Introduction to Statistical Learning

Notes by Tim Brown, Yao-ban Chan and Owen Jones

Module 3 Random Vectors

Contents

1 Random vectors

Random vectors

We must still do some more groundwork in order to analyse our linear models. Once we have done this, the theoretical results come out quite easily!

Normally we think of matrices and vectors as a bunch of numbers. However, they can also be a bunch of random variables.

We can then extend the traditional concepts of expectation, variance, etc. to random vectors.

1.1 Expectation - MM 2.2

Expectation

Traditionally random variables are denoted with capital letters. However we will denote them by lowercase according to linear algebra notation.

We define the expectation of a random vector $\mathbf{y}$ as follows:

$$\text{If} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix}, \quad \text{then} \quad E[\mathbf{y}] = \begin{bmatrix} E[y_1] \\ E[y_2] \\ \vdots \\ E[y_k] \end{bmatrix}.$$

1.2 Expectation properties - MM 2.2

Expectation properties

- If $\mathbf{a}$ is a vector of constants, then $E[\mathbf{a}] = \mathbf{a}$.
- If $\mathbf{a}$ is a vector of constants, then $E[\mathbf{a}^T\mathbf{y}] = \mathbf{a}^T E[\mathbf{y}]$.
- If A is a matrix of constants, then $E[A\mathbf{y}] = AE[\mathbf{y}]$.

Expectation properties

Example. Let

$$A = \begin{bmatrix} 2 & 3 \\ 1 & 4 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

and assume that $E[y_1] = 10$ and $E[y_2] = 20$.

Then

$$AE[\mathbf{y}] = \begin{bmatrix} 2 & 3 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} 10 \\ 20 \end{bmatrix} = \begin{bmatrix} 80 \\ 90 \end{bmatrix}.$$

Expectation properties

On the other hand,

$$\begin{aligned}
E[A\mathbf{y}] &= E\left[\begin{array}{c} 2y_1 + 3y_2 \\ y_1 + 4y_2 \end{array}\right] \\
&= \left[\begin{array}{c} E[2y_1 + 3y_2] \\ E[y_1 + 4y_2] \end{array}\right] \\
&= \left[\begin{array}{c} 2E[y_1] + 3E[y_2] \\ E[y_1] + 4E[y_2] \end{array}\right] \\
&= \left[\begin{array}{c} 80 \\ 90 \end{array}\right] = AE[\mathbf{y}].
\end{aligned}$$

1.3 Variance - MM 2.2

Variance

Defining the variance of a random vector is slightly trickier. We want to not just include the variance of the variables themselves, but also how the variables affect each other.

Recall that the variance of a random variable Y with mean μ is defined to be $E[(Y-\mu)^2]$.

We define the variance (or *covariance matrix*) of the random vector $\mathbf{y}$ to be

$$\text{var } \mathbf{y} = E[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^T]$$

where $\boldsymbol{\mu} = E[\mathbf{y}]$.

Variance

The diagonal elements of the covariance matrix are the variances of the elements of $\mathbf{y}$:

$$[\text{var } \mathbf{y}]_{ii} = \text{var } y_i, \qquad i = 1, 2, \ldots, k.$$

The off-diagonal elements are the covariances of the elements:

$$[\text{var } \mathbf{y}]_{ij} = \text{cov}(y_i, y_j) = E[(y_i - \mu_i)(y_j - \mu_j)].$$

This means that all covariance matrices are symmetric.

1.4 Variance properties - MM 2.2

Variance properties

Suppose that $\mathbf{y}$ is a random vector with var $\mathbf{y} = V$. Then:

- If $\mathbf{a}$ is a vector of constants, then var $\mathbf{a}^T\mathbf{y} = \mathbf{a}^T V \mathbf{a}$.
- If A is a matrix of constants, then var $A\mathbf{y} = AVA^T$.
- V is positive semidefinite.

1.5 Example- variance of a matrix times a random vector

Variance properties - example

Example. Continuing the expectation example

$$A = \begin{bmatrix} 2 & 3 \\ 1 & 4 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

and assume that $E[y_1] = 10$, $E[y_2] = 20$ and

$$V = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

.

Variance properties - example

Then

$$\begin{aligned} \text{var } [A\mathbf{y}] &= \begin{bmatrix} 2 & 3 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 3 & 4 \end{bmatrix} \\ &= \begin{bmatrix} 5 & 5 \\ 5 & 5 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 3 & 4 \end{bmatrix} \\ &= \begin{bmatrix} 25 & 25 \\ 25 & 25 \end{bmatrix}. \end{aligned}$$

1.6 Example - variance of $(X^T X)^{-1} X^T \mathbf{y}$

Variance properties

Example. Let

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

be a random vector with var $y_i = \sigma^2$ for all i, and the elements of $\mathbf{y}$ are independent.

Then the covariance matrix of $\mathbf{y}$ is

$$\text{var } \mathbf{y} = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix} = \sigma^2 I.$$

Variance properties

Example. Assume that X is a matrix of full rank (with more rows than columns), which implies that $X^T X$ is nonsingular. Let

$$\mathbf{z} = (X^T X)^{-1} X^T \mathbf{y} = A\mathbf{y}.$$

Then

$$\begin{aligned}
\text{var } \mathbf{z} &= AVA^T = [(X^TX)^{-1}X^T]\sigma^2 I[(X^TX)^{-1}X^T]^T \\
&= (X^TX)^{-1}X^T(X^T)^T[(X^TX)^{-1}]^T\sigma^2 \\
&= (X^TX)^{-1}X^TX[(X^TX)^T]^{-1}\sigma^2 \\
&= (X^TX)^{-1}\sigma^2.
\end{aligned}$$

We will be using this quite a bit later on!

1.7 Matrix Square Root - not in MM

Matrix square root

The square root of a matrix A is a matrix B such that $B^2 = A$. In general the square root is not unique.

If A is symmetric and positive semidefinite, there is a unique symmetric positive semidefinite square root, called the principal root, denoted $A^{1/2}$.

