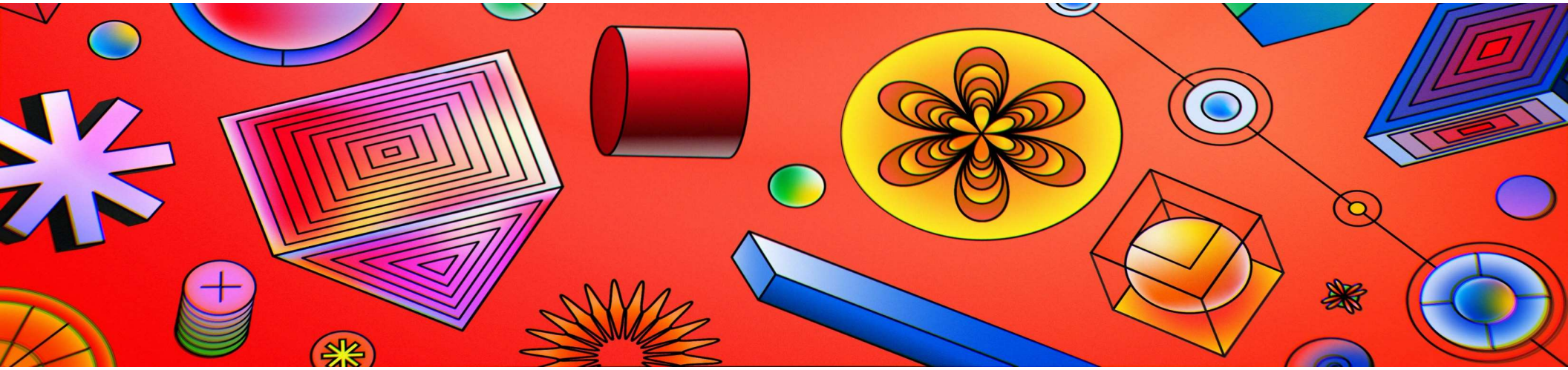
Fall 2023

BIF524/CSC463 Data Mining

Classification

Eileen Marie Hanna, *PhD*

12/10/2023



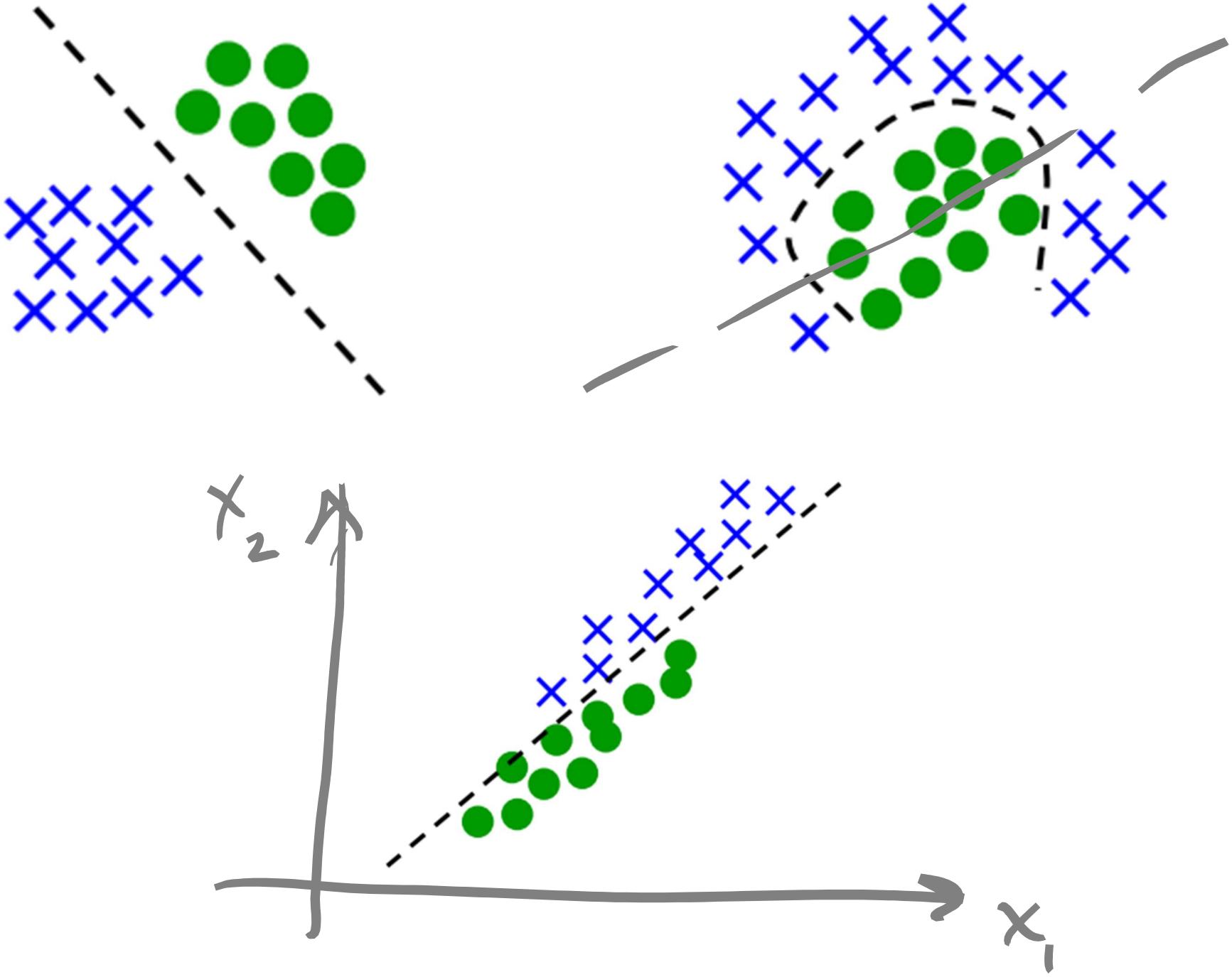
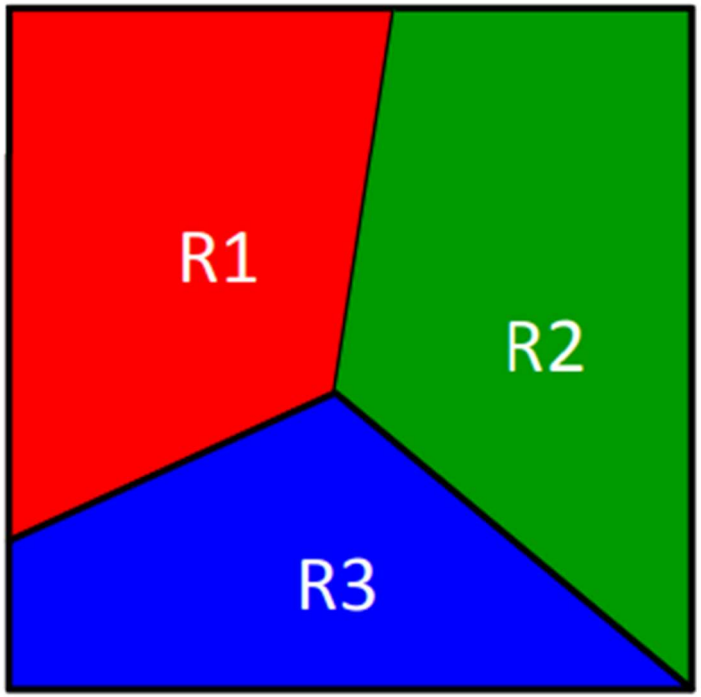


Fig. from R. Gutierrez-Osuna



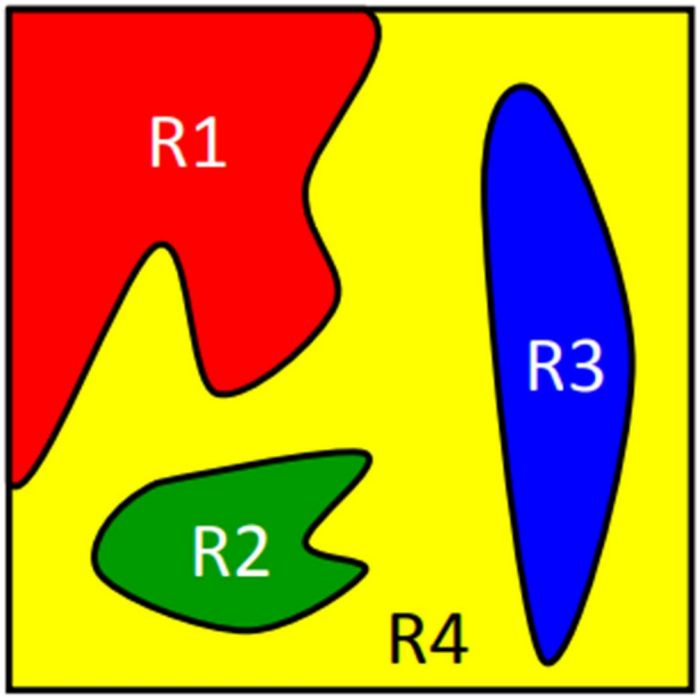
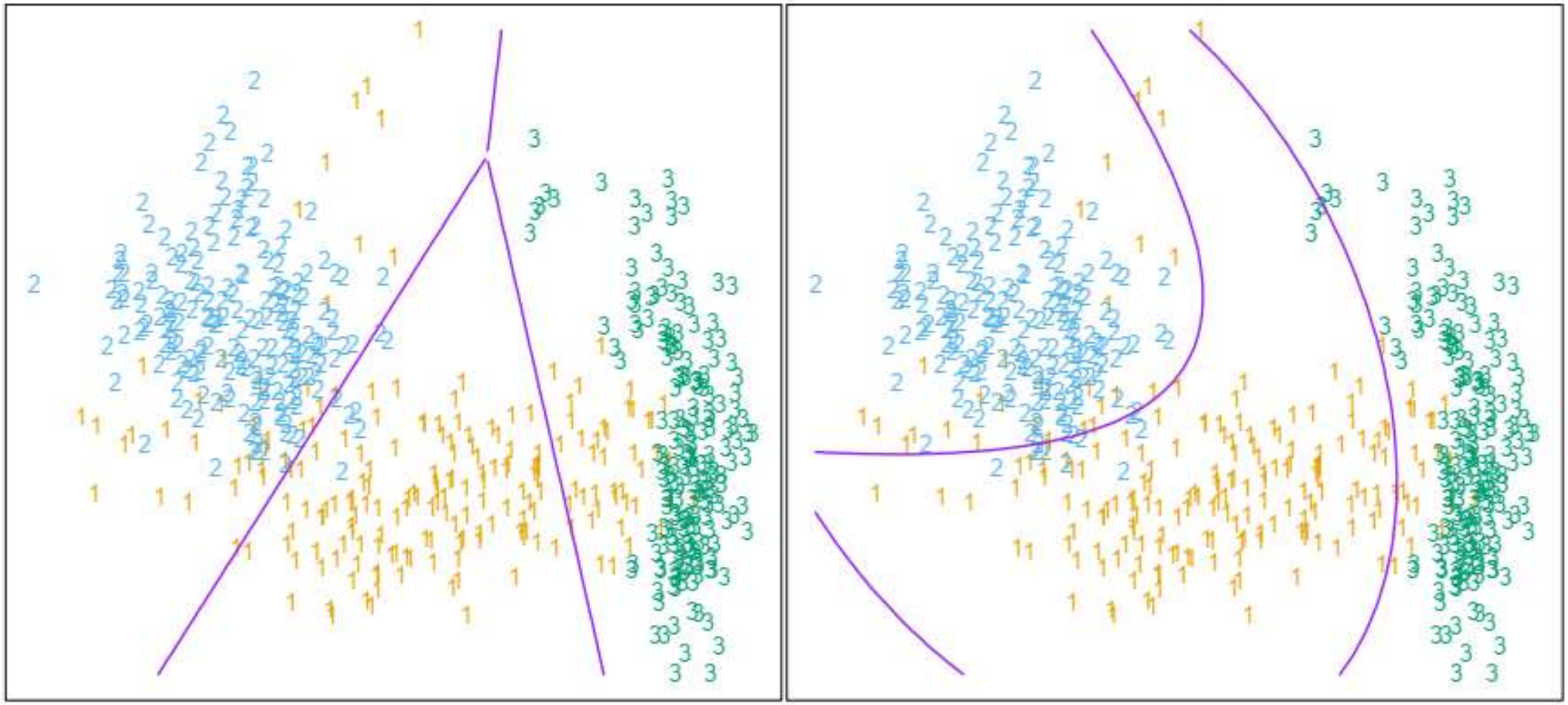


