Proteomics: Finding the Needle in the Haystack

Statistics: **Xuena Wang*, Kith Pradhan*, Chen Ji*, Xiangfeng Wu*, Yeming Ma**, Wei Zhu*;**

Operations Research: **Piyush Kumar*, Valentin Polishchuk*, Olaf Hall-Holt*, Joseph Mitchell*, Estie Arkin*;**

Computational Mathematics: **James Glimm*;**

Long Island Cancer Center: **John Kovach*****

*Department of Applied Mathematics and Statistics, ***Long Island Cancer Center; State University of New York at Stony Brook

**Medical Department, Brookhaven National Laboratory

Potentially the most profound advance in medical diagnostics in the last half century

- Cancer diagnostics
  - Ovarian, breast, colon, pancreatic, prostrate, bone, endometrial brain, liver, lung, head and neck
  - Subtypes
  - Stages, response to therapy
  - Identification of specific proteins and their function
- Alzheimer's disease
- Probably applicable to many conditions

The nature of the data extraction problem: large volume, low signal to noise ratio

- Data quantity
  - Many markers, many subjects, many groups in which subjects have been or will be placed, multiple measurements (time series) per subject
- Data quality
  - Individual marker does not always correlate perfectly with disease, so that multiple markers are needed to achieve high accuracy
- Analysis requirements
  - Near 100% accuracy for rare cancers (e.g. ovarian cancer)

Data Volume and Complexity

Data	Markers	Subjects	Groups	Measurements
4-3-02	15,154	216	3	1
8-7-02	15,154	153	2	1
Q* High resolution	368,750	216	2	1
Future	Millions	1000’s	Multiple	10’s

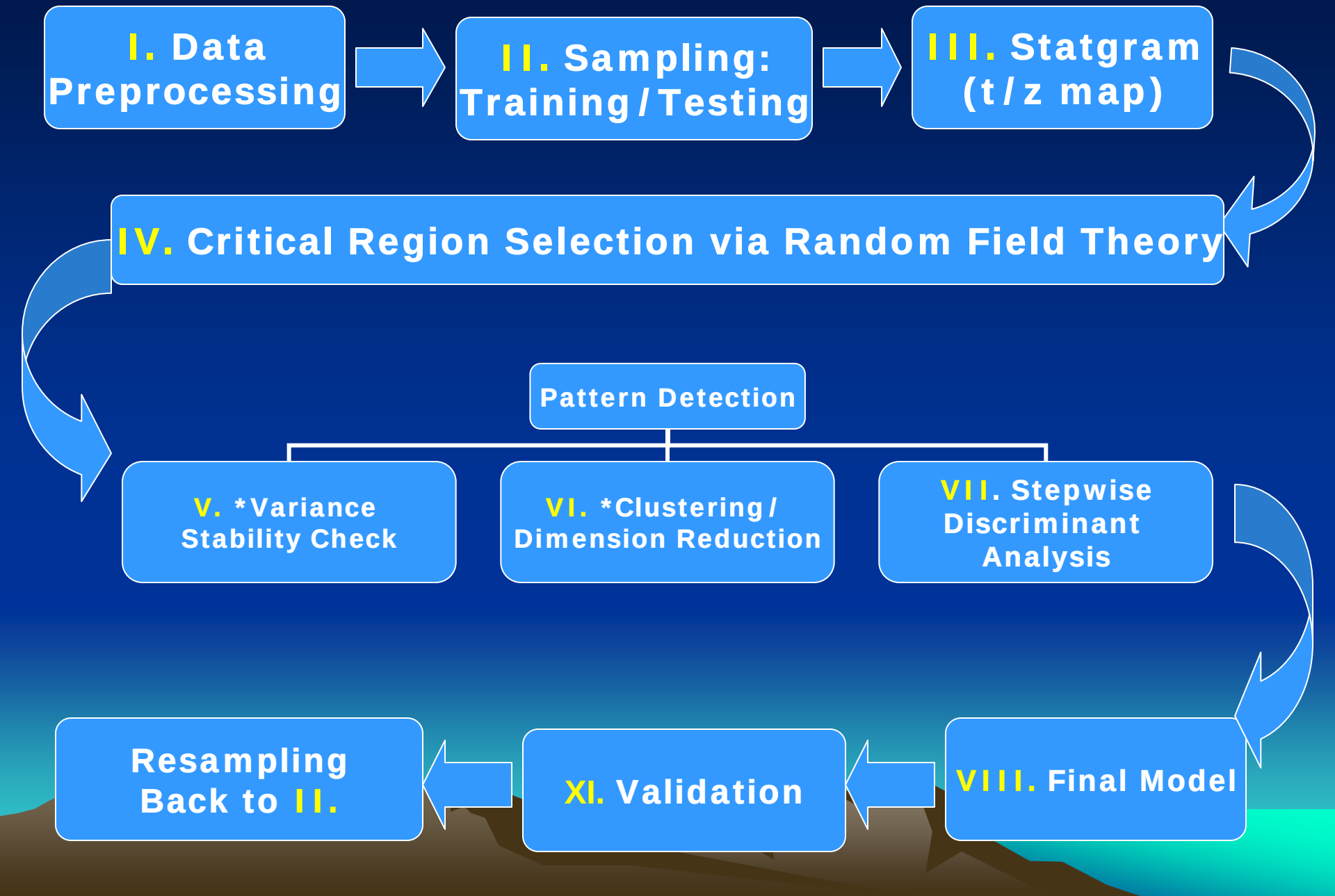
Q-star Ovarian Cancer Data

(NCI/FDA Ovarian High Resolution QqTof SELDI Data)
http://ncifdaproteomics.com/ppatterns.php

