Exercise 6

Applied Statistics, IT University of Copenhagen

T = Theoretical Exercise, R = R Exercise

Preparation

- Read pages 92, 96–99, 102–108, 111–113, 132–139 from Verzani (2014).

Problems

1. Hair vs. Eye Colour (T)

To investigate the relations between hair color and eye colour, the hair color and eye color of 5383 was recorded. The data are given in Table 1. Eye color is encoded by the values 1 (Light) and 2 (Dark), and hair color by 1 (Fair/red), 2 (Medium), and 3 (Dark/black). By dividing the numbers in the table by 5383, the tble is turned into a joint probability distribution for random variables X (hair color) taking values 1 to 3 and Y (eye color) taking values 1 and 2.

(a) Determine the joint and marginal probability distributions of X and Y.

The joint distribution of X and Y is the same table, but each number is divided by 5383 in order to obtain a valid probability distribution. See Table 2.

To determine the marginal distribution of X, we marginalize Y out, i.e. for each value x that X can take on, we find $P(X = x)$ by summing over all the values that Y can take. Similarly, to determine the marginal distribution of Y, we marginalize X out, i.e. for each value y that Y can take on, we find $P(Y = y)$ by summing over all the values that X can take:

Table 1: Relation between hair color and eye color.

	Hair color		
	Fair/Red	Medium	Dark/black
Eye color			
Light	1168	825	305
Dark	573	1312	1200

Table 2: Joint PMF of X and Y

	X		
	1	2	3
Y			
1	0.2169794	0.1532603	0.0566599
2	0.1064462	0.2437303	0.2229240

	X			
Y	1	2	3	Py
1	0,2169794	0,1532603	0,0566599	0,4268996
2	0,1064462	0,2437303	0,2229240	0,5731005
Px	0,3234256	0,3969906	0,2795839	1

(b) Find out whether X and Y are dependent or independent.

X and Y are independent if and only if $P(X = x, Y = y) = P(X = x)P(Y = y)$ for all x and y.

Let's check for $x = 1$ and $y = 1$

$$P(X = 1, Y = 1) = 0.217$$
$$P(X = 1)P(Y = 1) = 0.323 \cdot 0.427 = 0.138$$

Since $P(X = x, Y = y) \neq P(X = x)P(Y = y)$, X and Y are dependent.

2. Joint distribution (T)

Let X and Y be continuous random variables with the joint probability density function

$$f(x, y) = \begin{cases} cx + 1 & \text{if } x, y \geq 0, x + y < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

(a) Find the constant c.

We know that any continuous distribution must integrate to 1. To find the value of c, we therefore integrate out x and y.

x and y can only attain values in the interval [0,1[, as we know $x, y \geq 0$ and $x + y < 1$. The only hurdle here is the constraint that x and y must sum to 1. To satisfy this constraint, we fix a value for x between 0 and 1 in the integral and integrate y from 0 to $1 - x$:

$$
\begin{aligned}
&\int_0^1 \int_0^{1-x} cx+1 \text{ d}y \text{ d}x \\
=&\int_0^1 [cxy+y]_0^{1-x} \text{ d}x \\
=&\int_0^1 cx(1-x)+1-x \text{ d}x \\
=&\int_0^1 cx(1-x) \text{ d}x + \int_0^1 1-x \text{ d}x \\
=&\left[-c\left(\frac{1}{3}x^3-\frac{1}{2}x^2\right)\right]_0^1 + \left[x-\frac{1}{2}x^2\right]_0^1 \\
=&\frac{-c}{3}+\frac{c}{2}+\frac{1}{2} \\
=&\frac{c}{6}+\frac{1}{2}
\end{aligned}
$$

We can now isolate c:

$$
\begin{aligned}
\frac{c}{6}+\frac{1}{2}&=1 \\
\frac{c}{6}&=\frac{1}{2} \\
c&=3
\end{aligned}
$$

(b) Compute the marginal distribution $f_X(x)$.

We get the marginal distribution $f_X(x)$ by integrating $f(x,y)$ with respect to y in the interval [0,1-x] in order to satisfy the constraint that $x+y<1$. By now, we also know that $c=3$, so let's include that in the distribution:

$$
\begin{aligned}
f_X(x) &= \int_0^{1-x} 3x+1 \text{ d}y \\
&= [3xy+y]_0^{1-x} \\
&= 3x(1-x)+1-x
\end{aligned}
$$

(c) Compute $P(Y<2X^2)$.

