1 Some preliminaries

1.1 Formulas for expectation and covariance

For the expectation $E(X)$ we have the simple rules (lower-case character denoting constants)
1) $E(aX + b) = aE(X) + b$
2) $E(X + Y) = E(X) + E(Y)$.
Note that the property 2) holds whether the variables are independent or not.
For the covariance $C(X, Y) = E((X - E(X))(Y - E(Y)))$ we have the following properties
1) $V(X) = C(X, X)$
2) $C(X, Y) = C(Y, X)$
3) $C(aX, Y) = aC(X, Y)$
4) $C(X + a, Y) = C(X, Y)$
5) $C(X + Y, Z) = C(X, Z) + C(Y, Z)$
which are easily shown using the definition. Hence we have formulas like

$$C(aX + bY + c, dZ + eW) = adC(X, Z) + aeC(X, W) + bdC(Y, Z) + beC(Y, Z).$$

For a stochastic N-dimesnional vector $\boldsymbol{X} = (X_1, X_2, \cdots, X_N)^T$ we call
$E(\boldsymbol{X}) = (E(X_1), E(X_2), \cdots, E(X_N)^T)$ the expectation of the vector. We define the covariance matrix $C_{\boldsymbol{X}}$ of $\boldsymbol{X}$ by the $N \times N$ matrix having $C(X_i, X_j)$ as element (i, j). Note that the diagonal elements are the variances $V(X_1), V(X_2), \cdots, V(X_N)$. If $C(X_i, X_j) = 0$ we say that X_i and X_j are uncorrelated. If variables are independent then they are uncorrelated, but in general uncorrelated variables need not be independent. In the special and important case of multidimensional normal distribution the fact that variables are uncorrelated imply that they are independent.

A useful representation is that $C_{\boldsymbol{X}} = E((\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{X} - E(\boldsymbol{X}))^T)$.

Sometimes we denote $C_{\boldsymbol{X}} = V(\boldsymbol{X})$. We also define $C(\boldsymbol{X}, \boldsymbol{Y}) = E((\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{Y} - E(\boldsymbol{Y}))^T)$. By using the rules stated above it is easy to see that $E(A\boldsymbol{X} + b) = AE(\boldsymbol{X}) + b$ where A is a matrix and b is a vector. In addition we have $C(A\boldsymbol{X}, B\boldsymbol{Y}) = AC(\boldsymbol{X}, \boldsymbol{Y})B^T$ with the special case $V(A\boldsymbol{X}) = AV(\boldsymbol{X})A^T$.

1.2 Some linear algerbra

The covariance matrix $V(\boldsymbol{X})$ must be symmetric since $C(X_i, X_j) = C(X_j, X_i)$. A symmetric matrix C can be diagonalized, i.e. you can find a (orthonormal) coordinate system in which is represented by a diagonal matrix, i.e. one that has 0:s for the non-diagonal elements. Hence we can find an orthogonal matrix U (corresponding to a change of coordinate system - a rotation and possible change of coordinate directions) such that $U^T C U = D$ is diagonal , i.e.

$$D = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & 0 & 0 & \cdots & \lambda_N \end{pmatrix} = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_N).$$

Orthogonal matrices have the property that $UU^T = U^T U = I$, so $U^{-1} = U^T$. Therefore we also have $C = UDU^T$. The values $\lambda_1, \lambda_2, \cdots, \lambda_N$ are called eigenvalues and corresponding to these there are eigenvectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_N$ and $C\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i$. The eigenvectors corresponding to different eigenvalues are orthogonal. If several eigenvalues are identical there is some freedom in chosing the eigenvectors, but they can always be chosen as orthogonal. The eigenvectors hence can be the base vectors for an orthonormal (ON) coordinate system.

For a covariance matrix all eigenvalues are non-negative. To see this we let $\boldsymbol{b} = (b_1, b_2, \cdots, b_N)^T$ be arbitrary and form $Y = \boldsymbol{b}^T\boldsymbol{X} = \sum_1^n b_i X_i$. Denote $V(\boldsymbol{X}) = C$. Since variances are non-negative we have

$$0 \leq V(\boldsymbol{b}^T\boldsymbol{X}) = \boldsymbol{b}^T C\boldsymbol{b} = \boldsymbol{b}^T UDU^T\boldsymbol{b} = (U^T\boldsymbol{b})^T D(U^T\boldsymbol{b})$$

where we also used the fact that $(AB)^T = B^T A^T$ - note the reversed order. Since $\boldsymbol{b}$ was arbitrary we can write this as

$$0 \leq \boldsymbol{a}^T D\boldsymbol{a} = \sum_1^n \lambda_i a_i^2$$

for arbitrary $\boldsymbol{a}$-vectors (choose $\boldsymbol{a} = U^T\boldsymbol{b}$ i.e. $\boldsymbol{b} = U\boldsymbol{a}$). Since $\sum_1^n \lambda_i a_i^2 \geq 0$ for *all* choices of $\boldsymbol{a}$ this requires $\lambda_1, \lambda_2, \cdots, \lambda_N \geq 0$. E.g., we can take $\boldsymbol{a} = \boldsymbol{e}_i = (0, \cdots, 0, 1, 0, \cdots, 0)^T$ i.e. as the i:th unite vector. Hence all eigenvalues are non-negative. C is said to be positive semi-definite. If all eigenvalues are strictly positive C is said to be positive definite.