Suppose that P diagonalises A, that is $P^TAP = \Lambda$. Then

$$\begin{aligned}
A &= P\Lambda P^T \\
&= P\Lambda^{1/2}\Lambda^{1/2}P^T \\
&= P\Lambda^{1/2}P^TP\Lambda^{1/2}P^T \\
A^{1/2} &= P\Lambda^{1/2}P^T.
\end{aligned}$$

Multivariate normal distribution

1.8 Multivariate Normal Distribution - MM 2.3

Definition 3.1. Let $\mathbf{z}$ be a $k \times 1$ vector of independent standard normal random variables, A an $n \times k$ matrix, and $\mathbf{b}$ an $n \times 1$ vector. We say that

$$\mathbf{x} = A\mathbf{z} + \mathbf{b}$$

has (an n-dimensional) multivariate normal distribution, with mean $\boldsymbol{\mu} = \mathbf{b}$ and covariance matrix $\Sigma = AA^T$.

We write $\mathbf{x} \sim MVN(\boldsymbol{\mu}, \Sigma)$ or just $\mathbf{x} \sim N(\boldsymbol{\mu}, \Sigma)$.

Multivariate normal distribution

For any $\boldsymbol{\mu}$ and any symmetric positive semidefinite matrix Σ, let $\mathbf{z}$ be a vector of independent standard normals. Then

$$\boldsymbol{\mu} + \Sigma^{1/2}\mathbf{z} \sim MVN(\boldsymbol{\mu}, \Sigma).$$

If $\mathbf{x} \sim MVN(\boldsymbol{\mu}, \Sigma)$ and Σ is $k \times k$ positive definite, then $\mathbf{x}$ has the density

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}.$$

Compare this with the univariate normal density

$$f(x) = \frac{1}{(2\pi)^{1/2}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}.$$

1.9 Linear combination of MVN rvs is MVN - not in MM

Multivariate normal distribution

Any linear combination of multivariate normals results in another multivariate normal: if $\mathbf{x} \sim MVN(\boldsymbol{\mu}, \Sigma)$ is $k \times 1$, A is $n \times k$, and $\mathbf{b}$ is $n \times 1$, then

$$\mathbf{y} = A\mathbf{x} + \mathbf{b} \sim MVN(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^T).$$

To see why, put $\mathbf{x} = \Sigma^{1/2}\mathbf{z} + \boldsymbol{\mu}$, then

$$\mathbf{y} = A\Sigma^{1/2}\mathbf{z} + A\boldsymbol{\mu} + \mathbf{b}.$$

1.10 Uncorrelated vs. independence for normally distributed random variables - not in MM

Multivariate normal distribution

If $\mathbf{z} = (z_1, z_2)^T$ is multivariate normal, then z_1 and z_2 are independent if and only if they are uncorrelated.

In general, if z_1 and z_2 are normal random variables, $\mathbf{z} = (z_1, z_2)^T$ does not have to be multivariate normal. Moreover, z_1 and z_2 can be uncorrelated but not independent.

For example, suppose that $z_1 \sim N(0,1)$ and $u \sim U(-1,1)$, then $z_2 = z_1 \text{sign}(u) \sim N(0,1)$, but $\mathbf{z} = (z_1, z_2)^T$ is *not* multivariate normal. (Consider its support.) Moreover z_1 and z_2 are uncorrelated, but clearly dependent.

1.11 Simulating MVN random vectors- not in MM

Multivariate normals in R

To generate a sample of size 100 with distribution

$$MVN\left(\begin{bmatrix} 3 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}\right)$$

```
library(MASS)
a <- matrix(c(3, 1), 2, 1)
V <- matrix(c(1, .8, .8, 1), 2, 2)
y <- mvrnorm(100, mu = a, Sigma = V)
plot(y[,1], y[,2])
```

Multivariate normals in R

Alternatively, starting with standard normals:

```
P <- eigen(V)$vectors
sqrtV <- P %*% diag(sqrt(eigen(V)$values)) %*% t(P)
z <- matrix(rnorm(200), 2, 100)
y_new <- sqrtV %*% z + rep(a, 100)
plot(y[,1], y[,2])
points(y_new[1,], y_new[2,], col = "red")
```

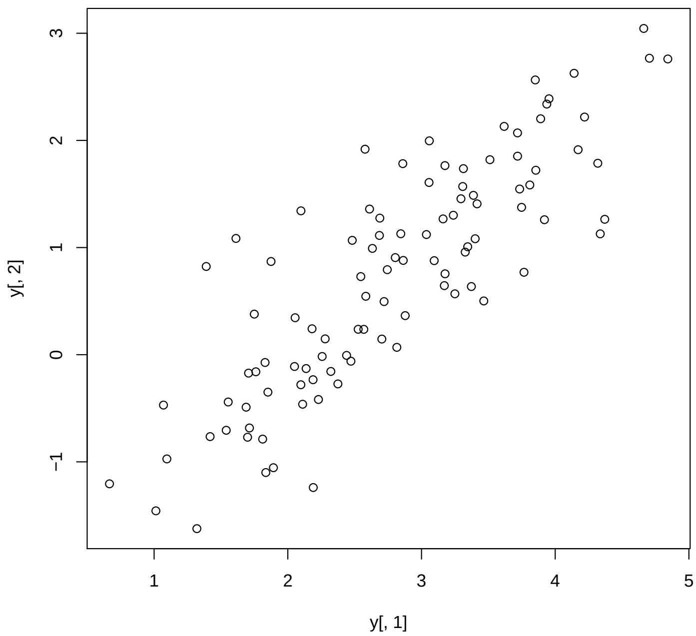

Figure 1: 100 bivariate normal random vectors using `mvrnorm`

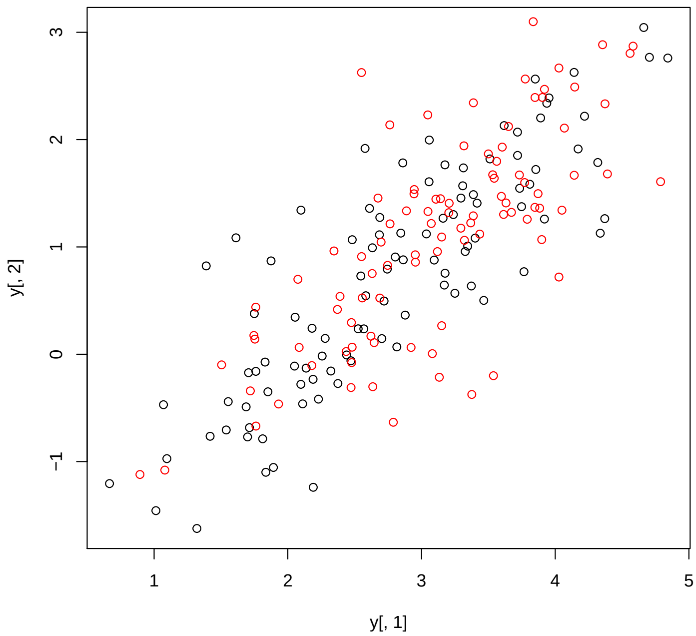

Figure 2: 100 extra with same distribution in red using definition and `rnorm`

2 Random quadratic forms

Random quadratic forms

We have seen that a matrix induces a quadratic form (multivariate function).

