301.62 in the third leftmost column), the cutpoint divides height=0 from the rest. That is, it indicates that the presence or absence of the peak is what's important. At other sites (e.g. 2995.1 in the top left corner), the proportion of peaks above a certain height (0.55) is important for classification ability.

## 2.6 (e) Feature extraction for a new spectrum

From the previous steps we have a set of common peaks, and an optimal discriminating split point for the height of each peak. To do class prediction for a new spectrum, we first construct a vector of binary features for that spectrum, one for each of the common peaks. Each feature equals one if a peak with height greater than the split point is found in the new spectrum, and zero otherwise. As before, a peak is considered to correspond to a common peak if its center lies within  $\log .005$  of the position of the common peak. In Figure 1, the first feature will equal 1 if the new spectrum contains a peak at 2995.12 higher than 0.55,, the second feature will equal 1 if the new spectrum contains a peak at 1053.85 higher than 0.80, and so on.

## 2.7 (f) Class prediction via nearest shrunken centroids

Here we show how to use the peak proportions  $\hat{p}_{ik}$  to classify a new spectrum into class 1 (healthy) or class 2(diseased). Given a spectrum from a new patient with peak heights  $y_1^*, y_2^*, \dots, y_p^*$ , let  $z_i^* = I[y_i^* > q(\hat{\alpha}(i), i)]$ . This is the binary feature vector from step (e), with a component equal to one if the spectrum has a peak above the cutpoint height at that site, and zero otherwise. We can then compare this binary profile to each of the probability centroid vectors  $(\hat{p}_{11}, \hat{p}_{21}, \dots, \hat{p}_{m1})$  and  $(\hat{p}_{12}, \hat{p}_{22}, \dots, \hat{p}_{m2})$  and predict to the class that is closest in overall squared distance (or some other metric)<sup>1</sup>. This is a kind of “nearest centroid” classification. However to select sites and potentially improve the prediction performance, we also consider shrinkage

---

<sup>1</sup>Our software also allows the use of absolute distance or binomial log-likelihood distance

of each pair of probabilities  $\hat{p}_{i1}, \hat{p}_{i2}$  towards their average.

Figure 5 shows a hypothetical example of nearest shrunken centroid classification in action.

Before giving details, the method is illustrated by the example shown in Table 1 (details in table caption).

Here are the details. Let  $s(t, \Delta) = \text{sign}(t)(|t| - \Delta)_+$ , the “soft-threshold” function. Here “+” means positive part. The soft-threshold function translates the value  $t$  towards zero by the amount  $\Delta$ , setting it to zero if  $|t| \leq \Delta$ . For example if  $\Delta = 0.5$ , then  $s(1.2, \Delta) = 0.7$ ,  $s(-1.2, \Delta) = -0.7$ ,  $s(.3, \Delta) = 0.0$ . Then we set  $\tilde{p}_{ik} = \bar{p}_i + s(\hat{p}_{ik} - \bar{p}_i, \Delta)$ , with  $\bar{p}_i = (\hat{p}_{i1} + \hat{p}_{i2})/2$ .

The parameter  $\Delta$  is chosen by tenfold cross-validation. That is, we divide the samples into 10 approximately equal sized parts. For each fixed value of  $\Delta$  we train the PPC algorithm on 9 parts of the data and then compute the error rate in predicting the class labels of the samples in the tenth part. This is done for each of the ten parts in turn, and the error rates added to give, the cross-validation error estimate for the value  $\Delta$ . This process is carried out for a grid of values of  $\Delta$ , to produce an error curve  $\text{cv}(\Delta)$ . Finally we examine this curve and choose  $\Delta$  to be its minimizer  $\hat{\Delta}$ .