Fig. from R. Gutierrez-Osuna



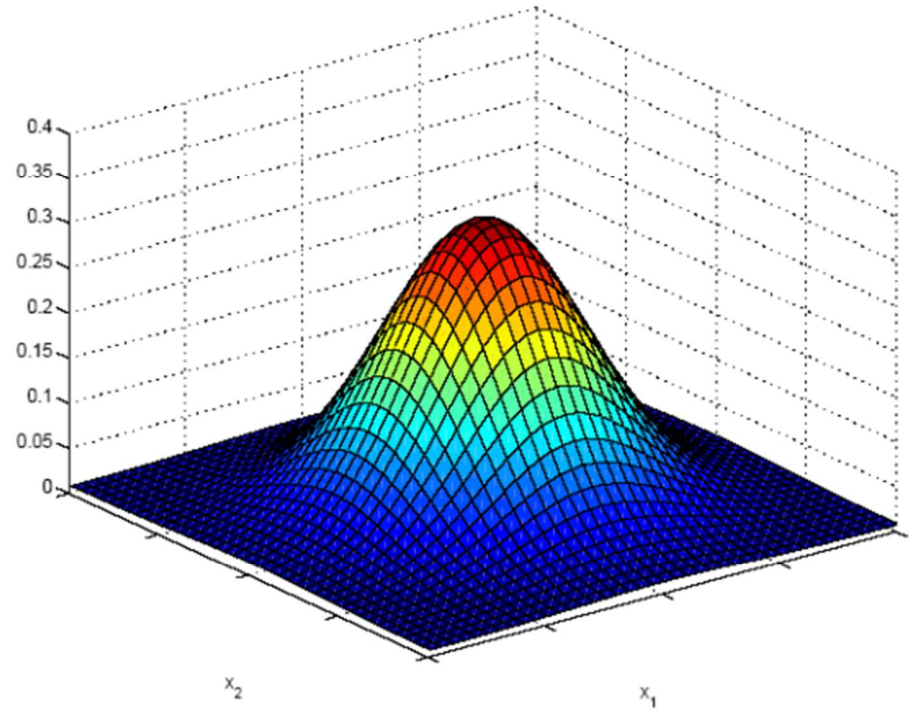
$$\mathbb{E}[f(x)] \xrightarrow{\text{discrete}} = \sum_x p(x) f(x)$$

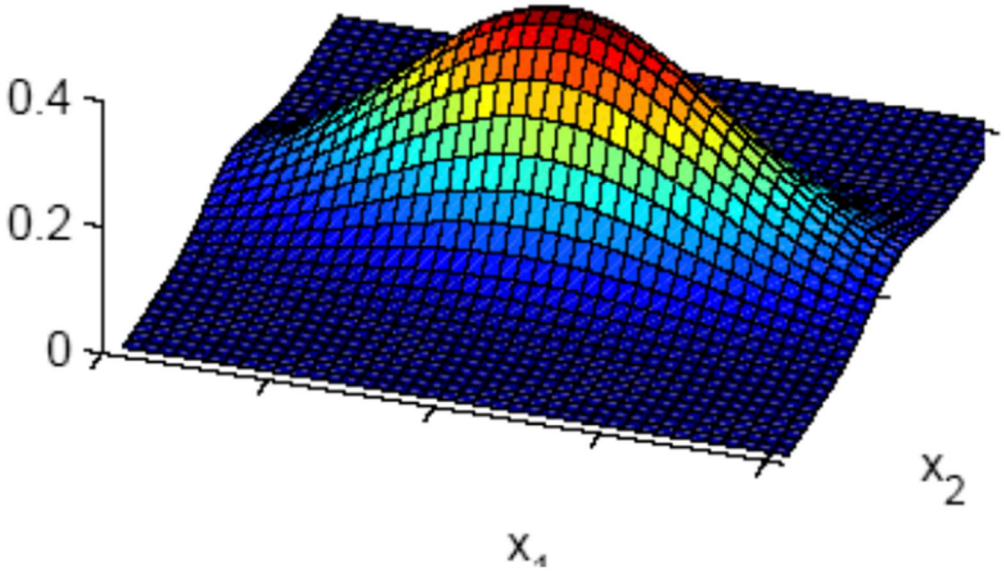
$$\xrightarrow{\text{cont.}} = \int p(x) f(x) \, dx$$

$$\mathrm{Var}[x] = \mathbb{E}\left[(x-\mu)^2\right] \quad \text{where } \mu = \mathbb{E}[x]$$

$$\simeq \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

$$\mathrm{Cov}[x_1, x_2] = \mathbb{E}_{x,y}\left[(x - \mathbb{E}[x_1])\{(y - \mathbb{E}[x_2])\right]$$





