Statistical methods in proteomics II

Example 1:

- Preprocessing and data reduction

Data Reduction

- 60831(time ticks) x 20 (fractions) x 41 (samples)

recombining fractions

60831 x 41

windowing and choosing max. intensity for each spectrum

~2000 x 41

discard windows that do not contain peak

506 x 41

Example 1:

- Preprocessing and data reduction
- T-test at each retained m/z
- Pick peaks that best separates training sample (mahalanobis distance is used to measure distance)

Def. The statistical distance or Mahalanobis distance between two points $x = (x_1, \ldots, x_p)^t$ and $y = (y_1, \ldots, y_p)^t$ in the p-dimensional space $\mathbb{R}^p$ is defined as

$$d_S(x, y) = \sqrt{(x - y)^t S^{-1} (x - y)}$$

and $d_S(x, 0) = \|x\|_S = \sqrt{x^t S^{-1} x}$ is the norm of x.

Example 1:

- Preprocessing and data reduction
- T-test at each retained m/z
- Pick peaks that best separates training sample (mahalanobis distance is used to measure distance)
- Fisher LDA
- Leave one out cross validation

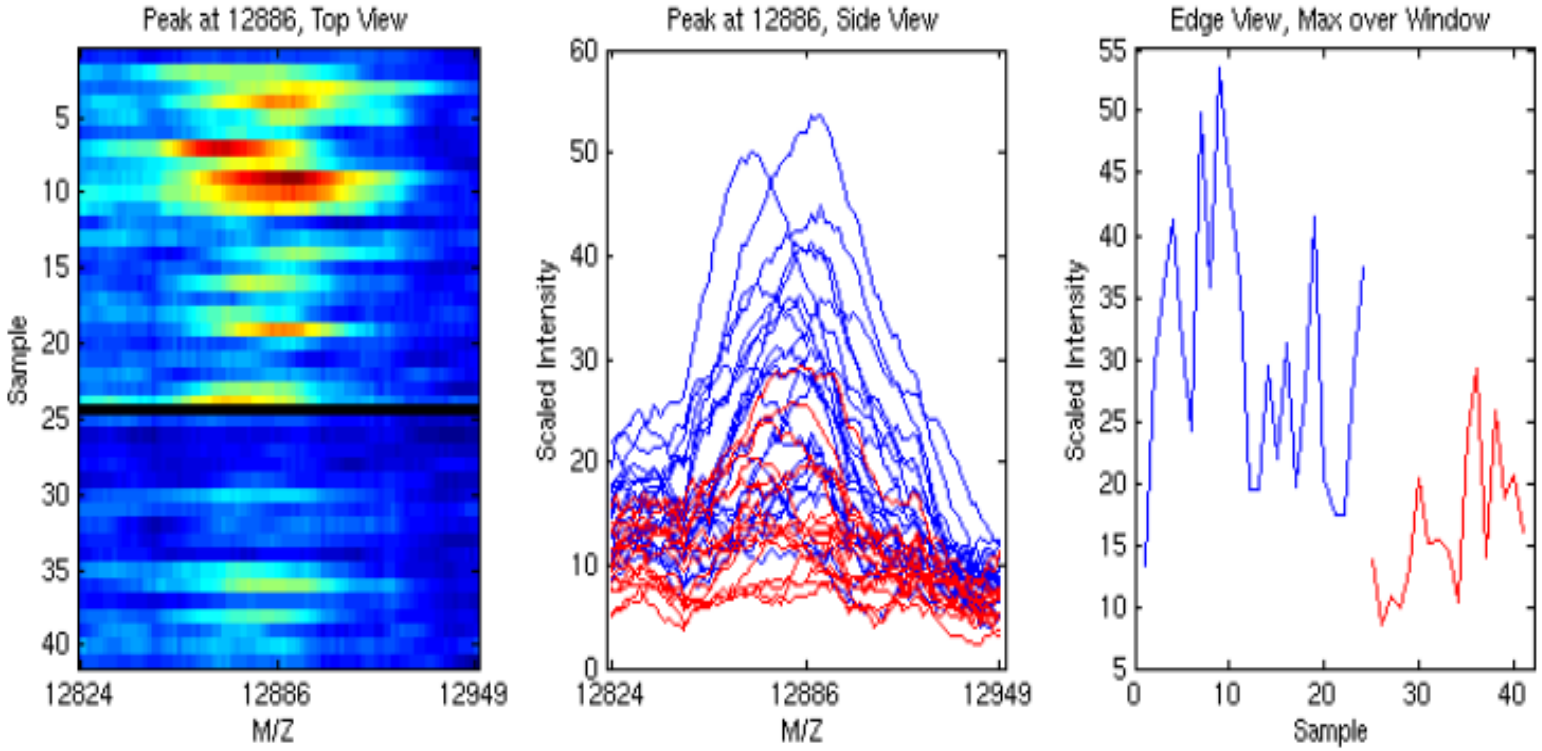
Table 1. Best sets of peaks for $N = 1, \ldots, 5$

Best peak set	Mahalanobis distance	Number misclassified	Expected number misclassified	Leave-one-out cross-validation	Empirical *p*-value
12 886	2.547	11	9	11	0.005
8840, 12 886	5.679	5	5	6	<0.01
3077, 12 886, 74 263	9.019	3	3	4	<0.01
5863, 8143, 8840, 12 886	12.585	3	2	3	<0.01
4125, 7000, 9010, 12 886, 74 263	23.108	1	1	1	<0.01

Expected misclassification: under assumption that the two samples are multivariate normal with common covariance structure: Prob(Z<.5 MD)

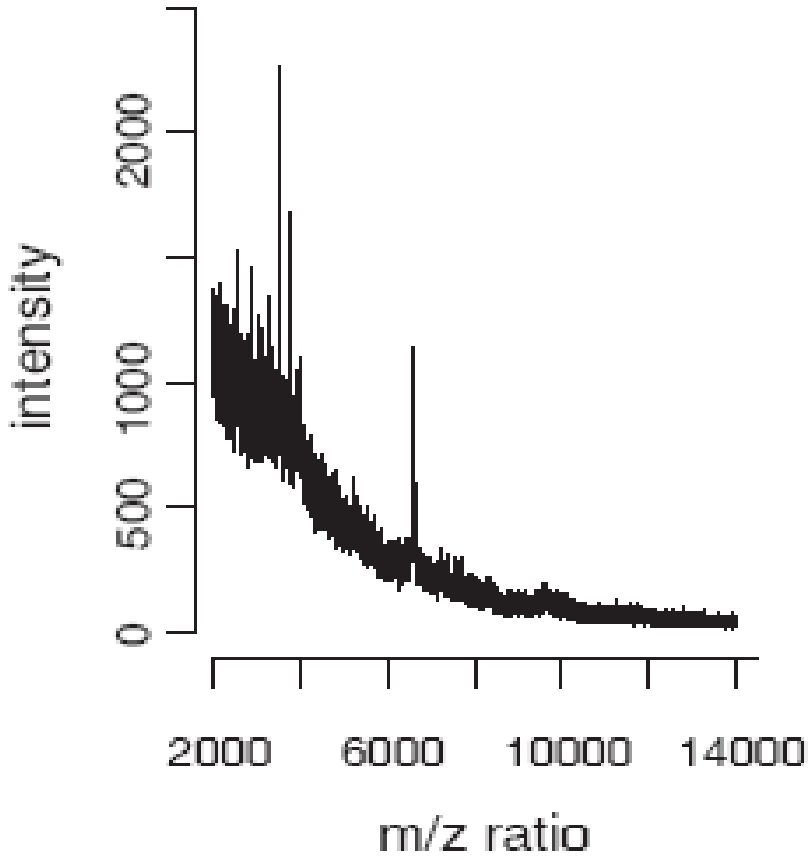
Z: N(0,1)

MD: Mahalanobis Distance


Peak at 12886, Top View
Sample
5
10
15
20
25
30
35
40
12824
12886
12949
M/Z
Peak at 12886, Side View
Scaled Intensity
60
50
40
30
20
10
0
12824
12886
12949
M/Z
Edge View, Max over Window
Scaled Intensity
55
50
45
40
35
30
25
20
15
10
5
0
10
20
30
40
Sample

- Example2: standardization and denoising algorithm for mass spectra to classify whole-organism bacterial specimens
  Satten et al, Bioinformatics 2004

- Unlike some applications in which complex samples are pre-fractionized to select a particular group of compounds of interest (like in our example in last lecture), mass spectra are obtained from whole organisms and more noisy



•There are peaks towering above baseline which are likely to correspond to real peptides