Don't know how to do this, but here's the solution from the solutions folder:

We first have to find the subset of values that satisfy that inequality, we do this by finding the intersection between the values below the curve $2X^2$ and the values below the line $1-X$, which is the original set. Having those in mind, we need to find where do they intersect:

$$
\begin{aligned}
y &= 1-x \\
y &= 2x^2 \\
1-x &= 2x^2 \\
2x^2-x+1 &= 0 \\
x &= \frac{1}{2}\text{(Ignoring the negative solution)}
\end{aligned}
$$

From this we know that the area we need to integrate is below $2X^2$ up to $\frac{1}{2}$, and then is below the original set, $1 - X$:

$$
\begin{aligned}
P(Y < 2X^2) &= \int_0^{\frac{1}{2}} \int_0^{2x^2} 3x + 1 \ dy \ dx + \int_{\frac{1}{2}}^{1} \int_0^{1-x} 3x + 1 \ dy \ dx = \\
&= \int_0^{\frac{1}{2}} \Big[(3xy + y)\Big]_{0 \text{ (On y)}}^{2x^2} dx + \int_{\frac{1}{2}}^{1} \Big[(3xy + y)\Big]_{0 \text{ (On y)}}^{1-x} dx = \\
&= \int_0^{\frac{1}{2}} 6x^3 + 2x^2 \ dx + \int_{\frac{1}{2}}^{1} -3x^2 + 2x + 1 \ dx = \\
&= \Big[\frac{6}{4}x^4 + \frac{2}{3}x^3\Big]_0^{\frac{1}{2}} + \Big[-x^3 + x^2 + x\Big]_{\frac{1}{2}}^{1} \\
&= \frac{17}{96} + \frac{3}{8} = \frac{53}{96} \approx 0.55
\end{aligned}
$$

3. Covariance and Correlation (T)

Show that the correlation between X and Y is simply the covariance of the corresponding standardised scores, i.e,

$$
\rho(X, Y) = \mathrm{Cov}\left[\frac{X - E[X]}{\sqrt{\mathrm{Var}[X]}}, \frac{Y - E[Y]}{\sqrt{\mathrm{Var}[Y]}}\right]. \tag{2}
$$

The correlation between X and Y is defined as

$$
\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}
$$

The covariance between X and Y is defined as

$$
\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]
$$

Thus, the covariance between the two quantities in the exercise definition is

$$
\begin{aligned}
&\mathrm{Cov}\left[\frac{X - E[X]}{\sqrt{\mathrm{Var}[X]}}, \frac{Y - E[Y]}{\sqrt{\mathrm{Var}[Y]}}\right] \\
=&E\left[\frac{(X - E[X])(Y - E[Y])}{\sqrt{\mathrm{Var}[X]}\sqrt{\mathrm{Var}[Y]}}\right] - E[X - E[E]]E[Y - E[Y]]
\end{aligned}
$$

The expectation of a random variable minus its expectation is 0, so we have

$\mathrm{E}[X-\mathrm{E}[X]] = \mathrm{E}[Y-\mathrm{E}[Y]] = 0.$

$$
\begin{aligned}
& E\left[\frac{(X-E[X])(Y-E[Y])}{\sqrt{\mathrm{Var}[X]}\sqrt{\mathrm{Var}[Y]}}\right] - E[X-E[E]]E[Y-E[Y]] \\
=& E\left[\frac{(X-E[X])(Y-E[Y])}{\sqrt{\mathrm{Var}[X]}\sqrt{\mathrm{Var}[Y]}}\right] - 0\cdot 0 \\
=& E\left[\frac{(X-E[X])(Y-E[Y])}{\sqrt{\mathrm{Var}[X]\mathrm{Var}(Y)}}\right]
\end{aligned}
$$

We can view $\frac{1}{\sqrt{\mathrm{Var}[X]\mathrm{Var}(Y)}}$ as a constant factor, and we can therefore use the property linearity of expectations:

$$
\begin{aligned}
& E\left[\frac{(X-E[X])(Y-E[Y])}{\sqrt{\mathrm{Var}[X]\mathrm{Var}(Y)}}\right] \\
=& \frac{1}{\sqrt{\mathrm{Var}[X]\mathrm{Var}(Y)}} E\left[(X-E[X])(Y-E[Y])\right] \\
=& \frac{E\left[(X-E[X])(Y-E[Y])\right]}{\sqrt{\mathrm{Var}[X]\mathrm{Var}(Y)}} \\
=& \frac{\mathrm{Cov(X, Y)}}{\sqrt{\mathrm{Var}[X]\mathrm{Var}(Y)}} \text{ (by the definition of covariance)}
\end{aligned}
$$