Notice that if the probabilities are shrunken so that they coincide, the site  $i$  no longer contributes to the nearest centroid rule. Sample classification is then done as follows. For a test set with peak heights  $y_1^*, y_2^*, \dots, y_p^*$ , let  $z_i^* = I[y_i^* > q(\hat{\alpha}(i), i)]$  and compute the distances  $d_k = \sum_i (z_i^* - \tilde{p}_{ik})^2$  (or absolute value). We predict to class 2 if  $d_2 < d_1$  and class 1 otherwise. Estimated class probabilities are also available, derived as in Tibshirani et al. (2003). Note that *all* of the training steps, including split-point estimation,

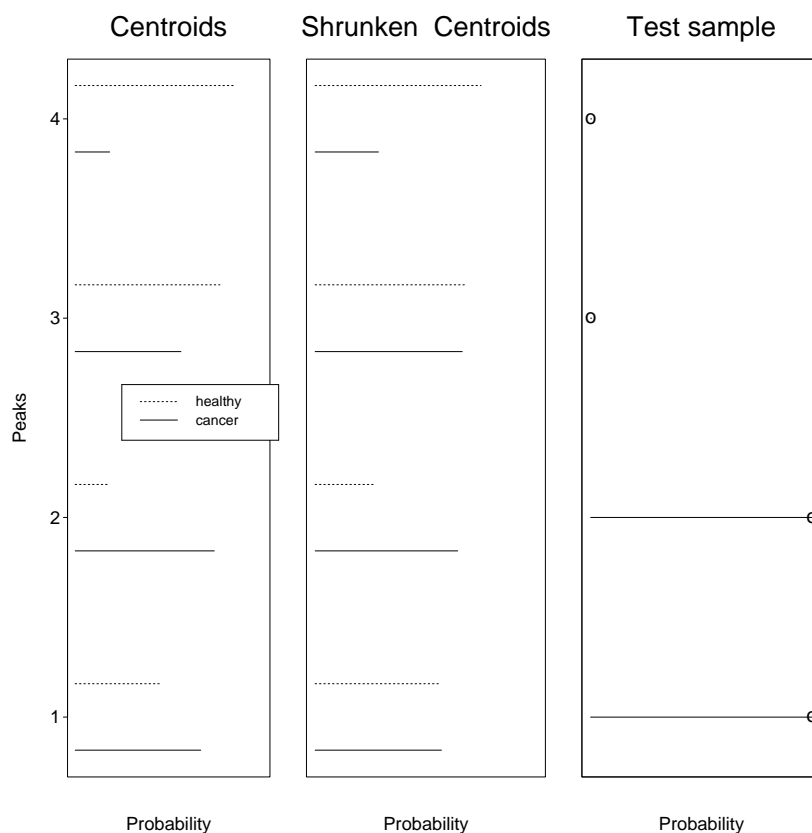


Figure 5: *Hypothetical example of nearest shrunken centroid classification in action. There are four peaks shown in the left panel: peaks 1 and 2 appear more often in the cancer group than the healthy group, while the reverse is true for peaks 3 and 4. In the middle panel the probabilities in the two classes have been shrunken towards each other. As a result, the probabilities for peaks 1 and 3 are now equal, and those peaks will not participate in the class prediction of a new spectrum. The right panel shows the feature vector (consisting of zeroes and ones) for a test spectrum: it has peaks 1 and 2, but not 3 or 4. To predict the class for this spectrum, we compare its feature vector to the healthy and cancer profiles in the middle panel, and find the closest one in squared distance. Here the closest is the cancer profile, and so the prediction is to class “cancer”.*

Peak #	Unshrunk centroids		Shrunk centroids		Feature vector
	Normal	Cancer	Normal	Cancer	
1	0.29	0.83	0.48	0.64	1
2	0.55	0.15	0.36	0.34	0
3	0.74	0.34	0.55	0.53	0
4	0.31	0.70	0.50	0.51	1
5	0.83	0.45	0.64	0.64	0
6	0.69	0.32	0.50	0.50	0
7	0.64	0.28	0.46	0.46	0
8	0.67	0.32	0.50	0.50	0
9	0.64	0.30	0.47	0.47	1

Table 1: *Illustrative example of nearest centroid classifier used in PPC method. The raw (unshrunk) class centroids, are shown in the 2nd and 3rd columns. These are the proportion of samples in each class with peaks higher than the optimal split point, at each site (in red type in figure 1). For illustration our example has only 9 peak sites. In reality there will usually be many more (192 in our ovarian cancer example). A typical feature vector from a new spectrum is shown in the rightmost column. This spectrum has a peak higher than the cutpoints at sites 1, 4 and 9. A nearest centroid classifier compares the feature vector to the two centroids, and predicts to the class to which it is closest in squared distance. In this case we predict to class “Cancer”. We can often improve upon this classifier by shrinking the centroids towards each other by an amount  $\Delta$ . Here we chose  $\Delta = .19$ , producing the shrunk centroids in columns 4 and 5. Our prediction is again based on nearest centroids, but now using the shrunk centroids. The probabilities at peaks 5–9 have been shrunk together, and so the prediction is based only on the first 4 peaks. In this case the prediction is still to class “Cancer”, but we have simplified the model.*

are repeated within each cross-validation fold.

