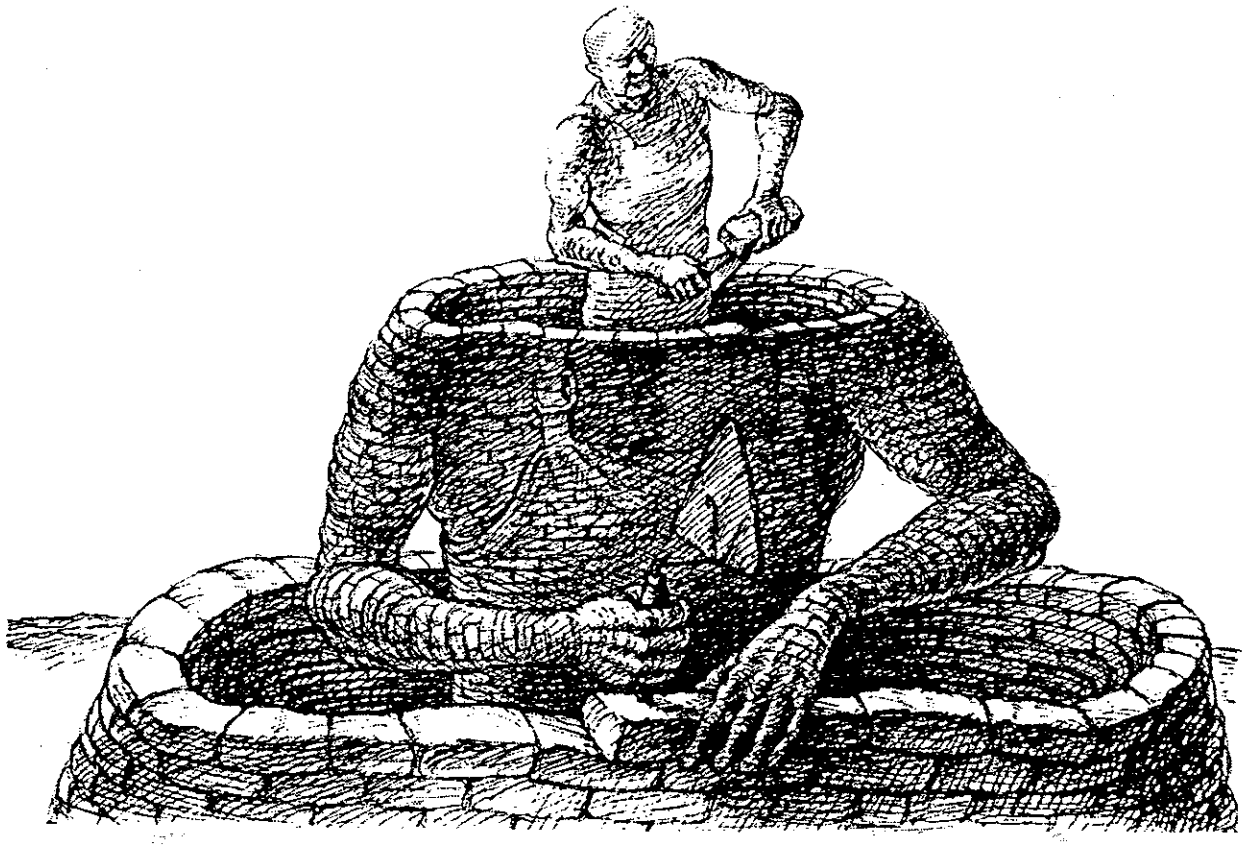


Science models are rather a simulation of human consciousness than the reality of the universe.

Chapter 5

Identification of Linear Relations in Noisy Data

Models are a matter of inspiration,
not of deduction.
Jan Willems.

Identification of linear relations from noisy data is one of the enduringly central problems in modelling systems. In this chapter, the close relationship between linearity, and orthogonality, linear models and noise will be highlighted. It contains material which is important both from a *conceptual* as from a *computational* point of view for the rest of this thesis. Besides a detailed analysis of some basic identification schemes that will be used in chapter 8 on identification of linear dynamic systems, this chapter also provides the introduction to chapter 6. There we shall introduce *the uncertainty principle of mathematical modelling*, which allows to characterize geometrically all linear static models that are linearly compatible with certain given data. In this chapter, the emphasis is laid on the analysis of identification schemes for *static* linear systems (i.e. non-dynamic, without memory). It will become clear in chapter 8 that the results and obtained insights carry over in a straightforward way to the identification of *dynamic* linear systems. As a matter of fact, to each identification scheme discussed in this chapter, we shall couple an identification scheme for a dynamic linear system in chapter 8.

This chapter is organized as follows. In section 5.1. attention is paid to the deep connection between orthogonality, linear models and noise. Two opposite modelling philosophies are discussed: The deduction and the inspiration framework. In section 5.2., we prove both mathematically and by experimental verification, that in a lot of applications the pure noise variables satisfy certain orthogonality properties with respect to the exact variables. Together with additional requirements concerning a sufficiently high signal-to-noise ratio, this allows to obtain consistent models if the noise is additive with respect to the exact data. This leads in section 5.3. to the analysis of several identification schemes including the linear least squares approach, and the total linear least squares algorithm. These techniques are shown to be extremes of a unifying identification scheme which essentially consists of rank one modification of the available data. The resulting models have the embarassing property that they model the noise in a very structured way, namely as a rank one matrix. Hence, full rank noise models are surveyed as well. Finally, in section 5.4. we pay attention to the computation of the intersection of column spaces of two matrices. The noise sensitivity of

these algorithms is analysed in detail. It is shown how the previously discussed identification algorithms do *not* allow to compute the intersection *consistently* but on the other hand, permit the computation of a certain *dual* model for the intersection.

5.1 Linear Relations, Orthogonality and Noise

5.1.1 Orthogonality and Linear Relations

Linearity is one of the most important organizational principles for which mathematics has provided a thorough understanding. The use of linear models and theories may be motivated by their *simplicity*: There is an extensive relevant theory available, with no corresponding wealth of transparant results in non-linear system theory. Linear problems are more easily solved than their non-linear counterparts, both from the conceptual as the computational point of view. As a matter of fact, *linearization* is a frequently applied tool for the analysis and modelling of non-linear systems. Think of Newton's famous law $F = ma$ which serves as a good approximation for the highly non-linear phenomenon of gravitation. In a lot of applications, non-linear problems are replaced or approximated by linear time-variant ones, such as in adaptive identification and control. A good example is Lyapunov's first method to investigate the stability of non-linear autonomous systems [17]. Linear theory may merely be used because of the lack of any better approach. This is a pragmatic viewpoint: an incomplete solution is better than none at all. In a stochastic framework, it is well known [10] [20] that *linear* estimation and prediction is the *optimal* modelling method if the random processes involved are gaussian or if the optimal estimate is restricted to be a linear functional of the observed random variables and the loss function to be optimized is quadratic. Hence, results obtainable by linear estimation can be bettered by nonlinear estimation only when the random processes are nongaussian and even then, only by considering at least third order statistics. Another important fact is the 'eigen'-property of gaussian densities with respect to linear operations: Linear functions (and therefore conditional expectations) on a gaussian random process are gaussian random variables. Gaussian random variables remain gaussian after passing them through a linear system. As a last example where linearity plays a most prominent role, consider the characterization of *predictability* of stochastic processes via the *Wold decomposition* [9]. Deterministic processes are defined here explicitly in terms of *linear predictability*. As a matter of fact, also the notion *orthogonality* plays a crucial in this definition, hence already revealing the close connection between linearity and orthogonality.

It would be a most exciting job to track historically the deep interconnection between linearity and the notion of orthogonality. Despite the challenge, we shall not include the resulting essay in this thesis, but only provide a short-cut by mentioning the required mathematical ingredients without working out the details.

What is needed for our purpose is the notion of a *pre-Hilbert space* [18, p.46].

Definition 1 Pre-Hilbert space
A pre-Hilbert space consists of :

1. *A vector space with two operations called addition and scalar multiplication.*
2. *An inner product, satisfying the conditions of symmetry and bilinearity and in most cases, nonnegativity.*

For a detailed exposition, the reader is referred to [18].

Depending on the application at the hand, the elements of the vector space are real numbers, n-tuples of real or complex numbers, functions, random variables, etc.... While the vector space axioms only describe algebraic properties of objects of that space such as addition, scalar multiplication and combinations of these, norms are the tools to characterize topological concepts such as openness, closure, convergence, completeness, etc .... Once an inner product rule has been fixed, the notions of norms, distances and of angles are automatically induced, including orthogonality:

Definition 2 *Two vectors are said to be orthogonal if and only if their inner product equals zero.*

A most salient feature is the fact that once orthogonality has been defined, it can be used on its turn to define precisely the notion of linear relation, or more specifically:

Orthogonality is absence of linear combinations!

Despite its generality, this statement is very precise and when applied rigorously, leads to the uncertainty principle of mathematical modelling to be presented in chapter 6. Its states that, when 2 vectors are not orthogonal, they are linearly related, in this sense that at least part of one of them can be explained, by a linear relation, by the other one and vice versa. Since the mathematical description of orthogonality requires the use of an inner product, the precise contents of the notion of linear relations via linear combinations, will also depend upon the specific choice of an inner product. Contrary to what is thought, the choice of an inner product is not trivial at all. Not that we shall provide precise arguments to motivate our use of the *Euclidean* inner product throughout this work, but the question is worth investigating. Many mathematicians and engineers believe that the choice of an inner product is arbitrary. However generally there are well defined motivations to prefer a certain inner product with respect to another:

- In our framework of oriented energy (chapter 4), the inner product between vectors is weighted by the inverse of the Grammian of the undesired vectorsequence, which acts as a pre-whitening operation, in a sense that it amplifies the desired signal in those directions where the undesired one is weak, while it weakens the desired signal with respect to the undesired one in those directions where the undesired vector sequence is strong. Note that this undesired vectorsequence is not necessarily 'noise', but may consists of observed signals, the influence of which is to be eliminated from other observations. A similar choice of inner product occurs in Gauss-Markov and mimimum variance estimation, where the covariance matrix of the noise serves as a weight matrix in the computation of the inner product [18, p.84]. When implemented recursively, these estimators together with a linear dynamical system, lead to the Kalman filter.

- In the theory of stochastic processes, the inner product is an integral of the product of the two random variables, weighted by their joint probability density function, that expresses the importance of certain values according to their relative occurence. The idea of regarding random variables as elements of a metric space, with the distance between two elements being the variance of their difference, is due to Wold and Kolomogorov [9].

- Certain polynomial functions satisfy a desirable orthogonality property if their appropriately defined inner product is weighted by a certain weighting function. Examples are Chebyshev, Laguerre, Hermite, etc...polynomials, which provide orthonormal bases for certain functional spaces. As a matter of fact, in a lot of modelling environment, it may be preferable to express functions in a coordinate system generated by certain polynomials. If moreover these polynomials are orthonormal, certain mathematical expressions, containing cross-products and norms, are most easily computed in this new basis.

- The motivation may be physical or based upon observed or assumed physical invariance principles, such as for instance the finiteness of the speed of light in the special relativity theory, which leads to the indefinite Minkowskian inner product. In the general theory of relativity, the inner product is Riemannian and is dependent upon the mass distribution and the curving of the space-time. Since the inner product is essentially needed to compute distances, angles, etc..., an appropriate knowledge of this inner product is absolutely necessary for reliable computations. Hence the intensive cosmological research that aims at finding the correct mass distribution of the universe.

The following observations concerning the notion of orthogonality are worth mentioning:

- Depending on the application, orthogonality may have a real physical meaning: Orthogonality in the theory of relativity, with its indefinite Minkowskian inner product, is closely linked to the interpretation of *simultaneity*, in the theory of electric and gravity fields to *equipotentiality*, in a modelling environment to *linear relations* and in a numerical environment to *optimal numerical stability.*

- In a probabilistic framework, orthogonality reduces to uncorrelatedness, which is equivalent with the absence of linear relations among stochastic processes. The connection between linearity and orthogonality is the basic idea behind the Wold decomposition [9], leading to Kolmogorov's work on linear predicition, which on its turn, later on lead to ARMAX modelling strategies and the Kalman filter.

- Orthogonality arises in a natural way when optimizing quadratic criteria (least squares). One of the fundamental insights is that the optimal estimate of a variable as a function of a known data is nothing else but the *orthogonal projection* of the variable on the space generated by the known data. Orthogonality and corresponding projections theorems play a most prominent role in the analysis and solution of optimization problems such as minimum norm optimization, minimum variance and Gauss-Markov estimates [18]. In general, orthogonality is the key issue in linear least squares estimation [9], although even in the sixties, strenuous efforts were made in many papers on linear least squares estimation to avoid the connection. As a matter of fact, it is an orthogonality principle that is the fundamental basis of Kalman's work on recursive linear least squares estimation [10].

- The Fourier Transform and its discrete versions both exploit the orthogonality of the goniometric functions. As a matter of fact, they correspond to an orthonormal change of basis, hence conserving energy contents as expressed by Parceval's theorem. Some operations are most easily studied in the frequency domain, such as band-pass filtering and sampling. Others lend themselves more to an analysis in the time domain.

- Soon after the introduction of the digital computing techniques, it was found that orthonormal numerical operations minimize the propagation of undesired effects, originating in the finite precision arithmetic [25]. This observation lead to powerful numerically robust algorithms [7], that are essentially based upon orthonormal matrix operation (Givens, Householder, QR,...).

- Without doubt, the reader will know of other applications and observations on orthogonality from his own experience and background.

5.1.2 What is noise ?

The main theme of this section is the demonstration of the fact that *noise* is really *defined* by the model that is applied. This approach emphasises that the notion of noise will always contain a certain degree of arbitraryness. Noise is simply a garbage trash, in which we dump all our ignorance concerning the precise contents of the behavior and structure of a system. The more we know about a system, the less will be left as unexplained. Hence, a rather general yet precise definition of noise is:

Noise is what is not explained by the model.

In other words, in any application, noise is implicitly modelled too! Let's now analyse the consequences of these insights with respect to linear models and orthogonality. Linearity is the basic feature of the models we shall develop in this work, By the paradigm that noise is what is not explained by the model, it follows immediately that *noise is absence of linear relations.* Having established the equivalence between absence of linear relations and orthogonality, it is intuitively clear that the only precise characterization of noise should be in terms of orthogonality. As described up to now, noise is almost only a matter of definition! However, it will be shown furtheron that the relation between noise and orthogonality is not purely a matter of definition but also of observation and experimental verification. In any case, there exists a close connection between what is viewed as a model and what is considered as noise.

In order to make our statements more precise, we shall discuss three important features of mathematical models:

1. First we shall discuss two (extreme) standpoints with respect to the theory of mathematical modelling: the *deductive* approach and the *inspiration* approach. Later on, it is up to the reader to decide whether he will share one of the extreme points of view or make for himself a convex combination of the two extremes.

2. Second, it will be discussed how the identification of a model (and hence the corresponding noise model) is always determined by the purpose itself of the model. A distinction will be made between *predictive* and *descriptive* modelling approaches.

3. Third, we shall analyse the desirable property of an identification scheme to be *consistent.* While this notion is well known in statistics (the deductive approach), we shall concentrate also on the meaning of consistency for the inspirationist.

Deduction versus inspiration.

The deductionist believes (he may have strong reasons to do so) that there exist 'true' linear laws, governing the behavior of the system under study. It is his task to discover these laws from measurements, that are however, in general, corrupted by perturbations, which, among other possible causes, originate in the *finiteness* of observation tools, both in precision and duration. Hence, the deductionist believes in the existence of two fundamental extremes: 'true' linear systems and nasty disturbances (or if you want, good and evil!). The manifest lack of *complete* information about the phenomenon under study, necessitates the use of statistics and requires a deep analysis of how Nature's randomness may corrupt the observation of the linear phenomenon. Kolmogorov's probabilistic framework is a mathematical attempt to quantify Nature's intrinsic fuzzyness via the postulate of the existence of stochastic processes and their characterization in terms of probability density functions. The deductionist's reasoning essentially consists of a (difficult) *perturbation analysis*: Starting from the 'true' linear system and postulating some characteristic quantities on the disturbances, he calculates how these disturbances affect the linear phenomenon and then compares his results with the observations. All this requires quite some (conceptual) machinery, such as ergodicity and finite sample statistics. However, if at the end, everything fits nicely into some a priori fixed confidence bounds, employing hypothesis testing tools, degrees of freedom concepts etc..., the deductionist lends back in his seat with satisfaction: He has found an explanation for the phenomenon. According to the principle that *noise is what is not explained by the model*, he will find a probability density function that characterizes the uncertainty of his obtained linear model, which in some sense, is complementary to the uncertainty characterized by the probability density function of the noise. Observe that in a lot of cases, the deductionist will need quite some a priori knowledge or assumptions on for instance, the noise characteristics, which are *not always* (euphemism for *almost never*) a priori verifiable. These are Kalman's famous *prejudices* [10].
The inspirationist does not believe that Nature could be so simplistic that it would ever function according to linear laws. He is convinced of the fact that non-linearity is a general, generic and fundamental property which is deeply rooted in Nature. However, he is quite aware of the unlimited richness both of the collection of non-linear models and of their properties. So he will quite arbitrarily prefer a linear model, above all non-linear ones, being attracted by the simplicity of linear models, by the availability of algorithms and a wealth of theoretical results or maybe merely by referring to the esthetic appeal of linearity [26]. Then he will ask himself how the data and observations are *falsified* by his choice. In order to make this question well defined, it is necessary to introduce a certain criterion, which measures the *goodness-of-fit* of the data by the model. Of course, the specified criterion possibly influences the number of models that are compatible with all the requirements. The obtained model may be *unique* or not. However, in all cases, the residuals (Latin for *what is left behind*) are then declared to be noise (*what is not explained by the model*). In the inspirationist's eyes, mathematical modelling is merely a matter of falsification, instead of deduction. The fundamental issue is that of *approximation* instead of perturbation.

Predictive versus descriptive models.

The ultimate purpose of the model will determine the identification technique that is employed, the model that will be obtained and consequently, what is considered to be noise.

From a *descriptive* point of view, one is mainly interested in explaining a certain behavior by a prespecified class of models. The model specification may have been based upon physical motivation or mathematical convenience. Since it is possible that the data are generated by a system that does not belong to the model class (in most practical situations this will be the case!), the modeller will have to choose a certain criterion, which will allow him to guarantee that the obtained model is the best possible among all approximations of the data in the considered model class. As an example, the total linear least squares technique, to be discussed in section 5.3, is *descriptive* when all data may be changed in order to obtain a linear static model. The linear least squares method only changes part of the data. It is in general not suited for a descriptive modelling approach, but, as a matter of fact, it is one of the preferred methods for a *predictive* modelling environment. Here one considers the already obtained data as containing information concerning the 'exact' system. It may be possible that new data carry additional information about the same 'exact' system. Exploiting this new information, the model may be updated by comparing the predicition obtained by the 'old' model with the most recent new information. For linear models, everything can be translated nicely into a mathematical framework, where orthogonality plays a most prominent role. This predictive approach is the basis of the Kalman filter, of Kailath's innovations approach [9] and Ljung's identification techniques [16]. Finally, note that a descriptive approach might be interesting, when one is essentially interested in a *simulation* of a certain data set. If however one is interested in a correct *prediction*, a predictive approach is more appropriate.

Consistency of identification schemes.

In the statistical analysis of estimators, *consistency* is, besides unbiasedness and efficiency, a desirable property for an estimator. We shall use consistency in the same sense as in statistics for the deduction point of view. However, for the inspirationist, consistency will have another meaning.

In analysing identification schemes, the deductionist in fact considers the reverse problem: How are data perturbed by noise and how can the modeller hope to recover, if necessary asymptotically, the exact system. From his analysis, the deductionist will find and investigate certain properties (such as orthogonality of noise and exact data, section 5.3), and he will declare an estimator to be *consistent*, if it is able to recover the original exact system (asymptotically) if some a priori conditions are fulfilled. Hence, the study of consistency is based upon the interaction between a priori imposed conditions, which may (tend to) be true (asymptotically), and the mechanism of an identification scheme. As an example, it is well known that linear least squares estimation is a consistent method, if only one of the variables is noisy and if that noise is not correlated with one of the exact variables [19, p.96]. Proving consistency of a method is not always an easy task. It requires concepts such as *convergence in probability*, *almost sure* and *mean-square convergence* (see [19, p.97]). For instance, probability limits refer to one particular way in which estimates may settle down as the number of observations they are based on is increased. Almost sure and mean convergence are stronger properties, that imply convergence in probability. For instance, mean square convergence is defined as convergence of the mean-square deviation of the estimate from the 'true' value as the number of observations tends to infinity. The demonstration of the lack of consistency is easier: simply consider an example in which all conditions are fulfilled and show that the estimator does not yield the correct answer.

The inspirationist does not bother about the existence of a true system. In essence, he will propose a certain criterion and a certain class of models. He will then choose that model of the model class that optimizes the criterion. Hence, for the inspirationist, modelling is a matter of optimal approximation. However, we shall give some examples to show that even the inspirationist is not completely free in the specification of the criteria he proposes. As a matter of fact, he is obliged to verify that his optimization criterion is *consistent*: This means that the optimization criterion should be such that the exact system is recovered if the data were really generated by a system that belongs precisely to the model class he has specified. This is a quite natural assumption: Any identification method (or in general any theory) should be such that, if the data are exact, the exact model is found. In other words, the approximation should be continuous with respect to the noise level as this goes to zero. Observe that noise in this context simply means everything what can not be explained by a model in the model class.

5.2 Additive noise models and identification

In this section, we shall consider additive noise models from a deductive point of view, i.e. we shall concentrate on a perturbation analysis of exact data that are linearly related. Although the obtained insights may be also considered from the inspirationist point of view, we shall mainly interpret them from the deductionist framework, leaving a pure inspirationist approach for chapter 6.

This section is organised as follows. In subsection 5.2.1, we describe the general problem formulation and the assumptions on the *genesis* of the data. In section 5.2.2, we derive an *orthogonality theorem*, which clearly demonstrates why some frequently used identification theorems can be succesfully applied. These identification schemes will be more closely investigated in section 5.3. In section 5.2.3, we derive the so-called *lever theorem*. It also provides insight in several identification schemes it will be used in the analysis of several approaches to compute the intersection of subspaces.

5.2.1 The problem formulation

In a lot of *linear* applications, the solution of the problem reduces to answering the following question:

Given an $m \times n$ matrix A, with $m \gg n$. Find a suitable rank $r < n$ matrix $\hat{A}$ and a pure noise matrix $\tilde{A}$ such that $A = \hat{A} + \tilde{A}$.

Of course, this problem is only meaningful when it is defined what is to be understood by 'suitable rank' and what are the conditions to split A into $\hat{A}$ and $\tilde{A}$. In order to get more insight into the question, we shall now adopt a deductionist point of view and start from the reverse problem:

Given an $m \times n$ exact matrix $\hat{A}$ with $r < n < m$ where r is the rank of $\hat{A}$, and a noise generating mechanism that generates the noise matrix $\tilde{A}$. How and under what conditions is it possible to recover the matrix $\hat{A}$ and/or certain of its properties from the matrix A?

Some remarks and comments are in order:

- The fact that the exact matrix $\hat{A}$ is rank deficient, is at the heart of all linear modelling strategies, both static and dynamic. *The rank deficiency guarantees linear dependence among the columns (and the rows) of $\hat{A}$ and vice versa.* Moreover, the number of linearly independent linear relations between the columns of $\hat{A}$ is equal to the so-called corank of $\hat{A}$ which is simply $n - r$. As a matter of fact, maximization of the corank is one of the crucial issues of our inspiration approach of chapter 6. However, in this section we shall think as a deductionist. Therefore, one of the first problems to be solved is: How does the noise term $\tilde{A}$ affect the rank r of the exact matrix $\hat{A}$.

- The rank deficiency at the same time implies the existence of a row and column space of the matrix $\hat{A}$, which are r-dimensional. The modeller is mainly interested in these subspaces and their structure. Examples are the solution of sets of linear equations (kernel of the matrix $\hat{A}$, this chapter), factor analysis and oriented energy (row space of the matrix $\hat{A}$, chapter 4), the state vector sequence of a dynamical system (subspace of the column space, chapter 8), the shift structure of the subspace (realization and identification theory, chapter 7), etc ....

- It is postulated here that the observed data were generated by *addition* of noise and exact data. We have to confess that, in the identification literature, this is mainly done for mathematical convenience. Only in *additive* noise models. it is possible to exploit the connection between linearity and orthogonality. However, multiplicative models are in some applications to be preferred or more appropriate. Recently, some nice philosophical-mathematical framework for multiplicative noise models has been put forward in [11].

- Some *regularity* conditions on the noise will be needed, mainly expressing the fact that the average elementwise energy of the noise is independent of the number m of rows of $\tilde{A}$ and independent of the elementwise energy of the elements of $\hat{A}$. This implies that data dependent noise models do not fit in this framework, as is the case if for instance the operation point of a measurement equipment is adapted depending on the scale of the data. In brief, the noise is assumed to be absolute and not relative with respect to the exact data.

- Observe that we assume that $m \gg n$. *Mutatis mutandis*, the results also apply of course for the cases where $m \ll n$. In any case, the fact itself of overdetermination, i.e. the fact that either m/n be very small or very large, is the crucial assumption. In order to threat both cases at the same time, we shall use the term *short space* for the space of the matrix corresponding to the smallest dimension $min(m, n)$ and the term *long space* for the other one. Hence, if $m \gg n$, the long space is the column space and the short space is the row space.

- It is natural to assume that $\tilde{A}$ is of full column rank n (in general, the short space of the noise matrix is the ambient space). This can be most easily seen from our general principle which states that *noise is absence of linear relations.* Since for linear models only rank deficiency guarantees the existence of linear relations, the noise matrix must be of full column rank. Yet, it will be shown how certain frequently used identification schemes violate this principle.

- We did not yet say anything about the conditions on the relative position of the long spaces of $\hat{A}$ and $\tilde{A}$. This will be the subject of the orthogonality theorem derived in section 5.2.2. The lever theorem of section 5.2.3, will provide additional information about the short spaces of $\hat{A}$ and $\tilde{A}$.

5.2.2 Orthogonality of an exact and a random vector.

The definition of noise is not completely arbitrary or dependent upon the definition of the model. In a lot of cases, observations and experiments force us to assume and accept that certain properties of (observation) noise with respect to exact data are intrinsic properties of Nature. In this section, we shall look at the modelling problem from the deductionist point of view. We shall derive under which conditions, noise sequences tend to be orthogonal to the exact data. As a consequence, the problem of identification reduces to remove from the observed data, models of that noise such that what remains (the model of the exact data) is orthogonal to what was removed (the noise). The statistically motivated reader will surely recognize throughout the discussion, the notion of *uncorrelatedness* or *probabilistic orthogonality*. However, one can also understand the derivation from a geometrical point of view. The results will be illustrated throughout with some convincing simulations.

Some intuition.

Assume that the rank r of the exact $m \times n$ matrix $\hat{A}$, with $m \gg n > r$, is independent of the number of observations of the exact data (which happens to be the case in all applications). As m increases, the dimension of the orthogonal complement of the r-dimensional column space of the matrix $\hat{A}$ increases linearly with m. Because it is assumed that there are no linear relations among the pure noise variables, the dimension of the column space of $\tilde{A}$ equals n and is independent of m. Now intuitively, it is obvious that for increasing overdetermination m, the probability that the column space of the noise matrix $\tilde{A}$ will be orthogonal to the column space of the matrix $\hat{A}$, will increase if the noise is independent from the exact data. This intuition is to be completed by a rigorous statement of the minimal number of conditions that may allow to confirm it mathematically. An example of such conditions is analysed in the next section.

A detailed analysis

First, we shall reduce the preceding general problem formulation to a more restricted case. We shall investigate the probability density function of the angle between one single random vector in m dimensions and a fixed given subspace of dimension r in the m-dimensional ambient space with $r < m$. Having found this density, one may invoke independency of the noise variables in order to establish the generalization to more than one random vector, i.e. the columns of the noise matrix $\tilde{A}$.

In principle, the solution is straightforward:

1. Specify probability density functions for the elements of the random vector. In most cases, only one type of density function is sufficient. It is assumed that the elements are independent random variables. Hence, the joint probability density function of the vector equals the product of the element probability density functions.

2. Compute for each m a representative hypervolume that contains all or a certain fraction of the possible regions in which the random vector may lie. This hypervolume is denoted by V_m and is obviously determined by the joint probability density function.

3. Consider a fixed r dimensional subspace of the m dimensional ambient space which represents the exact data.

4. From the density functions of the noise, compute the geometrical probability that the random vector will make an angle between α and $\alpha + d\alpha$ with the fixed r dimensional subspace. This is done by computing the volume $v(\alpha, r, m)$ generated by all vectors that make an angle between α and $\alpha + d\alpha$ with the r dimensional fixed subspace. The resulting density function $p(\alpha, r, m) = v(\alpha, r, m)/V_m$ will be called the the *directional density function.*

A most interesting case occurs if the directional density is independent of the considered subspace in relation to which the angles are computed:

Definition 3 Spherical probability density
A probability density function will be called spherical if its directional density function is independent of the coordinate system.

A vector in the m dimensional space can be described via a generalization of the classical polar coordinates with $m-1$ angles $\beta_1, \beta_2, \ldots, \beta_{m-1}$ and its length. If the probability density function of the vector is *spherical*, then all directions are equally alike and the density function of this vector, expressed in the new coordinates is uniform! Let's consider two clarifying examples in a two-dimensional ambient space: $m = 2$.

Example: Spherical density
Assume that the elements v_1 and v_2 of the vector v are independently and identically normally distributed with mean zero and variance σ^2. Consider as the volume V_2 the circle in the plane with radius σ: $V_2 = \pi\sigma^2$. From the independency of the element probability density functions, it is easily derived that, the probability that a vector makes an angle between α and $\alpha + d\alpha$ with an arbitrary direction $[cos\beta, sin\beta]$ is given by (figure 5.1.a):

$$p(\alpha, 1, 2)d\alpha = \frac{(\sigma d\alpha)(\sigma/2)}{\pi\sigma^2}$$

Hence, the directional density function is equal to:

$$p(\alpha, 1, 2) = \frac{1}{2\pi}$$

which is uniform. The distribution is spherical.

Example: Non-spherical density
Consider a one dimensional fixed subspace ($r = 1$) in a two dimensional ambient space $m = 2$. Let this subspace be generated by $\hat{a} = [\hat{a_1}\ \hat{a_2}]^t$ with $\hat{a_1}^2 + \hat{a_2}^2 = 1$. The elementwise noise probability density function is uniform between (-1) and $+1$. Obviously, an hypervolume V_2 in this case is simply the square centered around $(0, 0)$ with area equal to 4. The probability density function for an angle β, measured from the x-axis is given by:

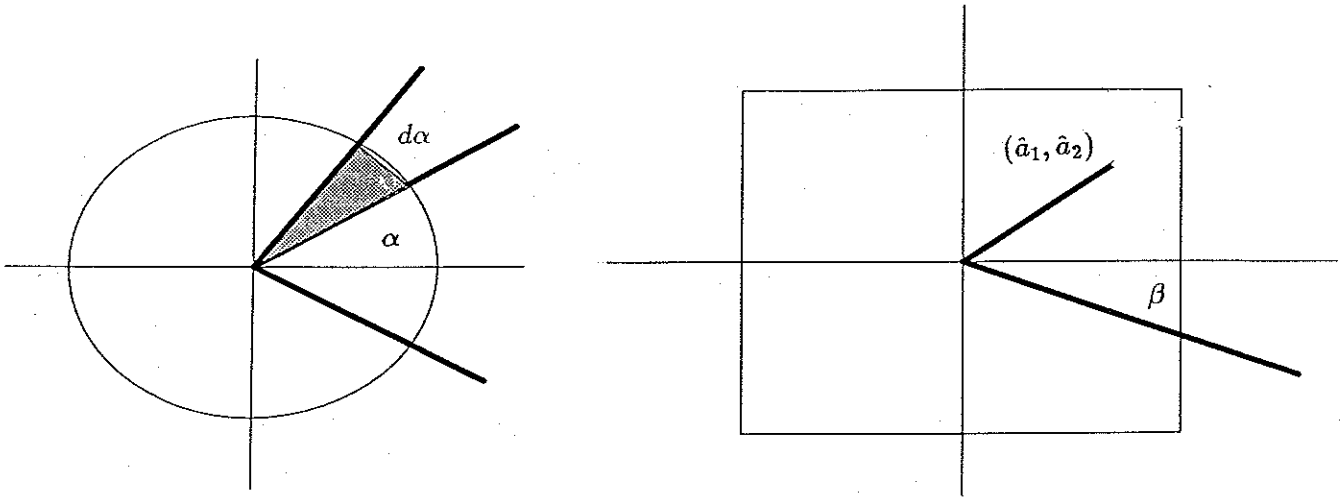


Figure 5.1: Derivation of the directional density functions.

- For $\beta \in [l\pi/2, (\pi/4 + l\pi/2)]\,, l = 0, 1, \ldots : p(\beta) = (8cos^2\beta)^{-1}$
- For $\beta \in [(\pi/4 + l\pi/2), (l+1)\pi/2]\,, l = 0, 1, \ldots : p(\beta) = (8sin^2\beta)^{-1}$

Let's now compute the directional density function representing the probability that a random vector makes an angle between α and $\alpha + d\alpha$ with $\hat{a}$. For a random vector v making an angle of β with the x-axis, this angle equals: $cos\alpha = \hat{a_1}cos(\beta) + \hat{a_2}sin(\beta)$. Hence, the required directional probability density is equal to :

$$p(\alpha, 1, 1) = arccos(\hat{a_1}cos(\beta) + \hat{a_2}sin(\beta))\, p(\beta)$$

Hence the directional density function $p(\alpha, 1, 2)$ explicitly depends upon the position of the fixed 1-dimensional subspace. The directional density function is not spherical.

Let's emphasize that *sphericity* is not equivalent with *isotropy* of the oriented energy distribution of a stochastic vector sequence, generated by the same probabilistic laws, unless the distributions are zero mean gaussian. As an example, consider again the second example with the uniform distribution. It is easily proved, that the oriented energy distribution of a 2-vector sequence with elementwise uniform distribution between $-\gamma$ and $+\gamma$ is *isotropic.* The expected value of the oriented energy of such a sequence in a direction measured by an angle α from the x-axis, is easily seen to be:

$$E_\alpha = \frac{1}{4\gamma^2} \int_{-\gamma}^{+\gamma} \int_{-\gamma}^{+\gamma} (xcos\alpha + ysin\alpha)^2 dxdy = \gamma^2/3$$

The oriented energy distribution is independent of the direction, hence is *isotropic.* Yet, the directional density is *non-spherical.*

We shall now derive the general directional probability density function of a random vector with a fixed r-dimensional subspace in an m-dimensional ambient space, under the assumption that:

All directions in the m-dimensional ambient space are equally alike for the random vector.

In other words, the directional density function is *spherical.* The main result is stated in the following theorem:

Theorem 1 Orthogonality and spherical densities.
Consider a fixed subspace $\mathcal{S}^r$ in the m dimensional vector space of m-tuples $\mathcal{R}^m$ ($m \gg r$). Assume that a random vector v is generated with equal probability for all directions in $\mathcal{R}^m$. Then the directional probability density function $p(\alpha, r, m)$ of the angle α between v and the fixed subspace $\mathcal{S}^r$ is given by:

$$p(\alpha, r, m) = K(r, m)(sin\ \alpha)^{m-r-1}(cos\ \alpha)^{r-1}$$

where $K(r, m) = 2\frac{C_r C_{m-r}}{C_m}$ and the C_k can be obtained recursively as:

- $C_1 = 1$
- $C_2 = \pi$
- $C_k = \frac{2\pi}{(k-2)} C_{k-2} \quad k > 2$

Proof: A proof, which is tedious though straightforward, was obtained in [23]. The reader is referred to appendix D. □

The result of this theorem will now be illustrated with a series of figures. In figure 5.2, $r = 2$ and the density function $p(\alpha, 2, m)$ is shown for 5 different values of m: (1) $m = 4$; (2) $m = 10$; (3) $m = 20$; (4) $m = 50$; (5) $m = 100$. One can see that the average angle α shifts towards orthogonality for increasing m. For instance, for $m = 100$, the probability that the angle α is less than 70^o , is almost zero. Also observe that the probability density function for $m = 2$ is symmetric with respect to 45^o. This is a general observation for every density function $p(\alpha, r, 2r)$. It expresses the fact that the random vector may belong to $\mathcal{S}^r$ and to $(\mathcal{S}^r)^\perp$ with equal probability. The qualitative explanation of figure 5.2 is intuitively straightforward. For fixed r and increasing m, the dimension of the orthogonal complement of the fixed subspace $\mathcal{S}^r$ increases. Since the random vector v has equal probability for all directions, the probability that it will have a bigger component in the orthogonal complement, increases with increasing overdetermination m/r. Also observe that for $m \to \infty$, the probability density function approaches more and more a Dirac impulse at 90^o. This is consistent with the definition of statistical orthogonality. Figure 5.3.a. allows to verify what happens if m/r is kept constant for increasing r and this for $m = 2r$. Observe that the density functions are symmetric around 45^o and that the standard deviation decreases for increasing r. Hence, in spaces of higher dimensionality, the random vectors tend to cluster around 45^o with a subspace of half the dimension. The same kind of result can be found in figure 5.3.b., but now for $m = 3r$. Observe now that the average angle converges to a value between 50 and 60 degrees.

