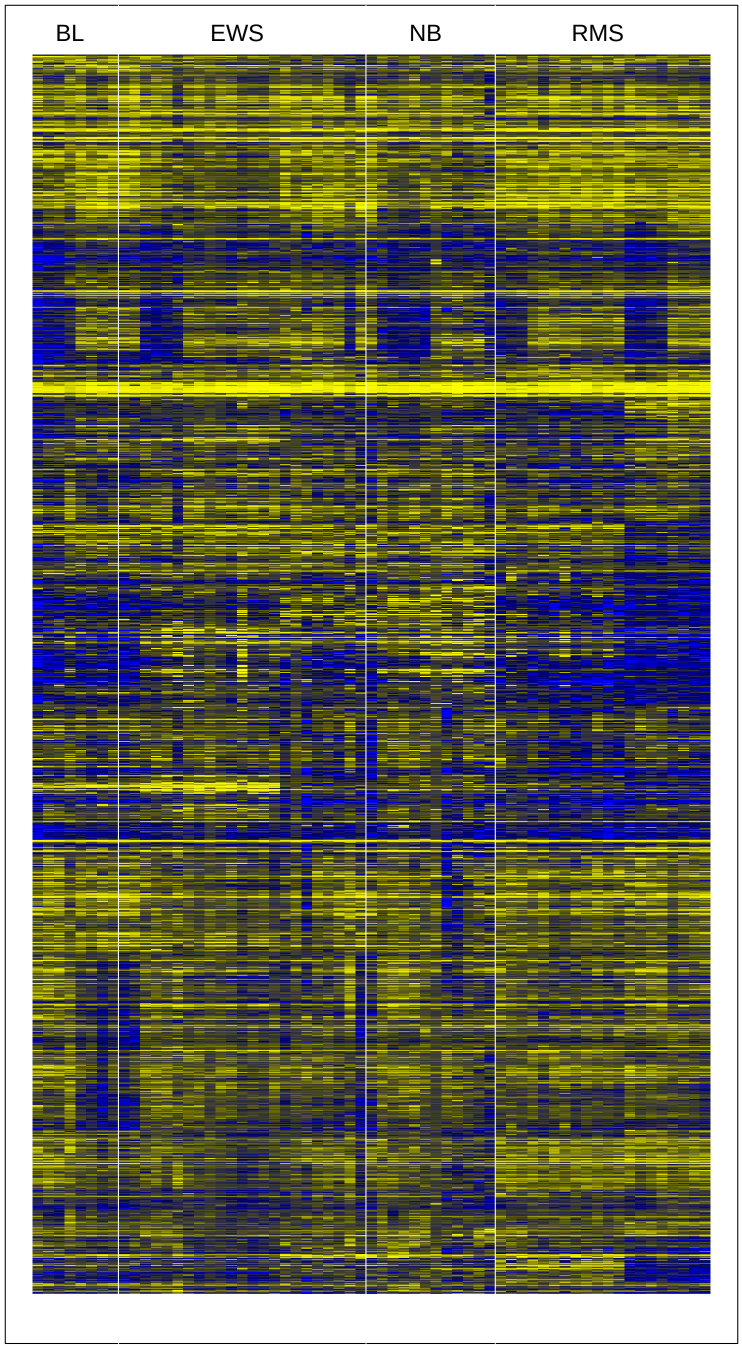
Classification of microarray samples

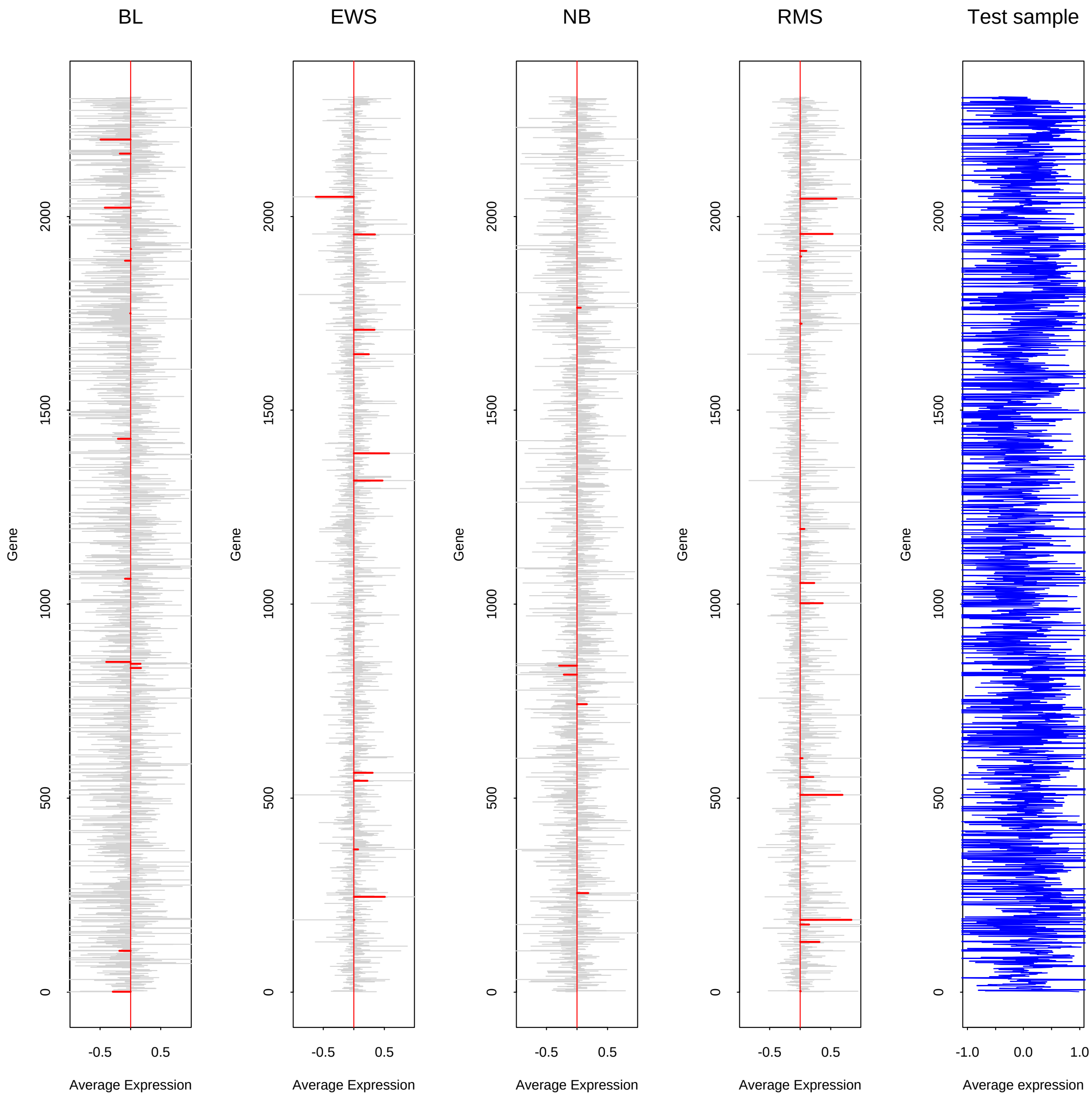
Example: small round blue cell tumors; Khan et al, Nature Medicine, 2001

- 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".
- Upon close examination, network is linear. It's essentially extracting linear principal components, and classifying in their subspace.
- But even principal components is unnecessarily complicated for this problem!

Khan data



Class centroids



Nearest Shrunken Centroids

Idea: shrink each class centroid towards the overall centroid. First normalize by the within-class standard deviation for each gene.

Details

- Let x_{ij} be the expression for genes $i = 1, 2, \ldots p$ and samples $j = 1, 2, \ldots n$.
- We have classes $1, 2, \ldots K$, and let C_k be indices of the n_k samples in class k.
- The ith 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 ith component of the overall centroid is $\bar{x}_i = \sum_{j=1}^{n} x_{ij}/n$.

