

Classification Continued

- Support Vector Machines
 - Ideas / Development
 - Microarray Issues / Example
- Nearest Shrunken Centroids
 - Ditto

Support Vector Machines

Separating Hyperplanes

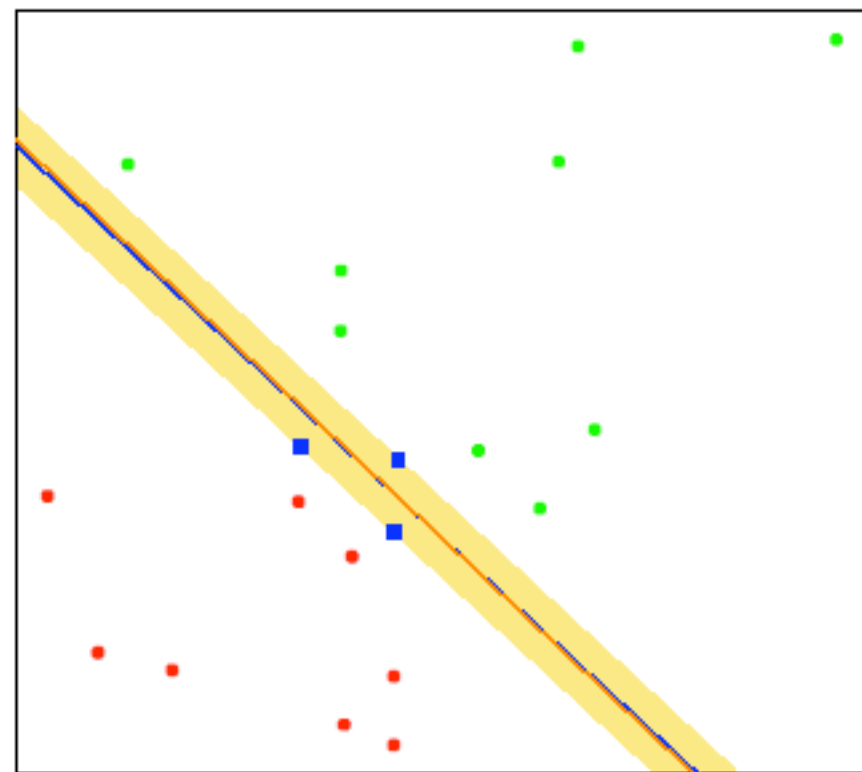
The separating hyperplane with **maximum margin** is likely to perform well on test data.

Here the separating hyperplane is almost identical to the more standard linear logistic regression boundary;

see



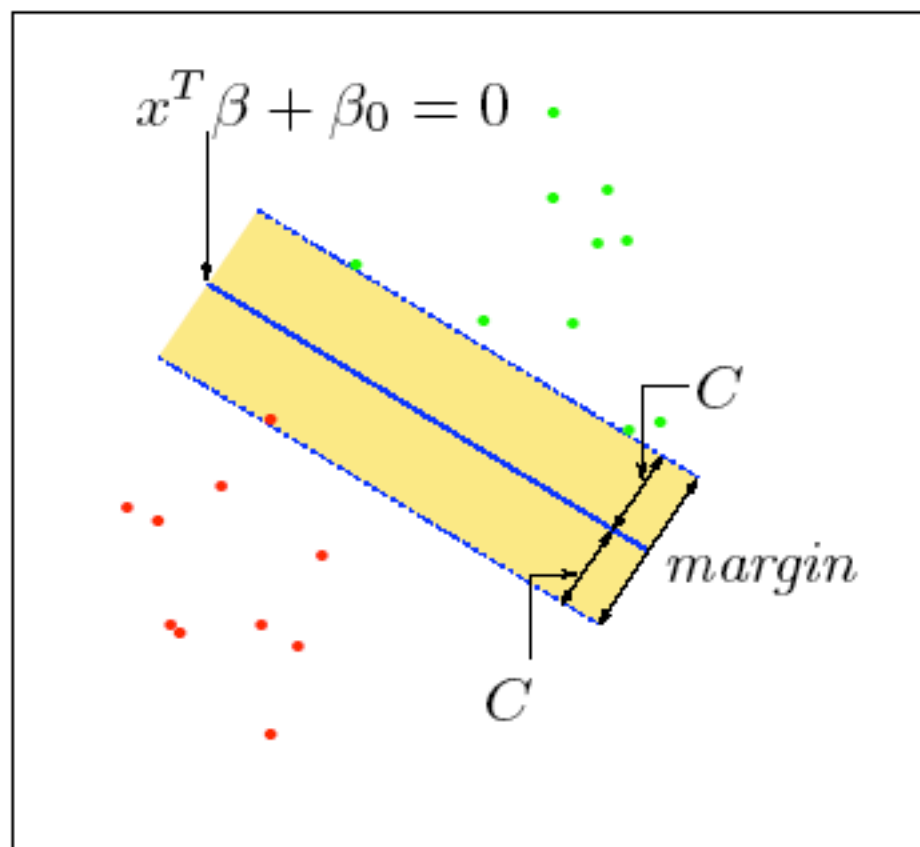
pp 95.



Maximum Margin Classifier

Vapnik(1995)

$$x_i \in \mathbb{R}^p, y_i \in \{-1, 1\}$$

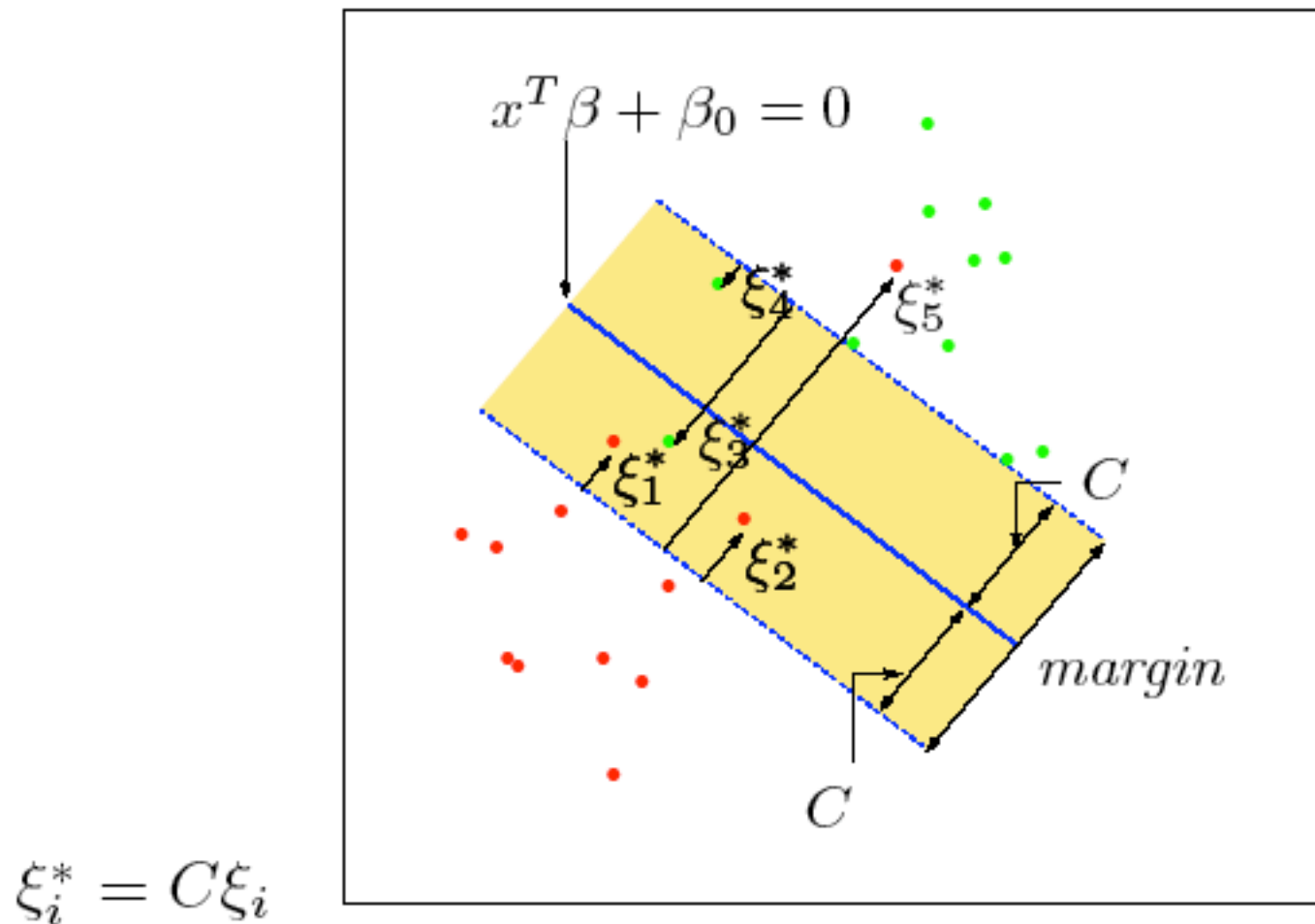


$$\max_{\beta, \beta_0, \|\beta\|=1} C$$

subject to $y_i(x_i^T \beta + \beta_0) \geq C, i = 1, \dots, N.$

Note: $y_i(x_i^T \beta + \beta_0)$ is distance from x_i to boundary.

