Linear classifiers

- Linear regression
- linear and quadatric discriminant functions
- example: gene expression arrays
- reduced rank LDA
- logistic regression
- separating hyperplanes

Linear classifiers

Some concepts:

- linear regression $\hat{f}_k(x) = \hat{\beta}_{k0} + \hat{\beta}_k x$
- decision boundary between classes k and ℓ:

$$\{x : \hat{f}_k(x) = \hat{f}_\ell(x)\}$$

- linear discriminant analysis, logistic regression

$$\log \frac{P(G = 1|X = x)}{P(G = 2|X = x)} = \beta_0 + \beta^T x$$

- explicit approaches: separating hyperplanes
- discriminant functions:

$$\begin{aligned} \delta_k(x), \quad & k = 1, 2, \ldots K \\ \hat{G}(x) \quad & = \text{argmin } \delta_k(x) \end{aligned}$$

Linear regression

Indicator response matrix

$$g = \begin{pmatrix} 3 \\ 1 \\ 4 \\ \vdots \\ 2 \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ \vdots & & & \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$\hat{F} = \hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{X}\hat{\beta}$$

$$\hat{f}(x) = \mathbf{X}\hat{\beta} = \begin{pmatrix} \hat{f}_1(x) \\ \hat{f}_2(x) \\ \vdots \\ \hat{f}_K(x) \end{pmatrix}$$

Note: $E(Y|X = x) = p(x) = \begin{pmatrix} p_1(x) \\ p_2(x) \\ \vdots \\ p_K(x) \end{pmatrix}$

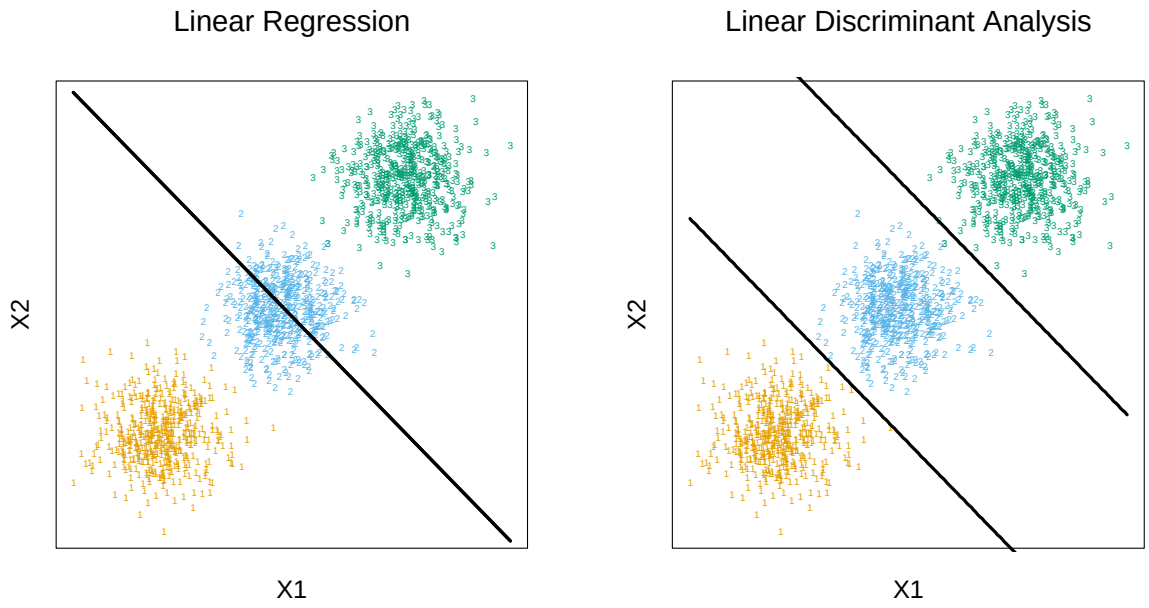
$p_k(x) = P(G = k|X = x)$

Targets: $\min_\beta \sum_{i=1}^N ||y_i - \beta^T x_i||^2$, y_i, x_i are ith rows of $\mathbf{Y}$ and $\mathbf{X}$.

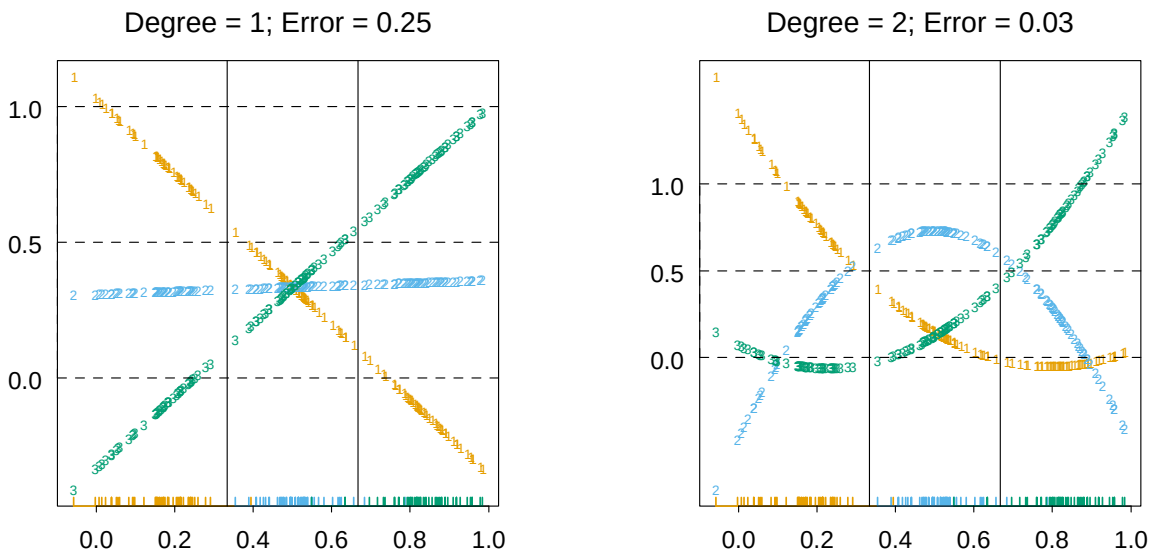
with $\hat{f}(x) = \hat{\beta}^T x$, $\hat{G}(x) = \text{argmin}_k ||\hat{f}(x) - t_k||^2$,

$t_k = (0, 0, \ldots 0, 1, 0, \ldots)$ (1 in kth position).

Masking problems with linear regression



The data come from three classes in R^2 and are easily separated by linear decision boundaries. The right plot shows the boundaries found by linear discriminant analysis. The left plot shows the boundaries found by linear regression of the indicator response variables. The middle class is completely masked (never dominates).



The effects of masking on linear regression in R for a three-class problem. The rug plot *at the base indicates the positions and class membership of each observation. The three curves in each panel are the fitted regressions to the three-class indicator variables; for example, for the red class,* y_{red} *is* 1 *for the red observations, and* 0 *for the green and blue. The fits are linear and quadratic polynomials. Above each plot is the training error rate. The Bayes error rate is* 0.025 *for this problem, as is the LDA error rate.*

Linear discriminant analysis

- $f_k(x)$= density of X in class $G = k$
- π_k class prior $Pr(G = k)$.
- Bayes theorem

$$Pr(G = k|X = x) = \frac{f_k(X)\pi_k}{\sum_{\ell=1}^{K} f_\ell(x)\pi_\ell}$$

- leads to LDA, QDA, MDA (mixture DA), Kernel DA, Naive Bayes
- LDA:

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-(1/2)(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\right)$$

- $\log \frac{Pr(G=k|x)}{Pr(G=\ell|x)} =$
  $\log \frac{\pi_k}{\pi_\ell} - (1/2)(\mu_k + \mu_\ell)^T\Sigma^{-1}(\mu_k - \mu_\ell) + x^T\Sigma^{-1}(\mu_k - \mu_\ell)$

More on LDA

- estimate μ_k by centroid in class k, and $\mathbf{\Sigma}$ by pooled within class covariance matrix
- estimated Bayes rule: classify to class k that maximizes the discriminant function