- Let

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

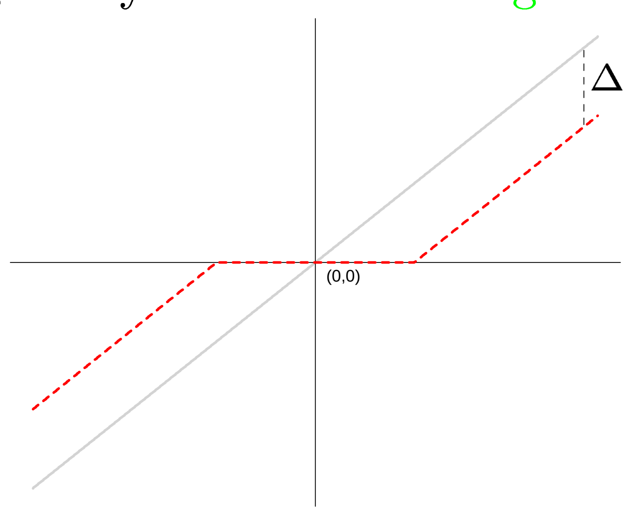
where s_i is the pooled within-class standard deviation for gene i:

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

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

$$\bar{x}'_{ik} = \bar{x}_i + s_i d'_{ik}$$

- The shrinkage is by soft-thresholding:



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

- Choose Δ by cross-validation.

K-Fold Cross-Validation

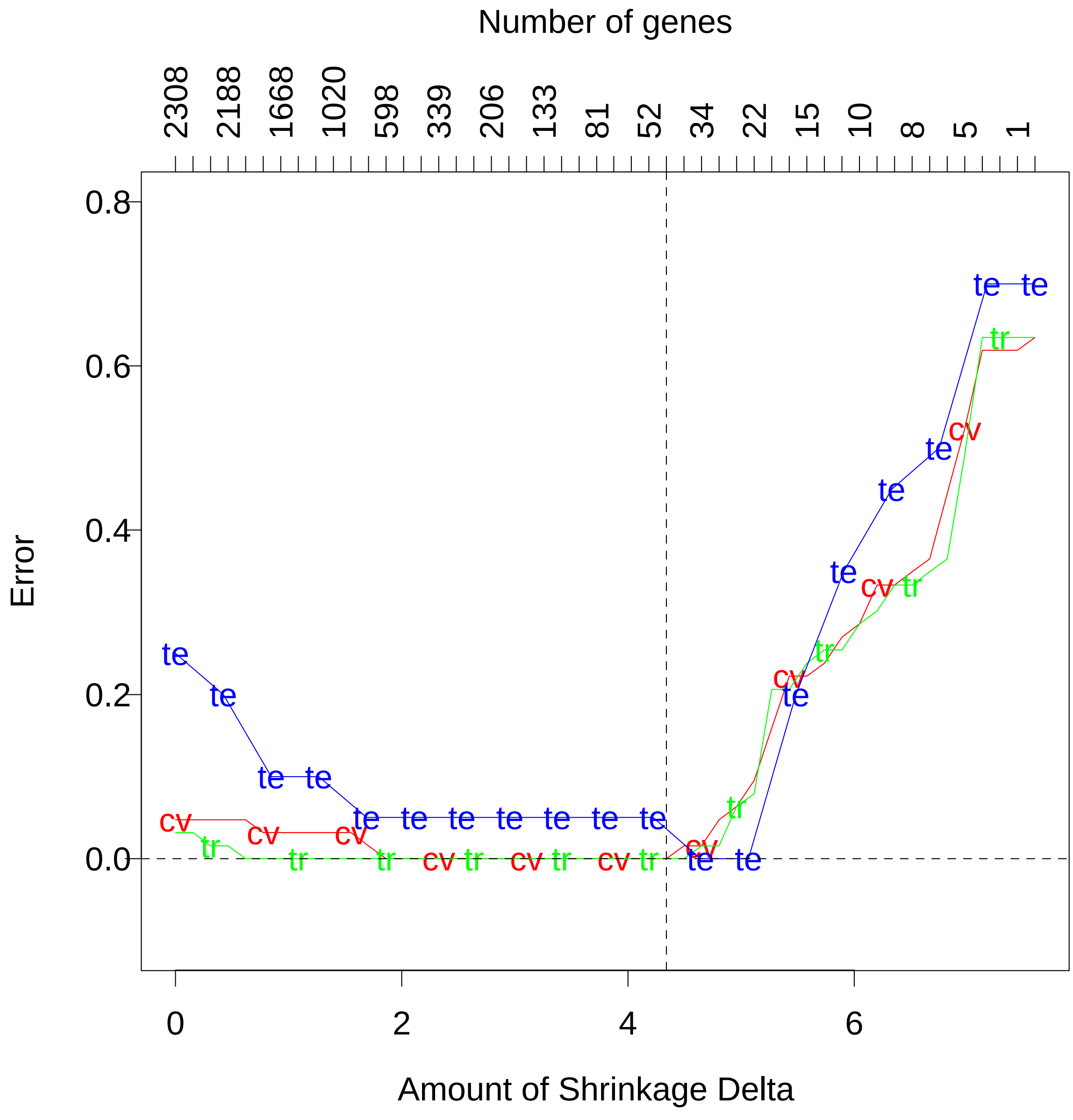
Primary method for estimating a tuning parameter λ.
Divide the data into K roughly equal parts.