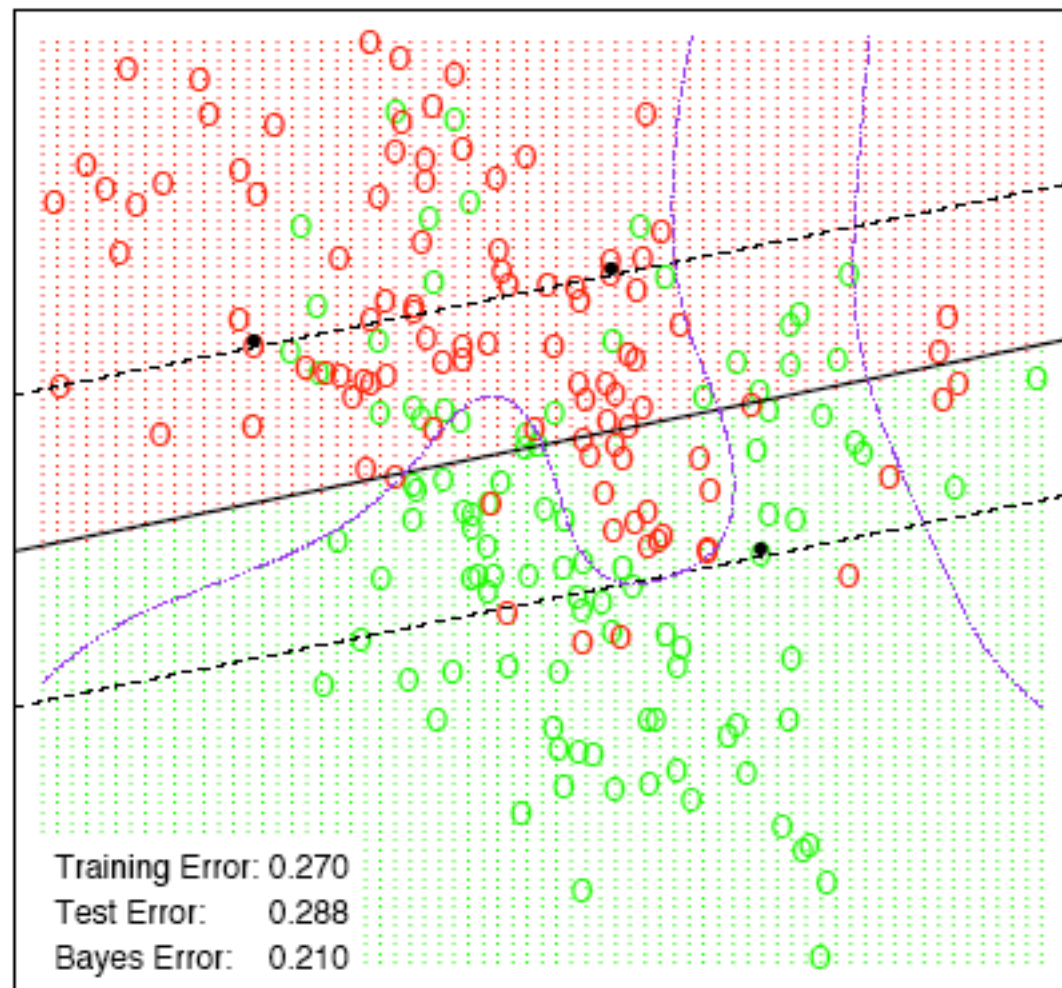
Overlapping Classes



$$\max_{\beta, \beta_0, \|\beta\|=1} C$$

subject to $y_i(x_i^T \beta + \beta_0) \geq C(1 - \xi_i)$, $\xi_i \geq 0$, $\sum_i \xi_i \leq B$

Example



Fitted function is $\hat{f}(x) = x^T \hat{\beta} + \hat{\beta}_0$

Resulting classifier is $\hat{G}(x) = \text{sign}[\hat{f}(x)]$

Quadratic Programming Solution

After a lot of *stuff* we arrive at a Lagrange dual

$$L_D = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{i'=1}^N \alpha_i \alpha_{i'} y_i y_{i'} x_i^T x_{i'}$$

which we maximize subject to constraints (involving B as well).

The solution is expressed in terms of fitted Lagrange multipliers $\hat{\alpha}_i$:

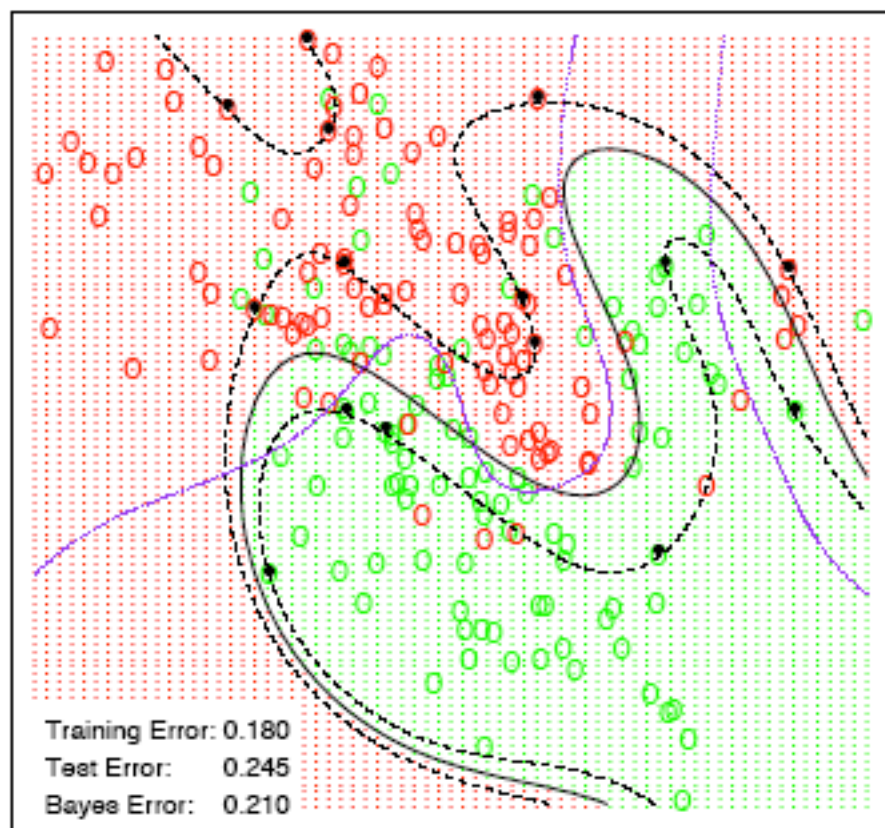
$$\hat{\beta} = \sum_{i=1}^N \hat{\alpha}_i y_i x_i$$

Some fraction of $\hat{\alpha}_i$ are exactly zero (from KKT conditions); the x_i for which $\hat{\alpha}_i > 0$ are called **support points** \mathcal{S} .

$$\hat{f}(x) = x^T \hat{\beta} + \hat{\beta}^0 = \sum_{i \in \mathcal{S}} \hat{\alpha}_i y_i x^T x_i + \hat{\beta}^0$$

Flexible Classifiers

SVM - Degree-4 Polynomial in Feature Space



Enlarge the feature space via basis expansions, e.g. polynomials of total degree 4. $h(x) = (h_1(x), h_2(x), \dots, h_M(x))$

$$\hat{f}(x) = h(x)^T \hat{\beta} + \hat{\beta}_0$$

SVM

$$L_D = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{i'=1}^N \alpha_i \alpha_{i'} y_i y_{i'} \langle h(x_i), h(x_{i'}) \rangle$$

