Ch4 Linear methods for classification

MATH 6312
Department of mathematics, UTA

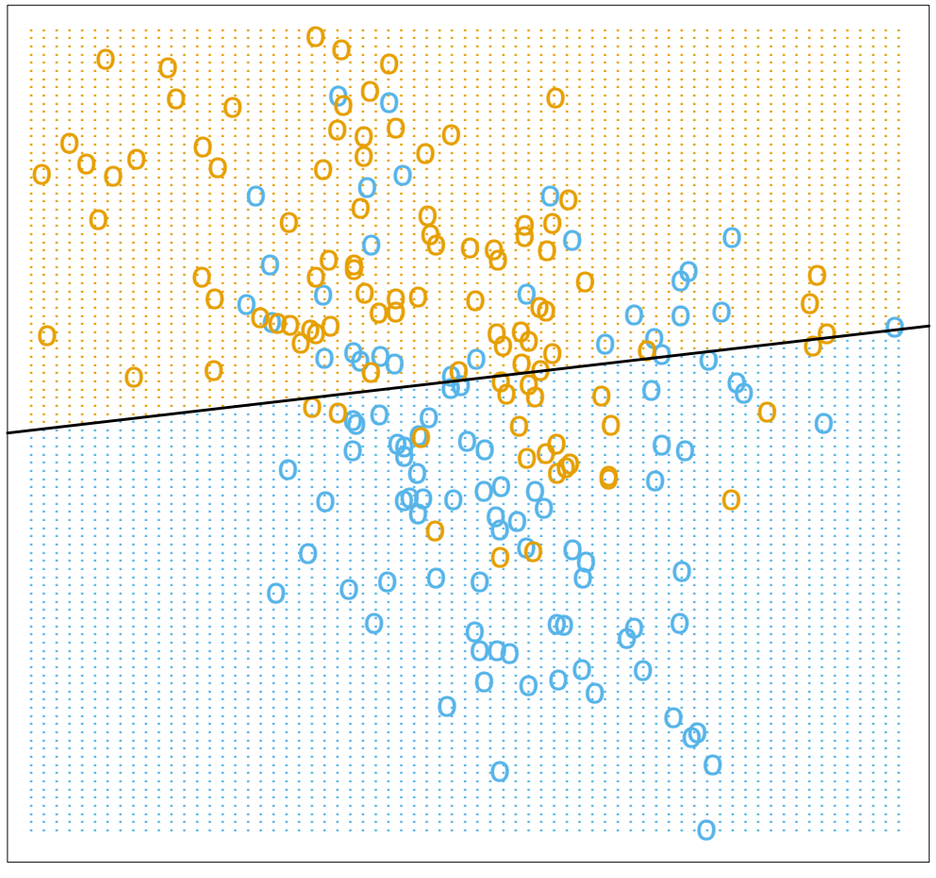
ESL 4.1, 4.3, 4.3.2-3, 4.4, 4.4.1-3, 4.4.5, 6.6.3
Optional reading: ESL 4.2

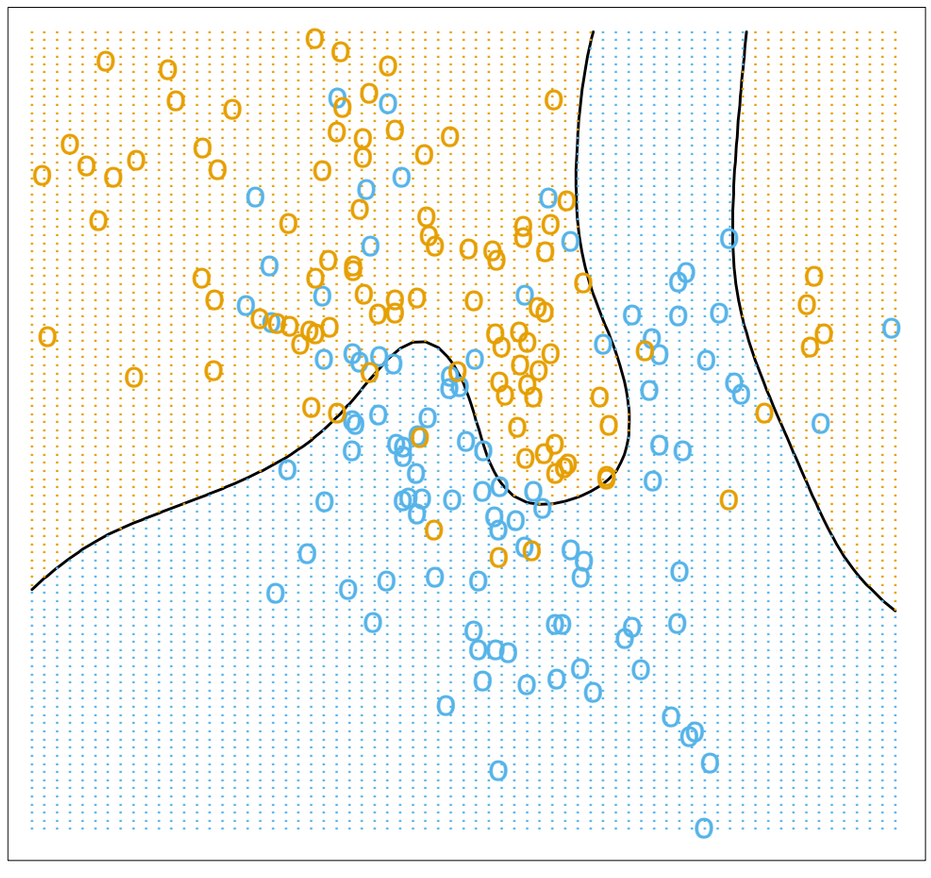
Classification

- Classification is a predictive task in which the response takes values across discrete categories G={1,…,K},
  - eye color∈{brown, blue, green}
  - email∈{spam, ham}.
- Given a feature vector X and a qualitative response Y taking values in the set G, the classification task is to build a classifier f(X) that takes as input the feature vector X and predicts its value for Y; i.e. f(X)∈G.
- Bayes classifier:
- $$f(x) = \underset{k=1,\ldots,K}{\arg\max} \Pr(G = k \mid X = x)$$

Decision boundary

- Decision boundary between two classes is a hyper surface that partitions the underlying space into two sets, one for each class.





Discriminative vs generative algorithm

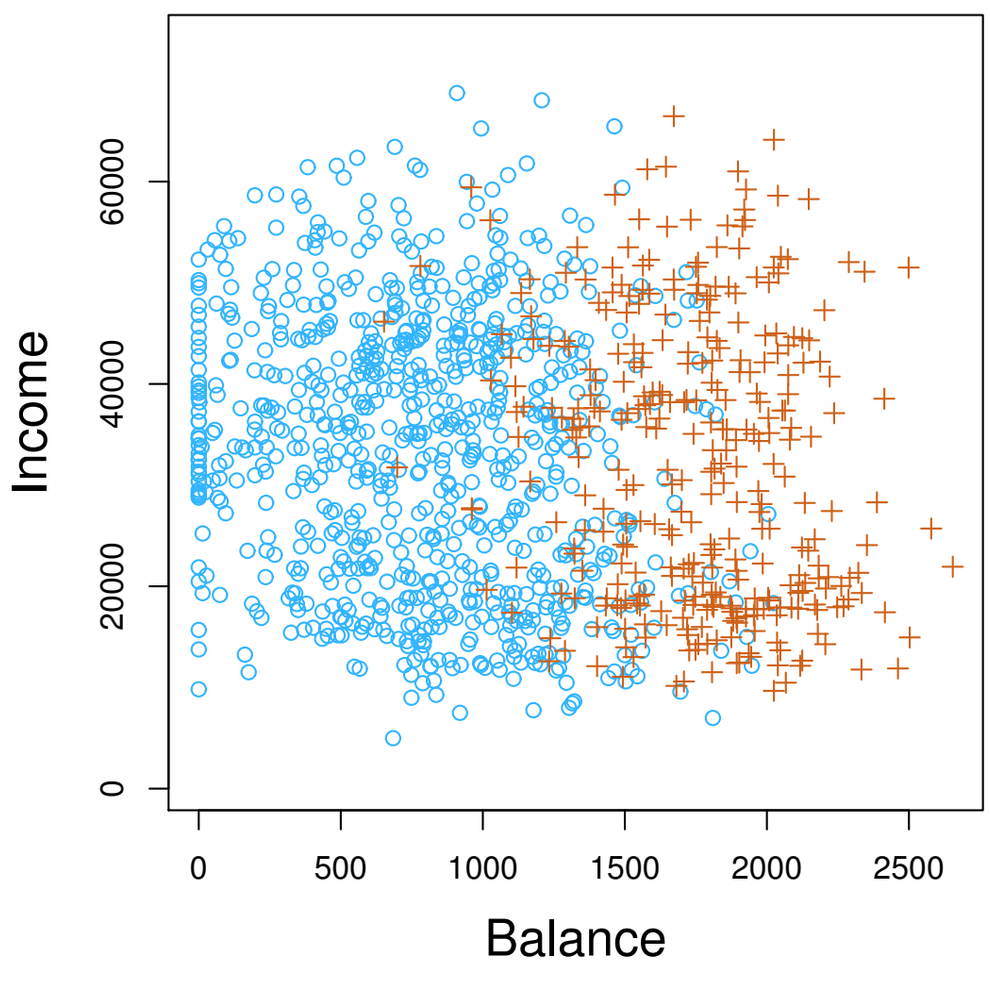
- A discriminative algorithm does not care about how the predictors are generated, it simply classifies the response.
  - Logistic regression: based on p(y|X).
- A generative algorithm models how the predictors are generated in order to classify the response.
  - Linear/Quadratic discriminant analysis (LDA/QDA), naive Bayes: based on p(y,X).