Unaffected Women	*95*
Women with ovarian cancer	*121*
Total	216

Low Resolution Data Sets

- http://clinicalproteomics.steem.com/download-ovar.php
- 4-3-02
  - 100 ovarian cancer patients
  - 116 unaffected individuals
- 8-7-02
  - 162 ovarian cancer patients
  - 91 unaffected individuals

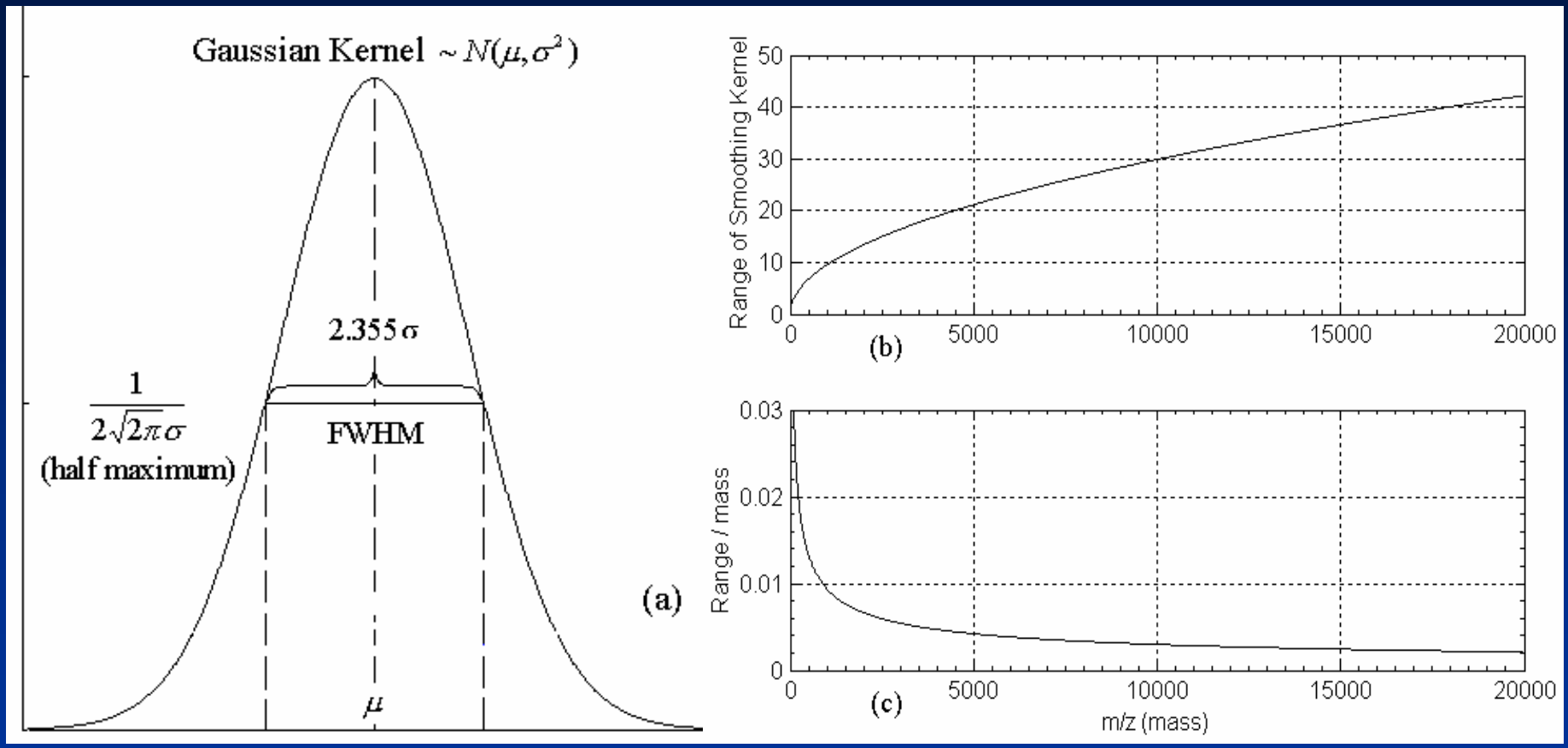

I. Data Preprocessing
II. Sampling: Training / Testing
III. Statgram (t / z map)
IV. Critical Region Selection via Random Field Theory
Pattern Detection
V. *Variance Stability Check
VI. *Clustering / Dimension Reduction
VII. Stepwise Discriminant Analysis
VIII. Final Model
XI. Validation
Resampling Back to II.

Ia. Data Preprocessing: Smoothing

(1) De-noise and thus enhance the signal to noise ratio;

(2) Smooth with Gaussian kernel to enable a multiple-test correction via the random field theory.