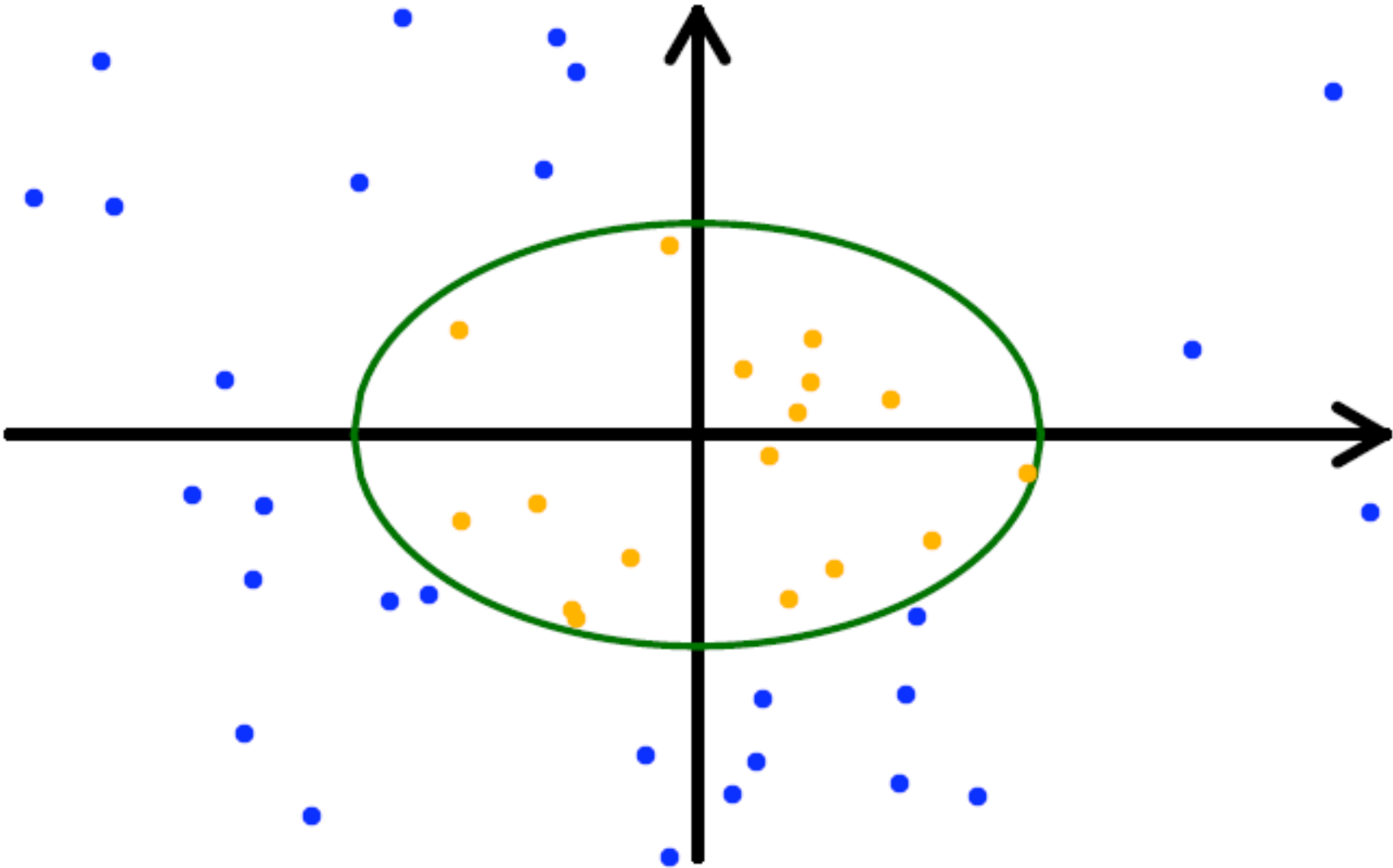
$$\begin{aligned} f(x) &= h(x)^T \beta + \beta_0 \\ &= \sum_{i=1}^N \alpha_i y_i \langle h(x), h(x_i) \rangle + \beta_0. \end{aligned}$$

L_D and constraints involve $h(x)$ only through inner-products

$$K(x, x') = \langle h(x), h(x') \rangle$$

Given a suitable positive kernel $K(x, x')$, don't need $h(x)$ at all!

$$\hat{f}(x) = \sum_{i \in \mathcal{S}} \hat{\alpha}_i y_i K(x, x_i) + \hat{\beta}_0$$



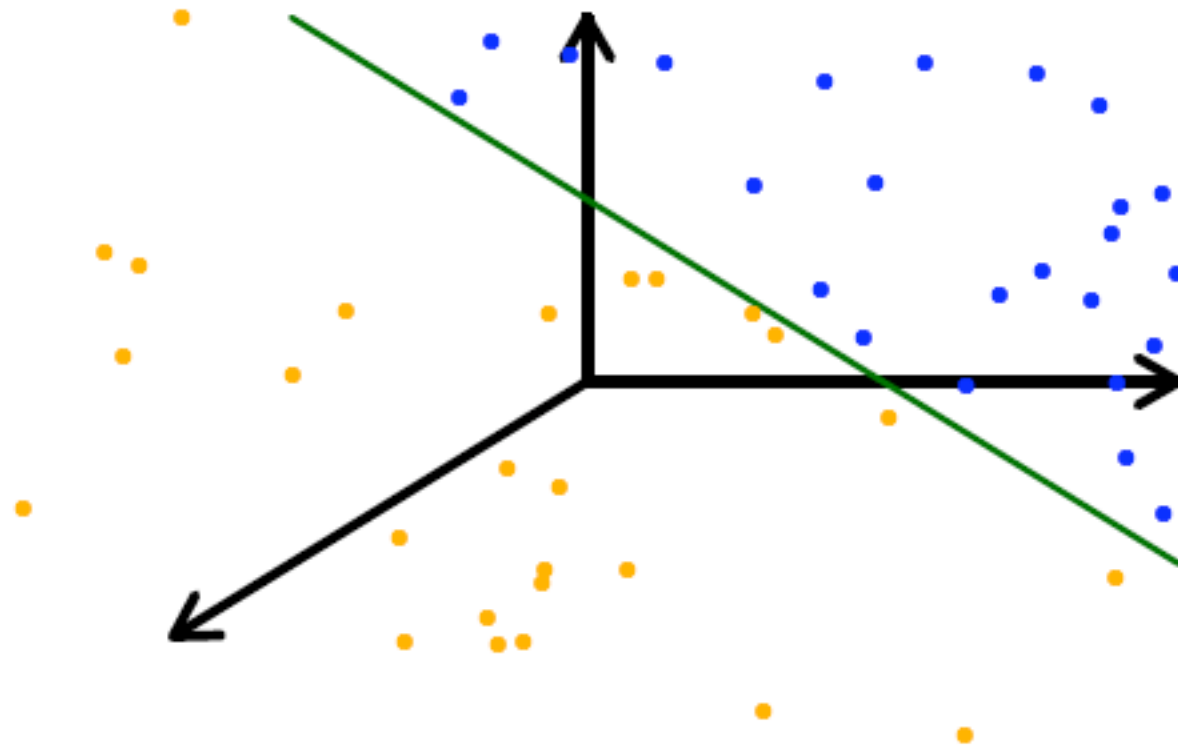
To deal with ellipsoidal boundary, transform
original features

x_1, x_2

to new features

$$z_1 = x_1^2, \quad z_2 = \sqrt{2} x_1 x_2, \quad z_3 = x_2^2$$

In the **new feature space** the boundary is a hyperplane!



Popular Kernels

$K(x, x')$ is a symmetric, positive (semi-)definite function.

*d*th deg. poly.: $K(x, x') = (1 + \langle x, x' \rangle)^d$

radial basis: $K(x, x') = \exp(-\|x - x'\|^2/c)$

Example: 2nd degree polynomial in \mathbb{R}^2 .

$$\begin{aligned} K(x, x') &= (1 + \langle x, x' \rangle)^2 \\ &= (1 + x_1x'_1 + x_2x'_2)^2 \\ &= 1 + 2x_1x'_1 + 2x_2x'_2 + (x_1x'_1)^2 + (x_2x'_2)^2 + 2x_1x'_1x_2x'_2 \end{aligned}$$

Then $M = 6$, and if we choose

$$h_1(x) = 1, h_2(x) = \sqrt{2}x_1, h_3(x) = \sqrt{2}x_2, h_4(x) = x_1^2, h_5(x) = x_2^2,$$

$$\text{and } h_6(x) = \sqrt{2}x_1x_2,$$

$$\text{then } K(x, x') = \langle h(x), h(x') \rangle.$$

The Kernel trick

- Linear regression model: given $n \times p$ model matrix X and response n -vector y , fitted values are given by

$$\hat{y} = X(X^T X)^{-1} X^T y$$

- When $X^T X$ is singular (e.g. if $p > n$), solution is not unique; **ridge regression** adds a positive constant to its diagonal:

$$\hat{y}_{rr} = X(X^T X + \lambda I)^{-1} X^T y$$

- Can rewrite above as

$$\hat{y}_{rr} = K(K + \lambda I)^{-1} y$$

where $K = X X^T$ is the $n \times n$ matrix of inner products between the feature vectors.

