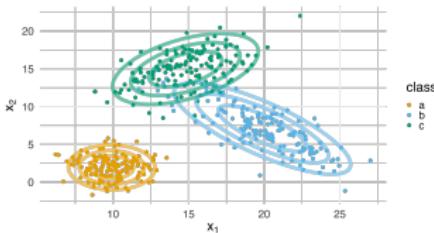


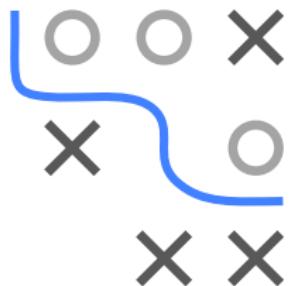
# Introduction to Machine Learning

## Classification Discriminant Analysis



### Learning goals

- LDA and QDA construction principle based on generative approach
- How are their parameters estimated
- Linear and quadratic decision boundaries



# LINEAR DISCRIMINANT ANALYSIS

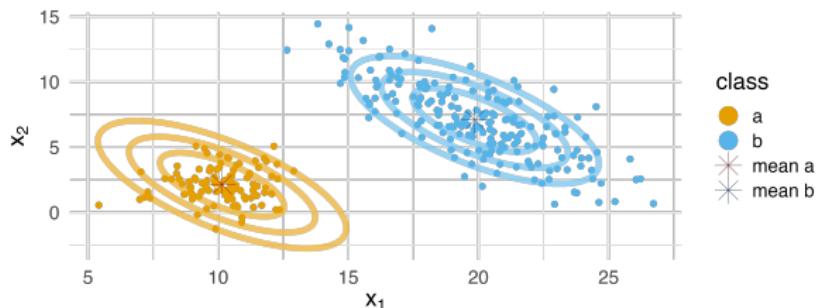
Generative approach, following Bayes' theorem:

$$\pi_k(\mathbf{x}) \approx \mathbb{P}(y = k \mid \mathbf{x}) = \frac{\mathbb{P}(\mathbf{x}|y = k)\mathbb{P}(y = k)}{\mathbb{P}(\mathbf{x})} = \frac{p(\mathbf{x}|y = k)\pi_k}{\sum_{j=1}^g p(\mathbf{x}|y = j)\pi_j}$$

Assume that distribution  $p(\mathbf{x}|y = k)$  per class is **multivariate Gaussian**:

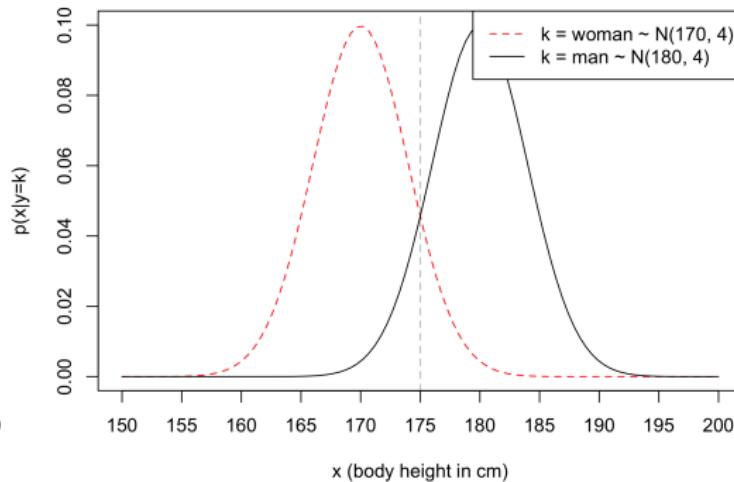
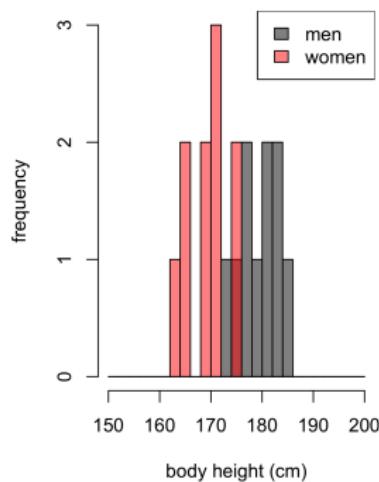
$$p(\mathbf{x}|y = k) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)\right)$$

with **equal covariance structure**, so  $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma} \quad \forall k$



