Sample classification from protein mass spectroscopy by "peak probability contrasts"

Robert Tibshirani

Depts of Health Research & Policy, and Statistics,
Stanford University

Email:tibs@stat.stanford.edu

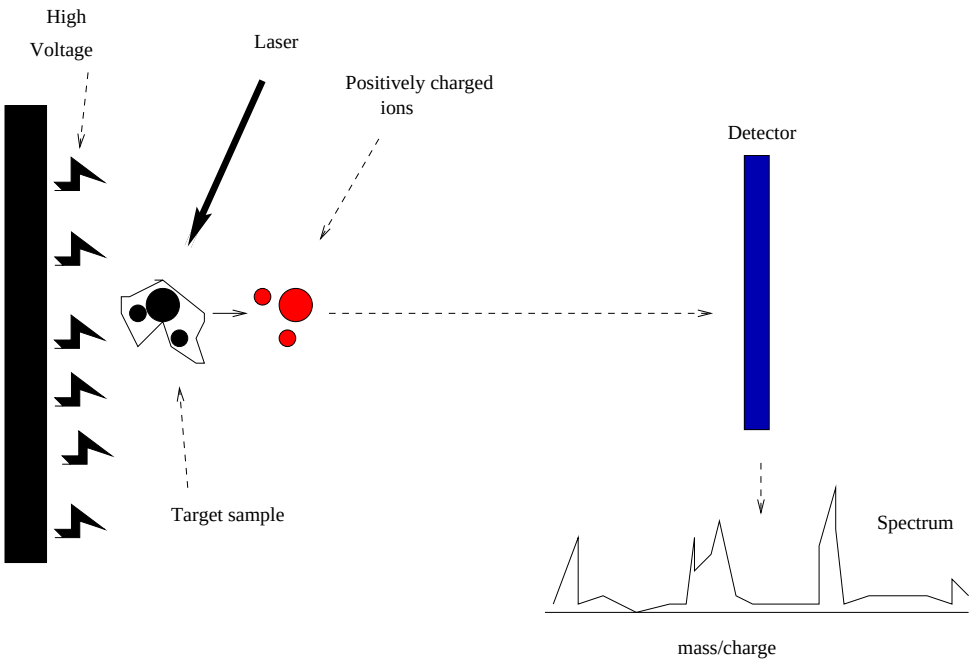
http://www-stat.stanford.edu/~tibs

Joint work with Trevor Hastie, Balasubramanian Narasimhan (Statistics/Biostatistics), Scott Soltys, Gongyi Shi, Albert Koong, Quynh Le (Radiation Oncology)

Protein mass spectroscopy

- Time-of-flight Mass spectrometry for measuring relative abundance of difference sized proteins in a blood sample.
- emerging as an important technology, a useful complement to gene expression arrays
- there are a number of popular systems including MALDI (matrix assisted laser desorption/ionization) and SELDI (Surface enhanced laser desorption/ionization). They refer to the way the sample is bound to a surface before being bombarded by a laser.