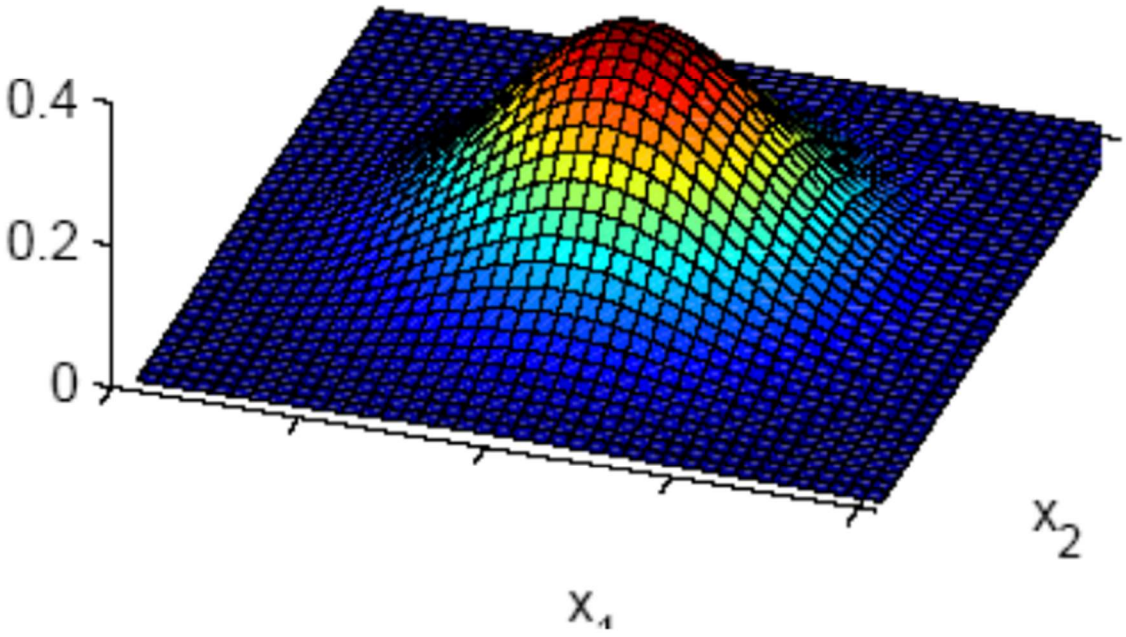


Fig. from R. Gutierrez-Osuna

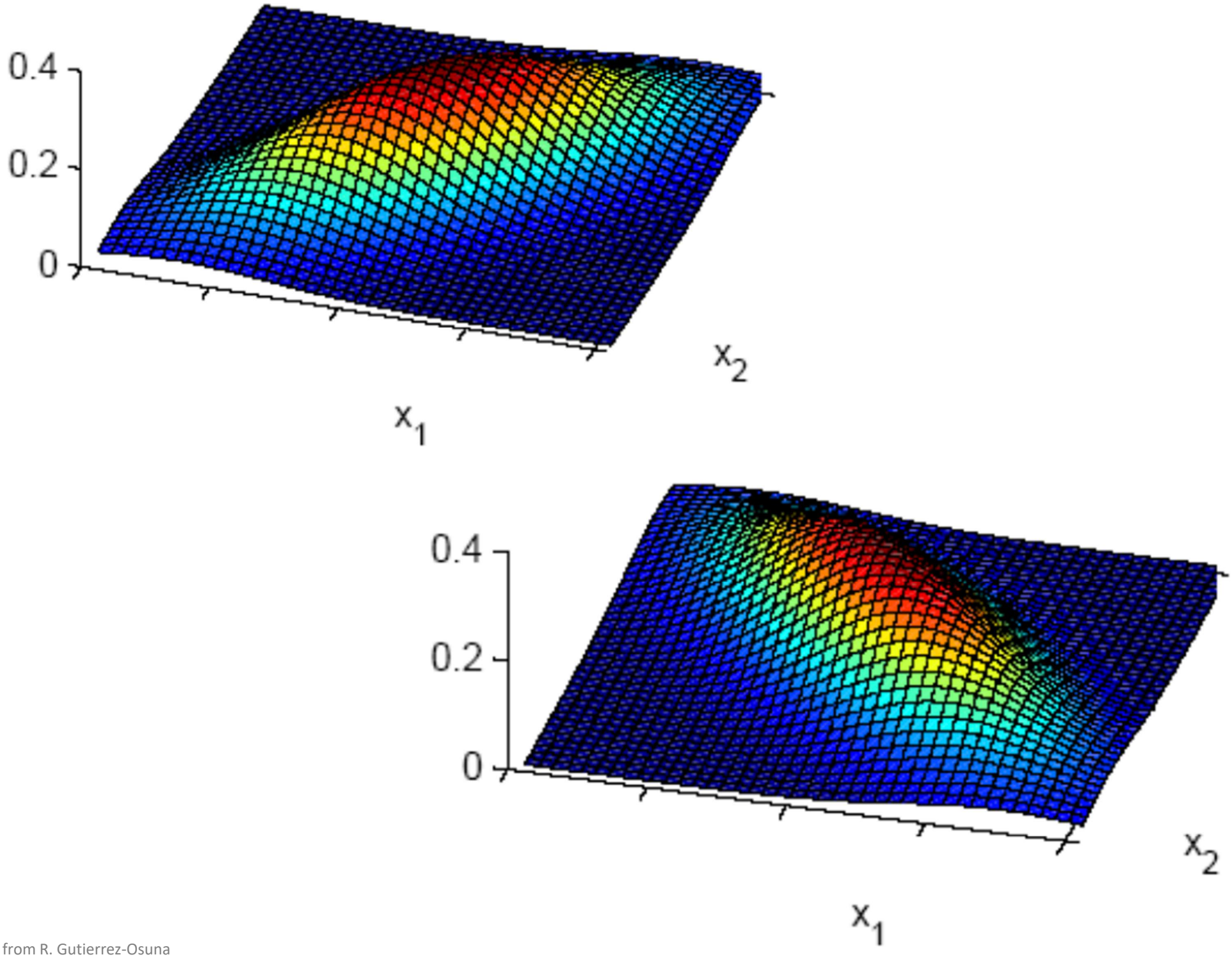


Fig. from R. Gutierrez-Osuna

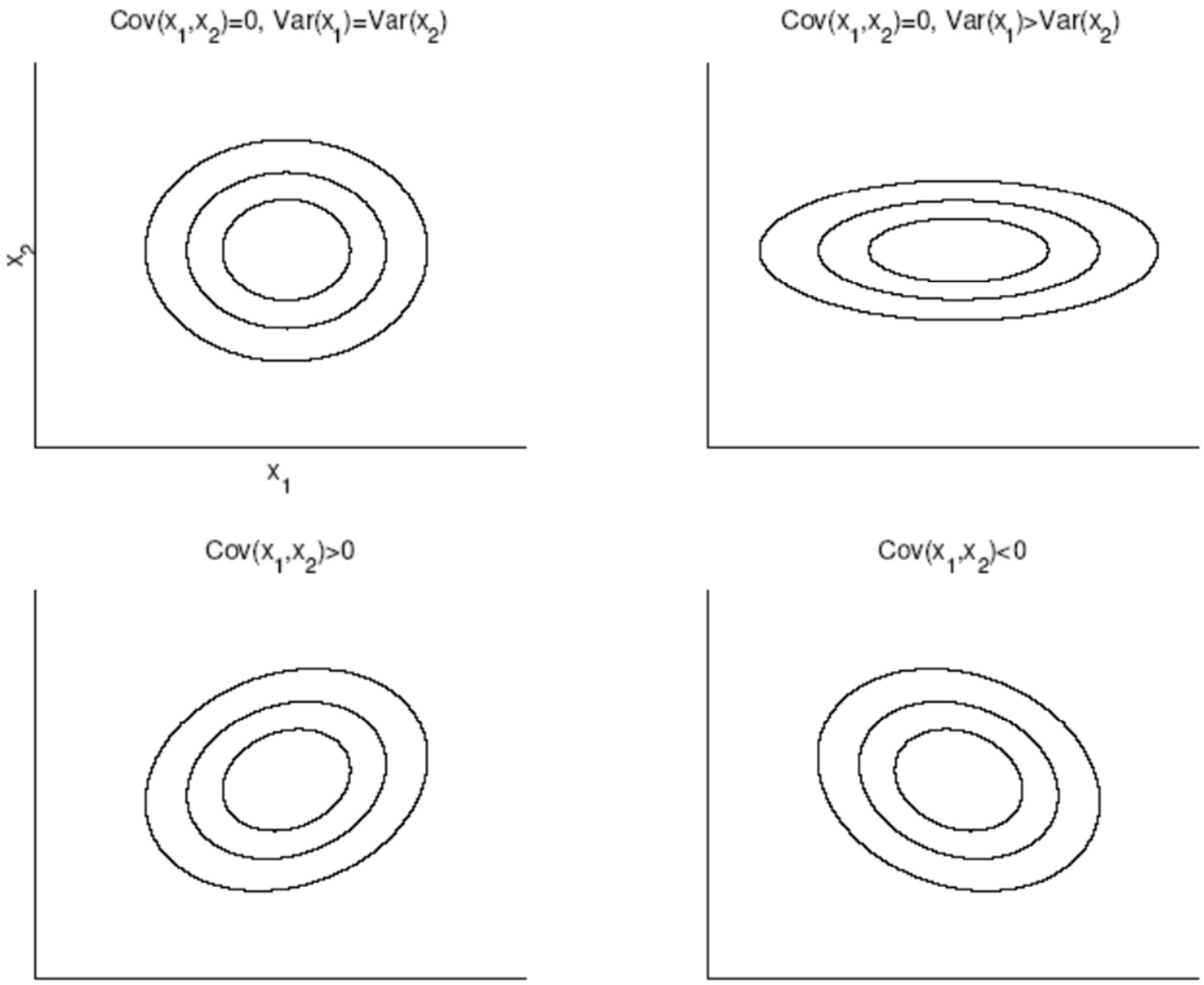
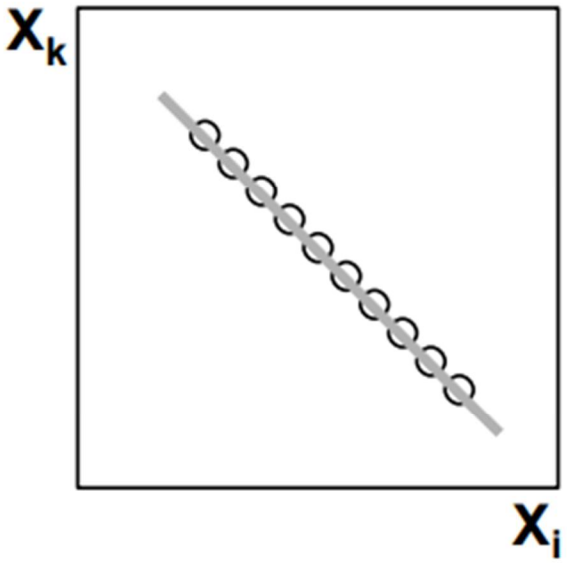
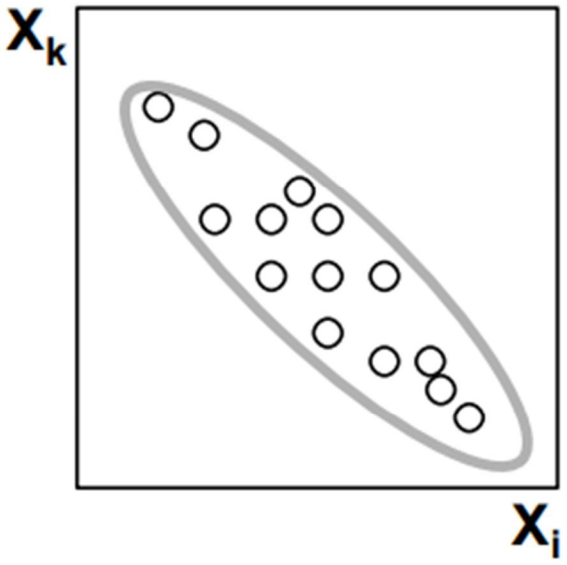


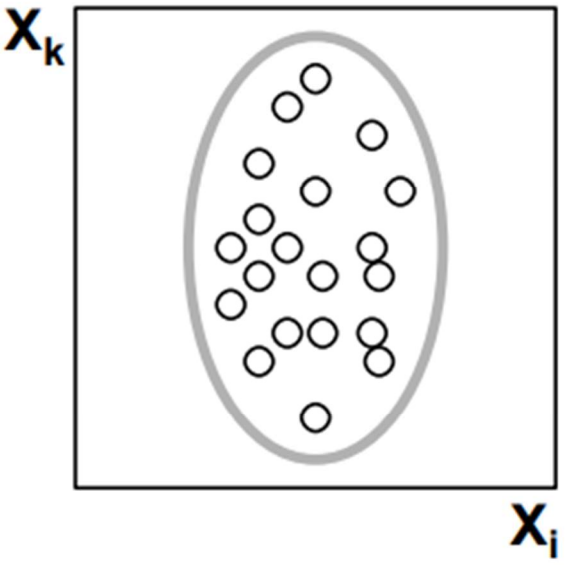
Fig. from R. Gutierrez-Osuna



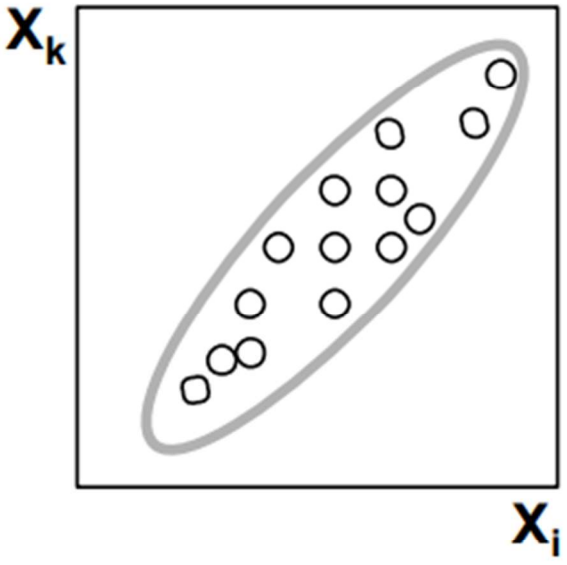
$C_{ik}=-\sigma_i\sigma_k$
$\rho_{ik}=-1$



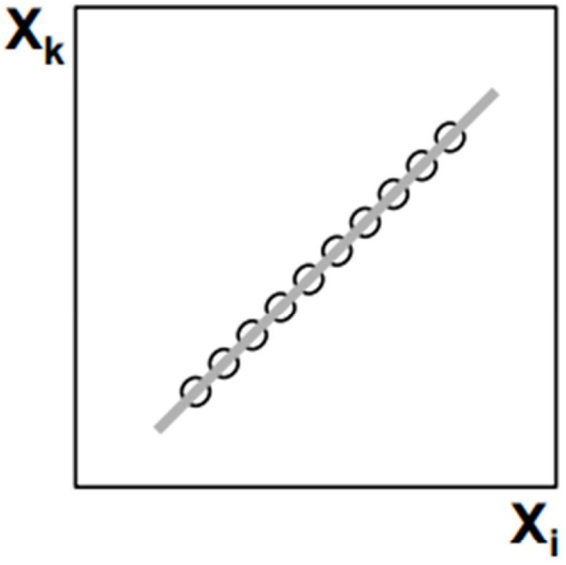
$C_{ik}=-\frac{1}{2}\sigma_i\sigma_k$
$\rho_{ik}=-\frac{1}{2}$