What happens when the variables in a quadratic form are random variables?

The form becomes a scalar function of random variables, so it is itself a random variable.

Random quadratic forms will pop up regularly in our theory of linear models - the sample variance for a random sample is an example. To analyse our models, their distributions are needed.

2.1 Expectation - MM 2.2

Random quadratic forms

Theorem 3.2. *Let* $\mathbf{y}$ *be a random vector with* $E[\mathbf{y}] = \boldsymbol{\mu}$ *and var* $\mathbf{y} = V$*, and let* A *be a matrix of constants. Then*

$$E[\mathbf{y}^T A\mathbf{y}] = tr(AV) + \boldsymbol{\mu}^T A\boldsymbol{\mu}.$$

Remark. Rearranging gives $tr(AV) = E[\mathbf{y}^T A\mathbf{y}] - \boldsymbol{\mu}^T A\boldsymbol{\mu}$, generalising $\text{cov}(ay, bz) = ab(E(yz) - E(y)E(z))$ for random variables y, z and constants a, b. The proof is based on this.

Proof. We denote the (i, j)th element of the matrix V by σ_{ij}. For $i \neq j$,

$$\sigma_{ij} = \text{cov}(y_i, y_j) = E[y_i y_j] - \mu_i \mu_j.$$

and for $i = j$

$$\sigma_{ii} = \text{var } y_i = E[y_i^2] - \mu_i^2.$$

Random quadratic forms

Suppose the length of $\mathbf{y}$ is k.

$$\begin{aligned} E[\mathbf{y}^T A\mathbf{y}] &= E\left[\sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij} y_i y_j\right] \\ &= \sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij} E[y_i y_j] \\ &= \sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}(\sigma_{ij} + \mu_i \mu_j) \\ &= \sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}\sigma_{ji} + \sum_{i=1}^{k}\sum_{j=1}^{k} \mu_i a_{ij} \mu_j \\ &= \sum_{i=1}^{k} [AV]_{ii} + \boldsymbol{\mu}^T A\boldsymbol{\mu} \\ &= tr(AV) + \boldsymbol{\mu}^T A\boldsymbol{\mu}. \end{aligned}$$

2.2 Example - expectation and variance of a random quadratic form

Random quadratic forms

Example. Let $\mathbf{y}$ be a 2×1 random vector with

$$\boldsymbol{\mu} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \quad V = \begin{bmatrix} 2 & 1 \\ 1 & 5 \end{bmatrix}.$$

Let

$$A = \begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix}.$$

Consider the quadratic form

$$\mathbf{y}^T A\mathbf{y} = 4y_1^2 + 2y_1 y_2 + 2y_2^2.$$

Random quadratic forms

The expectation of this form is

$$E[\mathbf{y}^T A\mathbf{y}] = 4E[y_1^2] + 2E[y_1 y_2] + 2E[y_2^2].$$

From the given covariance matrix,

$$\begin{aligned} 2 &= \text{var } y_1 = E[y_1^2] - E[y_1]^2 = E[y_1^2] - 1 \\ 5 &= \text{var } y_2 = E[y_2^2] - E[y_2]^2 = E[y_2^2] - 9 \end{aligned}$$

so $E[y_1^2] = 3$ and $E[y_2^2] = 14$. Also

$$1 = \text{cov}(y_1, y_2) = E[y_1 y_2] - E[y_1]E[y_2] = E[y_1 y_2] - 3$$

so $E[y_1 y_2] = 4$.

Random quadratic forms

This gives

$$E[\mathbf{y}^T A\mathbf{y}] = 4 \times 3 + 2 \times 4 + 2 \times 14 = 48.$$

But from Theorem 3.2,

$$\begin{aligned} E[\mathbf{y}^T A\mathbf{y}] &= tr(AV) + \boldsymbol{\mu}^T A\boldsymbol{\mu} \\ &= tr\left(\begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} 2 & 1 \\ 1 & 5 \end{bmatrix}\right) + \begin{bmatrix} 1 & 3 \end{bmatrix}\begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} 1 \\ 3 \end{bmatrix} \\ &= tr\left(\begin{bmatrix} 9 & 9 \\ 4 & 11 \end{bmatrix}\right) + \begin{bmatrix} 1 & 3 \end{bmatrix}\begin{bmatrix} 7 \\ 7 \end{bmatrix} \\ &= 9 + 11 + 7 + 21 = 48. \end{aligned}$$

Random quadratic forms

```
mu <- c(1,3)
V <- matrix(c(2,1,1,5),2,2)
y <- t(mvrnorm(100, mu = mu, Sigma = V))
A <- matrix(c(4,1,1,2),2,2)
sum(diag(A%*%V)) + t(mu)%*%A%*%mu

##      [,1]
## [1,]   48

quadform <- function(y, A) t(y) %*% A %*% y
mean(apply(y, 2, quadform, A = A))

## [1] 54.76954
```

2.3 Noncentral χ^2 distribution - MM 2.3

Noncentral χ^2 distribution

Definition 3.3. Let $\mathbf{y} \sim MVN(\boldsymbol{\mu}, I)$ be a $k \times 1$ random vector. Then

$$x = \mathbf{y}^T \mathbf{y} = \sum_{i=1}^{k} y_i^2$$

has a *noncentral* χ^2 *distribution* with k degrees of freedom and noncentrality parameter $\lambda = \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\mu}$. We write $x \sim \chi^2_{k,\lambda}$.

Warning: some authors (including R!) define λ to be $\boldsymbol{\mu}^T\boldsymbol{\mu}$.
Note that the distribution of x depends on $\boldsymbol{\mu}$ only through λ.

