Multivariate Normal

-Applied Multivariate Analysis & Statistical Learning-

MMA chapter 4.1 to 4.6

Lecturer: Darren Homrighausen, PhD

Preamble:

- Define normal distribution
- Give some properties
- Assessing normality
- Transformations to normality
- Outlier detection

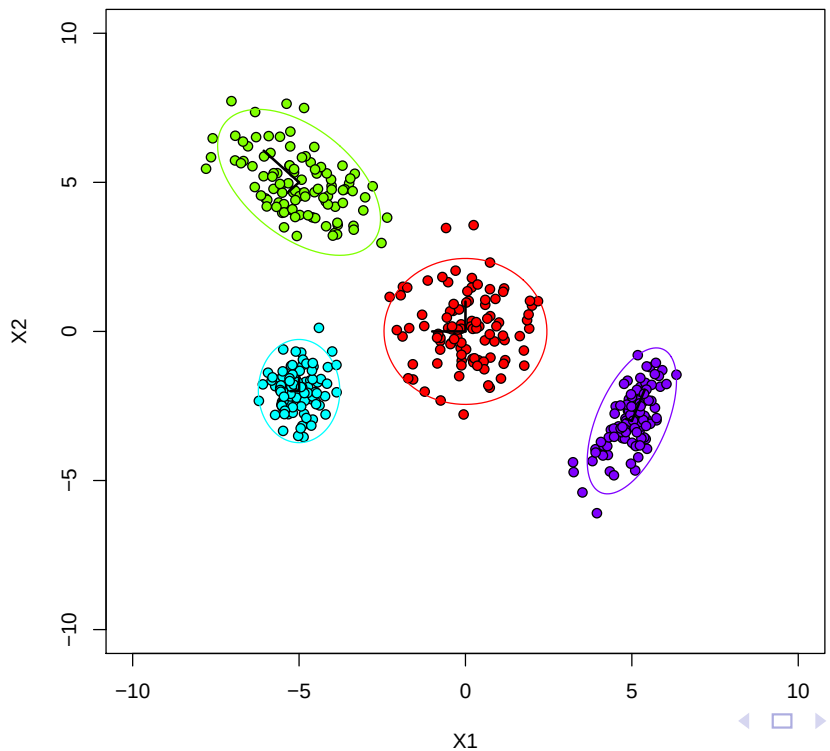
Multivariate Gaussian Distributions

What is a Gaussian?

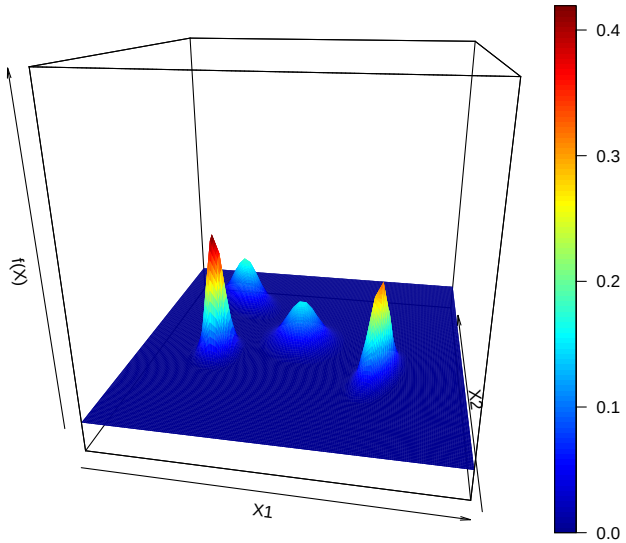
Suppose

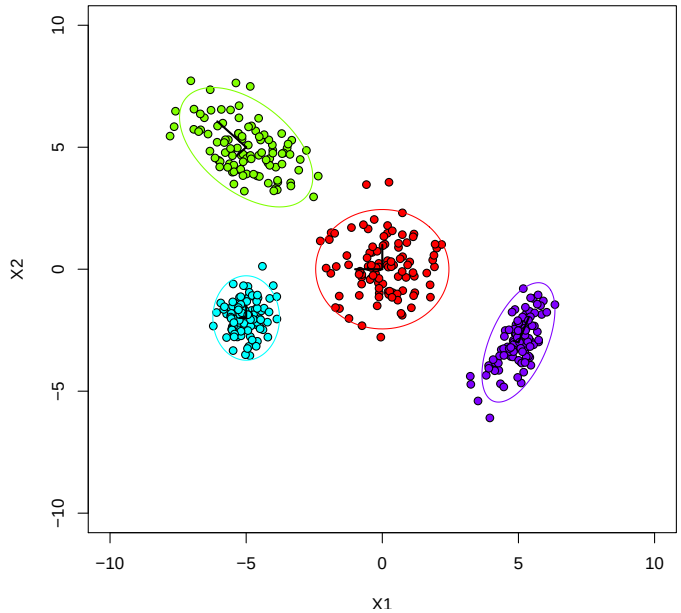
$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) \\ \mathrm{cov}(X_2, X_1) & \mathrm{var}(X_2) \end{bmatrix}\right)$$

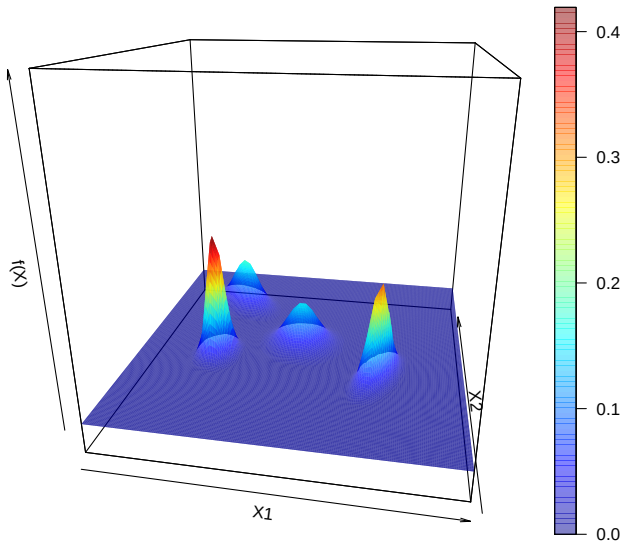
Here are $n = 100$ draws from four different Gaussian distributions.

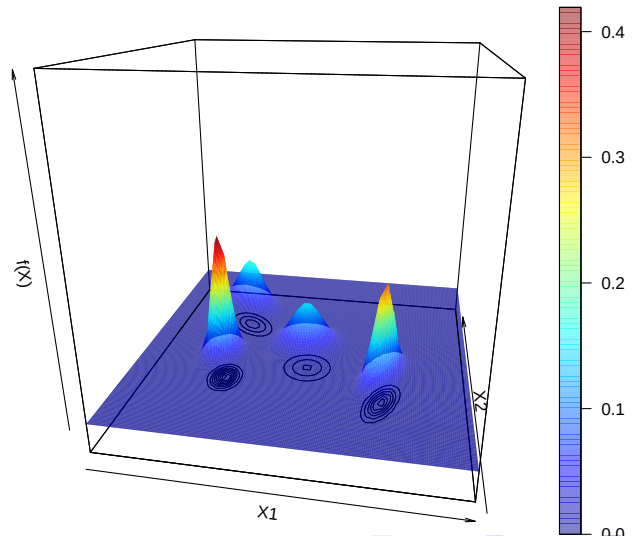


What is a Gaussian?

