Identification of Linear Systems in Noisy Data

by

Sekhar Chandra Tatikonda

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Bachelor of Science in Electrical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

January 1993

Author .....
Department of Electrical Engineering and Computer Science
January 20, 1993

Certified by .....
Sanjoy Mitter
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Leonard A. Gould
Chairman, Departmental Committee on Undergraduate Students

Identification of Linear Systems in Noisy Data

by

Sekhar Chandra Tatikonda

Submitted to the Department of Electrical Engineering and Computer Science
on January 20, 1993, in partial fulfillment of the
requirements for the degree of
Bachelor of Science in Electrical Engineering

Abstract

All methods used to determine the number of linear relations and the linear relations themselves in a cloud of data are based on assumptions. These assumptions can include specific information known about the system but in general the biases are *a priori* assumed without any justification accept that it makes the identification process computationally manageable. Recently Professor Rudolf Kalman discovered a bias free identification method which he calls generalized least squares (GLS). In this paper we explore the structure of all GLS solutions and relate these solutions to existing identification methods. We show that all GLS solutions are in fact total least squares (TLS) solutions. Finally we discuss the problems of using maximum likelihood as a performance criterion over the GLS scheme.

Thesis Supervisor: Sanjoy Mitter
Title: Professor of Electrical Engineering and Computer Science

Contents

Chapter 1

Introduction

1.1 Identification Problem and Results

The most basic identification problem is that of determining first the number of linear relations in a cloud of data and second determining what these relations are. Most existing identification methods rely upon assumptions. In fact we have already assumed that the relations are linear. Some of these assumptions incorporate specific information known about the model and are therefore reasonable. We can call such reasonable assumptions constraints. But in most identification cases the assumptions have no *a priori* explanations. The only reason one chooses these assumptions is either ignorance of the basic identification problem or ease of computation of a given solution scheme. In this thesis we hope to discern what the underlying assumptions are for different identification methods.

In the case where there is no noise in the data we can uniquely determine the number of relations and the relations themselves. When we add noise, though, it not clear that we can do this identification uniquely. By a continuity argument we believe that as the noise in the system goes to zero we should be able to determine the relations uniquely. We believe that if the noise in the system is smaller than some value then the system can be identified uniquely. For the case where the exact data lies on a hyperplane (an $n-1$ dimensional subspace in R^n) we can, given suitable noise constraints, determine the relationship exactly. That is we can determine the number

of relations to be one exactly. The exact relation itself may still be ambiguous. It is not so clear, though, how to determine the system when the exact system lies on lower-dimensional subspace of R^n.

The two fundamental questions involved in identifying linear relations in noisy data are: first, how many relations the exact data supports (here on called the rank of the system), and second, what are the relations themselves. These two questions are clearly not independent of each other. For example even if we are able to identify the rank of the system there are many different relations (i.e. planes) that have that same rank. On the other hand if we are able to determine the relations themselves then we can uniquely identify the rank. Similarly given the relation there are many different noises we can place on the data. But for a given noise there is still only one relation, the exact relation. In summary we may never know the exact relations but we may be able to identify the exact rank. For most of this paper we will be interested in answering the first question of determining rank independently of the second question of determining the exact relations.

The purpose of this paper to describe the assumptions of different identification problems. First we discuss the assumptions of existing identification methods. Then we examine some of the problems with the so-called Frisch identification method. Following the work of Professor R. Kalman (Kalman, 1990) we study Kalman's generalized least squares (GLS) method. We explore the structure and provide a new canonical structure for all GLS solutions. We then relate the GLS method to other known methods like total least squares (TLS) and principal component decomposition. Finally we give a description of the maximum likelihood method as a performance criterion over a scheme.

1.2 Problem Formalization

Let us formalize this identification problem.

Definition 1 *Define the cloud of data points to be the set* $\{x_t \in R^n \mid t = 1, ..., T\}$.

Here, n is the number of components in any measurement and T is the total number of measurements. We assume $T > n$. Let X be an $n \times T$ matrix of data. That is $X = [x_1, x_2, ..., x_T]$. Let $\Sigma = \frac{1}{T}\Sigma_t x_t x_t' = \frac{1}{T}XX'$. Later we will see that this is the covariance matrix of the sample data.

Following Kalman's notation let q equal the rank of the system. That is q is the number of linear relations supported by the data. For example if $q = 1$ then the exact data lies on an $n-1$ dimensional plane in R^n. If $q = n-1$ then the exact data lies on a one-dimensional subspace in R^n.

It is impossible for any identification scheme to have no underlying assumptions. We need to make some assumptions on the structure of the noise before we can hope to analyze this problem. The following assumptions will hold throughout this paper:

Assumption 1: Assume $0 < q < n$. That is assume the exact part of the data does indeed support q linear relations. We do not allow $q = 0$ or n because these are degenerate cases. When $q = 0$ there are no relations. The data lies in all of R^n. When $q = n$ the exact data lies on a point in R^n.

Assumption 2: We will not assume *a priori* any probabilistic nature to the noise in the system. The reason for this is we are not sure what is causing the noise. If the data represents repeated measurements of some physical experiment then we may assume the noise is probabilistic, say following a Gaussian distribution. But if the data is economic in nature then there is no notion of repeatability of the experiment. For different economic measurements it is very hard to keep the conditions the same, whereas for physical measurements it is not so hard to keep the conditions the same. Thus in the economic data example the noise may not be purely random but instead caused by changes in the overall system that have occurred between measurements.

Assumption 3: We will assume the noise is additive. This is a prejudice but most everyone accepts this assumption. Maybe in future analysis one can assume the noise has some other form. Additive noise implies we can decompose $x_t = \hat{x}_t + \tilde{x}_t$ for all t. Here $\hat{x}_t$ is the signal part of the data and $\tilde{x}_t$ is the noise part of the

data. Thus we also see that $X = \hat{X} + \tilde{X}$, where $\hat{X}$ and $\tilde{X}$ denote the true and noisy parts of the data respectively.

Assumption 4: We further assume the exact part of the data and the noisy part of the data are orthogonal. That is $\Sigma_t \hat{x}_t \tilde{x}_t' = \hat{X}\tilde{X}' = 0$. Note that this orthogonality in T-space. This is our strongest assumption. Ordinary least squares (OLS) assumes this data orthogonality implicitly. In fact for OLS this orthogonality result follows directly from the Gauss-Markov theorem. From now on we will call this assumption the *data orthogonality assumption*.

Assumption 5: We will assume that all the measurements have zero mean. That is $\Sigma_{t=1}^{T} x_t = 0$. Similarly $\Sigma_t \hat{x}_t = 0$ and $\Sigma_t \tilde{x}_t = 0$. The transformation from a given data set to a zero-mean data set is just an affine transformation. Thus this assumption is unnecessary for the calculations we will be making. It makes the problem, though, easier if we can identify an origin in this way. Note that Σ is a true sample covariance matrix. To see this note that the covariance matrix between the i^{th} and the j^{th} component equals $var(y_i, y_j) = \Sigma_t y_{it} y_{jt} - \bar{y}_i \bar{y}_j$. (Here y_i is a $T \times 1$ vector containing all the component i measurements.) But $\bar{y}_i \bar{y}_j = 0$. So Σ is the sample covariance.

By assumption 3 we see that the covariance matrix $\Sigma = \frac{1}{T} XX'$ can be decomposed into $\Sigma = \hat{\Sigma} + \tilde{\Sigma}$ where $\hat{\Sigma} = \frac{1}{T}\hat{X}\hat{X}'$ and $\tilde{\Sigma} = \frac{1}{T}\tilde{X}\tilde{X}'$. To see this note that $\Sigma = \frac{1}{T}XX' = \frac{1}{T}(\hat{X} + \tilde{X})(\hat{X} + \tilde{X})' = \frac{1}{T}\hat{X}\hat{X}' + \frac{1}{T}\tilde{X}\tilde{X}' + \frac{1}{T}\hat{X}\tilde{X}' + \frac{1}{T}\tilde{X}\hat{X}' = \hat{\Sigma} + \tilde{\Sigma}$ because the last two addends are zero by data orthogonality.