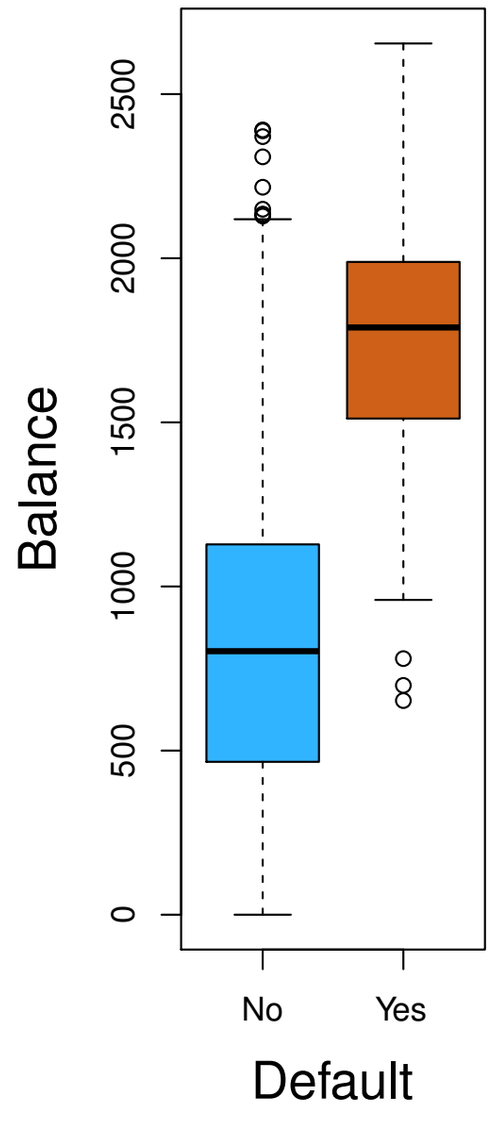
generative

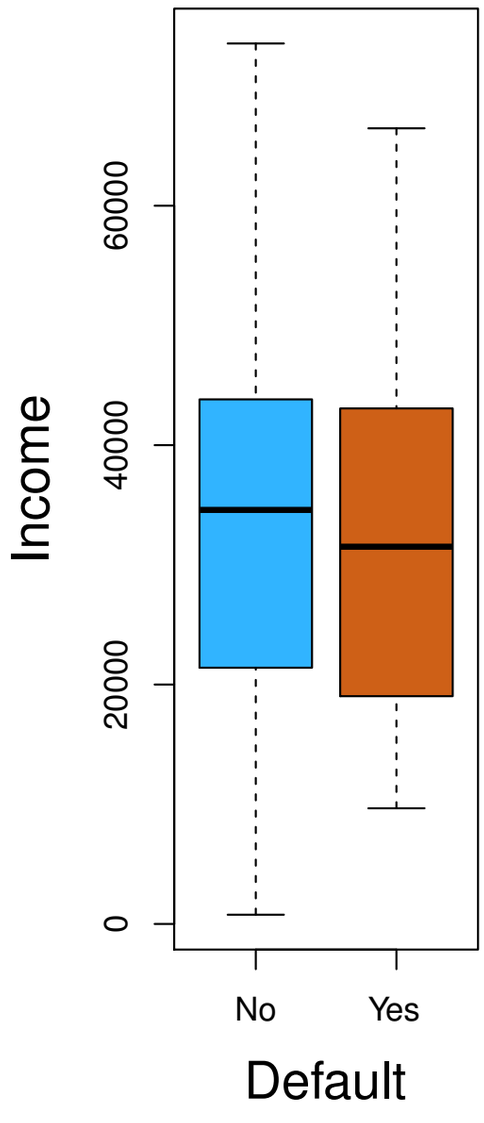
- p(y,X) = p(y|X)p(X) = p(X|y)p(y)

discriminative

Example: Credit card default



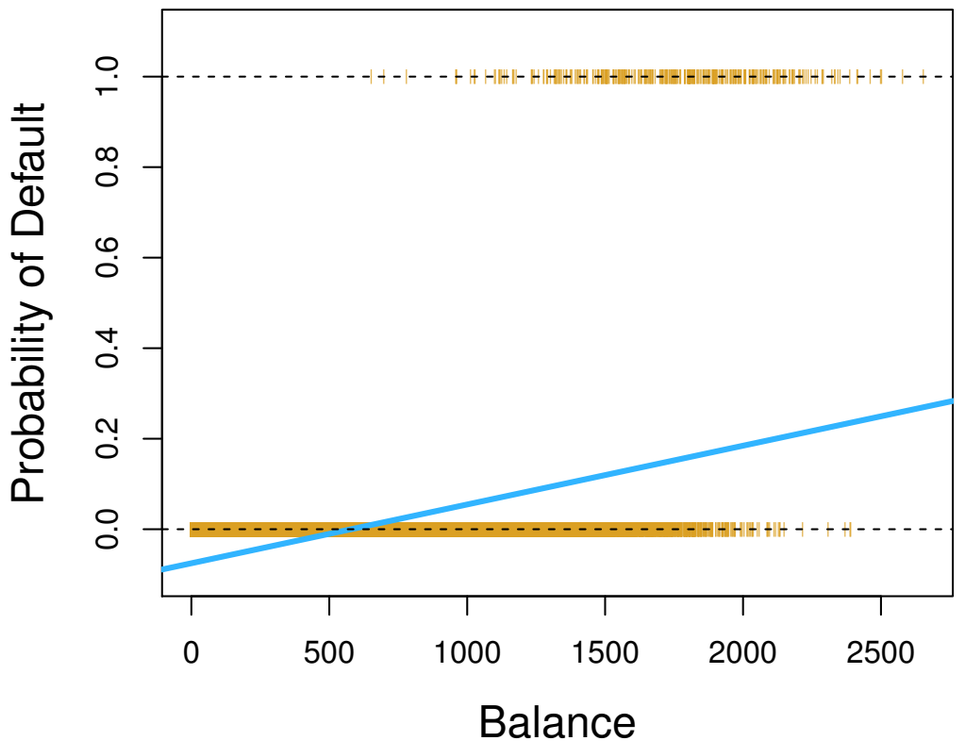


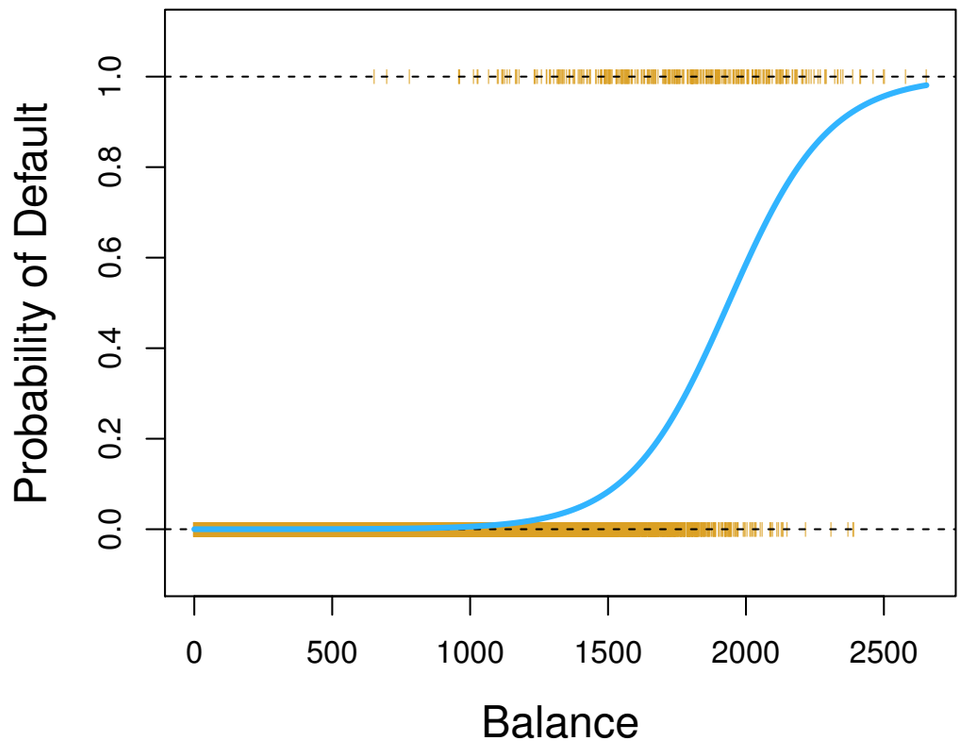


Can we use linear regression?

- Suppose we code Y= 1 if Default; Y = 0 otherwise.
- Can we simply perform a linear regression of Y on X and classify as Yes if the predicted Y > c (=threshold)?
- In this case of a binary outcome, linear regression does a good job as a classifier, and is equivalent to LDA which we discuss later.
- However, the linear regression might produce probabilities less than zero or bigger than one. Also, it often performs very poorly for K>2 classes. The logistic regression is more appropriate.

Linear vs logistic regression





- The orange marks indicate the response Y, either 0 or 1.
- Linear regression does not estimate $Pr(Y = 1|X)$ well.
- Logistic regression seems well suited to the task.

Odds ratio

- Odds of “sick” among treatment

$$\frac{p(sick \mid treatment)}{1-p(sick \mid treatment)}$$

- Odds ratio of “sick”

		outcome	
		sick	healthy
exposure	treatment	p_{11}	p_{12}
	control	p_{21}	p_{22}

$$OR = \frac{p(sick \mid treatment)/[1-p(sick \mid treatment)]}{p(sick \mid control)/[1-p(sick \mid control)]} = \frac{p_{11}p_{22}}{p_{12}p_{21}}$$

- OR = 1: exposure does not affect odds of outcome
- OR>1: exposure associated with higher odds of outcome
- OR<1: exposure associated with lower odds of outcome

Logistic regression (binary classification)

- Let p(x)=Pr(y=1|x). The logistic regression assumes that "log odds" is linear in x.

$$\text{logit}(\text{p}(\mathbf{x}))=\log\frac{p(\mathbf{x})}{1-p(\mathbf{x})}=\beta'\mathbf{x} \text{ where } \mathbf{x}=(1,x_1,\ldots,x_p)',\ \beta=(\beta_0,\beta_1,\ldots,\beta_p)'$$

- Solving the above for p(x), $\Pr(y=1|x)=\frac{1}{1+e^{-\beta'x}}$

- p(x)>1-p(x) if β′x>0; p(x)≤0.5 otherwise. i.e. we classify an observation to 1 if β′x>0; 0 if β′x≤0

- The decision boundary (β′x=0) is linear in x.