# UNIVARIATE EXAMPLE

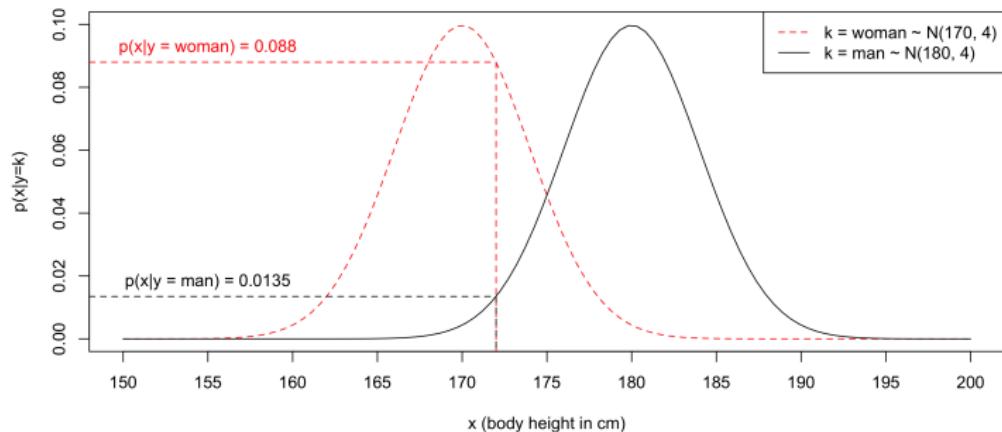
- Classify a new person as male or female based on their height  
(naive toy example, unrealistic in many ways)
- We will compute in the true DGP, so we assume we know all distributions and their params; we use the LDA setup



Optimal separation is located at the intersection (= decision boundary)!

# UNIVARIATE EXAMPLE: EQUAL CLASS SIZES

Let's compute posterior probability that a 172 cm tall person is male

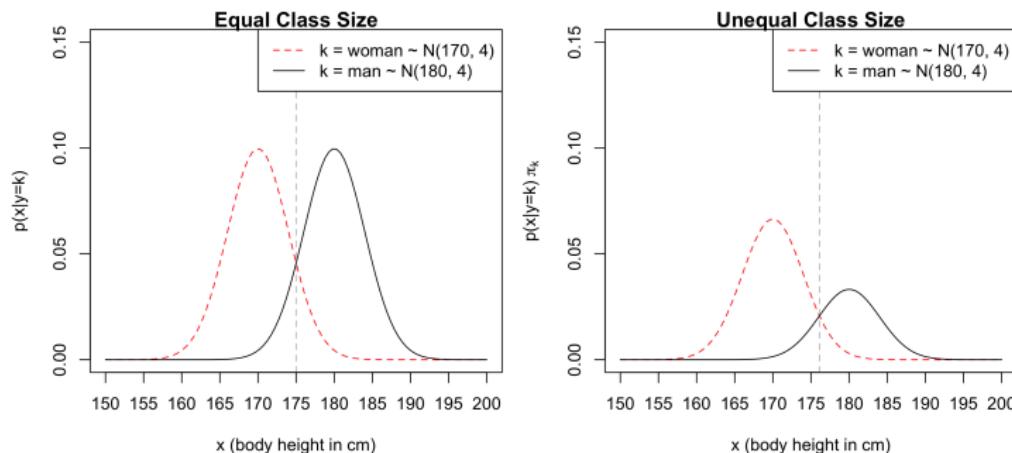


Assuming equal class sizes, prior probs  $\pi_k$  cancel out (since  $\pi_{\text{man}} = \pi_{\text{woman}}$ ):

$$\mathbb{P}(y = \text{man} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid y = \text{man})}{p(\mathbf{x} \mid y = \text{man}) + p(\mathbf{x} \mid y = \text{woman})} = \frac{0.0135}{0.0135 + 0.088} = 0.133$$