Let A be an $n \times q$ matrix with full rank q. For a given $\hat{\Sigma}$ and a given q if $A'\hat{\Sigma} = 0$ then we know the columns of the matrix A are q linearly independent basis vectors that span the null space of $\hat{\Sigma}$. The columns of A can be thought of as the normals to q different hyperplanes ($n-1$ dimensional planes in R^n). The intersection of these q hyperplanes contains the subspace that the exact data determine. The dimension of this intersection is $n - q$. This is also the rank of $\hat{\Sigma}$.

The basic identification problem can thus be restated: Given a sample data covariance matrix Σ determine uniquely if possible a factorization $\Sigma = \hat{\Sigma} + \tilde{\Sigma}$ that "best

fits" the data. From a given $\hat{\Sigma}$ we can determine $q = corank\hat{\Sigma}$ (corank = n – rank) and we can determine A the matrix of linear relations. A is just any matrix whose columns span the null space of $\hat{\Sigma}$. Note we cannot determine A uniquely. We can though determine A uniquely up to *similarity*.

Definition 2 *Two matrices $A^{n\times q}$ and $B^{n\times q}$ are called similar if there exists an invertible matrix $U^{q\times q}$ such that $A = BU$. We denote similarity by $A \sim B$.*

This definition can be found in Kalman, 1990, Definition 3.22.

Note that $0 = A'\hat{\Sigma} = UB'\hat{\Sigma} \Rightarrow B'\hat{\Sigma} = 0$. The null space of $\hat{\Sigma}$ determines the exact solution plane.

We can restate the problem more abstractly: For any positive definite, symmetric matrix Σ we want to decompose it into two nonnegative definite symmetric matrices $\hat{\Sigma}$ and $\tilde{\Sigma}$. We need to find the decomposition that "best fits" the data.

As stated now the problem has no unique solution (Kalman, 1990, Theorem 3.15). For any given Σ and any matrix $A^{n\times q}$ and any q we can find a decomposition of Σ such that $A'\hat{\Sigma} = 0$. We will prove this in Theorem 9. Obviously this result is of no use to us. We must describe what "best fit" means. To do this we need to add constraints to the problem. We call a set of constraints (that determine the *prejudice* of a given method) a *scheme*. Thus a scheme is a set of rules that help us decompose Σ. Sometimes schemes are determined by knowledge of the data but unfortunately most schemes are chosen arbitrarily. In most cases, though, there is no *a priori* reason for choosing a given scheme over another scheme. Schemes impose structure on the relationships in the data. For example the elementary least squares (ELS) scheme implies that $n-1$ of the variables are exact and only one of the variables is noisy. Thus we see the ELS scheme imposes a structure on the noise.

Given a scheme there may be many solutions to a given identification problem. We will need a way to distinguish the best solution over all the solutions in a given scheme. We will call any such criterion a *performance criterion*. That is we want to find the optimum solution over a scheme. For example one performance criterion is the norm of $\tilde{\Sigma}$ which can be minimized. Note that performance criteria are different

than schemes in that they do not impose any structural relations on the data. Instead they optimize some criterion over a set of possible solutions.

Chapter 2

Existing Schemes

This chapter will describe existing schemes for identification of systems in noisy data. We then discuss their respective biases. Most identification schemes rely on some sort of regression type method. Here we will discuss elementary least squares (ELS), ordinary least squares (OLS), and principal components (SVD). We will then discuss the Frisch scheme. (Much of this chapter is a summary of chapters three through five of Kalman's Nine Lectures on Identification).

2.1 Elementary Least Squares

The usual setup for an ELS problem is the following. Find the best $A^{(n-1)\times 1}$ such that we minimize the squared 2 norm of $Y^{T\times 1} - X^{T\times(n-1)}A$. We have assumed that of the n measured variables $n-1$ of them are exact. That is X is exact and Y is noisy. The well known solution is $A = (X'X)^{-1}X'Y$. Note that in the ELS scheme there are n different solutions. In general practical considerations will tell help us determine the noisy component but in the absence of such evidence we need a performance criterion to tell us which of the n solutions is "best." In our new notation the ELS scheme can be rewritten as: For any given Σ an ELS scheme is a decomposition $\Sigma = \hat{\Sigma}+\tilde{\Sigma}$. Where $\tilde{\Sigma}$ has all zero non-diagonal components and only one nonzero diagonal component. Thus we see that there are at most n ELS solutions. We will show later that there are indeed n ELS solutions satisfying this notational scheme. There are three main

assumptions in this scheme:

1. We assume that some of the variables are exact.

2. Often times people make the mistake of assuming the noisy variable is caused by the exact variables. They believe by choosing one variable inexact they are in some sense imposing causality on the system. In general we do not have any sense of causal relation in noisy data.

3. We assume that there is only one relation in the data. That is we are assuming *a priori* that the data lies on an $n-1$ dimensional subspace (i.e. $q=1$).

Note also that we need to decide which of the n variables is noisy. Again, this is often decided on practical considerations. This could also be considered a fourth assumption because we need to decide which one is the best solution. If we consider a performance criterion as part of the determination of unique solution then this may not be a assumption because the performance criterion will determine the best solution over a scheme.

2.2 Ordinary Least Squares

One way to deal with the $q=1$ bias is to use the OLS scheme. In this scheme we try to find that $A^{(n-q)\times q}$ that minimizes the squared two norm of $Y^{T\times q} - X^{T\times(n-q)}A$. That is we have q noisy variables dependent on $n-q$ exact variables. Once again the solution is $A=(X'X)^{-1}X'Y$. For a given q there are n combination q solutions. In our notation the OLS scheme is: for any Σ we want to find a decomposition such that $\tilde{\Sigma}$'s upper left $q\times q$ matrix is nonzero but the rest of $\tilde{\Sigma}$ must be zero. Note that OLS schemes contain the ELS schemes.

It is a remarkable fact that the solution to our notational scheme is the same solution as that of the OLS minimization problem. Kalman proves that the two methods are equivalent (Kalman, 1990, Theorem 3.3).

The OLS scheme also suffers from the same biases as ELS. We assume that some of the variables are exact, we choose which ones will be exact and inexact, and we choose the value of q all *a priori*.

The following theorem will shed some light on the structure of the matrices A.

Theorem 1 *The matrix Σ^{-1} contains the regression coefficients for the OLS solutions. That is any q rows of Σ^{-1} will be a regression solution. Let the columns of A contain any q rows of Σ^{-1}. In particular the rows of Σ^{-1} are the n different $q = 1$ solutions.*

Proof: See Kalman, 1990, Theorem 3.3.

Note that as a result of this theorem we see that all OLS solutions where $q > 1$ are just combinations of ELS solutions. That is the solution space for a given OLS scheme is in the intersection of q ELS solution spaces. We will later explore the simplex constructed from the positive linear combinations of the ELS regression coefficients.

2.3 Principal Components

Principal components scheme (SVD) does not assume some of the variables are exact. Instead it assumes each component is noisy. Because of this this scheme is sometimes called the total least squares (TLS) method or the error-in-variables method. The SVD scheme is formulated as follows. Let the eigenvalue-eigenvector decomposition of $\Sigma = U\Lambda U'$, where U is the orthonormal matrix of eigenvectors and Λ is the diagonal matrix of eigenvalues. In the SVD solution we decompose Λ into two diagonal matrices. For a given q we select q of the eigenvalues of Λ to represent those principle components that make up the noise of the observed signal and the remaining $n - q$ eigenvalues to represent those principle components that make up the true signal of the observed signal.. We can decompose $\Lambda = \hat{\Lambda} + \tilde{\Lambda}$. $\tilde{\Lambda}$ consists of q of the diagonal components of Λ. Thus $\hat{\Sigma} = U\hat{\Lambda}U'$ and $\tilde{\Sigma} = U\tilde{\Lambda}U'$.

Note that for a given q there are n combination q different SVD solutions that we can choose from. Most people, though, choose the q smallest eigenvalues of Σ to

be the q nonzero eigenvalues of $\tilde{\Sigma}$. One possible performance criterion over the SVD scheme would be to minimize the two norm of $\tilde{\Sigma}$.

In the SVD scheme we have been able to eliminate two of the biases of the OLS scheme. In the SVD scheme we no longer assume some of the variables are exact or inexact. Instead we assume all the variables are noisy and subtract out an error from each of them. We also remove the problem of deciding which q variables are noisy but we add the bias of assuming which q eigenvalues correspond to the noisy principal components. We also have the problem of determining q. In both OLS and SVD we have to *a priori* determine q. It turns out that if we have extra information about the system in the form of a norm limitation on the noise then we can determine the $q = 1$ case uniquely. That is if the exact data only supports one relation and we have suitable information about the size of the variances of the noisy variables we can determine $q = 1$ uniquely. We will prove this in Theorem 4.