Definition 4 Cumulative distribution function
The cumulative distribution function $P(\alpha, r, m)$ is defined as:

$$P(\alpha, r, m) = \int_{(\pi/2)-\alpha}^{\pi/2} p(x, r, m)dx$$

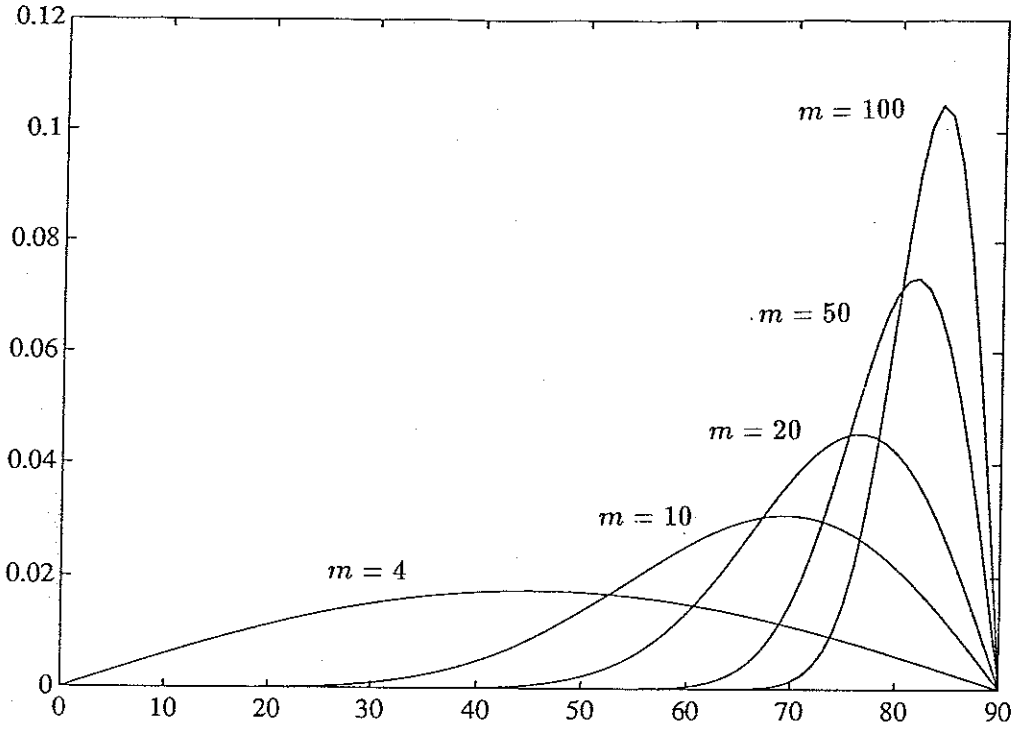


Figure 5.2: Probability density functions $p(\alpha, r, m)$ with fixed $r = 2$ and increasing m

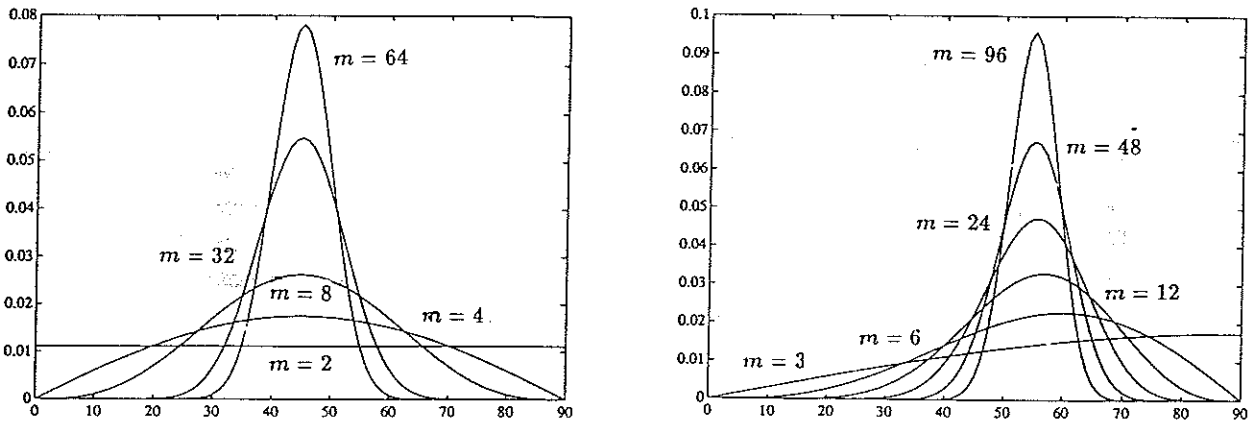


Figure 5.3: Probability density functions for constant ratio m/r. In figure 5.3.a., for $m = 2r$. In figure 5.3.b., for $m = 3r$.

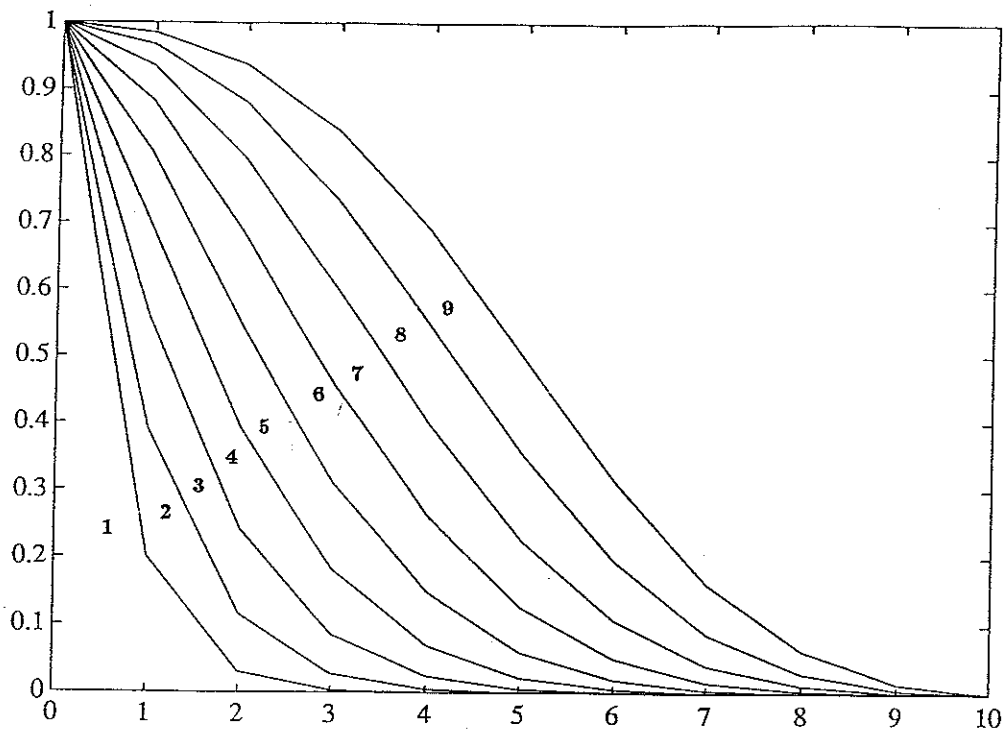


Figure 5.4: Cumulative distribution function $P(\alpha, r, m)$ for fixed $m = 10$ and increasing r in abscis. Plot k corresponds to $\alpha = k5^o$

Of course, $P(\alpha, r, m)$ is nothing else than the probability that the random vector v makes an angle between $\pi/2 - \alpha$ and $\pi/2$ with the fixed subspace $\mathcal{S}^r$. In figure 5.4., one finds in ordinate the probability that the angle is more than 45^o, 50^o, ..., 85^o for $m = 10$ with the dimension r of the fixed subspace in abscis.
In figure 5.5., we have chosen a fixed angle $\alpha = 45^o$. In ordinate one finds the probability that the random vector v makes an angle of more than 45^o with a fixed subspace of dimension r in absics. The different curves correspond to increasing values of m.
Finally, in figure 5.6., one finds an experimental verification of the results with the random number generator of Matlab. The continuous line is the theoretical probability that the random vector v makes an angle of more than 45^o with a subspace of dimension r (in abscis). The ambient space dimension $m = 10$. For each r, 100 random vectors are generated, the number of which the angle is larger than 45^o is devided by 100 and plotted with a $*$. This is repeated 10 times for each i.

Towards a generalization

The orthogonality theorem states that asymptotically, for increasing overdetermination and under the assuption of a *spherical* noise density function, the noise and the exact data column space will be orthogonal. Moreover, we have derived the probability density functions for all finite m as well. Generalization of this result is possible in two ways:

1. The proof is only valid for noise densities that are such that the probability for the random vector to lie in any direction of the ambient m-dimensional space is uniform

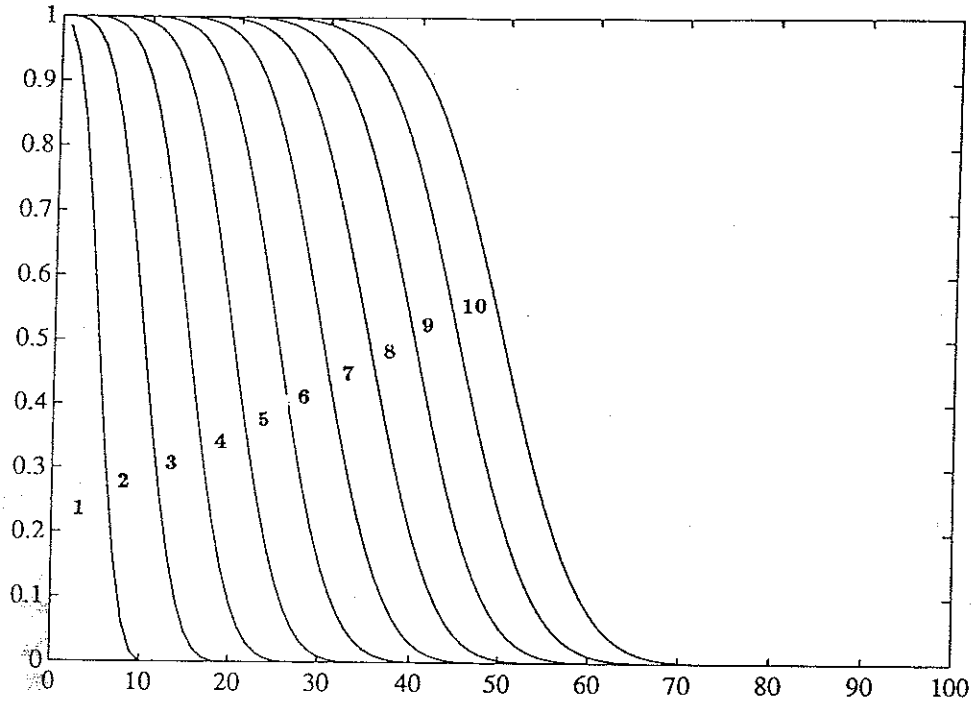


Figure 5.5: Probability (in ordinate) that a random vector makes an angle of more than 45^o degrees with a subspace of dimension r in abscis. The k-th plot corresponds to $j = 10k$.

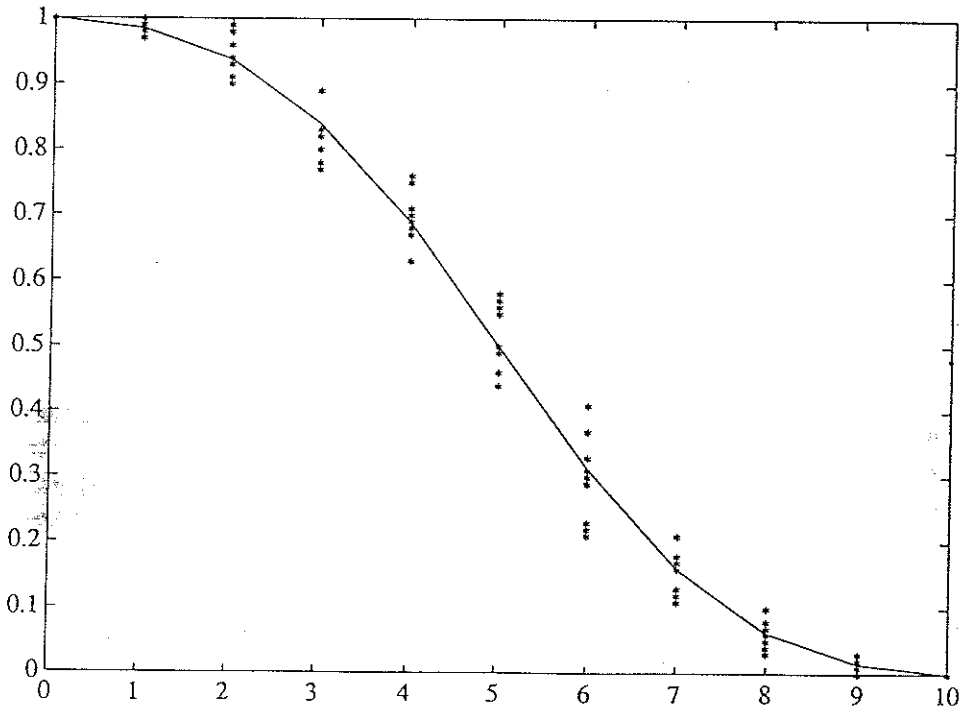


Figure 5.6: Verification with Matlab's normal distribution pseudo-random generator

The full line represents the computed probability that an arbitrary vector in R^{10} makes an angle of more than 45^o with an i-dimensional subspace (i in abscis). Each star represents 100 experiments: 100 arbitrary directions in R^{10} are generated, the number of them that makes an angle of more than 45^o with a fixed subspace S^i is divided by 100 and plotted.

over all directions. One example of such a density is of course the Gaussian density, which, by invoking the *central limit theorem*, happens to be a good noise model in practical applications. However, in principle, it must be possible to derive the same result for other density functions. As a matter of fact, it is intuitively seen that the precise form of the density function does not matter (as long as it has zero mean). The reason is that asymptotically, the orthogonality results from the increasing dimension of the orthogonal complement of the r-dimensional column space of $\hat{A}$ for increasing m. This is also confirmed by a closer look at the computation of the directional density function. Assume that an orthogonal coordinate system is chosen with coordinates $x_1, x_2, \ldots, x_m$ such that the first r coordinate axes are in the r-dimensional subspace $\mathcal{S}^r$ and choose the remaining $m-r$ in the orthogonal complement. Further let the joint probability density function be given as $p(x_1, x_2, \ldots, x_m)$ and determine a volume V_m that is representative for the geometrical distribution of the vectors. For a spherical distribution this is for instance a hypersphere with radius equal to the variance. For a uniform distribution, this is a rectangular hyperbox. Assuming that the angle is measured over the shortest arc, an expression for the expected value of the cosine of an angle α between one random direction and the r-dimensional subspace $\mathcal{S}^r$ is given by:

$$\begin{aligned} 0 \leq E(cos\alpha) &\leq \frac{1}{V_m} \int_{V_m} \frac{\sqrt{x_1^2 + \ldots + x_r^2}}{\sqrt{x_1^2 + \ldots + x_r^2 + x_{r+1}^2 + \ldots + x_m^2}} p(x_1, \ldots, x_m) dx_1 dx_2 \ldots dx_m \\ &\leq \frac{1}{V_m} \int_{V_m} \frac{\sqrt{x_1^2 + \ldots + x_r^2}}{\sqrt{x_1^2 + \ldots + x_r^2 + x_{r+1}^2 + \ldots + x_m^2}} dx_1 \ldots dx_m \end{aligned}$$

This expression suggests that, independent of the density function, $lim_{m \to \infty} cos\alpha = 0$.

2. In the case of a *spherical* density, the generalization towards several random vectors is straightforward by invoking the independence between the several columns of $\tilde{A}$. Indeed, it is the sphericity that allows to compute recursively the directional probability density function $p([\alpha_1, \ldots, \alpha_i], r+i-1, m)$ from $p([\alpha_1, \ldots, \alpha_{i-1}], r+i-2, m)$ for $i \geq 2$. As a matter of fact, what is really needed in general is the conditional distribution of the i-th random vector given $(i-1)$ random vectors. However, in the case of non-spherical distributions, the computation is highly complicated by the fact that the relative position of the r-dimensional column space of the matrix $\hat{A}$, will influence the results. The computation should therefore proceed via a *conditional density* procedure.

3. The orthogonality theorem provides also geometrical insight in the frequently met assumption that a covariance matrix of the noise is diagonal. This is equivalent with the fact that the columns of the noise matrix $\tilde{A}$ are mutually orthogonal.

5.2.3 The SVD of the sum of 2 matrices: the lever theorem.

Inspired by the results of the previous section, we shall now derive in a completely deterministic way, the SVD of the sum of two matrices whereby the first one contains an 'exact' signal and the second one represents the noise. While the orthogonality theorem is *in se* an asymptotic theorem, in order to derive the next theorem, we shall a priori assume that the noise sequence is orthogonal to the exact sequence. In this way, we hope to demonstrate

another asymptotic result, which reveals the structure of the singular value decomposition of a matrix when perturbed by noise. The derived result will allow to understand *geometrically* why some identification schemes may provide good results and others not.

Theorem 2 The lever theorem
Consider three $m \times n$ matrices A, B, C with $m > n$ and the following features:

1. $C = A + B$
2. $r = rank(A) < n < n + r \leq m$. *The SVD of A is given by $A = U_A S_A V_A^t$ where U_A is $m \times r$, S_A is $r \times r$ and V_A is $n \times r$. The orthogonal complement of the column space of V_A is generated by the columns of the $n \times (n - r)$ orthonormal matrix $V_A^\perp$.*
3. *B is of full column rank with SVD $B = U_B S_B V_B^t$ where U_B is $m \times n$, S_B and V_B are both $n \times n$.*
4. *The column spaces of A and B are orthogonal.*
5. $V_A^t(V_B S_B^t S_B V_B^t)V_A^\perp = 0$

then the SVD of C is given by:

$$C = U_C S_C V_C^t = (X_1 \; X_2) \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix} \begin{pmatrix} Y_1^t V_A^t \\ Y_2^t (V_A^\perp)^t \end{pmatrix}$$

where the matrices $X_1, S_1, Y_1, X_2, S_2, Y_2$ follow from the SVD of:

$$\begin{aligned} CV_A &= X_1 S_1 Y_1^t \\ CV_A^\perp &= X_2 S_2 Y_2^t \end{aligned}$$

Proof: Since $C = A + B$, it is easy to show that:

$$\begin{aligned} C &= U_A S_A V_A^t + U_B S_B V_B^t \\ &= U_A S_A V_A^t + U_B S_B (V_B^t V_A V_A^t) + U_B S_B (V_B^t (I_n - V_A^t V_A)) \\ &= (U_A S_A + U_B S_B V_B^t V_A) V_A^t + (U_B S_B V_B^t V_A^\perp)(V_A^\perp)^t \\ &= P_1 V_A^t + P_2 (V_A^\perp)^t \end{aligned}$$

with obvious definitions for P_1 and P_2. Observe that $P_1 = (A + B)V_A = CV_A$ and $P_2 = (A + B)V_A^\perp = CV_A^\perp$. Let these matrices on their turn have the SVD's:

$$P_1 = X_1 S_1 Y_1^t \qquad P_2 = X_2 S_2 V_2^t$$

then the matrix C can be written as:

$$C = [X_1 \; X_2] \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix} \begin{pmatrix} Y_1^t V_A^t \\ Y_2^t (V_A^\perp)^t \end{pmatrix}$$

Observe that this would be a valid singular value decomposition of C, up to a reordening of the singular values, if:

$$X_1^t X_2 = 0$$

or equivalently, if:

$$\begin{aligned} P_1^t P_2 &= (S_A^t U_A^t + V_A^t V_B S_B^t U_B^t)(U_B S_B V_B^t V_A^\perp) \\ &= V_A^t V_B S_B^t U_B^t U_B S_B V_B^t V_A^\perp \\ &= V_A^t V_B S_B^t S_B V_B^t V_A^\perp \\ &= V_A^t B^t B V_A^\perp \\ &= 0 \end{aligned}$$

which is 0, according to assumption 5. □

Let's analyse a little bit more in detail the seemingly strange assumption 5. First observe that $V_B S_B^t S_B V_B$ is nothing else but the Grammian of the oriented energy distribution of the row vector sequence of B. Since V_B is square orthonormal, there must exist orthonormal matrices Q_1 and Q_2 of appropriate dimensions such that $V_A = V_B Q_1$ and $V_A^\perp = V_B Q_2$. It is straightforward to prove that these matrices should be a solution to the following set of equations:

$$\begin{aligned} Q_1^t Q_1 &= I_r \\ Q_2^t Q_2 &= I_{m-r} \\ Q_1^t Q_2 &= 0 \\ Q_1^t S_B^t S_B Q_2 &= 0 \end{aligned}$$

If $S_B^t S_B$ is a multiple of the identity matrix, this condition is really not restrictive in this sense that the computed factorization of C will always be an SVD. This corresponds to the case that the oriented energy distribution of the row vector sequence of B is isotropic. For general $S_B^t S_B$ the solution looks as follows: Q_1 may consist of any set of r different columns of the identity matrix I_m (or a linear combination of them). Q_2 may contain the remaining $m - r$ columns of I_m (or a linear combination of them). In brief, Q_1 and Q_2 must have a complementary zero pattern. This implies that the columns of V_A are linear combinations of r singular vectors of U_B, while the columns of $V_A^\perp$ are linear combinations of the remaining $m - r$ singular vectors of U_B.

The following conclusions concerning the lever theorem are important:

- Loosely speaking, the lever theorem states that, under certain conditions on orthogonality between the exact and the pure noise space, it is possible to recover *exactly* (asymptotically) the features of the *short space*. This can be seen from the fact that in the theorem, the right singular vectors, corresponding to S_1, are a linear combination $V_A Y_1$ of the right singular vectors of the matrix A. An equivalent statement for the left singular vectors is not possible. This will also be obvious from an example to be presented below. This is the origin of the name *lever theorem*. By adding the matrix B, the *long space* is irreparably disturbed, while the *short space* is more robust.

- It can be seen that the singular value decomposition of the matrix C essentially consists of two singular value decompositions: One of the matrix $CV_A V_A^t$, which is the orthogonal projection of the rows of C onto the subspace generated by the right singular vectors contained in V_A. The other SVD is that of $CV_A^\perp (V_A^\perp)^t$ which is the projection of C onto the orthogonal complement of the row space of A.

- Since the derived factorization of C is an SVD up to a reordening of the singular value values, there is another criterion which is necessary. It confirms the intuition that the possibility of recovering the correct subspace V_A depends upon the signal-to-noise ratio: The separation should be possible from the singular values in S_1 and S_2. If one knows a priori that all singular values in S_1 are larger than those in S_2, then if the other conditions are satisfied, the first r right singular vectors of C will generate the subspace generated by the columns of V_A.

- Observe that when A represents exact data, and B pure noise data, then the orthogonality theorem of section 5.2.2. proves that this condition is asymptotically satisfied and that for large overdetermination m/n, it is a good approximation.

- Assumption 5 should be satisfied. However, we have demonstrated, that in case that the oriented energy distribution of the row vector sequence of the matrix B is isotropic, then assumption 5 is automatically satisfied. Fortunately, if the columns of B are zero mean independent identically distributed random variables, the corresponding oriented energy distribution is always approximately isotropic for sufficiently large m. As a matter of fact, assumption 5 provides a rationale for the so-called *Mahalanobis* transformation which essentially consists of a *prewhitening* of the data with the inverse of the noise covariance matrix. Obviously, this corresponds to using another inner product in which the symmetric positive definite weighting matrix is the inverse noise covariance matrix. With respect to this new weighting, the spatial oriented energy distribution of the noise vector sequence is isotropic, hence assumption 5 is satisfied. Hence we have demonstrated that the Mahalanobis transformation in fact makes the 'exact' data and the noise data orthogonal! This idea is at the basis of successful implementations as the Kalman filter and high resolution array processing.

Corollary 1 The lever theorem for Gaussian disturbances.
Let the entries of the matrix $m \times n$ matrix B be independently normally distributed with zero mean and variance σ^2. Assume that the $m \times n$ matrix A is of rank r with $r < n < n+r < m$, with SVD $A = U_A S_A V_A^t$, then, with increasing probability for increasing overdetermination m/n, the SVD of $C = A + B$ will approach the following 'limit' SVD:

$$C = ((U_A S_A + \sqrt{m}\sigma U_1)(S_A^2 + m\sigma^2 I_r)^{-1/2} \quad U_2^t) \begin{pmatrix} (S_A^2 + m\sigma^2 I_r)^{1/2} & 0 \\ 0 & \sqrt{m}\sigma I_{m-r} \end{pmatrix} \begin{pmatrix} V_A^t \\ P^t (V_A^\perp)^t \end{pmatrix}$$

where $U_1 = \frac{1}{\sqrt{m}\sigma} B V_A$ and $U_2 = \frac{1}{\sqrt{m}\sigma} B V_A^\perp$ and P is an arbitrary orthonormal $(m-r) \times (m-r)$ matrix.

Proof: Note that the covariance matrix of the row vectors of B equals $m\sigma^2 I_n$. The proof follows from a straightforward application of the orthogonality theorem 1 and the lever theorem 2. □

The following remarks apply:

- Observe that there is a Pythagoras-like squaring of scaled orthonormal matrices in the long space. This implies that it is impossible to recover the original exact columnspace of A from the SVD of C.

- Also note that there is a 'noise' threshold visible in the singular value spectrum of the matrix C which allows to estimate the rank of the matrix A.
- The results for the consistency of the *short space* estimate and for the singular spectrum, are well known in the literature on statistical estimation of linear relations [2]. However, the expression for the *long space*, in terms of the left singular vectors of the matrix C, is new. It will prove to be extremely important in the applications we are going to consider.

These remarks are now illustrated in the following example:

Example:
Consider a $m \times 3$ matrix A all of which the elements are equal to 1. This matrix is perturbed by a $m \times 3$ pure noise matrix B, of which the elemements follow a zero mean normal distribution with variance $\sigma^2 = 1$. According to corollary 1, the SVD of the sum $C = A + B$ will approach, for increasing m, the limit SVD:

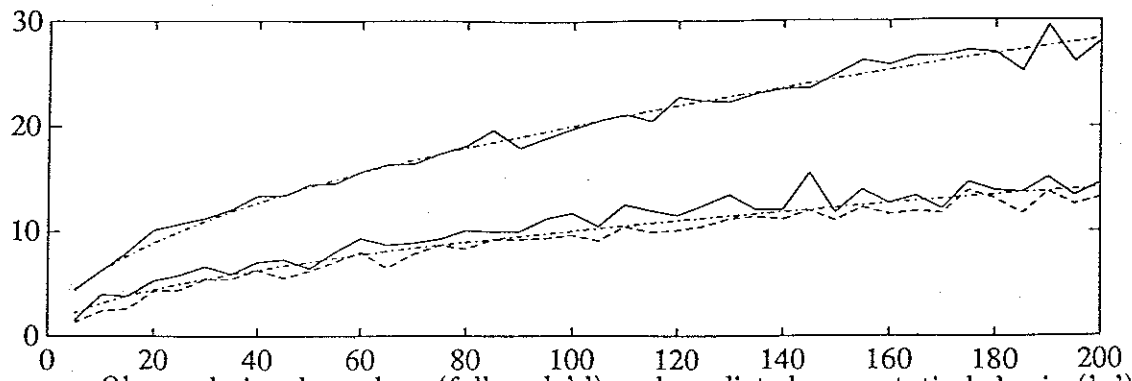
$$lim_{m\to\infty} C = \left(\begin{pmatrix} 1/\sqrt{m} \\ \cdots \\ 1/\sqrt{m} \end{pmatrix} \frac{\sqrt{3}}{2} + B \begin{pmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{pmatrix} \frac{1}{2\sqrt{m}} \quad \frac{1}{\sqrt{m}} B V_A^{\perp} P \right)$$
$$diag(2\sqrt{m}, \sqrt{m}, \sqrt{m} \begin{pmatrix} 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\ & (V_A^{\perp} P)^t & \end{pmatrix}$$

where P is an arbitrary orthonormal 2×2 matrix. In figure 5.7.a.(upper curve), one can find the canonical angles (for a definition see section 5.4) between $span_{col}(B)$ and $span_{col}(A)$, illustrating the orthogonality theorem. The singular values of C as a function of m are depicted in figure 5.7.b, together with their asymptotic behavior. The angle between the first right singular vectors of A and C is plotted in figure 5.7.a.(lower curve), while the angle between the first left singular vector of A and C can be found in figure 5.7.a.(middle curve). Obviously, the short space is consistently estimated. The long space however not. It may be verified that the angle between $span_{col}(A)$ and $span_{col}(C)$ will converge to 30^o.

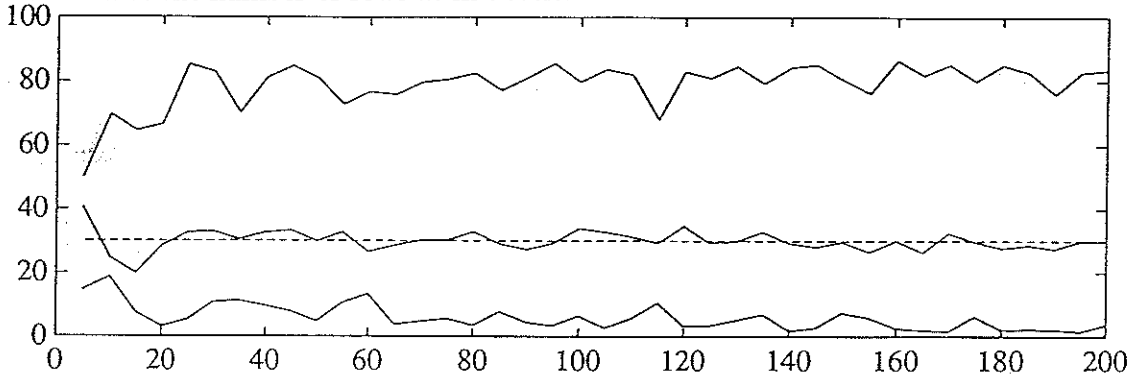
The same example permits to explain geometrically why the long space can not be estimated consistently. The reason is the Pythagoras-like squaring. The explanation can most easily be followed on figure 5.8. The first left singular vector can be considered as the middle line of an m-dimensional hypersphere S_m. Let's call this vector c. This middle line must be the sum of two orthogonal vectors $c = a + b$ where $a^t b = 0$. Hence a will lie on the hypersurface of the sphere. If the noise level σ is known a priori, the length of b equals $b^t b = m\sigma^2$. Hence b is a solution of $c^t b = m\sigma^2$ which represents an hyperplane, orthogonal to c but not through the origin. The intersection of the hypersphere S_m with this hyperplane is another hypersphere S_{m-1} in a lower dimensional space. It is easy to see that there will be infinitely many pairs a and b since it suffices that a lies on the surface of the hypersphere S_{m-1}. Already for $m = 2$, there are 2 solutions.

5.3 Identification schemes for static linear systems.

In this section, we shall investigate the problem of obtaining linear relations from noisy data in a *deductive* framework. The *inspiration* approach will be postponed until chapter 6. The



Observed singular values (full and '-') and predicted asymptotic behavior('.-') as a function of the number of rows m in abscis.



Observed canonical angles (full) and predicted asymptotic behavior as a function of the number of rows m.

Figure 5.7: Illustration of the lever theorem.

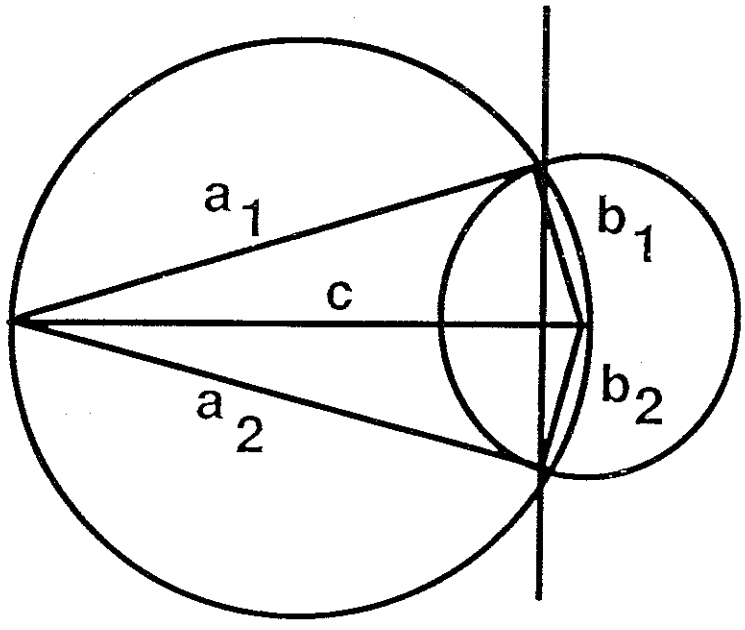


Figure 5.8: Inconsistency of the long space.

problem and its solutions will be considered from the point of view of matrix algebra. However, we shall also frequently exploit the geometrical insights obtained in the previous section: the orthogonality and the lever theorem.

In section 5.3.1, we briefly review the classical 'ordinary' least squares method. The total least squares method is shortly discussed in section 5.3.2. These derivations are well known. A less well known fact is that both approaches are special cases of a unifying theorem, which states how the rank of matrices can be lowered by one under certain conditions. This result is derived and discussed in section 5.3.3. In section 5.3.4. we discuss another modelling approach, based upon nonnegativeness requirements of certain matrices.

The matrix formulation of the problem that will be considered here, reduces to the following:

Given an $m \times n$ matrix A of rank n with $m > n$. Find a $m \times n$ matrix $\tilde{A}$ and $n \times 1$ vectors x such that:

1. $rank(A - \tilde{A}) \leq n - 1$ where we shall denote the difference $A - \tilde{A}$ as $\hat{A}$.
2. $\hat{A}x = 0$

These requirements are common to all the schemes that will be discussed. However, as we shall see, additional requirements may be imposed for several reasons. They constitute the essential difference, especially from the conceptual point of view, between the identification schemes.

- The orthogonality of the column spaces of $\hat{A}$ and $\tilde{A}$ may be required. That this condition is quite natural, was demonstrated with the orthogonality theorem. Hence, the assumption $\hat{A}^t\tilde{A} = 0$. It will be shown that all discussed identification schemes satisfy this condition.

- It may be required that $rank(\tilde{A}) = n$. The reason for this is, that when $\tilde{A}$ is rank deficient, there exist linear relations between the columns of the noise matrix, hence contradicting our general principle that noise should be absence of linear relations. Yet, as will be demonstrated, linear and total linear least squares violate this essential assumption, hence do not provide good noise models (unless in very restricted cases).

- Denoting $\Sigma = A^tA$, $\hat{\Sigma} = \hat{A}^t\hat{A}$, $\tilde{\Sigma} = \tilde{A}^t\tilde{A}$, then, if $\hat{A}^t\tilde{A} = 0$, we have that:

$$\Sigma = \hat{\Sigma} + \tilde{\Sigma}$$

  Hence the problem specialises to: Given the positive definite $n \times n$ matrix Σ, find a non-negative or positive definite matrix $\tilde{\Sigma}$, such that $\hat{\Sigma} = \Sigma - \tilde{\Sigma}$ is rank deficient and nonnegative. This last requirement is essential since $\hat{\Sigma}$ is to be the Grammian of a certain matrix $\hat{A}$.

- It may be required that $\tilde{\Sigma}$ be a diagonal matrix. The reason for this is again the orthogonality theorem, stating that asymptotically, all noise vectors will be orthogonal. Linear least squares satisfies this condition (although in a very restrictive way), total linear least squares not, the identity matrix approach does and also the models considered in chapter 6.

- It may be required that the so called corank of $\hat{A}$, which equals $n - rank(\hat{A})$ is to be minimized. Chapter 6 is completely devoted to this exciting, yet unsolved problem.

5.3.1 Linear Least Squares.

The Linear Least Squares (henceforth abbreviated as LLS) problem is a computational problem of primary importance in many applications. As is well known, the method of least squares was apparently first used by Gauss in 1795, though it was first published by Legendre in 1805. It is less well known that Adrian in America, unaware of these developments, independently developed the method in 1808 [9]. The principle is fairly general and applies to all kinds of frameworks such as stochastic processes, Fourier analysis, orthogonal polynomials, Wiener and Kalman filtering, the innovations approach, etc...Probably, it constitutes one of the most studied and analysed principles of mathematical engineering. In this thesis, the LLS problem will be discussed in terms of matrix algebra. However, *mutatis mutandis*, the results can be translated to other domains since the basic principles and resulting interpretations are very general.