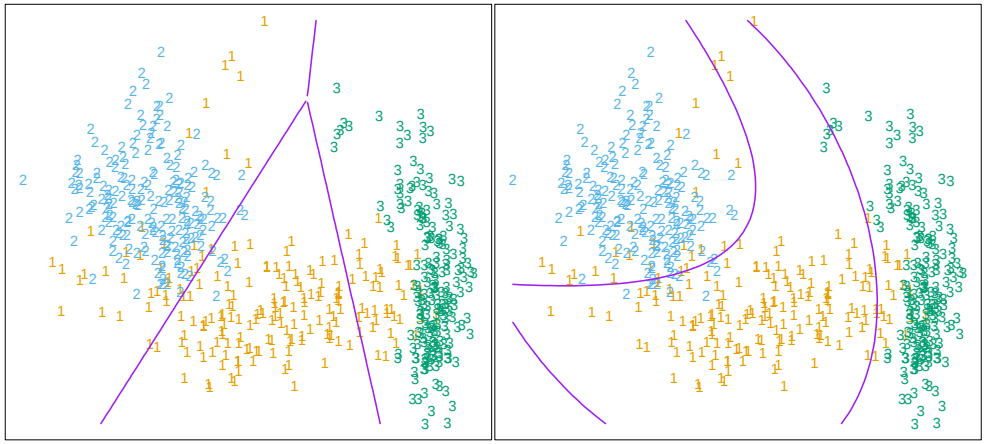
$$\delta_k(x) = x^T \hat{\mathbf{\Sigma}}^{-1} \mu_k - \frac{1}{2} \mu_k^T \hat{\mathbf{\Sigma}}^{-1} \mu_k + \log \pi_k \qquad (1)$$

- for two classes, we classify to class 2 if

$$x^T \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2} \hat{\mu}_2^T \hat{\mathbf{\Sigma}} \hat{\mu}_2 - \frac{1}{2} \hat{\mu}_1^T \hat{\mathbf{\Sigma}} \hat{\mu}_1 + \log(N_1/N) - \log(N_2/N)$$

where N_1, N_2 are number of observations in each class.

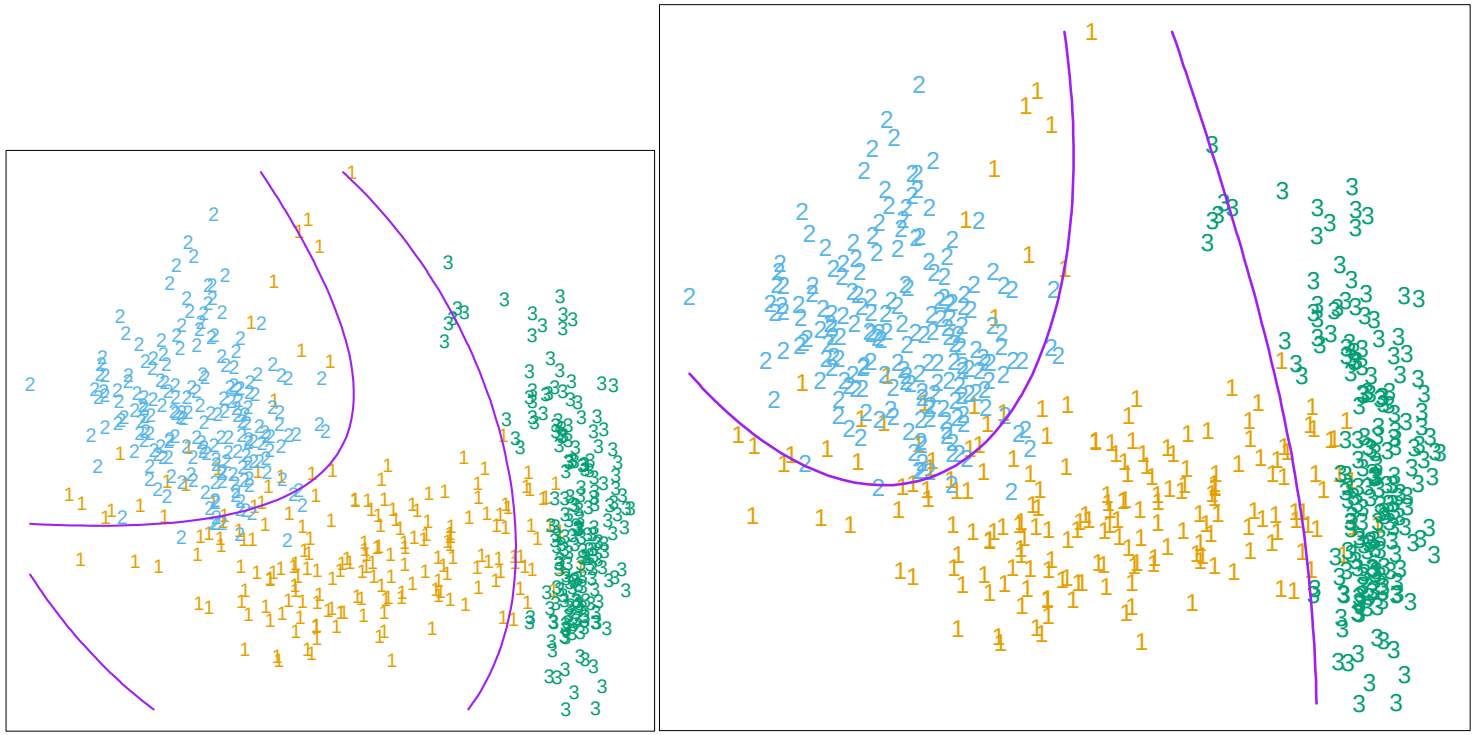
Linear boundaries and their projections



The left plot shows some data from three classes, with linear decision boundaries found by linear discriminant analysis. The right plot shows quadratic decision boundaries. These were obtained by finding linear boundaries in the five-dimensional space $X_1, X_2, X_1X_2, X_1^2, X_2^2$. *Linear inequalities in this space are quadratic inequalities in the original space.*

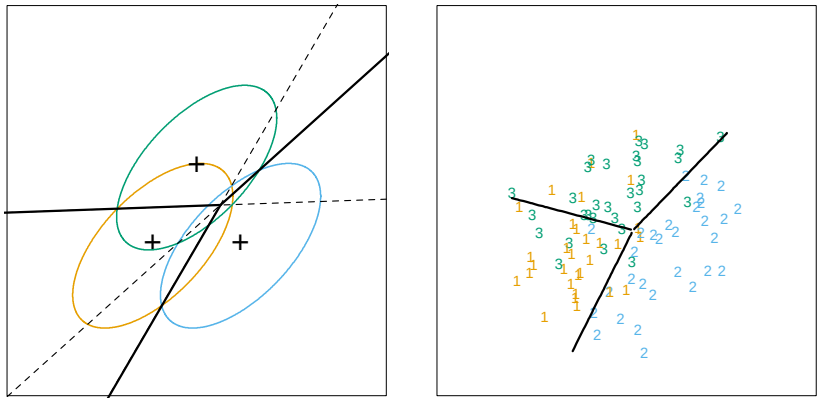
Quadratic discriminant analysis

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log\pi_k$$



Two methods for fitting quadratic boundaries. The left plot shows the quadratic decision boundaries for the data in the previous figure (obtained

using LDA in the five-dimensional space $X_1, X_2, X_1X_2, X_1^2, X_2^2$). The right plot shows the quadratic decision boundaries found by QDA. The differences are small, as is usually the case.