For now though let us see how the SVD solutions are related to the OLS solutions. Is the SVD solution a linear combination of the OLS solutions? A more general question we can ask is where does the SVD solution lie in the space of all solutions. (We know from Theorem 9 in chapter 3 that any matrix $A^{n \times q}$ is a solution.) The SVD solution is a linear combination of the ELS solutions. But this is obvious because the n different ELS solutions form a basis for the space of R^n. This is because Σ^{-1} is nonsingular.

Lemma 1 *The SVD solution does not in general lie in the simplex of OLS solutions.*

Proof: For a given q we need to find an SVD solution that does not lie in the simplex of the n combination q OLS solutions. In particular we will show that there exists a $q = 1$ SVD solution that does not lie in the simplex of the n ELS solutions. We know $A'\hat{\Sigma} = 0$ where A is an $n \times 1$ matrix. So $0 = A'\hat{\Sigma} = A'(\Sigma - \tilde{\Sigma}) \Rightarrow A'\Sigma = A'\tilde{\Sigma} \Rightarrow A = \Sigma^{-1}\tilde{\Sigma}A \Rightarrow A = \Sigma^{-1}(\lambda_i * u_i)$ where λ_i and u_i are the eigenvalue and eigenvector pair we have chosen for this SVD scheme. Now u_i need not contain all positive components; therefore A is not in the positive linear combination of the ELS solutions. (At least one of the eigenvectors of Σ cannot be all positive.) The SVD

solution therefore need not lie in the simplex of the ELS solutions.

2.4 Frisch Scheme

So far all the schemes we have discussed require us to *a priori* determine q. For certain Frisch scheme problems we can determine q uniquely. The Frisch scheme is any decomposition of Σ where $\tilde{\Sigma}$ is diagonal. This diagonalization constraint implies that even the noise variables are uncorrelated. With this new condition it becomes much harder to determine different q solutions. In fact we can find conditions where certain values of q cannot exist.

As we have discussed before the ELS regressions can be used to define a simplex. All solutions in this simplex are $q = 1$ Frisch solutions. Furthermore for any such Σ that supports such a solution there are no other $q > 1$ Frisch solutions. Thus for this simplex only $q = 1$ solutions can exist. The proof of this is given in Theorem 2.

It turns out that if Σ^{-1} has only positive components then a $q = 1$ Frisch solution is the only possible solution. That is no solutions where $q > 1$ can exist. We can also show that for any Σ there is a $q = 1$ solution. To see this note the ELS solution is a Frisch solution. It can also be shown that if Σ^{-1} does not have positive entries (or cannot be transformed into one) then there not only exists a $q = 1$ case but there also exists a $q > 1$ solution. See Kalman, 1982. Thus for the Frisch scheme we can uniquely determine a $q = 1$ solution when $q = 1$ is the only solution possible. All Σ's support a $q = 1$ solution though. If on the other hand we require that for a given Σ the best solution is the $\tilde{\Sigma}$ with maximum corank then the positivity constraint holds always for $q = 1$. If we require maximum corank solutions then all $q = 1$ solutions will have positive Σ^{-1} matrices and vice-versa. One reason for insisting on maximum corank solutions comes up when we deal with endogenous versus exogenous variables. We would like to explain the most data with the least number of independent variables. Thus by maximizing corank we are maximizing the number of endogenous variables and minimizing the number of exogenous variables. This is just a version of "Occam's Razor" or the principle of parsimony.

These results are summarized in the following theorem:

Theorem 2 *Suppose all the entries of $L\Sigma^{-1}L$, where L is a diagonal matrix consisting of 1s and -1s, are positive numbers.*

1) *Then the only solutions that exist are $q = 1$ solutions.*

2) *Every coefficient vector A associated with a solution of the problem has positive entries and is a positive linear combination of the elementary regression vectors, that is $A' = \Sigma b_j s^j$, $b_j > 0$ for $j = 1, ..., n$, $s^j = jth$ column of Σ^{-1}.*

3) *Any positive linear combination of the rows of Σ^{-1} induces a solution of the problem with $\tilde{\Sigma} = diag(\frac{b_1}{a_1}, ..., \frac{b_n}{a_n})$ where the b_j's are the same as in part 2.*

Proof: See Kalman, 1982, page 149. One can see that it depends strongly on the Perron-Frobenius theorem.

Note that the positivity of Σ^{-1} is not a necessary condition for the $q = 1$ case. The following example is a case where the existence of a $q = 1$ solution does not imply that Σ^{-1} has positive components.

Let

$$\Sigma = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

Σ is positive definite because its eigenvalues are $4, 1$, and 1.

We see then that

$$\Sigma^{-1} = \begin{pmatrix} .75 & -.25 & -.25 \\ -.25 & .75 & -.25 \\ -.25 & -.25 & .75 \end{pmatrix}.$$

This cannot be transformed into a strictly positive component matrix. We can easily find, though, a $q = 1$ solution. Let

$$\tilde{\Sigma} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Thus we see that $\hat{\Sigma}$ has corank = 1. We can also easily find a $q = 2$ solution by allowing

$$\tilde{\Sigma} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Once again if we require maximum corank then $q = 2$ is the only solution for the above example. Note that for $n = 3$ the Frisch scheme is completely solved. If the positivity constraint holds then $q = 1$; otherwise $q = 2$.

Complete Frisch solutions can be found in the $n = 3, q = 2$ case. But in general the solution spaces are very intractable. In the cases where Σ^{-1} is not transformable to a positive component matrix I explored the possibility of a norm limitation on the noise covariance matrix in determining whether $q = 1$ or $q = 2$ for the $n = 3$ case. The computations are difficult and laborious. I also do not think they are valuable. Note that even if Σ^{-1} has positive components and we know $q = 1$ solution is the only solution that can exist we still need to find the relation. There are sadly an infinite number of them. Any solution in the simplex is an acceptable solution. So even if we can determine q uniquely we cannot determine the relation uniquely. In fact the dimension of any solution space is $n-q(q+1)/2$. To see this note that in general there are $q(q+1)/2$ degrees of freedom when solving the decomposition problem. For the Frisch case we need only determine n values. (These are the diagonal values of $\tilde{\Sigma}$). Therefore the combined solution space has dimension $n - q(q+1)/2$. A performance criterion can help us determine which solution to choose.

The main assumption of the Frisch scheme is that the noise covariance matrix is diagonal. This is a very strong assumption. But as a result of this assumption we are able to sometimes determine q uniquely. This is a theme that will come up again in this paper. If one wants to determine a solution uniquely then one will have to impose strong constraints on the system. We will show in chapter four that the Frisch scheme also fails to yield linear projection matrix.

2.5 Nature of Solutions

The discussion of the Frisch scheme sheds light on the definition of a solution. We need to define what we mean by unique solution. In general there are two types of questions we need to ask:

1) What conditions on Σ (or Σ^{-1}) and $\tilde{\Sigma}$ are needed given all other constraints such that a $q = k$ solution exists?

2) What conditions are needed given all the constraints such that a $q = k$ solution exists and no other $q = i$ $(i \neq k)$ solutions exist?

The first question is an existence question whereas the second is a uniqueness question. For OLS and SVD we can find any q solutions. We cannot, though, find uniqueness of solutions. This may be the job of the performance criterion. For the Frisch scheme, though, we can find solutions that satisfy both questions above.

Chapter 3

Schemes in General

3.1 General Theorems

So far we have been discussing different identification schemes. We have shown, though, that they all have built-in assumptions (or prejudices.) Let us now find properties that all schemes must have. By doing this we hope to understand the basic structures of all solution schemes. In general a scheme is any method of decomposing $\Sigma = \hat{\Sigma} + \tilde{\Sigma}$ where $corank\hat{\Sigma} = q$.

We are trying to find conditions on the noise such that as the noise goes to zero we can determine uniquely $\hat{\Sigma}$ and q. To that end we describe three basic results that all schemes satisfy.