## 2.8 False discovery rates

The simple difference in proportions for peak  $i$

$$T_i = \hat{p}_{i2} - \hat{p}_{i1} \quad (1)$$

can be used to assess the significance of the peak. False discovery rates (Benjamini & Hochberg 1985), (Tusher et al. 2001), (Efron & Tibshirani 2002), (Storey & Tibshirani 2003) are a useful measure for this. For a given threshold  $t$  we compute the number of  $T_i$  that exceed  $t$  in absolute value. Then we randomly permute the class labels and apply the PPC procedure to the spectra with permuted labels, giving scores  $T_1^{*b}, T_1^{*b}, \dots, T_m^{*b}$ . This process is repeated  $B$  times, producing scores for  $b = 1, 2, \dots, B$ . Finally the false discovery rate is estimated by

$$\widehat{\text{FDR}}(t) = \frac{\sum_{i=1}^B I(|T_i^{*b}| > t)/B}{\sum_{i=1}^B I(|T_i| > t)}. \quad (2)$$

The numerator is an estimate of the number of false positive peaks, and hence the ratio estimates the proportion of false positives among the peaks called significant. We estimate  $\text{FDR}(t)$  in this way, for a range of values of the threshold  $t$ . From this, we find the threshold  $t$  giving a reasonable low FDR (say 5%) and call significant all peaks  $i$  that fall beyond this threshold. Note that the estimation of FDR is only for descriptive purposes, and is not used formally in the sample classification process.

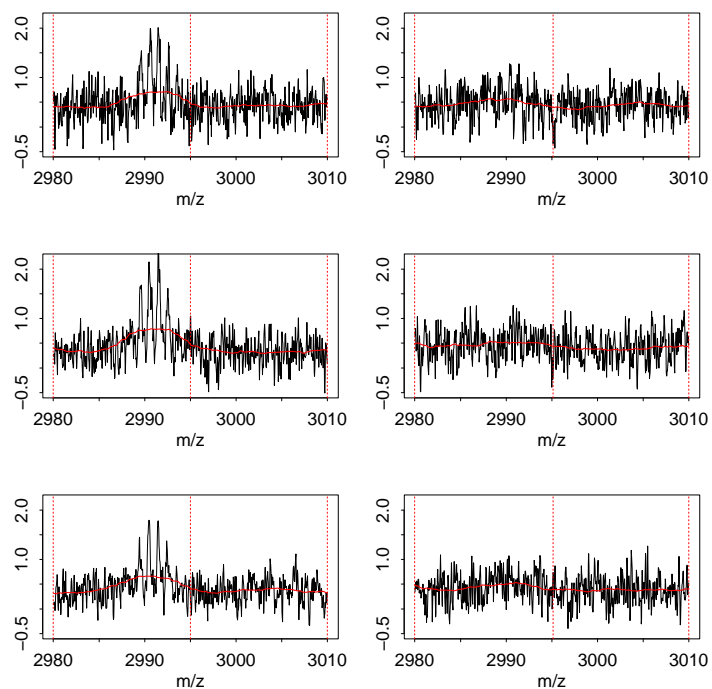


Figure 6: *Left Column: Three spectra from cancer patients having a peak higher than .55 at the site  $m/z = 2995.1$ . Both the raw (black) and smoothed (red) spectra are shown. In the right column, we show three spectra healthy patients without the peak, or whose peak is too low. The vertical dotted lines indicate the centroid 2995.1 and the outer limits for the peak position.*

## 2.9 Use of other classifiers

The features derived from the PPC method can also serve as useful inputs into other classifiers. We have chosen the nearest shrunken centroid method as our primary classifier, because of its simplicity and interpretability. But other methods have potential advantages in this context. For example the lasso (Tibshirani 1996), (Efron et al. 2002) is a multivariate fitting method that produces a sparse set of feature weights, and could potentially improve the prediction performance of nearest shrunken centroids in this setting. A binary decision tree (Breiman et al. 1984) can find subgroups of cancer or healthy patients, defined in terms of their the individual peak behavior.