Noncentral χ^2 distribution
Example. Let

$$\mathbf{y} \sim MVN\left(\begin{bmatrix} 4 \\ 2 \\ -2 \end{bmatrix}, I_3\right).$$

Then $y_1^2 + y_2^2 + y_3^2$ has a noncentral χ^2 distribution with 3 degrees of freedom and noncentrality parameter

$$\lambda = \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\mu} = \frac{1}{2}\begin{bmatrix} 4 & 2 & -2 \end{bmatrix}\begin{bmatrix} 4 \\ 2 \\ -2 \end{bmatrix} = 12.$$

Noncentral χ^2 distribution
Suppose $\mathbf{y} \sim MVN(\boldsymbol{\mu}, I_k)$ and $x = \mathbf{y}^T\mathbf{y} \sim \chi^2_{k,\lambda}$. Then from Theorem 3.2,

$$E[x] = tr(I_k) + \boldsymbol{\mu}^T\boldsymbol{\mu} = k + 2\lambda.$$

The noncentrality parameter $\lambda = \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\mu}$ is zero if and only if $\boldsymbol{\mu} = \mathbf{0}$, in which case x is just the sum of squares of k independent standard normals.
In this case, we already know that x has an ordinary (central) χ^2 distribution with k degrees of freedom.

Noncentral χ^2 distribution
The density of the noncentral χ^2 distribution is complicated; one expression is

$$f(x; k, \lambda) = \sum_{i=0}^{\infty} \frac{e^{-\lambda}\lambda^i}{i!} g(x; k + 2i)$$

where g is the density of the χ^2 distribution

$$g(x; k) = \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} x^{k/2-1} e^{-x/2}.$$

The density is a weighted sum of χ^2 densities whose weights are Poisson probabilities. This can be used to show that if $x \sim \chi^2_{k,\lambda}$, then

$$\text{var } x = 2k + 8\lambda.$$

Some densities are shown in Figure 3.

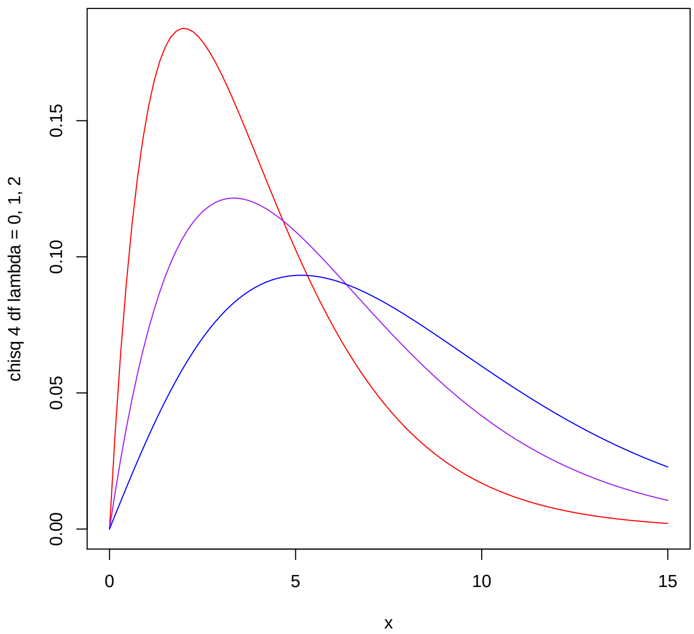

Figure 3: Densities of chisquare with 4 df and $\lambda = 0, 1, 2$

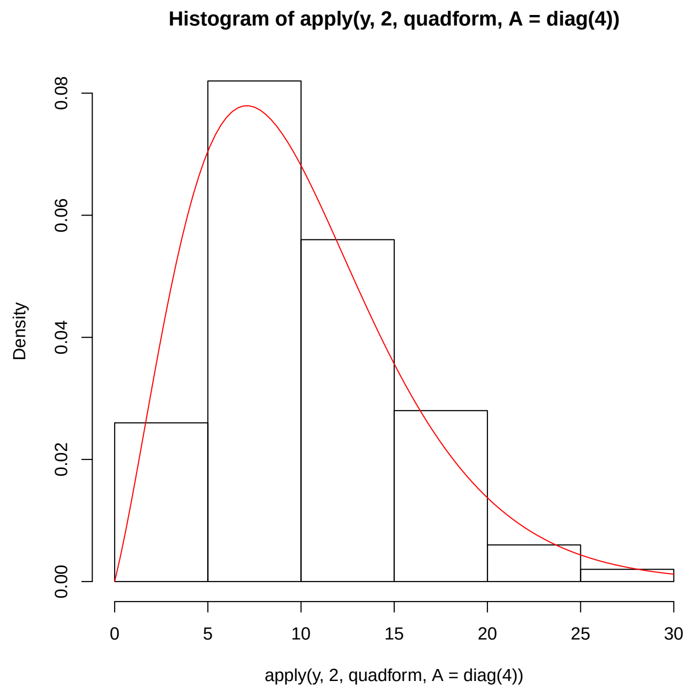

Figure 4: 100 Chisq rvs with df = 4, $\lambda = 3$

```
y <- t(mvrnorm(100, mu = c(1,2,1,0), Sigma = diag(4)))
hist(apply(y,2,quadform,A=diag(4)), freq=FALSE)
curve(dchisq(x,4,6), add=TRUE, col="red")
```

Histogram and density shown in Figure 4.

2.4 Adding noncentral χ^2 rvs - MM 2.3

Adding noncentral χ^2s

Theorem 3.4. *Let $X^2_{k_1,\lambda_1}, \ldots, X^2_{k_n,\lambda_n}$ be a collection of n independent noncentral χ^2 random variables, with $X^2_{k_i,\lambda_i} \sim \chi^2_{k_i,\lambda_i}$. Then*

$$\sum_{i=1}^{n} X^2_{k_i,\lambda_i}$$

has a noncentral χ^2 distribution with $\sum_{i=1}^n k_i$ degrees of freedom and noncentrality parameter $\sum_{i=1}^n \lambda_i$.