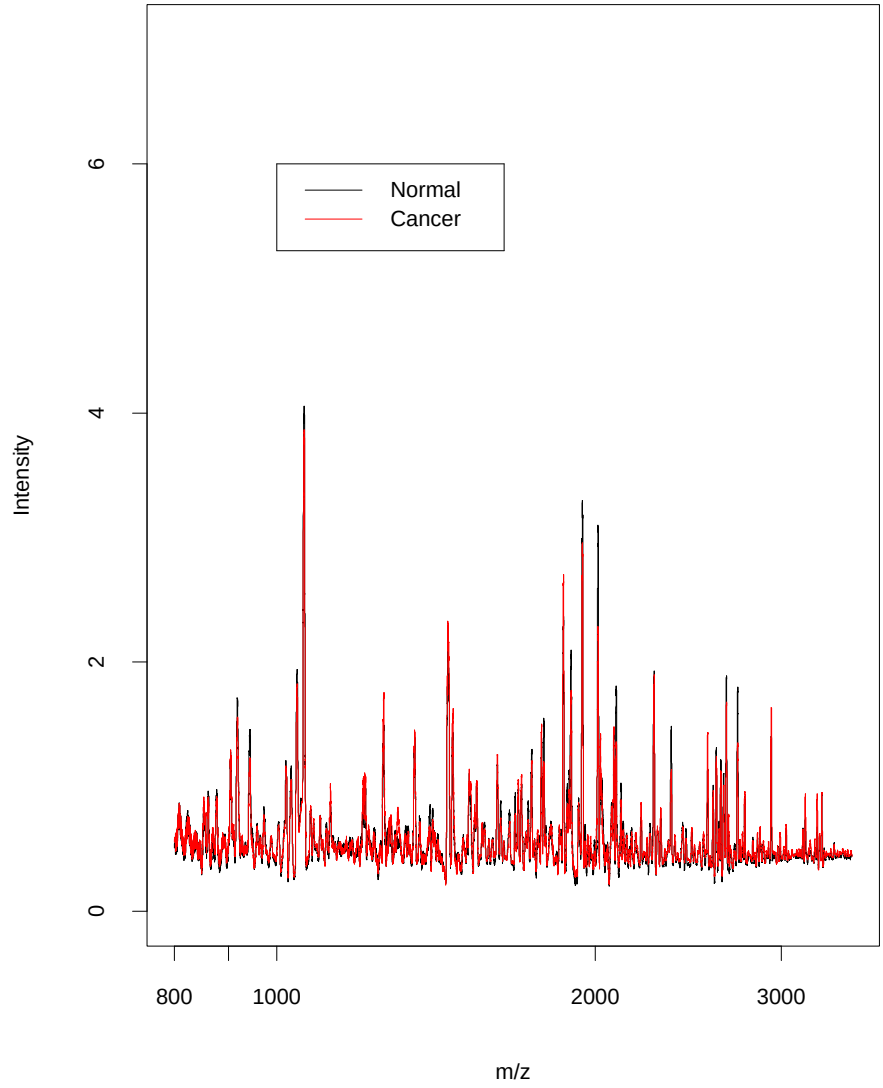
Mass spec process



Ovarian cancer MALDI dataset

- Wu et al. (2003)
- Training set- 89 patients- 42 normal, 47 with ovarian cancer
- serum samples measurements, each spectrum sampled at 91360 points

Average spectra



Existing classification methods

(for this problem)

- Support vector machines, trees, boosting, genetic algorithms
- Some well known papers have been flawed by poor experimental design and/or validation. Has created unreasonably high expectations for future experiments (eg 95% sensitivity and specificity)

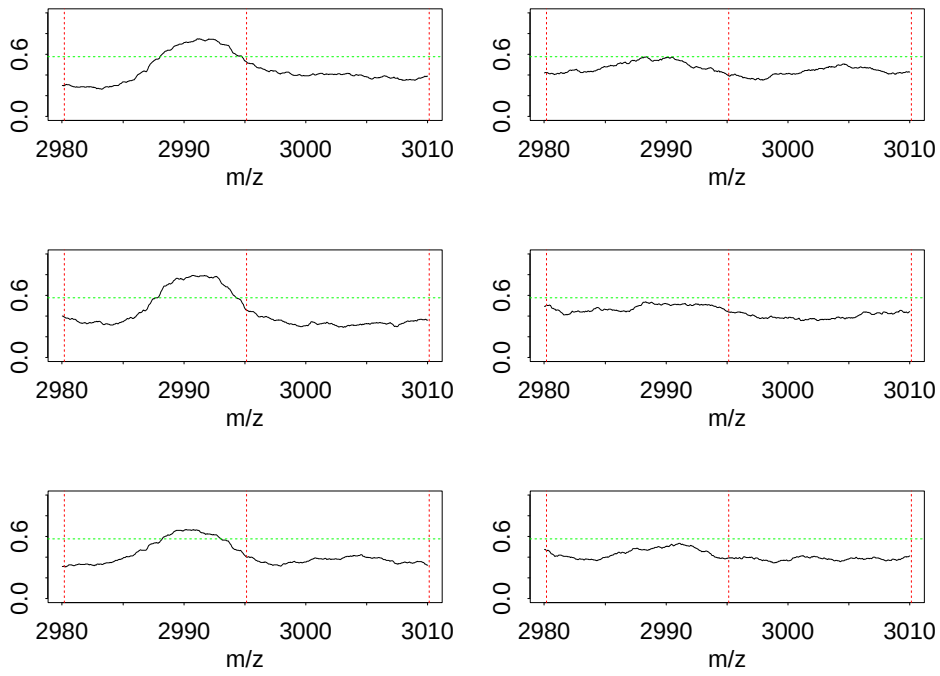
Desirable features for a classifier

It is important to discuss desirable properties for such a procedure:

