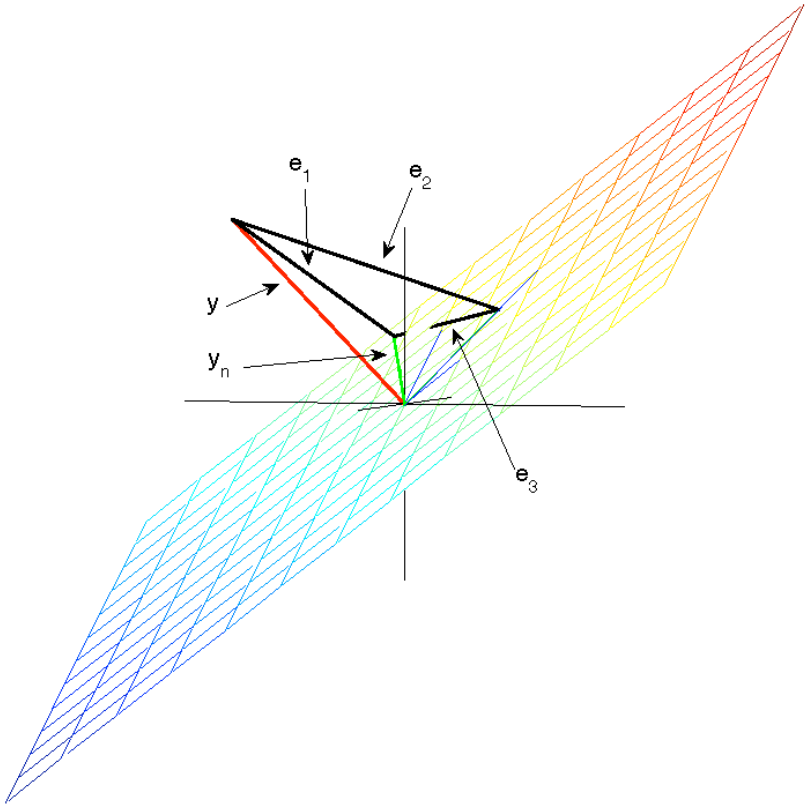
The General Linear Model in Functional MRI

Henrik BW Larsson

Functional Imaging Unit, Glostrup Hospital
University of Copenhagen

Part I



Preface

The General Linear Model (GLM) or multiple regression analysis is widely used in functional Magnetic Resonance Imaging (fMRI) in order to test whether one or more areas of the brain are involved in a certain task or not. Because many important results are based on the GLM, a thorough understanding of the math behind the GLM might be worthy.

Herein I try to provide the mathematical basis behind the GLM, which hopefully can ease the insight into the GLM. I use often a 3 dimensional (3D) perspective in order to visualize the fundamentals. The 3D perspective is easily generalized to any dimension, the math is the same. However, I should stress that this is not a book in math, and detailed proofs are therefore avoided, unless these are easily understandable and essential. It is my hope that this book fills out, what I think is a gab between an ordinary textbook in math, and many of the excellent books in fMRI.

In order to read this introduction, I assume that the reader has some acquaintance with the concept of vectors, vector addition, subtraction, multiplication, and also some knowledge about matrix operations. If this is not the case, I would recommend reading a basic book in linear algebra, before reading this.

I do not believe that this introduction is free from errors, or could not have been formulated in a more pedagogic way, so I will be happy for any kind of feedback and criticism.

Henrik BW Larsson
Functional Imaging Unit
Glostrup Hospital
Copenhagen University
Denmark
www.fiunit.dk

August 2011

Content

Part I

The General Linear Model used in fMRI

A model is a simplified representation of a given reality. Applying a model to a number of observations allows factors with a significant impact on the observations to be distinguished and quantified, because the model's validity can be tested.

The General Linear Model (GLM) is often used in fMRI, especially in BOLD imaging in order to make inferences about the significance of possible factors explaining the observed MR signal of the individual voxel during an fMRI experiment, e.g. the single run of an experiment.

The GLM approach treats data as a linear combination of model functions plus noise, i.e. the error term. The model functions are also called predictors or explanatory variables. The model functions are assumed to have known shapes and thus represent elements of our model, but their amplitudes are unknown and need to be estimated, that is we estimate the individual weight of all given explanatory variables. The error term is unpredictable and random, and we can expect this term to have a Gaussian distribution. In fact, one basic assumption for using regression analysis and GLM is that all variation of observed values are explained by the explanatory variables, and what is left, i.e. the error term, is Gaussian randomly distributed noise. This is equivalent with saying, that all observed values are independent, and all systematic variation of the observed values are due to our explanatory variables.

The evaluation of one voxel as a function of time, independent of adjacent voxels, is called univariate statistics. In part I, we will assume that our observed values of a single voxel, i.e. the time series of the MR signal of a voxel, are independent. In section II, we evaluate the situation if correlation between serial time points of a single voxel exists. Also, we will ignore the problem with multiple testing and the fact that adjacent voxels are correlated. However, both issues are important and are what really makes fMRI statistics different from ordinary statistics. In the following chapters, we will have a closer look at the math behind the univariate statistics and the corresponding equations and try to interpret them in an intuitive way.

The following equations are the mathematical representation of the general linear model:

$$
\begin{aligned}
y_1 &= \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \beta_4 x_{14} + \varepsilon_1 \\
y_2 &= \beta_1 x_{21} + \beta_2 x_{22} + \beta_3 x_{23} + \beta_4 x_{24} + \varepsilon_2 \\
y_3 &= \beta_1 x_{31} + \beta_2 x_{32} + \beta_3 x_{33} + \beta_4 x_{34} + \varepsilon_3 \qquad \text{Eq.1} \\
&\vdots \\
y_N &= \beta_1 x_{N1} + \beta_2 x_{N2} + \beta_3 x_{N3} + \beta_4 x_{N4} + \varepsilon_N
\end{aligned}
$$

Here, y is the observed MR signal of a voxel and is a function of discrete time points. In practice, y is indexed by the volume index (i.e. the frame number) as a subscript, running from 1 to N. Typically, the time spacing is equal to the Repetition Time (TR), which is around 0.5 s to 3 s. The x's are the explanatory variables. For example, the sequence x_{11}, x_{21}, x_{31}…. x_{N1} could represent the time course of resting and active blocks whereby the values would vary between zero and one. The sequence x_{11}, x_{21}, x_{31} ….x_{N1} constitute the first explanatory variable, where the first subscript is the discrete time point, and the second is the index of the number of the explanatory variable. The next explanatory variable is x_{12}, x_{22}, x_{32} ….x_{N2}. This

could represent some information about a translational movement in some direction. In the model above, there are 4 explanatory variables, and the model expression indicates that we believe that our observable variable y_i where the index i runs from 1 to N, is the result of a linear combination of the four explanatory variables, and where β_1 , β_2 , β_3 and β_4 express how much each explanatory variable contributes to the observed signal. The goal is to estimate these betas. The ε_1 to ε_N, is an error term that comprises what is left after we have described our signal with the four explanatory variables. Therefore, we expect that the elements of ε are randomly, i.e. independently, distributed, that is Gaussian distributed, which can be visualized graphically and which can be tested statistically.

Matrix and vector notation
Equation 1. can also be written in matrix notation, as:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ : \\ y_N \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \\ : & : & : & : \\ x_{N1} & x_{N2} & x_{N3} & x_{N4} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ : \\ \varepsilon_N \end{bmatrix} \qquad \text{Eq.2}$$

The *y*-column is called the *y*-vector, and the beta-column is called the beta-vector, and the *x*'s are together called the design matrix. Last we have the error-vector ε. Eq.1 and Eq.2 are exactly similar, and immediately show how to multiply the beta-vector with the x-matrix. Eq. 2 can also be written as:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ : \\ y_N \end{bmatrix} = \beta_1 \begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \\ : \\ x_{N1} \end{bmatrix} + \beta_2 \begin{bmatrix} x_{12} \\ x_{22} \\ x_{32} \\ : \\ x_{N2} \end{bmatrix} + \beta_3 \begin{bmatrix} x_{13} \\ x_{23} \\ x_{33} \\ : \\ x_{N3} \end{bmatrix} + \beta_4 \begin{bmatrix} x_{14} \\ x_{24} \\ x_{34} \\ : \\ x_{N4} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ : \\ \varepsilon_N \end{bmatrix} \qquad \text{Eq.2.2}$$

Eq. 2.2 really emphasizes that the observation of the vector ***y*** is modelled as the sum of 4 vectors weighted by the respective beta values (the element of the beta vector), plus a residual. Normally, a vector is written as ***v*** (bold) or $\vec{v}$. However, when appropriate, we will sometime use MATLAB notation for convenience so the ***y*** (column-) vector is written as *y*(:), the beta-vector as β(:). The explanatory variables are written as

$$\vec{x}_1 = \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{N1} \end{bmatrix} \quad \text{and} \quad \vec{x}_2 = \begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{N2} \end{bmatrix} \quad \text{etc.}$$

and these constitute the x-matrix written as *X*(:,:) or just *X* using capital letters for matrixes. So capital letters indicate a matrix, except M and N, which are reserved to indicate the total numbers of columns and the number of rows in a matrix. A matrix can also be written in turns of its vectors as $X = \left[\vec{x}_1 \quad \vec{x}_2 \quad \ldots \quad \vec{x}_N\right]$. Therefore, there are various ways we can write Eq.2:

$$\vec{y} = X\vec{\beta} + \vec{\varepsilon} \quad \text{or} \quad \vec{y}_N = X_{NM}\vec{\beta}_M + \vec{\varepsilon}_N \quad \text{or} \quad \boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \text{Eq.3}$$

Or, in MATLAB notation Eq.2.2 can be written as:

$$y(:) = \beta_1\ X(:,1) + \beta_2\ X(:,2) + \beta_3\ X(:,3) + \beta_4\ X(:,4) + \varepsilon(:)$$

or Eq.4

$$y(:) = X(:,:)\beta(:) + \varepsilon(:)$$

As stated above, this equation shows that we are explaining the ***y***-vector as a weighted sum of the four ***x*** column vectors, each of which is considered as an explanatory variable plus an error term. The weightings are the beta values. The individual beta value can be interpreted as the slope in a linear regression model and is by definition the change in *y* for a unit change in the corresponding *x*. Thus, if the *x* columns just have values between zero and one (as it often is in a simple fMRI study) then the individual beta is equivalent to the corresponding MR signal change when a x-value change from zero to one. However, in general, the beta-values are not necessarily equivalent to a change in the MR signal.

To give an example in 3D space, consider the following:

$$\begin{bmatrix}1\\4\\2\end{bmatrix} = \begin{bmatrix}1 & 0 & 1\\0 & 2 & 1\\1 & 1 & 1\end{bmatrix}\begin{bmatrix}\beta_1\\\beta_2\\\beta_3\end{bmatrix} = \beta_1\begin{bmatrix}1\\0\\1\end{bmatrix} + \beta_2\begin{bmatrix}0\\2\\1\end{bmatrix} + \beta_3\begin{bmatrix}1\\1\\1\end{bmatrix} \qquad \text{Eq.4}$$

We construct the vector $[1\ 4\ 2]^T$ from the sum of the three vectors: $[1\ 0\ 1]^T$, $[0\ 2\ 1]^T$, $[1\ 1\ 1]^T$, and the job is to find the three beta-values, which will satisfied Eq.4.
Note that the transpose of a vector is defined as:

$$\begin{bmatrix}1 & 4 & 2\end{bmatrix}^T = \begin{bmatrix}1\\4\\2\end{bmatrix}$$

and

$$\begin{bmatrix}1\\4\\2\end{bmatrix}^T = \begin{bmatrix}1 & 4 & 2\end{bmatrix}$$

A vector $\mathbf{a} = [a_1\ a_2\ a_3]$ is called a row vector, while a vector

$$\mathbf{b} = \begin{bmatrix}b_1\\b_2\\b_3\end{bmatrix}$$

is called a column vector.

Note that the three explanatory vectors are linearly independent, and the corresponding matrix has full rank (which is 3), meaning that one cannot construct, say $X(:,1)$ from the two others $\boldsymbol{x}$- vectors by multiplication with some scalar, addition or subtraction. That is to say that the following equation has no solution:

$$\begin{bmatrix}1\\0\\1\end{bmatrix} = v\begin{bmatrix}0\\2\\1\end{bmatrix} + w\begin{bmatrix}1\\1\\1\end{bmatrix} \text{ where } v \text{ and } w \text{ are scalars.}$$

The geometric situation is shown in Fig.1.

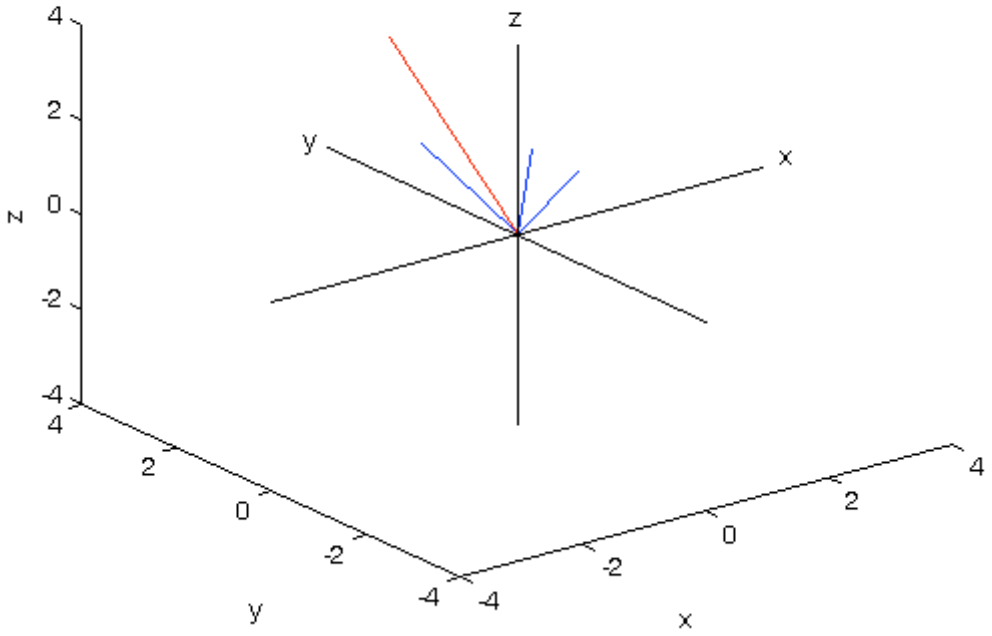


Figure 1. The first, second and third dimension is denoted x, y and z, and represent the first, second and third coordinate of any vector, and should not be confused with the $\boldsymbol{x}$-vector and the $\boldsymbol{y}$-vector in the text.

The red vector is $[1\ 4\ 2]^T$ and the blue vectors are the $\boldsymbol{x}$ column vectors. Try to identify each of the blue vectors.

As the X matrix has full rank we can easily find the beta values (using MATLAB) as:

$$\begin{bmatrix}\beta_1\\ \beta_2\\ \beta_3\end{bmatrix} = \begin{bmatrix}1 & 0 & 1\\ 0 & 2 & 1\\ 1 & 1 & 1\end{bmatrix}^{-1}\begin{bmatrix}1\\ 4\\ 2\end{bmatrix} = \begin{bmatrix}-1 & -1 & 2\\ -1 & 0 & 1\\ 2 & 1 & -2\end{bmatrix}\begin{bmatrix}1\\ 4\\ 2\end{bmatrix} = \begin{bmatrix}-1\\ 1\\ 2\end{bmatrix} \quad \text{Eq.5}$$

Therefore the red vector is a weighted sum of the three blue vectors, the weighting being the beta-values.

Eq. 4 also represents 'three equations with three unknowns' and if the three blue vectors are linearly independent, there will be exactly one solution. The 3 x 3 matrix $X(:,:)$ is said to be regular and its determinant is different from zero. This means that there exists an inverse matrix such that $X(:,:)\,X(:,:)^{-1} = X(:,:)^{-1}\,X(:,:) = I$, where I is the identity matrix, i.e.:

$$\begin{bmatrix}1 & 0 & 1\\ 0 & 2 & 1\\ 1 & 1 & 1\end{bmatrix}\begin{bmatrix}1 & 0 & 1\\ 0 & 2 & 1\\ 1 & 1 & 1\end{bmatrix}^{-1} = \begin{bmatrix}1 & 0 & 1\\ 0 & 2 & 1\\ 1 & 1 & 1\end{bmatrix}\begin{bmatrix}-1 & -1 & 2\\ -1 & 0 & 1\\ 2 & 1 & -2\end{bmatrix} = \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix} = I_{3,3}$$

Eq.4 can also be interpreted in another way. The beta vector is transformed by the X matrix to an 'image' : the ***y***-vector. So the beta-vector is the input, and the ***y***-vector is the output, and the X matrix is the transformer or operator, which operate on the beta-vector to produce the ***y***-vector.
In this situation, a point in the definition space (the space represented by the beta vector) has a corresponding point in the image space (the space represented by the ***y***-vector), and vice versa, and the matrix X, is associated with a transformation from definition space to image space. The inverse matrix X^{-1} is the transformation which brings a point in the image space back to the definition space. For this reason there is a one to one relation between image space and definition space. Obviously, if we first operate by using X and then X^{-1}, then we are back again and nothings happens, that is:

$$X^{-1}X\vec{\beta} = XX^{-1}\vec{\beta} = I\vec{\beta} = \vec{\beta}$$

Often, or always, in an fMRI run, the set of equations have many more rows than columns as we often have approximately 100 time points (number of volumes) but less than 10 explanatory variables. The situation can be exemplified as:

$$\begin{bmatrix}1\\ 4\\ 2\end{bmatrix} = \begin{bmatrix}1 & 1\\ 0 & 1\\ 1 & 1\end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix} = \beta_1\begin{bmatrix}1\\ 0\\ 1\end{bmatrix} + \beta_2\begin{bmatrix}1\\ 1\\ 1\end{bmatrix} \quad \text{Eq.6}$$

two columns but 3 rows.
The matrix equation is inconsistent, meaning that it has no solution. (A consistent system has one or more solutions). This is because one cannot find two beta-values in order to satisfy the three 'row' equations. The second row indicates that β_2 should be 4. Then the first row

indicates that β_1 should be -3, but this cannot satisfy the last row. This is very often the situation when we have e.g. 100 rows (observations). It is very unlikely that, say 7 explanatory variables, multiplied by their respective beta-values, will be able to produce the ***y***-vector exactly. It is the same as stating that the ***y***-vector $[1\ 4\ 2]^T$ is not in the (sub) space created by $X(:,1)$ and $X(:,2)$, or stated more explicit, the ***y***-vector is not in column space created by $X(:,1)$ and $X(:,2)$. The rank of the matrix in Eq.6 is 2. (Rank is defined precisely later). We also talk about the column rank and the row rank. The matrix in Eq.6 has full column rank, which is 2, because the two columns are linearly independent. However, the matrix does not have full row rank (which in this case is three), because the first and the last row of the design matrix are identical. The first two rows are linearly independent, so the row rank is 2. In general the row rank is always equal with the column rank, as just shown. However, a matrix can have full row rank (if the rank is equal with the numbers of rows), and/or full column rank (if the rank is equal with the number of columns).

If we want Eq.6 to be satisfied we have to add an error term:

$$\begin{bmatrix}1\\4\\2\end{bmatrix}=\begin{bmatrix}1 & 1\\0 & 1\\1 & 1\end{bmatrix}\begin{bmatrix}\beta_1\\\beta_2\end{bmatrix}+\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\varepsilon_3\end{bmatrix}=\beta_1\begin{bmatrix}1\\0\\1\end{bmatrix}+\beta_2\begin{bmatrix}1\\1\\1\end{bmatrix}+\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\varepsilon_3\end{bmatrix}\quad \text{Eq.6.2}$$

Now the equation is allowed some flexibility, and if the error term is chosen appropriately, the Eq.6.2 will be satisfied.
Obviously, the vector $\vec{\varepsilon}=\vec{y}-X\vec{\beta}$, must exist and differ from the zero vector. If this residual vector could be minimised, then we have found the 'best solution' in some sense. A measure of the closeness could be the length, also called the norm $\left|\vec{y}-X\vec{\beta}\right|$ or the squared distance $\left|\vec{y}-X\vec{\beta}\right|^2$ which is equal to $|\vec{\varepsilon}|$ and $|\vec{\varepsilon}|^2$, respectively. The situation is shown in Fig 2. The red vector is the $[1\ 4\ 2]^T$, and the two blue vectors are $[1\ 0\ 1]^T$ and $[1\ 1\ 1]^T$. The subspace created by the two blue vectors is a plane spanned by these two vectors. See Figure 3. This plane is created from every linear combination of $[1\ 0\ 1]^T$ and $[1\ 1\ 1]^T$, that is:

$$\vec{y}^{\,subspace}=v\begin{bmatrix}1\\0\\1\end{bmatrix}+w\begin{bmatrix}1\\1\\1\end{bmatrix}$$

where v and w can take any number.

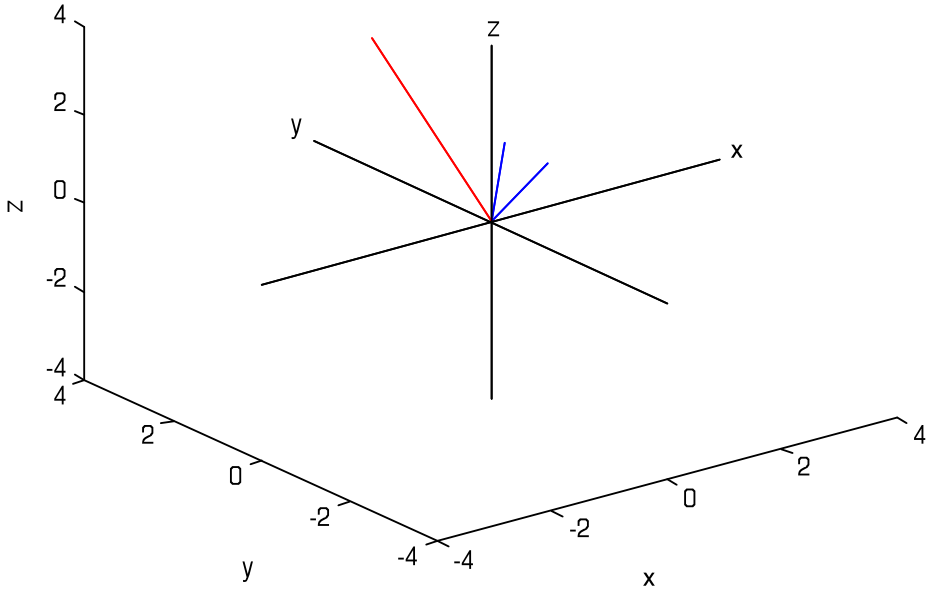


Figure 2

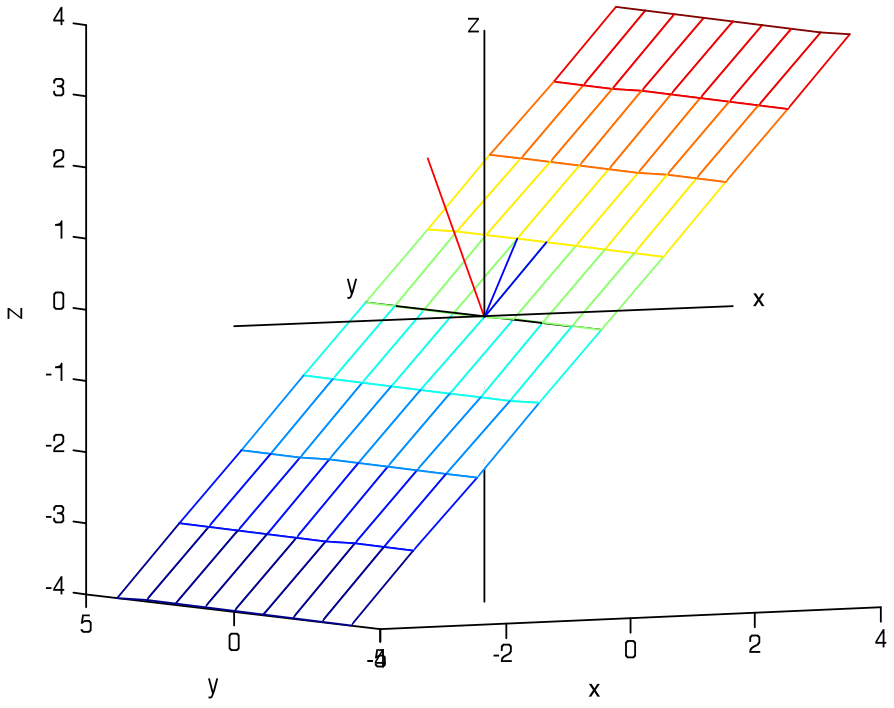


Figure 3.

We seek to find those betas, which will minimise the norm, i.e. the length, of the residual vector $\vec{\varepsilon}$. The solution must, obviously, exist in a subspace of 3D space. This subspace must be the plane which is defined by the two column vectors, i.e. the blue vectors in Fig. 3. By intuition, it can be seen that the point in this plane, which is as close as possible to the observed vector, i.e. the red one, can be found by an orthogonal projection of the red vector into the subspace created by $X(:,1)$ and $X(:,2)$ (i.e. $[1\ 0\ 1]^T$ and $[1\ 1\ 1]^T$, respectively). This will exactly minimize the length of the error vector. We will now study how such a projection can be done.

Projection of a vector into a subspace

Consider a 2 D space spanned by the two basis vectors:

$$\vec{q}_1 = \begin{bmatrix}1\\0\end{bmatrix}, \vec{q}_2 = \begin{bmatrix}0\\1\end{bmatrix}$$

These two vectors span the 2 D space because all points in this space can be constructed from a linear combination as:

$$v\begin{bmatrix}1\\0\end{bmatrix} + w\begin{bmatrix}0\\1\end{bmatrix}$$

where v and w can take any scalar value. The vectors are of unit length and are orthogonal. We say that $\vec{q}_1$ and $\vec{q}_2$ form an *orthonomal* basis for this 2D space. Another example of an orthonomal basis in 2D space is

$$\vec{q}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}, \vec{q}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix}-1\\1\end{bmatrix} \quad \text{where} \quad |\vec{q}_1| = 1 \quad \text{and} \quad |\vec{q}_2| = 1$$

The inner product (also called the dot product) of two vectors $\vec{a} = \begin{bmatrix}a_1 & a_2\end{bmatrix}^T$, $\vec{b} = \begin{bmatrix}b_1 & b_2\end{bmatrix}^T$ is defined as:

$$\vec{a}^T\vec{b} = \langle \vec{a},\vec{b} \rangle = \vec{a} \bullet \vec{b} = \begin{bmatrix}a_1 & a_2\end{bmatrix}\begin{bmatrix}b_1\\b_2\end{bmatrix} = a_1b_1 + a_2b_2 \quad \text{Eq.6.3}$$

This is clearly a scalar value.

Two (non-zero) vectors are orthogonal if and only if the so call inner product (the dot product) is zero:

$$\vec{a}^T\ \vec{b} = \langle \vec{a},\vec{b} \rangle = \vec{a} \bullet \vec{b} = 0 \quad \text{Eq. 6.4}$$

For example, given vectors $\vec{a} = \begin{bmatrix}1 & 0 & 0 & 0\end{bmatrix}^T$ and $\vec{b} = \begin{bmatrix}0 & 1 & 0 & 0\end{bmatrix}^T$ the inner product is:

$$\vec{a}^T\vec{b} = \begin{bmatrix}1 & 0 & 0 & 0\end{bmatrix}\begin{bmatrix}0\\1\\0\\0\end{bmatrix} = 0$$

It can be shown that Eq.6.3 is equivalent to the following expression:

$$\vec{a}^T\vec{b} = \cos(\alpha)\left|\vec{a}\right|\left|\vec{b}\right|$$

The geometric implication is shown in Fig.3b. The circle shown is the unit circle.