Generalized linear models (GLM)

$$E(y) = h^{-1}(X\beta) \Leftrightarrow h[E(y)] = X\beta$$

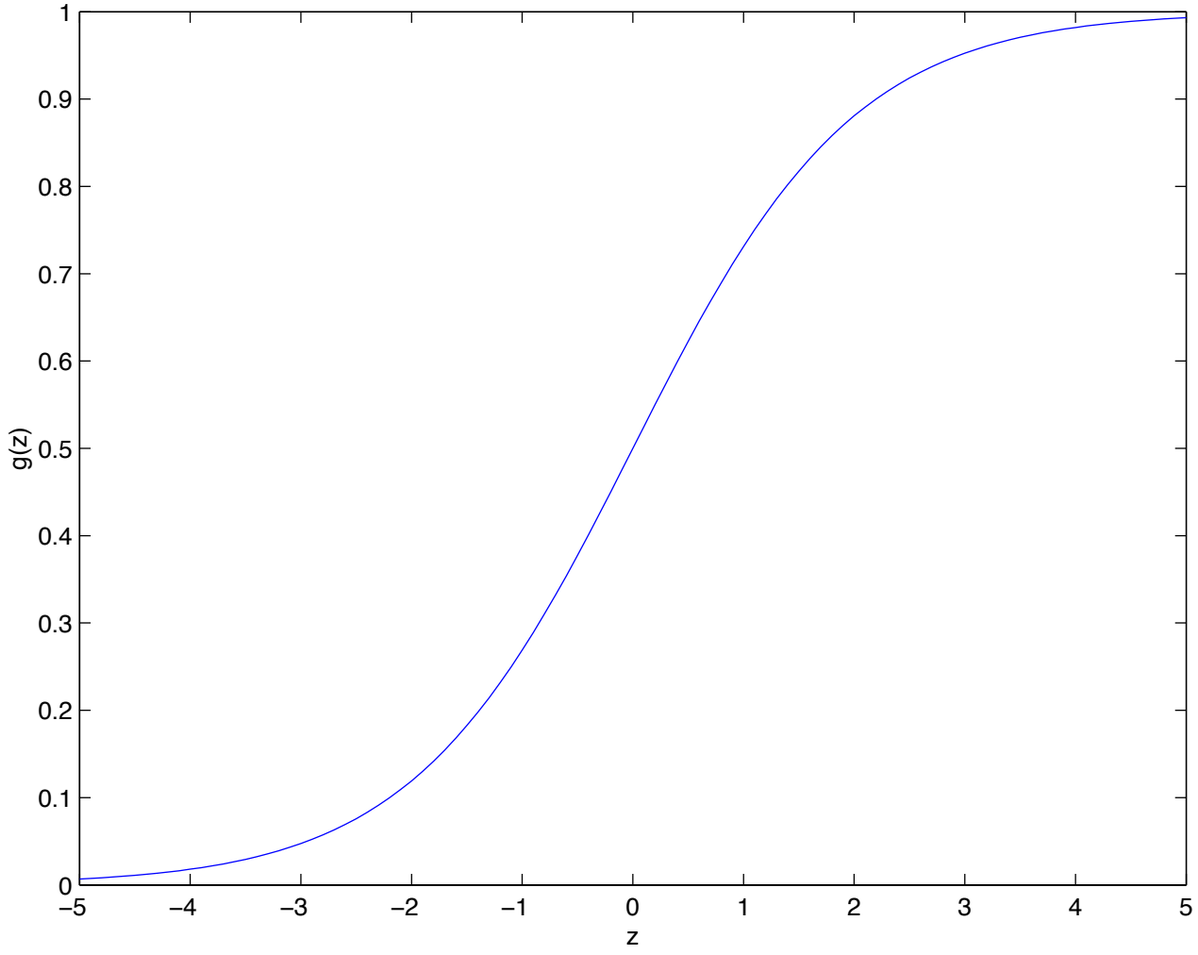
	Link function h()	distribution of y
Linear regression	identity	normal
Logistic regression	logit	binomial

and many more…

Logistic function

- The inverse of logit transformation is called a "logistic" or "sigmoid" function.

$$g(z=\beta'\mathbf{x}) = \text{logit}^{-1}(\beta'\mathbf{x}) = \frac{1}{1+e^{-\beta'\mathbf{x}}} = \frac{e^{\beta'\mathbf{x}}}{1+e^{\beta'\mathbf{x}}} = \Pr(y=1 \mid x)$$



How to fit the logistic regression model

- The mode is fitted by finding the maximum likelihood estimate (MLE) of the coefficient β.

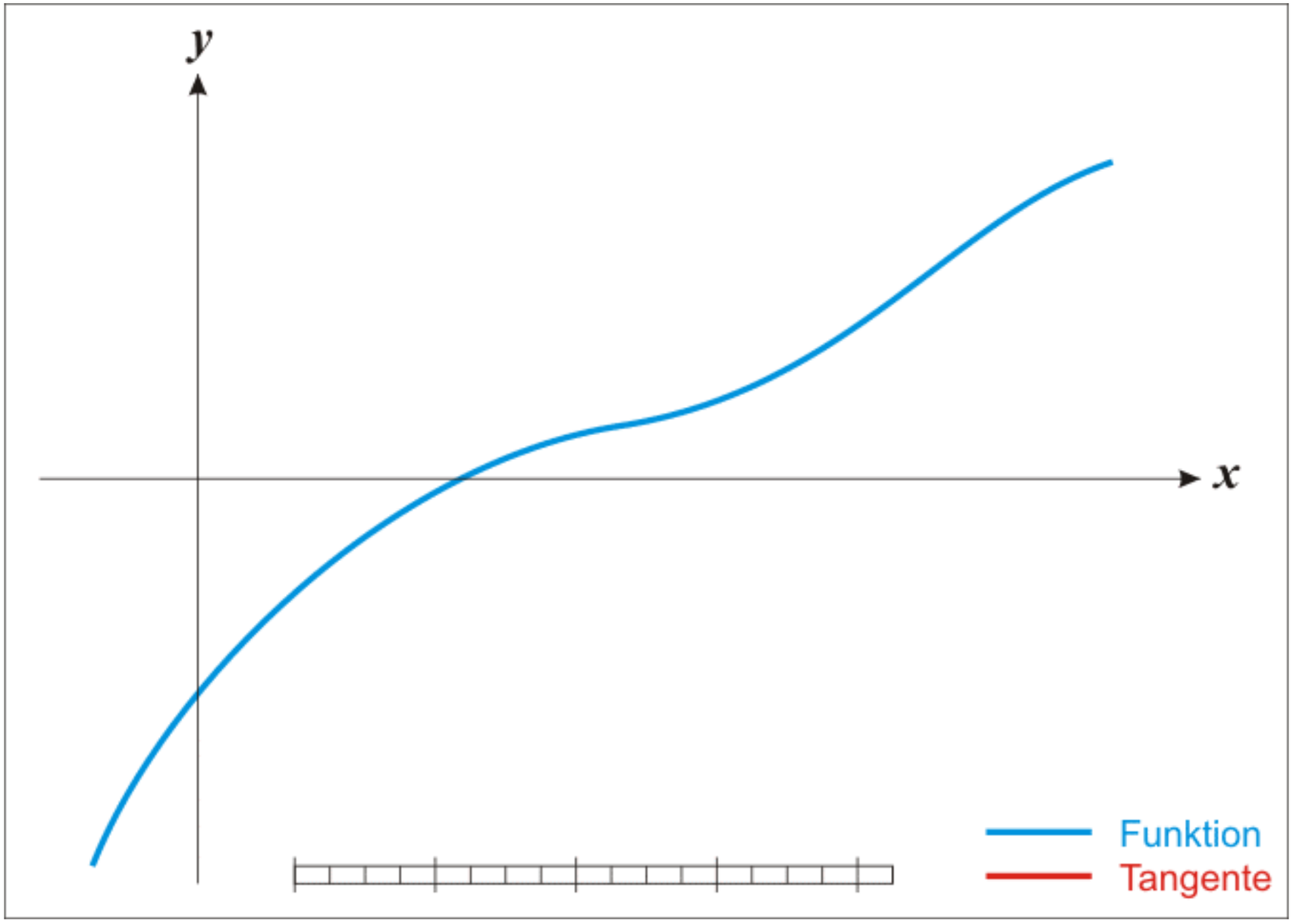
- We code $y_i \in \{0,1\}$ for two classes. So, $y_i \sim \text{bin}(1, p(x_i;\beta))$. Then, the log-likelihood of β is

$$\begin{aligned}\ell(\beta) &= \sum_{i=1}^{N}\left\{y_i \log p(x_i;\beta) + (1-y_i)\log(1-p(x_i;\beta))\right\} \\ &= \sum_{i=1}^{N}\left\{y_i \beta^T x_i - \log(1+e^{\beta^T x_i})\right\}.\end{aligned}$$

$$\Pr(y=1 \mid X=x_i) = p(x_i;\beta) = \frac{1}{1+e^{-\beta' x_i}}$$

- There is no closed form solution for MLE here. The numerical optimization method is used (Newton-Raphson).

- To maximize g(x), we need to find x such that $g'(x)=f(x)=0$. By the Taylor's expansion, $f(x^{new}) - f(x^{old}) = f'(x^{old})(x^{new} - x^{old})+o(|x^{new} - x^{old}|)$
- Newton-Raphson is an iterative method to update x repeatedly to find a solution for $f(x)=0$. The iteration stops when $|x^{old}-x^{new}|$ is very small (or $|g(x^{old})-g(x^{new})|$).



- https://commons.wikimedia.org/wiki/File%3ANewtonIteration_Ani.gif

- Back to the logistic regression: to apply the Newton-Raphson, one needs to calculate the 1st (gradient) and 2nd (Hessian) derivatives, and update β by

$$\beta^{\text{new}} = \beta^{\text{old}} - \left(\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} \right)^{-1} \frac{\partial \ell(\beta)}{\partial \beta}$$

- The update formula can be rewritten as

$$\beta^{new} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z} \qquad \mathbf{z} = \mathbf{X}\beta^{old} + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{p})$$

$$\mathbf{W} = [\,p(x;\beta^{old})(1 - p(x;\beta^{old}))]_{ii} \text{ is a } N \times N \text{ diagonal matrix of weights}$$

$$\mathbf{p} = [\,p(x_1;\beta^{old}),\ldots,p(x_N;\beta^{old})]'$$

- Iteratively reweighed least squares (IRLS): At each iteration, we update the weight matrix **W** and re-weight the response y to get the adjusted response z.