4. Correlation Coefficient (R)

(a) The data set `normtemp (UsingR)` contains body measurements for 130 healthy, randomly selected individuals. The variable `temperature` measures normal body temperature, and the variable `hr` measures resting heart rate. Make a scatter plot of the two variables. What does the plot show you? Find the Pearson correlation coefficient. How does the estimate relate to the scatter plot?

```
require(UsingR)
```

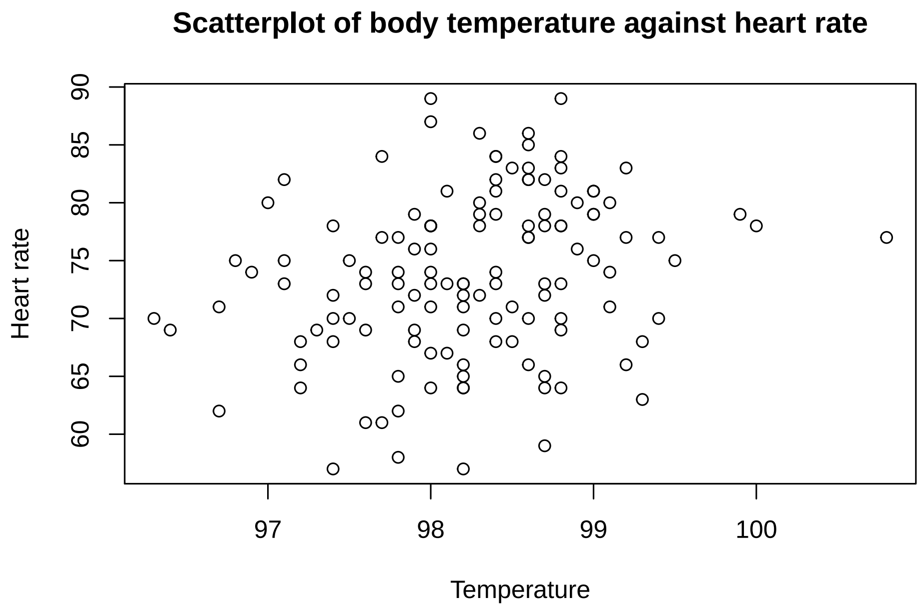
```
## Loading required package: UsingR

## Loading required package: MASS

## Loading required package: HistData

## Loading required package: Hmisc

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
##
## Attaching package: 'UsingR'
```

```
## The following object is masked from 'package:survival':
##
##     cancer
```

```r
plot(normtemp$temperature,
     normtemp$hr,
     main = "Scatterplot of body temperature against heart rate",
     xlab = "Temperature",
     ylab = "Heart rate")
```



The plot shows that there does not seem to be a strong correlation between body temperature and heart rate. There may be a slight positive correlation. Let's calculate the correlation coefficient to test this hypothesis:

```r
cor(normtemp$temperature,normtemp$hr)
```

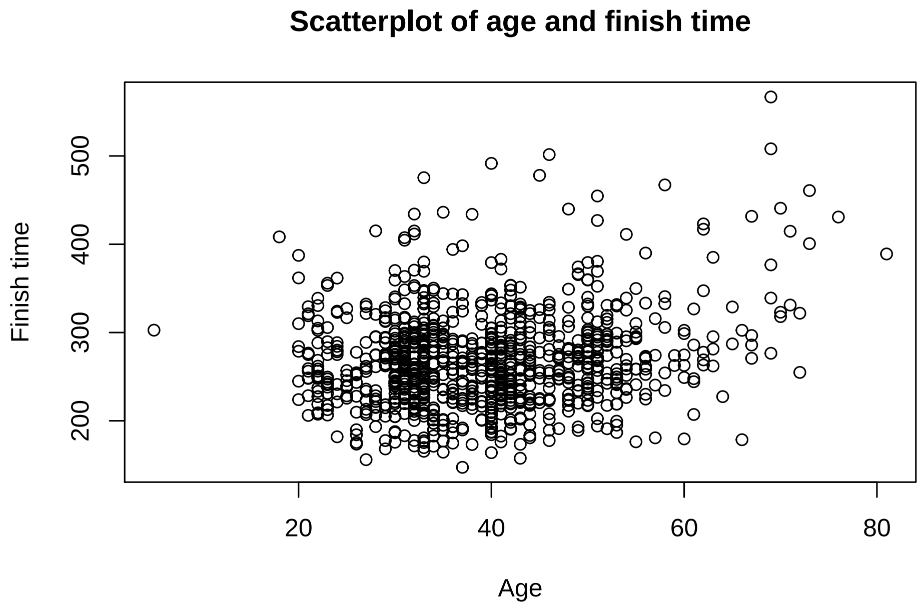
```
## [1] 0.2536564
```

We have a Pearson correlation of 0.25. A positive correlation coefficient means that an increase in body temperature will result in an increase in heart rate, which was also our suspicion.