Theorem 3 *Fix any q, $0 < q < n$. Let $\lambda_1(\Sigma) \leq ... \leq \lambda_n(\Sigma)$, where $\lambda_i(\Sigma)$ denotes the ith smallest eigenvalue of Σ. Then for all schemes such that corank $\hat{\Sigma} = q$ and rank $\tilde{\Sigma} = s$ ($s \geq q$, it will be later shown that if $s < q$ then no solution can exist) we have*

$$\lambda_1 + ... + \lambda_q \leq trace\tilde{\Sigma} \leq \lambda_{n-s+1} + ... + \lambda_n$$

and

$$\lambda_q \leq \| \tilde{\Sigma} \| \leq \lambda_n.$$

Proof: See Kalman, 1990, theorem 4.5.

Theorem 4 *Let* $\|\tilde{\Sigma}\| \leq \epsilon$ *for some* $\epsilon > 0$. *Then there exist* $q = 1, ..., k$ *solutions iff* $\epsilon \geq \lambda_k(\Sigma)$.

Proof: From the preceding theorem and the fact that the two norm is equal to the largest eigenvalue we see that if the largest eigenvalue of $\tilde{\Sigma}$ is λ_k then we can subtract from Σ those k principal components. Thus we have shown the existence of a $q = k$ solution.

Theorem 5 $\lambda_k(\Sigma) \geq \lambda_k(\hat{\Sigma}) \geq \lambda_k(\Sigma) - \lambda_n(\tilde{\Sigma})$.

Proof: This is an easy extension of Wilkinson's theorem (Wilkinson, 1965, page 102).

In theorems three and four equality holds if we are using the SVD scheme. For both of these two theorems we see that noise constraints can help us determine q. For example if $\lambda_1 < \epsilon < \lambda_2$ then we know $q = 1$ is the only value of q that can hold. The data will support only one relation. This relation though is not unique. Note that if $\lambda_1 < \epsilon < \lambda_3$, we cannot tell if there is a $q = 1$ or a $q = 2$ solution. Note that if we are insisting on maximum corank then $q = 2$ is the only solution. It is important to realize that if the true data supports only one relation then as we reduce the noise we will be able to determine $q = 1$ solution uniquely. This is because as the noise goes to zero the norm of the noise covariance matrix will eventually get smaller than the second smallest eigenvalue of $\hat{\Sigma}_{true}$. For a $q = 2$ solution we see that even if the noise goes to zero the norm of the noise covariance matrix will only get smaller than the third smallest eigenvalue of $\hat{\Sigma}_{true}$. We cannot know until the noise equals zero if none or one of the two smallest eigenvalues of Σ represents true data. Therefore we cannot determine whether $q = 1$ or $q = 2$. If we insist on maximum corank then $q = 2$ is the only solution for this case.

The following theorem explains a little bit about the structure of solutions. This theorem is an extension of one given in Kalman, 1982, page 149.

Theorem 6 *Suppose* Σ *and* $\tilde{\Sigma}$ *are given. Then the maximum eigenvalue of* $\tilde{\Sigma}\Sigma^{-1}$ *is* 1 *and it has geometric multiplicity = algebraic multiplicity = q. All its other eigenvalues belong to* $[0, 1)$.

Proof: First we show that the eigenvalues (the roots) of $\det(\Sigma - \lambda\tilde{\Sigma}) = 0$ are in bijective correspondence to the eigenvalues of $\det(\sigma I - \tilde{\Sigma}\Sigma^{-1}) = 0$.

$$\det(\Sigma - \lambda\tilde{\Sigma}) = \det(I - \lambda\tilde{\Sigma}\Sigma^{-1})\Sigma = \lambda \det\Sigma \det(\frac{1}{\lambda}I - \tilde{\Sigma}\Sigma^{-1}).$$

Let $\sigma = \frac{1}{\lambda}$. Therefore $\lambda_{min} = \sigma_{max}$. To show that $\sigma_{max} = 1$ we must show that $\lambda_{min} = 1$.

Assume toward a contradiction that $\lambda_{min} < 1$. If $x \in$ nullspace of $\tilde{\Sigma}$ then $x'\tilde{\Sigma}x = 0$ and $x'\lambda\tilde{\Sigma}x = 0$. If $x \notin$ nullspace of $\tilde{\Sigma}$ then $x'\lambda_{min}\tilde{\Sigma}x < x'\tilde{\Sigma}x$.

Thus we see that if $x \in$ nullspace of $\tilde{\Sigma}$ then $x'\hat{\Sigma}x = x'\Sigma x - x'\lambda\tilde{\Sigma}x > 0$. If $x \notin$ nullspace of $\tilde{\Sigma}$ then $x'\hat{\Sigma}x = x'\Sigma x - x'\lambda\tilde{\Sigma}x > x'\Sigma x - x'\tilde{\Sigma}x \geq 0$. Thus we see that $x'\hat{\Sigma}x > 0$ but this contradicts the fact that $\hat{\Sigma}$ is strictly nonnegative definite.

Since $\hat{\Sigma}$ has a nullspace of size q we see that the algebraic multiplicity of $\lambda = 1$ eigenvalue = geometric multiplicity of $\lambda = 1$ eigenvalue$= q$. All other values of $\lambda \in (1, \infty)$. Thus $\sigma = \frac{1}{\lambda} \in [0, 1)$. The proof is done.

Lemma 2 *The rank* $\tilde{\Sigma} \geq q$.

Proof: If this were not the case then $\tilde{\Sigma}\Sigma^{-1}$ could not have q eigenvalues equal to 1 by theorem six.

The matrix $\tilde{\Sigma}\Sigma^{-1}$ is very important because it represents a projection-like matrix. All its eigenvalues are between zero and one. If the rank $\tilde{\Sigma} = q$ then $\tilde{\Sigma}\Sigma^{-1}$ will have q eigenvalues $= 1$ and $n - q$ eigenvalues $= 0$. Thus we see that when $\tilde{\Sigma}$ has rank q the matrix $\tilde{\Sigma}\Sigma^{-1}$ will indeed be a projection matrix.

If $\tilde{\Sigma}\Sigma^{-1}$ contains only positive components then $q = 1$ is the only solution. This follows from the Perron-Frobenius theorem on positive matrices. In general this is not very useful because we don't know what $\tilde{\Sigma}$ is. But for the Frisch scheme we know that $\tilde{\Sigma}$ will always have nonnegative components. Thus we see how the positivity of Σ^{-1} can be exploited in the Frisch scheme.

Unfortunately there does not exist much theory on how to "force" a matrix to have its leading eigenvalue simple. One idea that is open for research is to treat $\tilde{\Sigma}$ as a perturbation matrix. We know the eigenvalues of Σ^{-1} and we would like to see

how the largest eigenvalue changes when we multiply it by the perturbation $\tilde{\Sigma}$. This is not an easy exercise because we cannot decompose $\tilde{\Sigma}$ into $I + \epsilon H$ where H is some perturbation matrix.

It turns out that we can reformulate the problem into solving a set of inequalities. The problem as we have it now is: Decompose Σ into two nonnegative definite matrices.

Lemma 3 $\hat{\Sigma} - \hat{\Sigma}\Sigma^{-1}\hat{\Sigma} = \tilde{\Sigma} - \tilde{\Sigma}\Sigma^{-1}\tilde{\Sigma}$.

Proof: See Kalman, 1990, page 2.5.

Theorem 7 $\Sigma = \hat{\Sigma} + \tilde{\Sigma}$, $\hat{\Sigma} \geq 0$, *and* $\tilde{\Sigma} \geq 0$ *iff* $\hat{\Sigma} \geq \hat{\Sigma}\Sigma^{-1}\hat{\Sigma}$ *iff* $\tilde{\Sigma} \geq \tilde{\Sigma}\Sigma^{-1}\tilde{\Sigma}$. *This is also called Bekker's Theorem.*

Proof: See Kalman, 1990, page 2.6.

This new formulation developed by Kalman will help us solve for $\tilde{\Sigma}$ in general.

The following theorem will help us reduce the space off all solutions and define the region of solutions where a projection matrix exists.

First we define a new scheme called generalized least squares (GLS). A GLS scheme is any scheme where $\tilde{\Sigma} = \tilde{\Sigma}\Sigma^{-1}\tilde{\Sigma}$. Thus we see by lemma 3 that that GLS scheme is equivalent to $\hat{\Sigma} = \hat{\Sigma}\Sigma^{-1}\hat{\Sigma}$.

