Discriminant Analysis

CS-309

Covariance

$$\text{Variance}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

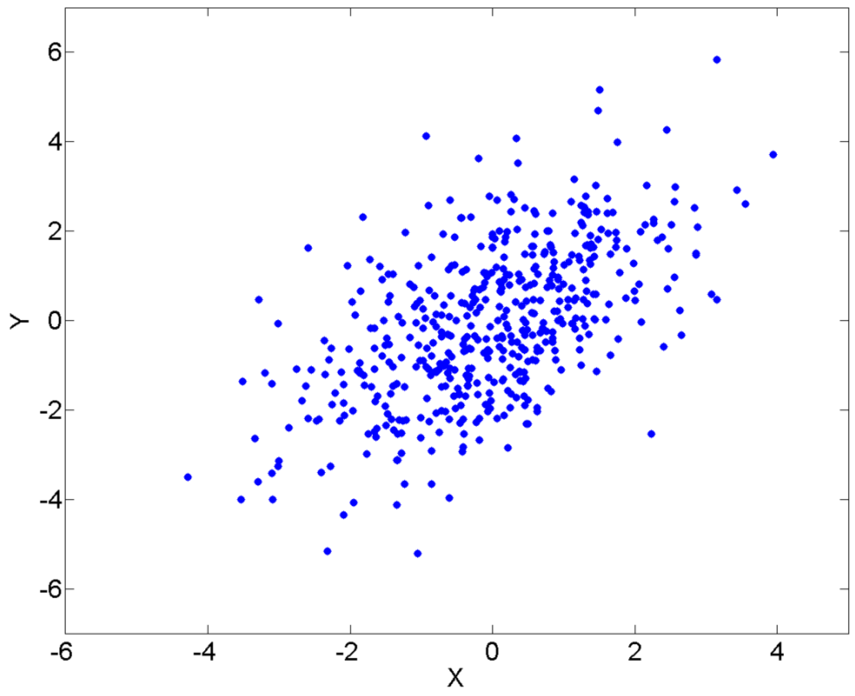
$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})$$

$$\text{Covariance}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- $\text{Covariance}(x, x) = \text{var}(x)$
- $\text{Covariance}(x, y) = \text{Covariance}(y, x)$

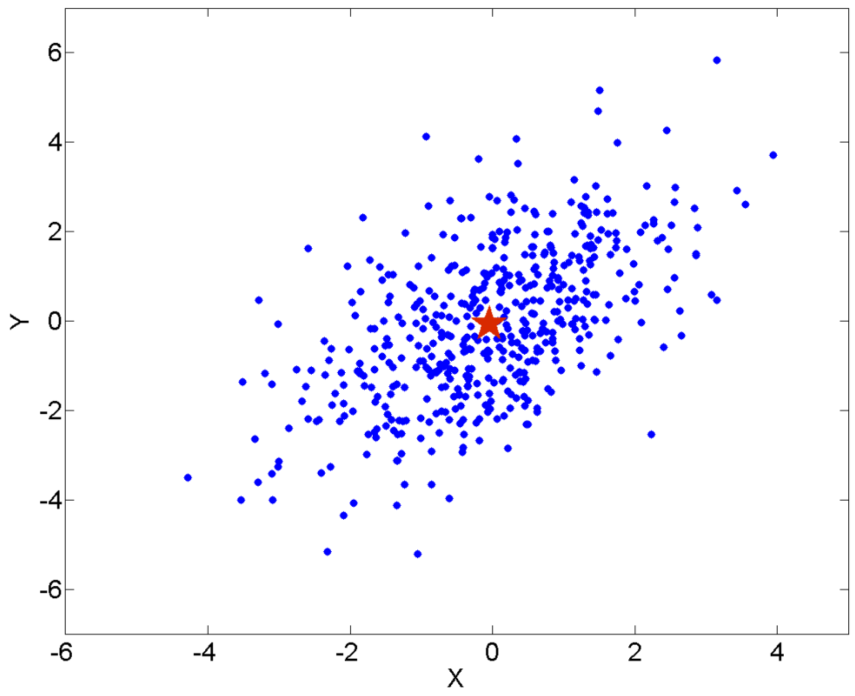
Covariance

$$\text{Covariance}(\mathrm{x},\mathrm{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$



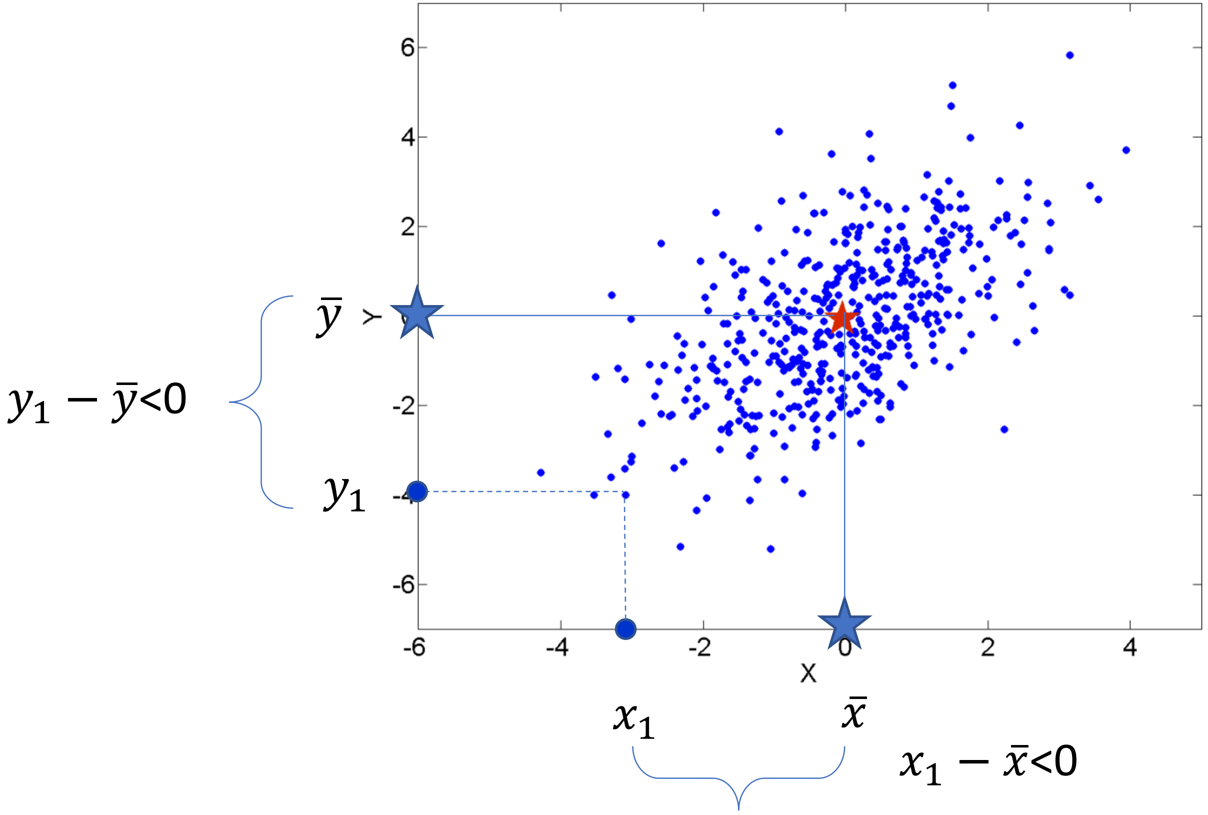
Covariance

$$\text{Covariance}(\mathrm{x}, \mathrm{y}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$