We now want to find a matrix A such that $AA^T = C$ where C is a covariance matrix (which therefore is positive semi-definite). This is tantamount to finding a "square root " of C. As for ordinary square roots it is not unique. One of many choices is the following. Write $C = UDU^T$ where $D = \text{diagonal}(\lambda_1, \lambda_2, \cdots, \lambda_N)$, i.e. diagonalize C. Let $D^{1/2} = \text{diagonal}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \cdots, \sqrt{\lambda_N})$. (Note that this is well-defined since $\lambda_i \geq 0$ for all i). We can choose $A = UD^{1/2}$ and then $AA^T = (UD^{1/2})(UD^{1/2})^T = UD^{1/2}D^{1/2}U^T$ since $D^{1/2}$ is symmetric. Hence we get $AA^T = UDU^T = C$. In fact there is a special choice $C^{1/2}$ as the (in fact unique) symmetric matrix which is itself positive semi -definite satisfying $C = C^{1/2}C^{1/2}$. Matlab finds this matrix with the function *sqrtm*.

A projection P is a symmetric matrix satisfying $PP = P$. Hence the eigenvalues are either 0 or 1. Since if $Px = \lambda x$ then $Px = P^2x = \lambda^2 x$ so $\lambda^2 = \lambda$ and hence λ is 0 or 1. Hence it is easy to see that P projects a vector onto a subspace with dimension equal to the number of eigenvalues which are 1.

A useful concept in this setting is the trace of a matrix, i.e. $\text{trace}(A) = \sum_{i=1}^N a_{ii}$ which is the sum of the diagonal elements of the matrix. In fact we have $\text{trace}(AB) = \text{trace}(BA)$ since the (i,i)-element of AB is $(AB)_{ii} = \sum_j a_{ij}b_{ji}$ so

$$\text{trace}(AB) = \sum_i (AB)_{ii} = \sum_i \sum_j a_{ij}b_{ji} = \sum_j \sum_i b_{ji}a_{ij} = \sum_j (BA)_{jj} = \text{trace}(BA).$$

Therefore for a covariance matrix C we get

$\text{trace}(C) = \text{trace}(UDU^T) = \text{trace}(U^TUD) = \text{trace}(D)$=sum of the eigenvalues.

In addition we see that projection P has $\text{trace}(P)$=dimension of the subspace it projects unto.

2 Multidimensional normal distributions

As is well known the standard (one-dimensional) normal distribution $N(0,1)$ has the density

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2), \ -\infty < x < \infty.$$

The general (one-dimensional) normal distribution $N(\mu, \sigma^2)$ (note that we let the second parameter be the variance and not the standard deviation) has the density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}), \ -\infty < x < \infty.$$

In fact X is $N(\mu,\sigma^2)$ if and only if $X=\mu+\sigma Z$ where Z is $N(0,1)$ so the general (one-dimensional) normal distribution could be defined as what you get when $X=\mu+\sigma Z$ and Z is $N(0,1)$.

We now define the multidimensional normal distribution in a similar manner.
Let $\boldsymbol{Z}=(Z_1,Z_2,\cdots,Z_m)^T$ be m independent $N(0,1)$-distributed variables. Notice that $E(\boldsymbol{Z})=\boldsymbol{0}$ and $V(\boldsymbol{Z})=I_m$. We now let $\boldsymbol{\mu}$ be a $N\times 1$-vector and A a $N\times m$ matrix and form

$$\boldsymbol{X}=\boldsymbol{\mu}+A\boldsymbol{Z}$$

so $\boldsymbol{X}$ is N-dimensional variable. We see that $E(\boldsymbol{X})=\boldsymbol{\mu}+AE(\boldsymbol{Z})=\boldsymbol{\mu}$ and $V(\boldsymbol{X})=V(\boldsymbol{\mu}+A\boldsymbol{Z})=AV(\boldsymbol{Z})A^T=AA^T$. We will say that $\boldsymbol{X}$ is $N(\boldsymbol{\mu},AA^T)$. Of course $AA^T=C$ is symmetric since its is a covariance matrix. The interesting fact is that if a covariance matrix C is given and we construct a matrix A such that $AA^T=C$ then the distribution of $\boldsymbol{X}=\boldsymbol{\mu}+A\boldsymbol{Z}$ depends only on C. Hence if $BB^T=C$ (so $AA^T=BB^T$) is another choice the distribution of $\boldsymbol{X'}=\boldsymbol{\mu}+B\boldsymbol{Z'}$ is the same as that of $\boldsymbol{X}=\boldsymbol{\mu}+A\boldsymbol{Z}$. If the rank of C, i.e. the number of linearly independent columns is r then we need at least r different Z-variables.

The construction above means that there are "building blocks" in the form of the Z-variables and that each X_i is a certain linear combination of these Z-variables. Notice that typically the different X-variables are dependent since the same Z-variables occur in these linear combinations. So if a certain Z_i has an extreme value all the X-variables containing it have extreme values.

For completeness we give the following theorem.

Theorem: The subspaces spanned by the columns of A and AA^T are identical.