If one wants to fit a linear model to given data, typically, one uses a greater number of measurements than the number of unknowns. According to Thurstone as cited in [15], it is a general principle that a scientific theory must be *overdetermined* by the data. Mathematically, this means that the data are supposed to be connected by certain relations so that they cannot be regarded as independent variables. Thurnstone's principle is intuitively understandable from the point of view of *uniqueness* of the solution of a problem. In the context of linear static identification however, it has as an important consequence that the orthogonality theorem can be invoked in order to analyse under which circumstances linear least squares estimation will provide a good solution. The linear least squares problem is the following:

Given an $m \times n$ matrix A with $m > n$ and b an m-vector. Find an n-vector x such that $\|b - Ax\|_2^2$ is minimized.

If rank(A)=n, then the LLS will be called *generic*. Although important, both from a conceptual as computational point of view, we shall not consider in detail the case where A is near rank deficient (as revealed by its singular values). For the numerical treatment of these inconveniences, which is the domain of *regularization* techniques, the interested reader is referred to the extensive literature on the subject [7] [14]. Instead we are interested in the conceptual properties of the generic (i.e. full column rank) linear least squares as a preparation to the conceptual results in chapter 6. The well known mathematical solution of the general LLS problem is summarized in the following theorem:

Theorem 3 Linear Least Squares. *The solution to the general LLS problem is:*

$$x = A^+ b$$

The residual $(b - AA^+b)$ satisfies the orthogonality property

$$A^t(b - AA^+b) = 0$$

where A^+ is the pseudo-inverse of the matrix A.

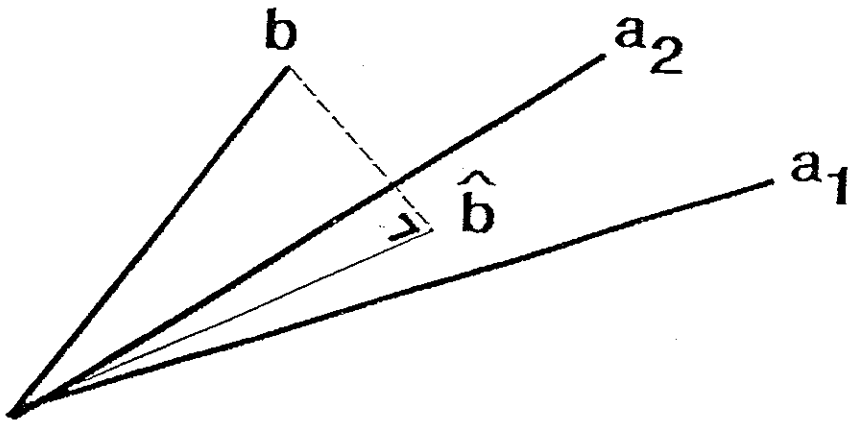


Figure 5.9: Linear least squares for a three variable problem

Proof : Trivial when employing singular value decomposition. □

Observe that the matrix AA^+ is the projection operator onto the column space of the matrix A.

Corollary 2 Generic linear least squares *The solution to the generic LLS problem with* $rank(A) = n$, *is given by:*

$$x = (A^t A)^{-1} A^t b$$

Verify that the vector $\hat{b} = Ax = A(A^tA)^{-1}A^tb$ is a vector in the column space of A. It is nothing else than the orthogonal projection of the vector b onto the columnspace of A. The geometrical interpretation of the linear least squares problem is depicted for a 2 variable problem in figure 5.9. Note that this geometrical picture applies *mutatis mutandis* to all mathematical disciplines where the LLS principle applies.

From the preceding discussion, it is obvious that in the solution of the linear least squares problem of estimating x in $Ax = b$, one acts as if only the right hand side b is subject to inaccuracies. Assume that the exact vector $\hat{b}$ lies in the columnspace of A but that only a noise corrupted vector b is available. If the noise is assumed to follow a zero mean Gaussian distribution, then a simple look at the orthogonality theorem 1 suffices to see that the probability that the pure noise vector $\tilde{b}$ will be orthogonal to the exact version $\hat{b}$ and hence to the columnspace of A, increases with increasing overdetermination according to the probability density functions derived in theorem 1. Hence we have proved *consistency* in a *deductive* framework. A formal proof of it can be found in almost all textbooks on linear least squares estimation and identification (e.g. [19, p.97]).

Obviously, linear least squares estimation also satisfies *consistency* from the *inspiration* point of view, since if the data are exact, it will provide the correct answer.

However, if all the data, i.e. the elements of A *and* those of b are measured with the same tool, or more generally, if all data are noisy, the column vectors a_i of A are to be treated

symmetrically with respect to the vector b : This means, that there is no (objective) reason to prefer the vector b as a right hand side instead of any other column vector of the matrix A.

Proposition:
Consider a matrix A with m rows and n columns ($m > n$) and rank(A)=n. Denote by A_i the $m \times (n-1)$ matrix obtained by deleting the i-th column a^i of A. Then, there are n linear least squares estimation problems with solution x^i , each of which is defined by:

$$min_{x^i} \|a^i - A_i x^i\| \quad i = 1, \ldots, n$$

Obviously,

$$x^i = (A_i^t A_i)^{-1} A_i^t a^i$$

The following theorem describes a remarkable geometrical insight.

Theorem 4 On all possible linear least squares solutions
Define $\Sigma = A^t A$ where A is a $m \times n$ matrix of rank n with $m > n$. Then the i-th linear least squares solution, defined in proposition 1 equals (up to a scalar) the i-th column s^i of $S = \Sigma^{-1}$. Moreover, the norm of the corresponding residu equals $(1/s_i^i)^2$ where s_i^i is element (i,i) of Σ^{-1}.

Two proofs of this result will be given. One here, based on a well known matrix inversion lemma and one in section 5.3.3.

Proof 1 : Without loss of generality, the result will be proved for $i = n$. The Grammian Σ may be partioned as :

$$\Sigma = \begin{pmatrix} A_n^t A_n & A_n^t a^n \\ (a^n)^t A_n & (a^n)^t a^n \end{pmatrix}$$

Since A is nonsingular, also A_n is nonsingular and a^n is not identically zero, hence the Grammian $A_n^t A_n$ is invertible. Now assume that A_n has a QR-decomposition $A_n = Q_n R_n$. Applying the matrix inversion lemma for partioned matrices (Appendix C), results in the following expression for the n-th column s^n of Σ^{-1}:

$$s^n = \begin{pmatrix} (A_n^t A_n)^{-1} A_n^t a^n \\ -1 \end{pmatrix} \alpha_n$$

where $\alpha_n = (a^n)^t (I_n - Q_n Q_n^t) a^n$. Applying theorem 3 then proves the theorem. □

Note that this theorem confirms the fact that, when all variables are noisy, linear least squares estimation delivers *biased* estimates!
In chapter 6, we shall discuss more conceptual ideas behind this important theorem. However, observe the essential difference in interpretation of the problem formulation and the solution provided by the theorems 3 and 4 which reduces to the distinction between *predictive* and *descriptive* modelling and identification:

- In theorem 3, one tries to write a *specified* vector b in the best possible way as a linear combinations of the columns of the matrix A. This point of view may be appropriate from a *predictive* standpoint, where it is assumed that our 'past' knowledge on the system is contained in the column space of the matrix A.

- In theorem 4, one tries to find linear relations among the columns of the matrix A. There is no special reason to prefer one column with respect to another one. This corresponds to a *descriptive* viewpoint. The disappointing fact is that in this case, linear least squares provides n different but equivalent linear models.

5.3.2 Total Linear Least Squares

The idea of applying another correction to the data than only a modification of the right hand side b in the set of linear equations is already old: The oldest efforts can be traced back as far as 1878 by Adcock and 1904 by Pearson [2]. Since then it has been *reinvented* many times, in particular in the statistical community where it is known under the name *errors - in - variables*. Recently, Golub and Van Loan [7] have proposed a solution procedure, based upon the singular value decomposition, which was analysed in depth in [24], mainly in a deductive framework. The total linear least squares via SVD is discussed from an inspirationist point of view in [26], in a modelling context based upon the paradigm of low complexity and low misfit models. One of the main reasons of the intensive interest in total linear least squares strategies, is that linear least squares fails to provide a *unique descriptive* linear model if all data are considered to be noisy as was expressed by theorem 4. It will now be demonstrated that the total least squares approach succeeds in resolving this problem, but fails with respect to another necessary feature of good descriptive methods: It provides a unique linear relation from a descriptive point of view, when all data are noisy but it is not consistent in its noise model.

The total linear least squares problem is the following:

Consider an $m \times n$ matrix A ($m > n$) containing m measurements on an n-vector signal. Find the vector $x \in \mathcal{R}^n$ that solves:

$$min_x \|A - \hat{A}\|_F^2 \ \ such \ \ that \ \ \hat{A}x = 0$$

Observe that, if $rank(A) = n$, the problem in essence reduces to finding in $\mathcal{R}^{m \times n}$, that matrix $\hat{A}$ of rank $n-1$ or lower, that is closest to A in Frobeniusnorm. The basic solution can be found in terms of singular value decomposition.

Theorem 5 Total linear least squares
Let A be an $m \times n$ matrix $(m \geq n)$ of full rank n with singular value decomposition:

$$A = \sum_{i=1}^{n} u_i \sigma_i v_i^t$$

Then, the solution to:

$$min_{\hat{A}} \|A - \hat{A}\|_F^2 \ \ such \ \ that \ \ rank(\hat{A}) = n - 1$$

is given by:

$$\hat{A} = \sum_{i-1}^{n-1} u_i \sigma_i v_i^t$$

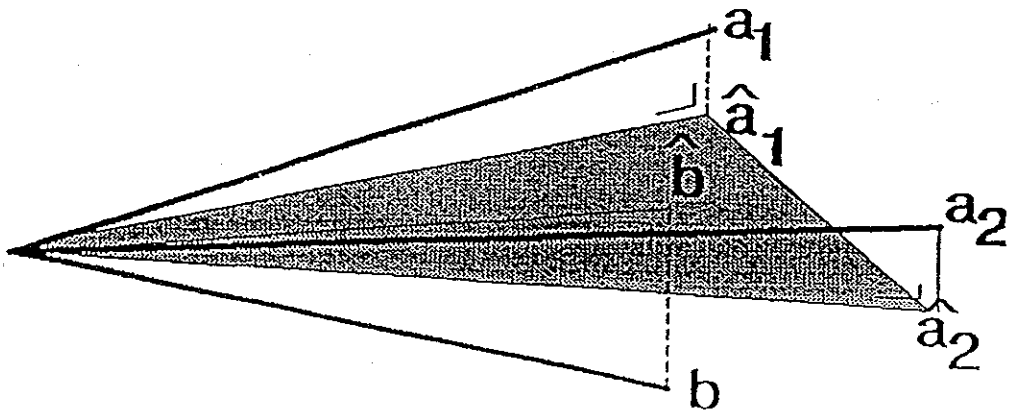


Figure 5.10: Total linear least squares with three variables.

Proof: [7]. □

An illustration of the geometry of the total least squares solution in the column space of A is depicted in figure 5.10. for a three variable example.
The following remarks are in order:

- Observe that we have stated the total linear least squares problem without mentioning a right hand side b. Our point of view is, that when the problem is stated as solving as good as possible $Ax \simeq b$, hence fixing a specified vector b as right hand side, this is in contradiction with the spirit and origin of total linear least squares as a *descriptive* modelling approach. The problems one might run into when trying to solve $Ax \simeq b$ arise when the last component v_{n+1}^{n+1} of the smallest singular vector v^{n+1} of the matrix $[A\ b]$ is almost zero. This corresponds to the so called non-generic TLLS [24]. However, in all practical applications, the linear relation described by the smallest singular vector v^{n+1} *an sich*, without normalization, is sufficient. It contains all necessary information that one could need *without the need for any normalization.* This conclusion will have considerable consequences.

- A largely unexplored difference between linear least squares and total linear least squares is the comparison of their generalization to more than one right hand side: Solve the $n \times p$ matrix X from $AX \simeq B$ where A is $m \times n$ and B is $m \times p$.

  Linear least squares solution: (generic)

  $$X = (A^t A)^{-1} A^t B$$

  It is easily seen (and it can be proved straightforwardly) that the p columns of X correspond to the solution of the p separate LLS problems. This *decoupling* property is a very salient feature in a lot of derivations.

  Total Linear Least Squares: (generic)

  $$X = V_{21} V_{22}^{-1}$$

where the $(n+p) \times p$ matrix V_2 contains the smallest p right singular vectors of $[A\ B]$. V_{21} is the upper $n \times p$ part while V_{22} is the lower $p \times p$ part of V_2. The relation of the columns of X with the p TLLS solutions of the separate problems is not obvious.

- The following result allows to treat identification problems with mixed exact and noisy data. It is a combination of linear least squares and total linear least squares, due to Golub [7].

  Lemma 1 *Assume that the first p columns of a $m \times n$ matrix A, with $m \geq n$, are noise free while the others are noisy, then the mixed least squares-total linear least squares solution to $Ax \simeq 0$ is obtained as follows:*

  - *Partition $A = (A_1 A_2)$ and accordingly $x = (x_1^t x_2^t)^t$, where A_1 is $m \times p$.*
  - *Compute the QR-factorization of A:*

$$Ax = Q \begin{pmatrix} R_1 & R_2 \\ 0 & R_3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \simeq 0$$

  - *Solve for x_2 the total least squares problem:*

$$R_3 x_2 \simeq 0$$

  and for x_1 the least squares problem:

$$R_1 x_1 = -R_2 x_2$$

- The total linear least squares solution consists of modifying the data matrix with a rank one matrix. The solution is the smallest singular vector. From the lever theorem, it is easily seen that the solution will be *consistent* in a *deductive* approach, in that it asymptotically will converge to the original one (if the exact data matrix was of rank $n-1$). This is also proved in [24]. The noise model provided by the TLLS scheme however, is *highly unrealistic*: It is a rank one matrix, hence there are $n-1$ linear relations among the noise variables. Remember however that we agreed on the fact that *noise should be absence of linear relations*! The fact that total linear least squares provides a good *descriptive* solution (namely a consistent one) when all data are perturbed by noise and when there exists an exact matrix of rank $n-1$, can be considered as *pure coincidence*! The reason is that its solution coincides with that of an identification scheme that is discussed in section 5.3.4., which provides a more realistic noise model (as a matter of fact, one of full rank!). The proof in [24] is a perfect copy of the classical proof of consistency of the identity matrix approach of section 5.3.4. Finally, note that from an *inspiration* point of view, the total least squares is consistent as well.

- The noise model provided by total linear least squares is not only not satisfactory because of its rank one property. An important drawback is that it also changes the data *within the column space of the matrix A*. This is in contradiction with the orthogonality theorem since this states that the original exact data are perturbed additively by the noise in a orthogonal way. Hence, the resulting noisy observations can never stay in the same subspace as the original exact data!

5.3.3 Rank one modifications.

While the previous sections contained results that are more or less well known, at least from the technical point of view, in this section we shall derive a new theorem that unifies linear and total linear least squares. The result essentially states that both schemes are special cases of the same general identification approach, which consists of identifying approximate linear relations from noisy data, by a rank one modification.

Rank one modification of the covariance

We shall now present a new theorem which provides a unification of identification schemes that are implicitely based upon a rank one modification of the matrix $\Sigma = A^t A$. At the same time, it includes a second proof for theorem 4. It is assumed that $\hat{A}^t \tilde{A} = 0$.

Theorem 6 Rank one modification of the covariance matrix
Let Σ be an $n \times n$ symmetric positive definite matrix. Let $\tilde{\Sigma}_p = p\sigma_p^{-1}p^t$ be a rank one matrix where p is an $n \times 1$ vector with unit norm $\|p\| = 1$ and σ_p a positive real number. Define the matrix $\hat{\Sigma} = \Sigma - \tilde{\Sigma}$. Then the conditions:

1. *$rank(\hat{\Sigma}) = n - 1$*
2. *$\hat{\Sigma}$ is nonnegative definite*

are satisfied if and only if $\sigma_p = p^t\Sigma^{-1}p$. The solution x_p of $(\Sigma - \tilde{\Sigma})x_p = 0$ is then given by $x_p = \Sigma^{-1}p$.

Proof:

If part: We first prove that $rank(\hat{\Sigma}) = n - 1$.

$$(\Sigma - \tilde{\Sigma})x_p = \Sigma\Sigma^{-1}p - \sigma_p^{-1}pp^t\Sigma^{-1}p = p - p = 0$$

Because $rank(\tilde{\Sigma}) = 1$ and $rank(\Sigma) = n$, it follows that $rank(\hat{\Sigma}) = n - 1$. The nonnegativeness of $\hat{\Sigma}$ follows from the fact that the matrix :

$$(\Sigma - \tilde{\Sigma})\Sigma^{-1}(\Sigma - \tilde{\Sigma})$$

is obviously nonnegative definite. It is an easy exercise to show that it is precisely equal to $\Sigma - \tilde{\Sigma}$.

Only if part: The condition:

$$(\Sigma - p\sigma_p^{-1}p^t)x_p = 0$$

implies that:

$$x_p = \Sigma^{-1}p\sigma_p^{-1}(p^t x_p)$$

Hence the solution must be proportional to the vector $\Sigma^{-1}p$. Put $x_p = \Sigma^{-1}p\alpha_p$ where α_p is a non-zero scalar. Then:

$$p\alpha_p = p\sigma_p^{-1}p^t\Sigma^{-1}p\alpha_p$$

Because $p^t p = 1$, left multiplication with p^t results in

$$\sigma_p = p^t\Sigma^{-1}p$$

□

This theorem has important conceptual consequences.

- The theorem provides a parametrization of all rank one modifications of a positive definite matrix such that the result is nonnegative definite. Once a vector p has been fixed:
  - The 'noise energy' σ^{-1} is determined via the quadratic form $\sigma_p = p^t \Sigma^{-1} p$
  - The solution x_p is a linear combination of the columns of Σ^{-1}. The weights are the elements of the vector p.
- In theorem 4, we have found that the n possible linear least squares solutions for the problem are precisely the columns of Σ^{-1}. Hence, by manipulation of the coefficients of the vector p, one can influence the geometrical position of the solution x_p with respect to the n linear least squares solutions. If it is a priori known that the variables 1 to k are noise free and that the other ones are perturbed by the same amount of noise, simply set $p^t = [0 \ldots 1 \ldots 1]$. As a special case, the i-th linear least squares solution is obtained by choosing $p = e^i$, the i-th column of the identity matrix I_n. This confirms the interpretation of linear least squares as that solution to the problem that only considers one of the variables to be noisy. Observe the the energy of the residual of the i-th linear least squares solution is given by $(e^i)^t \Sigma^{-1} e^i$, which corresponds to the result of theorem 4.
- Also the total linear least squares solution is easily found to be a special case. Simply choose $p = v^n$, the 'smallest' eigenvector of Σ. The corresponding solution follows from $x_{TLLS} = \Sigma^{-1} v_n = (\sigma_n)^{-2} v_n$ and hence can be taken to be equal to the smallest right singular vector of A, as was proved in theorem 5. The 'noise energy' equals the smallest eigenvalue of Σ.
- Finally observe the important role that is played by the inverse of the Grammian Σ. This prominence will be confirmed further in chapter 6.

Example:
Identify a linear relation of the form $y = x\gamma$ from the data depicted in figure 5.11.a with Grammian matrix given by:

$$\Sigma_1 = \begin{pmatrix} 2.8 & 3.6 \\ 3.6 & 8.2 \end{pmatrix}$$

The data have zero mean. The vector p of theorem 6 can be parametrized as $p = [\alpha \ , \ \pm\sqrt{1-\alpha^2}]$. The linear least squares results are found for $\alpha = 1$ ($\gamma = 2.2778$) and for $\alpha = 0$ ($\gamma = 1.2857$). The total least squares solution results in $\gamma = 2$. As could be expected this is in between the slopes of the linear least squares results. The noise energy σ_p as a function of α is depicted in figure 5.11.b. The upper curve corresponds to the noise being of a much higher energy level than the data, the lower curve to the reverse situation. The solution with minimum noise energy is the total linear least squares solution. The asymptote in figure 5.11.c. corresponds to that value of α where the identified slope changes sign.

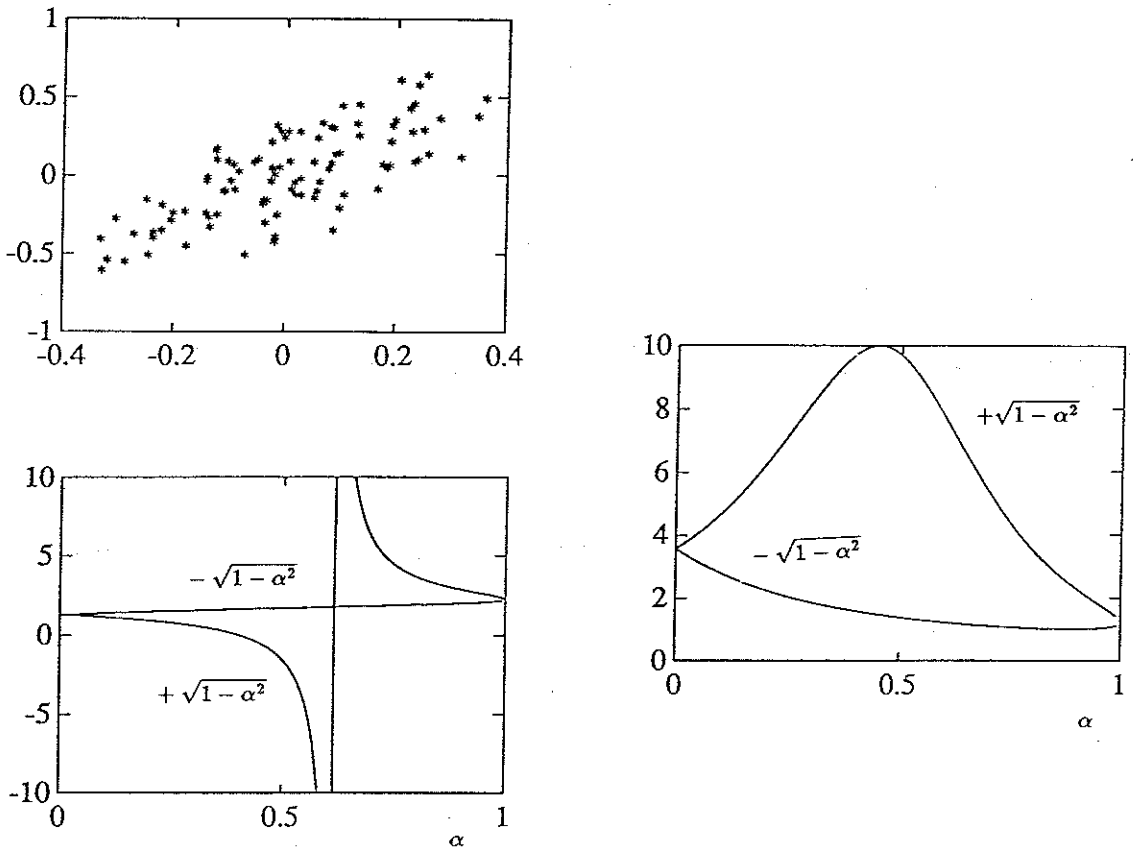


Figure 5.11: Data scatter (a), Noise energy (b) and identified slope (c) as a function of α.

Qualitatively the same results can be seen in figure 5.12. but now for a Grammian matrix:

$$\Sigma = \begin{pmatrix} 20.8 & 39.6 \\ 39.6 & 80.2 \end{pmatrix}$$

The quantitative differences between figure 5.11. and figure 5.12. are:

- The condition number in figure 5.11.is 10, while in figure 5.12. it equals 100. Hence, in figure 5.11. the scatter of the data is less than in figure 5.12. This results in a relative smaller difference between the linear least squares solutions in figure 5.12. with respect to figure 5.11.
- The eigenvectors are the same in the two examples, hence also the total linear least squares solutions. Observe that in figure 5.12., there is relatively less difference between all solutions than is the case in figure 5.11. The minimum of the noise curve is less explicit in figure 5.12.

Rank one modifications of a matrix

Having stated the previous theorem that unifies least squares and total least squares identification scheme, let's return to the problem formulation in terms of an overdetermined $m \times n$ matrix A of rank n with $m > n$. By considering rank one modifications of this matrix, we shall find back the same results as in theorem 6, but in addition it will be shown that **of all rank one modifications, total least squares is the most restrictive.** Hereto, we shall derive results in three consecutive stages:

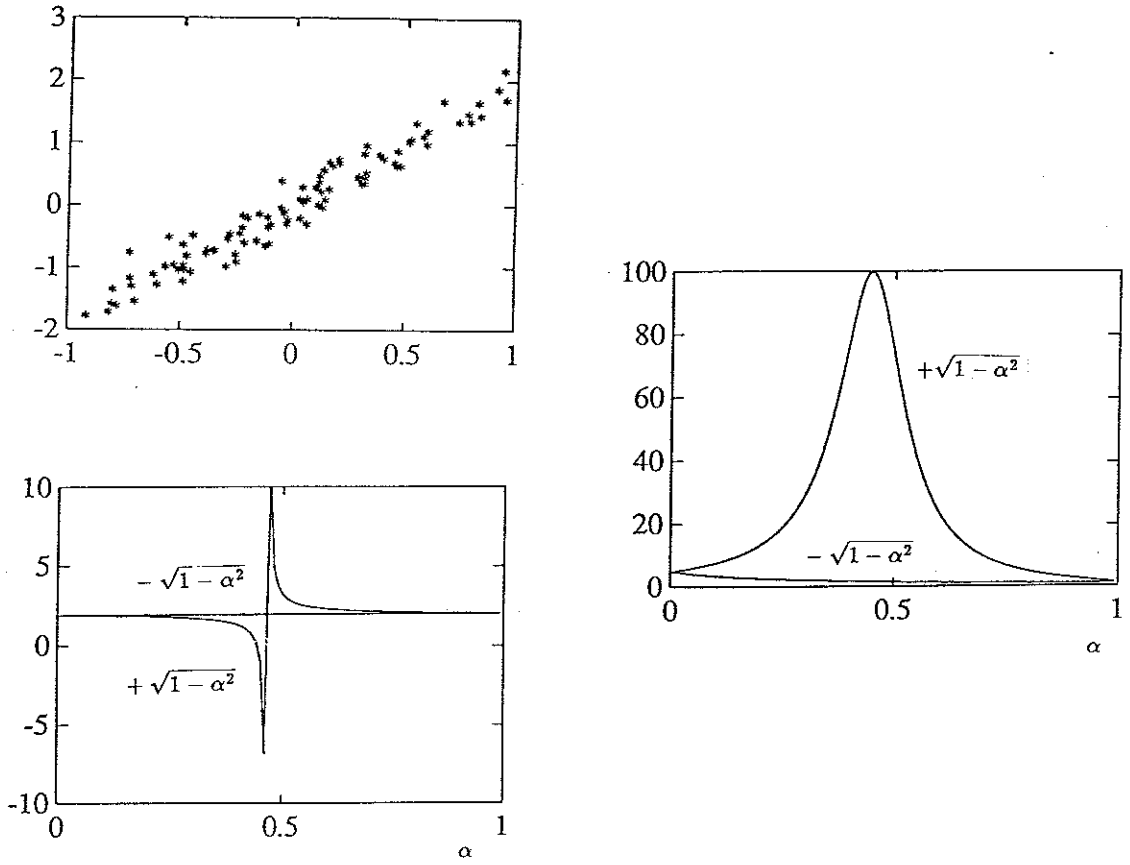


Figure 5.12: Data scatter (a), noise energy (b) and identified slope (c) as a function of α.

1. First some general properties of rank one modifications are stated.
2. Second, we introduce some orthogonality requirements.
3. Third, it is demonstrated that Total Linear Least Squares even demands for one additional orthogonality requirement.

The problem that will be considered is the following:

Given a $m \times n$ matrix A with $m > n$, of full column rank n. What are the conditions on vectors u $(m \times 1)$ and v $(n \times 1)$ and the scalar σ such that:

$$rank(A - u\sigma^{-1}v^t) = n - 1$$

or equivalently, there exists a $n \times 1$ vector x such that:

$$(A - u\sigma^{-1}v^t)x = 0$$

The results are summarized in the following theorem:

Theorem 7 General properties of rank one modifications.
Given an $m \times n$ matrix A, $m > n$, $rank(A)=n$;
Then, for an $m \times 1$ vector u and an $n \times 1$ vector v, both with unit norm, and a scalar σ:

$$rank(A - u\sigma^{-1}v^t) = n - 1$$

and there exists a vector x such that $(A - u\sigma^{-1}v^t)x = 0$:

1. *if* $v^t x \neq 0$ *(necessary).*
2. *iff* $u \in span_{col}(A)$
3. x *is proportional to* $y = (A^t A)^{-1} A^t u$
4. $u^t A (A^t A)^{-1} A^t u = 1$
5. $\sigma = v^t (A^t A)^{-1} A^t u$

Proof:

1. By contradiction, Assume $v^t x = 0$. From $(A - u\sigma^{-1} v^t)x = 0$, it follows immediately that $Ax = (v^t x \sigma^{-1})u$. Hence, $Ax = 0$ which is impossible because $rank(A) = n$.
2. Call $(v^t x)\sigma^{-1} = \alpha$. From $Ax = u\alpha$ it follows immediately that $u \in span_{col}(A)$.
3. Note that also $\alpha = u^t A x$ because u has unit norm. Now, since $rank(A) = n$, $(A^t A)^{-1}$ exists and x can be solved as $x = (A^t A)^{-1} A^t u \alpha$.
4. Follows from :
$$\alpha = (v^t x)\sigma^{-1} = u^t A x = u^t A (A^t A)^{-1} A^t u \alpha$$
5. Follows from:
$$(A - u\sigma^{-1} v^t)x = (A - u\sigma^{-1} v^t)(A^t A)^{-1} u = 0$$

□

While at first sight, these results look quite trivial, they learn us a lot about identification schemes based upon rank one modifications. In these applications, the matrix A will contain m measurements on n measurement channels and a linear relation among them is to be discovered. The 'noise' matrix $\tilde{A}$ is modelled as a rank one matrix, leaving behind the model for the 'exact' data $\hat{A} = A - \tilde{A} = A - u\sigma^{-1} v^t$.

- Observe that the vector u must lie in the column space of the matrix A.
- Once a vector u is fixed, the solution x is determined as well, because x is proportional to $y = (A^t A)^{-1} A^t u$. Observe that this represents the least squares expression for y in $Ay = u$. However, since u is in the column space of A, this overdetermined set of equations is perfectly well solvable.
- σ is determined by u and v. However, for each u there exist infinitely many v and σ.

Observe, that we did not require any additional assumption on the vectors u and v, such as orthogonality requirements. We know however from section 5.2.2., that such orthogonality conditions between noise and exact data are legitimate asy.mptotically. If these are introduced, we arrive at the following results:

Corollary 3 Rank one modifications with orthogonality requirements
Assume that, in addition to all conditions of theorem 7, it is required that the column space of the noise matrix $\tilde{A} = u\sigma^{-1} v^t$ *is orthogonal with respect to the column space of the matrix that models the exact data* $\hat{A} = A - \tilde{A}$. *Then:*

1. $A^t u = v\sigma^{-1}$
2. $\sigma^2 = v^t(A^tA)^{-1}v$
3. $\sigma^{-2} = u^tAA^tu$
4. $u = A(A^tA)^{-1}v\sigma^{-1}$

Proof :

1. From the orthogonality of the columnspaces, it follows that:

$$(A^t - v\sigma^{-1}u^t)u\sigma^{-1}v^t = 0$$

so that:

$$(A^tu - v\sigma^{-1})v^t = 0$$

Since $v \neq 0$, this is only possible if $A^tu = v\sigma^{-1}$.

2. Follows from theorem 7, from $s = v^t(A^tA)^{-1}A^tu$ and $A^tu = v\sigma^{-1}$.
3. Follows directly from $A^tu = v\sigma^{-1}$ and the fact that $v^tv = 1$.
4. From $Ay = u$, we get $A^tAy = A^tu = v\sigma^{-1}$ so that $y = (A^tA)^{-1}v\sigma^{-1}$.

□

Observe that once u has been fixed (necessarily in the column space of A), then both v and σ are now well determined, contrary to the result without orthogonality requirements. Hence, the mere introduction of the orthogonality requirement, reduces the number of possible rank one noise models. As a special example, consider *linear least squares identification* which reduces to a modification of one column of A, say the i-th, such that the modification and the result are orthogonal. Hence, in order to find the i-th LLS solution, simply require that $v = e^i$, the i-th column of I_n. Then it is an easy exercise to verify that $\sigma^2 = (e^i)^t(A^tA)^{-1}e^i$ which confirms the results of theorem 4 and 6. Moreover, for rank one modifications that satisfy the required orthogonality of the column spaces of $\hat{A}$ and $\tilde{A}$, it can be proved that:

Corollary 4 Rank one modifications and nonnegative definiteness
If $\hat{A}^t\tilde{A} = 0$ and $\tilde{A} = u\sigma^{-1}v^t$, then $A^tA - v\sigma^{-2}v^t$ is nonnegative definite.

Proof : Observe that

$$(A^tA - v\sigma^{-2}v^t)(A^tA)^{-1}(A^tA - v\sigma^{-2}v^t)$$

is nonnegative definite. A straightforward calculation shows that this expression equals precisely $A^tA - v\sigma^{-2}v^t$. □

Hence, all column orthogonal models satisfy at once the required nonnegative definitenness conditions of theorem 6. It will now be demonstrated, that one arrives at the total linear least squares solution, only if *in addition to all previous conditions, also row orthogonality is required!*

Corollary 5 Total linear least squares
Assume that in addition to all previous conditions, stated in theorem 7 and corollary 2 and 3, it is also required that:

$$\hat{A}\tilde{A}^t = 0$$

then the only possible rank one modifications are terms from the dyadic decomposition of A.

Proof : From the additional row orthogonality requirement, it follows that:

$$(A - u\sigma^{-1}v^t)v\sigma^{-1}u^t = 0$$

which implies:

$$Av = u\sigma^{-1}$$

Together, with the result of corollary 2 that $A^t u = v\sigma^{-1}$, this defines a Schmidt pair of the matrix A, hence the rank one modification is a dyadic decomposition term of A. □

Observe that of course only the term corresponding to the smallest singular value will satisfy the required nonnegative definiteness conditions of theorem 7. Also note that most of the results are readily generalized to rank higher than one modification, such as rank 2, 3, etc..., both from the computational as the conceptual point of view (e.g. with respect to the noise model interpretation). We mention this fact here explicitly, because total linear least squares is sometimes used in identification with several right hand sides. In the case of e.g. 2 right hand sides, the result of the identification is then a rank 2 noise model, which is of course again very structured!

In summary, the conceptual objections against linear and total linear least squares, for identification of linear relations among variables that are all noisy, are that:

1. The noise model is essentially a rank one modification of the data, hence the noise is very structured. Hence, the noise model is not *consistent*, both for linear as total linear least squares schemes and all possible cases in between (partial total linear least squares when some of the variables are a priori known to be exact).

2. The least squares solution is *not consistent* when all data are noisy. As a matter of fact, there are n linear least squares solutions.

3. Remarkably enough, the solution computed by the total linear least squares scheme is consistent, because it coincides with the solution of another identification scheme, which is consistent both in the noise model as in solution. Hence, when one is only interested in the solution, there is no danger in using total linear least squares, on the contrary, its numerical robustness is very favorable. If however one is also interested in the noise model (as will be the case in section 5.4.), total least squares is to be rejected.

5.3.4 The identity matrix as a noise model

It was demonstrated that both linear and total linear least squares provide a noise model that is completely unsatisfactory from the conceptual point of view, if all variables are a priori known to be subject to noise. The reason is the rank one property of the noise model. In this section, it will be shown how to obtain a solution satisfying the following conditions:

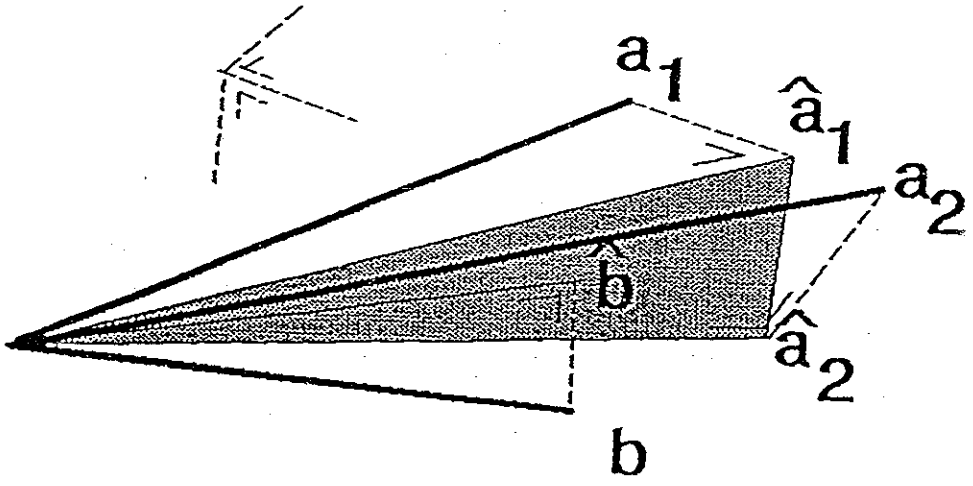


Figure 5.13: The identity matrix approach for a three variable example.