# UNIVARIATE EXAMPLE: UNEQUAL CLASS SIZES

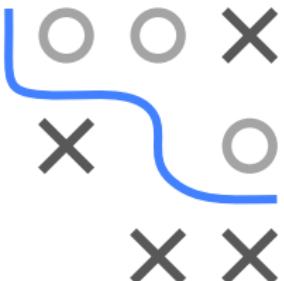
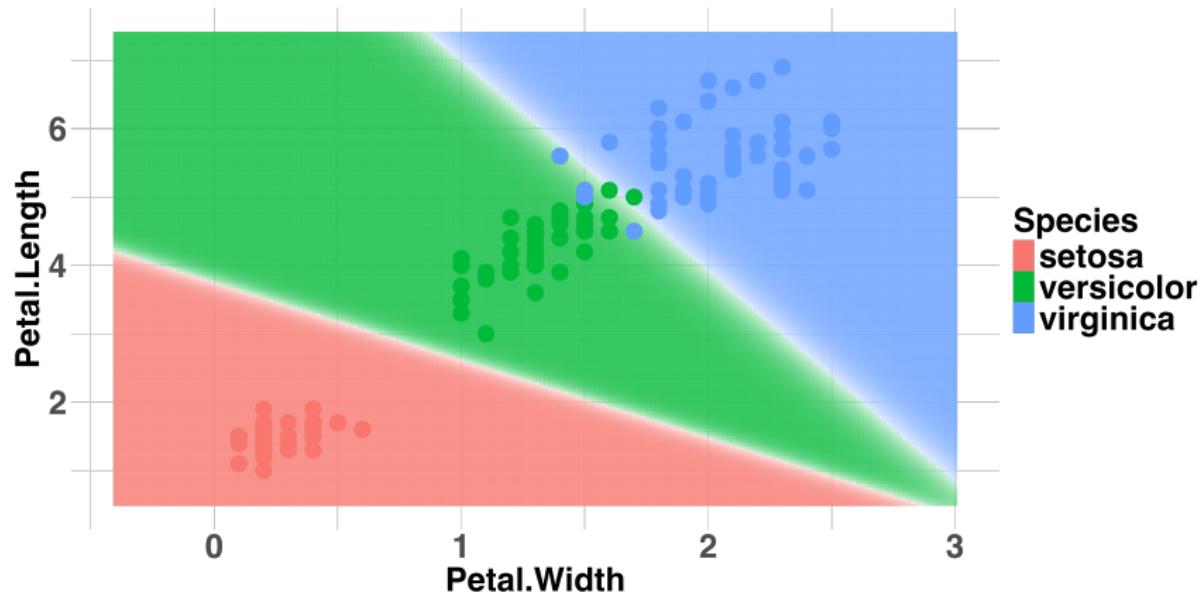
For unequal class sizes (e.g.,  $\pi_{woman} = 2\pi_{man}$ ), the prior probs matter and cause a shift of the decision boundary towards the smaller class



$$\begin{aligned}\mathbb{P}(y = \text{man} \mid \mathbf{x}) &= \frac{p(\mathbf{x} \mid y = \text{man})\pi_{\text{man}}}{p(\mathbf{x} \mid y = \text{man})\pi_{\text{man}} + p(\mathbf{x} \mid y = \text{woman})\pi_{\text{woman}}} \\ &= \frac{0.0135 \cdot \frac{1}{3}}{0.0135 \cdot \frac{1}{3} + 0.088 \cdot \frac{2}{3}} = 0.0712\end{aligned}$$