Covariance

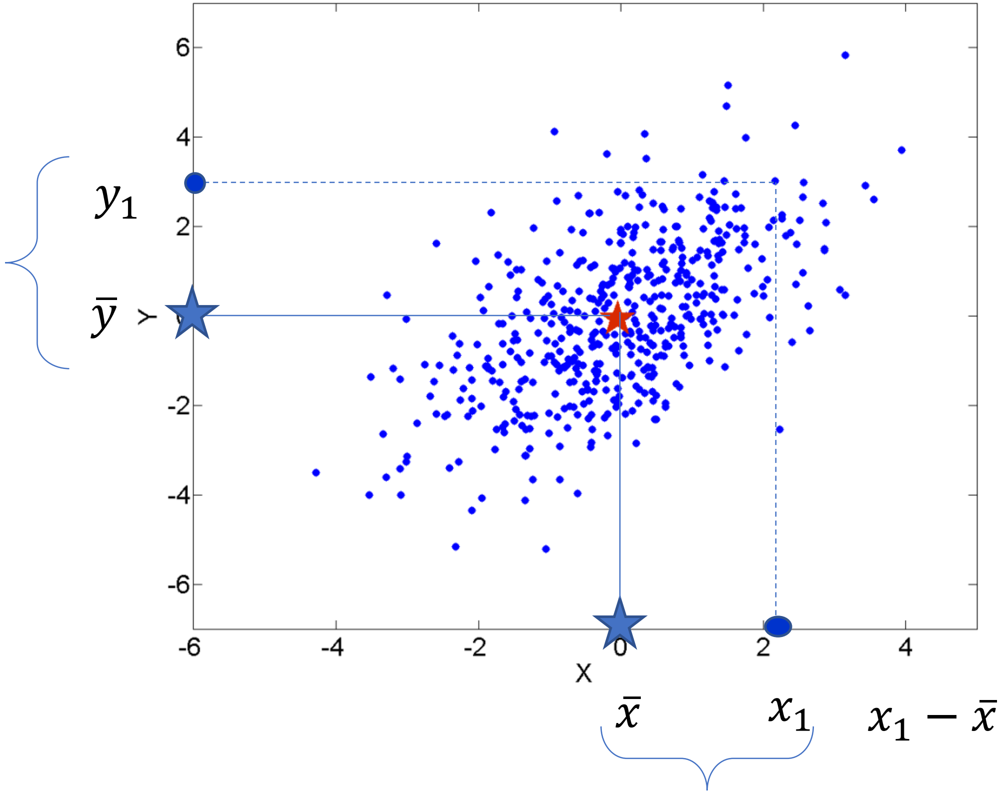
$$\text{Covariance}(\mathrm{x}, \mathrm{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$



