Norwegian University of Science and Technology
Department of Mathematical Sciences

TMA4267 Linear statistical models
Recommended exercises 2 – solutions

Problem 1 Principal components analysis

```
USArrests

pca <- prcomp(USArrests, scale = TRUE) # scale: variables are scaled

## a
# coefficients of PCs, called rotations or loadings
pca$rotation
pca$rotation %*% t(pca$rotation) # yes, it is orthogonal

# done "manually":
corrmatrix <-
  cor(USArrests) # cor, not cov, since covariates are scaled
corrmatrix
cov(scale(USArrests)) # the same
eigen(corrmatrix) # same as pcs$rotations - coincidentally same signs

## b
# sample variances of the PCs are the eigenvalues of s
pca$sdev
pca$sdev ^ 2 # compare with eigenvalues above

## c
# scores - values of the PCs
pca$x
cov(pca$x) # offdiag=0, ondiag=pca$sdev^2
pca$sdev ^ 2
# check that scores for Alabama are the linear combinations they should be:
t(pca$rotation) %*% t(scale(USArrests))[, "Alabama"]
# gives the 4 linear combinations that are the scores for Alabama - also
# t(scale(USArrests))[, 1] can be used for picking the first column

## d
# plot the scores of two first PCs against each other
```

```
plot(pca$x[, 1], pca$x[, 2], type = "n")
text(pca$x[, 1], pca$x[, 2], rownames(USArrests), cex = 0.6)
# The same with also loading of two first PCs plotted
biplot(pca, scale = 0, cex = 0.6)
# scale=0: arrows scaled to represent the loadings

# plot scaled original data for assault and rape, which are influental for PC1,
# then add line parallell with vector consisting
# of loadings for assault and rape for PC1
plot(scale(USArrests)[, 2], scale(USArrests)[, 4])
abline(0, pca$rotation[2, 1] / pca$rotation[4, 1], col = "red")
# agrees visually with line onto which projections have large variance

# the same for other pairs of PCs:
biplot(pca,
       choices = c(1, 3),
       scale = 0,
       cex = 0.6)

## e
# How many PCs do we need to capture a large part of the variability in the data?
summary(pca) # 2 to get 87%
plot(pca) # screeplot

## f
# the effect of scaling
pca.noscale <- prcomp(USArrests, scale = FALSE)
summary(pca.noscale) # 1
pca.noscale$rotation # PC1 dominated by Assault, PC2 by UrbanPop, etc.
cov(USArrests) # this explains why
cov(pca.noscale)
biplot(pca.noscale, scale = 0)
biplot(pca.noscale, choices = c(3, 4), scale = 0)
```

Problem 2 Multivariate transformation – the t distribution

a) Since U and V are independent, the joint pdf is the product of the marginal pdfs:

$$f_{U,V}(u,v) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \cdot \frac{1}{\Gamma(p/2)2^{p/2}} v^{p/2-1} e^{-v/2} = \frac{1}{2^{(p+1)/2}\Gamma(p/2)\sqrt{\pi}} v^{p/2-1} e^{-(u^2+v)/2}$$

The inverse of the transformation $t = u/\sqrt{v/p}$ and $w = v$ is $u = t\sqrt{w/p}$ and $v = w$, with Jacobian $\sqrt{w/p}$. This gives the joint distribution of T and W:

$$\begin{aligned}
f_{T,W}(t,w) &= f_{U,V}\left(t\sqrt{\frac{w}{p}}, w\right) \cdot \sqrt{\frac{w}{p}} \\
&= \frac{1}{2^{(p+1)/2}\,\Gamma(p/2)\sqrt{\pi}} w^{p/2-1} e^{-(t^2 w/p+w)/2} e^{-w/2} \left(\frac{w}{p}\right)^{1/2} \\
&= \frac{1}{2^{(p+1)/2}\,\Gamma(p/2)\sqrt{\pi p}} w^{(p-1)/2} e^{-(1+t^2/p)w/2}
\end{aligned}$$

(An alternative way to find the joint distribution of T and W is to use the conditional pdf, $f_{T|W=w}$, of T given $W = w$, that is, the pdf of $\sqrt{p/w}\,U$, which has the $N(0, p/w)$ distribution. Then $f_{T,W}(t,w) = f_{T|W=w}(t) f_W(w)$, where f_W is the pdf of W. But here the point was to use the transformation formula.)

b) The marginal pdf of T is

$$\begin{aligned}
f_T(t) &= \int_0^\infty f_{T,W}(t,w)\,dw \\
&= \frac{1}{2^{(p+1)/2}\,\Gamma(p/2)\sqrt{\pi p}} \int_0^\infty w^{(p+1)/2-1} e^{-(1+t^2/p)w/2} \\
&= \frac{\Gamma(\frac{p+1}{2})}{(1+\frac{t^2}{p})^{(p+1)/2}\,\Gamma(\frac{p}{2})\sqrt{\pi p}} \int_0^\infty \frac{1}{(\frac{2}{1+t^2/p})^{(p+1)/2}\Gamma(\frac{p+1}{2})} w^{(p+1)/2-1} e^{-(1+t^2/p)w/2} dw \\
&= \frac{\Gamma((p+1)/2)}{\Gamma(p/2)\sqrt{\pi p}} \left(1 + \frac{t^2}{p}\right)^{-(p+1)/2}.
\end{aligned}$$

The trick was to recognize that the integrand of the last integral is the pdf of the gamma distribution, $\frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}$, with parameters $\alpha = (p+1)/2$ and $\beta = 2/(1+t^2/p)$, so that the integral is 1.

