Syllabus

Parameter estimation and detection. Applications to Brain Imaging.

Dr. Jan C. de Munck
Dept PMT
VUmc
Tel.: 020-4440169
Email: jc.munck@vumc.nl

1 Introduction

The problem of estimating parameters from measured data appears in many branches of science, but also in day to day life. The reason is that it often happens that the quantity of interest cannot be accessed directly, but use is made of measurements on which that quantity is dependent. For instance, temperature can be determined by measuring the height of a column of mercury. Another example is a nowadays instrument to determine the speed of a bicycle consisting of a clock and a system that counts the number of rotations of the bicycle wheel. By assuming the circumference of the wheel is known, the number of rotations can be related to distance and in combination to the time measurements this yields the speed of the bike. Although these are simple examples, they already demonstrate that the extraction of the parameter of interest from the raw data depend on certain assumptions.

A field of research where parameter estimation problems abound are the medical and biological sciences, and particularly in brain imaging. Here the central goal is to obtain structural and functional information from the human brain by recording signals from the brain with techniques such as MRI, PET, EEG and MEG. Usually, the raw signals themselves do not immediately provide the physiological information of interest, but they are in some way dependent on parameters of interest. The question is how to extract the parameters of interest from the raw data in a meaningful, robust and reliable way, thereby accounting for both noises and artefacts.

The central theme of this course is to show how the parameter estimation problem can be addressed in various problems related to brain imaging. Different approaches to the same problem, may yield different results, and to a large extend these differences can be related to the underlying assumptions. It is attempted to discuss and formulate these assumptions as precisely as possible, in order to facilitate the interpretation of the results and so to avoid ambiguities. Although most examples presented in this course originate from brain imaging, the applied theory is applicable in far more general situations.

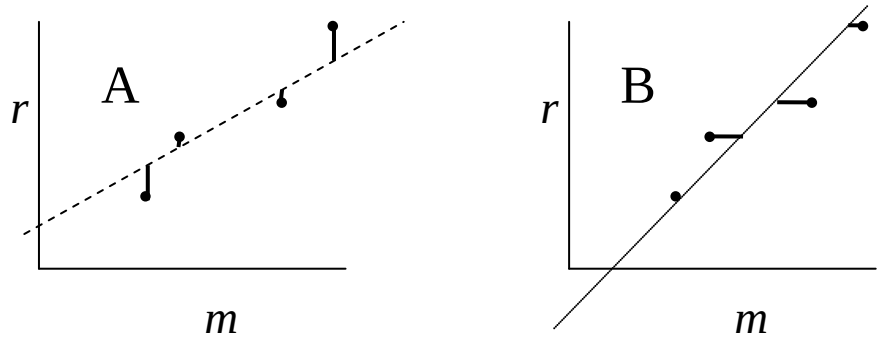

Figure 1.1 *Line fitting can be done using different methods, leading to different estimated parameters. In panel A the line vertical distances are minimized leading, in this example, to a positive line offset whereas in panel B horizontal differences are minimized resulting in a negative offset.*

1.1 Line fitting as a parameter estimation problem

One of the simplest examples of a parameter estimation problem is to fit a line through a set of measurements (m_n, r_n), n=0,…,N-1, see figure 1.1. There are different methods to address this problem. One could fit the best line "by eye", one could minimize the vertical distances between the measurement points and the unknown line, or minimize the horizontal distances. Moreover, one could give higher weights to the "good data points" and lower weights to the points of which it is a priori known that they have poor quality. Moreover, one could minimize a combination of

horizontal and vertical differences. Each choice will lead to a different line estimate, i.e. a different estimate of the line's offset and slope. In situations where the results strongly dependent on the method and when important (clinical) decisions are made, dependent on the estimated parameters, the method of parameter estimation should be selected with care. The main topic of this lecture is to provide insight into the theoretical considerations that can help with the selection of an optimal choice.
The methodology of parameter estimation and signal processing techniques can be roughly divided into *data driven* and *model driven* variants. These two approaches of the parameter estimation problem have different goals, merits and applicability. When reading and valuing scientific literature, it is important to recognise the differences in the underlying philosophy. Data driven techniques are directed towards the transformation and/or visualisation of the observed data in such a way that it becomes easier to interpret the observations. For instance, one has measured evoked potentials of some stimulus of tens of subjects and one presents the "grant average" (average over all subjects) as the most characteristic response of the experiment, thereby ignoring that there could be systematic inter-subject differences. Another example is the computation of a power spectrum of an EEG signal to present the relative strengths of different rhythms that are present in the EEG. So, data driven methodologies start with a computation, followed by a presentation (visualisation) and the end with a, possibly quantitative interpretation or a hypothesis.

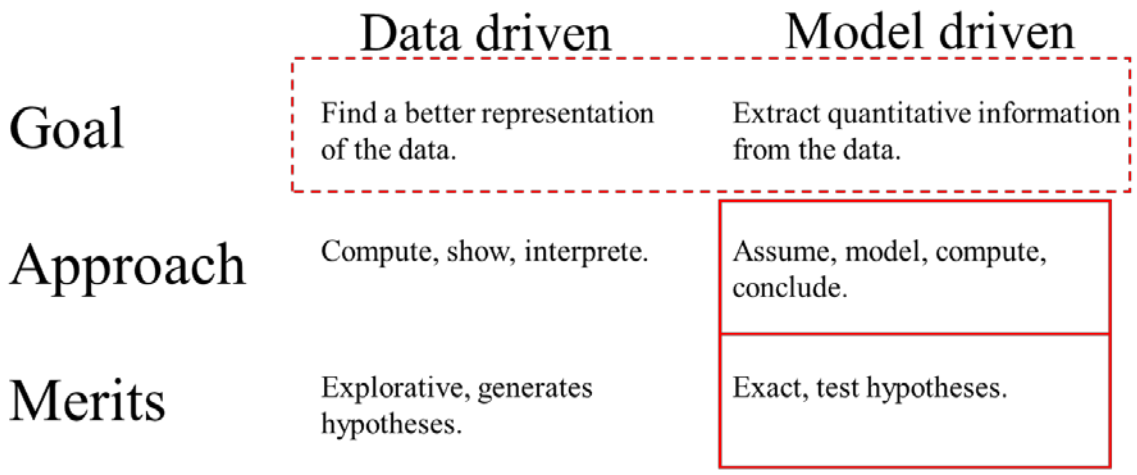

Figure 1.2. *Schematic view of differences between data driven and model driven approaches for extracting information from raw data.*

Model driven approaches start with a mathematical model describing the data (for instance a straight line plus measurement noise) and derive an estimate of the unknown parameters (e.g. slope and offset) based on some idealised statistical principle. Moreover, model driven approaches yield an estimate of the reliability of the estimated parameter, often in the form of a statistical test. Compared to data driven approaches, which often contain many ambiguities related to the most optimal way to extract information from the data, the attractiveness of model driven methods is that within validity of the assumed model, all such ambiguities are eliminated. Figure 1.2 summarizes the differences between data driven and model driven approaches.
As shall be demonstrated in this course, the ambiguity of the line fitting problem is solved by assuming that no measurement errors occur in the *m*-measurements (*x*-coordinates), the errors in the *r*- measurements (*y*-coordinates) have a Gaussian distribution and for the optimality criterion the so-called maximum likelihood criterion is adopted. If one or more of the underlying assumptions is violated, model driven approaches of parameter estimation often have a natural extension in the form of an alternative estimation algorithm, wherein alternative assumptions are

taken as starting point. Data driven approaches of the line fitting problem do not consider the variations in the precise parameter values very important. The most import aspect is that a cloud of points (complex data set) is replaced by a very simple one, i.e. a straight line with similar characteristics.
In medical and biological sciences, and particularly in brain imaging, both data and model driven approaches of parameter estimation and signal processing play a role. The focus if this course is however, more on the model driven approach.

1.2 Parameter estimation problems in brain imaging

The main brain imaging technologies are magnetic resonance imaging (MRI), positron emission tomography (PET) and magneto encephalography (MEG), which is closely related to the better known electro-encephalography (EEG). MRI and PET are techniques to make 3 dimensional (3D) images of the brain. By adapting the experimental setup, image acquisition can be done sequentially resulting in 4D (time varying) data sets. MEG and EEG consist of a limited number of sensors (typically 150 magnetometers and 70 electrodes, respectively), that record, from a certain measurement position, the magnetic field and electric potential of the brain. Therefore, both MEG and EEG contain essentially 1-dimensional temporal information and special data analysis techniques are required to extract spatial information from these data types.
Furthermore, large differences exist with regard to the physical and physiological meaning of brain signals recorded with EEG, MEG, PET or (f)MRI. The physiological quantity of interest (activation spot, functional correlation) is somewhere hidden in the data. Therefore, the extraction of the parameters of interest from the raw data often require a detailed mathematical description of the relation ship between the parameter(s) of interest and the raw measurements, in particular when model driven approaches are used. Also the large differences in spatial and temporal resolutions of different brain imaging modalities contribute to the wide variety of parameter estimation problems related to brain imaging. Below, a few of them are presented in more detail.

1.2.1 fMRI activation studies

Functional MRI is a brain imaging modality where an MR scanner is used to detect which parts of the brain are activated when a subject performs a certain task. For that purpose, the brain of the subject is scanned in a relatively fast mode (2 to 3 s. per volume) of low resolution (3 to 5 mm voxel size). Moreover, the resulting MR images have maximum sensitivity for the contrast between oxygenated and non-oxygenated blood. During the MR scanning, the subject alternates between the performance of a task (motion of fingers, perception of a visual stimulus, etc) and a rest condition. Both conditions have a typical duration of 15 to 60 s. In total, typically 100 to 400 scans are made of 50,000 to 100,000 voxels.
An fMRI data set is usually considered as a set of time series (Blood Oxygen Level Dependent-signals), and the relevant question is: which of these BOLD signals are influenced by the performance of the task? This question can be modelled (translated into mathematical terms) in different ways. One could detect those voxels for which there is a statistically significant difference between the average signal during activation and during rest. However, with such an approach it is difficult to

1. account for the fact that there are slow trends in the data, which influence the signal averages in both periods,
2. account for the fact that the effect of task performance does not translate directly in a change of signal. Instead there is a delay, and a gradual change in local oxygen consumption.
3. account for the fact that the fMRI signals are disturbed by correlated noise,

4. account for the fact that the fMRI signals are disturbed by head motion, heart beat and respiration.

For these reasons, the fMRI signals are expressed as the linear combination of know effects: a predicted hemodynamic response, with unknown amplitude, a predicted response due to motion (again with unknown amplitude) and other predicted effects with unknown amplification factors. The parameter estimation problem with fMRI analysis is to estimate these unknown parameters from the data, and to determine the relative size (with respect to noise) of the parameter of interest. When this parameter of interest deviates significantly from zero, it may be concluded that the voxel for which this conclusion is drawn, is involved in the task the subject is performing.

> **Exercise 1.1** Considering the problems discussed above regarding the determination of brain activation using fMRI, do you think a model or a data driven approach is most appropriate. Why?

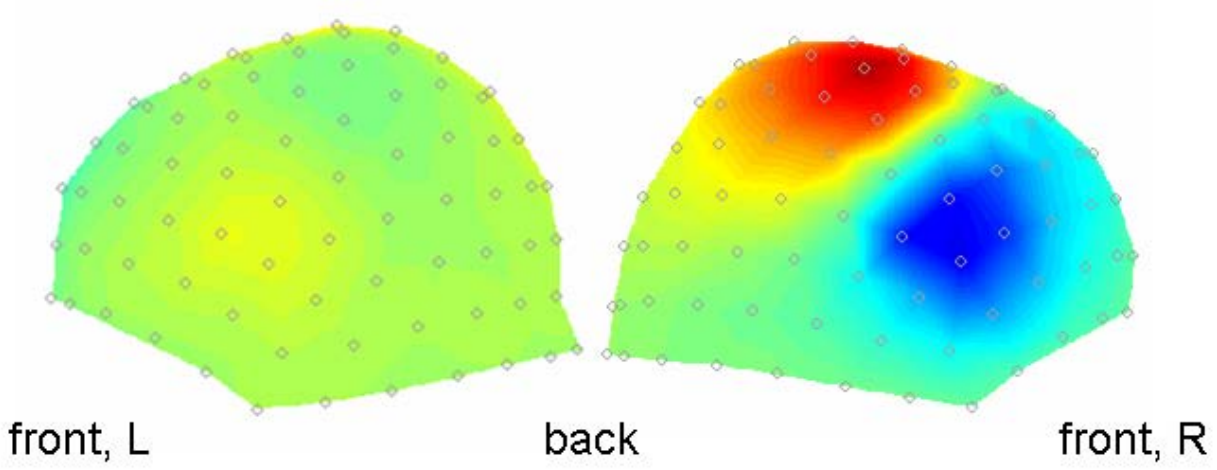

Figure 1.3. *The magnetic field generated by the brain at time t= 20 ms after stimulation is measured with a 151 channel MEG system. One of the parameter estimation challenges in MEG research is to determine where in the brain the current source is located that generates the observed magnetic field pattern. The black circles represent the positions of the MEG measurement sensors.*

MEG/EEG source localization

When the electric (EEG) or magnetic (MEG) field is known at multiple sites on the skin, respectively outside the head, an analysis of the spatial distribution of the electromagnetic field pattern can be used to determine the position and geometry of the underlying generator in the brain. To find the position of the generator using model driven approaches of parameter estimation, several assumptions need to be made, which have to be translated in a mathematical form. For instance, by assuming that the EEG/MEG signals result from dendrite currents at interacting synapses, a *current dipole model* can be postulated. This model predicts the entire electric and magnetic fields, and in particular the EEG/MEG, when the geometry of the head (including the electric conductivities of the different tissues), the position and orientation of the current dipole and the position of the sensors are known.

For a dipole in an infinite medium with constant conductivity the prediction formula for the potential is quite simple. If **x** is the dipole position, and **m** is its moment vector, then the electric potential at electrode *i*, with position vector $\mathbf{x}_i$ can be expressed as:

$$\tilde{\psi}(\mathbf{x}_i) = \frac{1}{4\pi\sigma} \frac{(\mathbf{x}-\mathbf{x}_i)\cdot\mathbf{m}}{\left|\mathbf{x}-\mathbf{x}_i\right|^3} \quad , \qquad (1.1)$$

where σ is the conductivity of the medium. When the magnetic induction is measured with a magnetometer at $\mathbf{x}_i$, oriented in the direction $\mathbf{n}_i$, the predicted measurement value due to the same dipole is

$$\tilde{B}_n(\mathbf{x}_i) = \frac{1}{4\pi\mu_0} \frac{(\mathbf{x}-\mathbf{x}_i)\times\mathbf{m}}{\left|\mathbf{x}-\mathbf{x}_i\right|^3} \cdot \mathbf{n}_i \quad , \qquad (1.2)$$

In infinite medium is a very unrealistic model to predict MEG or EEG measurements. In particular, the EEG is highly dependent on the geometry of the head and the different conductivities of the different compartments. However, it is beyond the scope of this course to discuss all details and the main point here is that a realistic model prediction can be derived. For MEG, equation (1.2) could also be used for a spherical symmetric head model, centred at the origin, when all the magnetometers are oriented such that they point to the origin ($\mathbf{n}_i = \mathbf{x}_i / \left|\mathbf{x}_i\right|$). For source localisation studies, where the goal is to find **x** and **m** from potential measurements ψ_i, or magnetic field measurements B_i, it is handy to summarize all six dipole parameters into a single vector of unknown parameters,

$$\mathbf{p} \equiv (p_0, p_1, ..., p_5)^{\mathrm{T}} = (x, y, z, m_x, m_y, m_z)^T \quad , \qquad (1.3)$$

where the transposition symbol ^{T} has been used to make a column vector from a row vector, and where $\mathbf{x} = (x,y,z)^{\mathrm{T}}$ is the dipole position and $\mathbf{m} = (m_x, m_y, m_z)^{\mathrm{T}}$ is the dipole moment. If the measured potential at the electrode *i* is given by ψ_i the source localisation problem can be expressed as the minimization of a cost function, e.g.

$$\mathrm{Cost}(p_0, p_1, ..., p_5) = \sum_i \left|\psi_i - \tilde{\psi}_i(p_0, p_1, ..., p_5)\right| \quad , \qquad (1.4)$$

which quantifies the difference between the recorded data and the data predicted by the model. The smaller the cost function, the better the predicted data resembles the observed data and therefore the dipole parameters have to be adjusted such that this cost function is as small as possible. In this sense, the localisation of electrical dipoles based on MEG or EEG is conceptually not very different from fitting a straight line through a set of measurement points.

> **Exercise 1.2** If there are the EEG data is recorded from 25 electrodes, how many current dipoles can at most be estimated from these EEG data? Hint: how many parameters are needed for each dipole?
> **Exercise 1.3** Source localization can be based on the minimization of a cost function. Does a small difference between data and model guarantee that the dipole parameters for which this minimum is obtained are good (realistic) estimates? Hint: consider the line fitting problem with either very many or very few data points.

The principle of MEG/EEG based source localization is almost always based on the minimization of a cost function. Important differences between existing methods are related to detailed modelling aspects of dipole models: How are dipoles varying over time (moving/rotating/amplitude variations)? How many dipoles are active? How is the (correlation of) the noise accounted for? How are MEG and EEG combined in a single model? Finally, as a quality assurance, one is interested in the reliability range (confidence intervals) of the dipole parameters.

Exercise 1.4 Suppose that multi-channel EEG is recorded from *I* electrodes (excluding reference electrode) and that a time window of *J* samples is selected for dipole analysis. How many data points are recorded? If a moving dipole is used, consisting of 2 dipoles, what is the number of parameters to be estimated from the data? Alternatively, if the dipole positions and orientations remain fixed in time, and only the amplitudes are varying in time, what is the number of estimated parameters?

1.2.2 MEG/EEG averaging and correlation analysis

Source localization with MEG/EEG data is often done in experiments where the same stimulus is presented to the subject many times repeatedly, and the resulting MEG/EEG signals are averaged, triggered by the stimulus onset. By computing this average, the ongoing EEG, which is caused by brain activity not related to the stimulus is averaged out, and only the part of the MEG/EEG related to the stimulus will "survive" the averaging procedure. Averaging of MEG/EEG data is often done as part of a pre-processing step with a intuitive justification, instead of having a formal mathematical reason of averaging the data. This may be fine to some extend, but if one is interested in alternative approaches to simple data averaging, e.g. to account for habituation effects (the fact that the response to stimulation may become weaker and weaker during the experimental sessions), the question ultimately pops up, of a good mathematical foundation for such an alternative. Furthermore, during the experiment, one must be certain that each following stimulus is applied only when the response of the preceding stimulus has died out. Whether this is the case or not can also be addressed using a formal mathematical model, wherein certain parameters are estimated and statistically tested against zero.
If it is assumed that each time the same stimulus is applied the brain yields the same MEG/EEG response, and that the only trial to trial variations in the measurements are due to noise, one can model the recorded MEG/EEG data as the sum of a constant signal (independent of stimulus trial) and (uncorrelated) background noise representing the ongoing EEG. The parameter estimation question is then to estimate the constant brain response, given the data and the model relating them. It then appears that simple averaging of the data is a *maximum likelihood estimator* of the brain response. In this way, the intuitive approach has obtained a "fundamental" basis, but also, the same basis can be extended to more complicated response models which account for trial to trial variations and for correlation of background noise Finally, the formal response model provides a framework to test whether the ending part of the brain response deviates from zero.
Apart from source localization, MEG and EEG can be used to study (changes in) functional correlations in the human brain. By recording spontaneous MEG/EEG in a group of normal controls and in a group of diseased subjects, it can be explored whether in the diseased group long range connections are impaired, by testing the correlations of the MEG/EEG signals between far away sensors. If significant changes in long range connections can indeed be detected on the basis of MEG signals, it may for instance become possible to use MEG as diagnostic tool for an early diagnosis of diseases like Alzheimer's disease. Also in these types of studies parameter estimation problems play a role, for example when a correlation co-efficient is estimated from the raw data. Moreover, the statistical significance of this estimated correlation co-efficient has to be tested in some way, because also noisy signals may have a correlation.

1.2.3 Image reconstruction

Medical images are usually acquired in a way that is very different from classical photography, where the image is the direct result of the amount of light that interacts with a photo reactive layer. With most medical images, there is an *indirect* relationship between the quantity that is measured and the resulting image. For instance, with a PET camera co-incidences of radioactive

decay are counted with detectors that are placed around the object to be scanned. Thus, the raw data consists of projected images of the 3D radioactivity distributions. The extraction of a 3D tracer image from projection data is an example of a parameter estimation problem. The unknown parameters are concentrations of the radioactive tracer, at each pre-defined cube representing a voxel of the reconstructed PET scan.

> **Exercise 1.5** Suppose a PET scan has a field of view of 30 cm by 30 cm, a pixel size of 9 mm2 and consists of 60 slices. How many parameters need to be estimated in the reconstruction problem?

The PET reconstruction problem is a very complicated one because the number of parameters is large, the geometry of the detectors must be accounted for, the photon scattering and attenuation, the Poisson distribution of the decay process, etc. Therefore, the reconstruction process is often approximated by a spatial filtering procedure that is applied to the recorded projection data. Also MR images are the result of a reconstruction algorithm. The very raw input data consists of RF echoes that are generated using either spin echo or gradient echo mechanisms. The image reconstruction is highly dependent on the way the object is scanned: which gradients are applied, in which order, with which duration and amplitude, etc. Furthermore, it depends on (the sensitivity of) the coil that is used to record echoes and on the number of these coils that are used. For many acquisition methods the raw data can be considered as a regular sampling of the image in *k*-space, and then the image reconstruction is equivalent to a spatial Fourier transform.

1.2.4 T_2- determination from MRI data

An MR image can be considered as a regular array of grey values that together form an image of a scanned object. Because of the many intermediate steps that are required to convert the raw acquired data into an MR image, and because of the high dependence on instrumental parameters, the *exact* physical meaning of these grey values can often not be given. However, by playing around with scanning parameters such *TR* and *TE*, MR images can be created that are selective sensitive for e.g. T_1, T_2 or proton density contrasts. For instance, if a spin/echo sequence is used and with a long *TR* (>>T_1), the MR signal *S* is proportional to

$$S \propto \rho \exp(-TE / T_2) \quad , \qquad (1.5)$$

where ρ is the proton density. When *TE* is of the order of the average T_2, the MR image becomes very sensitive to variations in T_2, i.e. the differences in grey values of different tissues is to a large extend caused by differences in the T_2 value. However, since also the proton density varies from tissue to tissue, the contrast in the image is a mixture of combined T_2 and proton density effects. Such images are called T_2-*weighted* images.
There are many clinical and scientific applications of quantitative T_2-maps, i.e. MR images where the grey value is directly proportional to the local T_2-value. In such maps the effect of proton density would be eliminated. This can be achieved by making a sequence of T_2-*weighted* images with different echo times. By combining these images, one obtains for each voxel a series of grey values of which the value is proportional to S_n, with

$$S_n \propto \rho \exp(-TE_n / T_2) \quad ,n=0,..,N-1 \qquad (1.6)$$

where TE_n is the echo time of the *n*-th image.

> **Exercise 1.6** How many echo times are needed at least to eliminate ρ? If two different echo times are applied (N=2), how can the pure T_2-value be extracted from the data?

In reality, the spatial variations of T_2 can be larger than the voxel size of the MR image. In such cases equations (1.5) and (1.6) are not valid anymore, and the signal at each voxel must be considered as a mixture of two or more different T_2-values. One solution to this problem could be to apply a sequence with a higher spatial resolution. However, higher resolution will generally require longer scanning time and will likely lead to a decreased the signal to noise ratio of the acquired scan. Another solution to the problem is to refine the exponential decay model to a multi-exponential decay model:

$$S_n = A_{short} \exp(-TE_n / T_{2,short}) + A_{long} \exp(-TE_n / T_{2,long}) \qquad ,n=0,..,N-1 \qquad (1.7)$$

Here the MRI grey value is described as a mixture of two T_2-values: $T_{2,\text{short}}$ and $T_{2,\text{long}}$. The parameter estimation problem is now extended to the estimation of four parameters: $A_{2,\text{short}}$, $A_{2,\text{long}}$, $T_{2,\text{short}}$ and $T_{2,\text{long}}$. Therefore, also the number of echoes has to be increased. A further complication of the parameter estimation problem is that all estimated parameters need to have the constraint that they are positive.

1.2.5 PET tracer kinetic modelling

When a radioactive tracer is injected into the body a PET camera can be used to follow the concentration of the tracer as a function of time. For that purpose, co-incidences of 512 kEV photon decay are detected in predefined time frames, and for each time frame a PET image is reconstructed and expressed as the number of tracer particles per volume. Dependent on the tracer used and the measurement setup, the tracer will distribute itself over different compartments: the arteries, the veins, a tissue where the tracer accumulates or some molecules to which the tracer will bound. Since these compartments can be much smaller than the voxels of the reconstructed PET images, the recorded concentrations represent an (unknown) mixture of different compartments.
One of the goals of the use of PET tracer studies is to obtain insight into the interaction of the different compartments. For that purpose, mathematical models are built describing the exchange and binding of the tracer material. In particular, the goal is to make images of the exchange and binding constants that play a role in the tracer dynamics. When the model consists of L compartments, the concentration time curve $C(t)$ will consists of a mixture of L exponential functions

$$C(t_n) = b_0 \, e^{-k_0 t_n} + ... + b_{L-1} \, e^{-k_{L-1} t_{L-1}} \qquad ,n=0,..,N-1 \qquad (1.8)$$

Here t_n represents the time at which the n-the frame was recorded. The parameter estimation problem is to determine (b_0,…b_{L-1}, k_0,…k_{L-1}) from the observed concentrations at all reconstructed voxels. Then a map or an image can be made of each of these parameters. Complicating factors in the parameter estimation problem are e.g. that 1.) meaningful parameters are positive 2.) the measurements are corrupted with Poisson noise (and not Gaussian noise).

1.2.6 Coordinate matching

The solution of the MEG/EEG source localization problem yields a set of dipole parameters representing the electric (dipole) activity that best explains the observed MEG or EEG recordings. Due to the mathematical model underlying the estimation procedure, these dipole parameters must be interpreted as coordinates in the framework attached to the MEG and/or EEG sensors. In other words, the source localization analysis only provides information about the position of the electrical activity with respect to the MEG/EEG sensors. To be clinically and scientifically meaningful, the source activity must be known with respect to an anatomical picture of the brain, e.g. an MR scan of the subject.

MEG/EEG dipoles can be matched onto subject's anatomical MR scan when a set of 3-D point pairs are given, that represent corresponding MEG/MRI points. In practice, such point pairs can be obtained by placing magnetic markers (energized coils) during MEG scanning and measuring their 3-D coordinates with respect to the MEG sensors. Then, before the MRI scan is made, the magnetic markers are removed and replaced by markers that are well visible on the MRI scan. Because the MR scan consists of rectangular voxels of known dimensions, the coordinates of the markers on the MR scan can be determined by counting voxels in *x*-*y*- and *z*-direction, and multiplying these counts with the voxel size. This procedure yields a series of *N* point pairs $(\mathbf{x}^{(0)}_{\mathrm{MEG}}, \mathbf{x}^{(0)}_{\mathrm{MRI}})$, $(\mathbf{x}^{(1)}_{\mathrm{MEG}}, \mathbf{x}^{(1)}_{\mathrm{MRI}})$, …, $(\mathbf{x}^{(\mathrm{N-1})}_{\mathrm{MEG}}, \mathbf{x}^{(\mathrm{N-1})}_{\mathrm{MRI}})$ that related by an unknown rotation *R* and an unknown translation **t**:

$$\mathbf{x}^{(n)}_{MEG} = R(\mathbf{x}^{(n)}_{MRI}) + \mathbf{t} \qquad ,n=0,..,N\text{-}1 \qquad (1.9)$$

Here *R* is a 3x3 rotation matrix, which depends on three rotation angles, and t is a 3-D translation vector, with three unknown components (t_x, t_y, t_z). The parameter estimation problem to be solved is to extract the three rotation angles and the three translation components from the recorded *N* point pairs and the theoretical relation (1.9).

> **Exercise 1.7** How many parameters need to be determined from the data in the coordinate matching problem? How many equations are represented in (1.9) for each pair of points? How many point pairs are needed at least to determine rotation and translation?
> **Exercise 1.8** What can you say about the accuracy of $\mathbf{x}^{(n)}_{\mathrm{MRI}}$?

Once the parameter estimation problem related to (1.9) is solved, an estimate of *R* and **t** is available, and then the transformation from MEG to MRI can be computed for an arbitrary point. In particular, one can apply *R* and **t** onto the estimated dipole parameters and convert them to the MRI coordinate system. Finally, by rounding off these coordinates to the most nearby voxel, the MEG/MRI data viewer programme knows where in the MRI scan the dipoles should be drawn. One of the problems that need to be solved in a practical application is how many markers are needed and where the markers should be placed, such that the transformation from one scan to the other is as accurately as possible at the region of interest. Intuitively one would expect that the accuracy is best near the gravity point of the markers, and the accuracy becomes poor for far away regions.

1.2.7 Image matching

A very common problem in (brain) imaging is that scans two different modalities need to be overlaid and compared. Such problems arise e.g. when one modality represents anatomy (MR or CT) and the other one represents functional brain activity (PET, SPECT, fMRI). By combining both scans, using image matching, it can be determined where in the brain certain functionality is localised. Overlaying or fusing images made on different scanners is complicated because the brain is scanned with different (rotated and translated) field of views and with different pixel sizes, see figure 1.4.