- $$\beta^{new} = \arg\min_{\beta} (z - X\beta)'W(z - X\beta)$$

Significant test (H_0: $\beta=b_0$ vs H_1: $\beta\neq b_0$)

- Fisher information: the inverse of Fisher information is used for asymptotic variance of MLE.
  - $I(\beta)=-E\left(\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta'}\right)$ $\qquad I(\beta)^{-1}=(\mathbf{X'WX})^{-1}\approx \mathrm{var}(\hat{\beta})$
  - For a sufficiently large N, $\hat{\beta}\sim N(\beta,(\mathbf{X'WX})^{-1})$

- Wald test

$$W=\frac{(\hat{\beta}_j-b_0)^2}{\mathrm{var}(\hat{\beta}_j)}\sim\chi_1^2$$

$\mathrm{var}(\hat{\beta}_j)$ is jth diagonal of $(\mathbf{X'WX})^{-1}$ evaluated at $\beta=\hat{\beta}$

Likelihood ratio test (LRT)

- Goodness-of-fit: how well do the observed data correspond to the fitted model?
  - Deviance D^2: Compare log-likelihoods of saturated and fitted model M_1. The saturated model uses y_i's to classify y_i. (p_1: # of parameters in M_1)

$$D^2(M_1) = 2(\log L(M_{saturated}) - \log L(M_1)) \sim \chi^2_{N-p_1}$$

- Based on the deviance, one can perform tests for nested models

$$G^2 = D^2(M_{smaller}) - D^2(M_{bigger}) \sim \chi^2_{df}$$

  - df = # of parameters in bigger - # of parameters in smaller.

Wald test vs LRT

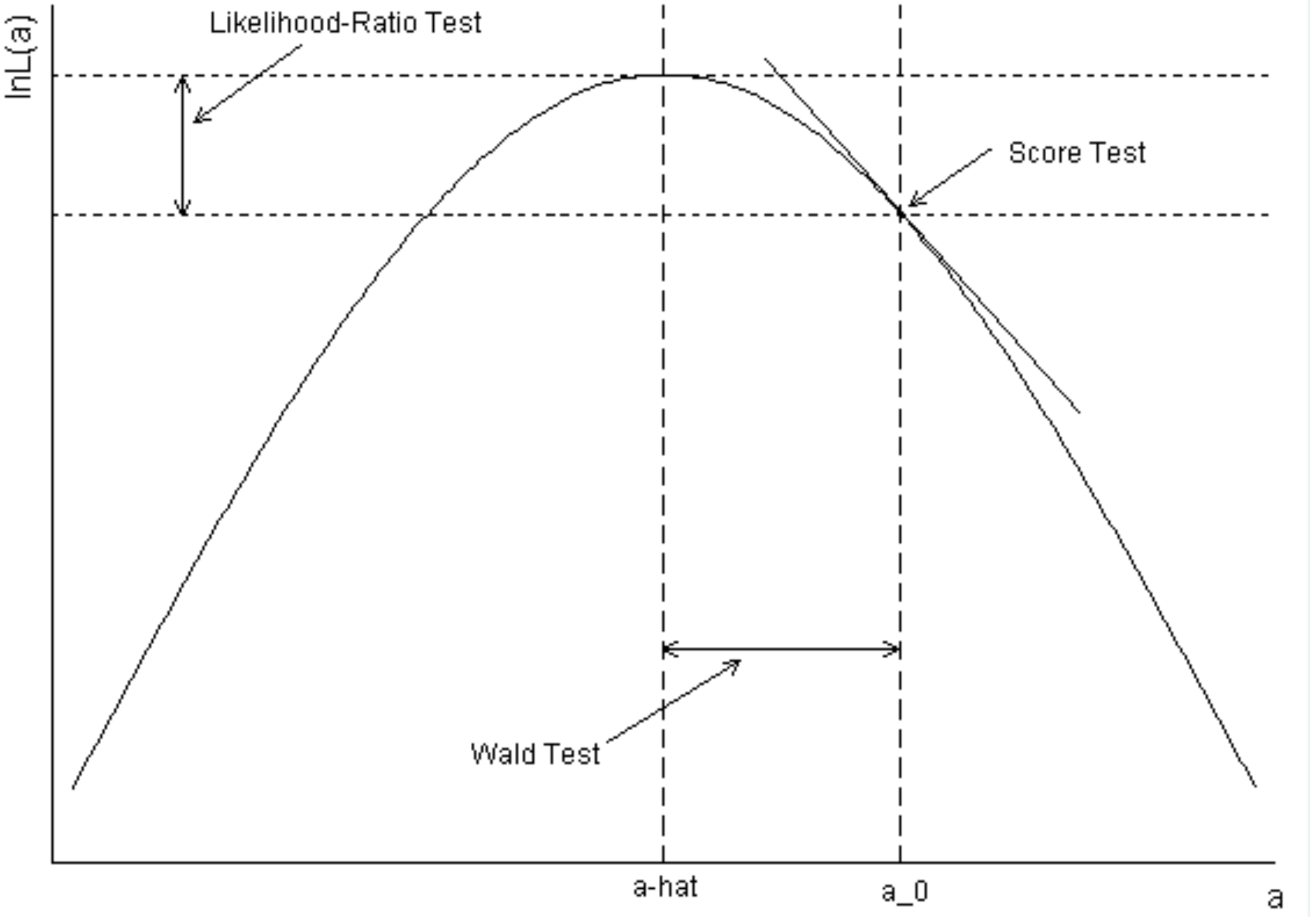
- The Wald approach enjoys popularity due to its simplicity (confidence intervals based on LRT are difficult to construct by hand).

- Both tests can be used to test multi-parameters simultaneously. The two approaches often agree quite well for large sample.

- Wald test uses the variance of estimate under the alternative hypothesis. Confidence interval based on the Wald test is symmetric due to the normal assumption.

- Tests and confidence intervals based on likelihood ratios are more accurate.

- For more about likelihood based test, see https://sites.google.com/site/xgsu00/stat-5370/files-stat5370/notes-10-GLM.pdf?attredirects=0&d=1


lnL(a)
Likelihood-Ratio Test
Score Test
Wald Test
a-hat
a_0
a

Pearson residuals

- Pearson residual: $e_i = \dfrac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}$ where $\hat{p}_i = p(\mathbf{x}_i, \hat{\beta})$
- Pearson's chi square (X^2) is another measure for the goodness-of-fit, and it has the same asymptotic distribution as the deviance.

$$X^2 = \sum_{i=1}^{N} e_i^2 \sim \chi^2_{N-p}$$

- Standardized Pearson residual: $|r_i|>2$ or 3 provides evidence of lack of fit

$$r_i = e_i / \sqrt{1-h_i},\ h_i = ith \text{ diagonal of } \mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2}$$

- Cook's distance (p=# of predictors): $D_i = r_i^2 \dfrac{h_i}{(1-h_i)p}$

Deviance residuals

- Deviance residual

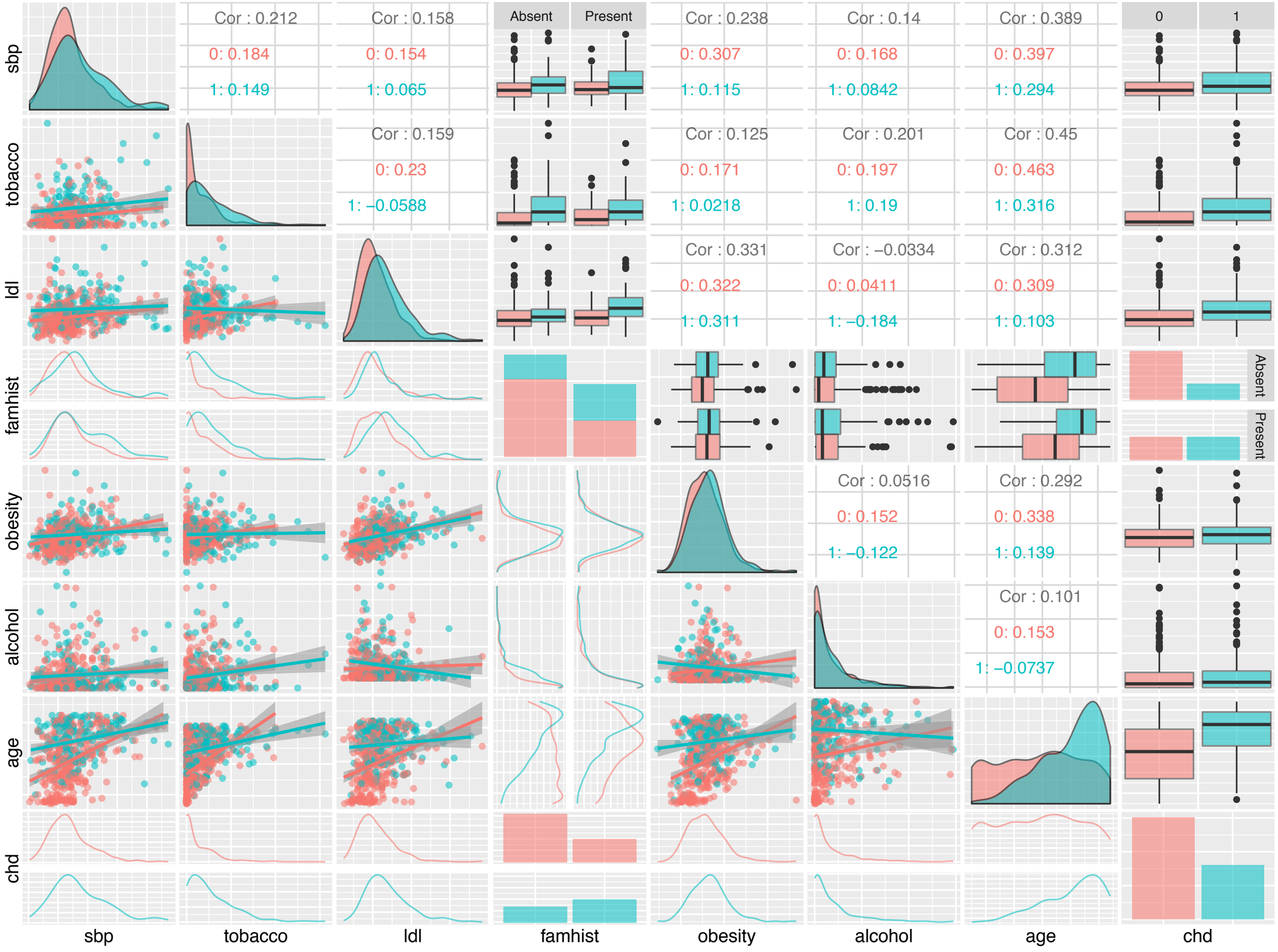
$$d_i = \mathrm{sgn}(y_i - \hat{p}_i)\sqrt{2[y_i \log\frac{y_i}{\hat{p}_i} + (1-y_i)\log\frac{1-y_i}{1-\hat{p}_i}]},\ \mathrm{D}^2 = \sum_{i=1}^{N} d_i^2$$

- It can be standardized similarly. $d_i / \sqrt{1-h_i}$

- Generally speaking, the standardized deviance residuals tend to be preferable because they are more symmetric than the standardized Pearson residuals, but both are commonly used.