- Suppose we now have $h(x)$, a vector of $p \gg n$ basis function in x . Suppose as well, $\mathcal{K}(x, x') = \langle h(x), h(x') \rangle$.
- We fit the model $f(x) = h(x)^T \beta$ by penalized least squares:

$$\min_{\beta} \sum_{i=1}^N (y_i - h(x_i)^T \beta)^2 + \lambda \beta^T \beta.$$

- It is easy to show that

$$\hat{f}(x) = \sum_{i=1}^N \hat{\alpha}_i \mathcal{K}(x, x_i),$$

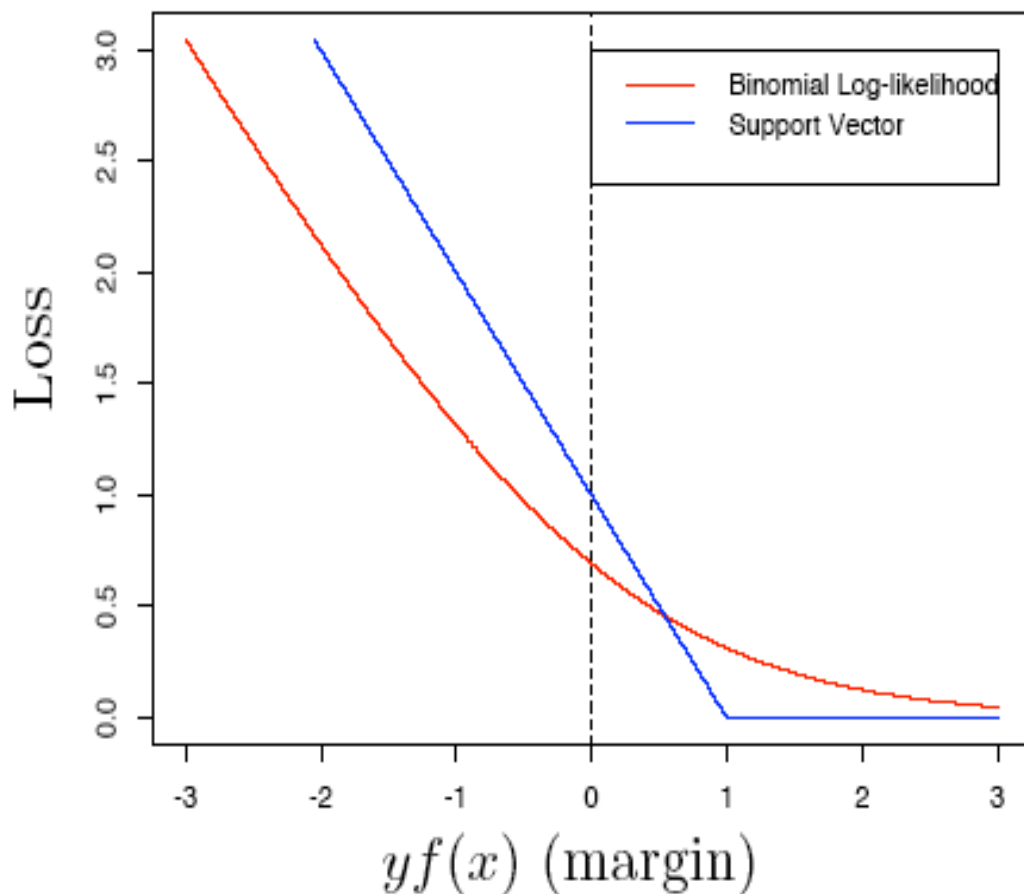
where

$$\hat{\alpha} = (K + \lambda I)^{-1} y,$$

and K is the $N \times N$ matrix $\{K\}_{ij} = \mathcal{K}(x_i, x_j)$.

- Hence we can fit a penalized regression model in any feature space for which we have an inner-product kernel.

SVM via Loss + Penalty



With $f(x) = h(x)^T \beta + \beta_0$ and $y_i \in \{-1, 1\}$, consider

$$\min_{\beta_0, \beta} \sum_{i=1}^N [1 - y_i f(x_i)]_+ + \lambda \|\beta\|^2$$

Solution identical to SVM solution, with $\lambda = \lambda(B)$.

$$\text{In general } \min_{\beta_0, \beta} \sum_{i=1}^N L[y_i, f(x_i)] + \lambda \|\beta\|^2$$

Loss Functions

For $Y \in \{-1, 1\}$

Log-likelihood: $L[Y, f(X)] = \log(1 + e^{-Y f(X)})$

- (negative) binomial log-likelihood or **deviance**.
- estimates the **logit**

$$f(X) = \log \frac{\Pr(Y = 1|X)}{\Pr(Y = -1|X)}$$

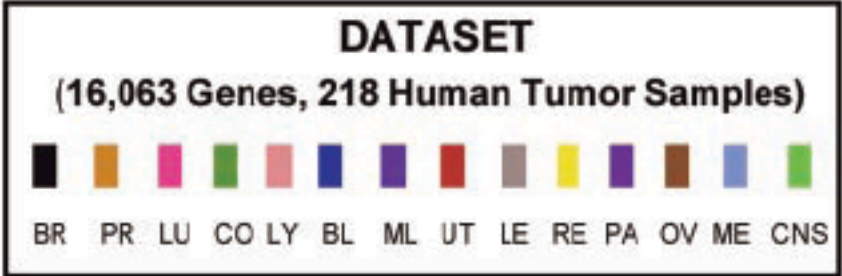
SVM: $L[Y, f(X)] = (1 - Y f(X))_+$.

- Called “**hinge loss**”
- Estimates the **classifier** (threshold)

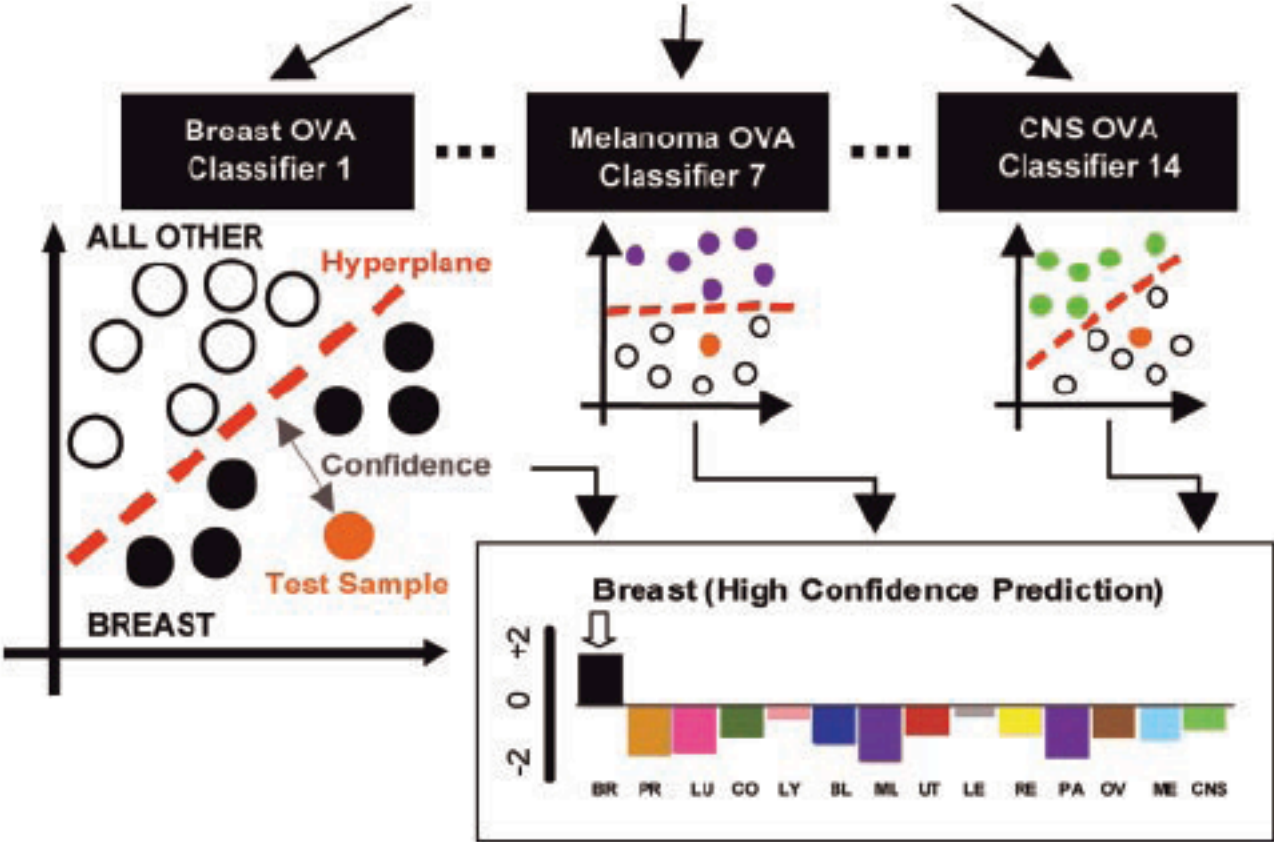
$$C(x) = \text{sign} \left(\Pr(Y = 1|X) - \frac{1}{2} \right)$$