1. $rank(\tilde{A}) = n$
2. $\hat{A}^t\tilde{A} = 0$
3. $\tilde{\Sigma} = \tilde{A}^t\tilde{A}$ is diagonal.
4. $\hat{\Sigma} = \Sigma - \tilde{\Sigma} = A^tA - \hat{A}^t\hat{A}$ is rank deficient and nonnegative definite.

A solution follows immediately from the eigenvalue decomposition of Σ:

Theorem 8 The identity matrix approach.
Let Σ be an $n \times n$ symmetric positive definite matrix with eigenvalue decomposition:

$$\Sigma = \sum_{i=1}^{n} v^i \sigma_i^2 (v^i)^t$$

with $\sigma_i \geq \sigma_{i+1}$. Then $\Sigma - \sigma_n^2 I_n$ is rank deficient and nonnegative definite.

Proof: Trivial. □

First observe that the σ_i are the singular values and the eigenvectors v^i are the right singular vectors of A. Again, one could make a distinction between the generic ($\sigma_{n-1} > \sigma_n$) and the non-generic case ($\sigma_{n-k} = \ldots = \sigma_n$). We shall only discuss the generic case. It is easy shown that there will be infinitely many matrices $\hat{A}$ and $\tilde{A}$ that satisfy all conditions. This can be seen from the following parametrization:

- Let $\tilde{U}$ and $\hat{U}$ be orthonormal $m \times n$ matrices that are such that $\hat{U}^t\tilde{U} = 0$.
- $\tilde{A} = \sigma_n\tilde{U}$ with $\tilde{U}^t\tilde{U} = I_n$
- $\hat{A} = \hat{U}(S - \sigma_n I_n)V^t$ where S contains the singular values of A

The following remarks are important from the conceptual point of view:

- Obviously, there are infinitely many pairs $\hat{U}$, $\tilde{U}$. The conceptual interpretation is however crucial: The existence of infinitely many noise models $\sigma_n \tilde{U}$ is appealing since it implies that it is *impossible to recover exactly the noise and the exact data!*. Hence, in the cycle *exact data* $\rightarrow$ *additive noise* $\rightarrow$ *identification*, the uncertainty about the exact data has increased. This observation is exploited in [11] to define a non-decreasing entropy function for this cycle. The reader should be aware of the fact that the non-uniqueness of the noise model, is in strong contrast with linear and total linear least squares, where, generically the noise model was *unique*. Of course, one could argue that the matrix $\tilde{\Sigma} = \sigma_n^2 I_n$ is also unique. This is so of course, but this uniqueness is less restricive than the uniqueness in the linear and total linear least squares. As a matter of fact, in chapter 6 we shall also remove the uniqueness of $\tilde{\Sigma}$!

- Despite the explicit non-uniqueness of $\hat{A}$, the linear relation found back by this scheme, is *unique*. It is the smallest right singular vector of A. It is easy to see from the lever theorem, that the identification scheme proposed here, will be *consistent* under appropriate conditions of the noise. Observe that the solution precisely coincides with that of the total linear least squares scheme, which explains the consistency of the total linear least squares scheme, despite its *wrong* noise model, and its succes in identifying linear relations from noisy data from a deductive point of view. The proof of consistency for the identity matrix approach from a *deductive* point of view, can be found in classical textbooks and articles [2] [19]. Note that also from the *inspiration* point of view, the method is *consistent*.

- The total linear least squares noise model is of rank one. However, this is not its only serious drawback. It will now be shown that the noise model depends upon the linear relation which is being identified. From a deductive point of view, this is of course a regrettable property.

Example:
In order to demonstrate the essential difference between total linear least squares and the identity matrix approach, let's consider the following extremely simple two variable identification experiment: Given m measurements of two signals, contained in the m-vectors x and y. Determine the constant α such that y can be linearly explained by x as $y \simeq x\alpha$. Both the TLLS and the identity matrix approach can be analysed in terms of the SVD of the $m \times 2$ matrix A:

$$A = [\, x \;\; y \,] = u_1 \sigma_1 v_1^t + u_2 \sigma_2 v_2^t$$

Assuming that $v_2^t = [v_{12} \;\; v_{22}]$, both TLLS and the identity matrix approach deliver as a solution for α:

$$\alpha = -v_{12}/v_{22}$$

where it is tacitly assumed that $v_{22} \not\simeq 0$. Hence both strategies deliver the same result for the linear relation. However, let's now have a look at the corresponding noise model. For the identity matrix approach, the noise model is given by:

$$[\, \tilde{x} \;\; \tilde{y} \,] = \sigma_2 \tilde{U}_2$$

where $\tilde{U}_2^t \tilde{U}_2 = I_2$. which states that the noise model is of rank two and the noise sequences $\tilde{x}$ and $\tilde{y}$ are orthogonal and of equal energy. Moreover, observe that the noise model is in this case independent of the linear relation. With the discussion of the orthogonality property of noise sequences with respect to exact data of section 5.2 in mind, it follows that this identification scheme is consistent both in the obtained model of the linear relation and the noise, whenever, from a deductive point of view, the data were obtained from an exact linear law, but with observations that are corrupted by additive noise. As a consequence, if the same experiment were performed but now with another scalar β instead of α, but with the same observation equipment to measure x and y, one would find the same noise level σ_2 for x and y. If however, the TLLS noise model is analysed, one finds:

$$[\ \tilde{x}\ \ \tilde{y}\ ] = u_2 \sigma_2 v_2^t$$

The noise is structured: it is of rank one. Moreover, observe that it is dependent upon the linear model, because it is easily verified that $\tilde{x} = \tilde{y}(-\alpha)$, indicating that if α is big, the energy of the noise in the x direction, is α^2 times larger than that of the noise in the y direction. Hence, if the same experiment were repeated with another scalar β, one would now find a noise vector in the x direction, which is $-\beta$ the noise vector in the y direction!

5.4 Computing intersections between spaces

In a lot of applications, the intersection of two spaces is of importance. As an example, it will be shown in chapter 7, how the structure of certain spaces can be computed from the intersection of other spaces (structure exploiting factor analysis). Some of the identification algorithms of chapter 8, are based upon an implicit computation of the intersection of two spaces. In this section, it will be shown how the previously discussed identification schemes can be applied to compute *approximate intersections* in case the available data are corrupted by noise. It is shown how only the linear least squares approach succeeds in approximating the correct intersection, under the assumptions that are necessary for a successful application of least squares for solving a set of linear equations. Total linear least squares and the identity matrix approach however, do **not** succeed in computing the correct intersection, even if the assumptions for a successful use of these techniques are fulfilled. Fortunately, in most applications the explicit computation of the intersection is not needed, but a *dual model* is sufficient. This can be consistently computed as will be shown, employing the lever theorem of section 5.2.3.

This section is organised as follows. Following a statement of the problem formulation, we review the notion of canonical angles between subspaces. A perturbation analysis is performed as to demonstrate the bias in canonical angles when the data are perturbed by noise. It is shown how the optimization of a certain approximation criterion allows to compute approximate intersections in terms of the generalized singular value decomposition. The results are clarified with several examples and illustrations.

5.4.1 The problem formulation.

Given a $m \times n$ matrix A and an $m \times p$ matrix B with $m \geq max(n,p)$. Compute a $m \times r$ matrix C such that the columns of C are a basis for the intersection of the column space of

the matrices A and B. Hence:

$$span_{col}(C) = span_{col}(A) \bigcap span_{col}(B)$$

Observe that the problem has two aspects:

1. Determine the dimension r of the intersecting subspace.
2. Determine r linear independent basis vectors for this subspace.

We want to emphasize the essential difference of this problem with the problems that can be solved via our oriented energy framework of chapter 4. Oriented energy and the computation of oriented signal-to-signal ratios, allow to analyse the geometrical distribution of vectors in the 'short' spaces while computing the intersection between spaces is posed as a problem in the 'long' space of a matrix.
Although several methods of computing an (approximate) intersection will be discussed, sooner or later they all end up in an interpretation in terms of canonical angles between subspaces. Therefore, first this issue is developed.

5.4.2 Canonical correlation analysis

The theory of canonical correlations and variables was developed independently by Hotelling [8] and Obukhov [9]. While Hotelling's original derivation was mainly in terms of matrix algebra. today, the notion of canonical variates has been extended and generalized. Applications include data analysis [5], random processes [4] [9] and minimal realization procedures of Markov processes [1] [13] [21]. A numerically stable method to compute the canonical structure via a singular value decomposition has been proposed by Golub [6].
Canonical correlation starts with the computation of an orthonormal basis for the column spaces of the matrices A and B. This can for instance be done in a numerically reliable way via QR-factorization or via singular value decomposition, which has the additional advantage that at the same time, the dimensions of the column spaces of A and B are estimated via the singular values, at the cost of more floating point operations. However, unless a rank revealing QR algorithm would be used, the QR-decomposition will deliver n orthonormal vectors for A and p for B, while the true dimensionality of the corresponding subspaces could be lower. Hence, we prefer to use the singular value decomposition.

$$\begin{aligned} A &= U_A \Sigma_A V_A^t \\ B &= U_B \Sigma_B V_B^t \end{aligned}$$

Here U_A is an $m \times r_A$ matrix and U_B an $m \times r_B$ matrix where r_A and r_B are the rank of A and B, and hence the dimension of their column space. The following definition provides a generalization of the concept of an angle between two vectors.

Definition 5 Canonical angles between subspaces.
Assume that U_A $(m \times r_A)$ and U_B $(m \times r_B)$ are orthonormal matrices, then the canonical angles $\theta_1, \ldots, \theta_{min(r_A, r_B)}$ between $span_{col}(U_A)$ and $span_{col}(U_B)$ are defined recursively as:

- $cos(\theta_1) = max(u^t v) = u_1^t v_1$ *with* $u \in span_{col}(U_A)$ *and* $v \in span_{col}(U_B)$ *and* $u^t u = 1 = v^t v$

- $cos(\theta_k) = max(u^t v) = u_k^t v_k$ *for* $k = 2, \ldots, min(r_A, r_B)$ *with* $u \in span_{col}(U_A)$, $v \in span_{col}(U_B)$, $u_k^t u_l = \delta_{kl} = v_k^t v_l$.

The directions u_i and v_i are termed the principal directions between the subspaces.

Note that θ_1 is the smallest canonical angle. In order to get more insight into the concept of canonical angles, let's consider the following theorem:

Theorem 9 Canonical angles and singular value decomposition
The cosines of the canonical angles $\theta_1, \ldots, \theta_{min(r_A, r_B)}$ between the columnspaces of the orthonormal $m \times r_A$ matrix U_A and the orthonormal $m \times r_B$ matrix B, are the singular values of $U_A^t U_B$:

$$cos(\theta_i) = \sigma_i \, (U_A^t U_B)$$

Proof: Since $u_1 \in span_{col}(U_A)$, there must exist a vector p_1 with $p_1^t p_1 = 1$ such that $u_1 = U_A p_1$. Similarly, there exists a vector q_1 with $q_1^t q_1 = 1$ such that $v_1 = U_B q_1$. From $u_1^t v_1 = p_1^t U_A^t U_B q_1$ it follows that this expression is maximized if p_1 is the largest left and q_1 the largest right singular vector of $U_A^t U_B$. Let σ_1 be the corresponding largest singular value. We still have to prove that $\sigma_1 \leq 1$ if it has to be the cosine of an angle. Hereto, rewrite U_B as:

$$U_B = U_A U_A^t U_B + (I_m - U_A U_A^t) U_B = U_A P_1 + U_A^{\perp} P_2$$

Obviously, $U_B^t U_B = I_{r_B} = P_1^t P_1 + P_2^t P_2$ so that $\sigma_i(P_2^t P_2) = 1 - \sigma_i(P_1^t P_1)$. From the nonnegative definiteness of $P_2^t P_2$ it then follows that

$$\sigma_i(P_1^t P_1) \leq 1$$

Since now $U_A^t U_B = P_1$, it follows that all singular values of $U_A^t U_B$ are equal or smaller than 1. □

First note that if U_A and U_B were vectors, everything reduces nicely to the conventional Euclidean inner product between vectors. Hence, the matrix product $U_A^t U_B$ represents a kind of generalized inner product. Observe that the extreme geometrical situations are obvious from the theorem. If the two column spaces $span_{col}(U_A)$ and $span_{col}(U_B)$ coincide, then there exists an orthonormal matrix P such that $U_B = U_A P$. Hence $U_A^t U_B = P$, which has all of its singular values equal to 1. Hence all canonical angles are 0! If the column spaces are orthogonal, then obviously $U_A^t U_B = 0$ a matrix with all of its singular values 0. Hence all canonical angles are equal to 90°!
The following result is important, especially from a conceptual point of view:

Theorem 10 Principal directions and pseudo-inverse.
The canonical angles between and the principal directions in the column spaces of two matrices A and B can be obtained from the singular value decomposition of the matrix:

$$AA^+ BB^+ = USV^t$$

The cosines of the canonical angles are the singular values of S, the principal directions in $span_{col}(A)$ are the left singular vectors and the principal directions in $span_{col}(B)$ are the right singular vectors.

Proof: The proof follows immediately from the singular value decomposition of A and B. □

Observe that a matrix of the form AA^+ is nothing else but *the projection operator* onto the column space of a matrix A.
It is easily verified that the pseudo-inverse of a vector a is given by:

$$a^+ = a^t/\|a\|^2$$

Hence, the principal directions between two vectors and their angle follow from:

$$\begin{aligned} aa^+bb^+ &= aa^tbb^t/(\|a\|^2\|b\|^2) \\ &= \frac{a}{\|a\|}(\frac{a^tb}{\|a\|\|b\|})\frac{b^t}{\|b\|} \end{aligned}$$

The singular value of this rank one matrix is nothing else than the well known expression for the cosine of the angle between two vectors in a Euclidean space.
The following corollary provides a numerical robust algorithm to compute the intersection of two spaces.

Corollary 6 Computing the intersection
The dimension r of the intersection of the column spaces of an orthonormal $m \times r_A$ matrix U_A and an orthonormal $m \times r_B$ matrix U_B is equal to the number of singular values of $U_A^tU_B$ equal to 1:

$$r = dim[span_{col}(U_A) \bigcap span_{col}(U_B)] = \#\{\sigma_i(U_A^tU_B) = 1\}$$

Let $U_A^tU_B$ have the SVD :

$$U_A^tU_B = PSQ^t = [\ P_1\ \ P_2\ ] \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix} \begin{pmatrix} Q_1^t \\ Q_2^t \end{pmatrix}$$

where S_1 is the $r \times r$ diagonal matrix containing the singular values equal to 1 and P_1 and Q_1 contain corresponding left and right singular vectors. Then the intersection is generated by the columns of the $m \times r$ matrix $C = U_AP_1$ and also by $C = U_BQ_1$.

Proof: see e.g.[7]. □

The major practical difficulty with this computation technique is the fact that the cosines of canonical angles close to 0^o are not very sensitive. Hence, it is sometimes difficult to decide whether an angle equals 0^o or not, from inspection of its cosine! Since one can also compute in a numerically reliable way (orthogonal projections onto) orthogonal complements of subspaces, it is sometimes better to exploit the following results:

Corollary 7 *Let $\theta_1, \ldots, \theta_{min(r_A,r_B)}$ be the canonical angles between the columnspaces of the $m \times r_A$ orthonormal matrix U_A and the $m \times r_B$ orthonormal matrix B. Let the column space of the $m \times (m - r_A)$ orthonormal matrix $U_A^\perp$ be the $(m - r_A)$ dimensional orthogonal complement of U_A and the column space of the $m \times (m - r_B)$ orthonormal matrix $U_B^\perp$ be the $(m - r_B)$ dimensional orthogonal complement of U_B. Then, there exist singular values such that:*

- $sin(\theta_i) = \sigma_i\,(U_A^t U_B^\perp) = \sigma_i\,(U_B^t U_A^\perp)$
- $tan(\theta_i) = \sigma_i\,(((U_B^\perp)^t U_A^\perp)((U_B^\perp)^t U_A))$

Proof: Trivial □

Observe that when for instance U_A is a 10×3 matrix and U_B is 10×6, then the matrix $U_A^t U_B$ will have three singular values. The matrix $U_B^t U_A^\perp$ will however have 6 singular values. The difference is easily accounted for by singular values equal to 0 and 1.

5.4.3 Intersection via a set of linear equations

We shall now demonstrate the theoretical equivalence between the intersection of the column space of two matrices and the existence of a solution of a certain set of linear equations.

The main observation

If there exists an r-dimensional intersection between the column spaces of the $m \times n$ matrix A and the $m \times p$ matrix B, then there must exist r pairs of non-zero vectors p and q of appropriate dimension such that:

$$Ap = Bq$$

such that $rank(\ A[p_1 \ldots p_r]\ ) = r$. This can be rewritten as :

$$(\ A\ \ B\ ) \begin{pmatrix} p \\ -q \end{pmatrix} = 0$$

However, some care must be taken when A and/or B are rank deficient. Hereto consider the following example:

Example:
Let a and b be 2 m-vectors with $a \neq b$.

$$A = [a \ \ -a] \qquad B = [b \ \ -b]$$

Then obviously:

$$(A\ B) \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} = 0$$

However, there is no intersection between the column spaces of A and B unless $a = b$. Observe that:

$$A \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} = B \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

which provides the zero vector as a trivial intersection.
In general, the solutions p and q in the kernel of A, resp. B will be called the *trivial solutions.*

Theorem 11 Intersection via sets of linear equations, exact data
Let A be a $m \times n$ matrix and B an $m \times p$ matrix. The r dimensional subspace that is the intersection of the columnspaces of A and B is generated by the columns of the $m \times r$ matrix C where:

1. $r = rank(A) + rank(B) - rank(A\ B)$
2. *Let $rank([A\ B]) = t$ and assume that P and Q follow from:*

$$(A\ B)\begin{pmatrix} P \\ Q \end{pmatrix} = 0$$

Then $span_{col}(C) = span_{col}(AP) = span_{col}(BQ)$ and $dim(span_{col}(C)) = r$.

Proof: The proof is a straightforward application of Grasmann's well known dimension lemma [3]. □

The theorem allows to compute the intersection via the solution of a set of linear equations which can be done via the singular value decomposition. However, observe that in order to compute a basis of $span_{col}(AP) = span_{col}(BQ)$, another SVD would be required. Especially when $m \gg max(n,p)$, this is a costly operation. Therefore, the following result allows to determine the intersection via an *economy size* smaller SVD:

Corollary 8 Economy size intersection computation.
Let A be a $m \times n$ matrix of rank $r_A \leq n$ and B an $m \times p$ matrix of rank $r_B \leq p$, with $m \geq max(n,p)$. Assume that $rank([A\ B]) = t$. Then, the column space of the $m \times r$ matrix C equals the intersection of $span_{col}(A)$ and $span_{col}(B)$ where:

- *The singular value decomposition of $(A\ B)$ is given by:*

$$(A\ B) = U\begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} V_{11}^t & V_{21}^t \\ V_{12}^t & V_{22}^t \end{pmatrix}$$

  where S is a $t \times t$ matrix, V_{11} is $n \times t$, V_{12} is $n \times (n+p-t)$, V_{21} is $p \times t$ and V_{22} is $p \times (n+p-t)$.

- *$rank(SV_{11}^t V_{12}) = r$. If the singular value decomposition of $SV_{11}^t V_{12}$ is given by:*

$$SV_{11}^t V_{12} = U_3 S_3 V_3^t$$

  where S_3 is a $r \times r$ matrix, then the matrix C can be derived as:

$$C = AV_{12}V_3$$

Proof: The SVD of $(A\ B)$ delivers an expression for AV_{12}:

$$\begin{aligned} AV_{12} &= U\begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} V_{11}^t \\ V_{12}^t \end{pmatrix} V_{12} \\ &= U\begin{pmatrix} SV_{11}^t V_{12} \\ 0 \end{pmatrix} \\ &= U\begin{pmatrix} U_3 S_3 V_3^t \\ 0 \end{pmatrix} \end{aligned}$$

That $rank(SV_{11}^t V_{12})$ provides the correct dimension r of the intersection, follows immediately from theorem 11, where it was shown that $rank(AV_{12}) = r$. □

Observe that the required SVD is that of a $t \times (n + p - t)$ matrix only, while AP is a $m \times (n + p - t)$ matrix. Especially when m is large, this economy size SVD computation may result in considerable computational savings. Note that there is also a connection to the oriented energy framework, since it holds that:

$$p^t A^t A p = p^t A^t B q = q^t B^t A p = q^t B^t B q$$

Observe however that p and q do not necessarily have the same norm. Finally, let's mention the fact that in a lot of applications, all required information can be retrieved from knowledge of the vectors p_i and q_i only, without the need to compute explicitly the intersection. This at first sight seemingly innocent observation is however extremely important when the data are perturbed by additive noise.

5.4.4 The deductive analysis of approximate intersections

If it is a priori certain that the provided data in the matrices A and B are exact, then both procedures, the canonical correlation analysis and the intersection computation via a set of linear equations, will allow to determine exactly the dimensionality of the intersection and an appropriate basis for the subspace. This can be done for instance via the singular value decomposition, up to machine precision. Observe however, that in the real $(m + n + p)$-dimensional space of all matrices A and B, the existence of an intersection between the column spaces becomes less and less *generic* for increasing overdetermination m/n and m/p. This can be most easily seen from the reformulation of the problem as the solution of a set of linear equations. If in addition the data are corrupted by noise, it is highly improbable that an intersection will even exist. However, the canonical angles in the first approach and the singular values in the second approach, may still suggest that an intersection is *not too far away*, in the sense that, if the matrices are perturbed, the corresponding column spaces might intersect again (partially). Hence one could think of trying to compute an approximate intersection by defining an appropriate approximation criterion.

While for exact data, the two methods will recover exactly (up to machine precision) the intersection if there is one, it will now be demonstrated that this equivalence disappears in favour of the second one, if the data are corrupted by additive noise. It will be assumed that the data of the exact matrices A and B are generated by a *stationary* process, which implies that their average mean and variance are approximately constant. This assumption is needed because, as will be shown, the precise choice of the dimensions of the matrices starts now playing an important role. Hence, a stationary *mechanism* (a system) is needed to provide us with new data if the matrix dimensions are enlarged. Fortunately, this is the case in all applications that are studied in this work. First, we shall analyse via the orthogonality and the lever theorem, what happens to the canonical angles and the singular values when the data are corrupted by additive noise. Second, it will be shown how a wealth of results can be derived from the optimization of a certain criterion subject to constraints.
In order to motivate our main result, consider the following example:

Example:
Assume that $\hat{A}$ and $\hat{B}$ are m-vectors with $\hat{A} = \hat{B}$. Let $\tilde{B}$ be exactly orthogonal to $\hat{B}$. Further assume that the elementwise variance of the exact process is $\sigma^2_{\hat{B}}$ and that of the noise elementwise is $\sigma^2_{\tilde{B}}$. This situation is depicted in figure 5.13.a. Under the given assumptions, the canonical angle between the vectors can be computed, as a function of m:

$$cos(\theta) = \frac{\sqrt{m}\sigma_{\hat{B}}}{\sqrt{m}\sqrt{\sigma^2_{\hat{B}} + \sigma^2_{\tilde{B}}}} = \frac{1}{\sqrt{1 + (\sigma_{\tilde{B}}/\sigma_{\hat{B}})^2}}$$

Of course, this expression is only justified theoretically *asymptotically*. One can see that the canonical angle will be determined by the elementwise signal-to-noise ratio. As a matter of fact, for general vectors $\hat{A}$ and B, this will be so with increasing probability for increasing values of m, keeping in mind the probability density functions of section 5.2.1. Moreover, as a consequence, the reached threshold is *independent of m*. In figure 5.13.c.(lower curve), one can find a simulation with Matlab's normal distribution random generator for a one dimensional intersection $\hat{a} = \hat{b} = [\ 1\ 1\ \ldots\ 1\ ]$. The components of the added noise vector $\tilde{b}$ follow a Gaussian distribution, zero mean with variance $\sigma^2_{\tilde{b}} = 1$. As predicted by the theory, the canonical angle between a and b converges to a threshold equal to 45^o.
If the dimension of the intersection is larger than 1, another effect has to be taken into account, namely *the conditioning of the exact $\hat{B}$ matrix plays a considerable role in the perturbing effect of the noise.* This is most easily demonstrated with another very simple example.

Example: Sensitivity
Consider an exact 4×2 matrix $\hat{B}$ and a perturbing orthogonal 4×2 matrix $\tilde{B}$ of which the column space is orthonormal to that of $\hat{B}$. By choosing an appropriate coordinate system, assume that the matrices take on the following from:

$$\hat{B} = \begin{pmatrix} 1 & \alpha \\ 0 & \beta \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \quad \tilde{B} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ \sigma\gamma & \sigma\delta \\ -\sigma\delta & \sigma\gamma \end{pmatrix} \quad B = \begin{pmatrix} 1 & \alpha \\ 0 & \beta \\ \sigma\gamma & \sigma\delta \\ -\sigma\delta & \sigma\gamma \end{pmatrix} \qquad \gamma^2 + \delta^2 = 1$$

Here, α and β are parameters. It is assumed that $\alpha^2 + \beta^2 = 1$. The parameter σ models the power of the noise. The condition number of the matrix $\hat{B}$ is easily verified to be:

$$\begin{aligned} K^2_{\hat{B}} &= \frac{(\alpha^2 + \beta^2 + 1) + \sqrt{(\alpha^2 + \beta^2 + 1)^2 - 4\beta^2}}{(\alpha^2 + \beta^2 + 1) - \sqrt{(\alpha^2 + \beta^2 + 1)^2 - 4\beta^2}} \\ &= \frac{1 + \sqrt{1 - \beta^2}}{1 - \sqrt{1 - \beta^2}} \end{aligned}$$

This condition number equals 1 for $\beta = 1$ and is infinite for $\beta = 0$. Let the matrix B be the sum of these two matrices $B = \hat{B} + \tilde{B}$. It is an easy exercise to show that the canonical angles between the column space of B and $\hat{B}$ are given by:

$$\theta_1 = arccos(\frac{1}{\sqrt{1 + \sigma^2}}) \qquad \theta_2 = arccos(\frac{\beta}{\sqrt{\beta^2 + \sigma^2}})$$

Observe that the first angle only depends upon the noise level. The second angle however aproaches 90^o if $\beta \to 0$.

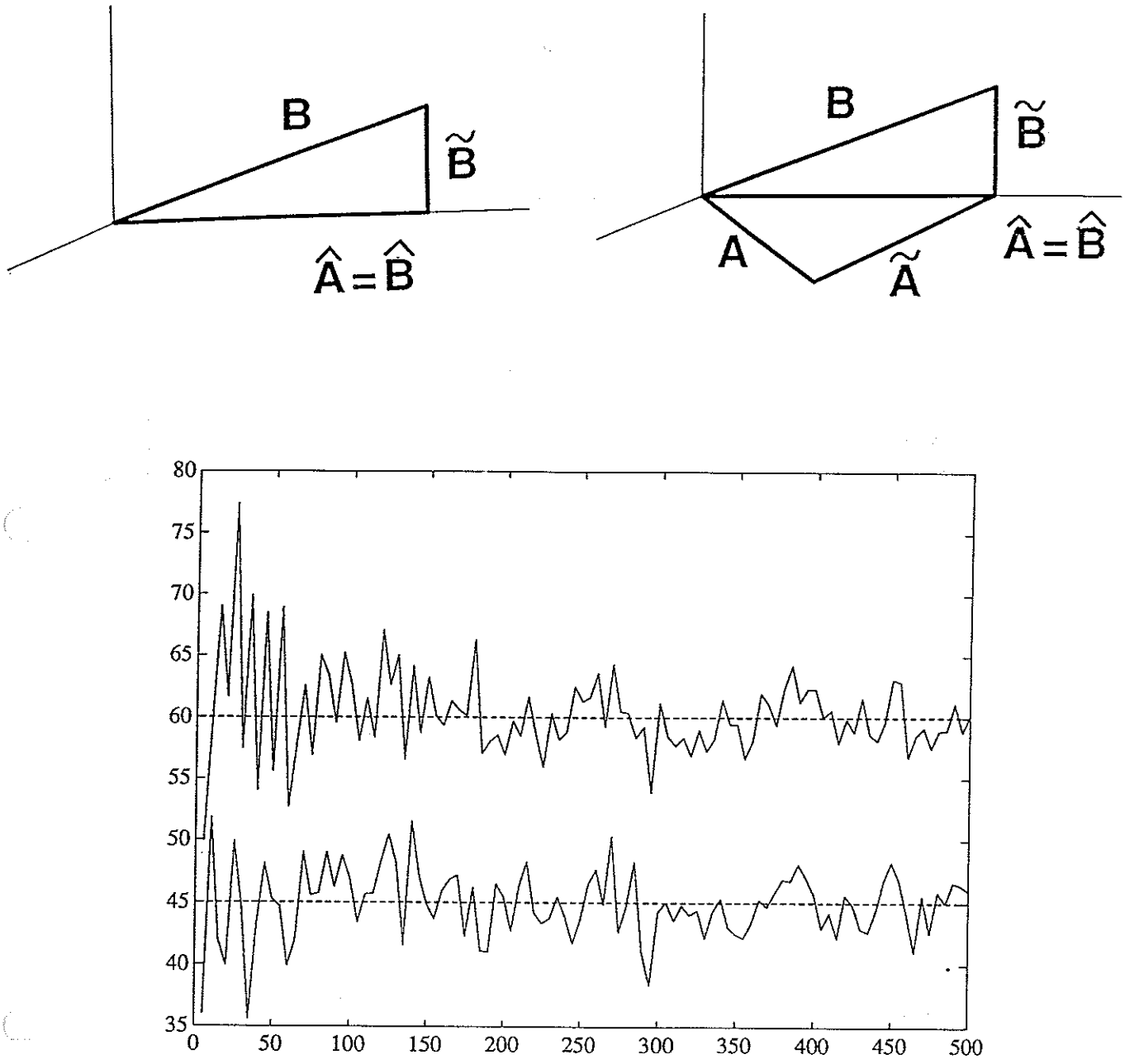


Figure 5.14: (a) Geometrical situation when one vector is noise corrupted. (b) Geometrical situation when both vectors are noise corrupted. (c) Simulation with Matlab: one vector noise corrupted (lower curve), two vectors noise corrupted (upper curve).

The next example illustrates the geometry with two noisy matrices A and B:

Example:
Assume that $\hat{A}$ and $\hat{B}$ are m-vectors with $\hat{A} = \hat{B}$. Hence the intersection is one dimensional. Further let $\tilde{A}$ be exactly orthogonal to $\hat{A}$ and $\tilde{B}$ be orthogonal to $\hat{B}$ and $\tilde{A}$. Assume that the elementwise variance of the exact process is $\sigma^2_{\hat{A}}$ and that of the noise elementwise is $\sigma^2_{\tilde{A}}$ and $\sigma^2_{\tilde{B}}$. This situation is depicted in figure 5.13.b. Under the given assumptions, the canonical angle between the vectors can be computed:

$$cos\theta = \frac{1}{(\sqrt{1+(\sigma_{\tilde{A}})^2/(\sigma_{\hat{A}})^2})(\sqrt{1+(\sigma_{\tilde{B}})^2/(\sigma_{\hat{B}})^2})}$$

As a matter of fact, keeping in mind the probability density functions of section 5.2.2, this will be so with increasing probability for increasing values of m. Moreover, as a consequence, the reached threshold is *independent of* m. In figure 5.13.c.(upper curve), one can find a simulation with Matlab's normal distribution random generator for a one dimensional intersection $\hat{a} = \hat{b} = [\,1\;1\;\ldots\;1\,]$ as a function of m, the number of components. The components of the added noise vectors $\tilde{a}$ and $\tilde{b}$ follow a Gaussian distribution, zero mean and variance $\sigma^2_{\tilde{b}} = \sigma^2_{\tilde{b}} = 1$. As predicted by the theory, the canonical angle between a and b converges to a threshold equal to 60°.

Observe that, when only one matrix is noisy and the other is noisefree, a linear least squares projection will allow to recover 'exactly' the original intersection. It is easily seen that this will be impossible when the two matrices are subject to additive noise. We are now ready to state the main result:

Theorem 12 Deductive analysis of the canonical angles between noisy subspaces. *Let $\hat{A}$ and $\hat{B}$ be exact $m \times n$ and $m \times p$ matrices of rank $r_{\hat{A}}$ and $r_{\hat{B}}$ with $m \geq max(n,p)$. Assume that the columns of the orthonormal $m \times r_{\tilde{A}}$ matrix $U_{\tilde{A}}$ form a basis for $span_{col}(\hat{A})$ and the columns of the $m \times r_{\tilde{B}}$ orthonormal matrix $U_{\tilde{B}}$ are a basis of $span_{col}(\hat{B})$. Assume that both $\hat{A}$ and $\hat{B}$ are perturbed by additive noise matrices $\tilde{A}$ and $\tilde{B}$, that are generated by a zero mean spherical density function with elementwise variances $\sigma^2_{\tilde{A}}$ and $\sigma^2_{\tilde{B}}$. Let $A = \hat{A} + \tilde{A}$ and $B = \hat{B} + \tilde{B}$. Then, as m increases, the cosines of the canonical angles between $span_{col}(A)$ and $span_{col}(B)$ will converge to the singular values of the matrix:*

$$(S^2_{\hat{A}} + m\sigma^2_{\tilde{A}} I_{r_{\hat{A}}})^{-1/2} S^t_{\hat{A}} U^t_{\hat{A}} U_{\hat{B}} S_{\hat{B}} (S^2_{\hat{B}} + m\sigma^2_{\tilde{B}} I_{r_{\hat{B}}})^{-1/2}$$

Proof: The proof follows immediately from the orthogonality theorem and the lever theorem for spherical distributions as follows: Asymptotically, it holds that:

$$\tilde{A}^t\hat{A} = 0 \qquad \tilde{A}^t\hat{B} = 0$$
$$\tilde{B}^t\tilde{A} = 0 \qquad \tilde{B}^t\hat{B} = 0$$

The result then follows immediately from corollary 1 (lever theorem for gaussian distributions). □

Observe that the canonical angles are asymptotically determined by:

- The signal-to-noise ratio
- The conditioning of the matrices $\hat{A}$ and $\hat{B}$.

In statistical literature, a lot of research has been invested in removing the *bias*, expressed by the theorem, from the 'sample' canonical correlations [5, p.23]. However, our geometrical approach in fact shows that the deviation from the 'true' values is really *irreversible*, due to the very fact that the *long space of a matrix is irreversibly lost!.*

Let's now have a glance at the behavior of the singular values of the matrix $[A\ B]$. If the intersection is r-dimensional, the matrix $[\hat{A}\ \hat{B}]$ has at least $(rank(\hat{A})+rank(\hat{B})-rank([\hat{A}\ \hat{B}])$ singular values equal to zero. Because of the additive noise $\tilde{A}$ and $\tilde{B}$, the intersection will be lost almost surely. Under the assumptions we have made on the generating mechanism of the data contained in the matrices $\hat{A}$ and $\hat{B}$ and of the assumptions on the noise matrices $\tilde{A}$, $\tilde{B}$, the prediction of the behavior of the singular values as a function of m follows immediately from the orthogonality and the lever theorem.
As a general result, the singular values that ought to be zero because of an existing intersecting subspace in the case of exact data, now increase as a function of m, according to a $\sqrt{m}$ law, as predicted by the orthogonality theorem and the lever theorem.
In summary,

- The bias on the canonical angles will considerably complicate the correct *a posteriori* estimation of the dimension of the intersection.
- Especially for bad signal-to-noise ratios, a singular value analysis of the matrix $[A\ B]$ may help in deciding about the dimensionality of the intersection.

5.4.5 Computation of the intersection

In the preceding discussion, we have mainly provided the necessary tools to estimate the dimension r of the intersecting subspace. It will now be shown how to compute the intersection via the optimization of a certain criterion.