Example: South African heart disease data

- There are n = 462 individuals broken up into 160 cases (those who have coronary heart disease) and 302 controls (those who don't). There are p = 7 variables measured on each individual.
  - sbp (systolic blood pressure)
  - tobacco (lifetime tobacco consumption in kg)
  - ldl (low density lipoprotein cholesterol)
  - famhist (family history of heart disease, present or absent)
  - obesity; alcohol; age.


sbp
tobacco
ldl
famhist
obesity
alcohol
age
chd
Cor : 0.212
0: 0.184
1: 0.149
Cor : 0.158
0: 0.154
1: 0.065
Absent
Present
Cor : 0.238
0: 0.307
1: 0.115
Cor : 0.14
0: 0.168
1: 0.0842
Cor : 0.389
0: 0.397
1: 0.294
0
1
Cor : 0.159
0: 0.23
1: −0.0588
Cor : 0.125
0: 0.171
1: 0.0218
Cor : 0.201
0: 0.197
1: 0.19
Cor : 0.45
0: 0.463
1: 0.316
Cor : 0.331
0: 0.322
1: 0.311
Cor : −0.0334
0: 0.0411
1: −0.184
Cor : 0.312
0: 0.309
1: 0.103
Absent
Present
Cor : 0.0516
0: 0.152
1: −0.122
Cor : 0.292
0: 0.338
1: 0.139
Cor : 0.101
0: 0.153
1: −0.0737
sbp
tobacco
ldl
famhist
obesity
alcohol
age
chd

	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

- Correlated variables can cause multicollinearity and interpretation problems.
  - sbp and obseity are not significant, and obesity has a negative sign. (Marginally, these are both significant and have positive signs)
- Model selection: find a subset of the variables that are sufficient for explaining their joint effect on the prevalence of chd.
  - (1) Drop the least significant coefficient, and refit the model.
  - (2) Repeat (1) until no further terms can be dropped from the model.

- The final model from the selection procedure.

	Coefficient	Std. Error	Z score
`(Intercept)`	−4.204	0.498	−8.45
`tobacco`	0.081	0.026	3.16
`ldl`	0.168	0.054	3.09
`famhist`	0.924	0.223	4.14
`age`	0.044	0.010	4.52

- Interpretation of tobacco coefficient
  - exp(0.081) is the odds ratio associated with a one-unit increase in tobacco.
  - No casual inference for the observational study!

Multiclass logistic regression

- Log odds for K-1 classes are linear in x, and computed using Pr(G=K|X=x) as a baseline.

$$\log \frac{\Pr(G=1|X=x)}{\Pr(G=K|X=x)} = \beta_{10} + \beta_1^T x$$

$$\log \frac{\Pr(G=2|X=x)}{\Pr(G=K|X=x)} = \beta_{20} + \beta_2^T x$$

$$\vdots$$

$$\log \frac{\Pr(G=K-1|X=x)}{\Pr(G=K|X=x)} = \beta_{(K-1)0} + \beta_{K-1}^T x.$$

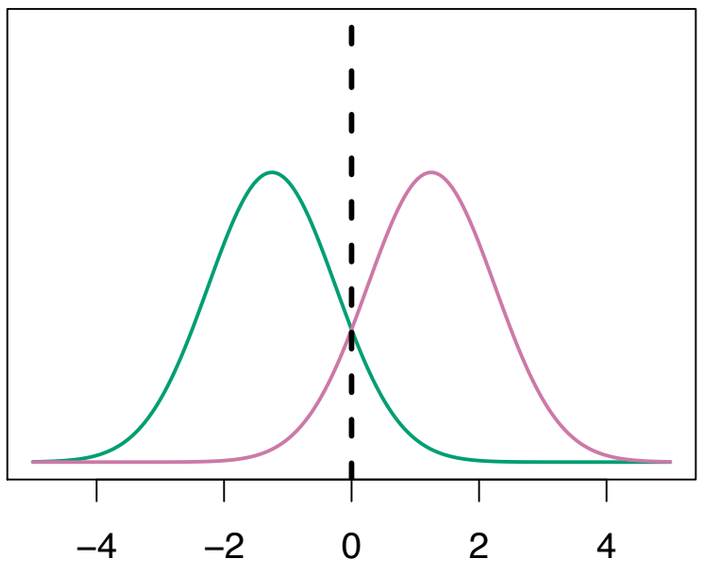
- An observation is classified to G=k if log odds for G=k is positive and the largest among K-1 log odds.

- If no log odds are positive, the observation is classified to G=K.

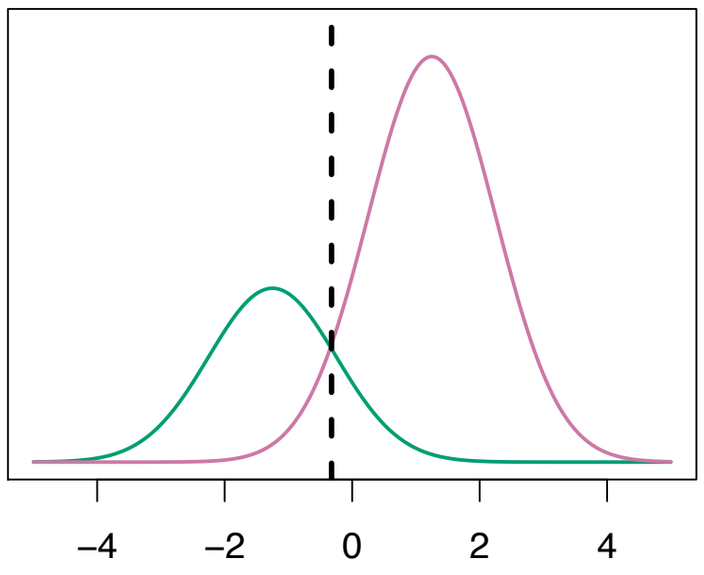
Linear discriminant analysis (LDA)

- $f_k(x)=\Pr(X=x|Y=k)$: density for X in class k. We assume $[X|Y=k] \sim N(\mu_k,\Sigma)$.
  - μ_k: centroid in class k. $[X=x|Y=k]$ has equal variance!
- $\pi_k=\Pr(Y = k)$: marginal or prior probability for class k.
- Bayes theorem: $$\Pr(Y=k|X=x)=\frac{\Pr(X=x|Y=k)\Pr(Y=k)}{\Pr(X=x)}=\frac{\pi_k f_k(x)}{\sum_{l=1}^{K}\pi_l f_l(x)}$$
-

π_1=.5, π_2=.5



π_1=.3, π_2=.7



- log odds for classes k and l:

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

- Discriminant function (it is linear in x!)

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

- Bayes classifier for LDA: $f^{LDA}(x) = \arg\max_{k=1,\ldots,K} \delta_k(x)$

- Priors and moments are estimated using sample moments and frequencies.

$\hat{\pi}_k = N_k/N$, where N_k is the number of class-k observations;

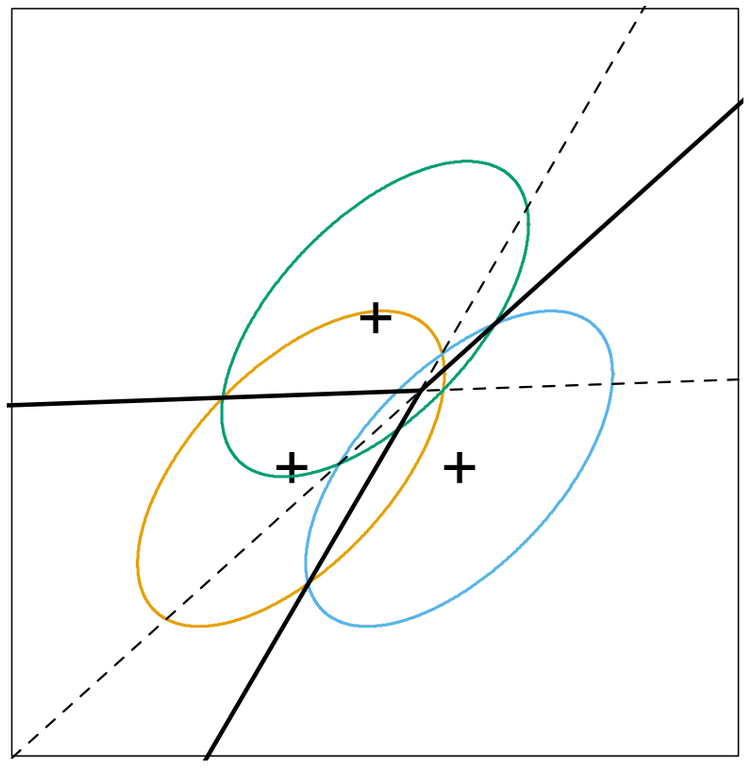
- $\hat{\mu}_k = \sum_{g_i=k} x_i/N_k$;

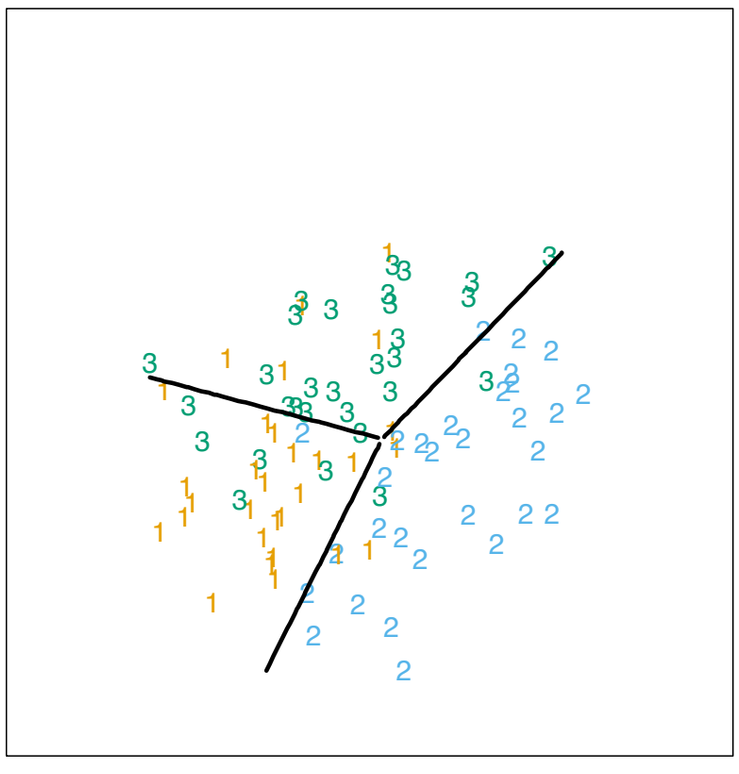
$\hat{\mathbf{\Sigma}} = \sum_{k=1}^{K} \sum_{g_i=k} (x_i-\hat{\mu}_k)(x_i-\hat{\mu}_k)^T/(N-K)$.