$C_{ik}=0$
$\rho_{ik}=0$

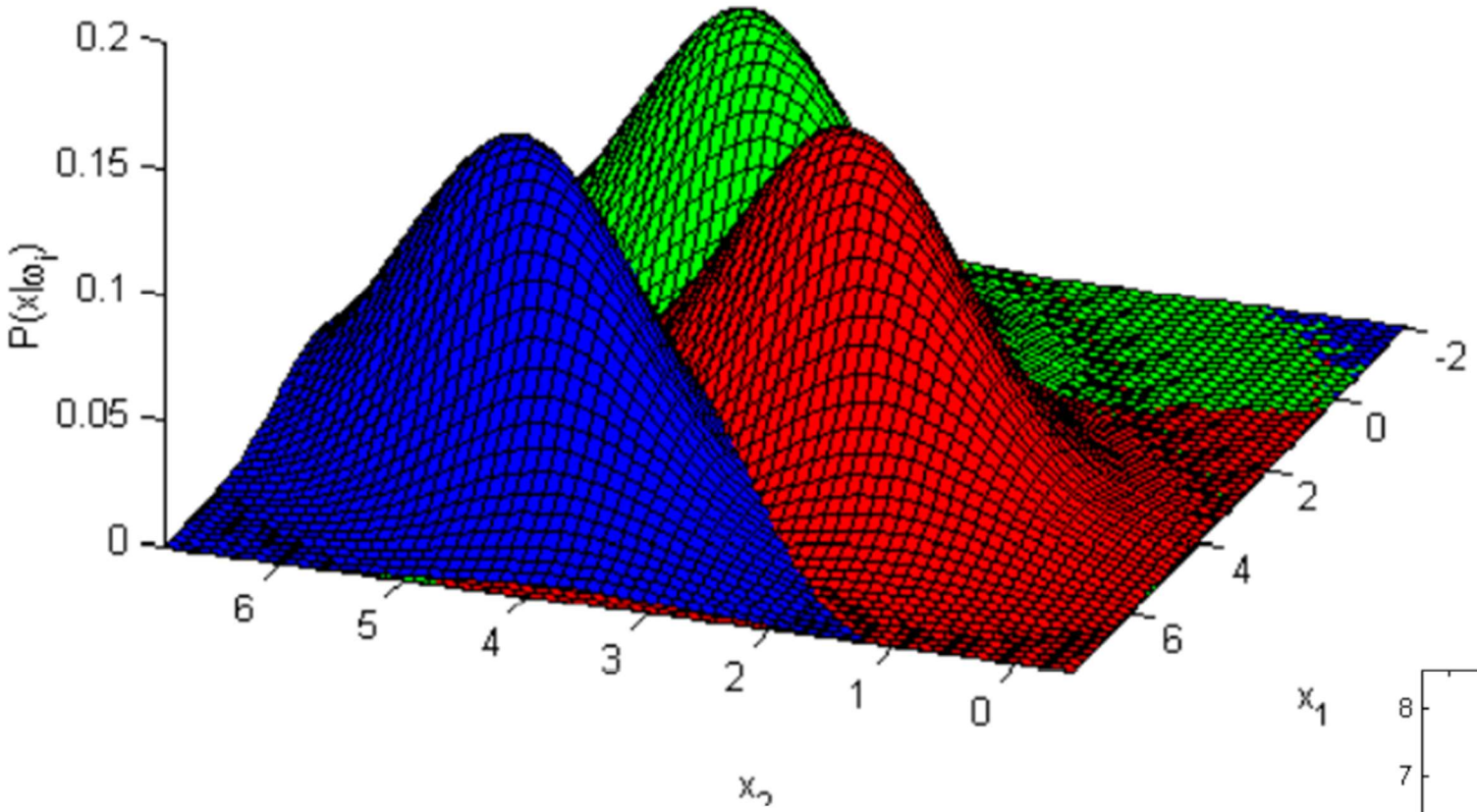


$C_{ik}=+\frac{1}{2}\sigma_i\sigma_k$
$\rho_{ik}=+\frac{1}{2}$



$C_{ik}=\sigma_i\sigma_k$
$\rho_{ik}=+1$

Fig. from R. Gutierrez-Osuna



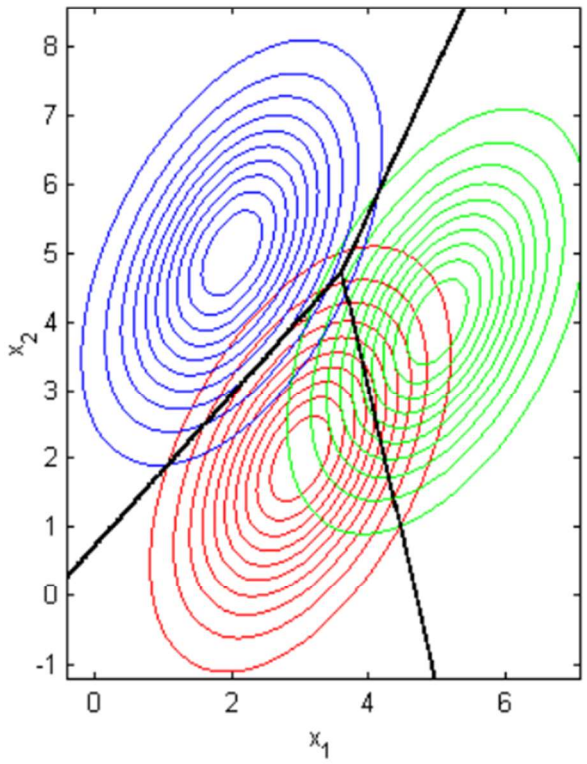
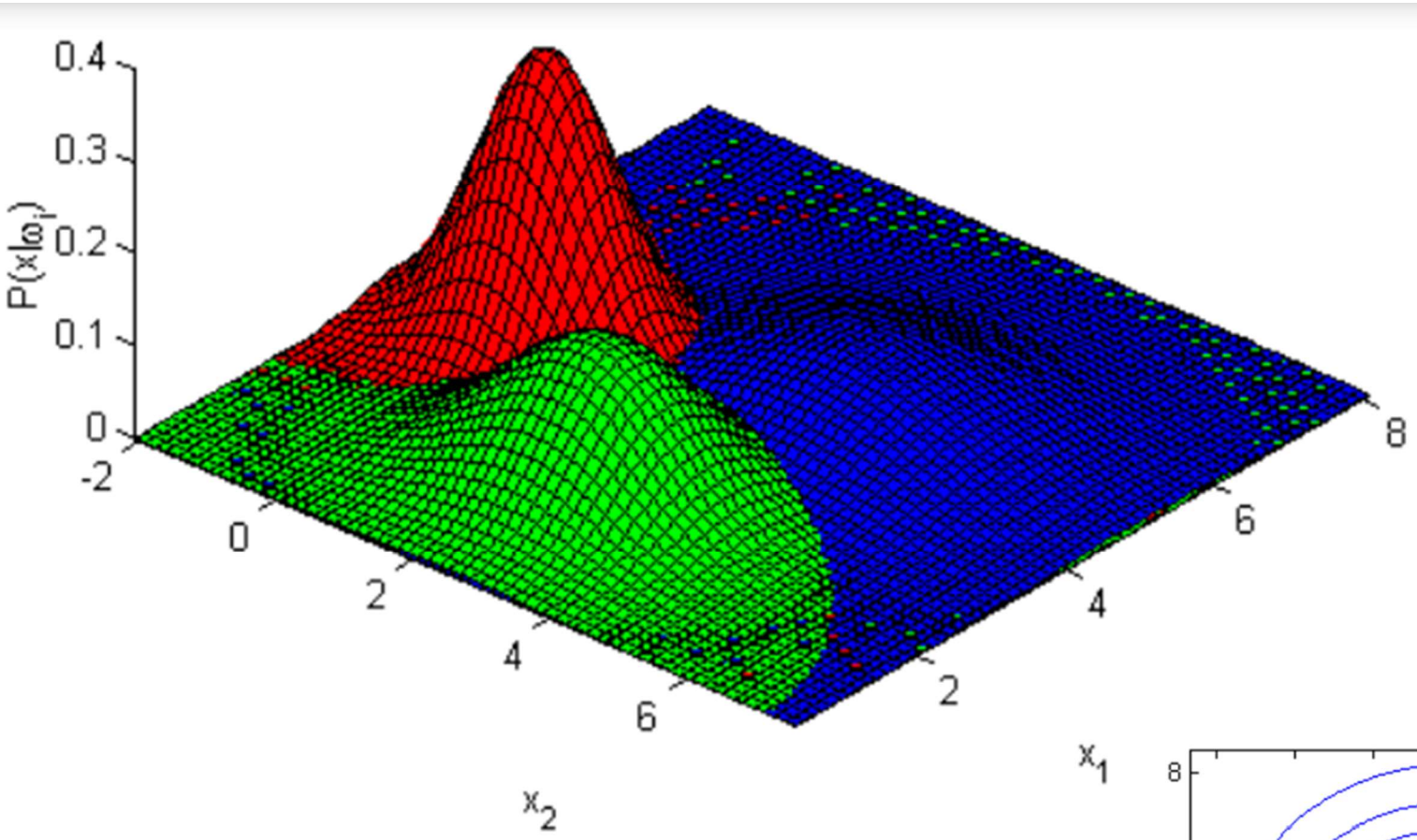


Fig. from R. Gutierrez-Osuna



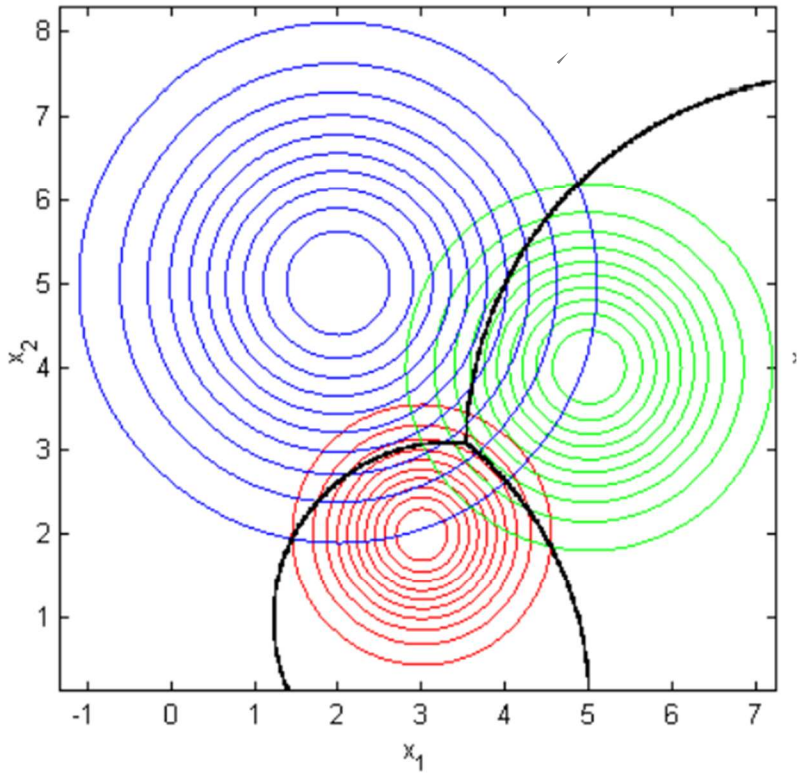


Fig. from R. Gutierrez-Osuna

Linear Discriminant Analysis (LDA)

- A **popular** approach when we have a response with **more than two classes**.
- **It separately models the distribution of predictors in each response class**.
    - And **then it applies the Bayes' theorem** to generate estimates for $\Pr(Y = y|X = x)$.
- It is **more stable** than logistic regression **when** the **classes are well separated**.
- It is also more stable **when the number of observations is small, and the distribution of predictors is approx. normal**.

The Bayes' theorem in the context of classification

- Let K be the set of unordered and distinct classes that a qualitative response can take.

$$\underbrace{\mathbf{Pr}(\boldsymbol{Y}=\boldsymbol{k}|\boldsymbol{X}=\boldsymbol{x})}=\frac{\mathbf{Pr}(\boldsymbol{X}=\boldsymbol{x}|\boldsymbol{Y}=\boldsymbol{k})\times\mathbf{Pr}(\boldsymbol{Y}=\boldsymbol{k})}{\mathbf{Pr}(\boldsymbol{X}=\boldsymbol{x})}$$

$Pr(Y=default|X=\{25,1,15\text{K}\})=0.214$

$Pr(Y=nonrdefault|X=\{25,1,15\text{K}\})=0.315$

The Bayes' theorem in the context of classification

- Let K be the set of unordered and distinct classes that a qualitative response can take.

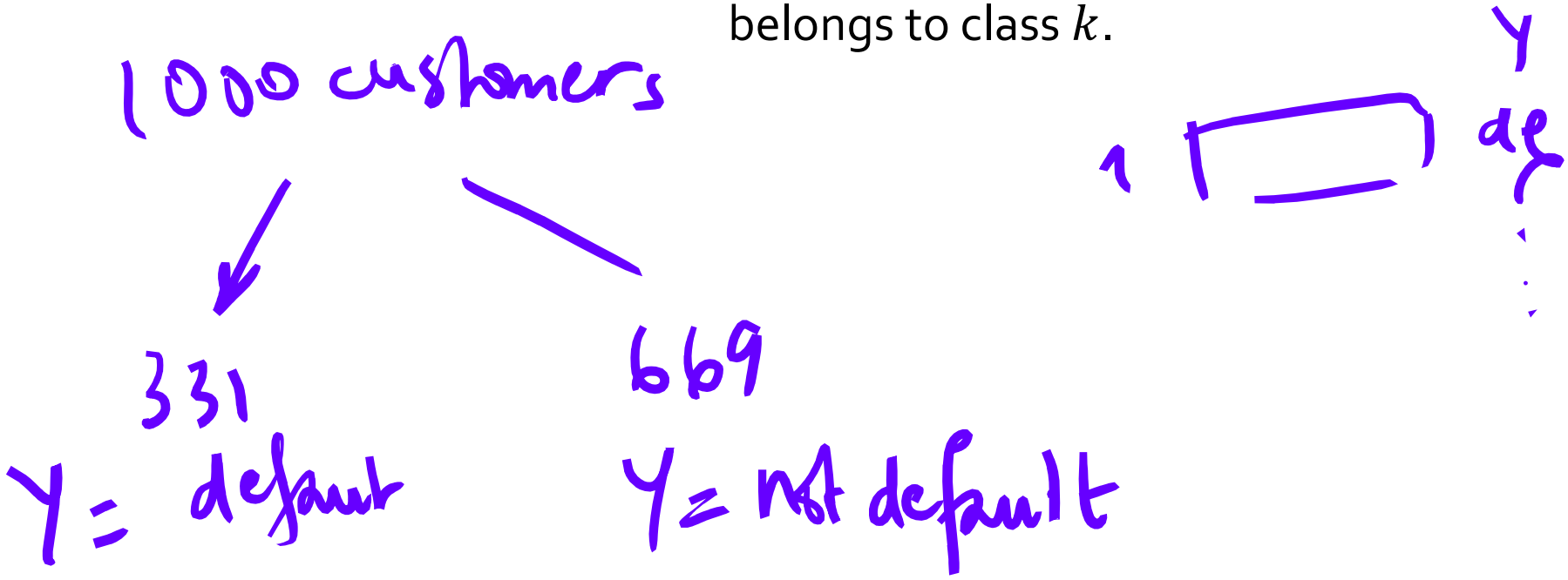
$$\mathbf{Pr}(\boldsymbol{Y = k | X = x}) = \frac{\mathbf{Pr}(\boldsymbol{X = x | Y = k}) \times \mathbf{Pr}(\boldsymbol{Y = k})}{\mathbf{Pr}(\boldsymbol{X = x})}$$

The Bayes' theorem in the context of classification

- Let K be the set of unordered and distinct classes that a qualitative response can take.

$$\Pr(Y = k|X = x) = \frac{\Pr(X = x|Y = k) \times \Pr(Y = k)}{\Pr(X = x)}$$

Let π_k be the **prior probability** that a randomly chosen observation belongs to class k.



