Norwegian University of Science and Technology
Department of Mathematical Sciences

Tentative Solutions to TMA4267 Linear Statistical Models August 2014

Problem 1 The Multivariate Normal Distribution

a) Since $\boldsymbol{Y}$ can be written as $\boldsymbol{AX}$ where $\boldsymbol{A} = \begin{pmatrix} 3 & -2 \\ 1 & 1 \end{pmatrix}$, we see that $\boldsymbol{Y}$ is a bivariate vector of linear combinations of $\boldsymbol{X}$. Since $\boldsymbol{X}$ is multivariate normal, then also $\boldsymbol{Y}$ will be multivariate normal. The mean and covariance of $\boldsymbol{Y}$ is given as:

$$\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{A}\boldsymbol{\mu} = \begin{pmatrix} 3 & -2 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} -1 \\ 3 \end{pmatrix}$$

$$\mathrm{Cov}(\boldsymbol{Y}) = \boldsymbol{A}\,\mathrm{Cov}(\boldsymbol{X})\boldsymbol{A}^T = \begin{pmatrix} 3 & -2 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}\begin{pmatrix} 3 & 1 \\ -2 & 1 \end{pmatrix} = \begin{pmatrix} 11 & -0.5 \\ -0.5 & 4 \end{pmatrix}$$

Since $\begin{pmatrix} Z \\ Y_2 \end{pmatrix}$ is bivariate normal (same reason as for $\boldsymbol{Y}$ above), then Z and Y_2 are independent if $\mathrm{Cov}(Z, Y_2) = 0$. Let $\begin{pmatrix} Z \\ Y_2 \end{pmatrix}$ be written as $\boldsymbol{BX}$ where $\boldsymbol{B} = \begin{pmatrix} 1 & a \\ 1 & 1 \end{pmatrix}$.

$$\mathrm{Cov}\begin{pmatrix} Z \\ Y_2 \end{pmatrix} = \boldsymbol{B}\,\mathrm{Cov}(\boldsymbol{X})\boldsymbol{B}^T = \begin{pmatrix} 1 & a \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ a & 1 \end{pmatrix} = \begin{pmatrix} 2a^2 + a + 1 & \frac{3}{2} + \frac{5}{2}a \\ \frac{3}{2} + \frac{5}{2}a & 4 \end{pmatrix}$$

Thus

$$\mathrm{Cov}(Z, Y_2) = \frac{3}{2} + \frac{5}{2}a = 0$$
$$a = -\frac{3}{5} = -0.6$$

b)

$$f(\boldsymbol{x}) = ce^{-\frac{1}{2}Q(x_1, x_2)}$$
$$c = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$
$$Q(x_1, x_2) = (\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$$

where $\mu_1 = 1$, $\mu_2 = 2$, $\sigma_1 = 1$, $\sigma_2 = \sqrt{2}$ and $\rho = \text{Cov}(X_1, X_2)/(\sigma_1\sigma_2) = \frac{1}{2\sqrt{2}}$, so that $c = 0.12$.

$$\begin{aligned} f(\boldsymbol{x}) &= ce^{-\frac{1}{2}Q(x_1,x_2)} = d \\ Q(x_1, x_2) &= 2(\ln(c) - \ln(d)) \\ (\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) &= 2(\ln(c) - \ln(d)) = D^2 \end{aligned}$$

This is an ellipse with principal axes along the eigenvectors of the covariance matrix $\boldsymbol{\Sigma}$, and half-lengths $D\sqrt{\lambda_i}$. See exercise E1P2 for details on this result.

Moreover, $(\boldsymbol{X} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$ is distributed as χ^2_2. And, the solid ellipsoid of $\boldsymbol{x}$ values satisfying

$$(\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \leq \chi^2_2(\alpha)$$

has probability $1 - \alpha$. Let $\alpha = 0.05$, and we find that $\chi^2_2(0.05) = 5.99$. So, $2(\ln(c) - \ln(d)) = 5.99$. Solving for d gives $d = \exp(\ln(c) - 5.99/2) = 0.006$.

In the drawing the halflengths are given as $\sqrt{5.99\lambda_i}$, where $\lambda_1 = 2.2$ and $\lambda_2 = 0.8$ from the R print-out. Halflengths are then 3.6 and 2.2. Thus, the points on the ellipse at the principal axes are for the first principal axis (-0.4,-1.4) and (2.4,5.4), and for the second principal axis (-1.0,2.8) and (3.0,1.2).