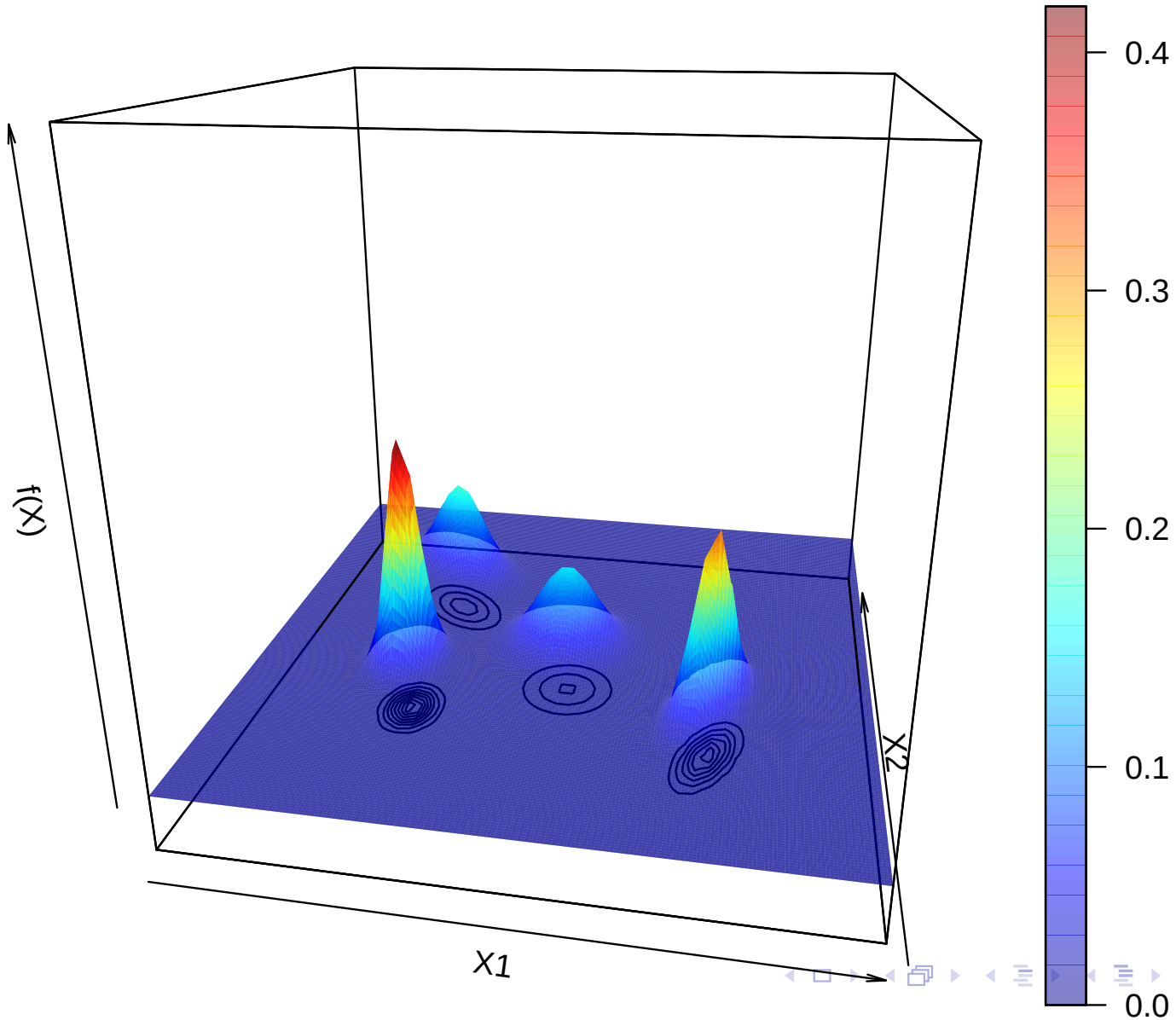








What is a Gaussian?



Definition of the normal distribution

The normal or Gaussian distribution is defined by the following pdf:

$$f(X) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-(X-\mu)^\top \Sigma^{-1}(X-\mu)/2}$$

where

- $X \sim N(\mu, \Sigma) \in \mathbb{R}^p$
- $|\Sigma|$ is the determinant of Σ

The term: $(X-\mu)^\top \Sigma^{-1}(X-\mu)$ plays a special role

---

MMA 4.1.2 & 2.11.3 (for definition of $|\Sigma|$)

Defining distance

Our usual notion of distance is based on the ℓ_2 or Euclidean norm:

$$||a - b||_2 = \sqrt{\sum_{j=1}^{p} (a_j - b_j)^2}$$

(Here, a and b are just two generic length p vectors)

We can rewrite this as

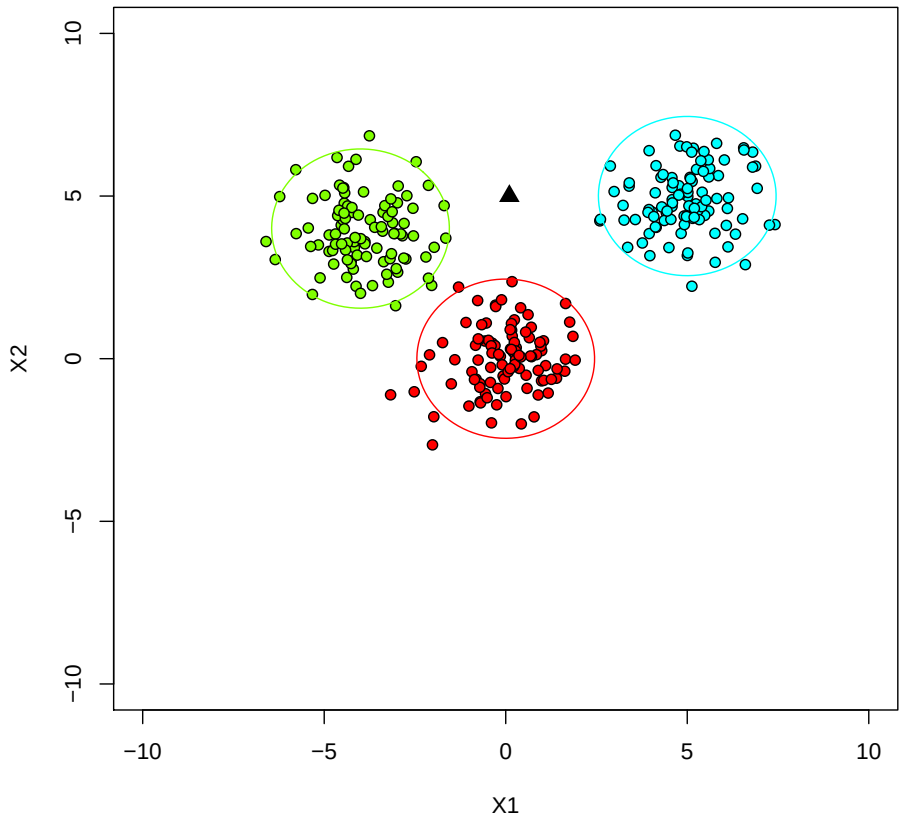
$$||a - b||_2^2 = (a - b)^\top (a - b)$$