The Bayes' theorem in the context of classification

- Let K be the set of unordered and distinct classes that a qualitative response can take.

$$\mathbf{Pr}(\boldsymbol{Y = k | X = x}) = \frac{\mathbf{Pr}(\boldsymbol{X = x | Y = k}) \times \mathbf{Pr}(\boldsymbol{Y = k})}{\mathbf{Pr}(\boldsymbol{X = x})}$$

$f_k(x)$ is the **density function of $\boldsymbol{X}$ for an observation that belongs to the $\boldsymbol{k^{th}}$ class,**

i.e., $f_k(x) \equiv \Pr(X = x | Y = k)$

$f_k(x)$ is **high when there is a relatively high probability that an observation in class $\boldsymbol{k}$ has $\boldsymbol{X \approx x}$**. It is very small otherwise.

	Bal	Inc	Student	Ethnicity	...	Default
(1)	10k	100k	Y	1	Y	
(2)	2k	50k	N	2	N	Yes / No
⋮	⋮	⋮	⋮	⋮	⋮	⋮
(1000)						

$$Pr(Y = Yes)$$

$$Pr(X = x \mid Y = k)$$

$$Pr\left(\{Bal = 10k, Inc = 50k, Stud. = N, Eth. = 2, \ldots\} \mid Y = No\right)$$

The Bayes' theorem in the context of classification

- Let K be the set of unordered and distinct classes that a qualitative response can take.

$$\mathbf{Pr}(\boldsymbol{Y = k | X = x}) = \frac{\mathbf{Pr}(\boldsymbol{X = x | Y = k}) \times \mathbf{Pr}(\boldsymbol{Y = k})}{\mathbf{Pr}(\boldsymbol{X = x})}$$

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

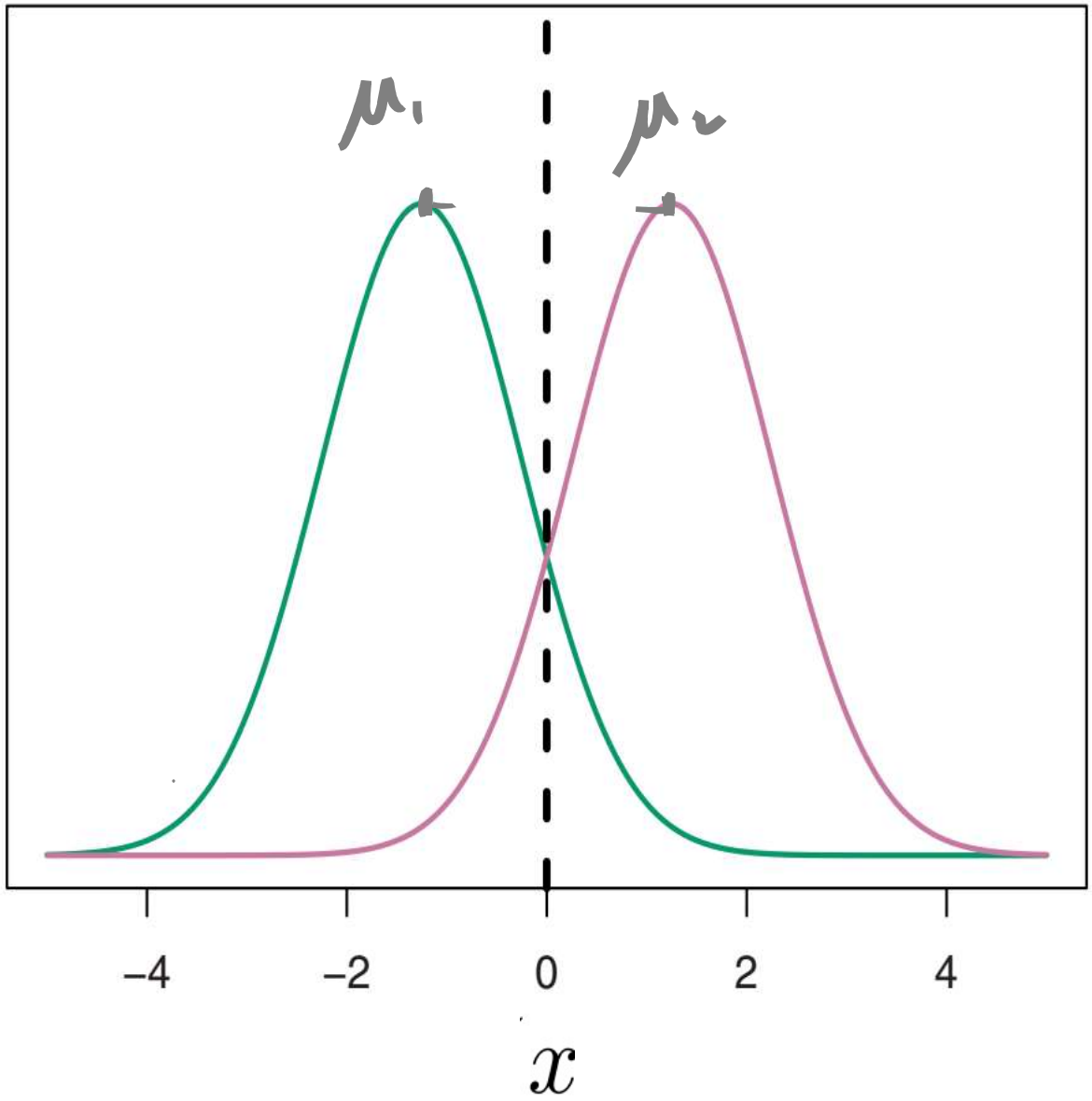
  - **<u>Key idea:</u>** We need an **estimate for $\boldsymbol{f_k(x)}$** that we can use in order **to estimate $\boldsymbol{p_k(x)}$**.

    - We can then **classify** an observation to the class for which $\boldsymbol{p_k(x)}$ **is greatest**.
    - In order to **estimate $\boldsymbol{f_k(x)}$**, we first need to **make** some **assumptions about its form**, e.g., normal.

σ_1

σ_2

$\pi_k f_k(x)$



The Bayes' theorem in the context of classification

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

- e.g., if we assume that $f_k(x)$ is **normal** or Guassian,

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)$$

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_l)^2\right)}$$

- Assuming that variance is constant across all K classes.

The Bayes' theorem in the context of classification

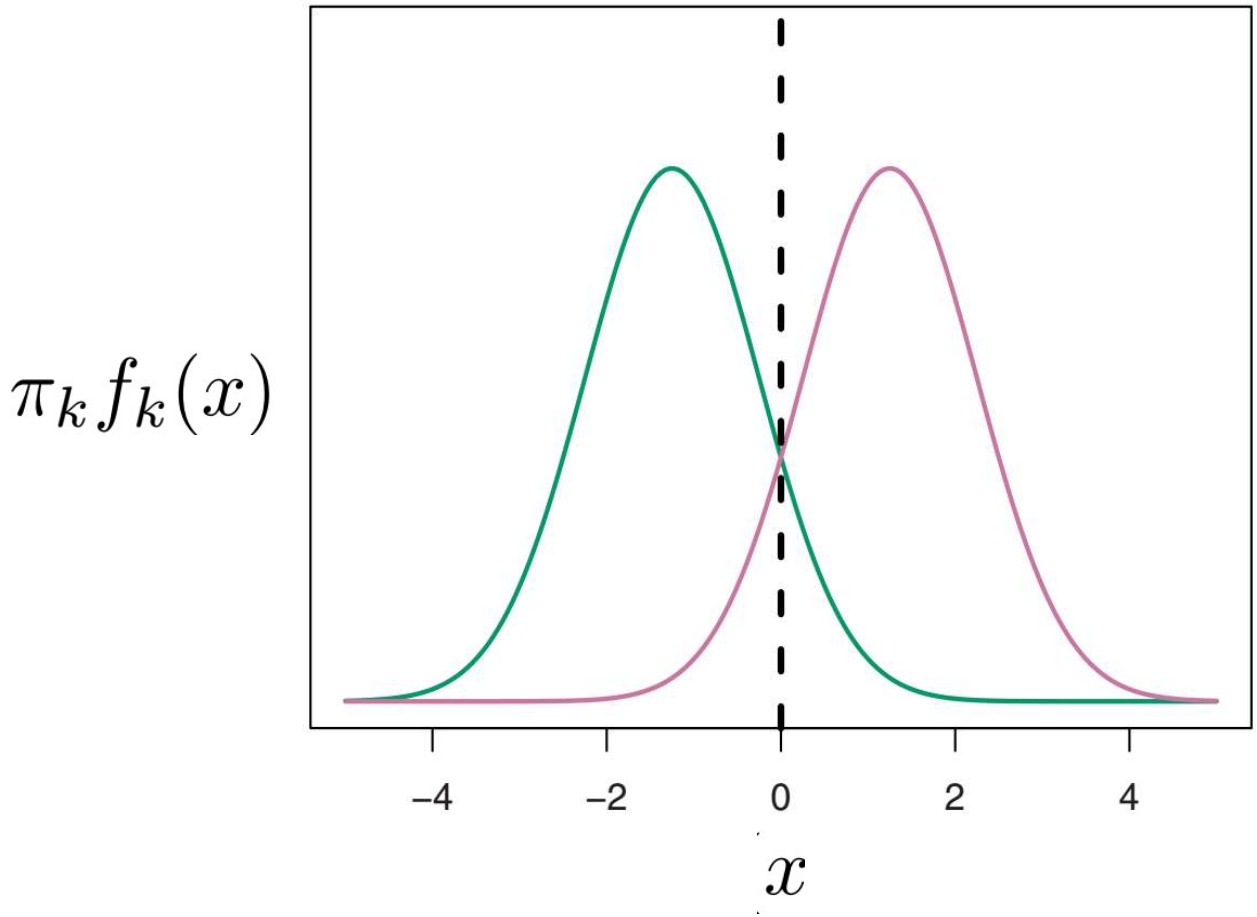
- First, using a single predictor (i.e., $p = 1$), and assuming that $\Pr(X = x|Y = y)$ is normal:
  - The **decision boundary** corresponds to the **midpoint between the sample means**:

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}$$

The Bayes' theorem in the context of classification

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}$$

- For instance, if $K = 2$ and $\Pr(Y = Class_1) = \Pr(Y = Class_2) = 0.5$



$$\mu_1 = -1.25,\ \mu_2 = 1.25,\ \text{and}\ \sigma_1^2 = \sigma_2^2 = 1$$

The Bayes' theorem in the context of classification

- In **real cases**, we are **not able to calculate the Bayes classifier**, even if we are sure that, for instance, the distribution is normal within each class.

- **LDA estimates the Bayes classifier by estimating π_k, μ_k, and σ^2.**

average of all training observations in class k

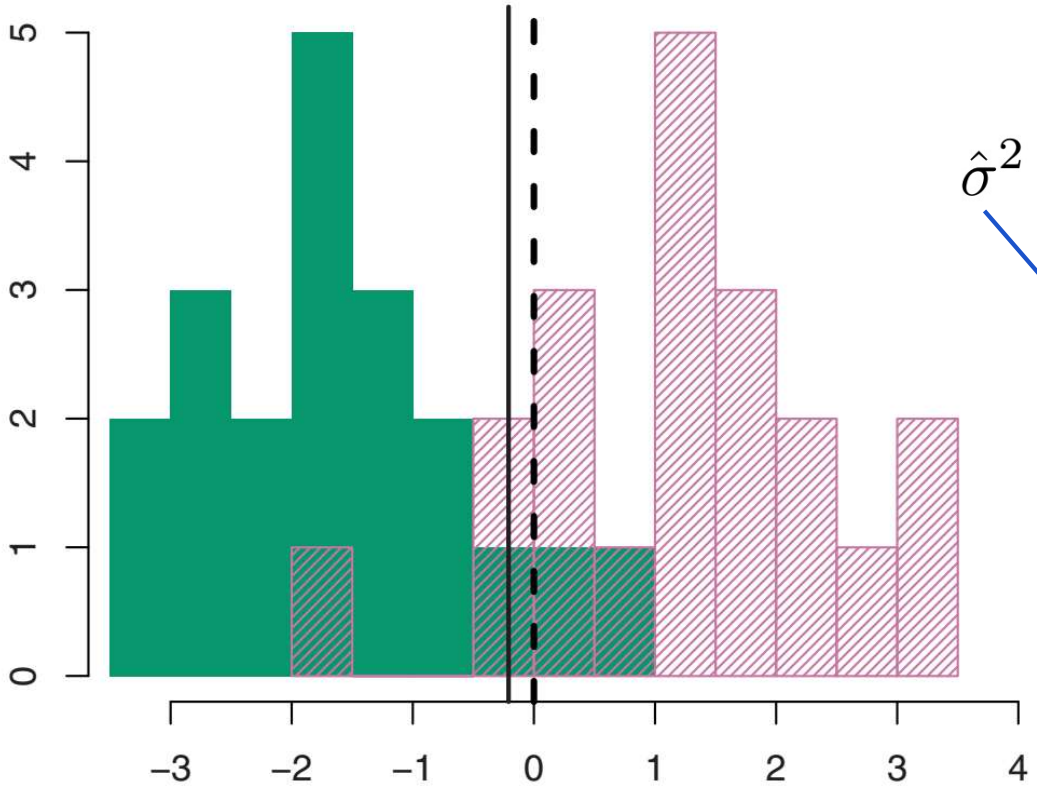
number of training observations in class k