(b) The data set `nym.2002 (UsingR)` contains information about the 2002 New York city marathon. What do you expect the correlation between age and finishing time to be? Make a scatter plot and compute the correlation coefficient. Does the result match your expectation?

I expect age to be positively correlated with the finishing time, i.e. the older you are, the slower you will run.

```
plot(nym.2002$age,
     nym.2002$time,
     main = "Scatterplot of age and finish time",
     xlab = "Age",
     ylab = "Finish time")
```




```
cor(nym.2002$age,nym.2002$time)
```

```
## [1] 0.1898672
```

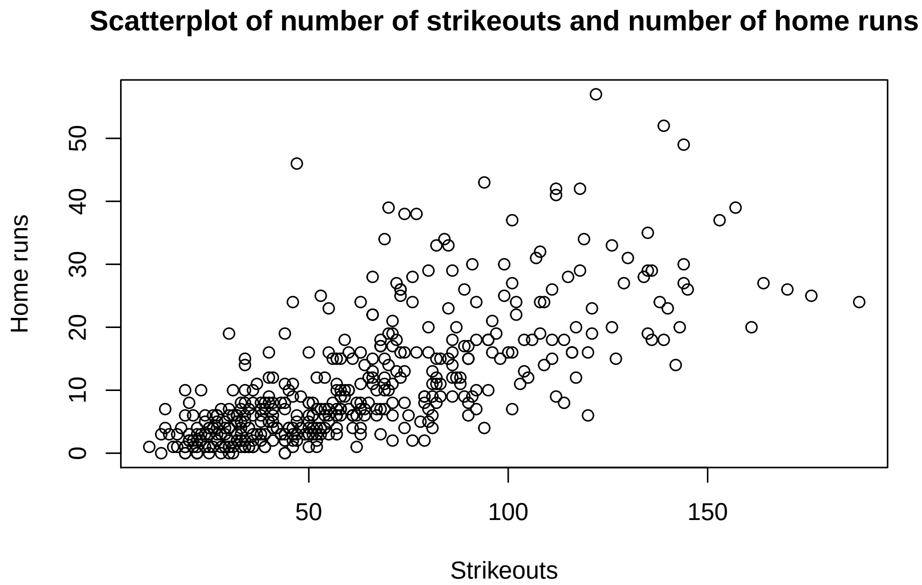
We see a positive correlation, which confirms the hypothesis.

(c) The `batting` set (`UsingR`) data set contains baseball statistics for the 2002 Major League Baseball season. What is the correlation between the number of strikeouts (`SO`) and the number of home runs (`HR`)? Make a scatter plot to see whether there is any trend. Does the data suggest that in order to hit a lot of home runs one should strike out a lot?

```
cor(batting$SO,batting$HR)
```

```
## [1] 0.7084697
```

```
plot(batting$SO,
     batting$HR,
     main = "Scatterplot of number of strikeouts and number of home runs",
     xlab = "Strikeouts",
     ylab = "Home runs")
```



With a correlation coefficient of 0.7, it seems that if you strike out a lot, you will also hit a lot of home runs. Wikipedia confirms this hypothesis: *Although a strikeout suggests that the pitcher dominated the batter, the free-swinging style that generates home runs also leaves batters susceptible to striking out. Some of the greatest home run hitters of all time—such as Alex Rodriguez, Reggie Jackson, and Jim Thome—were notorious for striking out.* (https://en.wikipedia.org/wiki/Strikeout)

5. Sampling from a Joint Probability Distribution (R)

Let X and be Y be continuous random variables with the joint probability density function $f(x, y)$. In general, one can draw samples $(x^{(n)}, y^{(n)})$ from the joint probability density of two random variables factoring the probabity density as $f(x, y) = f(y|x)f(x)$ and first drawing a sample $x^{(n)}$ for X from the marginal density $f(x)$ and then the sample $y^{(n)}$ for Y from the conditional density $f(y|x)$ conditioned on $x = x^{(n)}$.