Theorem 8 *A projection matrix exists iff we use the GLS scheme.*

Proof: Let $\Pi = \{\hat{P} \mid \hat{P}x_t = \hat{x}_t \text{ for } t = 1, ..., T\}$. We want to find that subset of the solution space (i.e. those decompositions) that allows for a projection matrix to exist.

Let $X \in R^{n*T}$ be the data matrix. Then we want to find all $\hat{P}$ in Π such that $\hat{P}X = \hat{X}$. Define $\tilde{P} = I - \hat{P}$. So

$\tilde{X} = \tilde{P}X$.

$\tilde{X}X' = \tilde{P}XX'$

$\tilde{X}(\tilde{X} + \hat{X})' = \tilde{P}\Sigma$

$\tilde{\Sigma} + \tilde{X}\hat{X}' = \tilde{P}\Sigma$

By data orthogonality the second addend on the left is zero. Therefore we see $\tilde{P} = \tilde{\Sigma}\Sigma^{-1}$. Now since $\tilde{P} = \tilde{P}^2$ by the definition of a projection matrix we see that $\tilde{\Sigma}\Sigma^{-1} = \tilde{\Sigma}\Sigma^{-1}\tilde{\Sigma}\Sigma^{-1}$ or $\tilde{\Sigma} = \tilde{\Sigma}\Sigma^{-1}\tilde{\Sigma}$. But this is the definition of the GLS scheme.

Note that by the projection principle the noise should be orthogonal to the data. That is $\hat{X}\tilde{X}^T = 0$. But this is just assumption four.

It is important to have a projection matrix for our schemes. The decomposition of $\Sigma = \hat{\Sigma} + \tilde{\Sigma}$ is only the first step in the identification process. Once we have discerned q and A' and $\tilde{\Sigma}$ we still need to know what a given measurement's exact and noisy parts are. A projection matrix will do this for us. If the scheme is not GLS then the projection of our measurement onto its exact part will be some nonlinear operator. Since we would like to stay with linear operators we see that the GLS scheme is the only way to go.

Let the noise covariance matrix of a GLS solution be denoted by $\tilde{\Sigma}^{GLS}$. The following lemmas describe the importance of GLS solutions.

Lemma 4 *$q = k$ and $q \neq i$ $(i \neq k)$ iff rank $\tilde{\Sigma}^{GLS} = k$.*

Proof: See Kalman, 1990, page 6.4.

Lemma 5 *A $q = k$ solution cannot exist if rank $\tilde{\Sigma} < k$. So $\tilde{\Sigma}^{GLS}$ has the smallest rank for a given q.*

Proof: If the rank of $\tilde{\Sigma} < k$ then the number of eigenvalues of $\tilde{\Sigma}\Sigma^{-1}$ equal to one is less than k. Therefore there can be no $q = k$ solutions.

Lemma 6 *For a given A' the solution scheme with the smallest rank, trace, and norm of $\tilde{\Sigma}$ is the GLS scheme.*

Proof: See Kalman, 1990, page 6.4.

In general the Frisch scheme does not fall under the GLS scheme. Since it does not fall under the GLS scheme it will not have a linear projection matrix. Take for example

$$\Sigma = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$$

and

$$\tilde{\Sigma} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Then it can be shown that the eigenvalues of $\tilde{\Sigma}\Sigma^{-1}$ are $1, .6124$, and 0. This is not a projection matrix, and therefore is not a GLS solution.

Chapter 4

GLS Solutions

Kalman has discovered the construction of all GLS solutions. What follows is a summary of chapter three of his Nine Lectures on Identification.

Kalman's definition of the GLS scheme is this: Given any symmetric $\Sigma > 0$ and any q with $0 < q < n$, a generalized least squares scheme is given by any pair $\hat{\Sigma} \geq 0$ and $\tilde{\Sigma} \geq 0$ such that $\Sigma = \hat{\Sigma} + \tilde{\Sigma}$, rank $\tilde{\Sigma} = q$, and $\tilde{\Sigma}\Sigma^{-1}$ is a projection matrix.

The last condition $\tilde{\Sigma}\Sigma^{-1}$ being a projection matrix is equivalent to $\tilde{\Sigma} = \tilde{\Sigma}\Sigma^{-1}\tilde{\Sigma}$. Kalman has derived a canonical construction for every $\tilde{\Sigma}$ that satisfies this last equality.

Theorem 9 *Given any symmetric positive definite matrix $\Sigma > 0$ and any q with $0 < q < n$, let C be any $n \times q$ matrix with rank $C = q$. Then*

$$\tilde{\Sigma} = C(C'\Sigma^{-1}C)^{-1}C',$$

is a GLS noise covariance matrix.

Proof: See Kalman, 1990, theorem 3.19.

It turns out that for a given C we can determine the matrix of linear relations A. $A' = C'\Sigma^{-1}$. To see why this works just substitute the above canonical formula for $\tilde{\Sigma}$ into the equation $A'\Sigma = A'\tilde{\Sigma}$. Up to the similarity of matrices as discussed before, A and C are in bijective correspondence. Kalman has also proved that up to similarity,

the matrix C appearing in $C(C'\Sigma^{-1}C)^{-1}C'$ is unique. (Kalman, 1990, Proposition 3.23).

Thus we see that for any given A there is one and only one C and one and only one $\tilde{\Sigma}^{GLS}$ that is related to it.

The set of all GLS solutions contains all the elementary least squares solutions, ordinary least squares solutions, and SVD solutions. Thus the set of GLS solutions is very large. It is in fact too large for us to determine q or A' uniquely. To see why note that the SVD solution is a GLS solution. We know that all $q, 0 < q < n$ solutions exist in the SVD scheme. Thus we cannot determine q uniquely.

In fact the GLS schemes really does not tell us much more than the two basic theorems three and four. The GLS scheme is an underconstrained one and thus needs more constraints if we want to get any real information out. Recall that in the Frisch case we were able to find conditions on Σ (actually Σ^{-1}) so that $q = 1$ is the only solution. For GLS there are no conditions we can put on Σ (or its inverse) that will help us determine q. This is because for GLS schemes the determination of q and A' comes from the matrix C. Remember the equation $C(C'\Sigma^{-1}C)^{-1}C'$ gets its rank value from C. There are no conditions on Σ that will help us determine q. This means that if we want to find conditions that will help us determine q we will need to examine $\tilde{\Sigma}$.

The most obvious choice is to put a norm constraint on the noise covariance matrix $\tilde{\Sigma}$. But in fact this gives us no more information than the general theorems three and four. This is because the SVD solution belongs to GLS and in some sense the SVD is the "best fit" in the absence of all other information.

Our derivation of the GLS scheme so far has been circuitous. Least squares implies the minimization of some squared error. In the next chapter we show what it is we are minimizing.

With the GLS scheme there is a one-to-one mapping between the plane we choose in R^n (i.e the matrix A') and its $\tilde{\Sigma}^{GLS}$ noise matrix. We can find a solution for every plane of every dimension ($0 < q < n$). Obviously some planes are better solutions than others. This makes the problem difficult. The data does not seem to tell us how

to choose q or $A^{'}$ that will best fit the data.

As an interesting side note let us describe the duality between the noise and the exact data. We know that $\tilde{\Sigma}$ has the following structure: $\tilde{\Sigma} = C(C'\Sigma C)^{-1}C'$. For any $D^{n\times(n-q)}$ such that $C'\Sigma^{-1}D = 0$ we have $\hat{\Sigma} = \Sigma - \tilde{\Sigma} = \Sigma - C(C'\Sigma C)^{-1}C' = D(D'\Sigma D)^{-1}D'$. This says that $\tilde{\Sigma}^{GLS}(C) = \hat{\Sigma}^{GLS}(D)$. (Kalman, 1990, page 6.9). Because of the implicit data orthogonality condition we are able to get this duality of the signal and the noise.

Chapter 5

Relation of GLS to TLS

5.1 Relation of D-Norm to GLS

So far we have shown that GLS solutions have nice properties. Until now, though, we have not been able to determine which GLS solution is best. We need to determine a suitable criteria for "best fit". In this chapter we will relate every GLS solution for a given Σ to a TLS error minimization problem. To my knowledge this has not been done before. Least squares methods imply a minimization. The minimization for generalized least squares can be thought of as a TLS minimization. The description of TLS given by Golub and Van Loan (Golub and Van Loan, 1980) assumes causality. We will assume a more general model of TLS. That is we will not *a priori* assume any causality in the variables.