(3) Alternate method: smooth with least squares polynomial fitting to remove noise



(a) Illustration of the Gaussian kernel. The smoothed intensity of the biomarker with m/z value is calculated as the weighted average, proportional to the Gaussian density, of the intensities of its neighboring biomarkers. (b) Relationship between the range of the 11 adjacent points within the Gaussian smoothing kernel (y-axis) and the median of the range (x-axis) when FWHM = 11. (c) Relationship between the ratio of the range over its median (y-axis) and the median of the range (x-axis) for FWHM = 11.

Ib. Data Preprocessing: Normalization

To ensure that the spectra are comparable across subjects/runs, we normalize each spectrum by first computing the average intensity of the entire spectrum, and then use the intensity for each m/z value divided by the average intensity.

* Prior to normalization, one must ensure that the baseline subtraction is done uniformly across the spectra.

Ic. Data Preprocessing: Outlier Detection

A two-step clustering analysis via (1) the K-mean algorithm and (2) hierarchical classification is performed to examine:

(a) Whether there are any sub-groups in each category (cancerous or normal).

(b) Whether there is any outlier or abnormal spectra in the data. The most obvious outliers are those ‘in a league of their own’ via the K-mean clustering.

II. Training/Testing Sets

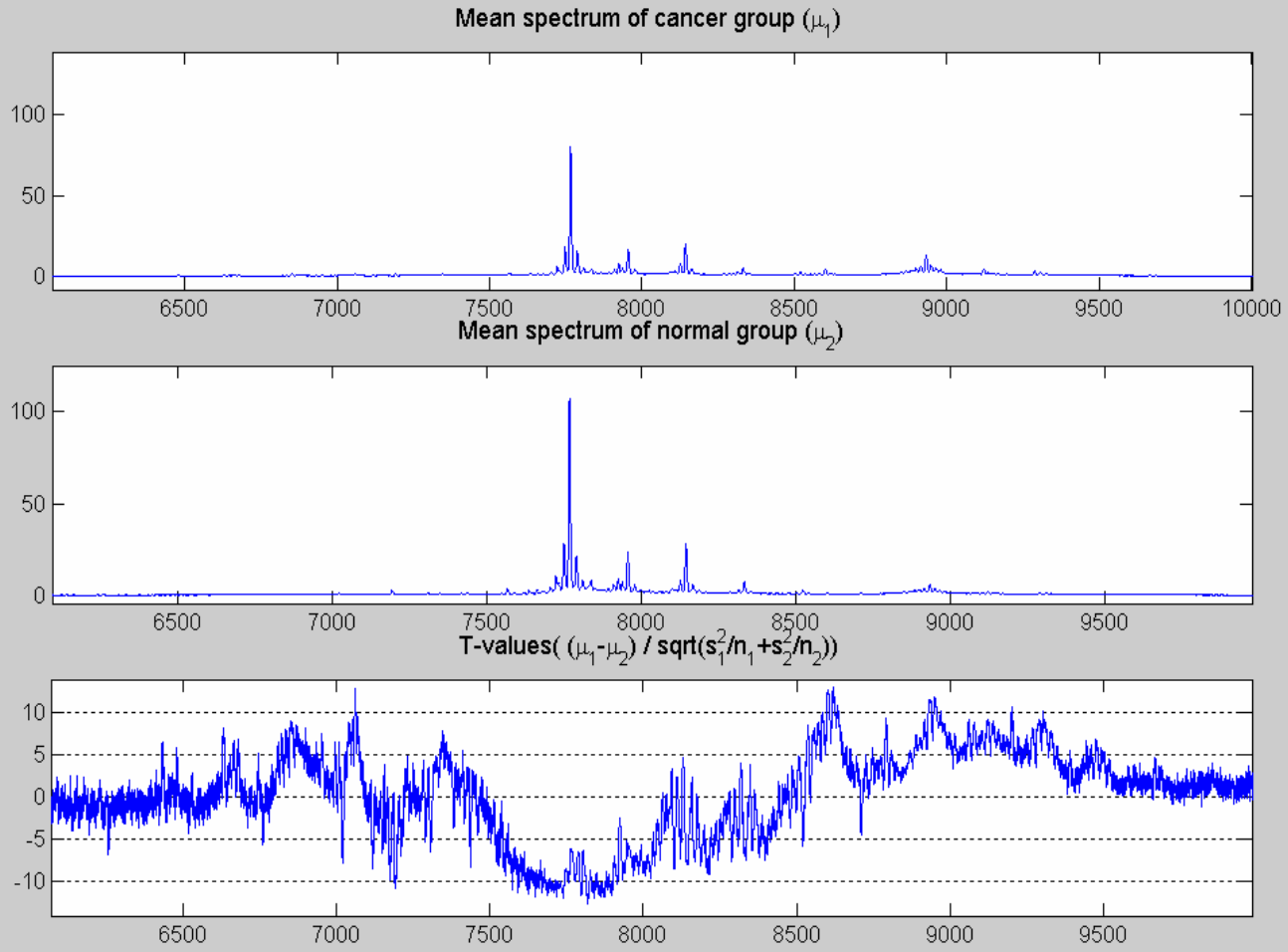
(1) Training Set: A certain number of subjects (e.g. half) selected at random from each group (cancerous or normal).

(2) Testing Set: The remaining subjects in each group.