Decision boundary

- The discriminant function is linear in x, so the decision boundary between two classes k and l becomes linear in x.

$$\delta_k(x) = \delta_l(x) \Rightarrow x'\Sigma'(\mu_k - \mu_l) = (\mu_k + \mu_l)'\Sigma(\mu_k - \mu_l) - \log\frac{\pi_k}{\pi_l}$$





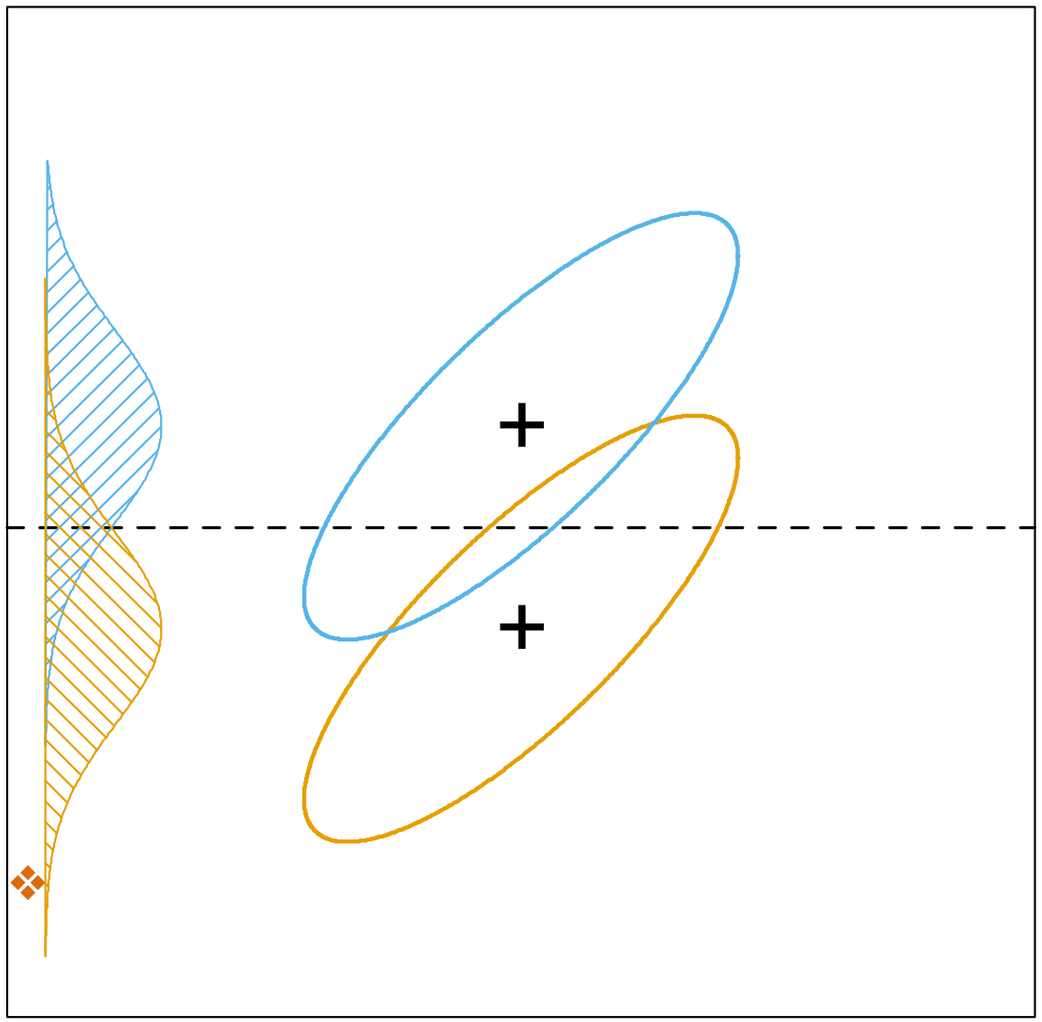
Reduced-rank LDA

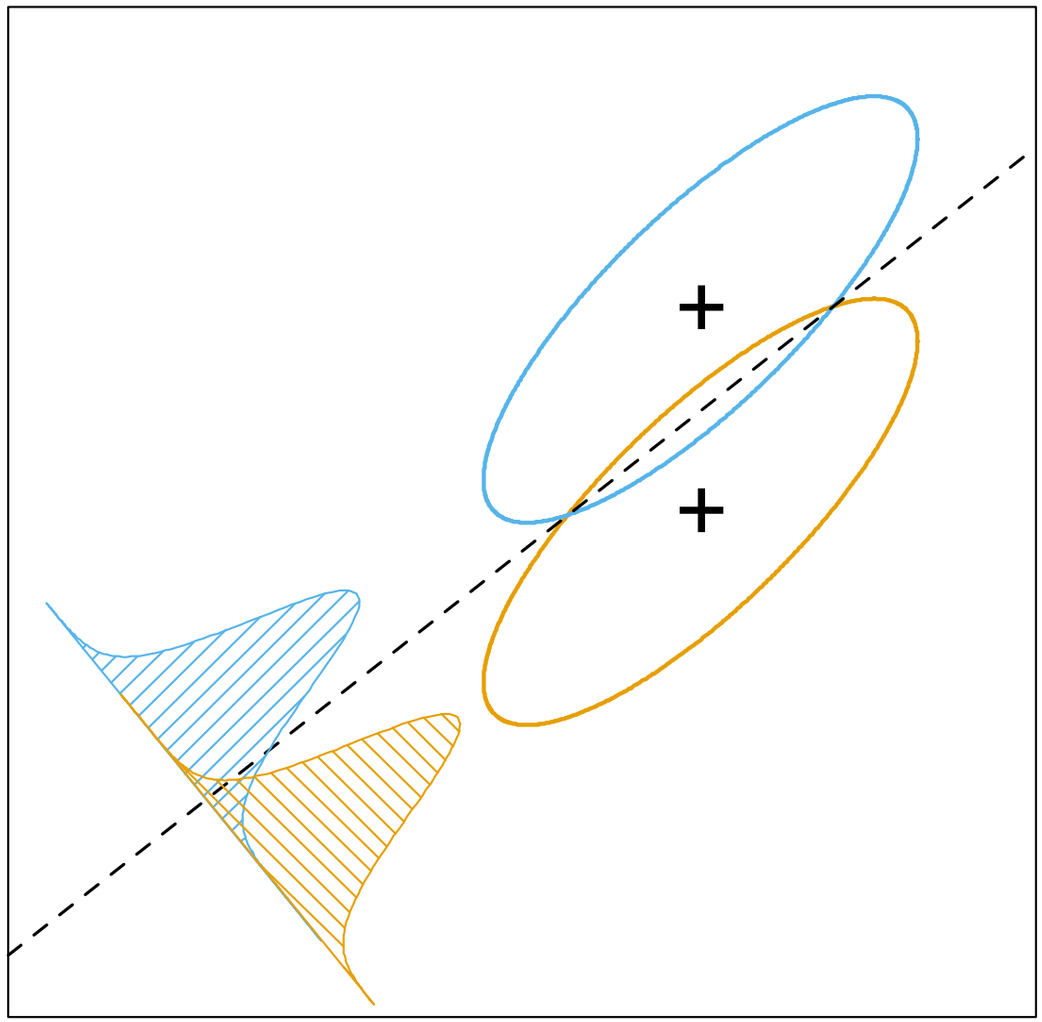
- Why might we want to reduce a dimension to L<K–1? To view informative low-dimensional projections of the data.
  - Visualization: view the data in a two-dimensional plot without losing too much the information needed for LDA classification.
  - Regularization: some dimensions may not providing a lot of separation between the classes, but just noise.
- Reduced-rank LDA is a nice way to project down to lower than K–1 dimensions. It chooses the lower dimensional subspaces so as to spread out the centroids as much as possible
- The dimension reduction from p to K–1 was exact, in that we didn't change the LDA rule at all.

- By taking the (within-class) covariance into account as well, a direction with minimum overlap can be found.

- For K>3, we can find K coordinates with the smallest to largest overlaps (also known as canonical variates or discriminant coordinates).

- Discriminant or canonical variables are obtained by projecting predictor variables onto these coordinates.





- **B**: between-class covariance, **W**: within-class covariance. **T**=**B+W**: total variance

- Fisher's problem: We want to find a direction that separate K centroids well, relatively to **W**. This can be found by maximizing the Rayleigh quotient:

$$\max_a \frac{a'\mathbf{B}a}{a'\mathbf{W}a} \Leftrightarrow \max_a a'\mathbf{B}a \text{ subject to } a'\mathbf{W}a = 1$$

- 1st eigenvalue (λ_1: largest eigenvalue) of $\mathbf{W}^{-1}\mathbf{B}$ is the solution to the problem. a_1 (=1st eigenvector of $\mathbf{W}^{-1}\mathbf{B}$) is the 1st discriminant coordinates.

- The j*th* eigenvector a_j is the j*th* discriminant coordinate. (i.e. a_j yields the j*th* largest quotient while it is orthogonal to other discriminant coordinates).

- In this way, the feature space can further decomposed into successively optimal subspaces in term of centroid separation.