SVMs for Expression Arrays

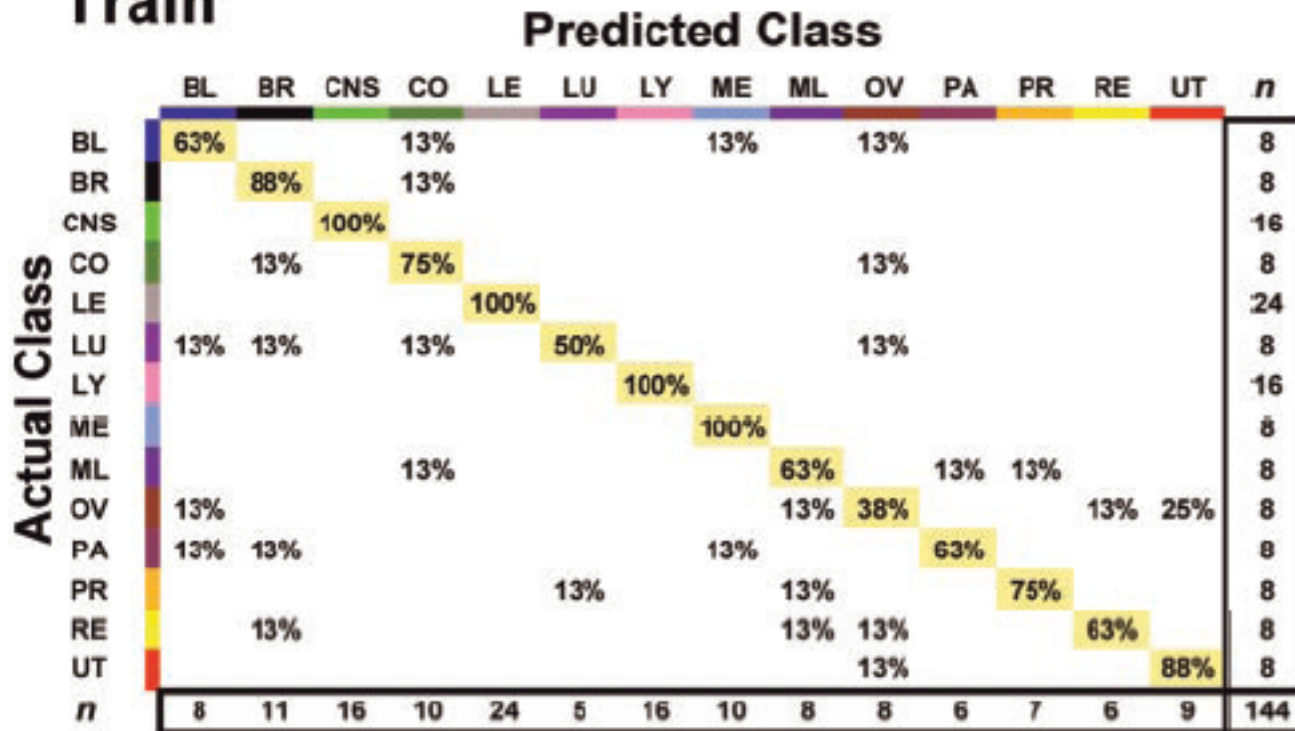
- Suppose we have 5000 genes and 50 samples, divided into two classes.
- Since we have many more variables than observations, there are infinitely many separating hyperplanes in 5000 dimensional feature space.
- SVMs provide the unique **maximal margin** separating hyperplane.
- Prediction performance can be good, but typically no better than simpler methods such as nearest centroid.
- All genes get a weight, so no gene selection.