Problem 3 The standard normal and chi-squared distributions

a) Denote by ϕ the $N(0,1)$ pdf and by f the pdf of U^2. Then

$$\begin{aligned}
P(U^2 \le x) &= P(-\sqrt{x} \le U \le \sqrt{x}) = P(U \le \sqrt{x}) - P(U < -\sqrt{x}), \\
f(x) &= \frac{d}{dx} P(U^2 \le x) = \frac{d}{dx}\Big(P(U \le \sqrt{x}) - P(U < -\sqrt{x})\Big) \\
&= \phi(\sqrt{x}) \frac{d}{dx}\sqrt{x} - \phi(-\sqrt{x}) \frac{d}{dx}(-\sqrt{x}) = \frac{1}{\sqrt{2\pi}} e^{-x/2} \frac{1}{2\sqrt{x}} + \frac{1}{\sqrt{2\pi}} e^{-x/2} \frac{1}{2\sqrt{x}} \\
&= \frac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2}.
\end{aligned}$$

The moment-generating function of U^2 is given by

$$\begin{aligned}
M(t) &= \int_{-\infty}^{\infty} e^{tu^2}\phi(u)\,du = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{tu^2} e^{-u^2/2} du = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-u^2(1-2t)/2} du \\
&= \frac{1}{\sqrt{1-2t}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-v^2/2} dv = \frac{1}{\sqrt{1-2t}}
\end{aligned}$$

for $t < \frac{1}{2}$. We used the substitution $v = u\sqrt{1-2t},\ \ dv = \sqrt{1-2t}\,du$.

b) First we show that

$$f(v) = \begin{cases} \frac{1}{2^{p/2}\Gamma(p/2)} v^{p/2-1} e^{-v/2} & \text{if } v \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

is really a pdf: Firstly, $f(v) \geq 0$ for all v, and secondly,

$$\begin{aligned}
\Gamma\left(\frac{p}{2}\right) &= \int_0^{\infty} u^{p/2-1} e^{-u} du \quad \text{(definition of gamma function)} \\
&= \int_0^{\infty} \frac{1}{2^{p/2}} v^{p/2-1} e^{-v/2} dv \quad \text{(substitution } v = 2u,\ \ dv = 2du\text{),}
\end{aligned}$$

so that $\int_{-\infty}^{\infty} f(v)\,dv = \frac{1}{\Gamma(p/2)} \int_0^{\infty} \frac{1}{2^{p/2}} v^{p/2-1} e^{-v/2} dv = \frac{1}{\Gamma(p/2)} \Gamma(p/2) = 1$.

Since V is the sum of p independent χ_1^2 variables, the MGF of V is the product of the MGF of p χ_1^2 variables,

$$\underbrace{M(t)M(t)\cdots M(t)}_{p \text{ factors}} = (1-2t)^{-p/2}, \quad t < \frac{1}{2}.$$

The MGF of a variable Y having pdf f is given by

$$\begin{aligned}
Ee^{tY} &= \int_{-\infty}^{\infty} e^{ty} f(y)\,dy = \int_0^{\infty} e^{ty} \frac{1}{2^{p/2}\Gamma(p/2)} y^{p/2-1} e^{-y/2} dy \\
&= \int_0^{\infty} \frac{1}{2^{p/2}\Gamma(p/2)} y^{p/2-1} e^{-(1-2t)y/2} dy \\
&= (1-2t)^{-p/2} \int_0^{\infty} \frac{1}{2^{p/2}\Gamma(p/2)} u^{p/2-1} e^{-u/2} dy \quad (u = (1-2t)y,\ \ du = (1-2t)dy) \\
&= (1-2t)^{-p/2} \int_0^{\infty} f(u)\,du = (1-2t)^{-p/2}.
\end{aligned}$$

We must assume $u > 0$, that is, $t < \frac{1}{2}$, for the integral to converge.

So V has the MGF of a variable having pdf f. So V has pdf f.

Problem 4 Normal and chi-squared distributions in R

```
## a
B <- 20
n <- 10
rnorm(B, mean = 0, sd = 1) # draw B standard normal variates
rnorm(B, 0, 1) # the same - 2nd argument is mean, 3rd is sd
rnorm(B) # the same - default values for mean is 0 and for sd 1
dchisq(1, 1) # density at 1 for chi-squared with df=1
pt(0, n - 1) # cdf at 0 for t with df=n-1
qf(0.05, 1, 2) # critical value with area 0.05 to the left
qf(0.05, 1, 2, lower.tail = FALSE) # critical value with area 0.05 to the right
qf(0.95, 1, 2) # same as above

## b
plot(dnorm, -4, 4)
abline(v = qnorm(0.05), col = "red")
abline(v = qnorm(0.95), col = "red")
# adding shades to tails:
xvalues <- seq(from = -4, to = qnorm(0.05), length = 101)
polygon(x = c(-4, xvalues, qnorm(0.05)), y = c(0, dnorm(xvalues), 0), col = "gray")
xvalues <- seq(from = qnorm(0.95), to = 4, length = 101)
polygon(x = c(qnorm(0.95), xvalues, 4), y = c(0, dnorm(xvalues), 0), col = "gray")

## c
B <- 10000
y <- rnorm(B, 0, 1) ^ 2
range(y)
hist(y, nclass = 100, freq = FALSE)
  # freq = FALSE probabilities, not frequencies, at y axis
dchisq1 <- function(x) dchisq(x, df = 1)
  # plot only takes a function with ONE argument,
  # needed to make a df=1 version of dchisq
plot(dchisq1, min(y), max(y), add = TRUE, col = "red")
# or more compactly, using a so-called anonymous function:
plot(function(x) dchisq(x, df = 1), min(y), max(y), add = TRUE, col = "red")
abline(v = qchisq(0.1, 1), col = 3)
abline(v = qchisq(0.9, 1), col = 3)
```