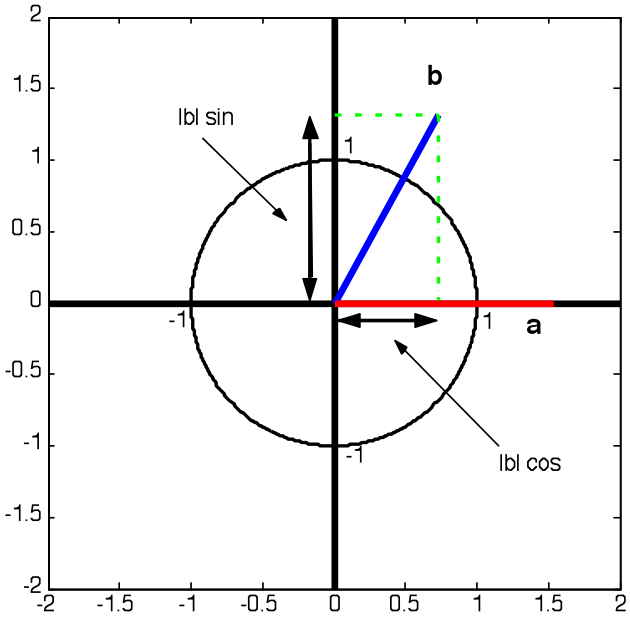


Figure 3b

Clearly, if both vectors are of unit length, then the product is just cosine to the angle between the vectors. If the **a**-vector (red) is of unit length, then the product is just the length of the projected **b**-vector projected orthogonally down on the **a**-vector, i.e. $|\mathbf{b}|\cos\alpha$. If both the **a**- and the **b**-vector are different from unit length, then the product is the length of the projection of the **b**-vector, where the **b**-vector is projected down on the **a**-vector multiplied with the length of the **a**-vector. Note, that this interpretation is consistent with the fact that the product of two orthogonal vectors is zero.

An N dimensional space can be divided into subspaces. Without going into too many details, consider a 3 D space spanned by the three basis vectors:

$$\vec{q}_1 = \begin{bmatrix}1\\0\\0\end{bmatrix},\quad \vec{q}_2 = \begin{bmatrix}0\\1\\0\end{bmatrix},\quad \vec{q}_3 = \begin{bmatrix}0\\0\\1\end{bmatrix}$$

These three vectors span the 3D space, because every point in the 3D space can be created by a linear combination of these three vectors. Now we consider the *subspace* spanned by $\vec{q}_1$ and $\vec{q}_2$. Clearly, the space spanned by these two vectors will be a two dimensional (2D) plane, created by a linear combination as:

$$v\begin{bmatrix}1\\0\\0\end{bmatrix}+w\begin{bmatrix}0\\1\\0\end{bmatrix}$$

where *v* and *w* can be any scalar value. Again, note that the column vectors of the design matrix in Eq.6, span a 2D planar subspace of the 3D space as:

$$v\begin{bmatrix}1\\0\\1\end{bmatrix}+w\begin{bmatrix}1\\1\\1\end{bmatrix}$$

where *v* and *w* can be any scalar value and that the ***y***-vector of Eq.6 will not be included in this subspace.

Suppose a vector ***y*** is N dimensional and we want to project this vector into a subspace with a dimension M where M < N. We can create an *orthonormal* basis for this subspace, the most simple being:

$$\vec{q}_1=\begin{bmatrix}1 & 0 & 0 & \cdots & 0\end{bmatrix}^T,\vec{q}_2=\begin{bmatrix}0 & 1 & 0 & \cdots & 0\end{bmatrix}^T,.....,\vec{q}_M=\begin{bmatrix}0 & 0 & \ldots & 1 & \ldots & 0\end{bmatrix}^T.$$

Note that these basis vectors are of unit length and orthogonal at each other.

Let us denote the projection of vector $\vec{y}$ into the subspace as vector $\vec{y}_n$, and the orthonormal basis for this subspace as the following vectors all of unit length and orthogonal at each other: $\vec{q}_1,\vec{q}_2,.....\vec{q}_M$ .Then the orthogonal projection is:

$$\vec{y}_n=\langle\vec{y},\vec{q}_1\rangle\vec{q}_1+\langle\vec{y},\vec{q}_2\rangle\vec{q}_2+...+\langle\vec{y},\vec{q}_M\rangle\vec{q}_M=\langle\vec{q}_1,\vec{y}\rangle\vec{q}_1+\langle\vec{q}_2,\vec{y}\rangle\vec{q}_2+...+\langle\vec{q}_M,\vec{y}\rangle\vec{q}_M=$$
$$\vec{q}_1\langle\vec{q}_1,\vec{y}\rangle+\vec{q}_2\langle\vec{q}_2,\vec{y}\rangle+...+\vec{q}_M\langle\vec{q}_M,\vec{y}\rangle=\vec{q}_1\vec{q}_1^T\vec{y}+\vec{q}_2\vec{q}_2^T\vec{y}+...+\vec{q}_M\vec{q}_M^T\vec{y}=$$
$$\left(\vec{q}_1\vec{q}_1^T+\vec{q}_2\vec{q}_2^T+...+\vec{q}_M\vec{q}_M^T\right)\vec{y}=$$
$$\begin{bmatrix}\vec{q}_1 & \vec{q}_2 & \cdots & \vec{q}_M\end{bmatrix}\begin{bmatrix}\vec{q}_1^T\\\vec{q}_2^T\\\vdots\\\vec{q}_M^T\end{bmatrix}\vec{y}=QQ^T\vec{y}$$

where

$$Q=[\vec{q}_1\quad\vec{q}_2\ldots\vec{q}_M]$$

Thus a projection of the vector into an orthonormal subspace results in a vector $\vec{y}_n$ and is:

$\vec{y}_n = QQ^T\vec{y}$ Eq.7

For completeness it should be noted that if the vector $\vec{y}$ is projected on the entire space, i.e. M=N, we get:

$$\vec{y} = (\vec{q}_1\vec{q}_1^T + \vec{q}_2\vec{q}_2^T + ...+ \vec{q}_N\vec{q}_N^T)\vec{y} = QQ^T\vec{y} \quad \Leftrightarrow \quad QQ^T = I$$

In Eq.6, we want to project the ***y***-vector into the subspace created by the two column vectors of the design matrix. We then first have to create an orthonormal subspace. The starting point is just to take one of the column vectors an normalize it to unit length:

$$\vec{q}_1 = \frac{\vec{x}_1}{|\vec{x}_1|} = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\0\\1\end{bmatrix}$$

An orthogonal unit vector is created as:

$$\vec{q}_2 = \frac{\vec{x}_2 - (\vec{x}_2 \bullet \vec{q}_1)\vec{q}_1}{|\vec{x}_2 - (\vec{x}_2 \bullet \vec{q}_1)\vec{q}_1|}$$

The nominator is:

$$\vec{x}_2 - (\vec{x}_2 \bullet \vec{q}_1)\vec{q}_1 = \begin{bmatrix}1\\1\\1\end{bmatrix} - \frac{2}{\sqrt{2}}\frac{1}{\sqrt{2}}\begin{bmatrix}1\\0\\1\end{bmatrix} = \begin{bmatrix}0\\1\\0\end{bmatrix}$$

the length is simply one, so:

$$\vec{q}_2 = \begin{bmatrix}0\\1\\0\end{bmatrix}$$

We have now constructed an orthonormal basis for the subspace spanned by the column vectors of the design matrix and can now perform a projection according to Eq.7, as:

$$\vec{y}_n = \begin{bmatrix}\vec{q}_1 & \vec{q}_2\end{bmatrix}\begin{bmatrix}\vec{q}_1^T\\ \vec{q}_2^T\end{bmatrix}\vec{y} = \begin{bmatrix}1/\sqrt{2} & 0\\ 0 & 1\\ 1/\sqrt{2} & 0\end{bmatrix}\begin{bmatrix}1/\sqrt{2} & 0 & 1/\sqrt{2}\\ 0 & 1 & 0\end{bmatrix}\vec{y} = \begin{bmatrix}1/2 & 0 & 1/2\\ 0 & 1 & 0\\ 1/2 & 0 & 1/2\end{bmatrix}\vec{y} = \begin{bmatrix}1.5\\4\\1.5\end{bmatrix}$$

However, if the dimension (N) is large, then it can be rather tedious to construct this orthonormal basis. In the following, we will show that we can bypass this step and utilize the design matrix directly.

We need the *QR factorization theorem*, which states that a matrix *X* (size : N x M), with full column rank (i.e. all columns are linearly independent i.e. the rank is equal to the number of columns) can be written as a product of an orthonormal matrix *Q* (same size as *X*) and a (non-

singular) upper triangular matrix R (size: M x M) such that $X = Q\,R$. An upper triangular matrix means that all entries below the diagonal are zero and all the entries of the diagonal are different from zero. In this situation, it is easy to calculate the determinant: it is just the product of the entries of the diagonal, and because all these are different form zero, then the determinant is also different form zero, i.e. the matrix has full rank and is non-singular and can be inverted.

As R is non-singular we may write:

$$Q = X R^{-1} \quad \text{Eq.8}$$

We also have:

$$X^T X = (QR)^T (QR) = R^T Q^T QR = R^T R \quad \text{Eq.9}$$

due to the fact that $Q^T Q = I$.

This can be shown by evaluating:

$$Q^T Q = \begin{bmatrix} \vec{q}_1^T \\ \vec{q}_2^T \\ \vdots \\ \vec{q}_M^T \end{bmatrix} \begin{bmatrix} \vec{q}_1 & \vec{q}_2 & .... & \vec{q}_M \end{bmatrix} = \begin{bmatrix} \vec{q}_1^T\vec{q}_1 & \vec{q}_1^T\vec{q}_2 & ... & \vec{q}_1^T\vec{q}_M \\ \vec{q}_2^T\vec{q}_1 & \vec{q}_2^T\vec{q}_2 & ... & \vec{q}_2^T\vec{q}_M \\ \vdots & \vdots & \ddots & \vdots \\ \vec{q}_M^T\vec{q}_1 & \vec{q}_M^T\vec{q}_2 & ... & \vec{q}_M^T\vec{q}_M \end{bmatrix} = \begin{bmatrix} 1 & 0 & ... & 0 \\ 0 & 1 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & 1 \end{bmatrix} = I$$

We can now calculate the projection of $\boldsymbol{y}$-vector into the subspace created by the two column vectors in Eq.6 :

$$\vec{y}_n = QQ^T\vec{y} = (XR^{-1})(XR^{-1})^T\vec{y} = XR^{-1}R^{-T}X^T\vec{y} = X(R^TR)^{-1}X^T\vec{y} = X(X^TX)^{-1}X^T\vec{y} \quad \text{Eq.10}$$

Here we have used the fact that for matrices:

$(A\,B)^T = B^T\,A^T$ and $(A\,B)^{-1} = B^{-1}\,A^{-1}$

Eq. 6 can be rewritten as $y(:) = X\beta(:)$ or $\boldsymbol{y} = X\boldsymbol{\beta}$. Remember that this system was inconsistent because X was rank deficient or singular. However, we immediately see that Eq.10 suggests an equation as $\boldsymbol{y}_n = X\boldsymbol{\beta}^{opt}$, where $\boldsymbol{\beta}^{opt} = (X^TX)^{-1}X^T\boldsymbol{y}$. Note that this optimal solution really exists even if X^{-1} does not exist!. Later we will show that if X has full column rank, then (X^TX) has full rank and has an inverse matrix $(X^TX)^{-1}$.

So to conclude, Eq.6 will make sense if we write:

$$\vec{y} = X\vec{\beta}^{opt} + \vec{e}$$

Because **e** has the minimum norm (shortest distance), it will be orthogonal to the space spanned by the column vectors of X.

Lets calculate the orthogonal projection into the subspace corresponding to Eq.6, using Eq.10:

$$\vec{y}_n = \begin{bmatrix}1 & 1\\ 0 & 1\\ 1 & 1\end{bmatrix}\left(\begin{bmatrix}1 & 0 & 1\\ 1 & 1 & 1\end{bmatrix}\begin{bmatrix}1 & 1\\ 0 & 1\\ 1 & 1\end{bmatrix}\right)^{-1}\begin{bmatrix}1 & 0 & 1\\ 1 & 1 & 1\end{bmatrix}\begin{bmatrix}1\\ 4\\ 2\end{bmatrix} = \begin{bmatrix}1 & 1\\ 0 & 1\\ 1 & 1\end{bmatrix}\begin{bmatrix}2 & 2\\ 2 & 3\end{bmatrix}^{-1}\begin{bmatrix}1 & 0 & 1\\ 1 & 1 & 1\end{bmatrix}\begin{bmatrix}1\\ 4\\ 2\end{bmatrix} =$$
$$\begin{bmatrix}1 & 1\\ 0 & 1\\ 1 & 1\end{bmatrix}\begin{bmatrix}1.5 & -1\\ -1 & 1\end{bmatrix}\begin{bmatrix}1 & 0 & 1\\ 1 & 1 & 1\end{bmatrix}\begin{bmatrix}1\\ 4\\ 2\end{bmatrix} = \begin{bmatrix}1 & 1\\ 0 & 1\\ 1 & 1\end{bmatrix}\begin{bmatrix}1.5 & -1\\ -1 & 1\end{bmatrix}\begin{bmatrix}3\\ 7\end{bmatrix} = \begin{bmatrix}0.5 & 0\\ -1 & 1\\ 0.5 & 0\end{bmatrix}\begin{bmatrix}3\\ 7\end{bmatrix} = \begin{bmatrix}1.5\\ 4\\ 1.5\end{bmatrix}$$

and the optimal beta values:

$$\vec{\beta}^{opt} = \begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix}^{opt} = \left(\begin{bmatrix}1 & 0 & 1\\ 1 & 1 & 1\end{bmatrix}\begin{bmatrix}1 & 1\\ 0 & 1\\ 1 & 1\end{bmatrix}\right)^{-1}\begin{bmatrix}1 & 0 & 1\\ 1 & 1 & 1\end{bmatrix}\begin{bmatrix}1\\ 4\\ 2\end{bmatrix} = \begin{bmatrix}1.5 & -1\\ -1 & 1\end{bmatrix}\begin{bmatrix}3\\ 7\end{bmatrix} = \begin{bmatrix}-2.5\\ 4\end{bmatrix}$$

To verify that the beta values are correct evaluate:

$$\vec{y}^n = \begin{bmatrix}1.5\\ 4\\ 1.5\end{bmatrix} = X\vec{\beta}^{opt} = \begin{bmatrix}1 & 1\\ 0 & 1\\ 1 & 1\end{bmatrix}\begin{bmatrix}-2.5\\ 4\end{bmatrix} = \begin{bmatrix}1.5\\ 4\\ 1.5\end{bmatrix}$$

The projection is shown in Figure 4.

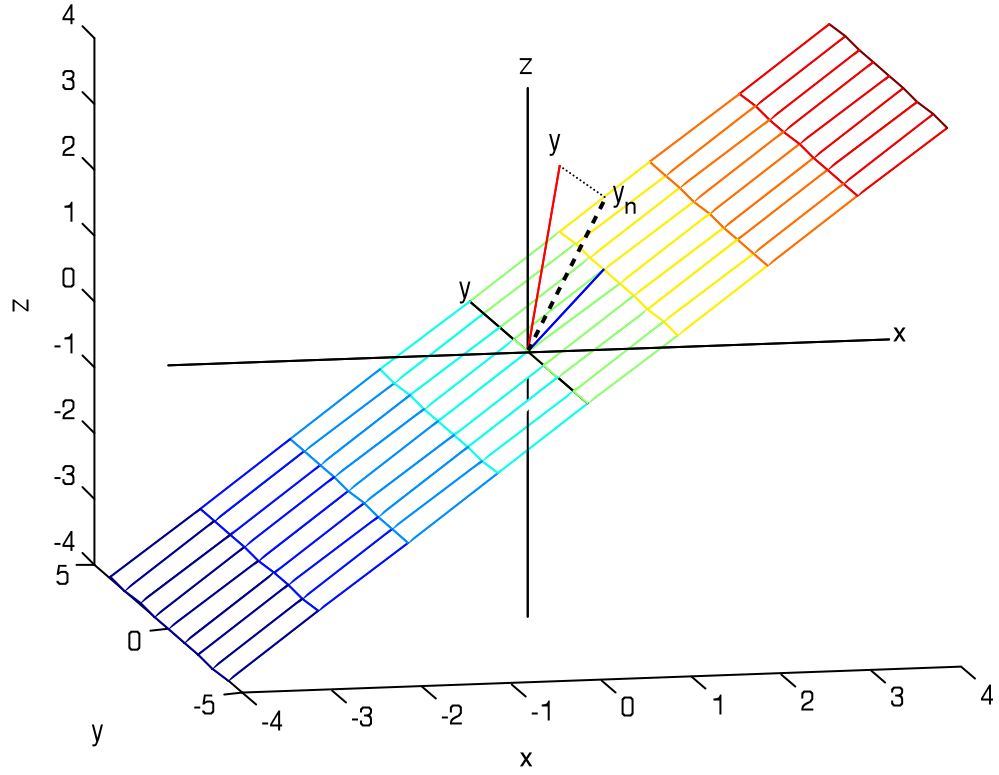


Figure 4

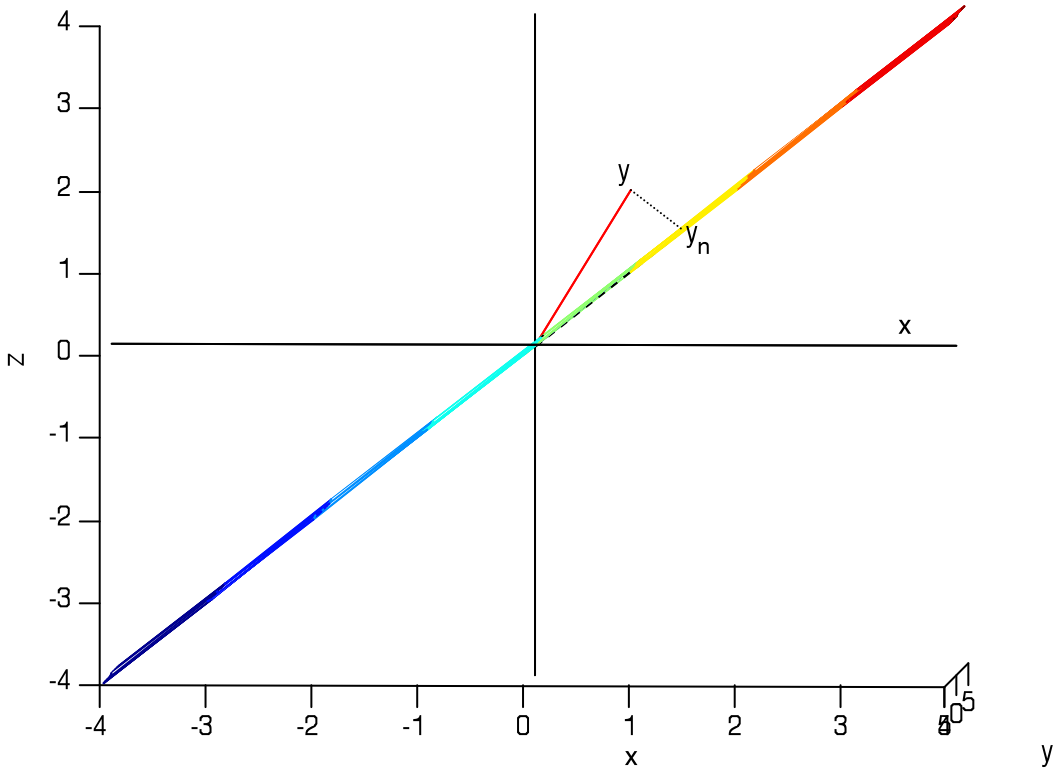


Figure 5

Figure 5 shows the same situation from a different perspective. It is clearly shown that the solution obtained really minimizes the distance between the ***y***-vector, i.e. the observation vector and the solution space.

Finding the betas

Now we have the first important result, which is a formula in order to find the optimal linear regression coefficients, the beta values:

$$\vec{\beta}^{opt} = (X^T X)^{-1} X^T \vec{y} \quad \text{Eq.11}$$

where X is the design matrix and $\vec{y}$ is the vector of observations.

One could also solve the problem of finding the beta-values (the element of the beta-vector) by using a fitting optimisation procedure, corresponding to Eq.2.2. Then a cost function has necessarily to be invoked. A commonly employed cost function is the sum of squares between observed values (y_i) and calculated values for different guesses of beta-values. The trick is then to minimise the sum of squares (SOSQ) in order to find the most optimal beta-values, i.e. minimise the following cost function:

$$\text{SOSQ}= \sum_{i=1}^{i=N} \left[y_i - (\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}) \right]^2 \quad \text{Eq.12}$$

However, it should be pointed out that this is exactly the squared distance between the observation vector ***y*** and the point in the solution space. And minimising this cost function is equivalent to finding the point in the solution space which minimises the distance. Thus, a fitting procedure will result in the same betas as provided by Eq.11.

Inference of beta's

The beta-vector, $\vec{\beta}^{opt} = \left[\beta_1 \quad \beta_2 \quad \cdots \quad \beta_M \right]^T$, is a multivariate stochastic variable, where each element (β_i) has a normal (Gaussian) distribution, centred around the 'true' value (β_i^{true}). We therefore say that the vector $\vec{\beta}^{opt}$ is normally distributed and write:

$$\vec{\beta}^{opt} = \text{Normal}(\vec{\beta}^{true}, \sigma^2 (X^T X)^{-1})$$

The variance is given by:

$$Var(\vec{\beta}^{opt}) = \sigma^2 (X^T X)^{-1}$$

The variance of an individual beta (β_i) is given as the i'th element of the diagonal of the matrix $(X^T X)^{-1}$ multiplied by the square of σ:

$$Var(\beta_i^{opt}) = \sigma^2 (X^T X)^{-1}_{(i,i)}$$

and the covariance between two different betas is given as:

$$Cov(\beta_i^{opt}, \beta_j^{opt}) = \sigma^2 (X^T X)^{-1}_{(i,j)}$$

If two different betas are uncorrelated, then the corresponding *of-diagonal* element is zero. However, the betas will often be correlated.

The residual variance σ^2 is simply calculated from the residual vector as:

$$\sigma^2 = \frac{e(:)^T e(:)}{N - rank(X)} \quad \text{Eq.13}$$

where N is the number of observations - e.g. the number of acquired volumes in a run.
In order to evaluate the significance of the individual betas, the student-t statistic can be used. With this method we can test whether a beta is significantly different from zero or a particular value *d*, which of course can be zero, by calculating a *t* value as:

$$t_{N-p} = \frac{\beta_i^{opt} - d}{\sqrt{\sigma^2 \left(X^T X\right)^{-1}_{(i,i)}}} \quad \text{Eq.14}$$

with N – rank(X) degrees of freedom.

Orthogonal projector

The matrix in the form of $A(A^TA)^{-1}A^T$ is sometimes called an orthogonal projector - see Eq.10 - and is here denoted as P. It has two characteristic properties mentioned below.
A symmetric matrix means symmetry with regard to the diagonal like:

$$\begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix}$$

and the transpose of a matrix means an interchange of rows and columns as:

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}^T = \begin{bmatrix} a & d & g \\ b & e & h \\ c & f & i \end{bmatrix} \quad \text{note that } X^{TT} = X$$

The two properties are:

1: $P^T = P$

indicating that P is a symmetric matrix; this is easily shown as:

$$\left[A(A^TA)^{-1}A^T\right]^T = A^{TT}(A^TA)^{-T}A^T = A(A^TA^{TT})^{-1}A^T = A(A^TA)^{-1}A^T$$

2: $P^2=P$

indicating that a vector already projected will not change when projected twice. That is if $\vec{a}_n = P\vec{a}$ is the projection of the a vector then an additional projection $\vec{a}_n = P\vec{a}_n$ will have no effect.

Also this feature can easily be shown as:

$$\left[A(A^TA)^{-1}A^T\right]^2 = \left[A(A^TA)^{-1}A^T\right]\left[A(A^TA)^{-1}A^T\right] = A(A^TA)^{-1}(A^TA)(A^TA)^{-1}A^T = A(A^TA)^{-1}A^T$$

Obviously, P must be singular, i.e. non-invertible, otherwise if P^{-1} exists then we would have:

$$P^2 = P \Leftrightarrow P^2P^{-1} = PP^{-1} = I \quad \wedge \quad P^2P^{-1} = P(PP^{-1}) = P \Leftrightarrow P = I$$

the identity matrix, which is uninteresting.
One last remark: from Fig.5 it can be seen that the observed vector $\vec{y}$ minus $\vec{y}_n$ is orthogonal to $\vec{y}_n$, that is:

$$\vec{y} - \vec{y}_n \perp \vec{y}_n \quad \text{and} \quad \vec{y} - \vec{y}_n = \vec{y} - P\vec{y} = (I - P)\vec{y}$$

Remember that $\vec{y} - \vec{y}_n$ is the residual, $\vec{e}$ vector, and it is seen that the matrix (*I-P*) is a projector that projects $\vec{y}$ into the residuals. Quite generally, the column spaces of *P* and *I-P* are mutually orthogonal vector spaces and the sum of the dimension of *P* and *I-P* is N. This means that every vector in the column space of *P* is orthogonal to every vector in the column space of *I-P*.
As an example, we could calculate the residual vector directly in the example corresponding to Fig.3. The P matrix is:

$$P = A(A^TA)^{-1}A^T = \begin{bmatrix}1 & 1\\ 0 & 1\\ 1 & 1\end{bmatrix}\left(\begin{bmatrix}1 & 0 & 1\\ 1 & 1 & 1\end{bmatrix}\begin{bmatrix}1 & 1\\ 0 & 1\\ 1 & 1\end{bmatrix}\right)^{-1}\begin{bmatrix}1 & 0 & 1\\ 1 & 1 & 1\end{bmatrix} = \begin{bmatrix}0.5 & 0 & 0.5\\ 0 & 1 & 0\\ 0.5 & 0 & 0.5\end{bmatrix}$$

It is immediately seen that the matrix is singular and symmetric. We can then find *I-P* as:

$$I - P = \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix} - \begin{bmatrix}0.5 & 0 & 0.5\\ 0 & 1 & 0\\ 0.5 & 0 & 0.5\end{bmatrix} = \begin{bmatrix}0.5 & 0 & -0.5\\ 0 & 0 & 0\\ -0.5 & 0 & 0.5\end{bmatrix}$$

You see that addition or subtraction of two matrices of the same size is performed separately for each entry. We can now calculate $\vec{y}_n$ as:

$$\vec{y}_n = P\vec{y} = \begin{bmatrix}0.5 & 0 & 0.5\\ 0 & 1 & 0\\ 0.5 & 0 & 0.5\end{bmatrix}\begin{bmatrix}1\\ 4\\ 2\end{bmatrix} = \begin{bmatrix}1.5\\ 4\\ 1.5\end{bmatrix}$$

as we previously have seen. The residual vector, or error vector, is calculated as:

$$\vec{e} = (I - P)\vec{y} = \begin{bmatrix} 0.5 & 0 & -0.5 \\ 0 & 0 & 0 \\ -0.5 & 0 & 0.5 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix} = \begin{bmatrix} -0.5 \\ 0 \\ 0.5 \end{bmatrix}$$

Note that $\vec{y} = \vec{y}_n + \vec{e}$, and that $\vec{y}_n$ and $\vec{e}$ are orthogonal vectors as predicted. Also note that the rank of *P* is 2 and the rank of *I-P* is 1, and that the sum is 3 as expected.