Intuitively, two points are close if their distance is small

Let's visually try to determine which of three candidate distributions is intutively closer to a point (▲)

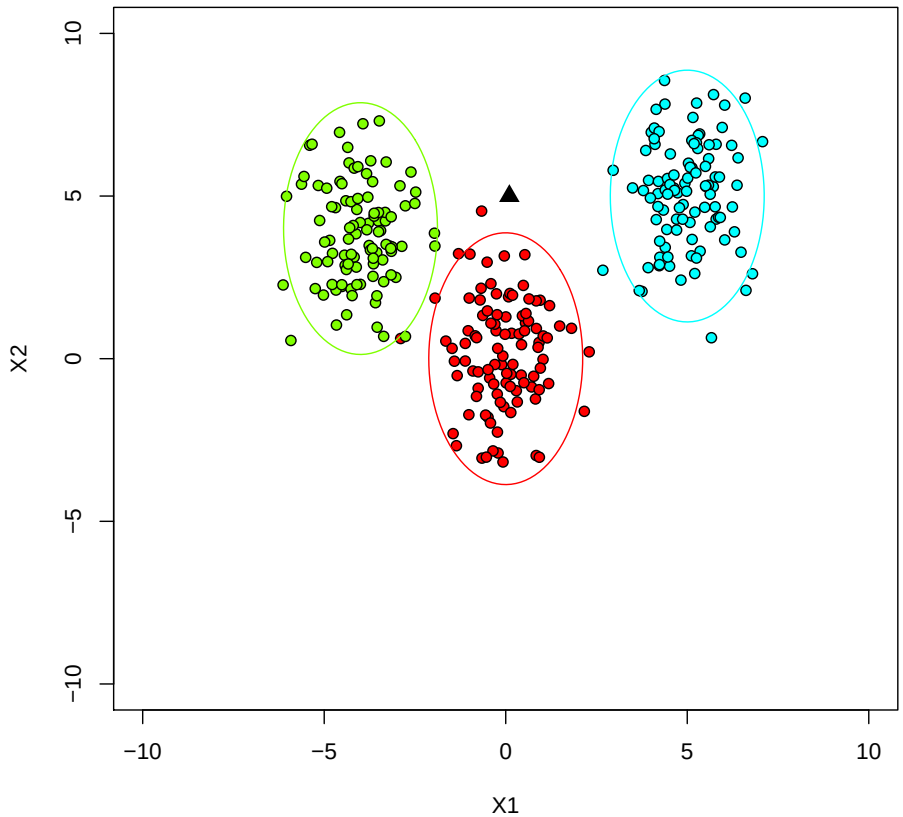
Intuition

What if the distributions looked like this?



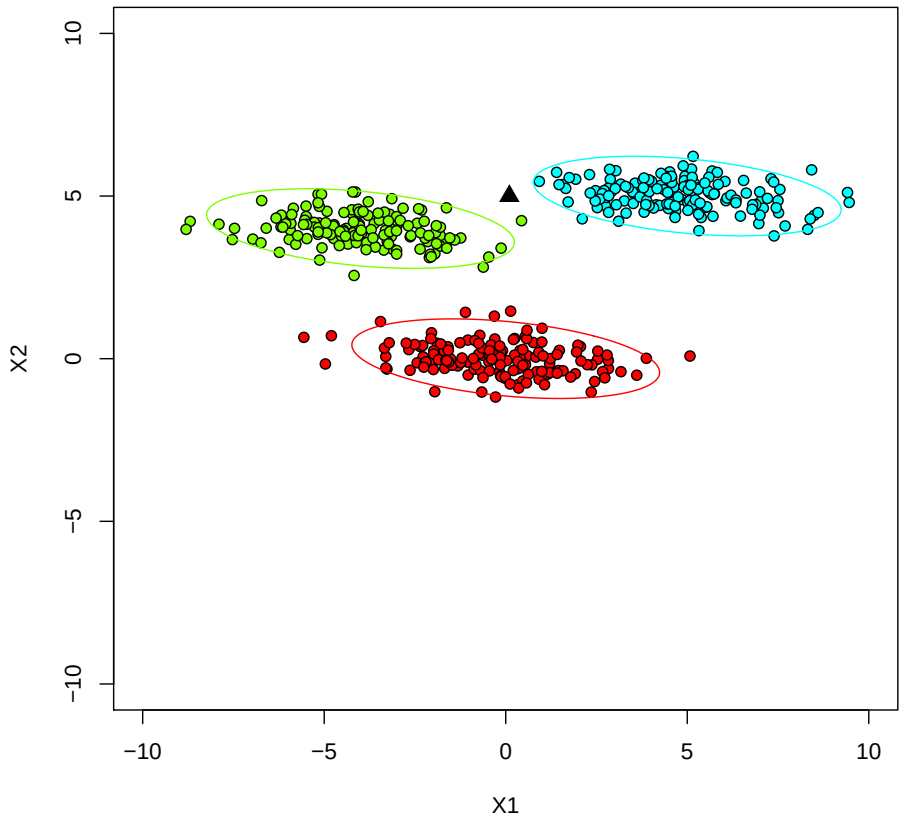
Intuition

Or this?



Intuition

How about this?



Mahalanobis distance

In each case, the Euclidean distance to the mean of each distribution is the same

However, changing the covariance made a big difference

With multivariate data, we need to take into account the covariance to define a sensible notion of distance

Mahalanobis distance

This is exactly what the term $\Delta^2 = (X - \mu)^\top \Sigma^{-1}(X - \mu)$ does

This is the definition of the Mahalanobis distance

(This distance comes up prominently in Linear Discriminant Analysis)

Mahalanobis distance can be derived from the ℓ_2 norm:

$$\begin{aligned}\Delta^2 &= (X - \mu)^\top \Sigma^{-1}(X - \mu) = (X - \mu)^\top \Sigma^{-1/2}\Sigma^{-1/2}(X - \mu) \\ &= (\Sigma^{-1/2}(X - \mu))^\top (\Sigma^{-1/2}(X - \mu)) = \left|\left|\Sigma^{-1/2}(X - \mu)\right|\right|_2^2\end{aligned}$$

MMA chapters 3.13 & 4.1.2

Why the Gaussian distribution?

Why the Gaussian distribution?

The Gaussian distribution is by far the most studied distribution

(For instance, the entire MMA book is devoted to its exploration)

This deserves some explanation

The reason is two-fold

- The Gaussian distribution has very useful properties
- The (multivariate) central limit theorem

Let's look at each of these in turn...

