The Singular Value Decomposition and Long and Short Spaces of Noisy Matrices

Bart De Moor, *Member, IEEE*

***Abstract*—Using geometrical, algebraic, and statistical arguments, it is clarified why and when the singular value decomposition is successful in so-called subspace methods. First we introduce the concepts of long and short spaces. We discuss a fundamental asymmetry in the consistency properties of the estimates: The model, which is associated with the short space, can be estimated consistently but the estimates of the original data, which follow from the long space, are always inconsistent. We find an expression for the asymptotic bias in terms of canonical angles, which can be estimated from the data. This allows us to describe all equivalent reconstructions of the original signals as a matrix ball, the center of which is the minimum variance estimate. Remarkably, the canonical angles also appear in the optimal weighting that is used in weighted subspace fitting approaches. The results are illustrated with a numerical simulation. Examples that are discussed include total linear least squares, the direction-of-arrival estimation algorithm ESPRIT, a biomedical signal processing application to separate the fetal ECG from that of the mother, and the identification of linear state space models from noisy input-output data.**

I. Introduction

An important area in model-based signal processing involves so-called inverse problems, the solution of which is very much related to the field of system identification. In many of these applications, the singular value decomposition (SVD) plays a key role:

Theorem 1: Every real $m \times n$ matrix A can be factorized as $A = U\Sigma V^t$ where $UU^t = I_m = U^tU$, $VV^t = I_n = V^tV$ and Σ is real $m \times n$ with its only nonzero elements on the diagonal. These elements are called the singular values and ordered as $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. The columns of U and V are the left, respectively, right singular vectors of A.

More details, including algorithms, on the SVD can be found in [13]. Basically, there are three frequently cited features that seem to justify the success of the SVD in so-called subspace methods: 1) The information on the underlying model (or data generating mechanism) is completely contained in certain subspaces of the data matrix, which can be determined from the SVD. 2) The complexity of the model is given by the (approximate) rank of the data matrix, which can be estimated from the singular values. 3) Since in most applications, the data are corrupted by additive noise, it is expected that the SVD has a certain noise filtering effect (taking into account the relative insensitivity of the SVD with respect to perturbations).

In most applications, the SVD is brought in after the analysis of an algorithm has been completed, just as a numerical tool that permits to avoid the explicit formation of covariance matrices (of the type A^tA). In this work, however, we will use algebraic, geometrical, and statistical arguments in order to show and illustrate the real power of the SVD. While mathematically, the properties of the SVD are well understood, we present here a new and original geometric framework that facilitates the engineering interpretation of the SVD in the context of model based signal processing.

The main results of this paper are the following: In Sections II-A and B, we give a detailed derivation of the algebraic and geometric conditions that allow to derive the exact model from the SVD of a data matrix. Using a multivariate version of the classical Pythagorean lemma for triangles, it is shown that the original, exact signals *cannot* be recovered when the exact data are modified by additive perturbations. We think that this asymmetry in the consistency of the estimates of the model and the signals is a new insight which is often overlooked. We also introduce the notions of long and short spaces of noisy matrices. The short space can be consistently estimated from the SVD of the data matrix under certain conditions, while this is not the case for the long space. In Section II-C, we characterize the inconsistency of the singular values of the data matrix. In Sections II-D and E, we will derive some extensions that will allow us to take into account possibly available *a priori* information about the noise covariance matrix. In Section III, we derive a minimum variance estimate of the original data and provide a geometrical interpretation. It is shown that when the geometrical assumptions of Section II are satisfied, the minimum variance estimate can be computed from the SVD of the data matrix. In Section IV, we show that the minimum variance signal estimate is the center of a matrix ball of equivalent signal reconstructions. This set of matrices represents the uncertainty in the estimation of the original signals, which originates in the inconsistency of the estimate of the long space. In Section V, we show that statistically, under

Manuscript received February 14, 1991; revised October 12, 1992. This work was supported in part by the Belgian Program on Interuniversity Attraction Poles initiated by the Belgian State Science Policy Programming (Prime Minister's Office) and the European Community Research Program ESPRIT, Basic Research Action 3280.