Proof: We denote these subspaces as $M(A)$ and $M(AA^T)$. Let A be a $n\times m$-matrix so AA^T is a $n\times n$-matrix. If we take an arbitrary vector $\boldsymbol{x}$ which is orthogonal to all the m columns in A, i.e. we have $\boldsymbol{x}^TA=\boldsymbol{0}^T$. Then $\boldsymbol{x}^TAA^T=\boldsymbol{0}^T$, i.e. $\boldsymbol{x}$ is orthogonal to all the n columns in AA^T. This means that $M(AA^T)$ is in $M(A)$.

On the other hand we can take an arbitrary n-dimensional column vector $\boldsymbol{x}$ which is orthogonal to all the n columns in $M(AA^T)$, i.e. $\boldsymbol{x}^TAA^T=\boldsymbol{0}^T$. Then we have $\boldsymbol{x}^TAA^T\boldsymbol{x}=0$ i.e. $(A^T\boldsymbol{x})^T(A^T\boldsymbol{x})=0$ which implies $A^T\boldsymbol{x}=\boldsymbol{0}$. This means that $\boldsymbol{x}^TA=\boldsymbol{0}^T$ implying that $\boldsymbol{x}$ is orthogonal to all the columns in$A$. Hence $M(A)$ is in $M(AA^T)$.

Since $M(A)$ contains $M(AA^T)$ and $M(AA^T)$ contains $M(A)$ we have $M(A)=M(AA^T)$, QED.

A consequence of this is that if $AA^T=BB^T$ then $M(A)=M(AA^T)=$(since $AA^T=BB^T$) $=M(BB^T)=M(B)$. The subspace we can get by forming $\boldsymbol{X}=A\boldsymbol{Z}$ and $\boldsymbol{X'}=B\boldsymbol{Z'}$ respectively are identical, i.e. the probability mass is on the same subspace whether we study $\boldsymbol{X}$ or $\boldsymbol{X'}$. This subspace has the dimension $r=$rank for A=rank for $C=AA^T$.

Choosing $m>r=$the rank for C means that we really have "unneccessary" Z-variables. Since the rank for $C=AA^T$ is r we could find an orthogonal transformation U such that UAA^TU^T looks like this

$$\begin{pmatrix} D_r & 0 \\ 0 & 0 \end{pmatrix}$$

where $D_r=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_r)$ and all these are >0, so D_r is invertible.
In fact if $V(\boldsymbol{X})=C$ is invertible and $r\times r$ it can be shown that the density of $\boldsymbol{X}$ is

$$f_{\boldsymbol{X}}(\boldsymbol{x})=\frac{1}{(2\pi)^{r/2}\sqrt{det(C)}}\exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^TC^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$$

so it only depends on C and $\boldsymbol{\mu}$.

In $\boldsymbol{X}$ is $N(E(\boldsymbol{X}),C)$ and $c_{ij}=C(X_i,X_j)=0$ then we will show below that X_i and X_j are independent.Hence for multidimensional normal distributions uncorrelated variables are independent.

A little more general we study $Y_1 = a_1^T \boldsymbol{X}$ and $Y_2 = a_2^T \boldsymbol{X}$. If $a_1^T C a_2 = 0$ then Y_1 and Y_2 are independent. This holds since then $C(Y_1, Y_2) = C(a_1^T \boldsymbol{X}, a_2^T \boldsymbol{X}) = a_1^T C a_2 = 0$ which means that

$$\boldsymbol{Y} = (Y_1, Y_2)^T \text{ is } N(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix})$$

where $\mu_1 = E(Y_1) = a_1^T E(\boldsymbol{X})$ and $\mu_2 = E(Y_2) = a_2^T E(\boldsymbol{X})$ and $\sigma_1^2 = V(Y_1) = a_1^T C a_1$ and $\sigma_2^2 = V(Y_2) = a_2^T C a_2$. This means that we can write $Y_1 = \mu_1 + \sigma_1 Z_1$ and $Y_2 = \mu_2 + \sigma_2 Z_2$ where Z_1 and Z_2 are independent standard normally distributed and therefore Y_1 and Y_2 are independent.

3 Estimation in the linear model

We have the linear model

$$\boldsymbol{Y} = A\boldsymbol{\theta} + \boldsymbol{\varepsilon} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$$

where $\dim(\boldsymbol{\theta}) = \text{Rank}(A) = k$ and ε is $N(\boldsymbol{0}, \sigma^2 I_N)$. This means that $\boldsymbol{\mu} = A\boldsymbol{\theta}$ lies in a k-dimensional subspace U_k of R^N which is the subspace spanned by the columns in A.

In estimating $\boldsymbol{\theta}$ we use the Least Square method (LS-method). Having the observations $\boldsymbol{y}$ this means that we try to minimize as a function of $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_k)^T$

$$Q(\boldsymbol{\theta}) = ||\boldsymbol{y} - A\boldsymbol{\theta}||^2 = \sum_{i=1}^{N} (y_i - \mu_i)^2 = \sum_{i=1}^{N} (y_i - \sum_{j=1}^{N} a_{ij}\theta_j)^2$$

First we give a geometric argument. When $\boldsymbol{\theta}$ varies the vector $\boldsymbol{\mu} = A\boldsymbol{\theta}$ spans the k-dimensional subspace U_k of R^N. The function Q is minimized when $\boldsymbol{\theta}$ is chosen such that $A\boldsymbol{\theta}$ is the orthogonal projection onto U_k. This means that the minimizing parameter vector $\widehat{\boldsymbol{\theta}}$ is such that $\boldsymbol{y} - A\widehat{\boldsymbol{\theta}}$ is orthogonal to each vector $\boldsymbol{\mu} = A\boldsymbol{\theta}$ in the subspace U_k. Hence we have the scalar product

$$(A\boldsymbol{\theta})^T (\boldsymbol{y} - A\widehat{\boldsymbol{\theta}}) = 0.$$

If this is to hold for all $\boldsymbol{\theta}$ this means $A^T(\boldsymbol{y} - A\widehat{\boldsymbol{\theta}}) = 0$ and hence

$$(A^T A)\widehat{\boldsymbol{\theta}} = A^T \boldsymbol{y}$$

which is sometimes called the normal equation.