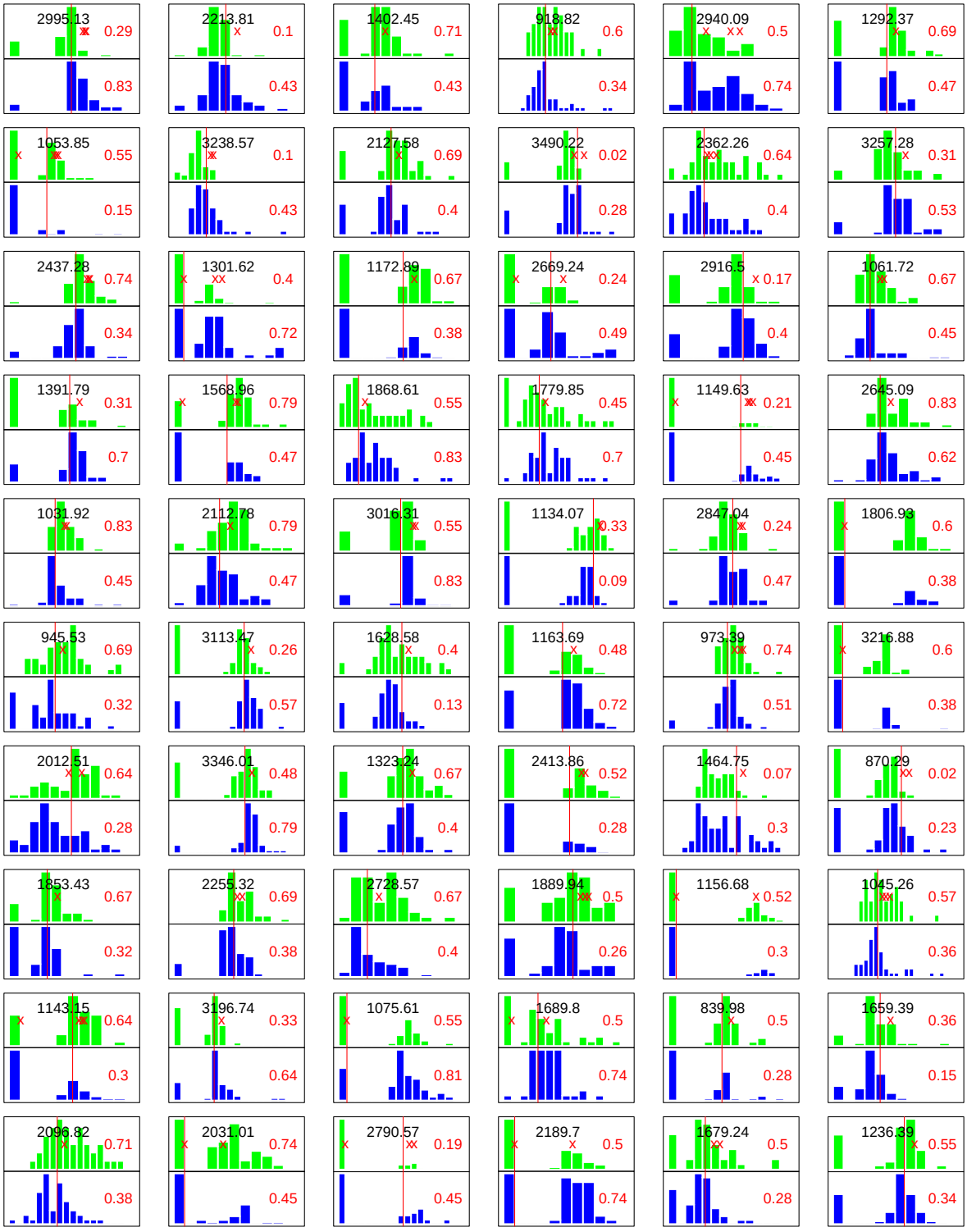
- It should focus on the peaks in the spectra, at least for the initial analysis.
- The method should account for the variation in the horizontal position and heights of the same biological peak in spectra.
- It should give a measure of importance for all peaks.
- If possible, the sample classification rule should use the peak information in a relatively simple way and provide a direct method for filtering out the less significant peaks.