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

$$\hat{\pi}_k = n_k/n$$

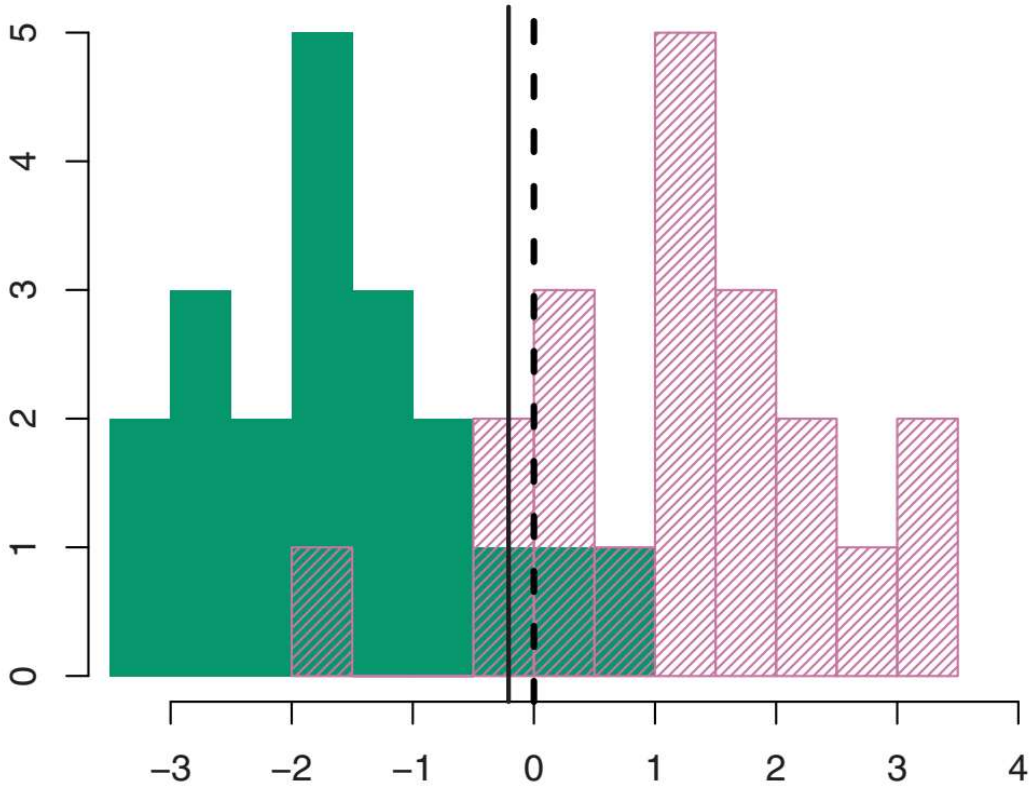
weighted average of the sample variances for each of the K classes



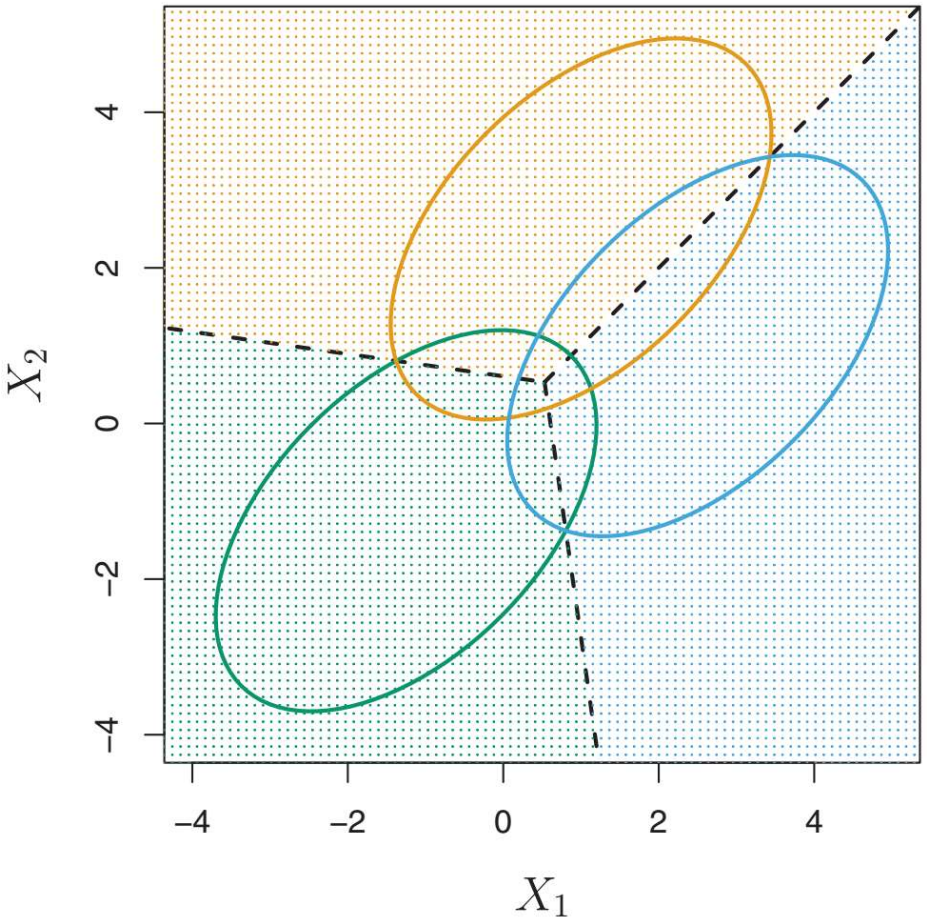
The Bayes' theorem in the context of classification

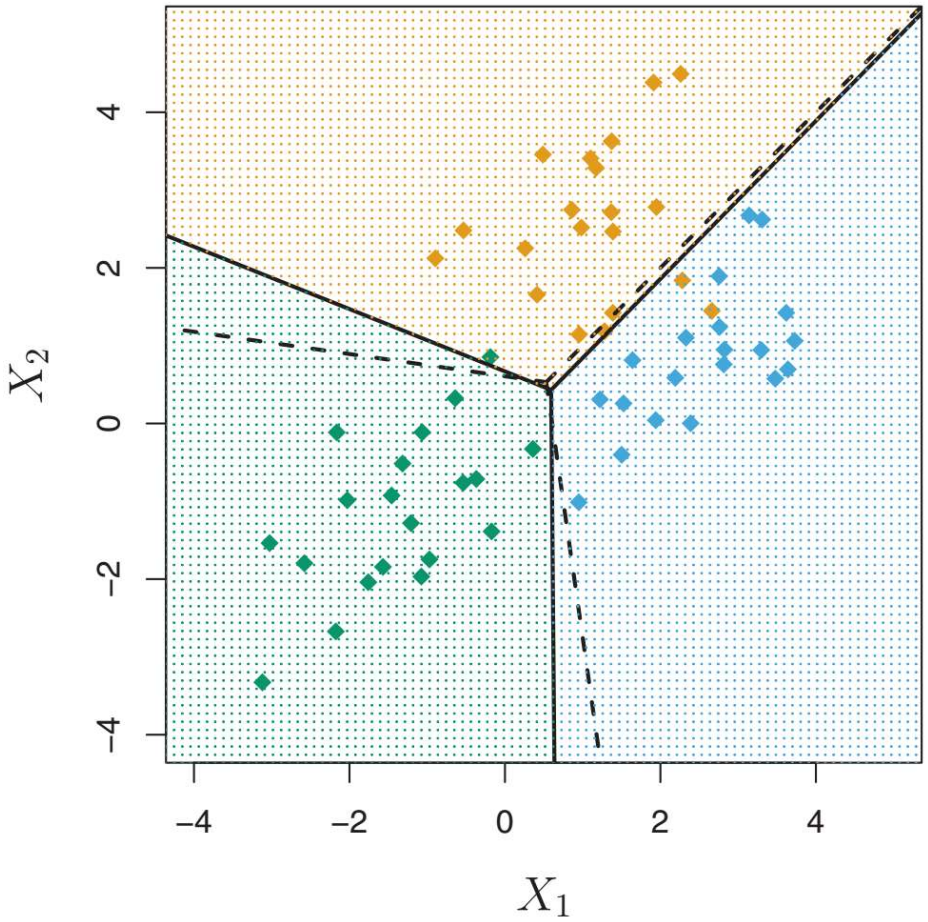
- LDA assigns x to the class for which the discriminant function is the largest:

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$



Observations drawn from a multivariate Gaussian distribution





- **Ellipses contain 95% of the probability for each class.**
- **Full black lines are the Bayes decision boundaries,** assuming that we know them. **Usually, we don't!**
- **LDA decision boundaries as dashed lines.**

How good is an LDA model fit?

- Consider LDA applied on the Default data of 10,000 observations, with credit card balance and student as predictors.
- The model gives a training error of 2.75%.

- **Two types of errors:**

  - incorrectly assign a person who defaulted to did not default class
  - incorrectly assign a person who did not default to defaulted class

Confusion matrix showing predictions

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

Sensitivity and Specificity

- Sensitivity is the percentage of true positives – i.e., true positive rate.
  - Here, the percentage of true defaulters that are identified.

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

Sensitivity and Specificity

- Specificity is the percentage of true negatives – i.e., true negative rate.
  - Here, the percentage of non-defaulters that are correctly identified.

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

LDA performance

- As we saw earlier, the Bayes classifier gives the lowest number of misclassifications.
    - Such misclassifications could be in **erroneously classifying a person as not default class or default class**.
- **The tolerance of errors depends on the purpose of the classification.**
    - e.g., for the Default data, a credit card company typically wants to know which customers have a default risk -> need to correctly classify persons who will default, with possible tolerance for errors in the not default class.
- **How can a classifier accommodate such assumptions?**

LDA performance

- The Bayes classifier assigns an observation to the class with the highest posterior probability.

$$\mathbf{Pr}(\boldsymbol{Y} = \boldsymbol{k}|\boldsymbol{X} = \boldsymbol{x}) = \frac{\mathbf{Pr}(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{Y} = \boldsymbol{k}) \times \mathbf{Pr}(\boldsymbol{Y} = \boldsymbol{k})}{\mathbf{Pr}(\boldsymbol{X} = \boldsymbol{x})}$$

- For the two-class response, a customer is classified as Default if:

$$\Pr(\texttt{default} = \texttt{Yes}|X = x) > 0.5$$

we can consider a stricter threshold, here a lower threshold, e.g.,

$$\Pr(\texttt{default} = \texttt{Yes}|X = x) > 0.2$$

LDA performance - Default

$$\Pr(\texttt{default} = \texttt{Yes} | X = x) > 0.5$$

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

persons erroneously predicted as not default (~75.7%)

$$\Pr(\texttt{default} = \texttt{Yes} | X = x) > 0.2$$

		True default status		
		No	Yes	Total
Predicted default status	No	9,432	138	9,570
	Yes	235	195	430
	Total	9,667	333	10,000

persons will default

persons erroneously predicted as not default (~41.4%)