We would like to discover a reason for choosing one GLS solution over another GLS solution. One performance criterion would be to minimize the error over a given norm. This norm will contain information the identifier has about the model.

Definition 3 *For any $n \times T$ matrix X and any symmetric, nonnegative definite, $n \times n$ matrix D, the D-Norm is written as $\| X \|_D = \| X'DX \|$.*

Problem Statement: Given a D we would like to find the noise covariance matrix $\tilde{\Sigma} = \frac{1}{T}\tilde{X}\tilde{X}'$ that minimizes $\| \tilde{X} \|_D$.

We assume D is a symmetric nonnegative definite matrix. It turns out that the relation between the D space and the $\tilde{\Sigma}$ space is many-to-one for a given q. Later we will try to reduce the size of the D space. The D matrix can be used to encode information we have about the system.

Note that the minimization of $\parallel \tilde{X}'D\tilde{X} \parallel$ is not weighted least squares (WLS). To see this note that a typical WLS problem involves the minimization of $e'Qe$ where e is the $T \times 1$ vector of noise measurements. That is e measures the noise across all T measurements. Here D weights each of the components whereas WLS weights each of the different measurements.

Theorem 10 *For every D there exists a minimization problem corresponding to one GLS solution. For every GLS solution there exists at least one D corresponding to a minimization problem.*

Proof: Carried out throughout this chapter.

Note that if $D = I$ then we are simply minimizing $\parallel \tilde{X}'\tilde{X} \parallel$ which equals the minimization of $\frac{1}{T} \parallel \tilde{X}\tilde{X}' \parallel = \parallel \tilde{\Sigma} \parallel$. This is just the optimal SVD solution. Note that if for a given q there are eigenvalues with multiplicity greater than one then we will have a problem choosing which of the repeated eigenvalues represent the signal and which of the repeated eigenvalues represent noise. We refer the reader to the Byrnes and Willems' paper.

Lemma 7 *Every $\tilde{\Sigma}$ can be factored into $\tilde{\Sigma} = BB'$, where $B^{n \times q}$ has full rank q, iff $B'\Sigma^{-1}B = I_q$.*

First we show the existence of one B. Let the eigenvalue-eigenvector decomposition of $\tilde{\Sigma} = U\Lambda U'$, where Λ is the diagonal matrix of eigenvalues $0, ..., 0, \sigma_1 \leq ... \leq \sigma_q$. Remember the rank of $\tilde{\Sigma}^{GLS} = q$. Then let $B = U\Gamma$ where

$$\Gamma = \begin{pmatrix} \sqrt{\sigma_1} & 0 & 0 & 0 & 0 \\ 0 & \sqrt{\sigma_2} & 0 & 0 & 0 \\ . & . & . & . & . \\ 0 & 0 & 0 & 0 & \sqrt{\sigma_q} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}^{n*q} .$$

Proof: $(\Rightarrow)$ $\tilde{\Sigma} = \tilde{\Sigma}\Sigma^{-1}\tilde{\Sigma} \Rightarrow BB' = BB'\Sigma^{-1}BB' \Rightarrow (B'\Sigma^{-1}B)^2 = (B'\Sigma^{-1}B)^3$. But the determinant of $(B'\Sigma^{-1}B) \neq 0$ so $B'\Sigma^{-1}B = I_q$.

$(\Leftarrow)$ $B'\Sigma^{-1}B = I_q \Rightarrow BB'\Sigma^{-1}BB' = BB' \Rightarrow \tilde{\Sigma}^{GLS} = BB'$ because any matrix Z that satisfies $Z = Z\Sigma^{-1}Z$ can be shown to be a GLS noise covariance matrix. (This follows from Bekker's Theorem.) See Kalman, 1990, Theorem 3.19.

Lemma 8 *Let U^+ denote the pseudo-inverse of U, where U is an $m \times n$ matrix. If $\| U \| = 1$ and $U^+U = I_n$ then $\| UR \| = \| R \|$. Similarly if $\| U \| = 1$ and $UU^+ = I_m$ then $\| RU \| = \| R \|$.*

Proof: By the Cauchy-Schwarz inequality we see $\| UR \| \leq \| U \| \| R \| = \| R \|$. Likewise $\| UR \| = \| U^+ \| \| UR \| \geq \| U^+UR \| = \| R \|$. Note first that since $U^+U = I_n$, it follows that $\| U^+ \| = 1$. These two statement prove $\| R \| = \| UR \|$. Similarly one can prove the second half of the lemma.

The problem at hand now is to find the $\tilde{X}$ that minimizes $\| \tilde{X}'D\tilde{X} \|$. Note that once we have $\tilde{X}$ we have $\tilde{\Sigma} = \frac{1}{T}\tilde{X}\tilde{X}'$. We also know that $\tilde{\Sigma}$ can be factored into $\tilde{\Sigma} = BB'$ where $B'\Sigma^{-1}B = I_q$.

Lemma 9 *For a given $\tilde{\Sigma} = \frac{1}{T}\tilde{X}\tilde{X}' = BB'$ we have $\| \tilde{X}'D\tilde{X} \| = \| B'DB \|$.*

Proof: Let the eigenvalue-eigenvector decomposition of $\tilde{\Sigma} = BB' = \frac{1}{T}\tilde{X}\tilde{X}' = U\Lambda U'$. Let the singular value decomposition of $\tilde{X} = RST'$. Note that $R = U$ and $\frac{1}{T}SS' = \Lambda$ because $U\Lambda U' = \frac{1}{T}\tilde{X}\tilde{X}' = \frac{1}{T}RST'TS'R' = \frac{1}{T}RSS'R'$ and $\frac{1}{T}SS'$ is diagonal. From before we know that $B = U\Gamma$.

Now $\| \frac{1}{T}\tilde{X}'D\tilde{X} \| = \| \frac{1}{T}TS'R^TDRST' \| = \| \frac{1}{T}S'R'DRS \| = \| \frac{1}{T}S'U'DUS \|$.

Now $\| B'DB \| = \| \Gamma U'DU\Gamma \|$.

Now $S = \sqrt{T}[\Gamma 0]$ where the Γ has dimensions $n \times q$ and the 0 is a matrix of zeroes with dimensions $n \times (T - q)$. The added $T - q$ zero-valued columns do not change the norm. Thus we see that $\| \frac{1}{T}\tilde{X}'D\tilde{X} \| = \| B'DB \|$.

Therefore we can restate the problem as: find a B satisfying $B'\Sigma^{-1}B = I_q$ and minimizing $\| B'DB \|$.

For $q = 1$ this calls for minimizing the scalar $(b'Db)^{\frac{1}{2}}$ given that $b'\Sigma^{-1}b = 1$. This suggests Lagrange multipliers. For $q > 1$ the matrix equations involved in the Lagrange multiplier method become unwieldy.

Let us reformulate the problem again. Let $C = \Sigma^{-1/2}B$, $(B = \Sigma^{1/2}C)$. Then $B'\Sigma^{-1}B = C'C$ and $B'DB = C'\Sigma^{1/2}D\Sigma^{1/2}C$. Let $\Sigma_{new} = \Sigma^{1/2}D\Sigma^{1/2}$. Essentially what we are doing is transforming the variables through a linear transformation.

The new minimization problem is thus to minimize $\| C'\Sigma_{new}C \|$ given $C'C = I_q$.

For $q = 1$ we see the minimization of $c'\Sigma_{new}c$ given $c'c = 1$ occurs when c equals the eigenvector related to the smallest eigenvalue of Σ_{new}.

Example 1: Principal Components. In the SVD solution we assume $D = I$. That is we assume that all the variables are equally noisy. $\Sigma_{new} = \Sigma = U\Lambda U'$. So $c = u_1$ is the smallest eigenvector of U. $Thenb = \Sigma^{1/2}u_1 \Rightarrow bb' = \sigma_1 u_1 u_1' = \tilde{\Sigma}^{GLS}$.