Let A be an $m \times n$ and B be an $m \times p$ matrix with $m \geq max(n,p)$. Then an approximate r-dimensional intersection can be defined as the column space of the $m \times r$ matrix C that optimizes:

$$min_{C,P,Q}\ \ \gamma\|C - AP\|_F^2 + (1-\gamma)\|C - BQ\|_F^2$$

subject to the rank constraints $rank(C) = rank(AP) = rank(BQ) = r$, where P is an $n \times r$ matrix and Q is an $p \times r$ matrix and γ provides a certain weighting of both terms.

The following remarks are in order:

- The criterion is intuitive in this sense, that one makes linear combinations of the columns of A and B and then tries to recover that column space C which is closest to these subspaces.
- The weighting, represented by γ, can be applied in case that one has more confidence in one of the two matrices which is for instance more exact than the other. Of course, if there is no such a priori information, simply set $\gamma = 1/2$. We shall derive our results for this case, obviously without loss of generality.

- Note that the rank condition on C is crucial since it will allow to deflate *trivial* solutions. The precise determination of a meaningful r can be done via the deductive analysis of the preceding section or via the results to be presented below.
- Observe that the proposed criterion is consistent from the point of view of the inspirationist: If the data are exact, the exact intersection will be recovered.

We shall now show how in addition, the matrix C can be computed once P and Q are known. The rest of the section will then be devoted to the computation of the matrices P and Q, subject to various constraints.
Observe that the criterion is equivalent with the following expression:

$$min_{C,P,Q}\ \ trace(2C^tC + P^tA^tAP + Q^tB^tBQ - C^tAP - P^tA^tC - C^tBQ - Q^tB^tC)$$

The trace operator only involves the diagonal elements of a matrix.

In order to avoid repeating in every statement the same notations, the following conventions apply from now on throughout:

- A is an $m \times n$ matrix of rank r_A. B is an $m \times p$ matrix of rank r_B. Everywhere $m \geq max(n,p)$. The matrix C is an $m \times r$ matrix of rank r. P is an $n \times r$ matrix and Q is $p \times r$. r is the dimension of the approximate intersection.
- Because the generalized singular value decomposition allows for an elegant formulation of our results, it will be needed to define a generalized SVD of the matrix pair $[A,B]$. Hereto, expand the matrix with the least number of columns with zeros until it has $max(n,p)$ columns. Hence, without loss of generality it may be assumed that *both A and B are $m \times n$ matrices*. Then, a GSVD of $[A,B]$ can be defined as:

  $$A = U_A D_A X^t \qquad B = U_B D_B X^t$$

  Here U_A is an $m \times r_A$ orthonormal matrix, U_B is $m \times r_B$ orthonormal, D_A is $r_A \times n$ diagonal, D_B is $r_B \times n$ diagonal, X is $n \times n$.
- The canonical angles and corresponding vectors between the column spaces of A and B can be computed from the singular value decomposition of $U_A^t U_B$:

  $$U_A^t U_B = RST^t$$

  The matrix R is $r_A \times r_A$ orthonormal, T is $r_B \times r_B$ orthonormal. The $r_A \times r_B$ diagonal matrix S contains the cosines of the canonical angles. The first r of these cosines and the corresponding left and right singular vectors are denoted by R_r, S_r, T_r.
- The diagonal matrices Γ and Λ are matrices of Lagrange multipliers throughout.

5.4.6 A heuristic approach.

Observe that, apart from the rank constraints, the minimization of the criterion reduces to the solution of a (mixed) total linear least squares problem of the form:

$$min_{P,Q}\ \ \|(A\ B)\begin{pmatrix} P \\ -Q \end{pmatrix}\|_F^2$$

This suggests the following heuristic procedure:

1. If both A and B are noisy, compute the SVD of $(A\ B)$ and obtain P and Q from the total linear least squares solution. If A is exact and B noisy, first perform a QR-factorization, and then apply the mixed linear least squares - total linear least squares solution/strategy discussed in lemma 1, in order to find the matrices P and Q.

2. Obtain the matrix C as $C = (AP + BQ)/2$ in case of noisy A and B, and as $C = AP$ in case of exact A and noisy B.

3. Since P and Q are formed from singular vectors of the matrix $(A\ B)$, they satisfy $P^tP + Q^tQ = I$. However, this is not sufficient for AP,BQ or their sum to be of a rank, equal to the (approximate) intersection dimension. In order to deflate the trivial solutions, one could:

   (a) Either compute the SVD of C and from its singular values estimate the dimension of an intersection. This can be done via the economy size SVD approach provided by corollary 8. Trivial solutions will be revealed by small singular values of the matrix C.

   (b) Either first deflate the trivial solutions from a singular value decomposition of the matrices A and B. A small singular value of either A or B indicates near rank deficiency, which is the cause of a trivial solution.

   The difficulty obviously arises from what is to be considered as small for a singular value.

While the described procedure is based on heuristics, it will be useful in obtaining additional insight in the computation of approximate intersections via the criterion optimization procedures with constraints described in the next sections. The reason is that in these methods, the dimension decision is based upon the canonical angles, which may be difficult because of the inherent *biasedness* of the estimates of the canonical angles in case of noisy data (deductive approach).

5.4.7 Least squares intersection.

If one of the 2 matrices is exact, it is obvious that the solution will be contained, at least implicitly, the orthogonal projection of the column space of the noisy matrix onto the column space of the exact matrix. This is formalized in the following theorem.

Theorem 13 Approximate Least Squares Intersection.
Assume that the $m \times n$ matrix A is exact, and the $m \times n$ matrix B is noisy. The optimization of the criterion:

$$min_{P,Q} \|BQ - AP\|_F^2$$

subject to

$$C^tC = \Delta$$

where Δ is a positive definite $r \times r$ diagonal matrix and $C = AP$, leads to an approximate r-dimensional intersection given by

$$C = AA^+BQ = U_A R_r \Delta^{1/2}$$

where $Q = X^{-t} D_B^+ T_r S_r^{-1/2} \Delta^{1/2}$.

Proof: Assume that the matrix Q is known. Then it is not too difficult to see that the minimization of the criterion reduces to the solution of r linear least squares problems, each in one of the columns $p^i, i = 1, \ldots, min(n, p)$, the solution of which can be summarized as:

$$P = A^+ BQ$$

Hence, the criterion reads:

$$min_Q \|(AA^+ - I_m)BQ\|_F^2$$

Observe that $AA^+ = U_A U_A^t$. In order to eliminate trivial solutions, it is ensured that $C^t C = \Delta = Q^t B^t U_A U_A^t BQ = P^t A^t AP$. Introducing a diagonal matrix Γ of Lagrange multipliers, differentiation with respect to each column of Q and setting the result equal to zero, results in the following generalized eigenvalue problem:

$$B^t(U_A U_A^t - I_m)BQ = B^t U_A U_A^t BQ\Gamma$$

which, using the GSVD and the canonical correlation decomposition reduces to:

$$\begin{aligned} B^t BQ &= B^t U_A U_A^t BQ(I - \Gamma) \\ XD_B^t D_B X^t Q &= XD_B^t U_B^t U_A U_A^t U_B D_B X^t Q(I - \Gamma) \\ XD_B^t(D_B X^t Q) &= XD_B^t T S^t S T^t (D_B X^t Q)(I - \Gamma) \end{aligned}$$

Obviously, the solution of this generalized eigenvalue problem consists of singular vectors contained in T. First, choose Q to be equal to $Q = X^{-t} D_B^+ T$ and substitute it in the optimization criterion:

$$\begin{aligned} \|AP - BQ\|_F^2 &= \|U_A U_A^t BQ - BQ\|_F^2 \\ &= \|U_A RST^t D_B X^t X^{-t} D_B^+ T - U_B D_B X^t X^{-t} D_B^+ T\|_F^2 \\ &= \|U_A RS - U_B T\|_F^2 \\ &= S^t S + I - S^t R^t U_A^t U_B T - T^t U_B^t U_A RS \\ &= S^t S + I - S^t S - S^t S \\ &= I - S^t S \end{aligned}$$

Obviously, the criterion is minimized if $Q = X^{-t} D_B^+ T_r$. When substituting this solution in the constraint, one finds:

$$Q^t B^t U_A U_A^t BQ = T_r^t T^t S^t S T T_r = S_r^2$$

Hence, the solution Q must be scaled appropriately in order to find:

$$Q = X^{-t} D_B^+ T_r S_r^{-1} \Delta^{1/2}$$

□

Example:
Let the k-th elements of the columns of the $m \times 5$ matrix $\hat{A}$ be given as: $a_k^1 = sin(0.2k)$, $a_k^2 =$

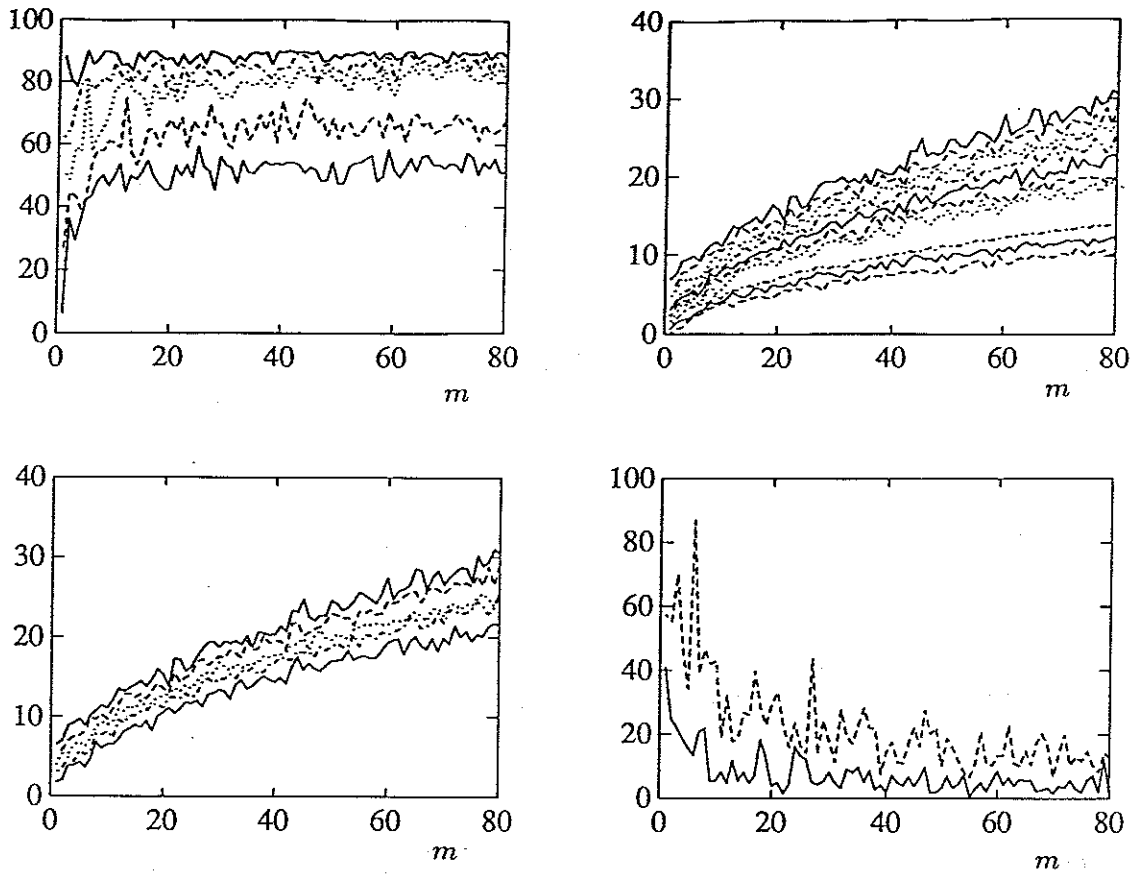


Figure 5.15: (a) Canonical angles between $span_{col}(\hat{A})$ and $span_{col}(B)$. (b) Singular Values of $[\hat{A}\ B]$. (c) Singular values of $\hat{B}$. (d) Canonical angles between $span_{col}(\hat{A})$ and $span_{col}(C)$.

$sin(0.3k)$, $a_k^3 = sin(0.4k)$, and a^4 and a^5 are generated as random noise with zero mean and variance 1. The matrix $\hat{B}$ consists of 5 columns, the first of which is generated as $b^1 = 0.5a^1 + 0.4a^2$, the second column as $b^2 = 0.8a^1 - 0.6a^2$, the third one is $b^3 = sin(0.5k)$ while columns b^4 and b^5 are again zero mean random noise with variance 1. Obviously, the intersection of the columnspace is of dimension two, as soon as $m \geq 3$. A noise matrix $\tilde{B}$ is added to $\hat{B}$. Its elements are zero mean from a gaussian distribution with variance 1. The canonical angles between $\hat{A}$ and B are depicted in figure 5.14.a. as a function of m while the singular values of $[\hat{A}\ B]$ as a function of m are found in figure 5.14.b. Observe that the matrix $\hat{B}$ is well conditioned. Its singular values as a function of m are depicted in figure 5.14.c. Hence, the canonical angles will reach a threshold which can be predicted by theorem 12. The intersection is estimated via the least squares approach as $C = U_{\hat{A}} U_{\hat{A}}^t BQ$. The canonical angles between $span_{col}(C)$ and $span_{col}(\hat{A})$ as a function of m are depicted in figure 5.14.d.

As a general conclusion for the least squares mechanism, one can state the following:

If only one matrix is corrupted by noise with a spherical directional density, the linear least squares method succeeds in computing an intersection, which is consistent.

Observe that the solution remains exactly similar if part of the matrix A and B are noisefree. Assume that the first n_1 columns of A and B are noisefree. When partitioned in an obvious

way, the optimization criterion becomes:

$$min_{P_1,P_2,Q_1,Q_2} \| [A_1 \; B_1] \begin{pmatrix} P_1 \\ -Q_1 \end{pmatrix} + [A_2 \; B_2] \begin{pmatrix} P_2 \\ -Q_2 \end{pmatrix} \|_F^2$$

which leads to a completely similar solution approach as derived above.

5.4.8 Total least squares intersection

One may be a little bit surprised about the seemingly trivial formulation of the general conclusion for the least squares intersection computation. However, it will now be demonstrated that the total linear least squares approach does *not* succeed in computing the correct intersection, even if all necessary assumptions to apply successfully the method for solving sets of linear equations are fulfilled!. As a matter of fact, when all data are subject to additive noise, *the loss of intersection is irreversible!* Somewhat loosely speaking, one could state that when both A and B are perturbed by noise, one should solve the intersection problem via total linear least squares. In this section, it will be shown how one can optimize a certain criterion in order to define a reasonable approximate intersection, from an *inspiration* point of view. However, the deductionist will be disappointed: *Even if it is certain that there is an intersection between two spaces, with exact data, it is impossible to find it back, even asymptotically, when all data are noise corrupted.* This follows immediately from the orthogonality theorem, the lever theorem and the discussion on the identity matrix approach. As an example, consider the following thought experiment:

Example:
Assume that an exact pair of vectors $\hat{a} = \hat{b}$ is perturbed by noise such that the result equals $a = \hat{a} + \tilde{a}$, $b = \hat{b} + \tilde{b}$. We shall now demonstrate that if all *reasonable* conditions are satisfied, it is impossible to find back the original intersection. In other words, there exist a infinite number of solutions that satisfy the conditions :

$$\tilde{a}^t\hat{a} = 0 \quad \tilde{b}^t\hat{b} = 0 \quad \tilde{b}^t\tilde{a}^t = 0 \quad \tilde{a}^t\tilde{a} = \tilde{b}^t\tilde{b}$$

The first two conditions express the orthogonality of noise and exact data, the third one is the orthogonality of the noise variables themselves and the last condition states the noise oriented energy is isotropic. As a numerical example, consider in 5 dimensions with 5 coordinates x_1, x_2, x_3, x_4, x_5, the exact vector $\hat{a} = \hat{b} = (0, 4, 0, 0, 0)$. The noise vectors are $\tilde{a} = (2, 0, 0, 0, 0)$, $\tilde{b} = (0, 0, 2, 0, 0)$. However, the modeller only observes the sums $a = (2, 4, 0, 0, 0)$ and $b = (0, 4, 2, 0, 0)$. The interested reader may wish to verify that all 'exact' vectors, that are compatible with the conditions are lying on an ellipsoid with equation:

$$\frac{(x_2 - 32/9)^2}{1/9} + x_4^2 + x_5^2 = (4/3)^2$$
$$x_1 = 8 - 2x_2 = x_3$$

Take for instance $x_4 = 4/3$, $x_5 = 0$ them $x_2 = 32/9$. The model for the exact vector is $\hat{a} = (8/9, 32/9, 8/9, 12/9, 0)$, for the noise vectors $\tilde{a} = (10/9, 4/9, -8/9, -12/9, 0)$, $\tilde{b} = (-8/9, 4/9, 10/9, -12/9, 0)$. Even if the noise variance were known a priori (i.e. the value of $\tilde{a}^t\tilde{a} = \tilde{b}^t\tilde{b}$), there would exist an infinite number of solutions.

Theorem 14 Total Linear Least Squares Approximate Intersection
Given an $m \times n$ matrix A and an $m \times n$ matrix B. An approximate r-dimensional intersection is generated by the columns of the matrix C:

$$C = \frac{1}{2}(AP + BQ)$$

where the $n \times r$ matrices P and Q follow from the optimization of the criterion:

$$min_{P,\,Q} \|AP - BQ\|_F^2$$

subject to the constraint:

$$C^t C = \Delta$$

where Δ is a $r \times r$ positive definite diagonal matrix (necessarily of rank r). The solutions are given by:

$$\begin{aligned} P &= \sqrt{2} X^{-t} D_A^+ R_r (I_r + S_r)^{1/2} \Delta^{1/2} \\ Q &= \sqrt{2} X^{-t} D_B^+ T_r (I_r + S_r)^{1/2} \Delta^{1/2} \end{aligned}$$

Proof: The expression for the matrix C follows from:

$$min_{\,P,Q,C} \|C - AP\|_F^2 + \|C - BQ\|_F^2$$

When this expression is differentated with respect to C and the result is set equal to zero, one easily finds the expression for C. When this is inserted in the general criterion, it follows that the optimization problem reduces to:

$$min_{P,Q} \|AP - BQ\|_F^2$$

Observe that the criterion is equivalent to:

$$min_{P,Q} trace(P^t A^t AP + Q^t B^t BQ - P^t A^t BQ - Q^t B^t AP)$$

subject to:

$$P^t A^t AP + Q^t B^t BQ + P^t A^t BQ + Q^t B^t AP = \Delta$$

Introducing a diagonal Lagrange multiplier matrix Γ results in the following coupled generalized eigenvalue problem:

$$\begin{aligned} A^t AP - A^t BQ &= (A^t AP + A^t BQ)\Gamma \\ B^t BQ - B^t AP &= (B^t BQ + B^t AP)\Gamma \end{aligned}$$

This can be rewritten as:

$$\begin{aligned} A^t AP(I - \Gamma) &= A^t BQ(I + \Gamma) \\ B^t BQ(I - \Gamma) &= B^t AP(I + \Gamma) \end{aligned}$$

Employing the GSVD of the matrix pair $[A, B]$, results in:

$$\begin{aligned} D_A^t (D_A X^t P)(I - \Gamma) &= D_A^t U_A^t U_B (D_B X^t Q)(I + \Gamma) \\ D_B^t (D_B X^t Q)(I - \Gamma) &= D_B^t U_B^t U_A (D_A X^t P)(I + \Gamma) \end{aligned}$$

Upon using the canonical correlation SVD of $U_A^t U_B = RST^t$, first set $P = X^{-t}D_A^+R$ and $Q = X^{-1}D_B^+T$. Substitute this into the criterion to derive that it is equal to:

$$trace(2(I - S))$$

Obviously, the optimizing matrices are given by:

$$P = X^{-t}D_A^+R_r \quad Q = X^{-t}D_B^+T_r$$

Substituting these expressions in the constraints expressions, results in the diagonal matrix:

$$2(I_r + S_r)$$

Obviously, this requires an adaptation of the solutions P and Q into:

$$\begin{aligned} P &= \sqrt{2}X^{-t}D_A^+R_r(I_r + S_r)^{1/2}\Delta^{1/2} \\ Q &= \sqrt{2}X^{-t}D_B^+T_r(I_r + S_r)^{1/2}\Delta^{1/2} \end{aligned}$$

□

Let's emphasize the following facts:

- From a deduction point of view, the intersection provided by the matrix C is *not* consistent. It does not deliver the unbiased estimate, even not asymptotically. However, the matrices P and Q will approach asymptotically, their correct counterparts $\hat{P}$ and $\hat{Q}$ where $\hat{A}\hat{P} = \hat{B}\hat{Q}$. Hence, the deductionist should concentrate all necessary information about the model in the *short space*, and not in the *long space*.
- However, as a most remarkable fact, also the inspirationist is bound to respect certain conditions. Of crucial importance, is the normalization requirement $C^tC = \Delta$.
  - Without this condition, the solution would be trivially $P = 0$ and $Q = 0$. This proves the necessity of normalization conditions.
  - With *this* condition, the solution reduces to a generalized total linear least squares problem, of which it is known that its solution is consistent if the data are exact. If for instance one chooses as a normalization condition $P^tP = I_r = Q^tQ$, the scheme does *not* provide the correct, intersection, even with exact data! As an example, consider the problem of computing the intersection between the column spaces of $\hat{A}$ and $\hat{B}$ where:

    $$\hat{A} = \begin{pmatrix} 1 & 0 \\ 0 & \alpha \end{pmatrix} \quad \hat{B} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \alpha \neq 0$$

    via the solution to $\hat{A}p = \hat{B}q$ where p and q are 2-vectors. The correct intersection is of course $\hat{B}$ itself, and this is also the solution found by total linear least squares. However, if other normalization conditions are imposed, such as:

    $$p^tp = 1 = q^tq$$

    it is easily verified that there is no solution unless $\alpha = 1$. Hence, methods with $P^tP = I_r = Q^tQ$ as normalization condition are not consistent, even from the inspiration point of view!

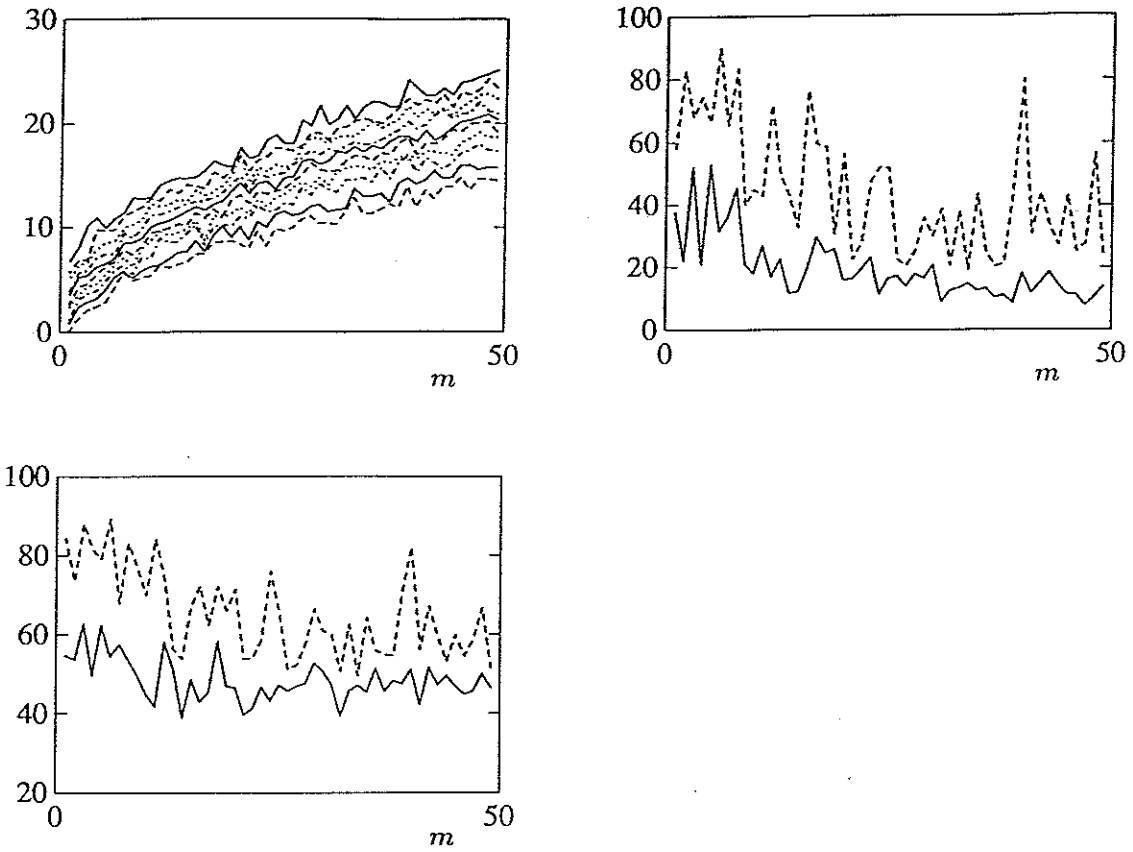


Figure 5.16: (a) Singular values of $[A\ B]$. (b) Canonical angles between $span_{col}(P)$ and $span_{col}(\hat{P})$. (c) Canonical angles between $span_{col}(C)$ and $span_{col}(\hat{A})$.

- Observe that in some applications, the data in the matrix $\hat{A}$ might be more sensitive to noise than those in B. Hence one could analyse the problem by applying some *weighting*. An example of this will be given in chapter 8, where it will be shown how the sensitivity depends upon the power spectrum of the signals.
- Note how the (generalized) singular value decomposition allows to solve most elegantly the seemingly difficult coupled generalized eigenvalue problem, and this in a numerically reliable way! It is easy to derive that the matrix of Lagrange multipliers is given by:

$$\Gamma = (I - S)(I + S)^{-1}$$

Example:
Consider the same example as in the case of linear least squares, but now add also noise to the matrix $\hat{A}$. The elements of $\tilde{A}$ are zero mean normally distributed with variance 1. The singular values of $[A\ B]$ are depicted in figure 5.15.a. The canonical angles between $span_{col}(P)$ and $span_{col}(\hat{P})$ are depicted in figure 5.15.b. The canonical angles between $span_{col}(C)$ and $span_{col}(\hat{A})$ can be found in figure 5.15.c.

5.4.9 Intersection via optimization of the RV-coefficient.

In the recent multivariate statistics literature, a new measure of similarity between two sets of random variables was introduced, called the RV-coefficient ([21] and the references therein). We shall shortly review this new measure and then show that as a matter of fact, *there is*

nothing new to it, since it essentially reduces to the solution we have been deriving in the previous section. Our approach is instructive, since the demonstration is constructive by employing the generalized singular value decomposition. Moreover, it emphazises the close connection between our oriented energy framework developed in chapter 4 and the notion of canonical angles.

Definition 6 Normalized oriented energy.
Consider an $m \times n$ matrix A, $(m \geq n)$ representing m vectors in an n-dimensional space. Then the normalized oriented energy distribution of the matrix A^tA is the oriented energy distribution of the matrix $A^tA/(\sqrt{trace((A^tA)^2)})$.

Insight in the normalized oriented energy distribution is most easily obtained from the singular value decomposition of the matrix $A = U_A S_A V_A^t$ with the singular values σ_i:

$$\frac{A^tA}{\sqrt{trace((A^tA)^2)}} = \frac{\sigma_1^2 + \ldots + \sigma_n^2}{\sqrt{\sigma_1^4 + \ldots + \sigma_n^4}}$$

Obviously, the normalized oriented energy distribution of any *isotropic* $m \times n$ vector sequence, coincides with the oriented energy distribution of the matrix $\sqrt{n}I_n$, hence showing the square root law of the lever theorem.

Definition 7 The RV-coefficient.
The RV-coefficient between two $m \times n$ matrices A and B, $m \geq n$ is defined as:

$$RV(A,B) = \frac{trace(A^tBB^tA)}{\sqrt{trace((A^tA)^2)trace((B^tB)^2)}}$$

It is easily seen that the RV-coefficient is a measure for the difference in normalized oriented energy of the two matrices, if the inner product between two matrices A and B is defined by $trace(A^tB)$:

$$\begin{aligned} \| \frac{A^tA}{\sqrt{trace((A^tA)^2)}} - \frac{B^tB}{\sqrt{trace((B^tB)^2)}} \|^2 &= 2(1 - \frac{trace(A^tBB^tA)}{\sqrt{trace((A^tA)^2)trace((B^tB)^2)}}) \\ &= 2(1 - RV(A,B)) \end{aligned}$$

In the special case of 2 vectors a and b with an angle θ, the RV-coefficient is straightforward to compute:

$$\| \frac{aa^t}{\sqrt{tr(aa^t)^2}} - \frac{bb^t}{\sqrt{tr(bb^t)^2}} \|^2 = 2(1 - cos^2(\theta))$$

Some simple-to-prove properties of $RV(A,B)$ are:

1. $RV(A,B) = RV(B,A)$.
2. $RV(A,B) = RV(AP,BQ)$ where P and Q are square orthonormal matrices.

From the preceding discussion, it is straightforward to formulate the computation of an approximate intersection between the column spaces of the matrices A and B as the maximization of the RV-coefficient between linear cominations of columns of A and B.

Theorem 15 Approximate intersection via RV-coefficient optimization.
An approximate r-dimensional intersection between the column spaces of two $m \times n$ matrices A and B, can be computed from the maximization of the RV-coefficient $RV(AP, BQ)$ over all possible $n \times r$ matrices P and Q, subject to the constraints:

$$P^t A^t A P = \Delta_A \qquad Q^t B^t B Q = \Delta_B$$

where Δ_A and Δ_B are $r \times r$ diagonal positive definite matrices. The solutions are:

$$P = X^{-t} D_A^+ R_r \Delta_A^{1/2} \qquad Q = X^{-t} D_B^+ T_r \Delta_B^{1/2}$$

Before deriving the proof, let us state the following remarks:

- One can take as the intersection either the column spaces of the matrix AP or BQ. The constraints ensure that these matrices will be of rank r. In our identification approach of chapter 8, this will correspond to fixing a forward (future) and backward (past) state predictor space. However, observe that the constraints do not necessarily guarantee that the average sum $(AP + BQ)/2$ is not rank deficient.
- Instead of requiring the matrices Δ_A and Δ_B to be diagonal positive definite, one could replace them also with general, not necessarily diagonal, positive definite matrices. However, this case is easily reduced via e.g. a Choleski decomposition preprocessing step to the diagonal case. Moreover, the diagonality ensures the most pure linear indepence between the columns of AP and BQ (uncorrelatedness in a statistical framework).
- The reader should verify the remarkable ressemblance between the solution presented in theorem 15 and that of theorem 14. Yet, the optimized criteria and the constraints are quite different. One can also check that the coupled generalized eigenvalue problem in the proof below differs from the coupled generalized eigenvalue problem in the proof of theorem 14.

Proof: Observe that by the constraints, one can get rid of the denominators of the criterion in the RV-coefficient. Introducing two diagonal Lagrange multipliers matrices Γ and Λ and differentiating with respect to the columns of P and Q results in the following coupled generalized eigenvalue problem:

$$\begin{aligned} A^t B Q Q^t B^t A P &= A^t A P \Lambda \\ B^t A P P^t A^t B Q &= B^t B Q \Gamma \end{aligned}$$

Upon substition of the GSVD of the matrix pair $[A, B]$, one easily finds:

$$\begin{aligned} X D_A^t U_A^t U_B (D_B X^t Q)(D_B X^t Q)^t U_B^t U_A (D_A X^t P) &= X D_A^t (D_A X^t P) \Lambda \\ X D_B^t U_B^t U_A (D_A X^t P)(D_A X^t P)^t U_A^t U_B (D_B X^t Q) &= X D_B^t (D_B X^t Q) \Gamma \end{aligned}$$

Insert the canonical correlation SVD of $U_A^t U_B = RST^t$ and first set $(D_B X^t Q) = T$ and $(D_A X^t P) = R$. Obviously:

$$\begin{aligned} X D_A^t R S S^t &= X D_A^t R \Lambda \\ X D_B^t T S S^t &= X D_B^t T \Gamma \end{aligned}$$

Hence, corresponding columns of T and R solve the problem provided that $\Lambda = \Gamma = SS^t$. So we find that:

$$P = X^{-t}D_A^+R \qquad Q = X^{-t}D_B^+T$$

When these expressions are inserted in the numerator of the optimization criterion, its value equals:

$$s_1^2 + \ldots + s_n^2$$

Here $s_i, i = 1, \ldots, n$, are the singular values of S. They are the cosines of the canonical angles between the column spaces of A and B. Obviously, the optimal solutions are given by:

$$P = X^{-t}D_A^+R_r\Delta_A^{1/2} \qquad Q = X^{-t}D_B^+T_r\Delta_B^{1/2}$$

The scaling with the diagonal matrices arises from the constraints, but does not change anything substantially to the optimization procedure. □

In case that the diagonal matrices $\Delta_A = \Delta_B = I_r$. the solution reduces to the classical canonical variate analysis as it is met in the statistical literature [5, p.15]. For the purpose of completeness, we summarize an alternative demonstration, which clarifies the well known canonical structure of the covariance matrix:

$$\begin{pmatrix} A^tA & A^tB \\ B^tA & B^tB \end{pmatrix} \longleftrightarrow \begin{pmatrix} I_n & S \\ S & I_n \end{pmatrix}$$

This canonical decomposition can be directly obtained from the GSVD of the matrix pair $[A, B]$ and the SVD of the matrix $U_A^tU_B = RST^t$:

$$\begin{pmatrix} A^tA & A^tB \\ B^tA & B^tB \end{pmatrix} = \begin{pmatrix} XD_A^tR & 0 \\ 0 & XD_B^tT \end{pmatrix} \begin{pmatrix} I_n & S \\ S & I_n \end{pmatrix} \begin{pmatrix} R^tD_AX^t & 0 \\ 0 & T^tD_BX^t \end{pmatrix}$$

Under the constraints $P^tA^tAP = I_r = Q^tB^tBQ$, the optimal solution is easily find to be $P = X^{-t}D_A^+R_r$ and $Q = X^{-t}D_B^+T_r$.

5.4.10 Unnormalized computation of the intersection.

The several approaches that have been discussed in the preceding sections, all lead sooner or later to an analysis of the canonical angles between the subspaces. However, as could be concluded from the deductive analysis of the *biasedness* of the canonical angles, it can be rather difficult to estimate the exact dimension of the intersection. The reason is that the canonical correlation coefficients squared (the squared cosines of the canonical angles), represent *the energy shared by a pair of* **unit** *directions in both spaces.* However, because of the inherent normalization that is performed in canonical correlation analysis, a pair of vectors with the smallest canonical angle, does not necessarily represent a significant *enrgy* contribution in both spaces. Loosely speaking, *high energetic and low energetic contributions are treated on the same basis* whenever canonical correlation analysis is invoked to compute an approximate intersection.

In order to take into account the relative energy contributions, we will now propose yet another criterion for the computation of an appropriate intersection. The source of inspiration

is some recent work of Arun and Kung [12] and Ramos and Verriest [21]. Moreover, it will be demonstrated how the solution can also be interpreted as a special case of the RV-coefficient optimization philosophy.

Let A and B be $m \times n$ matrices with $m \geq n$. Instead of using the GSVD of the matrix pair (A, B), we will employ the generalized singular value decomposition of the matrix pair (A^t, A^tB) given by:

$$A^t = XD_AU_A^t \quad A^tB = XD_{AB}U_{AB}^t$$

It is assumed that the vectors of X are ordered corresponding to the diagonal elements of $D_{AB}D_A^+$ in decreasing order. Recall that the pseudo-inverse of 0 equals 0. Denote by D_r the diagonal matrix corresponding to the r largest diagonal elements of $D_{AB}D_A^+$, by D_{Ar} and D_{ABr} the diagonal matrices with corresponding diagonal elements of D_A and D_{AB}, by X_r the matrix with the corresponding columns of X and by U_{ABr} the corresponding r columns of the matrix U_{AB}. The matrix $Y = X^{-t}$ while Y_r corresponds in the same way to the appropriate generalized singular values of the matrix pair (A^t, A^tB).

The proposed intersection is the best possible subspace, contained in the column space of A, that minimizes the criterion:

$$min_{P,Q} \ \|BQ - AP(P^tA^tAP)^{-1}P^tA^tBQ\|_F^2$$

subject to the constraints:

$$P^tA^tAP = \Delta_A \quad Q^tB^tBQ = \Delta_B$$

where Δ_A and Δ_B are positive definite diagonal matrices. The rationale behind this criterion is the following: $AP(P^tA^tAP)^{-1}BQ$ is nothing else than the orthogonal projection of the columns of BQ onto the column space of AP. BQ represents the best possible approximation of the intersection in the column space of B, while AP is the best possible approximation in the column space of A.