Covariance

$$\text{Covariance}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

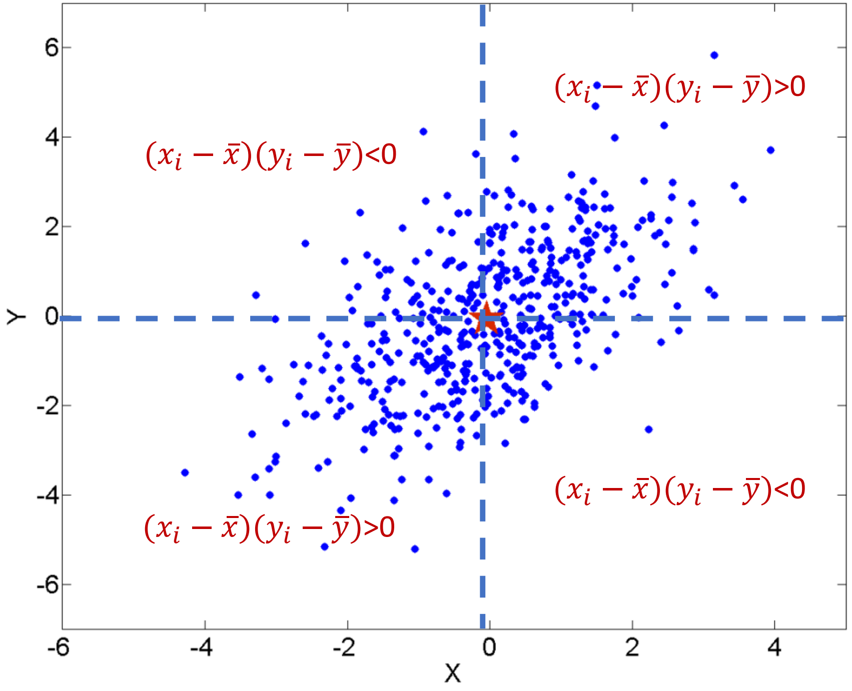
$y_1 - \bar{y} > 0$



$x_1 - \bar{x} > 0$

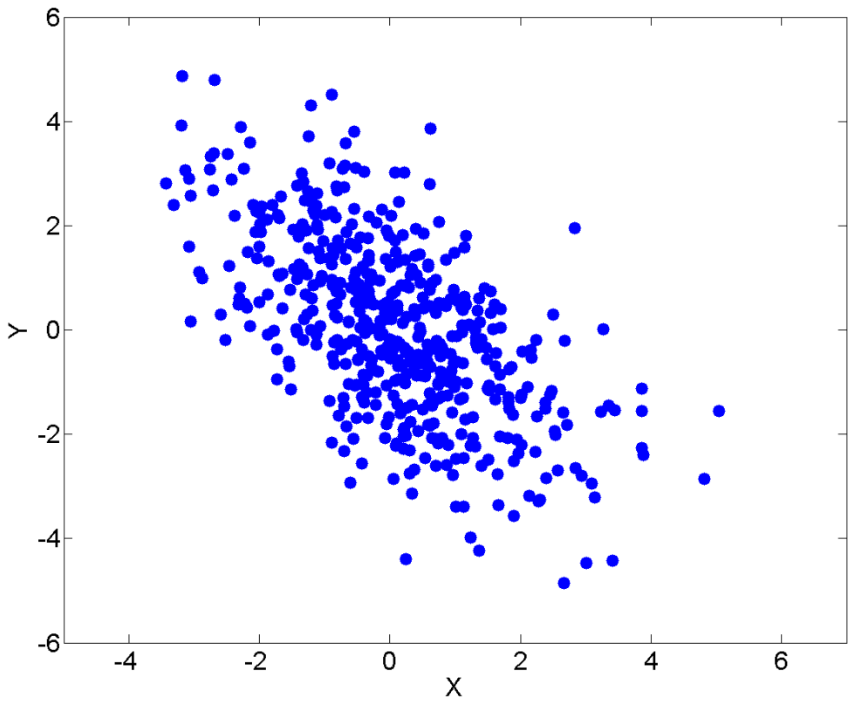
Covariance

$$\text{Covariance}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$



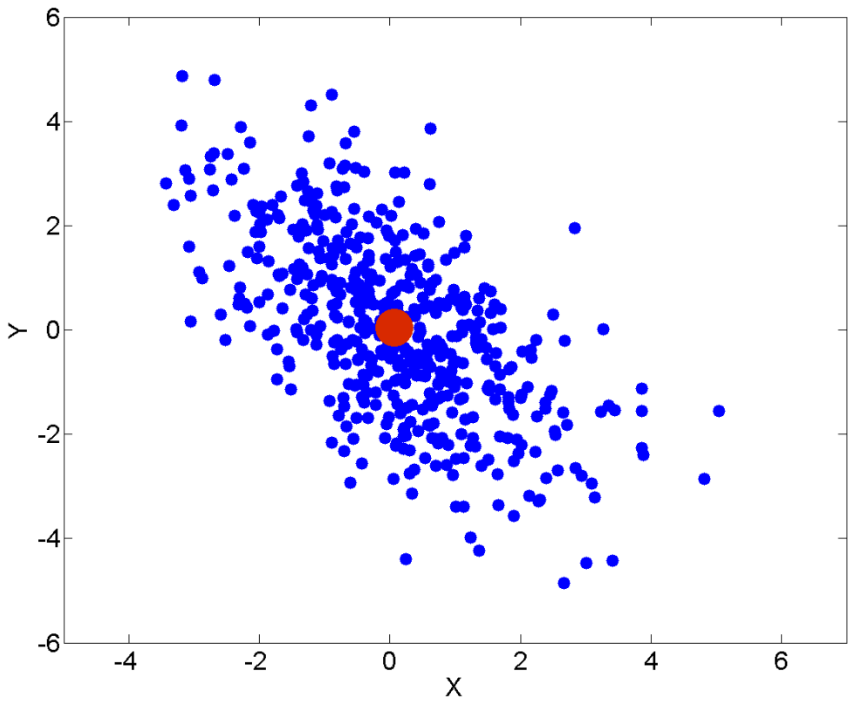
Covariance

$$\text{Covariance}(\mathrm{x}, \mathrm{y}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$



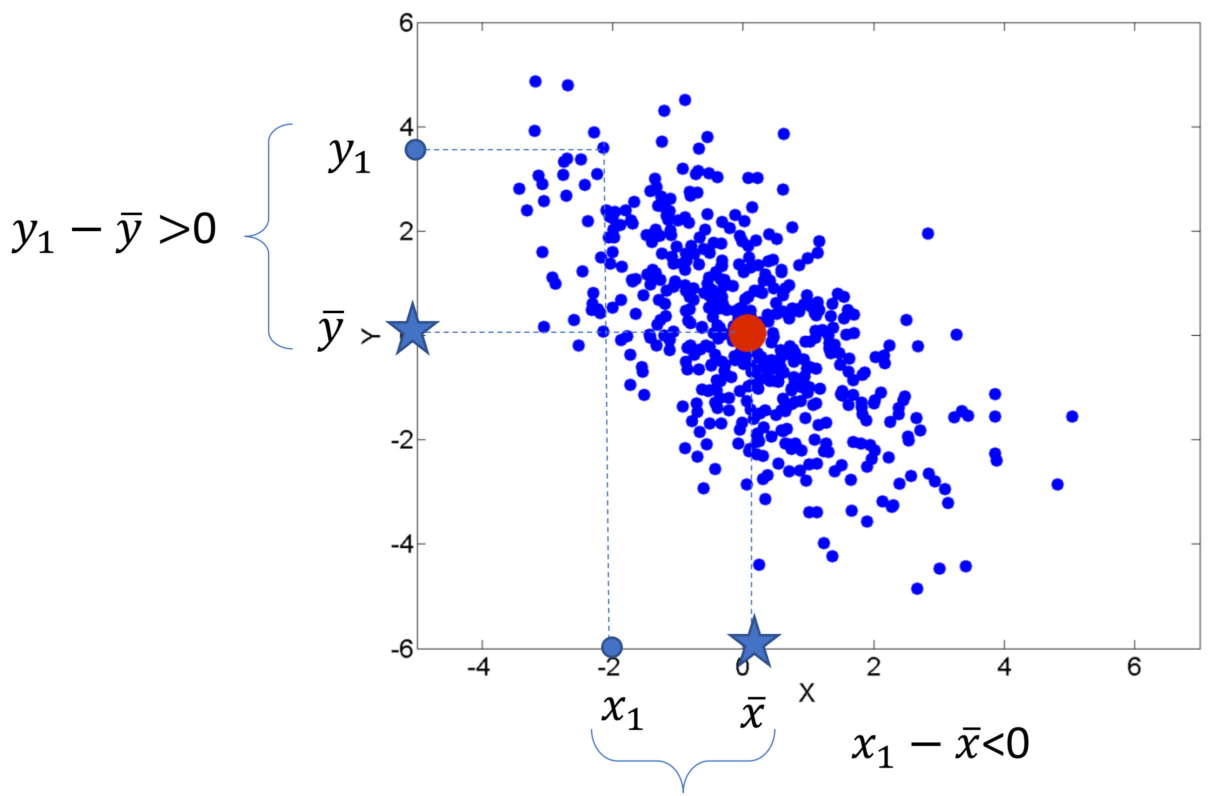
Covariance

$$\text{Covariance}(\mathrm{x},\mathrm{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$