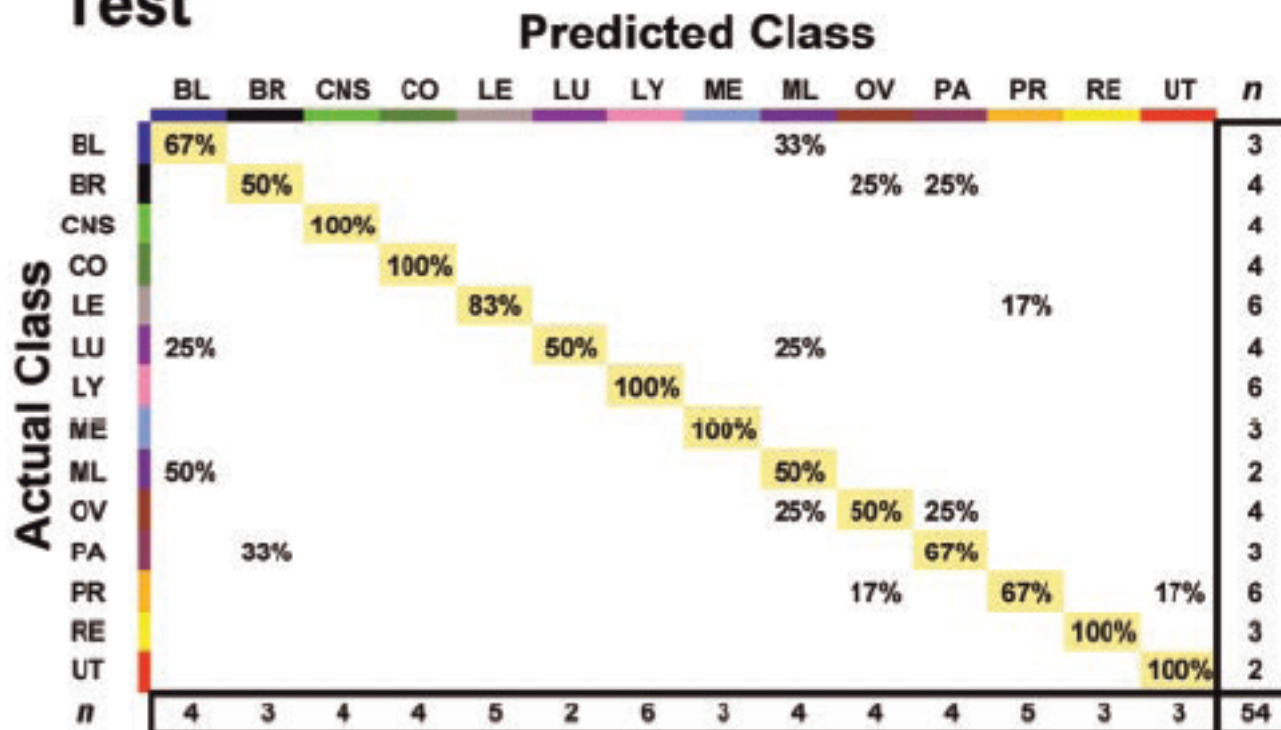
MULTICLASS PREDICTION



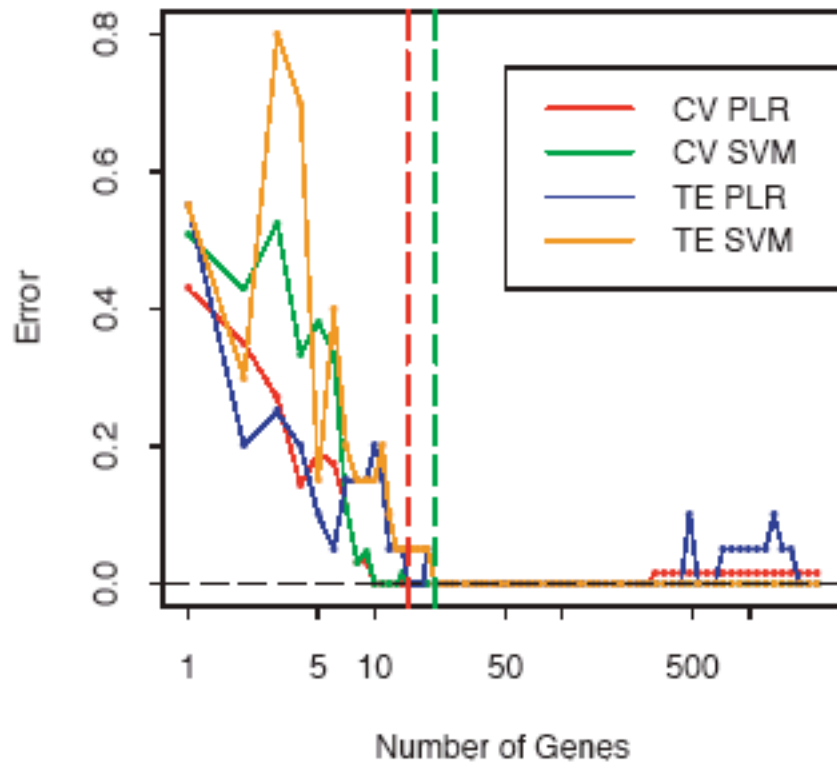
Train



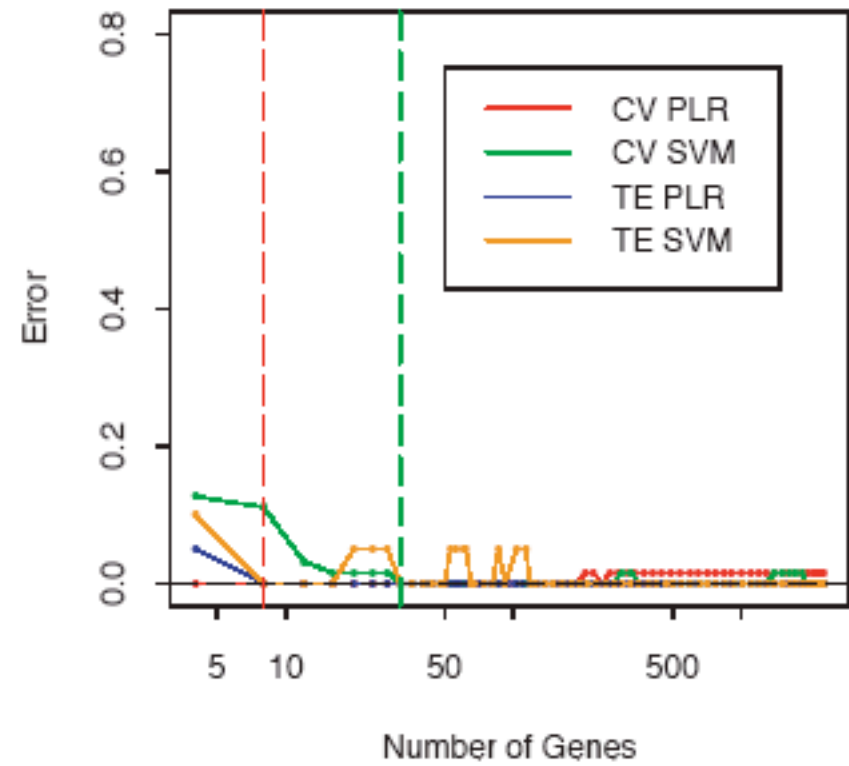
Test



Use UR



Use RFE

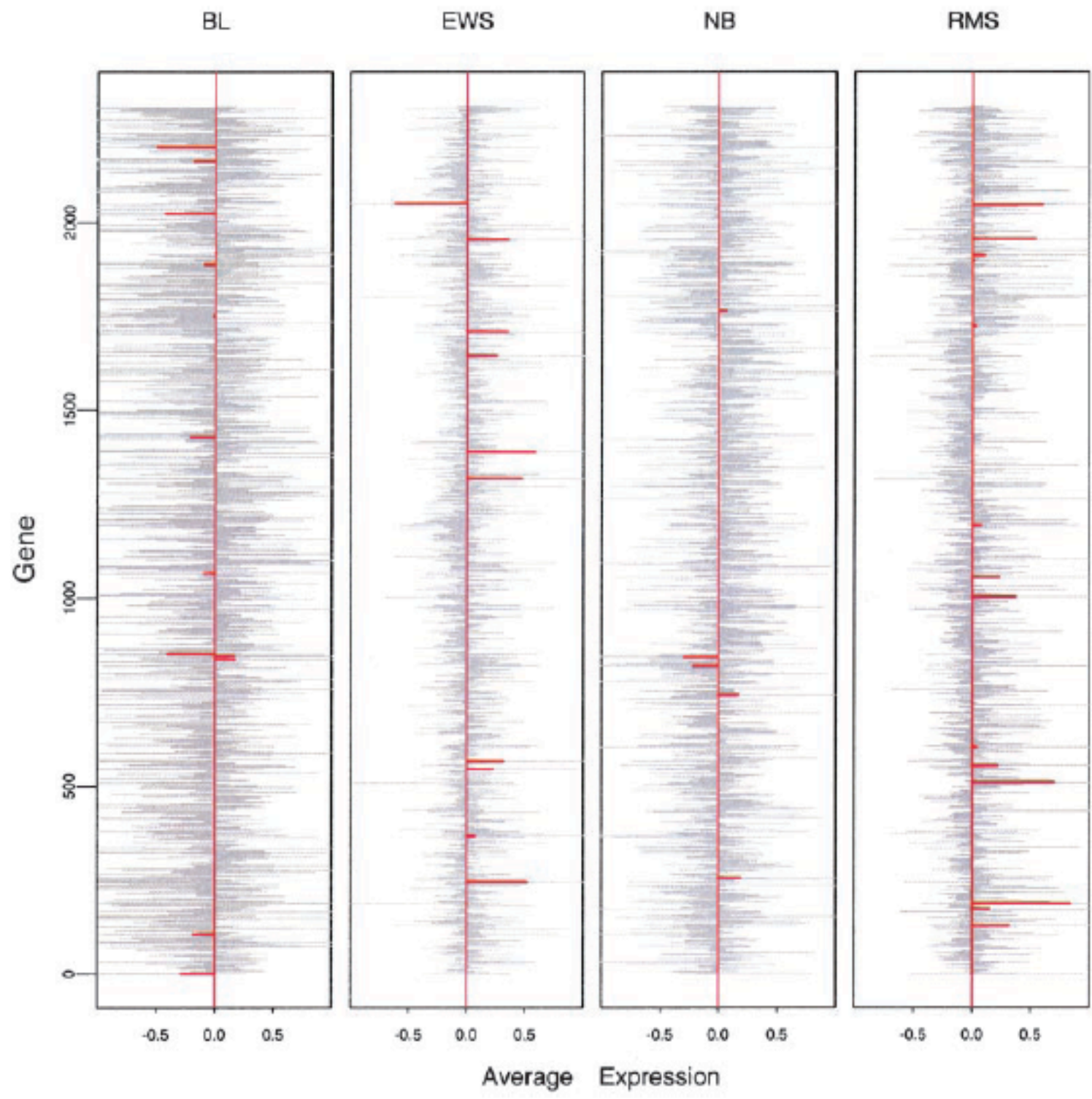


Method	10-fold CV error	Test error	No. of genes
SVM UR	0/63	0/20	21
PLR UR	0/63	0/20	15
SVM RFE	0/63	0/20	32
PLR RFE	0/63	0/20	8

Nearest Shrunken Centroids (PAM)

Example: small round blue cell tumors; Khan et al, Nature Medicine, 2001

- Tumors classified as **BL** (Burkitt lymphoma), **EWS** (Ewing), **NB** (neuroblastoma) and **RMS** (rhabdomyosarcoma).
- There are 63 training samples and 25 test samples, although five of the latter were not SRBCTs. 2308 genes
- Khan et al report zero training and test errors, using a complex neural network model. Decided that 96 genes were “important”.
- Too complicated!



- Idea: shrink each class centroid towards the overall centroid. First normalize by the within class-standard deviation for each gene.
- Let x_{ij} be the expression for genes $i = 1, 2, \dots, p$ and samples $j = 1, 2, \dots, n$.
- We have classes $1, 2, \dots, K$, and let C_k be indices of the n_k samples in class k .
- The i th component of the centroid for class k is $\bar{x}_{ik} = \sum_{j \in C_k} x_{ij} / n_k$, the mean expression value in class k for gene i ; the i th component of the overall centroid is $\bar{x}_i = \sum_{j=1}^n x_{ij} / n$.

- Let

$$d_{ik} = (\bar{x}_{ik} - \bar{x}_i) / s_i, \quad (1)$$

where s_i is the pooled within class standard deviation for gene i :

$$s_i^2 = \frac{1}{n - K} \sum_k \sum_{i \in C_k} (x_{ij} - \bar{x}_{ik})^2. \quad (2)$$

- Shrink each d_{ik} towards zero, giving d'_{ik} and new shrunken centroids or prototypes

$$\bar{x}'_{ik} = \bar{x}_i + s_i d'_{ik} \quad (3)$$

- The shrinkage is *soft-thresholding*: each d_{ik} is reduced by an amount Δ in absolute value, and is set to zero if its absolute value is less than zero. Algebraically, this is expressed as