Orthogonalization of explanatory variables

When a paradigm is created, the columns of the design matrix are defined. It is an advantage, if possible, to create these explanatory vectors in such a way that these are mutually orthogonal, meaning that the corresponding (inner) vector product (dot product) is zero. Sometime this is not possible, just think of inclusion of vectors related to motion. It is easy to see that if two explanatory vectors are orthogonal, then the correlation between the corresponding regression coefficients (β_i and β_j) is zero. Recall that the covariance between two different betas is given as:

$$Cov(\beta_i^{opt}, \beta_j^{opt}) = \sigma^2 (X^T X)^{-1}_{(i,j)}$$

Thus, if the explanatory vector *X*(:,i) is orthogonal to explanatory variable *X*(:,j), then the entry element with index (i,j) of (X^TX) is zero, and also the corresponding entry of $(X^TX)^{-1}$ is zero. Therefor, orthogonalization of the columns of the design matrix secures that the corresponding regression coefficients are not correlated. If two explanatory vectors of the design matrix are not orthogonal, one can orthogonalize them before creating the design matrix by using Eq.7. In the following, we will show this using a very simple example where results can be derived directly by visual inspection.

Consider the following linear system:

$$\begin{bmatrix} 15 \\ 5 \\ 5 \end{bmatrix} = \beta_1 \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + \beta_2 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \vec{\varepsilon}$$

In principle, we could have many more rows, but the story would be the same. In order to show things in 3D space we choose three rows and two explanatory variables. We want to find the optimal solution with regards to the betas. We first note that the design matrix has full column rank, i.e. two, but can also see that the explanatory variables are not orthogonal. In order to solve the equation we first make an orthogonal projection of our observation vector ***y*** = $[15\ 5\ 5]^T$ into the space spanned the columns of the design matrix, using Eq.10:

$$\vec{y}_n = X(X^T X)^{-1} X^T \vec{y} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix} \left(\begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 15 \\ 5 \\ 5 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}^{-1} \begin{bmatrix} 20 \\ 15 \end{bmatrix} =$$

$$\begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 20 \\ 15 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 5 \\ 10 \end{bmatrix} = \begin{bmatrix} 15 \\ 5 \\ 0 \end{bmatrix}$$

The beta vector can then be calculated from:

$$\begin{bmatrix} 15 \\ 5 \\ 0 \end{bmatrix} = \beta_1 \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + \beta_2 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 10 \end{bmatrix}$$

We could also have used Eq.11 directly.

The error vector is simply:

$$\vec{\varepsilon} = \vec{y} - \vec{y}_n = \begin{bmatrix} 15 \\ 5 \\ 5 \end{bmatrix} - \begin{bmatrix} 15 \\ 5 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 5 \end{bmatrix}$$

The situation is shown in Fig.5.2

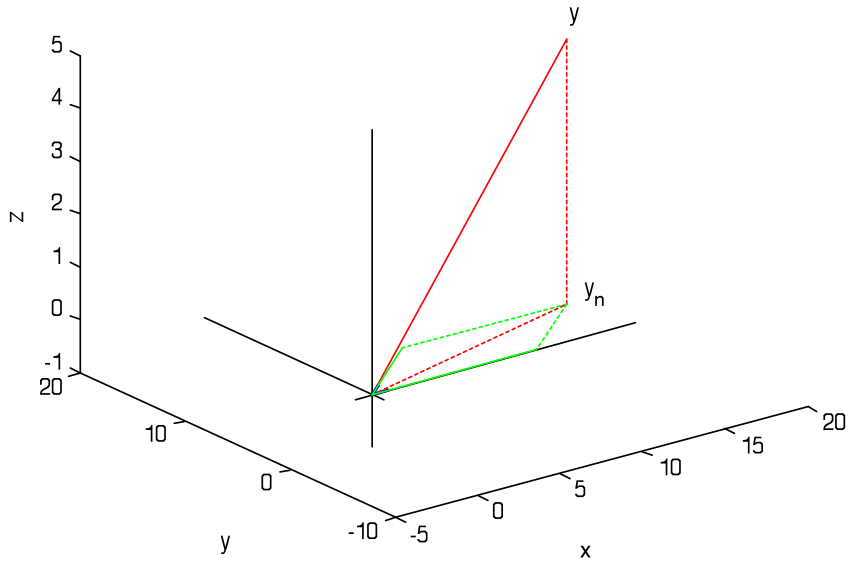


Figure 5.2. ***y*** and $\boldsymbol{y}_n$ is shown.

In Fig.5.3 the situation is seen from above.

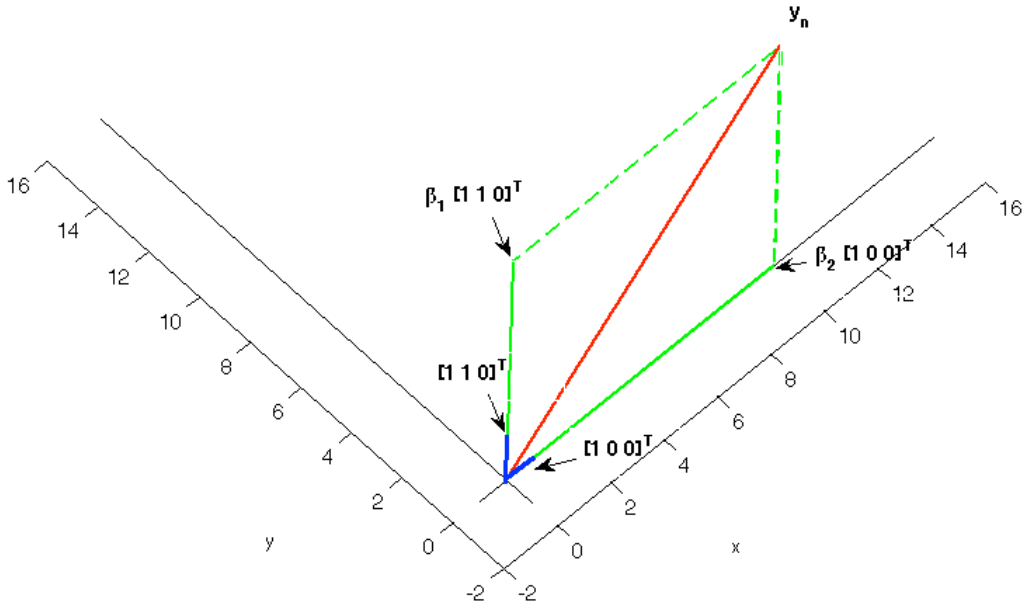


Figure 5.3. The explanatory vectors are indicated in blue. The interpretation of the value of the various betas, should be clear from the figure.

The variance is given by Eq.13:

$$\sigma^2 = \vec{\varepsilon}^T \vec{\varepsilon} / (N - rank(X)) = 25/(3-2) = 25$$

We have previously stated that the variance of the beta vector is given by:

$$Var(\vec{\beta}^{opt}) = \sigma^2 (X^T X)^{-1}$$

which should give:

$$Var(\vec{\beta}^{opt}) = 25 \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}$$

Clearly, the off-diagonal entries are not zero, indicating that β_1 and β_2 are correlated. This can in fact be seen from the figure because if the vector $\boldsymbol{y}$, and therefor $\boldsymbol{y}_n$, moves along and parallel with the 'y'-axis (second dimension) then β_1 increases as β_2 decreases, see Fig.5.4.

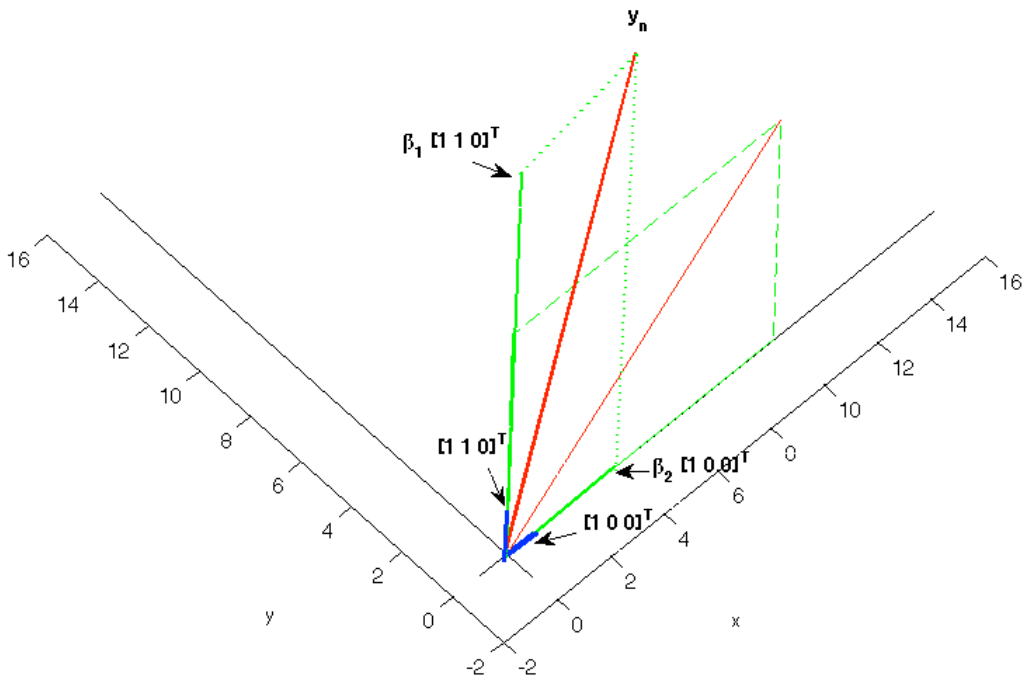


Figure 5.4. An example on covariance between beta-values when explanatory variables are not orthogonal.

Also note that the variance of β_2 is larger than β_1. This is because the norm (length) of the second explanatory vector is smaller than the length of the first explanatory vector: $|X(:,1)| = \sqrt{2}$, $|X(:,2)| = 1$. Generally, the larger the norm of an explanatory variable, the smaller the uncertainty, and an explanatory variable with a smaller norm relative to the others will be more undetermined. Therefore, it is advisable to construct explanatory variables with equal norm, if possible. The important point to make here is that the off-diagonal entries are not zero, indicating a non-zero covariance between β_1 and β_2. Now, what if we have orthogonalized the first column vector $\boldsymbol{x}_1$ to the second $\boldsymbol{x}_2$ in the design matrix X before calculation? So let us find a new $\vec{x}_1'$ which is orthogonal to $\vec{x}_2$. We first find the orthogonal projection of $\vec{x}_1$ to $\vec{x}_2$, using e.g. Eq.7:

$$\vec{x}_1^{\,n} = QQ^T\vec{x}_1 = \vec{x}_2\vec{x}_2^T\vec{x}_1 = \begin{bmatrix}1\\0\\0\end{bmatrix}\begin{bmatrix}1 & 0 & 0\end{bmatrix}\begin{bmatrix}1\\1\\0\end{bmatrix} = \begin{bmatrix}1 & 0 & 0\\0 & 0 & 0\\0 & 0 & 0\end{bmatrix}\begin{bmatrix}1\\1\\0\end{bmatrix} = \begin{bmatrix}1\\0\\0\end{bmatrix}$$

Then our new vector $\vec{x}_1^{\,'}$ is:

$$\vec{x}_1^{\,'} = \vec{x}_1 - \vec{x}_1^{\,n} = \begin{bmatrix}1\\1\\0\end{bmatrix} - \begin{bmatrix}1\\0\\0\end{bmatrix} = \begin{bmatrix}0\\1\\0\end{bmatrix}$$

a result which should not be a surprise. Now we have a new linear system:

$$\begin{bmatrix}15\\5\\5\end{bmatrix} = \beta_1\begin{bmatrix}0\\1\\0\end{bmatrix} + \beta_2\begin{bmatrix}1\\0\\0\end{bmatrix} + \begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\varepsilon_3\end{bmatrix} = \begin{bmatrix}0 & 1\\1 & 0\\0 & 0\end{bmatrix}\begin{bmatrix}\beta_1\\\beta_2\end{bmatrix} + \vec{\varepsilon}$$

which can be solved as before, or by using Eq.11. We see that the new design matrix not only has a full column rank (2), as before, but in addition has columns which are mutually orthogonal. The betas can be calculated from Eq.11:

$$\vec{\beta}^{\,opt} = \left(X^T X\right)^{-1} X^T\vec{y} = \left(\begin{bmatrix}0 & 1 & 0\\1 & 0 & 0\end{bmatrix}\begin{bmatrix}0 & 1\\1 & 0\\0 & 0\end{bmatrix}\right)^{-1}\begin{bmatrix}0 & 1 & 0\\1 & 0 & 0\end{bmatrix}\begin{bmatrix}15\\5\\5\end{bmatrix} = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}^{-1}\begin{bmatrix}5\\15\end{bmatrix} = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\begin{bmatrix}5\\15\end{bmatrix} = \begin{bmatrix}5\\15\end{bmatrix}$$

which should be compared with the previously found beta vector. The situation is shown in Fig.5.5.

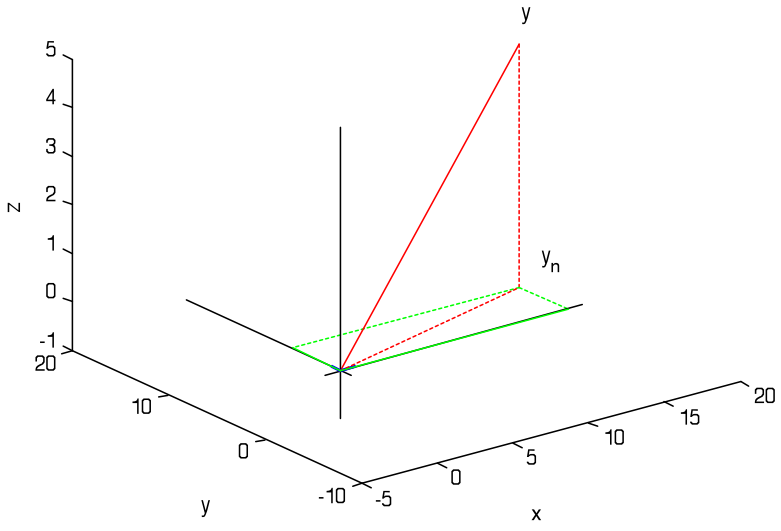


Figure 5.5. The situation after orthogonalization.

Fig.5.6 shows the situation from above.

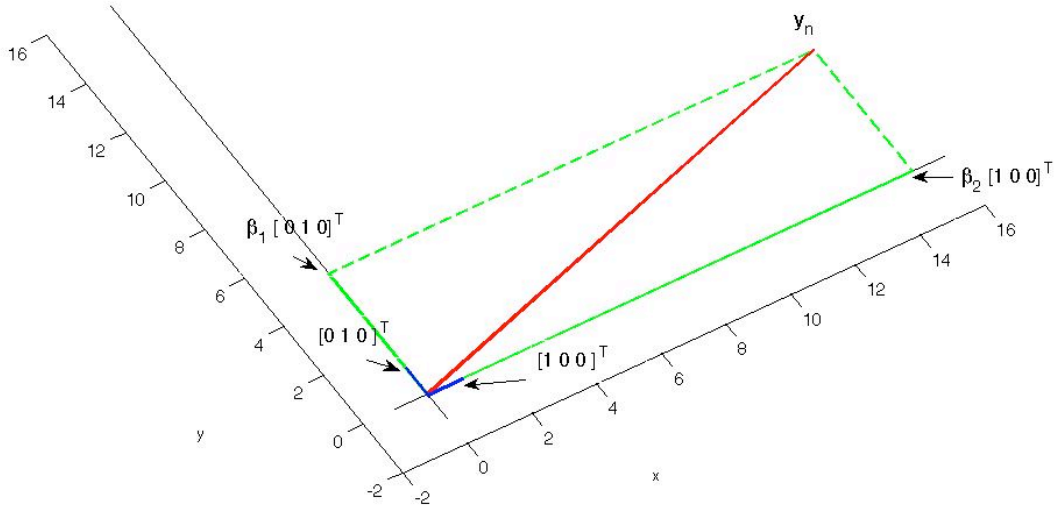


Figure 5.6. The explanatory vectors in blue. Note that if $\boldsymbol{y}_n$ moves along the y-axis (second dimension), the projection on the x-axis (first dimension) is unaffected, and if $\boldsymbol{y}_n$ moves along the x-axis the projection on the y-axis is also unaffected, thus β_1 and β_2 are not correlated.

The variance estimation of the beta vector is:

$$Var(\vec{\beta}^{opt}) = \sigma^2 (X^T X)^{-1} = 25\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Because the off-diagonal entries are zero, we see that the covariance between β_1 and β_2 is zero, and also that β_1 and β_2 have the equal variance which is related to the fact that the two explanatory vectors now have an equal norm (length).

Thus, orthogonalizing of an explanatory vector ***a*** to an explanatory vector ***b*** and then use the new vector ***a***' (which is orthogonal to vector ***b***) in the GLM, now estimates a new set of beta-values, where the new beta-value corresponding to ***b*** quantify how much vector ***b*** explain our

observed data, $\vec{y}$, independently of what can be explained by vector ***a***. The new beta-value corresponding to ***a***' quantifies the contribution of ***a*** over and above ***b***.

At this point we will stress that this form of correlation between the beta-values, govern by non-orthogonal explanatory vectors, should not be confused with correlation between our observed values, that is correlation between the elements of the ***y***-vector. This is a much more serious situation, which we will deal with in Part II.

An overparameterised system

In order to explain the observation $y(:)$ by a model, one constructs a design matrix, where each column, $X(:,i)$, i = 1:M, represents an explanatory variable. The total number of explanatory variables is M. However, one cannot just include as many explanatory variables as one might wishes. The reason is that the determination of the individual beta will not be unique, i.e. that more than one $\beta(:)$ vector will fit the general linear model. The system has become overparameterised. Again, in order to show the essence of this situation, a three-dimensional example will illustrate this point.
Consider the following matrix equation:

$$\begin{bmatrix}2\\4\\4\end{bmatrix} = \begin{bmatrix}1 & 0 & 2\\0 & 2 & 2\\1 & 1 & 3\end{bmatrix}\begin{bmatrix}\beta_1\\\beta_2\\\beta_3\end{bmatrix} = \beta_1\begin{bmatrix}1\\0\\1\end{bmatrix} + \beta_2\begin{bmatrix}0\\2\\1\end{bmatrix} + \beta_3\begin{bmatrix}2\\2\\3\end{bmatrix} \qquad \text{Eq.15}$$

We can determine that the column rank is 2 (because $2\vec{x}_1 + \vec{x}_2 = \vec{x}_3$), and also the row rank is of course 2 (because $X(1,:) + ½\, X(2,:) = X(3,:)$). This is equivalent with stating that the colums of X are linear dependent, meaning that it is possible to create one of the columns from the others by some multiplicative factor, subtraction or addition. The linear dependency among the columns of the design matrix can be tested formaly by an investigation of whether or not the following equation has a solution (which is not the trivial solution):

$$\begin{bmatrix}1 & 0 & 2\\0 & 2 & 2\\1 & 1 & 3\end{bmatrix}\begin{bmatrix}x\\y\\z\end{bmatrix} = \begin{bmatrix}0\\0\\0\end{bmatrix} \qquad \text{Eq.16}$$

Lets us call $(x\ y\ z)^T = \boldsymbol{v}$**.** The trivial solution is $\boldsymbol{v} = \mathbf{0}$, which is uninteresting, but of course represent a solution. Generally, from linear algebra we know that the homogeneous system $X\,\boldsymbol{v} = \mathbf{0}$ has nontrivial solutions if and only if the rank of X is less than the number of columns of X. Then the matrix X is called singular. The matrix X is singular if and only if its set of columns (or rows) is a linearly dependent set. Equation 16 is equivalent to the following equations:

$$x + 2z = 0$$
$$2y + 2z = 0$$
$$x + y + 3z = 0$$

Solving this set of equations gives $[x\ y\ z]^T = \mathbf{0}^T$ or $[x\ y\ z]^T = t\,[2\ 1\ -1]^T$, where t is a real number. If the matrix X had full rank we could just have inverted the matrix:

$$\vec{v} = X^{-1}\vec{0} = \vec{0}$$

and we would have only the trivial solution.

Now back to Eq.15. A linear system $\vec{y} = X\vec{\beta}$ is called consistent if it has one or more solutions. If it does not have a solution the system is said to be inconsistent. Furthermore, the linear system is consistent if and only if rank(X) = rank(X , $\boldsymbol{y}$), that is:

$$rank\begin{bmatrix}1 & 0 & 2\\ 0 & 2 & 2\\ 1 & 1 & 3\end{bmatrix} = rank\begin{bmatrix}1 & 0 & 2 & 2\\ 0 & 2 & 2 & 4\\ 1 & 1 & 3 & 4\end{bmatrix}$$

It can be seen that the rank of both matrices is 2. Obviously, this statement is equivalent to stating that a system $\vec{y} = X\vec{\beta}$ is consistent if and only if $\vec{y}$ is within the column space of X:

$$\vec{y} \in col(X)$$

Now suppose that the system $\vec{y} = X\vec{\beta}$ is consistent. Then $\vec{y} = X\vec{\beta}$ has a unique solution if and only if the corresponding homogeneous system, $X\vec{\beta} = \vec{0}$, only has the trivial solution. This is because X will then have full rank and X^{-1} exists and we could immediately calculate the unique solution: $\vec{\beta} = X^{-1}\vec{y}$.
From the above, we can conclude that Eq.15 is consistent, because $\boldsymbol{y}$ is in the column space of the design matrix but the design matrix is singular, and therefore there must exist more than one solution. In order to illustrate this, let's solve Eq.15:

$$\begin{bmatrix}\beta_1 + 2\beta_3 = 2\\ 2\beta_2 + 2\beta_3 = 4\\ \beta_1 + \beta_2 + 3\beta_3 = 4\end{bmatrix} \Leftrightarrow \begin{bmatrix}\beta_1 = 2 - 2t\\ \beta_2 = \frac{1}{2}(4 - 2\beta_3) = 2 - t\\ \beta_3 = t\end{bmatrix} \quad t \in \Re$$

Thus, for any real number t, solutions exist in the form of:

$$\begin{bmatrix}\beta_1\\ \beta_2\\ \beta_3\end{bmatrix} = \begin{bmatrix}2 - 2t\\ 2 - t\\ t\end{bmatrix}$$

Clearly, the system has many solutions. It is overparameterised, and one cannot make any relevant inference about the individual beta. In 3 D it is easy to show the geometrical implications of Eq.15.

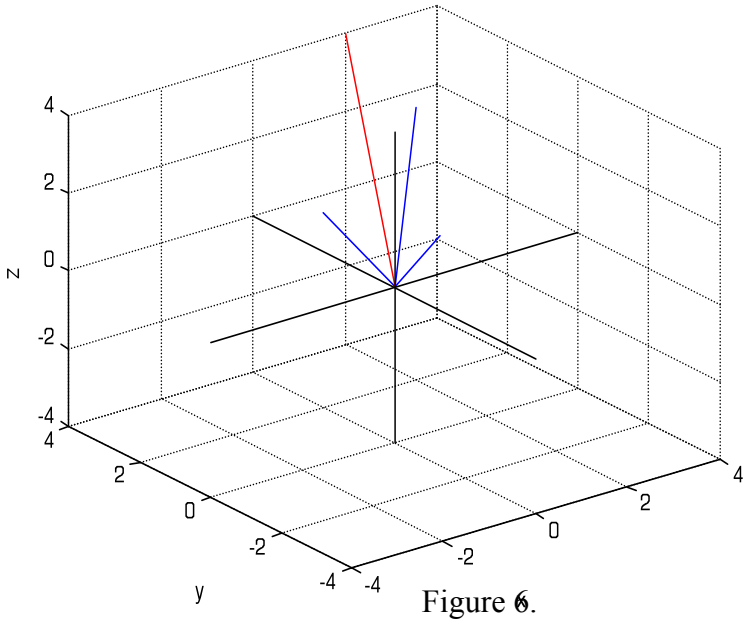


Figure 6.

All the vectors (the three blue corresponding to the three column vectors of the design matrix and the red ***y*** vector) are in the same plane. This is highlighted in the next Fig.7.

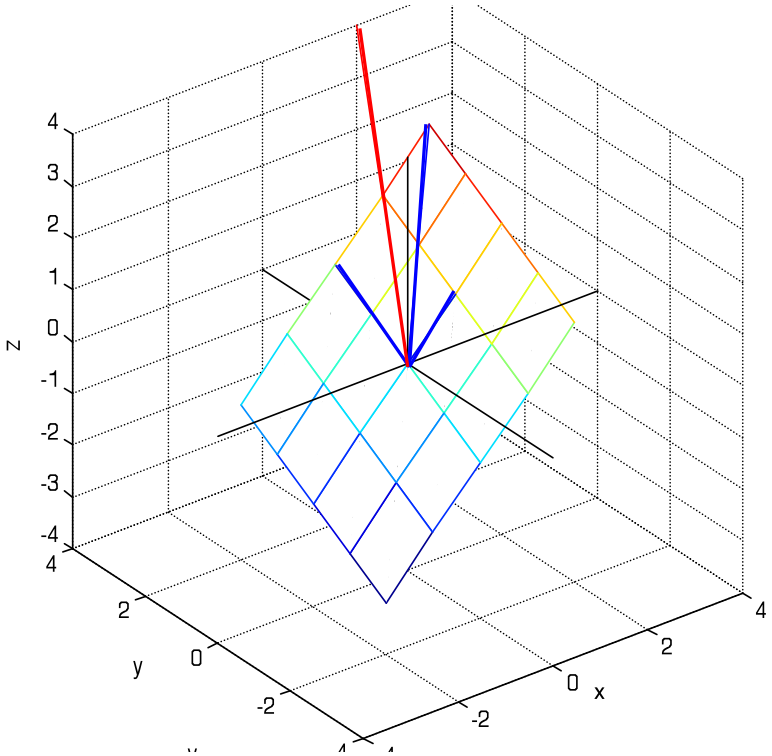


Figure 7

Looking at the last figure, it should be quite clear that $\vec{y}$ is within the column space of the design matrix, and furthermore, the $\vec{y}$ can be created by the three column vectors of the *X* matrix in infinitively many ways.

Dimension, Definition Space, Image Space, Rank and Pseudoinverse

In the most general situation we have a combination of an overparameterised system where the observed ***y***-vector is not within the column space spanned by the explanatory variables, i.e. is not in the column space created by the column vectors of the design matrix. The reason is that we often have 100 measures of the ***y***-vector, and perhaps 7 or more explanatory variables, which may be linear dependent to each other. In order to interpret the situation let's review the first situation, which was associated with Eq.4. However, first we will emphasise the definition of the *dimension* of a vector space: The dimension of a vector space is the *number of linearly independent vectors, which span that vector space*. Also, the definition of the *rank* of a matrix is important: The rank of the matrix *X* is equal to the *dimension of the image (vector) space*, corresponding to the space spanned by *X*, i.e. the space created by *X* multiplied by all possible $\boldsymbol{\beta}$-vectors (i.e. the entries of $\boldsymbol{\beta}$ running through possible values). The situation corresponding to Eq.4 is depicted in figure 8.