A more direct and less abstract approach would be to explicitely write down

$$Q(\boldsymbol{\theta}) = ||\boldsymbol{y} - A\boldsymbol{\theta}||^2 = \sum_{i=1}^{N} (y_i - \mu_i)^2 = \sum_{i=1}^{N} (y_i - \sum_{j=1}^{N} a_{ij}\theta_j)^2.$$

Taking the partial derivative with respect to θ_j we get (denoting $a_{ij} = a_{ji}^T$)

$$\frac{\partial Q}{\partial \theta_j} = -2 \sum_{i=1}^{N} (y_i - \sum_{j=1}^{N} a_{ij}\theta_j) a_{ij} = -2 \sum_{i=1}^{N} a_{ji}^T (y_i - \sum_{j=1} a_{ij}\theta_j)$$

which is the j:th component of $-2A^T(\boldsymbol{y} - A\boldsymbol{\theta})$ so we again get the equation $(A^T A)\boldsymbol{\theta} = A^T \boldsymbol{y}$.

If A has full rank, i.e. the columns of A are linearly independent then $A^T A$ is invertible and we have the estimate of $\boldsymbol{\theta}$ is $\widehat{\boldsymbol{\theta}} = (A^T A)^{-1} A^T \boldsymbol{Y}$. The predictor is $\widehat{\boldsymbol{\mu}} = A\widehat{\boldsymbol{\theta}} = A(A^T A)^{-1} A^T \boldsymbol{Y}$ which lies in U_k. Letting $P_{U_k} = A(A^T A)^{-1} A^T$ we see that it is a projection onto U_k. Note that $P_{U_k} = P_{U_k}^2$ and that it is symmetric.

4 Distribution of the estimates

The estimate is $\widehat{\boldsymbol{\theta}} = (A^TA)^{-1}A^T\boldsymbol{Y}$ and since $\boldsymbol{Y}$ is $N(A\boldsymbol{\theta}, \sigma^2 I_N)$ so we get

$$E(\widehat{\boldsymbol{\theta}}) = E((A^TA)^{-1}A^T\boldsymbol{Y}) = (A^TA)^{-1}A^TE(\boldsymbol{Y}) = (A^TA)^{-1}A^TA\boldsymbol{\theta} = \boldsymbol{\theta}$$

so the LS-estimate is unbiased. We also have $E(\widehat{\boldsymbol{\mu}}) = E(A\widehat{\boldsymbol{\theta}}) = AE(\widehat{\boldsymbol{\theta}}) = A\boldsymbol{\theta} = \boldsymbol{\mu}$ and therefore $\widehat{\boldsymbol{\mu}}$ is an unbiased estimate of $\boldsymbol{\mu}$. The covariance matrix is

$$V(\widehat{\boldsymbol{\theta}}) = V((A^TA)^{-1}A^T\boldsymbol{Y}) = (A^TA)^{-1}A^TV\boldsymbol{Y}((A^TA)^{-1}A^T)^T =$$

$$(A^TA)^{-1}A^TV\sigma^2 I_N(A^TA)^{-1}A^T)^T = \sigma^2(A^TA)^{-1}.$$

so $\widehat{\boldsymbol{\theta}}$ is $N(\boldsymbol{\theta}, \sigma^2(A^TA)^{-1})$. If $C(\widehat{\theta}_i, \widehat{\theta}_j) = (A^TA)^{-1}_{ij} = 0$ then $\widehat{\theta}_i$ and $\widehat{\theta}_j$ are independent.

5 Gauss-Markov theorem

We now know how to estimate $\boldsymbol{\theta}$ by $\widehat{\boldsymbol{\theta}}$, but what if we want to estimate $b^T\boldsymbol{\theta}$, i.e. a linear combination of the θ-parameters? One obvious choice is to estimate it by $b^T\widehat{\boldsymbol{\theta}}$, i.e. by the corresponding linear combination of the $\boldsymbol{\theta}$-estimate. The Gauss-Markov theorem states that in fact this is the best way (least variance) to estimate $b^T\boldsymbol{\theta}$ of all unbiased estimates which are <u>linear</u> in $\boldsymbol{Y}$. So assume we have an estimate $c^T\boldsymbol{Y}$ of $b^T\boldsymbol{\theta}$. If it is to be unbiased we must have $E(c^T\boldsymbol{Y}) = b^T\boldsymbol{\theta}$ and hence since $E(c^T\boldsymbol{Y}) = c^TE(\boldsymbol{Y}) = c^TA\boldsymbol{\theta}$ and this has to be $b^T\boldsymbol{\theta}$ for <u>all</u> $\boldsymbol{\theta}$ we must have $c^TA = b^T$.