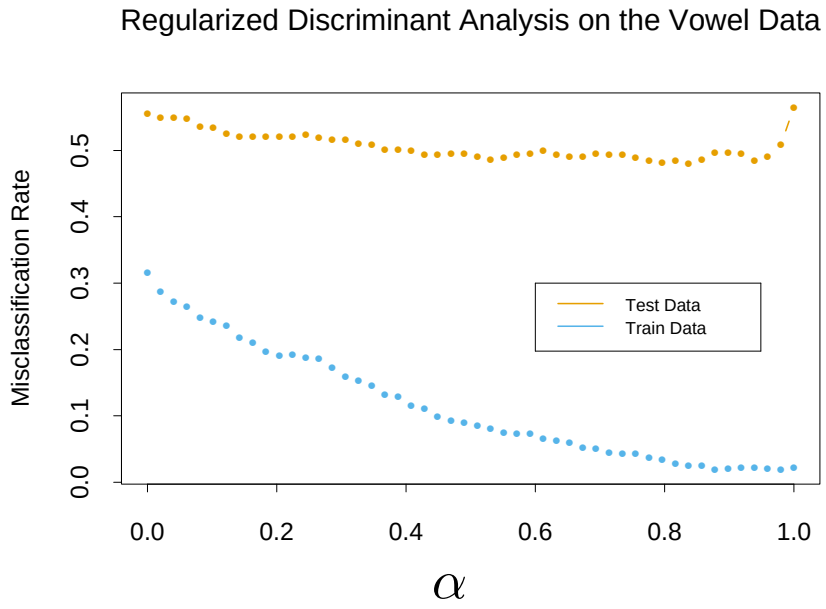
The left panel shows three Gaussian distributions, with the same covariance and different means. Included are the contours of constant density enclosing 95% of the probability in each case. The Bayes decision boundaries between each pair of classes are shown (broken straight lines), and the Bayes decision boundaries separating all three classes are the thicker solid lines (a subset of the former). On the right we see a sample of 30 *drawn from each Gaussian distribution, and the fitted LDA decision boundaries.*

Regularized discriminant analysis

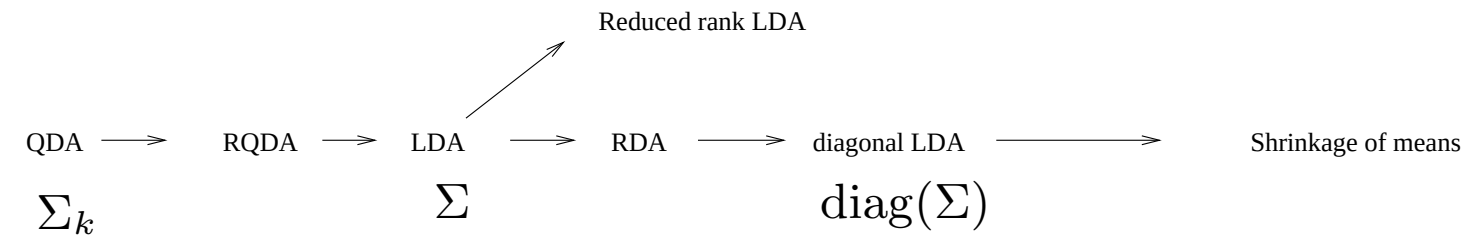
- Regularized QDA $\hat{\Sigma} = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma}$
- Regularized LDA $\hat{\Sigma}(\gamma) = \gamma\hat{\Sigma} + (1-\gamma)\hat{\sigma}^2 I$
- Together $\rightarrow \hat{\Sigma}(\alpha, \gamma)$
- could use $\hat{\Sigma}(\gamma) = \gamma\hat{\Sigma} + (1-\gamma)\text{diag}(\hat{\Sigma})$
- in recent microarray work we use

$$\delta_k(x) = \sum_{j=1}^{p} \frac{(x_j - \hat{\mu}'_{jk})^2}{s_j^2} - (1/2)\log \pi_k$$

where μ'_{jk} is a shrunken centroid. Details later



Test and training errors for the vowel data, using regularized discriminant analysis with a series of values of $\alpha \in [0, 1]$. *The optimum for the test data occurs around* $\alpha = 0.9$, *close to quadratic discriminant analysis.*


Reduced rank LDA
QDA
RQDA
LDA
RDA
diagonal LDA
Shrinkage of means
Σ_k
Σ
$\mathrm{diag}(\Sigma)$
Low Bias
High Variance
High Bias
Low Variance

Classification in high dimensions

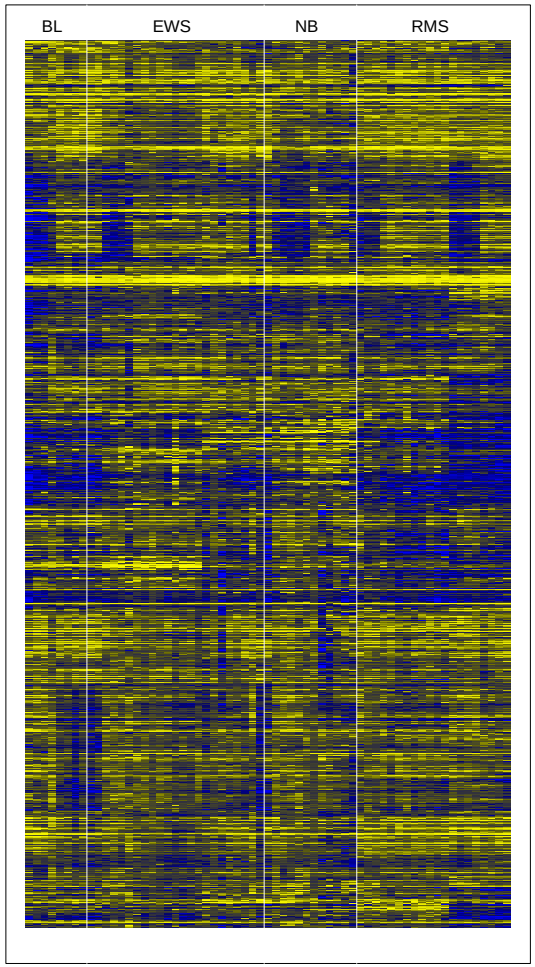
- important for gene expression microarray problems and other genomics problems
- Starting point: diagonal LDA which uses $\text{diag}(\hat{\Sigma})$
- nearest centroid classification, on standardized features, is equivalent to diagonal LDA
- nearest shrunken centroids regularizes further, by discarding noisy features

Classification of microarray samples

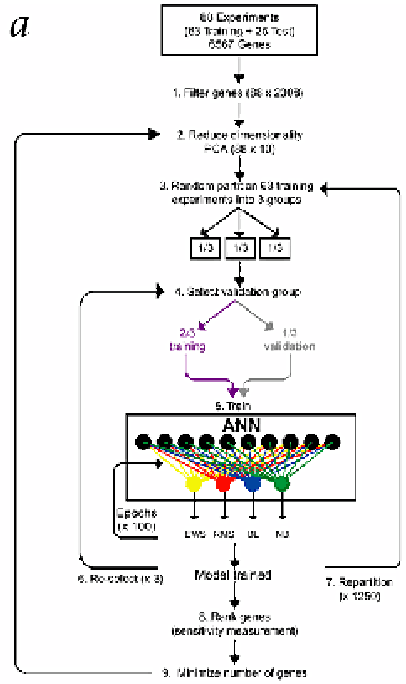
Example: small round blue cell tumors; Khan et al, Nature Medicine, 2001

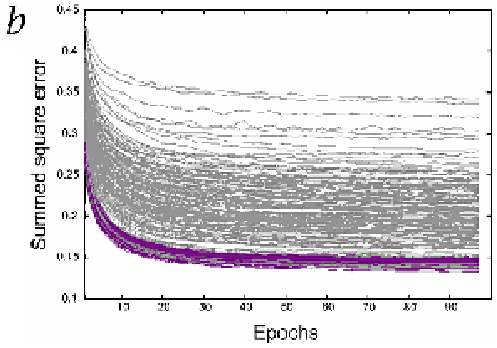
- Tumors classified as BL (Burkitt lymphoma), EWS (Ewing), NB (neuroblastoma) and RMS (rhabdomyosarcoma).
- There are 63 training samples and 25 test samples, although five of the latter were not SRBCTs. 2308 genes
- Khan et al report zero training and test errors, using a complex neural network model.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