The author is with the Belgian National Function for Scientific Research, Belgium, and the Department of Electrical Engineering, Katholieke Universiteit Leuven, B-3001 Leuven, Belgium.

IEEE Log Number 9210125.

some generic assumptions, the geometric conditions described in Section II are satisfied asymptotically. The noise variance can be estimated from the singular values. We also obtain an explicit characterization of the uncertainty on the estimated signals in the long space via estimates of the canonical angles. In Section V-C, we give another geometrical-statistical interpretation of the singular values of the minimum variance estimate of the original signals. It turns out that they coincide with the optimal weighting matrix in weighted subspace fitting methods. In Section VI, we give a numerical example illustrating the main points of the previous sections. A brief summary of applications is given in Section VII: total linear least squares, the direction-of-arrival algorithm ESPRIT, biomedical signal processing with the separation of the fetal ECG and the maternal ECG, and finally, a subspace algorithm for the identification of linear state space models from noisy input-output data.

We have chosen to work out in detail the exact geometrical and algebraic theory as opposed to a more statistically oriented approach (which is, however, possible). We assume that the necessary assumptions are exactly satisfied and not just in a statistical sense, which makes our derivation easier and allows us to present a clear geometrical picture. Moreover, a finite sample statistical analysis is very complicated, if not impossible. Therefore we confine ourselves to a demonstration that statistically, the geometrical assumptions hold asymptotically, as the number of measurements goes to infinity.

II. Subspace Methods and the SVD

In this section, we first formulate the main problem that can be solved with subspace methods. Next we use the SVD to derive some sufficient assumptions for the techniques to be successful. Finally, we give some geometrical insight in the results.

A. Linear Subspace Methods and Long and Short Spaces

In many applications, one observes or constructs a matrix M of measurements, where $M \in \mathbb{R}^{p \times q}$. Here, q is the number of measurement channels and p is the number of measurements over these channels. Typically, p is much larger than q. One of the basic assumptions is that the observed data are generated by two unknown matrices: an exact data matrix E and additive perturbations N so that

$$M = E + N. \tag{1}$$

The matrix N contains the noise, which can have several causes (model uncertainty, measurement inaccuracies, etc.). Typically rank $(N) = q$ and also rank $(M) = q$. For the moment, however, we make no statistical assumptions on N but simply treat it algebraically as a matrix. In applications that are essentially based upon linearity, the matrix E, which contains the exact noise-free data, will be rank deficient:

$$\text{rank } (E) = r < q. \tag{2}$$

The number $q - r$ is the number of linearly independent linear relations between the columns of the matrix E. Let $V_e \in \mathbb{R}^{q \times q}$ be an orthogonal matrix (i.e., $V_e^t V_e = I_q = V_e V_e^t$) which is partitioned as

$$V_e = \begin{pmatrix} \overset{r}{V_{e1}} & \overset{q-r}{V_{e2}} \end{pmatrix} \tag{3}$$

so that the row space of E coincides with the column space of V_{e1}[1]: $R(E^t) = R(V_{e1})$ and the columns of V_{e2} generate an orthogonal[2] basis for the null space of E:

$$EV_{e2} = 0. \tag{4}$$

The matrix V_e could be obtained directly from the SVD of E as

$$E = (U_{e1} \quad U_{e2}) \begin{pmatrix} S_{e1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_{e1}^t \\ V_{e2}^t \end{pmatrix} \tag{5}$$

where $U_{e1} \in \mathbb{R}^{p \times r}$, $S_{e1} \in \mathbb{R}^{r \times r}$. Recall that $E \in \mathbb{R}^{p \times q}$ with $p >> q$. In this case, we will call the column space of E, its long space while the row space is its short space. The mnemonic trick is to look at the "size" of the vectors. The vectors of the long space are "longer" (they have more components) than those of the short space. The main reason for introducing these names is the fact that in some applications, it is common to represent the standard model with matrices $M = E + N$ that have much more columns than rows. The long space is in this case the row space while the short space is the column space.