•Non-uniform background (baseline) that is decreasing overall

•Higher variance associated with higer baseline

• NOT SEEN in this plot: variability in maximum intensity across spectra

Peaks diminish if log transformation is taken without removing background

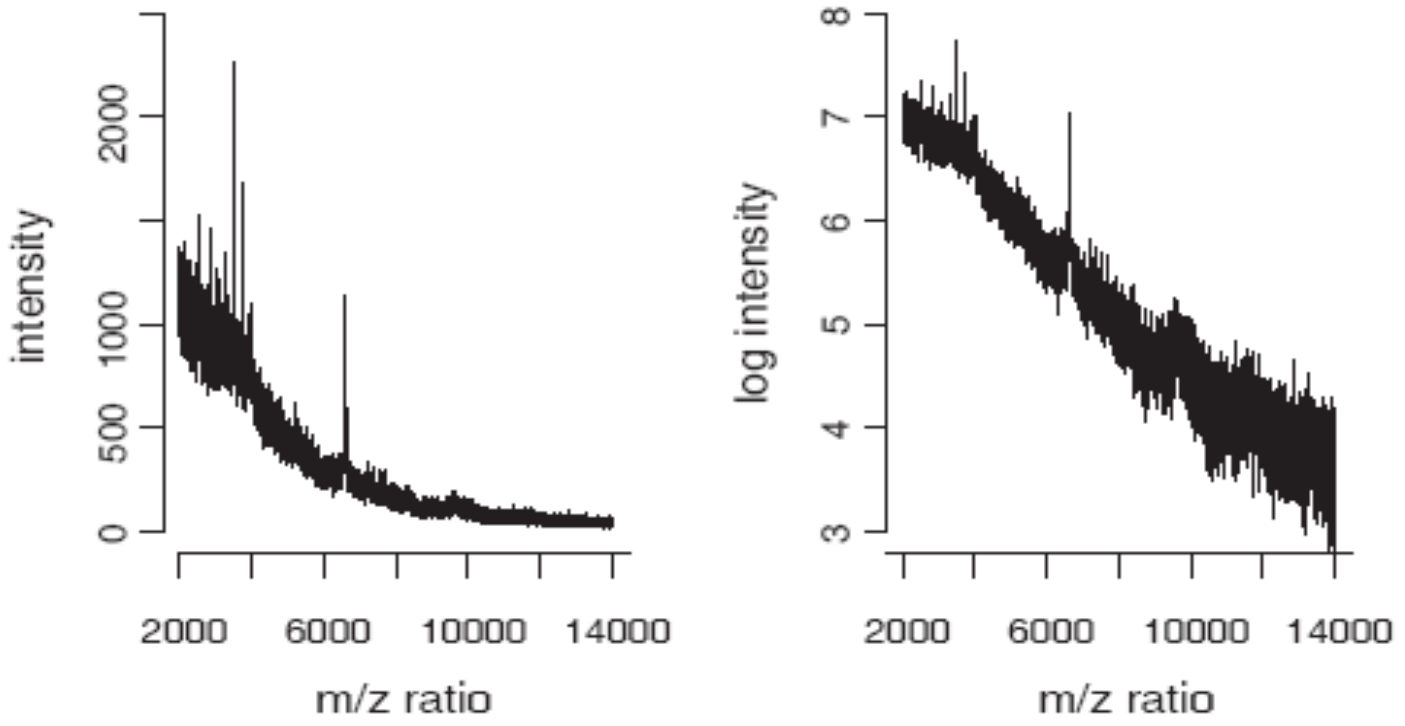


Fig. 1. A raw *B.anthracis* spectrum (left) and its (natural) log-transformed version (right).

Standardization:

- To account for the non-uniform baseline and variability of maximum intensity

$$y_i^* = \frac{y_i - Q_{0.5}(x_i)}{Q_{0.75}(x_i) - Q_{0.25}(x_i)}$$

$Y_{i:}$ observed intensity at m/z ratio x_i

Y_i^*: standardized intensity

$Q_a(x_i)$: estimate of a-th quantile of spectra intensities around m/z x_i

How to define “local” and estimate the quantile when we actually observe one y?

$Q_\alpha(x)$ is weighted average of

$$Q_\alpha^+(x) = \min_y \{W_h(x,y) \geq \alpha\},$$

$$Q_\alpha^-(x) = \max_y \{W_h(x,y) \leq \alpha\}$$

, where h defines neighborhood of x and

$$W_h(x_i, y) = \frac{\sum_{j=\max(1,i-h)}^{\min(n,i+h)} I\{y_j \leq y\} \left\{1 - \frac{(i-j)^2}{h^2}\right\}}{\sum_{j=\max(1,i-h)}^{\min(n,i+h)} \left\{1 - \frac{(i-j)^2}{h^2}\right\}},$$

Define quantiles: +/- because of discrete y

- Start from bottom up, the smallest y such that w≥α
- The largest y such that w≤α

Function of y given x_i:

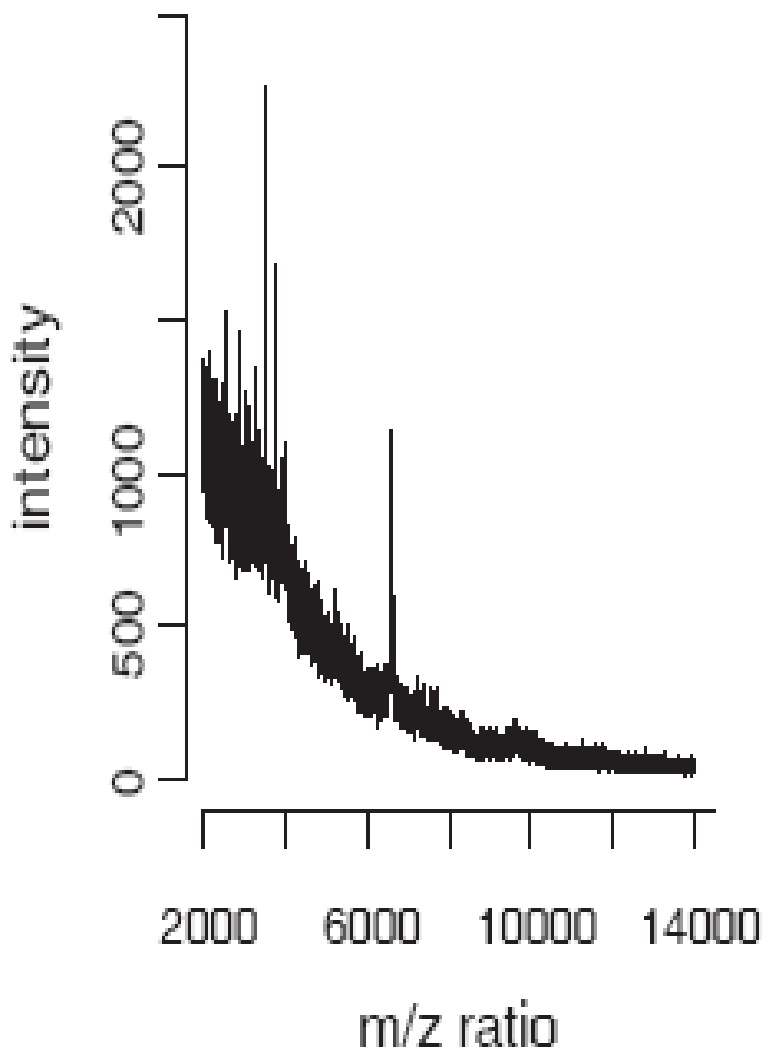
Basically counting the number of intensities less than y within a window, but with weights close to 1 if near center and close to 0 if near ends.

- If $Q_\alpha^+(x) = Q_\alpha^-(x)$ then use the common value to define $Q_\alpha(x)$.
- If $Q_\alpha^+(x) > Q_\alpha^-(x)$, let $c^\pm = W_h(x, Q_\alpha^\pm(x)$ and define $Q_\alpha(x) = \frac{c^+ - \alpha}{c^+ - c^-} Q_\alpha^-(x) + \frac{\alpha - c^-}{c^+ - c^-} Q_\alpha^+(x)$

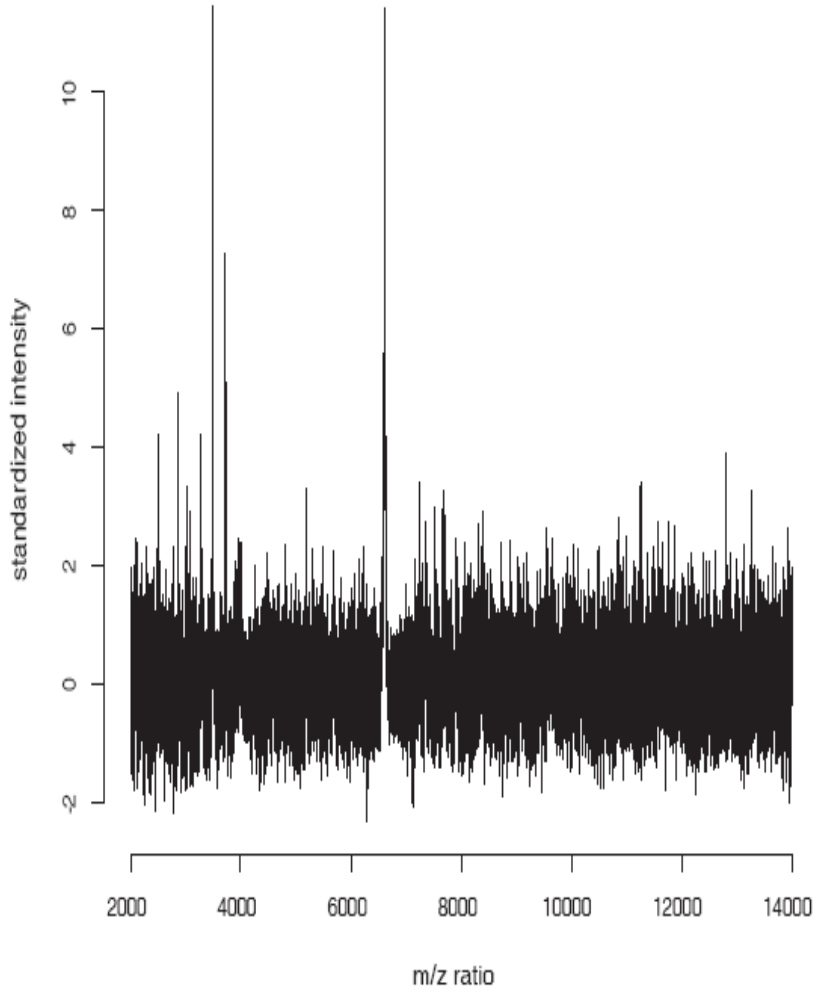
$$y_i^* = \frac{y_i - Q_{0.5}(x_i)}{Q_{0.75}(x_i) - Q_{0.25}(x_i)}$$