# 3 Results

## 3.1 Ovarian cancer MALDI dataset

The peak extraction step found a total of 14,067 peaks, or an average of 158 peaks per spectrum. These were then clustered into 192 groups of peaks.

Figure 1 summarizes the peak information in the first 30 of these 192 peak clusters 1. Each box shows a histogram of peaks at the given  $m/z$  site, in the non-cancer (green) and cancer (blue) classes. The different sites are arranged from strongest to weakest, starting at the top left, and moving down the left column.

Figure 2 show an exploded, annotated view of the top left box. The optimal split point for each site is indicated by a vertical red line, and the resulting proportions to the right of that split point are shown in the red

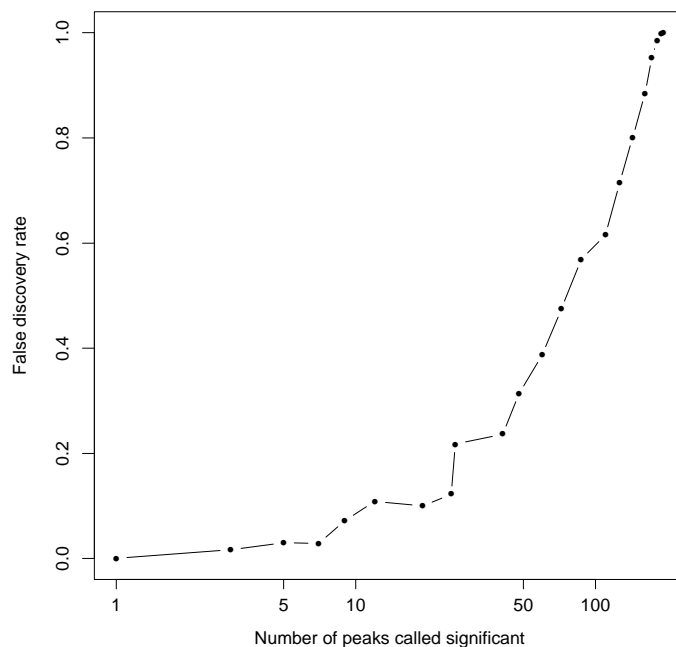


Figure 7: *Estimated false discovery rate, as a function of the number of peaks called significant.*

numbers in the box. For example, the strongest site is at  $m/z = 2995.1$ , with a much larger proportion of cancer patients having peaks above that split point, compared to control patients (0.70 vs. 0.17). (The split at this site actually corresponds to a peak height of .58). Figure 6 shows an example of three spectra in each group, at the strongest site  $m/z = 2995.1$ . Figure 7 displays the false discovery rate, as the number of significant sites is varied. The false discovery rate starts to rise above 0.05 after the first 7 or 10 peaks. Hence the strongest 10 peaks are very likely to be significant, but we are less certain about the peaks farther down the list.

Method	Cross Validation errors/89 (se)	Number of sites
(1) PPC	23(1.1)	7
(2) PPC/width .0025	27(1.7)	3
(3) PPC/pres-abs	30(1.8)	133
(4) PPC/lasso	25(1.5)	192
(5) LDA/t-15	31(1.4)	15
(6) SVM/t-15	27(1.6)	15
(7) SVM	21(1.4)	91360
(8) Wavelets	26(1.3)	91360

Table 2: *Results for ovarian cancer example. Methods are 1) peak probability contrast with default peak width of 0.005; 2) PPC with peak width .0025; 3) PPC with splits restricted (i.e. peak present or absent), 4) lasso applied to binary features from PPC method; 5) linear discriminant analysis using the top 15 sites as ranked by the t-statistic, 6) support vector machine using these same 15 sites, 7) support vector machine applied to all sites, 8) nearest shrunken centroid classifier applied to wavelet coefficients.*

Figure 8 shows a heatmap display of the top 7 peaks in the 89 samples.