By adding and subtracting $b^T\widehat{\boldsymbol{\theta}}$ we get

$$V(c^T\boldsymbol{Y}) = V((c^T\boldsymbol{Y} - b^T\widehat{\boldsymbol{\theta}}) + b^T\widehat{\boldsymbol{\theta}}) =$$

$$V(c^T\boldsymbol{Y} - b^T\widehat{\boldsymbol{\theta}}) + V(b^T\widehat{\boldsymbol{\theta}}) + 2C(c^T\boldsymbol{Y} - b^T\widehat{\boldsymbol{\theta}}, b^T\widehat{\boldsymbol{\theta}})$$

using the formula $V(X+Y) = V(X) + V(Y) + 2C(X,Y)$. As we show below $C(c^T\boldsymbol{Y} - b^T\widehat{\boldsymbol{\theta}}, b^T\widehat{\boldsymbol{\theta}}) = 0$ so $V(c^T\boldsymbol{Y}) \geq V(b^T\widehat{\boldsymbol{\theta}})$ since $V(c^T\boldsymbol{Y} - b^T\widehat{\boldsymbol{\theta}}) \geq 0$.

What remains is showing that the covariance is 0. We get using $C(A\boldsymbol{X}, B\boldsymbol{Y}) = AC(\boldsymbol{X}, \boldsymbol{Y})B^T$

$$C(c^T\boldsymbol{Y} - b^T\widehat{\boldsymbol{\theta}}, b^T\widehat{\boldsymbol{\theta}}) = C((c^T - b^T(A^TA)^{-1}A^T)\boldsymbol{Y}, b^T(A^TA)^{-1}A^T)\boldsymbol{Y}) =$$

$$(c^T - b^T(A^TA)^{-1}A^T)\sigma^2 I_N(b^T(A^TA)^{-1}A^T)^T = (c^TA - b^T)(A^TA)^{-1}b\sigma^2 = 0$$

since $c^TA = b^T$ by the requirement that $c^T\boldsymbol{Y}$ should be an unbiased estimate of $b^T\boldsymbol{\theta}$.

6 Estimating σ^2

We have $I = I - P_{U_k} + P_{U_k}$ and they are both projections (onto U_k and onto the rest of R^N respectively) and they are orthogonal since $P_{U_k}(I - P_{U_k}) = P_{U_k} - P_{U_k}^2 = 0$. By multiplying $\boldsymbol{Y} - \boldsymbol{\mu}$ we get

$$\boldsymbol{Y} - \boldsymbol{\mu} = (I - P_{U_k})(\boldsymbol{Y} - \boldsymbol{\mu}) + P_{U_k}(\boldsymbol{Y} - \boldsymbol{\mu}) = (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}) + (\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})$$

since $\boldsymbol{\mu} \in U_k$ and therefore $P_{U_k}\boldsymbol{\mu} = \boldsymbol{\mu}$ and $\widehat{\boldsymbol{\mu}} = P_{U_k}\boldsymbol{Y}$. Since the terms are orthogonal and normally distributed they are independent. We get

$$||\boldsymbol{Y} - \boldsymbol{\mu}||^2 = ||\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}||^2 + ||\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}||^2$$

and the terms are independent. We see that since $\boldsymbol{Y}$ is $N(\boldsymbol{\mu}, \sigma^2 I_N)$ we have $\boldsymbol{Y} = \boldsymbol{\mu} + \sigma \boldsymbol{Z}$ where $\boldsymbol{Z}$ is $N(\boldsymbol{0}, I_N)$ so

$$||\boldsymbol{Y} - \boldsymbol{\mu}||^2/\sigma^2 = \sum_{i=1}^{N} Z_i^2$$

and is therefore $\chi^2(N)$-distributed.

Orthogonal transformations (rotations) do not change lengths e.g. $||\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}||^2$ or $||\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}||^2$. E.g. if $\boldsymbol{Y}' = U\boldsymbol{Y}$ and $\boldsymbol{\mu}' = U\boldsymbol{\mu}$ then (note that $UU^T = U^T U = I_N$) we get

$$||\boldsymbol{Y} - \boldsymbol{\mu}||^2 = (\boldsymbol{Y} - \boldsymbol{\mu})^T(\boldsymbol{Y} - \boldsymbol{\mu}) = (\boldsymbol{Y} - \boldsymbol{\mu})^T U^T U(\boldsymbol{Y} - \boldsymbol{\mu}) = (U(\boldsymbol{Y} - \boldsymbol{\mu}))^T(U(\boldsymbol{Y} - \boldsymbol{\mu})) = ||\boldsymbol{Y}' - \boldsymbol{\mu}'||^2$$