Example 2: Elementary Least Squares:

This time we let

$$D = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ . & . & . & . \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

This means we believe that only the first component is noisy. We know from chapter two that

$$\tilde{\Sigma} = \begin{pmatrix} s_{11}^{-1} & 0 & 0 \\ . & . & . \\ 0 & 0 & 0 \end{pmatrix}.$$

(Here s_{11} is the upper left corner component of the matrix $S = \Sigma^{-1}$.) Thus we see that $b' = [s_{11}^{-1/2}, 0, ..., 0]$. It is easy to see that $b'\Sigma^{-1}b = 1$. Now we need to show that $c = \Sigma^{-1/2}b$ is the smallest eigenvector of Σ_{new}. Since D has rank $n - 1$ and $\Sigma^{1/2}$ has

rank n we see that the smallest eigenvalue of $\Sigma_{new} = \Sigma^{1/2} D \Sigma^{1/2}$ is zero. Thus $\Sigma_{new} c$ should equal zero, as can be verified.

In general for the $q = 1$ case we need to show that the smallest eigenvector of Σ_{new} is indeed an acceptable solution. Let the eigenvalue-eigenvector decomposition of $\Sigma_{new} = U \Lambda U'$. Then $c = u_1 \Rightarrow b = \Sigma^{1/2} u_1$. We need to show that $\tilde{\Sigma} = \tilde{\Sigma} \Sigma^{-1} \tilde{\Sigma}$. But $bb' \Sigma^{-1} bb' = \Sigma^{1/2} u_1 u_1' \Sigma^{1/2} \Sigma^{-1} \Sigma^{1/2} u_1 u_1' \Sigma^{1/2} = \Sigma^{1/2} u_1 u_1' \Sigma^{1/2} = bb'$, proving the assertion.

Now we will tackle the problem for $q > 1$. We want to know what $C^{n \times q}$ will minimize $\parallel C' \Sigma_{new} C \parallel$ given $C'C = I_q$.

Lemma 10 $\parallel C' \Sigma_{new} C \parallel \geq \lambda_q(\Sigma_{new})$ *given* $C'C = I_q$.

Proof: By definition, $\parallel C' \Sigma_{new} C \parallel = max_x x' C' \Sigma_{new} C x$ subject to $x'x = 1$. Let $y = Cx$. Then we want to maximize $y' \Sigma_{new} y$ given $y'y = x'C'Cx = 1$. Since C has rank q, its columns must span a subspace of dimension q. That means y is a linear combination of q vectors in R^n. We want to maximize over y and minimize over C the scalar $y' \Sigma_{new} y$. Minimization over C implies that the columns of C should be the q smallest eigenvectors of Σ_{new}. Maximization over y implies $y = u_q$. This in turn implies the minimum over C of $\parallel C' \Sigma_{new} C \parallel = \lambda_q(\Sigma_{new})$.

Therefore we see that C should contain the smallest q eigenvectors of Σ_{new}. (Once again we run into trouble if we have repeated eigenvalues. Let us assume we do not.) This is a direct generalization of the $q = 1$ case. We need to show that $\tilde{\Sigma} = BB' = \Sigma^{\frac{1}{2}} CC' \Sigma^{\frac{1}{2}}$ satisfies $\tilde{\Sigma} = \tilde{\Sigma} \Sigma^{-1} \tilde{\Sigma}$. Now $\tilde{\Sigma} \Sigma^{-1} \tilde{\Sigma} = \Sigma^{\frac{1}{2}} CC' \Sigma^{\frac{1}{2}} \Sigma^{-1} \Sigma^{\frac{1}{2}} CC' \Sigma^{\frac{1}{2}}$. But $C'C = I_q$, so the preceding implies $\Sigma^{\frac{1}{2}} CC' \Sigma^{\frac{1}{2}} = \tilde{\Sigma}$. Every $\tilde{\Sigma}^{GLS}$ can be written in the form $\tilde{\Sigma} = \Sigma^{1/2} CC' \Sigma^{1/2}$. To see this note that every GLS solution has a corresponding B matrix where $\tilde{\Sigma} = BB'$. The matrices B and C are in one-to-one correspondence because $B = \Sigma^{\frac{1}{2}} C$ and Σ has full rank.

Thus we have shown that for any D we can determine a unique $\tilde{\Sigma}^{GLS}$. Now we need to show that every $\tilde{\Sigma}$ has a corresponding D. (Actually this reverse map is one-to-many for a fixed q.)

Let $\tilde{\Sigma} = BB'$. Then $C = \Sigma^{-1/2} B$. Note $C'C = I_q$ by definition of B. To

construct D let $\Sigma_{new} = \Sigma^{1/2} D \Sigma^{1/2} = U \Lambda U'$, where U contains the q columns of C and the remaining $n - q$ columns are chosen orthogonal to the first q. Choose Λ such that its first q eigenvalues are smaller than the remaining $n - q$ eigenvalues. Thus $D = \Sigma^{-1/2} U \Lambda U' \Sigma^{-1/2}$.

5.2 Relationship of GLS to the Byrnes and Willems' Paper

When minimizing the D-norm we formulated the following problem:

Minimize $\parallel C' \Sigma_{new} C \parallel$ given $C'C = I_q$, where $\Sigma_{new} = \Sigma^{1/2} D \Sigma^{1/2}$. (Remember we have to determine q *a priori*.)

It turns out that the minimization of this problem occurs when C contains the smallest q eigenvectors of Σ_{new}. That is if the eigenvalue-eigenvector decomposition of $\Sigma_{new} = U \Lambda U'$, where $\sigma_1 \leq \sigma_2 \leq ... \leq \sigma_n$, then $C = U_q$. If Σ_{new} has distinct eigenvalues then we can find a unique C and therefore we can find a unique $\tilde{\Sigma} = \Sigma^{1/2} C C' \Sigma^{1/2}$. If the eigenvalues of Σ_{new} are not distinct then it may be the case that we cannot determine C and thus $\tilde{\Sigma}$ uniquely. This is reminiscent of the problem dealt with in Byrnes and Willems' paper.

We can restate the problem as:

Minimize $g_{\Sigma_{new}}(\pi) = \parallel \Sigma_{new} - Q_\pi \Sigma_{new} \parallel^2$ where $g_{\Sigma_{new}}(\pi)$ is the error function over every plane π in R^n and Q_π is the projection matrix onto the plane π.

The solution to this problem is $Q_\pi = U_d U_d'$ where $d = n - q$ and U_d contains the d eigenvectors related to the largest d eigenvalues of Σ_{new}. Note this formulation of the problem results in the same solution as the problem stated above. To see this note that $P_\pi = I - Q_\pi = U_q U_q'$, where this time U_q contains the q smallest eigenvectors of Σ^{new}. This new formulation is the same formulation used by Byrnes and Willems. Thus we can use their results on Σ_{new}.

Essentially all we have done is transform our given data matrix Σ into a new one Σ_{new} that takes into account our weighting matrix D. Thus as shown in Byrnes

and Willems' paper the solution set to min $g_{\Sigma_{new}}$ is a connected submanifold of the grassmanian $G(d,n)$. After our analysis on Σ_{new} we must reformulate the projection matrix Q_π in terms of Σ and not Σ_{new}.

For example the noise projection matrix for the Σ_{new} problem is just CC', whereas in the original Σ problem the noise projection matrix is $\tilde{\Sigma}\Sigma^{-1} = \Sigma^{1/2}CC'\Sigma^{1/2}\Sigma^{-1}$.

The solution plane for the Σ_{new} problem is just all x such that $C'x = 0$. The solution plane for the original Σ problem then becomes all x such that $C'\Sigma^{-\frac{1}{2}}x = 0$.

In summary we can use Byrnes and Willems' results if we transform our original data matrix Σ to $\Sigma_{new} = \Sigma^{1/2}D\Sigma^{1/2}$.

Chapter 6

Problems with Common Performance Criteria

There are many obvious performance criteria. Sadly many of them fail to help us determine a "best fit" solution. They work fine if we have a very structured scheme or we predetermine q. For GLS though we have a less structured solution space and we do not predetermine q. Also the SVD solution is a member of the GLS solution. We know in the absence of all other information the SVD solution has nice characteristics.

6.1 Maximization of Signal-to-Noise Ratio

When trying to maximize the signal-to-noise ratio (SNR) we try to maximize $trace(\hat{\Sigma})/trace(\tilde{\Sigma})$. If we maximize this with respect to q then we see that this maximization always implies $q = 1$. This is because we want to both maximize $trace(\hat{\Sigma})$ and minimize $trace(\tilde{\Sigma})$. The best way to do this is to use an SVD solution scheme where we let $\tilde{\Sigma}$ contain the smallest eigenvalue of Σ. The problem with this performance criterion is that it tries to blindly maximize the SNR. In this process it does nothing to determine which parts of the data are true signal or noise. It just tries to account for the most data. But accounting for the most data possible is exactly the same thing as assuming the $q = 1$ minimum eigenvalue SVD solution. (Kalman, 1989, pages 12-14.)