The tenfold cross-validated misclassification rate for the PPC method is shown in line (1) of Table 2. The minimum CV error is achieved with 7 peaks. The cross-validated sensitivity and specificity were 35/47 and 30/42 respectively. In Line (2) we halved the peak width to 0.0025: this seems to hurt the prediction accuracy. In line (3) we have restricted the splits so that the contrast represents presence versus absence of a peak. The error rate has increased. In line (4) we have applied the lasso to the binary features from

the PPC method. This does not seem to improve prediction accuracy of the PPC method in this problem.

Lines (5) and (6) of the table represent two of the best performing methods among those of those studied by Wu et al. (2003). Both methods start with the 15 sites having the largest t-statistics in absolute value. The first (LDA) is a linear discriminant analysis based on these 15 features, while the second (SVM) is a support vector classifier. For the SVM we optimized over the choice of its cost parameter. Both LDA and SVM perform worse than the PPC method here. In line (7) we applied SVM to all sites. Its prediction performance might be a little better than that of PPC. In line (8) we applied a discrete wavelet transform with Daubuchies compact wavelets to each spectrum, using the Wavethresh3 package in R (Nason 1998). We then applied the nearest shrunken classifier to the resulting wavelet coefficients. The classifier used only 6 wavelet coefficients, but when transformed back to the original domain, it resulted in non-zero weights for all 91360 features. This procedure is analogous to the PPC method in line (1), but uses a different feature extraction (encoding). We see that the error rate is no better than that of PPC, and it uses many more features.

Figure 9 shows the cross-validation error curves for the PPC methods, as a function of the threshold parameter  $\Delta$ . The number of sites is indicated along the top of the figure. We have include the PPC/lasso method on the plot, using the number of non-zero sites as the plotting abscissa.

We note that the CV error for LDA and SVM/t-15 reported in Wu et al. (2003) averaged about 12–15%, or 12–14 errors out of 89. This is far better than the results in table 2. But in their study, these authors used all 89

	Number in quartile					Total
	0	1	2	3	4	
Healthy	6	16	12	4	4	42
Cancer	3	4	8	16	16	47

Table 3: *Training set results for peak at  $m/z = 2995.1$ . Number of samples in quartiles of peak heights; quartile “0” means no peak.*

samples to choose the 15 sites, and then applied cross-validation keeping the 15 sites fixed (personal communication). This produces an unrealistically low error rate that does not accurately estimate the true test error rate.

The strongest peak used by PPC was at  $m/z = 2995.1$ . The corresponding peak heights are shown in Table 3 and show a strong trend towards for larger peaks in the cancer spectra. The t-statistic at  $m/z = 2995.1$  was 3.19. Among the 91360 t-statistics, the value 3.19 ranks as only the 4196th largest. Hence it is not clear that screening on the value of the t-statistics is a good way to choose features in this example.

### 3.2 An artificial spiking experiment

To assess the performance of the PPC algorithm, we created artificially “spiked” spectra from our original data. First we created artificial control and cancer datasets, each consisting of approximately half of the original control and cancer patients, respectively. By construction these artificial datasets were heterogeneous but were similar to each other. We then chose two sets of 5 sites:

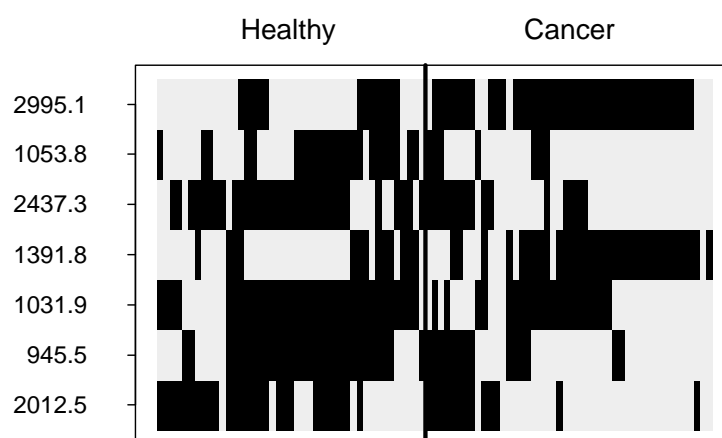


Figure 8: *Binary heatmap of top 7 training set features. Each row corresponds to one of the 7 peak centroids, and each column corresponds to a training spectra. A pixel is dark if the peak at that site exceeds the threshold determined by PPC. The rows (peaks) are ordered by decreasing strength from top to bottom.*