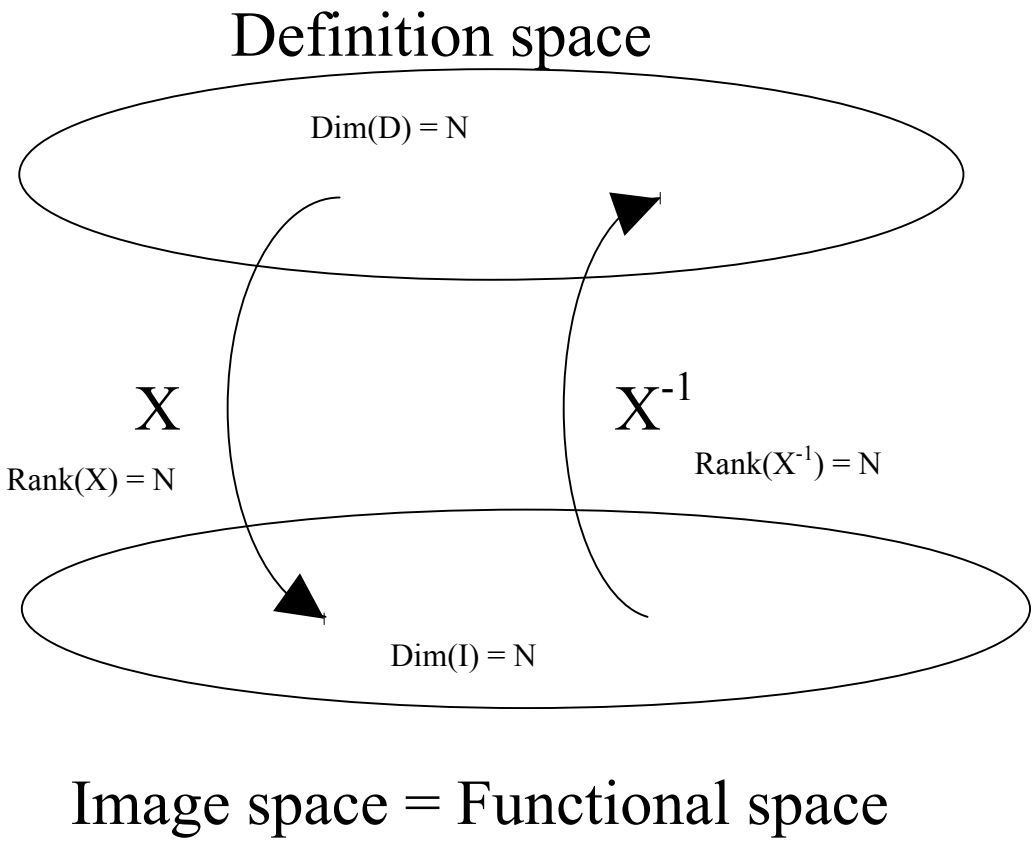


Figure 8

The definition space contains all possible beta-vectors. The image space contains all possible observations, i.e. the ***y***-vectors. Because all ***y***-vectors can be synthesized by a linear combination of the column vectors of the *X* matrix, the entire image space corresponds to the space spanned by these column vectors of the design matrix. In this situation, there is a one to one correspondence between a point in the definition space, i.e. the beta-vector and the ***y***-vector in the image space. If one selects a point in the image space, i.e. a *y*(:), then there exists only one beta-vector in the definition space. The design matrix *X* makes the transformation from the beta-vector (a point) in the definition space and transforms it into a point in the image space, i.e. $\vec{y} = X\vec{\beta}$. The matrix X^{-1} brings a point in the image space back to the original point in the definition space, i.e. $\vec{\beta} = X^{-1}\vec{y}$.

Now, let us look at the next situation we have encountered previously; this corresponds to the situation represented by Eq.6. The situation is depicted in Fig. 9:

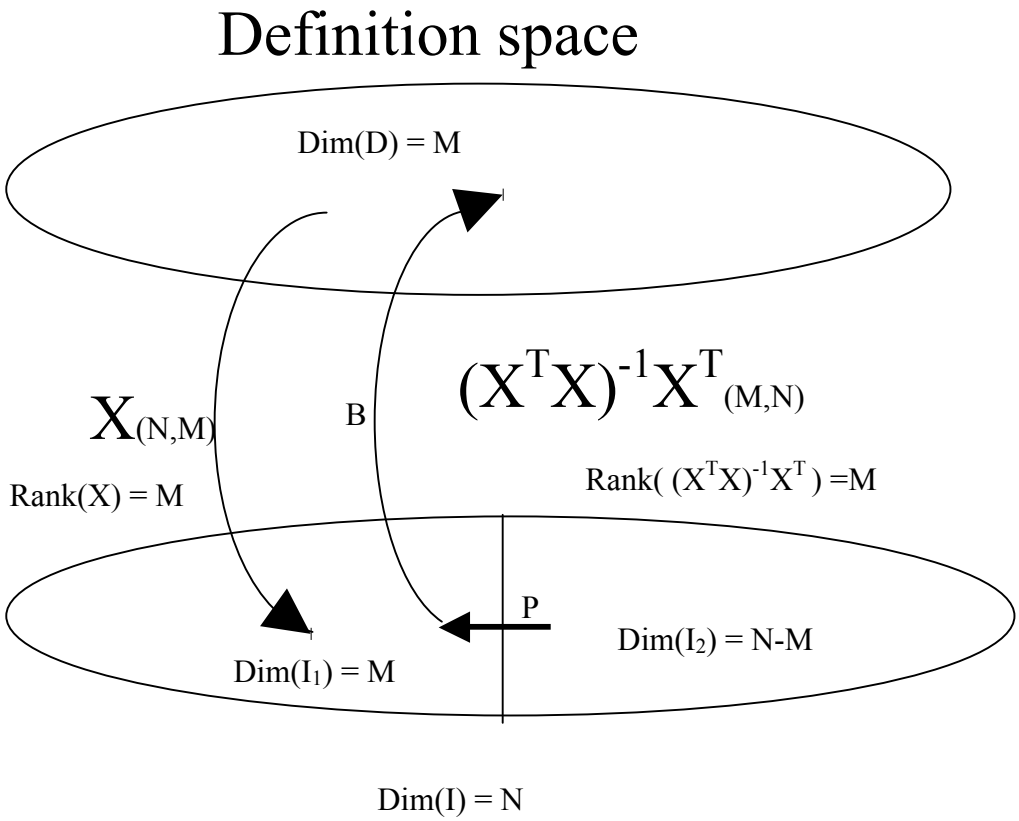


Figure 9

Eq.6 represented a situation where there was no solution, unless we included an error term. We saw that by making an orthogonal projection, we could obtain an optimal solution by minimizing the residual term or the error term.

The definition space contains the beta vectors. The functional space (I) is now the sum of two mutual orthogonal subspaces: $I = I_1 + I_2$. I_1 is the image space (see previous definition) and is the space spanned by the column vectors of the design matrix *X*, and this matrix brings a point in the definition space to the image space I_1. However, an observed point ($\vec{y}$) in the functional space is not likely to be within I_1, as an observation is always associated with noise. However, any vector in the functional space, under the assumption of the model, can be written as the sum of a vector in I_1 and a vector in I_2 as $\vec{y} = \vec{y}_n + \vec{y}_e$, where $\vec{y}_n$ and $\vec{y}_e$ are orthogonal at each other. $\vec{y}_e$ is, of course equal to the error vector, $\vec{e}$. For a given observation ($\vec{y}$) in I, we find the most optimal beta-vector in the definition space by the combined process depicted in Figure 9. It consist of an orthogonal projection along I_2, and then back to the definition space, i.e.:

$$B \oplus P = (X^T X)^{-1} X^T$$

where

$$P = X(X^T X)^{-1} X^T$$

The situation corresponding to an overparameterised system, exemplified by Eq.15, is depicted in Figure 10. Note that M>N. The definition space D is now the sum of two orthogonal subspaces, D_1 and D_2. The image of a point in the subspace D_2 is the **0** – vector (which has the dimension zero). The subspace D_2 is called the *null-space* and has the dimension r. In the example corresponding to Eq.15 the dimension of the null-space was one. Furthermore, any vector, $\vec{\beta}$, in the definition space can be resolved in two orthogonal vectors as $\vec{\beta} = \vec{\beta}_1 + \vec{\beta}_2$, where $\vec{\beta}_1$ belongs to D_1 and $\vec{\beta}_2$ belongs to D_2. The image of $\vec{\beta}$ is $\vec{y}$:

$$\vec{y} = X\vec{\beta} = X(\vec{\beta}_1 + \vec{\beta}_2) = X\vec{\beta}_1$$

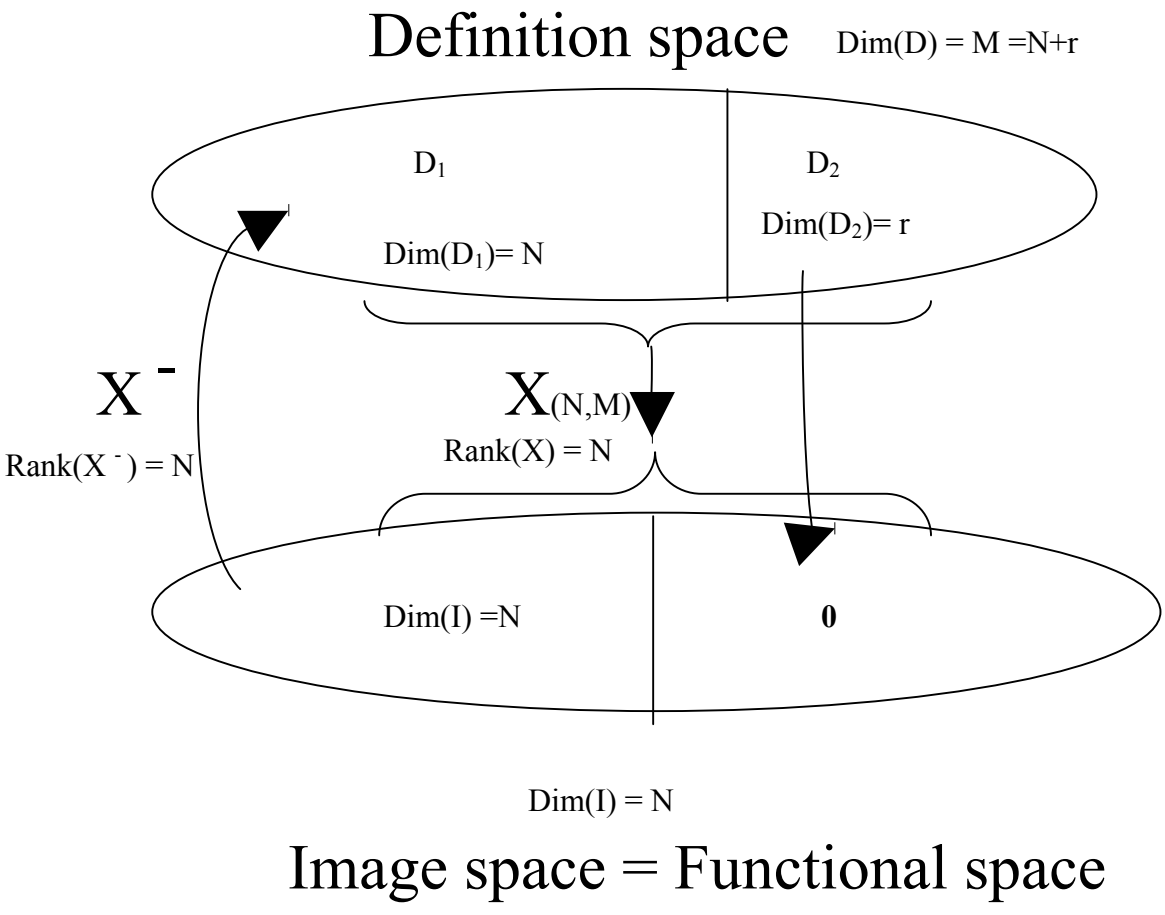


Figure 10

Note that the functional space and the image space are identical, and that the dimension of the zero-vector space is zero.

The pseudoinverse, also called the generalized inverse, is the process where a point in the image space is transformed back to the definition space, but restricted to D_1. The corresponding matrix is denoted X^-. Importantly, for a matrix without full column rank, as in Eq.15, the matrix X^TX is singular, and $(X^TX)^{-1}$ is not defined. This means that we cannot make a simple projection, P, in order to find a solution as we did before.

In the most general situation we have a combination of Figure 9 and 10. This is shown in Figure 11.

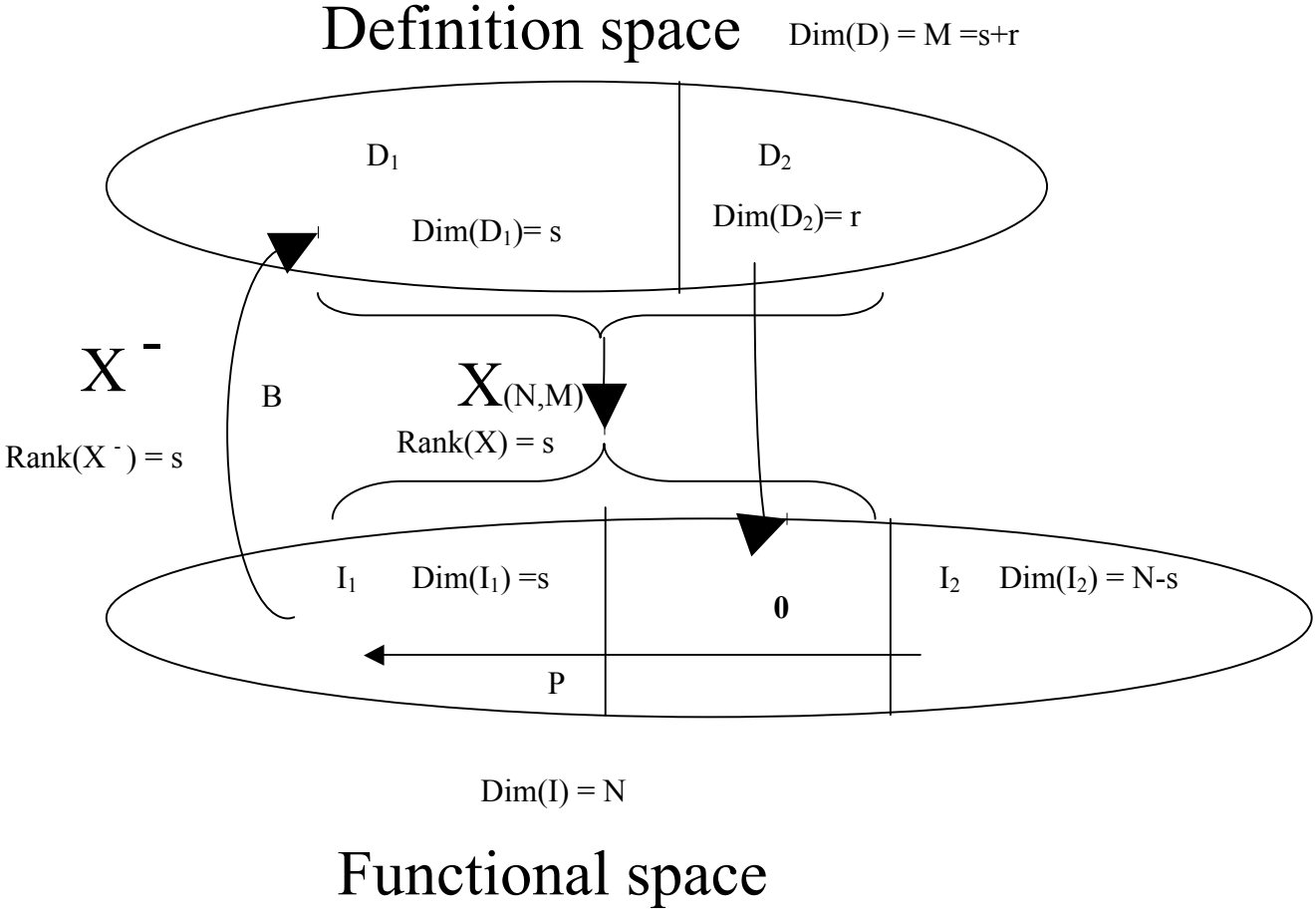


Figure 11

D_1 and D_2 are orthogonal subspaces of the definition space. I_1 and I_2 are orthogonal subspaces of the functional space, and I_1 (plus the zero-vector space) is the image space. D_2 is the null-space of D which is written as null(D). The rank of the X matrix is: rank(X) = s, and the dimension of the null space is: dim(null(D))= M-s = r. Furthermore, dim(I) ≥ dim(D), i.e. N ≥M which corresponds to the situation where we may have 100 observations and perhaps 10 explanatory variables.

A pseudoinverse is then constructed as a combined operation of an orthogonal projection along I_2 down on I_1, and then from I_1 to D_1. That is $X^- = B \oplus P$.

In order to give an example Eq.15 is modified to:

$$\begin{bmatrix} -4 \\ 4 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 2 & 2 \\ 1 & 1 & 3 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \beta_1 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} + \beta_2 \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix} + \beta_3 \begin{bmatrix} 2 \\ 2 \\ 3 \end{bmatrix} \quad \text{Eq.17}$$

which corresponds to the equation: $\boldsymbol{y} = X\boldsymbol{\beta}$. From previous inspection of X, we know rank(X) was 2, and thus the equation $X\boldsymbol{\beta} = \mathbf{0}$ has one or more solutions. We found that null(D) = $t\begin{bmatrix} 2 & 1 & -1 \end{bmatrix}$, where $t \in \Re$ (t belongs to all reel numbers), and clearly dim(null(D)) = 1. Additionally we find that:

$$rank\begin{bmatrix} 1 & 0 & 2 \\ 0 & 2 & 2 \\ 1 & 1 & 3 \end{bmatrix} \neq rank\begin{bmatrix} 1 & 0 & 2 & -4 \\ 0 & 2 & 2 & 4 \\ 1 & 1 & 3 & 4 \end{bmatrix}$$

because the rank of the last matrix is 3, and therefore the $\boldsymbol{y}$-vector is not in the space spanned by the three column vectors of X. That is to say that one cannot synthesize the $\boldsymbol{y}$-vector by any linear combination of the column vectors of the X matrix. The situation is depicted in Figure 12. Note that in this particular situation N=M.

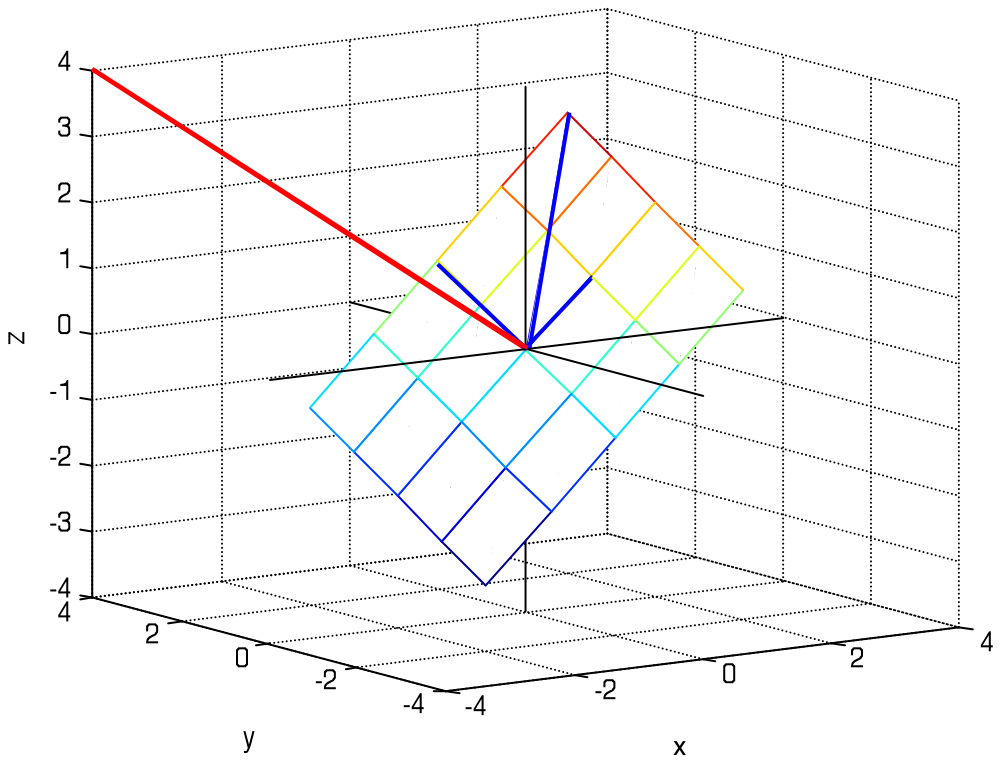


Figure12

Thus, the space D_2 is the space spanned by the vector:

$$\vec{d}_3' = [2 \quad 1 \quad -1]^T \text{ which belongs to the null space.}$$

Clearly, any of the row vectors of the matrix *X*, will be in the subspace D_1. Why ?, recall that $\vec{d}_3' = [2 \quad 1 \quad -1]^T$ represent the null-space and therefore the image of this vector is the zero-vector. Therefore, $\vec{d}_3' = [2 \quad 1 \quad -1]^T$ is orthogonal to any row vector of the *X* matrix. As a vector in D_1, lets choose

$$\vec{d}_2' = [1 \quad 0 \quad 2]^T$$

This arbitrary choice is part of the non-unique solution for the pseudo-inverse procedure. We then find a new vector, which is mutually orthogonal to these two vectors as:

$$\vec{d}_1' = \vec{d}_2' \times \vec{d}_3' = \begin{bmatrix} 2 & -5 & -1 \end{bmatrix}^T$$

where we have used that the vector cross product of two vectors results in a new vector orthogonal to the plan spanned by these two vectors. We could also have used the principle of orthogonal projection corresponding to Eq. 10. Recall that $P=A(A^TA)^{-1}A^T$ projects into the space spanned by the matrix *A*, and that (*I*-*P*) project into a orthogonal space to that spanned by *A*. Thus, we can also find a representation of the last dimension (denoted $\vec{d}_1''$)by:

$$\vec{d}_1'' = (I-P)\begin{bmatrix}1\\1\\1\end{bmatrix} = \left(\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix} - \begin{bmatrix}2&1\\1&0\\-1&2\end{bmatrix}\left(\begin{bmatrix}2&1&-1\\1&0&2\end{bmatrix}\begin{bmatrix}2&1\\1&0\\-1&2\end{bmatrix}\right)^{-1}\begin{bmatrix}2&1&-1\\1&0&2\end{bmatrix}\right)\begin{bmatrix}1\\1\\1\end{bmatrix} =$$

$$\left(\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix} - \begin{bmatrix}2&1\\1&0\\-1&2\end{bmatrix}\begin{pmatrix}6&0\\0&5\end{pmatrix}^{-1}\begin{bmatrix}2&1&-1\\1&0&2\end{bmatrix}\right)\begin{bmatrix}1\\1\\1\end{bmatrix} =$$

$$\left(\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix} - \begin{bmatrix}2&1\\1&0\\-1&2\end{bmatrix}\begin{pmatrix}1/6&0\\0&1/5\end{pmatrix}\begin{bmatrix}2&1&-1\\1&0&2\end{bmatrix}\right)\begin{bmatrix}1\\1\\1\end{bmatrix} =$$

$$\left(\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix} - \begin{bmatrix}13/15&1/3&1/15\\1/3&1/6&-1/6\\1/15&-1/6&29/30\end{bmatrix}\right)\begin{bmatrix}1\\1\\1\end{bmatrix} =$$

$$\begin{bmatrix}2/15&-1/3&-1/15\\-1/3&5/6&1/6\\-1/15&1/6&1/30\end{bmatrix}\begin{bmatrix}1\\1\\1\end{bmatrix} = \begin{bmatrix}-8/30\\20/30\\4/30\end{bmatrix}$$

Notice that $\vec{d}_1' = -\frac{30}{4}\vec{d}_1''$, so these vectors belong to the same subspace. Also note that the vector we choose for this projection is $[1\ 1\ 1]^T$, where all entries are different from zero.

Therefore, D_1 is spanned by $\vec{d}_1'$ and $\vec{d}_2'$ meaning that any linear combination of these two vectors will be within D_1. D_1 and D_2 are mutual orthogonal subspaces of D. Concerning the image space, I_1 is the space spanned by for example the first two column vectors of the *X* matrix:

$$\vec{i}_1' = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \quad \text{and} \quad \vec{i}_2' = \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}$$

These are linearly independent. We could also have chosen the two last column vectors, which are also linearly independent. Both set span the same vector space. Finally, I_2 is found as the space spanned by:

$$\vec{i}_3' = \vec{i}_1' \times \vec{i}_2' = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \times \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} -2 \\ -1 \\ 2 \end{bmatrix}$$

Again, it is seen that I_1 and I_2 are mutual orthogonal subspaces of the functional space I.

Now any vector beta in the definition space can be resolved as:

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} \vec{d}_1' & \vec{d}_2' & \vec{d}_3' \end{bmatrix} \begin{bmatrix} \beta_1' \\ \beta_2' \\ \beta_3' \end{bmatrix} = \begin{bmatrix} 2 & 1 & 2 \\ -5 & 0 & 1 \\ -1 & 2 & -1 \end{bmatrix} \begin{bmatrix} \beta_1' \\ \beta_2' \\ \beta_3' \end{bmatrix}$$

A shorter notation for this equation is denoted as:

$$\vec{\beta} = K\vec{\beta}' \quad \text{Eq.18}$$

We now have two representations of the definition space: D and D´ and we choose the base vectors for D as:

$$\text{D space}: \quad \vec{d}_1 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^T \quad \vec{d}_2 = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^T \quad \vec{d}_3 = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T$$

where $\vec{\beta}$ is the corresponding vector representation in this (conventional) orthonormal coordinate system.
The base vectors for the new coordinate system D are:

$$\text{D' space}: \quad \vec{d}_1' = \begin{bmatrix} 2 & -5 & -1 \end{bmatrix}^T \quad \vec{d}_2' = \begin{bmatrix} 1 & 0 & 2 \end{bmatrix}^T \quad \vec{d}_3' = \begin{bmatrix} 2 & 1 & -1 \end{bmatrix}^T$$

where $\vec{\beta}'$ is the corresponding vector representation in this coordinate system. The space D and D´are identical. A single space can be described by different representations. The two

representations D and D´ and the transition between them are given by Eq.18. We could also have normalized the base vectors in D´. D´ would then be an orthonormal base system. However, for the present purpose we will deliberately avoid this.

Likewise, any vector in the image space can be resolved as:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} \vec{i}_1' & \vec{i}_2' & \vec{i}_3' \end{bmatrix} \begin{bmatrix} y_1' \\ y_2' \\ y_3' \end{bmatrix} = \begin{bmatrix} 1 & 0 & -2 \\ 0 & 2 & -1 \\ 1 & 1 & 2 \end{bmatrix} \begin{bmatrix} y_1' \\ y_2' \\ y_3' \end{bmatrix}$$

For short this equation is denoted as:

$$\vec{y} = S\vec{y}' \quad \text{Eq.19}$$

We now have two representations of the image space: I and I´ , and the base vectors for I are:

$$\text{I space}: \quad \vec{i}_1 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^T \quad \vec{i}_2 = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^T \quad \vec{i}_3 = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T$$

where $\vec{y}$ is the corresponding vector representation in this (conventional) orthonormal coordinate system. The base vectors for the new coordinate system I´ are:

$$\text{I' space}: \quad \vec{i}_1' = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix}^T \quad \vec{i}_2' = \begin{bmatrix} 0 & 2 & 1 \end{bmatrix}^T \quad \vec{i}_3' = \begin{bmatrix} -2 & -1 & 2 \end{bmatrix}^T$$

where $\vec{y}'$ is the corresponding vector representation in this coordinate system. The space I and I´ are identical, but has two different representations. The transition between these two representations is given by Eq.19. We could also have normalized the base vectors and orthogonalized $\vec{i}_1'$ and $\vec{i}_2'$ in I´ , and I´ would then be an orthonormal base system. However, we avoid this for the following reason.