6.2 Minimize Norm of $\tilde{\Sigma}$

This performance criteria tries to minimize $\| \tilde{\Sigma} \|$. Once again we see that the this performance criterion does not take into account the structure in the data. It just tries to account for the most data possible. This is the same as assuming $q = 1$ and that the noise lies on the smallest eigenvector of Σ.

Note that if we do have other conditions on $\| \tilde{\Sigma} \|$ like a norm restriction on the noise then we can as in chapter four sometimes determine q. In general though we do not have any information on how large the noise can get so we cannot determine q. This is just a restatement of theorem 9. Even as the noise goes to zero we can always identify any q solution. It is only when the noise equals zero that we know q uniquely.

6.3 Maximum Likelihood as a Performance Criterion

Maximum likelihood is often used to estimate unknown parameters such as the mean and variance of a random variable. We will show that it may be possible to use maximum likelihood as a performance criterion if we make some new assumptions. Note, though, that there has been some controversy over whether maximum likelihood makes "sense" for our identification problem. In this chapter we will discuss the problems of using maximum likelihood as a performance criterion over the GLS scheme.

We now include the assumption that the noise is probabilistic in nature. Specifically we assume that $x = \hat{x} + \tilde{x}$, where $\tilde{x}$ is a gaussian random variable with zero mean and covariance $\tilde{\Sigma}$ and $\hat{x}$ is unknown but deterministic. The maximum likelihood procedure determines the parameters that best explains the data. In our problem we are given x_i for $i = 1, ..., T$ and want to determine the $\hat{x}_i$ for $i = 1, ..., T$.

In his lecture notes Kalman states that the maximum likelihood method fails as a performance criterion. According to him one cannot use maximum likelihood meth-

ods to determine $\tilde{\Sigma}$ unless $q = 1$. And even in the case where $q = 1$ the maximum likelihood solution is just the SVD solution. Kalman attempts to maximize the likelihood of $L(\tilde{x}_i; \tilde{\Sigma}_{Pop})$ where $\tilde{\Sigma}_{Pop}$ is the noise population covariance matrix. Kalman opts to determine the noise. He uses a two step procedure. We give a summary of Kalman's method:

Assumptions on the data: (Kalman, 1990, page 7.2.)

Assumption 1: $x_i = \hat{x}_i + \tilde{x}_t$ for $i = 1, ..., T$ are T independent random samples of a Gaussian random vector x.

Assumption 2: For any x_i, $\hat{x}_i$ is the unknown mean of x such that there exists a matrix A where $A'\hat{x}_i = 0$ for all i.

Assumption 3: $cov\tilde{x} = \tilde{\Sigma}_{Pop}$, where Pop means population. These are independent of t and they are nonnegative definite.

Assumption 4: $E(\tilde{x}_i) = 0$ for all i.

The maximum likelihood function L is just the product of the probabilities of each data vector. This follows from assumption one which states that the different measurements are independent. From Anderson (Anderson, 1958, page 45) and Kalman (Kalman,1990) we see that the function L is

$$L(\tilde{x}_i; \tilde{\Sigma}_{Pop}) = (2\pi)^{-nT/2}(det\tilde{\Sigma}_{Pop})^{-T/2} \times exp[(-\frac{1}{2}) \sum_i \parallel \tilde{x}_i \parallel^2_{\tilde{\Sigma}^{-1}_{Pop}}].$$

(Here T is the number of independent measurements.) This is equivalent to $L(\tilde{x}_i; \tilde{\Sigma}_{Pop}) = (2\pi)^{-nT/2}(det\tilde{\Sigma}_{Pop})^{-T/2} \times exp[(-\frac{1}{2})trace\tilde{\Sigma}_{Data}\tilde{\Sigma}^{-1}_{Pop}]$.

To maximize L we need to maximize with respect to both $\tilde{\Sigma}_{Pop}$ and $\tilde{x}_i$ (or $\tilde{\Sigma}_{Data}$.) First we will maximize with respect to $\tilde{\Sigma}_{Pop}$. Then we will maximize with respect to the data. This maximization procedure is used by Kalman. Maximizing with respect to $\tilde{\Sigma}_{Pop}$ and then maximizing with respect to $\tilde{\Sigma}_{Data}$ may lead to problems because the optimal $\tilde{\Sigma}_{Pop}$ depends on $\tilde{\Sigma}_{Data}$. Specifically $\tilde{\Sigma}^{optimum}_{Pop} = \frac{1}{T}\tilde{\Sigma}_{Data}$ (Anderson, 1958, lemma 3.2.2).

I believe that one should incorporate the model straight into the likelihood function. In this way we do not need to run two different maximizations on the likelihood function. A critique of this whole procedure, though, is beyond the scope of this thesis. The maximum likelihood method represents a promising direction of research.

Chapter 7

Conclusions

In this paper we have shown that the GLS solution scheme has many nice properties. Among them we have shown that a linear projection matrix will exist iff we are using the GLS scheme. We have also shown that every GLS problem is a TLS problem. That is every GLS solution is an SVD solution under a suitable transformation, or weighting, of the measured variables. We use the D-norm to carry out the weighting. We have also related GLS to the work of Byrnes and Willems. We can use their work to help us deal with repeated eigenvalues.

We have also shown that maximum likelihood may be used as a performance criterion over different schemes. It is not clear that Kalman's formulation of the maximum likelihood scheme is the most appropriate one. Kalman chooses to maximize with respect to both the signal and the model. I believe that any such maximum likelihood method should some how incorporate the model into the likelihood function itself. It appears that this latter method may be a topic of fruitful research.

Finally we have given a new canonical decomposition of all GLS solutions, $\tilde{\Sigma} = \Sigma^{1/2}CC'\Sigma^{1/2}$, where C is any $n \times q$ matrix such that $C'C = I_q$. Thus we see that the number of GLS solutions for a given q is equal to the number of q-dimensional subspaces of R^n.

There are many open questions that need to be resolved. We still do not have a firm understanding of the structure of GLS solutions. We do not have a nonarbitrary criteria for determining q. This research might entail the need for defining something

like noisy rank.

It is not clear how to determine which GLS solution is the correct solution to choose from. We need outside information to make this decision. The structure of this outside information needs to be determined.

We need to determine criteria so that we can determine the maximum corank solutions in the Frisch case. We have criteria only for the $q = 1$ case. We also need to prove or disprove the statement: we can determine the model uniquely as the noise goes to zero.

Finally it is my belief that the space of regressors is equal to the space of GLS solutions. That is we can construct any GLS solution from a diagonal D matrix. A diagonal D matrix represents a weighting where we have different variances on each component. There are no cross variances among the components in this weighting scheme. The space of regressors is obviously smaller than or equal to the GLS space. It needs to be shown that it is either equal or strictly smaller than GLS.

In conclusion it appears that we cannot in general unambiguously determine q or A without some outside information or outside criteria. This whole problem arose because we wanted to examine what the underlying assumptions behind regressions were. In this thesis we have discussed different identification schemes and different performance criteria.

Anderson, T. W. (1958) An Introduction to Multivariate Statistical Analysis, John Wiley and Sons, New York.

Anderson, T. W. and Rubin, Herman. (1956) "Statistical Inference in Factor analysis," Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. V, University of California Press, Berkeley, pp 111-150.

Byrnes, Christopher and Willems, Jan. "Least Squares Estimation, Linear Programming and Momentum."

Golub, Gene and Van Loan, Charles. (1980) "An Analysis of the Total Least Squares Problem," SIAM J. Numer. Anal. 17, pp 883-893.

Kalman, Rudolf. (1982) "System Identification From Noisy Data," Dynamical Systems II, edited by Bednarek and Cesari, Academic Press, pp 331-342.

Kalman, Rudolf. (1989) "A Theory for the Identification of Linear Relations," Proceedings Colloque LIONS, edited by Brezis and Ciarlet.

Kalman, Rudolf. (1990) Nine Lectures on Identification, Springer Lecture Notes on Economics and Mathematical Systems.

Wilkinson, J. H. (1965) The Algebraic Eigenvalue Problem, Oxford University Press, London.