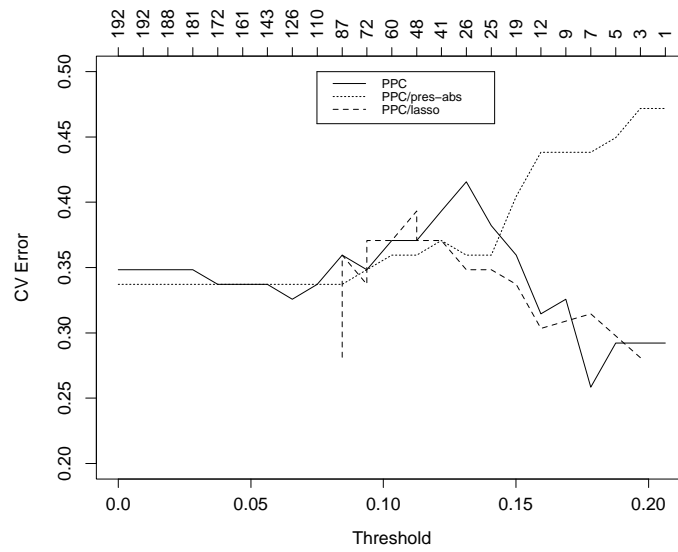


Figure 9: *Tenfold cross-validation error for PPC, as a function of the threshold parameter  $\Delta$ . The corresponding number of peaks used is shown along the top of the figure. We have included the PPC/lasso method on the plot, using the number of non-zero sites as the plotting abscissa.*

$f$	10 site model		full model	
	# sites found	Test errors /45	# sites found	Test errors /45
2	7	0	10	20
1	4	3	8	24
0.5	3	8	10	21

Table 4: *Results for artificial spiking experiment.*

820.0 1106.7 1680.0 2540.0 3113.3 for control samples

1393.3 1966.7 2253.3 2826.7 3400.0 for cancer samples

An artificial peak was spiked into each control spectra at the first 5 sites, and spiked into the cancer spectra at the second five sites. In each case this peak was a narrow spike of width 1. To simulate actual conditions, the position of the spike was also “jittered” by 0.0025 from the target site.

In detail, if  $h(x)$  is the intensity of the spectrum at  $x = \log(m/z)$ , then the height of the spectrum at  $x'$  after spiking was defined to be

$$h(x') + \bar{h}(x') \cdot f \quad (3)$$

where  $\bar{h}(x')$  is the average intensity at  $x'$  for all spectra, and  $f$  is a fraction equal to 2, 1, or 0.5. Here  $x'$  is the jittered version of  $x$ , that is,  $x' = x + U$ , where  $U$  is uniformly distributed on  $[-0.0025, 0.0025]$

After creating the dataset, we randomly split it into a training set (2/3) and a test set (1/3). We applied the PPC procedure giving the results in Table 4. In the 2nd and 3rd columns, the PPC model was shrunken down to its top ten sites, and we report how many of the actual ten spiking sites appeared in among these ten and the resulting test set error rate. By “ap-

peared” we mean that the actual spiking site was within 0.005 units of one of the centroids. Similarly, the last two columns report the results for the full (unshrunk) PPC model. This model typically had on the order of 600 centroid sites.

As expected, as we spike with smaller peaks, both the ability to detect these peaks and the ability to use them for prediction, tends to decrease. However with peaks twice as high as the average intensity ( $f = 1$ ), the test error of the shrunken model is low (3/45) and the procedure finds 4 of the 10 spiking sites. The full model is better at finding the spiking sites (among others), but by retaining noisy sites it pays a price in test error.

## 4 Discussion

Sample classification from proteomic data is often difficult because the signal intensity for each  $m/z$  point can be affected by both biological processes and experimental condition variabilities. The preprocessing steps of MS output are critical for the overall analysis of proteomic data. Peak normalization, identification and alignment can all affect the performance of class prediction using conventional classification approaches. The proposed peak probability contrast method first extracts clusters of peaks in the spectra. In other experiments in our lab, we found that this extraction step appears to be robust and reproducible when tested on spectra obtained from different runs using the same plasma sample. It can help to minimize experimental variability.