We now rotate by using an orthogonal matix U such that the first k coordinates are in U_k and the rest of the $N - k$ coordinates are in the rest of R^N. We then get $\boldsymbol{Y}' = U\boldsymbol{Y}$ and $\boldsymbol{\mu}' = U\boldsymbol{\mu}$. In fact we have $\widehat{\boldsymbol{\mu}}' = (Y_1', Y_2', \cdots, Y_k', 0, 0, \cdots, 0)^T$ (a projection onto the first k coordinates) and $\boldsymbol{\mu}' = (\mu_1', \mu_2', \cdots, \mu_k', 0, 0, \cdots, 0)^T$ and therefore

$$||\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}||^2/\sigma^2 = ||\boldsymbol{Y}' - \widehat{\boldsymbol{\mu}}'||^2/\sigma^2 = \sum_{i=k+1}^{N} (\frac{Y_i' - \mu_i'}{\sigma})^2 = \sum_{i=k+1}^{N} Z_i^2$$

where the Z_i are independent $N(0,1)$ and hence it has a $\chi^2(N-k)$-distribution. In the same manner

$$||\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}||^2/\sigma^2 = ||\widehat{\boldsymbol{\mu}}' - \boldsymbol{\mu}'||^2/\sigma^2 = \sum_{i=1}^{k} (\frac{Y_i' - \mu_i'}{\sigma})^2 = \sum_{i=1}^{k} Z_i^2$$

so it has a $\chi^2(k)$-distribution. The $\chi^2(f)$-distribution has mean f so we see that

$$\widehat{\sigma}^2 = \frac{||\boldsymbol{y} - \widehat{\boldsymbol{\mu}}||^2}{N - k} \text{ is an unbiased estimate of } \sigma^2.$$

Notice also that the terms are independent since they are sums of non-overlapping independent Z-variables

7 Test of linear hypothesis

We have the basic linear model

$$\boldsymbol{Y} = A\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

where $\dim(\boldsymbol{\theta})$ =Rank$(A) = k$ and ε is $N(\boldsymbol{0}, \sigma^2 I_N)$. For simplicity we assume that is is a full-rank model, i.e. that the columns in A are lineraly independent.

In a hypothesis model H_0 we have q linearly independent restrictions on $\boldsymbol{\theta}$. We have k θ-variables and q of these can be expressed in the remaining $k - q = l$ when H_0 is true. So in fact we have l "genuine" parameters in the hypothesis model. Let these be $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_l)^T$ and hence $\boldsymbol{\theta} = B\boldsymbol{\lambda}$ for some matrix B. Note that in the hypothesis model $\boldsymbol{\mu}$ lies in a l-dimensional subspace U_l of the k-dimensional subspace U_k which is what is spanned by the columns of A.

Therefore the hypothesis model is $\boldsymbol{Y} = (AB)\boldsymbol{\lambda} + \boldsymbol{\varepsilon}$ so the design matrix is $AB = C$ which gives the LS-estimate

$$\widehat{\widehat{\boldsymbol{\lambda}}} = ((AB)^T(AB))^{-1}(AB)^T\boldsymbol{Y} = (C^T C)^{-1}C^T\boldsymbol{Y}$$

and the predictor

$$\widehat{\widehat{\boldsymbol{\mu}}} = AB\widehat{\widehat{\boldsymbol{\lambda}}}.$$

Notice that $C(C^TC)^{-1}C^T = P_{U_l}$ is a projection onto the l-dimensional subspace U_l. We have U_l as a subspace in U_k and that therefore $P_{U_l}P_{U_k} = P_{U_l}$.

We have $I_N = (I_N - P_{U_k}) + (P_{U_k} - P_{U_l}) + P_{U_l}$ and these terms are orthogonal since they project onto orthogonal subspaces. If we assume that H_0 is true then $P_{U_l}\boldsymbol{\mu} = \boldsymbol{\mu}$. Then we have

$$\boldsymbol{Y} - \boldsymbol{\mu} = (I_N - P_{U_k})(\boldsymbol{Y} - \boldsymbol{\mu}) + (P_{U_k} - P_{U_l})(\boldsymbol{Y} - \boldsymbol{\mu}) + P_{U_l}(\boldsymbol{Y} - \boldsymbol{\mu}) =$$

$$(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}) + (\widehat{\boldsymbol{\mu}} - \widehat{\widehat{\boldsymbol{\mu}}}) + (\widehat{\widehat{\boldsymbol{\mu}}} - \boldsymbol{\mu})$$

which are orthogonal and hence

$$||\boldsymbol{Y} - \boldsymbol{\mu}||^2 = ||\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}||^2 + ||\widehat{\boldsymbol{\mu}} - \widehat{\widehat{\boldsymbol{\mu}}}||^2 + ||\widehat{\widehat{\boldsymbol{\mu}}} - \boldsymbol{\mu}||^2.$$

We now choose a change of coordinates by use of an orthogonal matrix U such that the first l coordinates are in U_l, the next $k - l$ are in the rest of U_k and the remaining $N - k$ coordinates complete the system to the whole of R^N. This does not change the lengths of the vectors so with $\boldsymbol{Y}' = U\boldsymbol{Y}$ and $\boldsymbol{\mu}' = U\boldsymbol{\mu}$ we get

$$||\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}||^2 = ||\boldsymbol{Y}' - \widehat{\boldsymbol{\mu}'}||^2 = \sum_{i=k+1}^{N} Y_i'^2$$

$$||\widehat{\boldsymbol{\mu}} - \widehat{\widehat{\boldsymbol{\mu}}}||^2 = ||\widehat{\boldsymbol{\mu}'} - \widehat{\widehat{\boldsymbol{\mu}'}}||^2 = \sum_{i=l+1}^{k} Y_i'^2$$

$$||\widehat{\widehat{\boldsymbol{\mu}}} - \boldsymbol{\mu}||^2 = ||\widehat{\widehat{\boldsymbol{\mu}'}} - \boldsymbol{\mu}'||^2 = \sum_{i=1}^{l}(Y_i' - \mu_i')^2$$

which therefore are $\sigma^2\chi^2(N-k)$, $\sigma^2\chi^2(k-l)$ and $\sigma^2\chi^2(l)$ and independent.