Properties of Gaussians

MMA chapter 4.2

Important properties

If $X \sim N(\mu, \Sigma)$, then

- Linear combinations are Gaussian

$$\mathbb{A}X \sim N(\mathbb{A}\mu, \mathbb{A}\sigma\mathbb{A}^\top) \in \mathbb{R}$$

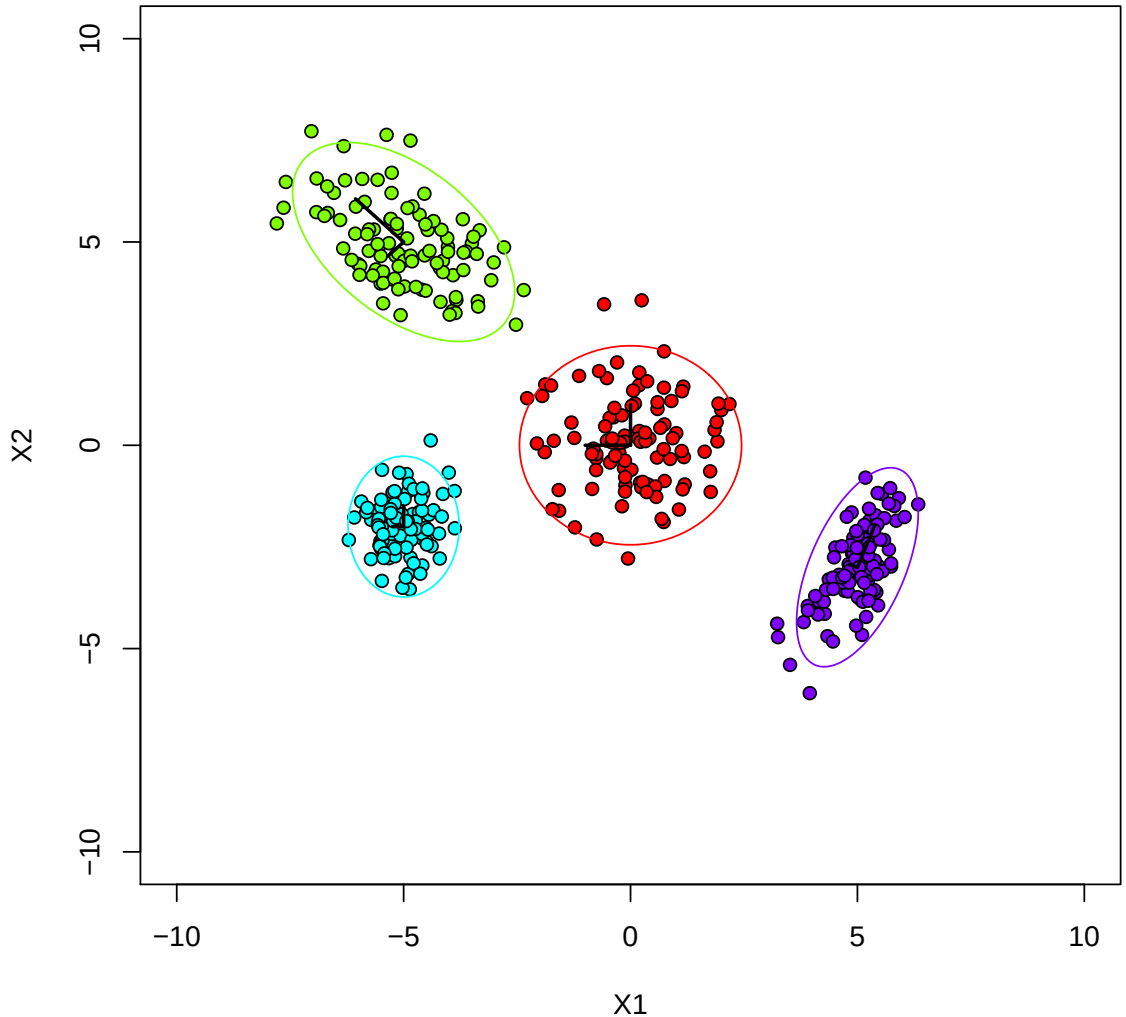
- The Mahalanobis distance between X and μ follows a χ^2

$$\Delta^2 \sim \chi^2_p$$

Setting $\Delta^2 = c^2$ defines an ellipse that has axes $\pm c d_j u_j$
(Where $\Sigma = UD^2U^\top$)

Going back to a previous plot..

Ellipsoids



Ellipsoids: contours

That plotted contour is a 95% confidence area given by

$$\mathbb{P}\left(\Delta^2 \leq \chi^2_p(1-\alpha)\right) = 1-\alpha$$

where $\chi^2_p(1-\alpha)$ is the $(1-\alpha)100\%$ of the χ^2_p distribution

```
> qchisq(.95,1)
[1] 3.841459
> sqrt(qchisq(.95,1))
[1] 1.959964
> qchisq(.95,2)
[1] 5.991465
```

Important properties (continued)

- All marginal distributions are Gaussian

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

  Then $X_k \sim N(\mu_k, \Sigma_k)$ for $k = 1, 2$

  (Note that the converse is not true)

- Two Gaussians are independent if and only if their covariance is 0
- Two Gaussians are conditionally independent given everything else if and only if the corresponding term in Σ^{-1} is zero

  (Note that Σ^{-1} is commonly referred to as the precision matrix)

Estimation

MMA chapter 4.3

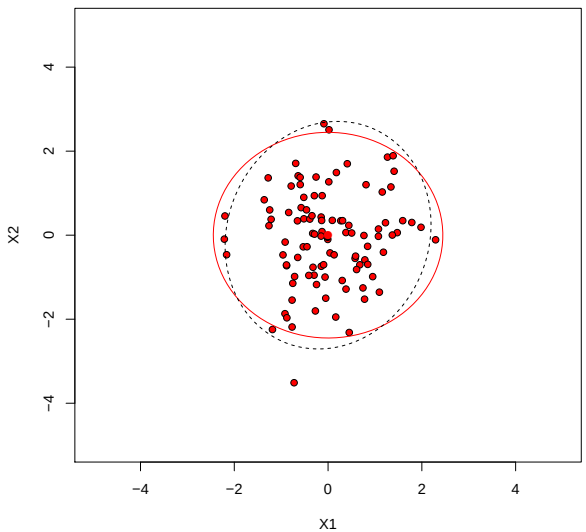
Estimating μ and Σ?

Suppose we make $n = 100$ independent observations

$$X_1, \ldots, X_{100} \sim N\left(\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \begin{bmatrix} 0.0012 \\ 0.001 \end{bmatrix}$$

$$\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})^\top = \frac{1}{n-1}(\mathbb{X} - \overline{\mathbb{X}})^\top(\mathbb{X} - \overline{\mathbb{X}}) = \begin{bmatrix} .8 & .1 \\ .1 & 1.2 \end{bmatrix}$$



Estimating μ and Σ?