Khan data

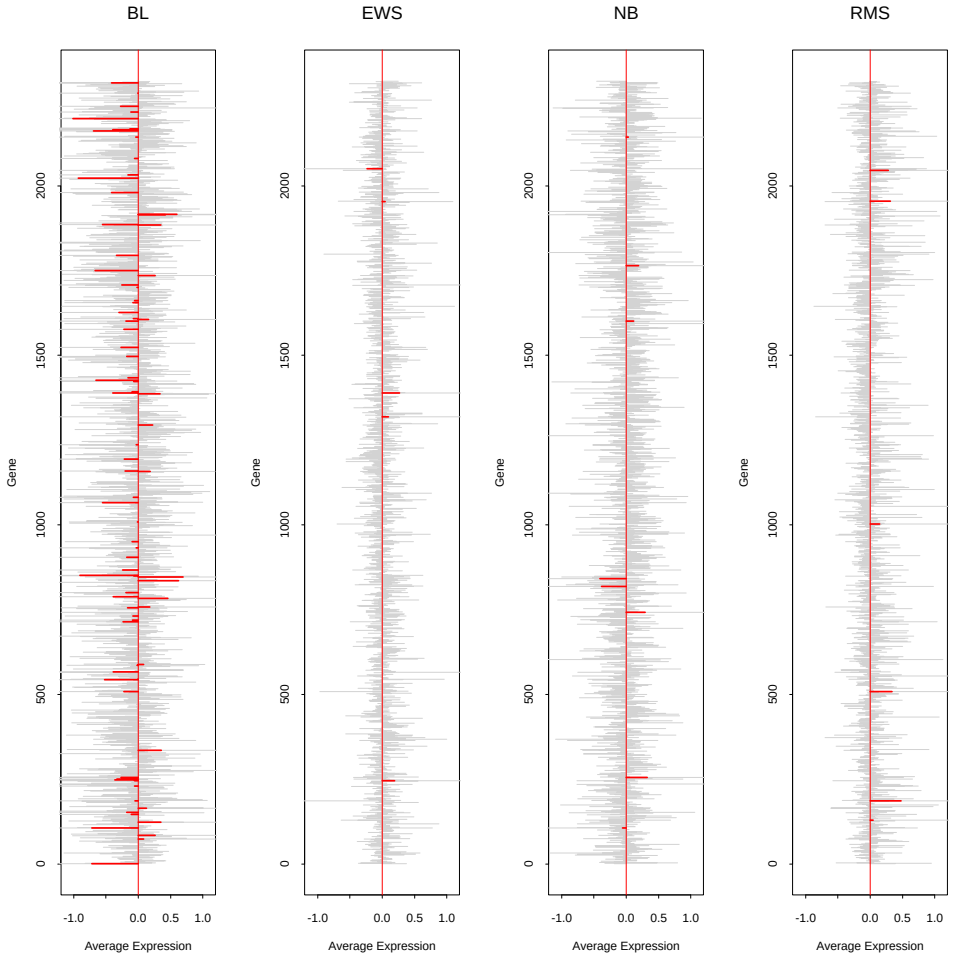


Neural network approach





Class centroids



Shrunken centroids

- 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 $i = 1, 2, \ldots p$ 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

$$d_{ik} = (\bar{x}_{ik} - \bar{x}_i)/s_i, \tag{2}$$

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

$$s_i^2 = \frac{1}{n-K} \sum_k \sum_{i \in C_k} (x_{ij} - \bar{x}_{ik})^2. \tag{3}$$

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

$$\bar{x}'_{ik} = \bar{x}_i + s_i d'_{ik} \tag{4}$$

- 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

$$d'_{ik} = \text{sign}(d_{ik})(|d_{ik}| - \Delta)_+ \tag{5}$$

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.

Class probabilities

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

$$\delta_k(x^*) = \sum_{i=1}^{p} \frac{(x_i^* - \bar{x}'_{ik})^2}{s_i^2} - 2\log \pi_k \tag{6}$$

- The classification rule is then

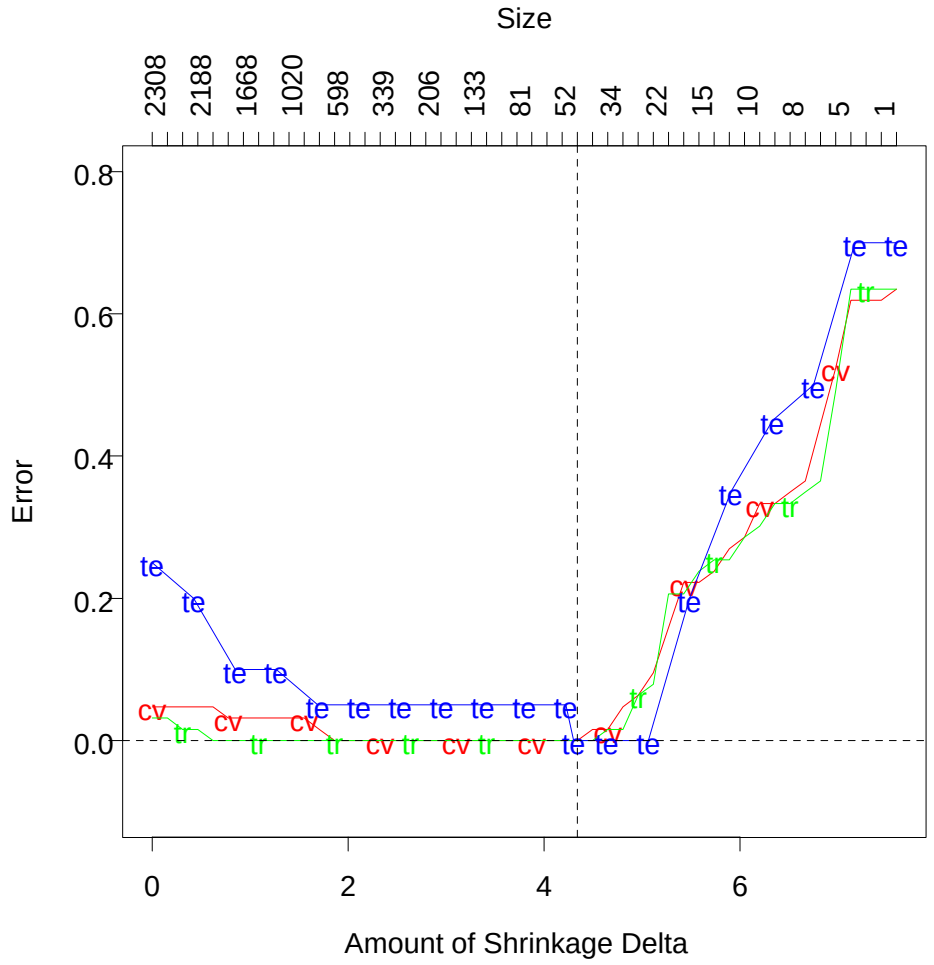
$$C(x^*) = \ell \text{ if } \delta_\ell(x^*) = \min_k \delta_k(x^*) \tag{7}$$

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

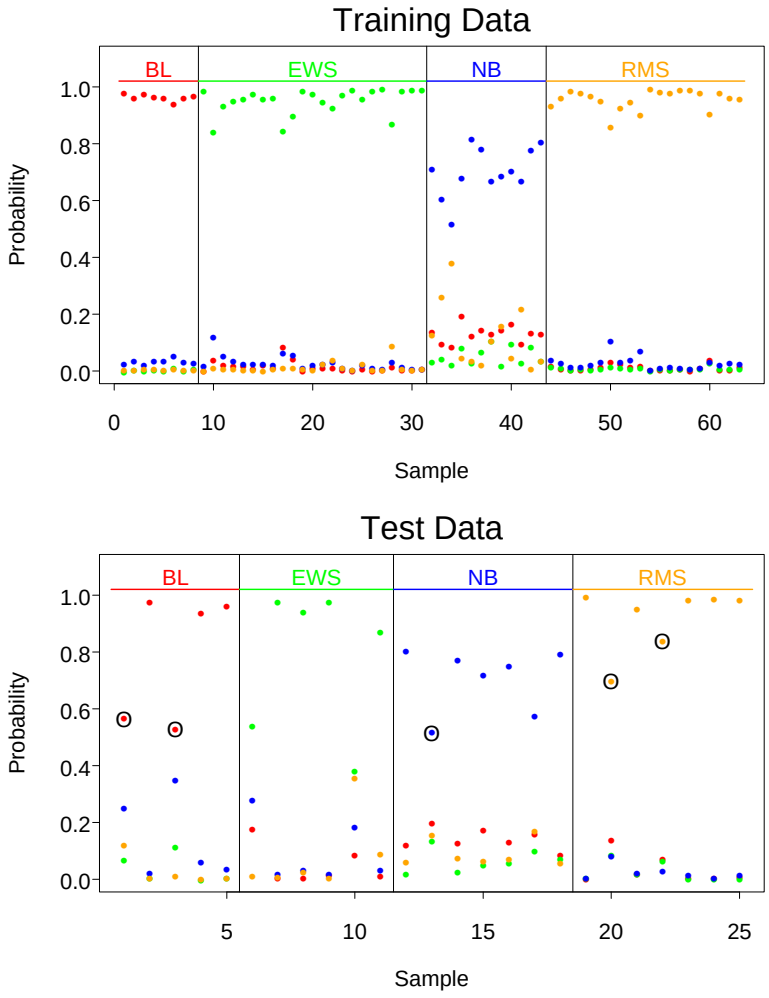
$$\hat{p}_k(x^*) = \frac{e^{-\frac{1}{2}\delta_k(x^*)}}{\sum_{\ell=1}^{K} e^{-\frac{1}{2}\delta_\ell(x^*)}} \tag{8}$$