We have since $I_N - P_{U_l} = (I_N - P_{U_k}) + (P_{U_k} - P_{U_l})$ that

$$||\boldsymbol{Y} - \widehat{\widehat{\boldsymbol{\mu}}}||^2 = ||\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}||^2 + ||\widehat{\boldsymbol{\mu}} - \widehat{\widehat{\boldsymbol{\mu}}}||^2$$

whether H_0 is true or not. If H_0 is true then

$$\widehat{\widehat{\sigma}}^2 = \frac{||\widehat{\boldsymbol{\mu}} - \widehat{\widehat{\boldsymbol{\mu}}}||^2}{k-l}$$

and

$$\widehat{\sigma}^2 = \frac{||\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}||^2}{N-k}$$

are independent estimates unbiased estimates of σ^2 on $k-l$ and $N-k$ degrees of freedom. Their quotient is therefore $F(k-l, N-k)$-distributed when H_0 is true and the observed value and can be used as a test quotient for testing H_0.

We have $\widehat{\boldsymbol{\mu}} - \widehat{\widehat{\boldsymbol{\mu}}} = (P_{U_k} - P_{U_l})\boldsymbol{Y}$ and therefore

$$||\widehat{\boldsymbol{\mu}} - \widehat{\widehat{\boldsymbol{\mu}}}||^2 = ||(P_{U_k} - P_{U_l})\boldsymbol{Y}||^2 = ((P_{U_k} - P_{U_l})\boldsymbol{Y})^T(P_{U_k} - P_{U_l})\boldsymbol{Y}) = \boldsymbol{Y}^T(P_{U_k} - P_{U_l})\boldsymbol{Y}$$

since $(P_{U_k} - P_{U_l})^2 = (P_{U_k} - P_{U_l})$ because it is a projection.

Now we solve problem 1.6 which in a) states that if $\boldsymbol{Y}$ is $N(\boldsymbol{0}, I_N)$ then

$$E(\boldsymbol{Y}^T A\boldsymbol{Y}) = trace(A) = \sum_{i=1}^{N} a_{ii}.$$

This is true since $(A\boldsymbol{Y})_i = \sum_{j=1}^{N} a_{ij}Y_j$ and therefore

$$\boldsymbol{Y}^T A\boldsymbol{Y} = \sum_{i=1}^{N} Y_i(A\boldsymbol{Y})_i = \sum_{i,j} = a_{ij}Y_iY_j$$

and therefore

$$E(\boldsymbol{Y}^T A\boldsymbol{Y}) = \sum_{i,j} a_{ij}E(Y_iY_j) = (E(Y_iY_j) = 0 \text{ if } i \neq j \text{ and 1 otherwise}) = \sum_{i=1}^{N} a_{ii} = trace(A).$$

If $\boldsymbol{Y}$ has the distribution $N(\boldsymbol{\mu}, \sigma^2 I_N)$ we get since $\boldsymbol{Y} = \boldsymbol{\mu} + \sigma\boldsymbol{Z}$ where $\boldsymbol{Z}$ is $N(\boldsymbol{0}, I_N)$

$$\boldsymbol{Y}^T A\boldsymbol{Y} = (\boldsymbol{\mu} + \sigma\boldsymbol{Z})^T A(\boldsymbol{\mu} + \sigma\boldsymbol{Z}) = \boldsymbol{\mu}^T A\boldsymbol{\mu} + \sigma\boldsymbol{\mu}^T A\boldsymbol{Z} + \sigma\boldsymbol{Z}^T A\boldsymbol{\mu} + \sigma^2\boldsymbol{Z}^T A\boldsymbol{Z}.$$

We therefore get

$$E(\boldsymbol{Y}^T A\boldsymbol{Y}) = \boldsymbol{\mu}^T A\boldsymbol{\mu} + 0 + 0 + \sigma^2 trace(A)$$

since $E(\boldsymbol{Z}) = \boldsymbol{0}$. Using this result for $A = P_{U_k} - P_{U_l}$ which has trace$(A) = k - l$ since it is a projection unto a subspace of dimension $k - l$ we get

$$E(||\widehat{\boldsymbol{\mu}} - \widehat{\widehat{\boldsymbol{\mu}}}||^2) = \sigma^2(k-l) + \boldsymbol{\mu}^T(P_{U_k} - P_{U_l})\boldsymbol{\mu} = \sigma^2(k-l) + \boldsymbol{\mu}^T(I_N - P_{U_l})\boldsymbol{\mu}.$$

This is due to the fact that $P_{U_k}\boldsymbol{\mu} = \boldsymbol{\mu}$ whether H_0 is true or not. If H_0 is true then $P_{U_l}\boldsymbol{\mu} = \boldsymbol{\mu}$ because then $\boldsymbol{\mu}$ lies in U_l. If H_0 is not true then we call $P_{U_l}\boldsymbol{\mu} = \boldsymbol{\mu}_0$ and we get

$$E(\widehat{\widehat{\sigma}}^2) = \sigma^2 + \frac{||\boldsymbol{\mu} - \boldsymbol{\mu}_0||^2}{k-l} = \sigma^2 + \frac{||E(\widehat{\boldsymbol{\mu}}) - E(\widehat{\widehat{\boldsymbol{\mu}}})||^2}{k-l}$$