After extraction of peak clusters, PPC uses resulting features in a nearest centroid classifier. These features can also serve as usefully inputs into other

classifiers such as a binary decision tree, or lasso model. Comparison with other classifiers that operate on the raw spectra, PPC's performance is just as competitive while providing an advantage of generating a simple, more interpretable set of features for further investigation. The efficiency of this method in finding a relative small number of peak clusters for class prediction will facilitate future identification of biologically significant and relevant proteins for tumor development and progression. Discovery of these proteins will result in novel targets for cancer prevention and antitumor therapies.

The concept of low molecular weight (LMW) serum proteome was recently introduced through the increasing popular mass spectral based proteomic analysis. Its importance was demonstrated by a number of studies (Petricoin et al. 2002b), (Kozak et al. 2003). However, because these LMW markers were mainly identified through bioinformatic/statistical analysis, their identities remain elusive. Two other ovarian cancer studies yield two panels of markers each with five  $m/z$  values (534, 989, 2111, 2251 and 2465 (Petricoin et al. 2002b); 4.4k, 15.9k, 18.9k 23.0k and 30.1k (Kozak et al. 2003). None of them show identical  $m/z$  value as the panel of markers we have shown here. This can be easily explained by different samples, handling process, instruments, and statistical tools used by these studies. Although not necessarily straight forward, there are ways to purify those serum markers for identification through tryptic peptide mapping (Rai et al. 2002) or amino acid sequencing (Klade et al. 2001).

Software for performing the Peak Probability Contrast analysis will be available at

<http://www-stat.stanford.edu/~tibs/PPC>

**Acknowledgments:** we would like to thank Baolin Wu and Hongyu Zhao for sharing their data on ovarian cancer, and Yutaka Yasui for providing details on his peak finding algorithm. We would also like to thank an Editor and reviewers for comments that led to substantial improvements in the manuscript. Tibshirani was partially supported by National Science Foundation Grant DMS-9971405 and National Institutes of Health Contract N01-HV-28183. Le was partially supported by the PHS Grant Number CA67166, awarded by the National Cancer Institute.

## Appendix: the multi-class case

The PPC procedure can be easily generalized to problems with more than two classes. In the notation of section 2.5, let  $G_k$  be the indices of observations in group  $k$  each of size  $n_k$ , for  $k = 1, 2, \dots, K$ . Let

$$p_{ik}(\alpha) = \sum_{j \in G_k} I[y_{ij} > q(\alpha, i)]/n_k, \quad k = 1, 2, \dots, K,$$

and  $\bar{p}_i(\alpha) = \sum_k n_k p_{ik}(\alpha) / \sum_k n_k$

We choose  $\alpha(i)$  to maximize  $\sum_k |p_{ik}(\alpha) - \bar{p}_i(\alpha)|$  for each site  $i$ , and then set  $\hat{p}_{ik} = p_{ik}(\hat{\alpha}(i))$ . Centroid shrinkage and classification then proceeds exactly as in Section 2.7.

## References

Adam, B.-L., Qu, Y., Davis, J. W., Ward, M. D., Clements, M. A., Cazares, L. H., Semmes, O. J., Schellhammer, P. F., Yasui, Y., Feng, Z. & Jr.,

- G. L. W. (2003), ‘Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy mean’, *Cancer Research* **63**(10), 3609–3614.
- Benjamini, Y. & Hochberg, Y. (1985), ‘Controlling the false discovery rate: a practical and powerful approach to multiple testing’, *J. Royal. Stat. Soc. B.* **85**, 289–300.
- Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984), *Classification and Regression Trees*, Wadsworth.
- Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2002), Least angle regression, Technical report, Stanford University.
- Efron, B. & Tibshirani, R. (2002), ‘Microarrays, empirical bayes methods, and false discovery rates’, *Gen. Epi.* .
- Hanash, S. (2003), ‘Disease proteomics’, *Nature* **422**, 226–232.
- Hansen, H. & Pedersen, A. (1986), ‘Tumor markers in patients with lung cancer’, *Chest* **89**, 219–224.
- Hutchens, T. W. & Yip, T. T. (1993), ‘New desorption strategies for the mass-spectrometric analysis of macromolecules’, *Rapid Communications in Mass Spectrometry* **7**, 576–580.
- Klade, C., Voss, T., Krystek, E., Ahorn, H., Zatloukaa, I K. Pummer, K. & Adolf, G. (2001), ‘Identification of tumor antigens in renal cell carcinoma by serological proteome analysis’, *Proteomics* **1**, 890–8.