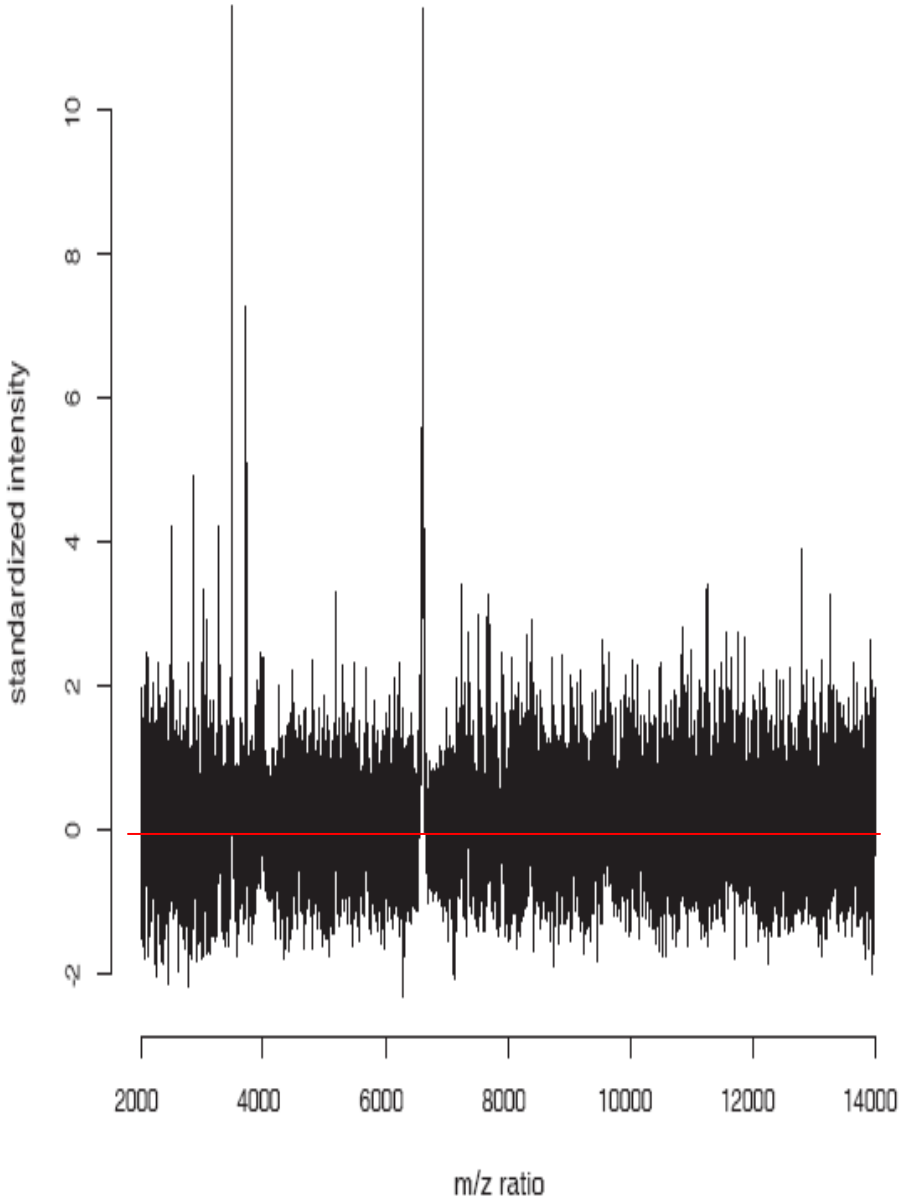
Compare with $\tilde{y}_i = \frac{y_i - y_{min}}{y_{max} - y_{min}}$ used by other authors (Baggerly, 2004)

Raw data



After standardization





- Flat baseline
- High peaks
- Constant Noise intensity

Negative standardized intensities are likely to be pure noise

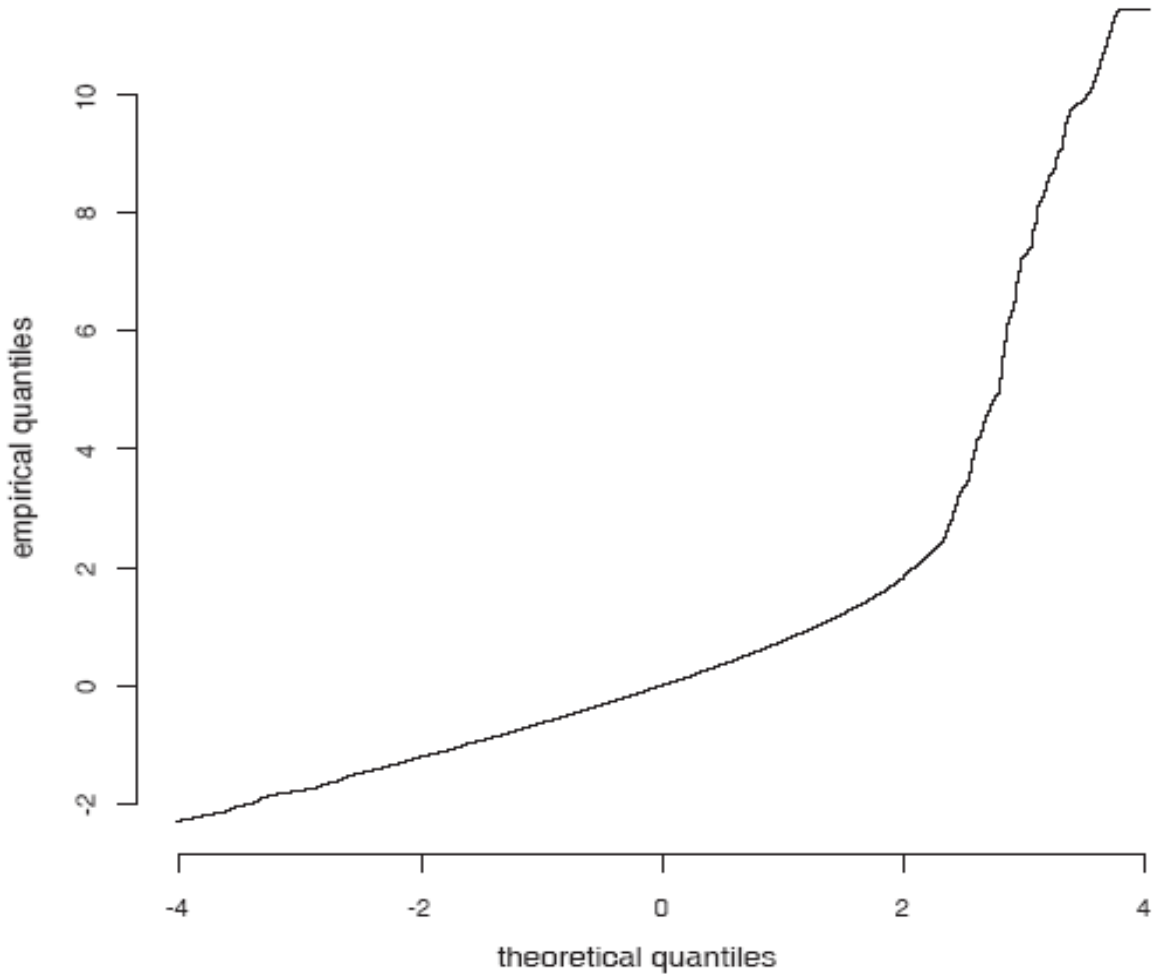


Fig. 3. A standard normal q–q plot of the standardized spectrum values for the *B.anthracis* spectrum of Figures 1 and 2.