# LDA AS LINEAR CLASSIFIER

Because of the equal covariance structure of all class-specific Gaussians, the decision boundaries of LDA are always linear



## LDA AS LINEAR CLASSIFIER

Can easily prove this by showing that posteriors can be written as affine-linear functions - up to rank-preserving transformation:

$$\pi_k(\mathbf{x}) = \frac{\pi_k \cdot p(\mathbf{x}|y=k)}{p(\mathbf{x})} = \frac{\pi_k \cdot p(\mathbf{x}|y=k)}{\sum_{j=1}^g \pi_j \cdot p(\mathbf{x}|y=j)}$$



As the denominator is the same for all classes we only need to consider

$$\pi_k \cdot p(\mathbf{x}|y=k)$$

and show that this can be written as a linear function of  $\mathbf{x}$ .

# LDA AS LINEAR CLASSIFIER

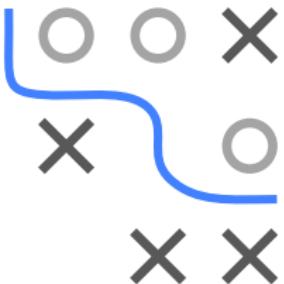
$$\begin{aligned} & \pi_k \cdot p(\mathbf{x}|y=k) \\ \propto & \pi_k \exp\left(-\frac{1}{2}\mathbf{x}^T \Sigma^{-1} \mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_k^T \Sigma^{-1} \boldsymbol{\mu}_k + \mathbf{x}^T \Sigma^{-1} \boldsymbol{\mu}_k\right) \\ = & \exp\left(\log \pi_k - \frac{1}{2}\boldsymbol{\mu}_k^T \Sigma^{-1} \boldsymbol{\mu}_k + \mathbf{x}^T \Sigma^{-1} \boldsymbol{\mu}_k\right) \exp\left(-\frac{1}{2}\mathbf{x}^T \Sigma^{-1} \mathbf{x}\right) \\ = & \exp(w_{0k} + \mathbf{x}^T \mathbf{w}_k) \exp\left(-\frac{1}{2}\mathbf{x}^T \Sigma^{-1} \mathbf{x}\right) \\ \propto & \exp(w_{0k} + \mathbf{x}^T \mathbf{w}_k) \end{aligned}$$

by defining  $w_{0k} := \log \pi_k - \frac{1}{2}\boldsymbol{\mu}_k^T \Sigma^{-1} \boldsymbol{\mu}_k$  and  $\mathbf{w}_k := \Sigma^{-1} \boldsymbol{\mu}_k$ .

By finally taking the log, we can write our transformed scores as linear:

$$f_k(\mathbf{x}) = w_{0k} + \mathbf{x}^T \mathbf{w}_k$$

- The above is a little bit “lax” so let’s carefully check
- We left out several (pos) multiplicative constants
- $\exp\left(-\frac{1}{2}\mathbf{x}^T \Sigma^{-1} \mathbf{x}\right)$  contains  $\mathbf{x}$  but is the same for all classes
- $\log(at + b)$  is still isotonic for  $a > 0$