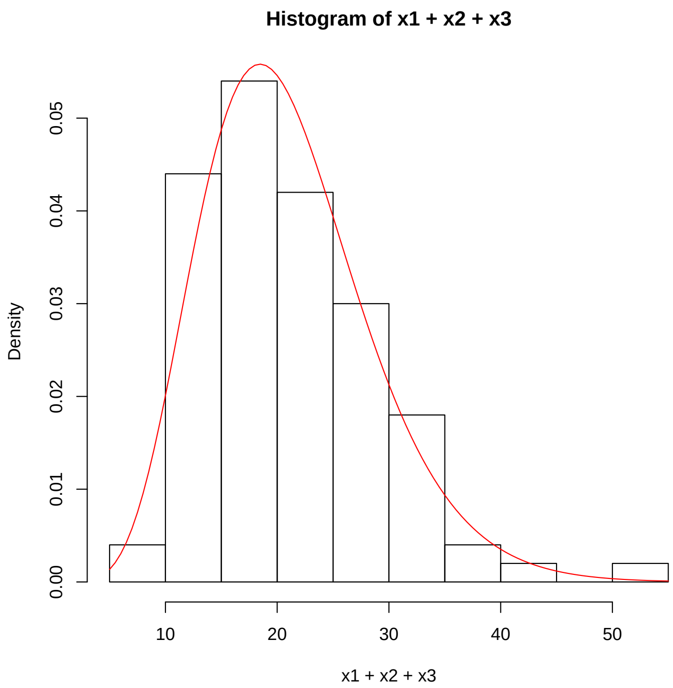

Figure 5: Adding 100 noncentral χ^2 rvs

If we set $\lambda_i = 0$ for all i, we get the result from MAST90105 that the sum of independent χ^2 variables is another χ^2 variable whose degrees of freedom is the sum of the degrees of freedom of the components. This result can be proved using moment generating functions.

Adding noncentral χ^2s

```
x1 <- rchisq(100,3,0)
x2 <- rchisq(100,6,2)
x3 <- rchisq(100,5,5)
hist(x1+x2+x3,freq=FALSE)
curve(dchisq(x,14,7),add=TRUE,col='red')
```

Histogram and density are shown in Figure 5.

2.5 Distribution of quadratic form for MVN - MM 2.3

Theorem 3.5. *Let* $\mathbf{y} \sim MVN(\boldsymbol{\mu}, I)$ *be a* $n \times 1$ *random vector and let* A *be a* $n \times n$ *symmetric matrix. Then* $\mathbf{y}^T A\mathbf{y}$ *has a noncentral* χ^2 *distribution with* k *degrees of freedom and noncentrality parameter* $\lambda = \frac{1}{2}\boldsymbol{\mu}^T A\boldsymbol{\mu}$ *if and only if* A *is idempotent and has rank* k*.*

Proof. We only demonstrate $\Leftarrow$ as this is the one that will be most important for us. Assume that A is idempotent and has rank k. Because it is symmetric, it can be diagonalised. Let the (orthogonal) diagonalising matrix be P.

Since A is symmetric and idempotent, all its eigenvalues are either 0 or 1 (Theorem 2.2). Moreover, Theorem 2.3 tells us that the sum of the eigenvalues is

$$tr(A) = r(A) = k.$$

Therefore, A must have k eigenvalues of 1 and $n-k$ eigenvalues of 0.

Recalling that the columns of P are the unit length eigenvectors corresponding to the eigenvalues, it is possible to arrange the columns of P so that all the 1 eigenvalues are first. Then we can partition the diagonalisation of A as

$$P^T A P = \left[\begin{array}{c|c} I_k & 0 \\ \hline 0 & 0 \end{array}\right].$$

Now define the random vector $\mathbf{z} = P^T\mathbf{y} \sim MVN(P^T\boldsymbol{\mu}, I_n)$. Partition the vectors/matrices as

$$\mathbf{z} = \left[\begin{array}{c} \mathbf{z}_1 \\ \hline \mathbf{z}_2 \end{array}\right], \quad P = \left[\begin{array}{c|c} P_1 & P_2 \end{array}\right]$$

where $\mathbf{z}_1$ is $k \times 1$ and P_1 is $n \times k$.

Then $\mathbf{z}_1 = P_1^T\mathbf{y} \sim MVN(P_1^T\boldsymbol{\mu}, I_k)$ and

$$\begin{aligned}
\mathbf{y}^T A\mathbf{y} &= (P\mathbf{z})^T A(P\mathbf{z}) = \mathbf{z}^T P^T AP\mathbf{z} \\
&= \left[\begin{array}{c|c} \mathbf{z}_1^T & \mathbf{z}_2^T \end{array}\right] \left[\begin{array}{c|c} I_k & 0 \\ \hline 0 & 0 \end{array}\right] \left[\begin{array}{c} \mathbf{z}_1 \\ \hline \mathbf{z}_2 \end{array}\right] \\
&= \left[\begin{array}{c|c} \mathbf{z}_1^T & \mathbf{z}_2^T \end{array}\right] \left[\begin{array}{c} \mathbf{z}_1 \\ \hline 0 \end{array}\right] \\
&= \mathbf{z}_1^T\mathbf{z}_1.
\end{aligned}$$

Therefore, $\mathbf{y}^T A\mathbf{y} = \mathbf{z}_1^T\mathbf{z}_1$ has a noncentral χ^2 distribution with k degrees of freedom and noncentrality parameter

$$\lambda = \frac{1}{2}\boldsymbol{\mu}^T P_1 P_1^T\boldsymbol{\mu}.$$

Since

$$\begin{aligned}
A &= \left[\begin{array}{c|c} P_1 & P_2 \end{array}\right] \left[\begin{array}{c|c} I & 0 \\ \hline 0 & 0 \end{array}\right] \left[\begin{array}{c} P_1^T \\ \hline P_2^T \end{array}\right] \\
&= \left[\begin{array}{c|c} P_1 & P_2 \end{array}\right] \left[\begin{array}{c} P_1^T \\ \hline 0 \end{array}\right] \\
&= P_1 P_1^T,
\end{aligned}$$

we have

$$\lambda = \frac{1}{2}\boldsymbol{\mu}^T A\boldsymbol{\mu}.$$

1. The following remarks are only intended to help paint a picture of the underlying geometry - the calculations above just use the linear algebra results without the geometric picture.

2. A matrix is symmetric and idempotent if, and only if, it is an (orthogonal) projection on to some subspace (eg the x-axis, the xy- plane in 3 space or, more generally, a line or a plane in 3 space).