Noise seems to be normally distributed

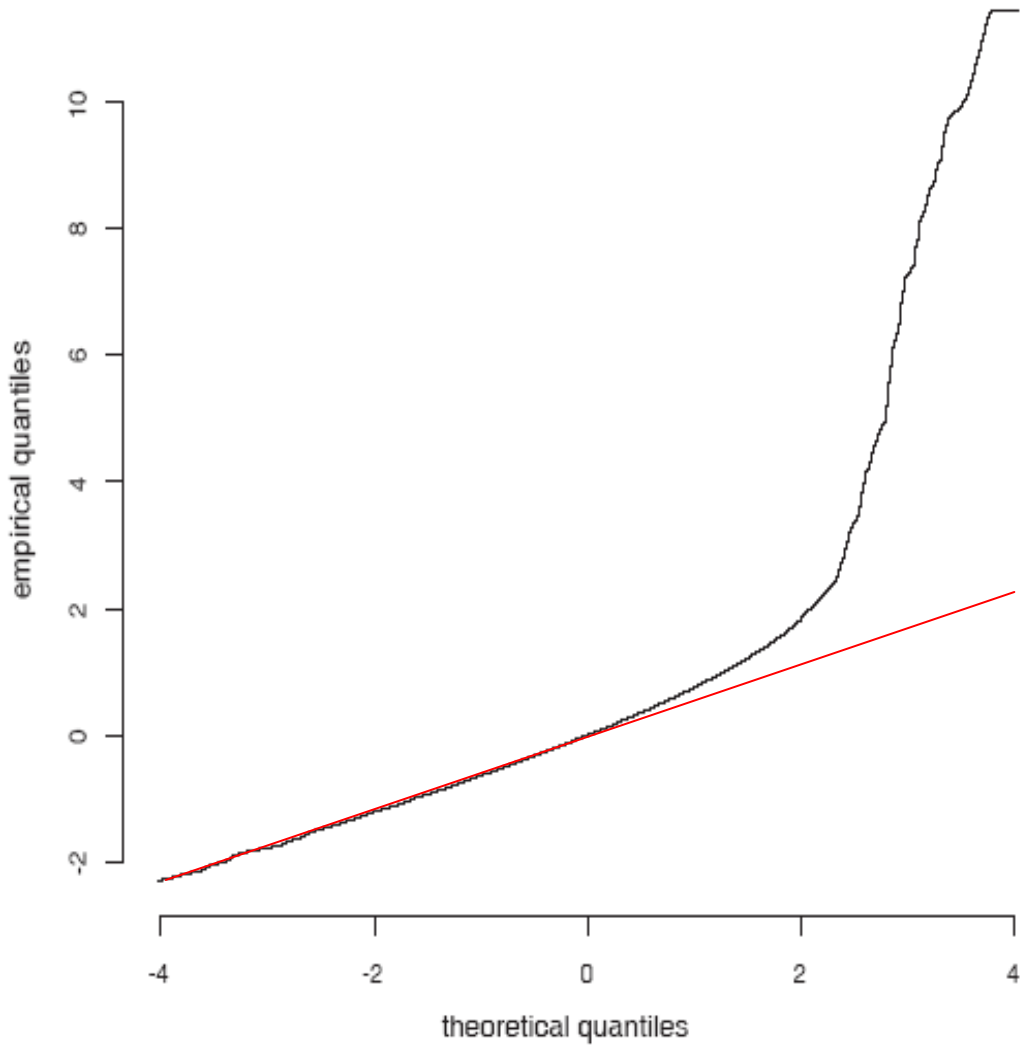


Fig. 3. A standard normal q–q plot of the standardized spectrum values for the *B.anthracis* spectrum of Figures 1 and 2.

Denoising

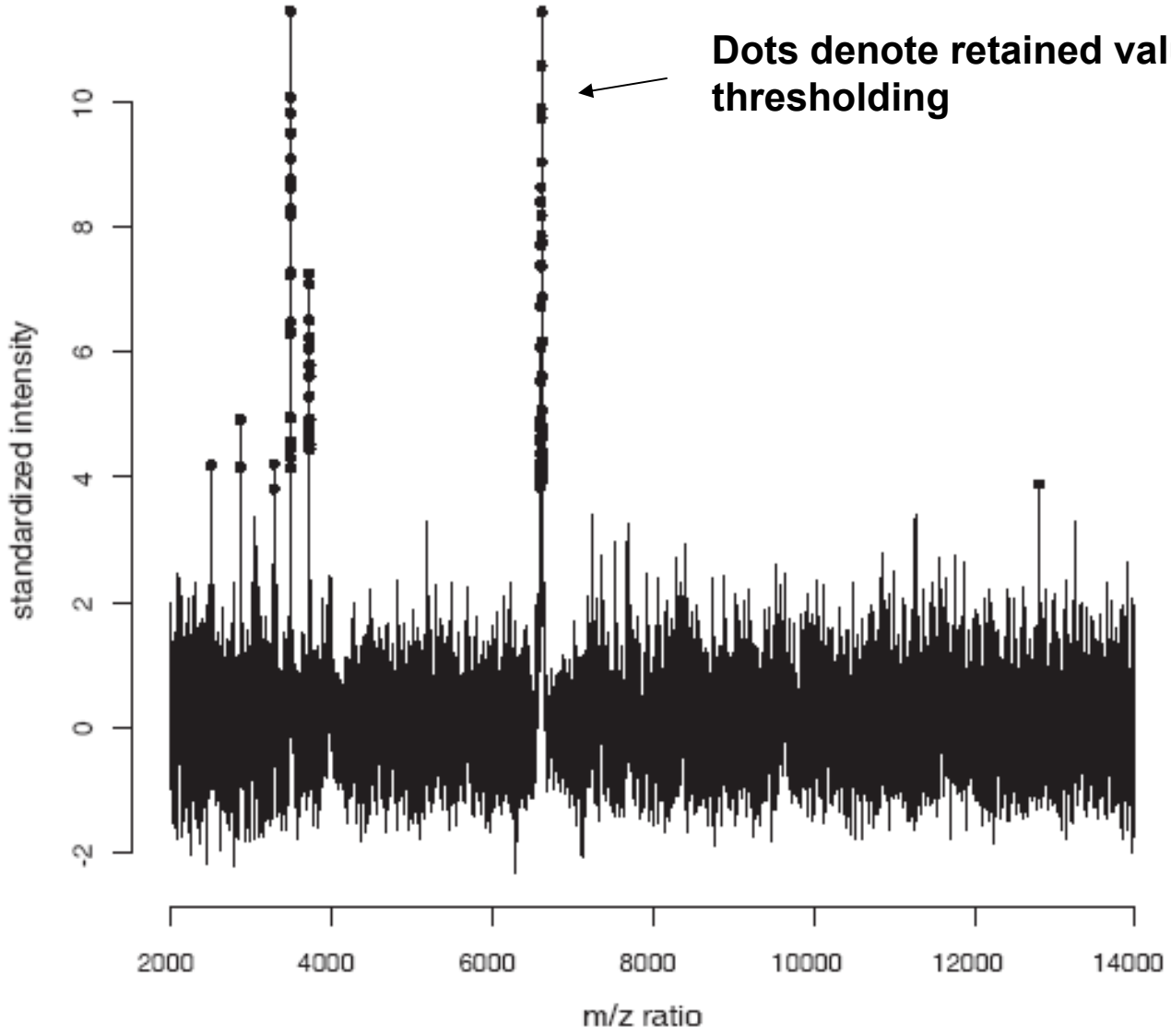
- Negative standardized values are considered as pure noise and used to estimate noise variance σ
- Thresholding (have we seen this before?):
  - Hard

$$\tilde{y}(x) = \begin{cases} y^*(x) & y^*(x) \geq 6\hat{\sigma} \\ 0 & otherwise \end{cases}$$

  - Soft

$$\tilde{y}(x) = \max(0, y^*(x) - 6\hat{\sigma})$$

denoised, thresholded spectrum, with 6σ cutoff



Defining peaks Yasui et al, Biostatistics 2003

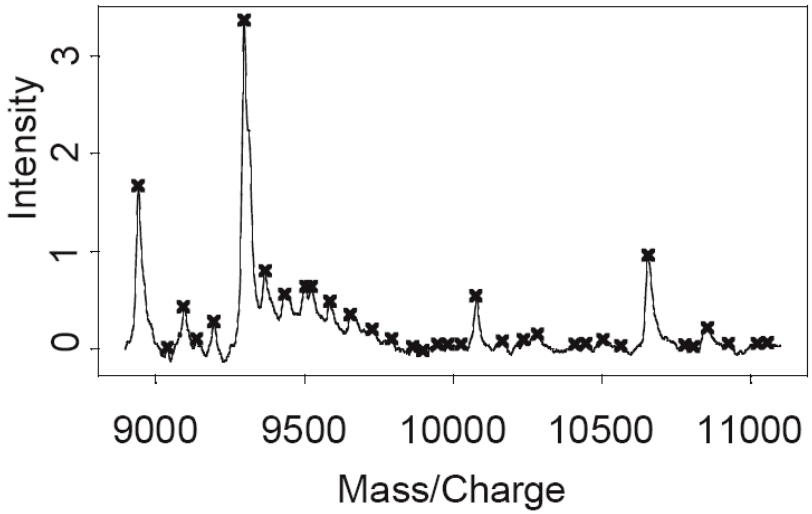
Data from SELDI experiment

- Absolute intensity values are highly variable. Relative intensity values show less variability.
- Intensity values depend on energy applied, while keeping protein abundance constant.
- Mass/charge axis may shift between experiments (by 0.1-0.2%)
- Proposal to reduce continuous spectrum to binary (peak/no-peak) data and to perform mass/charge axis alignment.

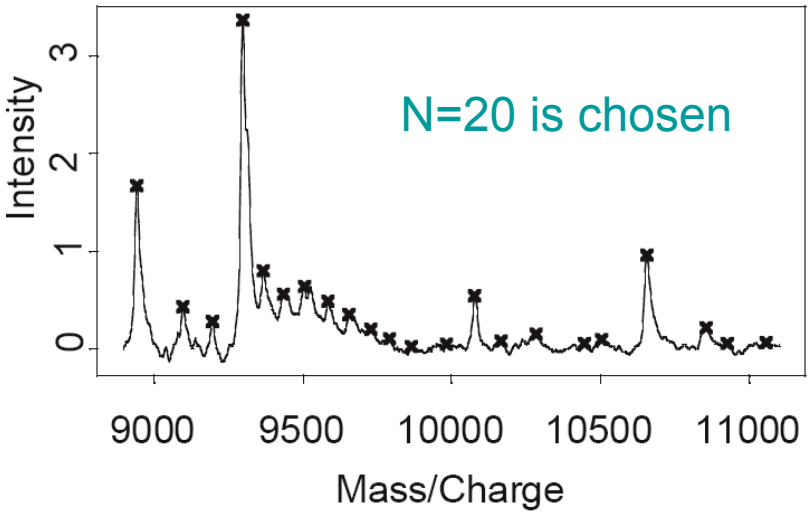
Defining peaks, contd Yasui et al, Biostatistics 2003

- A value $y(x)$ is defined as "peak" if it is the highest in a $\pm N$-point neighborhood of x in the set of observed m/z values. (N chosen $=$ 10, 20,30, 40).
- The definition can be augmented by requiring that $y(x)$ is higher than average intensity in *broad* neighborhood,defined via *super-smoother*, e.g. using smoothing window consisting of 5% of all points
- Note: a non-peak does not imply absence of protein at that m/z value.
- m/z axis alignment addressed by labeling as "peak" all points within $\pm 0.2\%$ of each peak.