Now, the nice thing is that any vector in D´ can be written as:

$$\vec{\beta}' = \begin{bmatrix} \beta_1' \\ \beta_2' \\ \beta_3' \end{bmatrix} = \begin{bmatrix} \beta_1' \\ \beta_2' \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \beta_3' \end{bmatrix} = \vec{\beta}_{D_1}' + \vec{\beta}_{D_2}'$$

and we know that only $\vec{\beta}_{D_1}'$ will 'survive' a transformation from the definition space to the image space. Likewise, any vector in I´ can be written as:

$$\vec{y}' = \begin{bmatrix} y_1' \\ y_2' \\ y_3' \end{bmatrix} = \begin{bmatrix} y_1' \\ y_2' \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ y_3' \end{bmatrix} = \vec{y}_{I_1}' + \vec{y}_{I_2}'$$

where $\vec{y}_{I_1}'$ always represents the orthogonal projection of $\vec{y}'$ (if not already in I_1).

Obviously, there should be a one to one relationship between $\vec{\beta}'_{D_1}$ and $\vec{y}'_{I_1}$ which can be found by:

$$\vec{y} = X\vec{\beta} \quad \wedge \quad \vec{\beta} = K\vec{\beta}' \quad \wedge \quad \vec{y} = S\vec{y}' \Rightarrow \quad \vec{y}' = S^{-1}XK\vec{\beta}'$$

The matrix *S* has full rank so:

$$S^{-1} = \begin{bmatrix} \frac{5}{9} & \frac{-2}{9} & \frac{4}{9} \\ \frac{-1}{9} & \frac{4}{9} & \frac{1}{9} \\ \frac{-2}{9} & \frac{-1}{9} & \frac{2}{9} \end{bmatrix} \quad \text{and} \quad S^{-1}XK = \begin{bmatrix} 0 & 5 & 0 \\ -6 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Therefore

$$\begin{bmatrix} y'_1 \\ y'_2 \\ y'_3 \end{bmatrix} = \begin{bmatrix} 0 & 5 & 0 \\ -6 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta'_1 \\ \beta'_2 \\ \beta'_3 \end{bmatrix} \quad \text{or for short} \quad \vec{y}' = X'\vec{\beta}' \qquad \text{Eq.20}$$

If we just look at the relationship between the space D_1' and I_1' it can be specified as:

$$\begin{bmatrix} y'_1 \\ y'_2 \end{bmatrix} = \begin{bmatrix} 0 & 5 \\ -6 & 2 \end{bmatrix} \begin{bmatrix} \beta'_1 \\ \beta'_2 \end{bmatrix}$$

An inverse is then easily found as:

$$\begin{bmatrix} \beta'_1 \\ \beta'_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{15} & \frac{-1}{6} \\ \frac{1}{5} & 0 \end{bmatrix} \begin{bmatrix} y'_1 \\ y'_2 \end{bmatrix}$$

We can include the entire I′ and D′ space and get:

$$\begin{bmatrix} \beta'_1 \\ \beta'_2 \\ \beta'_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{15} & \frac{-1}{6} & 0 \\ \frac{1}{5} & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} y'_1 \\ y'_2 \\ y'_3 \end{bmatrix} \quad \text{or for short} \quad \vec{\beta}' = X'^{-}\vec{y}' \qquad \text{Eq.21}$$

X'^{-} is the pseudoinverse relating to the two prime coordinate systems. However, we wish to express the pseudoinverse in the two conventional coordinate systems. This can be done as:

$$\vec{\beta}' = X'^{-}\vec{y}' \quad \wedge \quad \vec{\beta} = K\vec{\beta}' \quad \wedge \quad \vec{y}' = S^{-1}\vec{y} \quad \Rightarrow \quad \vec{\beta} = KX'^{-}S^{-1}\vec{y} \quad \text{Eq.22}$$

Thus a pseudoinverse relating I to D is then:

$$X^{-} = KX'^{-}S^{-1} = \begin{bmatrix} 2 & 1 & 2 \\ -5 & 0 & 1 \\ -1 & 2 & -1 \end{bmatrix}\begin{bmatrix} 1/15 & -1/6 & 0 \\ 1/5 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} 5/9 & -2/9 & 4/9 \\ -1/9 & 4/9 & 1/9 \\ -2/9 & -1/9 & 2/9 \end{bmatrix} = \begin{bmatrix} 2/9 & -2/9 & 1/9 \\ -5/18 & 4/9 & -1/18 \\ 1/6 & 0 & 1/6 \end{bmatrix}$$

This is the final pseudoinverse. Therefore we have:

$$\vec{\beta}^{opt} = X^{-}\vec{y} \quad \Rightarrow \quad \vec{y}^{pred} = X\vec{\beta}^{opt} = XX^{-}\vec{y}$$

The procedure described above results in a solution with minimum norm, that is:

$\left\|\vec{y} - X\vec{\beta}^{opt}\right\| = \left\|\vec{y} - XX^{-}\vec{y}\right\|$ is minimum. This is equivalent to the fact that $P = XX^{-}$ is an orthogonal projector into the space spanned by X. If X has full column rank then $X^{-} = (X^{T}X)^{-1}X^{T}$ and the orthogonal projector is given as $P = X(X^{T}X)^{-1}X^{T}$ as shown previously.

For the pseudoinverse with minimum norm the following relations holds:

$$XX^{-}X = X, \quad X^{-}XX^{-} = X^{-}, \quad \left(X^{-}X\right)^{H} = X^{-}X, \quad \left(XX^{-}\right)^{H} = XX^{-} \quad \text{Eq.23}$$

(We will come back to the superscript H later, and also the importance of these relations)
As an example we get:

$$XX^{-}X = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 2 & 2 \\ 1 & 1 & 3 \end{bmatrix}\begin{bmatrix} 2/9 & -2/9 & 1/9 \\ -5/18 & 4/9 & -1/18 \\ 1/6 & 0 & 1/6 \end{bmatrix}\begin{bmatrix} 1 & 0 & 2 \\ 0 & 2 & 2 \\ 1 & 1 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 2 & 2 \\ 1 & 1 & 3 \end{bmatrix}$$

Furthermore, a solution corresponding to Eq.17 is then given by:

$$\vec{\beta}^{opt} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 2/9 & -2/9 & 1/9 \\ -5/18 & 4/9 & -1/18 \\ 1/6 & 0 & 1/6 \end{bmatrix}\begin{bmatrix} -4 \\ 4 \\ 4 \end{bmatrix} = \begin{bmatrix} -4/3 \\ 8/3 \\ 0 \end{bmatrix}$$

and the predicted image is:

$$\vec{y}^{pred} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 2 & 2 \\ 1 & 1 & 3 \end{bmatrix}\begin{bmatrix} -4/3 \\ 8/3 \\ 0 \end{bmatrix} = \begin{bmatrix} -4/3 \\ 16/3 \\ 4/3 \end{bmatrix} = \frac{2}{3}\begin{bmatrix} -2 \\ 8 \\ 2 \end{bmatrix}$$

It can be shown that this predicted vector is in the space spanned by the column vectors of the matrix X. If the last column is multiplied by –1 and added to 5 times the second column, then one gets $\vec{y}^{\,pred}$. $\vec{y}^{\,pred}$ is shown on Figure 13.

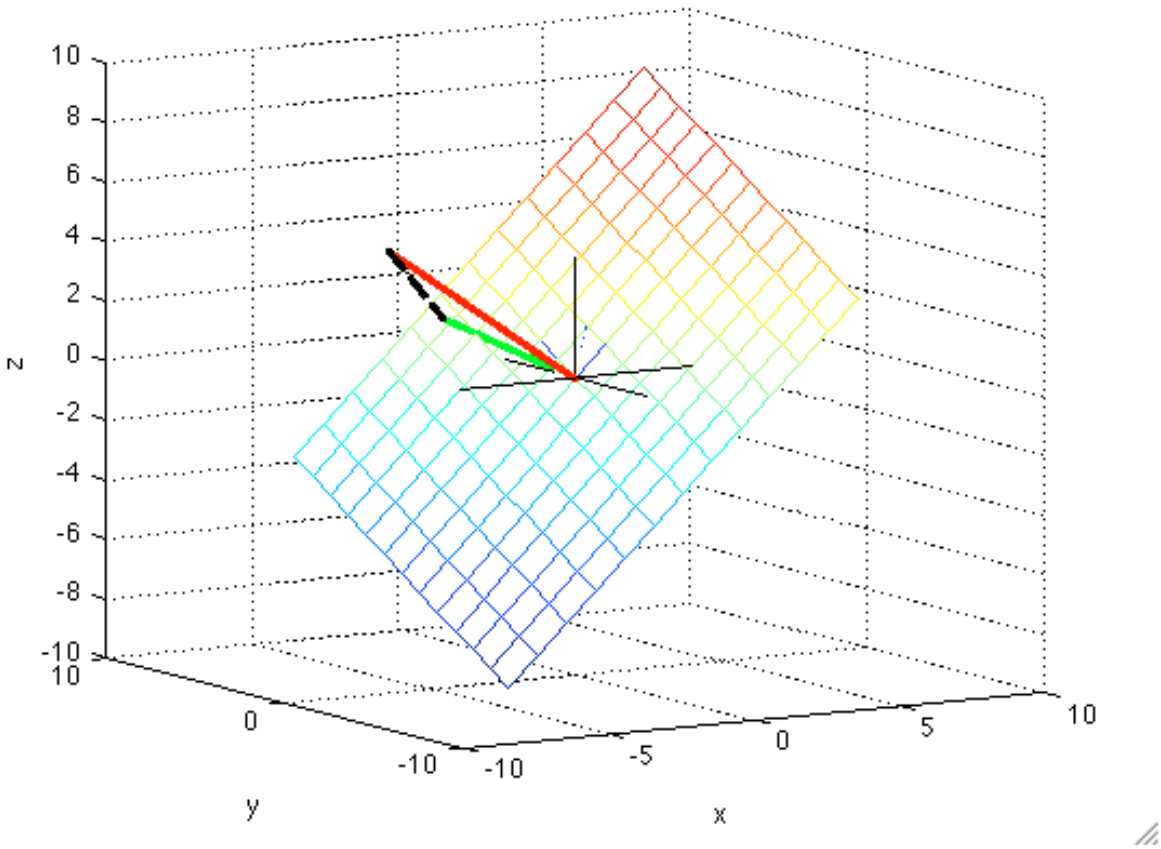


Figure 13

In Figure 13, the red vector is the original observation, i.e. $\vec{y}$, and the green vector is the predicted vector, i.e. $\vec{y}^{\,pred}$. Note that this vector is really a projection down to the subspace spanned by the column vectors of the original design matrix X.

It is rather tedious to construct the pseudoinverse by the definition, but the procedure above gives a clear picture of how to construct the pseudoinverse. However, in practice one can do a trick in the following way.
A matrix X, with N rows and M columns ($X_{N,M}$), having the rank s, can be written in terms of submatrices as:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

where A has the full rank of s and has s rows and s columns. Then A^{-1} exist. Now, a pseudoinverse is just:

$$\begin{bmatrix} A^{-1} & O \\ O & O \end{bmatrix}$$

with the size M rows and N columns. *O* are matrices of zeros, i.e. each entry is zero.

If this is correct then Eq.23 should be correct, that is:

$$X X^{-} X = \begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} A^{-1} & O \\ O & O \end{bmatrix}\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} I & A^{-1}B \\ O & O \end{bmatrix} = \begin{bmatrix} A & B \\ C & CA^{-1}B \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

which requires that $D = C A^{-1} B$.

This is in fact correct because the last (M-s) columns of *X* can be written as a linear combination of the first s columns as:

$$\begin{bmatrix} B \\ D \end{bmatrix} = \begin{bmatrix} A \\ C \end{bmatrix} E \Leftrightarrow \begin{Bmatrix} B = AE \\ D = CE \end{Bmatrix} \Rightarrow D = CA^{-1}B$$

In the example corresponding to Eq.17 a pseudoinverse could be:

$$X^{-} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{because} \quad \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$$

and a solution to Eq.17 would then be:

$$\vec{\beta}^{\,opt} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} -4 \\ 4 \\ 4 \end{bmatrix} = \begin{bmatrix} -4 \\ 2 \\ 0 \end{bmatrix}$$

and

$$\vec{y}^{\,pred} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 2 & 2 \\ 1 & 1 & 3 \end{bmatrix}\begin{bmatrix} -4 \\ 2 \\ 0 \end{bmatrix} = \begin{bmatrix} -4 \\ 4 \\ -2 \end{bmatrix}$$

and $\vec{y}^{\,pred}$ is within the space spanned by the column vectors of *X*, i.e. in I_1 (see Figure 11) as expected: multiply the last column by –2 and add 4 times the second column of *X*. However,

this way of constructing a pseudoinverse does not necessarily result in a solution with minimum norm!.

In Figure 14 this new solution is added.

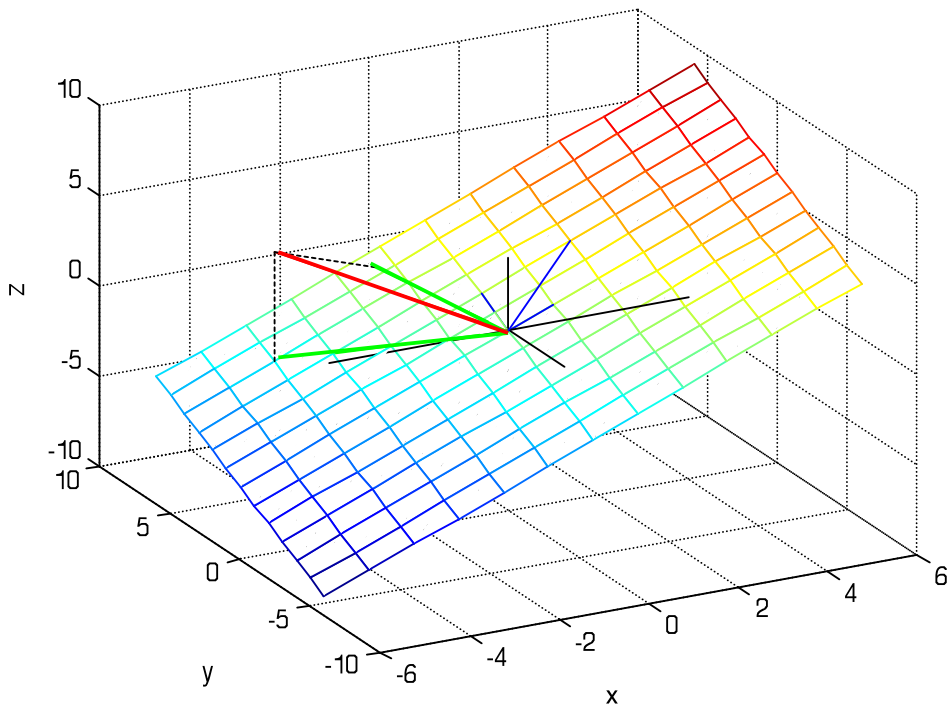


Figure 14

In MATLAB one can always find the pseudoinverse having the least square norm (minimum distance to the solution space) by the command pinv(*X*).

Contrasts

Sometimes the difference between beta-values might be of interest. Consider $\vec{\beta} = \begin{bmatrix} \beta_1 & \beta_2 & \dots & \beta_M \end{bmatrix}^T$. We might be interested in whether $(\beta_i - \beta_j)$ is (statistically) significant. Remember, the individual beta-value indicates how much an explanatory variable contributes to the observed data $\vec{y}$. And we can test whether this contribution is significant. Imagine that we have a paradigm where we move the right index finger, alternating with movement of all fingers on the right side. Then $(\beta_{index\ finger} - \beta_{all\ fingers})$ would indicate whether activation (for a particular voxel) is different in these two situations, and a fMRI map would show where this difference is significant. Note that it would be difficult to construct a single explanatory variable and a corresponding beta, which inherently incorporates this information.

Previously, we saw that the general linear model might be overparameterised, meaning that the solution is not unique concerning the betas. Therefore, difference of betas would naturally also be non-unique. However, this is in fact not always true. We might estimate differences of betas which are unique even though the betas themselves are not unique.

In order to proceed, the contrast vector $\vec{c}$ is introduced. This vector pulls out the contrast we want to investigate, such as:

$$\vec{c}^T\vec{\beta} = \begin{bmatrix} 0 & \cdots & 1 & \cdots & -1 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_i \\ \vdots \\ \beta_j \\ \vdots \\ \beta_M \end{bmatrix} = \beta_i - \beta_j$$

If one can find a new vector $\vec{s}$ such that:

$\vec{c}^T = \vec{s}^T X$ then we have:

$\vec{c}^T\vec{\beta} = \vec{s}^T X\vec{\beta} = \vec{s}^T\vec{y}$ with all the terms being scalars. The last product is a projection of $\vec{y}$ on $\vec{s}$. The relationships show that we get the desired contrast by a linear combination of the entries of the ***y***-vector. The specific combination is determined by the s-vector. This should become clear from an example below.
Also note that $\vec{c}^T = \vec{s}^T X \Leftrightarrow \vec{c} = X^T\vec{s}$ indicates that $\vec{c}$ is just a linear combination of the rows of *X*. We furthermore require that:

$$\vec{c}^T\vec{\beta}^{opt} = \vec{s}^T\vec{y}_n = \vec{s}^T\vec{y}$$

where $\vec{y}_n$ is the orthogonally projected $\vec{y}$ into the space spanned by *X*. Thus, the contrast should be independent of the residual variance, i.e. the noise. Furthermore:

$$\vec{c}^T\vec{\beta}^{opt} = \vec{s}^T\vec{y}_n = \vec{s}^T\vec{y} \Rightarrow \vec{s}^T P\vec{y} = \vec{s}^T\vec{y} \Rightarrow \vec{s}^T P = \vec{s}^T \Leftrightarrow P\vec{s} = \vec{s}$$

These equations show $\vec{s}$ is within the space spanned by the column vectors of X (see section dealing with orthogonal projector). The last conversion above utilize that P is symmetric, so $P^T=P$.
One can test whether the contrast vector gives a unique solution by noting that:

$$\vec{c} = X^T\vec{s} = X^T P\vec{s} = X^T XX^- \vec{s} = X^T XX^- X^{-T}\vec{c} = X^T X(X^T X)^-\vec{c}$$

Thus the contrast vector should be unchanged by multiplication with $X^T X(X^T X)^-$. In conclusion, we have the following conditions that must be satisfied in order to estimate a unique contrast:

$$\vec{c}^T = \vec{s}^T X \Leftrightarrow \vec{c} = X^T\vec{s} \quad \text{Eq.24}$$

which shows that the contrast vector is within the space spanned by X^T, and:

$$\vec{s}^T\vec{y}_n = \vec{s}^T\vec{y} \Leftrightarrow P\vec{s} = \vec{s} \quad \text{Eq.25}$$

which shows that $\vec{s}$ is within the space spanned by X. Equations 24 and Eq.25 lead to the following condition:

$$\vec{c} = X^T X(X^T X)^-\vec{c} \quad \text{Eq.26}$$

which shows that the contrast vector is also within the space spanned by X^TX. In order to illustrate the subtleties of these considerations, we will look at some examples.

In the example corresponding to Eq.15 we would like to estimate the difference between β_1 and β_2. So we have the contrast vector:

$$\vec{c} = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} \text{ and } \vec{c}^T\vec{\beta} = \beta_1 - \beta_2 \text{ and } \vec{s}^T = \vec{c}^T X^-$$

One pseudoinverse for X was previously found to be: $X^- = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{bmatrix}$

We then get: $\vec{c}^T\vec{\beta} = \beta_1 - \beta_2 = \vec{s}^T\vec{y} = \vec{c}^T X^-\vec{y} = \begin{bmatrix} 1 & -1 & 0 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} 2 \\ 4 \\ 4 \end{bmatrix} = 0$

Now we ask the question whether this is a unique solution, so lets us see what happens if we use the other pseudoinverse we have found:

$$X^{-} = \begin{bmatrix} 2/9 & -2/9 & 1/9 \\ -5/18 & 4/9 & -1/18 \\ 1/6 & 0 & 1/6 \end{bmatrix}$$

We get:

$$\vec{c}^{T}\vec{\beta} = \beta_1 - \beta_2 = \vec{s}^{T}\vec{y} = \vec{c}^{T}X^{-}\vec{y} = \begin{bmatrix} 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} 2/9 & -2/9 & 1/9 \\ -5/18 & 4/9 & -1/18 \\ 1/6 & 0 & 1/6 \end{bmatrix} \begin{bmatrix} 2 \\ 4 \\ 4 \end{bmatrix} = -1$$

Clearly, the contrast is not unique.

We also note that:

$$\vec{c} = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} \neq X^{T}X(X^{T}X)^{-}\vec{c} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & 1 \\ 2 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 2 \\ 0 & 2 & 2 \\ 1 & 1 & 3 \end{bmatrix} \left(\begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & 1 \\ 2 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 2 \\ 0 & 2 & 2 \\ 1 & 1 & 3 \end{bmatrix} \right)^{-} \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} = \begin{bmatrix} 2/3 \\ -7/6 \\ 1/6 \end{bmatrix}$$

Clearly, this contrast cannot be estimated uniquely for the given design matrix *X*. Recall that we previously found a solution for the equation:

$$\begin{bmatrix} 2 \\ 4 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 2 & 2 \\ 1 & 1 & 3 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \beta_1 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} + \beta_2 \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix} + \beta_3 \begin{bmatrix} 2 \\ 2 \\ 3 \end{bmatrix} \quad \text{Eq.15}$$

where

$$\vec{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 2 - 2t \\ 2 - t \\ t \end{bmatrix}$$

This could indicate that a contrast vector as:

$$\vec{c} = \begin{bmatrix} 1 \\ -2 \\ 0 \end{bmatrix}$$

would estimate a contrast corresponding to $\beta_1 - 2\beta_2$ uniquely. Let us estimate this value as before. We get:

$$\vec{c}^T\vec{\beta} = \beta_1 - 2\beta_2 = \vec{s}^T\vec{y} = \vec{c}^T X^- \vec{y} = \begin{bmatrix}1 & -2 & 0\end{bmatrix}\begin{bmatrix}2/9 & -2/9 & 1/9\\ -5/18 & 4/9 & -1/18\\ 1/6 & 0 & 1/6\end{bmatrix}\begin{bmatrix}2\\4\\4\end{bmatrix} = -2$$

and

$$\vec{c}^T\vec{\beta} = \beta_1 - 2\beta_2 = \vec{s}^T\vec{y} = \vec{c}^T X^- \vec{y} = \begin{bmatrix}1 & -2 & 0\end{bmatrix}\begin{bmatrix}1 & 0 & 0\\ 0 & \frac{1}{2} & 0\\ 0 & 0 & 0\end{bmatrix}\begin{bmatrix}2\\4\\4\end{bmatrix} = -2$$

We also find that:

$$\vec{c} = \begin{bmatrix}1\\-2\\0\end{bmatrix} = X^T X (X^T X)^- \vec{c} = \begin{bmatrix}1 & 0 & 1\\ 0 & 2 & 1\\ 2 & 2 & 3\end{bmatrix}\begin{bmatrix}1 & 0 & 2\\ 0 & 2 & 2\\ 1 & 1 & 3\end{bmatrix}\left(\begin{bmatrix}1 & 0 & 1\\ 0 & 2 & 1\\ 2 & 2 & 3\end{bmatrix}\begin{bmatrix}1 & 0 & 2\\ 0 & 2 & 2\\ 1 & 1 & 3\end{bmatrix}\right)^-\begin{bmatrix}1\\-2\\0\end{bmatrix} = \begin{bmatrix}1\\-2\\0\end{bmatrix}$$

Now let us have a look at the example corresponding to Eq.17. First, $\vec{y} = \begin{bmatrix}-4 & 4 & 4\end{bmatrix}^T$ is projected orthogonally down into the subspace spanned by the column vectors of *X*. An orthogonal projector is given by:

$P = X(X^T X)^{-1} X^T$ see Eq.10

However, because *X* does not have the full column rank, $(X^T X)^{-1}$ does not exist. Instead we use:

$P = XX^-$

which is an orthogonal projector into the space spanned by the column vectors of *X*, if we use the X^- with the least square norm !. This projector was previously found as:

$$P = \begin{bmatrix}1 & 0 & 2\\ 0 & 2 & 2\\ 1 & 1 & 3\end{bmatrix}\begin{bmatrix}2/9 & -2/9 & 1/9\\ -5/18 & 4/9 & -1/18\\ 1/6 & 0 & 1/6\end{bmatrix} = \begin{bmatrix}5/9 & -2/9 & 4/9\\ -2/9 & 8/9 & 2/9\\ 4/9 & 2/9 & 5/9\end{bmatrix}$$

The orthogonal projection is then calculated as:

$$\vec{y}_n = P\vec{y} = \begin{bmatrix} 5/9 & -2/9 & 4/9 \\ -2/9 & 8/9 & 2/9 \\ 4/9 & 2/9 & 5/9 \end{bmatrix} \begin{bmatrix} -4 \\ 4 \\ 4 \end{bmatrix} = \begin{bmatrix} -4/3 \\ 16/3 \\ 4/3 \end{bmatrix}$$

Again we can find a unique contrast with a contrast vector $\vec{c} = [1 \quad -2 \quad 0]^T$ as:

$$\vec{c}^T\vec{\beta} = \beta_1 - 2\beta_2 = \vec{s}^T\vec{y} = \vec{c}^T X^- \vec{y} = \begin{bmatrix} 1 & -2 & 0 \end{bmatrix} \begin{bmatrix} 2/9 & -2/9 & 1/9 \\ -5/18 & 4/9 & -1/18 \\ 1/6 & 0 & 1/6 \end{bmatrix} \begin{bmatrix} -4/3 \\ 16/3 \\ 4/3 \end{bmatrix} = -20/3$$

or

$$\vec{c}^T\vec{\beta} = \beta_1 - 2\beta_2 = \vec{s}^T\vec{y} = \vec{c}^T X^- \vec{y} = \begin{bmatrix} 1 & -2 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} -4/3 \\ 16/3 \\ 4/3 \end{bmatrix} = -20/3$$

We note that this particular contrast is unique and does not depend on the specific pseudoinverse employed. Note, that we first projected ***y*** orthogonally into the space spanned by the column vectors of *X*. If we had applied a pseudoinverse directly on ***y***, which was not the one with the least square norm, the projection would not necessary be an orthogonal projection, and the contrast calculated from the actual contrast vector $\vec{c} = \begin{bmatrix} 1 & -2 & 0 \end{bmatrix}^T$ would not be identical for the two situations. Also recall that the orthogonal subspace to the space spanned by *X*, i.e. (*I*-*P*)***y***, represents the variance not explained by the model, and is therefore interpreted as the noise. Also note that the contrast –20/3 could be calculated directly from the solution when using the pseudoinverse with the least square norm, (see above), where we found:

$$\vec{\beta}^{opt} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 2/9 & -2/9 & 1/9 \\ -5/18 & 4/9 & -1/18 \\ 1/6 & 0 & 1/6 \end{bmatrix} \begin{bmatrix} -4 \\ 4 \\ 4 \end{bmatrix} = \begin{bmatrix} -4/3 \\ 8/3 \\ 0 \end{bmatrix}$$

We directly find that $\beta_1 - 2\beta_2 = -4/3 - 2 \cdot 8/3 = -20/3$

Normally, we are not interested in a contrast vector like the one above, but the examples above show that a unique contrast is dependent on the design matrix *X*. In the examples the design matrix *X* was more or less randomly selected. In practice a typical design matrix and the typical linear model could look like:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \quad \text{Eq.27}$$

The first column of the design matrix could correspond to mean of the MR signal, while the next column could correspond to movement of the index finger, and the last column could correspond to movement of all fingers. Clearly, this design matrix is rank deficient as the first column is equal to the sum of the last two, so the matrix does not have the full column rank and the model is overparameterised. Therefore, a unique estimate of the individual beta's is meaningless. As an example, imagine that we had observed $\vec{y}_n = [2 \quad 5 \quad 2 \quad 5 \quad 2 \quad 5]^T$ (i.e. the observed data after an orthogonal projection into the space spanned by the design matrix). Then the sets of possible solutions are:

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} t \\ 2-t \\ 5-t \end{bmatrix} \quad \wedge \quad t \in \Re$$

Obviously, the only unique parameter, which can be estimated, is $\beta_2 - \beta_3$ corresponding to the contrast vector: $\vec{c} = \begin{bmatrix} 0 & 1 & -1 \end{bmatrix}^T$. This can be verified by using Eq.26. So, even though the system is overparameterised, there is still a useful linear combination of parameters. We will interpret this contrast as the difference between activation of the index finger and all fingers. From this example, it is easy to see the meaning of the s-vector (see Eq. 25). We get:

$$\vec{s} = \frac{1}{3}\begin{bmatrix} -1 & 1 & -1 & 1 & -1 & 1 \end{bmatrix}^T$$

Thus, the effect of the *s*-vector is simply a reorganisation of the linear system to extract the contrast in question. In the example, it corresponds to performing three sets of measurements. Each set consists of a subtraction of a measurement with movement of all fingers from another measurement with only movement of the index fingers. For completeness, the contrast $\vec{c} = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}^T$ is also possible, but the interpretation is dubious (the difference between the global mean and movement of the index finger). If the mean value had been subtracted from all observations, leaving us with the following model:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \quad \text{Eq.28}$$

then the individual betas are uniquely defined. This is also consistent with the fact that the design matrix now have full column rank (=2), and Eq.26 is satisfied.