Peak probability contrasts

1. Take logs of m/z axis. We'll consider approximate width of a peak to be log(.005).
2. Extract peak positions and heights from individual spectra, using either mass spec software, or a home grown procedure. [we adapted the procedure of Yasui et al. (2003), looking for local maxima]
3. Apply 1-dimensional complete linkage hierarchical clustering, to the collection of all 14,067 peaks. Cut off dendrogram at height log(.005). This gave 192 centroids.
4. Find optimal split for each centroid site, for discriminating normal from cancer.
5. For each spectrum i and site j, compute features $z_{ij} = 1$ if spectrum has a peak above split point at site j, and zero otherwise.
6. Apply nearest shrunken centroid classifier to features z_{ij}.



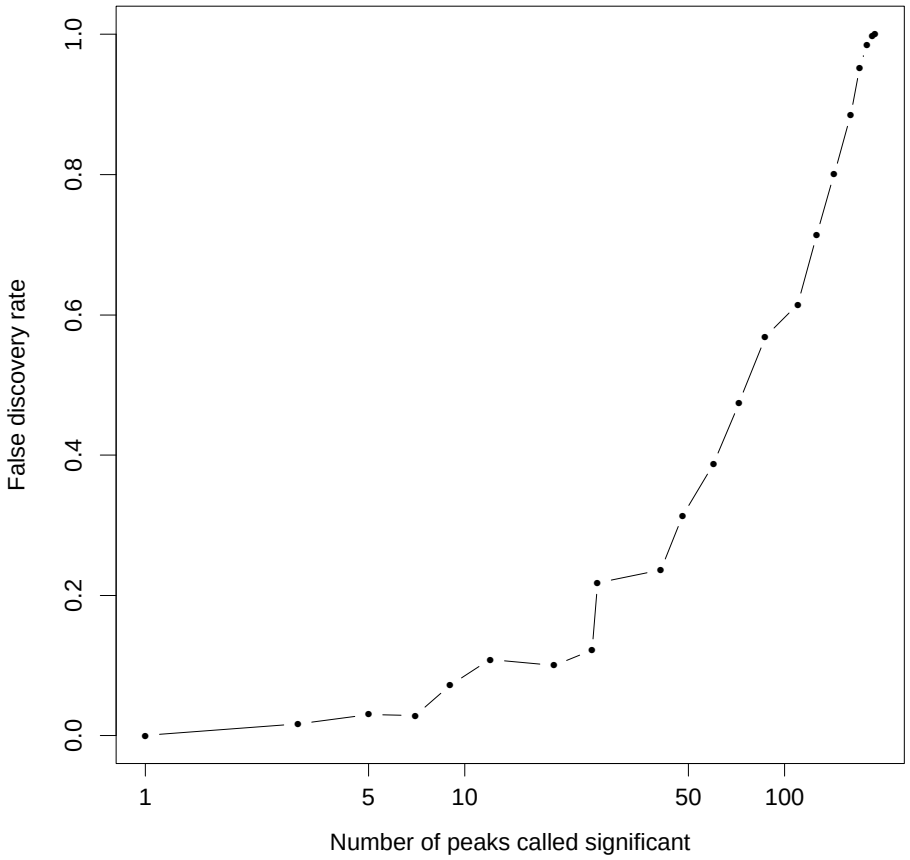
Left Column: Three spectra from cancer patients having a peak higher than 6 at the site $m/z = 2995.1$ *; right column: three spectra healthy patients without the peak, or whose peak is too low. The vertical dotted lines indicate the centroid* 2995.1 *and the outer limits for the peak position.*