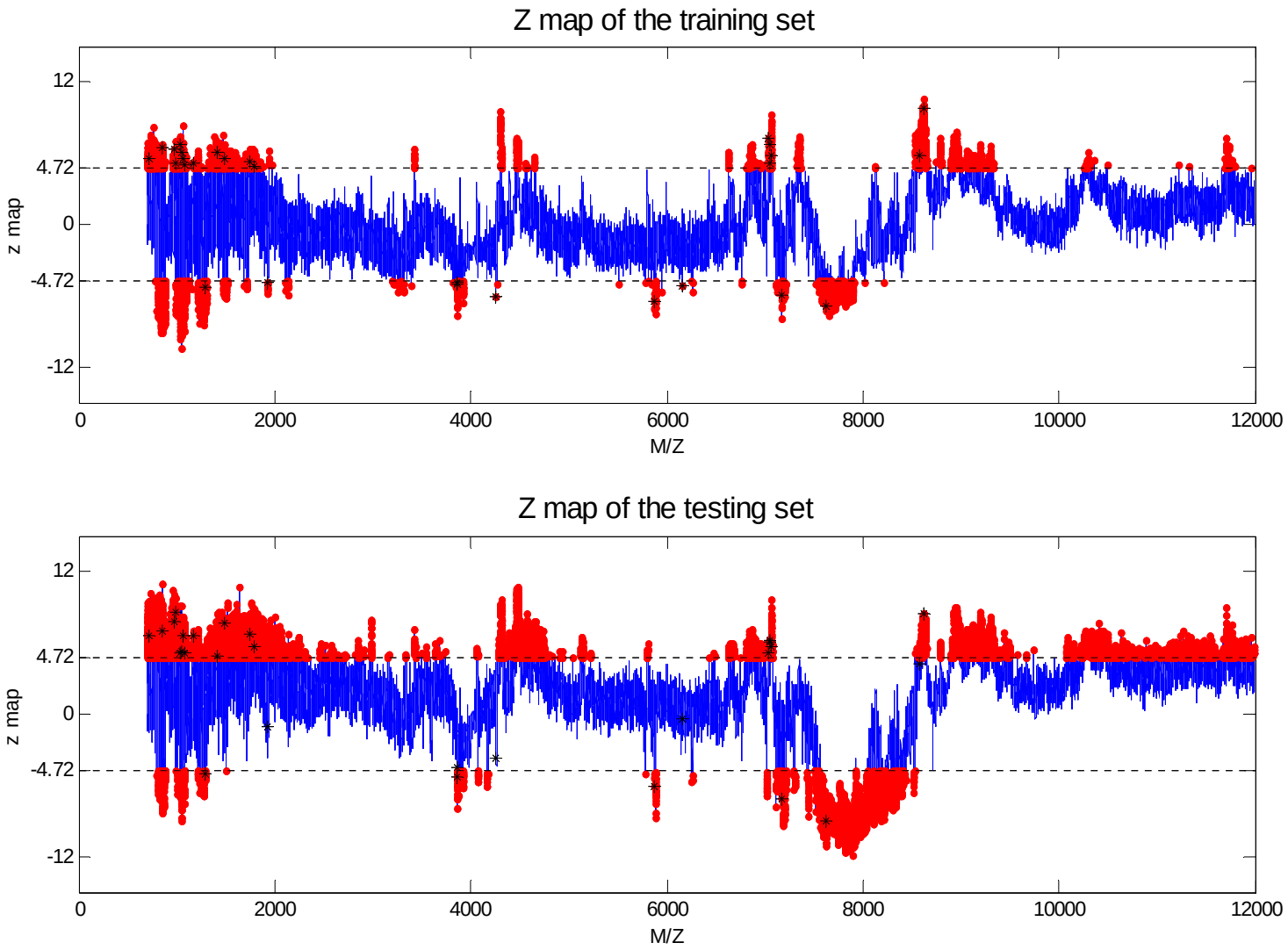
Q* Ovarian Cancer Data

Group	*Training Data*	*Testing Data*
Cancer	*60*	*61*
Normal	*47*	*48*
Total	107	109

III. Statgram (t/z map)



III. Statgram (t/z map)



t test for statistical significance

$$t = \frac{\overline{y_1(x)} - \overline{y_2(x)}}{\sqrt{s_1^2(x)/n_1 + s_2^2/n_2}}$$