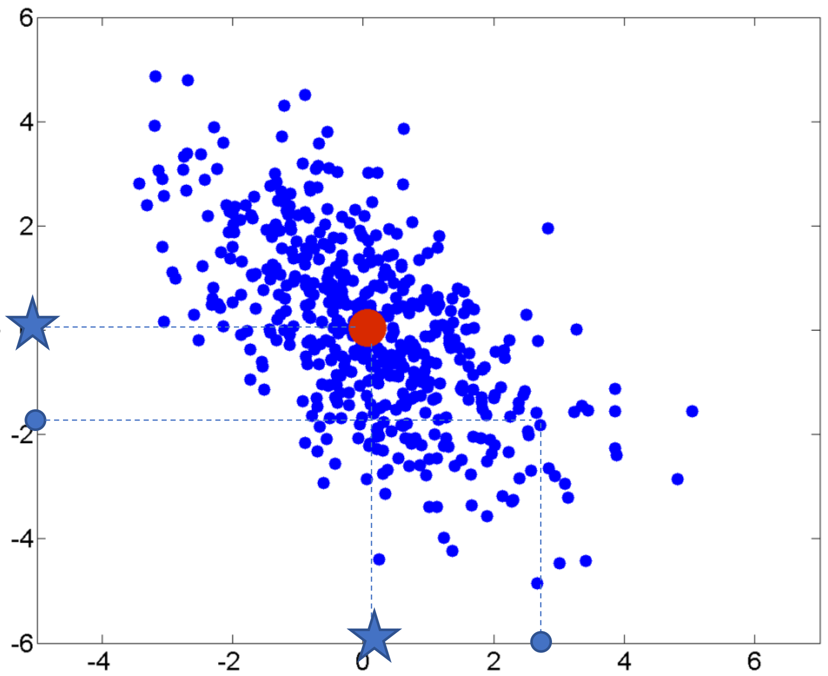
Covariance

$$\text{Covariance}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$



Covariance

$$\text{Covariance}(\mathrm{x}, \mathrm{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$



$y_1 - \bar{y}$<0

$\bar{y}$

y_1

$\bar{x}$

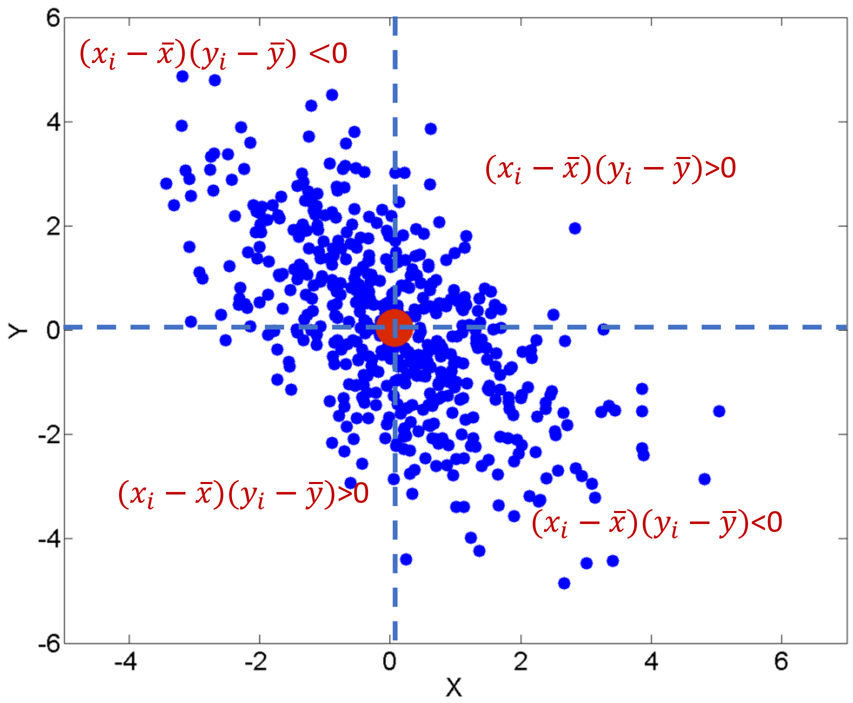
x_1

$x_1 - \bar{x}$>0

Covariance

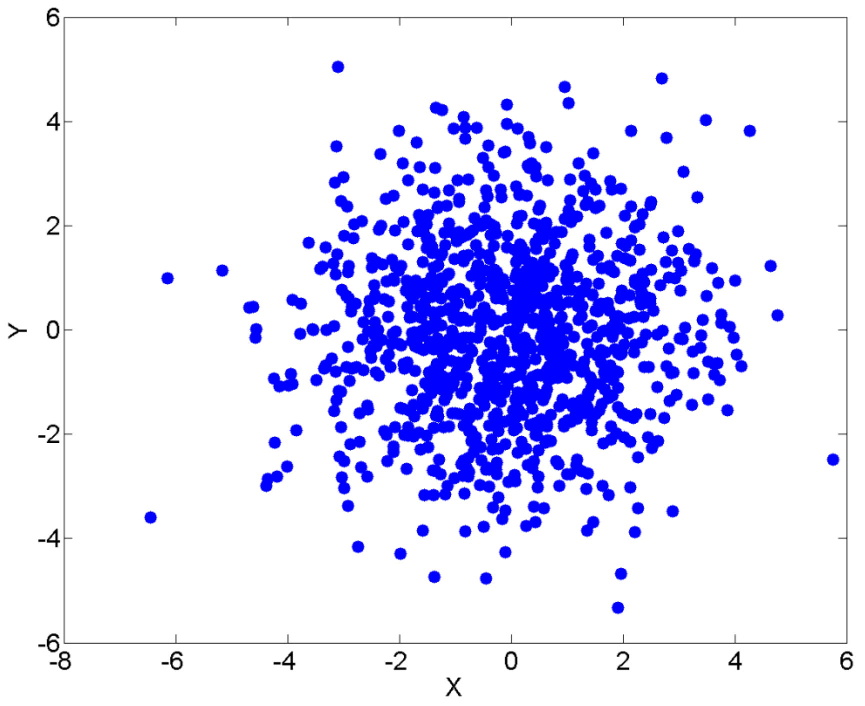
$$\mathrm{Covariance}(\mathrm{x},\mathrm{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Negative Relation