2995.13 0.29 0.83
2213.81 0.1 0.43
1402.45 0.71 0.43
918.82 0.6 0.34
2940.09 0.5 0.74
1292.37 0.69 0.47
1053.85 0.55 0.15
3238.57 0.1 0.43
2127.58 0.69 0.4
3490.22 0.02 0.28
2362.26 0.64 0.4
3257.28 0.31 0.53
2437.28 0.74 0.34
1301.62 0.4 0.72
1172.89 0.67 0.38
2669.24 0.24 0.49
2916.5 0.17 0.4
1061.72 0.67 0.45
1391.79 0.31 0.7
1568.96 0.79 0.47
1868.61 0.55 0.83
1779.85 0.45 0.7
1149.63 0.21 0.45
2645.09 0.83 0.62
1031.92 0.83 0.45
2112.78 0.79 0.47
3016.31 0.55 0.83
1134.07 0.33 0.09
2847.04 0.24 0.47
1806.93 0.6 0.38
945.53 0.69 0.32
3113.47 0.26 0.57
1628.58 0.4 0.13
1163.69 0.48 0.72
973.39 0.74 0.51
3216.88 0.6 0.38
2012.51 0.64 0.28
3346.01 0.48 0.79
1323.24 0.67 0.4
2413.86 0.52 0.28
1464.75 0.07 0.3
870.29 0.02 0.23
1853.43 0.67 0.32
2255.32 0.69 0.38
2728.57 0.67 0.4
1889.94 0.5 0.26
1156.68 0.52 0.3
1045.26 0.57 0.36
1143.15 0.64 0.3
3196.74 0.33 0.64
1075.61 0.55 0.81
1689.8 0.5 0.74
839.98 0.5 0.28
1659.39 0.36 0.15
2096.82 0.71 0.38
2031.01 0.74 0.45
2790.57 0.19 0.45
2189.7 0.5 0.74
1679.24 0.5 0.28
1236.39 0.55 0.34

Estimation of False discovery rates

- Benjamini & Hochberg (1985), Storey (2002)
- let $\hat{p}_{ij}$ be proportion of class j samples with a peak at site i that is above threshold. Denote the shrunken version by $\tilde{p}_{ij}$.
- permute sample labels, and repeat entire PPC fitting process
- estimate # of false positives by # of times a difference as large as $\tilde{p}_{i2} - \tilde{p}_{i1}$ is obtained.
- use this to estimate the FDR

False discovery rates



Nearest shrunken centroids

- Tibshirani et al. (2001), designed especially for gene expression studies
- Compute centroids for each class. Shrink them towards overall centroids.
- Without shrinkage, equivalent to nearest centroids and diagonal LDA (see e.g. Dudoit et al. (2001)). Shrinkage selects features and can improve classification performance

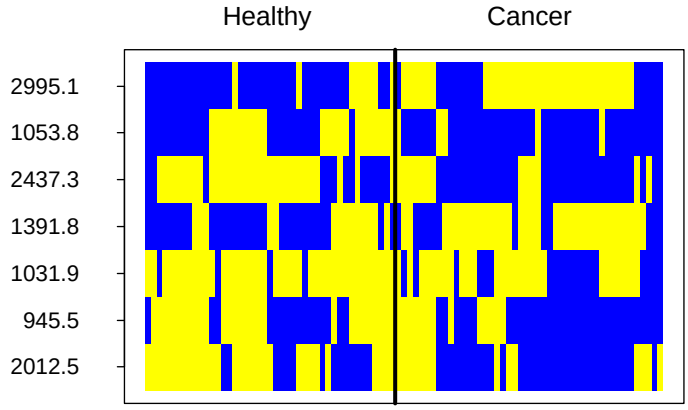
Results

Method	CV errors/89 (se)	# sites
(1) PPC	23(1.1)	7
(2) PPC/pres-abs	30(1.8)	133
(3) PPC/lasso	25(1.5)	192
(4) LDA/t-15	31(1.4)	15
(5) SVM/t-15	27(1.6)	15
(6) SVM	21(1.4)	91360

PPC top peak is at 2995.1

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.

Heatmap



Artificial spiking experiment

- started with random samples of actual spectra
- "spiked" in 5 different artificial peaks in each of cancer and control spectra. f= signal to background ratio.

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

(

Discussion

- Understanding differential peaks in serum as a difficult problem. Signals tend to be small and can easily be overwhelmed by experimental variation
- Peak probability contrast method is potentially useful- gives overview of all peaks and their disciminatory power.
- An Excel/R package will be available soon, using the powerful language interface developed by Balasubramanian Narasimhan.

References

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.

Dudoit, S., Fridlyand, J. & Speed, T. (2001), 'Comparison of discrimination methods for the classification of tumors using gene expression data', *J. Amer. Statist. Assoc* pp. 1151–1160.

Storey, J. D. (n.d.), A direct approach to false discovery rates. Submitted. Available at `http://www-stat.stanford.edu/~jstorey/`.

Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. (2001), 'Diagnosis of multiple cancer types by shrunken centroids of gene expression', *Proc. Natl. Acad. Sci.* **99**, 6567–6572.

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.