Problem 2 Predicting fat content in meat

a) The null- and alternative hypotheses are:

$$H_0 : \beta_{100} = 0 \text{vs.} H_1 : \beta_{100} \neq 0$$

The t-statistic is related to a t-distribution with $n - p = 215 - 101 = 114$ degrees of freedom. The value of the t-statistics is -0.343. The 2.5% quantile of the t-distribution with 114 degrees of freedom is approximately 1.98 from the statistical tables. This means that the p-value of the test will be much larger than 0.05 and we will not reject the null hypothesis.

How would you judge the model fit?
The regression is significant, and explains 95% of the variability in the data. A number of the 100 coefficients looks to be significant (from Figure 4). The residual plots shows no obvious trend, but the tail behaviour of the qq-plot looks a bit off the normal line. However, the Andreson-Darling test will not reject the null hypothesis of normal data. So, the model fit seems ok wrt residual analysis.

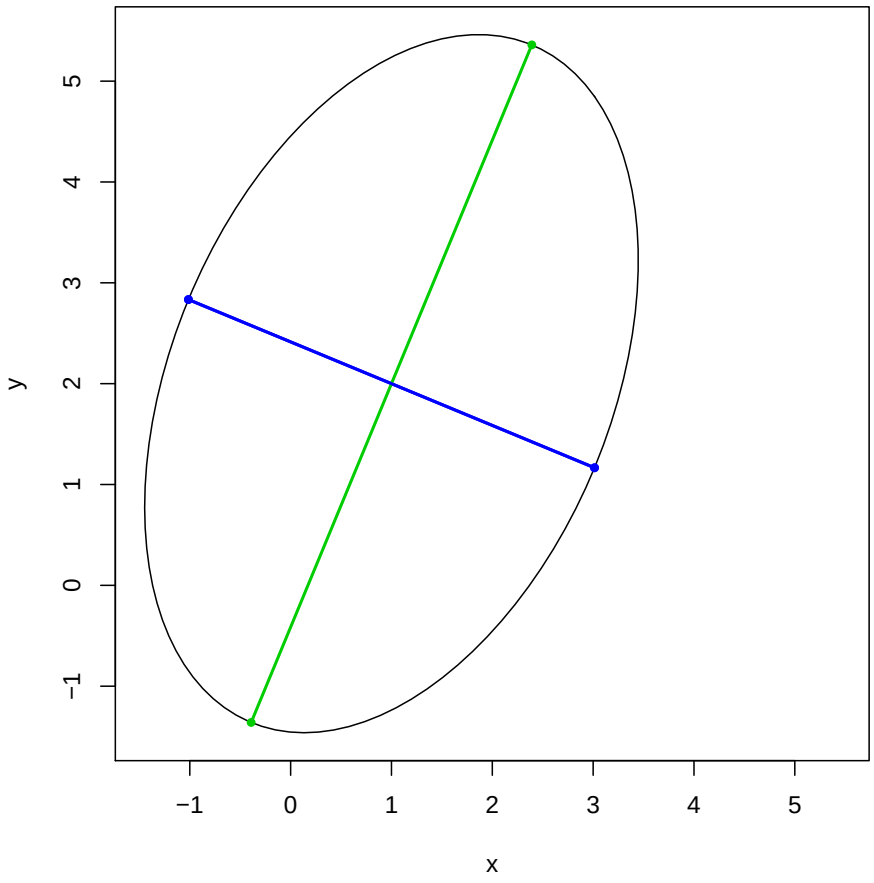

Figure 1: Contour in Problem 1b

What is overfitting? When there are many covariates as compared to observations (yes, 100 covariates is much compared to 215 observations) we have the potential problem of not just fitting the signal in the data, but also the noise. Fitting the data noise is called overfitting. Might that be a problem here? Yes, possibly. From the plot of the estimated coefficients (Figure 4) we see that some of the estimated coefficients are very large, which might point towards overfitting.

b) What is the mathematical definition of the principal component loadings and scores?