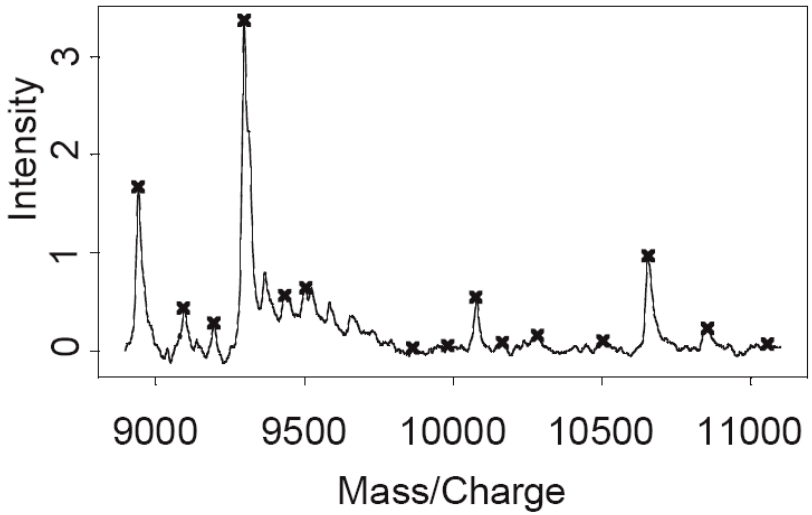
+/- 10 points



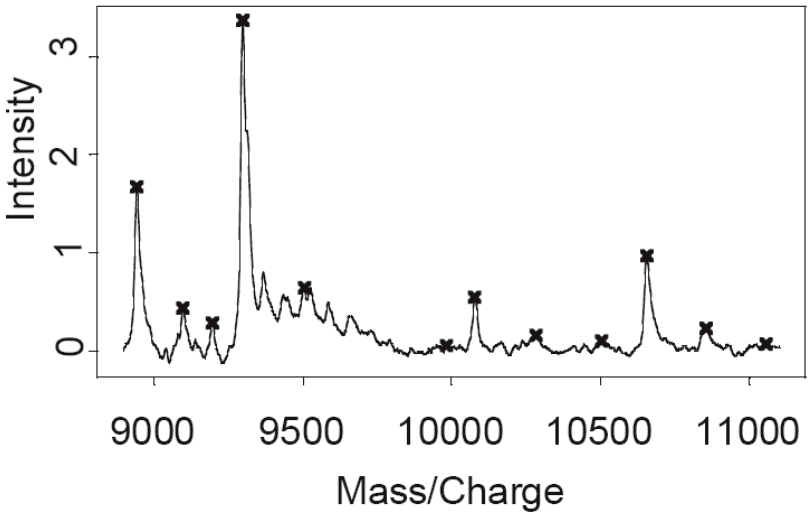
+/- 20 points



+/- 30 points



+/- 40 points



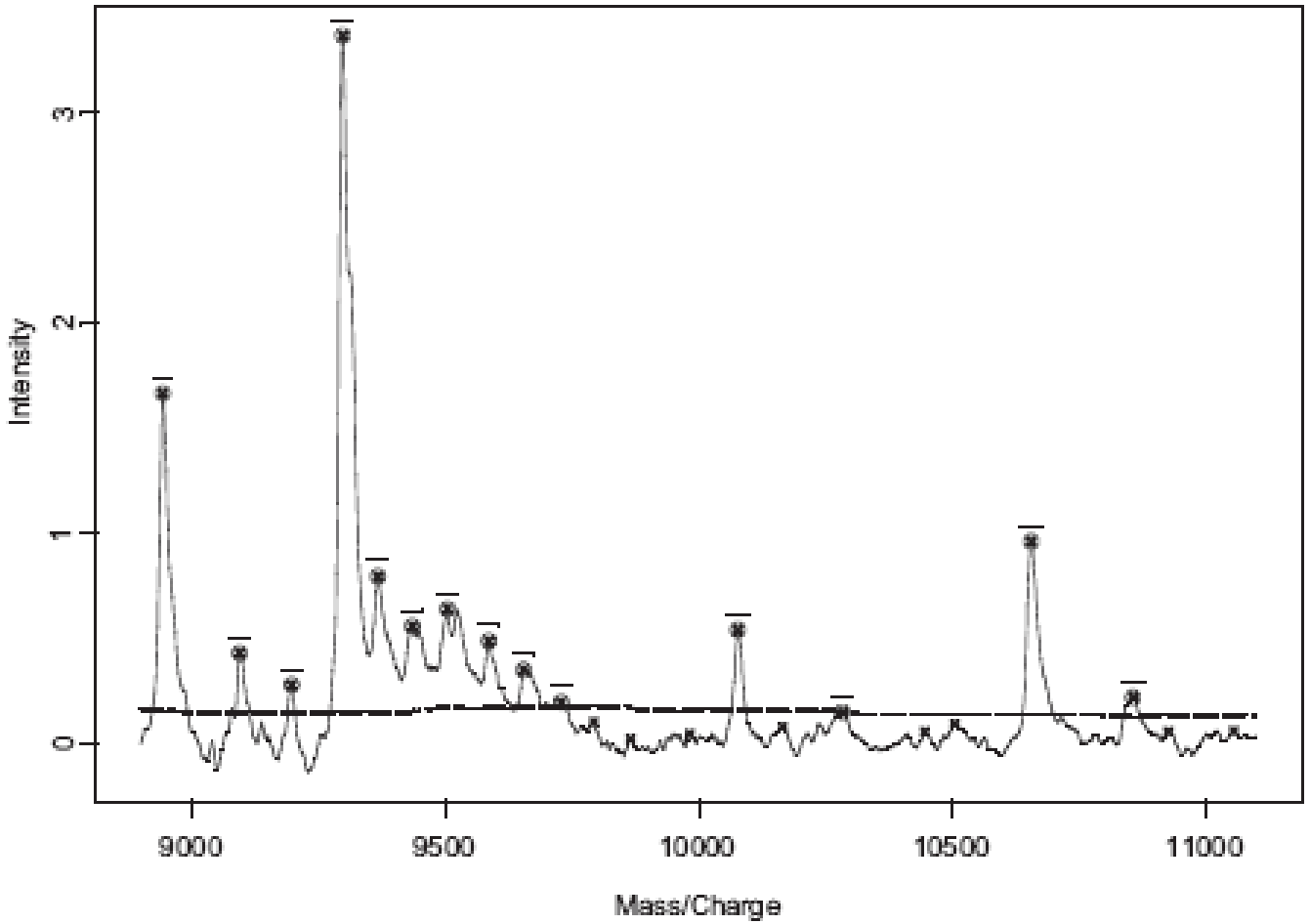
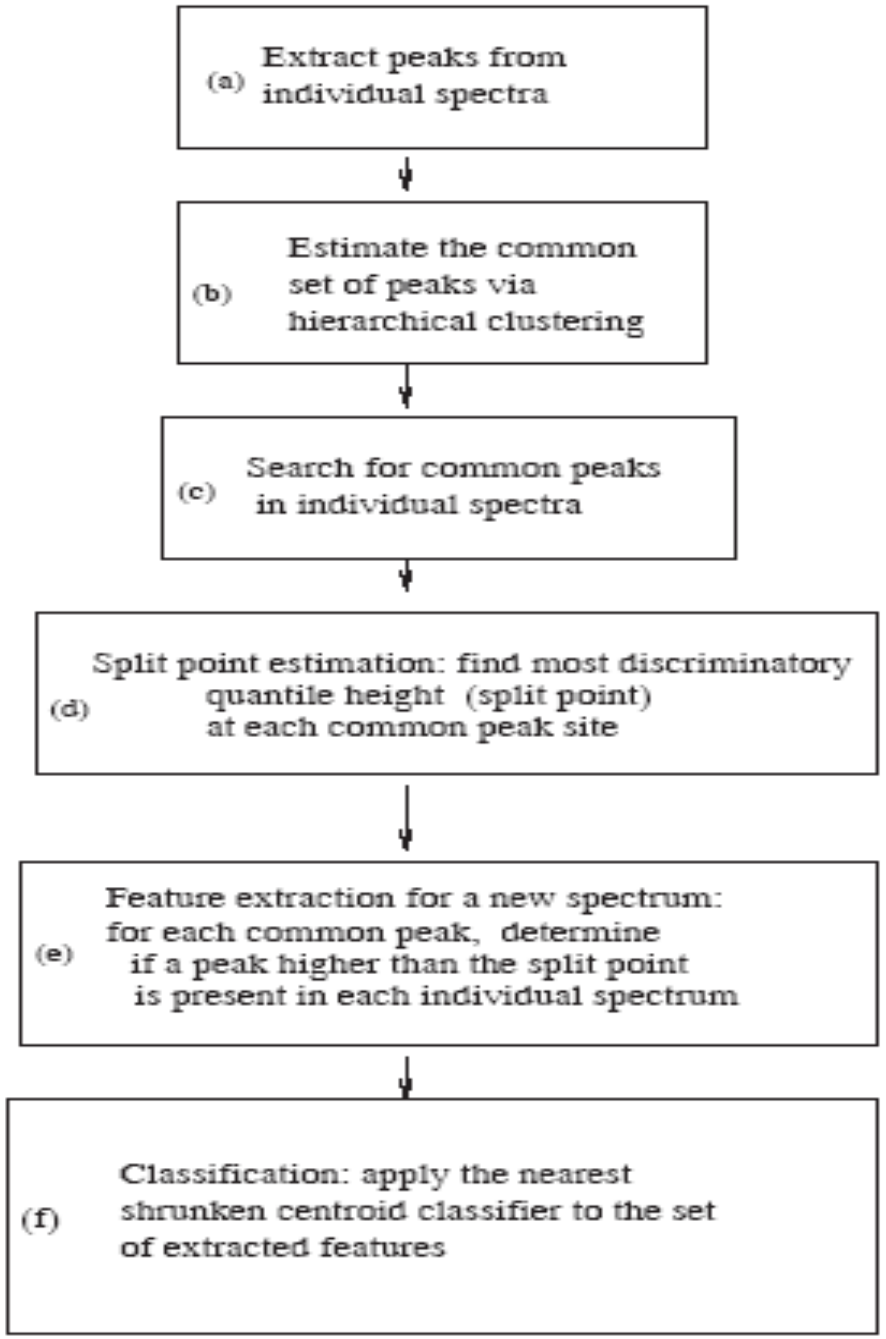


Fig. 3. An example of SELDI data after the pre-analysis processing. The broken line indicates the average intensity calculated by the super-smoother method using 5% of all data points (i.e. 5% of 11 175 points) as the smoothing window: the figure shows the region corresponding to about 10% of the 11 175 points. The points marked by '⊗' are the peaks, points that have intensity values that are higher than the averages in their broad neighborhoods and also the highest in their respective nearest ±20-point neighborhood sets. A horizontal bar above each peak shows an interval corresponding ±0.2% of the peak's mass/charge value.