$\overline{y_1}, \overline{y_2}$ = means

s_1, s_2 = STD

n_1, n_2 = number of subjects

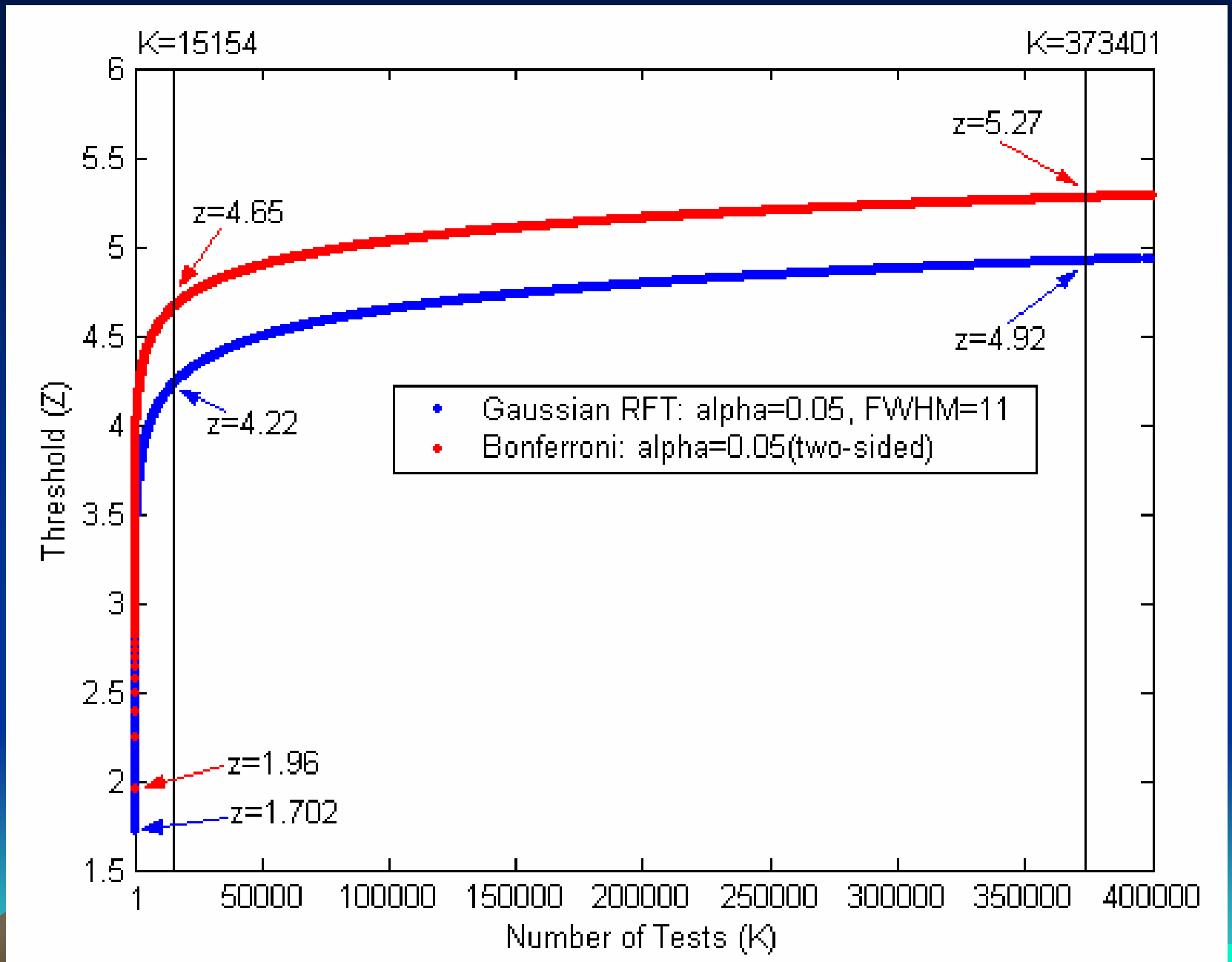
$x = m/z$ value

IV. Critical region selection & multiple-test correction via the random field theory (z-test)

$$\alpha \approx \int_t^{\infty} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du + \frac{K\sqrt{\ln 2}}{\pi(FWHM)} e^{-t^2/2}$$

IV. Critical region selection & multiple-test correction via the random field theory (t-test with v degrees of freedom)

$$\alpha \approx \int_t^{\infty} \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)\sqrt{v\pi}}\left(1+\frac{u^2}{v}\right)^{-\frac{v+1}{2}} du + \frac{K\sqrt{\ln 2}}{\pi(FWHM)}\left(1+\frac{t^2}{v}\right)^{-\frac{v+1}{2}}$$


K=15154
K=373401
z=5.27
z=4.65
z=4.92
z=4.22
z=1.96
z=1.702
Gaussian RFT: alpha=0.05, FWHM=11
Bonferroni: alpha=0.05(two-sided)
Threshold (Z)
Number of Tests (K)
6
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1
50000
100000
150000
200000
250000
300000
350000
400000