Now, assume that $f(x) = 1$ for $0 < x < 1$, $f(x) = 0$ otherwise; and

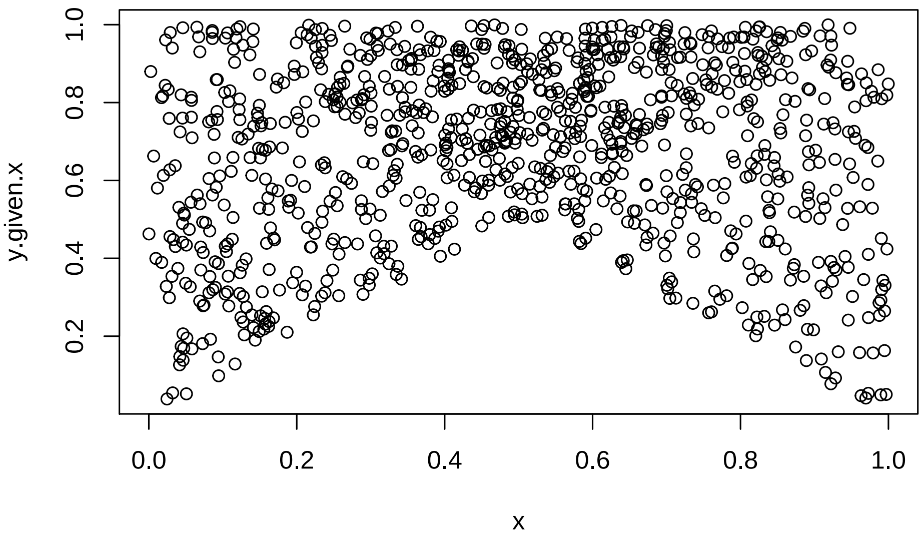
$$f(y|x) = \begin{cases} y - x + 1 & \text{if } -1 + x \leq y < x \\ -y + x + 1 & \text{if } x \leq y < 1 + x \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Implement a computer program that draws samples from $f(x, y)$ by using the random number generator for uniformly distributed random variables in R.

This is probably wrong... I guess $f(y|x)$ is a PMF? Also I feel like we need to know the distribution of y in order to sample from $f(y|x)$. I'll assume that it is a U(0,1) distribution

```
N.SAMPLES <- 1000
```

```r
x <- runif(N.SAMPLES)
y.given.x <- rep(0,N.SAMPLES)
for (i in 1:N.SAMPLES){
  y <- runif(1)
  if (-1+x[i]<=y && y<x[i]){
    y.given.x[i] <- y-x[i]+1
  }
  else if (x[i]<=y && y<1+x[i]){
    y.given.x[i] <- -y+x[i]+1
  }
  else y.given.x[i] <- 0
}
plot(x,y.given.x)
```



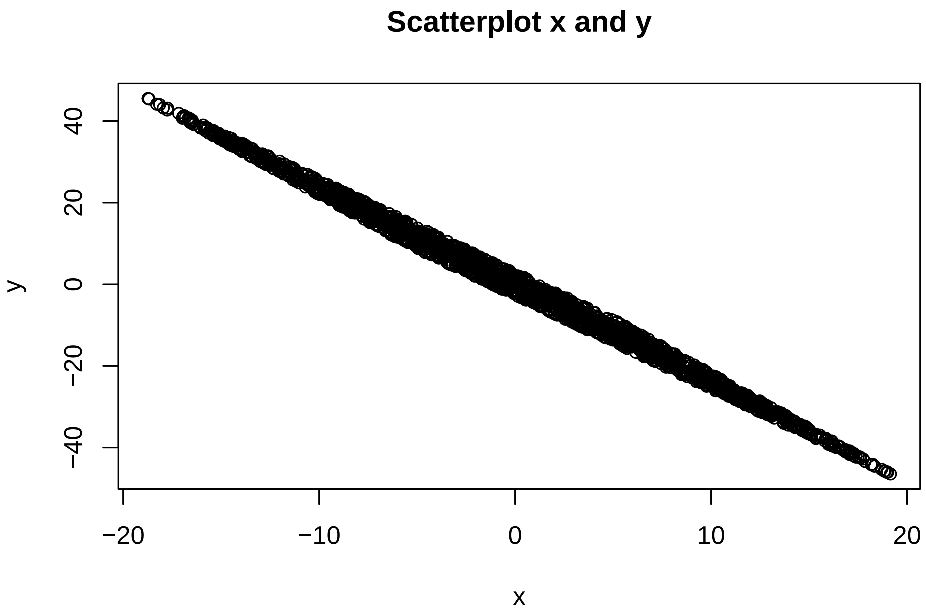
6. Covariance, Correlatedness and Independence (R)