There are two cases for $\overline{X}$

- If $X \sim N(\mu, \Sigma)$, then $\overline{X} \sim N(\mu, \Sigma/n)$, which is exactly like the univariate result
- Central limit theorem (CLT): Regardless of the source distribution, independent draws will, for fixed p,

$$\sqrt{n}(\overline{X} - \mu) \to N(0, \Sigma) \text{ as } n \to \infty$$

As far as the estimate S,

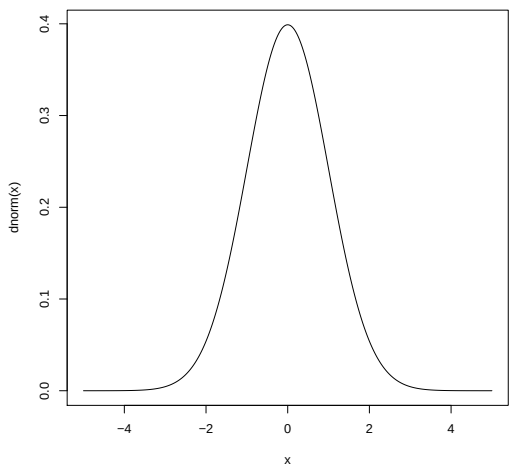
$$\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})^\top \sim W_p(n-1, \Sigma)$$

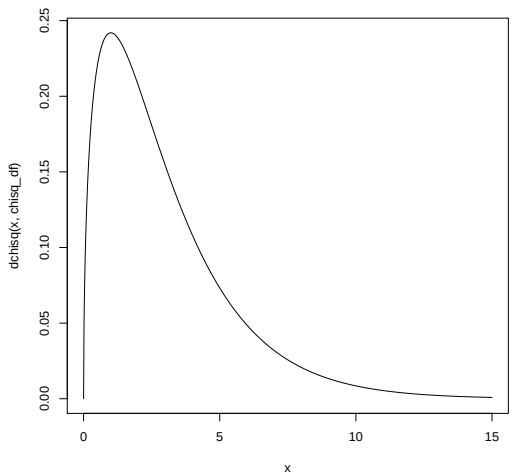
where W_p is the Wishart distribution

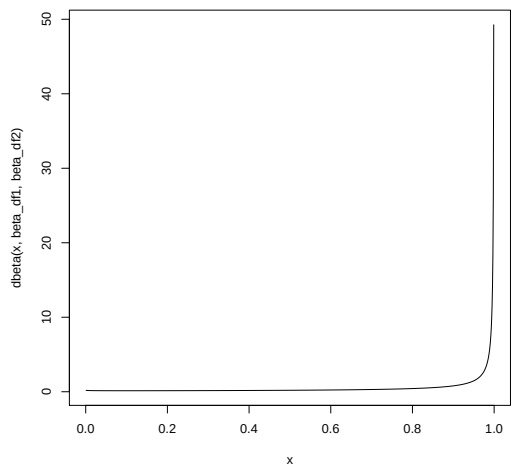
Checking for normality

Univariate normals

By far the most common approach is to look at a QQ-plot...







Univariate normals

By far the most common approach is to look at a QQ-plot...

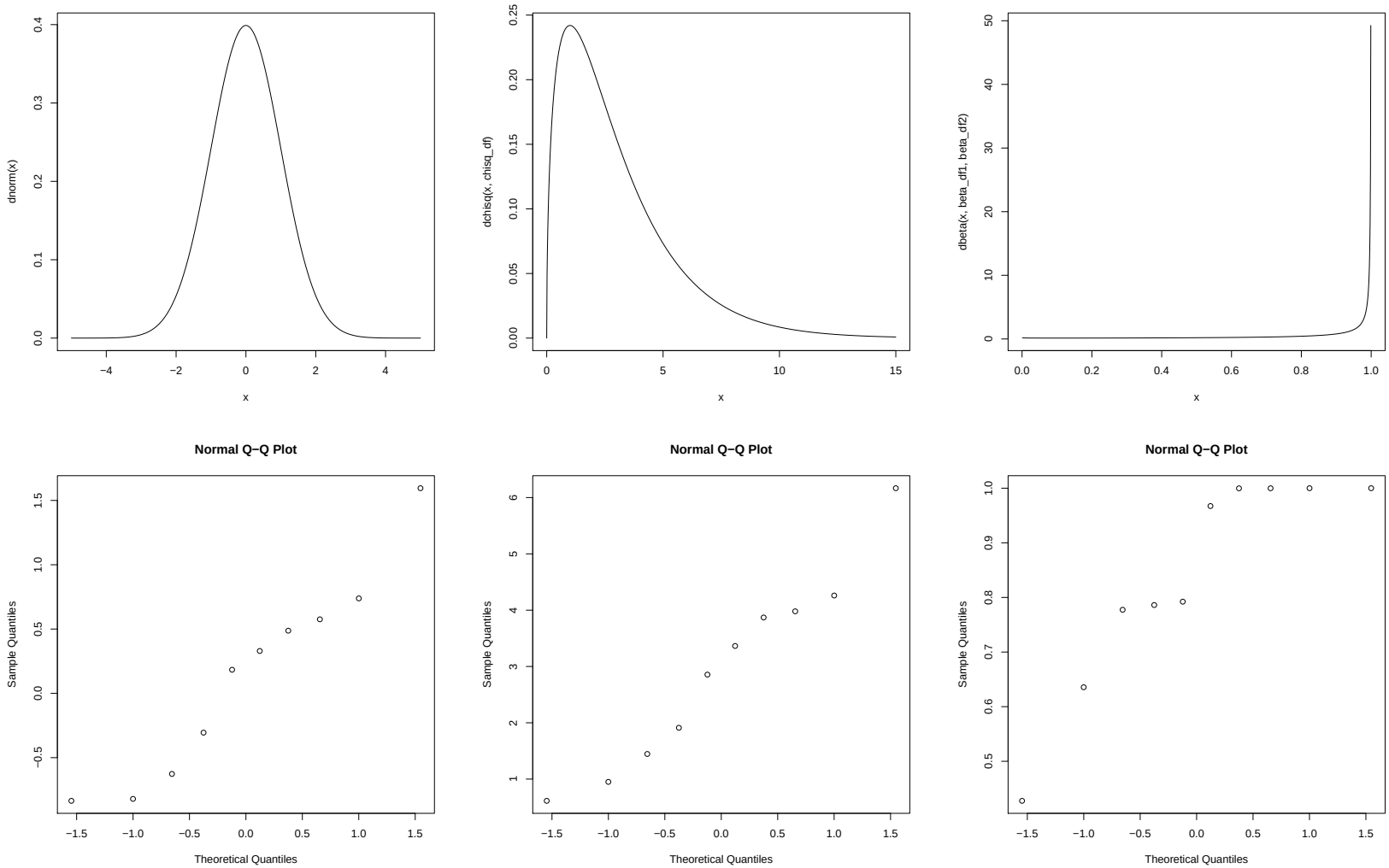


Figure: $n = 10$

Univariate normals

By far the most common approach is to look at a QQ-plot...