We are now going to discuss the following question: What assumptions are needed so that we can recover from the SVD of the matrix M, the rank r of the matrix E, and the subspaces $R(V_{e1})$ and $R(V_{e2})$? Can we also recover the column space $R(E)$?

The general answer for the rank r and the subspaces $R(V_{e1})$ and $R(V_{e2})$ is *yes*, under certain conditions. For the column space $R(E)$ the answer is *no*. In other words, *Under certain assumptions, the short space of the exact matrix E can be estimated consistently, while the long space cannot be recovered from the SVD of M, not even asymptotically.*

In the remainder of this section, we investigate the problem from the algebraic and geometrical point of view. In Section V, it is shown how the conditions can be satisfied statistically (at least asymptotically as $p \rightarrow \infty$).

B. Algebraic and Geometric Assumptions

For the time being, we assume that the rank r of E and its SVD as in (5) are known. Using the matrices V_{e1} and V_{e2}, we can write M as

$$\begin{aligned} M = E + N &= E + NV_{e1}V_{e1}^t + NV_{e2}V_{e2}^t \\ &= (EV_{e1} + NV_{e1})V_{e1}^t + (NV_{e2})V_{e2}^t. \end{aligned}$$

[1] $R(\cdot)$ denotes the range (column space) of the matrix between brackets.
[2] We use the word "orthogonal" for rectangular matrices V_{e2} that satisfy the property $V_{e2}^t V_{e2} = I$.

Let the matrices between brackets have SVD's: $EV_{e1} + NV_{e1} = P_1 S_1 Q_1^t$ and $NV_{e2} = P_2 S_2 Q_2^t$. Then we can write M as

$$M = P_1 S_1 Q_1^t V_{e1}^t + P_2 S_2 Q_2^t V_{e2}^t$$
$$= (P_1 \quad P_2) \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix} \begin{pmatrix} Q_1^t V_{e1}^t \\ Q_2^t V_{e2}^t \end{pmatrix}. \qquad (6)$$

This is an SVD of M if $P_1^t P_2 = 0$. But $R(P_1) = R((E + N)V_{e1})$ and $R(P_2) = R(NV_{e2})$ so that

$$P_1^t P_2 = 0 \Leftrightarrow V_{e1}^t (E^T + N^t) N V_{e2} = 0. \qquad (7)$$

This condition is necessary for (6) to be an SVD of M. We can strengthen it somehow by splitting it up into two conditions:

$$E^t N = 0 \qquad (8)$$

and

$$V_{e1}^t N^t N V_{e2} = 0. \qquad (9)$$

Why do we select conditions (8) and (9) and not other ones that would also be special cases of (7), for instance, $E^t N = \sigma I_q$ and $N^t N = I_q$? The reason is simple: We will show below that in many engineering applications, there are precisely conditions (8) and (9) that occur. While (7) is perfectly general, conditions (8) and (9) represent the "generic" situation as will be demonstrated below.

Hence, we have shown that we can find from the SVD of M, the subspaces generated by V_{e1} and V_{e2}, which are the row space and null space of E, when the following sufficient conditions are satisfied:

1) The exact data should be "orthogonal" to the noise in the sense that $E^t N = 0$.

2) The matrices V_{e1} and V_{e2} must be orthogonal to each other in the inner product generated by the matrix $N^t N$: $V_{e1}^t N^t N V_{e2} = 0$. This is, for instance, the case when $N^t N$ is a scalar multiple of the identity matrix, which will be treated in Section II-C. If $N^t N$ is not a scalar multiple of the identity, one cannot recover the subspaces V_{e1} and V_{e2} from the SVD of M alone. If, however, $N^t N$ is known up to within a scalar, one can obtain similar results as explained in Section II-D.

3) The smallest singular value of S_1 must be larger than the largest singular value of S_2. Otherwise we cannot separate the subspace generated by V_{e1} from the one generated by V_{e2}. This can be seen from (6). The ratio σ_r / σ_{r+1} could serve as a measure of the (signal + noise)-to-noise ratio.