Normally, one is not very interested in the difference itself between various betas, but in whether or not the difference (contrast) is significant. It can be shown that the linear compound $\vec{c}^T\vec{\beta}$ can be tested by a student t-test as:

$$t_{N-p} = \frac{\vec{c}^T\vec{\beta}^{opt} - d}{\sqrt{\sigma^2\vec{c}^T\left(X^TX\right)^{-1}\vec{c}}} \quad \text{Eq.29}$$

with N-p degrees of freedom, where N is the number of observation, p is the rank of X. Equation 29 tests whether the linear compound $\vec{c}^T\vec{\beta}^{opt}$ is significantly different from the scalar d. Most often d is zero. If X does not have the full column rank then $\left(X^TX\right)^-$ is used instead.

The use of F-tests

Sometimes one wants to investigate whether one model is better than another model in fMRI. The model is per definition given by the design matrix X. Imaging that we have a full model given by X_1 with a corresponding residual variance, Var_1, which is given by Eq 13:

$$Var_1 = \frac{SOSQ_1}{N - rank(X_1)} = \frac{e_1(:)^T e_1(:)}{N - rank(X_1)}$$

where SOSQ is the residual sum of squares. We create a reduced model, called X_2, with a residual variance, Var_2, by removing one or more columns of the design matrix X_1. Var_2 is also given by Eq.13:

$$Var_2 = \frac{SOSQ_2}{N - rank(X_2)} = \frac{e_2(:)^T e_2(:)}{N - rank(X_2)}$$

If we want to test whether the full model is significantly better than the reduced model in regards to explaining the observed data, an F-test can be used:

$$F_{rank(X_1)-rank(X_2),N-rank(X_1)} = \frac{(SOSQ_2 - SOSQ_1)/(rank(X_1) - rank(X_2))}{SOSQ_1/(N - rank(X_1))} =$$

$$\frac{SOSQ_2 - SOSQ_1}{SOSQ_1} \quad \frac{N - rank(X_1)}{rank(X_1) - rank(X_2)}$$

The nominator is equal to the extra increase of the residual sum of squares due to the reduced model after using the full model. The larger F gets the more unlikely it is that the full model and the reduced model equally explain the observed data ***y***.

The procedure above can be reformulated in order to give a more explicit understanding of the F-test. Recall that the error vector $\vec{e}$ can be calculated directly by an orthogonal projection as

$$\vec{e} = (I - P)\vec{y} = (I - XX^-)\vec{y} = R\vec{y}$$

where quit generally $P = XX^-$. (If X has, as previously mentioned, full column rank, then $P = XX^- = X(X^TX)^{-1}X^T$). Therefore, the error vector can be calculated directly for the full model and the reduced model as:

$$\vec{e}_1 = (I - X_1X_1^-)\vec{y}$$

and

$$\vec{e}_2 = (I - X_2X_2^-)\vec{y}$$

The difference error vector can be calculated directly as:

$$\vec{e}_3 = \vec{e}_2 - \vec{e}_1 = (I - X_2X_2^-)\vec{y} - (I - X_1X_1^-)\vec{y} = ((I - X_2X_2^-) - (I - X_1X_1^-))\vec{y} = S\vec{y}$$

The *S* matrix is a new matrix which directly projects $\vec{y}$ into the difference error vector between the two models. As a projector *S* obeys the following two rules:

$$SS = S$$
$$S^T = S$$

Therefore we can reformulate the F-test as:

$$F = \frac{SOSQ_2 - SOSQ_1}{SOSQ_1} = \frac{SOSQ_3}{SOSQ_1} = \frac{\vec{e}_3^T\vec{e}_3}{\vec{e}_1^T\vec{e}_1} = \frac{(S\vec{y})^T S\vec{y}}{(R_1\vec{y})^T R_1\vec{y}} = \frac{\vec{y}^T S^T S\vec{y}}{\vec{y}^T R_1^T R_1\vec{y}} = \frac{\vec{y}^T S\vec{y}}{\vec{y}^T R_1\vec{y}}$$

where $R_1 = (I - X_1X_1^-)$ and the appropriate weighting with the number of freedom is ignored.

The correct expression is:

$$F_{rank(S),N-rank(X_1)} = \frac{\vec{y}^T S \vec{y}}{\vec{y}^T R_1 \vec{y}} \frac{N - rank(X_1)}{rank(S)}$$

Obviously $\boldsymbol{e}_1$ is orthogonal to the space spanned by X_1, and $\boldsymbol{e}_2$ is orthogonal to the subspace spanned by X_2. Because X_2 is a subspace of X_1, $\boldsymbol{e}_3$ must lie within a subspace of X_1 and $\vec{e}_2 = \vec{e}_1 + \vec{e}_3$. Finally, the S matrix projects $\boldsymbol{y}$ into a subspace of X_1. Therefore, we can substitute the $\boldsymbol{y}$-vector in the nominator with the optimal betas obtained from the full model as:

$$F_{rank(S),N-rank(X_1)} = \frac{\vec{y}^T S \vec{y}}{\vec{y}^T R_1 \vec{y}} \frac{N - rank(X_1)}{rank(S)} = \frac{\vec{\beta}^{opt^T} X^T S X \vec{\beta}^{opt}}{\vec{y}^T R_1 \vec{y}} \frac{N - rank(X_1)}{rank(S)} \quad \text{Eq.30}$$

Therefore, if we have estimated all beta-values for all voxels for the full model, and we want to evaluate a reduced model, then we need not estimate a new set of beta-values for all voxels. We only have to calculate the S matrix and its rank, and use the previously obtained beta-values as stated in Eq.30. This, therefore constitute a time saving procedure.
Again, an example in 3D space is appropriate. Looking at Eq. 17:

$$\vec{y} = X_1 \vec{\beta} + \vec{e}_1 = \begin{bmatrix} -4 \\ 4 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 2 & 2 \\ 1 & 1 & 3 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \vec{e}_1 \quad \text{Eq.17}$$

where X_1 represents the full model. We now ask: “do we need all the columns of the design matrix, or could we just use a design matrix with one column”, e.g. the second column. Therefore, the reduced model looks like:

$$\vec{y} = X_2 \vec{\beta} + \vec{e}_2 = \begin{bmatrix} -4 \\ 4 \\ 4 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix} \vec{\beta} + \vec{e}_2 \quad \wedge \quad \vec{\beta} = \beta_2$$

Now in essence what we have to do is to compare the resulting residuals of the two models. We find:

$$\vec{e}_1 = (I - X_1 X_1^-)\vec{y} = \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 0 & 2 \\ 0 & 2 & 2 \\ 1 & 1 & 3 \end{bmatrix} \begin{bmatrix} 2/9 & -2/9 & 1/9 \\ -5/18 & 4/9 & -1/18 \\ 1/6 & 0 & 1/6 \end{bmatrix} \right) \begin{bmatrix} -4 \\ 4 \\ 4 \end{bmatrix} =$$

$$\left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 5/9 & -2/9 & 4/9 \\ -2/9 & 8/9 & 2/9 \\ 4/9 & 2/9 & 5/9 \end{bmatrix} \right) \begin{bmatrix} -4 \\ 4 \\ 4 \end{bmatrix} = \begin{bmatrix} 4/9 & 2/9 & -4/9 \\ 2/9 & 1/9 & -2/9 \\ -4/9 & -2/9 & 4/9 \end{bmatrix} \begin{bmatrix} -4 \\ 4 \\ 4 \end{bmatrix} = \begin{bmatrix} -8/3 \\ -4/3 \\ 8/3 \end{bmatrix}$$

where

$$R_1 = I - X_1 X_1^- = \begin{bmatrix} 4/9 & 2/9 & -4/9 \\ 2/9 & 1/9 & -2/9 \\ -4/9 & -2/9 & 4/9 \end{bmatrix}$$

Likewise, we find for the reduced model:

$$\vec{e}_2 = (I - X_2 X_2^-)\vec{y} = \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix} \begin{bmatrix} 0 & 2/5 & 1/5 \end{bmatrix} \right) \begin{bmatrix} -4 \\ 4 \\ 4 \end{bmatrix} =$$

$$\left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & 4/5 & 2/5 \\ 0 & 2/5 & 1/5 \end{bmatrix} \right) \begin{bmatrix} -4 \\ 4 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/5 & -2/5 \\ 0 & -2/5 & 4/5 \end{bmatrix} \begin{bmatrix} -4 \\ 4 \\ 4 \end{bmatrix} = \begin{bmatrix} -4 \\ -4/5 \\ 8/5 \end{bmatrix}$$

where

$$R_2 = I - X_2 X_2^- = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/5 & -2/5 \\ 0 & -2/5 & 4/5 \end{bmatrix}$$

In fact, we need not to calculate the error-vectors explicitly, only R_1 and R_2.
The S matrix is then:

$$S = R_2 - R_1 = (I - X_2 X_2^-) - (I - X_1 X_1^-) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/5 & -2/5 \\ 0 & -2/5 & 4/5 \end{bmatrix} - \begin{bmatrix} 4/9 & 2/9 & -4/9 \\ 2/9 & 1/9 & -2/9 \\ -4/9 & -2/9 & 4/9 \end{bmatrix} =$$

$$\begin{bmatrix} 5/9 & -2/9 & 4/9 \\ -2/9 & 4/45 & -8/45 \\ 4/9 & -8/45 & 16/45 \end{bmatrix}$$

Now the estimation of F:

$$F_{1,1} = \frac{\vec{y}^T S \vec{y}}{\vec{y}^T R_1 \vec{y}} \frac{N - rank(X)}{rank(S)} =$$

$$\frac{\begin{bmatrix} -4 & 4 & 4 \end{bmatrix} \begin{bmatrix} 5/9 & -2/9 & 4/9 \\ -2/9 & 4/45 & -8/45 \\ 4/9 & -8/45 & 16/45 \end{bmatrix} \begin{bmatrix} -4 \\ 4 \\ 4 \end{bmatrix}}{\begin{bmatrix} -4 & 4 & 4 \end{bmatrix} \begin{bmatrix} 4/9 & 2/9 & -4/9 \\ 2/9 & 1/9 & -2/9 \\ -4/9 & -2/9 & 4/9 \end{bmatrix} \begin{bmatrix} -4 \\ 4 \\ 4 \end{bmatrix}} \frac{3-2}{1} = \frac{1}{5}$$

This is not significant, so if our initial null hypothesis was that the full model and the reduced model explain the observed data, ***y***, equally well, then the F-test supports the null hypothesis, and we would then only use the reduced model in further considerations. Normally, we have, of course, many observations so N may be in the hundreds. If the same tendency as in the 3D example had applied to all the observations then the F-test had likely generated a higher value and would thus become significant. But, the 3D example shows the principal features.

Now, let's try another reduced model by taking just the last column as explanatory variable as an example:

$$\vec{y} = X_2\vec{\beta} + \vec{e}_2 = \begin{bmatrix} -4 \\ 4 \\ 4 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 3 \end{bmatrix}\vec{\beta} + \vec{e}_2 \quad \wedge \quad \vec{\beta} = \beta_3$$

We calculate a new error vector corresponding to the reduced model:

$$\vec{e}_2 = (I - X_2X_2^-)\vec{y} = \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 2 \\ 2 \\ 3 \end{bmatrix}\begin{bmatrix} 0.1176 & 0.1176 & 0.1765 \end{bmatrix}\right)\begin{bmatrix} -4 \\ 4 \\ 4 \end{bmatrix} =$$

$$\left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0.2353 & 0.2353 & 0.3529 \\ 0.2353 & 0.2353 & 0.3529 \\ 0.3529 & 0.3529 & 0.5294 \end{bmatrix}\right)\begin{bmatrix} -4 \\ 4 \\ 4 \end{bmatrix} = \begin{bmatrix} 0.7647 & -0.2353 & -0.3529 \\ -0.2353 & 0.7647 & -0.3529 \\ -0.3529 & -0.3529 & 0.4706 \end{bmatrix}\begin{bmatrix} -4 \\ 4 \\ 4 \end{bmatrix} = \begin{bmatrix} -5.4118 \\ 2.5882 \\ 1.8824 \end{bmatrix}$$

(We need not calculate this error vector explicitly).
And

$$R_2 = I - X_2X_2^- = \begin{bmatrix} 0.7647 & -0.2353 & -0.3529 \\ -0.2353 & 0.7647 & -0.3529 \\ -0.3529 & -0.3529 & 0.4706 \end{bmatrix}$$

Then, we calculate the *S* matrix again:

$$S = R_2 - R_1 = (I - X_2 X_2^-) - (I - X_1 X_1^-) = \begin{bmatrix} 0.7647 & -0.2353 & -0.3529 \\ -0.2353 & 0.7647 & -0.3529 \\ -0.3529 & -0.3529 & 0.4706 \end{bmatrix} - \begin{bmatrix} 4/9 & 2/9 & -4/9 \\ 2/9 & 1/9 & -2/9 \\ -4/9 & -2/9 & 4/9 \end{bmatrix} =$$

$$\begin{bmatrix} 0.3203 & -0.4575 & 0.0915 \\ -0.4575 & 0.6536 & -0.1307 \\ 0.0915 & -0.1307 & 0.0261 \end{bmatrix}$$

This matrix is of course the same for all voxels.
And the F-test (which can be performed voxelwise):

$$F_{1,1} = \frac{\vec{y}^T S \vec{y}}{\vec{y}^T R_1 \vec{y}} \frac{N - rank(X)}{rank(S)} =$$

$$\frac{\begin{bmatrix} -4 & 4 & 4 \end{bmatrix} \begin{bmatrix} 0.3203 & -0.4575 & 0.0915 \\ -0.4575 & 0.6536 & -0.1307 \\ 0.0915 & -0.1307 & 0.0261 \end{bmatrix} \begin{bmatrix} -4 \\ 4 \\ 4 \end{bmatrix}}{\begin{bmatrix} -4 & 4 & 4 \end{bmatrix} \begin{bmatrix} 4/9 & 2/9 & -4/9 \\ 2/9 & 1/9 & -2/9 \\ -4/9 & -2/9 & 4/9 \end{bmatrix} \begin{bmatrix} -4 \\ 4 \\ 4 \end{bmatrix}} \frac{3-2}{1} = 1.47$$

which is not significant (due to the small number of observations). The situation is shown in figure 15a and 15b (from another angle of view). The red vector is the observed data, ***y***. This vector is projected into the space spanned by the full model. The projection corresponds to the thick green line. The difference between ***y*** and its projection gives $\boldsymbol{e}_1$. ***y*** is also projected into the space spanned by the third column vector which corresponds to the thin green line. The difference between ***y*** and its projection gives $\boldsymbol{e}_2$. The difference between $\boldsymbol{e}_2$ and $\boldsymbol{e}_1$ gives $\boldsymbol{e}_3$. It is clearly shown that when a reduced model is used, the error vector increases, and this increase is part of the nominator of the F-test.

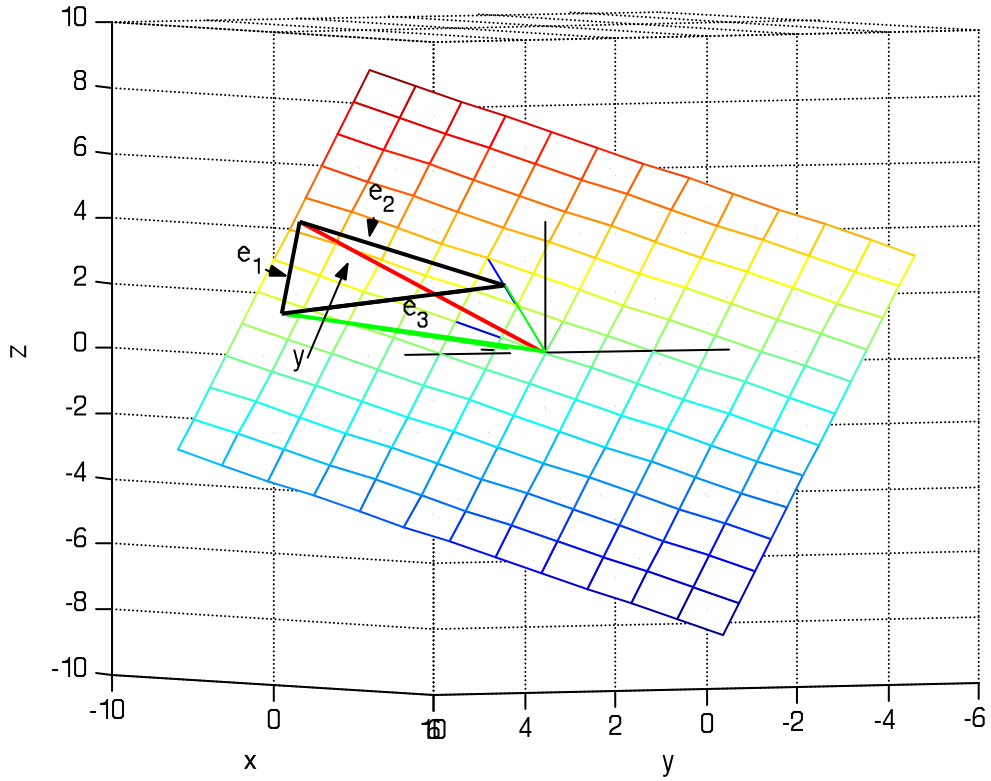


Figure 15a

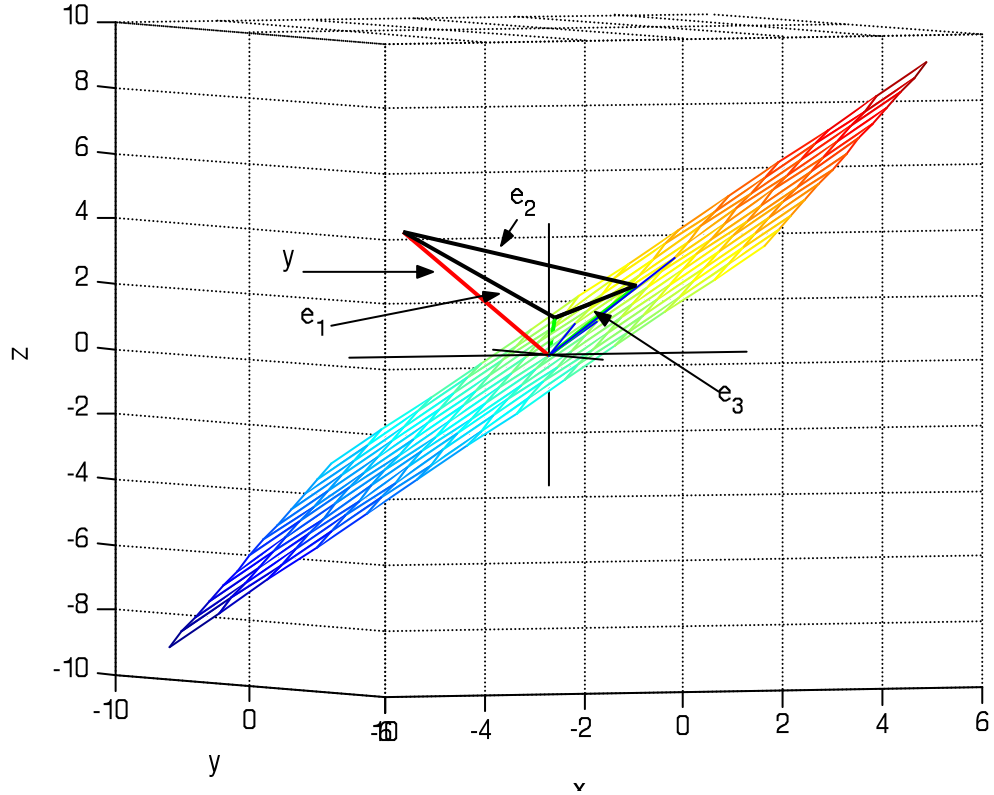


Figure 15b

Eigenvectors and eigenvalues

Consider a square matrix X_{NxN} and a nonzero vector $\boldsymbol{\beta}$. We want to find the solutions to the equation:

$$X\vec{\beta} = \lambda\vec{\beta} \quad \text{Eq.30}$$

The geometrical meaning of multiplying vector $\boldsymbol{\beta}$ by X is a vector, which should fulfil Eq.30; i.e. the resulting vector has the same direction as the vector itself, but is scaled by a scalar λ. We want to find all those beta vectors that fulfil this criterion. The solution vector $\boldsymbol{\beta}$ is called an eigenvector and the scalar is called an eigenvalue. One or more eigenvectors and eigenvalues exist for a given square matrix X. Eq.30 is equivalent with:

$$(X - \lambda I)\vec{\beta} = \vec{0}, \quad \vec{\beta} \neq \vec{0} \quad \text{Eq.31}$$

where I is the identity matrix. The solution to this equation is obviously found for situations where the matrix (X-λI) is singular. Otherwise, if the matrix (X-λI) is non-singular, meaning that it can be inverted directly, we then have:

$$\vec{\beta} = (X - \lambda I)^{-1}\vec{0} = \vec{0}, \quad \vec{\beta} \neq \vec{0}$$

which contradicts our definition. A matrix is singular only if the determinant is zero:

$$\det(X - \lambda I) = 0 \quad \text{Eq.32}$$

This determinant is a polynomial in λ of degree N, and is called the characteristic polynomial of the matrix X. The polynomial has N roots, which may be complex values. The same distinct value of a root may appear more than once. If the same root appears 'r' times, we say that the *algebraic multiplicity* of this root is 'r'.

It can be shown that:

$$\det(X) = \lambda_1\lambda_2 \cdots \lambda_N \quad \text{Eq.33}$$

$$\text{tr}(X) = \lambda_1 + \lambda_2 + .. + .. + \lambda_N \quad \text{Eq.34}$$

where 'tr' stands for the trace, and is defined as the sum of the diagonal entries of a matrix. Therefore, the matrix X *itself* is singular if and only if one or more of the eigenvalues are zero.

For each eigenvalue λ of X, there exists a set of nonzero vectors, eigenvectors, $\boldsymbol{\beta}$ such that Eq.30 is satisfied. The set of all the eigenvectors, corresponding to the eigenvalue λ, together with the zero vector, is a vector space called the *eigenspace* of λ, and may be written as E(λ). Even though the entries of X can be real values, both the eigenvalues and eigenvectors might have complex entries.

Note that the eigenspace E(λ) corresponds to the null space of (X- λI):

$$\text{E}(\lambda) = null(X - \lambda I)$$

and

$$\dim(E(\lambda)) = N - rank(X - \lambda I) \quad \text{Eq.35}$$

Do you see why? Recall that the rank of a matrix denotes the dimension of the image space. The dimension of the definition space is N. The matrix (X-λI) has a certain rank, let's call it 'k'. If k < N, it means that N-k dimensions collapse, and this is exactly the dimension of the eigenspace of λ, E(λ) (see Figure 10). In Figure 10, the dimension of the definition space was set to M, and the rank of X was set to N, so here M-N is the dimension of the null-space and would here correspond to the dimension of the eigenspace).

For each of the eigenvalues of X, which might appear more than once, the dimension 'N-k' of the associated eigenspace is called the *geometric multiplicity* associated with this eigenvalue. Note that if an eigenvalue only appears once, then only one eigenvector will be associated with this eigenvalue, and the geometric multiplicity is one that is the dimension of this eigenspace is one. However, if an eigenvalue has an algebraic multiplicity greater than one, say two, then the corresponding eigenspace is spanned by a corresponding number of eigenvectors, e.g. two. It is the dimension of this space, which is denoted as the *geometric multiplicity* corresponding to the specific λ. It can be shown that, the *geometric multiplicity* is always equal to or smaller than the *algebraic multiplicity*.

It can be shown that if the matrix X has 's' distinct eigenvalues, then the corresponding 's' eigenvectors are linearly independent. Especially if the matrix X_{NxN} has N distinct eigenvalues, then all the eigenvectors are linearly independent. Such a matrix, having 'N' linearly independent eigenvectors, is called *simple*, and we say that the matrix has a complete, or full, set of eigenvectors. Matrices that are not simple are called *defective.*

However, a matrix can be simple (meaning having N linearly independent eigenvectors), even though the matrix does not have N distinct eigenvalues, that is some eigenvalues appear more than once.

Diagonalizable matrices

Consider again a square matrix $X_{\mathrm{N,N}}$. Clearly, we have N eigenvectors ($\boldsymbol{\beta}$) and N eigenvalues (λ). If we now define a matrix S and a matrix Λ such that:

$$S = \left[\vec{\beta}_1 \quad \vec{\beta}_2 \quad \dots \quad \vec{\beta}_N\right]$$

and

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix}$$

then Eq.30 is equivalent with:

$$XS = S\Lambda \quad \text{Eq.36}$$

Obviously, if the matrix X is simple, which means that it has N linearly independent eigenvectors and therefore S has full rank, then we can divide Eq.36 with S^{-1} on both side:

$$S^{-1}XS = \Lambda \quad \text{Eq.37}$$

The invertible matrix S in Eq.37 is said to diagonalize the matrix X. Then X is simple if and only if X is diagonalizable for some S. S need not to be unique, e.g. one can just make a permutation of the columns of S. Also, if the algebraic multiplicity of a distinct eigenvalue is two for example (i.e. the eigenvalue appears two time as a root in the characteristic polynomial), and the geometric multiplicity is also two (i.e. corresponding to this eigenvalue we have two linearly independent eigenvectors), then the sum of these two eigenvectors is also an eigenvector, because:

$$X\vec{v}_1 = \lambda\vec{v}_1 \quad \text{and} \quad X\vec{v}_2 = \lambda\vec{v}_2 \Rightarrow X(\vec{v}_1 + \vec{v}_2) = X\vec{v}_1 + X\vec{v}_2 = \lambda\vec{v}_1 + \lambda\vec{v}_2 = \lambda(\vec{v}_1 + \vec{v}_2)$$

In fact every linear combination of these two eigenvectors is also an eigenvector.