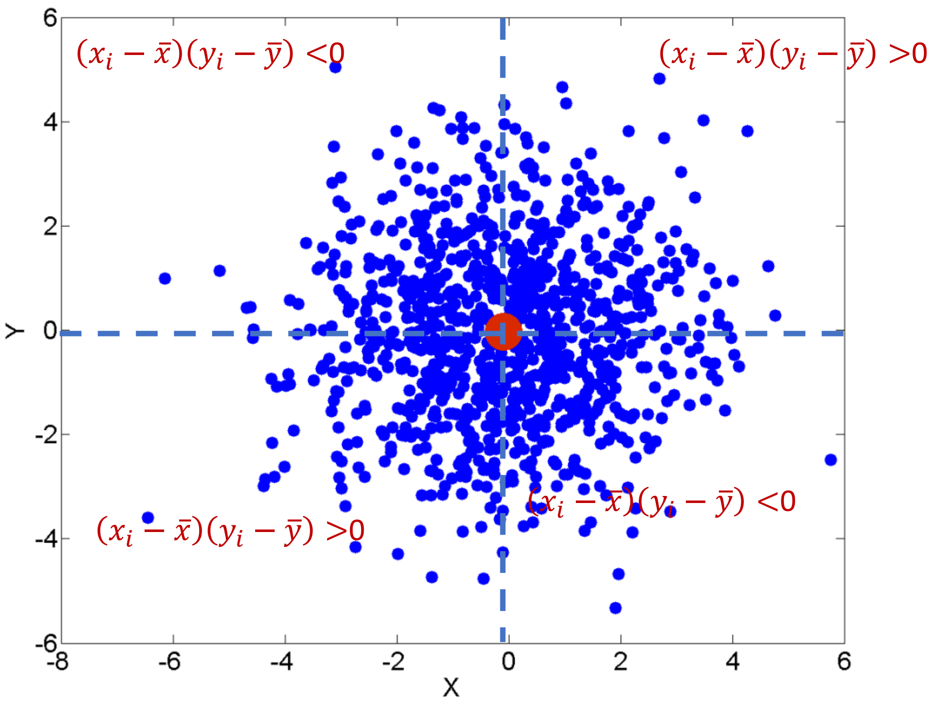
Covariance

$$\text{Covariance}(\mathrm{x}, \mathrm{y}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$



Covariance

$$\text{Covariance}(\mathrm{x}, \mathrm{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$



Covariance Matrix

$$Cov\ (\Sigma) = \begin{bmatrix} cov(x_1, x_1) & cov(x_1, x_2) & \cdots & cov(x_1, x_m) \\ cov(x_2, x_1) & cov(x_2, x_2) & \cdots & cov(x_2, x_m) \\ \vdots & \vdots & \vdots & \vdots \\ cov(x_m, x_1) & cov(x_m, x_2) & \cdots & cov(x_m, x_m) \end{bmatrix}$$

$$Cov\ (\Sigma) = \frac{1}{n}(X - \bar{X})(X - \bar{X})^T; where\ X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

Covariance Matrix

$$Cov\ (\Sigma) = \begin{bmatrix} cov(x_1, x_1) & cov(x_1, x_2) & \cdots & cov(x_1, x_m) \\ cov(x_2, x_1) & cov(x_2, x_2) & \cdots & cov(x_2, x_m) \\ \vdots & \vdots & \vdots & \vdots \\ cov(x_m, x_1) & cov(x_m, x_2) & \cdots & cov(x_m, x_m) \end{bmatrix}$$

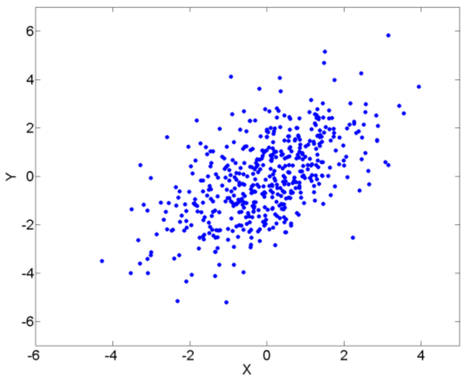
- Diagonal elements are variances, i.e. Cov(x, x)=$var(x)$.
- Covariance Matrix is symmetric.
- It is a positive semi-definite matrix.

Covariance Matrix

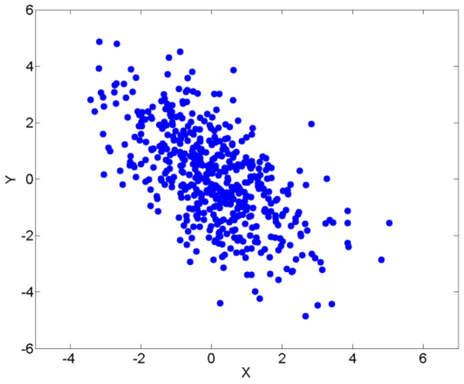
- Covariance is a real symmetric positive semi-definite matrix.
  - All eigenvalues must be real
  - Eigenvectors corresponding to different eigenvalues are orthogonal
  - All eigenvalues are greater than or equal to zero
  - Covariance matrix can be diagonalized,

    i.e. $\mathbf{Cov = PDP^T}$

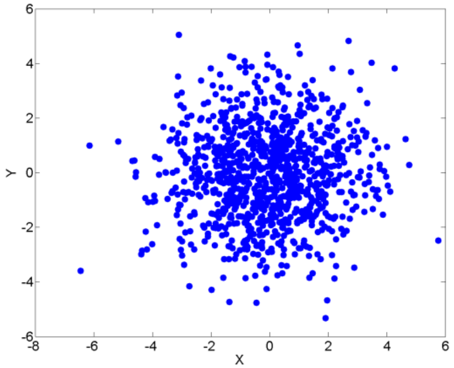
Correlation



Positive relation



Negative relation



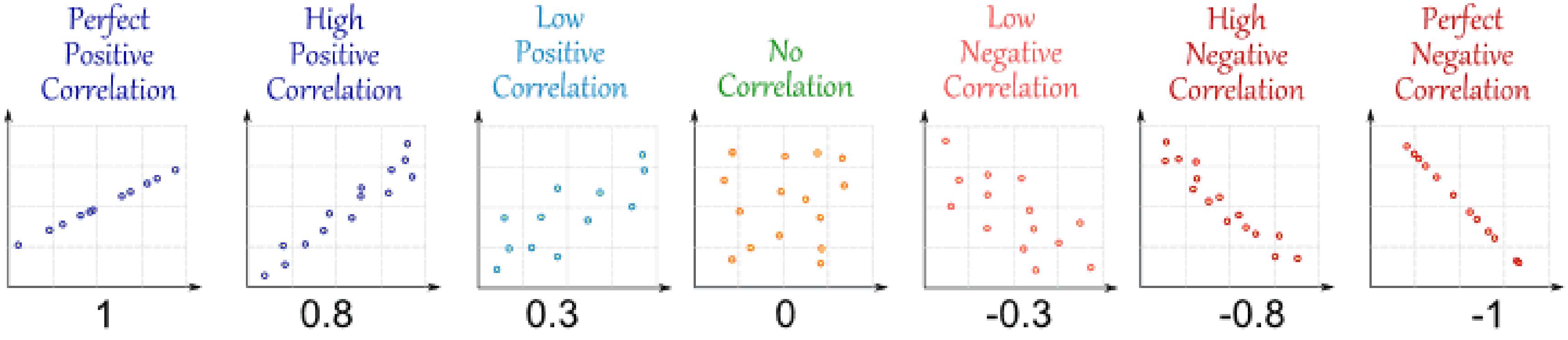
No relation

- Covariance determines whether relation is positive or negative, but it was impossible to measure the degree to which the variables are related.
- Correlation is another way to determine how two variables are related.
- In addition to whether variables are positively or negatively related, correlation also tells the degree to which the variables are related each other.