LDA performance - Default

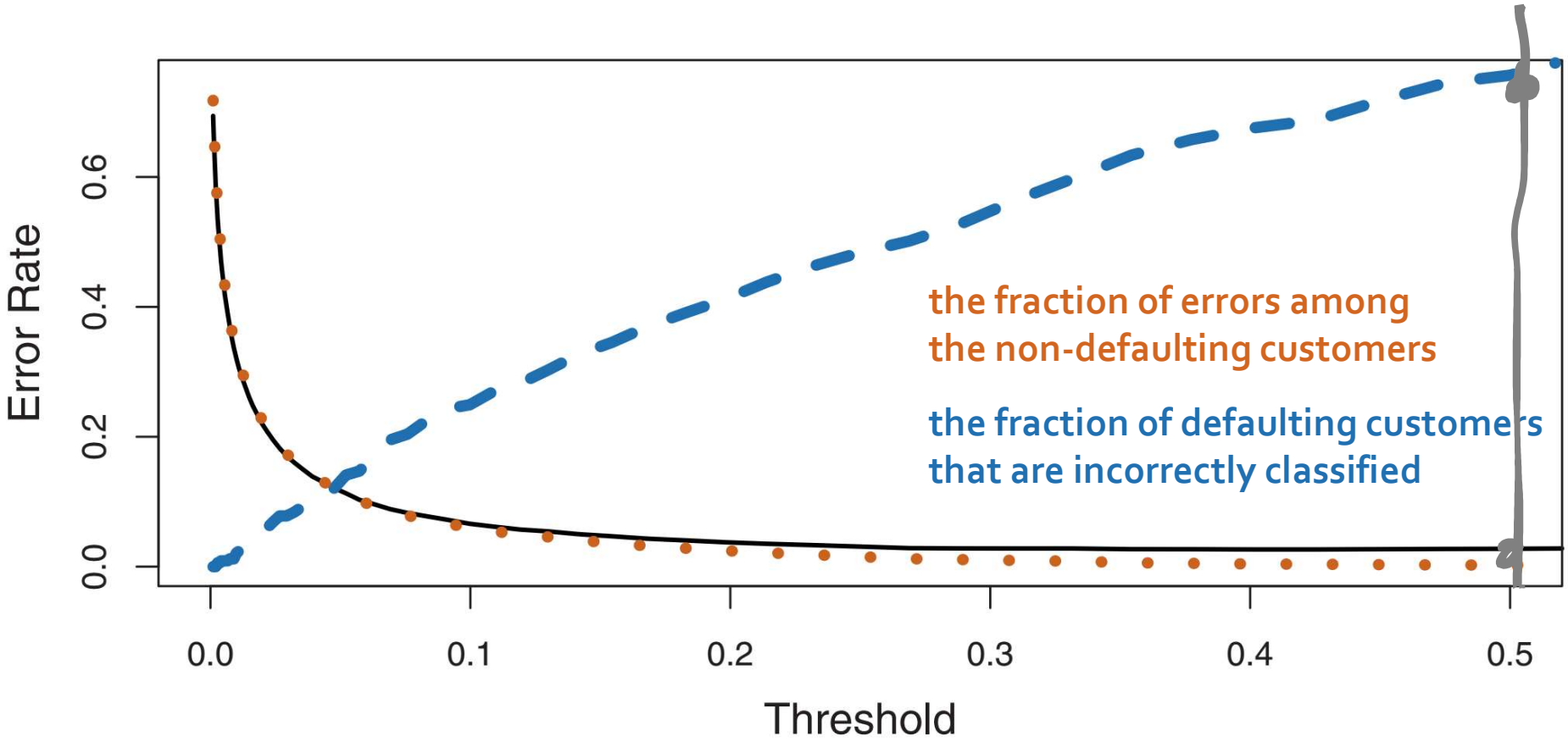
$$\Pr(\texttt{default} = \texttt{Yes} | X = x) > 0.5$$

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

$$\Pr(\texttt{default} = \texttt{Yes} | X = x) > 0.2$$

		True default status		
		No	Yes	Total
Predicted default status	No	9,432	138	9,570
	Yes	235	195	430
	Total	9,667	333	10,000

Error rates as a function of the threshold value of the posterior probability



Threshold of 0.5 -> lowest error rate (which is known for Bayes classifier) but high error rate for persons who default

As threshold is decreased -> lower error rate for persons who default but the error rate for persons who do not default increases

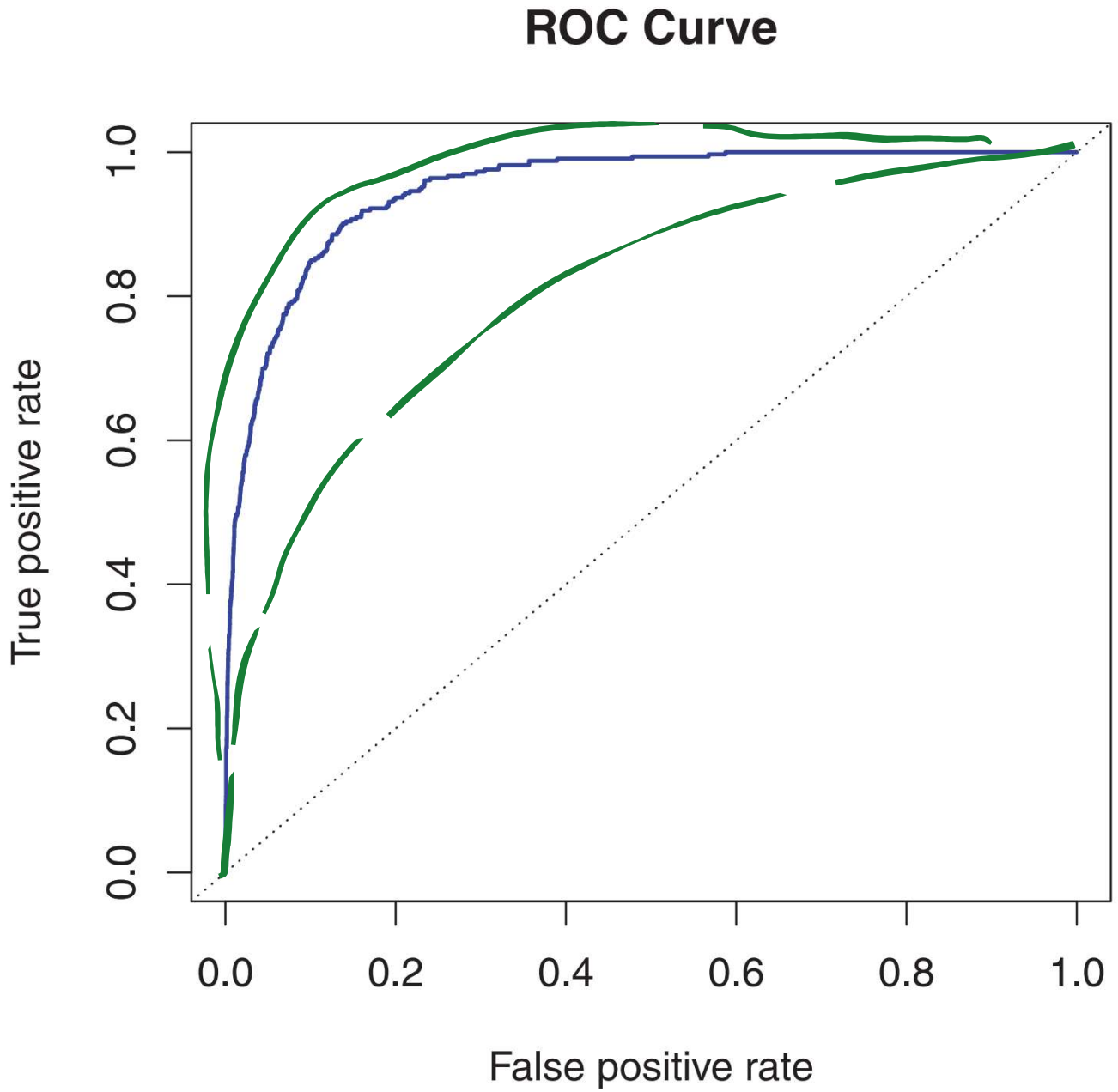
		Predicted class		
		− or Null	+ or Non-null	Total
True class	− or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N*	P*	

		Predicted class		
		non-disease	**disease**	Total
True class	**non-disease**	True Neg. (TN)	False Pos. (FP)	**total non-disease**
	disease	False Neg. (FN)	True Pos. (TP)	**total disease**
	Total	**total predicted non-disease**	**total predicted disease**	

		Predicted class		
		− or Null	+ or Non-null	Total
True class	− or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N*	P*	

		Predicted class		
		non-disease	**disease**	Total
True class	**non-disease**	True Neg. (TN)	False Pos. (FP)	**total non-disease**
	disease	False Neg. (FN)	True Pos. (TP)	**total disease**
	Total	**total predicted non-disease**	**total predicted disease**	

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N*	



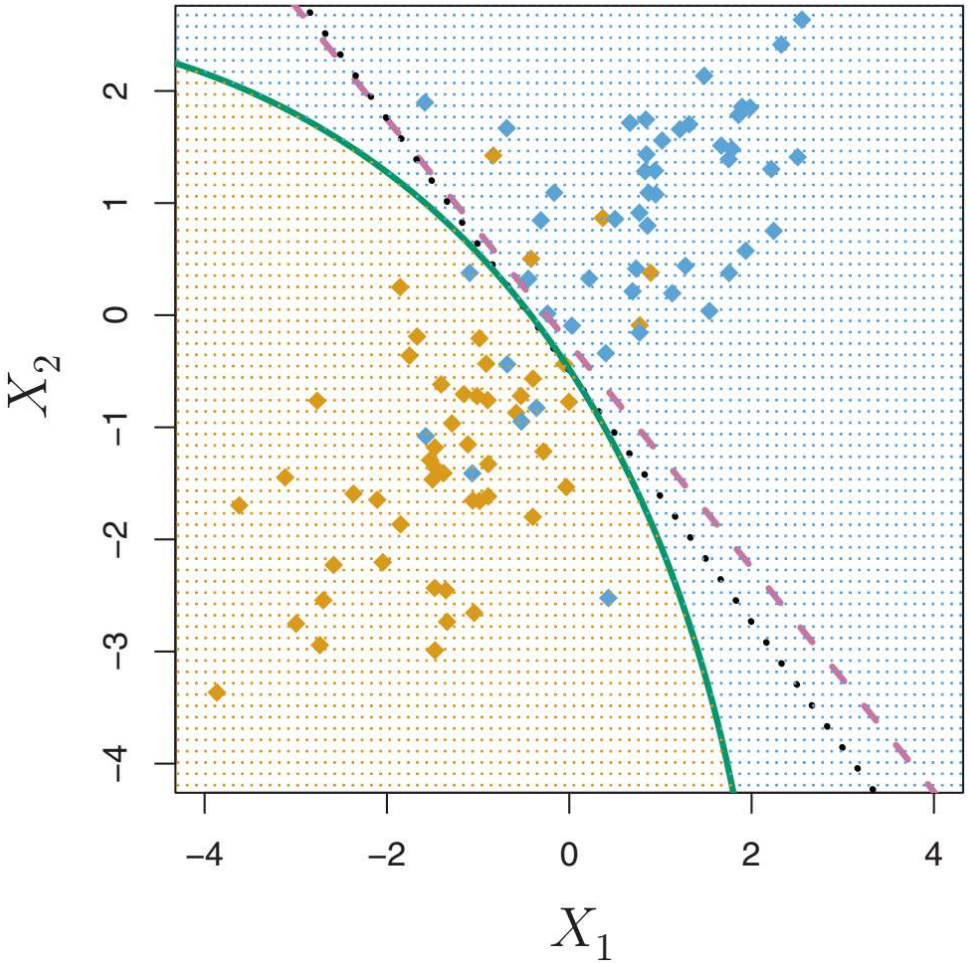
Receiver operating characteristics (ROC) curve plots the two types of errors for all possible thresholds

Ideally, the larger the area under curve (AUC), the better the classifier.