1	2	3	4	5
Test	Train	Train	Train	Train

- for each $k = 1, 2, \ldots K$, fit the model with parameter λ to the other $K - 1$ parts, and compute its error in predicting the kth part. Average this error over the K parts to give the estimate $CV(\lambda)$.
- do this for many values of λ. Draw the curve $CV(\lambda)$ and choose the value of λ that makes $CV(\lambda)$ smallest.

Typically we use $K = 5$ or 10.

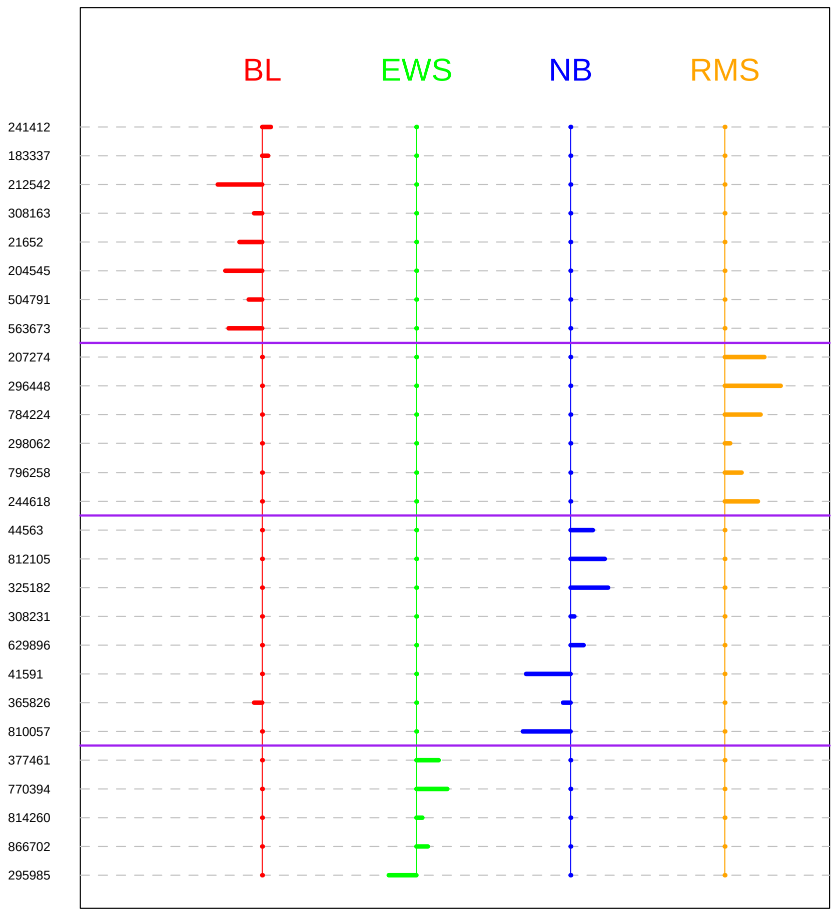
Results


Number of genes
2308
2188
1668
1020
598
339
206
133
81
52
34
22
15
10
8
5
1
Error
0.8
0.6
0.4
0.2
0.0
te
cv
tr
0
2
4
6
Amount of Shrinkage Delta