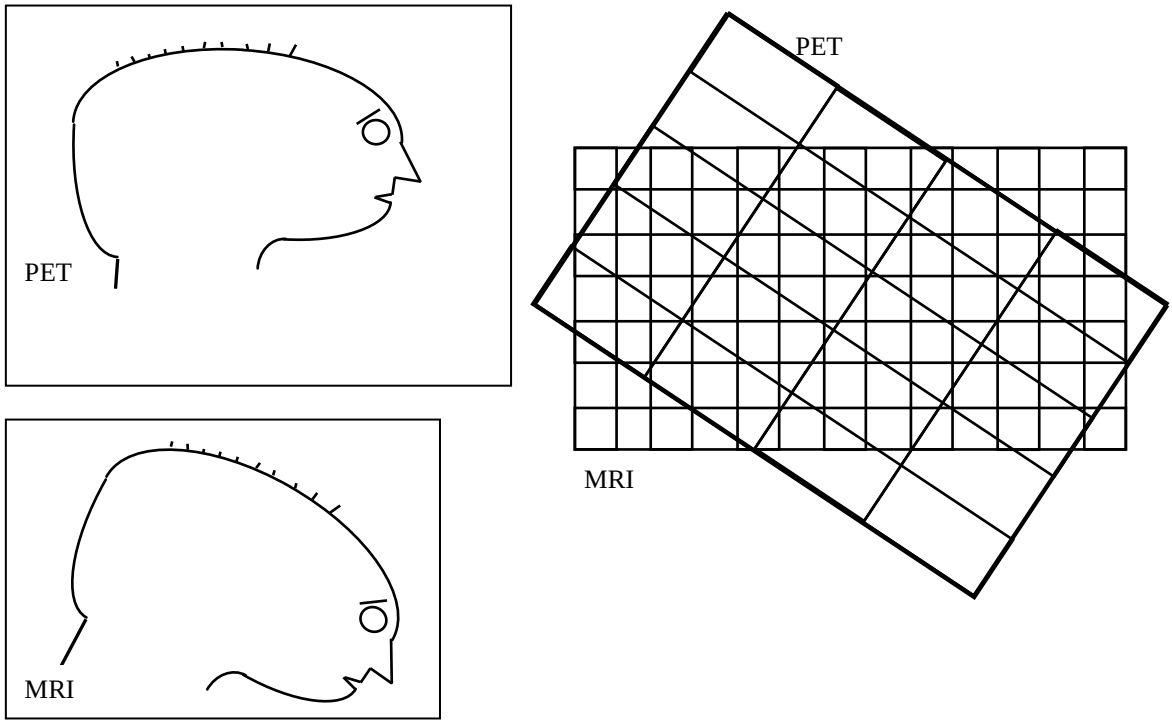

Figure 1.4. *Image Fusion. When the same subject is scanned on different modalities that are to be compared, several problems have to be solved. First, the field of views are generally different on both modalities, second, the resolutions are different and third, the orientation of the subject w.r.t. the field of views are different. These problems can be solved with data interpolation and image matching techniques. This figure presents the example of a head that is scanned with PET (low resolution) and MR (high resolution).*

One solution to the image fusion problem is to identify a set of corresponding anatomical points on both scans and to use the co-ordinate matching approach described above. However, this approach is not always possible because on the functional scan clear anatomical landmarks can not always be identified. Sometimes it is possible to use artificial markers that are attached to the object and which are visible on both scanner modalities. Then these markers can be used as point pairs, provided that these markers do not move with respect to the head during the time between the scans.
A completely different approach to point pair matching is to interpolate both scans to the same mesh, and to determine a correlation coefficient between both images. By shifting and rotating one of the images with respect to the other, and computing again the correlation coefficient, it can be determined whether the transformed image gives a higher or a lower correlation coefficient. When an appropriate correlation coefficient is used, a higher correlation will indicate a better match, i.e. a better projection of functional information onto anatomy. This procedure can be automated using a function maximisation algorithm.

> **Exercise 1.9** Is the image matching algorithm more based on a model or a data driven approach?

Image matching algorithms also play a role e.g. when dynamical PET scans or fMRI time series are analysed. Despite the instruction given to the subject to lay still, there will always be small movement. Ignoring this motion in the analysis of PET and fMRI time series will result in sub-optimal results. Therefore one usually applies a motion correction on the raw data. For that

purpose, an image matching algorithm is applied between each scan and the first one. This procedure results in a series of aligned images wherein motion effects are reduced.

Exercise 1.10 Suppose that an fMRI series consists of 100 3D images on which motion correction is applied. How many motion parameters are to be estimated in total?
Exercise 1.11 Suppose that an fMRI series consists of several 3D images for which motion correction is applied. How could one assess the consistency of the obtained parameter estimates?

2 Parameter estimation principles

Usually different methods to extract parameters of interest from the raw data exist. When different solutions methods yield different results the question presents itself: which one is right (if any)? To illustrate this dilemma we consider the estimation of the speed of (a car), based on a series of N position measurements $s_0, s_1, s_2,\ldots, s_{N-1}$ at times $t_0, t_1, t_2,\ldots, t_{N-1}$. When the car moves at constant speed v and if at t=0 its position s is p, then the theoretical relation between position and time would be

$$s = p + vt \quad . \qquad (2.1)$$

One strategy to find an estimate of v is to fit a straight line through the data points (s_n, t_n). This can be done by minimizing the sum of squared differences between the theoretical values $p+vt_n$ and the observed values s_n. In other words, one would minimize the cost function

$$\text{Cost} = \min_{(p,v)} \sum_n (s_n - p - vt_n)^2 \quad . \qquad (2.2)$$

Here, the cost function must be considered as a function of the unknown parameters p and v. The values $\hat{p}$ and $\hat{v}$ for which this cost function is minimal yield the best reproduction of the measurements and therefore, these values can be considered as the best possible estimates of the true parameters. Note that in this approach there are two unknown parameters to be estimated. One of them is the parameter of interest (v), and the other one is a nuisance parameter (p), which nevertheless has to be added because there is no a priori knowledge on the position of the car at t=0.

Minimization of the cost function can be done by computing the partial derivatives of Cost w.r.t. p and v, setting them to zero (i.e. setting $\frac{\partial \text{Cost}}{\partial p} = 0$ and $\frac{\partial \text{Cost}}{\partial v} = 0$). This yields two linear equations in p and v that can be solved.

Exercise 2.1 Verify that the solution of the minimization problem posed in (2.2) is given by

$$\hat{p} = \frac{\sum_n (t_n)^2 \sum_n s_n - \sum_n t_n \sum_n t_n s_n}{\sum_n (t_n)^2 N - \sum_n t_n \sum_n t_n}$$
$$\hat{v} = \frac{-\sum_n t_n \sum_n s_n + N \sum_n t_n s_n}{\sum_n (t_n)^2 N - \sum_n t_n \sum_n t_n} \quad . \qquad (2.3)$$

The ambiguity of the result presented in (2.3) raises from the alternative line fitting strategy, resulting from the observation that equation (2.1) is equivalent to

$$t = -p/v + s/v \quad . \qquad (2.4)$$

An alternative estimation would result from minimizing

$$\text{Cost}' = \min_{(p,v)} \sum_{n} (t_n + \frac{p}{v} - \frac{s_n}{v})^2 \quad . \qquad (2.5)$$

This alternative cost function results in different estimates of the speed and offset. So here is the question pops up, which one is right?

> **Exercise 2.2** Show that minimization of (2.5) results in a different solution than (2.3).
> **2.2A** Show this in a numerical example, using e.g. Office Excel.
> **2.2B** Show this theoretically. Hint1: first introduce two new parameters *A* and *B*, such that the minimization problem becomes more similar to (2.2). For example, take $A=-p/v$ and $B=1/v$. Then one can first minimize over *A* and *B*, to find $\hat{A}$ and $\hat{B}$, and the alternative estimates of *p* and *v* can be found from $\hat{p}' = -\hat{A}/\hat{B}$ and $\hat{v}' = 1/\hat{B}$. Hint2: the minimization of (2.5) over *A* and *B* is equivalent to the minimization over *p* and *v* of (2.2). Swap the roles of t_n and s_n, and identify *v* to *A* and *p* to *B*.

The underlying cause leading to the two different estimates is that in the first approach the "vertical" differences between line and data points were minimized (i.e. in the *s*-direction), whereas in the second approach the "horizontal" differences were minimized (i.e. in the *t*-direction). One could provide an argumentation for each of these possibilities, by assuming in the first approach that the time measurements contain no measurement error and that all measurement errors in the data were in the position measurements. In the second approach the implicit assumptions are reversed.
But if so, which is the preferred estimator? The answer depends on the precise way the measurements are done. The most realistic situation measurement setup is that cars are detected at fixed points along the highway. Then the positions of the detection points (s_0, s_1, s_2,…, s_{N-1}) can determined very accurately, whereas the detection time may depend on the detection method, and therefore vary from time to time. In this situation largest errors are in the time measurements and the second approach would be the method of choice.
However, even if it is possible to say that measurement errors dominate in position or time, the line fit is not unambiguous. In equations (2.2) and (2.5) the optimal line fits were defined in terms of the minimization of a sum of squared differences. If the cost function had been define as

$$\text{Cost} = \min_{(p,v)} \sum_{n} |s_n - p - vt_n| \qquad (2.6)$$

the solution of the minimization problem would be different from (2.3).
To find the minimum of the absolute value based cost function of (2.6), it is noted that for fixed *p*, Cost() as a function of *v* is a sum of modulus functions and therefore it consists of line segments that jump to another slope whenever one of the terms equals zero, i.e. whenever $s_n - p - v\,t_n = 0$, see figure 2.1. Similarly, when *v* is kept fixed Cost() as function of *p* will make jumps whenever one of the arguments absolute values is zero. The minimum of Cost() will never occur in the middle of a linear line segment, and therefore we know that at the minimum we have, for some *m* and *n*

$$\begin{cases} s_n - p - vt_n = 0 \\ s_m - p - vt_m = 0 \end{cases} \quad . \qquad (2.6A)$$

In other words, the minimum of Cost(*p*,*v*) occurs when the string line we are trying to fit, passes exactly through two data points (s_n, t_n) and (s_m, t_m). Beforehand, we do not know which two points give the lowest Cost(), but in principle we could just test all point pairs, evaluate Cost()

and pick out the best one. In practice, much faster algorithms exist, in particular for data sets with large number of points and more than two parameters. Here it suffices to indicate how a solution can be found and what its main characteristics are. The fact that the solution passes exactly through two of the points, implies that the best solution it is insensitive to outliers in the data.

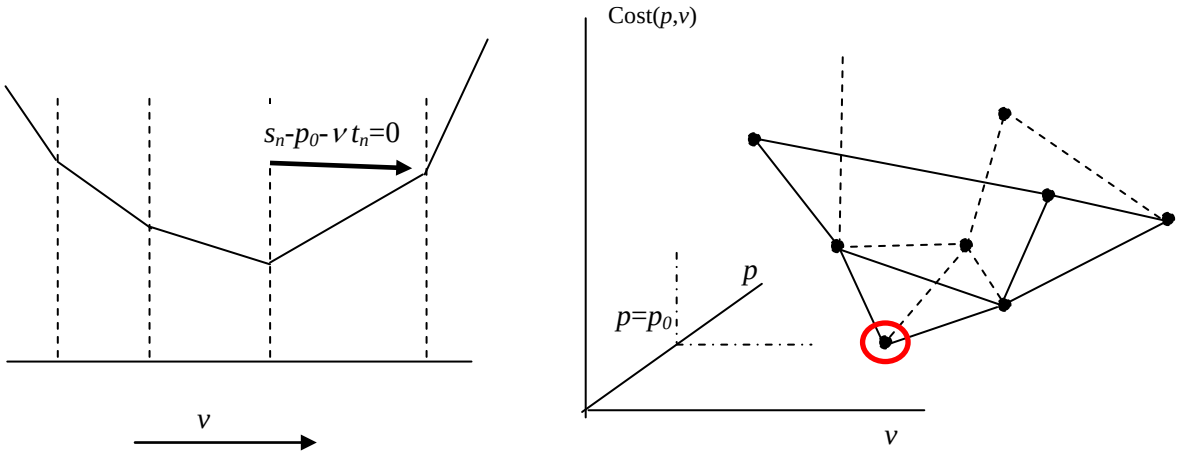

Figure 2.1. *Cost function based on absolute values. On the left a cross section is presented of Cost(p,v) for fixed $p=p_0$ and varying v. On the right the 3D-version of Cost() is visualized. Since the curve on the left is a superposition of scaled and horizontally shifted curves of the form |v|, the curve consists of straight line segments that alter direction each time that one of the arguments of |v| is 0. Since the minimum of Cost() must occur at such a crossing point, a strategy to find the minimum of Cost() would consist of testing and comparing all these crossing points.*

However, there is a very practical reason to choose the least squares cost function and that is that this cost function can be minimised analytically by setting the partial derivatives of the cost function equal to zero, and solving the resulting equations. The minimisation of (2.6) is more difficult. Such a practical argumentation is however very “opportunistic” and in many cases, where the different cost functions have a large effect, it is important to have also a theoretical argumentation to chose for one cost function or another. These arguments should be founded on modelling assumptions and different assumptions logically imply different methods and lead to different results. The great advantage of such a theoretical framework is that a discussion of assumptions is usually more objective than a discussion afterwards, wherein the merits of different estimation methods are evaluated, purely on the plausibility of the different results that these methods lead to. This is also the main argument for model driven approaches compared to data driven approaches of parameter estimation.
The theoretical framework that will be explored in this course is so-called Maximum Likelihood Estimation (ML estimation). This framework describes the measurements in terms of a model containing the parameters of interest and the measurement noise. The latter follows a stochastic (non-deterministic) distribution that is assumed to be partly known. ML also provides a measure to determine the most reliable parameter estimates and it can even provide confidence intervals of each parameter. These are the “error bars” or the boundaries within which alternative good parameter estimates will reside. The ML approach is also very fruitful in applications where one is interested in the question whether a parameter is different from zero, which is interpreted as the detection of “an effect”. Since it would be a very large coincidence if an estimated parameter would turn out to be zero exactly, and therefore “different from zero” should be read as “so far

off from zero that it exceeds chance level". Within the field of brain imaging, the ML approach is applicable e.g. in fMRI data analysis, dipole fitting and some image reconstruction methods. For image fusion, or problems where there are a priori constraints on the estimated parameters, methods are chosen on more practical grounds.

2.1 Maximum Likelihood estimation

Here we consider the situation that one is interested in the estimation of a number of M parameters, collected in an M-dimensional vector $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_{M-1})^T$ that cannot be measured directly. Instead, N measurements have been done $(r_0, r_1, \ldots, r_{N-1})$ that are related to the measurements by the N model predictions $\tilde{r}_n(\boldsymbol{\theta})$, n=0,…,*N-1*. The bold face representation of $\boldsymbol{\theta}$ indicates that it is a (column) vector. In a practical situation, the index n may refer to time or sensor, or to a certain repetition of a measurement. Furthermore, the measurements r_n of $\tilde{r}_n(\boldsymbol{\theta})$ are embedded in the noise η_n. This noise is unpredictable and therefore it is modelled as a random variable, with a certain probability density distribution, on which certain assumptions can be made. This situation can be expressed as

$$r_n = \tilde{r}_n(\boldsymbol{\theta}) + \eta_n \qquad , n=0,\ldots N\text{-}1. \qquad (2.7)$$

In other words, we have recorded the data r_n, which contain information on the parameters of interest $\boldsymbol{\theta}$ and which are disturbed by the random noise η_n. Equation (2.7) is equivalent to

$$\eta_n = r_n - \tilde{r}_n(\boldsymbol{\theta}) \qquad , n=0,\ldots N\text{-}1. \qquad (2.8)$$

Once the data are recorded, and once we have *assumed* a certain model (i.e. we have specified how $\tilde{r}_n(\boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$), we can compute for each $\boldsymbol{\theta}$ the realisation of the noise. Generally, not every realisation is equally probable: this depends on the *assumed* distribution of the noise. For instance, when η_n would have a (multivariate) Gaussian distribution, i.e. with density function $f_\eta(\eta)$, with zero mean and no cross correlation,

$$f_\eta(\eta_n) = f_\eta\left(r_n - \tilde{r}_n(\boldsymbol{\theta})\right) \sim \frac{e^{-\frac{1}{2\sigma^2}\sum_n (r_n - \tilde{r}_n(\boldsymbol{\theta}))^2}}{(2\pi)^{N/2}(\sigma)^N} \qquad . \qquad (2.9)$$

If the Gaussian distribution is an appropriate model for the noise, then equation (2.9) can be used to compute the *most-likely* distribution of the observed residuals, by determining for which parameters $\boldsymbol{\theta}$ the right hand side is maximum. The parameter vector $\hat{\boldsymbol{\theta}}$ for which that occurs is called the *maximum likelihood* (ML) estimate of the true value $\boldsymbol{\theta}$. So $\hat{\boldsymbol{\theta}}$ is found by solving the following maximisation problem:

$$\text{Gain}_{\text{ML}} = \max_{\boldsymbol{\theta}} \frac{e^{-\frac{1}{2\sigma^2}\sum_n (r_n - \tilde{r}_n(\boldsymbol{\theta}))^2}}{(2\pi)^{N/2}(\sigma)^N} \qquad . \qquad (2.10)$$

The solution of this problem is simpler than it would seem at first sight, because the denominator is independent of $\boldsymbol{\theta}$. Because of the minus sign in the exponent, the maximisation problem posed in (2.10) is equivalent to the following minimisation problem:

$$\text{Cost}_{\text{OLS}} = \min_{(\boldsymbol{\theta})} \sum_n (r_n - \tilde{r}_n(\boldsymbol{\theta}))^2 \qquad , \qquad (2.11)$$

where one can recognise the Ordinary Least Squares estimate (OLS).

Exercise 2.3 Show how the line fit problem of exercise 2.1 fits in the theory of section 2.1. What is $\boldsymbol{\theta}$, what is s_n and what is t_n?

If instead of multivariate Gaussian, the noise would be distributed according to independent double exponential distributions, the distribution function would look like

$$\eta_n = r_n - \tilde{r}_n(\boldsymbol{\theta}) \sim \frac{e^{-\frac{1}{\lambda}\sum_n |r_n - \tilde{r}_n(\boldsymbol{\theta})|}}{(2\lambda)^N} \quad . \qquad (2.11B)$$

Applying the maximum likelihood estimation principle with this noise distribution yields the absolute value based cost function, i.e.

$$\text{Cost}_{\text{ABS}} = \min_{(\boldsymbol{\theta})} \sum_n |r_n - \tilde{r}_n(\boldsymbol{\theta})| \quad , \qquad (2.11C)$$

Therefore, similar to the OLS cost function the ABS function can be given a firm statistical basis if one is willing to assume that the noise has a double exponential instead of a Gaussian distribution. In reality, considering that noise is often caused by the superposition of independent processes, is a often more realistic to assume a multivariate Gaussian distribution. Compared to a Gaussian distribution, in a bi-exponential distribution outliers are much more likely, see figure 2.2.

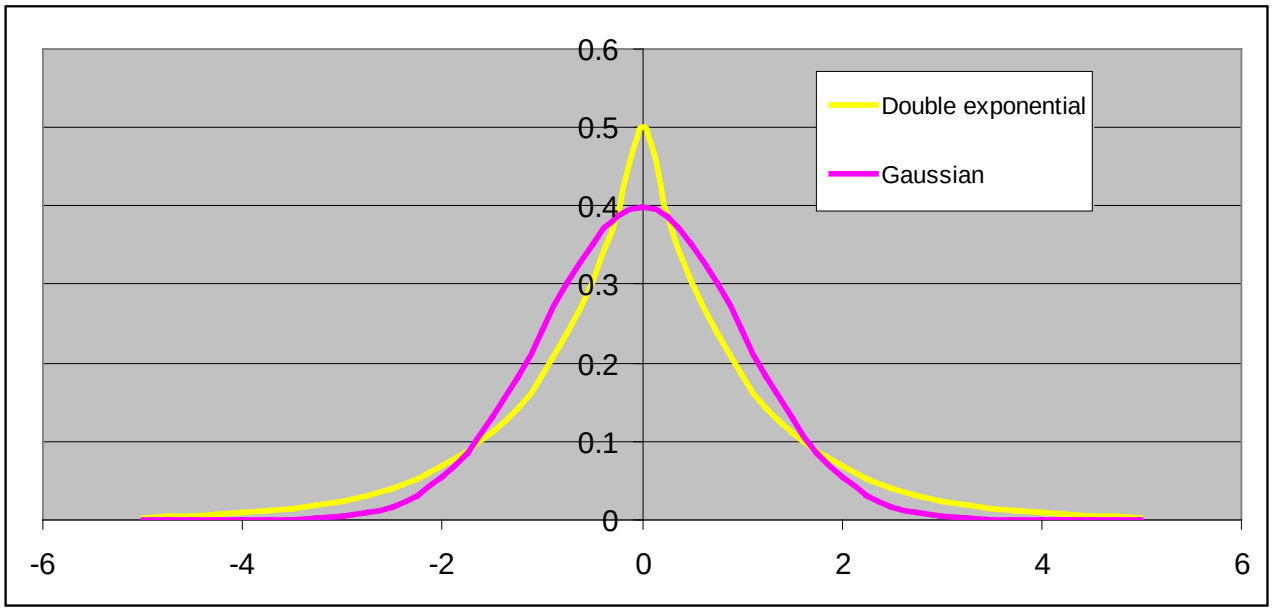

Figure 2.2. *The double exponential and Gaussian distributions are presented in a single plot. One observes that in the double exponential distribution large negative and large positive values are more likely than in the Gaussian distribution.*

2.2 The OLS estimate of linear models

In the previous section it has been shown that ML estimation starts off from a maximisation problem, which can be transformed into a minimisation. We now make distinction between the case that all model parameters $\boldsymbol{\theta}$ are linear, and the case that they are non-linear. In the former case,

$$\tilde{r}_n(\theta_0,\theta_1,\theta_2,...,\theta_{Z-1}) = \sum_{m=0}^{M-1} B_{nm}\theta_m$$

or , n=0,…,N-1 . (2.12)

$$\tilde{\mathbf{r}}(\boldsymbol{\theta}) = B\boldsymbol{\theta}$$

Here the matrix-vector notation has been used, which is explained in appendix A. The essence of linear models is that $\tilde{\mathbf{r}}(\boldsymbol{\theta})$ can be expressed as a matrix vector multiplication, such that the matrix is independent of $\boldsymbol{\theta}$.

Exercise 2.4 Show how the line fit problem of exercise 2.1 is a linear model. Give the explicit form of the matrix $\tilde{B}$.

Exercise 2.5 Suppose one has recorded the height h_n of an object at times t=0T, 1T, 2T,…, (N-1)T. Assuming free motion en neglecting the resistance of the air, give a model of the height as function of time. Give a presentation of the data model in terms of the matrix B assuming the initial height, the initial speed and the acceleration of gravity are the unknown parameters.

The attractive aspect of linear models is that OLS estimator can be found analytically. This means that an expression can be derived, which gives the estimator in terms of the data. For that purpose, the cost function is differentiated with respect to each of the components of $\boldsymbol{\theta}$, setting all these partial derivatives equal to zero, and solving the resulting system of M equations and M unknowns. Doing so for all components θ_m with m=0,…,M-1 one finds:

$$\begin{aligned}
\frac{\partial}{\partial\theta_m}(\mathrm{Cost}_{\mathrm{OLS}}) &= \frac{\partial}{\partial\theta_m}\left(\frac{1}{N}\sum_{\mathrm{n}}(r_n-\tilde{r}_n(\boldsymbol{\theta}))^2\right)\\
&= -2\left(\frac{1}{N}\sum_{\mathrm{n}}(r_n-\tilde{r}_n(\boldsymbol{\theta}))\frac{\partial\tilde{r}_n(\boldsymbol{\theta})}{\partial\theta_m}\right)\\
&= \frac{-2}{N}\sum_{\mathrm{n}}(r_n-\tilde{r}_n(\boldsymbol{\theta}))\frac{\partial}{\partial\theta_m}\left(\sum_k B_{nk}\theta_k\right)\\
&= \frac{-2}{N}\sum_{\mathrm{n}}(r_n-\tilde{r}_n(\boldsymbol{\theta}))\sum_k B_{nk}\frac{\partial\theta_k}{\partial\theta_m} \qquad , m=0,\ldots,M\text{-}1. \qquad (2.12\mathrm{B})\\
&= \frac{-2}{N}\sum_{\mathrm{n}}(r_n-\tilde{r}_n(\boldsymbol{\theta}))\sum_k B_{nk}\delta_{km}\\
&= \frac{-2}{N}\sum_{\mathrm{n}}(r_n-\tilde{r}_n(\boldsymbol{\theta}))B_{nm}\\
&= \frac{-2}{N}\sum_{\mathrm{n}}(r_n-\sum_{m'}B_{nm'}\theta_{m'})B_{nm} = 0
\end{aligned}$$

Here the *Kronecker delta* symbol has been used:

$$\delta_{k,m} \equiv \begin{cases} 1 & \text{if} \quad k=m \\ 0 & \text{if} \quad k\neq m \end{cases} \qquad . \qquad (2.13)$$

This symbol appears in the derivation where the derivative of θ_k with respect to θ_m is needed (fourth line of equation 2.12B). When k and m are different, the result is 0 (similarly to a function that only depends on x and which is differentiated w.r.t. y) and when k and m refer to

the same component, the result is 1 (similar to the function $f(x)=x$, which is differentiated w.r.t. x).
When the partial derivatives are set to zero, a set of equations is obtained, which is satisfied by the estimator of $\boldsymbol{\theta}$ that we are looking for. Therefore the last line in equation (2.12B) represents a set of equations in $\hat{\boldsymbol{\theta}}$, and this set of equations is linear**:**

$$\begin{aligned}\sum_{m'}\left(\sum_{n} B_{nm'}B_{nm}\right)\hat{\theta}_{m'} &= \sum_{n} B_{nm} r_n \\ \left(B^T B\right)\hat{\boldsymbol{\theta}}_{OLS} &= B^T \mathbf{r} \\ \hat{\boldsymbol{\theta}}_{OLS} &= (B^T B)^{inv} B^T \mathbf{r}\end{aligned} \quad . \qquad (2.14)$$

The meaning of these equations is that the unknown parameters can be found by solving a (linear) system of M equations and M unknowns (the M components of $\hat{\boldsymbol{\theta}}$). The solution of this system of equations is not explicitly derived (such as was done in exercise 2.1), but instead left and right hand side of the equations is multiplied with $(B^{\mathrm{T}}B)^{\mathrm{inv}}$ to obtain a formal solution. The explicit algorithm to find this matrix inverse from $(B^{\mathrm{T}}B)$ is not discussed in this course, we suffice to note that under certain conditions this matrix can be computed, i.e. when the matrix $B^{\mathrm{T}}B$ has an inverse. This is the case what all columns of B are linearly independent. If that is not the case, the system of equations may have multiple solutions meaning in the present context that there is more than one parameter vector $\boldsymbol{\theta}$, that all lead to residuals η_{n} that have the same (maximum) likelihood. Linear dependence occurs for instance, when there are more unknowns than measurements, i.e. when $M>N$. In terms of the example in exercise 2.5, this means that we need at least three height measurements.

Exercise 2.6 What goes wrong in the derivation from (2.12B) to (2.14) in case of nonlinear models?
Exercise 2.7 How can equation (2.14) be simplified when there is 1 linear parameter?
Exercise 2.8 Use the result of exercise 2.7 for the line fit problem posed by eq. (2.2) in case the offset is zero.
Exercise 2.9 Use the result of exercise 2.7 for the line fit problem posed by eq. (2.2) in case the slope is zero.

2.3 Generalizations

2.3.1 Estimation of the noise level

Until here, nothing has been said about the noise level σ, appearing in the assumed noise distribution, equation (2.9). Apparently, the ML estimator is independent of the noise level, at least when the noise is uncorrelated and it has a Gaussian distribution. Nevertheless, it is possible to estimate also the noise level from the data and to do so in the same framework of modelling assumptions as is used for the parameter estimates. The way to obtain an estimator $\hat{\sigma}$ for the noise is quite similar to the way the estimator of $\boldsymbol{\theta}$ has been derived. Starting from the gain function (2.10), one searches for that value of $\hat{\sigma}$ that gives the most like distribution of $\boldsymbol{\eta}$, i.e. the one that gives maximum gain.
Therefore, an estimate of the noise level can be obtained by setting the partial derivatives of (2.10) to zero, but this time these partial derivatives include the derivative to σ. Then, these derivatives are set to zero and the resulting equations are solved. We have seen already that the

estimator of $\boldsymbol{\theta}$ was independent of σ. Therefore, we can substitute $\hat{\boldsymbol{\theta}}$ in equation (2.10), and maximize with respect to σ:

$$\text{Gain}_{\text{ML}} = \max_{\sigma} \frac{e^{-\frac{1}{2\sigma^2}\sum_n \left(r_n - \tilde{r}_n(\hat{\boldsymbol{\theta}})\right)^2}}{(2\pi)^{N/2}(\sigma)^N} \quad . \tag{2.15}$$

The result of this maximization is:

$$\hat{\sigma}^2_{OLS} = \frac{1}{N}\sum_{\text{n}} (r_n - \tilde{r}_n(\hat{\boldsymbol{\theta}}))^2 \quad . \tag{2.16}$$

Exercise 2.10 Verify that equation (2.16) is correct.

Equation (2.16) has a simple interpretation. To find the noise level, one only has to subtract the data from the (best possible) model $\tilde{r}_n(\hat{\boldsymbol{\theta}})$, and to compute the average of squared differences. This is similar to what one would do intuitively, but here the validity of this approach is proved mathematically, within certain modelling assumptions.

2.3.2 Weighted least squares

One of the assumptions about the noise that has been used until here is that all measurements are done with the same accuracy. In other words, the noise level in all measurements was assumed to be equal to the same value σ. This assumption might be unrealistic in several situations. For instance, measurements might be done with different instruments, with different accuracies. Another example is a situation where data from different modalities are analyzed with the same model. For instance, one might be interested in solving the source localisation problem on the basis of simultaneous MEG and EEG data. In that case, the data unit of MEG is *T* and the unit of EEG is *V* and therefore the numerical value of the noise must be different for different sensors, even if all MEG sensors and all EEG sensors have a comparable relative noise level.
When of each of the *N* measurement the relative noise levels $\sigma_0\sigma_1\sigma_2,\ldots,\sigma_{N-1}$ are known, the same approach as for the uncorrelated case can be followed, after a slight generalization. Here with a relative noise level it is meant that for each measurement, the noise level is known up to a multiplication factor σ. In other words, for sensor *n* the noise level is $\sigma\,\sigma_n$, where σ is unknown, and σ_n is known (e.g. from test measurements that are done prior to the experiment of interest, or from theoretical considerations). We then have for the distribution of the residuals that

$$f_{\boldsymbol{\eta}}(\eta_n) = f_{\boldsymbol{\eta}}(r_n - \tilde{r}_n(\boldsymbol{\theta})) \sim \frac{e^{-\frac{1}{2\sigma^2}\sum_n \left(\frac{r_n - \tilde{r}_n(\boldsymbol{\theta})}{\sigma_n}\right)^2}}{(2\pi)^{N/2}(\sigma)^N \prod_m \sigma_m} \quad , \tag{2.17}$$

see appendix B. Here, $\prod_m \sigma_m$ is the product of all relative noise levels. Note that if all relative noise levels are equal to 1, the same noise distribution function as in equation (2.9) is obtained. Similar to the OLS case, the parameter estimator is derived by finding that $\hat{\boldsymbol{\theta}}$ for which the maximum likelihood of the residuals is maximum. Also here the estimation of $\boldsymbol{\theta}$ and σ are decoupled, because for each σ maximum gain is achieved when the following cost is minimal:

$$\text{Cost}_{\text{WLS}} = \min_{(\boldsymbol{\theta})} \sum_{\text{n}} \left(\frac{r_n - \tilde{r}_n(\boldsymbol{\theta})}{\sigma_n} \right)^2 \quad , \tag{2.18}$$

This cost function is commonly called *weighted least squares*, because the contribution of each measurement to the cost function is weighted by relative noise level of the measurement. The higher the noise level, the lower the contribution and therefore high quality measurements have more impact on the parameter estimate than low quality measurements. When for example MEG and EEG are used simultaneously for the estimation of the dipole parameters, one can use the cost function (2.18), and let n=0,…,N_{MEG}-1 refer to MEG channels and noise levels, whereas n=N_{MEG},…,N-1 would refer to the EEG channels.

The solution $\hat{\boldsymbol{\theta}}_{\text{WLS}}$ of the minimization problem posed in equation (2.18) can be found in a way very similar to the derivation of equation (2.14), provided that the model is linear. Also the noise level can be found in an analogous way. However, the solution will be presented in the next section, wherein the parameter estimation problem is generalized further.