Load the data set by copying the data file from course webpage to your working directory and typing `load("mypnts.Rdata")`.

```r
load("mypnts.Rdata")
```

(a) Make a scatter plot of the points. Are the x and y coordinates correlated?

```r
plot(mypnts$x,
     mypnts$y,
     main = "Scatterplot x and y",
     xlab = "x",
     ylab = "y")
```



They look negatively correlated. Let's check the correlation coefficient:

```
cor(mypnts$x,mypnts$y)
```

```
## [1] -0.9976884
```

Yep

(b) Estimate the *covariance matrix* of the data set. It is a 2×2 matrix containing the all the pairwise covariances, i.e.,

$$\mathbf{C} = \begin{pmatrix} \mathrm{Cov}(X,X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(Y,X) & \mathrm{Cov}(Y,Y) \end{pmatrix} \tag{4}$$

What can you see from the covariance matrix estimate?

```
cov(mypnts)
```

```
##             x          y
## x   68.20358 -162.3543
## y -162.35430  388.2671
```

We see that x has a variance of 68, y has a variance of 388, and x and y have a covariance of -162.

(c) Apply the mapping

$$x' = ax + by \quad \text{and} \quad y' = cx + dy, \tag{5}$$

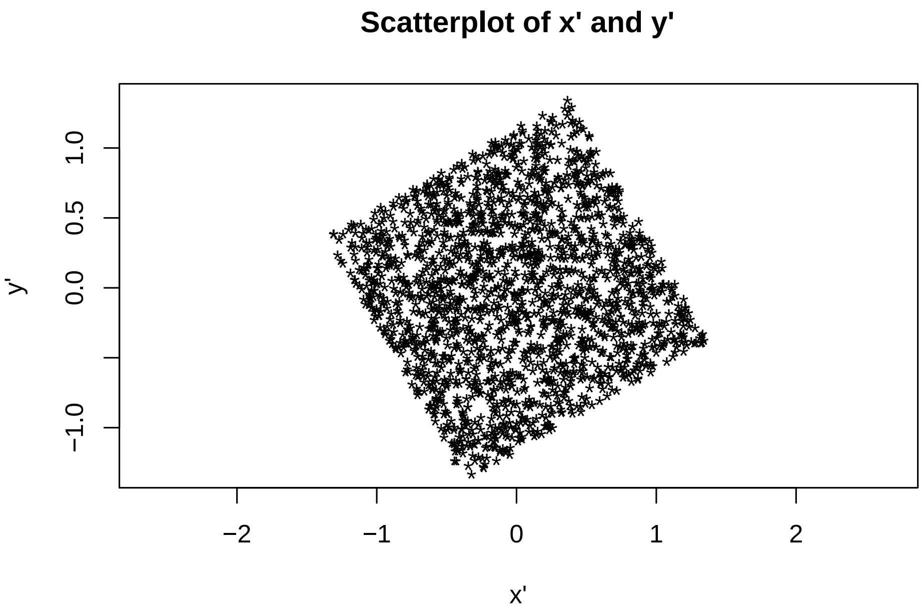
to the points where $a = 0.07$, $b = 0$, $c = 1$,and $d = 0.42$. This process is called *whitening*.

```
a <- 0.07
b <- 0
c <- 1
d <- 0.42
```

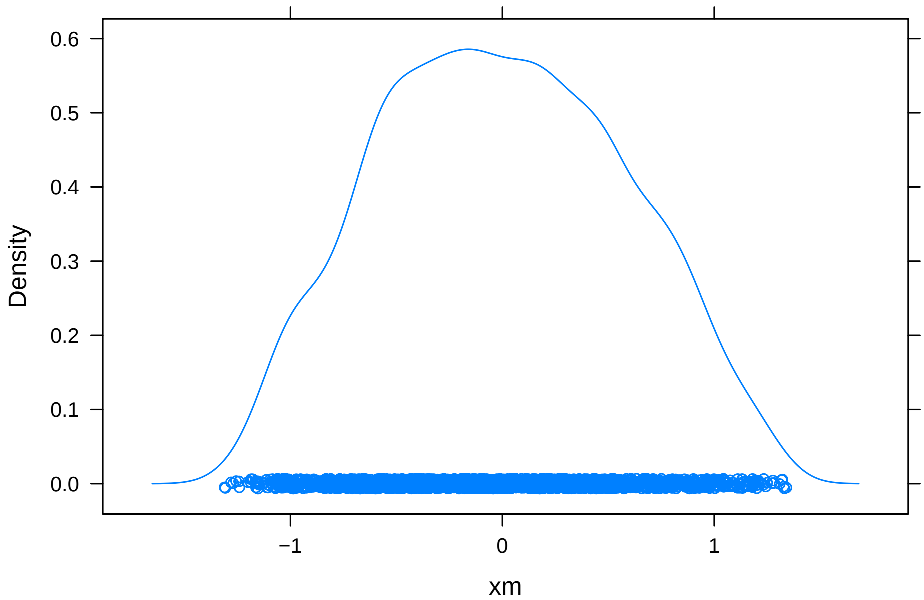
```r
xm <- a*mypnts$x + b*mypnts$y
ym <- c*mypnts$x + d*mypnts$y
```