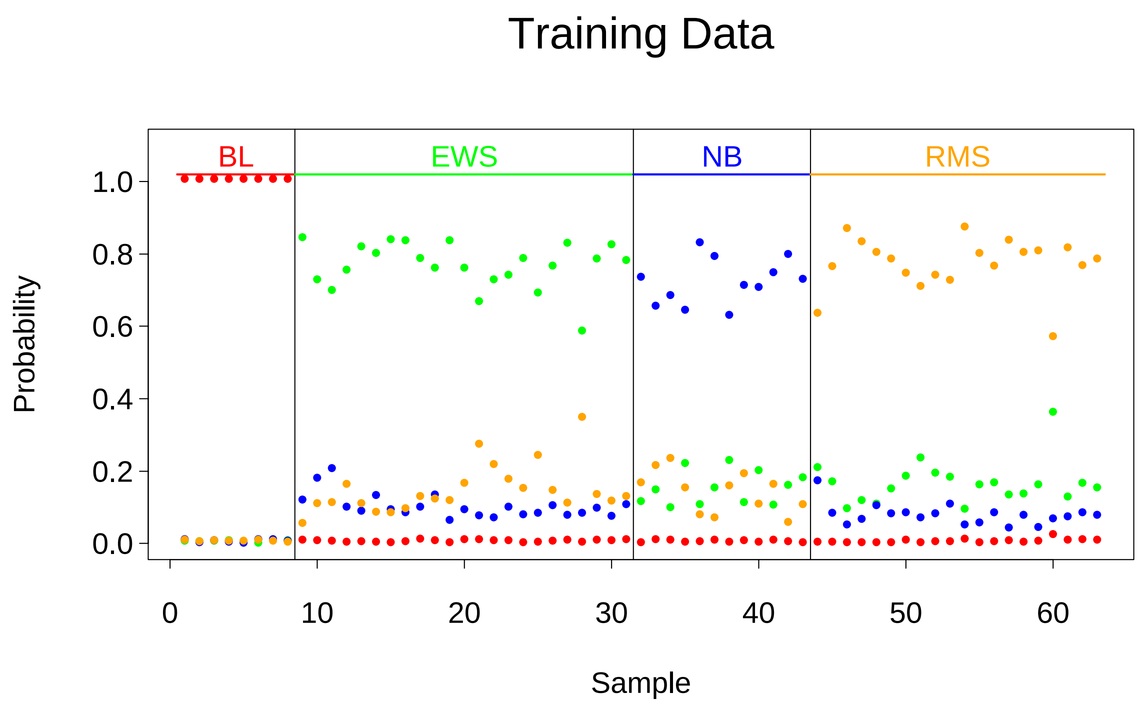
Advantages

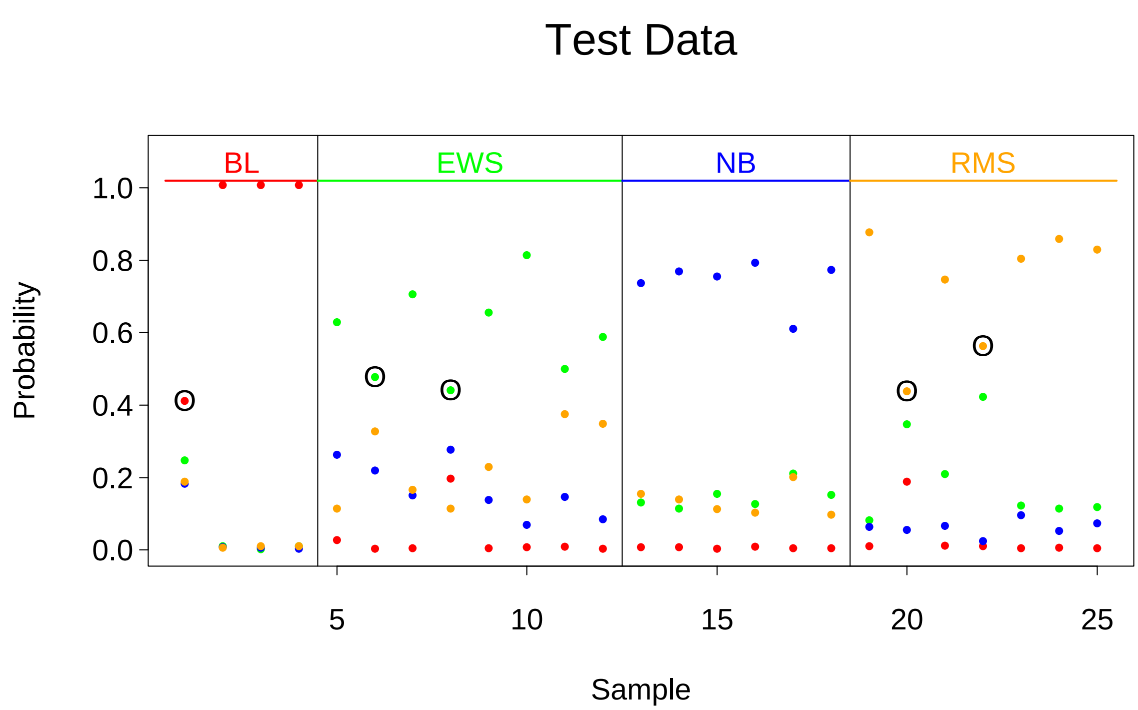
- Simple, includes nearest centroid classifier as a special case.
- Thresholding denoises large effects, and sets small ones to zero, thereby selecting genes.
- with more than two classes, method can select different genes, and different numbers of genes for each class.

The genes that matter



Estimated Class Probabilities





Class probabilities

- For a test sample $x^* = (x_1^*, x_2^*, \ldots x_p^*)$. We define the discriminant score for class k

$$\delta_k(x^*) = \sum_{i=1}^{p} \frac{(x_i^* - \bar{x}'_{ik})^2}{s_i^2} - 2\log \pi_k$$

- The classification rule is then

$$C(x^*) = \ell \text{ if } \delta_\ell(x^*) = \min_k \delta_k(x^*)$$

- estimates of the class probabilities, by analogy to Gaussian linear discriminant analysis, are

$$\hat{p}_k(x^*) = \frac{e^{-\frac{1}{2}\delta_k(x^*)}}{\sum_{\ell=1}^{K} e^{-\frac{1}{2}\delta_\ell(x^*)}}$$

- Still very simple. In statistical parlance, this is a restricted version of a naive Bayes classifier (also called idiot's Bayes!)

Adaptive threshold scaling

- idea: define class-dependent scaling factors θ_k for each class:

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_k \theta_k \cdot s_i}. \tag{1}$$

- Use smaller factors for hard-to-classify classes => same test error with fewer total number of genes
- Adaptive procedure: start with all $\theta_k = 1$, and then reduce θ_k by 10% for the class k with largest area under training error curve.
- repeat 20 times and choose solution with smallest area under curve for all classes
- can dramatically reduce total number of genes used, without increasing error rate

Lymphoma data

Scaling factors changed from $(1, 1, 1)$ to $(1.9, 1, 1.5)$