2.3.3 Generalized least squares

One of the assumptions made in the previous section is still that the noise is uncorrelated from one measurement to another. When the measurements represent subsequent time samples, this assumption is not always realistic because the signal of interest might be corrupted by a noise that tends to take on the same value from one sample to the next. For example, if one records an evoked potential using an EEG system with a sampling frequency of 200 Hz, the potential might be interfered by ongoing EEG activity such as the 10 Hz alpha rhythm. When the ongoing activity, including the alpha rhythm is considered as noise, the noise is correlated over nearby time samples, as well as over time samples that are recorded with a 100 ms interval. In this case, a parameter estimation problem based on the EEG data should treat the background noise as correlated noise. However, correlated noise is not constrained to the temporal domain. It also plays a role in the spatial domain, e.g. in the dipole fitting problem based on MEG and/or EEG data. Since after averaging responses to a train of identical stimuli the noise mainly consists residual background noise of the ongoing MEG/EEG, this background noise is correlated over sensors. This is so, because the generators of the background noise have an influence on multiple MEG/EEG sensors, and therefore nearby sensors record similar noise signals. Therefore, also in the source localisation problem one would ideally treat the background noise as correlated noise and the OLS estimator would result in sub-optimal estimates.

The Maximum Likelihood estimation paradigm can also be applied to cases where the noise is correlated. The main idea remains that the unknown parameters are sought such that the residuals (data minus model) are as likely as possible. In order to apply this principle to the case of correlated noise, one has to write down an expression for the distribution of correlated noise. This is done for the case of Gaussian noise in equation (2.19) (see appendix B):

$$f_{\boldsymbol{\eta}}(\boldsymbol{\eta}) = f_{\boldsymbol{\eta}}(\mathbf{r} - \tilde{\mathbf{r}}(\boldsymbol{\theta})) \sim \frac{e^{-\frac{(\mathbf{r}-\tilde{\mathbf{r}}(\boldsymbol{\theta}))^T C^{inv} (\mathbf{r}-\tilde{\mathbf{r}}(\boldsymbol{\theta}))}{2\sigma^2}}}{(2\pi)^{N/2} \sigma^N (\det(C))^{1/2}} \quad . \tag{2.19}$$

Exercise 2.11 Show that (2.17) is a special case of (2.19). Hint: assume that C is a diagonal matrix.

Exercise 2.12 Given N measurements, r_0, r_1,…,$r_{\text{N-1}}$ of which we theoretically assume a constant value. Each measurement is disturbed by uncorrelated Gaussian noise η_{n}, with a standard deviation $\sigma\sigma_{\text{n}}$ and where σ is unknown and σ_{n} are given levels. Derive the ML

estimator for the theoretical constant level and for σ. Hint: $\mathbf{r} = \theta\mathbf{b} + \boldsymbol{\eta}$, what is $\mathbf{b}$, what is the ML estimator of θ?

Here, the covariance matrix of the noise $\boldsymbol{\eta}$ equals $\sigma^2 C$, where σ is an unknown noise level parameter and C is the **known** covariance pattern (i.e. the noise covariance apart from a common scaling factor σ^2). In practice, this pattern is determined from calibration experiments or it is based on theoretical assumptions that predict a certain covariance. In the MEG/EEG case for instance, one could record some spontaneous brain activity without applying a stimulus. From these data one would estimate the covariance matrix and then, in the evoked potential experiment one would assume that the background noise has the same correlation pattern, but with a possibly different level σ.
Alternatively, the covariance pattern can be known from theoretical considerations. In the time domain it is often very realistic to assume that the noise is *stationary*. This implies that the noise characteristics of two measurements at t_1 and t_2 only depend on the time difference t_1-t_2. For regularly sampled signals this means that the noise covariance between the the j_1-th and the j_2–th sample, only depends on j_1-j_2. Therefore, the covariance matrix of a stationary process has the following form

$$C_{stationary} = \begin{pmatrix} c_0 & c_1 & c_2 & & c_{N-1} \\ c_1 & c_0 & c_1 & c_2 & \\ & c_1 & c_0 & & \\ & & & & c_1 \\ c_{N-1} & & & c_1 & c_0 \end{pmatrix} \quad , \qquad (2.20)$$

i.e. all sub diagonals are constant. Such a matrix is called a (symmetric) *Toeplitz matrix*. The first row of this matrix c_j is the correlation time function and this completely specifies the rest of the matrix. Usually, the correlation will decrease for samples separated further and further in time. When these long range correlations are neglected, a *banded* Toeplitz matrix is obtained

$$C_{banded} = \begin{pmatrix} c_0 & c_1 & c_2 & & 0 \\ c_1 & c_0 & c_1 & c_2 & \\ c_2 & c_1 & c_0 & & \\ 0 & & & & c_1 \\ 0 & 0 & & c_1 & c_0 \end{pmatrix} \quad . \qquad (2.21)$$

Many different methods are described in the literature to extract estimates of banded Toeplitz covariance matrices from noise measurements.
The notion of noise stationarity of the noise can also be applied in the space domain. Considering the spatial covariance of MEG sensors, one could argue that correlation between channels is mainly caused by ongoing brain activity, whereby random interactions at many different synapses each cause a magnetic field at multiple neighbouring sensors. This situation is sketched in figure 2.3 (left) assuming a spherically symmetric head. Stationarity in the space domain implies that, because of the assumed spherical symmetry, noise covariance only depends on sensor distance and not on the spatial positions of the sensors themselves. If these assumptions are translated into mathematical equations, it becomes possible to predict the noise covariance as a function of sensor distance (figure 2.3 right). In this way, when all sensor positions are known, one can also compute theoretical covariance for all sensor pairs and this can be used to obtain the covariance matrix C. Since it is unknown how many synapses are active and what their

contribution to the noise is, one adds an unknown scaling factor σ^2 in the covariance matrix of the real experiment.

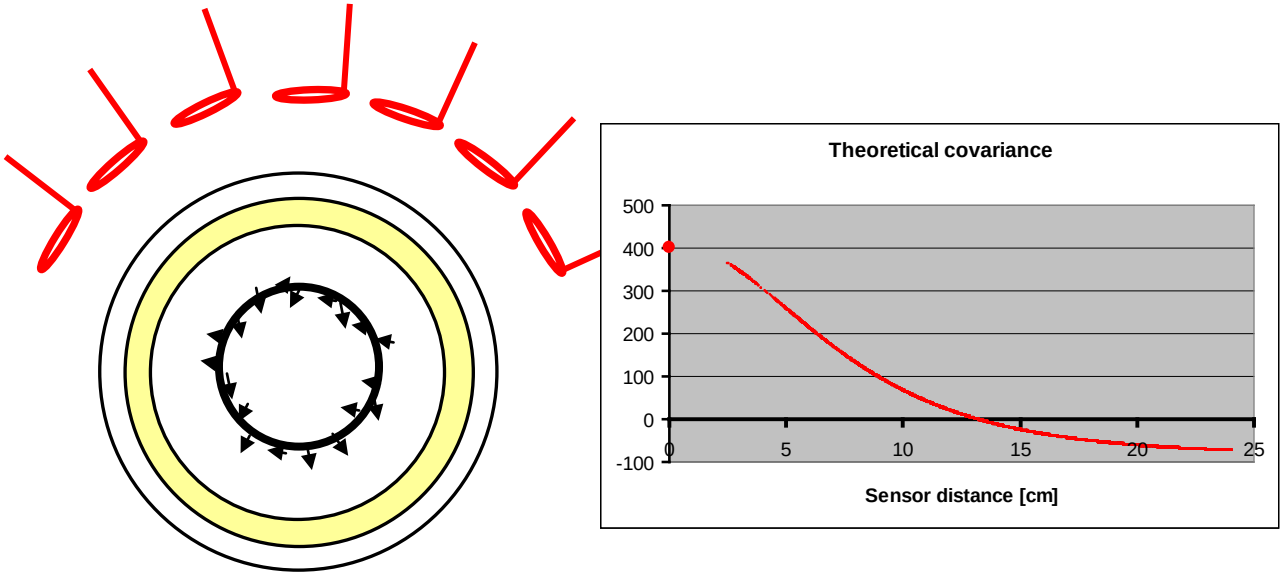

Figure 2.3. *When using MEG or EEG for source localisation, a theoretical model for the spatial covariance matrix can be obtained by assuming that the noise is generated by the ongoing brain activity at the synapses. Each tiny post-synaptic current contributes to the MEG (or EEG) signal (left) and because of the dipolar B-pattern of each of these currents, the noise at neighbouring sensors becomes correlated. With several additional assumptions it becomes possible to predict the noise covariance of sensor distance (right) and to obtain a theoretical noise covariance matrix.*

In case of correlated noise, with known covariance pattern, the parameters and the noise level can be found from the ML criterion by maximizing

$$\text{Gain}_{\text{ML}} = \max_{(\boldsymbol{\theta},\sigma)} \frac{e^{-\frac{(\mathbf{r}-\tilde{\mathbf{r}}(\boldsymbol{\theta}))^T C^{inv} (\mathbf{r}-\tilde{\mathbf{r}}(\boldsymbol{\theta}))}{2\sigma^2}}}{(2\pi)^{N/2}\sigma^N (\det(C))^{1/2}} \quad . \qquad (2.22)$$

This maximisation can again be split into a maximisation over $\boldsymbol{\theta}$ and another one over σ. For all σ fixed, maximisation over $\boldsymbol{\theta}$ is equivalent to the minimisation of

$$\text{Cost}_{\text{GLS}} = \min_{(\boldsymbol{\theta})} \left(\mathbf{r} - \tilde{\mathbf{r}}(\boldsymbol{\theta})\right)^T C^{inv} \left(\mathbf{r} - \tilde{\mathbf{r}}(\boldsymbol{\theta})\right) \quad , \qquad (2.23)$$

This cost function is generally known as *generalized least squares* because it is the generalization of OLS and WLS. For the noise level, one finds

$$\hat{\sigma}^2_{GLS} = \frac{(\mathbf{r} - \tilde{\mathbf{r}}(\hat{\boldsymbol{\theta}}_{GLS}))^T C^{inv} (\mathbf{r} - \tilde{\mathbf{r}}(\hat{\boldsymbol{\theta}}_{GLS}))}{N} \quad , \qquad (2.24)$$

where $\hat{\boldsymbol{\theta}}_{GLS}$ is the solution of the minimization of (2.23).

Exercise 2.13 Derive from equation (2.24) the estimated noise level in case of uncorrelated noise of which the noise level varies over measurements. Hint: how does *C* look like in this case?
Exercise 2.14 Show that equation (2.16) is consistent with equation (2.24).
Exercise 2.15 Assume that the background noise is correlated in time and a banded noise covariance matrix is assumed, where all noise correlations of between measurements separated more than one sample are ignored. How would you estimate both the noise covariance and the unknown parameters using the ML criterion?

To find the GLS solution of the minimization problem posed in equation (2.23), we note that since *C* is positive definite, there exist a matrix *W*, such that

$$C^{-1} = WW^T \quad . \tag{2.25}$$

The matrix *W* is not necessarily unique, i.e. there can be more than one *W* satisfying equation (2.25). What is important here is that at least one matrix *W* can be found. It appears that the decomposition presented in equation (2.25) makes it possible find an expression for the GLS estimator, without the need to know *W* itself.
First, equation (2.25) is inserted in the expression for the cost function equation (2.23) and then both the data vector **r** and the model vector $\tilde{\mathbf{r}}(\boldsymbol{\theta})$ are combined with W^T:

$$\begin{aligned} \text{Cost}_{\text{GLS}} &= \min_{(\boldsymbol{\theta})} \quad (\mathbf{r} - \tilde{\mathbf{r}}(\boldsymbol{\theta}))^T \mathbf{W}\mathbf{W}^T (\mathbf{r} - \tilde{\mathbf{r}}(\boldsymbol{\theta})) \\ &= \min_{(\boldsymbol{\theta})} \quad (\mathbf{W}^T\mathbf{r} - \mathbf{W}^T\tilde{\mathbf{r}}(\boldsymbol{\theta}))^T (\mathbf{W}^T\mathbf{r} - \mathbf{W}^T\tilde{\mathbf{r}}(\boldsymbol{\theta})) \quad , \\ &= \min_{(\boldsymbol{\theta})} \quad (\mathbf{r}'-\tilde{\mathbf{r}}'(\boldsymbol{\theta}))^T (\mathbf{r}'-\tilde{\mathbf{r}}'(\boldsymbol{\theta})) \end{aligned} \tag{2.26}$$

where the primed symbols **r'** and $\tilde{\mathbf{r}}'(\boldsymbol{\theta})$ are abbreviations for $W^{\mathbf{T}}\mathbf{r}$ and $W^{\mathbf{T}}\tilde{\mathbf{r}}(\boldsymbol{\theta})$. Then it appears that the last line in equation (2.26) is identical to the problem posed in equation (2.11). This implies that by modifying the data and similarly modifying the model through the pre-multiplication with W^T, a GLS problem can be transformed into a simpler OLS problem. This modification is usually called *pre-whitening*, because the noise is made "white", or uncorrelated. The consequence of this equivalence is that in the case of a linear parameter model, one can copy the solution of the OLS problem, presented in equation (2.14), to the GLS problem. Note that in case of the linear parameter model the pre-whitened model $W^{\mathbf{T}}\tilde{\mathbf{r}}(\boldsymbol{\theta})$ looks like

$$W^{\mathbf{T}}\tilde{\mathbf{r}}(\boldsymbol{\theta}) = W^{\mathbf{T}}B\boldsymbol{\theta} \quad , \tag{2.27}$$

because in the linear case the relationship between the predicted data and the model parameter consists of a matrix multiplication of model parameters and a fixed matrix *B*. Therefore, if *B'* is defined ass

$$B' = W^{\mathbf{T}}B \quad , \tag{2.28}$$

we find for the solution of the GLS estimator that

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{GLS} &= \left(B'^T\ B'\right)^{inv} B'^T\ \mathbf{r}' \\ &= \left((W^T B)^T W^T B\right)^{inv} \left(W^T B\right)^T W^T \mathbf{r} \\ &= \left(B^T W W^T B\right)^{inv} B^T W W^T \mathbf{r} \\ &= \left(B^T C^{inv} B\right)^{inv} B^T C^{inv} \mathbf{r}\end{aligned} \quad . \qquad (2.29)$$

Here the first step consists of replacing the data and model symbols of equation (2.14) by the primed ones of the GLS problem. Then, the primed symbols are replaced by $W^T\mathbf{r}$ and $W^T B$ and finally pairs of pre-whiting matrices are taken together to form the covariance matrix C. Surprisingly, the expression for the parameters estimated by GLS does not contain the pre-whiting matrix itself, but only the (inverse of) the noise covariance matrix C.

Exercise 2.16 Use equation (2.29) to find an expression for the WLS estimator of $\boldsymbol{\theta}$.

Because of the very powerful pre-whitening argument presented above, we will from here on only consider ML estimation under the assumption that the background noise is uncorrelated. If in a practical example it appears to be more realistic to assume correlated noise, and if an adequate distribution for the noise covariance is available, we can always switch to correlated noise by pre-whiten data and the model, similar to what is done in equation (2.27) and (2.29).

2.3.4 Geometrical interpretation of linear parameter fitting

In appendix A a brief review of the basics of matrix vector algebra is given. This algebra is very useful in order to derived OLS or GLS estimators in very condensed form, as can e.g. be appreciated when comparing equations (2.3) to (2.14). Moreover, the representation of a discrete set of measurements or model predictions in the form of a higher dimensional vector, can often facilitate the interpretation of many of the formulas we need in the theory of parameter estimation. For instance, if $\mathbf{a}=(a_0, a_1,\ldots,a_{N-1})$, then the root mean square (RMS) of this signal is $|\mathbf{a}| = \sqrt{\sum_n a_n^2}$, which is the same as the length of the **a**, when interpreted as an N-dimensional vector. For zero mean signals **a** and **b** one can compute the correlation co-efficient as

$$\mathrm{corr}(\mathbf{a},\mathbf{b}) = \frac{\sum_n a_n b_n}{\sqrt{\sum_n a_n^2}\sqrt{\sum_n b_n^2}} = \frac{\mathbf{a}^T\mathbf{b}}{|\mathbf{a}||\mathbf{b}|} = \cos(\omega_{\mathbf{a},\mathbf{b}}) \quad , \qquad (2.28AA)$$

where $\omega_{\mathbf{a},\mathbf{b}}$ is the angle between the vectors **a** and **b**, observed from the origin. When **a** and **b** are not zero mean, the same analogy is obtained by first applying a projection operator on (which remove the averages) and computing the cosine angle between the projected versions of **a** and **b**. Figure 2.4 presents the geometrical analogue of the computation of the OLS estimate in the linear case in one or two dimensions. In figures (2.4A) and (2.4B) the data **r** and the regressor to fit **b** are represented as a graph (vector component as function of n) whereas in figures (2.4C) and (2.4D) they are depicted in abstract vector space. From figure (2.4C) it becomes clear that the OLS estimation is obtained by determining how much the vector **b** has to be extended in order to get as close to the data as possible. Obviously, the best point is obtained by *projecting* the point **r** at right angles onto the vector **b**. In more than one dimensions (figure 2.4D), one considers the hyperplane spanned by the columns of B (equation 2.12) and projects the point **r** onto that hyperplane. Equation (2.14) implies that the OLS estimator of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}} = (B^T B)^{inv} B^T \mathbf{r}$

and therefore, within the model assumptions, the closest point to **r**, is $B\hat{\boldsymbol{\theta}} = B(B^T B)^{inv} B^T \mathbf{r}$ and therefore also we may conclude that the projection operator that projects the point r onto the requested hyperplane is given by the square matrix $B(B^T B)^{inv} B^T$.

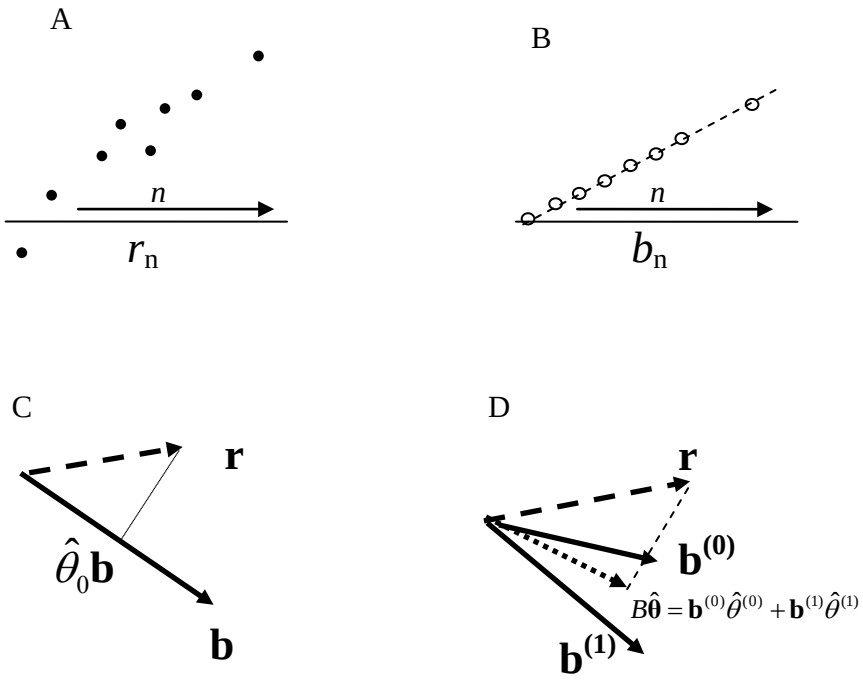

Figure 2.4. *The fit of a one parameter(θ_0) linear model to a set of data points r_n can be given a simple graphical interpretation when the OLS criterion is used. In panel A and B the data and model vectors are represented in the "direct" way: value as function of measurement number n. The parameter estimation problem implies by which factor θ_o the model values have to be multiplied such that they get as close as possible to the data. In panel C and D the data and model vectors are represented as "abstract" vectors in an N dimensional space. One observes that the solution of the parameter estimation problem corresponds to that point on the line through **b** that is closest to the point **r**. That point can be found by applying an orthogonal projection of **r** onto the line through **b** (panel C). For higher dimensional linear problems (panel D), a similar analogue can be formulated. Instead of project **r** onto a line through **b**, **r** is projected into the (hyper-)plane through the columns of B = (**b**$^{(0)}$, **b**$^{(1)}$, …).*

> **Exercise 2.17**. In analogy of figure 2.4, give a graphical representation of the difference of computing the best linear fit of **b** to **r** compared to **r** to **b** (cf. **Exercise 2.2**). Under which conditions are the differences between both options expected to be the largest? Verify with a numerical example that this expectation is correct.

2.3.5 Final remarks

In this chapter several implicit and explicit assumptions are made which are used to derive estimators of the parameters of interest. These assumptions are the foundations on which the theory is built. However, it is good to realize that these assumptions are not the only sensible assumptions. There are alternatives that should be considered when the present ones do not suffice are not applicable.

1. In equation (2.7), the parameters $\boldsymbol{\theta}$ are viewed as *deterministic* and fixed parameters that adopt a "true", yet unknown value. Only the noise is random and we only specify a probability distribution for the noise. The parameter estimation procedure eventually yields an estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, i.e. an algorithm is derived yielding $\hat{\boldsymbol{\theta}}$ as a function of the recorded data. But since the data is contaminated with noise, also $\hat{\boldsymbol{\theta}}$ must be considered as a random variable. The distribution of $\hat{\boldsymbol{\theta}}$ is completely determined by our assumed distribution of the noise.
   There are alternative approaches, in which also $\boldsymbol{\theta}$ is considered as a random variable. In one of those paradigms, called Bayesian estimation, the investigator has *a priori* knowledge of $\boldsymbol{\theta}$ in the form of a probability density distribution. Once the measurements are done, the a priori distribution yields an *a posteriori* distribution, due to the new information provided by the measurements.
   There is no general agreement with regards to the applicability of the Bayesian philosophy to parameter estimation problems. Some argue that any available knowledge should be added as a priori information, even if it is in the form of a probability density distribution of the parameters of interest. In this way, one would reach the strongest possible conclusion from the available data. Others argue that a priori information in the form of a probability distribution is subjective (because this distribution is not experimentally accessible), and such information will never lead to a solid conclusion.

2. In equation (2.7) only measurement errors that are *superimposed* on the model are considered. However, this model does not cover all practical situations. Suppose for instance, that one records the electrical potential on a set of *N* different electrodes, and one aims to find the underlying dipole source by fitting a dipole model onto these data. Then for each electrode, the data model $\tilde{r}_n(\boldsymbol{\theta})$ would predict the potential of a current dipole, recorded at electrode *n*. When equation (2.7) is applied, one only considers errors in the recorded potentials, whereas errors in the electrode positions are ignored. A more explicit description of the data would be given by a model $\tilde{r}(\mathbf{x}_n + \boldsymbol{\varepsilon}_n, \boldsymbol{\theta})$, where $\mathbf{x}_n$ is the *n*-th electrode position, and $\boldsymbol{\varepsilon}_\mathbf{n}$ is the electrode position error. However, the analysis of such model is much more complicated than the situation described by equation (2.7), even when the electrode position errors are assumed to have a Gaussian distribution. Therefore, despite its relevance, situations where also the dependent variable contains a random error are ignored.
   In the linear case, and in particular in the line fitting example, solution methods are available where the minimization of the deviation of the theoretical line from the measurement point is both in the *x*- and *y*-direction. The problem is known as *total least squares*.

3. In the derivation of the parameter estimates, the assumption was made that one knows the covariance matrix of the noise, apart from a multiplication factor σ^2. Such assumption simplifies the derivation, but in some cases no reasonable assumption on *C* can be made. For instance, in the case of MEG/EEG one is investigating the underling sources of an evoked potential due to a visual stimulus. In this case the ongoing MEG/EEG signals are the most important sources of noise. One could determine the covariance of the background noise by doing a separate experiment wherein the subject is not stimulated to obtain an estimate of the covariance of the noise. Then this covariance matrix could be used as input of the GLS estimator when the evoked potentials are analysed. However, it is questionable whether such an approach is correct, because the ongoing MEG/EEG, and the alpha rhythm in particular, is also influenced by the presence of a visual stimulus, and this influence takes place in an unknown way. Therefore, the covariance matrix determined in the rest condition might be non-representative for the condition where subject processes a visual stimulus. A better

approach would be to extract the covariance pattern form the raw (unaveraged) evoked potential data, simultaneously with the dipole source parameters. In this approach, the ML paradigm can still be followed, because one can also maximise the likelihood of the observed residuals by varying over C (or a parameterisation of it), instead of only varying over σ and θ. However, in that case the denominator of (2.20) is no longer fixed and the nice separation of σ and θ gets lost, resulting in a much more difficult maximisation problem. Moreover, many of the statistical tests derived in the sequel are not (strictly) valid anymore.

3 Properties of the estimated parameters. Linear models.

In the previous chapter it was discussed how model parameters can be estimated from observed data. Based on several precise assumptions, the estimated parameters are optimal in the sense of maximum likelihood of the resulting residuals. Although this optimality property gives a guarantee that the data are treated in the best possible way, so far we still lack a measure of reliability of the estimated parameters. In particular, we would like to extract some kind of reliability range around each estimated parameter, such that we are sure that it is "very likely" that the true parameter is within this range. When applied to the source localisation problem based on MEG/EEG, the parameter ranges could be interpreted as error measures. In other applications, such as fMRI, we would apply the parameter range to test whether the interval contains zero. If that is not the case, we would conclude that there is a "significant effect", i.e. a significant deviation from zero, or equivalently, as we shall see, a significant correlation between fMRI signal and the applied stimulus.

The mathematical modelling approach of parameter estimation provides a way to estimate, in addition to each parameter, also a *confidence interval*. In particular when the model parameters are linear, these confidence intervals and their statistical meaning is relatively easy to obtain. For that purpose, the estimated parameter $\hat{\boldsymbol{\theta}}$ is considered as random variable, similar to the noise part of the model. So, the true parameters $\boldsymbol{\theta}$ are fixed, deterministic and unknown parameters, but their estimates $\hat{\boldsymbol{\theta}}$ are influenced by the noise and therefore have a certain statistical distribution. The statistical distribution of $\hat{\boldsymbol{\theta}}$ depends on the method that was used to estimate it. In case of uncorrelated noise and the OLS estimator (equation (2.14)), one can express $\hat{\boldsymbol{\theta}}_{\mathrm{OLS}}$ as

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{OLS} &= (B^T B)^{inv} B^T \mathbf{r} \\ &= (B^T B)^{inv} B^T (B\boldsymbol{\theta} + \boldsymbol{\eta}) \\ &= \boldsymbol{\theta} + (B^T B)^{inv} B^T \boldsymbol{\eta}\end{aligned} \quad . \qquad (3.1)$$

Therefore, to study the distribution of $\hat{\boldsymbol{\theta}}$, one needs to consider how the noise propagates through the estimator. Since $\boldsymbol{\eta}$ has a multivariate Gaussian distribution with zero mean, and covariance matrix $\sigma^2 I_{\mathrm{N}}$ and because $\hat{\boldsymbol{\theta}}_{\mathrm{OLS}}$ is a linear transformation of $\boldsymbol{\eta}$ (see 3.1), we can apply the theory of appendix B to obtain the distribution of $\hat{\boldsymbol{\theta}}_{\mathrm{OLS}}$. Because of the linear transformation $\hat{\boldsymbol{\theta}}_{\mathrm{OLS}}$ also has a Gaussian distribution. In particular, we can obtain its expected value and covariance matrix. For the expected value of the estimated parameter vector one obtains

$$\mathrm{E}\{\hat{\boldsymbol{\theta}}_{OLS}\} = \boldsymbol{\theta} + (B^T B)^{inv} B^T \mathrm{E}\{\boldsymbol{\eta}\} = \boldsymbol{\theta} \quad . \qquad (3.2)$$

Equation (3.2) implies that the OLS estimator is *unbiased*, i.e. its expected value equals the true value. In other words, if the whole experiment would be repeated over and over again, with different noise realisations, the mean of the OLS estimators in all experiments would converge to the true one.

The covariance matrix of the estimated parameter vector can be found by starting of from its definition, and substituting (3.1). In this way one obtains

$$
\begin{aligned}
\mathrm{Cov}(\hat{\boldsymbol{\theta}}_{OLS}) &\equiv \mathrm{E}\left\{\left(\hat{\boldsymbol{\theta}}_{OLS}-\boldsymbol{\theta}\right)\left(\hat{\boldsymbol{\theta}}_{OLS}-\boldsymbol{\theta}\right)^{T}\right\} \\
&= (B^{T}B)^{inv}B^{T}\mathrm{E}\{\boldsymbol{\eta}\boldsymbol{\eta}^{T}\}B(B^{T}B)^{inv} \\
&= (B^{T}B)^{inv}B^{T}\sigma^{2}I_{N}B(B^{T}B)^{inv} \\
&= \sigma^{2}(B^{T}B)^{inv}B^{T}B(B^{T}B)^{inv} \\
&= \sigma^{2}(B^{T}B)^{inv}
\end{aligned}
\quad . \qquad (3.3)
$$

Equation (3.3) gives the covariance matrix of the estimated parameters, and in particular the diagonal of this matrix gives the variance of the estimated parameters. As argued in appendix B, the square roots of these variances can be interpreted as the standard deviations of the estimated parameters.

> **Exercise 3.1** Give a vector matrix expression of the GLS estimator of $\boldsymbol{\theta}$ in case of correlated noise with covariance matrix C.
> **Exercise 3.2** Suppose that despite the fact that the noise is correlated in reality, the parameter vector $\boldsymbol{\theta}$ is estimated using the OLS algorithm. Does this way of treating the data result in a biased estimator of $\boldsymbol{\theta}$? Hint: investigate what will change in the analysis of (3.1) and (3.2) when noise is correlated.

Despite the relatively simple argument underlying (3.2) and (3.3), it should be kept in mind that equations (3.2) and (3.3) and their interpretations are strictly speaking only valid within all modelling assumptions: linear model, no systematic errors, random errors appear only in **r**, Gaussian distribution of noise and noise is uncorrelated. When one of these assumptions is violated, conclusions should be reconsidered.
Furthermore, the applicability of (3.3) to derive confidence intervals is limited because it still depends on the noise level σ. Only in situations where this noise level is known, we can apply (3.3). In other cases, we would have to estimate σ first. Such an estimate has already been derived in (2.16), where the data and model prediction have been subtracted, squared and averaged. To obtain an expression in which the dependence on $\hat{\boldsymbol{\theta}}_{\mathrm{OLS}}$ is eliminated, equation (2.16) is rewritten in matrix form and the expression for $\hat{\boldsymbol{\theta}}_{\mathrm{OLS}}$ (2.14) is substituted:

$$
\begin{aligned}
\hat{\sigma}^{2}_{OLS} &= \frac{1}{N}\sum_{\mathrm{n}}(r_n-\tilde{r}_n(\hat{\boldsymbol{\theta}}))^{2} \\
&= \frac{\left(\mathbf{r}-B\hat{\boldsymbol{\theta}}\right)^{T}\left(\mathbf{r}-B\hat{\boldsymbol{\theta}}\right)}{N} \\
&= \frac{\left(\mathbf{r}-B(B^{T}B)^{inv}B^{T}\mathbf{r}\right)^{T}\left(\mathbf{r}-B(B^{T}B)^{inv}B^{T}\mathbf{r}\right)}{N} \\
&= \frac{\left(\mathbf{r}^{T}-\mathbf{r}^{T}B(B^{T}B)^{inv}B^{T}\right)\left(\mathbf{r}-B(B^{T}B)^{inv}B^{T}\mathbf{r}\right)}{N} \\
&= \frac{\mathbf{r}^{T}\left(I-B(B^{T}B)^{inv}B^{T}\right)\mathbf{r}}{N} \\
&= \frac{\mathbf{r}^{T}P_{B}\,\mathbf{r}}{N}
\end{aligned}
\quad , \qquad (3.4)
$$

where it has been used that P_B with

$$P_B \equiv I\text{-}B(B^T B)^{inv} B^T \quad , \quad (3.5)$$

is a symmetric projection matrix (see appendix A). The advantage of ((3.4), compared to (2.16), is that the noise level estimate is directly expressed in known quantities: the data and the model matrix, without the intermediate step of $\hat{\boldsymbol{\theta}}_{\mathrm{OLS}}$.