The main result is summarized in the following theorem:

Theorem 16 Unnormalized computation of the intersection.
The solution P that minimizes the criterion:

$$\|BQ - AP(P^tA^tAP)^{-1}P^tA^tBQ\|_F^2$$

subject to $P^tA^tAP = \Delta_A$, $Q^tB^tBQ = \Delta_B$, both positive definite diagonal, is given by:

$$P = Y_rD_{Ar}^{-2}D_{ABr}$$

and the matrix Q follows from the solution of the generalized eigenvalue problem:

$$U_{ABr}D_r^2U_{ABr}^tQ = B^tBQ\Lambda$$

Proof: Observe that the given constraint for Q implies that the matrix BQ is orthogonal. Hence, the criterion reduces to:

$$max_{P,Q} \ trace(Q^tB^tAP\Delta_A^{-1}P^tA^tBQ)$$

Assume that Q is fixed and a differentation with respect to P results in the following expression, in terms of the GSVD:

$$P = X^{-t} D_A^{+} Z \Delta_A^{1/2}$$

where Z is still an arbitrary matrix satisfying $ZZ^t = I_r$. Substitution in the criterion and using the expression for A^tB in terms of the GSVD, results in:

$$Q^t B^t A P \Delta_A^{-1} P^t A^t B Q = Q^t U_{AB} D_{AB} D_A^{+} Z Z^t D_A^{+} D_{AB} U_{AB}^t Q$$

Obviously, since Q is still undetermined, this expression is maximized for $Z = I_r$ which results in the expression for the optimal matrix P. The criterion now reads:

$$Q^t U_{AB_r} D_r^2 U_{AB_r}^t Q$$

Together with the constraint for Q, this results in the generalized eigenvalue problem for Q upon introducing a diagonal matrix of Lagrange multipliers Λ. □

The following remarks are in order:

- Observe that no canonical angles are involved.
- The decision for the dimension of the approximate intersection follows from inspection of the generalized singular values of the matrix pair (A^t, A^tB).
- The proposed solution computes an approximate intersection which is contained in the column space of A. The reason is the explicit least squares projection of BQ on AP. Of course, a similar result could also be found in the column space of B by optimizing the criterion:

  $$min_{P,Q} \ \|AP - BQ(Q^tB^tBQ)^{-1}Q^tB^tAP\|_F^2$$

  subject to the same constraints. However, still further research is needed to propose a criterion that finds a kind of average intersection of the form $(AP + BQ)/2$.

The following example demonstrates that these generalized singular values allow for a decision based upon relative energy contribution.

Example:

Consider two cosines $a_1(k) = 5cos(0.8 * k)$ and $a_2(k) = 5cos(0.3 * k)$. A new signal is generated as a convex combination of these signals of the form $a = a_1\alpha + a_2(1 - \alpha)$ where $0 \leq \alpha \leq 1$. With this signal, one constructs a 40×6 Hankel matrix $\hat{A}$. Two noise corrupted versions A and B are constructed from this Hankel matrix by adding random noise, zero mean with standard deviation 0.01 to the exact signal. Observe that the signal-to-noise ratio is high, but the point we want to demonstrate is the difference between the generalized singular values of the pair (A^t, A^tB) (figure 6.16.a) and the canonical angles (figure 6.16.b). As can be seen from these figures, the generalized singular values take into account the relative energy of the two modes while the canonical angles normalize both modes. In figure 6.16.a., the 4 largest generalized singular values are depicted (the 2 others are relatively small), while in figure 6.16.b., one can see that the four smallest canonical angles are all very close to 0^o. The 2 others, caused by the noise, are close to 90^o as could be expected from the *lever theorem*

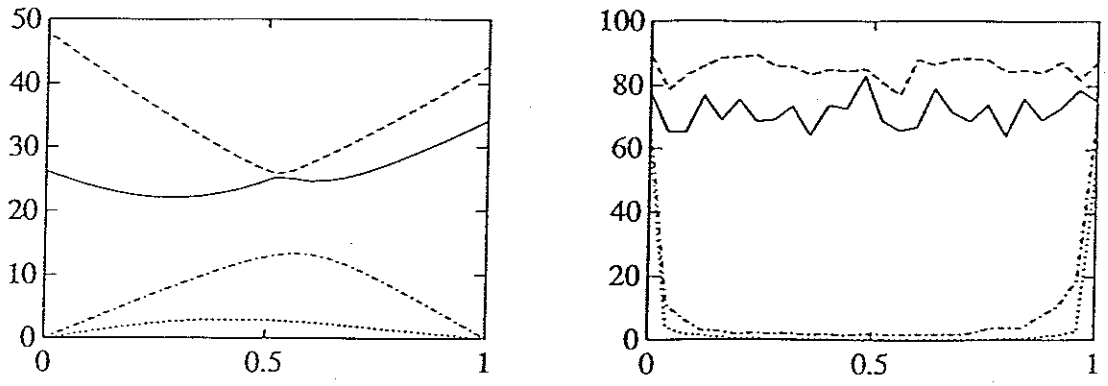


Figure 5.17: Generalized Singular Values of the matrix pair (A^t, A^tB) (a) and the canonical angles between the column spaces (b), as a function of α.

for gaussian distributions.

Let's now demonstrate that the preceding result is nothing else but a special case of the RV-coefficient:

Corollary 9 *The optimizing solution P of*

$$RV(AP, B) = \frac{trace(P^tA^tBB^tAP)}{\sqrt{trace((P^tA^tAP)^2)trace((B^tB)^2}}$$

subject to the constraint:

$$P^tA^tAP = \Delta_A$$

where Δ_A is an $r \times r$ positive definite diagonal matrix, is given by:

$$P = X_r^{-t} D_{Ar}^{+} \Delta^{1/2}$$

Proof: Trivial from the GSVD of the matrix pair (A^t, A^tB). □

This demonstrates that the RV-coefficient approach, allows to unify all optimization strategies, whether they first perform a *normalization* of the data, such as is the case for canonical correlation analysis, or whether they use an *unnormalizing* criterion, such as the result of theorem 16.

5.4.11 A computational consideration.

At first sight, the computation itself of the GSVD seems to be a technical objection against a flexible use of the obtained insights, especially when the matrices A and B are very rectangular (much more columns than rows), as will be the case for our identification approaches to be presented in chapter 8. However, observe that a QR-decomposition preprocessing step can

reduce considerably the amount of required flops.
Hereto, consider the QR-decomposition:

$$(A\ B) = Q_{AB} \begin{pmatrix} R_1 & R_2 \\ 0 & R_3 \end{pmatrix}$$

When the solution is presented in terms of the GSVD of the matrix pair (A, B) as was the case in subsections 6.4.7, 6.4.8, 6.4.9:

$$A = U_A D_A X^t \quad B = U_B D_B X^t$$

then, compute the following GSVD on matrices of much smaller dimension if A and B are 'very' rectangular:

$$\begin{pmatrix} R_1 \\ 0 \end{pmatrix} = U_1 D_1 Y^t \quad \begin{pmatrix} R_2 \\ R_3 \end{pmatrix} = U_2 D_2 Y^t$$

Obviously:

$$A = (QU_1) D_1 Y^t \quad B = (QU_2) D_2 Y^t$$

equals the GSVD of the matrix pair (A, B). If the solution is presented in terms of the GSVD of the matrix pair $(A^t, A^t B$ as was the case in subsection 6.4.10, then: compute the GSVD:

$$(R_1^t\ 0) = Y D_1 U_1^t \quad R_1^t R_2^t = Y D_2 U_2^t$$

then obviously:

$$A^t = Y D_1 (QU_1)^t \quad A^t B = Y D_2 (QU_2)^t$$

is the GSVD of the matrix pair $(A^t, A^t B)$.

5.5 Conclusions

The results obtained in this section can be divided into four main conclusions:

1. First, we have dwelt upon the close connection between linearity, orthogonality and the *definition* of noise, as absence of linear relations. Two different points of view concerning mathematical modelling were highlighted: the deduction and the inspiration approach.

2. Via a statistical perturbations analysis it was shown how additive noise influences the singular value decomposition of a matrix. Two important geometrical results were derived: the orthogonality and the lever theorem.

3. Several identification schemes for detecting linear relations in noisy data have been analysed and discussed with respect to the model of the linear relations and the noise, mainly from a deductive point of view. A unification was presented between the linear and total linear least squares identification approach. It was shown how the provided noise model is completely unsatisfactory.

4. Finally, it was shown how in the case of noisy data, one can define a certain criterion that unifies all existing strategies to compute an approximate intersection between 'long' spaces of matrices. The generalized singular value decomposition plays an important role. However, in most applications the computed intersection will be irreversibly biased.

One has to concentrate all required information about the system in the 'short' space. We have also discussed a criterion, which takes into account explicitly the oriented energy of the data, without pre-normalization as is the case with all aproaches based upon canonical angles between subspaces.

Bibliography

[1] Akaike H. *Stochastic Theory of Minimal Realization.* IEEE Trans. on Automatic Control, Vol.19, pp.667-674, 1974.

[2] Anderson T.W. *The 1982 Wald memorial lectures: Estimating Linear Statistical Relationships.* The Annals of Statistics, 1984, Vol.12, No.1,pp.1-45.

[3] Gantmacher F.R. *Theorie des Matrices.* Tome 1/2, Dunod, Paris, 1966.

[4] Gelfand I.M., Yaglom A.M. *Calculation of the amount of information about a random function contained in another such function.* Amer. Math. Soc. Trans., Series (2), Vol.12., pp.199-246, 1959.

[5] Gittins R. *Canonical Analysis; A review with applications in Ecology.* Biomathematics Volume 12, Springer Verlag, Berlin, 1985.

[6] Golub G. *Matrix Decompositions and Statistical Calculations.* Statistical Computation, R.C. Milton And J.A. Nelder, eds., New York: Academic Press, pp.365-397, 1969.

[7] Golub G., Van Loan C. *Matrix Computations.* John Hopkins University Press, North Oxford Academic, 1983.

[8] Hotelling H. *Relations Between Two Sets of Variates.* Biometrika, Vol.28, pp.321-372, 1936.

[9] Kailath T.*A view of three decades of linear filtering theory.* IEEE Trans. Information Theory, IT-20, 2, pp.146-181, 1974.

[10] Kalman R.E. *Identification from real data.* Lecture presented at the 25th Anniversary Symposium of Econometric Institute, Erasmus University, Rotterdam, The Netherlands, January 15, 1982.

[11] Kovanic P. *A New Theoretical and Algorithmical Basis for Estimation, Identification and Control.* Automatica, Vol.22, N0.6, pp.657-674, 1986.

[12] Arun K.S., Kung S.Y. *Generalized Principal Component Analysis and its Application in Approximate Stochastic Realization.* in 'Modelling and Applications of Stochastic Processes.', Eds: U.B. Desai., Kluwer Academic Publishers, 1986, pp.75-104.

[13] Larimore W.E. *System Identification, Reduced Order Filtering and Modeling via Canonical Variate Analysis.* Proc. of the 1983 American Control Conference, June 22-24, San Francisco, California, USA, 1983.

[14] Lawson C.L., Hanson R.J. *Solving Least Squares Problems.* Prentice Hall, Inc. Englewood Cliffs, New Jersey, 1974.

[15] Ledermann W. *On the rank of the reduced correlational matrix in multiple factor analysis.* Psychometrika, Vol.2, No.2, June, 1937.

[16] Ljung L. *System Identification - Theory for the User.* Prentice Hall, Engelwood Cliffs, NJ, 1987.

[17] Luenberger D. *Introduction to Dynamic Systems, Theory, Models and Applications.* John Wiley and Sons, New York, 1979.

[18] Luenberger D.G. *Optimization by vector space methods.* John Wiley and Sons, Inc., New York, 1969.

[19] Norton J.P. *An introduction to identification.* Academic Press, London, 1986.

[20] Papoulis A. *Probability, Random Variables and Stochastic Processes.* McGraw-Hill Series in Electrical Engineering, McGraw-Hill Inc, 1984.

[21] Ramos J.A., Verriest E.I. *A unifying tool for comparing stochastic realization algorithms and model reduction techniques.* Proc. of the 1984 Automatic Control Conference, San Diego, California, June 1984, pp.150-155.

[22] Stigler S.M. *Gauss and the invention of least squares.* The Annals of Statistics, 1981, Vol.9, No.3, pp.465-474.

[23] Vandenberghe L., Van Mieghem P. *Een Nieuw Algoritme voor de Identifikatie van Lineaire Multivariabele Systemen.* Master Thesis, ESAT, Departement Elektrotechniek, Katholieke Universiteit Leuven, 1987, UDC:681.5.015(043).

[24] Van Huffel S. *Analysis of the total least squares problem and its use in parameter estimation.* Doctoral Thesis, Department of Electrical Engineering, Katholieke Universiteit Leuven, June 1987.

[25] Wilkinson J. *The Algebraic Eigenvalue Problem.* Oxford University Press, 1965.

[26] Willems J.C. *From data to models.* Automatica. Part I: vol.22, no.5, pp.561-580, 1986. Part II:vol.22, no.6, pp.675-694, 1986. Part III: Vol.23, no.1, pp.87-115, 1987.

A view of paradigms in science

Chapter 6

The Uncertainty Principle of Mathematical Modelling

Allez en avant, la foi vous viendra
Jean Le Rond d'Alembert

6.1 Problem formulation

6.1.1 The Philosophy of Identification

Identification of mathematical models from observations and measured data is one of the enduringly central problems in most scientific disciplines. It has profound implications in all branches of applied sciences. Mathematical models are employed for purposes of predicition, control, simulation, and as (black box) analysis tools to gain deeper insight in phenomena of which the human intellect did not yet perveive a fine understanding. However, as Willems remarks in [39]:

> Notwithstanding the fact that identification theory and time series analysis have produced some very useful algorithms and important applications, it can be stated that there is a need of putting a clear and rational foundation under the problem of obtaining models from time series. It is very much of an area where some of the first principles still need to be sorted out. In particular, one should, in our opinion, start by formalizing what is meant by an optimal (approximate) model.

In this chapter, one will find as the key result a confirmation of this statement:

Identification is a matter of definition

As one overlooks that part of system theory which is commonly denoted as 'System Identification', one can not avoid the impression that it consists of a *bag of tricks* within a *Fiddler's Paradise* [4]. The usual approach is to postulate a set of equations or equivalently, a model structure containing yet some unspecified parameters (e.g. AR, ARX, (D) ARMAX,CARIMA, NARMAX models). Most models (if not all) are unable to explain the observations exactly. The philosophy of statistics and probability theory, summarized in Kolmogorov's brilliant framework, then serves to guarantee that every (finite) observed data set can occur with a certain probability. Deviations between model and data are now claimed to

be caused by statistical sampling and tons of literature have been devoted to this problem, resulting in the desirable properties of unbiasedness, efficiency, confidence intervals, tests of significance etc.... Moreover, in a lot of cases the model is validated and concluded to be 'true', on the basis of assumptions which are a priori unverifiable and certainly not inspired by the available data. However, one may not forget that it is usually (tacitly) assumed that the data are generated by some simple probabibilistic mechanism, a distribution law, that can be described by a single distribution function, which on its turn can be quantified by some *summarizing* quantities. For instance, the most extreme (simple) case being the Gaussian distribition. A crucial hypothesis is that there is a 'true' value that can be recognized to be a particularly striking feature of this distribution function: its mean, median or variance etc .... Stochastic models will almost always be compatible with almost every reasonable set of observations, if the a prior assumptions are made appropriately. However, as Willems notes in [40] :" *...in most applications the regularity - the stability of the relative frequency of occurrence of the disturbance terms implied by the stochastic nature of a model is difficult, if not impossible to justify.*" Let's add a citation from Kalman [23]:

> For an objective outsider, much of the historical development of statistics is a long series of attempts to dodge the inevitable implications of uncertainty. Whenever the conventional statistical treatment of a problem gives a unique (certain) answer, as in maximum likelihood, least squares,..., common sense should tell us that such a miracle is possible only if additional assumptions (deus ex machina) are imposed on the data which somehow succeed in neutralizing the intrinsic uncertainty. We shall use the technical term "prejudice" for such assumptions. In other words, statistical methodology has been handicapped because statisticians have become mesmerized by the deep seated hope of giving certain answers to problems where the uncertainty is intrinsic. This is politics, not science.

Hence good models are such that:

Uncertain data imply uncertain models.

But what are uncertain data? The classical appoach to uncertainty suffers from à misleading antropocentrism: Not the model we chose is wrong, but the mismatch is due to several causes that corrupt the data: measurement noise, randomness and related side effects such as (statistical) sampling etc.... However, it is hard to realize that maybe one consciously uses a model whose structure is unable to capture the complexity of the phenomenon under consideration. Hence, it is not Nature, the data, that are to be blamed. Instead, we ourselves are to be blamed since we simplify too much reality by proposing simple models and then hope that these will describe satisfactory Nature's complex behavior. However, since models and laws are postulated by the modeller, they are neither obtained by deduction nor induction, but by inspiration. It is only later, that a black box modeller may find that the data contain some mechanism that explains them.

Generically and in general, the model will never exactly explain the available data. However, since the model is really put forward voluntarily by the user, at the same time, it defines the noise environment, the uncertainties that are not explained by the model. This postulated noise environment must of course be reasonable in terms of what is a priori known about the data generation process. But a lot of identification recipes create a noise environment that

can simply not be accepted as reasonable : think of linear least squares, total linear least squares which cause rank 1 noise models. The conclusion of this discussion is simple but deep and farreaching:

Noise is not caused by external effects, it is postulated

Only by postulating a specific noise environment and then applying the postulate to specific data, one can deduce quantitative information concerning the noise and the data. Stated in simple terms, noise is what is left unexplained by the postulated model. Hence, fixing the model and identification scheme, is at the same time defining the noise or what is to be considered as noise !

Let's summarize for clarity the conclusions of the preceeding discussion:

1. **Identification is a matter of definition**
2. **Uncertain data imply uncertain models.**
3. **Noise is defined once the model is defined**

6.1.2 The mathematics of linear modelling

The basic question that will be treated in this chapter concerns *linear static models.* It witnesses of an amazing (yet misleading) simplicity.

Can observed values of a finite family of variables be 'explained' by some underlying linear relations between the variables ?

Using elementary linear algebra, the answer is trivial when the data are noisefree, but the noisy problem is highly non-trivial both from the conceptual as from the mathematical point of view. *Noise* may mean one or all of many things: inaccuracy of the model, measurement errors, unknown effects, non-linearities (when dealing with linear models), any causal or random factors which cannot be modeled, of which no further information is available,.... This indicates that the origin of the *noise* may not be clear: Is it due to our ignorance of infinitely precise data or equivalently, is 'noise' the manifestation of our lack of complete information which is caused by the limited precision of our measurement equipment ? Or are the phenomena that we can observe in Nature so complex that they cannot be modeled by something as simple as linear relations, although good approximations might exist? The first question corresponds to the classical view of descriptive modelling : Nature operates consistently according to some universal laws and ours is the task to discover these. However, as is indicated in [39], the cornerstone of the philosophy of science is *falsification* rather than deduction. *Models and laws are postulated,* based on criteria like simplicity, esthetic appeal or by parallels with other disciplines and applications and it is only later that one finds out that they could also have been deduced from already existing knowledge. Therefore, the following statement is fundamental in that it really defines what is meant by noise (in any application !):

Noise = what is not explained by the model.

Hence, once the class of models has been fixed, at the same time, the notion of noise is well defined. If it is the user's desire to model the phenomenon under study by *linear relations* that are to be discovered in the data, one has immediately :

Noise is absence of linear relations.

Remains the question : How to define linear relations? First, observe that linearity is not really a question of fact nor of evaluation, but a self-imposed limitation on the types of operations or devices that are to be used [8]. One of those tools that are to be chosen is the metric. Many mathematicians believe that they have freedom in the choice of a metric for their mathematical model. This is true for pure mathematics (where the choice of norm can be dictated by pure 'intellectual' motivations such as for instance solvability of the problem) but it may no longer be true for the mathematical modelling of real processes. In a lot of cases, there are physical invariants that imply the use of a certain metric (as an example consider the theory of special relativity with its indefinite metric). Taking into account the necessary invariance principles that a metric should satisfy, there are strong indications that for the purpose studied in this paper - the identification of linear relations from noisy data -, the ordinary Euclidean metric is appropriate [30] as defined in the usual way: If x, y are real vectors with components x_i, y_i then the inner product is the real number

$$x_1 y_1 + \ldots + x_n y_n$$

Two vectors are orthogonal if their inner product equals 0 and the geometrical concept of an angle can be defined. Formally, one can now state that :

Orthogonality = absence of linear combinations

When 2 vectors are not orthogonal, at least part of one of them can be explained to be 'proportional' to the other. Observe however, that there may be convincing reasons to prefer a certain metric with respect to another. Such a preference could for instance be motivated by a priori information about the geometrical distribution of the noise in the space of unknowns, as could follow from an analysis of the oriented energy of a disturbing variable. Hence, *orthogonality* is to be understood not in the physical, Euclidean sense (the daily life orthogonality) but in the sense of its mathematical meaning via an inner product. However, *mutatis mutandis*, such an a priori knowledge can easily be dealt with in the present framework, hence, without loss of generality, we restrict ourselves to the *Euclidean inner product.*

Let us now consider the problem of identifying linear relations between measured data. Suppose n variables are measured over m time instants. The measurements are aggregated in a $m \times n$ matrix A. It is assumed that the number of measurements exceeds the number of variables: $m > n$ so that the matrix A has more rows than columns. This assumption of overdetermination is of course necessary since otherwise the problem would be trivial. An existing linear relation will reveal itself via an n-vector x that belongs to the kernel of the matrix A:

$$Ax = 0$$

The number of independent linear relations is indicated by the algebraic rank r of A. The *corank* of A is defined as $n - rank(A)$. The corank equals the number of linearly independent relations between the variables. Now when the data are really measurements on some multichannel phenomenon, generically there will be no linear relations between the data :

$$rank(A) = n \quad \text{or} \quad corank(A) = 0$$

A fundamental assumption is that this noise corrupts the data in an additive way (this of course is not only a matter of taste, but also of simplicity). The measured matrix A can then be written as:

$$A = \hat{A} + \tilde{A}$$

where "^" denotes the obtained model of the data and "~" denotes the obtained model of the noise. The noise variables cannot be linearly related with each other. Because if they did, they would satisfy linear relations, hence contradicting the very definition of the noise as absence of linear relations. Hence :

$$\tilde{A}^t \tilde{A} = \text{diagonal}$$

and $ker(\tilde{A}) = 0$. There can also exist no linear relation between the noise and the exact data, hence :

$$\hat{A}^t \tilde{A} = \tilde{A}^t \hat{A} = 0$$

Define the measured, exact and noise Grammians as:

$$\Sigma = A^t A \quad \hat{\Sigma} = \hat{A}^t \hat{A} \quad \tilde{\Sigma} = \tilde{A}^t \tilde{A}$$

Obviously, Σ is positive definite and $\hat{\Sigma}$ is nonnegative definite. $\tilde{\Sigma}$ is diagonal with nonnegative diagonal elemenents. The mathematical problem formulation of the identification of linear relations reduces to :

The Frisch scheme:

Given a positive definite $n \times n$ matrix Σ. Find all nonnegative diagonal matrices $\tilde{\Sigma}$ and all n-vectors x such that:

1. $\hat{\Sigma} = \Sigma - \tilde{\Sigma}$ is nonnegative definite
2. corank($\hat{\Sigma}$) is maximal
3. $\hat{\Sigma} x = 0$

The maximisation of the corank is necessary since of course one is interested in *the maximum number of linear relations*, which are described by the vectors x. From now on, the maximal achievable corank, which is part of the solution of the Frisch scheme, is denoted by **maxcor**. The corresponding minimal rank is abbreviated as **minr**. The minimization of the rank should be compared to the more conventional minimizations of certain norms, in commonly used identification approaches. As it will be shown, it is precisely this 'unconventional' criterion which makes the whole difference, both in conceptual interpretation as in mathematics. As a matter of fact, the maximal corank is the only objectively identifiable invariant of the problem.

There is a slight technical condition that is to be satisfied: There is a lower bound on the number of data if one does not want to exclude a priori certain coranks. Since the pure noise matrix $\tilde{A}$ is of full column rank n and must be orthogonal to the column space of $\hat{A}$, which is r dimensional, it is easy to see that:

$$m - r \geq n$$

This results in the fact that:

$$minr(\Sigma) \leq (m-n)$$

Since

$$1 \leq r \leq n-1$$

one finds that:

$$(2n-1) \leq m$$

Maximisation of the corank, will lead to:

The uncertainty principle of mathematical modelling:

Uncertain Data imply Uncertain Models.

At first sight, this looks trivial: Any statistician will subscribe its contents. However, our approach will turn out to be considerably different from the statistical one, where one usually pre-determines a certain norm criterion, which is to be optimized. In a lot of cases, additional assumptions are imposed (such as uncorrelatedness, maximum likelihood,...) in order to simplify the mathematics involved. What will be found is a model, which is essentially *unique* but on which the uncertainty is characterized by some probability distributions, the precise character of which is of course largely influenced by the a priori (unverifiable) assumptions. In our approach, the uncertainty on the models will not have a probabilistic character. Instead, the uncertainty will be simply a matter of *mathematics*, not of assumptions.

Another motivation for the maximisation of the corank is that this maximisation reflects any modeller's preference for *simple* models. Hereto observe that the corank is nothing else than the number of linearly independent relations that exist between the data. Hence, when the corank is high, the data are explained by a lot of linear relations. Mathematically, linear relations reduce to vectors in the kernel of the exact data matrix $\hat{\Sigma}$. However, the more there are vectors in the kernel, the less there are exact data matrices of which the row space will be orthogonal to these linear relations. Hence the simpler will be the behavior. Observe that this corresponds to a minimisation of the complexity when translated in the modelling framework of Willems [39].

For several reasons, the above problem is called the 'Frisch scheme' in honour of the 1969 Economics Nobel prize winner Frisch [17]. Of course, the problem looks very similar to what happens in statistics, where also concepts such as uncorrelatedness and statistical orthogonality are exploited to facilitate the identification problem. In fact, it was Wold who used the idea of regarding random variables as elements of a metric space with the distance between two elements as the variance of their difference. This geometric interpretation made it natural to interprete least squares estimation as projection onto a subspace. In the framework of stochastic processes as introduced by Kolmogorov [22], it is possible to split a process in a unique deterministic process and one that can be written as linear combinations (moving average) of a white noise process, which itself is a process with uncorrelated components. Both processes are uncorrelated. This is the so-called *Wold decomposition* and it may be interesting to mention that Wold was influenced by the work of Frisch [22]. There is indeed a big similarity between the characteristics of the Frisch scheme and the Wold decomposition. For instance, the Wold decomposition is restrictive since only *linear* regressions on past data

are considered. For processes generated in some non-linear way, the Wold decomposition may be artificial and disguise the essential simplicity of the process. In fact, it is possible to have processes that are purely non-deterministic in the sense of the Wold decomposition but which can be predicted *exactly* by a suitable nonlinear prediction [20]. Does this limitation not originate in the antropocentristic preference of the modeller or the statistician for *linear models?*
The Frisch scheme provides a conceptual framework, based upon the Euclidean inner product, that takes explicitly into account the double finiteness of the number of data (finite number of sample realizations and finite number of data in one realization). Hence, no so-called statistical 'sampling' problems occur. Moreover, we won't use any extra assumptions whatsoever, unless the very few ones stated previously. Observe that these assumptions have lead us to a mathematical 'second order' problem, where everything is stated in terms of Grammians. This second order statement, reminds of another modelling stragy referred to in statistics as *factor analysis* [3] [6]. This problem originates in the work of Spearman, a psychologist from the beginning of this century, who was interested in characterizing quantitatively general intelligence. In a famous paper [38], he used the *hierarchy of correlations* as a criterion. It was Spearman's ambition to explain human intelligence by as few factors as possible. We hope that at the end of this chapter, the reader will understand why he failed. An historical review with references can be found in [6].

This chapter is organized as follows:

- First, some classical identification schemes will be discussed in terms of their properties with respect to the Frisch scheme (section 2). The necessary mathematical insights have already been obtained in chapter 5. Attention will be paid to the properties that may be helpful in characterizing the maximal corank. Attention will be paid to both all diagonal matrices $\tilde{\Sigma}$ and all vectors x that fulfill the requirements of the Frisch scheme. It will be discussed how the 'volume' of the polyhedral cones that characterize the solution sets, are measures of the incompatibility of the data with the imposed model. Moreover, the polyhedral cones serve as a characterization of the fundamental non-uniqueness of the solutions to the problem.

- In section 3, we discuss a new geometrical framework in which the requirements of the Frisch scheme are discussed one by one. New geometrical notions are introduced such as orthant and null invariance. It is expected that this framework will lead to a general solution of the Frisch scheme, which up to date, remains unsolved in its full generality.

- A major drawback (if it really is one, it could also be considered as a natural limitation) of the Frisch scheme, will be discussed in section 4. As a function of n, there is an upper bound (the Wilson-Ledermann bound) on the generically achievable maximal corank of the problem, even when some other information as for instance the singular values strongly suggest a higher corank. This solves at the same time some important stability questions.

In order to conclude this introduction, it is worth mentioning that we find the story that is told in this chapter, exciting for several reasons:

1. At first sight, the problem looks very simple. This apparent simplicity is provoking and invites for a quick analysis. However, soon one finds out that it is highly intractable.

As a matter of fact, it is still unsolved in its full generality. Kalman even claims in a 1983 paper [24]:

> It is impossible to avoid the conclusion that the lack of progress on and the present confusion surrounding Frisch's ideas are due to mathematical rather than conceptual difficulties.

2. The problem has a long history, starting from the beginning of this century. Among the actors of the play are famous mathematical engineers such as Kalman, ... and celebrated econometricians, such as Frisch and Koopmans (both Nobel prize winners Economics).

3. The problem is at the heart of what could be called linear modelling. The problem itself acts like a tunnel where several scientific disciplines such as psychometry, econometry, statistics, mathemetical engineering and others, meet. It leads to deep methodological, conceptual and philosophical insights.

6.2 A Historical Review.

This section is organized as follows. In subsection 6.2.1., we describe the two-variable case, which provides some nice insights that are however of little use for the general case. Subsection 6.2.2. describes the history of the so called 'communalities', which consists of attempts to solve the problem via algebraic manipulations. Subsections 6.2.3. summarizes the results that exist about an extreme case, namely the **maxcor**$= n-1$ case. The linear least squares solutions are reviewed in 6.2.4., while 6.2.5. situates the identity matrix approach. In 6.2.6, we summarize some well known results on the **maxcor**$= 1$ case while in 6.2.7, we investigate an existence theorem for the **maxcor**> 1 case.

6.2.1 The 2 variable case

For 2 variables, the problem reduces to characterizing all nonnegative real numbers $\tilde{\sigma_1}$ and $\tilde{\sigma_2}$ such that the difference matrix:

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} - \begin{pmatrix} \tilde{\sigma_1} & 0 \\ 0 & \tilde{\sigma_2} \end{pmatrix}$$

is nonnegative definite and its corank is maximized. This problem is algebraically treatable because of the limited number of variables involded. If there are no trivial linear relations, we shall have that

$$\sigma_{11}\sigma_{22} > \sigma_{12}^2$$

Obviously, whenever $\sigma_{12} = 0$, there are no linear relations between the variables because in that case the two columns of the data matrix A are orthogonal to each other. Hence, let's analyse the case with $\sigma_{12} \neq 0$. This requirement however ensures that **maxcor**=1. Hence, setting the determinant of the difference matrix to zero, we find that:

$$(\sigma_{11} - \tilde{\sigma_1})(\sigma_{22} - \tilde{\sigma_2}) = \sigma_{12}^2$$

Since it is also required that the 'exact' diagonal elements are nonnegative:

$$\hat{\sigma}_{ii} = \sigma_{ii} - \tilde{\sigma}_i \geq 0$$

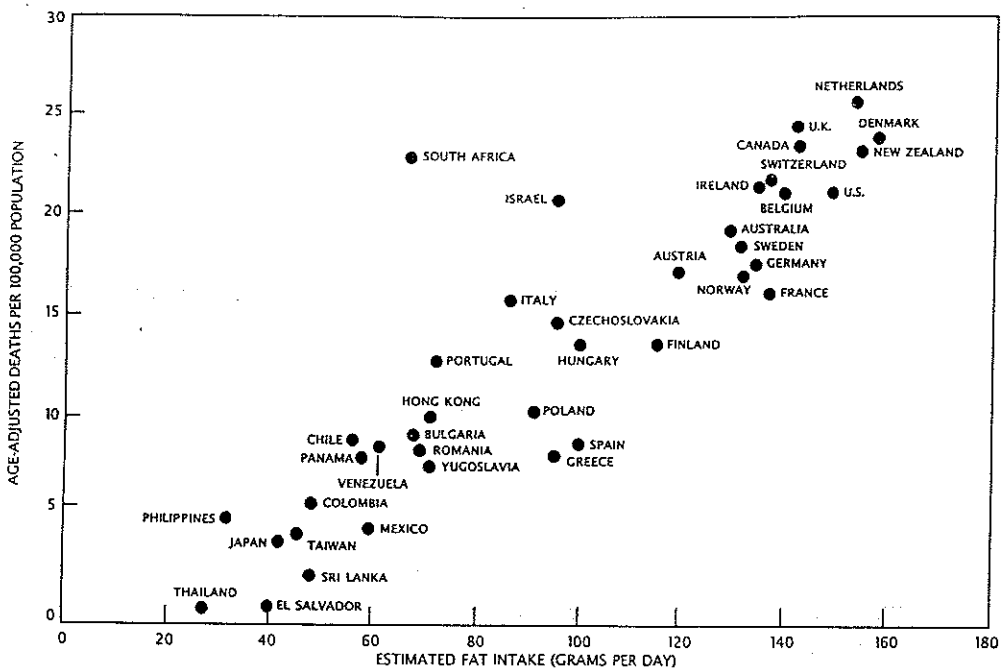


Figure 6.1: Fat intake (abscis) versus cancer deaths (ordinate) (from Scientific American, november 1987)

we find that the allowed exact diagonal elements are lying on the closed segment of a hyperbola bounded by the lines $\hat{\sigma}_{ii} = \sigma_{ii}$. If the relation between the variables (the columns of the 'exact' data matrix $\hat{A}$) is written as: $\hat{a}_1 = \hat{a}_2\alpha$, we find that:

$$\frac{\mid \sigma_{12} \mid}{\sigma_{11}} \leq \mid \alpha \mid \leq \frac{\sigma_{22}}{\mid \sigma_{12} \mid}$$

The following observations are crucial and will be generalized furtheron:

- The solution of the problem is *non-unique*. All coefficients α satisfying the given inequalities, are compatible with the requirements of the Frisch scheme. Hence, we have a first confirmation of our uncertainty principle without requiring any statistics!

- The bounds on the solutions α are the classical linear least squares solutions for the problem. Any solution in between these two solutions is allowed! This implies that the correct 'regression line' is located between the 2 elementary least squares regression lines.

- Stated somewhat vaguely, the uncertainty is 'proportional' to the distance between the least squares solutions. The reason is of course that data can not be explained by something as simple as a linear relation. Observe that we don't bother (how should we ever know?) about the real causes of this deviation from linearity. We only *postulate* a linear model and then observe how the model is falsified by the data. In order to illustrate this, consider figure 6.1. Depicted is the estimated occurence of breast cancer versus the estimated fat intake for several countries. The only conclusion that can be drawn from this picture is that there is a correlation between the two. However, one may not assume a causal relationship as both may be symptons of an underlying unmeasured variable.

6.2.2 Communalities

Generalizing the approach of the previous section, several attempts have been undertaken in the past, to solve the requirements of the Frisch scheme via the analysis of the determinant of the data matrix Σ, when the diagonal elements are changed. This is the so called analysis of the *communalities*. Mathematically, the problem is equivalent to the study of the function:

$$\mathbf{minr}(\Sigma) = \min\{rank(\Sigma - \tilde{\Sigma}) \mid (\Sigma - \tilde{\Sigma}) \text{ and } \tilde{\Sigma} \text{ NND}\}$$

Several authors have made contributions to the algebraic analysis of forementioned problem. However, since a survey of these results, starting with the work of Albert [1] in the forties, is of minor importance for our purpose, the interested reader is referred to the survey contained in [6]. An important genericity result about the communalities, will be treated in section 6.4. As is not so difficult to see, the analysis of the problem via the investigation of the communalities that are such that all conditions are satisfied, leads to algebraic manipulations of all kinds of formulas like determinantal conditions, nonnegativeness conditions, proving independency of (quadratic) equations etc.... Therefore, instead of following this (already) explored path, we prefer to turn our attention to the largely unexplored *geometrical* counterpart of the problem, which, instead of analysing the communalities, concentrates on the allowed solution vectors x.
However, the following result, which was derived in [5], is worth mentioning because it is appealing by its simplicity and elegancy and its generalizes the result on the hyperbola of the 2 variable case. [1]
The problem considered in [5] is the following:

Determine the subset χ_1, i.e. the family of all the noise covariance matrices $\tilde{\Sigma}$ leading to an exact data covariance matrix $\hat{\Sigma} = \Sigma - \tilde{\Sigma}$ with corank (mind: not maximal corank !) equal to 1.