Naturally, if the number of linearly independent eigenvectors corresponding to each eigenvalue is exactly equal to the algebraic multiplicity of that eigenvalue, then the matrix is simple. Importantly, real symmetric matrices are examples of simple matrices, a point we will come back to later. Also idempotent matrices (i.e. $P^2 = P$) are examples of simple matrices. Why? Clue: consider the images of all possible types vectors.

If rank(X) = r < N, then the rank of the matrix $S^{-1}XS$ is also 'r' (rank($S^{-1}XS$) = r), because (it can be shown that) the rank of X is unaltered by pre- or postmultiplication by a non-singular matrix. This tells us, that X will have 'r' nonzero eigenvalues.

The hermitian inner product, unitary matrices and Schur's theorem

Recall, that if $Q_{N,N}$ represented an orthonormal matrix, i.e.:

$$Q_{N,N} = \begin{bmatrix} \vec{q}_1 & \vec{q}_2 & \cdots & \vec{q}_N \end{bmatrix} \quad \text{where} \quad \vec{q}_i = \vec{q}_{N,i}$$

stating that each $\boldsymbol{q}$-vector is a column vector with N entries, then:

$$QQ^T = Q^TQ = I \quad \text{Eq.38}$$

that is an N x N identity matrix.

If we allow entries to be complex valued, then we can find a broader group of matrices which have the property corresponding to Eq.38. Consider such a matrix B, then we define the *hermitian* transpose of B as:

$$\text{The hermitian transpose of } B : B^H = \left(B^*\right)^T$$

that is we first conjugate all entries (if the entry is real then this operation has no effect), and then we transpose the matrix. We also define the *hermitian inner product* of vectors with possible complex entries:

$$\langle \vec{x}, \vec{y} \rangle = x_1^* y_1 + x_2^* y_2 + ... + x_N^* y_N \quad \text{Eq.39}$$

A general consequence of this definition is that $\langle \vec{x}, \vec{y} \rangle \neq \langle \vec{y}, \vec{x} \rangle$ but $\langle \vec{x}, \vec{y} \rangle^* = \langle \vec{y}, \vec{x} \rangle$.

Clearly, we could write this as:

$$\langle \vec{x}, \vec{y} \rangle = x_1^* y_1 + x_2^* y_2 + ... + x_N^* y_N = \begin{bmatrix} x_1^* & x_2^* & \cdots & x_N^* \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \vec{x}^H \vec{y}$$

Eventionally, a matrix U is called *unitary* if its columns form an orthonomal set with respect to the hermitian inner product. Therefore:

$$UU^H = U^H U = I \quad \text{Eq.40}$$

Note that this equation implies that:

$$U^H = U^{-1} \quad \textit{and} \quad (U^H)^H = U$$

If the matrix U only has real entries, then the columns of U form an orthonormal set and of course obey Eq.38. It is straightforward to show that unitary matrices preserve norms and inner products in the hermitian sense. First note that:

$$\langle U\vec{x}, U\vec{y} \rangle = (U\vec{x})^H U\vec{y} = \vec{x}^H U^H U\vec{y} = \vec{x}^H \vec{y}$$

and

$$\|U\vec{x}\|^2 = \langle U\vec{x}, U\vec{x} \rangle = (U\vec{x})^H U\vec{x} = \vec{x}^H U^H U\vec{x} = \vec{x}^H \vec{x} = \|\vec{x}\|^2 \quad \text{Eq.41}$$

If U is unitary and λ is an eigenvalue of U, then $|\lambda|=1$ and $|\det(U)| =1$, because:

$$U\vec{x} = \lambda\vec{x} \quad \text{implies:} \quad \langle U\vec{x}, U\vec{x} \rangle = \langle \lambda\vec{x}, \lambda\vec{x} \rangle = \lambda^2 \vec{x}^H \vec{x} = \lambda^2 \|\vec{x}\|^2 \quad \text{Eq.42}$$

which together with Eq.41 shows that $\lambda^2 = 1$. This must apply to every possible eigenvalue for U and because of Eq.33, $|\det(U)| =1$.

If U is a unitary matrix, then for some matrix A:

$$U^H AU = B \quad \text{then} \quad UBU^H = A \quad \text{Eq.43}$$

because:

$$U^H AU = B \Leftrightarrow UU^H AU = UB \Leftrightarrow AU = UB \Leftrightarrow AUU^H = UBU^H \Leftrightarrow A = UBU^H$$

We say that B is *unitarily similar* to A and A is unitarily similar to B, and call the process corresponding to Eq.43 a unitary transformation. If A and B are unitarily similar to each other, then A and B have the same eigenvalues. The eigenvalues and eigenvectors can be written as in Eq. 36:

$$AS = S\Lambda \quad \text{and} \quad A = UBU^H \Rightarrow UBU^H S = S\Lambda \Leftrightarrow BU^H S = U^H S\Lambda$$

Thus, B and A have the same eigenvalues. If S is invertible, then:

$$S^{-1}AS = \Lambda \quad \text{and} \quad \left(U^H S\right)^{-1} BU^H S = \Lambda$$

and we can take $U^H S$ as the diagonalizing matrix of B.

Finally, remember if a matrix $A_{N,N}$ is simple, i.e. it has N linearly independent eigenvectors, then we can diagonalize it, see Eq.37. This is not possible for defective matrices. However, it can be proved that for any matrix $A_{N,N}$, simple or defective, one can always find a unitary matrix U and an upper triangular matrix T such that:

$$\text{Schur's theorem:} \quad U^H AU = T \qquad \text{Eq.44}$$

An upper triangular matrix means that all entries below the diagonal are zero. As mentioned above, because A and T are unitarily similar to each other, they have the same eigenvalues. However, the eigenvalues are particularly easy to estimate for a triangular matrix, because it is simply the diagonal entries of T. This is because we only have to multiply the diagonal entries in order to calculate the determinant, which gives $\prod_{i=1}^{N}(x_{ii} - \lambda)$, since all other terms of the determinant are zero. It should be mentioned that Schur's theorem does not necessarily provide the diagonalization of A.

Hermitian matrices

A matrix A, where the hermitian transpose is equal to itself, is called an *hermitian matrix*. That is:

$$\text{An hermitian matrix:} \quad A^H = A \qquad \text{Eq.45}$$

Obviously, every real symmetric matrix is hermitian. This apparently innocent property turns out to be of major significance in multivariat statistic and image processing. The reason is that the so-called dispersion matrix, which describes the variance structure of a multivariate stochastic variable, is in fact, always symmetric. The implication of Eq.45 is mentioned in the following.

First, a hermitian matrix must have a diagonal with real entries since (a+ib)* = a-ib. In order to fullfill Eq.45, the imaginary part has to be zero.

The eigenvalues of a hermitian matrix A are real, and there exists a unitary matrix U such that:

$$U^H AU = \Lambda \quad \text{Eq.46}$$

because for every matrix $A_{N,N}$, we can use Schur's theorem, and then take the hermitian transpose and finally use that $A^H=A$:

$$U^H AU = T \Leftrightarrow \left(U^H AU\right)^H = T^H \Leftrightarrow U^H A^H U^{HH} = T^H \Leftrightarrow U^H A^H U = T^H \Leftrightarrow U^H AU = T^H$$

We therefore have:

$$U^H AU = T \quad \text{and} \quad U^H AU = T^H \Rightarrow T^H = T$$

which is possible only if T is a diagonal matrix. Thus, a hermitian matrix fulfils Eq.46. Also note that the diagonal entries of T have to be real which means that the eigenvalues are real. If we compare Eq.46 with Eq.37 then we see that a hermitian matrix must be simple!. Furthermore, a hermitian matrix has an orthogonal set of normalized eigenvectors spanning the N dimensional complex vector space. This is because $U^HU=I$.

If A is a real symmetic matrix then it is of course hermitian, and there exists a real orthonormal set of eigenvectors spanning the N dimensional real vector space. This is because all calculations can be done using only real values. However, it is important to note that Eq.46 does not guarantee that all sets of eigenvectors are mutually orthogonal!. However, one can always find a new set of eigenvectors which are mutually orthogonal, as shown in the following example.

In some cases, however, the eigenvectors of A are automatically orthogonal. If a hermitian matrix A has eigenpairs $(\lambda,\vec{u})$ and $(\mu,\vec{v})$ where $\lambda \neq \mu$, then the two eigenvectors are orthogonal, because:

$$\left(A\vec{u}\right)^H \vec{v} = \left(\lambda\vec{u}\right)^H \vec{v} = \lambda\vec{u}^H\vec{v}$$

and

$$\left(A\vec{u}\right)^H \vec{v} = \vec{u}^H A^H \vec{v} = \vec{u}^H A\vec{v} = \vec{u}^H\left(A\vec{v}\right) = \vec{u}^H\left(\mu\vec{v}\right) = \mu\,\vec{u}^H\vec{v}$$

This implies that:

$$\lambda\vec{u}^H\vec{v} = \mu\,\vec{u}^H\vec{v}$$

which is possible only if the two eigenvectors are orthogonal. Consequently, if a hermitian matrix has distinct eigenvalues, then the corresponding eigenvectors are mutually orthogonal.

Let us study a simple, but very instructive example in the following. We will look at the matrix equation:

$$\vec{y} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \vec{x} \Leftrightarrow \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

We call the matrix A:

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

We notice that the definition space is three-dimensional, however rank(A)=1. This means, if you remember, that the image space is one-dimensional. It that true? We can try to set the ***x***-vector to different values:

$$\vec{x} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \Rightarrow \vec{y} = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix}$$

or

$$\vec{x} = \begin{bmatrix} -1 \\ 10 \\ 3 \end{bmatrix} \Rightarrow \vec{y} = \begin{bmatrix} 12 \\ 12 \\ 12 \end{bmatrix}$$

or

$$\vec{x} = \begin{bmatrix} -1 \\ -2 \\ -3 \end{bmatrix} \Rightarrow \vec{y} = \begin{bmatrix} -6 \\ -6 \\ -6 \end{bmatrix}$$

or

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \Rightarrow \vec{y} = \begin{bmatrix} x_1 + x_2 + x_3 \\ x_1 + x_2 + x_3 \\ x_1 + x_2 + x_3 \end{bmatrix}$$

Do you see what happens?. No matter which vector enters the equation, the result will always be on the form:

$\vec{y} = k\begin{bmatrix}1\\1\\1\end{bmatrix}$ where k can be any real number. This is certainly a one-dimensional image space.

Now we ask if we can find vectors, where the image of the vector is a scaled version of itself. This is analogous to solving Eq.30. First we notice that A is a real symmetric matrix and by definition it is also hermitian. Next, we find the eigenvalues from the following equation:

$$\det(A-\lambda I)=\det\begin{vmatrix}1-\lambda & 1 & 1\\ 1 & 1-\lambda & 1\\ 1 & 1 & 1-\lambda\end{vmatrix}=(1-\lambda)^3+1+1-(1-\lambda)-(1-\lambda)-(1-\lambda)=(-\lambda)^2(\lambda-3)=0$$

The eigenvalues are then $\lambda_1=3$, $\lambda_2=0$, $\lambda_3=0$. Note that the algebraic multiplicity of the eigenvalue '0' is two, and the algebraic multiplicity of the eigenvalue '3' is one. We now search for the eigenvectors. If we start with $\lambda_1=3$, then we have to solve the following equation (corresponding to Eq.31):

$$\begin{bmatrix}1-3 & 1 & 1\\ 1 & 1-3 & 1\\ 1 & 1 & 1-3\end{bmatrix}\begin{bmatrix}x_{11}\\x_{21}\\x_{31}\end{bmatrix}=\begin{bmatrix}0\\0\\0\end{bmatrix}\Leftrightarrow\begin{bmatrix}-2 & 1 & 1\\ 1 & -2 & 1\\ 1 & 1 & -2\end{bmatrix}\begin{bmatrix}x_{11}\\x_{21}\\x_{31}\end{bmatrix}=\begin{bmatrix}0\\0\\0\end{bmatrix}$$

(The second subscript of x denotes the first eigenvector). We immediately note that the rank of this matrix is two, telling us that the image space is two-dimensional. Therefore, the dimension of the null-space is 3-2=1, which is also the dimension of the corresponding eigenspace, see Eq.35. Let us solve the equation:

$$\begin{aligned}-2x_{11}+x_{12}+x_{13}&=0\\ +x_{11}-2x_{12}+x_{13}&=0\\ +x_{11}+x_{12}-2x_{13}&=0\end{aligned}$$

Obviously the solution must have the form:

$\begin{bmatrix}x_{11}\\x_{21}\\x_{31}\end{bmatrix}=s\begin{bmatrix}1\\1\\1\end{bmatrix}$ where s belongs to the real numbers. So we can choose the first eigenvector as:

$$\vec{v}_1 = \begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \end{bmatrix} = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

where we have normalized to unit length. What is the geometic meaning of this?. If we take any vector $\vec{x}$, where all entries have the same value, then the output using matrix *A*, is a vector in the same direction. In fact we could foresee this initially if you go back.

Now let us solve Eq.31, corresponding to the two other eigenvalues (which gives the same value in this case):

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_{12} \\ x_{22} \\ x_{32} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \text{ and } \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_{13} \\ x_{23} \\ x_{33} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

We notice (again) the rank is one, and therefore the dimension of the null-space, which also is the dimension of the corresponding eigenspace, is 3-1=2. This means that the geometric multiplicity corresponding to the eigenvalue '0' is two, and we then know that this space is spanned by two linearly independent vectors. Let us solve the first one:

$$\begin{aligned} x_{12} + x_{22} + x_{32} &= 0 \\ x_{12} + x_{22} + x_{32} &= 0 \\ x_{12} + x_{22} + x_{32} &= 0 \end{aligned}$$

Clearly, one solution could be:

$$\vec{v}_2 = \begin{bmatrix} x_{12} \\ x_{22} \\ x_{32} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}$$

Corresponding, to the last eigenvalue (which is also 0), it will not bring us further if we choose the same solution, but clearly we can choose another solution as:

$$\vec{v}_3 = \begin{bmatrix} x_{13} \\ x_{23} \\ x_{33} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

What is the geometric meaning of this finding ? The output when using matrix *A* is the zero vector for any vector within the space spanned by the eigenvectors $\vec{v}_2$ and $\vec{v}_3$.

As predicted for hermitian matrices we note that the three eigenvectors are linearly independent, and that $\vec{v}_1$ is orthogonal to both $\vec{v}_2$ and $\vec{v}_3$. However, $\vec{v}_2$ and $\vec{v}_3$ are linearly independent, but not orthogonal. We can, however, just orthogonalize these vectors. The orthogonal projection of $\vec{v}_2$ on $\vec{v}_3$ is a vector (see Eq.7):

$$\vec{v}_2^{\,n} = QQ^T\vec{v}_2 = \vec{v}_3\vec{v}_3^{\,T}\vec{v}_2$$

An orthogonal vector $\vec{v}_2^{'}$ to $\vec{v}_3$ is then:

$$\vec{v}_2^{'} = \vec{v}_2 - \vec{v}_2^{n} = \vec{v}_2 - QQ^T\vec{v}_2 = \vec{v}_2 - \vec{v}_3\vec{v}_3^T\vec{v}_2 = (I - \vec{v}_3\vec{v}_3^T)\vec{v}_2 =$$

$$\left(\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix} - \frac{1}{2}\begin{bmatrix}1\\0\\-1\end{bmatrix}\begin{bmatrix}1&0&-1\end{bmatrix}\right)\begin{bmatrix}1\\-1\\0\end{bmatrix}\frac{1}{\sqrt{2}} = \left(\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix} - \frac{1}{2}\begin{bmatrix}1&0&-1\\0&0&0\\-1&0&1\end{bmatrix}\right)\begin{bmatrix}1\\-1\\0\end{bmatrix}\frac{1}{\sqrt{2}} =$$

$$\frac{1}{\sqrt{2}}\begin{bmatrix}0.5&0&0.5\\0&1&0\\0.5&0&-0.5\end{bmatrix}\begin{bmatrix}1\\-1\\0\end{bmatrix} = \frac{1}{\sqrt{2}}\begin{bmatrix}0.5\\-1\\0.5\end{bmatrix} = \frac{1}{2\sqrt{2}}\begin{bmatrix}1\\-2\\1\end{bmatrix}$$

We normalize $\vec{v}_2^{'}$ to unit length:

$$\vec{v}_2^{'} = \frac{1}{\sqrt{6}}\begin{bmatrix}1\\-2\\1\end{bmatrix}$$

As a control we can now test Schur's theorem (Eq.46) using $\vec{v}_1$, $\vec{v}_2^{'}$ and $\vec{v}_3$ as the unitary matrix U:

$$U^H AU = \begin{bmatrix}\frac{1}{\sqrt{3}}&\frac{1}{\sqrt{6}}&\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{3}}&\frac{-2}{\sqrt{6}}&0\\\frac{1}{\sqrt{3}}&\frac{1}{\sqrt{6}}&\frac{-1}{\sqrt{2}}\end{bmatrix}^H\begin{bmatrix}1&1&1\\1&1&1\\1&1&1\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{3}}&\frac{1}{\sqrt{6}}&\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{3}}&\frac{-2}{\sqrt{6}}&0\\\frac{1}{\sqrt{3}}&\frac{1}{\sqrt{6}}&\frac{-1}{\sqrt{2}}\end{bmatrix} =$$

$$\begin{bmatrix}\frac{1}{\sqrt{3}}&\frac{1}{\sqrt{3}}&\frac{1}{\sqrt{3}}\\\frac{1}{\sqrt{6}}&\frac{-2}{\sqrt{6}}&\frac{1}{\sqrt{6}}\\\frac{1}{\sqrt{2}}&0&\frac{-1}{\sqrt{2}}\end{bmatrix}\begin{bmatrix}1&1&1\\1&1&1\\1&1&1\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{3}}&\frac{1}{\sqrt{6}}&\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{3}}&\frac{-2}{\sqrt{6}}&0\\\frac{1}{\sqrt{3}}&\frac{1}{\sqrt{6}}&\frac{-1}{\sqrt{2}}\end{bmatrix} = \begin{bmatrix}\frac{1}{\sqrt{3}}&\frac{1}{\sqrt{3}}&\frac{1}{\sqrt{3}}\\\frac{1}{\sqrt{6}}&\frac{-2}{\sqrt{6}}&\frac{1}{\sqrt{6}}\\\frac{1}{\sqrt{2}}&0&\frac{-1}{\sqrt{2}}\end{bmatrix}\begin{bmatrix}\frac{3}{\sqrt{3}}&0&0\\\frac{3}{\sqrt{3}}&0&0\\\frac{3}{\sqrt{3}}&0&0\end{bmatrix} = \begin{bmatrix}3&0&0\\0&0&0\\0&0&0\end{bmatrix}$$

Thus we see that we get as expected:

$$\Lambda = \begin{bmatrix}3&0&0\\0&0&0\\0&0&0\end{bmatrix}$$

As indicated above, if we have used $\vec{v}_1$, $\vec{v}_2$ and $\vec{v}_3$ (which strictly speaking not form a unitary matrix, because $\vec{v}_2$ and $\vec{v}_3$ are not orthogonal) in an equation as Eq.46 we would also get:

$$\begin{bmatrix}\vec{v}_1 & \vec{v}_2 & \vec{v}_3\end{bmatrix}^H \begin{bmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{bmatrix} \begin{bmatrix}\vec{v}_1 & \vec{v}_2 & \vec{v}_3\end{bmatrix} = \begin{bmatrix}\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & 0\\ \frac{1}{\sqrt{3}} & 0 & -\frac{1}{\sqrt{2}}\end{bmatrix}^H \begin{bmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{bmatrix} \begin{bmatrix}\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & 0\\ \frac{1}{\sqrt{3}} & 0 & -\frac{1}{\sqrt{2}}\end{bmatrix} =$$

$$\begin{bmatrix}\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}}\\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0\\ \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}}\end{bmatrix} \begin{bmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{bmatrix} \begin{bmatrix}\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & 0\\ \frac{1}{\sqrt{3}} & 0 & -\frac{1}{\sqrt{2}}\end{bmatrix} = \begin{bmatrix}\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}}\\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0\\ \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}}\end{bmatrix} \begin{bmatrix}\frac{3}{\sqrt{3}} & 0 & 0\\ \frac{3}{\sqrt{3}} & 0 & 0\\ \frac{3}{\sqrt{3}} & 0 & 0\end{bmatrix} =$$

$$\begin{bmatrix}3 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix} = \Lambda$$

Normal matrices

We define a matrix as normal if and only if:

$$A^H A = AA^H \quad \text{Eq.47}$$

A normal matrix A will obey Eq.46, even in situations where A is not hermitian. This can be proven in the following way. We know that Schur's theorem is valid for any matix $A_{N,N}$:

$$U^H AU = T \Leftrightarrow U^H A^H U = T^H$$

If we evaluate the following two products we get:

$$\left(U^H AU\right)\left(U^H A^H U\right) = TT^H \Leftrightarrow U^H AA^H U = TT^H$$

$$\left(U^H A^H U\right)\left(U^H AU\right) = T^H T \Leftrightarrow U^H A^H AU = T^H T$$

Eq.47 implies that $T^H T = TT^H$ and this is only possible if T is a diagonal matrix. On the other hand, if the matrix A obeys Eq.46, i.e.:

$$U^H AU = \Lambda \Leftrightarrow U^H A^H U = \Lambda^H$$

then we can evaluate the following two products and we get:

$$\left(U^H AU\right)\left(U^H A^H U\right) = \Lambda\Lambda^H \Leftrightarrow U^H AA^H U = \Lambda\Lambda^H$$

$$\left(U^H A^H U\right)\left(U^H AU\right) = \Lambda^H\Lambda \Leftrightarrow U^H A^H AU = \Lambda^H\Lambda$$

Because $\Lambda\Lambda^H$ must be equal to $\Lambda^H\Lambda$, it follows that $AA^H = A^H A$.

Positive definite forms

We define a quadratic form 'q' of vector $\vec{x}$ as:

$$q(\vec{x}) = \vec{x}^T A\vec{x} = \left\langle \vec{x}, A\vec{x} \right\rangle = \left\langle A\vec{x}, \vec{x} \right\rangle \quad \text{Eq.48}$$

where *A* is a N x N matrix and vector ***x*** is a column vector consisting of N (real) variables. The form or function 'q' is a real value function (results in a scalar value), and can be viewed as a real-valued quadratic function in N variables. Let $\vec{x} = \begin{bmatrix} x & y \end{bmatrix}^T$ and as an example:

$$q(x,y) = \begin{bmatrix} x & y \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & -2 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x & y \end{bmatrix}\begin{bmatrix} x+y \\ x-2y \end{bmatrix} = x^2 + xy + xy - 2y^2 = x^2 + 2xy - 2y^2$$

or:

$$q(x,y) = \begin{bmatrix} x & y \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x & y \end{bmatrix}\begin{bmatrix} x \\ -2y \end{bmatrix} = x^2 - 2y^2$$

or:

$$q(x,y) = \begin{bmatrix} x & y \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x & y \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = x^2 + y^2$$

A form 'q' and the corresponding matrix *A* is said to be *positive definite* if $q(\vec{x}) > 0$ for all $\vec{x} \neq \vec{0}$. A form 'q' and the corresponding matrix *A* is called *positive semidefinite* if $q(\vec{x}) \geq 0$ for all $\vec{x}$. (So a positive definite form is always positive semidefinite). If $q(\vec{x}) \geq 0$ for all $\vec{x}$, and $q(\vec{x}_0) = 0$ where $\vec{x}_0 \neq \vec{0}$, the form is (of course) positive semidefinite and singular, because the matrix *A* is singular. Why? In this case, there exist an $\vec{x}_0 \neq \vec{0}$, so $A\vec{x}_0 = \vec{0}$. This is equivalent with the fact that the columns (or rows) of matrix *A* are linearly dependent. Negative definite and negative semidefinite are defined in a similar way. For completeness, the form 'q' is said to be indefinite if there exists $\vec{x}_1$ and $\vec{x}_2$ so $q\left(\vec{x}_1\right) > \vec{0}$ and $q\left(\vec{x}_2\right) < \vec{0}$.

If *A* is a real symmetric matrix then there exists a real orthogonal set of eigenvectors such that:

$$Q^{-1}AQ = \Lambda \Leftrightarrow Q^T AQ = \Lambda \quad \text{Eq.49}$$

because $Q^T Q = QQ^T = I \Rightarrow Q^T = Q^{-1}$. The quadratic form 'q' can then be evaluated by use of the corresponding eigenvalues. If we define $\vec{x} = Q\vec{y}$ then we get:

$$q(\vec{x}) = \vec{x}^T A\vec{x} = (Q\vec{y})^T A(Q\vec{y}) = \vec{y}Q^T AQ\vec{y} = \vec{y}^T \Lambda\vec{y} = \lambda_1 y_1^2 + \lambda_2 y_2^2 + ... + \lambda_N y_N^2 \quad \text{Eq.50}$$

In this case, matrix *A* is positive definite if and only if all eigenvalues are positive. Further, *A* is positive semidefinite if and only if the eigenvalues of *A* are non-negative. Finally, *A* is positive semidefinite and singular if and only if the eigenvalues of *A* are non-negative and at least one eigenvalue is zero.

If the matrix *A* is positive semidefinite, then there exists a unique positive semidefinite matrix $A^{1/2}$ such that $(A^{1/2})^2=A$. Part of the proof is alluded to in the following. Since *A* is real symmetric, we have Eq.49:

$$Q^T AQ = \Lambda \Leftrightarrow Q\Lambda Q^T = A$$

and we define $\Lambda^{1/2}$ as:

$$\Lambda^{1/2} = \begin{bmatrix} \lambda_1^{1/2} & 0 & \cdots & 0 \\ 0 & \lambda_2^{1/2} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \lambda_N^{1/2} \end{bmatrix}$$

We set $A^{1/2} = Q\Lambda^{1/2}Q^T$ and evaluate:

$$(A^{1/2})^2 = \left(Q\Lambda^{1/2}Q^T\right)\left(Q\Lambda^{1/2}Q^T\right) = Q\Lambda^{1/2}Q^T Q\Lambda^{1/2}Q^T = Q\Lambda Q^T = A$$

Note that because *A* is positive semidefinite, we know that the eigenvalues are nonnegative and, therefore the square root of the each eigenvalue exists. It follows that $A^{1/2}$ is also positive semidefinite and:

$$A^{1/2} = Q\Lambda^{1/2}Q^T \Leftrightarrow Q^T A^{1/2} Q = \Lambda^{1/2} \quad \text{Eq.51}$$

Singular values and singular vectors

Let *A* be a real or complex N x M matrix with rank(*A*) = r. We look at the matrix AA^H and A^HA, and make the crucial discovery that both are hermitian!:

$$(A^H A)^H = A^H A \quad \text{and} \quad (AA^H)^H = AA^H$$

As an example, evaluate:

$$A=\begin{bmatrix}1 & 1\\ 2 & 0\\ 1 & 1\end{bmatrix} \quad \text{then} \quad A^H A=\begin{bmatrix}1 & 2 & 1\\ 1 & 0 & 1\end{bmatrix}\begin{bmatrix}1 & 1\\ 2 & 0\\ 1 & 1\end{bmatrix}=\begin{bmatrix}6 & 2\\ 2 & 2\end{bmatrix},$$

$$AA^H=\begin{bmatrix}1 & 1\\ 2 & 0\\ 1 & 1\end{bmatrix}\begin{bmatrix}1 & 2 & 1\\ 1 & 0 & 1\end{bmatrix}=\begin{bmatrix}2 & 2 & 2\\ 2 & 4 & 2\\ 2 & 2 & 2\end{bmatrix}$$

Note that A has full column rank, rank(A)=2, and also that AA^H and A^HA are of rank 2. Another example with complex entries:

$$A=\begin{bmatrix}1-2i & 3+i\\ 2+i & 1\end{bmatrix} \quad \text{then} \quad A^H A=\begin{bmatrix}1+2i & 2-i\\ 3-i & 1\end{bmatrix}\begin{bmatrix}1-2i & 3+i\\ 2+i & 1\end{bmatrix}=\begin{bmatrix}10 & 3+6i\\ 3-6i & 11\end{bmatrix},$$

$$AA^H=\begin{bmatrix}1-2i & 3+i\\ 2+i & 1\end{bmatrix}\begin{bmatrix}1+2i & 2-i\\ 3-i & 1\end{bmatrix}=\begin{bmatrix}15 & 3-4i\\ 3+4i & 6\end{bmatrix}$$

It is easily seen that $(A^HA)^H=A^HA$ and $(AA^H)^H=AA^H$.