We would now to substitute equation (3.4), which is an estimator for σ, into (3.3) in order to obtain an *estimated* covariance of the e*stimated parameters*. The problem is however, that the noise estimator presented in (3.4) is *biased*. To demonstrate that the ML estimator of the noise level is indeed biased, we can proceed very similarly to the way the distribution of $\hat{\boldsymbol{\theta}}_{\mathrm{OLS}}$ was analyzed. Therefore, the model ($\mathbf{r}=B\boldsymbol{\theta}+\boldsymbol{\eta}$) is substituted into the estimator $\hat{\sigma}^2{}_{\mathrm{OLS}}$ and then we consider the expected value of $\hat{\sigma}^2{}_{\mathrm{OLS}}$. Starting from (3.4) one obtains, and using that $P_B B$=0 (why?),

$$\begin{aligned}
\mathrm{E}\left\{\hat{\sigma}^2_{OLS}\right\} &= \mathrm{E}\left\{\frac{\mathbf{r}^T P_B\,\mathbf{r}}{N}\right\} \\
&= \mathrm{E}\left\{\frac{\left(\boldsymbol{\eta}^T + \boldsymbol{\theta}^T B^T\right) P_B \left(B\boldsymbol{\theta}+\boldsymbol{\eta}\right)}{N}\right\} \\
&= \frac{1}{N}\mathrm{E}\left\{\boldsymbol{\eta}^T P_B \boldsymbol{\eta}\right\} \\
&= \frac{1}{N}\mathrm{E}\left\{\sum_{ij}\left(P_B\right)_{ij}\eta_i\eta_j\right\} \\
&= \frac{\sigma^2}{N}\mathrm{E}\left\{\sum_{i}\left(P_B\right)_{ii}\right\} \\
&= \frac{\sigma^2}{N}\mathrm{Tr}\left\{P_B\right\} \\
&= \sigma^2\,\frac{N-M}{N}
\end{aligned} \quad , \quad (3.6)$$

where in the fourth line it has been used that, since the noise is uncorrelated, that $\mathrm{E}\{\eta_i\eta_j\}=\sigma^2\delta_{ij}$. This property transfers the double sum into a single sum over the diagonal elements of P_B i.e. the trace of P_B. In appendix A it is discussed that the trace of a projection matrix equals its *rank*, which is, assuming that B has M linearly independent columns, equal to N-M. The geometrical interpretation thereof is that a projection operator puts all N-dimensional vectors in a subspace (hyper-plane) of smaller than N dimensions. In 3D, a (non-trivial) projector will map 3D space on a line or onto a plane. Similarly, projector P_B removes all "parts" of a vector that are linearly dependent on the columns of B.

The result of this analysis is, since M>0, that $\mathrm{E}\left\{\hat{\sigma}^2{}_{\mathrm{OLS}}\right\} \neq \sigma^2$ and that the OLS estimated noise variance is systematically a little bit too small. To correct for that bias, we use $\hat{\sigma}^2{}_{unbiased}$

$$\hat{\sigma}^2_{unbiased} = \frac{\mathbf{r}^T P_B\,\mathbf{r}}{N-M} \quad , \quad (3.7)$$

which is not the ML estimator, but which is unbiased. When N is large compared to M the practical consequences of using $\hat{\sigma}^2{}_{\text{unb}}$ instead of $\hat{\sigma}^2{}_{\text{OLS}}$ are small. Nevertheless, from here on we will use (3.7) as noise estimate and the subscript "unbiased" will generally be omitted.
When (3.7) is used as an estimate of the noise, one can express the estimated covariance of $\hat{\boldsymbol{\theta}}_{\text{OLS}}$ as

$$\text{EstCov}(\hat{\boldsymbol{\theta}}_{OLS}) \equiv \hat{\sigma}^2_{unbiased} (B^T B)^{inv} = \frac{\mathbf{r}^T P_B \mathbf{r}}{N-M} (B^T B)^{inv} \quad . \qquad (3.8)$$

Equation (3.8) is a matrix that gives a quality check for all estimated parameters. In particular, all estimated standard deviations squared are present on the diagonal of this matrix:

$$\left(\text{EstCov}(\hat{\boldsymbol{\theta}}_{OLS})\right)_{m1,m2} = \begin{pmatrix} \hat{\sigma}^2_{\theta 0} & \text{EstCov}_{\theta 0,\theta 1} & \text{EstCov}_{\theta 0,\theta 2} & \cdots & \\ \text{EstCov}_{\theta 1,\theta 0} & \hat{\sigma}^2_{\theta 1} & & & \\ \vdots & \ddots & \hat{\sigma}^2_{\theta 2} & & \\ & & & & \vdots \\ & & & \cdots & \hat{\sigma}^2_{\theta M-1} \end{pmatrix} \quad . \qquad (3.9)$$

In particular, to compute the estimated standard deviation of the first parameter $\hat{\theta}_0$ one uses

$$\left(\hat{\sigma}_{\hat{\theta}_0}\right)^2 \equiv \frac{\mathbf{r}^T P_B \mathbf{r}}{N-M} \left\{(B^T B)^{inv}\right\}_{0,0} \quad . \qquad (3.10)$$

Note that these expressions only give an *estimate* of the variance of the parameters, because the true noise level σ is not known. The non-diagonal matrix elements determine the covariation of the different estimated parameters and together they determine the confidence ellipse.
What is still lacking is a precise statistical meaning of these error measures. In other words, with the theory explained so far, one cannot yet test with which probability the estimated parameter deviates from the true one. Furthermore, there are many situations in which one is only interested in a subset of the estimated parameters. The other parameters, usually named *nuisance parameters*, have to be added to obtain a complete description of the data and to avoid systematic errors. For instance, one is interested in the determination of a linear increase of s_n as a function of t_n, but measurement conditions cause a substantial and unknown offset in the data. In this case, a straight line is fit through the data points and the slope acts as the parameter of interest, and the offset is the nuisance parameter. In another situation, one is interested in an equilibrium level that a system assumes, but an unknown drift in the data precludes the use of a simple averaging procedure to estimate the equilibrium level. One can apply the same line fit procedure, but this time with the offset as parameter of interest and the slope as nuisance parameter.
In brain imaging many examples of the use of nuisance parameters can be given. For instance, with fMRI activation studies, slow trends in the data that need to be removed when correlating fMRI signals to the applied activation curve. The shape of these trends are known (constant, linear, proportional to the respiratory signal) but the sizes of these trends are unknown and need to be extracted from the data by including them as nuisance parameters in the mathematical model describing the data.
A little more formally, the effect of the noise on the estimated parameters can be described as if the estimated parameter vector $\hat{\boldsymbol{\theta}}$ is a random variable with a certain statistical distribution $f_{\hat{\boldsymbol{\theta}}}(\hat{\boldsymbol{\theta}})$. To focus on the parameters of interest, we integrate away the nuisance parameters. For

example, if there are two parameters x and y, and y is the nuisance parameter, we consider the distribution function $f_x(x)$, with

$$f_x(x) = \int_{-\infty}^{\infty} f_{xy}(x, y)dy \quad . \tag{3.11}$$

This situation is further illustrated in figure 3.1. It appears that in case of Gaussian noise in combination with linear models the integration of nuisance parameters is very easily treated when the parameter vector is split into parameters of interest and nuisance parameters right from the start.

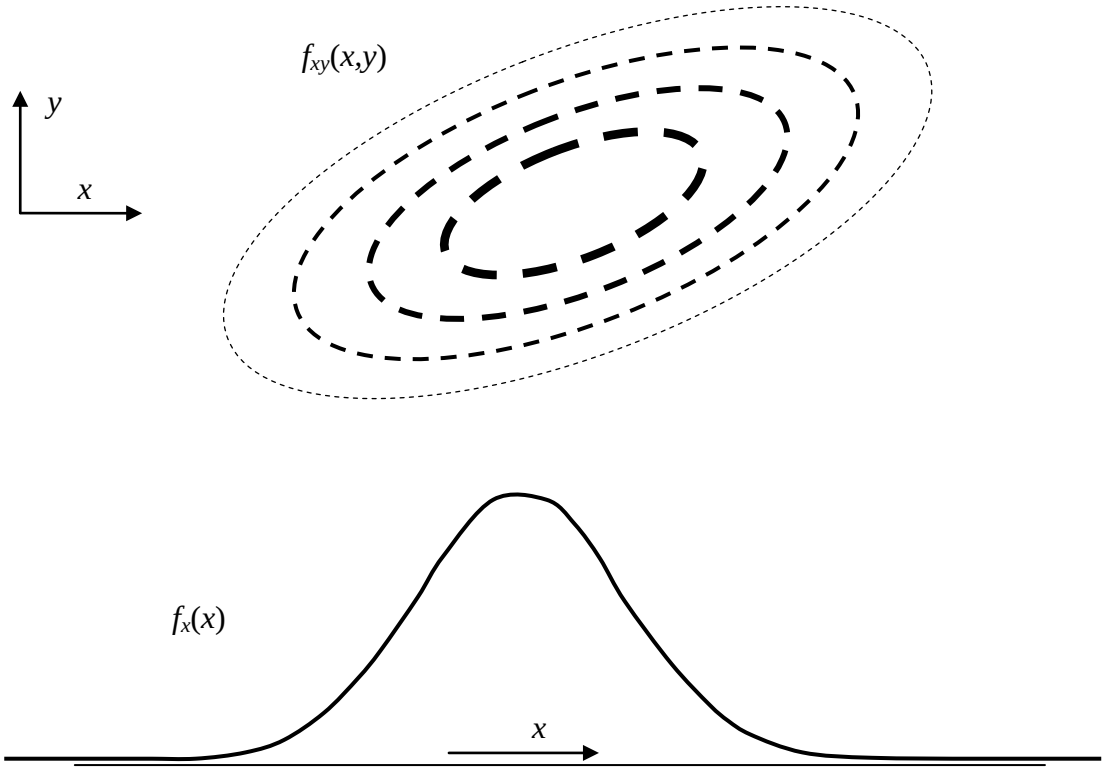

Figure 3.1. *Top. The (Gaussian) distribution of x and y is sketched. Here thicker lines represent higher values. Bottom. The distribution of x, thereby ignoring the effect of y, is obtained by integration $f_{xy}(x,y)$ over y from -∞ to +∞.*

3.1 A single parameter of interest

When there is a single parameter of interest θ and L nuisance parameters φ_0, φ_1, φ_2,…,φ_{L-1} the linear parameter model can be expressed as

$$r_n = \theta b_n + \sum_{l=0}^{L-1} S_{nl}\varphi_l + \eta_n$$

or , n=0,…N-1. (3.12)

$$\mathbf{r} = \theta\mathbf{b} + S\boldsymbol{\varphi} + \boldsymbol{\eta}$$

In the context of fMRI data analysis, b_n would indicate the predicted activation function at time n, θ is the strength of the effect, and S_{n0}, S_{n1}, S_{n2}, $S_{n,L-1}$ are offsets and linear and quadratic trends in the data, each with an unknown strength of φ_0, φ_1, φ_2,…, φ_{L-1}. The nuisance parameter $\boldsymbol{\varphi}$ must be included in the data model to allow a proper estimation of the parameter(s) in interest (in this case θ), but they do not have a (physiological) interest of themselves.
Parameters of (3.12) could be estimated by storing all unknown parameters in a single parameter vector $\boldsymbol{\theta}=(\theta, \varphi_0, \varphi_1, \varphi_2,\ldots, \varphi_{L-1})$ and storing all coefficients in a single matrix $B = (\mathbf{b}, S)$ (the columns of S preceded by an extra column vector $\mathbf{b}$) and extracting the first component of the estimated vector $\hat{\boldsymbol{\theta}}$ to obtain the ML-estimator of the parameter of interest θ. We would only apply (2.15) and be finished. Although the result is correct, this approach does not provide insight into the role of the nuisance parameters.

> **Exercise 3.3** Show that in the example of exercise 2.1 it matters that the nuisance parameter is included into the model. Hint: what estimator for v is obtained when it is assumed that p has a fixed value p_0 (for example p_0=0).

By setting the partial derivatives of the OLS cost function corresponding to (3.12) to zero, one finds the following estimate for θ and for $\boldsymbol{\varphi}$

$$\hat{\theta} = \frac{\mathbf{b}^T\mathbf{r} - \mathbf{b}^T S\left(S^T S\right)^{inv} S^T\mathbf{r}}{\mathbf{b}^T\mathbf{b} - \mathbf{b}^T S\left(S^T S\right)^{inv} S^T\mathbf{b}}$$

$$\hat{\boldsymbol{\varphi}} = \left(S^T(I - \mathbf{b}(\mathbf{b}^T\mathbf{b})^{inv}\mathbf{b}^T)S\right)^{inv}\left(S^T(I - \mathbf{b}(\mathbf{b}^T\mathbf{b})^{inv}\mathbf{b}^T)\mathbf{r}\right) \quad . \qquad (3.13)$$

These expressions can be simplified to

$$\hat{\theta} = \frac{\mathbf{b}^T P_{\mathbf{S}}\mathbf{r}}{\mathbf{b}^T P_{\mathbf{S}}\mathbf{b}}$$

$$\hat{\boldsymbol{\varphi}} = \left(S^T P_{\mathbf{b}} S\right)^{inv}\left(S^T P_{\mathbf{b}}\mathbf{r}\right) \qquad (3.14)$$

using the following abbreviations

$$P_{\mathbf{S}} = I_N - S\left(S^T S\right)^{inv} S^T$$
$$P_{\mathbf{b}} = I_N - \frac{\mathbf{b}\mathbf{b}^T}{\mathbf{b}^T\mathbf{b}} \quad . \qquad (3.15)$$

> **Exercise 3.4**. Suppose that no nuisance parameters are accounted for in the model, e.g. assumed that $\mathbf{r}=\theta\mathbf{b}+\boldsymbol{\eta}$. Show that the corresponding ML-estimator of θ is given by (3.16) and compare this to (3.14)

$$\hat{\theta} = \frac{\mathbf{b}^T\mathbf{r}}{\mathbf{b}^T\mathbf{b}} \quad . \qquad (3.16)$$

The matrices $P_{\mathbf{S}}$ and $P_{\mathbf{b}}$ in (3.14)-(3.15) are symmetric projection matrices, or projectors. They have the properties that (see appendix A)

$$\begin{aligned}(P_S)^2 &= P_S = (P_S)^T \\ (P_{\mathbf{b}})^2 &= P_{\mathbf{b}} = (P_{\mathbf{b}})^T\end{aligned} \quad . \qquad (3.17)$$

These projection matrices have the effect that, when applied upon a vector (interpreted as a signal), to remove all components that a linearly dependent of a certain set of vectors. For instance, with $P_{\mathbf{b}}$ this set of vectors consist only of **b**, and $P_{\mathbf{b}}$ will remove all components proportional to the vector **b**. If $\mathbf{b_0}=(1,1,\ldots,1)^T$, $P_{\mathbf{b0}}$ will remove the offset, and if $\mathbf{b_1}=(0,..,n,\ldots,N-1)^T$, $P_{\mathbf{b1}}$ will remove the linear trend. To remove both trend and offset, a projector P_S can be constructed, where S consists of two column vectors $\mathbf{b_0}$ and $\mathbf{b_1}$: $S = (\mathbf{b_0}, \mathbf{b_1})$.

Exercise 3.5 Show that

$$P_S S = 0 \quad . \qquad (3.18)$$

Explain this result.

3.1.1 Correspondence between nuisance effects and correlated noise

With these properties of projectors in mind, we can give the following interpretation of equation (3.14):

$$\begin{aligned}\hat{\theta} &= \frac{\mathbf{b}^T P_{\mathbf{S}} P_{\mathbf{S}} \mathbf{r}}{\mathbf{b}^T P_{\mathbf{S}} P_{\mathbf{S}} \mathbf{b}} \\ &= \frac{\mathbf{b}^T P_{\mathbf{S}}{}^T P_{\mathbf{S}} \mathbf{r}}{\mathbf{b}^T P_{\mathbf{S}}{}^T P_{\mathbf{S}} \mathbf{b}} \\ &= \frac{(P_{\mathbf{S}}\mathbf{b})^T P_{\mathbf{S}} \mathbf{r}}{(P_S \mathbf{b})^T P_{\mathbf{S}} \mathbf{b}} \\ &= \frac{\mathbf{b'}^T \mathbf{r'}}{\mathbf{b'}^T \mathbf{b'}}\end{aligned} \quad , \qquad (3.19)$$

with $\mathbf{b'} \equiv P_S \mathbf{b}$ and $\mathbf{r'} = P_S \mathbf{r}$. So the estimator of the θ parameter in case of nuisance effects (modelled as columns of the matrix S) can simply be computed by first removing nuisance effects from both the data ($\mathbf{r'} \equiv P_S \mathbf{r}$) and from the model ($\mathbf{b'} \equiv P_S \mathbf{b}$), and proceeding with these modified vectors as if no nuisance effects were present. It appears that this procedure is equally valid in case of multiple parameters of interest.

In this respect, the situation is very similar to the generalisation from OLS to GLS, where correlated noise was accounted for by first pre-whitening the data and pre-whitening the model, and then applying the OLS formulas to the pre-whitened vectors. Comparing both situations in detail, one observes that the whitening matrix W, i.e. "square root" of the inverse of the noise covariance matrix, plays a very similar role as the nuisance projection matrix P_S in the presence of nuisance parameters.

Despite the mathematical similarity between the use of nuisance parameters and the pre-whitening operator to improve the modelling of the observed data, one should realize that both approaches are based on fundamentally different assumptions. In case of nuisance parameters the "model of interest" is disturbed by deterministic effects of which the pattern can be measured

precisely (columns of S) and pre-whitening is to be applied when the model disturbance has a stochastic nature, of which the covariance pattern is known. In the practice of brain imaging it is not always obvious which approach is best and for instance different methodologies co-exist in the literature describing the removal of heart beat effects from fMRI-signals.

3.2 Multiple parameters of interest

When there are both multiple parameters of interest and multiple nuisance parameters, the theory of the previous section must be slightly generalized. The case of multiple parameters of interest appears in fMRI paradigms where multiple stimulus conditions are compared, or when a non-canonical hemodynamic response function is estimated from the data. Instead of equation (3.12), one has

$$r_n = \sum_{m=0}^{M-1} B_{nm}\theta_m + \sum_{l=0}^{L-1} S_{nl}\varphi_l + \eta_n$$

or

$$\mathbf{r} = B\boldsymbol{\theta} + S\boldsymbol{\varphi} + \boldsymbol{\eta} \quad , n=0,\ldots,N\text{-}1. \qquad (3.20)$$

As OLS-estimators, one finds

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{OLS} &= \left(B^T P_S B\right)^{inv}\left(B^T P_S \mathbf{r}\right) \\ \hat{\boldsymbol{\varphi}}_{OLS} &= \left(S^T P_B S\right)^{inv}\left(S^T P_B \mathbf{r}\right) \end{aligned} \quad , \qquad (3.21)$$

where the projector P_B is used:

$$P_B \equiv I - B\left(B^T B\right)^{inv} B^T \quad . \qquad (3.22)$$

The expression for the projection matrix P_S is the same as equation (3.15). Analogous to the one parameter case, $\hat{\boldsymbol{\theta}}_{OLS}$ can be expressed as

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{OLS} &= \left(B'^T B'\right)^{inv} B'^T \mathbf{r}' \\ B' &= P_S B' \quad \text{and} \quad \mathbf{r}' = P_S \mathbf{r} \end{aligned} \quad . \qquad (3.23)$$

In other words, if nuisance effects are removed from both data and model, the computation of $\hat{\boldsymbol{\theta}}_{OLS}$ proceeds similar to the case that no nuisance effects are present (cf 2.14).

Exercise 3.6 Show that the first of both equations in (3.21) is a generalization of (3.14).
Exercise 3.7 Suppose one is analyzing fMRI data using the general linear model (3.20) with no nuisance parameters. After visual inspection of the data it appears that the data is rather noisy. One decides to remove this noise by applying a linear filter on the data. The effect of the filtering procedure can be described as applying a projector P_S. The filtered data is passed to the "fMRI toolbox", in order to obtain parameter estimates at each fMRI voxel, just as if data were unfiltered. Is this a correct way to obtain OLS an estimate of the parameter of interest? If not, what is wrong?
Exercise 3.8 Suppose one is analyzing fMRI data using the general linear model (3.18). The noise is correlated with a covariance matrix σC, where C is known. There are L nuisance parameters. One can find an expression for the ML-estimator of $\boldsymbol{\theta}$ using a pre-whitening of the data based on W (with $WW^T=C^{inv}$) and a trend-removal based on P_S. Which one should be applied first W or P_S? Why?

3.2.1 Nuisance effects and pre-processing filters

It is common practice when analyzing raw data to "clean" the raw data in order to remove artefacts that would have an unreasonable effect on the quality of the estimated parameters. For instance, EEG or MEG may contain eye blinks or fifty Hz. oscillations due to power line interference ("hum"). Another example is that one may be interested in the behaviour of EEG exclusively in a certain frequency band, e.g. the oscillations in the 8 to 10 Hz alpha band. A third example is that fMRI data may contain low and high frequency oscillations that are not related to the stimulus paradigm. In such cases it is a well accepted procedure to pre-process the raw data, using a dedicated filter before applying the model. However, from a theoretical point of view it is worthwhile to integrate these pre-processing steps with the application of the data model as much as possible. When the pre-processing filter can mathematically be described as the application of an orthogonal projection operator, these advantages can be nicely illustrated.
For instance, an EEG channel is recorded at 200 Hz during 10 s (so that this channel contains N=2000 data points) and one wishes to remove all frequency components from 49.5 Hz to 50.5 Hz. Each removed frequency component f consists of a cosine part, proportional to $\cos(2\pi f\,\tau\, n)$ and a sine component, proportional to $\sin(2\pi f\,\tau\, n)$, where t is the sampling time and n is the sample number. When all $2F$ frequency components to be removed are assembled in an Nx$2F$ matrix S

$$S = \begin{pmatrix} \sin(\pi\frac{f_0 0}{100}) & \cos(\pi\frac{f_0 0}{100}) & \cdots & \cdots & \cos(\pi\frac{f_F 0}{100}) \\ \sin(\pi\frac{f_0 1}{100}) & \cos(\pi\frac{f_0 1}{100}) & & & \cos(\pi\frac{f_F 1}{100}) \\ \vdots & & & & \vdots \\ \sin(\pi\frac{f_0 (N-1)}{100}) & \cos(\pi\frac{f_0 (N-1)}{100}) & \cdots & \cdots & \cos(\pi\frac{f_F (N-1)}{100}) \end{pmatrix}, \qquad (3.24)$$

where f_0=49.5 Hz, f_1=(49.5+0.1) Hz, f_2=(49.5+0.2) Hz, and the maximum removed frequency is 50.5 Hz. The reason why these frequencies show jumps in 0.1 Hz is that the frequency resolution of a 10 s time window is 1/10 = 0.1 Hz. To remove the 50 Hz power line noise one computes the projection operator $P_S \equiv I - S\left(S^T S\right)^{inv} S^T$ and applies P_S to the EEG data vector **r** to obtain "clean" EEG **r'**. If the clean EEG is used to estimate some model parameters $\hat{\boldsymbol{\theta}}_{OLS}$, one should realize that the resulting parameters are (slightly) different from the case that the fifty Hz components (columns of S) would be added to the model as nuisance effects. The difference is that the pre-processing case, $\hat{\boldsymbol{\theta}}_{OLS}$ would be computed as $\hat{\boldsymbol{\theta}}_{OLS} = \left(B^T B\right)^{inv} B^T \mathbf{r}' = \left(B^T B\right)^{inv} B'^T \mathbf{r}'$, whereas in the nuisance parameter case, one would compute $\hat{\boldsymbol{\theta}}_{OLS} = \left(B'^T B'\right)^{inv} B'^T \mathbf{r}'$, i.e. both data and model are "cleaned".
Note that the example given above, where a projection operator is derived for band-stop filtering, is very artificial and at a conceptual level because many technical aspects have been omitted. For instance, all columns in (3.24) are mutually orthogonal, implying that $S^T S = I$ and that the computation of P_S can be highly simplified. Application of P_S is equivalent to Fourier transforming the data vector, set to zero all frequencies between 49.5 and 50.5 Hz, and transforming the result back into the time domain. This is certainly a way to remove unwanted frequency components, but it should also be kept in mind that there are better ways to do it (better in the sense that they better mimic the behaviour of "hardware" filters, applied before sampling the data). However, these aspects are beyond the scope of this course and the main goal here is to lay a formal connection to between band filtering and nuisance effects.

Exercise 3.9 Compare the difference in computing a single linear parameter of interest by either pre-processing the data with P_S or adding nuisance effects to model, consisting of columns of *S*. What can you say about the absolute value of the estimated parameter in both situations?

Exercise 3.9A Describe how the effect of the reference electrode in the dipole fitting problem can be described in terms of a nuisance parameter. How does the ML estimation problem look like, when likelihood is optimized w.r.t. this nuisance parameter? Hint: the need for a reference electrode implies that the potential at electrode *n* is determined up to an unknown constant.

3.2.2 Unbiased noise estimation in case of nuisance parameters

In the above sections a very simple receipt was presented to compute OLS estimates of parameters of interest when also nuisance effects are present in the data. One simply cleans the data and the model by applying P_S and proceeds as if no nuisance effects were present. But what about the estimated noise level, can one apply a similar receipt? One would perhaps think that one could just use (3.7) and replace all data and model vectors by their "cleaned" ones. This idea is almost correct, except that one has to account for an extra bias in $\hat{\sigma}^2$, caused by the estimation of the *L* nuisance parameters. Therefore, the correct expression appears to be given by

$$\hat{\sigma}^2_{unbiased} = \frac{\mathbf{r'}^T P_{B'} \mathbf{r'}}{N - M - L} \quad , \qquad (3.25)$$

where $P_{B'} \equiv I - B'\left(B'^T B'\right)^{inv} B'^T$ is computed from the "cleaned" model.

3.2.3 Geometrical interpretation

The expression for $\hat{\sigma}_{OLS}$ derived in equation (3.4) is equivalent to

$$\left|P_B \mathbf{r}\right| = \sqrt{N}\hat{\sigma}_{OLS} \quad . \qquad (3.25A)$$

Here the vector $P_B\mathbf{r}$ consists of those parts of r that are orthogonal to the linear space spanned by *B*. The length of this vector determines how close a linear combination of the columns of *B* can get to the data vector **r**. So the length of $P_B\mathbf{r}$ is a measure of the noise level and the formal relationship between these quantities is given by (3.25A)
The effect of nuisance parameters and the role of the projection operator P_S can also be understood when a geometrical interpretation is made of the data and model vectors **r** and **b**. When no nuisance effects would be present, $\hat{\theta}_{OLS}$ would be computed as the factor by which **b** has to be extended to obtain the orthogonal projection of **r** on **b** (figure 2.4B and 3.2). In the case of nuisance effects, one first projects out the columns of S from both the data and the model, obtaining **r'=** P_S **r** and **b'=** P_S **b** and $\hat{\theta}_{OLS}$ is computed using the orthogonal projector (I-$P_{\mathbf{b}'}$). If by mistake only the data would be cleaned and not the model, one would obtain the projection of **r'** onto **b** as projection (red vector in figure 3.2).

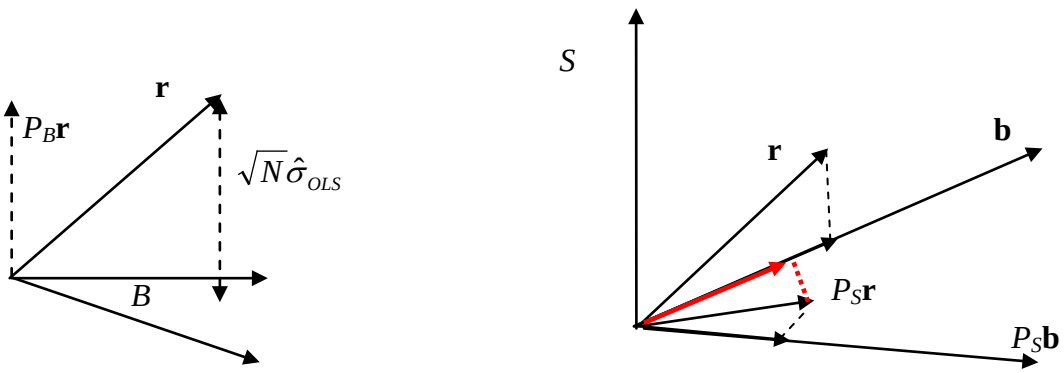

Figure 3.2. *On the left a geometrical interpretation of the OLS estimated noise level is given in case of a linear mode. On the right the linear space of nuisance effects of represented as a single vector pointing in vertical direction. When the associated projector P_S would only be applied upon the data and not on the model, the estimator would be equal to the ratio of* **b** *and the vector plotted in red.*

3.3 Testing the quality of fit

In practical situations, one is not only interested in the values of the estimated parameters themselves, but also in the quality of the parameter fit. How reliable are the parameters, what are the chances that the true parameters are close to the estimated ones? These questions can be addressed by considering also the confidence intervals, as discussed before. However, what is still lacking is the proper statistical interpretation of these intervals. Another important and related question is how well the model fits that data. This question can be quantified in different ways. For instance, one could compute the cost function obtained with the ML-estimated parameters, and relate this to "the amount of signal in the data", $\mathbf{r}^T\mathbf{r}$. An alternative would be to compute the correlation co-efficient between the data **r** and its reconstruction $\tilde{\mathbf{r}} = B\boldsymbol{\theta} + S\boldsymbol{\varphi}$. One could also compute this correlation and first remove the "unwanted effects", i.e. one could compute the correlation between $P_S\mathbf{r}$ and $P_S\tilde{\mathbf{r}} = P_S B\boldsymbol{\theta}$. It appears that all these questions and approaches are mutually related, at least within the framework of linear models and Gaussian noise. In the next sections these relationships will be elucidated, first for the one parameter case and then for the general linear model. Furthermore, we will first assume that there are no nuisance parameters.