3. For a wonderfully geometric discussion of this, see Chapters 1 to 3 of Gilbert Strang's book Linear Algebra and its Applications leading to pp. 153 to 162. This gives a geometric view of some key ideas in Module 4.

4. The result in 2 is on page 158 of Strang.

5. Recall that a basis is a set of linearly independent vectors for which every vector is a linear combination of basis vectors. For example, the columns of an orthogonal matrix are a basis.

6. A matrix times a vector is a linear transformation. The matrix of a linear transformation is different with different bases. The symmetric matrix, A, changes to $P^T AP = \Lambda$, if you change the basis to the columns of P (See Strang p.305).

7. If the matrix is idempotent as well, it is a projection onto the space which is linear combinations of the columns of P which have unit eigenvalues (the rest are 0).

8. The columns of P are a basis in which we are projecting on to the k axes corresponding to the columns of P with unit eigenvalues.

Corrolaries

Corollary 3.6. *Let $\mathbf{y} \sim MVN(\mathbf{0}, I_n)$ be a $n \times 1$ random vector and let A be a $n \times n$ symmetric matrix. Then $\mathbf{y}^T A\mathbf{y}$ has a (central) χ^2 distribution with k degrees of freedom if and only if A is idempotent and has rank k.*

Corollary 3.7. *Let $\mathbf{y} \sim MVN(\boldsymbol{\mu}, \sigma^2 I_n)$ be a $n \times 1$ random vector and let A be a $n \times n$ symmetric matrix. Then $\frac{1}{\sigma^2}\mathbf{y}^T A\mathbf{y}$ has a noncentral χ^2 distribution with k degrees of freedom and noncentrality parameter $\lambda = \frac{1}{2\sigma^2}\boldsymbol{\mu}^T A\boldsymbol{\mu}$ if and only if A is idempotent and has rank k.*

Example

Example. Let y_1 and y_2 be independent normal random variables with means 3 and -2 respectively and variance 1. Let

$$A = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

It is easy to verify that A is symmetric and idempotent, and has rank 1. Therefore

$$\mathbf{y}^T A\mathbf{y} = \frac{1}{2}\begin{bmatrix} y_1 & y_2 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \frac{1}{2}y_1^2 + y_1 y_2 + \frac{1}{2}y_2^2$$

has a noncentral χ^2 distribution with 1 degree of freedom and noncentrality parameter

$$\lambda = \frac{1}{4}\begin{bmatrix} 3 & -2 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 3 \\ -2 \end{bmatrix} = \frac{1}{4}.$$

Distribution of quadratic forms

What happens if $\mathbf{y}$ does not have variance I?

Theorem 3.8. *Let $\mathbf{y} \sim MVN(\boldsymbol{\mu}, V)$ be a $n \times 1$ random vector, and let A be a $n \times n$ symmetric matrix. Then $\mathbf{y}^T A\mathbf{y}$ has a noncentral χ^2 distribution with k degrees of freedom and noncentrality parameter $\lambda = \frac{1}{2}\boldsymbol{\mu}^T A\boldsymbol{\mu}$ if and only if AV is idempotent and has rank k.*

Corollaries

Corollary 3.9. *Let $\mathbf{y} \sim MVN(\mathbf{0}, V)$ be a $n \times 1$ random vector and let A be a $n \times n$ symmetric matrix. Then $\mathbf{y}^T A\mathbf{y}$ has a (central) χ^2 distribution with k degrees of freedom if and only if AV is idempotent and has rank k.*

Corollary 3.10. *Let $\mathbf{y} \sim MVN(\boldsymbol{\mu}, V)$ be a $n \times 1$ random vector where the variance V is nonsingular. Then $\mathbf{y}^T V^{-1}\mathbf{y}$ has a noncentral χ^2 distribution with n degrees of freedom and noncentrality parameter $\lambda = \frac{1}{2}\boldsymbol{\mu}^T V^{-1}\boldsymbol{\mu}$.*

R example: noncentral χ^2

Let

$$\mathbf{y} \sim MVN\left(\mathbf{a} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}, V = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}\right).$$

```
y <- t(mvrnorm(100, mu=a, Sigma=V))
x <- apply(y, 2, quadform, A = solve(V))
(lambda <- t(a) %*% solve(V) %*% a / 2)

##           [,1]
## [1,] 2.277778

mean(x)

## [1] 6.56662
```

R example ctd: noncentral χ^2

```
2 + 2*lambda

##           [,1]
## [1,] 6.555556
```

```
hist(x, freq=F)
curve(dchisq(x, 2, 2*lambda), add = TRUE, col='red')
```

100 Bivariate normals generating non-central χ^2 rv's with density shown in Figure 6.

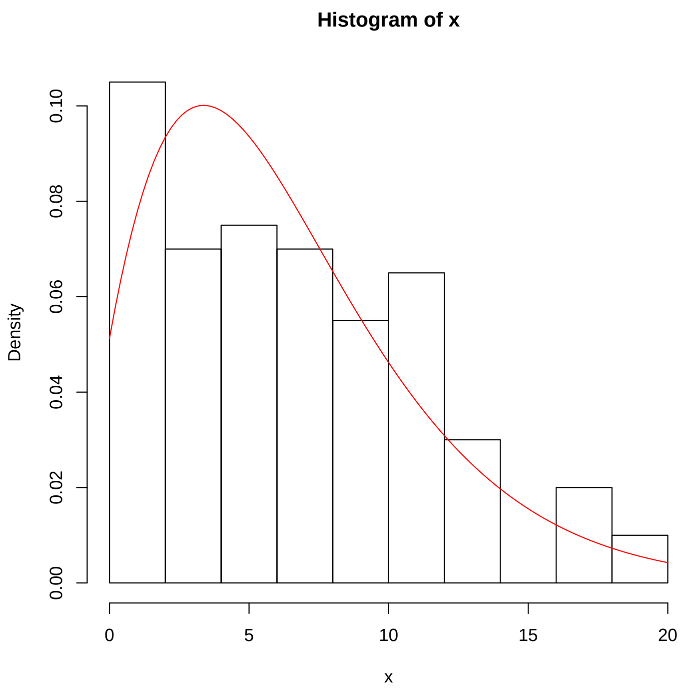

Figure 6: 100 Bivariate Normal Generate Noncentral χ^2

Noncentral χ^2

Example. Let y_1 and y_2 follow a multivariate normal distribution with means -1 and 4 respectively, and covariance matrix $V = \begin{bmatrix} 3 & 2 \\ 2 & 2 \end{bmatrix}$. Then

$$V^{-1} = \frac{1}{3 \times 2 - 2 \times 2} \begin{bmatrix} 2 & -2 \\ -2 & 3 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ -1 & 3/2 \end{bmatrix},$$

and the quadratic form

$$\mathbf{y}^T V^{-1} \mathbf{y} = y_1^2 - 2y_1 y_2 + \frac{3}{2} y_2^2$$

has a noncentral χ^2 distribution with 2 degrees of freedom and noncentrality parameter

$$\lambda = \frac{1}{2} \begin{bmatrix} -1 & 4 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -1 & 3/2 \end{bmatrix} \begin{bmatrix} -1 \\ 4 \end{bmatrix} = \frac{33}{2}.$$