Results on Khan data

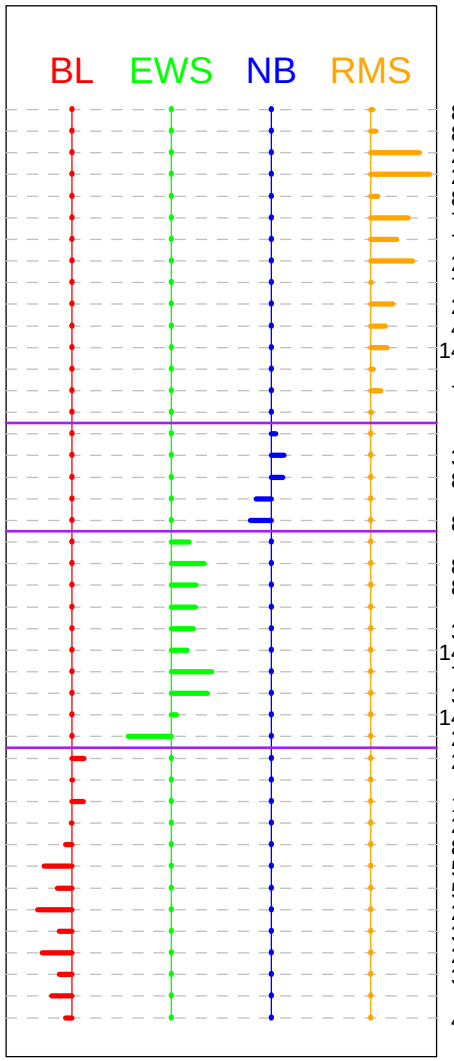
At optimal point, there are 43 active genes



Predictions

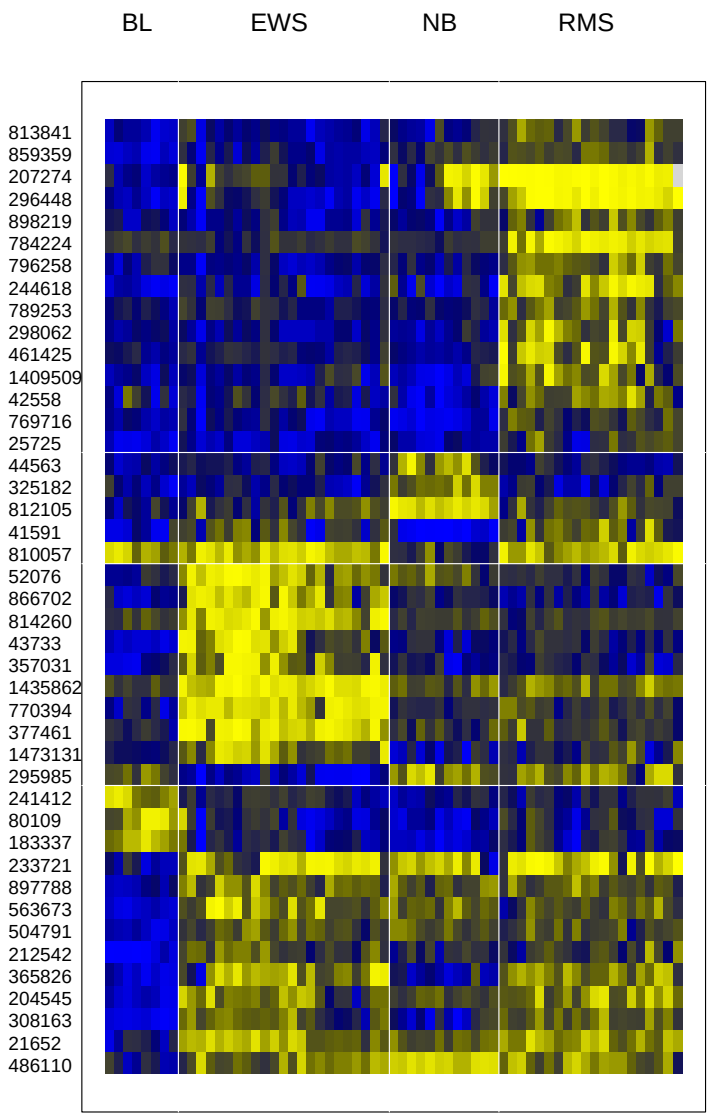


The genes that matter



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

Heatmap of selected genes

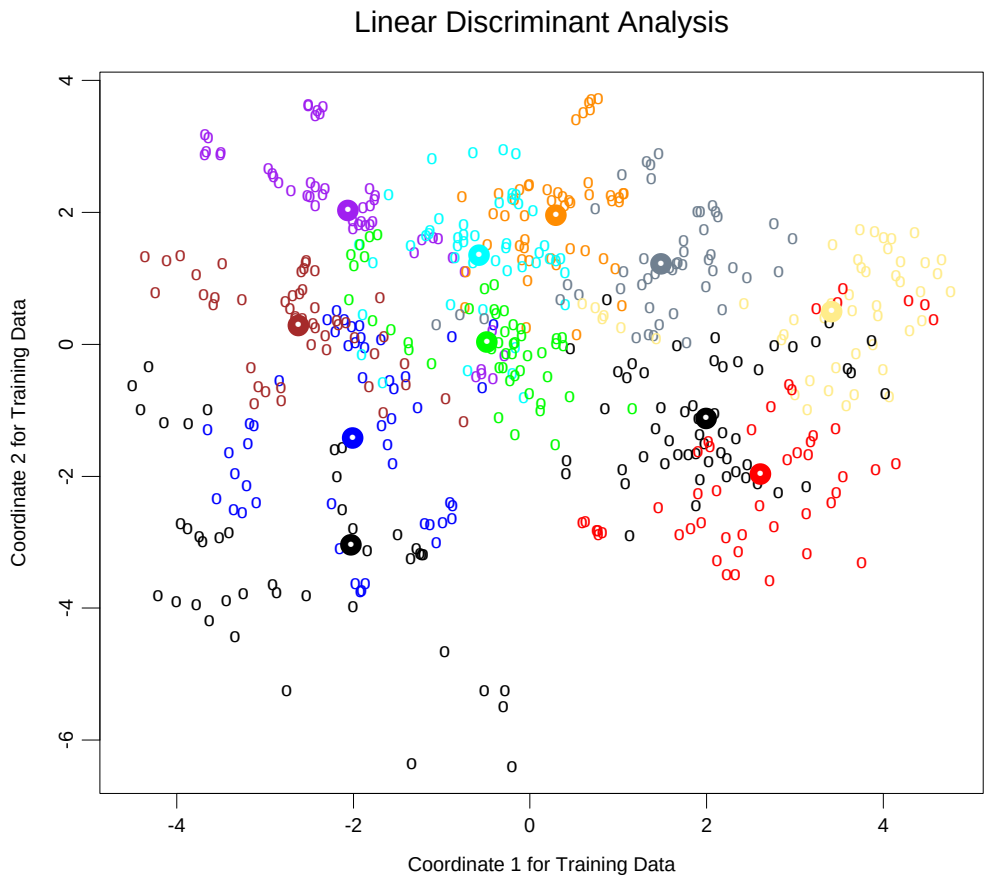

BL
EWS
NB
RMS
813841
859359
207274
296448
898219
784224
796258
244618
789253
298062
461425
1409509
42558
769716
25725
44563
325182
812105
41591
810057
52076
866702
814260
43733
357031
1435862
770394
377461
1473131
295985
241412
80109
183337
233721
897788
563673
504791
212542
365826
204545
308163
21652
486110

Reduced rank LDA

- let $\hat{\Sigma} = UDU^T$ (eigendecomposition)
- let $x^* = D^{-1/2}U^T x = \hat{\Sigma}^{-1/2}x$
- $\hat{\mu}_k^* = D^{-1/2}U^T\hat{\mu}_k$
- LDA: $\delta_k(x) = (1/2)||x^* - \hat{\mu}_k^*||^2 - \log\hat{\pi}_k$ (closest centroid in sphered space, apart from last term)
- hence if $p > K-1$, can project data onto $K-1$ dim space spanned by $\hat{\mu}_k^*$ and lose nothing