A numerical algorithmic procedure to obtain the points of χ_1 with any required degree of accuracy is described in [5, p.5.]. However, it is the following result we would like to mention, together with its proof:

Theorem 1 *χ_1 is a continuous and convex hypersurface.*

Proof : Obviously, χ_1 can be parametrized by n nonnegative reals $\tilde{\sigma}_i$, $i = 1, \ldots, n$. Consider in this n-dimensional space the intersection of χ_1 with the plane $\chi_{n,n-1}$, parallel to the coordinate plane associated to the last two reals $\tilde{\sigma}_{n-1}$ and $\tilde{\sigma}_n$. The fact that we consider the last two coordinates only does not limit the generality of the result since by an appropriate reordening any two coordinates can be chosen. Parametrize the points of this intersection by the n-tuple $(\tilde{\sigma}_1, \ldots, \tilde{\sigma}_{n-2}, \tilde{\sigma}_{n-1}, \tilde{\sigma}_n)$. Requiring that the corank $= 1$, is equivalent to the condition that :

$$det(\hat{\Sigma}) = det(\Sigma - diag(\tilde{\sigma}_1, \ldots, \tilde{\sigma}_{n-2}, \tilde{\sigma}_{n-1}, \tilde{\sigma}_n))$$

Define now a partition as :

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{11} & \Sigma_{12} \\ \Sigma_{12}^t & \hat{\Sigma}_{22} \end{pmatrix}$$

[1] We would like to thank Prof. Roberto Guidorzi of the University of Bologna for providing us with the English translation of the original Italian version of the article

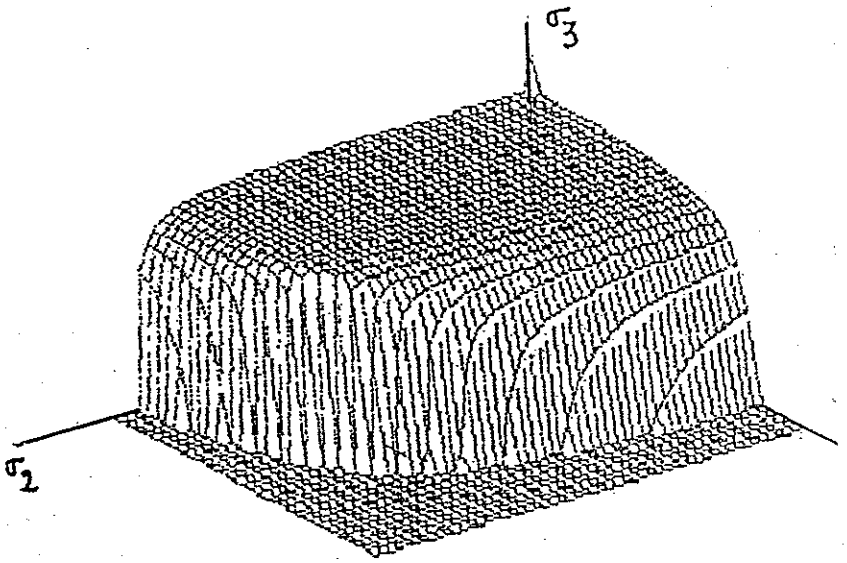


Figure 6.2: Hyperbolic noise surface for a three variable example.

where $\hat{\Sigma}_{22}$ is a 2×2 matrix. Assuming that $\hat{\Sigma}_{11}$ is nonsingular, one can apply the determinantal formula for partioned matrices (appendix C) in order to derive that :

$$det(\hat{\Sigma}) = det(\hat{\Sigma}_{11})det(\hat{\Sigma}_{22} - \Sigma_{12}^t \hat{\Sigma}_{11}^{-1} \Sigma_{12})$$

The searched coordinates $\tilde{\sigma}_{n-1}$ and $\tilde{\sigma}_n$ satisfy a relation of the type :

$$det\left(\begin{pmatrix} (\alpha - x_{n-1}) & \beta \\ \beta & (\gamma - x_n) \end{pmatrix} - \begin{pmatrix} a & b \\ b & c \end{pmatrix}\right) = 0$$

or also :

$$(\alpha^* - x_{n-1})(\gamma^* - x_n) - (\beta^*)^2 = 0$$

Since α^* and γ^* are positive, this relation describes an hyperbola which is concave towards the origin. □

An example of such an hyperbolic surface for a three variable case is depicted in figure 6.2. for the data matrix:

$$\Sigma = \begin{pmatrix} 9 & 4 & 2 \\ 4 & 17 & 2 \\ 2 & 2 & 8 \end{pmatrix}$$

The maximal corank of this example is 2.

6.2.3 Spearman matrices

An extreme case occurs, if the maximal corank **maxcor**=$n-1$. For historical reasons [6], this case is named the *Spearman case*, who hoped for an explanation of human intelligence in the form of one common factor between several intelligency tests. Fortunately, he failed. The precise reason will be the subject of section 6.4.

Definition 1 Spearman matrix
A matrix Σ is a Spearman matrix if $\mathbf{minr}(\Sigma) = 1$.

Theorem 2 *A positive definite, irreducible matrix Σ with elements σ_{ij} is a Spearman matrix if and only if , after sign changes of rows and corresponding columns, all its elements are positive and such that*

$$\sigma_{ik}\sigma_{jl} - \sigma_{il}\sigma_{jk} = 0 \quad \textit{and} \quad \sigma_{ik}\sigma_{ji} - \sigma_{ii}\sigma_{jk} \leq 0$$

for all quadruples $(i \neq j, k, l; j \neq k, l; k \neq l)$

The difference of elements of Σ are commonly referred to as the *tetrad differences* in statistical literature.

Proof : [6, p.127] □

The correct formulation of the conditions for the Spearman case took about 30 years and its history is riddled with errors and imprecisions [6, p.126]. There have been various attempts to generalize the approach of the theorem but very little has been achieved in terms of general results. The conclusion is however that, if the matrix A^tA is a Spearman matrix, the maximal corank **maxcor**of the Frisch scheme is $n - 1$: There exists a nonnegative definite diagonal matrix $\tilde{\Sigma}$ such that the difference $\Sigma - \tilde{\Sigma}$ is nonnegative and of rank 1. There are then $n - 1$ linearly independent linear relations.

6.2.4 Linear Least Squares

Already in chapter 5, we have described how the n linear least squares (LLS) solutions to the identification problem as proposed by the Frisch scheme, are related to the n column vectors of the inverse of the data covariance matrix Σ. We shall now briefly derive how the corresponding noise matrices look like. For the sake of completeness, let's restate the theorem:

Theorem 3 Least Squares Solutions.
Let Σ be a $n \times n$ positive definite data covariance matrix. Then the i-th column s^i of $S = \Sigma^{-1}$ is the i-th 'classical' linear least squares solution of the Frisch scheme. The real number $(1/s^i_i)^2$ is the norm of the residual of the i-th least square solution

Proof: For a proof, the reader is referred to theorem 4 and theorem 6 of chapter 5. □

Denote by $\tilde{\Sigma}_i$ the noise covariance matrix corresponding to the i-th LLS solution s^i. Then:

$$(\Sigma - \tilde{\Sigma}_i)s^i = 0$$

But because of the fact that s^i is the i-th column of Σ^{-1}, it holds also that $\Sigma s^i = e^i$ where e^i is the unit vector with zeros everywhere except for component i, which equals 1.

$$(\Sigma - diag(0 \ldots 1/s^i_i \ldots 0))s^i = 0$$

which identifies $\tilde{\Sigma}_i = diag(0 \ \ldots 1/s^i_i \ldots 0)$. In the light of the geometry discussed in chapter 5 the interpretation of this result is of course straightforward. The noise matrix corresponding to the i-th LLS solution has only one non-zero diagonal element, which equals precisely the norm of the corresponding residual vector. Only measurements on the i-th variable are considered to be noisy.

6.2.5 The identity matrix approach.

Observe that a zero-element diagonal element of the noise matrix $\tilde{\Sigma}$ reflects two possibilities:

1. Either it suggests that a corresponding column of the matrix A is noise-free. If the modeller would have this information a priori, he could require that the corresponding diagonal element of $\tilde{\Sigma}$ is zero.

2. If however all data are known to be noisy, a zero diagonal elementof $\tilde{\Sigma}$ indicates rank-deficiency of the noise and there exists linear relations between the noise columns of the matrix $\tilde{A}$.

It is however straightforward to show that the Frisch scheme *always* has at least one solution with a full rank diagonal positive definite noise matrix $\tilde{\Sigma}$. Simply compute the smallest eigenvalue λ_n of Σ with its corresponding eigenvector v_n. Take the noise covariance matrix $\tilde{\Sigma} = \lambda_n I_n$ where I_n is the $n \times n$ identity matrix . Then, is is easy to prove that:

1. $\Sigma - \tilde{\Sigma}$ is nonnegative definite.

2. corank $(\Sigma - \tilde{\Sigma}) = 1$ (generically)

3. $(\Sigma - \tilde{\Sigma})v_n = 0$

Note that this scheme provides the same linear relation as in the case of total least squares. However, the total linear least squares noise model does not fit into the Frisch Scheme requirements because its noise model is not diagonal. The identity matrix noise model is however more realistic, though not the most general. The matrix $\tilde{\Sigma} = \lambda_n I_n$ is not rankdeficient and all noise 'energies' (the diagonal elements of $\tilde{\Sigma}$) are equal. The column vectors of $\hat{A}$ and $\tilde{A}$ are undetermined in the sense that there exists infinitely many models for $\hat{A}$ and $\tilde{A}$ as long as the orthogonality requirements are fulfilled together with the Grammian conditions. Although the relation is unique, at least the models for the 'exact' data and the noise are not, as was explained in section 5.3.4. However, the identity matrix approach confirms that:

Linear Modelling is always possible.

Whatever mechanism generated the data, they can always be explained partially by at least one linear relation, given by the smallest eigenvector of the data matrix Σ. Hence, linear modelling obeying the rules of the Frisch scheme is *always* possible, whether the studied phenomenon is linear or not! Recall a similar statetement contained in the Wold decomposition. However, the identity matrix approach only represents *one* solution (that is only justifiable as *the unique* solution when additional *statistical* assumptions are imposed).

6.2.6 The polyhedral Cone with the Least Squares Solutions as Vertices.

Already in 1934, Frisch conjectured that a similar solution result would hold for the n-variable case as that which he had derived for the 2-variable case. However, it was Koopmans [29] who was the first to present the n-variable analog of Frisch's result, although his proof is complicated. Before stating his main result, let's first present some introductory definitions.

Definition 2 Positive (nonnegative) matrix
A matrix will be called positive (nonnegative) if its elements are all positive (nonnegative).

Definition 3 Inverse positiveness (nonnegativeness)
A square nonsingular matrix is called inverse-positive (-nonnegative) if its inverse is positive (nonnegative).

Definition 4 Sign similarity
A square matrix P is sign similar to a square matrix Q if there exist a diagonal sign matrix $L = \text{diag}(\pm 1)$ such that $P = LQL$

We are now ready to state the main result of this section.

Theorem 4 The maxcor= 1 case
If and only if the covariance matrix Σ is (sign-similar to) an inverse positive matrix, **maxcor**(Σ) *= 1. Every solution vector x is a nonnegative linear combination of the linear least squares solution vectors. The solution set is the convex polyhedral cone generated by the least squares solution vectors as extremal rays.*

Proof: For formal proofs, the reader is referred to [24] [6]. Kalman uses the Perron-Frobenius theorem, while Bekker succeeds in a proof without the need for the Perron-Frobeniustheorem. For a review of this important theorem, the reader is referred to appendix C. □

All proofs of this result (see references in [6]) after Koopman's make use of the remarkable Perron-Frobeniustheorem, including Kalman [24], who however extended the result into an existence theorem for the higher than corank one case. The link with the Perron-Frobeniustheorem is natural, since this theorem essentially states that the eigenvector, corresponding to the largest eigenvalue of a positive matrix, is unique. The close connection can be seen as follows:

$$\begin{aligned} \Sigma x &= \tilde{\Sigma} x \\ \Sigma^{-1}\tilde{\Sigma} x &= x \end{aligned}$$

Hence, when Σ is inverse positive, also the matrix $\Sigma^{-1}\tilde{\Sigma}$ will be positive (if all elements of $\tilde{\Sigma}$ are positive). Hence the largest eigenvalue of $\Sigma^{-1}\tilde{\Sigma}$ is unique and the corresponding eigenvector is positive. Observe however that it is still to be proved that that largest eigenvalue equals 1 [24].

Recently, Bekker [6] succeeded in deriving an algebraic proof without any need for the Perron-Frobenius theorem.

Observe that the theorem above contains both *necessary* and *sufficient* conditions for a matrix to have **maxcor**=1. Moreover, as every positive definite 2×2 is (sign-similar) to an inverse positive matrix, the theorem also 'explains' the stated result for the 2-variable case.

The theorem covers both a *quantitative* and *qualitative* aspect of the uncertainty principle. *Qualitatively*, it states that, even if there is maximally only one linear relation 'hidden' in the data, the solution is intrinsically non-unique. All vectors that are nonnegative linear combinations of the least squares vectors are candidates to describe the linear relations. The maximal corank is 1 if and only if by appropriate sign changes, all least squares solutions can be brought in the first orthant. However, also *quantitatively* this result is very attractive:

While a formal derivation is still lacking, experiments have shown that the cone spanned by the least squares vectors shrinks to a single point when the data tend towards the noise free case (recall that noise is to be interpreted in a broad sense as everything what is not compatible with linear relations). On the other hand, when more noise is artificially added (worse signal to noise ratio), the polyhedral cone generated by the least squares solution enlarges. Hence there is a direct relation between the amount of noise on the data (which could be both model mismatch and measurement inaccuracy, in brief, any incompatibility with the linearity assumption) and the uncertainty in the solution set, characterized by the 'volume' of the cone. If the noise is increased,the least squares solutions reach orthant planes: the situation changes and the inverse positiveness condition of the theorem ceases to be satisfied. The treatment of this situation will be postponed until section 6.3.

Example: A deductive experiment.
Consider as 'exact' data the 50×3 matrix:

$$\hat{A} = \begin{pmatrix} 1 & 50 & 51 \\ 2 & 49 & 51 \\ 3 & 48 & 51 \\ \cdots & \cdots & \cdots \\ 49 & 2 & 51 \\ 50 & 1 & 51 \end{pmatrix}$$

The pure noise matrix $\tilde{A}$ is generated with PC-Matlab's normal distribution random generator (zero mean). The sequences have varying standard deviation 2, 4, 6, 8, 10. The results of the identification and the resulting polyhedral cones are shown in figure 6.3. The linear relations are normalized such that the third component equals -1. Note that for small noise variances, which correspond to high signal-to-noise ratios, the 3 linear least squares solutions, are close to each other. The solution provided by the identity matrix approach (the smallest eigenvector of Σ) is depicted as a point. Observe that this point remains 'nicely' in the middle of the polyhedral cone, which confirms (experimentally) the well known statistically consistency of this estimator (which is of course only of any value in this deductive example!) -
The following result shows that in some cases it is possible to verify the fact that **maxcor**=1 by inspection of the matrix Σ itself instead of its inverse.

Definition 5 Metzler matrix.
A matrix M with elements m_{ij} is a Metzler matrix if $m_{ij} \geq 0$ for all $i \neq j$.

Lemma 1 Inverse positivity of a Metzler matrix.
Let M be a Metzler matrix. Then $-M^{-1}$ exists and is positive if and only if all of the eigenvalues of M are strictly within the left half complex plane.

Proof: [32]. □

Lemma 2 The maximal corank of Metzler matrices.
If the matrix Σ of the Frisch scheme is such that $-\Sigma$ is (sign-similar to) a Metzler matrix, then $\mathbf{maxcor}(\Sigma) = 1$.

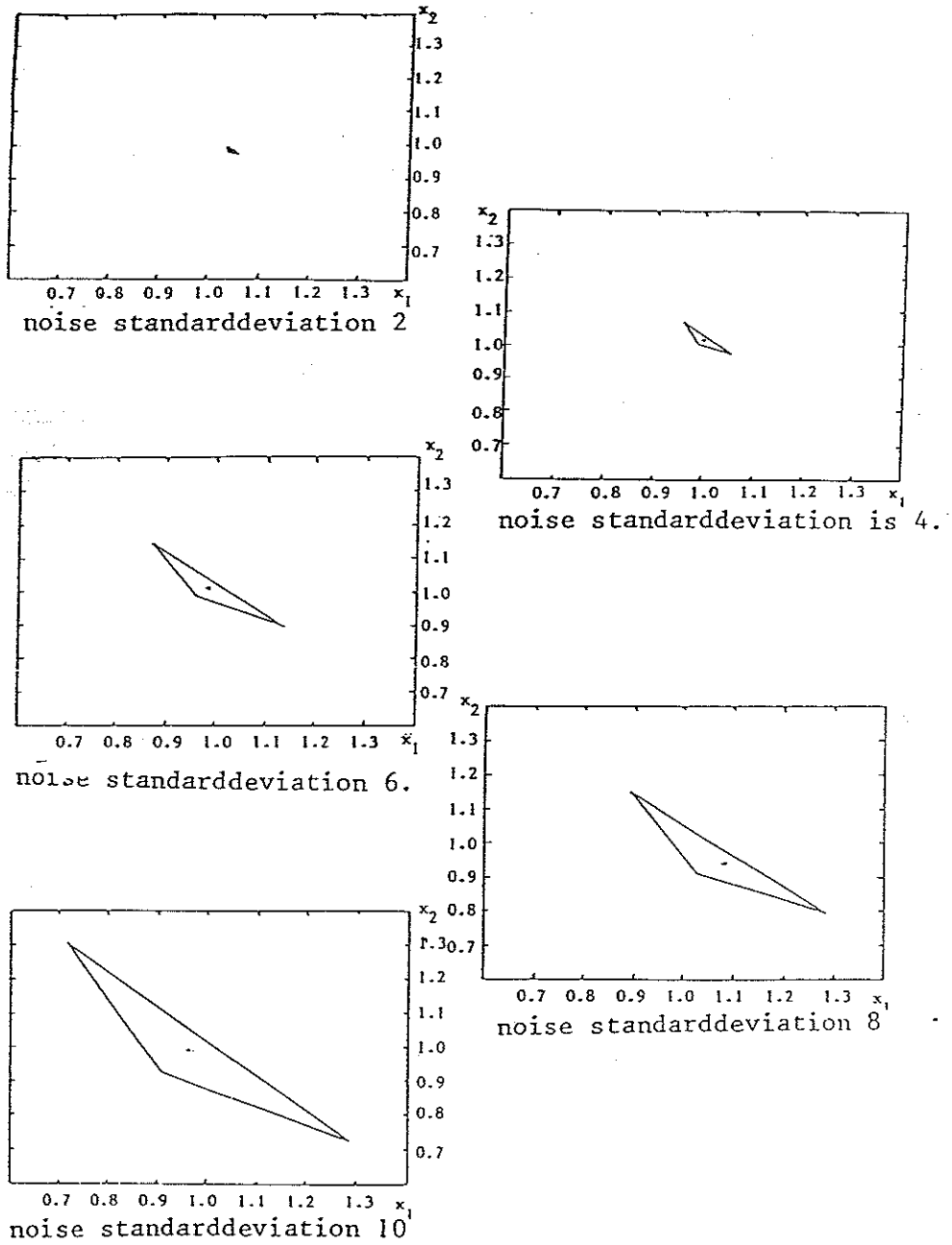


Figure 6.3: Intersection of the polyhedral solution cone with a the hyperplane $x_3 = -1$ for a three variable identification experiment, increasing noise levels.

Proof: Trivial from lemma 1 and theorem 4. □

The trouble however with the Perron-Frobenius theorem based approach as e.g. followed in [24] and the approach followed in [6], is that they are both algebraic. Moreover, the Perron-Frobeniustheorem is of no use whatsoever in characterizing the solution set if the data matrix Σ is *not* sign-similar to an inverse positive matrix. Thereto, we shall prefer a mixed geometrical-algebraic treatment of the problem. But first, let's analyse what happens if the inverse positiveness condition is not satisfied.

The existence theorem for maxcor> 1

If the data covariance matrix is not (sign similar to) an inverse positive matrix, the conditions for the **maxcor**$= 1$ case are not satisfied. While the geometrical description of the solution set is largely 'terra incognita', Reiersol proved the following existence result in 1941 [35]. His proof however is not very transparant. Kalman presents a clearer constructive proof in [24].

Theorem 5 *If and only if the data covariance matrix is* not *sign similar to an inverse positive matrix,* **maxcor**> 1 .

Observe that the theorem is an existence theorem in that it provides the condition under which there will exist at least two linearly independent solutions. Moreover, our proof will be (partially) constructive: we shall explicitly construct two linearly independent solutions but it is important to remember, that there are much more solutions than the two that will be constructed. The geometrical construction of the complete solution set will be the subject of section 6.3.
Contrary to the proof of the **maxcor**$=1$ case, we shall provide the reader with the proof of this theorem, because it provides some geometrical insight and makes use of some interesting matrix lemmas, which are stated and proved first. In order to avoid repeating the same statement every time, it is assumed that:

1. Σ is an $n \times n$ square symmetric positive definite matrix partitioned as:

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^t & \Sigma_{22} \end{pmatrix}$$

   where Σ_{11} is an $r \times r$ matrix ($r \leq n$).

2. $S = \Sigma^{-1}$ is partitioned accordingly:

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{12}^t & S_{22} \end{pmatrix}$$

3. $\tilde{\Sigma}_1$ is a $r \times r$ diagonal positive definite matrix.

4. The $n \times n$ matrix $\tilde{\Sigma}$ is a nonnegative diagonal matrix partitioned as:

$$\tilde{\Sigma} = \begin{pmatrix} \tilde{\Sigma}_1 & 0 \\ 0 & 0 \end{pmatrix}$$

Lemma 3 Positive definiteness of partitioned matrices.
Both Σ_{11} and Σ_{22} are positive definite.

Proof: Follows immediately from the eigenvalue interlacing property [18]. □

Lemma 4 Nonnegative Definiteness of Partitioned Matrices.
If Σ is nonnegative definite, then Σ_{11} and/or Σ_{22} are either positive or nonnegative definite [18].

Proof: Eigenvalue interlacing theorem □

Lemma 5 Rank Property.
If $corank(\Sigma - \tilde{\Sigma}) = n - r$, then $corank(S_{11}^{-1} - \tilde{\Sigma}_1) = n - r$.

Proof: Let X be a partitioned $n \times (n-r)$ matrix of rank $n-r$, satisfying:

$$\begin{pmatrix} \Sigma_{11} - \tilde{\Sigma}_1 & \Sigma_{12} \\ \Sigma_{12}^t & \Sigma_{22} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = 0$$

Because of the positive definiteness of Σ, by lemma 3, Σ_{22} is invertible and it is easily verified that:

$$X_2 = -\Sigma_{22}^{-1}\Sigma_{12}^t X_1$$

and hence:

$$\left[(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^t) - \tilde{\Sigma}_1\right] X_1 = 0$$

However, observe that:

$$rank \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = rank \left[\begin{pmatrix} I_{n-r} \\ -\Sigma_{22}^{-1}\Sigma_{12}^t \end{pmatrix} X_1 \right] = n - r$$

which shows that $rank(X_1) = n - r$. Moreover, via the matrix inversion lemma for partioned matrices it is easily verified that:

$$S_{11}^{-1} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^t$$

This proves the theorem. □

Lemma 6 Positive Definiteness of the Schur Complement.
If Σ is positive definite (nonnegative definite), the Schur complement $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^t$ is positive definite (or nonnegative definite if Σ_{22} is invertible).

Proof: Because Σ is positive definite, its inverse is also positive definite. From lemma 3, it follows that also S_{11} is positive definite. But this on its turn implies that S_{11}^{-1} is positive definite which equals precisely the Schur complement. □

Lemma 7 *Let $\Sigma - \tilde{\Sigma}$ be nonnegative definite of corank $n - r$. Then:*

$$(S_{11}^{-1} - \tilde{\Sigma}_1)NND \longleftrightarrow (\Sigma - \tilde{\Sigma})NND$$

Proof: Follows from a straightforward combination of lemma 3, 4, 5 and 6. □

We are now ready to prove the existence theorem for the higher than corank 1 case.
Proof: It will be proved that, if the matrix Σ is not sign similar to an inverse positive matrix, then there exist at least two linearly independent solutions.
1. Zero elements
First consider the case where at least two elements of S are 0, say the elements $s_{12} = s_{21} = 0$. Then:

$$S_{11} = \begin{pmatrix} s_{11} & 0 \\ 0 & s_{22} \end{pmatrix}$$

Then chose the noise matrix $\tilde{\Sigma}_1 = S_{11}^{-1}$. From lemma 5, it then follows that:

$$corank(S_{11}^{-1} - \tilde{\Sigma}_1) = 2 \longleftrightarrow corank(\Sigma - \tilde{\Sigma}) = 2$$

Moreover, both matrices are nonnegative definite from lemma 7.
2. No zero elements
Assume there are no zeros but that Σ is not sign similar to an inverse positive matrix. All possible situations can be converted to the following one. Let S_{11} be a 3×3 matrix with at least one negative element, say $-s_{13} = -s_{31}$ (we denote this explicitly by the minus sign). Then choose the following diagonal matrix and correspondingly, two columns of X_1 as:

$$\tilde{\Sigma}_1 = \begin{pmatrix} \frac{s_{23}}{s_{11}s_{23}+s_{12}s_{13}} & 0 & 0 \\ 0 & \frac{s_{13}}{s_{12}s_{23}+s_{13}s_{22}} & 0 \\ 0 & 0 & \frac{s_{12}}{s_{13}s_{23}+s_{12}s_{33}} \end{pmatrix} \quad X_1 = \begin{pmatrix} s_{11}s_{23} + s_{12}s_{13} & 0 \\ s_{12}s_{23} + s_{22}s_{13} & s_{12}s_{23} + s_{22}s_{13} \\ 0 & s_{13}s_{23} + s_{12}s_{33} \end{pmatrix}$$

It is easily verified that:

$$(S_{11}^{-1} - \tilde{\Sigma}_1)X_1 = 0$$

X_2 can be constructed as $X_2 = -\Sigma_{22}^{-1}\Sigma_{12}^t X_1$. Obviously, the columns of:

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

are linearly independent (since the columns of X_1 are). Applying lemma 7, then proves the theorem. □

Notwithstanding the constructive proof of the existence theorem, it delivers not very much insight in the solution set of the problem whenever the corank is larger than or equal to 2. We shall now concentrate on a procedure developed by Reiersol, which however will also not provide a completely satisfactory answer. Then we shall turn our attention to a purely geometrical treatment of the problem.

6.2.7 The Reiersol tree search procedure

Reiersol proved the following result in 1950 [36]. It gives an upper bound on the maximal corank in terms of the largest submatrix that is sign similar to an inverse positive matrix.

Theorem 6 The Reiersol tree search
Let ν_i denote subsets of the set of n variables and let Σ_{ν_i} denote the submatrix of the data covariance matrix Σ corresponding to the variables in ν_i. If n' is the maximum order of any submatrix Σ_{ν_i} which is sign similar to an inverse positive matrix, then :

$$\begin{aligned} \mathbf{maxcor}(\Sigma) &\leq n-(n'-1) \\ n-\mathbf{minr}(\Sigma) &\leq n-(n'-1) \\ \mathbf{minr}(\Sigma) &\geq n'-1 \end{aligned}$$

Proof : see [36] □

As an illustration, consider the following example, which we have borrowed from [28].

$$\Sigma = \begin{pmatrix} 4 & 3 & 0 & -1 & 1 \\ 3 & 8 & -1 & -3 & 1 \\ 0 & -1 & 4 & 3 & 1 \\ -1 & -3 & 3 & 6 & 1 \\ 1 & 1 & 1 & 1 & 3 \end{pmatrix}$$

None of the possible 4×4 is sign-similar to an inverse positive matrix. The submatrices $\Sigma_{245}, \Sigma_{235}, \Sigma_{234}, \ldots$ however are sign-similar to an inverse positive matrix. All 2x2 submatrices are (sign-similar to) an inverse elementwise positive matrix. Hence maximal corank ≤ 3!

6.3 A Geometrical Treatment of The Frisch Scheme.

Most of the investigations on the Frisch scheme, conducted till recently have concentrated on the problem of the communalities, using a more or less algebraic formalism (determinantal identities, using well known matrix lemmas, etc...). Among them, Kalman's use of the Perron - Frobeniustheorem and the **maxcor**$= 1$ case with the polyhedral cone generated by the least squares solutions, belong to those results that are more apt to geometrical interpretation.
The path initiated in our research exploits both some simple algebraic observations together with geometrical insights. Instead of trying to characterize the solution in terms of the communalities, we follow a complementary path and confine our attention to the solution vectors of the Frisch scheme. As will be clarified, this path is largely unexplored and neglected but yet provides powerful additional insight into the problem : The main result consists of an algorithmic approach that computes the maximal corank and some corresponding vectors, which satisfy some desirable and natural properties (orthant null invariant vectors). Moreover, it is conjectured that the complete solution of the Frisch scheme, i.e. the complete geometrical characterization of the solution set, can be done with the least squares vectors and the orthant null invariant vectors as a basic framework.

Subsection 6.3.1 and 6.3.2. will describe some necessary properties of the solution vector such as orthant and null invariance. In subsection 6.3.4., we compute an upper bound on the

noise energy of each measurement channel. In subsection 6.3.5., it is analysed how convex combinations of least squares vectors, provide additional insight in the geometry. In turns out that solution vectors with zeros are important, since they allow to compute the maximal corank. This idea is exploited in subsection 6.3.6. and leads to an algorithmic procedure, which reminds of the Reiersol tree search procedure.

6.3.1 Orthant invariance

An orthant of the n dimensional vector space $\mathcal{R}^n$ will be characterized by a diagonal matrix E with +1 and -1 along the diagonal: $E = \text{diag}(\pm 1)$. There are 2^n different orthants. The nonnegative (first) orthant is denoted by the $n \times n$ unit matrix I_n. A vector x is said to belong to orthant E, denoted by $x \in E$, if the vector Ex belongs to the first orthant : $Ex \in I_n$.

Definition 6 Orthant Invariant vectors.
A vector $x \in E$ is orthant invariant for the matrix Σ if $\Sigma x \in E$.

It is an easy exercise to verify that:

Theorem 7 Orthant Invariance in the Frisch Scheme.
All solutions x of the Frisch scheme are orthant invariant.

Proof : This follows directly from the diagonality and nonnegative definiteness of $\tilde{\Sigma}$ from $\Sigma x = \tilde{\Sigma} x$. □

Define $y = \Sigma x$. If $x \in E$ and x is orthant invariant, also $y \in E$ and equivalently : $Ey = (E\Sigma E)(Ex)$ where now both (Ey) and (Ex) are nonnegative vectors. This observation permits to compute explicitly and characterize geometrically all orthant invariant vectors of the matrix Σ :

Theorem 8 Computation of orthant invariant vectors.
In orthant E, the orthant invariant vectors can be obtained from the nonnegative solution of the set of linear equations:

$$(E\Sigma E \quad -I_n) \begin{bmatrix} Ex \\ Ey \end{bmatrix} = 0$$

Hence, the set of orthant invariant vectors is a convex polyhedral cone .

Proof : Trivial from the results in chapter 2. □

Geometrically, the solution of the above problem is the intersection of the kernel of the matrix $[(E\Sigma E) \ (-I_n)]$ with the first orthant. An algorithm to obtain all solutions to a set of nonnegative linear equations, may be found in chapter 2. Hence, we have the following algorithm which computes all orthant invariant vectors of the Frisch scheme:

- For each orthant E:
- Solve nonnegatively for all vectors z_1 and z_2:

$$(E\Sigma E \quad -I_n) \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = 0$$

- The solution set of the vectors Ez_1 constitutes the set of orthant invariant vectors for the matrix Σ in orthant E.

Observe that only half of total number of orthants is to be checked because of the sign-symmetry.
The property of orthant invariance ensures that if x and $y = \Sigma x$ are in the same orthant, there will exists a diagonal matrix D with nonnegative elements such that $y = Dx$. However, in order for this diagonal matrix D to be a valid diagonal matrix $\tilde{\Sigma}$ as required by the Frisch scheme, there are additional conditions that are to be satisfied.

6.3.2 Null invariance

A second straightforward observation will appear to be crucial in the computation of the maximal corank. Suppose that the i-th component x_i of a solution vector x is zero : $x_i = 0$. This implies that also the i-th component of the product Σx will be zero.

Definition 7 Null Invariance.
The vector x with zero component $x_i = 0$ will be said to be null invariant *for component i with respect to the matrix Σ if from $x_i = 0$ it follows that $(\Sigma x)_i = 0$.*

Then we have:

Theorem 9 Null Invariance for the Frisch scheme.
All solution vectors of the Frisch scheme with a zero component are null invariant for that component with respect to the matrix Σ.

Proof: Trivial. □

Observe that this condition is necessary. Hence, if an orthant invariant vector has a zero component which is not orthant invariant, it cannot be a solution of the Frisch scheme. Observe that in the described algorithm for the computation of all orthant invariant vectors in a certain orthant, it is very easy to verify whether a vector is null-invariant. Simply compare the zero patterns in the solution vectors z_1 and z_2. If z_2 has zero components whenever a component of z_1 is zero, the vector is null-invariant and satisfies a second necessary condition of a solution vector of the Frisch scheme.

6.3.3 Allowed vectors

The properties of orthant and null invariance are necessary for a solution vector but not yet sufficient. In other words, there exist orthant null invariant vectors that are no solution to the problem, simply because they 'cause' a diagonal matrix $\tilde{\Sigma}$ such that the difference $\Sigma - \tilde{\Sigma}$ is not nonnegative definite but indefinite. The corresponding diagonal matrix for an orthant null invariant vector x can be computed by the following scheme:

x is orthant null invariant with components x_i:

- if $x_i = 0$, set $\tilde{\sigma}_i = 0$ (i-th diagonal element of $\tilde{\Sigma}$)
- if $x_i \neq 0$, compute $y = \Sigma x$, set $\tilde{\sigma}_i = y_i / x_i$

An orthant null invariant vector x will be called *allowed* if the difference matrix $\Sigma - \tilde{\Sigma}$ is nonnegative definite.

While this last requirement of allowedness has not been solved yet in full generality, it has been solved for some special cases (corank 1 for instance). The geometrical framework however permits to find some more allowed vectors. Hereto we confine now the attention to some remarkable properties satisfied by the least squares vectors.

6.3.4 What is the maximal amount of noise on each channel?

Assume that x is an allowed orthant null invariant vector. The noise matrix $\tilde{\Sigma}$ that corresponds to x can be computed as follows:

1. Compute $y = \Sigma x$.
2. The matrix $\tilde{\Sigma}$ can now be computed from:

$$y = \Sigma x = \tilde{\Sigma} x$$

Hence we have that:

$$\begin{aligned} \tilde{\sigma}_i &= y_i/x_i \quad if \quad x_i \neq 0 \\ \tilde{\sigma}_i &= \text{'arbitrary'} \quad if \quad x_i = 0 \end{aligned}$$

Observe that the choice of $\tilde{\sigma}_i$ when $x_i = 0$ is not completely arbitrary. It must be ensured that $\Sigma - \tilde{\Sigma}$ is nonnegative definite.

In [5], one finds the following result on the maximal allowable amount of noise that can be present in one variable. Obviously, this represents an upper bound on the diagonal elements of $\tilde{\Sigma}$.

Theorem 10 The maximal noise theorem
Let the i-th diagonal element of $\tilde{\Sigma}$ be $\tilde{\sigma}_i$. Then:

$$0 \leq \sigma_i \leq det(\Sigma)/det(\Sigma_i)$$

where Σ_i is the matrix obtained from Σ by deleting its i-th row and column.

Proof : Follows directly from the requirements of the Frisch scheme and a well known theorem on the determinant of a partitioned matrix [5]. □

Note that the maximum for $\tilde{\sigma}_i$ is reached for the i-th linear least squares solution, as in theorem 1. If s_i^i is the i-th element of Σ^{-1} then it is easy to prove that:

$$1/s_i^i = det(\Sigma)/det(\Sigma_i)$$

The interpretation of this results confirms in some sense the intuition, that a least squares solution implicitly considers all data, except one variable, to be noisy. The identified amount of noise will be larger, than when all data are considered to be noisy.