3.3.1 Goodness of fit, one parameter

When there are no nuisance parameters, the data model simplifies to

$$\mathbf{r} = \mathbf{b}\theta + \boldsymbol{\eta} \quad . \tag{3.26}$$

Here **r** is the data vector, **b** is a known model vector, θ the unknown parameter and $\boldsymbol{\eta}$ is uncorrelated Gaussian noise. A maximum likelihood estimator of θ is obtained by minimizing the sum of squared differences between data **r** and model **b** θ. To express the quality of fit, one might compute how low the OLS cost function becomes when θ adopts its optimal value $\hat{\theta}_{OLS}$. In other words, we might compute the sum of squared differences between data **r** and the model prediction $\hat{\mathbf{r}}_{\mathrm{OLS}}$, with

$$\begin{aligned}\hat{\mathbf{r}}_{\text{OLS}} &\equiv \mathbf{b}\hat{\theta}_{\text{OLS}} \\ &= \mathbf{b}\frac{\mathbf{b}^T\mathbf{r}}{\mathbf{b}^T\mathbf{b}} \\ &= \frac{\mathbf{b}\mathbf{b}^T}{\mathbf{b}^T\mathbf{b}}\mathbf{r} \\ &= (I - P_{\mathbf{b}})\mathbf{r}\end{aligned} \quad . \qquad (3.27)$$

In this derivation, the estimate for $\hat{\theta}_{OLS}$ has been substituted (equation (3.13)) and the result has been expressed in terms of the projection matrix $P_{\mathbf{b}}$, defined in (3.15). One may explain equation (3.27) as follows. The effect of applying $P_{\mathbf{b}}$ to a vector **r** is to remove that part of **r** that is proportional (linear dependent of) to **b**. Similarly, the effect of (I-P_{b}) on **r** is the reverse: to keep the part of **r** proportional (linear dependent of) to **b**, and to remove all other components. In the context of the model assumption (3.26), this means that the model prediction consists of that part of the data vector **r**, that is proportional to **b** (see fig 5C).
With this expression for the model prediction, one finds for the minimum cost

$$\begin{aligned}\text{Cost}_{\text{Min}} &= (\mathbf{r} - \hat{\mathbf{r}}_{\text{OLS}})^T(\mathbf{r} - \hat{\mathbf{r}}_{\text{OLS}}) \\ &= (\mathbf{r} - \mathbf{r} + P_{\mathbf{b}}\mathbf{r})^T(\mathbf{r} - \mathbf{r} + P_{\mathbf{b}}\mathbf{r}) \\ &= (P_{\mathbf{b}}\mathbf{r})^T(P_{\mathbf{b}}\mathbf{r}) \\ &= \mathbf{r}^T P_{\mathbf{b}}{}^T P_{\mathbf{b}}\mathbf{r} \\ &= \mathbf{r}^T P_{\mathbf{b}}\mathbf{r}\end{aligned} \quad . \qquad (3.28)$$

Instead of the minimum cost, one usually considers the so-called *goodness of fit* (gof) statistic. The gof is a relative measure, which expresses how well the model fits the data. The relativity is brought in by dividing the minimum cost by the *data power* $\mathbf{r}^{\text{T}}\mathbf{r}$, which is defined as the sum of squared observations r_{n}. To convert "cost" in "goodness", the relative cost is subtracted from 1 (or from 100%):

$$\begin{aligned}\text{gof} &\equiv 1 - \frac{\text{Cost}_{\text{Min}}}{\mathbf{r}^T\mathbf{r}} \\ &= 1 - \frac{\mathbf{r}^T P_{\mathbf{b}}\mathbf{r}}{\mathbf{r}^T\mathbf{r}} \\ &= \frac{\mathbf{r}^T\mathbf{r} - \mathbf{r}^T P_{\mathbf{b}}\mathbf{r}}{\mathbf{r}^T\mathbf{r}} \\ &= \frac{\mathbf{r}^T(I - P_{\mathbf{b}})\mathbf{r}}{\mathbf{r}^T\mathbf{r}} \times 100\% \\ &= \frac{\hat{\mathbf{r}}^T\hat{\mathbf{r}}}{\mathbf{r}^T\mathbf{r}} \times 100\%\end{aligned} \quad . \qquad (3.29)$$

The last form of this expression shows that gof can be interpreted as the percentage of the data power that is explained by the data power of the model prediction.

3.3.2 Correlation between data and model prediction, one parameter

An alternative quantity to express the explanatory power of the model is to compute the correlation co-efficient ρ between the data **r** and the model prediction $\hat{\mathbf{r}}$.

$$\rho \equiv \frac{\hat{\mathbf{r}}^T \mathbf{r}}{\sqrt{(\mathbf{r}^T \mathbf{r})(\hat{\mathbf{r}}^T \hat{\mathbf{r}})}} \quad . \tag{3.30}$$

This equation can be considered as the definition of the correlation co-efficient of two N-dimensional vectors $\mathbf{r}$ and $\hat{\mathbf{r}}$. The rational of this measure is that ρ^2 varies between 0 and 1, and that the better the shape of $\hat{\mathbf{r}}$ resembles the shape of $\mathbf{r}$, the higher ρ^2.

Exercise 3.10 Use definition (3.27) to show that (3.30) is equal to

$$\rho = \frac{\mathbf{b}^T \mathbf{r}}{\sqrt{(\mathbf{r}^T \mathbf{r})(\mathbf{b}^T \mathbf{b})}} \quad . \tag{3.31}$$

By substitution of the expression for the model prediction, the correlation co-efficient squared in the one parameter model can be expressed in terms of $P_{\mathbf{b}}$

$$\begin{aligned} \rho^2 &= \frac{(\mathbf{r}^T (I - P_{\mathbf{b}})^T \mathbf{r})^2}{(\mathbf{r}^T \mathbf{r})(\mathbf{r}^T (I - P_{\mathbf{b}})^T (I - P_{\mathbf{b}})\mathbf{r})} \\ &= \frac{(\mathbf{r}^T (I - P_{\mathbf{b}})^T \mathbf{r})^2}{(\mathbf{r}^T \mathbf{r})(\mathbf{r}^T (I - P_{\mathbf{b}})\mathbf{r})} \\ &= \frac{\mathbf{r}^T (I - P_{\mathbf{b}})\mathbf{r}}{(\mathbf{r}^T \mathbf{r})} \\ &= \text{gof} \end{aligned} \quad . \tag{3.32}$$

In the last step, the one but last expression of (3.29) was used. The conclusion from this derivation is that, at least in the one parameter case, the squared correlation co-efficient and the gof are identical.

3.3.3 Student-*t*, one parameter

We now discussed two fit measures, the goodness of fit and the squared correlation coefficient. We will add a third one, which is defined in terms of the estimated parameter $\hat{\theta}$ and its estimated standard deviation $\hat{\sigma}_{\hat{\theta}}$. For the one parameter case, this standard deviation can be obtained from (3.8), by assuming a single parameter. Then the matrix B consists of a single column $\mathbf{b}$ and ($B^{\mathrm{T}}B$) becomes a scalar $\mathbf{b}^{\mathrm{T}}\mathbf{b}$. Furthermore, in the one parameter case the covariance matrix of the estimated parameter is a single number, indicating the variance of the estimated parameter, i.e. the square of $\hat{\sigma}_{\hat{\theta}}$. One finds

$$\mathrm{EstCov}(\hat{\boldsymbol{\theta}}_{OLS}) \equiv \hat{\sigma}^2_{unbiased} (B^T B)^{inv} = \frac{\mathbf{r}^T P_B \mathbf{r}}{N - M} (B^T B)^{inv}$$

$$\left(\hat{\sigma}_{\hat{\theta}}\right)^2 = \frac{\hat{\sigma}^2_{unbiased}}{\mathbf{b}^{\mathrm{T}}\mathbf{b}} = \frac{1}{N-1} \frac{\mathbf{r}^T P_b \mathbf{r}}{\mathbf{b}^{\mathrm{T}}\mathbf{b}} \quad . \tag{3.33}$$

The statistic that is here considered is proportional to the estimated parameter divided by its standard deviation. In the absence of nuisance parameters, it is defined as

$$t_{N-1} \equiv \frac{\hat{\theta}}{\hat{\sigma}_{\hat{\theta}}} \quad . \tag{3.34}$$

Exercise 3.11 Explain why expression (3.34) yields a dimensionless quantity.

The subscript N-1 in this definition indicates that, since we have used (3.7) with M=1 parameter, t_{N-1} is distributed as a Student-t distribution with N-1 degrees of freedom. This will be explained in a following section. Here we note that t_{N-1} measures the quality of fit of parameter θ. The smaller the standard deviation with respect to the parameter itself, the higher the quality of fit, and therefore, this quality is measured by the size of the computed t-value.
It appears that also the t-value has a relation with gof, despite the fact that gof is a measure of model prediction, and t-value is a measure of parameter quality. To find the relationship, t_{N-1} is squared and both $\hat{\theta}$ and $\hat{\sigma}_{\hat{\theta}}$ are expressed in terms of data and projection matrix P_b:

$$\begin{aligned}
\left(t_{N-1}\right)^2 &= \frac{\hat{\theta}^2}{\hat{\sigma}_{\hat{\theta}}^2} \\
&= \frac{\left(\dfrac{\mathbf{b}^T\mathbf{r}}{\mathbf{b}^T\mathbf{b}}\right)^2}{\dfrac{1}{N-1}\dfrac{\mathbf{r}^T P_{\mathbf{b}}\mathbf{r}}{\mathbf{b}^T\mathbf{b}}} \\
&= (N-1)\frac{\dfrac{\mathbf{r}^T\mathbf{b}\mathbf{b}^T\mathbf{r}}{\mathbf{b}^T\mathbf{b}}}{\mathbf{r}^T P_{\mathbf{b}}\mathbf{r}} \\
&= (N-1)\frac{\mathbf{r}^T(I-P_{\mathbf{b}})\mathbf{r}}{\mathbf{r}^T P_{\mathbf{b}}\mathbf{r}} \\
&= (N-1)\frac{\mathbf{r}^T\mathbf{r}-\mathbf{r}^T P_{\mathbf{b}}\mathbf{r}}{\mathbf{r}^T P_{\mathbf{b}}\mathbf{r}} \\
&= (N-1)\frac{\text{gof}}{1-\text{gof}} \\
&= (N-1)\frac{\rho^2}{1-\rho^2}
\end{aligned} \quad . \qquad (3.35)$$

This equation shows that there is a one to one relationship between gof and t_{N-1}. Moreover, this relationship has the form of an increasing function, reflecting the fact that both quantities measure the quality of fit.

Exercise 3.12 From one of the intermediate steps, it follows that t_{N-1} squared can be expressed as

$$\begin{aligned}
\left(t_{N-1}\right)^2 &= (N-1)\frac{\mathbf{r}^T(I-P_{\mathbf{b}})\mathbf{r}}{\mathbf{r}^T P_{\mathbf{b}}\mathbf{r}} \\
&= (N-1)\frac{\mathbf{r}^T(I-P_{\mathbf{b}})^2\mathbf{r}}{\mathbf{r}^T P_{\mathbf{b}}\mathbf{r}} \\
&= (N-1)\frac{\hat{\mathbf{r}}^T\hat{\mathbf{r}}}{\mathbf{r}^T P_{\mathbf{b}}\mathbf{r}}
\end{aligned} \quad . \qquad (3.36)$$

Give an interpretation of the numerator and the denominator of this form. Hint: see (3.27).

3.3.4 Geometrical interpretation of $\hat{\theta}$, gof, ρ and *t*

A geometrical interpretation of the single estimated parameter $\hat{\theta}$ can given as the factor by which the vector **b** has to be extended in order to get as close to **r** as possible. When this analogue is followed in more detail, one can also obtain a more intuitive interpretation of the meaning of the *t*-statistic and its relation to gof and the correlation coefficient ρ. For simplicity the effect of nuisance parameters is omitted here.
When the angle between **r** and **b** is ω (see figure 3.3), one may compute the length of $(I\text{-}P_{\mathbf{b}})\mathbf{r}$ as $|\mathbf{r}|\cos\omega$ (see figure 3.3). Therefore, assuming $\omega<\frac{1}{2}\pi$ and using that $\cos\omega=\mathbf{r}^{T}\mathbf{b}/(|\mathbf{r}||\mathbf{b}|)$, the ratio between $|\mathbf{b}|$ and $|(I\text{-}P_{\mathbf{b}})\mathbf{r}|$ can be computed as

$$\begin{aligned}\hat{\theta} &= \frac{|\,(I-P_{\mathbf{b}})\mathbf{r}\,|}{|\,\mathbf{b}\,|} \\ &= \frac{|\,\mathbf{r}\,|\cos\omega}{|\,\mathbf{b}\,|} \\ &= \frac{\mathbf{r}^{T}\mathbf{b}}{\mathbf{b}^{T}\mathbf{b}}\end{aligned} \quad , \qquad (3.36A)$$

which is identical to (3.16).
The goodness of fit (gof) equals the ratio of the squared lengths of the closed point $(I\text{-}P_{\mathbf{b}})\mathbf{r}$ and the data vector **r**. Considering figure 3.2, one observes that this ratio also equals the cosine of ω, which is the correlation co-efficient. Therefore,

$$\begin{aligned}\text{gof} &= \frac{|\,(I-P_{\mathbf{b}})\mathbf{r}\,|^{2}}{|\,\mathbf{r}\,|^{2}} \\ &= \cos^{2}\omega \\ &= \rho^{2}\end{aligned} \quad . \qquad (3.36B)$$

The estimated standard deviation in $\hat{\theta}$ equals $\hat{\sigma}_{\hat{\theta}} = \frac{1}{|\,\mathbf{b}\,|\sqrt{N-1}} \times \left|P_{\mathbf{b}}\mathbf{r}\right|$, see (3.33). Note that this implies that $\hat{\sigma}_{\hat{\theta}}$ is proportional to the line segment opposing ω, whereas $\hat{\theta}$ is proportional to the line segment adjacent to ω, see figure 3.3. Therefore,

$$
\begin{aligned}
t &= \frac{\hat{\theta}}{\hat{\sigma}_{\hat{\theta}}} \\
&= \frac{|(I-P_{\mathbf{b}})\mathbf{r}|}{|\mathbf{b}|} \cdot \frac{|\mathbf{b}|\sqrt{N-1}}{|P_{\mathbf{b}}\mathbf{r}|} \\
&= \sqrt{N-1}\frac{|(I-P_{\mathbf{b}})\mathbf{r}|}{|P_{\mathbf{b}}\mathbf{r}|} \\
&= \sqrt{N-1}\ \frac{\cos(\omega)}{\sin(\omega)} \\
&= \sqrt{N-1}\ \cot an(\omega) \\
&= \sqrt{N-1}\ \frac{\rho}{\sqrt{1-\rho^2}}
\end{aligned}
\qquad . \qquad (3.36C)
$$

Here the geometrical interpretation of figure 3.3 has been applied. The cotan() of ω equals the ratio of the two orthogonal projections of **r** onto **b** and its complement. The last line gives a direct relation between the correlation coefficient and the *t*-statistic, equivalent to (3.35).

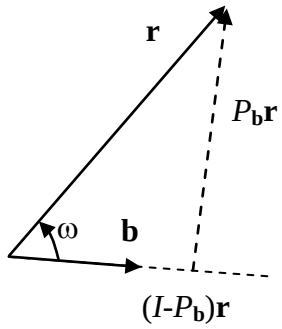

Figure 3.3. *The data vector* ***r*** *is projected onto be and to the space perpendicular to* ***b****. The ratio of the lengths of these two projections is proportional the t-statistic.*

3.3.5 One parameter model, in presence of nuisance parameters

All the expressions derived in the previous sections have a limited value, because they are only derived for the very simplified case of a single parameter of interest, and no nuisance parameters. However, in section 3.1 the following observation was made when comparing the estimator of θ in the presence and absence of nuisance parameters. It appeared that by introducing new symbols **r'** and **b'**, obtained from the original ones by removing the nuisance effects (i.e. $\mathbf{r'}=P_S\mathbf{r}$ and $\mathbf{b'}=P_S\mathbf{b}$), and replacing the unprimed symbols by the primed ones, the ML-estimator for θ in presence of nuisance parameters is obtained from the ML-estimator for θ in absence of these parameters. In other words, it appeared that a general case nuisance parameters can be derived from the special case of no such parameters.

When there are L nuisance parameters, collected in an L-dimensional vector $\boldsymbol{\varphi}$, the data model reads

$$
\mathbf{r} = \mathbf{b}\theta + S\boldsymbol{\varphi} + \boldsymbol{\eta} \qquad . \qquad (3.37)
$$

When the parameter of interest and the nuisance parameters would be collected in a single parameter vector $\boldsymbol{\theta}$, one would obtain an ML estimator $\hat{\boldsymbol{\theta}}$ having a Gaussian distribution centred

at $\boldsymbol{\theta}$ and with a covariance matrix equal to $\sigma^2\left((\mathbf{b},S)^{\mathrm{T}}(\mathbf{b},S)\right)^{inv}$. The recipe to obtain from here the statistical distribution of the parameter of interest would be to integrate the distribution of $\hat{\boldsymbol{\theta}}$ over all nuisance components, from –∞to +∞, as explained in figure 3.1. The result of this integration implies that the variance of $\hat{\theta}$ equals $\frac{\sigma^2}{\mathbf{b}^T P_S \mathbf{b}}$, the upper left element of the matrix $\sigma^2\left((\mathbf{b},S)^{\mathrm{T}}(\mathbf{b},S)\right)^{inv}$. Therefore, in the presence of nuisance parameters, the standard deviation of $\hat{\theta}$ can consistently be computed using $\mathbf{b}'=P_S\mathbf{b}$ instead of $\mathbf{b}$. However, when the *estimated* standard deviation $\hat{\sigma}_{\hat{\theta}}$ is computed using the primed symbols, an additional bias correction has to be applied, because of the increased number of estimated parameters (L+1 instead of 1). The precise mathematical reason for this adaptation is not discussed in this course. As a consequence, the t-statistic becomes

$$\begin{aligned}\left(t_{N-L-1}\right)^2 &\equiv \frac{\hat{\theta}^2}{\hat{\sigma}^2_{\hat{\theta}}} \\ &= (N-L-1)\frac{\mathbf{r}'^T (I-P_{b'})\mathbf{r}'}{\mathbf{r}'^T P_{b'}\mathbf{r}'}\end{aligned} \quad , \qquad (3.38)$$

where

$$\begin{cases}\mathbf{b}' \equiv P_S\mathbf{b} \\ \mathbf{r}' \equiv P_S\mathbf{r}\end{cases} \quad , \qquad (3.39)$$

Note that for L=0 (no nuisance parameters), equation (3.38) becomes (3.36). These two equations give a receipt to compute the quality of the fitted parameter. First remove the signals of no interest from data and model, compute $P_{\mathbf{b}'}$ and substitute these formulas in equation (3.37). Therefore, the addition of nuisance parameters does not make matters more complicated. Also, it appears that the relationship between gof, ρ^2 and t-statistic remain valid (apart from the adapted proportionality factor in the modified t).

Exercise 3.12A Does the partial correlation always decrease when nuisance effects are included in the model?

Exercise 3.13 From equation (3.38) one might argue that the inclusion of more nuisance parameters leads to a larger L and hence to a smaller multiplication factor and therefore to a smaller t-statistic. Is this a valid argument? Why or why not?

3.3.6 Multiple parameters

When there are multiple parameters of interest, the analysis of fit measures goes very similar to the single parameter case. First, it is assumed that there are no nuisance parameters. Then the data model reads as

$$\mathbf{r} = B\boldsymbol{\theta} + \boldsymbol{\eta} \quad , \qquad (3.40)$$

Where B is a N by M model matrix and $\boldsymbol{\theta}$ is an M-vector with the unknown parameters of interest. The OLS estimator of $\boldsymbol{\theta}$ is given by equation (2.14). The model prediction $\hat{\mathbf{r}}_{\mathrm{OLS}}$ can in this case be expressed as

$$\begin{aligned}\hat{\mathbf{r}}_{\mathrm{OLS}} &\equiv B\hat{\boldsymbol{\theta}}_{\mathrm{OLS}} \\ &= B(B^T B)^{inv} B^T \mathbf{r} \\ &= (I - P_B)\mathbf{r}\end{aligned} \quad . \qquad (3.41)$$

We see that by using the concept of a projection matrix, the multiple parameter case is similar to the single parameter case, the only difference being the way the projection matrix is computed. Therefore, with P_{B} instead of $P_{\mathbf{b}}$, we also find similar expressions for the minimum possible cost value and the goodness of fit:

$$\begin{aligned}\mathrm{Cost}_{\mathrm{Min}} &= (\mathbf{r} - \hat{\mathbf{r}}_{\mathrm{OLS}})^T (\mathbf{r} - \hat{\mathbf{r}}_{\mathrm{OLS}}) \\ &= \mathbf{r}^T P_B \mathbf{r}\end{aligned} \quad , \qquad (3.42)$$

and

$$\mathrm{gof} = \frac{\mathbf{r}^T (I - P_B)\mathbf{r}}{\mathbf{r}^T \mathbf{r}} \times 100\% \quad . \qquad (3.43)$$

Exercise 3.14 What is the relation between the ρ^2 and gof in case of multiple parameters?

What is different in the multiple parameter case is that we cannot generalize the definition of the *t*-statistic, when it is defined as the ratio of an estimated parameter and its estimated standard deviation. Instead, we generalize *t* by starting from an alternative interpretation, based on t^2, presented in equation (3.36). This generalization is indicated with the symbol *F*:

$$\begin{aligned}F &\equiv \frac{N-M}{M} \frac{\hat{\mathbf{r}}^T \hat{\mathbf{r}}}{\mathrm{Cost}_{\mathrm{min}}} \\ &= \frac{N-M}{M} \frac{\mathbf{r}^T (I - P_B)\mathbf{r}}{\mathbf{r}^T P_B \mathbf{r}}\end{aligned} \quad . \qquad (3.44)$$

This fit measure can be interpreted as the ratio of the power of the reconstructed signal $\hat{\mathbf{r}}$ and the resulting cost function. Since the lower the cost, the higher the quality of fit, it is clear that the cost must appear in the denominator, if *F* would express the quality of fit. The reason of having $\hat{\mathbf{r}}^T \hat{\mathbf{r}}$ in the numerator is not that obvious in this stage. We could only say that since $\hat{\mathbf{r}}^T \hat{\mathbf{r}} = \hat{\boldsymbol{\theta}}^T B^T B \hat{\boldsymbol{\theta}}$, the numerator is an indirect measure of the amplitude of the estimated parameter $\hat{\boldsymbol{\theta}}$. So the larger the amplitude of $\hat{\boldsymbol{\theta}}$ w.r.t. the cost function, the larger *F*. The relevance of the precise mathematical form of (3.38) appears when the statistical distribution of *F* is studied.

Exercise 3.15 Show that in the multiple parameter case, the relationship between *F* and gof is

$$F = \frac{N-M}{M} \frac{\mathrm{gof}}{1 - \mathrm{gof}} \quad . \qquad (3.45)$$

Exercise 3.16 Suppose *M*=1. What is the relationship between *t* and *F*?

When there are both multiple parameters of interest (M>1) and more than one nuisance parameter (L>0) these results can be generalized by the same primed/non-primed trick as used before. In that case the model reads as

$$\mathbf{r} = B\boldsymbol{\theta} + S\boldsymbol{\varphi} + \boldsymbol{\eta} \quad . \tag{3.46}$$

With

$$\begin{cases} B' \equiv P_S B \\ \mathbf{r}' \equiv P_S \mathbf{r} \end{cases} , \tag{3.47}$$

the F-statistic can be generalized to

$$F = \frac{N-L-M}{M} \frac{\mathbf{r}'^T (I - P_{B'})\mathbf{r}'}{\mathbf{r}'^T P_{B'}\mathbf{r}'} \quad . \tag{3.48}$$

Note that also the multiplication factor has been adapted.

3.3.7 Geometrical interpretation of *F*

The meaning of the F statistic can be better understood when a geometrical interpretation is given. In figure 3.4 the data vector **r'** and model space B', spanned by the column vectors (**b**$_0$**'**, **b**$_1$**'**,…) are depicted. The primed symbols indicate that both data and model have been "cleaned" using a projection operator P_S. The data vector is projected onto the model space in order to obtain the closest point to **r**' that is still a linear combination of (**b**$_0$**'**, **b**$_1$**'**,…). That point is indicated with
(I-$P_{B'}$)**r**', and, viewed as a vector, it makes an angle ω with respect to the data vector. The residual vector, i.e. the difference between the data and the closest point is $P_{B'}$**r**'. The F-statistic is proportional to the ratio of the squared lengths of the closest point vector and the residual vector. From figure 3.4 one observes that this ratio equals the squared cotangent of ω. Furthermore, since the correlation coefficient ρ equals the cosine of ω, we also have the following relationship,

$$\begin{aligned} F &= \frac{N-L-M}{M} \frac{\mathbf{r}'^T (I - P_{B'})\mathbf{r}'}{\mathbf{r}'^T P_{B'}\mathbf{r}'} \\ &= \frac{N-L-M}{M} \cot an^2(\omega) \\ &= \frac{N-L-M}{M} \frac{\rho^2}{1-\rho^2} \end{aligned} \quad , \tag{3.48A}$$

where

$$\rho = \frac{\mathbf{r}'^T (I - P_{B'})\mathbf{r}'}{\left\|\mathbf{r}'\right\| I - P_{B'})\mathbf{r}'\right|} \quad . \tag{3.48B}$$

Note that the F-statistic is a generalization of the t-statistic. Whereas Student-t is used to test the statistical significance of a single parameter, F tests the significance of a group of parameters. Furthermore, in case the t-statistic the "horizontal component" (I-$P_{B'}$)**r**' is proportional to the estimated parameter and "vertical component" $P_{B'}$**r**' is proportional to its standard deviation. In case of the F-statistic, such relations are not applicable. Note furthermore, that, since F (and t) are only dependent on the correlation coefficient (not on the data or model vectors separately), these statistics remain the same when the model vectors (**b**$_0$**'**, **b**$_1$**'**,…) are replaces by another set of model vectors that span the same linear space.

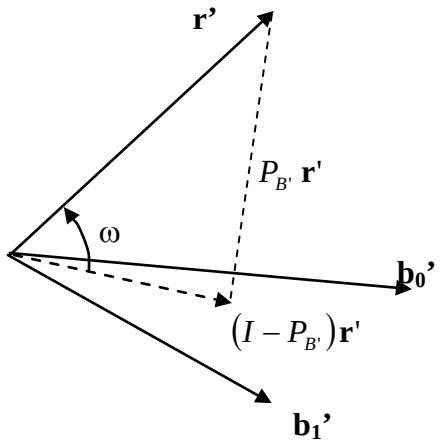

Figure 3.4. *The data vector* ***r'*** *is projected onto the linear space spanned by (****b0', b1',…****). The projected vectors determine the goodness of fit and the F-statistic.*

3.3.8 (*)Statistical distribution of *t* and *F*

In the previous sections, various mutually related quality of fit measures were derived. Although it was argued that these quantities indeed measure the quality of fit, what is missing is the precise statistical interpretation of in particular *t* and *F*. Such interpretation will be derived in the present section and it will be used to test the hypothesis that θ in the one parameter case, or $\boldsymbol{\theta}$ in the multiple parameter case are 0. So, in the multiple parameter case this means that the hypothesis to be tested is that all parameters θ_0, θ_1, θ_2 and θ_{M-1} are all 0. If this hypothesis is true, then it is expected that the fit measure will be poor: a low gof, a low ρ^2 a low *t* or a low *F*. However, because of the noise on the measurements, there will always be a chance that the observed fit measure will coincidentally not be so low. For that reason, we study the distribution of *t* and *F* in particular (the others being equivalent) under the assumption that $\boldsymbol{\theta}$=0.
When $\boldsymbol{\theta}$=0, the data vector equals

$$\begin{aligned}\mathbf{r} &= S\boldsymbol{\varphi} + \boldsymbol{\eta} \\ \mathbf{r}' &= P_S\mathbf{r} = P_S\boldsymbol{\eta}\end{aligned} \quad . \qquad (3.49)$$

This is true both in the single and multiple parameter case. To find the distribution of *t*, under the null hypothesis that θ=0, this expression must be substituted into (3.34), with the adaptation for the nuisance parameters

$$
\begin{aligned}
t_{N-L-1} &= \frac{\hat{\theta}}{\hat{\sigma}_{\hat{\theta}}} \\
&= \sqrt{N-L-1}\,\frac{1}{\sqrt{\mathbf{b}'^T\,\mathbf{b}'}}\,\frac{\mathbf{b}'^T\,\mathbf{r}'}{\sqrt{\mathbf{r}'^T\,P_{b'}\mathbf{r}'}} \\
&= \sqrt{N-L-1}\,\frac{1}{\sqrt{(P_S\mathbf{b})^T\,P_S\mathbf{b}}}\,\frac{\mathbf{b}^T P_S\boldsymbol{\eta}}{\sqrt{(P_{b'}P_S\boldsymbol{\eta})^T\,P_{b'}P_S\boldsymbol{\eta}}} \\
&= \frac{\eta}{\sqrt{\dfrac{(P_{b'}P_S\boldsymbol{\eta})^T\,P_{b'}P_S\boldsymbol{\eta}}{N-L-1}}}
\end{aligned}
\quad , \qquad (3.50)
$$

with

$$
\eta \equiv \frac{\mathbf{b}^T P_S\boldsymbol{\eta}}{\sigma\sqrt{\mathbf{b}^T P_S\mathbf{b}}} \quad . \qquad (3.51)
$$

This expression for t can be considered as the ratio of a standard Gaussian variable η and the square root of an independent chi-squared variable, with *N-L-1* degrees of freedom. If this can be verified, it follows that t has a Student-t distribution with *N-L-1* degrees of freedom. Therefore, the distribution of t is known, meaning that when a certain t-value is observed (computed from the data), the likelihood can be computed that the observed t-value is purely due to chance. However, before this conclusion can be drawn, it must be verified (1) that numerator and denominator are statistically independent, and (2) that the denominator can be expressed as the sum of *N*-*L*-1 squared Gaussian independent variables. These variables are independent because

$$
\left(\mathbf{b}^T P_S\right)\left(P_S P_{b'}\right) = \mathbf{b}^T P_S\left(P_S - \frac{P_S\mathbf{b}\mathbf{b}^T P_S}{\mathbf{b}^T P_S\mathbf{b}}\right) = \mathbf{b}^T P_S - \frac{\mathbf{b}^T P_S\mathbf{b}\mathbf{b}^T P_S}{\mathbf{b}^T P_S\mathbf{b}} = \mathbf{b}^T P_S - \mathbf{b}^T P_S = 0 \quad . \qquad (3.52)
$$

Furthermore, since $P_{b'}P_S$ is a symmetric projection matrix and

$$
\begin{aligned}
\mathrm{Rank}\left(P_{b'}P_S\right) &= \mathrm{Rank}\left(P_S - \frac{P_S\mathbf{b}\mathbf{b}^T P_S}{\mathbf{b}^T P_S\mathbf{b}}\right) \\
&= \mathrm{Trace}(P_S) - \mathrm{Trace}\left(\frac{P_S\mathbf{b}\mathbf{b}^T P_S}{\mathbf{b}^T P_S\mathbf{b}}\right) = N-L-1
\end{aligned}
\quad . \qquad (3.53)
$$

The projection matrix $P_{b'}P_S$ has *L+1* zero eigenvalues and *N-L-1* eigenvalues equal to 1. Hence, it can be decomposed as $P_{b'}P_S = \mathbf{U}\mathbf{U}^T$, where $\mathbf{U}=(\mathbf{u_0}, \mathbf{u_1}, \mathbf{u_2},\ldots, \mathbf{u_{N-L-2}})$ consists of *N-L-1* orthonormal columns. As a consequence, $(P_{b'}P_S\boldsymbol{\eta})^T\,P_{b'}P_S\boldsymbol{\eta}/\sigma^2$ can be expressed as a sum of *N-L-1* independent squared Gaussian variables $\mathbf{u}_k^{\mathrm{T}}\boldsymbol{\eta}$. Therefore, one may conclude that under the NULL hypothesis,
$t_{\mathrm{N\text{-}L\text{-}1}}$ has a Student-t distribution with *N-L-1* degrees of freedom.