- It is easy to show that

  - 1) $\mathbf{W}^{1/2}a_k$ is an eigenvector of $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$;

  - 2) covariance matrices $\mathbf{B}_Z$ and $\mathbf{W}_Z$ calculated based on discriminant variables $\{z_1=\mathbf{X}a_1,\ldots,z_{K-1}=\mathbf{X}a_{K-1}\}$ are diagonal matrices; and

  - 3) $\mathbf{W}_Z$ is an identity matrix.

- In other words, we transform the data points into the sphered space.

- Let **A**=(a_1,…,a_{K-1}), **Z**=**XA** be a new feature matrix of the discriminant variables, $\tilde{\mu}_k = \mu_k^T A$ be a new centroid transformed to the sphere space.

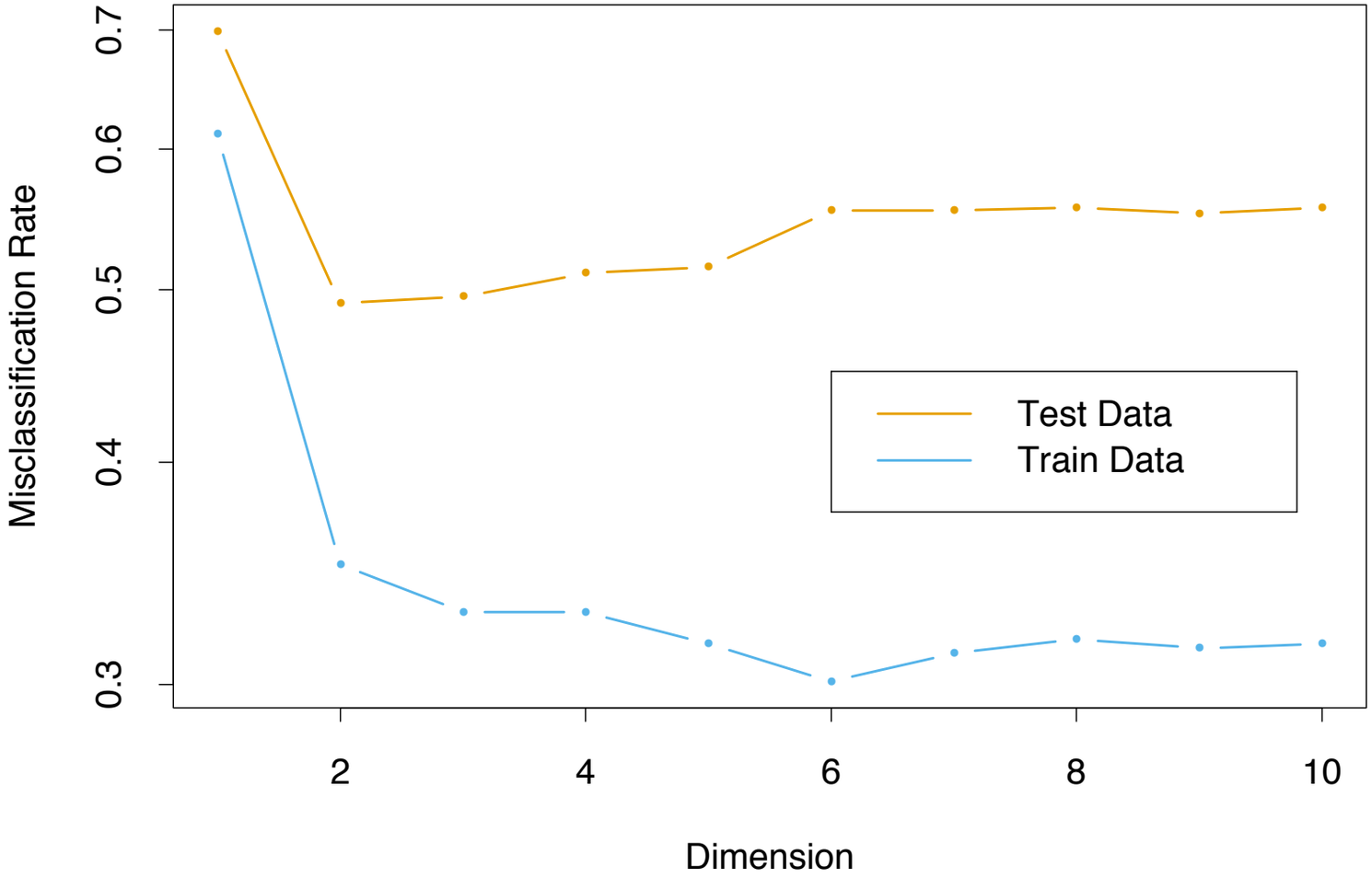
- In the sphered space, the LDA classifies a point z to the closest centroid (modulo logπ_k).

$$\hat{f}^{LDA}(z) = \arg\min_{j=1,\ldots,K} \frac{1}{2}\left\| z - \tilde{\mu}_j \right\|_2^2 - \log \pi_j$$

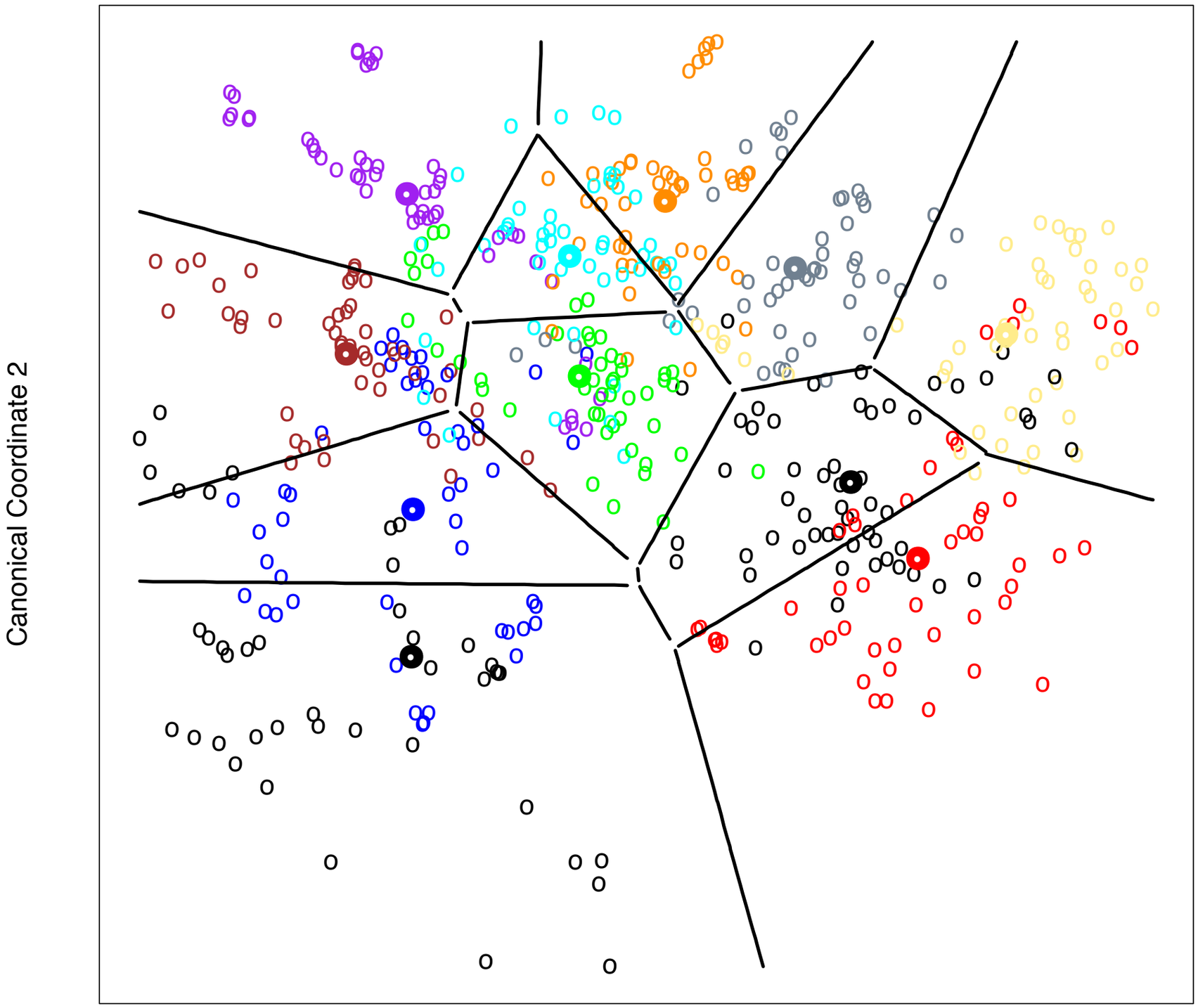
- The transformation was a bit complicated. In this way, however, the classification in the LDA sounds much simpler and more intuitive :)

- Now, how can you choose a dimension for a reduced-rank LDA?

- Vowel data example: n = 528 instances of spoken vowels (K=11), and p = 10 features measured.
  - 10 possible dimensions for the LDA classifier.
  - We choose # of coordinates with the minimized LDA misclassification rate in the test sample.