3 Independence

3.1 Independence of quadratic forms - MM 2.4

Independence of quadratic forms

Sometimes we will want to know when two quadratic forms are independent. The next theorem tells us when this happens.

Theorem 3.11. *Let $\mathbf{y} \sim MVN(\boldsymbol{\mu}, V)$ be a $n \times 1$ random vector with nonsingular variance V, and let A and B be symmetric $n \times n$ matrices. Then $\mathbf{y}^T A \mathbf{y}$ and $\mathbf{y}^T B \mathbf{y}$ are independent if and only if*

$$AVB = 0.$$

Proof. Again we only prove $\Leftarrow$ because this is most important for us. Suppose that $AVB = 0$. Since V is symmetric and positive definite we have $V = C^2$ for some symmetric, positive-definite C (see Section 1.7 of Module 3), thus $ACCB = 0$.

Let

$$R = CAC, \quad S = CBC,$$

then

$$RS = CACCBC = 0.$$

Because A, B, and C are symmetric, R and S are also symmetric. Thus

$$SR = S^T R^T = (RS)^T = 0 = RS.$$

By Theorem 2.4, we can find an orthogonal matrix P which diagonalises R and S simultaneously.

Since C is nonsingular, $r(R) = r(CAC) = r(A)$. Call the common rank r.

Thus

$$P^T R P = \left[\begin{array}{c|c} D_1 & 0 \\ \hline 0 & 0 \end{array} \right]$$

where D_1 has dimension $r \times r$.

The first r diagonal entries in D_1 are the non-zero eigenvalues of R and the first r columns of P are the corresponding unit length eigenvectors.

The remaining $n - r$ diagonal entries are 0 since the rank, r, is the number of non-zero eigenvalues. The second $n - r$ columns of P are orthogonal unit length eigenvectors corresponding to the eigenvalue 0 for R.

Suppose $\mathbf{x}$ is one of the first r columns of P, so that $R\mathbf{x} = \lambda\mathbf{x}$ for some non-zero λ.

Then

$$\begin{aligned} S\mathbf{x} &= \frac{1}{\lambda} S\lambda\mathbf{x} \\ &= \frac{1}{\lambda} SR\mathbf{x} \\ &= \mathbf{0}, \end{aligned}$$

since $SR = 0$, so that $\mathbf{x}$ is an eigenvector of S with eigenvalue 0.

A similar argument applied to S shows that each of the non-zero eigenvalues of S has an eigenvector which is one of the second $n - r$ columns of P because it is a column corresponding to a 0 eigenvalue of R.

Hence

$$P^T SP = \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & D_2 \end{array}\right]$$

where D_2 is diagonal (possibly with some zero entries) and the partition has the same dimensions as that for $P^T RP$ (ie D_2 is $n - r \times n - r$).

Now define

$$\mathbf{z} = P^T C^{-1} \mathbf{y}.$$

Then $\mathbf{z}$ is multivariate normal with

$$E[\mathbf{z}] = P^T C^{-1} \boldsymbol{\mu}$$

and

$$\operatorname{var} \mathbf{z} = P^T C^{-1} V C^{-1} P = P^T P = I.$$

Thus the elements of $\mathbf{z}$ are independent.

Partition $\mathbf{z}$ into

$$\mathbf{z} = \left[\begin{array}{c} \mathbf{z}_1 \\ \hline \mathbf{z}_2 \end{array}\right]$$

where $\mathbf{z}_1$ has dimension $r(A) \times 1$.

By rewriting our equations, we see that

$$\begin{array}{rcl} \mathbf{y} &=& CP\mathbf{z} \\ A &=& C^{-1} R C^{-1} \\ B &=& C^{-1} S C^{-1} \end{array}$$

So

$$\begin{array}{rcl} \mathbf{y}^T A \mathbf{y} &=& \mathbf{z}^T P^T C C^{-1} R C^{-1} C P \mathbf{z} \\ &=& \mathbf{z}^T P^T R P \mathbf{z} \\ &=& \left[\begin{array}{c|c} \mathbf{z}_1^T & \mathbf{z}_2^T \end{array}\right] \left[\begin{array}{c|c} D_1 & 0 \\ \hline 0 & 0 \end{array}\right] \left[\begin{array}{c} \mathbf{z}_1 \\ \hline \mathbf{z}_2 \end{array}\right] \\ &=& \mathbf{z}_1^T D_1 \mathbf{z}_1 \end{array}$$

and similarly

$$\mathbf{y}^T B\mathbf{y} = \mathbf{z}_2^T D_2 \mathbf{z}_2.$$

But $\mathbf{z}_1$ and $\mathbf{z}_2$ are mutually independent of each other, since all elements of $\mathbf{z}$ are independent. Therefore $\mathbf{y}^T A\mathbf{y}$ and $\mathbf{y}^T B\mathbf{y}$ are independent.

3.2 Example for independence of quadratic forms - MM 2.4

Let y_1 and y_2 follow a multivariate normal distribution with covariance matrix

$$V = \begin{bmatrix} a & b \\ b & c \end{bmatrix}.$$

Consider the symmetric matrices

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

Then

$$\mathbf{y}^T A\mathbf{y} = y_1^2, \quad \mathbf{y}^T B\mathbf{y} = y_2^2.$$

These quadratic forms will be independent if and only if

$$\begin{aligned} AVB &= \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} a & b \\ b & c \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & b \\ 0 & c \end{bmatrix} \\ &= \begin{bmatrix} 0 & b \\ 0 & 0 \end{bmatrix} \end{aligned}$$

is the 0 matrix.