Exercise 3.17(*) Show that $P_{b'}P_S$ is a symmetric projection matrix.

Similarly, it can be shown that F has the F-distribution "with M and N-L-M degrees of freedom". For that purpose, we start from the most general expression (3.48) and substitute the data model, under the null-hypothesis that $\boldsymbol{\theta}=0$, i.e. $\mathbf{r'}=P_S\boldsymbol{\eta}$. So it is found that

$$F = \frac{N-L-M}{M}\frac{\boldsymbol{\eta}^T P_S B'(B'^T B')^{inv} B'^T P_S \boldsymbol{\eta}}{\boldsymbol{\eta}^T\left(P_S - P_S B'(B'^T B')^{inv} B'^T P_S\right)\boldsymbol{\eta}}$$
$$= \frac{\frac{\boldsymbol{\eta}^T P_M \boldsymbol{\eta}}{M}}{\frac{\boldsymbol{\eta}^T P_{N-L-M}\boldsymbol{\eta}}{N-L-M}} \quad , \tag{3.54}$$

with

$$\begin{cases} P_M \equiv P_S B'(B'^T B')^{inv} B'^T P_S \\ P_{N-L-M} = P_S - P_M \end{cases} . \tag{3.55}$$

To demonstrate that F has an F-distribution, it must be shown that it is the ratio of two independent chi-squared distributions with M and N-L-M degrees of freedom. That both numerator and denominators are indeed chi-squared distributions with the requested degrees of freedom follows from the same argument as used in the derivation of the distribution of t. That they are independent follows from the observation that $P_M P_{N\text{-}L\text{-}N} = 0$.

Exercise 3.18$^{(*)}$ Show that $P_M P_{N\text{-}L\text{-}M}=0$.

3.3.9 *Student t*-distribution for large *N*

When the student t-distribution is plotted for different values of N, it appears, see figure 3.5, that for N larger than about 20, the distribution is hardly dependent on N. Moreover, for larger and larger N the Student t-distribution looks more and more like the standard normal distribution (Gaussian with mean 0 and standard deviation 1). This fact can be proven with mathematical rigour, but the interpretation is that, in the context for linear models and Gaussian noise, for larger and larger N we obtain better and better estimates of the standard deviation σ. As a result, the distinction between the true standard deviation s and its estimate $\hat{\sigma}$ disappears and the t-statistic $t = \hat{\theta}/\hat{\sigma}_{\hat{\theta}}$ approaches $t \approx \hat{\theta}/\sigma_\theta$. Therefore, in this limit, t will have a Gaussian distribution with unit standard deviation.
A practical consequence of this limiting case is that p-values of t-statistics can be approximately computed as

$$p \approx \frac{1}{\sqrt{2\pi}}\int_t^\infty \mathrm{e}^{-s^2/2}\mathrm{d}s = \mathrm{erf}(t) \quad , \tag{3.55B}$$

where erf(t) is the error function. The integral appearing in erf() cannot be computed analytically, in practice one has to use tables or other approximations.

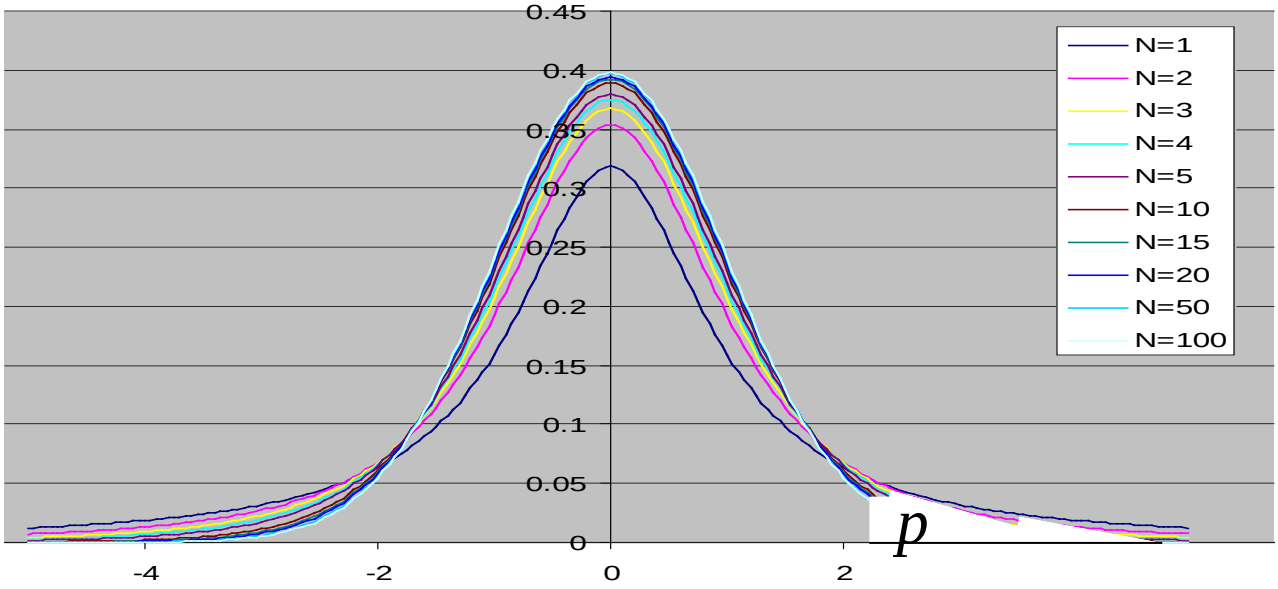

Figure 3.5. *The t-distribution is plotted for several degrees of freedom N. It can be observed that for N larger than about 20, these curves are indistinguishable from each other and are very close to the standard normal distribution.*

3.3.10 Statistical tests with *t* and *F*

As explained in the previous section, the practical implication of the *t* and *F*-statistics is that they can be used for statistical testing of a certain hypothesis. This hypothesis implies that a certain parameter, or a group of parameters is zero. By separating the estimated parameters into two groups: parameters of interest and nuisance parameters, it becomes possible to test a subgroup of parameters. A very common example that illustrates the usefulness of this approach is the case that a number of measurements are done (τ_0,r_0), (τ_1,r_1),…,(τ_{N-1},r_{N-1}), a straight line is fitted through these points and one wonders whether this line passes through the origin, yes or no. After deciding that the largest errors occur in the *r*-direction, one would estimate two parameters: an offset and a slope. The estimation of these parameters is a linear parameter estimation problem, for which the above theory is applicable.

Exercise 3.19 In the example sketched above, which parameter is the parameter of interest, and which one is the nuisance parameter?
Exercise 3.20 When the above example is placed in the framework of equation (3.37), what is **r**, what is **b**, what is θ, what is S, what is $\boldsymbol{\varphi}$ and what is L? Hint: $r_n = 1\times\text{offset} + \tau_n\times\text{slope} + \text{noise}$.
Exercise 3.21 Describe which quantities you would compute, using which equations to test whether the fitted line passes through the origin.
Exercise 3.22 How would you apply the theory if the question were to test whether there is a significant trend in the data points?
Exercise 3.23 The correlation between data and model depends on the inclusion of nuisance parameters. If the nuisance effect is co-linear with the data and/or model, then it is to be expected that the correlation decreases when nuisance effects are projected out. Is the effect of nuisance parameters always in the same direction? In other words, could it happen that correlation increases by including nuisance effects?

It should be noted that there is an important distinction between the *t*-test and the *F*-test. The *t*-test measures how far away a single estimated parameter is away from the origin. The *t*-statistic can be either positive or negative, and therefore the *t*-test can be performed two-sided or single sided. When it is performed two-sided, the null-hypothesis is rejected when *t* is highly positive or highly negative. This means that the estimated parameter will be declared significant, no matter whether it is positive or negative. When it is performed single-sided, one will test for either very large positive values, or very large negative values. The choice depends on the physical/physiological a priori knowledge one has about the parameter.
With an *F*-test, one tests how far off the combined effect of a set of estimated parameters is from zero. More precisely, the amplitude of the model prediction $\hat{\mathbf{r}} = B\hat{\boldsymbol{\theta}}$ is tested, with respect to the cost function. If the null-hypothesis is rejected with an *F*-test, this means that either one of the components of $\hat{\boldsymbol{\theta}}$, or a few of them, or all are significantly different from zero. The test does neither tell us, which of these alternatives is true and nor do we know whether the deviations are positive or negative.

> **Exercise 3.24** When an *F*-test is applied on a single parameter, is it equivalent to a single sided or a double sided *t*-test, or none of these?

3.3.10.1 The meaning of "degrees of freedom"

What is still somewhat unclear in the context of parameter testing is the notion of "degrees of freedom", as it appears in the *t*- and *F*-statistics. When there is one parameter of interest, the term "degrees of freedom" can be explained as follows. The data consists of *N* data points and we try to estimate *L+1* parameters: *L* nuisance parameters and one parameter of interest, θ. One could say that we try to solve a system of *N* equations with *L+1* unknowns, because every data point gives one constraint, or one equation. If we deal with a case that $N>L+1$, there are more equations than unknowns and we call the system of equations *over-determined*. Such system has no solution, but by minimizing the distance between model and data, we can account for the background noise and derive a parameter estimate. It is intuitively clear that for a fixed model complexity *L*, the more data, the larger *N* and the better the quality of the fit. The reverse seems also true. If *N* is fixed and *L* is increased the solution will eventually become instable and non-solvable. From both arguments it follows that, the degrees of freedom (*N-L-1*), refer to the level of over-determinedness. The higher the degrees of freedom, the higher the level of over-determinedness and the more reliable parameter estimate we may expect. This terminology is consistent with the fact that for a fixed *t*-value, the *p*-value (area under the curve from *t* to infinity) decreases with an increasing number of degrees of freedom. In other words, the more data points or the less parameters, the higher the significance (the smaller the *p*-value) of an observed *t*-value.
From this discussion it becomes clear that the nuisance parameters must be accounted for when performing a *t*-test. Although the nuisance parameters themselves are not of interest, they can be important for a correct description of the data. Furthermore, they "eat statistical power" from the data because part of the data "must be used to estimate the nuisance parameters and is not available for the parameter of interest". The reduction of statistical power results from a reduction in the degrees of freedom *N-L-1*.

3.3.10.2 When the null hypothesis is not 0

In the previous sections it was demonstrated how the *t*- and *F*-tests can be applied to test whether a certain estimated parameter, or a group of parameters, deviates from zero. In some

applications, one might be more interested to test whether a certain parameter deviates from θ_0, where θ_0 is a known value. In those situations, the same theory as described before can be applied, with a small adaptation which consists of subtracting θ_0 from θ to form a new variable θ'. Starting from equation (3.8) it follows that we get a new data vector **r'**=**r**- θ_0**b**

$$\begin{aligned}\mathbf{r}-\theta_0\mathbf{b} &= (\theta-\theta_0)\mathbf{b}+\mathbf{S}\boldsymbol{\varphi}+\boldsymbol{\eta}\\ \mathbf{r}' &= \theta'\mathbf{b}+\mathbf{S}\boldsymbol{\varphi}+\boldsymbol{\eta}\end{aligned} \quad . \tag{3.56}$$

If we now impose the null hypothesis θ'=0, and derive the *t*-test from **r'** instead of **r**, this is equivalent to imposing the hypothesis θ= θ_0 on the original data.

3.3.10.3 Comparing two averages using the *t*-statistic

The use of the *t*-statistic might be more familiar to the reader in the context of testing whether the averages of two sets of observations are significantly different. It will be shown here that this situation is a special case of the theory presented here. Suppose the two data sets are graphically presented as in figure 3.6. The N_0 black dots represent the first group and the N_1 open dots represent the second group. We here assume that the data of both groups are collected in a row r_n, of which the first N_0 elements belong to the first group and the remaining N_1 elements belong to the second group. Then, in a more familiar use of the *t*-test, one would compute the average A_0 and estimated standard deviation $\hat{\sigma}_0$ of the first group, and similarly A_1 and $\hat{\sigma}_1$ of the second group,

$$\begin{cases} A_0 = \dfrac{1}{N_0}\displaystyle\sum_{n=0}^{N0-1} r_n & \text{and} \quad \hat{\sigma}_0 = \sqrt{\dfrac{1}{N_0}\displaystyle\sum_{n=0}^{N0-1} {r_n}^2 - {A_0}^2} \\ A_1 = \dfrac{1}{N_1}\displaystyle\sum_{n=N0}^{N1+N0-1} r_n & \text{and} \quad \hat{\sigma}_1 = \sqrt{\dfrac{1}{N_1}\displaystyle\sum_{n=N0}^{N1+N0-1} {r_n}^2 - {A_1}^2} \end{cases} . \tag{3.57}$$

and one would compute the *t*-statistic with *N*-2 degrees of freedom as

$$t_{N-2} = \frac{\sqrt{N-2}}{\sqrt{\dfrac{1}{N_0}+\dfrac{1}{N_1}}}\frac{A_0-A_1}{\sqrt{N_0\hat{\sigma}_0^2+N_1\hat{\sigma}_1^2}} \quad , \tag{3.58}$$

with N=N_0+N_1.

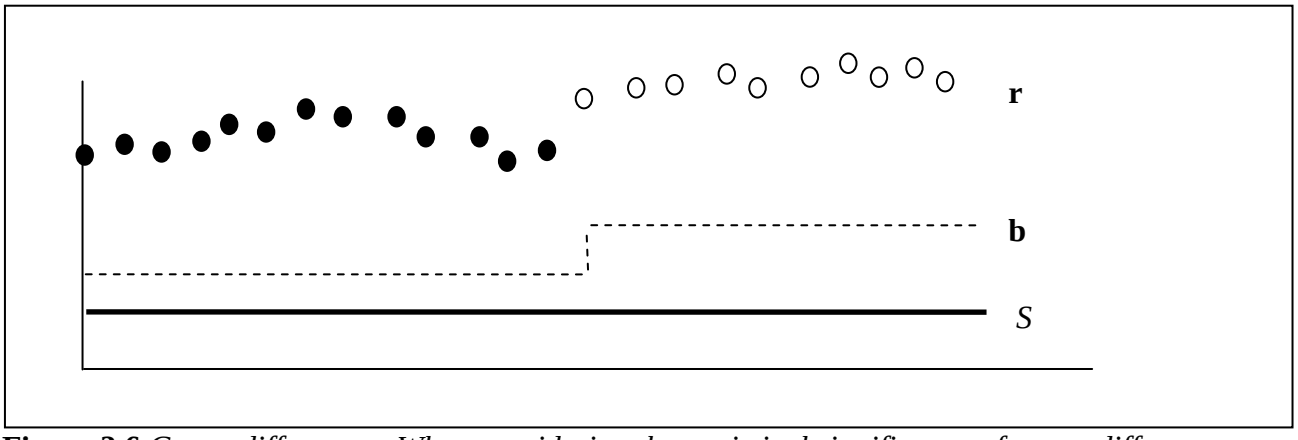

Figure 3.6 *Group differences. When considering the statistical significance of group differences, two approaches are equivalent: applying a t-test on the group averages, or applying a linear model with a jump-parameter of interest and a constant as nuisance parameter.*

An alternative approach would be to consider the data points r_n as a time series. What interests us is whether there is a jump in the time series if we pass from the first N_0 elements to the last N_1 ones. Such a jump could be modelled by the jump-vector **b**, represented in figure 3.6. We have a priori no idea how large the jump will be and therefore it is described with an unknown parameter θ that has to be multiplied with the jump function **b**. A complication is that we also do not know the offset, the vertical level that is common to both groups of points. The jump occurs with respect to the unknown offset φ which is multiplied with the constant vector $\mathbf{s}=(1,1,1\ldots1)^T$. So

$$b_n \equiv \begin{cases} 1 & \text{if} \quad n < N_0 \\ 0 & \text{if} \quad n \geq N_0 \end{cases} \quad \text{and} \quad s_n = 1 \qquad . \tag{3.59}$$

With this view, the question whether the data of group 0 are different from those in group 1 translates into the question whether the parameter $\hat{\theta}$ is different from zero, when estimated with the data model

$$\mathbf{r} = \theta\mathbf{b} + \varphi\mathbf{s} + \boldsymbol{\eta} \qquad , \tag{3.60}$$

where φ is a nuisance parameter. Therefore, we have L=1 and we compute the t-statistic according to (3.49)

$$t_{N-2} = \sqrt{N-2}\,\frac{1}{\sqrt{\mathbf{b}'^T\,\mathbf{b}'}}\,\frac{\mathbf{b}'^T\,\mathbf{r}'}{\sqrt{\mathbf{r}'^T\,P_{b'}\mathbf{r}'}} \qquad , \tag{3.61}$$

where the primed symbols result from removing the average, or $\mathbf{r'}=P_S\mathbf{r}$ and $\mathbf{b'}=P_S\mathbf{b}$. It appears that equation (3.61) is equivalent to (3.58) when the detailed substitutions are done. This is left as an exercise for the interested student. The following intermediate results might be helpful in this.

$$\begin{cases} \mathbf{b}'^T\,\mathbf{b}' = \mathbf{b}^T\mathbf{b} - \dfrac{\mathbf{b}^T\mathbf{s}\mathbf{s}^T\mathbf{b}}{\mathbf{s}^T\mathbf{s}} = N_0 - \dfrac{{N_0}^2}{N} = \dfrac{N_0 N_1}{N} \\ \mathbf{b}'^T\,\mathbf{r}' = \mathbf{b}^T\mathbf{r} - \dfrac{\mathbf{b}^T\mathbf{s}\mathbf{s}^T\mathbf{r}}{\mathbf{s}^T\mathbf{s}} = \sum_n^{N0-1} r_n - \dfrac{N_0}{N}\sum_{n=0}^{N} r_n = \dfrac{N_0 N_1}{N}(A_0 - A_1) \\ \mathbf{r}'^T\,\mathbf{r}' = \mathbf{r}^T\mathbf{r} - \dfrac{\mathbf{r}^T\mathbf{s}\mathbf{s}^T\mathbf{r}}{\mathbf{s}^T\mathbf{s}} = \sum_{n=0}^{N-1} {r_n}^2 - \dfrac{1}{N}\left(\sum_{n=0}^{N-1} r_n\right)^2 = N\left(\dfrac{1}{N}\sum_{n=0}^{N-1} {r_n}^2 - \left(\dfrac{1}{N}\sum_{n=0}^{N-1} r_n\right)^2\right) \\ \mathbf{r}'^T\,P_{b'}\mathbf{r}' = \mathbf{r}^T P_S\mathbf{r} - \dfrac{\left(\mathbf{b}^T P_S\mathbf{r}\right)^2}{\mathbf{b}^T P_S\mathbf{b}} = \sum_{n=0}^{N-1} {r_n}^2 - \dfrac{1}{N}\left(\sum_{n=0}^{N-1} r_n\right)^2 - \dfrac{N_0 N_1}{N}(A_0 - A_1)^2 \\ \qquad = N_0\left(\dfrac{1}{N_0}\sum_{n=0}^{N-1} {r_n}^2 - {A_0}^2\right) + N_1\left(\dfrac{1}{N_1}\sum_{n=N0}^{N-1} {r_n}^2 - {A_1}^2\right) \\ \qquad = N_0\hat{\sigma}_0^2 + N_1\hat{\sigma}_1^2 \end{cases} \tag{3.62}$$

Exercise 3.25 In the analysis of the group differences, does it matter whether we take s_n=1, s_n=2 or s_n=3? Why (not)?

Exercise 3.26 In the analysis of the group differences, a jump function **b** was used. Does it matter whether this function jumps from 0 to 1, or from 1 to 0, or from 2 to 3? Hint: In the equation for the *t*-statistic only **b'** appears.
Exercise 3.27 In the "standard" application of the *t*-test, two parameters are computed from the data, $\hat{\sigma}_0$ and $\hat{\sigma}_1$, which could be interpreted as estimates of the standard deviations in both groups. Does this imply that the *t*-test is based on the assumption that the true standard deviations of both groups are different? If not so, how could group-varying standard deviations be incorporated, assumed that their relative ratios are known a priori? Hint: what is assumed on $\boldsymbol{\eta}$ in equation (3.60)?
Exercise 3.28 In the example presented in figure 3.6 a *t*-test is applied to detect differences between two averages. How would this method be applied to fMRI data, when attempting to detect contrasts between scans made in activation and in rest condition?
Exercise 3.29 Two equivalent ways of using a *t*-test for group differences were presented. Suppose one also assumes that there is a linear trend in the data. How would you test group differences? In the case if figure 3.6, would the assumption of a trend make the group differences more significant or less significant?

3.3.10.4 Application of t- tests with evoked potentials

When MEG or EEG is recorded in an evoked potential/evoked magnetic field paradigm, the same stimulus is applied repeatedly, and the resulting signals are averaged, triggered by the stimulus. The presentation of the stimuli should ideally by presented with large enough intervals to prevent that that the tail of the *n*-th response overlaps with the head of the (*n*+1)-th response. On the other hand, the larger the inter-stimulus interval, the longer the duration of the whole experiment. For these reasons, it is important to study the duration of the response.
The response duration can be determined by testing whether a significant response is still present after some fixed time after the stimulus, or at a certain interval after the stimulus. Because the brain response on the stimulus is interfered by the ongoing MEG/EEG, which is modelled noise, the brain response can only be studied in statistical terms. When *t*- or *F*-tests are applied, one first needs a parameter model describing the brain responses in detail.
The simplest model to describe brain responses is to assume that each time a stimulus is applied, the brain region responsible for processing the stimulus reacts identical and the only reason why the signal looks different is because of the noise. This assumption is valid for any channel and for any moment after the stimulus (assuming no overlap). If the MEG/EEG signal at a certain channel, at a certain time after stimulus *n* is indicated by r_n, the constant brain response is given by θ and the noise is given by η_n, we have $r_n=\theta+\eta_n$. Therefore, the response estimation problem can be classified as the linear, single parameter case, and no parameters of interest, i.e. equation (3.12) with $\mathbf{b}=(1,1,\ldots,1)^T$. Hence, from (3.19) with $P_S=I$ one obtains

$$\hat{\theta} = \frac{\mathbf{b}^T\mathbf{r}}{\mathbf{b}^T\mathbf{b}} = \frac{\sum_n r_n}{N} = \frac{1}{N}\sum_n r_n \quad , \qquad (3.63)$$

confirming that the constant response assumption results in a simple averaging algorithm to find its ML-estimate. The *t*-statistic can be computed as

$$t_{N-1} = \frac{\hat{\theta}}{\hat{\sigma}_{\hat{\theta}}} = \sqrt{\frac{N-1}{\sum_n r_n^2 - \left(\sum_n r_n\right)^2}} \sum_n r_n \quad . \tag{3.64}$$

Exercise 3.30 Should the *t*-test of (3.64) be applied single or double sided?

To determine the response duration, one could start at the last sample, perform a *t*-test and if the NULL-hypothesis is not rejected, continue with the preceding sample until the first sample with a significant response is detected. The stimulus duration could be estimated as the interval after the stimulus, from the first sample to the one where a significant signal had been detected.

4 Application to fMRI

4.1 The principles of fMRI analysis

Parameter estimation theory is very well applicable to fMRI data analysis and in particular its linear variant in combination with parameter testing. As described before, in fMRI experiments, a series of (fast) scans is made from a subject, who alternately is performing a task or is at rest. The key idea is to perform a statistical test on the difference of scans made during rest and during activation. However, for several reasons a direct *t*-test between the averages of these scan groups is not recommendable. Firstly, scan time series are affected by slowly varying trends and by physiological artifacts such as heart beat, respiration and motion. Secondly, it should be accounted for the fact that the BOLD (Blood Oxygen Level Dependent) signal does not appear instantaneously, but shows a delay and a gradual build-up. Therefore the effect of activation varies over time. A direct *t*-test would ignore all this and lead to invalid detections of activity.

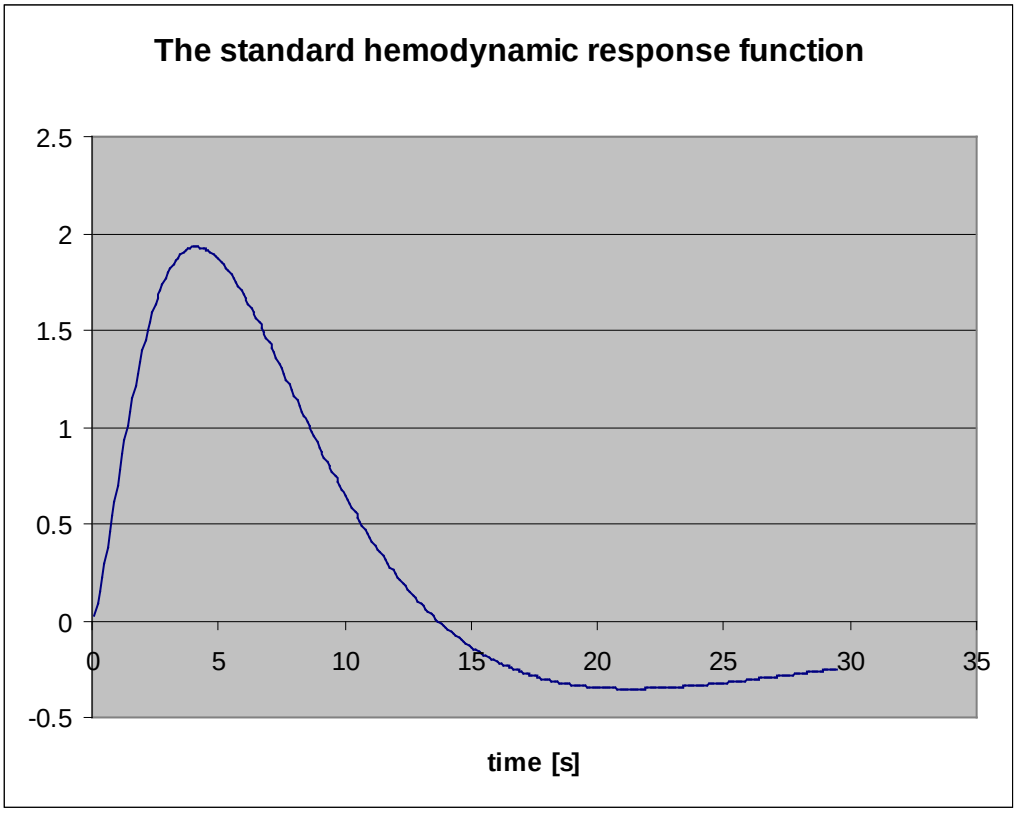

Figure 4.1 *The standard hemodynamic response function is shown. This function described the expected BOLD signal upon one unit of activation, given to the subject at time t=0.*

4.1.1 The hemodynamic response function

The build-up of the fMRI response is usually described with a hemodynamic response function (HRF). An HRF describes the course of the fMRI signal, as a response on a unit of activity, presented at *t*=0. The delay of the fMRI signal with respect to electrical activation is represented by the shift of the peak of the HRF with respect to the origin. An example of a HRF is plotted in figure 4.1. Although it is known that HRFs tend to vary over subject, brain region and nature of

activation, in fMRI analysis often the standard (canonical) HRF is used. The standard HRF is mathematically represented by

$$h(t) = c_1\left(\frac{t}{t_1}\right)^{b_1} \exp(-t/t_1) + c_2\left(\frac{t}{t_2}\right)^{b_2} \exp(-t/t_2) \quad . \qquad (4.1)$$

Here t_1, t_2, b_1, b_2, c_1, and c_2 are fixed parameters that determine the shape of the HRF. Standard values are: t_1=5 s, t_2=10 s, b_1=0.8, b_2=0.9, c_1=1, c_2=-0.3. The negative value of c_2 represents the effect that the initial increase of drained oxygenated blood, is followed by a decrease.

> **Exercise 4.1** Explore the dependence of $h(t)$ on t_1, t_2, b_1, b_2, c_1, and c_2. Try to model hemodynamic responses with similar shapes but different delays.

One of the central assumptions in most fMRI analyses is that the hemodynamic system behaves as a linear, time invariant (LTI) filter applied to the activation function (the function that indicates when the subject is at rest or activated)

$$a_n = \begin{cases} 0 & \text{during rest} \\ 1 & \text{during activation} \end{cases} \quad . \qquad (4.2)$$

In practice, the function a_n typically consists of a block function, with 5 to 10 zeroes, followed by the same number of ones, and so on. The LTI assumption makes it possible to predict the response of the brain hemodynamics upon and arbitrary a_n by computing a convolution between a_n and the discretized HRF:

$$b_n = \sum_{m=0}^{M-1} h_m a_{n-m} \quad , \qquad (4.3)$$

with

$$h_m = h(m\,TR) \quad . \qquad (4.4)$$

Here, TR is the time between two samples (fMRI repetition time). The meaning of equation (4.3) is that at time n the hemodynamic response consists of several contributions that can be summed directly. First, there is the immediate response $a_n h_0$ on a_n. Then, there is the ongoing response $a_{n-1}h_1$ on the input a_{n-1} one sample earlier. The subscript of h is now increased by one because one unit of time has passed between present and the response one time sample earlier. The other contributions to b_n have a similar origin: the further we look back in time, the lower the index of a, and the further its response has evolved and the larger the index of h. The LTI-assumption ensures that the simple summation of all the effects from the past predict the present signal at the output of the filter.

> **Exercise 4.2** What is the range of n in equation (4.3)?
> **Exercise 4.3** When TR=3s, and the HRF of figure 4.1 is adopted, how large should M in equation (4.3) be?

4.1.2 Rejecting the null-hypothesis

When the activation function is set and a hemodynamic response function is chosen, the fMRI signal can be described by

$$r_n = \theta\, b_n + \sum_l S_{nl}\varphi_l + \eta_n$$
$$\mathbf{r} = \mathbf{b}\theta + S\boldsymbol{\varphi} + \boldsymbol{\eta} \quad . \qquad (4.5)$$

where S_{nl} represent effects that we wish to disregard from the correlation analysis. Thus here θ is the parameter of interest and φ_l represent the parameters of no interest. This signal description is valid for all voxels, although the parameters θ and φ_l vary over voxels. To detect the activated regions, the statistical significance of $\hat{\theta}$ is tested using a *t*-test. This yields a *t*-value for every voxel, or a *t*-image, which is thresholded at a certain value corresponding to a certain *p*-value and an image (called *statistical map*) is obtained indicating all "active" voxels. There can also be voxels with a large negative *t*-statistic, and if such voxels survive the threshold, it indicates a significant "de-activation".

Exercise 4.4 Which assumptions are made so far in the correlation of the background noise $\boldsymbol{\eta}$ between different voxels? What about the noise levels at different voxels?
Exercise 4.5 Suppose that one would forget to include an offset (constant level) as confounder in the fMRI analysis. What would happen? Would one find an abnormal amount of activation, deactivation or no activation at all?
Exercise 4.6 If in the hemodynamic response model (4.1) both c_1 and c_2 are multiplied with the same number, does this have an effect on the results of the fMRI analysis?