Quadratic Discriminant Analysis (QDA)

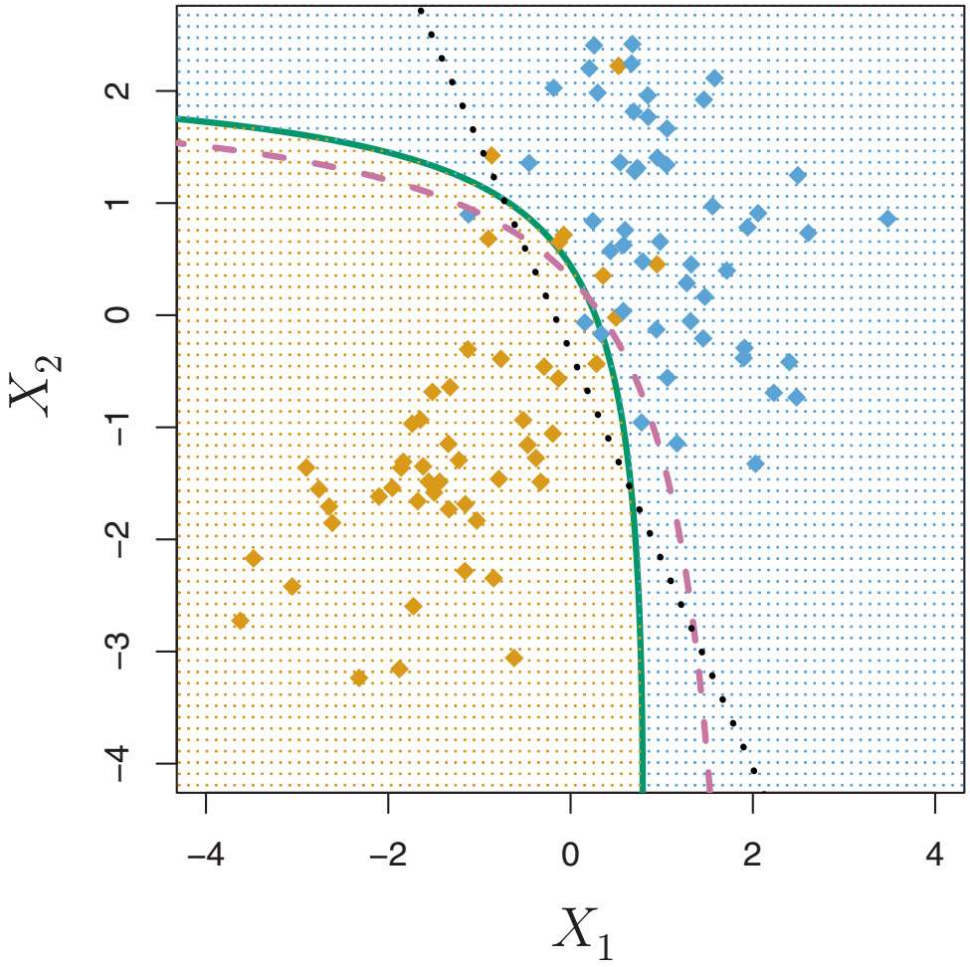
Unlike in LDA, **QDA assumes** that observations are drawn from **classes** also with **Gaussian distribution but with different covariance matrices** ($\sum_k$).

$\sum_1 = \sum_2$



Bayes decision boundary

$\sum_1 \neq \sum_2$

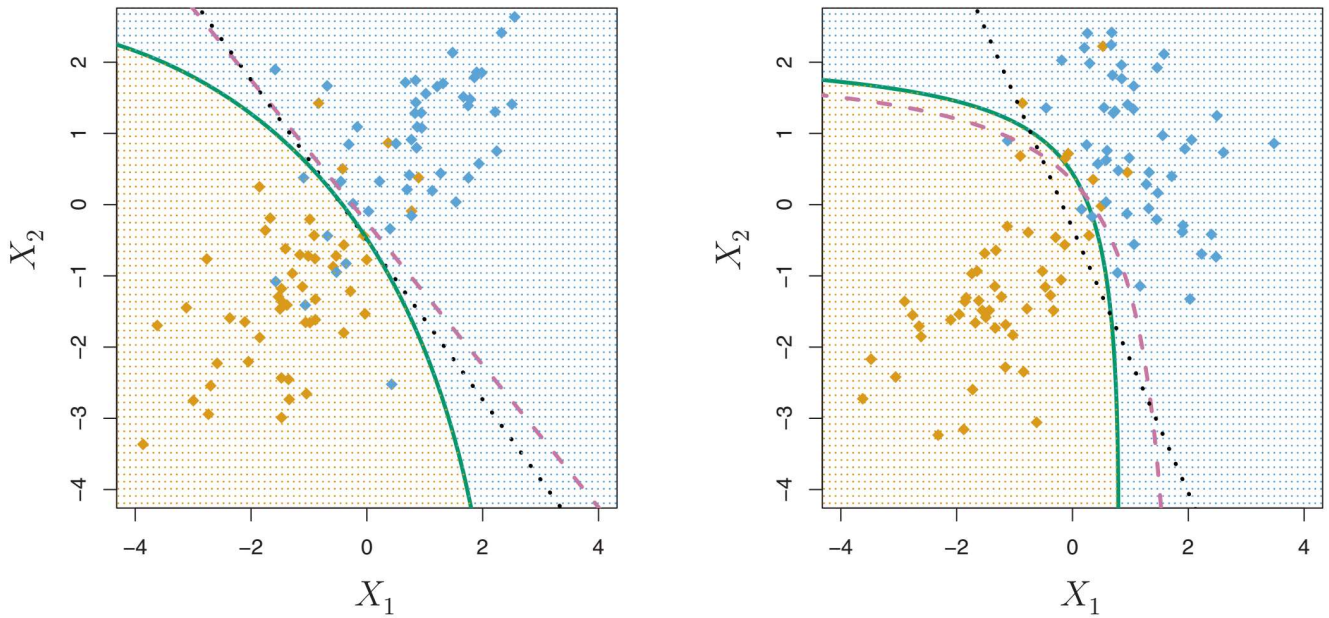


LDA decision boundary

QDA decision boundary

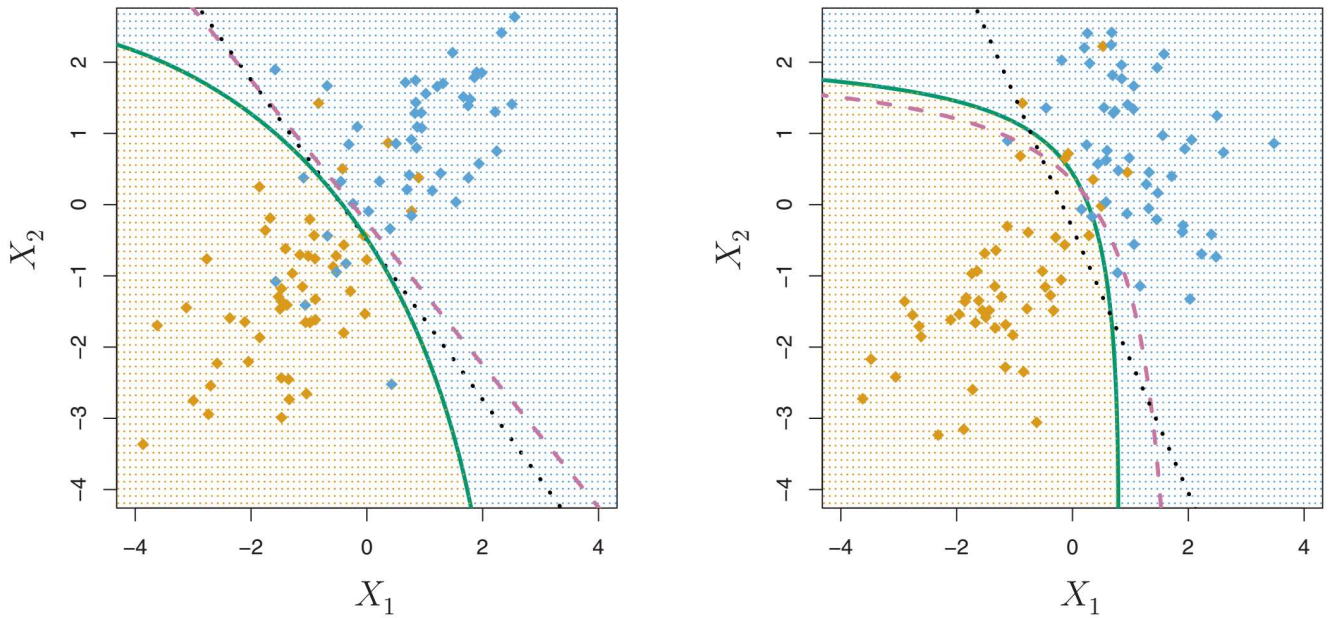
Quadratic Discriminant Analysis (QDA)

- As p increases, the number of covariance matrices to be estimated by QDA is $Kp(p+1)/2$.
- **LDA is less flexible than QDA, but it has a lower variance.**
- **LDA tends to be better when the number of training instances is relatively low**, i.e. when the variance needs to be minimized.
- **QDA tends to perform better with for very large training sets**, i.e. variance is not a concern.

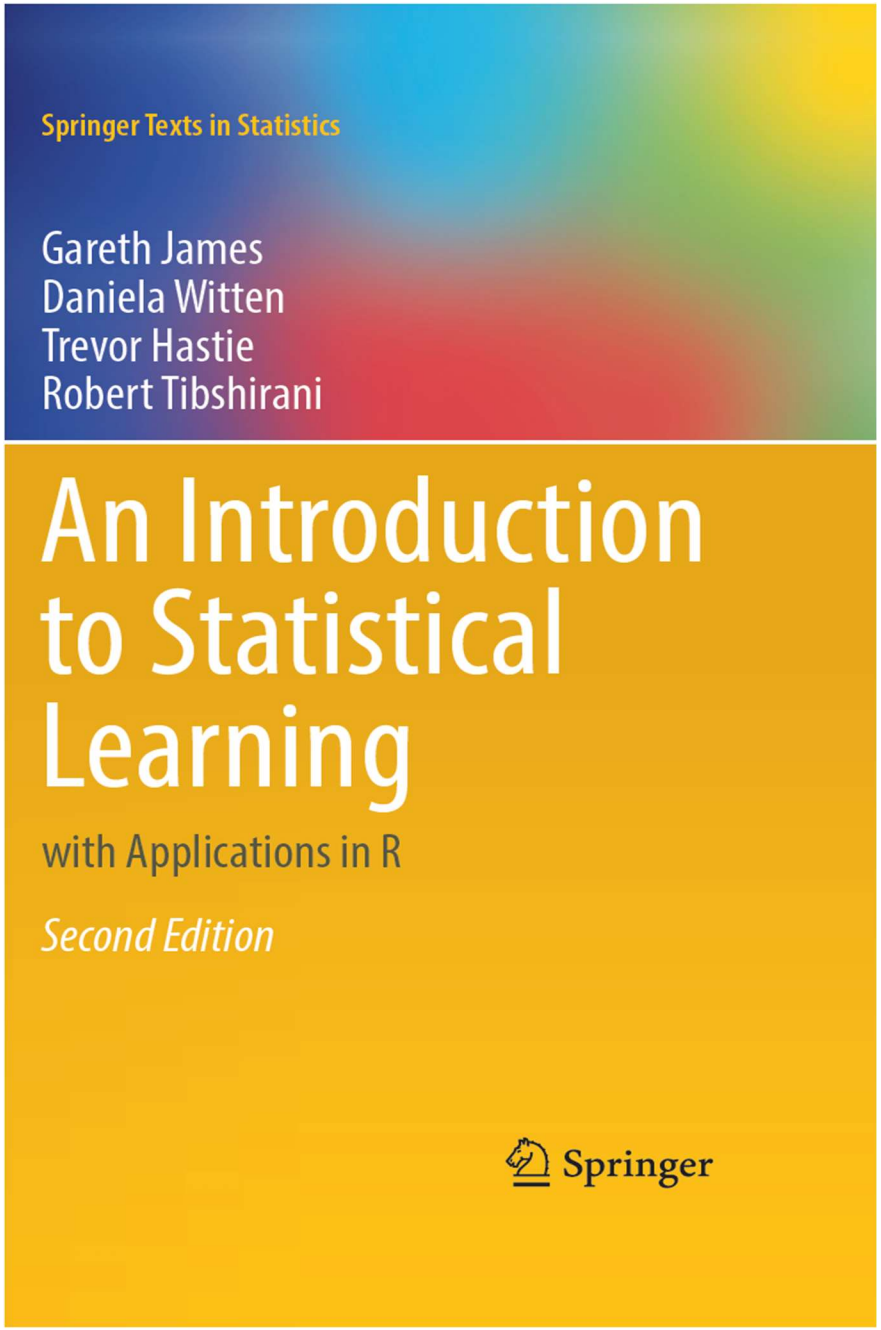


Quadratic Discriminant Analysis (QDA)

- When the true decision boundaries are linear -> LDA and logistic regression methods tend to perform well.
- In cases of moderately non-linear boundaries -> QDA may yield better results.
- For more complicated decision boundaries -> non-parametric approaches such as KNN can be better.



Reference


Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Second Edition
Springer