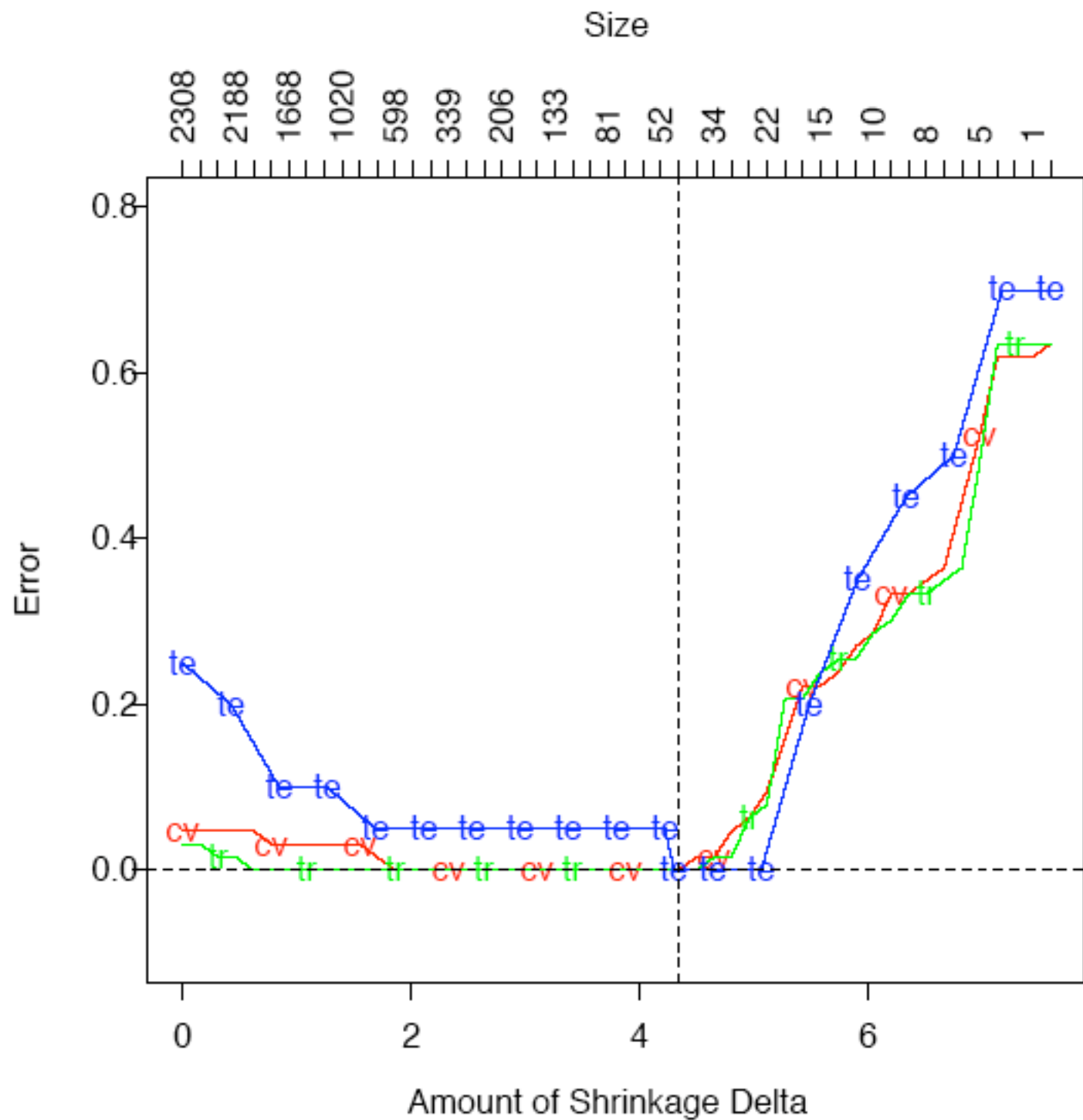
$$d'_{ik} = \text{sign}(d_{ik})(|d_{ik}| - \Delta)_+ \quad (4)$$

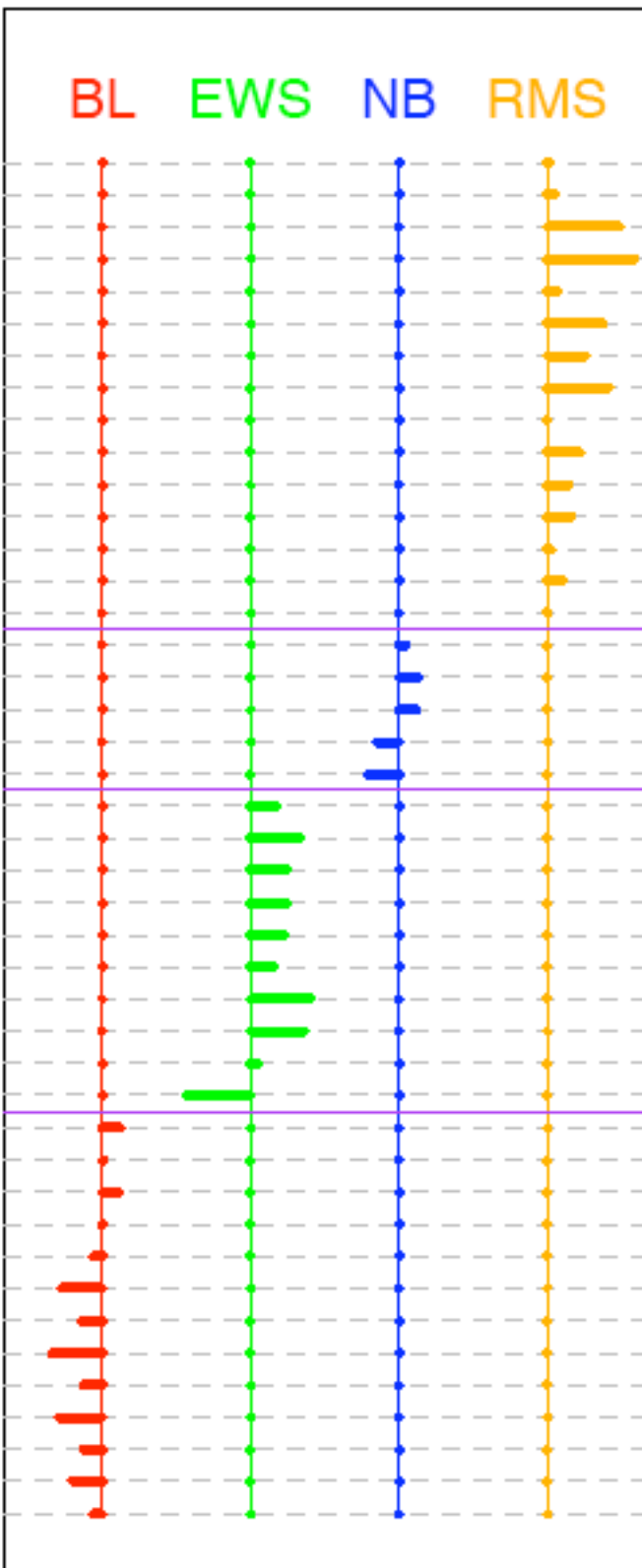
where $+$ means *positive part* ($t_+ = t$ if $t > 0$, and zero otherwise).

- Choose Δ by cross-validation.

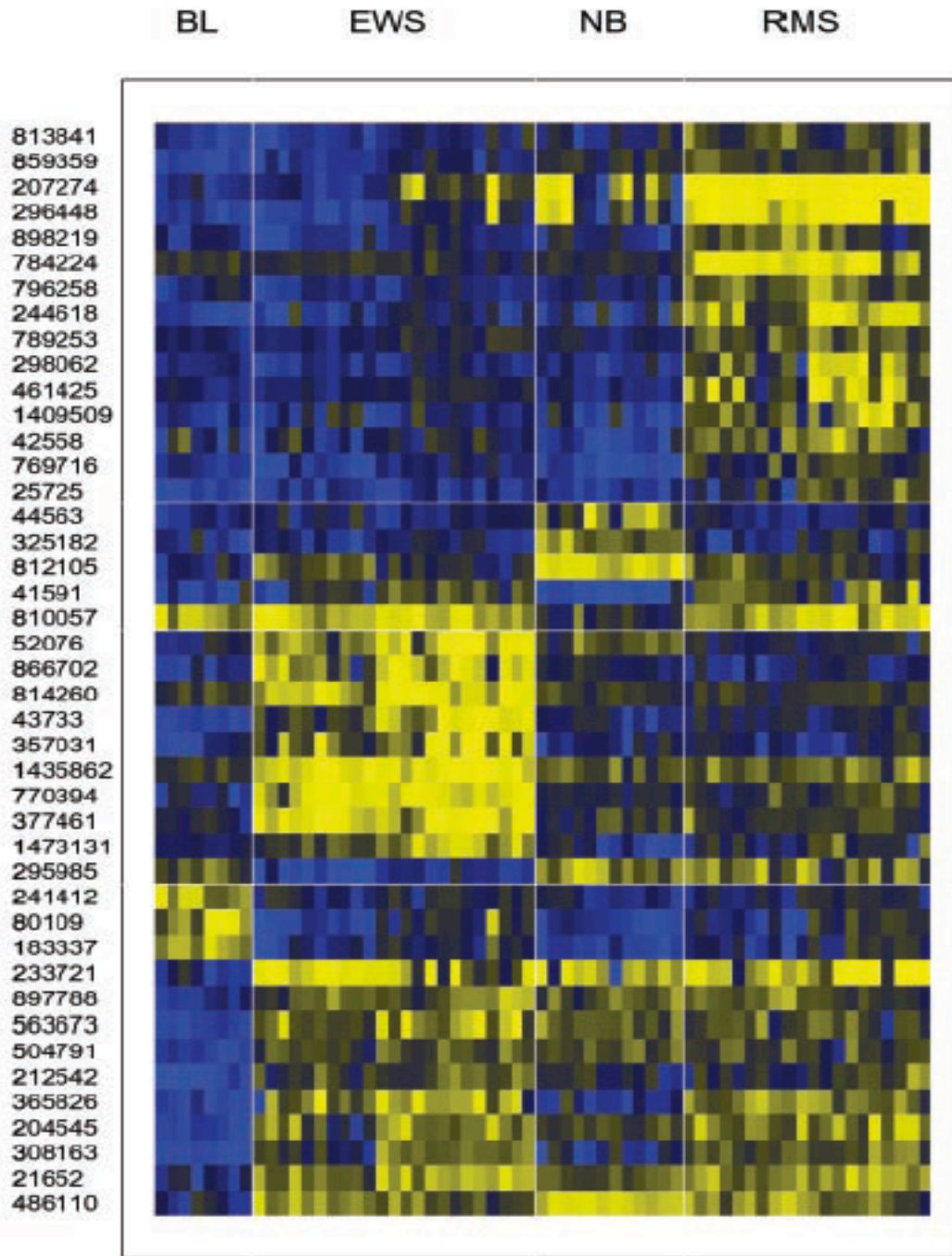
Advantages

- Simple, includes nearest centroid classifier as a special case.
- Thresholding denoises large effects, and sets small ones to zero—thereby selecting genes
- with more than two classes, method can select different genes, and different numbers of genes for each class.