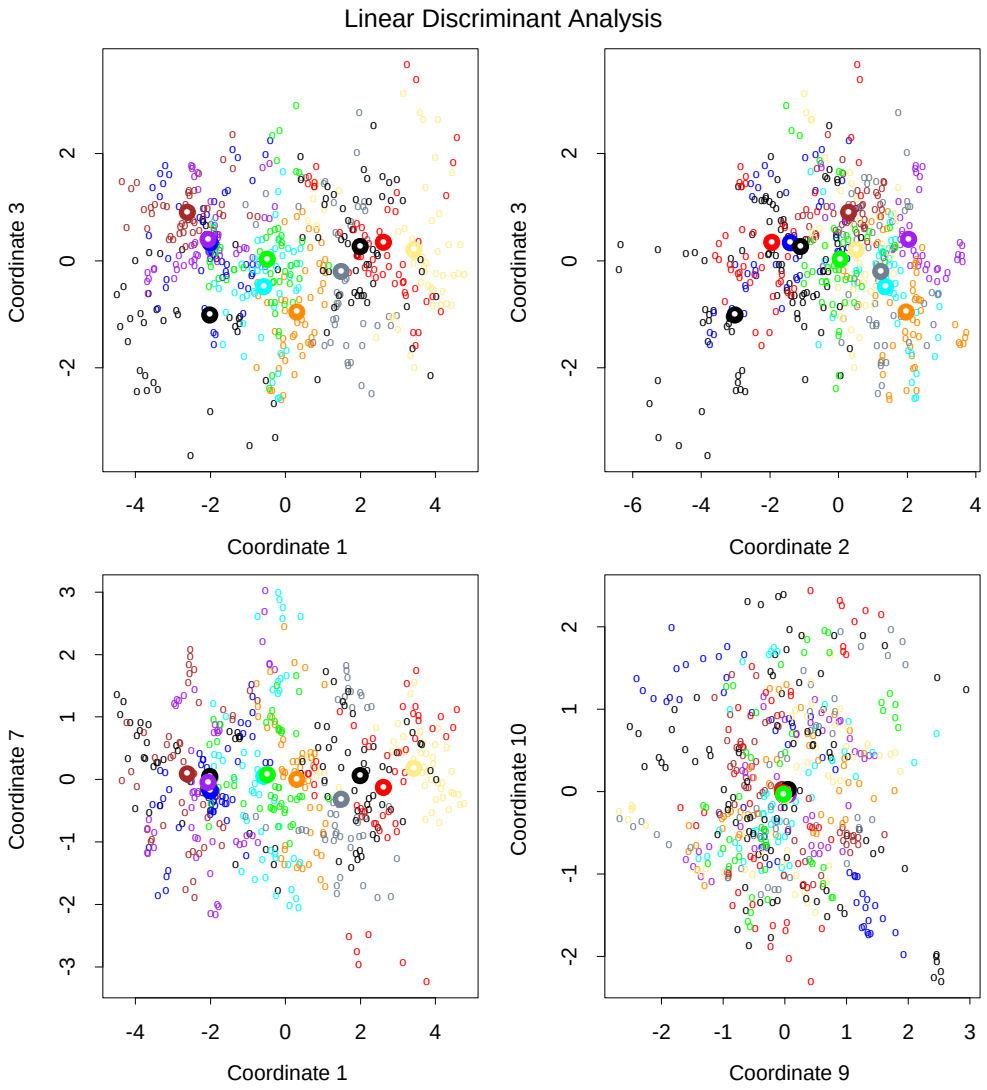
Can project onto even lower dimensions, using the principal components of $\hat{\mu}_k^*$:

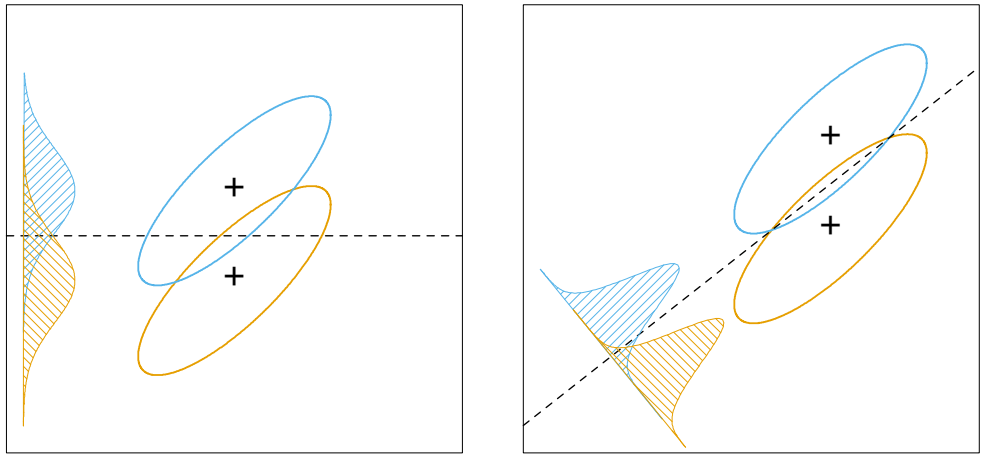
- compute $M = K \times p$ matrix of centroids, $W = \hat{\Sigma}$, $M^* = MW^{-1/2}$
- compute B^*= covariance matrix of M^*, svd $B^* = V^* D_B V^{*T}$
- $z_\ell = v_\ell^T x$ is the ℓth discriminant (or canonical) variable, with $v_\ell = W^{-1/2} v_\ell^*$



A two-dimensional plot of the vowel training data. There are eleven classes with $X \in R^{10}$, and this is the best view in terms of a LDA model. The heavy circles are the projected mean vectors for each class. The class overlap is considerable.

Projections onto pairs of discriminant variates





Although the line joining the centroids defines the direction of greatest centroid spread, the projected data overlap because of the covariance (left panel). The discriminant direction minimizes this overlap for Gaussian data (right panel).

Fisher's formulation of discriminant analysis

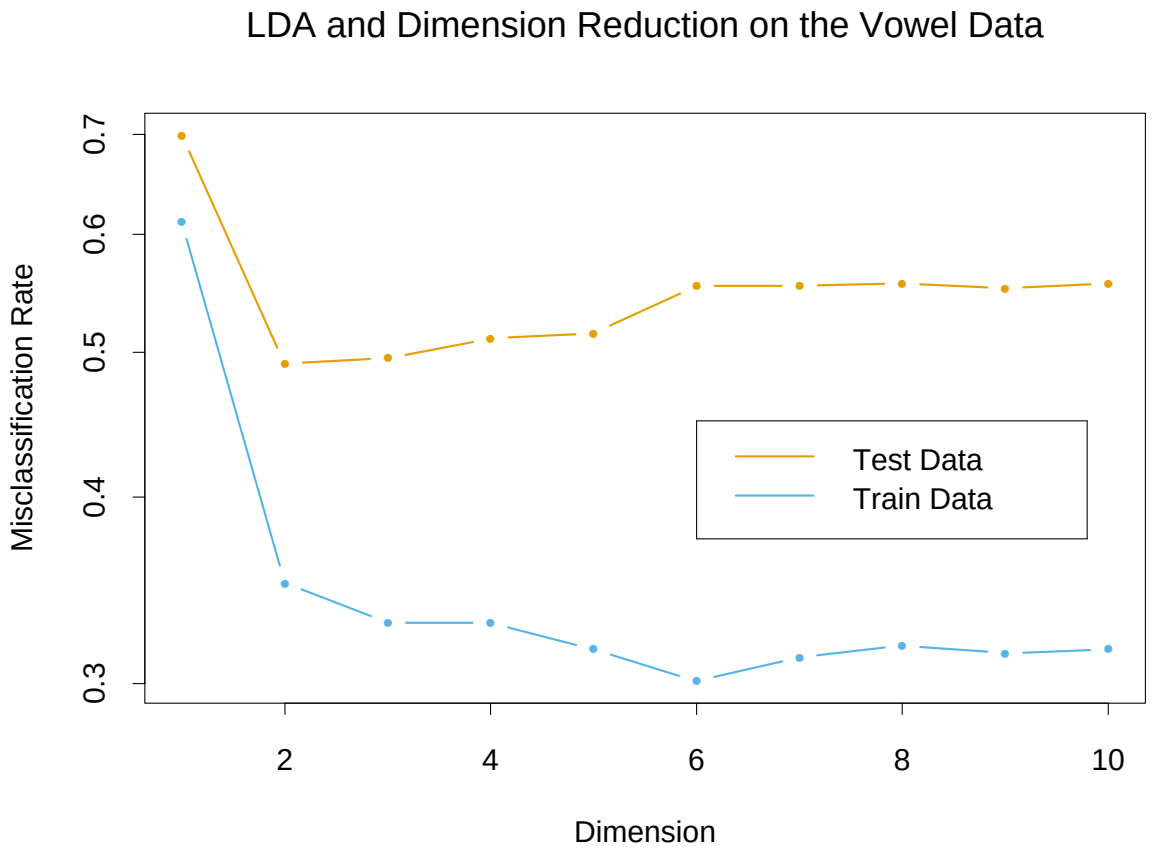
- Find $z = a^T x$ such that the between class variance: is maximized relative to within class variance

$$\max_a \frac{a^T B a}{a^T W a}$$

or

$$\max_a a^T B a \text{ such that } a^T W a = 1$$

- gives $a = v_1$; then do
$\max_{a_2} a_2^T B a_2$ such that $a_2^T W a_2 = 1, a_2 W a = 0$
- gives $a_2 = v_2$ etc



Training and test error rates for the vowel data, as a function of the dimension of the discriminant subspace. In this case the best error rate is for dimension 2.