"Peak probability contrasts" Clas

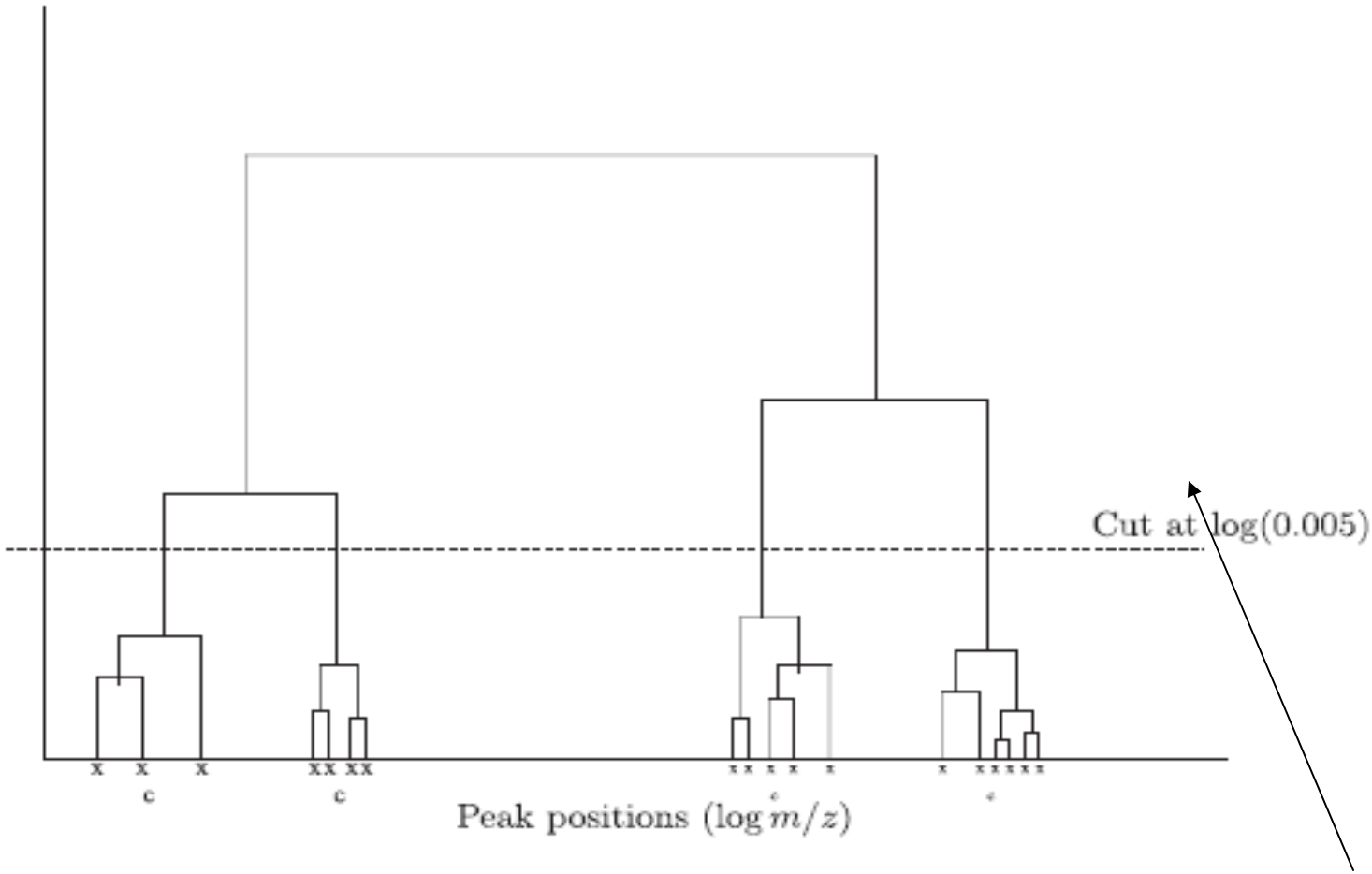
Tibshirani et al, Bioinformatics, 2004



PPC contd

- Analyzed ovarian cancer data (47 cancer patients, 42 controls). Spectra of 91360 m/z values, 0.019 Da apart.
- Data log-transformed, baseline subtracted, normalized.
- Peak extraction follows approach similar to Yasui et al. (N=100 in application)
- Complete linkage hierarchical clustering of full set of peaks from all spectra. One dimensional distance on m/z axis. "Tight" clusters may represent same peak, shifted across spectra.
- Individual spectra searched to determine which common peaks they contain.

Search for common peaks



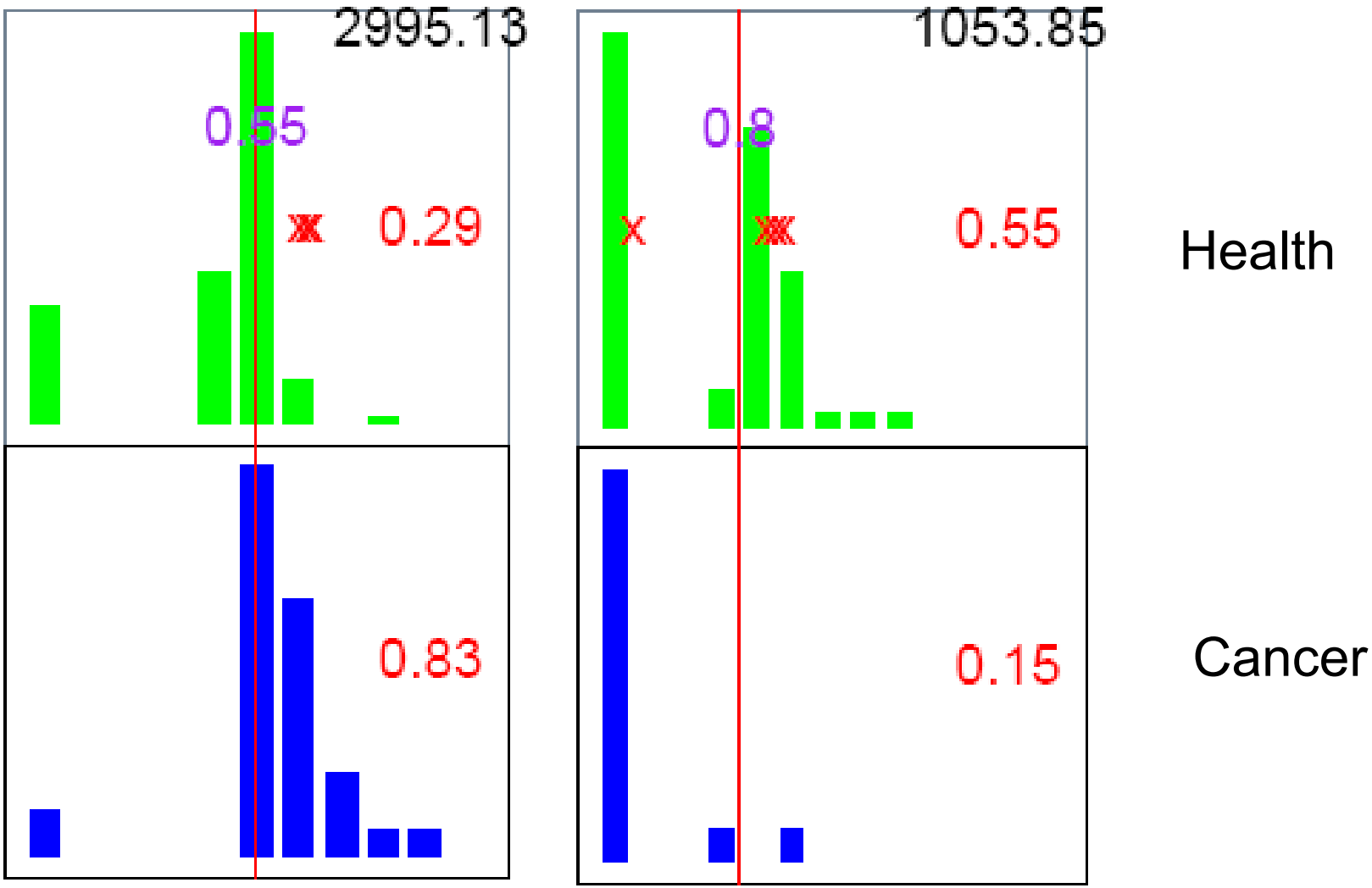
Complete linkage in clustering:
Cutoff guarantees each peak in cluster is within 0.005 from any other peak in same cluster