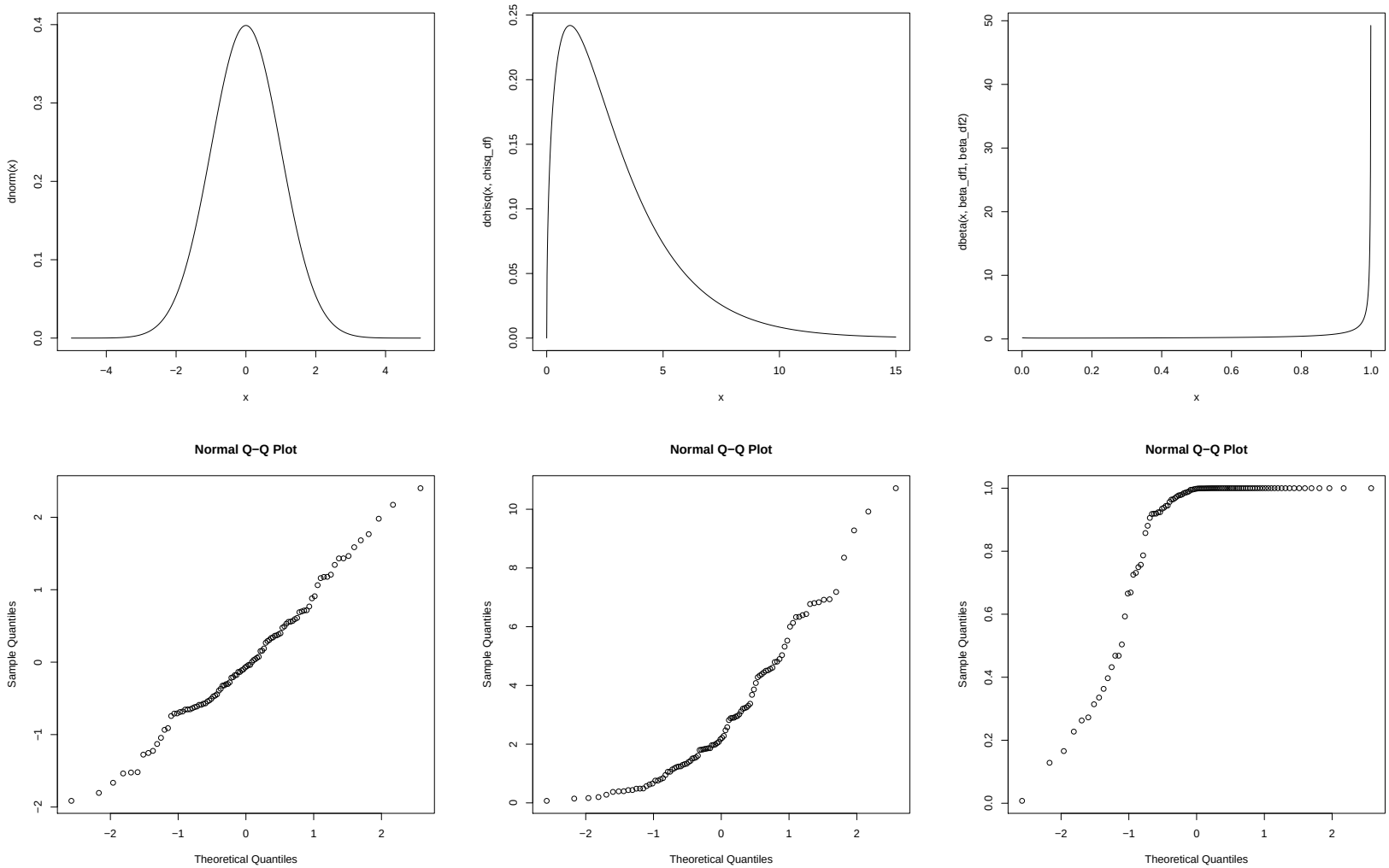


Figure: $n = 100$

Multivariate normals

Checking multivariate normality is more delicate

Reminder: All marginal distributions are Gaussian

It is necessary (but not sufficient) for all the marginal/bivariate distributions must be normal

Hence, checking these lower dimensions is sometimes the best you can do

Multivariate normals

Taking the sample Mahalanobis distance

$$D_i^2 = (X_i - \overline{X})^\top S^{-1}(X_i - \overline{X})$$

Again, a necessary (but not sufficient) condition says that if X is normal then

$$u_i = \frac{nD_i^2}{(n-1)^2} \sim \text{beta}(a = p/2, b = (n-p-1)/2)$$

(The beta distribution is a two parameter family)

Hence a QQplot can be created using the quantiles of the beta distribution

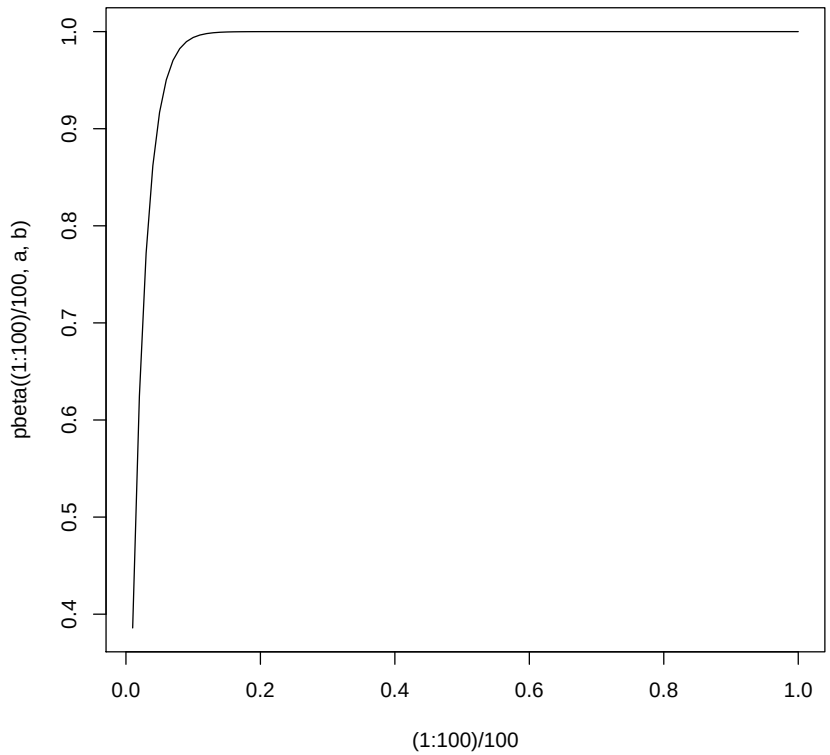
Multivariate normals

To compute a QQplot, we need to compare the empirical quantiles to the theoretical ones

```
alp = (p-2)/(2*p)
bet = (n-p-3)/(2*(n - p -1))

a = p/2
b = (n-p-1)/2

probs = (1:n - alp)/(n - alp - bet + 1)
quantiles = qbeta(probs,a,b)
```

Multivariate normals

The reference beta(a,b) distribution for a problem with $n = 100$ and $p = 2$ is:



Multivariate normals

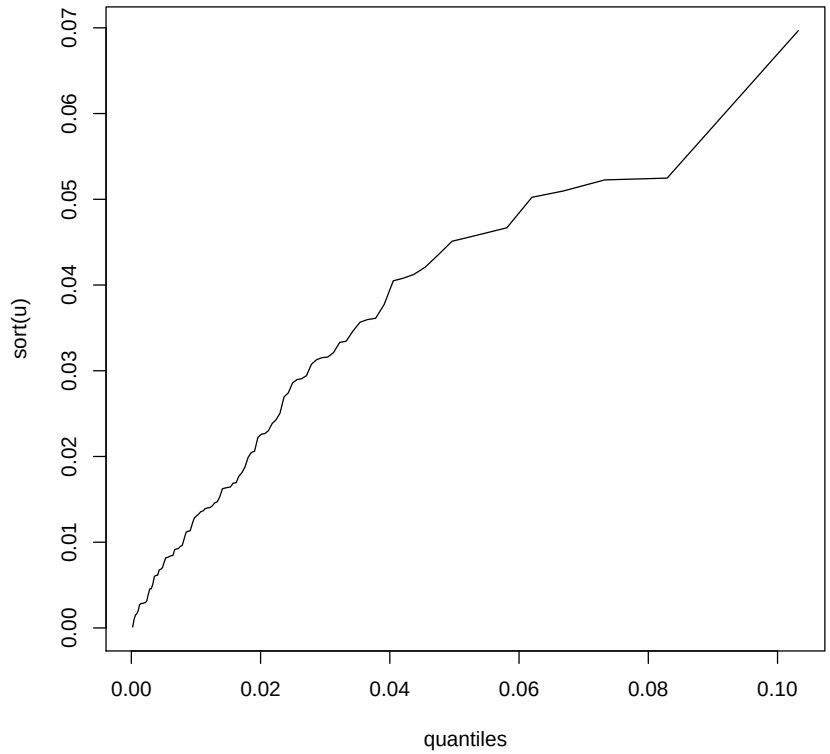
Let's simulation some normal random variables and create the QQplot

```
X        = matrix(rnorm(n * p), nrow= n)
Xcenter = scale(X,center=TRUE,scale=FALSE)

S = 1/(n-1)* t(Xcenter) %*% Xcenter

Dsq = rep(0,n)
for(i in 1:n){
Dsq[i] = t(Xcenter[i,]) %*% solve(S) %*% Xcenter[i,]
}

u    = n * Dsq / (n-1)**2
```

(There are more efficient ways of doing this, I just wanted to make the code as transparent as possible)

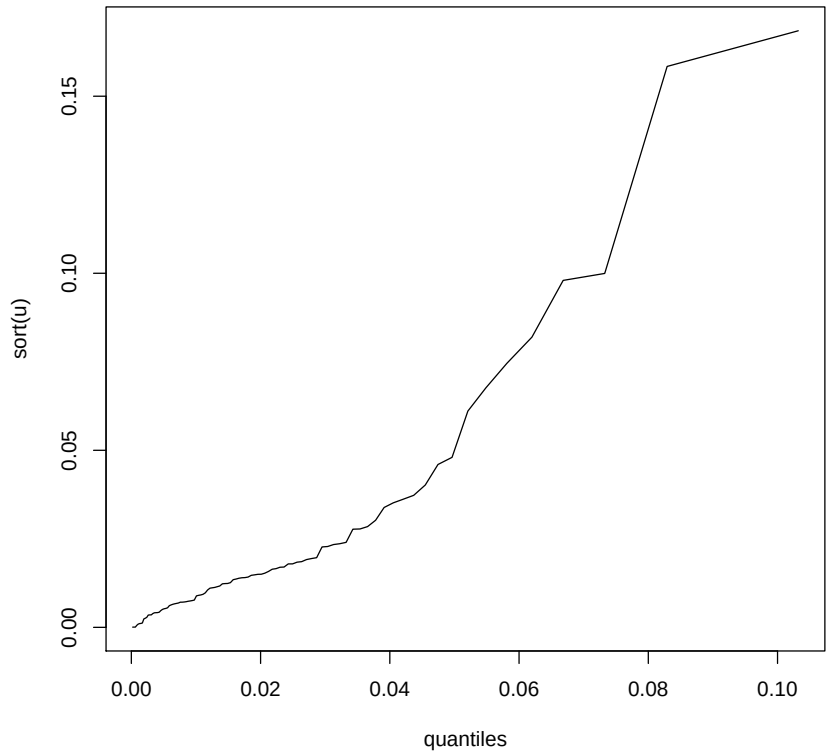
Multivariate normals

For Normal data:



Multivariate normals

For exponential data:



Transformations to normality

Central limit theorem

We already mentioned that for large enough n, the sample mean is normally distributed

Sometimes it is necessary to transform your data so that it looks more normal for smaller n

Box-Cox transformation

The workhorse method for this is the Box-Cox transformation, creating transformed data:

$$X_j^{(\lambda_j)} = \begin{cases} \frac{X_j^{\lambda_j} - 1}{\lambda_j} & \text{for } \lambda \neq 0 \\ \log(X_j) & \text{for } \lambda = 0 \end{cases}$$

where $\lambda = (\lambda_1, \ldots, \lambda_p)$ maximizes:

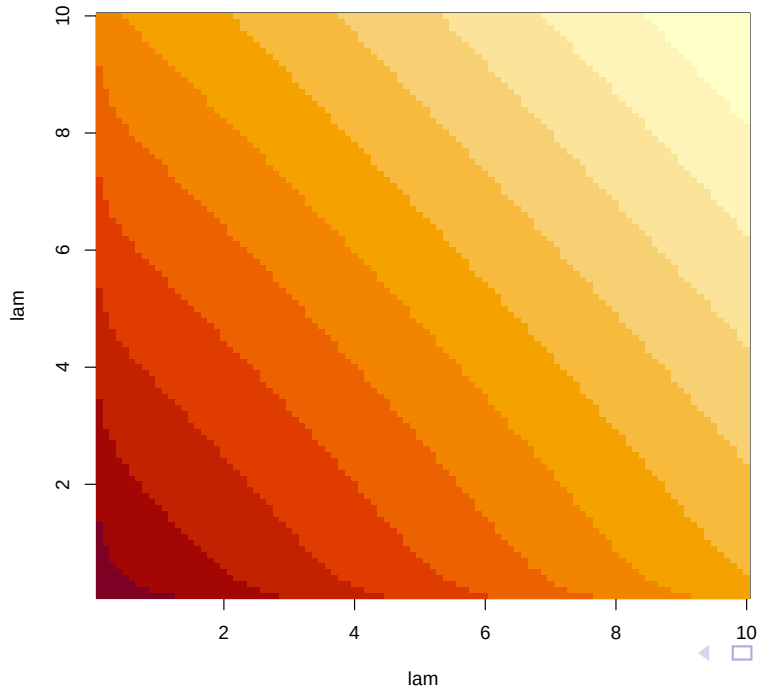
$$\ell(\lambda) = -\frac{n}{2} \log(|S_\lambda|) + \sum_{j=1}^{p} \left[(\lambda_j - 1) \sum_{i=1}^{n} \log(X_{ij}) \right]$$

This can be numerically maximized via e.g. LBFGS or via grid search

Box-Cox transformation

```
##write a logLike function## Exercise!

lam = (1:100)/10
lamGrid = expand.grid(lam,lam)
out = apply(lamGrid,1,logLikeF)
```

Exponential



Outlier detection

Summary of outlier detection

In my experience, the formal testing methods (e.g. MMA 4.6.2) don't work very well in practice

Hence, I want to restrict our attention to the following outlier detection methods