Linear Logistic Regression

$$\text{logit}P(x) \equiv \log \frac{P(x)}{1-P(x)} = \eta(x) = \beta^T x$$

$$\text{Log-Likelihood} = \sum_{i=1}^{n} \{y_i \log p_i + (1-y_i)\log(1-p_i)\}$$

IRLS algorithm

1. Initialize β.
2. Form linearized responses

$$z_i = \beta^T x_i + (y_i - p_i)/\{p_i(1-p_i)\}$$

3. Form weights $w_i = p_i(1-p_i)$
4. Update β by weighted LS of z_i on x_i with weights w_i.

Steps 2-4 are repeated until convergence.

IRLS is equivalent to Newton-Raphson procedure

$\frac{\partial \ell(\beta)}{\partial \beta} = \mathbf{X}^T(\mathbf{y} - \mathbf{p})$ and $\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} = -\mathbf{X}^T\mathbf{W}\mathbf{X}$.

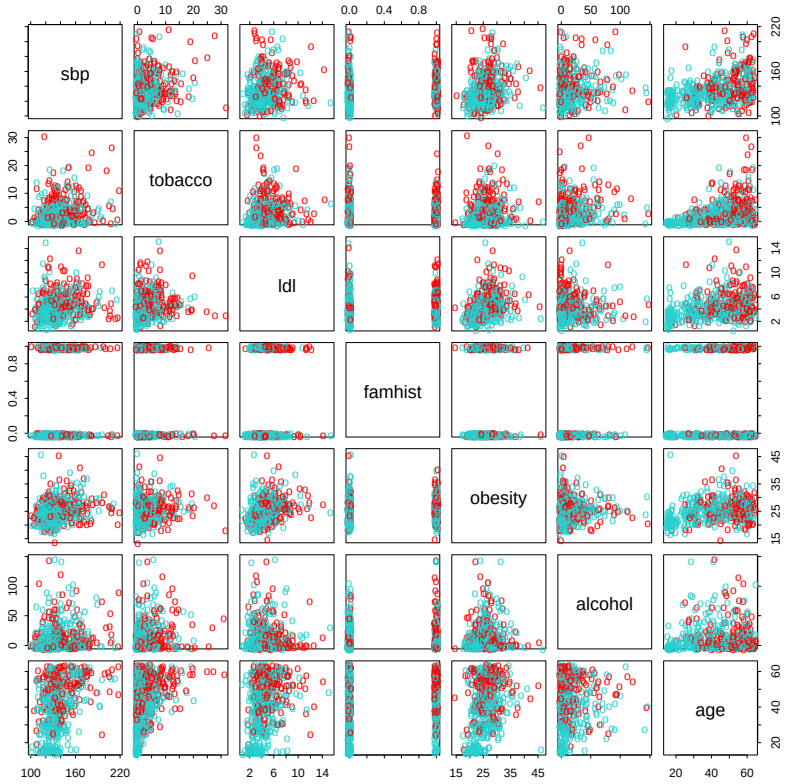
The Newton-Raphson step is thus

$$\begin{aligned} \beta^{new} &= \beta^{old} + (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y} - \mathbf{p}) \\ &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\left(\mathbf{X}\beta^{old} + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{p})\right) \\ &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z}. \end{aligned} \tag{9}$$

In the second and third line we have re-expressed the Newton-Raphson step as a weighted least squares step, with the response

$$\mathbf{z} = \mathbf{X}\beta^{old} + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{p}), \tag{10}$$

South African Heart disease data



A scatterplot matrix of the South African heart disease data. Each plot shows a pair of risk factors, and the cases and controls are color coded (red is a case). The variable famhist family history of heart disease *is binary (yes or no).*

Results from a logistic regression fit to the South African heart disease data.

	Coefficient	Std. Error	Z Score
(Intercept)	−4.130	0.964	−4.285
sbp	0.006	0.006	1.023
tobacco	0.080	0.026	3.034
ldl	0.185	0.057	3.219
famhist	0.939	0.225	4.178
obesity	-0.035	0.029	−1.187
alcohol	0.001	0.004	0.136
age	0.043	0.010	4.184

Model building

- deviance $\text{dev}(y, \hat{p}) = 2\ell(\hat{\beta})$
- H_0 : β has q only non-zero components
- H_1 : β is unrestricted
- under H_0 $\text{dev}(y, \hat{p}_0) - \text{dev}(y, \hat{p}_1) \sim \chi^2_{p-q}$ asymptotically

Logistic regression or LDA?

- LDA:

$$\begin{aligned}\log\frac{\Pr(G=k|X=x)}{\Pr(G=K|X=x)} &= \log\frac{\pi_k}{\pi_K}-\frac{1}{2}(\mu_k+\mu_K)^T\mathbf{\Sigma}^{-1}(\mu_k-\mu_K)\\ &\quad +x^T\mathbf{\Sigma}^{-1}(\mu_k-\mu_K)\\ &= \alpha_{k0}+\alpha_k^T x. \qquad (11)\end{aligned}$$

  This linearity is a consequence of the Gaussian assumption for the class densities, as well as the assumption of a common covariance matrix.

- Logistic model:

$$\log\frac{\Pr(G=k|X=x)}{\Pr(G=K|X=x)}=\beta_{k0}+\beta_k^T x. \qquad (12)$$

  They use the same form for the logits

- Discriminative vs informative learning: logistic regression uses

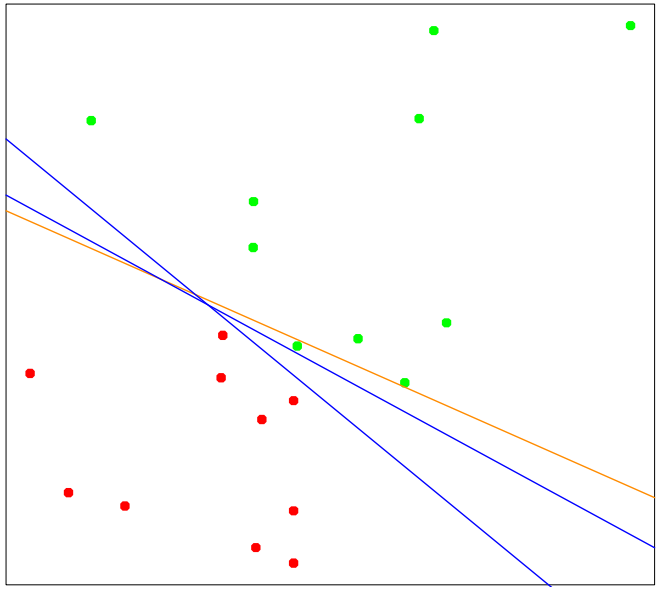
the conditional distribution of Y given x to estimate parameters, while LDA uses the full joint distribution (assuming normality).

$$\Pr(X, G = k) = \Pr(X) \Pr(G = k|X),$$

- If normality holds, LDA is up to 30% more efficient; o/w logistic regression can be more robust. But the methods are similar in practice.

Separating hyperplanes

$$\{x : \beta_0 + \beta_1 x_1 + \beta_2 x_2 = 0\}$$



A toy example with two classes separable by a hyperplane. The orange line is the least squares solution, which misclassifies one of the training points. Also shown are two blue separating hyperplanes found by the perceptron learning algorithm with different random starts.