Two dimensional LDA solution



Quadratic discriminant analysis (QDA)

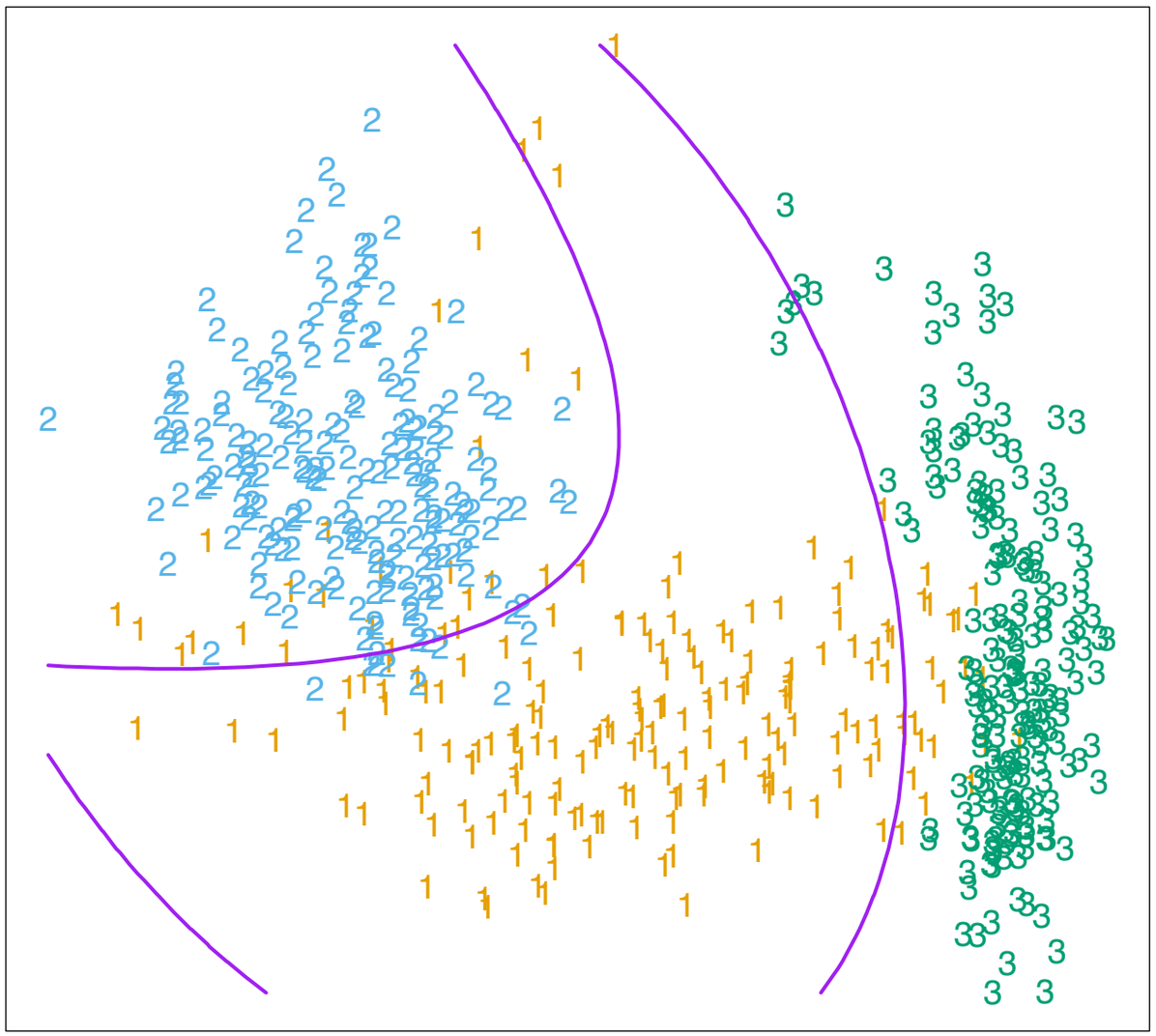
- QDA assumes different covariance matrix Σ_k of X for each class: $[X|Y=k] \sim N(\mu_k, \Sigma_k)$

- Quadratic discriminant function

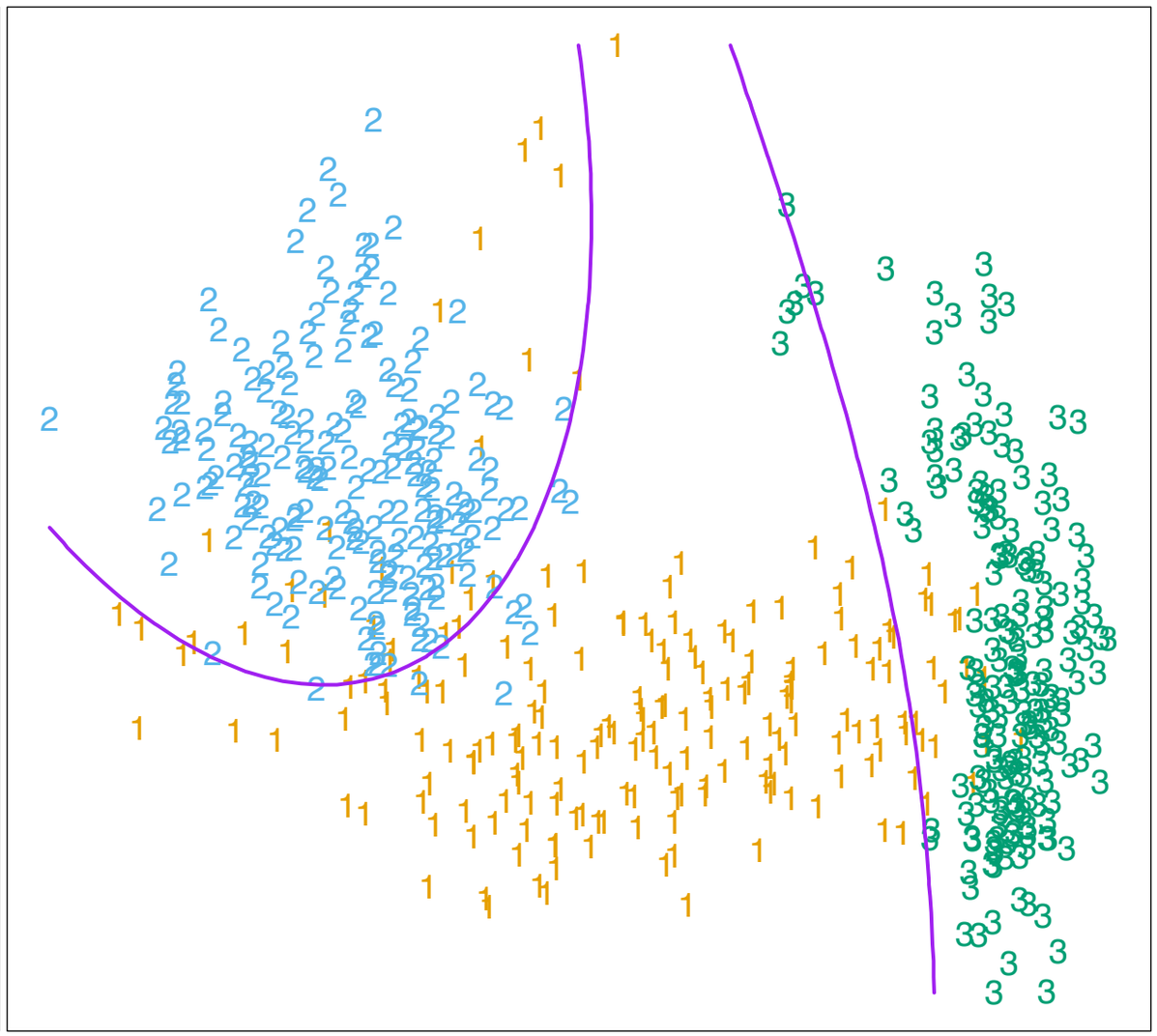
$$\delta_k(x) = -\frac{1}{2}\log|\mathbf{\Sigma}_k| - \frac{1}{2}(x-\mu_k)^T\mathbf{\Sigma}_k^{-1}(x-\mu_k) + \log\pi_k.$$

- Another way to get quadratic decision boundaries is to include quadratic terms in LDA. (i.e. $X_1, X_2, X_1X_2, X_1^2, X_2^2$)

QDA



LDA with quadratic terms



LDA vs logistic regression

- The logistic regression has no assumption on P(X) nor p(X|y=k), so it is more robust. When classes are well separated, however, the logistic regression is very unstable.

- The model fitting for the LDA (sample estimates) has less computational burden than the logistic regression (IRLS).

- Due to the normality assumption, LDA and QDA refer to as the Gaussian discriminant analysis (GDA).

- When LDA performs better than the logistic regression?

  - When the assumption on p(X|y=k) is correct, the LDA is more efficient, especially at small N.

Naive Bayes

- In GDA, predictors are assumed to be continuous (i.e. normal). GDA can be applied for discrete predictors, but it would be less efficient.

- In the LDA, one need to estimate p(p+1)/2 elements for the covariance matrix. (Kp(p+1)/2 for the QDA). For large p, the estimate may be very unstable.

- We assume predictors are independent given Y=k.

  - Often, it performs better than the GDA with non-normal predictors for large p.

- Assumption:

$$f_j(X) = P((X_1, X_2, \ldots, X_p)|Y = j) = \prod_{k=1}^{p} P(X_k|Y = j) = \prod_{k=1}^{p} f_{jk}(X_k)$$

  - This assumption is strong, but useful for large p. One can assume different densities for mixed predictors (collection of qualitative and quantitative predictors).

  - Typically, the univariate density estimate technique is applied (high-dimensional density estimation is unattractive and unstable).

$$\begin{aligned}
\log \frac{\Pr(G = \ell|X)}{\Pr(G = J|X)} &= \log \frac{\pi_\ell f_\ell(X)}{\pi_J f_J(X)} \\
&= \log \frac{\pi_\ell \prod_{k=1}^{p} f_{\ell k}(X_k)}{\pi_J \prod_{k=1}^{p} f_{Jk}(X_k)} \\
&= \log \frac{\pi_\ell}{\pi_J} + \sum_{k=1}^{p} \log \frac{f_{\ell k}(X_k)}{f_{Jk}(X_k)} \\
&= \alpha_\ell + \sum_{k=1}^{p} g_{\ell k}(X_k).
\end{aligned}$$

Confusion matrix

- Accuracy (= 1-misclassification rate)
  - (tp + tn) / (tp + tn + fp + fn)
- True positive rate (=sensitivity =recall): tp / (tp + fn)
- False positive rate (= 1-specificity): fp / (fp + tn)
- Precision: tp / (tp + fp)
-

Prediction	True condition: positive	True condition: negative
positive	true positive (tp)	false positive (fp)
negative	false negative (fn)	true negative (tn)

- high recall and low precision (or specificity)
  - many predicted positive (high sensitive), but most of them are incorrect.
- low recall and high precision (or specificity)
  - very few predicted positive (low sensitive), but most of them are correct.

- Tradeoff between
  - sensitivity and specificity (or 1-specificity): ROC curve
  - recall (=sensitivity) and precision: PR curve

- Precision are useful when classes are imbalanced. i.e., positive class is rare or more interesting (e.g. rare disease, airport scanning).

balanced

True condition

Prediction	positive	negative
positive	50	10
negative	10	50

imbalanced

True condition

Prediction	positive	negative
positive	50	100
negative	10	500

- Specificity remains unchanged between balanced and imbalanced, while precision changed.