Let $\boldsymbol{X}$ be the random variable under study and let $\boldsymbol{\Sigma}$ be the population covariance matrix. Further, let $(\lambda_i, \boldsymbol{e}_i)$ be eigenvalue/vector pairs from a spectral decomposition of the covariance matrix. The pairs are ordered so that $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p$. The linear combination $\boldsymbol{e}_i^T \boldsymbol{X}$ is the $i$the principal component, and the entries in $\boldsymbol{e}_i$ are called the loadings and the values $\boldsymbol{e}_i^T \boldsymbol{X}$ (with the data here) are called the scores.

Alternatively: The principal components are uncorrelated linear combinations of the original variables, found to maximize the variance of the compontents. The PCA loadings are the weights in the linear combinations, and are given as the eigenvectors of the covariance or correlation matrix of the population (or data) under study. The PCA scores are the numerical values of the linear combinations when applied to the samples. If the number of principal components are p there are p scores for each individual sample.

In our data the first principal component have nearly equal loadings for all original covariates, and can be seen as an average effect. The second principal component gives a high value for the first absorbances, then low, then high and then low again. The third principal component gives positive values for the first absorbances, and then negative values for the last absorbances. Observe the high degree of smoothness in the loadings.

The percentage of total variance explained by the first three principal components are 99.875%.

The PC scores can be used as covariates into a regression, taking the place of the original covariates. If all scores are used the regression in the PCs will be mathematically the same as the regression in the original variables. If less than p scores are used, then this will result in a shrinkage effect on the coefficients, similar to performin a rigde regression analysis. Using a few PCs as covariates might help towards possible overfitting (and also towards multicollinearity).

Problem 3 Design of experiments

a) Design matrix:

	A	B	C	D	AB
1	−1	−1	−1	1	1
2	1	−1	−1	−1	−1
3	−1	1	−1	−1	−1
4	1	1	−1	1	1
5	−1	−1	1	1	1
6	1	−1	1	−1	−1
7	−1	1	1	−1	−1
8	1	1	1	1	1

What type of experiment is this?
We see that we have a full factorial design in the factors A, B, C, but there is a fourth factor D added. This is a half fraction of a 2^4 design, also called a 2^{4-1}-design.

What is the generator and the defining relation for the experiment?
The generator for the design is D=AB (which is seen from the table above after the AB column is added). The defining relation is then I=ABD.

What is the resolution of the experiment?
The resolution of the design equals the number of letters in the defining relation, thus the resolution is III.

Write down the alias structure of the experiment.
A=BD, B=AD, C=ABCD, D=AB
AC=BCD, BC=ACD, CD=ABC
I=ABD

Problem 4 Multiple linear regression

Define the matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

a) $\boldsymbol{H}$ is a symmetric projection matrix, since $\boldsymbol{H}^T = \boldsymbol{H}$ and $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$. For a symmetric and idempotent matrix the rank is equal to the trace. The rank of $\boldsymbol{H}$ is p. See proof below.

$$
\begin{aligned}
\boldsymbol{H} &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \\
\boldsymbol{H}^T &= (\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)^T = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{H} \\
\boldsymbol{H}^2 &= (\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{H} \\
\mathrm{tr}(\boldsymbol{H}) &= \mathrm{tr}(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T) = \mathrm{tr}(\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}) = \mathrm{tr}(\boldsymbol{I}_{p\times p}) = p
\end{aligned}
$$

Graphically: The vector $\boldsymbol{HY}$ is a projection of the vector $\boldsymbol{Y}$ onto the space spanned by the columns of $\boldsymbol{X}$.

The matrix $\boldsymbol{I}-\boldsymbol{H}$ is also symmetric and idempotent, and thus a symmetric projection matrix. The rank of $\boldsymbol{I}-\boldsymbol{H}$ is $n-p$. See proof below.

$$
\begin{aligned}
(\boldsymbol{I}-\boldsymbol{H})^T &= \boldsymbol{I}-\boldsymbol{H}^T = \boldsymbol{I}-\boldsymbol{H} \\
(\boldsymbol{I}-\boldsymbol{H})^2 &= (\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{I}-\boldsymbol{H}) = \boldsymbol{I}-2\boldsymbol{H}+\boldsymbol{H}^2 = \boldsymbol{I}-\boldsymbol{H} \\
\mathrm{tr}(\boldsymbol{I}-\boldsymbol{H}) &= \mathrm{tr}(\boldsymbol{I}_{n\times n}) - \mathrm{tr}(\boldsymbol{H}) = n-p
\end{aligned}
$$

Graphically: The vector $(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}$ is a projection of the vector $\boldsymbol{Y}$ onto the space othogonal to the space spanned by the columns of $\boldsymbol{X}$.

b) Let $\mathrm{SSE} = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}$. Derive the distribution of SSE.