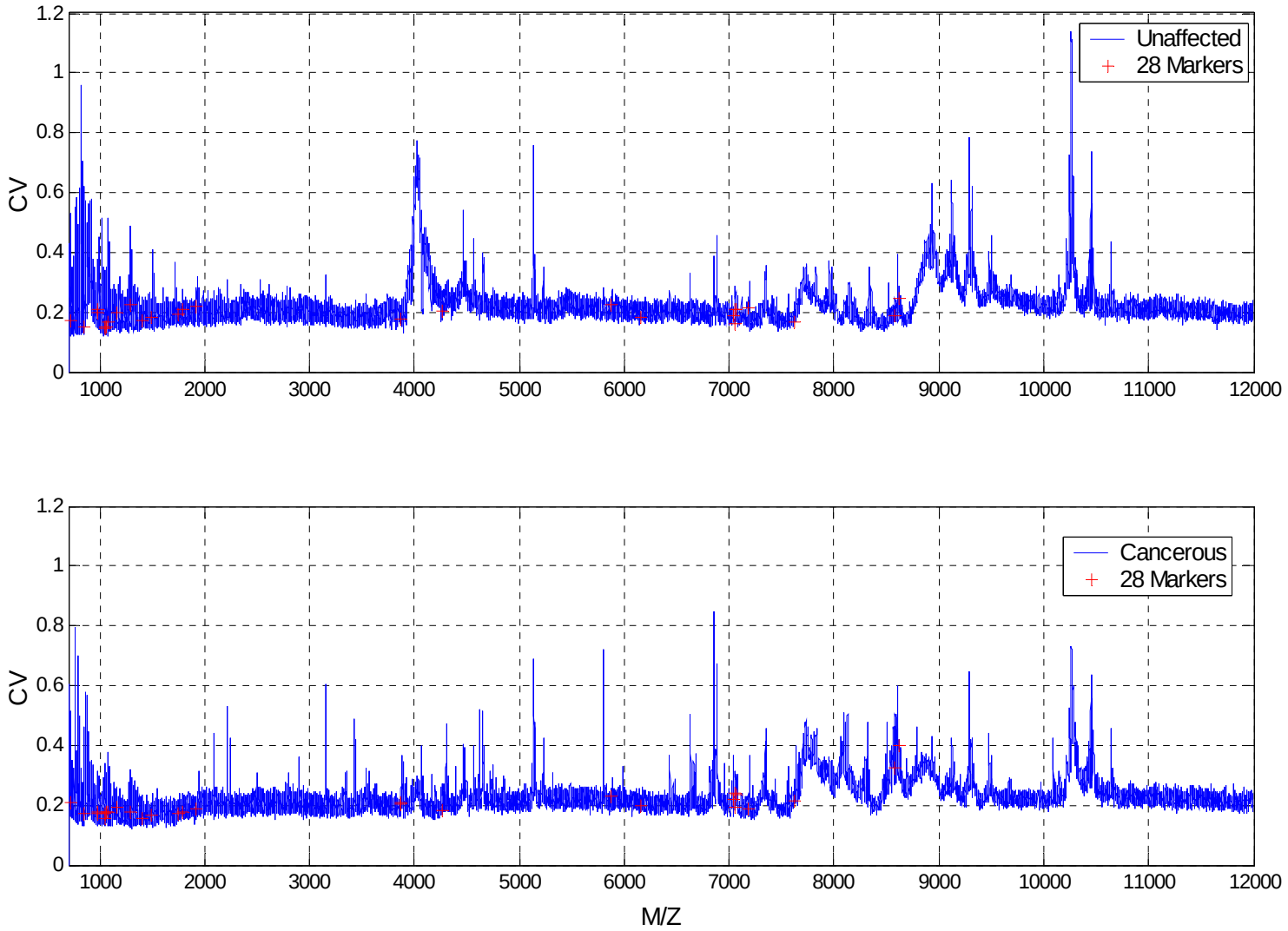
Thresholding Step Completed

- All selected markers are statistically significant
- Improved chance of validity when applied to large population samples
- Improved chance of success when searching for protein giving rise to marker

V. Noisy Markers

- CV = coefficient of variation = STD/mean
- Large CV indicates noisy marker
- Possible causes of large CV
  - Marker distinguishes between subgroups that have been lumped together in study
  - Marker is subject to random variation within the population
- Avoid large CV markers for diagnosis but worth understanding for science