Rosenblatt's Perceptron Learning Algorithm

If a response $y_i = 1$ is misclassified, then $x_i^T\beta + \beta_0 < 0$, and the opposite for a misclassified response with $y_i = -1$. The goal is to minimize

$$D(\beta, \beta_0) = -\sum_{i \in \mathcal{M}} y_i(x_i^T\beta + \beta_0), \tag{13}$$

over $||\beta|| = 1$, where $\mathcal{M}$ indexes the set of misclassified points.

$$\partial\frac{D(\beta, \beta_0)}{\partial\beta} = -\sum_{i \in \mathcal{M}} y_i x_i, \tag{14}$$

$$\partial\frac{D(\beta, \beta_0)}{\partial\beta_0} = -\sum_{i \in \mathcal{M}} y_i. \tag{15}$$

Stochastic gradient descent:

$$\begin{pmatrix} \beta \\ \beta_0 \end{pmatrix} \leftarrow \begin{pmatrix} \beta \\ \beta_0 \end{pmatrix} + \rho \begin{pmatrix} y_i x_i \\ y_i \end{pmatrix} . \tag{16}$$

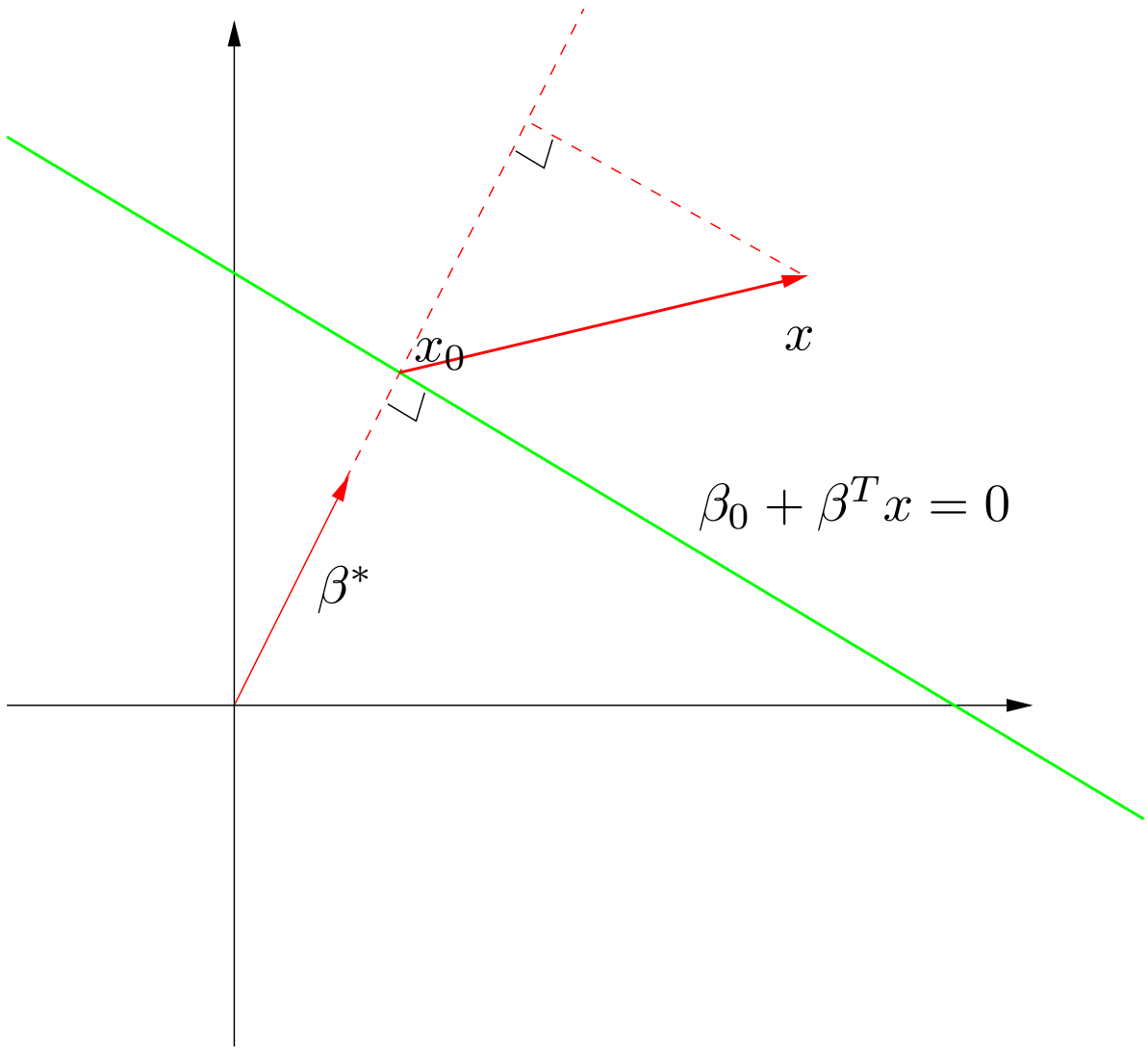
Converges if data are separable (Ex 4.6)

Geometry

Let L be the line $f(x) = \beta_0 + \beta_x = 0$.

The signed distance of any point x to L is given by

$$\begin{aligned} \beta^{*T}(x - x_0) &= \frac{1}{||\beta||}(\beta^T x + \beta_0) \\ &= \frac{1}{||f'(x)||} f(x). \end{aligned} \tag{17}$$



The linear algebra of a hyperplane (affine set).

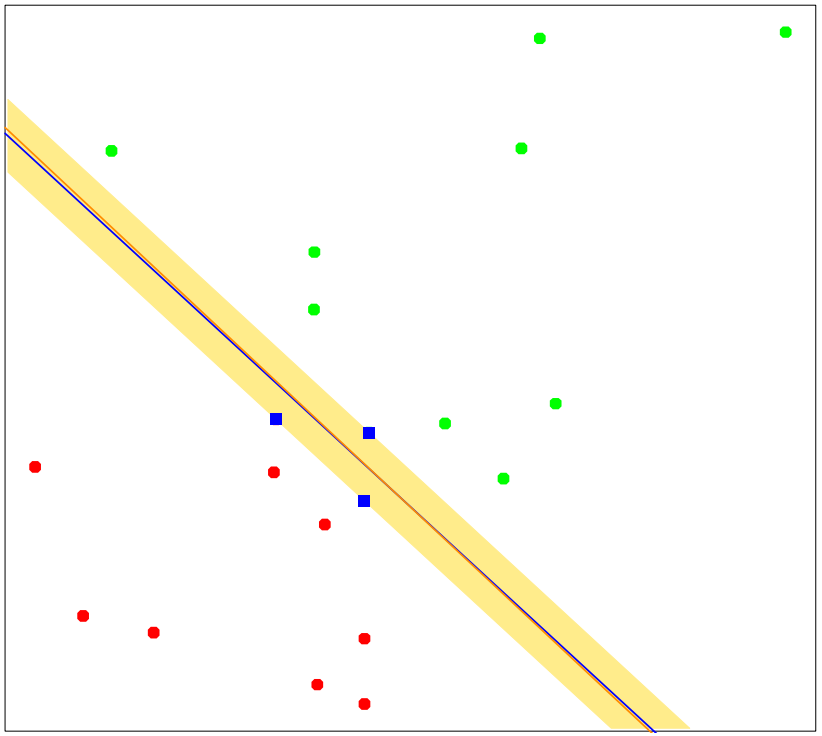
Optimal Separating Hyperplanes

Problem

$$\begin{aligned} &\max_{\beta,\beta_0,||\beta||=1} C \\ &\text{subject to } y_i(x_i^T\beta+\beta_0) \geq C,\ i=1,\ldots,N. \end{aligned} \tag{18}$$

$$\hat{\beta} = \sum_{i=1}^{N} \hat{\alpha}_i y_i x_i \tag{19}$$

$\hat{\alpha}_i > 0$ if x_i is on boundary, else 0. Such boundary points are called support points



The same toy example. The shaded region delineates the maximum margin separating the two classes. There are three support points indicated, which lie on the boundary of the margin, and the optimal separating hyperplane (blue line) bisects the slab. Included in the figure is the boundary found using logistic regression (red line), which is very close to the optimal separating hyperplane (see Chapter 12 of ESL).