The success of SVD-based subspace methods critically depends on these three assumptions, which in practice, however, are never satisfied exactly (except maybe for condition 3). A nice feature of the SVD, however, is its robustness with respect to (mild) violations of these conditions. If, for instance, $\|E^t N\|$ is small (where $\|\cdot\|$ is any unitarily invariant norm), the SVD of M will still deliver good approximations to $R(V_{e1})$ and $R(V_{e2})$. The smaller $\|E^t N\|$ gets, the better will be the approximations.

C. *The SVD of M in Terms of the SVD of E*

So far we have found three sufficient conditions that allow us to recover the row space of the exact matrix E from the SVD of the data matrix M. In this section, we give an explicit expression for the SVD of M in terms of the SVD of E in the case that

$$N^t N = \sigma^2 I_q \quad \text{and} \quad N^t E = 0.$$

The first condition implies that N itself is an orthogonal matrix and that every column of N has norm σ. The second condition reflects the orthogonality of the column spaces of N and E. We can now write for the SVD of M:

$$\begin{aligned} M &= E + N \\ &= U_{e1} S_{e1} V_{e1}^t + N V_{e1} V_{e1}^t + N V_{e2} V_{e2}^t \\ &= ((U_{e1} S_{e1} + N V_{e1})(S_{e1}^2 + \sigma^2 I_r)^{-1/2} \quad N V_{e2} \sigma^{-1}) \\ &\quad \cdot \begin{pmatrix} \sqrt{S_{e1}^2 + \sigma^2 I_r} & 0 \\ 0 & \sigma I_{q-r} \end{pmatrix} \begin{pmatrix} V_{e1}^t \\ V_{e2}^t \end{pmatrix} \\ &= (U_{m1} \quad U_{m2}) \begin{pmatrix} S_{m1} & 0 \\ 0 & S_{m2} \end{pmatrix} \begin{pmatrix} V_{m1}^t \\ V_{m2}^t \end{pmatrix} \quad \text{(say)}. \end{aligned} \qquad (10)$$

The second line follows directly from the orthogonality of each column of N with respect to those of E as is explained below. We now make the following observations:

There is a gap in the singular spectrum. The smallest singular value in

$$\sqrt{S_{e1}^2 + \sigma^2 I_r}$$

is larger than the largest one in σI_{q-r}. The $q - r$ smallest singular values are all equal and can be interpreted as a "noise threshold," which permits estimating the noise variance from

$$S_{m2} = \sigma I_{q-r}. \qquad (11)$$

The "exact" singular values (the singular values of E) can then be calculated from

$$S_{e1} = \sqrt{S_{m1}^2 - \sigma^2 I_r}. \qquad (12)$$

Let σ_r be the smallest value of S_{e1}. One could then define a signal-to-noise ratio (SNR) in decibels as

$$\text{SNR} = \left[20 \log_{10} \frac{\sigma_r}{\sigma} \right] \quad \text{dB}. \qquad (13)$$

Another possible definition of the signal-to-noise ratio uses all nonzero singular values σ_i of S_{e1}:

$$\begin{aligned} \text{SNR} &= \left[\log_{10} \frac{\sigma_1^2 + \cdots + \sigma_r^2}{q\sigma^2} \right] \text{dB} \\ &= 20 \log_{10} \frac{\|E\|_F}{\|N\|_F} \quad \text{dB}. \end{aligned} \qquad (14)$$

The row space of E can be recovered. The r right singular vectors corresponding to the largest singular values of M

are precisely the columns of V_{e1}. Hence under the given assumptions, we recover the row space of E (and hence also its null space) exactly.