V. Variance Stability Check*



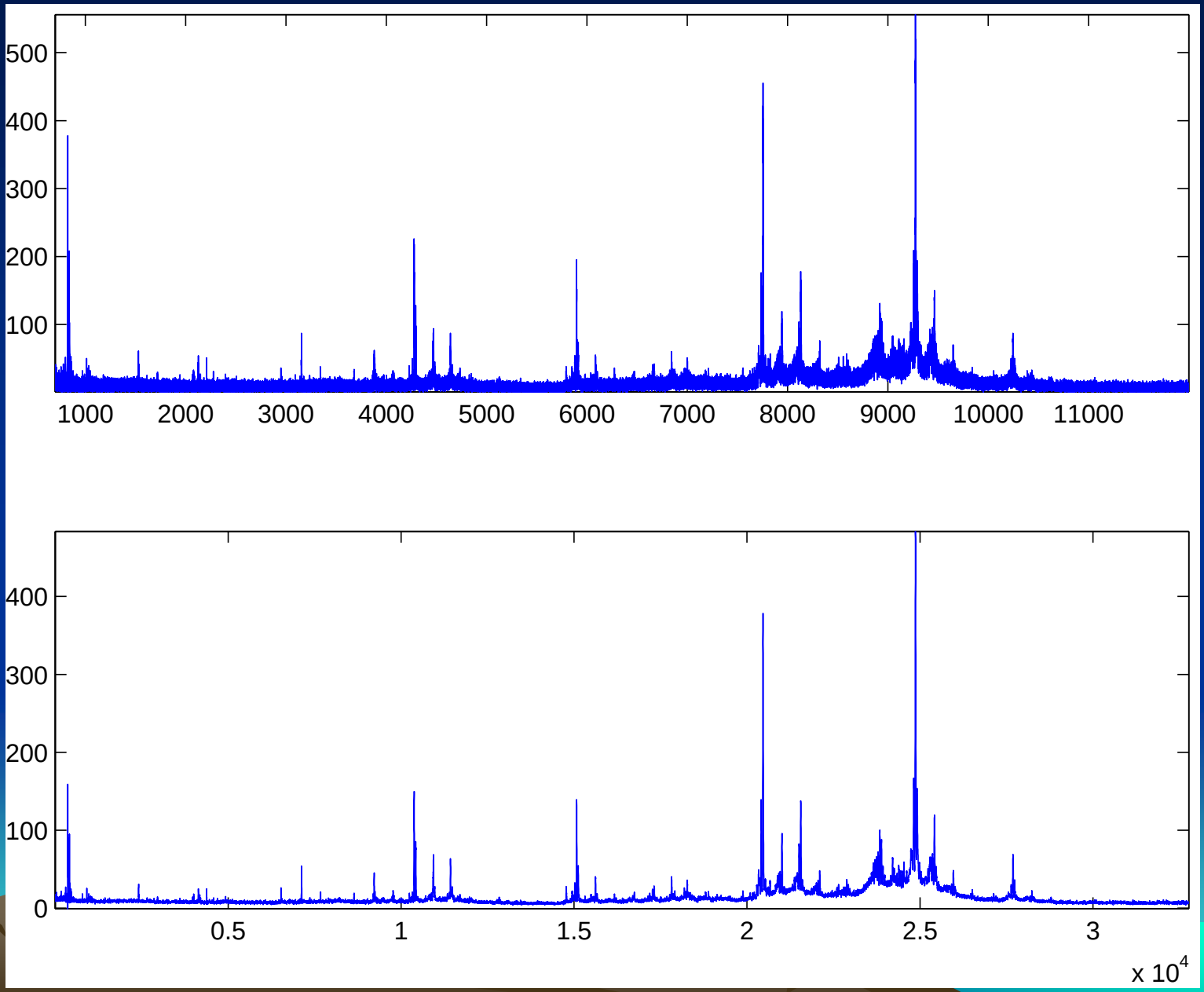
VI. Clustering of Markers*

- We use the K-mean clustering algorithm. Each new marker is near some existing cluster and is added to it or -- if not near -- it starts its own new cluster.
- Correlation between markers is computed across subjects and equals covariance / mean x mean
- Select one marker per cluster
- This is an optional procedure often done when the number of markers selected is too large.
- **Alternate method:** double discrete cosine transformation provides an averaging-data compression method to reduce the number of markers

Cosine Dimension Reduction

Original (368750)

Reduced (32768)



VII. Stepwise Discriminant Analysis

- Marker pool: Significant markers selected via the random field theory or Mahalanobis and (Fisher) linear discriminant analysis
- Markers can be selected and de-selected during the stepwise procedures
- A marker set is selected as the final model if it achieves 100% cross-validation in the training data set.

Evaluate Features Selected Using Cross Validation

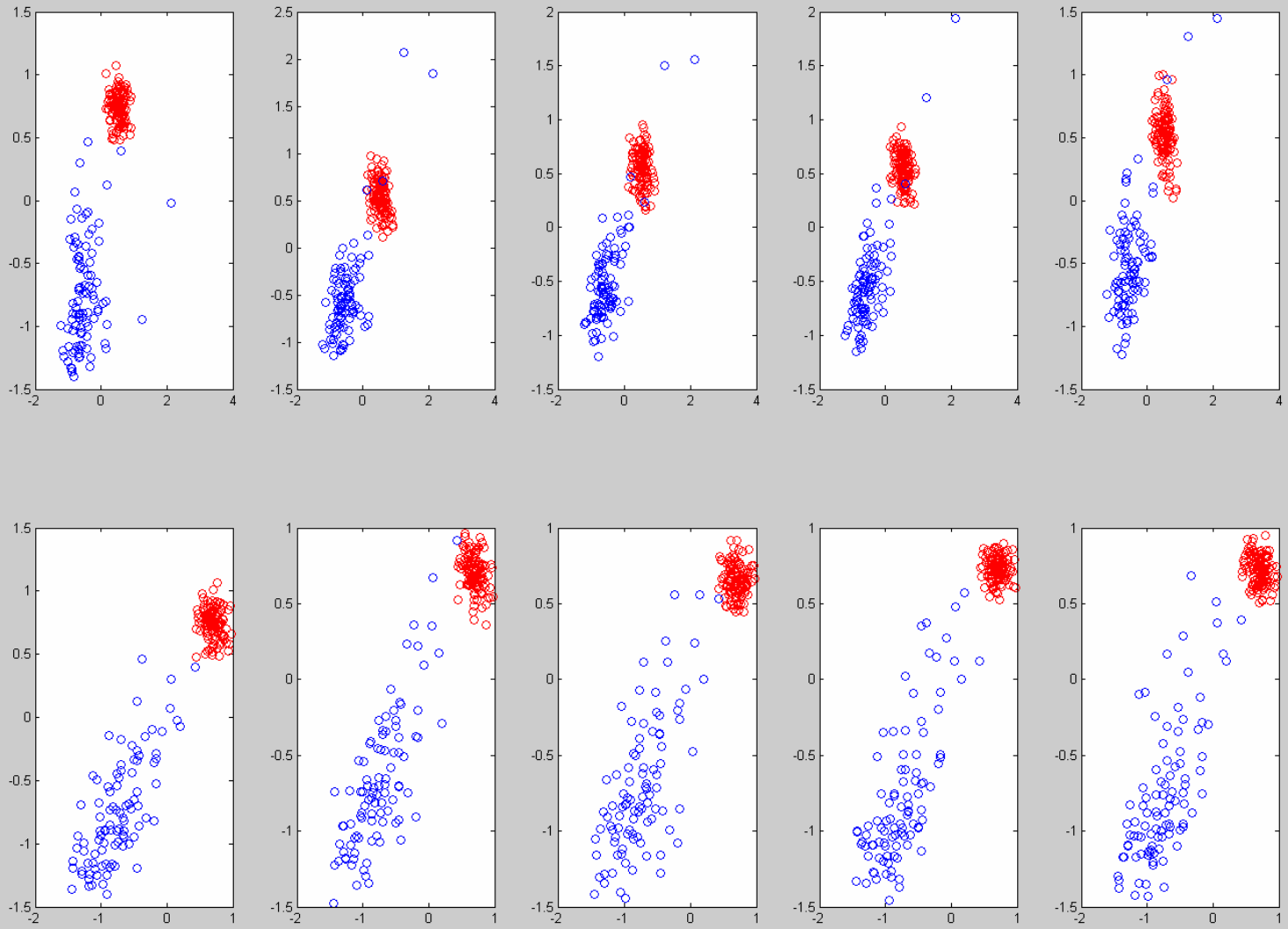
1. K-Nearest Neighbors (K-NN)
2. Neural Network (NN)
3. Support Vector Machine (SVM)