. Whether or not H_0 is true we have

$$E(\widehat{\sigma}^2) = \frac{||\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}||^2}{N-k} = \sigma^2.$$

8 Simultaneous confidence intervals

We know that

$$||\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}||^2/\sigma^2 = ||A(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})||^2/\sigma^2 = \frac{(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^T A^T A(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})}{\sigma^2}$$

is $\chi^2(k)$ and

$$\frac{(N-k)\widehat{\sigma^2}}{\sigma^2} = \frac{||\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}||^2}{\sigma^2} = \frac{||\boldsymbol{Y} - A\widehat{\boldsymbol{\theta}}||^2}{\sigma^2}$$

is $\chi^2(N-k)$ and they are independent. Therefore

$$\frac{(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^T(A^T A)(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})}{k\widehat{\sigma^2}}$$

is $F(k, N-k)$-distributed. This means that

$$P(\frac{(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^T(A^T A)(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})}{k\widehat{\sigma^2}} \leq F_p(k, N-k)) = 1 - p$$

Setting $k\widehat{\sigma}^2 F_p(k, N-k) = b$ this means that the probability is $1-p$ that $\boldsymbol{\theta}$ is in the ellipsoid E (centered at $\widehat{\boldsymbol{\theta}}$) defined by

$$(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^T(A^T A)(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) \leq b$$

So if $\boldsymbol{\theta}$ is in E then for an arbitrary function $g(\boldsymbol{\theta})$ in a function class we see that

$$I = (\min_{\boldsymbol{\theta}\in E} g(\boldsymbol{\theta}), \max_{\boldsymbol{\theta}\in E} g(\boldsymbol{\theta}))$$

is confidence interval with confidence (at least) $1-p$ for all such g-functions. These are called simultaneous confidence intervals since they hold for all g-functions in the class of functions.

We want to study $g(\boldsymbol{\theta}) = c^T\boldsymbol{\theta}$ for all vectors c, i e. for all linear combinations of the parameters. E.g. we might be interested in the pairwise differences $\theta_i - \theta_j$. For notational convenience letting $\boldsymbol{\psi} = \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}$ and $B = A^TA$ we have

$$P(\boldsymbol{\psi}^T B\boldsymbol{\psi} \le k\widehat{\sigma}^2 F_p(k, N-k)) = P(\boldsymbol{\psi}^T B\boldsymbol{\psi} \le b) = 1-p.$$

$\boldsymbol{\psi}^T B\boldsymbol{\psi} \le b$ describes an ellisoid (centered at the origin). We want to get a confidence interval for $c^T\boldsymbol{\theta}$ so we want to find the extremal values for it when $\boldsymbol{\psi}$ belongs to the ellipsoid.

Having $B = A^TA$ we can construct a matrix $B^{1/2}$ and since we assume A^TA is invertible (we have a full rank model) then we have also $B^{-1/2}$.

We have $c^T\boldsymbol{\psi} = c^TB^{-1/2}B^{1/2}\boldsymbol{\psi}$ and using the Cauchy-Schwartz inequality $|x^Ty| \le ||x||\,||y||$, (i.e. $|\sum_i x_iy_i| \le \sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}$) we get

$$|c^T\boldsymbol{\psi}|^2 = |(B^{-1/2}c)^T(B^{1/2}\boldsymbol{\psi})|^2 \le ||B^{-1/2}c||^2||B^{1/2}\boldsymbol{\psi}||^2 = (c^TB^{-1}c)(\boldsymbol{\psi}^TB\boldsymbol{\psi}) \le b(c^TB^{-1}c)$$

and hence $|c^T\boldsymbol{\psi}| = |c^T\boldsymbol{\theta} - c^T\widehat{\boldsymbol{\theta}}| \le \sqrt{b(c^TB^{-1}c)}$ which means that

$$c^T\widehat{\boldsymbol{\theta}} \pm \sqrt{b(c^TB^{-1}c)} = c^T\widehat{\boldsymbol{\theta}} \pm \widehat{\sigma}\sqrt{kF_p(k, N-k)(c^TB^{-1}c)}$$

is a confidence interval for $c^T\boldsymbol{\theta}$ with confidence level $1-p$.

If c is such that $\sum_i^k c_i = 0$ we call $c^T\boldsymbol{\theta}$ a contrast. E.g. the pairwise differences $\theta_i - \theta_j$ are such contrasts, but of course they can be e.g. $\theta_1 - (\theta_2+\theta_3)/2$. In this case the linear forms $c^T\boldsymbol{\theta}$ are confined to a $k-1$-dimensional subspace of the k-dimensional subspace. So the contrasts can be expressed using the $k-1$ parameters $\lambda_i = \theta_i - \theta_k,\ i = 1, 2, \cdots, k-1$. The parameter estimate $\widehat{\boldsymbol{\lambda}}$ is $(k-1)$-dimensional so the dimension changes from k to $k-1$. We therefore have the confidence interval

$$c^T\widehat{\boldsymbol{\theta}} \pm \sqrt{b(c^TB^{-1}c)} = c^T\widehat{\boldsymbol{\theta}} \pm \widehat{\sigma}\sqrt{(k-1)F_p(k-1, N-k)(c^TB^{-1}c)}.$$

Notice that $N-k$ is not changed since the $\widehat{\sigma}^2$-estimate is still based on $N-k$ degrees of freedom.