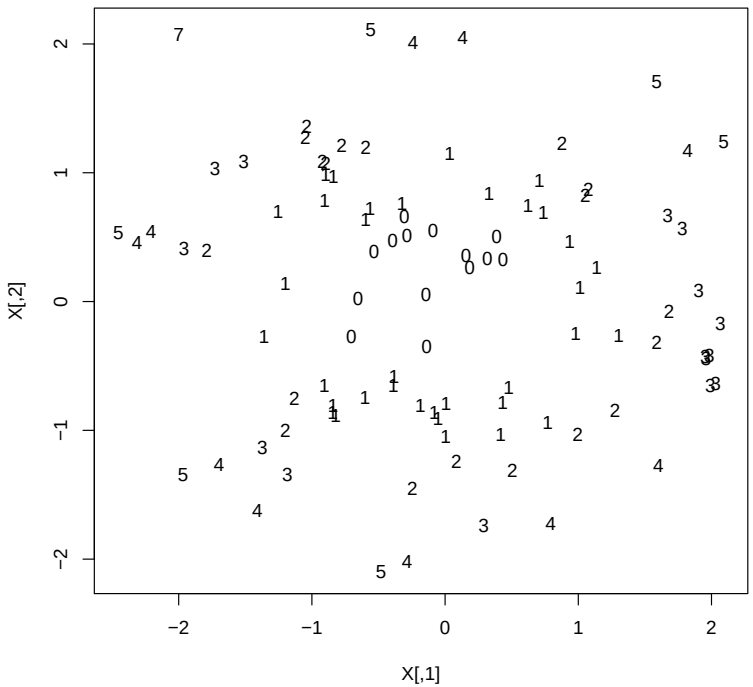
1. Using the sample Mahalanobis distance
2. Principal components analysis
3. Single linkage hierarchical clustering

We will discuss 1. now and leave 2. and 3. for when we talk about the underlying concept later

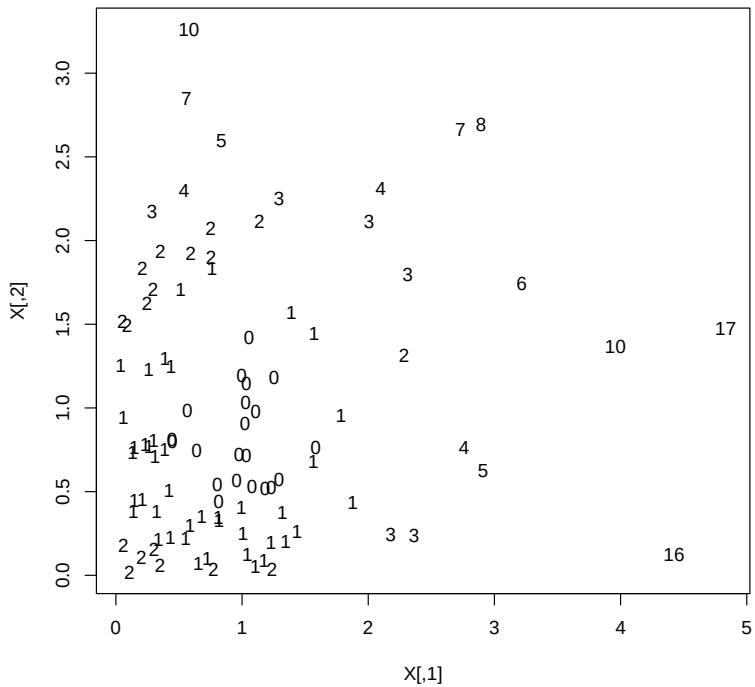
Outliers based on D_i^2

Returning to the previous two examples:

Normal



Exponential

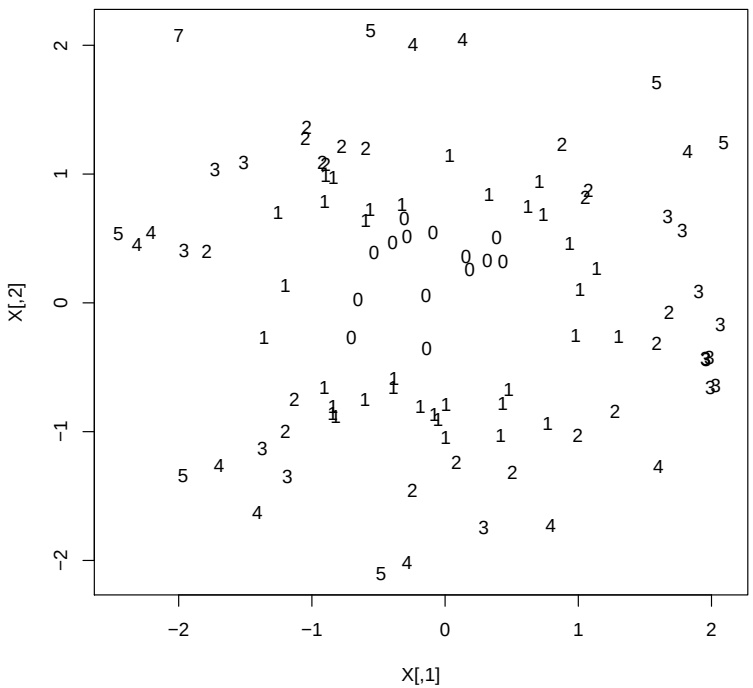


Outliers based on D_i^2

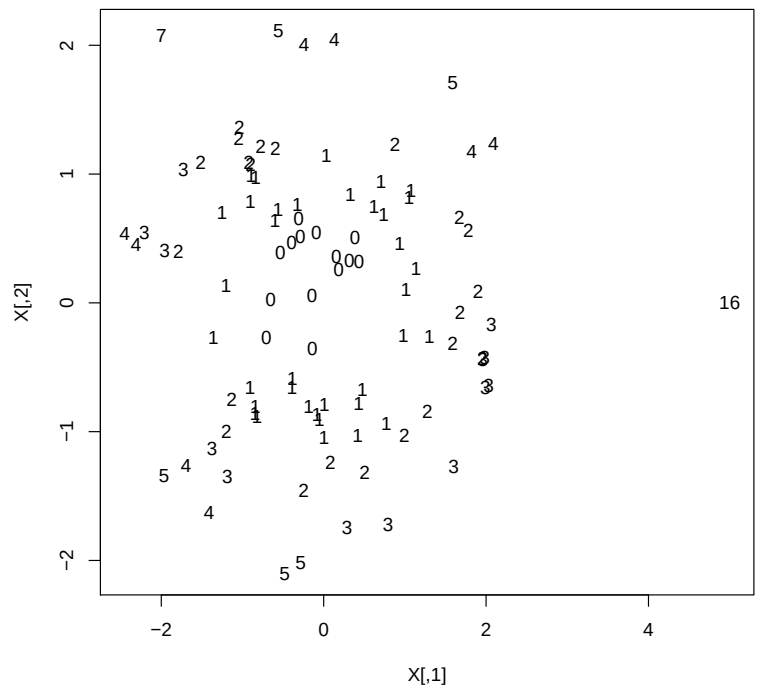
Let's contaminate the data to see what that would look like

```
X        = matrix(rnorm(n * p), nrow= n)
X[100,] = c(5,0)
```

Normal



Contaminated



Postamble:

- Define normal distribution

  (It can be defined in many ways. One major way is as the limit point for the sample mean)
- Give some properties (Most notable: independence is implied by uncorrelated)
- Assessing normality

  (It is necessary for the marginal distributions to be normal. It is also necessary for the sample Mahalanobis distance to be distributed beta)
- Transformations to normality

  (Covered the Box-Cox transformation)
- Outlier detection

  (We informally defined outlier detection via D_i^2. We will return to it later)