One of the key theorems of this course (theorem 3.26 in Bingham&Fry), state that if $\boldsymbol{D}$ is a symmetric and idempotent matrix with rank r and $\boldsymbol{Z} \sim N_n(0, \sigma^2\boldsymbol{I})$, then $\boldsymbol{Z}^T\boldsymbol{D}\boldsymbol{Z} \sim \sigma^2\chi^2_r$.

We have $\boldsymbol{D} = (\boldsymbol{I}-\boldsymbol{H})$ symmetric and idempotent with rank $n-p$, and $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$. To use the theorem we need to look at $\boldsymbol{Y}^* = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$.

$$
\begin{aligned}
(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}^* &= (\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) \\
&= (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y} - (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{X}\boldsymbol{\beta} = (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y} - (\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{H}\boldsymbol{X}\boldsymbol{\beta}) = \\
&= (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y} - (\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}
\end{aligned}
$$

since $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{X}$. Projecting $\boldsymbol{X}$ onto the space spanned by the columns of $\boldsymbol{X}$ gives $\boldsymbol{X}$.

Thus, we have shown that $\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y} = \boldsymbol{Y}^{*T}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}^*$, and we may use the theorem to conclude that $\mathrm{SSE} \sim \sigma^2\chi^2_{n-p}$.

The mean of a χ^2-distributed variable equals the number of degrees of freedom, so

$$
\begin{aligned}
\mathrm{E}(\frac{\mathrm{SSE}}{\sigma^2}) &= n-p \\
\mathrm{E}(\mathrm{SSE}) &= (n-p)\sigma^2 \\
\mathrm{E}(\frac{\mathrm{SSE}}{n-p}) &= \sigma^2
\end{aligned}
$$

Thus, $\hat{\sigma}^2 = \frac{\text{SSE}}{n-p}$ will be an unbiased estimator for σ^2.

Variance:

$$\begin{aligned}\text{Var}\left(\frac{\text{SSE}}{n-p}\right) &= \frac{1}{(n-p)^2}\text{Var}(\text{SSE}) = \frac{1}{(n-p)^2}\text{Var}(\sigma^2\frac{\text{SSE}}{\sigma^2}) \\ &= \frac{1}{(n-p)^2}2(n-p)\sigma^4 = \frac{2\sigma^4}{n-p}\end{aligned}$$

c)

$$\begin{aligned}\boldsymbol{A} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \\ \boldsymbol{B} &= \boldsymbol{I} - \boldsymbol{H} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\end{aligned}$$

Here $\boldsymbol{A}$ is a $p \times n$ matrix (since $\boldsymbol{X}$ is $n \times p$), and $\boldsymbol{B}$ is the same dimension as $\boldsymbol{H}$, that is, $n \times n$, and is symmetric and idempotent (found previously).

Since $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ then $\boldsymbol{AY}$ and $\boldsymbol{BY}$ are independent random variables if $\sigma^2\boldsymbol{A}\boldsymbol{B}^T = \boldsymbol{0}$.

$$\begin{aligned}\boldsymbol{A}\boldsymbol{B}^T &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T) \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = 0\end{aligned}$$

We have proven that $\boldsymbol{Z}_1 = \boldsymbol{AY}$ and $\boldsymbol{Z}_2 = \boldsymbol{BY}$ are independent random variables. Then it follows that $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2^T\boldsymbol{Z}_2$ are also independent random variables. Since $\boldsymbol{Z}_1 = \hat{\boldsymbol{\beta}}$ and $\boldsymbol{Z}_2^T\boldsymbol{Z}_2 = \boldsymbol{Y}^T\boldsymbol{BY} = \text{SSE}$, we have proven that $\hat{\boldsymbol{\beta}}$ and SSE are independent random variables.

Use in MLR: The independence of $\hat{\boldsymbol{\beta}}$ and SSE is used in the construction of a t-test for hypothesis about $\boldsymbol{\beta}$.