(d) Plot the mapped points (use the option 'asp=1' that gives unity aspect ratio), and their marginal distributions (`densityplot`) on both x' and y' axis. Compute the covariance matrix estimate for the mapped points. Are the mapped points uncorrelated? How about independent? Can you see why the mapping is called whitening?

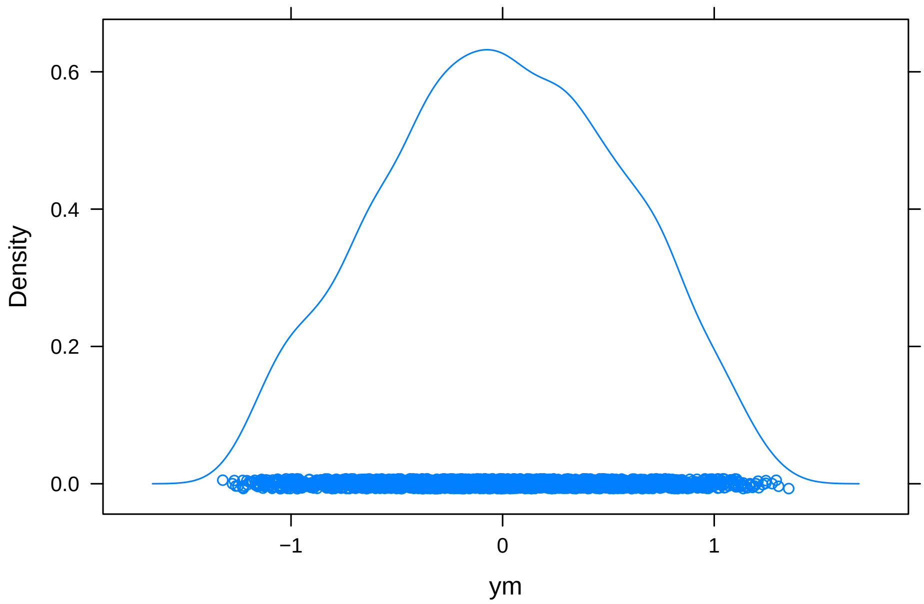
```r
require(lattice)
plot(xm,ym,
     main = "Scatterplot of x' and y'",
     xlab = "x'",
     ylab = "y'",
     asp=1,
     pch="*")
```




```r
densityplot(xm)
```




```
densityplot(ym)
```



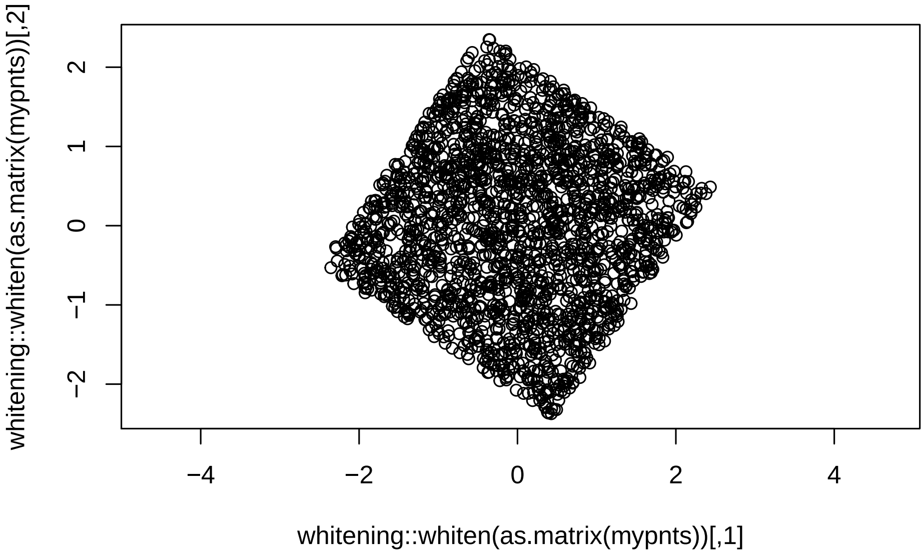

```
cov(data.frame(xm,ym))
```

```
##             xm          ym
## xm 0.334197555 0.001034248
## ym 0.001034248 0.316283824
```