- Kozak, K., Amneus, M., Pusey, S., Su, F., Luong, M., Luong, S., Reddy, S. & Farias-Eisner, R. (2003), 'Identification of biomarkers for ovarian cancer using strong anion-exchange proteinchips: potential use in diagnosis and prognosis', *Proc. Natl Acad Sci USA* **100**(1), 12343–8.
- Li, J., Zhang, Z., Rosenzweig, J., Wang, Y. Y. & Chan, D. W. (2002), 'Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer', *Clinical Chemistry* **48**, 296–1304.
- Merchant, M. & Weinberger, S. R. (2000), 'Recent advancements in surface-enhanced laser desorption/ionization-time of flight-mass spectroscopy', *Electrophoresis* **21**, 1164–1177.
- Nason, G. (1998), Wavethresh3 software, Technical report, Department of Mathematics, University of Bristol, Bristol, UK.
- Petricoin, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V., Steinberg, S. M., Mills, G. B., Simone, C., Fishman, D. A., Kohn, E. & Liotta, L. A. (2002b), 'Use of proteomic patterns in serum to identify ovarian cancer', *Lancet* **359**, 572–577.
- Petricoin, E. F., Zoon, K. C., Kohn, E. C., Barrett, J. C. & Liotta, L. A. (2002a), 'Clinical proteomics: translating benchside promise into bedside reality', *Drug discovery* **1**, 683–695.
- Qu, Y., Adam, B. L., Yasui, Y., Ward, M. D., Cazares, L. H., Schellhammer, P. F., Feng, Z., Semmes, O. J. & Wright, G. L., J. (2002), 'Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass

- spectral serum profiles discriminates prostate cancer from noncancer patients', *Clinical Chemistry* **48**, 1835–1843.
- Rai, A. J., Zhang, Z., Rosenzweig, J., Shih Ie, M. and Pham, T., Fung, E. T., Sokoll, L. J. & Chan, D. W. (2002), 'Proteomic approaches to tumor marker discovery', *Archives of Pathology and Laboratory Medicine* **126**, 1518–1526.
- Smith, R., Cokkinides, V. & Eyre, H. (2003), 'American cancer society guidelines for the early detection of cancer, 2003', *CA Cancer J Clin.* **53**(1), 27–43.
- Storey, J. & Tibshirani, R. (2003), 'Statistical significance for genomewide studies', *Proc. Natl. Acad. Sci.* **100**, 9440–5.
- Swenson, S., Jett, J., Sloan, J., Midthun, D., Hartman, T., Sykes, A., Aughenbaugh, G., Zink, F., Hillman, S., Noetzel, G., Marks, R., Clayton, A. & Paiolero, P. (2002), 'Screening for lung cancer with low-dose spiral computed tomography', *Am J Respir Crit Care Med* **165**(4), 508–513.
- Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *J. Royal. Statist. Soc. B.* **58**, 267–288.
- Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. (2003), 'Class prediction by nearest shrunken centroids, with applications to dna microarrays', *Statistical Science* pp. xxx–xxx.
- Tusher, V., Tibshirani, R. & Chu, G. (2001), 'Significance analysis of microarrays applied to transcriptional responses to ionizing radiation', *Proc. Natl. Acad. Sci. USA.* **98**, 5116–5121.

- Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K., & Zhao, H. (2003), ‘Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data’, *Bioinformatics* pp. 1636–1643.
- Yasui, Y., Pepe, M., Thompson, M. L., Adam, B.-L., Wright, G. L., Jr., Qu, Y., Potter, J. D., Winget, M., Thornquist, M., & Feng, Z. (2003), ‘A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection’, *Biostatistics* **4**, 449–463.