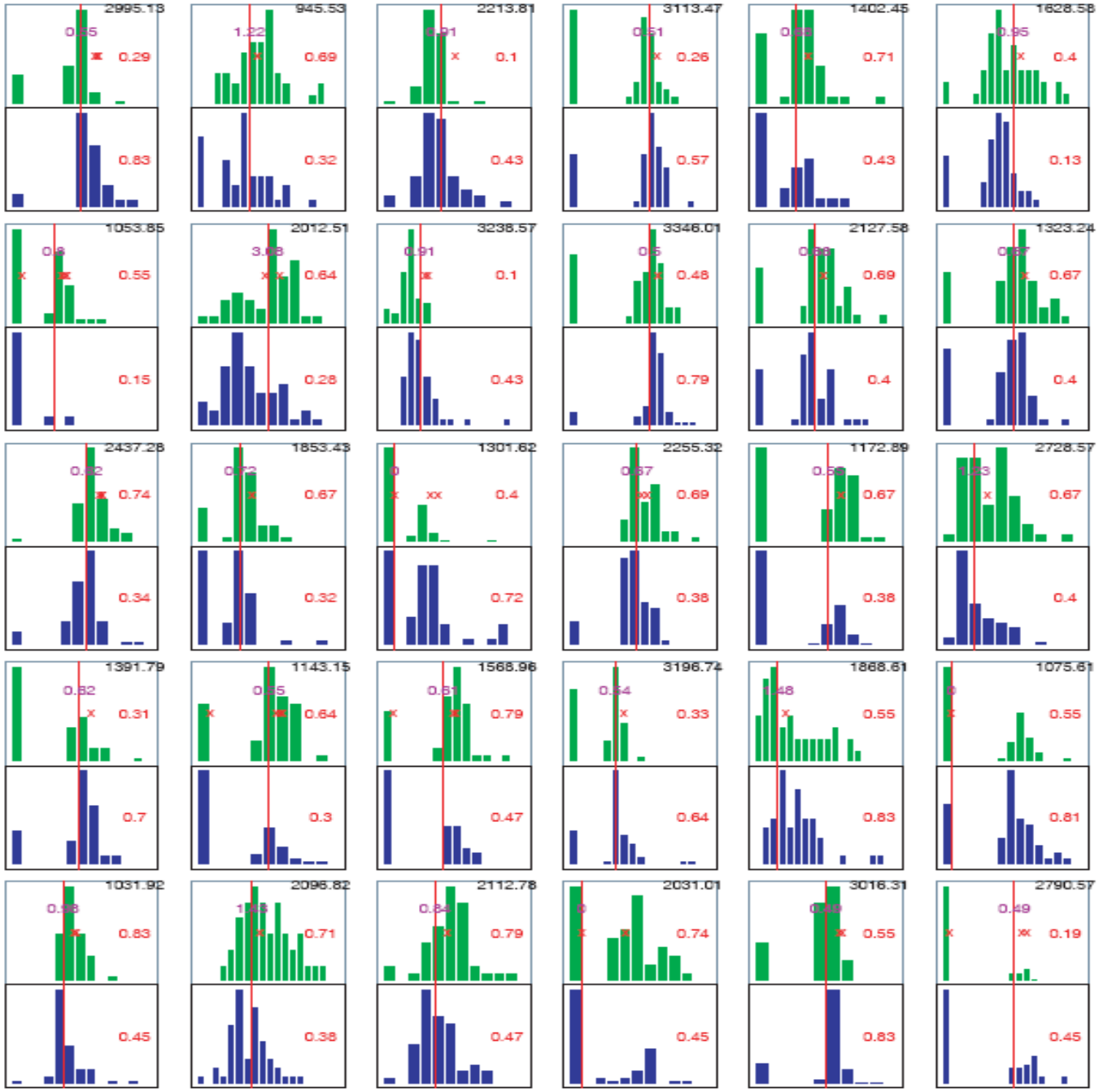
Split point estimation for each peak

- Each peak height cut at quantile giving optimal discrimination between two groups.
- Peak height of j-th observation at i-th site $= y_{ij}$.
- $q(i, \alpha) = \alpha$ quantile of peaks at i-th site. Define

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

  , the proportion of spectra in group k (=1,2) larger than $q(i, \alpha)$.
- Quantile is chosen as $\alpha\hat{(i)}$ maximizing $|p_{i2}(\alpha) - p_{i1}(\alpha)|$ and $\hat{p}_{ik} = p_{ik}(\hat{\alpha}(i))$


2995.13
0.55
0.29
0.83
1053.85
0.8
0.55
0.15
Health
Cancer



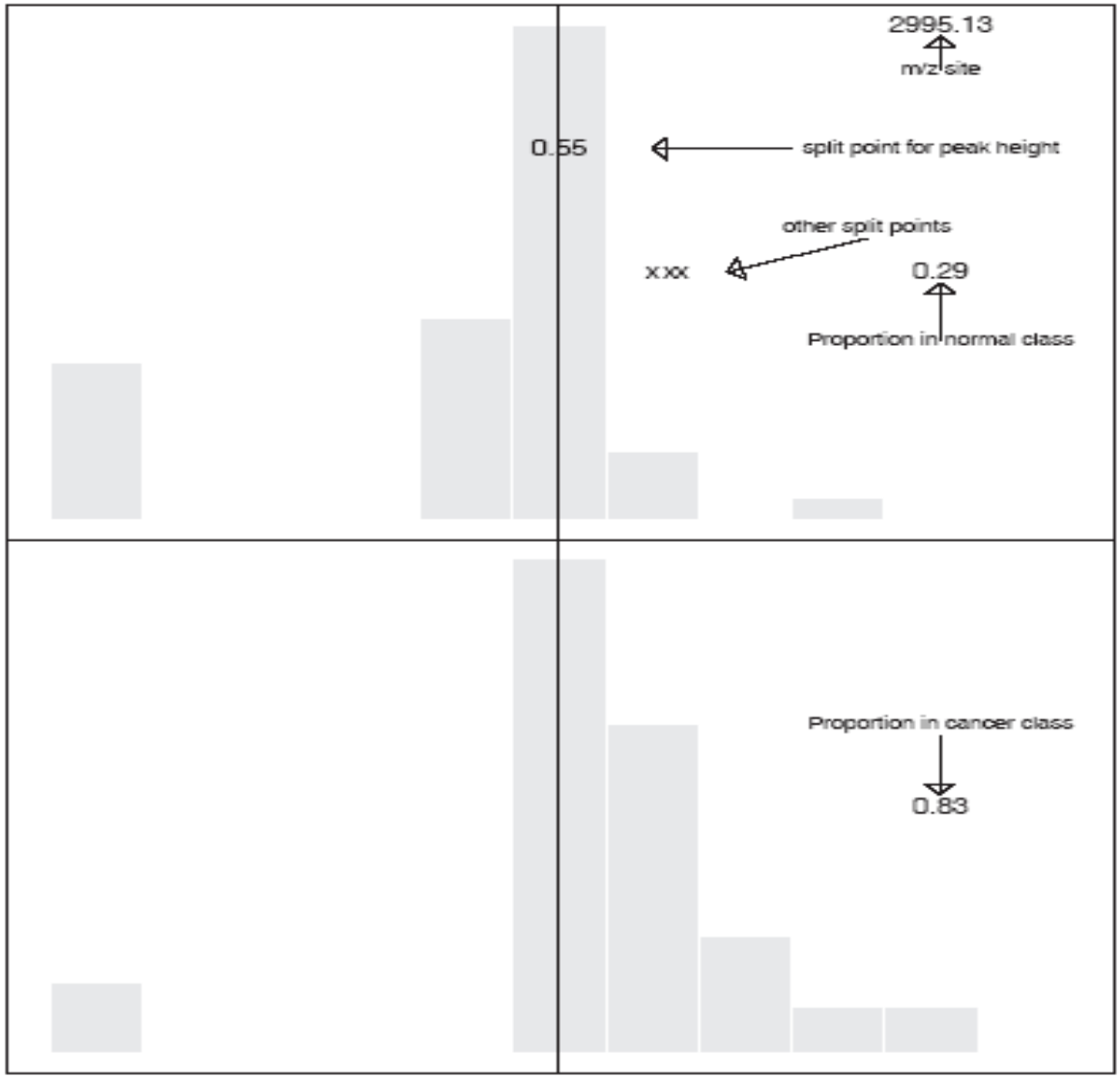


Fig. 2. Exploded view of the top left panel of Figure 1, with a legend detailing the format. The vertical line shows the estimated optimal height split point. The proportions of samples in each class having peaks higher than the split point are indicated. The 'x's indicate the horizontal positions of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

- To classify new spectrum,first form binary feature set indicating if peak with height greater than split point exists in the spectrum. So for new spectrum with peak heights $y_1^*, ..., y_p^*$, define indicators $z_i^* = I[y_i^* > q(i, \hat{\alpha}(i)]$.
- Then compare $z*$ vector to the vectors $(\hat{p}_{i1}, ...\hat{p}_{m1})$, $(\hat{p}_{i2}, ...\hat{p}_{m2})$ and assign spectrum to the nearest class in some metric.
- "Shrunken centroids" method is applied in actual algorithm and software. Other classifiers can also be used.
- Estimates of FDR obtained from $T_i = \hat{p}_{i2} - \hat{p}_{i1}$ via permutation of labels:

$$F\hat{D}R(t) = \frac{\sum_{i=1}^{B} I[|T_i^{*b}| > t]/B}{\sum_{i=1}^{B} I[|T_i| > t]}$$