A multivariate extension of the Pythagorean lemma. There is a Pythagoras-like squaring in the expression for the left singular vectors of M, given by $U_{m1} = (U_{e1}S_{e1} + NV_{e1})(S_{e1}^2 + \sigma^2 I_r)^{-1/2}$. Indeed, $U_{e1}S_{e1}$ consists of column vectors with norms given by the singular values in S_{e1}. NV_{e1} consists of column vectors with norms given by σ. Furthermore, the columns of U_{e1} are orthogonal to the columns of NV_{e1} because $E'N = 0$. Hence, the columns of $U_{e1}S_{e1} + NV_{e1}$ have a norm given by the diagonal elements of

$$\sqrt{S_{e1}^2 + \sigma^2 I_r}$$

which is a multivariate generalization of the Pythagorean lemma.

The canonical angles between exact and noisy subspaces. It is impossible to recover the original noisefree column space of E, which is represented by the matrix U_{e1}, from the SVD of M in (10). But we will now show that the canonical angles between the column space of E and the column space $R(U_{m1})$ generated by the r first left singular vectors of M can be computed. Recall that the cosines of the canonical angles between the column spaces $R(U_{e1})$ and $R(U_{m1})$ of two orthogonal matrices U_{e1} and U_{m1} are the singular values of the product $U_{e1}'U_{m1}$ (see, e.g., [13]). Here we find

$$\begin{aligned} U_{e1}'U_{m1} &= U_{e1}'(U_{e1}S_{e1} + NV_{e1}')(S_{e1}^2 + \sigma^2 I_r)^{-1/2} \\ &= S_{e1}(S_{e1}^2 + \sigma^2 I_r)^{-1/2} \qquad (15) \\ &= (I_r + \sigma^2 S_{e1}^{-2})^{-1/2} = S_{e1}/S_{m1} = C \quad \text{(say)}. \qquad (16) \end{aligned}$$

The cosines of the canonical angles are the diagonal elements of the diagonal matrix C in (16). From (16), we see that the canonical angles between the column space of E and the left principal singular subspace of M depend on the signal-to-noise ratio. The larger the signal-to-noise ratio, the smaller will be the canonical angles. In the noisefree case, for $\sigma = 0$, all angles are 0.

The fact that the column space of E cannot be recovered from the SVD of M can most easily be explained for the simple case where $r = 1$ and $p = 2$. Suppose a vector $M \in \mathbb{R}^2$ is given as in Fig. 1 where it is the sum of two orthogonal but unknown vectors E and N. The question of interest is the following: In how many ways can we decompose M in two orthogonal vectors E and N such that $M = E + N$ and $E'N = 0$. The classical construction is depicted in Fig. 2(a). We have to draw a circle that has M as its diameter. Every point of the circle has the property that it generates an orthogonal decomposition, the sum of which is precisely M. Two of this infinite number of decompositions are depicted in Fig. 2(a). Even if the norm of N would be known, say $\|N\| = \sigma_N$, still, for $p = 2$ there are two solutions as shown in Fig. 2(b). If $p = 3$, there is a circle of solutions which can be easily visualized

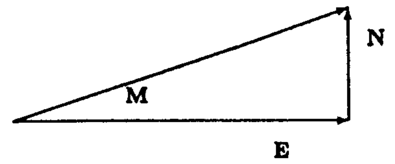


Fig. 1. The vector M is the orthogonal sum of the vectors E and N.

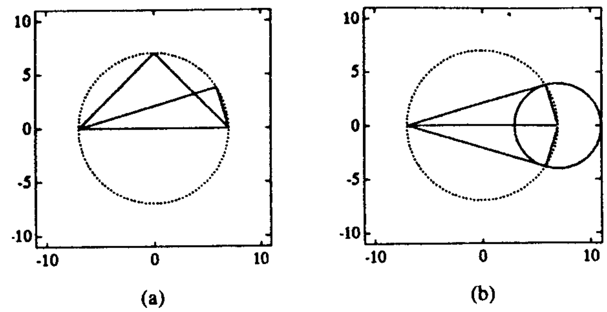


Fig. 2. (a) The horizontal diameter is the vector M. Every point on the circle generates an orthogonal decomposition of M. Two such decompositions are drawn. (b) The small circle has a radius σ_N which is the norm of N. Even when σ_N is known, there are still two orthogonal decompositions of M in two dimensions. In three dimensions this becomes a circle of solutions.