6.3.5 About vectors that are convex combinations of least squares vectors

In theorem 4, it was shown that the columns of the inverse matrix Σ^{-1} are the linear least squares solutions to the problem. Theorem 3 states that whenever the inverse matrix Σ^{-1} is (sign-similar to an) elementwise positive, all linear least squares vectors can be brought by appropriate sign changes into one orthant. The solution set of vectors that satisfy the Frisch scheme conditions is then the convex polyhedral cone generated by the least squares vectors. If Σ^{-1} is not (sign-similar to an) inverse positive elementwise however, the least squares solutions can never be 'moved' into one orthant. If now two least squares vectors belong to two different orthants, the line segment that connects these two vectors must have an intersection with at least one 'orthant' plane. The corresponding intersection vector hence will have at least one zero. Recall the constructive proof of the existence theorem 5, which exploited precisely this observation. The following theorem states the conditions for such a vector to be allowed [11].

Theorem 11 About convex combinations of 2 least squares solutions.
Let E be a diagonal sign matrix. Let s_E^i and s_E^j be the i-th and j-th column of the matrix $E\Sigma^{-1}E$. Let x be a convex combination of s_E^i and s_E^j, such that $x_k = 0$ and let x be orthant null invariant. Then:

$$x = \alpha s_E^i + (1-\alpha)s_E^j \quad with \quad \alpha = (s_E^j)_k / ((s_E^j)_k - (s_E^i)_k)$$

If x is orthant null invariant, then:

- *x is allowed if and only if $(s_E^i)_j = (s_E^j)_i \geq 0$*
- *x is not allowed if and only if $(s_E^i)_j = (s_E^j)_i < 0$*

Proof :
Proof for $(s_E^i)_j = (s_E^j)_i < 0$
Since x is orthant null invariant for $E\Sigma E$, there exists a nonnegative diagonal matrix $\tilde{\Sigma}$ such that:

$$\Sigma(Ex) = \tilde{\Sigma}(Ex)$$

Define the vector $y = E\Sigma x$, then obviously:

$$y = \alpha(Ee^i) + (1-\alpha)(Ee^j)$$

where e^i and e^j are the i-th and j-th column of the identity matrix I_n. The diagonal elements $\tilde{\sigma}_i$ and $\tilde{\sigma}_j$ can be computed as:

$$\tilde{\sigma}_i = y_i/x_i = \frac{\alpha}{\alpha(s_E^i)_i + (1-\alpha)(s_E^j)_i}$$

and

$$\tilde{\sigma}_j = y_j/x_j = \frac{1-\alpha}{\alpha(s_E^j)_i + (1-\alpha)(s_E^j)_j}$$

Now it is easy verified that, if $(s_E^i)_j = (s_E^j)_i < 0$, we have that:

$$\tilde{\sigma}_i > \frac{1}{(s_E^i)_i} \qquad \tilde{\sigma}_j > \frac{1}{(s_E^j)_j}$$

Hence, the maximal noise theorem is violated. The vector x is not allowed.

Proof for $(s_E^i)_j = (s_E^j)_i \geq 0$
Without loss of generality, assume that $i = 1$ and $j = 2$. Then define a diagonal matrix $\tilde{\Sigma}$ as:

$$\tilde{\Sigma} = \begin{pmatrix} \tilde{\Sigma}_1 & 0 \\ 0 & 0 \end{pmatrix}$$

where $\tilde{\Sigma}$ is a 2×2 diagonal matrix:

$$\tilde{\Sigma} = \begin{pmatrix} \tilde{\sigma}_1 & 0 \\ 0 & \tilde{\sigma}_2 \end{pmatrix}$$

Here $\tilde{\sigma}_1$ and $\tilde{\sigma}_2$ are given by:

$$\tilde{\sigma}_1 = \frac{\alpha}{\alpha(s_E^1)_1 + (1-\alpha)(s_E^2)_1}$$

and

$$\tilde{\sigma}_2 = \frac{1-\alpha}{\alpha(s_E^2)_1 + (1-\alpha)(s_E^2)_2}$$

Let S_{11} be the upper 2×2 part of $S = \Sigma^{-1}$.

$$S_{11} = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix}$$

From lemma 3 it follows that S_{11} is positive definite, hence invertible. A straightforward computation shows that:

$$S_{11}^{-1} - \tilde{\Sigma}_1 = \frac{s_{ij}}{det} \begin{pmatrix} \frac{x_j}{x_i} & -1 \\ -1 & \frac{x_i}{x_j} \end{pmatrix}$$

where $det = det(S_{11})$. It is easily verified that this matrix is of corank 1 and that it is nonnegative definite. The fact that then $\Sigma - \tilde{\Sigma}$ is nonnegative definite follows from lemma 7. □

An important conclusion is that in the case where Σ is not sign similar to an inverse positive matrix, not only the linear least squares vectors seem to play an important role. Their allowed convex combinations with at least one zero component are important as well. The theorem then provides a simple test to check whether such a convex combination will be allowed or not: Only those least squares solution i and j, when reduced to orthant E, may be convexly combined, when $(s_E^i)^j = (s_E^j)_i \geq 0$. This can be verified by simple inspection.
Observe that as a matter of fact, we have proved something more than was strictly required: Using the same notations as in theorem 11, we have that:

Corollary 1
- *No convex combination of s_E^i and s_E^j whatsoever is allowed if $(s_E^i)_j = (s_E^j)_i < 0$.*
- *Every convex combination of s_E^i and s_E^j is allowed as long as:*

$$\alpha \geq min\left(\frac{(s_E^i)_j}{(s_E^i)_j - (s_E^i)_i}, \frac{(s_E^j)_j}{(s_E^j)_j - (s_E^i)_j}\right)$$

Proof: The first part follows immediately from the proof of theorem 11. The second part follows from the second part of the proof of theorem 11, by requiring that $\tilde{\sigma}_i, \tilde{\sigma}_j \geq 0$. □

6.3.6 Recognition of the maximal corank

Assume that the maximal corank of Σ equals $n-r$. Then we have the following theorem:

Theorem 12 The relation between zeros in the solution vectors and the maximal corank.
If the maximal corank of Σ equals $n-r$, there exist $n-r$ linearly independent vectors X, with at least $n-r-1$ zeros.

Proof: Since **maxcor**$=n-r$, there exists a diagonal matrix $\tilde{\Sigma}$ such that $rank(\Sigma - \tilde{\Sigma}) = r$ and there exists an $n \times (n-r)$ matrix X of rank $n-r$, such that $\Sigma X = \tilde{\Sigma} X$. An example of such a diagonal matrix, is the $(n-r) \times (n-r)$ matrix $\tilde{\Sigma}_2$, obtained from the least squares solution, which considers the first r variables as noisefree and the last $n-r$ as noisy. With an obvious partitioning of Σ and X, it holds that:

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^t & (\Sigma_{22} - \tilde{\Sigma}_2) \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = 0$$

From lemma 3 follows the positive definiteness of Σ_{11}. Hence:

$$X_1 = -\Sigma_{11}^{-1} \Sigma_{12} X_2$$

and hence:

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} -\Sigma_{11}^{-1}\Sigma_{12} \\ I_{n-r} \end{pmatrix} X_2$$

Because X is of rank $n-r$, also X_2 is of rank $n-r$. Postmultiplication with X_2^{-1} delivers $n-r$ linearly independent solutions with at least $n-r-1$ zeros. □

These vectors are of course to be found among the orthant null invariant vectors of Σ.

6.3.7 The global solution set of the identification

The solution set consists of those vectors x that are orthant null invariant and allowed with respect to the matrix Σ. As a special case, the linear least squares solutions belong to this class of vectors. They play a prominent role in the maximal corank = 1 case. However, they play also a fundamental role in the description of the corank higher than 1 solution set. We conjecture that:

The general solution set of the Frisch scheme consists of a collection of polyhedral cones.

While there is no general proof yet of this statement, we have already provided a lot of necessary tools and results that seem to sustain this proposition. The vertices of these polyhedral solution cones are the least squares solutions and allowed orthant null invariant vectors with zero components. The maximal corank is $n-r$ if there exist maximally $n-r$ linear independent allowed orthant null invariant vectors each with $n-r-1$ zeros.

In order to investigate the solution of the Frisch scheme for a given $n \times n$ matrix Σ, the following *algorithmic* tools are available, derived from the results in sections 6.3.1. to 6.3.6.:

The Telescoping Approach:

1. Check the inverse positiveness condition.
2. If Σ^{-1} is not sign similar to a positive matrix, go to step 3. If Σ^{-1} is sign similar to a positive matrix. Assume that E is a diagonal sign matrix such that $E\Sigma^{-1}E$ is positive. Then the solution set is the polyhedral solution cone generated by all convex combinations of the least squares vectors, which are the columns of $E\Sigma^{-1}E$. The maximal corank is then one. The complete solution has been found.
3. Investigate the matrix via the Reiersol tree search procedure. This provides an upper bound for the maximal corank.
4. Start investigating for each of the 2^{n-1} orthants, the sign structure of $E\Sigma^{-1}E$ via theorem 11. Every allowed orthant null invariant vector with at least one zero component gives raise to a reduced $(n-1)\times(n-1)$ problem, which can on its turn be investigated. Proceed this telescoping procedure until a submatrix of Σ is reached, which is sign similar to an inverse positive matrix.
5. Solutions found for the smaller dimensional matrices, are automatically solution of the $n \times n$ problem by a simple application of lemma 5 and lemma 7.
6. The maximal corank is $n-r$ if via this telescoping procedure, $n-r$ linearly independent vectors can be found with $n-r-1$ zeros. This is an immediate consequence of theorem 12.

The brute force approach.

1. For each orthant, compute the orthant null invariant vectors and check there allowedness. For vectors that are convex combinations of least squares solutions, this can be done readily by a simple application of theorem 11. For others this has to be done numerically.
2. Among the allowed orthant null invariant vectors, detect these with the maximal number of zeros. Assume that this number is p. Then, if over all orthants, one finds $p+1$ linearly independent allowed orthant null invariant vectors, the maximal corank is p.

Obviously, both algorithms succeed in:

- Computation of the maximal corank $\mathbf{maxcor}\Sigma$.
- Finding vectors others than those provided by the least squares, and identity matrix approach, that are solutions to the Frisch scheme.
- Providing geometrical insight into the solution set.

Recall that the Frisch scheme consists of two main requirements:

1. Determine the maximal corank
2. Characterize the complete solution set

Having claimed that we have solved (at least from the algorithmic point of view) the determination of the maximal corank, let's now analyse what is lacking for a similar claim for the second part:

1. Is the set of normalized orthant null invariant vectors closed? Since the solutions are determined up to a scalar, it suffices to investigate vectors x that are lying on the unit hypersphere. These candidate solution vectors are called *normalized.* Obviously, the set of normalized orthant invariant vectors is closed. If the set of orthant null invariant vectors were closed, it would be legitimate to eliminate immediately vectors that are not null invariant. Because if it were open, than a vector which is orthant invariant but *not* null invariant could act as a vertix, which is itself not included in the final solution set but for instance, certain convex combinations with other vectors, that are orthant null invariant, could be orthant null invariant. Our conjecture is that the set of orthant null invariant vectors is closed.

2. How does the solution set looks like in one orthant? We have some good indications that the solution set is connected, i.e. when projected on the unit sphere, one obtains a connected pattern of closed polyhedral (spherical) objects. Of course, the connection could be via one point only.
   As a matter of fact, for the three variable case, we have the following theorem, which describes geometrically, the only case different from the **maxcor**=1 case:

Theorem 13 The corank=2 case with 3 variables.
Let Σ be a 3×3 symmetric positive definite matrix. If Σ is not sign similar to an inverse positive matrix, then the solution set of vectors x satisfying the requirements of the Frisch scheme, consists of 6 polyhedral cones in 6 orthants.

Proof: Assume that:

$$a^t = (a_1\ a_2\ a_3) \quad b^t = (b_1\ b_2\ b_3)$$

are two allowed orthant null invariant vectors of the same orthant E with corresponding noise matrices:

$$\tilde{\Sigma}_a = \begin{pmatrix} \alpha_1 & 0 & 0 \\ 0 & \alpha_2 & 0 \\ 0 & 0 & \alpha_3 \end{pmatrix} \quad \tilde{\Sigma}_b = \begin{pmatrix} \beta_1 & 0 & 0 \\ 0 & \beta_2 & 0 \\ 0 & 0 & \beta_3 \end{pmatrix}$$

Assume that $a_1 \neq 0$, $a_2 \neq 0$, $b_1 \neq 0$, $b_2 \neq 0$. Then from

$$\Sigma a = \tilde{\Sigma}_a a$$

it follows that:

$$\begin{aligned} \alpha_1 &= (\sigma_{11}a_1 + \sigma_{12}a_2 + \sigma_{13}a_3)/a_1 \\ \alpha_2 &= (\sigma_{21}a_1 + \sigma_{22}a_2 + \sigma_{23}a_3)/a_2 \end{aligned}$$

and

$$\begin{aligned}\sigma_{11} - \alpha_1 &= -(\sigma_{12}a_2 + \sigma_{13}a_3)/a_1 \\ \sigma_{22} - \alpha_2 &= -(\sigma_{21}a_1 + \sigma_{23}a_3)/a_2\end{aligned}$$

Because a is allowed, $\Sigma - \tilde{\Sigma}_a$ is nonnegative definite and hence its principal 2×2 minors are nonnegative:

$$(\sigma_{11} - \alpha_1)(\sigma_{22} - \alpha_2) - \sigma_{12}\sigma_{21} \geq 0$$

Substituting the expressions for α_1 and α_2 results in:

$$\sigma_{12}\sigma_{13}\frac{a_3}{a_1} + \sigma_{21}\sigma_{13}\frac{a_3}{a_2} + \sigma_{13}\sigma_{23}\frac{a_3^2}{a_1 a_2} \geq 0$$

Hence, if $a_1 a_2 >< 0$, we have that:

$$a_3(\sigma_{13}\sigma_{12}a_1 + \sigma_{12}\sigma_{23}a_2 + \sigma_{13}\sigma_{23}a_3) >< 0$$

Similar expressions hold for b. It is then straightforward to show that a similar expression holds for a convex combination $\gamma a + (1-\gamma)b$. □

Observe however that the solution set in one orthant is not necessarily convex for more than 3 variables. As an example, the reader may wish to analyse the following example in detail, using the algorithmic elements that were derived before.

Example:
Consider the 4×4 positive definite matrix:

$$\Sigma = \begin{pmatrix} 6 & -1 & 7 & 0 \\ -1 & 3 & -1 & -2 \\ 7 & -1 & 13 & -1 \\ 0 & -2 & -1 & 9 \end{pmatrix}$$

The solution set consists of several polyhedral cones in 8 different orthants. Because of the sign symmetry, we only have to give the solutions in 4 orthants. The solutions are numbered. If a solution and its negative are needed, the second one has the same number, preceded by a minus sign.

Orthant $E = diag(1\ 1\ 1\ 1)$**:** In this orthant there is 1 least squares solution and there are 3 orthant null invariant vectors with 1 zero each. Every nonnegative combination is an allowed vector.

1	2	3	4
11	0	13	0
63	7	85	21
0	1	1	3
14	6	19	5

Every nonnegative combination of the solutions 1 and 3 provides a $corank = 2$ solution.

Orthant $E = diag(1\ 1\ -1\ 1)$**:** In this orthant there are only 2 allowed orthant null invariant vectors. Every nonnegative combination is allowed as well.

1	5
11	27
63	7
0	-14
14	0

Orthant $E = diag(1\ 1\ -1\ -1)$**:** There is 1 least squares solution and 3 allowed orthant null invariant vectors with each 1 zero. The generate a cone of allowed vectors. The line **5-6** is a $corank = 2$ line.

5	6	7	8
27	9	11	283
7	0	0	39
-14	-5	-9	-150
0	-2	-1	-8

Orthant $E = diag(1\ -1\ -1\ -1)$**:** There are 2 least squares solutions in this orthant and 4 allowed orthant null invariant vectors with each 1 zero. The line **4-6** is not allowed. Hence, the solution set in this orthant consists of the union of two polyhedral cones: **2-4-7-9-10** and **2-6-7-9-10.**

9	10	-2	-4	6	7
8	50	0	0	9	11
-57	-1	-7	-21	0	0
-15	-43	-1	-3	-5	-9
-82	-5	-6	-5	-2	-1

Our conjecture is that the final, complete solution set can be completely described via the vectors with zeros, that are determined via the telescoping algorithm or via the brute force approach. The difficulty however arises in the reconstruction of the solution set from these vectors, One could of course try out numerically all possible convex combinations and check if they satisfy all necessary conditions. But still then, it is to be proved that the vectors with zeros found via the two procedures, provide a sufficient framework to reconstruct the global solution set.

3. Another question concerns the precise relation between the telescoping and the brute force approach. More specifically, do they both find the same set of allowed orthant null invariant vectors with zeros? In the above example, it is remarkable that in the orthant in which there are two least squares solutions, there are two polyhedral solution cones. Is there any relation?

6.3.8 A third proof of the maxcor=1 case.

In this section, it will be shown how our mixed algebraic-geometrical results, provides a third proof of the maximal corank=1 case, besides the ones derived in [6] [24]. Hereto, let's first state some properties.

Lemma 8 Geometrical characterization of positiveness.
If a symmetric non-positive matrix is sign similar to a positive matrix, its columns belong to 2 opposite orthants.

Proof: The proof is a simple algebraic exercise. □

Corollary 2 Allowed orthant null invariant vectors with at least one zero.
If the maximal corank of Σ is equal to or larger than 2, there exist allowed orthant null invariant vectors with at least one zero, that are obtained as convex combinations of 2 columns of $E\Sigma^{-1}E$, where E is a sign matrix.

Proof: Is an immediate consequence of the constructive proof of the existence theorem 5. □

The reverse statement of this corollary reads: If and only if there exists no allowed orthant null invariant vector as a convex combination of 2 columns of $E\Sigma^{-1}E$ for a certain sign matrix E, the maximal corank is one.

Theorem 14 Alternative proof for maxcor=1 case.
If and only if Σ is sign-similar to an inverse positive matrix, the maximal corank of the Frisch scheme is 1.

Proof: First assume that Σ^{-1} is positive. Then there is no convex combination of 2 of its columns that possibly contains a zero. Next, assume that Σ is not positive, but sign similar to a positive matrix. Then, from lemma 8, it follows that the columns belong to two opposite orthants. A zero can only be obtained by a convex combination of two columns that have an opposite sign pattern, say the i-th, s^i, and the j-th one, s^j. But it is easily verified that for such columns, always $s^i_j = s^j_i < 0$. From corollary 1, it follows that no convex combination of s^i and s^j will ever be allowed. □

6.4 The Wilson-Lederman bound

In this section, an important *genericity* result will be discussed about the solvability of the problem. On the one hand it contains a disappointing message while on the other hand it seems to indicate that Nature does not allow for too optimistic or naive modelling intentions.

Let's first perform some heuristic mathematics: We shall count the number of unknowns and the number of equations, following Thurstone's general principle of *overdeterminacy*, which states that " a scientific theory must be overdetermined by the data"[31]. Write $\Sigma = \hat{\Sigma} + \tilde{\Sigma}$ and assume that rank($\hat{\Sigma}$)=r. According to a theorem, historically attributed to Kronecker, the matrix $\hat{\Sigma}$ is of rank r if a r-rowed minor is non-zero whereas all 'bordered' minors of order $r+1$ vanish. The number of these bordered determinants is $(n-r)^2$. But because of the symmetry of $\hat{\Sigma}$, the number of conditions can be reduced to:

$$k_r = \frac{(n-r)(n-r+1)}{2}$$

The number of unknowns is n, namely equal to the number of communalities. The set of conditions can, *in general*, only be satisfied if $n \geq k_r$ or if :

$$n \geq \frac{(n-r)(n-r+1)}{2}$$

which, after some elementary manipulations becomes :

$$r \geq \mathrm{WL}(n) \equiv 1/2\{2n+1-\sqrt{8n+1}\}$$

The smallest possible value of r is therefore :

$$r \geq \mathrm{WL}(n) \equiv \lceil 1/2\{2n+1-\sqrt{8n+1}\} \rceil$$

where $\lceil (z) \rceil$ denotes the smallest integer greater than or equal to z. The abbreviation WLstands for 'Wilson-Ledermann'. Wilson was among the first to derive in 1929 the above expression [6],despite his own warning:

> There is perhaps no more tricky part of mathematics than that involved in counting equations and variables to determine whether or not the equations can in general be solved. Today this kind of mathematics is, among pure mathematicians, taboo except as a heuristic device.

Ledermann has succeeded in putting the derivation on a somewhat more rigorous footing [31] in 1937: The above deduction of the bound is incomplete because it must be shown that the k_r equations are independent, i.e. that none of them is a consequence of the others. The filling of this gap constitutes from a mathematical point of view the essential contribution of Ledermann's paper [31]. The bound is given for some small values of n in the following table.

n	2	3	4	5	6	7	8	9	10	11	12
WL(n)	1	1	2	3	3	4	5	6	6	7	8

However, the most general result about the Wilson-Ledermann bound, which has been proved rigorously is due to Shapiro :

Theorem 15 Generic Solvability of the Frisch Scheme.
With probability one, the rank of $\Sigma - \tilde{\Sigma}$ *cannot be reduced below the so-called Wilson-Ledermann bound :*

$$\mathrm{WL}(n) = [(2n+1) - \sqrt{8n+1}]/2$$

Proof : A complete and general proof can be found in [37] . □

A proof by Baratchart and Kalman, using Thom's topological transversality theorem has been announced [25]. WL(n) can be considered as almost surely a lower bound on the reduced rank of Σ. The set of symmetric $n \times n$ matrices for which the rank can be reduced below the Wilson-Ledermann bound WL(n) is *thin*, or to be more specific, it has *Lebesgue measure zero*. Shapiro has obtained this result omitting the nonnegative definiteness requirements of the Frisch scheme. Hence, specifically for the Frisch scheme, the lower bound on the minimal rank could still be more restrictive! The Wilson - Ledermann bound can be interpreted as counting the degrees of freedom for a certain rank to be generically (i.e. for arbitrary data

covariance Σ) attainable as a function of the number of variables. For a n-variable data covariance matrix, a rank r is attainable if and only if the number :

$$f(n,r) = n - k_r = 1/2(-n^2 + n(2r+1) - (r^2 - r)) \geq 0$$

If $f(n,r) = 0$, there are as many equations as unknowns and a finite number of solutions (for the communalities) is to be expected. If $f(n,r) > 0$ there will be an infinity of solutions for the communalities, the general solution having $f(n,r)$ 'degrees of freedom', i.e. involving arbitrary parameters. On the other hand, if $f(n,r) < 0$ no solution can be obtained unless the elements of Σ satisfy at least $-f(n,r)$ relations. As an example, consider the degrees of freedom of a problem with $n = 600$ variables. The case of **maxcor**= 35 or equivalently, **minr**= 565 is generically impossible, because $f(600,565) = -30$. However, **maxcor**= 34 (**minr**= 566) will generically have $f(600,566) = 5$ degrees of freedom. Table 6.1. summarizes the Wilson - Ledermann bound for low values of the number of variables $1 \leq n \leq 45$.

In relation to the Frisch scheme, the Wilson-Lederman bound and Shapiro's theorem, have three important consequences:

1. **Generic Solvability:** Suppose that one performs the following deductive simulation experiment. Start from an exact $m \times n$, $(m > n)$ matrix $\hat{A}$ of rank r and add some random noise to it in the form of a noise matrix $\tilde{A}$ so that the elementwise signal-to-noise ratio is really high. Then generically, one will not be able to recover the exact rank r from the measurement matrix $\Sigma = A^t A$, where $A = \hat{A} + \tilde{A}$, if r is smaller than the Wilson-Ledermann bound WL(n). Hence, for instance for the case $n = 10$, the maximal corank that is generically obtainable, is 4. However, especially when the signal to noise ratio is high, some other measures such as for instances the singular values of A (or the eigenvalues of Σ), may indicate that the matrix is very close (in Frobeniusnorm) to one of rank r. This shows that a severe requirement for the Frisch scheme to be useful in realistic identification problems, one has to have some a priori knowledge of the expected (co-)rank. Because if one chooses the number of measurement channels n too high, also the Wilson-Ledermann bound may become too high so that a low rank can not be achieved generically. Another point of view however could be that *linear relations between measurements* are only *generically* possible for sufficiently low coranks. Hence, a model which consists of a lot of linearly independent linear relations, might be too unrealistic, when a phenomenon is analysed via several measurements variables. Hence, the Wilson-Ledermann bound gives an lower bound on the complexity of a model!

2. **Stability of the solution set:** It also immediately follows that a reduced rank less than the Ledermann bound, cannot be stable in the sense that, when the elements of Σ are slightly changed, the matrix Σ can generically not be adjusted so that the reduced rank is preserved. Stability is usually expected for a rank $r \geq$ WL(n) because if Σ is a matrix for which there exists a matrix Σ such that $\Sigma - \tilde{\Sigma}$ satisfies the Frisch scheme requirements and is of rank $r \geq$ WL(n), then the whole neighbourhood of Σ is reducible to rank r [37]. This is not the case when the Wilson - Ledermann bound is not satisfied.

3. **Uniqueness of the noise covariance matrix $\tilde{\Sigma}$:** When $f(n,r) = 0$, there is a zero dimension solution (i.e. only distinct points) for the so-called communalities (the diagonal elements $\tilde{\sigma}_i$ of $\tilde{\Sigma}$). For example, $f(3,1) = 0$, hence there is a unique solution, which in this case, consists of one point in the parameterspace of the σ_i. Wilson [41] gives an

n	*attained corank* $= n - r$								
	1	2	3	4	5	6	7	8	9
$n = 1$	0	-	-	-	-	-	-	-	-
$n = 2$	1	-1	-	-	-	-	-	-	-
$n = 3$	2	0	-3	-	-	-	-	-	-
$n = 4$	3	1	-2	-6	-	-	-	-	-
$n = 5$	4	2	-1	-5	-10	-	-	-	-
$n = 6$	5	3	0	-4	-9	-15	-	-	-
$n = 7$	6	4	1	-3	-8	-14	-21	-	-
$n = 8$	7	5	2	-2	-7	-13	-20	-28	-
$n = 9$	8	6	3	-1	-6	-12	-19	-27	-36
$n = 10$	9	7	4	0	-5	-11	-18	-26	-35
$n = 11$	10	8	5	1	-4	-10	-17	-25	-34
$n = 12$	11	9	6	2	-3	-9	-16	-24	-33
$n = 13$	12	10	7	3	-2	-8	-15	-23	-32
$n = 14$	13	11	8	4	-1	-7	-14	-22	-31
$n = 15$	14	12	9	5	0	-6	-13	-21	-30
$n = 16$	15	13	10	6	1	-5	-12	-20	-29
$n = 17$	16	14	11	7	2	-4	-11	-19	-28
$n = 18$	17	15	12	8	3	-3	-10	-18	-27
$n = 19$	18	16	13	9	4	-2	-9	-17	-26
$n = 20$	19	17	14	10	5	-1	-8	-16	-25
$n = 21$	20	18	15	11	6	0	-7	-15	-24
$n = 22$	21	19	16	12	7	1	-6	-14	-23
$n = 23$	22	20	17	13	8	2	-5	-13	-22
$n = 24$	23	21	18	14	9	3	-4	-12	-21
$n = 25$	24	22	19	15	10	4	-3	-11	-20
$n = 26$	25	23	20	16	11	5	-2	-10	-19
$n = 27$	26	24	21	17	12	6	-1	-9	-18
$n = 28$	27	25	22	18	13	7	0	-8	-17
$n = 29$	28	26	23	19	14	8	1	-7	-16
$n = 30$	29	27	24	20	15	9	2	-6	-15
$n = 31$	30	28	25	21	16	10	3	-5	-14
$n = 32$	31	29	26	22	17	11	4	-4	-13
$n = 33$	32	30	27	23	18	12	5	-3	-12
$n = 34$	33	31	28	24	19	13	6	-2	-11
$n = 35$	34	32	29	25	20	14	7	-1	-10
$n = 36$	35	33	30	26	21	15	8	0	-9
$n = 37$	36	34	31	27	22	16	9	1	-8
$n = 38$	37	35	32	28	23	17	10	2	-7
$n = 39$	38	36	33	29	24	18	11	3	-6
$n = 40$	39	37	34	30	25	19	12	4	-5
$n = 41$	40	38	35	31	26	20	13	5	-4
$n = 42$	41	39	36	32	27	21	14	6	-3
$n = 43$	42	40	37	33	28	22	15	7	-2
$n = 44$	43	41	38	34	29	23	16	8	-1
$n = 45$	44	42	39	35	30	24	17	9	0

Table 6.1: Dimension (degrees of freedom) of the solution for the communalities as a function of the number of variables and the attained corank

example $f(6,3) = 0$, where there are 2 distinct numerical solutions. From the point of view of our uncertainty principle, this is rather disappointing. On the one hand, the Frisch scheme delivers us a set of infinitely many solutions (conjectured bounded and closed when normalized), but on the other hand the solutions that correspond to a maximal corank, have a set of communalities that are discrete points in the communality space.

Despite the fact that the Wilson-Lederman bound puts severe restritions on the number of linear relations that may generically be identified from linear relations, it may contain a serious warning about the use of *linear* models. Recall that, from the inspiration point of view, linearity is not generic in Nature, but that it is only a simplifying assumption of the modeller, that imposes some limitations on the techniques that are to be used. Does the Wilson-Lederman bound not suggest that the modeller can not exaggerate in this oversimplification of relatity?

However, in the light of the characterization of general intelligence, as considered by Spearman's factor analysis approach, the Wilson-Ledermann bound is reassuring. Frankly speaking, I would be scared if your and mine intelligence could be generically explained in terms of only one commom factor! In this sense, the Wilson-Ledermann bound ensures that human beings are *generically* not so simple. Of course, there may exist exceptions but they are certainly not stable.

6.5 Conclusions

In this chapter, we have been developing some algorithmic tools to describe th solution vectors that satisfy all necessary requirements of the Frisch scheme. The basic algorithms are the *telescoping approach* and the *brute force approach.* In the telescoping approach, one employs some well known results from linear algebra, in order to analyse the maximal corank, while in the brute force approach, we compute all orthant null invariant vectors with our algorithm described in chapter 2, for the computation of all nonnegative solutions to a set of linear equalities. Furthermore, we have summarized the known genericity results about the existence of a solution. Despite significant additional insight can be gained via these tools, into the geometrical nature of the solution set, Further research is needed in order to clarify completely the geometry.

Let's finally mention that recently the problem has been analysed in terms of dynamic systems [2] [10] [21] [34].

Bibliography

[1] Albert A.A. *The Minimum Rank of a Correlation Matrix.* Proc.Nat.Acad.Sci., 30, pp.144, 1944.

[2] Anderson B.D.O., Deistler M. *Identifiability in dynamic errors-in-variables models.* J. Time Series Analysis, 5, pp.1-13.,1984.

[3] Anderson T.W. *The 1982 Wald Memorial Lectures: Estimating Linear Statistical Relationships.* The Annals of Statistics, Vol.12, no.1, pp.1-45.

[4] Astrom K., Eykhoff P. *System Identification: A survey* Automatica, Vol.7., p.123-162.

[5] Beghelli S., Guidorzi R. *Problemi di Stima da Dati Affetti Da Rumore.* Atti dell'incontro nazionale dei ricercatori del progetto nazionale M.P.I., Como, Villa Olmo, June 1985.

[6] Bekker P.A., De Leeuw J. *The rank of reduced dispersion matrices.* Psychometrica, Vol.52, no.1, pp.125-135, March 1987.

[7] Berman A., Plemmons R.J. *Nonnegative matrices in the mathematical sciences.* Academic Press, New York, 1979.

[8] Bode H.W. and C.E. Shannon . *A simplified derivation of linear least square smoothing and prediction theory.* Proc. of the I.R.E. , 38, 417 - 425 (1950).

[9] Chatfield C. , Collins A. *Introduction to multivariate analysis.* Chapman and Hall Ltd., London 1980.

[10] Deistler M. *Linear Errors-in-variables systems.* In 'Time Series and Linear Systems.', ed. Sergio Bittanti. Lecture Notes in Control and Information Sciences (M.Thoma and A. Wyner, Eds), Springer-Verlag, 1986.

[11] De Moor B., Vandewalle J.. *A geometrical approach to the maximal corank problem in the analysis of linear relations.* Proc. of the 25th IEEE CDC Conference, Athens, Dec. 1986.

[12] De Moor B., Vandewalle J.. *The uniqueness versus the non-uniqueness principle in the identification of linear relations from noisy data.* Proc. of the 25th IEEE CDC Conference, Athens, Dec.1986.

[13] De Moor B., Vandewalle J. *All nonnegative solutions to sets of linear equalities and inequalities.* Proc. of the first International Conference on Industrial and Applied Mathematics, Paris, 29/06 - 3/07 1987.

[14] De Moor B., Vandewalle J.*A unifying theorem for linear and total linear least squares identification.*, June 1987, Accepted for IEEE Trans. on Automatic Control.

[15] Eckart C., Young G. *The approximation of one matrix by another of lower rank.* Psychometrika, Vol.1., 211-218, 1936.

[16] Eykhoff P. *System Identification, Parameters and State Estimation.* John Wiley and Sons, New York, 1974.

[17] Frisch R. *Statistical confluence analysis by means of complete regression systems.* Publication no.5.,University of Oslo Economic Institute, 192 pages, 1934.

[18] Gantmacher F.R.*Theorie des Matrices.* Tome 1/2, Dunod, Paris, 1966.

[19] Golub G., Van Loan C. *Matrix Computations.* North Oxford Academic, J.Hopkins Univ. Press, 1983.

[20] Goodwin G.C., Sin K.S. *Adaptive filtering.* Prentice Hall, Information and System Sciences, T.Kailath (Series Editor), Englewood Cliffs, NJ, 1984.

[21] Green M., Anderson B.D.O. *Identification of multivariable errors-in-variables with dynamics.* IEEE Trans. Automatic Control, 1987.

[22] Kailath T. (Editor).*Linear Least Squares estimation.* Benchmark papers in Electrical Engineering and Computer Science/17, Dowden, Hutchinson and Ross, Inc., 1977.

[23] Kalman R.E. *Identification from real data.* In ' Current developments in the Interface : Economics, Econometrics, Mathematics '(edited by M.Hazewinkel and A.H.G. Rinrooy Kan), D.Reidel Publishing Co., Dordrecht, pages 161-196, 1982.

[24] Kalman R.E. *System Identification from noisy data.* Proc. Int. Symp.on Dynamical Systems (Gainesville, Florida 1981), edited by A.Bednarek, Academic Press, 1982.

[25] Kalman R.E., Los C.A. *The Prejudices of Least Squares, Principal Components and Common Factor Schemes.* 6th International Conference on Mathematical Modeling, St. Louis, Missouri, USA, August 1987.

[26] Kalman R.E. *Identifiability and modeling in econometrics.* In P.R. Krishnaiah (Ed.). Developments in Statistics 4. New York, Academic Press, 1983.

[27] Kalman R.E. *We can do something about multicollinearity!* Communications in Statistics - Theory and Methods. 13, 115 - 125, 1984.

[28] Kalman R.E., Los C.A. *The Prejudices of Least Squares, Principal Components and Common Factor Schemes.* 6-th International Conference on Mathematical Modelling, St.Louis, Missouri, USA, August 1987.

[29] Koopmans T.C. *Linear regression analysis of economic time series.* De Erven F.Bohn NV, Haarlem, 1937.

[30] P. Kovanic. *A new theoretical and algorithmic basis for estimation, identification and control.* Automatica, Vol.22, no.6, 1986, pp.657-674.

[31] Ledermann W. *On the rank of the reduced correlation matrix in multiple factor analysis.* Psychometrika, 1937, 2, 85-93.

[32] Luenberger D.G. *Introduction to dynamic systems: Theory, models and applications.* John Wiley and Sons, New York, 1979.

[33] Pearson K. *On lines and planes of closest fit to points in space.* Philosophical Magazine, 2 , 559-572, 1901.

[34] Picci G., Pinzoni S. *A new class of dynamic models for stationary time series.* In 'Time Series and Linear Systems.' Ed. Sergio Bittanti. Lecture Notes in Control and Information Sciences (M.Thoma, A Wyner Eds.), Springer Verlag, 1986.

[35] Reiersol O. *Confluence analysis by means of lag moments and other methods of confluence analysis.* Econometrica, 9, 1-24, 1941.

[36] Reiersol O. *On the identifiability of parameters in Thurstone's multiple factor analysis.* Psychometrika, Vol.15, no.2, June 1950.

[37] Shapiro A. *Rank reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis.* Psychometrika, 47-2, 187-199, 1982.

[38] Spearman C.E. *General intelligence objectively measured and defined.* American Journal of Psychology, 15, pp.17-19, 1904.

[39] Willems J.C. *From data to models.* Automatica. Part I:vol.22, no.5, pp.561-580, 1986. Part II:vol.22, no.6, pp. 675-694, 1986. Part III:vol.23, no.1, pp.87-115, 1987.

[40] Willems J.C. *Models for dynamics.* Subm. to Dynamics Reported. October 1986.

[41] Wilson E.B., Worcester J. *The resolution of six tests into three general factors.* Proc. of the National Academy of Sciences, 25:73-77, 1929.