Estimates proportion of false positives among "significant" peaks

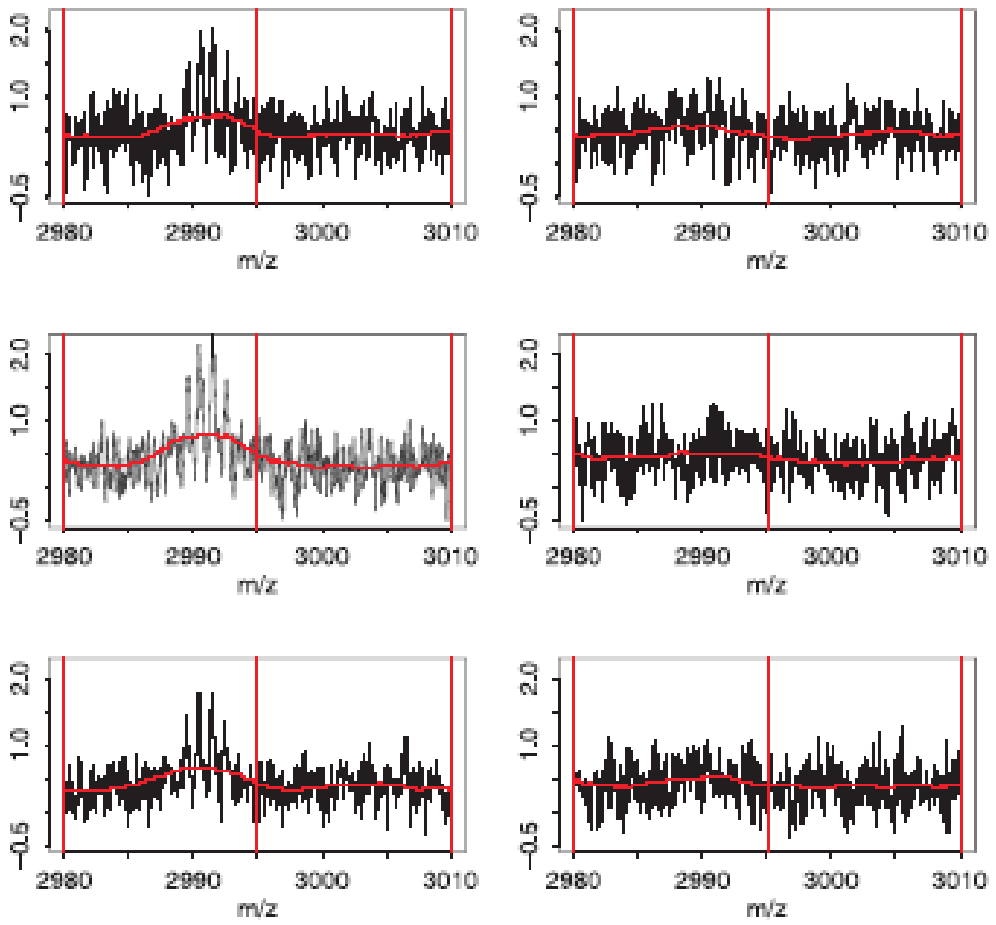
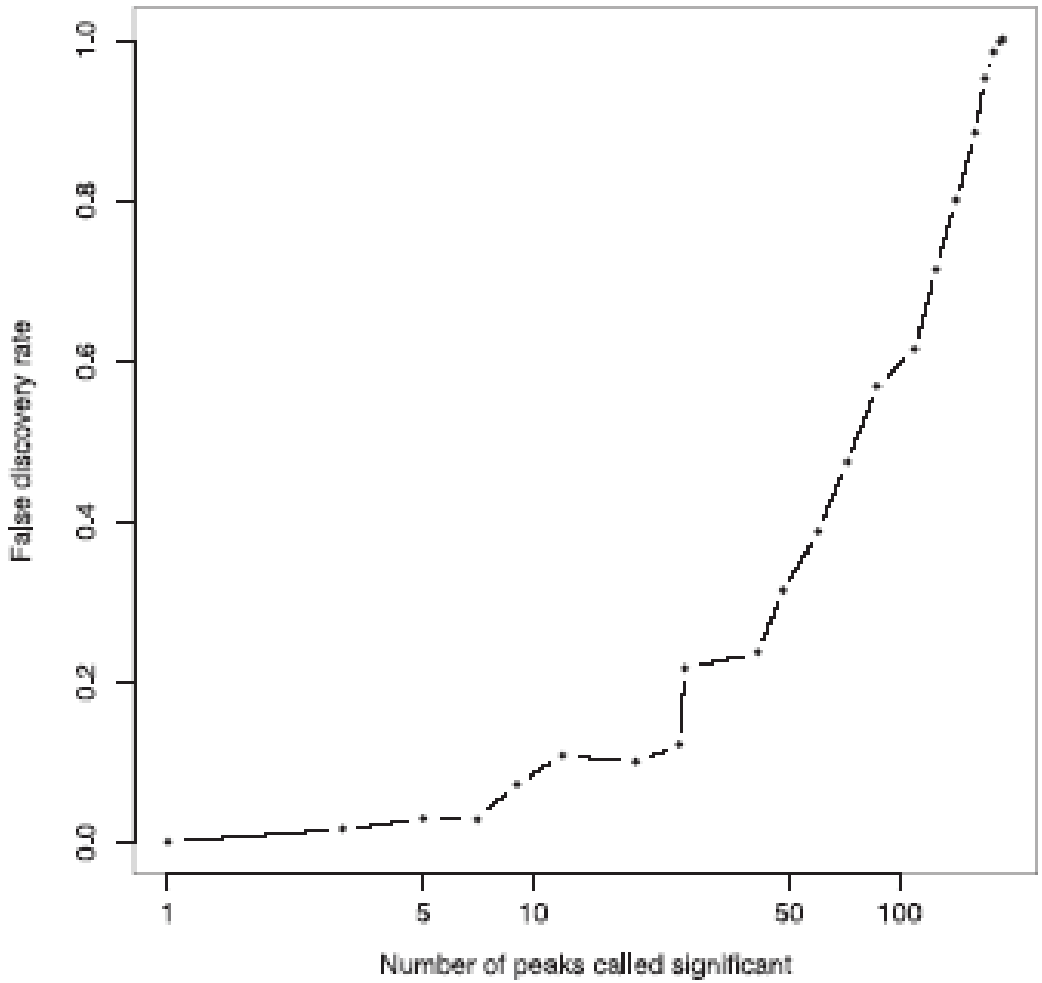
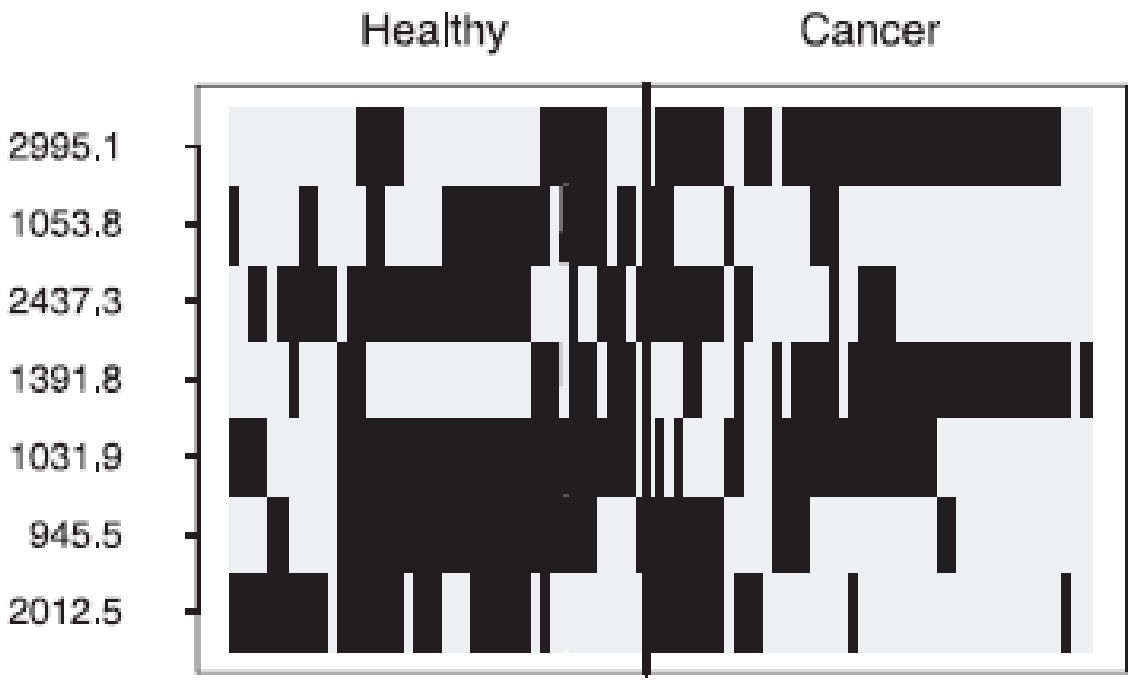


Fig. 6. Left column: three spectra from cancer patients having a peak higher than 0.55 at the site $m/z = 2995.1$. Both the raw (black) and smoothed (red) spectra are shown. In the right column, we show three spectra from healthy patients without the peak, or whose peak is too low. The vertical dotted lines indicate the centroid 2995.1 and the outer limits for the peak position.


False discovery rate
0.0
0.2
0.4
0.6
0.8
1.0
1
5
10
50
100
Number of peaks called significant



Heatmap (binary) of top seven training set peaks in ovarian data

Strongest peak is at m/z= 2995.1