Correlation

$$\rho_{xy} = Correlation\ (x, y) = \frac{cov(x, y)}{\sqrt{var(x)}\sqrt{var(y)}}.$$

$$-1 \leq Correlation(x, y) \leq +1$$



Linear Discriminant Analysis (LDA)

- Logistic regression involves directly modeling $\Pr(Y = k|X = x)$ using the logistic function.
- LDA considers an alternative and less direct approach to estimating these probabilities.
- In this alternative approach, we model the distribution of the predictors X separately in each of the response classes (i.e. given Y).
- Then use Bayes' theorem to flip these around into estimates for $\Pr(Y = k|X = x)$.

Bayes theorem for classification

Thomas Bayes was a famous mathematician whose name represents a big subfield of statistical and probabilistic modeling. Here we focus on a simple result, known as Bayes theorem:

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

One writes this slightly differently for discriminant analysis:

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

- $f_k(x) = \Pr(X = x|Y = k)$ is the *density* for X in class k. Here we will use normal densities for these, separately in each class.
- $\pi_k = \Pr(Y = k)$ is the marginal or *prior* probability for class k.

- Let $f_k(X) \equiv \Pr(X = x|Y = k)$ denotes the *density function* of X for an observation that comes from the kth class. In other words, $f_k(X)$ is relatively large if there is a high probability that an observation in the kth class has $X \approx x$, and $f_k(X)$ is small if it is very unlikely that an observation in the kth class has $X \approx x$.
- we will use the abbreviation $p_k(X) = \Pr(Y = k|X)$ and referred as posterior probability
- If we can simply plug in estimates of π_k and $f_k(X)$ into following equation then we can estimate $p_k(X)$

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

- In general, estimating π_k is easy if we have a random sample of Y s from the population: we simply compute the fraction of the training observations that belong to the kth class.
- However, estimating $f_k(X)$ tends to be more challenging, unless we assume some simple forms for these densities.

Linear Discriminant Analysis for p = 1

- For now, assume that $p = 1$—that is, we have only one predictor.
- We would like to obtain an estimate for $f_k(x)$ that we can plug into Eq (1) in order to estimate $p_k(x)$.
- We will then classify an observation to the class k for which $p_k(x)$ is greatest.
- Suppose we assume that $f_k(x)$ is *normal* or *Gaussian*.

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

- where μ_k and σ_k^2 are the mean and variance parameters for the kth class.
- For now, let us further assume that $\sigma_1^2 = . . . = \sigma_k^2$: that is, there is a shared variance term across all K classes, which for simplicity we can denote by σ^2.

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

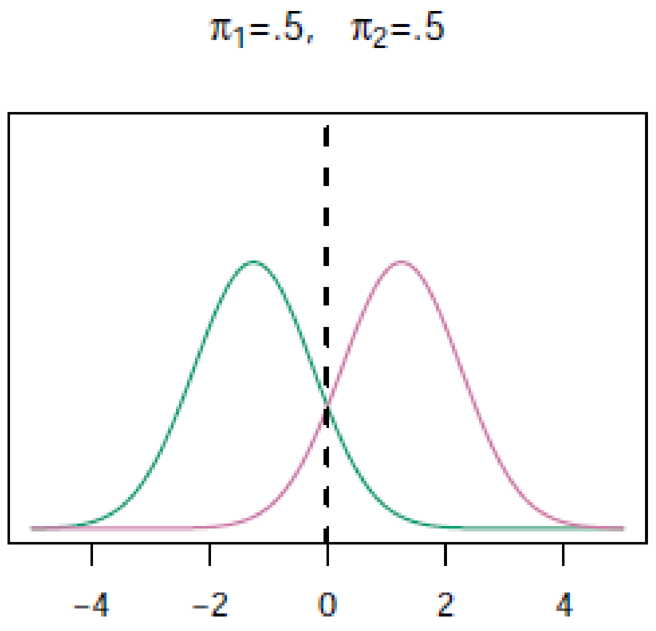
- Taking the log of last equation and ignoring those terms which does not involve k

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) \qquad \text{--Eq (2)}$$

it is not hard to show that this is equivalent to assigning the observation to the class for which $\delta_k(x)$ is the largest.

- For instance, if $K = 2$ and $\pi_1 = \pi_2$ then
- $\delta_1(x) - \delta_2(x) = x.\frac{\mu_1}{\sigma^2} - \frac{{\mu_1}^2}{2\sigma^2} - x.\frac{\mu_2}{\sigma^2} + \frac{{\mu_2}^2}{2\sigma^2} = \frac{1}{2\sigma^2}(2x(\mu_1 - \mu_2) - ({\mu_1}^2 - {\mu_2}^2))$

  $\Rightarrow \delta_1(x) > \delta_2(x)$ when $2x(\mu_1 - \mu_2) > ({\mu_1}^2 - {\mu_2}^2)$
- In this case, the Bayes decision boundary corresponds to the point where
- $X = \frac{{\mu_1}^2 - {\mu_2}^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}$ ---(Eq 3)



- The two normal density functions that are displayed, $f_1(x)$ and $f_2(x)$, represent two distinct classes.
- The mean and variance parameters for the two density functions are $\mu_1 = -1.25$, $\mu_2 = 1.25$, and $\sigma_1^2 = \sigma_2^2 = 1$.
- We also assumes $\pi_1 = \pi_2 = 0.5$
- by inspection of Eq (3), we see that the Bayes classifier assigns the observation to class 1 if $x < 0$ and class 2 otherwise.

- Note that in this case, we can compute the Bayes classifier because we know that X is drawn from a Gaussian distribution within each class, and we know all of the parameters involved.
- In practice, even if we are quite certain of our assumption that X is drawn from a Gaussian distribution within each class, we still have to estimate the parameters $\mu_1, \ldots, \mu_K, \pi_1, \ldots, \pi_K$, and σ^2.
- The *linear discriminant analysis* (LDA) method approximates the Bayes classifier by plugging estimates for π_k, μ_K, and σ^2 into Eq (2).