A very general point of concern is the meaning of a rejected NULL-hypothesis. It would be tempting to assume there is "an effect". In the fMRI application, one would conclude that at a certain voxel, the corresponding brain area would be involved in the performance of the task or the processing of the stimulus given to the subject. However, one should realise that this conclusion would be based on the assumptions underlying equation (1): the number and shape of the trends and the shape of the reference function (which depends on the assumed hemodynamic response function). Furthermore, we have assumed that the noise is Gaussian and that the covariance pattern is known. Very often the assumption of uncorrelated background noise is adopted. When rejecting the NULL-hypothesis, one should also consider the validity of all the other assumptions, before drawing far-reaching conclusions. In particular, the assumption of uncorrelated noise gives the tendency of drawing too optimistic conclusions.
Finally, even if all modelling assumptions are satisfied, it should be realised that the whole analysis is nothing more than a *correlation analysis*. If an effect is found this means in fMRI context that there is a significant correlation between the activation function and a certain location in the brain. However, this does not automatically mean that the activation function is also the direct *cause* of local brain activity, acting through the BOLD response on local changes of electrical activity. Alternative explanations should be explored. For instance, if fMRI is used to detect the contrast between an emotional and a neutral stimulus, it should be kept in mind that also the heart rate might be different in both conditions. And since a varying heart rate might also have an effect on the hemodynamics of the brain and therefore on the fMRI signal, it cannot be excluded that the detected brain region does not represent a BOLD response on electrical activation but in reality is a hemodynamic response to heart rate, without modulation of local electrical brain activity.
Similarly, when motor activation is studied, e.g. in clinical patients, it should be kept in mind that motor activity may be correlated to head motion, and head motion may cause signal distortions that have nothing to do with the BOLD effect or to brain activity. In other words, one should be very much aware that activated brain regions may just be artefacts.

4.1.3 Preprocessing of fMRI data

The statistical analysis of fMRI data is usually not performed directly on the raw data because this version of the data may contain errors that cannot simply be treated by adding parameters of no-interest. Instead the data is "pre-processed" in order to obtain "clean" data that are suitable for statistical analysis.

As mentioned in the chapter one, image matching algorithms can be used to detect and remove motion effects from the raw data. In this analysis, the rotation and translation of each scan is determined with respect to (for example) the first scan, and using these transformations each scan is interpolated onto a fixed grid. In this way, a new series of scans is obtained wherein motion effects are eliminated, at least in principle. However, this is of course only true to some extend. One of the assumptions made in the motion correction step is for instance, that the raw data consists of a series of scans that differ, apart from the noise, only by a ridged transformation. This means that it is implicitly assumed that the motion takes place within two scans and not during a scan. However, in most fMRI scan sequences there is no or only a very small dead time between scans, and therefore the motion assumption is very unrealistic. As a result, motion corrected fMRI data will contain at most a reduced motion effect, but motion will not be eliminated completely.

The motion correction algorithm yields apart from the corrected data, six time series of motion parameters. To reduce motion effects still further, these time series are usually included as effects of no interest in the statistical analysis of fMRI data. Furthermore, visual inspection of the motion parameters can indicate that at some point in time motion might be exceptionally large (typically more than 1 mm). One might decide to remove the corresponding scans from the analysis.

> **Exercise 4.7** The motion parameters can be expressed in several ways. One possibility is to express the transformation from each scan to the first, another one is to recomputed them as the transformation from one scan to the next. Which of the two approaches is most appropriate to used as effects of no interest?

Another pre-processing step is to spatially smooth fMRI data. This implies that each voxel is replaced by a weighted sum of the voxel itself and its nearest neighbours. The further away the neighbour, the smaller its weight. Most often a Gaussian weighing profile is adopted. The rationale for this processing step is that it is assumed the noise is uncorrelated over voxels, and by computing the spatial average, the noise level decreases. The price one has to pay is a decreased spatial resolution. The problem is however, that no complete theory has been formulated yet which in detail links the optimal amount of smoothing to the amount of spatial correlation of the background noise.

4.2 The multiple comparison problem

> **Exercise 4.8** Suppose an fMRI scan consists of 50.000 voxels. On each voxel a *t*-test is performed with a *p*-value of 5 %. How many activated voxels would be detected on average, when the data consisted of noise only? How large volume does it represent?

If you answered Exercise 4.8 correctly, you should realise that large part of the brain can be detected as correlated to the stimulus, whereas in reality nothing happened there. The central cause of this problem is that many statistical tests are done, at a moderate *p*-level. In other words, if you look long enough you will always find the desired effect in the noise. This problem is a very general one in scientific studies, where multiple statistical test are done. If only the positive

tests are reported, and the failing tests are ignored, the scientific report is useless. The problem is often referred to as the “multiple comparison problem”

> **Exercise 4.9** In section 3.3.7.4 multiple *t*-tests were applied to detect the response interval. Does it mean that here also the multiple comparison problem is involved?

A simple and effective way to avoid this problem is to consider the “family wise error” (FWE), which is the probability that *in any* of the *K* voxels, a false positive detection occurs. This probability is smaller than the chance on an FP in voxel 0, plus the chance on an FP in voxel 1,…, plus the chance on an FP in voxel *K-1*. If all voxels are tested at a probability *p,* then FWE<*Kp*. This conclusion leads the so-called Bonferroni-correction. When all tests are performed at a level of *p*/*K*, one can be sure that not any false positive occurs in the thresholded statistical map, with probability *p*.
The drawback of the Bonferroni-correction is that it is very severe and therefore it becomes difficult to demonstrate an effect. In other words, the *statistical power* of Bonferroni-correction is very low. One remedy is to apply a Bonferroni-correction on a part of the image, instead of the whole scan. This will reduce *K*, and improve statistical power. For instance, we know that with fMRI no activation will occur outside the brain. So if the statistical testing is restricted to the brain area, we can reduce the correction *K* by a factor of 2, dependent on field of view of the scan. If one is only interested in activation in the motor cortex, a region of interest (ROI) could be defined and statistical testing would be applied to even fewer voxels. However, to be “fair”, the ROI should be defined a priori, before any data is analysed, and the ROI may not be changed anymore in the exploration phase of the data analysis.
Several alternatives have been proposed in the literature, to obtain an intermediate position between Bonferroni-correction and no correction at all. One approach is to try to account for the fact that statistical tests at different voxels are not completely independent. In particular, when spatial smoothing has been applied on fMRI data, a positive test at one voxel would tend to predict a positive outcome at its neighbouring voxel. When knowledge on spatial correlations are brought into account, a corrected *p*-value needs not to be so severe as in Bonferroni, where all tests are assumed to be independent. Methods that are based on such considerations can be very involved.

> **Exercise 4.10** Suppose that MEG data is used to detect functional connectivity between different parts of the brain. For that purpose, the correlations between all pairs of sensors are computed, as well as the *p*-value representing the chance that the computed correlation value would be exceeded if the data consisted of pure noise. Suppose there are 100 sensors involved. If a Bonferroni-correction would be applied on the observed correlations, how large should *K* be? (*K* is the value through which the uncorrected *p*-level should be divided).

An alternative approach is to leave the concept of family wise error because control of FWE does not allow any false positive detection in the whole region of interest. Instead one could control the expected number of false positives as a fraction *q* of all detections. Here a detection is a voxel at which an effect is detected, truly or falsely. The method of Benjamini and Hochberg to determine the active voxels at false discovery rate (FDR) *q* is relatively easy to explain, though it is much more difficult to explain *why* it works. The latter question is therefore beyond the scope of this course.
First, one determines for each voxel *k* the probability p_k that in absence on an effect, a statistic would exceed the observed one. In other words, if *t*-tests are applied, for each voxel the quantity *t* would be computed according to equation (3.27), and using standard mathematical tables or

software libraries of the Student-*t* distribution, these *t*-values would be converted to *p*-values p_k. Then, instead of thresholding this p_k-map, the *p*-values are ordered from small to large. These ordered *p*-values are indicated as follows: $p_{(0)}$, $p_{(1)}$, $p_{(2)}$, $p_{(3)}$,…$p_{(K-1)}$. So the voxel with the most significant effect is $v_{(0)}$, etc. Then, the index *i* is determined, as being the largest ordered index for which

$$p_{(i)} \leq (i+1)\frac{q}{K} \quad . \tag{4.6}$$

Then all voxels with $p_k \leq p_{(i)}$ are declared active. The mathematical theorem is that the expected value of the false discovery rate, of all voxels declared active using this method, is at most equal to *q*.
This theory can at best be explained with a small example. Suppose that there are K=5 voxels, for which it is found that p_0=0.10, p_1=0.02, p_2=0.17, p_3=0.21 and p_4=0.05. Suppose one is satisfied with an expected FDR of q=20%. Then we have q/K=0.04, and ordering the *p*-values gives:

$p_{(4)}=p_3=0.21>5q/K = 0.20$.
$p_{(3)}=p_2=0.17>4q/K = 0.16$,
$p_{(2)}=p_0=0.10\leq 3q/K = 0.12$; (4.7)
$p_{(1)}=p_4=0.05\leq 2q/K = 0.08$;
$p_{(0)}=p_1=0.02\leq q/K = 0.04$;

Hence, only the voxels with the lowest three *p*-values are declared active: v_0, v_1 and v_4. Of these three voxels, 20% is expected to be declared active, falsely. Of course, 20% of three voxels is less than a single voxel, but this is due the illustration of the theory by means of a small numerical example.
It should be noted that, strictly speaking, according to the Benjamini and Hochberg-theory, $p_{(1)}$ and $p_{(0)}$ do not need to be tested anymore, because $p_{(3)}$ already passed the test of equation (4.6). In the above example it would not matter, but suppose that we had p_1=0.045 (instead of 0.02). Then, p_1 would stay the smallest observed *p*-value, so one would remain that $p_{(0)}=p_1$. However, in that situation, we would have that $p_{(0)} > q/K$. But since the *p*-value of i=2 would stay the largest *p*-value for which the test is passed, we would maintain that the voxels v_0, v_1 and v_4 were active, with the same false detection rate.
Although this theory is here demonstrated for the analysis of fMRI data one should keep in mind that it is relevant for many more studies, where multiple statistical tests are performed. Generally, the omission of a multiple comparison correction leads to too optimistic conclusions, which are more based on pure chance than on solid scientific data analysis.

4.3 Estimation of the hemodynamic response

In most fMRI applications a priori assumptions about the shape of the hemodynamic response are made. The activation function a_n is then convolved with the canonical the hemodynamic response function (HRF) (4.1), assuming certain parameters t_1, t_2, b_1, b_2, c_1, and c_2. From studies where different parameters settings are used and compared it is known that the HRF varies over subjects (e.g. age), task and brain region. To explore the shape of the HRF more systematically, one could attempt to estimate it from the data, instead of forcing a known shape. This goal can be achieved in different ways. One way is to treat the HRF parameters (t_1, t_2, b_1, b_2, c_1, and c_2) as

unknowns, i.e. determine maximum likelihood by exploring all combinations of these parameters. However, as will be detailed in the next chapter, this leads to a non-linear optimization problems, which are much harder to solve than linear parameter estimation problems discussed in chapter 3.
An alternative solution is disregard the parameterization of the HRF, and to treat the parameters h_n, appearing in the convolution (4.3) as unknown parameters of interest. Then the data model takes the following shape:

$$r_n = \sum_{m=0}^{M-1} h_m a_{n-m} + \sum_l S_{nl}\varphi_l + \eta_n \quad . \qquad (4.8)$$

To interpret this model in terms of the theory presented in chapter 3, the shifted activation function a_{n-m} is presented as B_{nm}, and the HRF h_m is presented as θ_m. With these substitutions one obtains:

$$r_n = \sum_{m=0}^{M-1} B_{nm}\theta_m + \sum_l S_{nl}\varphi_l + \eta_n$$

or . (4.9)

$$\mathbf{r} = B\boldsymbol{\theta} + S\boldsymbol{\varphi} + \boldsymbol{\eta}$$

Therefore, the HRF can be estimated the standard theory and the significance of $\hat{\boldsymbol{\theta}}$ can be tested with an *F*-test.

> **Exercise 4.11** At first sight it would seem that estimating the HRF by using h_m as free parameters not any a priori assumption at all is required on the HRF. Is that true?
> **Exercise 4.12** In this section two methods are described to estimate the shape of the HRF. Compare the number of estimated parameters in both methods.
> **Exercise 4.13** When the HRF estimation method is applied in practice, one will find that at some or many voxels the HRF is not significant. What does that mean?
> **Exercise 4.14** One of the central assumptions of fMRI analysis is that the recorded BOLD signal is a filtered version of the activation time function a_n. Furthermore, it is assumed that this filter is linear and time invariant, which allows us to use equation (4.3). Describe an experimental setup to verify in how much this assumption is correct.

4.3.1 Non-causal HRF

One assumption in fMRI analysis is that the activation function a_n *precedes* the BOLD-response. This assumption seems trivially true. However, its validity depends on the precise meaning of a_n. If a_n indicates at which intervals a subject is stimulated then it surely must be true, unless there would be some anticipation effect of the subject's brain, i.e. unless the subject knows beforehand when the next stimulus will be presented. If, on the other hand, a_n represents the moments in time that the subject has pressed on a button, there might be brain activity, related to the decision to press the button, that precedes the activation function resulting in a "non-causal" HRF.

> **Exercise 4.15** Where and in which equation is the assumption exploited that the BOLD signal *follows* the activation function?

If we are interested in "non-causal" effects, we need to extend the estimation model slightly. Instead of (4.8) we use

$$r_n = \sum_{m=-M_{prior}}^{-M_{prior}+M-1} h_m a_{n-m} + \sum_l S_{nl}\varphi_l + \eta_n \quad . \qquad (4.10)$$

The extension implies that the activation function is not only shifted to the left, but also to the right. The right shift is M_{prior} samples whereas the total number of parameters of interest is still indicated with M. In other words, the recorded signal r_{n} at time n, does not only depend on time samples of a_{n} preceding n, but also on later time samples. So, in principle, equation (4.10) completes the model because non-causal effects are incorporated. However, when an F-test is done on the estimated HRF $\hat{\boldsymbol{\theta}}$ one tests the significance of the *complete* HRF. In other words, a rejection of the NULL-hypothesis implies that *some* components, deviate significantly from zero. It does not tell us which ones. In particular, one still does not know if there are non-causal effects.
To test specifically for non-causal effects, only these parameters should be used as parameters of interest and the other ones should be added to the nuisance parameters. In a formula, one should use

$$\begin{aligned} \boldsymbol{\theta} &= (h_{-M_{prior}}, h_{-M_{prior}+1}, ..., h_{-1})^T \\ \boldsymbol{\varphi} &= (h_0, h_1, ..., h_{-M_{prior}+M-1}, \varphi_0, \varphi_1, ..., \varphi_{L-1})^T \end{aligned} \quad . \qquad (4.10)$$

Exercise 4.16 When testing for non-causal effects, which B and S should be used?
Exercise 4.17 When testing for non-causal effects, how many parameters of interest are there, how many parameters of non-interest and how many data points?
Exercise 4.18 Describe precisely how you would analyze fMRI data to test whether the HRF extends beyond 20 s.

The method to applied to test for non-causal effects applicable in a much more general settings. It often happens that it is unknown how many parameters must be included into a model. In the case of HRF estimation, there is no physiological law that tell is how large M should be. The approach followed here was to enlarge M, and to test whether the extra parameters deviate significantly from zero. Another example from brain imaging is the source localisation based on MEG/EEG. A model for the recorded MEG/EEG is assumed with a certain number of dipoles. By testing the significance of the amplitude of each of the dipoles, one obtains insight into the number of dipoles that are required in the model.

5 Appendix A. Some matrix-vector algebra

A vector is denoted with a bold face type, e.g. **v**, and a vector is always a column vector. The components of an *N*-dimensional vector are numbered from 0 to *N*-1. A column vector can be converted to a row vector by using the transposition operator $^{\mathrm{T}}$. So, if $\mathbf{v} = \begin{pmatrix} v_0 \\ \vdots \\ v_{N-1} \end{pmatrix}$ then $\mathbf{v}^{\mathrm{T}}=(v_0,\ldots,v_{N-1})$. Applying the transposition onto a row vector gives the corresponding column vector. So, in particular, one has $\mathbf{v}=(\mathbf{v}^{\mathrm{T}})^{\mathrm{T}}$. Furthermore, two column vectors $\mathbf{v_1}$ and $\mathbf{v_2}$ can be concatenated as $\mathbf{v} = (\mathbf{v_1}^{\mathrm{T}}, \mathbf{v_2}^{\mathrm{T}})^{\mathrm{T}}$.

A matrix is usually denoted by a capital. An *N*x*M* matrix *A* with matrix elements $\{A\}_{\mathrm{nm}}$ (*n*=0,…,*N-1* and *m*=0,…*M-1*) refers to the structure

$$A = \begin{pmatrix} A_{00} & \cdots & A_{0,M-1} \\ \vdots & \ddots & \vdots \\ A_{N-1,0} & \cdots & A_{N-1,M-1} \end{pmatrix} \quad . \qquad \text{(A.1)}$$

So, the matrix has *N* rows of length *M*, or *M* columns of length *N*. In this view a vector is a special type of a matrix, i.e. a vector is a matrix with only one column. The transposition operator is also defined for matrices:

$$A^T = \begin{pmatrix} A_{00} & \cdots & A_{N-1,0} \\ \vdots & \ddots & \vdots \\ A_{0,M-1} & \cdots & A_{N-1,M-1} \end{pmatrix} \quad . \qquad \text{(A.2)}$$

Therefore, if *A* is “tall” (more rows than columns), then A^T will be “broad”(more columns than rows), and vice verse.

Special types of matrices are *square* matrices, for which *M*=*N*. Such matrices are called *symmetric* if $\{\mathrm{A}\}_{\mathrm{nm}}=\{\mathrm{A}\}_{\mathrm{mn}}$. So a symmetric matrix invariant under “reflections” in the main diagonal, the diagonal from top-left to bottom-right. In terms of matrix transposition, a square matrix is symmetric if $A=A^T$. Furthermore, a square matrix is called a diagonal matrix, if all elements except the main diagonal are zero:

$$A = \begin{pmatrix} d_0 & 0 & 0 & 0 \\ 0 & d_1 & 0 & 0 \\ 0 & 0 & ... & ... \\ 0 & 0 & ... & d_{M-1} \end{pmatrix} \qquad \text{(A.3)}$$

When all diagonal elements of a diagonal matrix are 1, this matrix is called the identity matrix, and is indicated by I_N, where *N* refers to the number of rows or columns. Sometimes, when the dimensions are clear from the context, the subscript *N* is omitted.

Exercise A.1 Matrix *C* is a diagonal matrix. Is *C* symmetric?

Exercise A.2 Matrix *C* represents the covariance matrix of a vector $\boldsymbol{\eta}$. Why is *C* symmetric?

Exercise A.3 Suppose *C* is the covariance matrix of a vector $\boldsymbol{\eta}$. Which assumptions on the distribution of $\boldsymbol{\eta}$ are necessary to make *C* a diagonal matrix?

When A is matrix with M columns, and B a matrix with M rows, the matrix product $C = A$xB is defined as the NxK matrix:

$$C = \begin{pmatrix} \sum_{m=0}^{M-1} A_{0m}B_{m0} & \cdots & \sum_{m=0}^{M-1} A_{0m}B_{m,K-1} \\ \vdots & \ddots & \vdots \\ \sum_{m=0}^{M-1} A_{N-1,m}B_{m0} & \cdots & \sum_{m=0}^{M-1} A_{N-1,m}B_{m,K-1} \end{pmatrix} , \qquad \text{(A.4)}$$

where N is the number of rows of A, and K is the number of columns of B. In other words, the matrix elements of C consist of the inner products, of the rows of A and the columns of B. For matrix multiplication, one can verify that it satisfies the associative property, i.e. $(AB)C = A(BC)$, as long as the matrix multiplications are defined (i.e. the number of columns matches the number of rows in the next matrix), of course. The *commutative* property, valid for real numbers, is generally not valid for matrices. So AB will generally yield a different matrix than BA. It might even be the case that although AB is defined, that BA is not.

> **Exercise A.4** Give an example of a matrix A and a matrix B, such that AB is defined and BA is not.

If a matrix is multiplied with a scalar λ, it is meant that every matrix element is multiplied to λ:

$$\lambda C = \begin{pmatrix} \lambda C_{00} & \cdots & \lambda C_{0,M-1} \\ \vdots & \ddots & \vdots \\ \lambda C_{N-1,0} & \cdots & \lambda C_{N-1,M-1} \end{pmatrix} = C\lambda \qquad . \qquad \text{(A.5)}$$

Therefore, pre- and post-multiplying of a matrix with a scalar gives the same result. This is also true for vectors. One should be aware here that the scalar can sometimes take the form an inner product. For instance, when $\lambda = \mathbf{w}^T\mathbf{v}$, where $\mathbf{v}$ and $\mathbf{w}$ are vectors of the same dimensions, λ is a scalar and it is true that

$$\mathbf{v}^T\mathbf{w}C = C\mathbf{v}^T\mathbf{w} \qquad . \qquad \text{(A.6)}$$

At first glance, it would seem that (A.6) is wrong but simply because $\mathbf{w}^T\mathbf{v}$ is a scalar, it is right. If the transposition operator is applied onto a product of two matrices, the result is equivalent to applying the transposition on each matrix individually, and changing the order of matrix multiplication:

$$(AB)^T = B^T A^T \qquad . \qquad \text{(A.7)}$$

This can be verified by applying the definitions of matrix multiplication and transposition. Inner products between two N-dimensional vectors $\mathbf{v}$ and $\mathbf{w}$ can also be considered as matrix multiplications. A vector is simply a matrix consisting of one column, and therefore, $\mathbf{v}^T\mathbf{w} = \mathbf{w}^T\mathbf{v}$ is a 1x1 matrix, representing the inner product of $\mathbf{v}$ and $\mathbf{w}$. For any two vectors $\mathbf{v}$ and $\mathbf{w}$, one can also define the products $\mathbf{v}\mathbf{w}^T$ and $\mathbf{w}\mathbf{v}^T$. These products can be distinguished from the inner products by the fact that the transposition symbol is applied on the second vector of the pair. If $\mathbf{v}$ is an N-dimensional vector and $\mathbf{w}$ is an M-dimensional vector, the product $\mathbf{v}\mathbf{w}^T$ is an NxM matrix, containing all possible combinations of products of elements of $\mathbf{v}$ and elements of $\mathbf{w}$:

$$\mathbf{v}\mathbf{w}^T = \begin{pmatrix} v_0 w_0 & v_0 w_1 & \cdots & v_0 w_{M-1} \\ v_1 w_0 & v_1 w_1 & & v_1 w_{M-1} \\ \vdots & & \ddots & \vdots \\ v_{N-1} w_0 & v_{N-1} w_1 & \cdots & v_{N-1} w_{M-1} \end{pmatrix} = \left(\mathbf{w}\mathbf{v}^T\right)^T \quad . \qquad \text{(A.8)}$$

Exercise A.5 Show that the last equality of (A.8) is true. Hint: use (A.7) or express it in components.

When A is a square NxN matrix, the NxN matrix A^{inv} is defined if $A^{inv}A = AA^{inv} = I_N$. Not all square matrices have an inverse, only those for which all rows or all columns are linearly independent. For instance, a matrix with all elements equal to 1 does not have an inverse (N>1). For the inverse of a product of two matrices, once has a rule similar to the rule for the transposition of a product:

$$(AB)^{inv} = B^{inv} A^{inv} \quad , \qquad \text{(A.9)}$$

provided that the inverses of both A and B exist.
The inverse of a matrix can always be computed numerically, when it exists (and when it is not extremely large). Algorithms to do so are beyond the scope of this course. Some matrices have simple expressions for their inverse. An a example is a diagonal matrix:

$$A = \begin{pmatrix} d_0 & 0 & 0 & 0 \\ 0 & d_1 & 0 & 0 \\ 0 & 0 & ... & ... \\ 0 & 0 & ... & d_{M-1} \end{pmatrix} \Leftrightarrow A^{-1} = \begin{pmatrix} 1/d_0 & 0 & 0 & 0 \\ 0 & 1/d_1 & 0 & 0 \\ 0 & 0 & ... & ... \\ 0 & 0 & ... & 1/d_{M-1} \end{pmatrix} \qquad \text{(A.10)}$$

Exercise A.6 Show that (A.10) is true. Under which conditions does the inverse of a diagonal matrix exist?

In the case of independent Gaussian variables, with different variances, one has to deal with expressions like $v_0w_0/\sigma_0^2 + v_1w_1/\sigma_1^2 + v_2w_2/\sigma_2^2 + \ldots + + v_{N-1}w_{N-1}/\sigma_{N-1}^2$. Using diagonal matrices, this expression can be presented more concise as

$$\sum_{m=0}^{M-1} \frac{v_m w_m}{\sigma_m^2} = \mathbf{v}^T C^{inv} \mathbf{w} \qquad \text{with} \qquad \text{(A.11)}$$

$$\mathbf{v} = \begin{pmatrix} v_0 \\ v_1 \\ \\ v_{M-1} \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w_0 \\ w_1 \\ \\ w_{M-1} \end{pmatrix} \quad \text{and} \quad C = \begin{pmatrix} \sigma_0^2 & 0 & 0 & 0 \\ 0 & \sigma_1^2 & 0 & 0 \\ 0 & 0 & ... & ... \\ 0 & 0 & ... & \sigma_{M-1}^2 \end{pmatrix} \qquad \text{(A.12)}$$

Exercise A.7 Show that (A.11) is true.

The use of matrix vector notation can simplify many other expression that appear in parameter estimation theory. Suppose one has a signal s_n, n=0,…,*N-1* and one removes the offset, i.e. one subtracts from each measurement the average of all measurements:

$$s_n' = s_n - \frac{1}{N}\sum_m s_m \quad . \qquad \text{(A.13)}$$

This operation can be expressed in terms of matrix vector multiplications as follows. First, the N dimensional vector **s** and **e** are defined: s = $(s_0,\ldots,s_{N-1})^T$ and **e**=$(1,1,\ldots,1)^T$. Then the sum over all elements s_n can be expressed as $\mathbf{e}^T\mathbf{s}$ and the number of measurements N= $\mathbf{e}^T\mathbf{e}$. Therefore, the average of the signal equals $\mathbf{e}^T\mathbf{s}/\mathbf{e}^T\mathbf{e}$. This same number is subtracted from each element s_n. In vector terms, this means that the average times **e** is subtracted:

$$\begin{aligned}\mathbf{s}' &= \mathbf{s} - \mathbf{e}\frac{\mathbf{e}^T\mathbf{s}}{\mathbf{e}^T\mathbf{e}} \\ &= \mathbf{s} - \frac{\mathbf{e}\mathbf{e}^T}{\mathbf{e}^T\mathbf{e}}\mathbf{s} \\ &= (I_N - \frac{\mathbf{e}\mathbf{e}^T}{\mathbf{e}^T\mathbf{e}})\mathbf{s} \\ &= P\mathbf{s}\end{aligned} \quad , \qquad \text{(A.14)}$$

with

$$P = (I_N - \frac{\mathbf{e}\mathbf{e}^T}{\mathbf{e}^T\mathbf{e}}) \quad , \qquad \text{(A.15)}$$

Here, first the associative property was applied, $\mathbf{e}(\mathbf{e}^T\mathbf{s}) = (\mathbf{e}\mathbf{e}^T)\mathbf{s}$, to create a matrix $\mathbf{e}\mathbf{e}^T/\mathbf{e}^T\mathbf{e}$. Then, using $\mathbf{s}=I_N\mathbf{s}$, this matrix was combined with the identity matrix, to create a new matrix P. Applying P to the signal vector yield a new signal vector, with the offset removed. The matrix P defined in this way is an example of a *symmetric projection matrix*, for which $P=P^T$ and $P^2=P$. That the latter is true, one can understand because if the offset is removed from a signal, the averaged matrix elements are set to zero, and therefore applying the same operator a second time has no effect. It can also more formally be shown using the associative property of matrix multiplication:

$$\begin{aligned}P^2 &= (I_N - \frac{\mathbf{e}\mathbf{e}^T}{\mathbf{e}^T\mathbf{e}})(I_N - \frac{\mathbf{e}\mathbf{e}^T}{\mathbf{e}^T\mathbf{e}}) \\ &= (I_N - \frac{\mathbf{e}\mathbf{e}^T}{\mathbf{e}^T\mathbf{e}} - \frac{\mathbf{e}\mathbf{e}^T}{\mathbf{e}^T\mathbf{e}} + \frac{\mathbf{e}\mathbf{e}^T}{\mathbf{e}^T\mathbf{e}}\frac{\mathbf{e}\mathbf{e}^T}{\mathbf{e}^T\mathbf{e}}) \\ &= (I_N - \frac{\mathbf{e}\mathbf{e}^T}{\mathbf{e}^T\mathbf{e}} - \frac{\mathbf{e}\mathbf{e}^T}{\mathbf{e}^T\mathbf{e}} + \frac{\mathbf{e}(\mathbf{e}^T\mathbf{e})\mathbf{e}^T}{(\mathbf{e}^T\mathbf{e})^2}) \\ &= (I_N - \frac{\mathbf{e}\mathbf{e}^T}{\mathbf{e}^T\mathbf{e}} - \frac{\mathbf{e}\mathbf{e}^T}{\mathbf{e}^T\mathbf{e}} + \frac{\mathbf{e}\mathbf{e}^T}{\mathbf{e}^T\mathbf{e}}) \\ &= (I_N - \frac{\mathbf{e}\mathbf{e}^T}{\mathbf{e}^T\mathbf{e}}) \\ &= P\end{aligned} \quad . \qquad \text{(A.16)}$$

Exercise A.8 Given a series of N paired measurements (s_0,t_0), (s_1,t_1), (s_2,t_2),…,(s_{N-1},t_{N-1}). Then the correlation co-efficient between s and t is given as

$$\rho = \frac{\sum_n (s_n - \bar{s})(t_n - \bar{t})}{\sqrt{\sum_n (s_n - \bar{s})^2}\sqrt{\sum_n (t_n - \bar{t})^2}} \quad \text{with} \quad . \qquad \text{(A.17)}$$

$$\bar{s} = \frac{1}{N}\sum_n s_n \quad \text{and} \quad \bar{t} = \frac{1}{N}\sum_n t_n$$

Show how this expression can be simplified to $\rho = \frac{\mathbf{s}^T P \mathbf{t}}{\sqrt{\mathbf{s}^T P \mathbf{s} \;\; \mathbf{t}^T P \mathbf{t}}}$ using a vector notation and using the projection operator defined in (A.15).

Another example of a symmetric projection matrix is $A(A^TA)^{inv}A^T$, provided that the inverse of (A^TA) exists. Also, if P is a symmetric projection, then so is I_N-P. Therefore, also I_N-$A(A^TA)^{inv}A^T$ is a projector. The example given above is a special case hereof, wherein A consists of a single column.

Exercise A.9 Give an example of a matrix A such that A^TA does not have an inverse.
Exercise A.10 Show that if $(A^TA)^{inv}$ exists, $A(A^TA)^{inv}A^T$ is a symmetric projection matrix. Hint: show that the matrix is symmetric and show that applying the matrix twice is identical to applying it once.
Exercise A.11 Show that if P is a symmetric projection matrix, so is (I-P).

In the example of (A.13) a projector was used to remove the offset from a signal s_n. However, projectors can also be used to remove other components, such as trends or certain frequency components. When a single component is to be removed, proportional to **b**, the projector P_b with

$$P_{\mathbf{b}} = (I_N - \frac{\mathbf{b}\mathbf{b}^T}{\mathbf{b}^T\mathbf{b}}) \qquad , \qquad \text{(A.18)}$$

will remove all components from s that are proportional to **b**. Multiple components ($\mathbf{b}_0$,$\mathbf{b}_1$,..,$\mathbf{b}_{M-1}$) can removed from a signal **s**, by forming the projection matrix P_B with

$$P_B = (I_N - B(B^TB)^{inv}B^T) \qquad \text{(A.19)}$$

and

$$B = (\mathbf{b}_0, \mathbf{b}_1, ...\mathbf{b}_{M-1}) \qquad . \qquad \text{(A.20)}$$

In other words, the columns of B consist of all the column vectors $\mathbf{b}_0$ to $\mathbf{b}_{M-1}$.