Since we have supposedly whitened the data, we should have obtained an identity matrix. We have thus not really whitened the data

Using the r package 'whitening' (needs to be installed), we get a different data matrix after transforming the data:

```
plot(whitening::whiten(as.matrix(mypnts)),asp=1)
```




```
cov(whitening::whiten(as.matrix(mypnts)))
```

```
##                [,1]          [,2]
## [1,]  1.000000e+00 -3.287396e-15
## [2,] -3.287396e-15  1.000000e+00
```

This is how the data should have been transformed

Nonetheless, x' and y' are now somewhat uncorrelated. However, they are not independent, as for instance knowing that $x' = -1$ narrows down the possibilities for y by a lot.

The mapping is called whitening, because we tranform the data into white noise vectors.

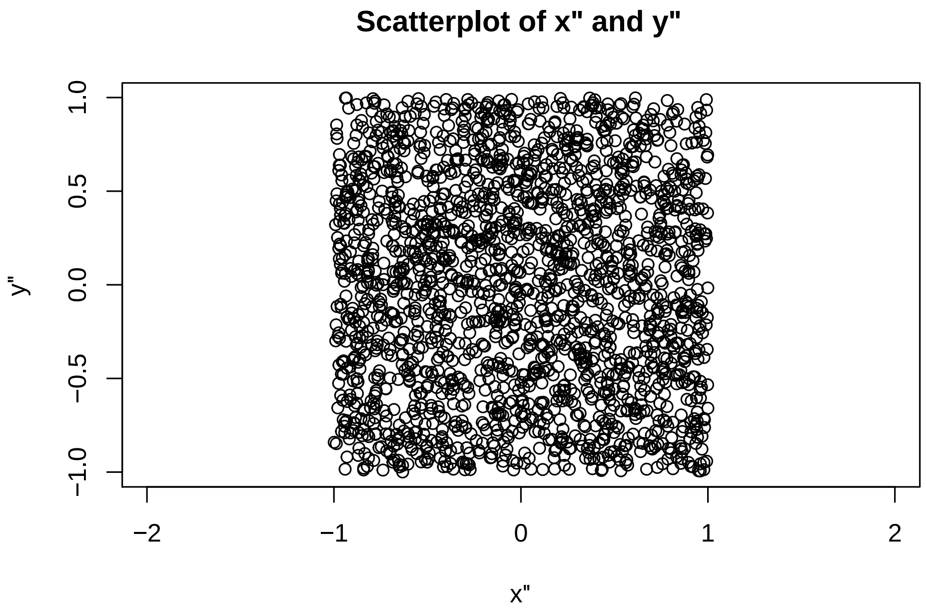
(e) Apply rotation to the mapped points

$$x'' = cos(\alpha)x' - sin(\alpha)y' \quad \text{and} \quad y'' = sin(\alpha)x' + cos(\alpha)y', \tag{6}$$

where $\alpha = -\pi/6$.

Plot the rotated, mapped points, and the marginals on the new axes. Are the rotated, mapped coordinates uncorrelated? How about independent?

```
alpha = -pi/6
xmm<- cos(alpha)*xm - sin(alpha)*ym
ymm<- sin(alpha)*xm + cos(alpha)*ym
plot(xmm,
     ymm,
     main = "Scatterplot of x'' and y''",
     xlab = "x''",
     ylab = "y''",
     asp = 1)
```




```
cor(data.frame(xmm,ymm))
```

```
##              xmm         ymm
## xmm  1.00000000 -0.02226271
## ymm -0.02226271  1.00000000
```

Now they are almost uncorrelated. They are also now independent, as the value of x’ ’ tells you nothing about the value of y’ ’.

(f) What did you learn from this exercise? Does the result generalise?

You can transform a highly correlated variables into uncorrelated, independent variables with a mean of 0 and a variance of 0 and yeah, I think the result does generalize

Verzani, John. 2014. *Using R for Introductory Statistics*. CRC Press.