But this means $b = 0$, i.e. y_1 and y_2 have zero covariance.

We've just shown that for multivariate random normals, uncorrelatedness gives independence of the squares of the components - it is also true that the random variables are independent as noted in Section 1.10.

3.3 Corollary and Example - MM 2.4

Corollary 3.12. *Let $\mathbf{y} \sim MVN(\boldsymbol{\mu}, \sigma^2 I)$ be a random vector and let A and B be symmetric matrices. Then $\mathbf{y}^T A\mathbf{y}$ and $\mathbf{y}^T B\mathbf{y}$ are independent if and only if $AB = 0$.*

```
A <- matrix(1, 2, 2)
B <- matrix(c(1,-1,-1,1), 2, 2)
A %*% B
```

```
##      [,1] [,2]
## [1,]    0    0
## [2,]    0    0
```

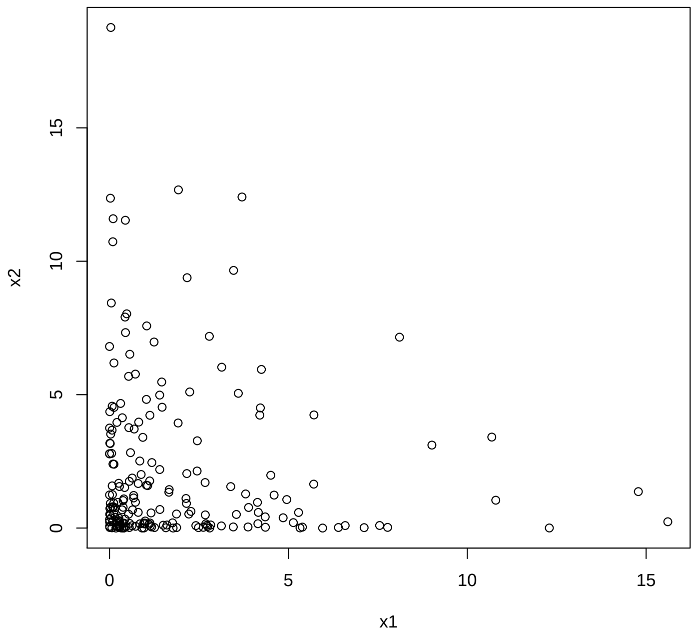

Figure 7: 2 sets of 200 Independent χ^2 rvs

```r
y <- t(mvrnorm(200, c(0, 0), diag(2)))
x1 <- apply(y, 2, quadform, A = A)
x2 <- apply(y, 2, quadform, A = B)
cor(x1, x2)
```

```
## [1] -0.08779417
```

```r
plot(x1, x2)
```

Plot in Figure 7

3.4 Independence of linear combination and quadratic form - MM 2.4

Next we consider when a quadratic form is independent of a random vector. Firstly, we define a random variable to be independent of a random vector if and only if it is independent of all elements of that vector.

Theorem 3.13. *Let* $\mathbf{y} \sim MVN(\boldsymbol{\mu}, V)$ *be a* $n \times 1$ *random vector, and let* A *be a* $n \times n$ *symmetric matrix and* B *a* $m \times n$ *matrix. Then* $\mathbf{y}^T A\mathbf{y}$ *and* $B\mathbf{y}$ *are independent if and only if* $BVA = 0$.

Lastly, we can combine several of the theorems we have seen before to tell when a group of quadratic forms (more than two) are independent.

3.5 A number of quadratic forms - MM 2.4

Theorem 3.14. *Let* $\mathbf{y} \sim MVN(\boldsymbol{\mu}, I)$ *be a random vector, and let* $A_1, \ldots, A_m$ *be a set of symmetric matrices. If any two of the following statements are true:*

- *All* A_i *are idempotent;*
- $\sum_{i=1}^m A_i$ *is idempotent;*
- $A_i A_j = 0$ *for all* $i \neq j$*;*

then so is the third, and

- *For all* i, $\mathbf{y}^T A_i \mathbf{y}$ *has a noncentral* χ^2 *distribution with* $r(A_i)$ *d.f. and noncentrality parameter* $\lambda_i = \frac{1}{2}\boldsymbol{\mu}^T A_i \boldsymbol{\mu}$*;*
- $\mathbf{y}^T A_i \mathbf{y}$ *and* $\mathbf{y}^T A_j \mathbf{y}$ *are independent for* $i \neq j$*; and*
- $\sum_{i=1}^m r(A_i) = r\left(\sum_{i=1}^m A_i\right)$.

When $\sum_i A_i = I$, the previous result is related to the following theorem (which we will not prove):

Theorem 3.15 (Cochran-Fisher Theorem)**.** *Let* $\mathbf{y} \sim MVN(\boldsymbol{\mu}, \sigma^2 I)$ *be a* $n \times 1$ *random vector. Decompose the sum of squares of* $\mathbf{y}/\sigma$ *into the quadratic forms*

$$\frac{1}{\sigma^2}\mathbf{y}^T\mathbf{y} = \sum_{i=1}^m \frac{1}{\sigma^2}\mathbf{y}^T A_i \mathbf{y}.$$

Then the quadratic forms are independent and have noncentral χ^2 *distributions with parameters* $r(A_i)$ *and* $\frac{1}{2\sigma^2}\boldsymbol{\mu}^T A_i \boldsymbol{\mu}$*, respectively, if and only if*

$$\sum_{i=1}^m r(A_i) = n.$$

3.6 Big Picture

Big picture - recall

You will recall that last semester there were a number of examples where χ^2 rv's were claimed to be independent (for example, the sums of squares in analysis of variance for the normal location problem (t-distribution), regression).

In each case, we separated the sum of squares of the response variable (around the parameters) into a sum of two quadratic forms:

$$\mathbf{y}^T\mathbf{y} = \mathbf{y}^T A_2 \mathbf{y} + \mathbf{y}^T A_2.$$

Big picture - recall

Since $\sum_{i=1}^{2} A_i = I$, it is idempotent. Then if we can establish one of the other two conditions given in the theorem (most commonly that A_i is idempotent), we can use the theorem to determine the distribution of the quadratic forms, and also conclude their independence.

This will justify the claims of independence in all cases so far, and many more that are central in techniques of machine learning.