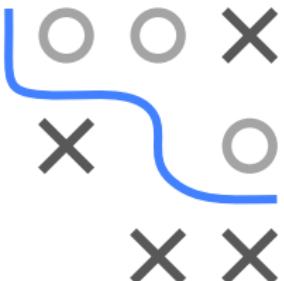
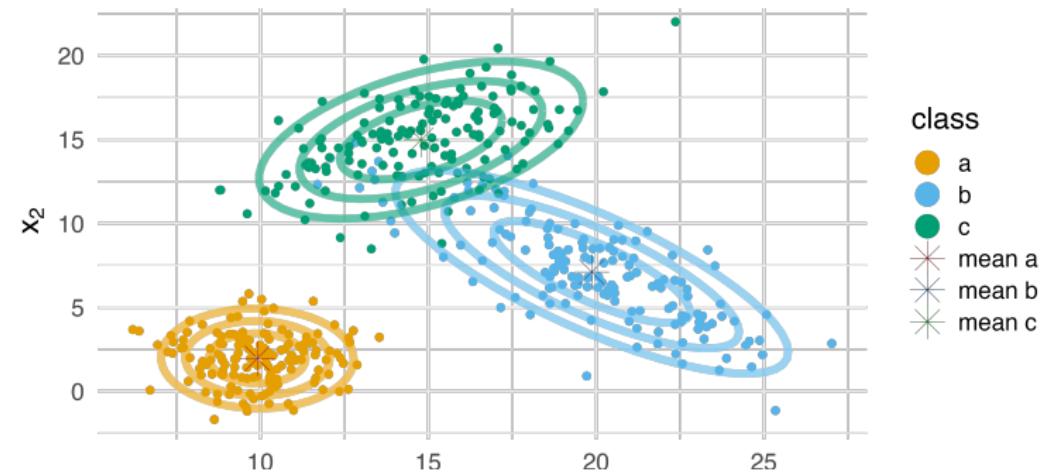


# QUADRATIC DISCRIMINANT ANALYSIS

Doesn't assume equal covariances  $\Sigma_k$  per class, so generalizes LDA:

$$p(\mathbf{x}|y = k) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp \left( -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right)$$

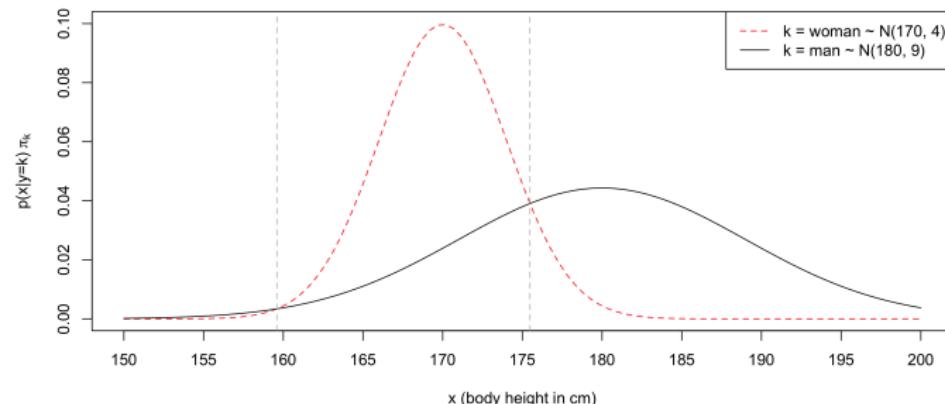
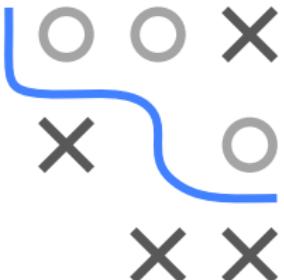
⇒ Better data fit but **requires estimation of more parameters** ( $\Sigma_k$ )!



# UNIVARIATE EXAMPLE WITH QDA

Different covariance matrices lead to multiple classification rules:

- $x < 159.6$  is being assigned to class *man*.
- $159.6 < x < 175.5$  is being assigned to class *woman*.
- $x > 175.5$  is being assigned to class *man*.