The K-NN Algorithm

There are five major steps to this algorithm:

1. Data pre-processing (entire data);
2. Sampling (divide the data into training/testing);
3. Pre-marker selection via the random field theory (training data);
4. Final marker selection via stepwise discriminant analysis (training data);
5. Validation of the final model via KNN (testing data).

Example Features



VIII. Validation via the Testing Set

- Binary decision: Use the K-nearest neighbor algorithm (e.g. K = 11) to score the testing subjects as cancerous or not
- Probability based scoring: Computes the likelihood ratio of cancer versus normal for each subject in the testing set.
- Alternate method: SVM, Neural Nets

K-Nearest Neighbor Scoring

- Each testing subject finds its K (e.g. K = 11) nearest neighbors in the training set.
- The disease state of the majority of these K training set subjects determines the predicted state of the testing subject.
- The algorithm depends on a distance between subjects, and this distance depends on the selected set of markers.
- Mahalanobis distance used to determine nearest neighbors.

Mahalanobis distance

- This is the covariance adjusted distance of a test subject scores to the mean marker scores of the cancerous and non-cancerous populations as defined by the training set.

$$z_i = (y - \overline{x_i})^T S_i^{-1} (y - \overline{x_i})$$

$i = 1$: Cancerous

$i = 2$: Non-cancerous

$S_i =$ covariance of K selected markers

in the $i = 1, 2$ population of the training group

Probability Based Scoring

L_i = probability of having scores z_i given disease state i.

$$\ln L_1 / L_2 = 0.5\left[\ln |S_2| - \ln |S_1| - (z_1 - z_2)\right]$$

$|S|$ = determinant S

Convert to probability of disease state given score by use of Bayes' formula

Results on the Q* Ovarian Data

10 training-testing set pairs

Average number of markers: 52

Success rate: 121.4/122 = 99.51% correct

Results

- Example Marker set (26) in reduced dimension (That resulted in 100% accuracy)

26	40	125	221	17844
17846	17847	17848	17849	22920
22921	22922	22923	22924	
22925	22926	23893	22910	
22911	22912	22913	22914	
22915	22916	22917	22918	

Low Resolution Data 1, 2

- Our algorithm achieved 100% sensitivity and 100% specificity for both data sets
- 18 protein biomarkers identified for Data 4-3-02 can also achieve perfect discrimination for all subjects in Data 8-7-02 via cross-validation.
- Problems with data:
  - (a) differential intensities at some markers inconsistent between two data sets;
  - (b) Data 8-7-02 is “easy” to classify, including with use of very low m/z markers, i.e. m/z = 2.79;
  - (c) offset of peaks between two data sets in high m/z region

Low Resolution, Continued

- Statistics for repeated choices of training/testing sets
  - Data 4-3-02 (cancer versus normal): many perfect results; mean success rate 96% (1200 training-testing pairs sampled).
  - Data 8-7-02 (cancer versus normal): many perfect results; mean success rate near 100%
  - Furthermore, marker sets selected from the 4-3-02 training data (50 cancer + 50 normal) can classify the entire 8-7-02 data via cross-validation with a mean success rate of 96%.
- Need to re-examine results in view of new questions regarding the data
  - Shift to align peaks between data
  - Compensate for effects of baseline subtraction
  - Statistical stability of cross data set comparison
  - Biological meaning of the markers to be determined in all cases
- Future work will emphasize high resolution data

Ovarian Cancer Screening

- Prevalence of ovarian cancer is 1 in 2500
- Screening with a test of 97% specificity would result in 75 false positives for every true positive case identified

- A suitable screening test must have near 100% specificity

Acknowledgement

- Parts of work were published in the Proceedings of the National Academy of Science.
- Support from the University at Stony Brook

- We are currently working on
  - Longitudinal profiling based on proteomic mass spectrometry
  - Cross-sectional classification and profiling for multiple diseases
  - Correlating proteomic markers with other covariates

Cancer Incidence