- 813841 tissue plasminogen activator
- 859359 quinone oxidoreductase homolog
- 207274 insulin-like growth factor 2
- 296448 insulin-like growth factor 2 (somatomedin A)
- 898219 homolog of mouse mesoderm specific transcript
- 784224 fibroblast growth factor receptor 4
- 796258 sarcoglycan alpha (dystrophin-associated glycoprotein)
- 244618 EST
- 789253 presenilin 2 (Alzheimer disease 4)
- 298062 troponin T2, cardiac muscle isoforms
- 461425 myosin MYL4
- 1409509 troponin T1, slow skeletal muscle isoforms
- 42558 L-arginine:glycine amidinotransferase
- 769716 neurofibromin 2 (mutated in neurofibromatosis type 2)
- 25725 farnesyl-diphosphate farnesyltransferase 1
- 44563 growth associated protein 43 (GAP43)
- 325182 N-cadherin (neuronal)
- 812105 ALL1-fused gene from chromosome 1q
- 41591 meningioma 1 (disrupted in balanced translocation)
- 810057 cold shock domain protein A
- 52076 neuroblastoma protein (NOE1)
- 866702 Fas-associated protein tyrosine phosphatase 1
- 814260 follicular lymphoma variant translocation protein 1
- 43733 glycogenin 2
- 357031 tumor necrosis factor alpha-induced protein 6
- 1435862 MIC2 surface antigen (CD99)
- 770394 IgG Fc fragment receptor transporter, alpha chain
- 377461 caveolin 1 (caveolae protein)
- 1473131 transducin-like enhancer of split 2
- 295985 EST
- 241412 E74-like factor 1 (ets domain transcription factor)
- 80109 major histocompatibility complex, class II, DQ alpha 1
- 183337 major histocompatibility complex, class II, DM alpha
- 233721 insulin-like growth factor binding protein 2
- 897788 receptor type protein tyrosine phosphatase F
- 563673 antiquitin 1
- 504791 glutathione S-transferase A4
- 212542 cDNA DKFZp586J2118
- 365826 growth arrest-specific protein 1
- 204545 EST
- 308163 EST
- 21652 alpha 1 catenin (cadherin-associated protein)
- 486110 profilin 2

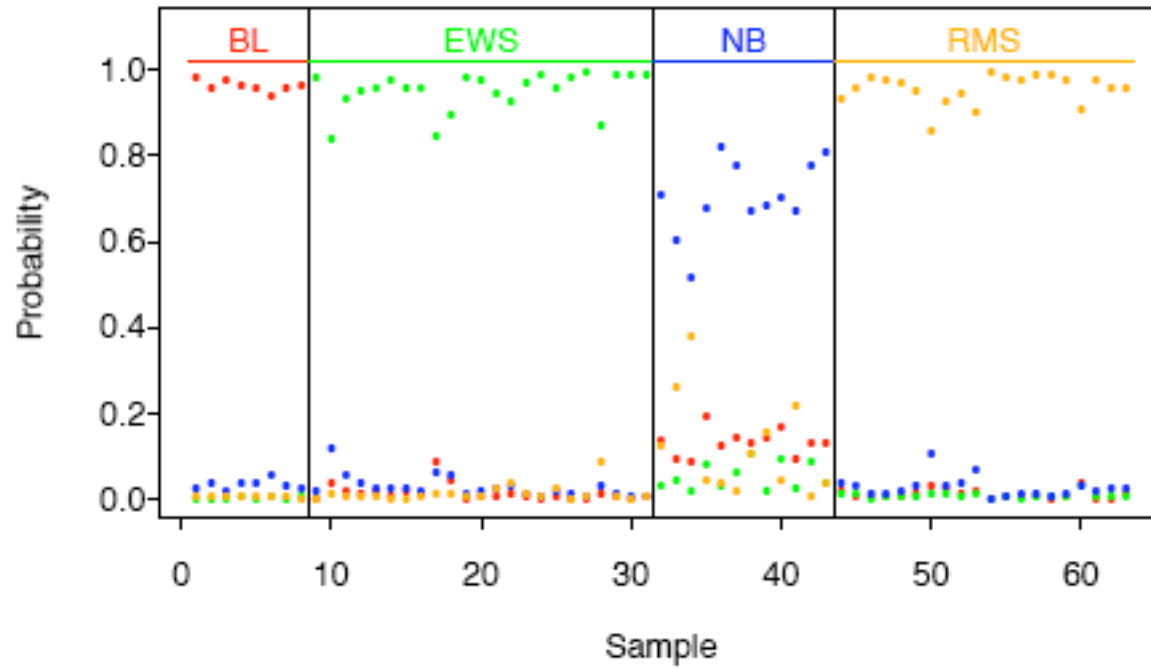


Heat map of the chosen 43 genes. Within each of the horizontal partitions, genes are ordered by hierarchical clustering, and similarly for the samples within each vertical partition.

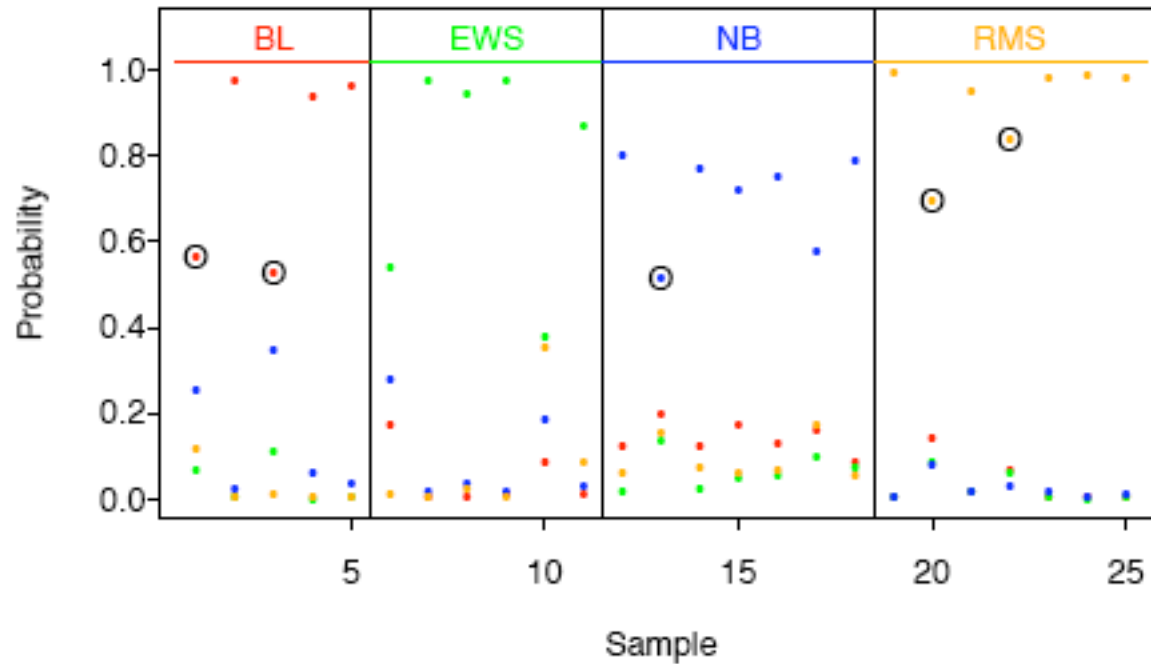


Heat map of three genes reported to characterize SRBCT tumors. They are c-myc (*Top*), CD45 (*Middle*) and myogenin (*Bottom*).

Training Data



Test Data



Leukemia classification

Golub et al 1999, Science. They use a “voting” procedure for each gene, where votes are based on a t-like statistic

Method	CV err	Test err
Golub	3/38	4/34
PAM	1/38	2/34

Breast Cancer classification

Hedenfalk et al 2001, NEJM. They use a “compound predictor” $\sum_j w_j x_j$, where the weights are t-statistics.

Method	BRCA1+	BRCA1-	BRCA2+	BRCA2-
Heden <i>et. al.</i>	3/7	2/15	3/8	1/14
PAM	2/7	1/15	2/8	0/14

Software



- `svm(...)` [in package `e1071`]
- `svmpath(...)`
- `pamr(...)` [also available as excel plug-in]