⇒ The separation function is quadratic, we learn a curved decision boundary  
(in 1D a little bit weird, as we learn an interval)

# QDA DECISION BOUNDARIES

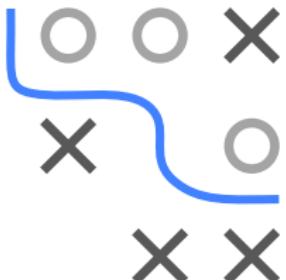
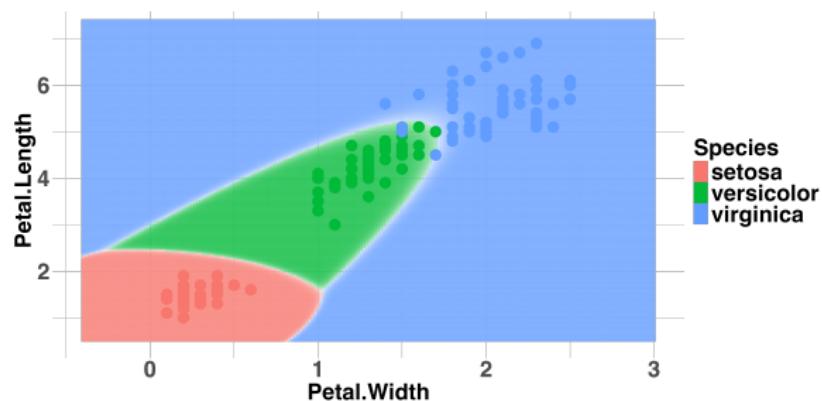
$$\pi_k(\mathbf{x}) \propto \pi_k \cdot p(\mathbf{x}|y=k)$$

$$\propto \pi_k |\Sigma_k|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \mathbf{x}^T \Sigma_k^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_k^T \Sigma_k^{-1} \boldsymbol{\mu}_k + \mathbf{x}^T \Sigma_k^{-1} \boldsymbol{\mu}_k\right)$$

Taking log, we get a quadratic discriminant function in  $x$ :

$$\log \pi_k - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2} \boldsymbol{\mu}_k^T \Sigma_k^{-1} \boldsymbol{\mu}_k + \mathbf{x}^T \Sigma_k^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \mathbf{x}^T \Sigma_k^{-1} \mathbf{x}$$

Allowing for curved decision boundaries:



# PARAMETER ESTIMATION

Parameters  $\theta$  are estimated in a straightforward manner by:

$$\hat{\pi}_k = \frac{n_k}{n}, \text{ where } n_k \text{ is the number of class-}k \text{ observations}$$

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y^{(i)}=k} \mathbf{x}^{(i)}$$

$$\hat{\Sigma}_k = \frac{1}{n_k - 1} \sum_{i:y^{(i)}=k} (\mathbf{x}^{(i)} - \hat{\mu}_k)(\mathbf{x}^{(i)} - \hat{\mu}_k)^T \quad (\text{QDA})$$

$$\hat{\Sigma} = \frac{1}{n - g} \sum_{k=1}^g \sum_{i:y^{(i)}=k} (\mathbf{x}^{(i)} - \hat{\mu}_k)(\mathbf{x}^{(i)} - \hat{\mu}_k)^T \quad (\text{LDA})$$



As  $\hat{\Sigma}_k, \hat{\Sigma}$  are  $p \times p$  matrices (for  $p$  features), estimating all  $\hat{\Sigma}_k$  involves  $\frac{p(p+1)}{2} \cdot g$  parameters across  $g$  classes (vs. just  $\frac{p(p+1)}{2}$  for LDA's  $\hat{\Sigma}$ )  
(in addition to estimating priors and class means)

# QDA PARAMETER ESTIMATION EXAMPLE

E.g., for a simple two-class, 2-dimensional dataset:

$$\text{Class 1: } \mathbf{x}_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} 2 \\ 3 \end{pmatrix}, \text{ Class 2: } \mathbf{x}_3 = \begin{pmatrix} 6 \\ 8 \end{pmatrix}, \mathbf{x}_4 = \begin{pmatrix} 7 \\ 9 \end{pmatrix}, \mathbf{x}_5 = \begin{pmatrix} 8 \\ 10 \end{pmatrix}$$