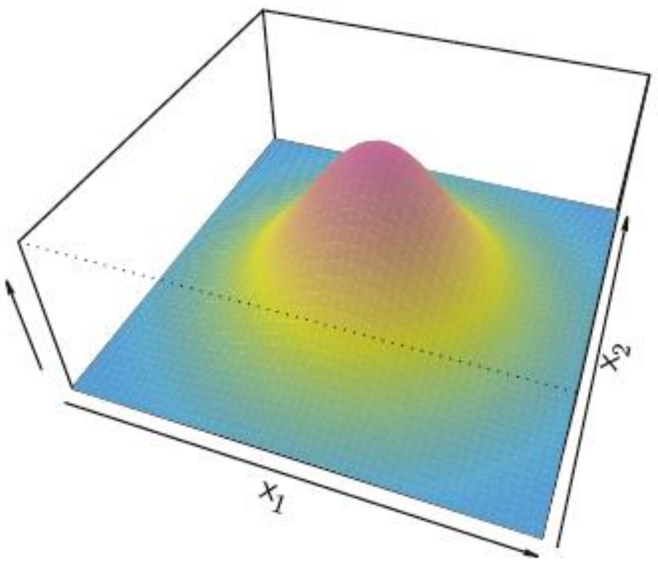
$$
\begin{aligned}
\hat{\mu}_k &= \frac{1}{n_k} \sum_{i:y_i=k} x_i \\
\hat{\sigma}^2 &= \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2
\end{aligned}
$$

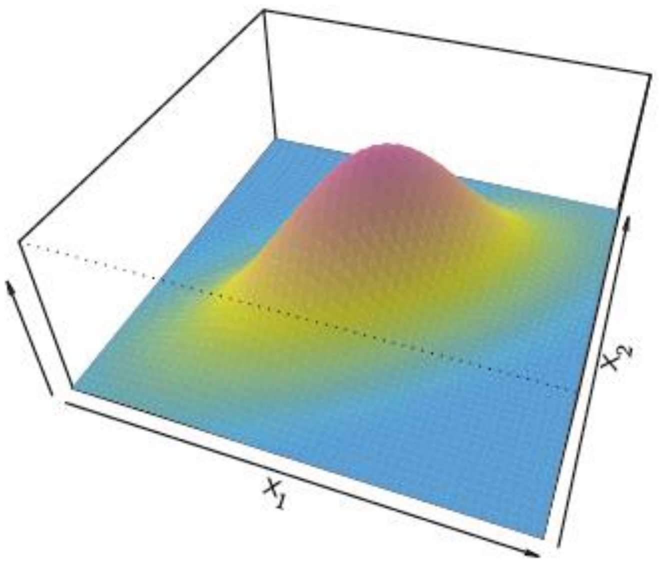
$$\hat{\pi}_k = n_k/n.$$

where n is the total number of training observations, and n_k is the number of training observations in the kth class.

Linear Discriminant Analysis for p >1

- we will assume that $X = (X_1, X_2, \ldots, X_p)$ is drawn from a *multivariate Gaussian* (or multivariate normal) distribution with a class-specific mean vector and a common covariance matrix.
- The multivariate Gaussian distribution assumes that each individual predictor follows a one-dimensional normal distribution, with some correlation between each pair of predictors.
- To indicate that a p-dimensional random variable X has a multivariate Gaussian distribution, we write $X \sim N(\mu, \mathbf{\Sigma})$.
- Here $E(X) = \mu$ is the mean of X, and $\text{Cov}(X) = \mathbf{\Sigma}$ is the $p \times p$ covariance matrix of X.





Two multivariate Gaussian density functions are shown, with p = 2. Left: *The two predictors are uncorrelated.* Right: *The two variables have a correlation of* 0.7*.*

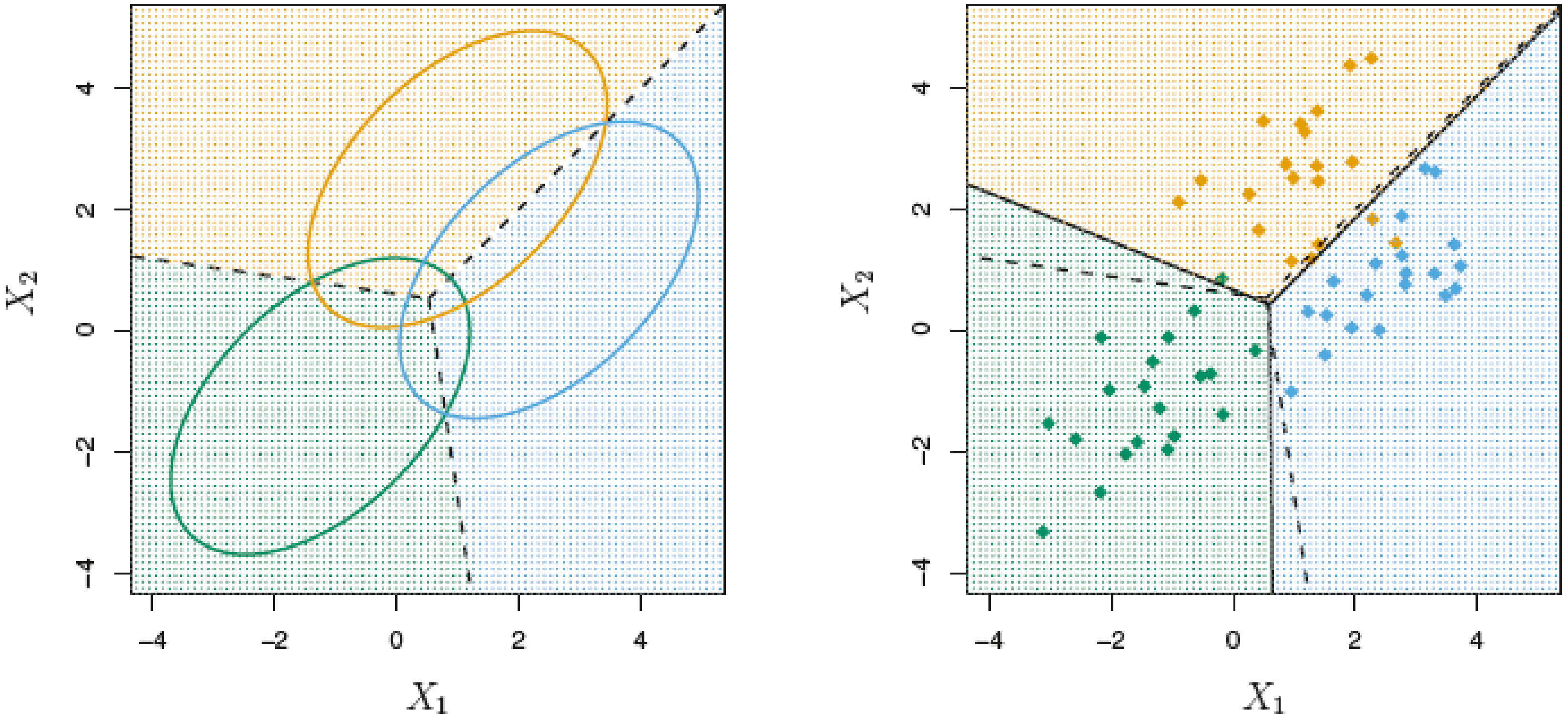
- Formally, the multivariate Gaussian density is defined as