AA^H and A^HA are positive semidefinite because:

$$\vec{x}^H AA^H\vec{x}=\left\langle A^H\vec{x}, A^H\vec{x}\right\rangle=\left\|A^H\vec{x}\right\|^2\geq 0$$

$$\vec{x}^H A^H A\vec{x}=\left\langle A\vec{x}, A\vec{x}\right\rangle=\left\|A\vec{x}\right\|^2\geq 0$$

The implication is that AA^H and A^HA have nonnegative eigenvalues. Two additional properties related to AA^H and A^HA can be shown. First, if rank(A)= r, then the rank of AA^H and A^HA is also 'r'. As an example, suppose A is of size N x M where M<N, and that the rank(A) = M (as in the example above with real entries). The matrix A^HA are of size M x M and will have full rank (2 in the example). The matrix AA^H is of size N x N, but still has a rank of 'M' (2 in the example above). Thus, it is clear that AA^H will have N-M zero eigenvalues (in the example above one eigenvalue is zero). Second, it can be shown that AA^H and A^HA share the same positive eigenvalues and have equal geometric multiplicity. This is indicated in the following.

The equation:

$$A^H A\vec{v}=\sigma^2\vec{v} \quad \text{Eq.52}$$

shows that vector $\boldsymbol{v}$ is an eigenvector of A^HA and that σ^2 is an eigenvalue of A^HA. Because we know that A^HA is positive semidefinite, and, therefore has only nonnegative eigenvalues, no loss of generalisation is introduced by writing the eigenvalue as a squared value. We now define a new vector '$\boldsymbol{u}$' by:

$$\sigma\vec{u}\equiv A\vec{v} \quad \text{Eq.53}$$

If we combine Eq.52 with Eq.53 we get:

$$A^H\vec{u} = \sigma\vec{v} \quad \text{Eq.54}$$

Now multiply with A on both right and left side and use Eq. 53 and obtain:

$$AA^H\vec{u} = A\sigma\vec{v} = \sigma A\vec{v} = \sigma^2\vec{u} \Leftrightarrow AA^H\vec{u} = \sigma^2\vec{u} \quad \text{Eq.55}$$

Indeed we see that AA^H and A^HA share the same nonnegative eigenvalues. If $\sigma > 0$, then

$$\vec{u} = \frac{1}{\sigma}A\vec{v} \quad \text{and} \quad \vec{v} = \frac{1}{\sigma}A^H\vec{u} \quad \text{Eq.56}$$

We conclude that the geometric multiplicity for a distinct positive eigenvalue, must be the same for AA^H and A^HA. If an eigenvalue $\sigma^2 = 0$ for A^HA, then $A\vec{v} = \vec{0}$. If an eigenvalue $\sigma^2 = 0$ for AA^H then $A^H\vec{u} = \vec{0}$.

Since A^HA is positive semidefinite, $(A^HA)^{1/2}$ exists and its eigenvalue are the nonnegative square root of the eigenvalues of A^HA. We call the positive eigenvalues of $(A^HA)^{1/2}$ the *singular values* of A. That is if σ is a singular value of A, it is the positive eigenvalue of $(A^HA)^{1/2}$ and σ^2 is an eigenvalue of A^HA and AA^H. Importantly, it can be readily shown that rank(A) is the number of singular values of A. If the algebraic multiplicity of a distinct singular value σ is repeated 'k' times, then there exists 'k' linearly independent eigenvectors $\left[\vec{v}_1 \quad \vec{v}_2 \quad \cdots \quad \vec{v}_k\right]$ of A^HA corresponding to the distinct eigenvalue σ^2. We call these vectors *right singular vectors* of A corresponding to the singular value, and $(\sigma, \vec{v}_i)$ a right singular pair. Likewise, there exists 'k' linearly independent eigenvectors $\left[\vec{u}_1 \quad \vec{u}_2 \quad \cdots \quad \vec{u}_k\right]$ of AA^H corresponding to the distinct eigenvalue σ^2. We call these vectors *left singular vectors* of A corresponding to the singular value, and $(\sigma, \vec{u}_i)$ a left singular pair. As mentioned previously, any linear combination of a set of eigenvectors corresponding to a distinct eigenvalue, is also an eigenvector. Therefore the set $\left[\vec{v}_1 \quad \vec{v}_2 \quad \cdots \quad \vec{v}_k\right]$ can be orthogonalized and normalized, such that all vectors are mutual orthogonal. This implies that also $\left[\vec{u}_1 \quad \vec{u}_2 \quad \cdots \quad \vec{u}_k\right]$ is mutually orthogonal and normalized because:

$$\sigma^2\delta_{ij} = \sigma^2\left\langle\vec{v}_i, \vec{v}_j\right\rangle = \left\langle\sigma\vec{v}_i, \sigma\vec{v}_j\right\rangle = \left\langle A^H\vec{u}_i, A^H\vec{u}_j\right\rangle = \left\langle\vec{u}_i, AA^H\vec{u}_j\right\rangle = \left\langle\vec{u}_i, \sigma^2\vec{u}_j\right\rangle = \sigma^2\left\langle\vec{u}_i, \vec{u}_j\right\rangle$$

where $\delta_{ij} = 0$ for $i \neq j$ and $\delta_{ij} = 1$ for $i = j$.

Singular-value decomposition

We are now in a position to prove the singular value decomposition of a matrix A. Let A be of size N x M, and, to keep things general let rank(A) = r ≤ M ≤ N or rank(A) = r ≤ N ≤ M. There exist unitary matrices $U_{N,N}$ and $V_{M,M}$ such that:

$$A = USV^H \quad \text{Eq.57}$$

where S is an N x M matrix with the singular values of A displayed along its diagonal and zeros elsewhere. In Eq.57 we can easily isolate S:

$$A = USV^H \Leftrightarrow U^HA = U^HUSV^H \Leftrightarrow U^HAV = U^HUSV^HV \Leftrightarrow U^HAV = S$$

So

$$U^H AV = S \quad \text{Eq.58}$$

That such an equation is possible is almost indicated by looking at Eq.53. $A^H A$ is hermitian and has the corresponding eigenvectors $V = \begin{bmatrix} \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_r & \vec{v}_{r+1} & \cdots & \vec{v}_M \end{bmatrix}$ organised so that these are mutualy orthogonal and have a corresponding set of eigenvalues, where $\sigma^2_{1\to r} > 0$ and $\sigma^2_{r+1\to M} = 0$, referring to each eigenvalue as many times as the individual algebraic multiplicity indicates. Similar, AA^H is hermitian and has the corresponding eigenvectors $U = \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_r & \vec{u}_{r+1} & \cdots & \vec{u}_N \end{bmatrix}$, organised so these are mutualy orthogonal, and have a corresponding set of eigenvalues, where $\sigma^2_{1\to r} > 0$ and $\sigma^2_{r+1\to N} = 0$. We now evaluate:

$$U^H AV = U^H A\begin{bmatrix} \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_r & \vec{v}_{r+1} & \cdots & \vec{v}_M \end{bmatrix} =$$

$$U^H \begin{bmatrix} A\vec{v}_1 & A\vec{v}_2 & \cdots & A\vec{v}_r & A\vec{v}_{r+1} & \cdots & A\vec{v}_M \end{bmatrix} =$$

$$U^H \begin{bmatrix} \sigma_1\vec{u}_1 & \sigma_2\vec{u}_2 & \cdots & \sigma_r\vec{u}_r & \vec{0}_{r+1} & \cdots & \vec{0}_M \end{bmatrix} =$$

$$\begin{bmatrix} \sigma_1 U^H\vec{u}_1 & \sigma_2 U^H\vec{u}_2 & \cdots & \sigma_r U^H\vec{u}_r & \vec{0}_{r+1} & \cdots & \vec{0}_M \end{bmatrix} =$$

$$\begin{bmatrix} \sigma_1\vec{q}_1 & \sigma_2\vec{q}_2 & \ldots & \sigma_r\vec{q}_r & \vec{0}_{r+1} & \cdots & \vec{0}_M \end{bmatrix} = S_{N\,x\,M}$$

where $\vec{q}_1 = \begin{bmatrix} 1 & 0 & \cdots & 0_N \end{bmatrix}^T$ and $\vec{q}_2 = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0_N \end{bmatrix}^T$ etc.

Of course if A itself is hermitian ($A^H = A$), then $A^H A = A^2 = AA^H$, and $A^H A$ and AA^H have the same set of eigenvectors, i.e. $V=U$. Furthermore, the singular values of A is also the eigenvalues of A, so Eq.58 becomes equivalent with Eq.46.

The pseudo-inverse

We are now able to introduce the *generalised inverse*, also called a pseudo-inverse of the matrix A regardless if the matrix is singular, invertible, square or rectangular. It can be shown that for a matrix $A_{N,M}$ there exists a unique generalised inversed matrix, A^-, such that:

1) $AA^-A = A$
2) $A^-AA^- = A^-$
3) $\left(A^-A\right)^H = A^-A$ Eq.59
4) $\left(AA^-\right)^H = AA^-$

Generally, $AA^{-} \neq I$. Only if $A^{-} = A^{-1}$, then $AA^{-} = AA^{-1} = I$.
It can easily be shown that S:

$$S = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix}_{N\,x\,M}$$

and S^{-} defined as:

$$S^{-} = \begin{bmatrix} 1/\sigma_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sigma_r & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix}_{M\,x\,N}$$

satisfy all parts of Eq.59. Again, note that $SS^{-} \neq I$, but $SS^{-}S = S$. We can now construct the pseudo-inverse of A by:

$$A^{-} = VS^{-}U^{H} \quad \text{Eq.60}$$

and show that this definition also satisfy Eq.59. For example we can prove part 1):

$$AA^{-}A = \left(USV^{H}\right)\left(VS^{-}U^{H}\right)\left(USV^{H}\right) = USS^{-}SV^{H} = USV^{H} = A$$

and part 3):

$$A^{-} = VS^{-}U^{H} \Leftrightarrow A^{-}A = (VS^{-}U^{H})(USV^{H}) = VS^{-}SV^{H}$$

and

$$(A^{-}A)^{H} = (VS^{-}SV^{H})^{H} = V(S^{-}S)^{H}V^{H} = VS^{-}SV^{H}$$

$$\Rightarrow (A^{-}A)^{H} = A^{-}A$$

As mentioned A^{-} is called the generalised inverse, or the p-inverse, as a shorthand for the pseudo-inverse, or Moore-Penrose inverse of A. This is the inverse one gets when using MATLAB. Note that if A is invertible, i.e. $A^{-} = A^{-1}$, then this A^{-1} also satisfies all parts of Eq.59. Also note that if A has full column rank, then $A^{-} = (A^{H}A)^{-1}A^{H}$, and again all parts of Eq. 59 are satisfied, e.g. part 2):

$$A^- = A^- AA^- = ((A^H A)^{-1} A^H) A((A^H A)^{-1} A^H) = (A^H A)^{-1}(A^H A)(A^H A)^{-1} A^H = (A^H A)^{-1} A^H$$

Recall our previous talk about the orthogonal projector P. We saw P was symmetric ($P^T = P$) and $P^2 = P$. As mentioned above, all type of matrices obey Eq.57 and Eq.60, but in addition P is hermitian so we get:

$$P = USV^H = U\Lambda U^H$$

and

$$P^- = VS^- U^H = U\Lambda^- U^H$$

Now, what can be said about the eigenvalues of an orthogonal projector P ? Imagine a vector, which is already within the vector subspace, which P projects into. Clearly, the projector has no effect, that is $P\vec{\beta} = \vec{\beta}$. Also, the corresponding eigenvalue is 1 and the beta-vector constitutes an eigenvector. What will happen with a vector orthogonal to the vector subspace P projects into ? Clearly, the projection will be the zero-vector, i.e. $P\vec{\beta} = 0\vec{\beta}$. Thus, the eigenvalue will be zero, and the beta-vector will also be an eigenvector. All other vectors will be projected orthogonal into the vector subspace, and will not constitute an eigenvector. We can therefore conclude, that the eigenvalues of an orthogonal projector is either 1 or 0. This obviously indicates that $\Lambda = \Lambda^-$, and therefore we get the peculiar result that the pseudo-inverse of an orthogonal projector is the projector itself:

$$P = P^- \quad \text{Eq.61}$$

It is then easy to verify all parts of Eq.57. Recall the number of eigenvalues different from zero is equal to the rank of the matrix. So, in this situation, the sum of the eigenvalues correspond to rank(P), which again is the dimension of the subspace P projects into. Above we showed that if a matrix A, has full column rank, then $A^- = (A^H A)^{-1} A^H$. Also recall that an orthogonal projection (given by matrix P) into the space spanned by the column vectors of A, was $A(A^T A)^{-1} A^T$. We then see that:

$$P = AA^- \quad \text{Eq.62}$$

and that P is a projector into the column space of A. Quit generally, P where A^- is define as above, and where A^- satisfy the condition corresponding to Eq.59, is a projector into the column space of A, because P satisfy the two properties of an orthogonal projector: $P^H = P$ and $P^2 = P$. For example:

$$P^2 = (AA^-)(AA^-) = A(A^- AA^-) = AA^- = P$$

and an equation $\vec{y} = A\vec{\beta}$ with a solution $\vec{\beta}^{opt} = A^- \vec{y}$ shows that $\vec{y}_n$ is indeed a projection of $\vec{y}$ into the column space of A, because $\vec{y}_n = A\vec{\beta}^{opt} = A(A^- \vec{y})$.

To sum up, an equation $\vec{y} = X\vec{\beta}$ is said to be consistent, if it has one or more solutions. If we have an equation as $\vec{y} = X_{N,N}\vec{\beta}$ where X is invertible, then the unique solution is given by

$\vec{\beta} = X^{-1}\vec{y}$. The determinant of X will be different from zero, and X will have full rank, and the columns (and rows) of X will be linearly independent. If the determinant is zero, then X is not invertible, and there will be either no solution or more than one solution, i.e. no unique solution.

Generally, an equation $\vec{y} = X_{N,M}\vec{\beta}$ is consistent if and only if $rank(X) = rank(X,\vec{y})$, that is the rank does not change if we include the ***y***-vector. This is again, equivalent to saying that the ***y***-vector is within the space spanned by X. In order to differentiate between one or more solutions of the equation $\vec{y} = X_{N,M}\vec{\beta}$, we can state that if the equation is consistent, then it has a unique solution if and only if X has full column rank.

Finally, if $rank(X) < rank(X,\vec{y})$, then the ***y***-vector is not within the space spanned by X, and there is no solution, and the equation is inconsistent. However, we can always find an optimal beta-vector given by $\vec{\beta}^{opt} = X^{-}\vec{y}$, corresponding to the equation $\vec{y} = X_{N,M}\vec{\beta}$, which has a minimum norm, defined as $|\vec{y} - X_{NxM}\vec{\beta}^{opt}|$, when using a pseudo-inverse as defined above. This is evident if the equation $\vec{y} = X_{N,M}\vec{\beta}$ is consistent because then $|\vec{y} - X_{NxM}\vec{\beta}^{opt}| = 0$. If the equation is inconsistent, but X has full column rank, then the optimal solution is given by $\vec{\beta}^{opt} = X^{-}\vec{y} = (X^{H}X)^{-1}X^{H}\vec{y}$. We have previously seen that this also resulted in a minimum norm, that is a minimal distance between the ***y***-vector and the solution plane. As mentioned, even in the situations where X does not have full column rank, the defined pseudo-inverse results in a minimum norm solution.

We have previously looked at the equation:

$$\begin{bmatrix}1\\4\\2\end{bmatrix} = \begin{bmatrix}1 & 1\\0 & 1\\1 & 1\end{bmatrix}\begin{bmatrix}\beta_1\\\beta_2\end{bmatrix}$$

The equation is inconsistent, but the matrix has full column rank. We saw that an orthogonal projector projecting into the subspace spanned by the column vectors of the design matrix was:

$$P = \begin{bmatrix}0.5 & 0 & 0.5\\0 & 1 & 0\\0.5 & 0 & 0.5\end{bmatrix}$$

P is a real symmetric matrix, and the eigenvalues are found from:

$$\det\begin{vmatrix}0.5-\lambda & 0 & 0.5\\0 & 1-\lambda & 0\\0.5 & 0 & 0.5-\lambda\end{vmatrix} = (1-\lambda)(0.5-\lambda)^2 - 0.25(1-\lambda) = 0 \Leftrightarrow$$

$$-\lambda(\lambda-1)^2 = 0$$

Thus the eigenvalues are $\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 0$, respectively. The algebraic multiplicity of the eigenvalue '1' is two, and the dimension corresponding to this eigenspace is two (see eq.35). We can the construct the following matrices which all are identical:

$$S = \Lambda = S^{-} = \Lambda^{-} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

The first eigenvector is found from:

$$\begin{bmatrix} -0.5 & 0 & 0.5 \\ 0 & 0 & 0 \\ 0.5 & 0 & -0.5 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Leftrightarrow \begin{bmatrix} x \\ y \\ z \end{bmatrix} = t \begin{bmatrix} 1 \\ s \\ 1 \end{bmatrix} \quad \text{where} \quad t, s \in \Re$$

We see that this eigenvector is exactly within the subspace spanned by $[1\ 0\ 1]^T$ and $[1\ 1\ 1]^T$. In fact both are representative of eigenvectors corresponding to the eigenvalue '1'.
Also $[1\ 0\ 1]^T$ and $[1\ 1\ 1]^T$ are linearly independent (but not orthogonal). It is easy to verify that P does not change these eigenvectors:

$$P \begin{bmatrix} t \\ ts \\ t \end{bmatrix} = \begin{bmatrix} t \\ ts \\ t \end{bmatrix}$$

The eigenvector for $\lambda = 0$ is found from:

$$\begin{bmatrix} 0.5 & 0 & 0.5 \\ 0 & 1 & 0 \\ 0.5 & 0 & 0.5 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Leftrightarrow \begin{bmatrix} x \\ y \\ z \end{bmatrix} = t \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \quad \text{where} \quad t \in \Re$$

The dimension of this eigenspace is one (see eq. 35). It is easy to see that this eigenvector will always be orthogonal to the subspace spanned by $[1\ 0\ 1]^T$ and $[1\ 1\ 1]^T$, and also that:

$$P \begin{bmatrix} t \\ 0 \\ -t \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

If our goal is to find the most optimal solution in a least square sense of the equation:

$$\begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \quad \text{setting } X = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}$$

we can use the singular-value decomposition of X, in order to find this solution. We first calculate XX^H:

$$XX^H = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 2 \\ 1 & 1 & 1 \\ 2 & 1 & 2 \end{bmatrix}$$

and the eigenvalues of XX^H are calculated from:

$$\det \begin{vmatrix} 2-\lambda & 1 & 2 \\ 1 & 1-\lambda & 1 \\ 2 & 1 & 2-\lambda \end{vmatrix} = 0 \Leftrightarrow \lambda^3 - 5\lambda^2 + 2\lambda = 0 \Rightarrow \lambda_1 = 4.5616, \lambda_2 = 0.4384, \lambda_3 = 0$$

and the singular values of X is $\sigma_1 = \sqrt{\lambda_1} = 2.1358, \sigma_2 = \sqrt{\lambda_2} = 0.6621$. The corresponding left singular vectors are:

$$\vec{u}_1 = \begin{bmatrix} 0.6572 & 0.3690 & 0.6572 \end{bmatrix}^T$$
$$\vec{u}_2 = \begin{bmatrix} 0.2610 & -0.9294 & 0.2610 \end{bmatrix}^T$$
$$\vec{u}_3 = \begin{bmatrix} 0.7071 & 0 & -0.7071 \end{bmatrix}^T$$

Note that corresponding to the three distinct eigenvalues we have three orthogonal eigenvectors. We have used the MATLAB function ‘eig’ in order to calculate these eigenvalues and eigenvectors. MATLAB automatically returns the normalised eigenvectors normalised to unit length. We can now construct the U matrix:

$$U = \begin{bmatrix} \frac{\vec{u}_1}{|\vec{u}_1|} & \frac{\vec{u}_2}{|\vec{u}_2|} & \frac{\vec{u}_3}{|\vec{u}_3|} \end{bmatrix} = \begin{bmatrix} 0.6572 & 0.2610 & 0.7071 \\ 0.3690 & -0.9294 & 0 \\ 0.6572 & 0.2610 & -0.7071 \end{bmatrix}$$

We now calculate the right singular vectors corresponding to X^HX from eq.54:

$$\vec{v}_1 = \frac{1}{\sigma_1} X^H \vec{u}_1 = \frac{1}{2.1358} \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}^H \begin{bmatrix} 0.6572 \\ 0.3690 \\ 0.6572 \end{bmatrix} = \begin{bmatrix} 0.6154 \\ 0.7882 \end{bmatrix}$$

and

$$\vec{v}_2 = \frac{1}{\sigma_2} X^H \vec{u}_2 = \frac{1}{0.6621} \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}^H \begin{bmatrix} 0.2610 \\ -0.9294 \\ 0.2610 \end{bmatrix} = \begin{bmatrix} 0.7884 \\ -0.6153 \end{bmatrix}$$

and construct V:

$$V = \begin{bmatrix} \frac{\vec{v}_1}{|\vec{v}_1|} & \frac{\vec{v}_2}{|\vec{v}_2|} \end{bmatrix} = \begin{bmatrix} 0.6154 & 0.7884 \\ 0.7884 & -0.6154 \end{bmatrix}$$

We can now verify the singular value decomposition of the matrix X. According to eq.57:

$$X = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} = USV^H = \begin{bmatrix} 0.6572 & 0.2610 & 0.7071 \\ 0.3690 & -0.9294 & 0 \\ 0.6572 & 0.2610 & -0.7071 \end{bmatrix} \begin{bmatrix} 2.1358 & 0 \\ 0 & 0.6621 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0.6154 & 0.7884 \\ 0.7884 & -0.6154 \end{bmatrix}^H =$$

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}$$

In MATLAB one can perform the singular value decomposition by using the function ‘svd’. We can use the singular value decomposition in order to find the most optimal value of the beta-vector from the perspective of a minimum norm. We get according to eq.60:

$$\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = X^- \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix} = VS^- U^H \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix} =$$

$$\begin{bmatrix} 0.6154 & 0.7884 \\ 0.7884 & -0.6154 \end{bmatrix} \begin{bmatrix} \frac{1}{2.1358} & 0 & 0 \\ 0 & \frac{1}{0.6621} & 0 \end{bmatrix} \begin{bmatrix} 0.6572 & 0.2610 & 0.7071 \\ 0.3690 & -0.9294 & 0 \\ 0.6572 & 0.2610 & -0.7071 \end{bmatrix}^H \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix} =$$

$$\begin{bmatrix} 0.5 & -1 & 0.5 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix} = \begin{bmatrix} -2.5 \\ 4 \end{bmatrix}$$

We have previously obtained this result, based on the fact that the matrix X had full column rank. However, the procedure described above works also in situations where X does not have full column rank, and from that perspective the concept of pseudo-inverse is a more general procedure. In MATLAB the pseudo-inverse can be calculated by using the function ‘pinv’, which is based on the singular value decomposition.

The rank of a large sized matrix

Normally, it is a relatively easy job to find the rank of a matrix. The rank of a matrix is equal to the number of non-zero singular values. However, for large matrices, e.g. 100 x 100, it can be difficult to decide the rank because often some of the columns or rows are more or less similar resulting in some of the singular values being very small. One can judge the rank by calculating the singular values of the matrix and ordering these along the diagonal in decreasing order. Then, it is possible to choose a threshold, e.g. 0.001, and below this threshold we set the singular values to zero. So, in fact we choose the rank of the matrix based on the singular values. Interestingly, we notice that if we allow very small singular values, then these singular values will turn into very high values if we want to find a pseudo-inverse because of the use of the reciprocal singular values in this situation. This means that the result may give spurious values and become unstable.

Part II

Dependency in fMRI

So far we have ignored a problem related to our observed values, that is the element of the y vector. As stated in the beginning of section I, a basic requirement for linear regression is that our observed values are independent, however, it is sufficient to require that the elements of $\vec{y}$ are uncorrelated. If the elements of $\vec{y}$ are correlated, it means that we have some additional systematic variation, not accounted for by the model that is systematic variation not explained by the explanatory variables. This will be disclosed if we look at the distribution of the error term, $\vec{e}$, which should have a Gaussian distribution if our model has captured all variation. If our model is incomplete, we will notice a non-Gaussian distribution. Whether or not this requirement is satisfied, can be tested using the Durbin-Watson statistic. It turns out, that our model is often not complete in BOLD imaging. The reasons are related to respiration, cardiac pulsation and likely many others causes. It has been shown that if not taken into proper consideration it will definitively reduce the validity of our statistic tests. So taking 'serial time correlation' into account is important. Strangely, this kind of dependency may occur more often then expected in medical statistic. A god example is treatment responses, measured several times over a time periods in groups of patients. The response of patient A at time 1, will be correlated with the response of patient A at time 2.

In order to understand how we can incorporate this effect, we will need some basic statistic related firstly to a one-dimensional probability distribution, followed by a multi-dimensional probability distribution.

Let y be a one-dimensional stochastic variable with a mean of $\bar{y}$ and a standard variation of σ_y, having a Gaussian distribution: $y \in \mathrm{N}(\bar{y},\sigma_y)$, the probability density distribution is:

$$p(y) = \frac{1}{\sigma_y \sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y - \bar{y}}{\sigma_y} \right)^2 \right) \wedge -\infty < y < \infty \quad \text{Eq.63}$$

This kind of function is shown in Figure 16, for $\bar{y} = 6$, and $\sigma_y = 2$.

Note that p(y) is a density, i.e. the probability of getting the value of x per unit of x. If we asked, what is the probability of p(y= a), where a is a specific value, that answer is zero. Furthermore, the probability of:

$$y \le a \quad \text{is} \quad P(y \le a) = \int_{-\infty}^{a} p(y)dy$$

and probability of:

$$y \ge a \quad \text{is} \quad P(y \ge a) = 1 - \int_{-\infty}^{a} p(y)dy$$

and the probability of:

$$b \le y \le a \quad \text{is} \quad P(b \le y \le a) = \int_{b}^{a} p(y)dy$$

Note, that if the limits a and b approach each other, eventually defining a point, the probability becomes zero.

Selected topics

Determinant
Durbin-Watson statistic
Rank
Subspace