$$\text{Class priors: } \hat{\pi}_1 = \frac{n_1}{n} = \frac{2}{5} = 0.4, \quad \hat{\pi}_2 = \frac{n_2}{n} = \frac{3}{5} = 0.6$$

$$\text{Class means: } \hat{\mu}_1 = \frac{1}{2} (\mathbf{x}_1 + \mathbf{x}_2) = \begin{pmatrix} 1.5 \\ 2.5 \end{pmatrix}, \quad \hat{\mu}_2 = \frac{1}{3} (\mathbf{x}_3 + \mathbf{x}_4 + \mathbf{x}_5) = \begin{pmatrix} 7 \\ 9 \end{pmatrix}$$

Class covariances:

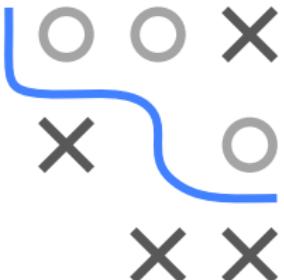
$$(\mathbf{x}_1 - \hat{\mu}_1)(\mathbf{x}_1 - \hat{\mu}_1)^\top = \begin{pmatrix} 0.25 & 0.25 \\ 0.25 & 0.25 \end{pmatrix} = (\mathbf{x}_2 - \hat{\mu}_1)(\mathbf{x}_2 - \hat{\mu}_1)^\top$$

$$\Rightarrow \hat{\Sigma}_1 = \frac{1}{1} \left( \begin{pmatrix} 0.25 & 0.25 \\ 0.25 & 0.25 \end{pmatrix} + \begin{pmatrix} 0.25 & 0.25 \\ 0.25 & 0.25 \end{pmatrix} \right) = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$$

$$(\mathbf{x}_3 - \hat{\mu}_2)(\mathbf{x}_3 - \hat{\mu}_2)^\top = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = (\mathbf{x}_5 - \hat{\mu}_2)(\mathbf{x}_5 - \hat{\mu}_2)^\top,$$

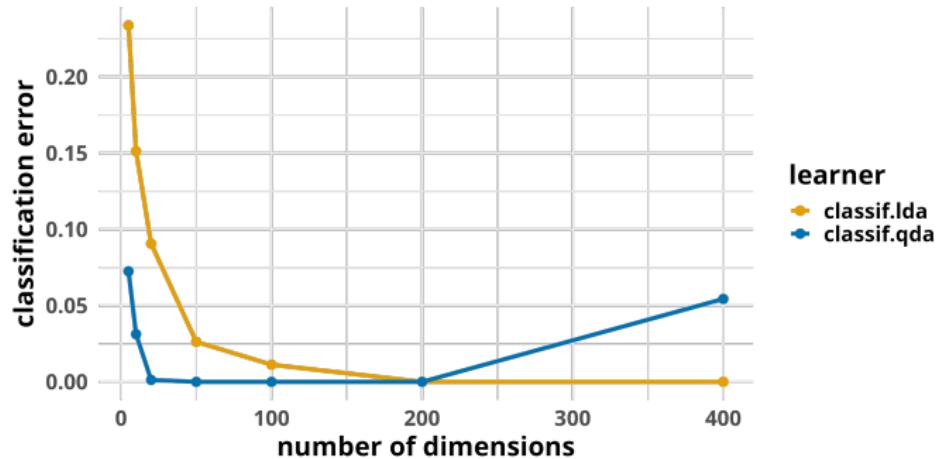
$$(\mathbf{x}_4 - \hat{\mu}_2)(\mathbf{x}_4 - \hat{\mu}_2)^\top = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

$$\Rightarrow \hat{\Sigma}_2 = \frac{1}{2} \left( \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \right) = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$



# DISCRIMINANT ANALYSIS COMPARISON

- We benchmark on simple toy data set(s)
- Normally distributed data per class, but unequal cov matrices
- And then increase dimensionality
- We might assume that QDA always wins here ...



⇒ LDA might be preferable over QDA in higher dimensions!