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

- In the case of $p > 1$ predictors, the LDA classifier assumes that the observations in the kth class are drawn from a multivariate Gaussian distribution $N(\mu_k, \mathbf{\Sigma})$, where μ_k is a class-specific mean vector, and $\mathbf{\Sigma}$ is a covariance matrix that is common to all K classes.
- Plugging the density function for the kth class, $f_k(X = x)$, into Eq (1) and performing a little bit of algebra reveals that the Bayes classifier assigns an observation $X = x$ to the class for which

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

- Once again, we need to estimate the unknown parameters $\mu_1, . . . , \mu_K$, $\pi_1, . . . , \pi_K$, and $\mathbf{\Sigma}$;
- Estimation is done in similar way, we estimated for p=1



An example with three classes. The observations from each class are drawn from a multivariate Gaussian distribution with p = 2, with a class-specific mean vector and a common covariance matrix. Left: *Ellipses that contain 95% of the probability for each of the three classes are shown. The dashed lines are the Bayes decision boundaries.* Right: 20 *observations were generated from each class, and the corresponding LDA decision boundaries are indicated using solid black lines. The Bayes decision boundaries are once again shown as dashed lines.*

- The test error rates for the Bayes and LDA classifiers are 0.0746 and 0.0770, respectively.

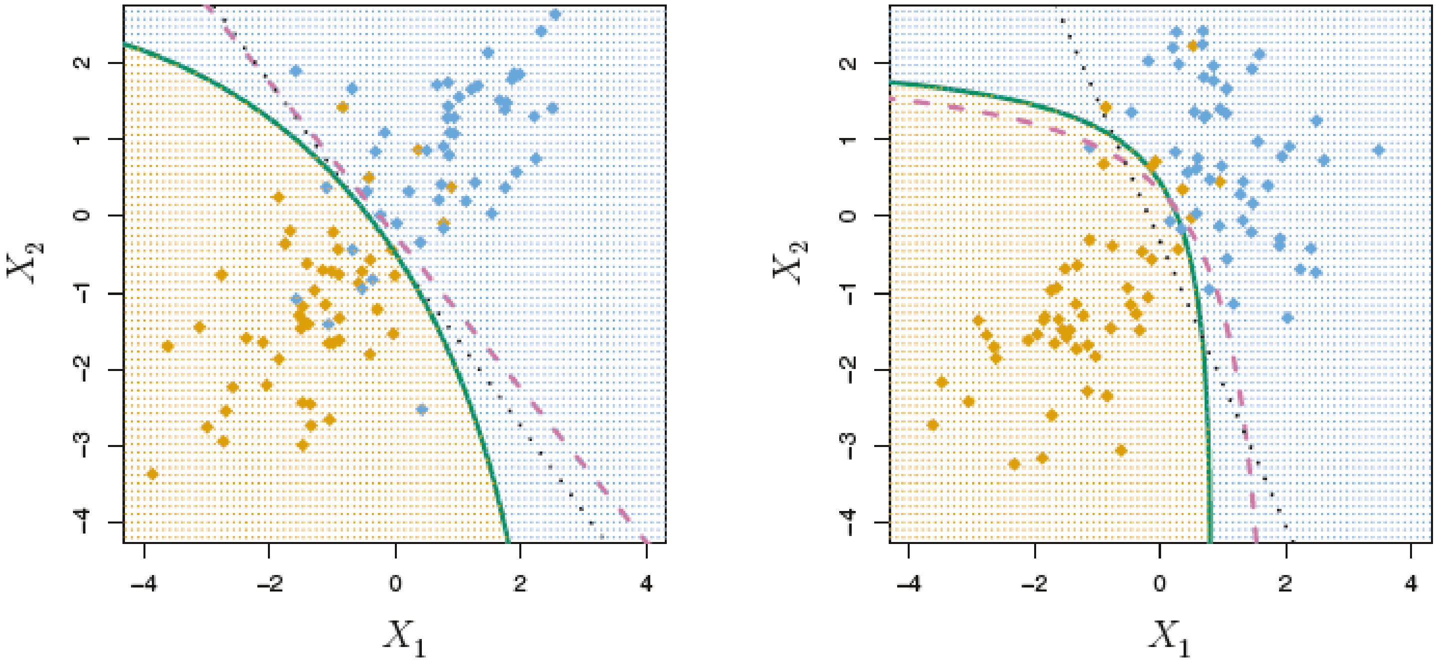
Quadratic Discriminant Analysis

- LDA assumes that the observations within each class are drawn from a multivariate Gaussian distribution with a class specific mean vector and a covariance matrix that is common to all *K* classes.
- Like LDA, the QDA classifier results from assuming that the observations from each class are drawn from a Gaussian distribution, and plugging estimates for the parameters into Bayes' theorem in order to perform prediction.
- However, unlike LDA, QDA assumes that each class has its own covariance matrix.

why would one prefer LDA to QDA, or vice-versa

- The answer lies in the bias-variance trade-off. When there are p predictors, then estimating a covariance matrix requires estimating $p(p+1)/2$ parameters.
- QDA estimates a separate covariance matrix for each class, for a total of $Kp(p+1)/2$ parameters.
- Consequently, LDA is a much less flexible classifier than QDA, and so has substantially lower variance.
- This can potentially lead to improved prediction performance. But there is a trade-off: if LDA's assumption that the K classes share a common covariance matrix is badly off, then LDA can suffer from high bias.

- LDA tends to be a better bet than QDA if there are relatively few training observations and so reducing variance is crucial.
- In contrast, QDA is recommended if the training set is very large.



Why discriminant analysis?

- When the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable. Linear discriminant analysis does not suffer from this problem.
- If n is small and the distribution of the predictors X is approximately normal in each of the classes, the linear discriminant model is again more stable than the logistic regression model.
- Linear discriminant analysis is popular when we have more than two response classes, because it also provides low-dimensional views of the data.