Exercise A.12 Consider the line L through the origin and the point B with coordinates (1,2,3). Compute the coordinates of projection of the point Q = (5,6,7) onto the line L.

An application of (A.19) is the design of a digital filter that removes all 50 Hz components from a measured signal. When the sampling frequency τ is known, one defines the n-th component of $\{\mathbf{b}_0\}_n$ and $\{\mathbf{b}_1\}_n$ as

$$\{\mathbf{b}_0\}_n = \cos(2\pi \times 50 \times \tau\, n)$$
$$\{\mathbf{b}_1\}_n = \sin(2\pi \times 50 \times \tau\, n) \qquad , \qquad \text{(A.20)}$$

and uses (A.19) and (A.20) to obtain a (elementary) 50 Hz suppression filter. In signal processing a filter that removes specific frequency components is often called a *notch filter*.

> **Exercise A.13** Why are *two* vectors needed (and not one) in (A.20) to remove a single frequency component?

When A is a square MxM matrix, the determinant of is defined and denoted by det(A). When M=1, det(A)=a_{00}. For M>1, det(A) is defined recursively, by summing the products of the elements of the first column, and the determinants of the (M-1)x(M-1) matrices, consisting of A from which the first column and the row under consideration are deleted. Such determinant is insensitive to the addition of linear combinations of rows and columns, and it can be shown that det(A)≠0 if and only if the rows (or columns) are linearly independent.
Furthermore, it can be shown that, e.g.

$$\det\begin{pmatrix} d_0 & 0 & 0 & 0 \\ 0 & d_1 & 0 & 0 \\ 0 & 0 & ... & ... \\ 0 & 0 & ... & d_{M-1} \end{pmatrix} = d_0 d_1 ... d_{M-1} = \prod_{m=0} d_m \qquad \text{(A.21)}$$

So for diagonal matrices, the determinant equals the product of the diagonal elements. Finally, we have e.g. that

$$\det(\lambda A) = \lambda^M \det(A) \quad . \qquad \text{(A.22)}$$

Another scalar that is defined for square matrices is the trace of a matrix, denoted by Tr{A}. It equals the sum of the diagonal elements:

$$\mathrm{Tr}\{A\} \equiv \sum_{\mathrm{n}=0}^{\mathrm{N-1}} a_{\mathrm{nn}} \quad , \qquad \text{(A.23)}$$

One can verify that, since the trace is the sum of the diagonal elements, it is invariant for the application of matrix transposition,

$$\mathrm{Tr}\{\ AB^T\ \} = \mathrm{Tr}\{\ (AB^T)^T\ \} = \mathrm{Tr}\{\ BA^T\ \}. \quad , \qquad \text{(A.24)}$$

An eigenvector $\mathbf{v}$ of a matrix A is a vector such that

$$A\mathbf{v} = \lambda\mathbf{v} \quad , \qquad \text{(A.25)}$$

where the number λ is called the eigenvalue. When A is a symmetric NxN matrix, it can be shown that A has N orthonormal eigenvectors $\mathbf{v}_\mathrm{n}$, with real eigenvalues. This means that

$$\mathbf{v}_\mathrm{n}{}^\mathrm{T}\mathbf{v}_\mathrm{m} = \begin{cases} 0 & \text{if } n \neq m \\ 1 & \text{if } n = m \end{cases} \quad , \qquad \text{(A.26)}$$

As a consequence, if A is symmetric, it can be expressed as

$$A = V\Lambda V^T \quad , \qquad \text{(A.27)}$$

where Λ is a diagonal matrix containing the N eigenvalues λ_n and V is an orthonormal matrix ($VV^T = V^TV = I_N$) containing the N eigenvectors $\mathbf{v}_\mathrm{n}$ as columns.

For a symmetric projection matrix P, the eigenvalues must be either 0 or 1. This is true, because if $\mathbf{v}$ is an eigenvector of P, we have

$$\begin{aligned} P^2\mathbf{v} &= P\mathbf{v} = \lambda\mathbf{v} \\ &= \lambda^2\mathbf{v} \end{aligned} \quad . \qquad \text{(A.28)}$$

Therefore $(\lambda-\lambda^2)\mathbf{v}=0$ for any $\mathbf{v}$, and hence $\lambda=0$ or $\lambda=1$. Hence, in case of a symmetric projector, the decomposition equation (A.27) yields a diagonal matrix Λ with ones and zeroes. As a consequence, the rank of these types of matrices equals its trace:

$$\begin{aligned} \text{Rank}(P) &= \text{Rank}(V\Lambda V^T) \\ &= \text{Rank}(\Lambda V^T) \\ &= \text{Rank}(\Lambda) \\ &= \text{Tr}(\Lambda) \\ &= \text{Tr}(\Lambda V^T V) \\ &= \text{Tr}(V\Lambda V^T) = \text{Tr}(P) \end{aligned} \quad . \qquad \text{(A.29)}$$

Here it has been used that the rank of a matrix is not changed when it is pre- or post-multiplied with a regular matrix (such as V, which is orthonormal). For general matrices, it is quite time consuming to compute its rank, but for projectors one only needs to sum its diagonal elements. Applying this rule to the projector I_N-$A(A^TA)^{inv}A^T$, one finds e.g.

$$\begin{aligned} \text{Rank}(\text{I}_\text{N} - A(A^\text{T}A)^{\text{inv}}A^\text{T}) &= \text{Tr}(\text{I}_\text{N} - A(A^\text{T}A)^{\text{inv}}A^\text{T}) \\ &= \text{Tr}(\text{I}_\text{N}) - \text{Tr}(A(A^\text{T}A)^{\text{inv}}A^\text{T}) \\ &= N - \text{Tr}((A^\text{T}A)^{\text{inv}}A^\text{T}A) \\ &= N - M \end{aligned} \quad ,\text{(A.30)}$$

where M is the number of columns of A (Note that is was implicitly assumed that all these columns are linearly independent). In the fMRI context, this equation implies that for any trend that is included in the model, one degree of freedom is removed from the t- or F-statistic, thereby decreasing the significance with a small amount.

6 Appendix B. The multivariate Gaussian distribution and related distributions

A random variable is associated with a probability density function (pdf), representing the relative occurrence of the values taken by that random variable in an experiment wherein these values are repeatedly realised. Random variables are therefore very appropriate to describe measurement noise. In mathematical models on measurements one usually assumes that the observed data equals the true value plus a random variable representing the noise. In order to be able to make statements on the distribution of estimated parameters, assumptions need to be made on the probability density functions, or the distributions, of the noise. In this appendix several distributions are summarized and a few properties are discussed, as far as they are needed for the comprehension of this course on parameter estimation.
One of the most common distributions is the univariate Gaussian distribution. It is given by

$$f(\eta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\eta-\mu)^2}{2\sigma^2}} \quad . \qquad \text{(B.1)}$$

From a distribution one can derive the mean value, indicated by E{η}, and the variance, which is the expected value of $(\eta-\mu)^2$, or E{$(\eta-\mu)^2$}. In case of the Gaussian distribution one obtains

$$\mathrm{E}\{\eta\} \equiv \int_{-\infty}^{\infty} \eta \, f(\eta) \mathrm{d}\eta = \mu \qquad \text{(B.2)}$$

and

$$\mathrm{E}\{(\eta-\mu)^2\} \equiv \int_{-\infty}^{\infty} (\eta-\mu)^2 \, f(\eta) \mathrm{d}\eta = \sigma^2 \quad . \qquad \text{(B.3)}$$

The mathematical details of computing these integrals for the Gaussian distribution are beyond the scope of this course. The square root of E{$(\eta-\mu)^2$} is called standard deviation. In case of a Gaussian distribution, the standard deviation equals σ.
The probability that η is larger than η_0 can be expressed as

$$P(\eta > \eta_0) = \int_{\eta_0}^{\infty} f(\eta) \mathrm{d}\eta \qquad \text{(B.4)}$$

Exercise B.1 How can you express the probability that η is smaller than η_0?
Exercise B.2 How can you express the probability that the absolute value of η is smaller than η_0?
Exercise B.3 What is expected value of $\eta-\mu$?

Distributions can also be defined for vectors, instead of single values. When each component of a vector has a Gaussian distribution and all components are statistically independent of each other, the pdf of the vector consists of the product of the single component pdfs:

$$\begin{aligned} f(\eta_0,\ldots,\eta_{N-1}) &= \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(\eta_0-\mu_0)^2}{2\sigma_0{}^2}} \times \ldots \times \frac{1}{\sqrt{2\pi}\sigma_{N-1}} e^{-\frac{(\eta_{N-1}-\mu_{N-1})^2}{2\sigma_{n-1}{}^2}} \\ &= \frac{1}{(2\pi)^{N/2}\prod_n \sigma_n} e^{-\sum_n \frac{(\eta_n-\mu_n)^2}{2\sigma_n{}^2}} \end{aligned} \quad . \qquad \text{(B.5)}$$

Here μ_n and σ_n are the mean and standard deviations of the *n*-th components. The summation in exponent of equation (B.5) appears here because of the multiplication of the exponential functions corresponding to the distributions of the individual components. In this way one can understand how the assumption of uncorrelated Gaussian noise results in the weighted least squares cost function.
When the components of $\boldsymbol{\eta}$ have a Gaussian distribution, but are not statistically independent, the pdf is called the multivariate Gaussian distribution and its distribution can be presented as

$$f(\boldsymbol{\eta}) = \frac{e^{-\frac{(\boldsymbol{\eta}-\boldsymbol{\mu})^T C^{inv}(\boldsymbol{\eta}-\boldsymbol{\mu})}{2}}}{(2\pi)^{\mathrm{N}/2}(\det(C))^{1/2}} \quad , \qquad \text{(B.6)}$$

In this equation the matrix vector notation has been used. Here the meaning of $\boldsymbol{\mu}$ remains the mean of each of the components of $\boldsymbol{\eta}$. The meaning of the matrix C appears to be the covariance matrix of the components of $\boldsymbol{\eta}$. The covariance matrix is defined as the expected value of one component of $\boldsymbol{\eta}-\boldsymbol{\mu}$ multiplied with another component. If these components are indicated by η_{n1} and η_{n2}, we have

$$\mathrm{E}\{(\eta_{n1}-\mu_{n1})(\eta_{n2}-\mu_{n2})\} = \int_{R^N}(\eta_{n1}-\mu_{n1})(\eta_{n2}-\mu_{n2})\,f(\boldsymbol{\eta})\mathrm{d}\boldsymbol{\eta} = \ldots = C_{n1,n2} \quad . \qquad \text{(B.7)}$$

Again, details of computing the multidimensional integral is beyond the scope of this course.

Exercise B.4 What is the meaning of the diagonal elements of the covariance matrix?.
Exercise B.5 Show that with pdf given by (B.5) the covariance between different components is null. Hint: use the result of exercise (B.3).
Exercise B.6 Show that (B.5) is a special case of (B.6). Hint: use (A.11) and (A.23) and the fact that in the independent case the covariance matrix is a diagonal matrix.

The diagonal elements of the covariance matrix represent the variances of the components of $\boldsymbol{\eta}$ and are indicated by σ^2_{n}

$$\sigma_n^2 \equiv \mathrm{E}\{(\eta_n-\mu_n)^2\} = C_{n,n} \quad . \qquad \text{(B.8)}$$

The variances indicate the variation of the variable around the mean. The non-diagonal elements of C represent covariances of different components of $\boldsymbol{\eta}$. The larger the covariance, the stronger the tendency of the components to assume the same value. This tendency is even better expressed by the correlation co-efficient ρ between two components. This coefficient is obtained by normalising the covariance matrix using the standard deviations:

$$\rho_{n_1 n_2} \equiv \frac{C_{n_1 n_2}}{\sigma_{n_1}\sigma_{n_2}} \quad . \qquad \text{(B.9)}$$

This correlation coefficient varies between -1 and 1 and represents the “theoretical” correlation, i.e. it can only be determined if the true distribution of $\boldsymbol{\eta}$ is known. Other correlation coefficients appearing in this course are computed from observed data and must be considered as correlation estimates.
Equation (B.7) can also be presented in vector form using (A.8), because (A.8) is a way to express all combinations of products of vector components that are needed in (B.7). One obtains

$$\mathrm{E}\{(\boldsymbol{\eta}-\boldsymbol{\mu})(\boldsymbol{\eta}-\boldsymbol{\mu})^T\} = C \quad . \qquad \text{(B.10)}$$

What happens in the theoretical analysis parameter estimation is that the random variables representing measurement noise are in some way converted to new random variables, representing parameter estimates and test statistics. Dependent on the algorithm or equations that are used in this conversion, one can determine the distribution of these new random variables. For instance, with linear models and Gaussian noise, we have found in equation (3.1) that

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{OLS} &= (B^T B)^{inv} B^T \mathbf{r} \\ &= \boldsymbol{\theta} + (B^T B)^{inv} B^T \boldsymbol{\eta}\end{aligned} \quad . \qquad \text{(B.11)}$$

This implies that the OLS estimator $\hat{\boldsymbol{\theta}}$ is obtained by a linear transformation of a random variable with multivariate Gaussian distribution. It can be shown that therefore $\hat{\boldsymbol{\theta}}$ also has a multivariate Gaussian distribution.
In general, when $\boldsymbol{\eta}$ is distributed as (B.6) with $\boldsymbol{\mu}$=0 and when

$$\boldsymbol{\varepsilon} = A\boldsymbol{\eta} + \mathbf{b} \quad , \qquad \text{(B.12)}$$

$\boldsymbol{\varepsilon}$ will also have a distribution of the form (B.6), but with mean and covariance given by (B13) and (B.13)

$$\begin{aligned}\mathrm{E}\{\boldsymbol{\varepsilon}\} &= \mathrm{E}\{A\boldsymbol{\eta} + \mathbf{b}\} \\ &= A\mathrm{E}\{\boldsymbol{\eta}\} + \mathbf{b} \\ &= \mathbf{b}\end{aligned} \qquad \text{(B.13)}$$

$$\begin{aligned}\mathrm{E}\{(\boldsymbol{\varepsilon} - \mathbf{b})(\boldsymbol{\varepsilon} - \mathbf{b})^T\} &= \mathrm{E}\{(A\boldsymbol{\eta})(A\boldsymbol{\eta})^T\} \\ &= A\mathrm{E}\{\boldsymbol{\eta}\boldsymbol{\eta}^T\}A^T \\ &= ACA^T\end{aligned} \qquad \text{(B.14)}$$

What happened in the derivation of (B.13) and (B.14) is that the expectation operator E{} and the matrix multiplication are interchanged. That this is all right is beyond the scope of this course. When $\boldsymbol{\mu}\neq$0, similar results can be derived.

Exercise B.7 Assume the noise covariance is $\sigma^2 I_{\mathrm{N}}$ and apply (B.12) to (B.14) to find the mean and the covariance of $\hat{\boldsymbol{\theta}}$ defined by (B.11).
Exercise B.8 The same exercise as (B.7), but now for general covariance *C*.

The great applicability of (B.14) is that it can be used to obtain the covariance matrix of the estimated parameters. In particular, by square rooting the diagonal, we obtain the standard deviations of each of the estimated parameters.
The linear transformation of (B.12) is relatively simple because it transforms one multivariate Gaussian variable into another one. Other meaningful distributions can be obtained from Gaussian variables in order to quantify the distributions of *t* and *F*, which are used for statistical testing of estimated parameters.
For instance, if *N* independent Gaussian variables η_{n}, with mean 0 and standard deviation σ, are squared and added, the resulting sum Q has a χ^2-distribution with *N* degrees of freedom. In a formula, if Q is

$$Q \equiv \frac{1}{\sigma^2}\sum_{n=0}^{N-1} {\eta_n}^2 \qquad \text{(B.15)}$$

then Q is distributed according to

$$Q \sim \frac{1}{2^{N/2}\Gamma(N/2)} Q^{N/2-1} e^{-Q/2} \qquad \text{(B.16)}$$

Here $\Gamma()$ is the gamma function. In (B.16) it is used as a normalisation constant. It is a generalisation of the faculty (!) operator for non-integers.

When both η and η_n have a Gaussian distribution with zero mean and standard deviation σ, and when these variables are all independent, the variable *t*, defined by

$$t \equiv \frac{\eta}{\sqrt{\frac{1}{N}\sum_{n=0}^{N-1}\eta_n^2}} \tag{B.17}$$

has a Student-*t* distrubution, with "*N* degrees of freedom"

$$t \sim \frac{\left(1+\frac{t^2}{N}\right)^{-(N+1)/2}}{\sqrt{N}\frac{\Gamma(1/2)\Gamma(N/2)}{\Gamma((N+1)/2)}} \tag{B.18}$$

Here the denominator is merely a normalisation factor, required to make the total probability 1. The dependence on *t* is relatively simple and does not contain exponentials or special functions. Note that since both the numerator and the denominator in (B.17) have the same standard deviation σ, the definition (B.17) and the distribution (B.18) of *t* are both independent of σ. This property makes the statistic *t* very appropriate in statistical tests where the true standard deviation is not known. In the definition given in equation (3.27), which appears to be equivalent to (B.17), the same cancellation of σ-dependence occurs.
Finally, when ξ_m and η_n have a Gaussian distribution with zero mean and standard deviation σ, and when they are all independent, one can define the variable *F* as follows

$$F \equiv \frac{\frac{1}{N}\sum_{n=0}^{N-1}\eta_n^2}{\frac{1}{M}\sum_{m=0}^{M-1}\xi_m^2} \quad . \tag{B.19}$$

In words, *F* is, apart from the constant factor *M*/*N*, equal to the ratio of two independent χ^2-distributions with *N* and *M* degrees of freedom. The distribution function of *F* appears to be

$$F \sim \frac{\left(\frac{N}{M}\right)^{N/2}}{\frac{\Gamma(N/2)\Gamma(M/2)}{\Gamma((N+M)/2)}} \frac{F^{N/2-1}}{\left(1+\frac{N}{M}F\right)^{(N+M)/2}} \tag{B.20}$$

The statistics of *F* is highly related to that of the correlation co-efficient, defined as

$$\rho^2 \equiv \frac{NF}{M+NF} \tag{B.21}$$

and distributed according to

$$\rho^2 \sim \frac{\Gamma\left(\frac{N+M}{2}\right)}{\Gamma\left(\frac{N}{2}\right)\Gamma\left(\frac{M}{2}\right)}\left(\rho^2\right)^{N/2-1}\left(1-\rho^2\right)^{M/2-1} \tag{B.22}$$

Exercise B.9 If the variable F has an F-distribution, what can you say about the distribution of $1/F$?

7 Appendix C. Matlab Exercises

Question 1
The goal of this exercise is to study different variants of a line fit procedure.
The theoretical relationship between x and y is:
$y_n = ax_n + b$.
A. The x- and y-points are present in the files `x.dat` and `y.dat`. Import these data in MatLab using

```
x   = csvread('x.dat')';
y   = csvread('y.dat')';
```

Here the transpose sign "`'`" make x and y, which are stored on a single line, into column vectors. Create a constant time function using

```
const  = ones(size(y));
```

Combine these time functions in a model matrix B, using

```
B      = [const x];
```

and compute the OLS estimator of slope and offset by implementing the formula

$$\boldsymbol{\theta} = \left(B^T B\right)^{inv} B^T \mathbf{y}.$$

Note that in MatLab the transposed of a vector or matrix is obtained using the `'`-sign.

B. Compute the OLS-estimator using the MatLab function `fminsearch`(). This function works as follows.

```
thetabest = fminsearch(@(par) Cost(par, x, y, costfu), [0.1;0.1]);
```

`fminsearch`() uses a sophisticated algorithm to search in a parameter space (the size(`par`)-dimensional space) for the minimum of the user defined cost function, `Cost(par, x, y, costfu)`. `fminsearch`() returns the best combination of parameters and assigns them here to the output variable named `thetabest`. In order to compute `Cost()` at different parameter combinations, `Cost()` needs access to the data vector, model vector and some other parameters. These additional parameters are also arguments of `Cost()`, in our case the x-coordinates are set in the vector x, y-coordinates in vector `y` and `costfu` as a parameter specifying the cost function type to be used.
Because `fminsearch`() needs to know which of these arguments is the parameter over which the minimum need to be sought, the first argument of `fminsearch`() is `@(par)` where `par` contains the name of the parameter vector. The last argument of `fminsearch`(), here the vector `[0.1;0.1]` are the starting values used by `fminsearch()`.

Complete the following implementation of `Cost` () such that `fminsearch()` finds the OLS estimator of the offset and slope of the line passing through the *x*- and *y*-coordinates. Verify the results of 1A.

```
function [Cost] = Cost(theta, x, y, costfu)

N      = size(x);
const  = ones(size(x));

if(costfu=='L2')
    model  =
    Cost   =
end
end
```

C. Adapt the functions of exercise **1B** such that the L1-norm is minimized instead of L2. Keep the L2-option working for later use.

D. Extend the functions of exercise **1C** such that horizontal distances are minimized instead of vertical ones. For that purpose, use the inverted relationship between *x* and *y* and add another parameter to `Cost`().

$x_n = (y_n - b)/a$.

E. The file `yout.dat` contains a single outlier. Verify that by making a plot. Use the functions derived under **1B-D** to complete the following table:

		Offset	Slope
Vertical differences	L2		
	L2, outlier		
	L1		
	L1, outlier		
Horizontal differences	L2		
	L2, outlier		
	L1		
	L1, outlier		

Which rules of thumb you can derive from these findings, regarding robustness?

F. Make the outlier in `yout.dat` much bigger, recomputed the L1 estimate (vertical differences) and compare the estimated parameters to the original `yout.dat`. Explain the results. Do you get similar results when horizontal differences are minimized?

Question 2

The goal of this exercise is to use projectors for artifact removal from EEG-signals. In the given example the EEG is disturbed by a sudden eye movement, which causes deflections of several hundreds of ms. A simple idea is to record in addition to the EEG a bi-polar EOG (Electro Oculargram), which is particularly sensitive to eye motions.

A. Read in EEG and ECOG signals from file using

```
EEG  = csvread('EEG.dat' )';
EOG  = csvread('EOG.dat' )';
```

and plot them using `plot(EEG(1, :), 'r')`, etc.

B. Build a MatLab function that returns a projector that removes given time functions. Start from the following function body

```
function [Projector] = MakeProjector(TimeFuncs)

    Projector =
end
```

Test your function using a well chosen input vector.
C. The simplest idea is to create a projector from the EOG and apply it on the EEG. Does that work?

D. The EOG contains a large drift. Try to clean the EOG with a projector based on the vector `1:501` Alternatively, use the vector `-250:250` as input vector of `MakeProjector()`. Is there a difference? Explain.

E. Repeat question **C** using the EOG modified in **D**. Do the results improve?

F. Create a new projector using the EOG, an offset and a trend. Apply this new projector to the EEG and give your comment of the results.

Question 3
The goal of this question is to estimate one or more fMRI activation parameters when the signal is embedded in correlated or uncorrelated noise. This fMRI signal will be constructed using a simulation of Gaussian noise. Three MatLab functions are provided:

`MakeCovar(N, K)`	Creates an *N*x*N* covariance matrix, of which the amount of correlation between subsequent samples is regulated by *K*. Samples separated more than *K* samples are uncorrelated. For *K* = 0 the identity matrix is constructed.
`MakeNoise(N, sigma, Cov)`	Creates a column vector of height *N* consisting of Gaussian noise with covariance `Cov`. The noise level is scaled by the parameter sigma. Each time this function is called, different noise realizations are generated.
`MakeProjector(Regs, Remove)`	Creates a projector matrix, using the columns of `Regs`. Applying this projector onto a given vector ***v*** will either remove the best possible linear combinations of the columns of `Regs` from ***v***, or it will project the vector ***v*** onto the space spanned by these columns. This behavior depends on `Remove` parameter.

A. Choose *N*=200 and create covariance matrices for different values of K. These matrices can be viewed using `imshow()`. The model used in `MakeCovar()` always produces constant subdiagonals. What is the physical meaning thereof?

B. Create two noise signals with different levels of correlation and plot them in the time domain using

```
figure, plot(Noise1,'r.-')
```

Is the appearance of these plots consistent with the choice of the covariance matrices?

C. Under some assumptions (not treated in the lectures) the covariance of the noise can be estimated from noisy data using the following formula

$$\hat{C} = \frac{1}{M}\sum_m \boldsymbol{\eta}_m \boldsymbol{\eta}_m^T$$

where M is the number of (independent) realizations of the noise vector $\boldsymbol{\eta}_m$. A simplified interpretation of this formula is that the covariance is the expected value of a product of two quantities and therefore it can be estimated by computing the average of these quantities realized in real data.
Create a covariance matrix C, implement a loop over M noise realizations and compute its estimate using the above formula. Determine how many noise realizations are needed for an "adequate" estimate of the true covariance.

D. We simulate an fMRI signal from two shifted block time functions with different amplitudes. The block functions can be created with the following code fragment

```
Block1 = zeros(N,1);
Block2 = zeros(N,1);
for k=1:N
    n1 = mod( k   , 30);
    n2 = mod((k-1), 30);
    Block1(k) = (n1<20);
    Block2(k) = (n2<20);
end
```

Create these block functions and verify the result. Compute the correlation coefficient.

E. Complete the following MatLab function such that it simulates noisy fMRI data and returns the OLS-estimated parameters of interest.

```
function [theta] = EstimateFMRIpars(N, Sigm, Cov, Block1, Block2)

    Noise = MakeNoise(N, Sigm, Cov);  % simulate noise
    fMRI  = ones(N,1) + ((1:N)')*3/N + 2*Block1 + Block2 + Noise; % simulate
                                                                  % fMRI data

    B  = [Block1 Block2];             % Regressors of interest
    S  = [ones(N,1)  (1:N)'];         % Nuisance regressors
    PS = MakeProjector(S, true);      % Projector

    fMRIClean =
    BClean    =
    theta     =
end
```

What are the true parameters in the above example? Call this function in a loop over different noise realizations and make a scatter plot of the estimated parameters. From visual inspection of

this scatter plot, are the estimated parameters correlated? Is the estimator biased? Is the estimator biased when the noise is correlated and the parameters are estimated with OLS? Explain.

F. Extend the above function by adding the option for a GLM estimate of the parameters as opposed to the OLS estimator. If `CompGLS` is `false` the function should return the OLS estimate and when it is `true`, it should return the GLS estimate.

```
function [theta] = EstimateFMRIpars(N, Sigm, Cov, Block1, Block2, CompGLS)
```

Hint: You can use the Cholesky decomposition as pre-whitening matrix. This decomposition can be computed as

```
    Prew   = chol(inv(Cov));
```

Multiplication of this matrix with a noise vector generated with covariance matrix `Cov` will turn into a noise vector where samples are uncorrelated. Verify this with an example.

G. Use the function developed in **E.** to compare the behavior of GLS and OLS estimators on the basis of visual inspection of scatter plots.

H. Extend the function under **E.** by also returning an OLS or GLS estimate of the parameter noise covariance.

```
function [theta CovThe] = EstimateFMRIpars(N, Sigm, Cov, Block1, Block2,
CompGLS)
```

The ML estimator for the parameter covariance matrix is

$$\text{EstCov}\left(\hat{\boldsymbol{\theta}}\right) = \hat{\sigma}^2 \left(B'^T B'\right)^{inv} = \frac{1}{N}\sum_{\text{n}} \left(r_{\text{n}} - r_{\text{n}}(\hat{\boldsymbol{\theta}})\right)^2 \left(B'^T B'\right)^{inv}$$

Explore whether, on average, these estimates are adequate. Is it adequate when OLS estimators are used although the noise is correlated?

8 Appendix D. Formulas Chart

Basic matrix vector algebra

$$(AB)^T = B^T A^T$$

$$(AB)^{inv} = B^{inv} A^{inv}$$

$$\sum_{m=0}^{M-1} \frac{v_m w_m}{\sigma_m^2} = \mathbf{v}^T C^{inv} \mathbf{w} \qquad \text{where} \qquad C = \begin{pmatrix} \sigma_0^2 & 0 & 0 & 0 \\ 0 & \sigma_1^2 & 0 & 0 \\ 0 & 0 & ... & ... \\ 0 & 0 & ... & \sigma_{M-1}^2 \end{pmatrix}$$

Multivariate Gaussian distribution

$$f(\boldsymbol{\eta}) = \frac{e^{-\frac{(\boldsymbol{\eta}-\boldsymbol{\mu})^T \mathbf{C}^{-1}(\boldsymbol{\eta}-\boldsymbol{\mu}))}{2}}}{(2\pi)^{N/2}(\det(\mathbf{C}))^{1/2}}$$

Double exponential distribution

$$f(\eta) \propto e^{-\lambda|\eta-\mu|}$$

OLS estimator /GLS estimator

$$\hat{\boldsymbol{\theta}}_{OLS} = (B^T B)^{inv} B^T \mathbf{r} \qquad \hat{\boldsymbol{\theta}}_{GLS} = (B^T C^{inv} B)^{inv} B^T C^{inv} \mathbf{r}$$

$$\hat{\sigma}^2{}_{OLS} = \frac{1}{N} \sum_{\mathrm{n}} (r_n - \tilde{r}_n(\hat{\boldsymbol{\theta}}))^2$$

Parameter covariance estimate

$$\mathrm{EstCov}(\hat{\boldsymbol{\theta}}_{OLS}) = \hat{\sigma}^2 (B^T B)^{inv}$$

Projector

$$\mathbf{P_S} = \mathbf{I}_N - \mathbf{S}\left(\mathbf{S}^T \mathbf{S}\right)^{inv} \mathbf{S}^T$$

t-test

$$t_{N-L-1} \equiv \frac{\hat{\theta}}{\hat{\sigma}_\theta} = \sqrt{N-L-1} \frac{\rho}{\sqrt{1-\rho^2}}$$

$$p \approx \frac{1}{\sqrt{2\pi}} \int_t^\infty \mathrm{e}^{-s^2/2} \mathrm{ds} = \mathrm{erf}(t) \;\; (\text{large } N)$$

F-test

$$F = \frac{N-L-M}{M} \frac{\mathbf{r'}^T (I - P_{B'}) \mathbf{r'}}{\mathbf{r'}^T P_{B'} \mathbf{r'}} = \frac{N-L-M}{M} \frac{\rho^2}{1-\rho^2}$$

(*L* nuisance regressors, *M* regressors of interest)

Correlation coefficient

$$\rho = \frac{\mathbf{b}^T P_S \mathbf{r}}{\sqrt{(\mathbf{r}^T P_S \mathbf{r})(\mathbf{b}^T P_S \mathbf{b})}}$$

Convolution with nuisance regressors

$$r_n = \sum_{m=-M_{prior}}^{-M_{prior}+M-1} h_m a_{n-m} + \sum_l S_{nl} \varphi_l + \eta_n$$

Toeplitz matrix

$$C_{Toeplitz} = \begin{pmatrix} c_0 & c_1 & c_2 & & c_{N-1} \\ c_1 & c_0 & c_1 & c_2 & \\ & c_1 & c_0 & & \\ & & & & c_1 \\ c_{N-1} & & & c_1 & c_0 \end{pmatrix}$$