from Fig. 2(b) by rotating the circles in the third dimension, so that two spheres are generated. Their intersection is then a circle with radius σ_N. In p dimensions, there is a $(p - 1)$-dimensional hypersphere of solutions. The conclusion is that the orthogonal decomposition of M into two vectors E and N is certainly not unique, for $r = 1$, let alone for $r > 1$, even if $\|N\| = \sigma_N$ is given. From Fig. 2, we can easily compute the angles between the observed vector M and all vectors E if $\|N\| = \sigma_N$ is given. Indeed, for all vectors E, we have

$$\cos \alpha = \frac{\|E\|}{\|M\|} = \frac{\|E\|}{\sqrt{\|E\|^2 + \|N\|^2}} = \frac{1}{\sqrt{1 + (\|N\|^2/\|E\|^2)}}. \qquad (17)$$

(Note that $\tan \alpha = \|N\|/\|E\|$.) Observe that $\cos \alpha$ is a function of the signal-to-noise ratio (14). The higher the SNR, the larger will be $\cos \alpha$, hence the smaller will be α. In the noiseless case when $\sigma_N = 0$, $\alpha = 0$ as well.

Expression (16) is the multivariate generalization of (17) and we will come back to it in Section IV.

D. *What to Do if $N'N$ Is Not a Multiple of the Identity Matrix*

There is a price to be paid if condition (9) is not satisfied: In that case, we must know the matrix $N'N$ up to within an unknown real scalar multiple λ, i.e., $\Sigma_N = N'N\lambda$. Assume that Σ_N, which is a positive definite matrix, is factored as $\Sigma_N = R'R$ where $R \in \mathbb{R}^{q \times q}$ is a nonsingular (false) square root. In this case, we can still recover the appropriate subspaces by considering the SVD of the matrix $M_* = MR^{-1}$. This can be seen as follows. First define E_* and N_* via $M_* = E_* + N_* = ER^{-1} + NR^{-1}$. Observe that

$$N_*^t N_* = R^{-t}N^tNR^{-1} = 1/\lambda I_q$$

which is a scalar multiple of the identity matrix. Define an orthogonal matrix $Y \in \mathbb{R}^{q\times q}$ which is partitioned as

$$Y = \begin{pmatrix} \overset{r}{Y_1} & \overset{q-r}{Y_2} \end{pmatrix}$$

so that $R(Y_1) = R(R^{-t}V_{e1})$ and $R(Y_2) = R(RV_{e2})$. Then $R(Y_1) = R(E_*^t)$ and $E_*Y_2 = 0$. We now apply the same analysis as the one in Section II-B to our transformed model $M_* = E_* + N_*$:

$$\begin{aligned} M_* &= E_* + N_* = [(E_* + N_*)Y_1]Y_1^t + N_*Y_2Y_2^t \\ &= P_1S_1Q_1^tY_1^t + P_2S_2Q_2^tY_2^t \\ &= (P_1 \quad P_2)\begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix}\begin{pmatrix} Q_1^tY_1^t \\ Q_2^tY_2^t \end{pmatrix} \end{aligned}$$

which is an SVD of M_* if $E^tN = 0$. This can be seen by observing that $R(P_1) = R[(E_* + N_*)Y_1]$ and $R(P_2) = R(N_*Y_2)$ so that

$$\begin{aligned} P_1^tP_2 = 0 &\Leftrightarrow Y_1^t(E_*^t + N_*^t)N_*Y_2 = 0 \\ &\Leftrightarrow V_{e1}^tR^{-1}(R^{-t}E^t + R^{-t}N^t)NR^{-1}RV_{e2} = 0 \\ &\Leftrightarrow V_{e1}^tR^{-1}R^{-t}N^tNV_{e2} = 0 \\ &\Leftrightarrow V_{e1}^tV_{e2} = 0 \end{aligned}$$

which follows from the orthogonality of (3).

The fact that N^tN should be known up to within a scalar is not a trivial requirement. In some cases, this matrix can be estimated during the absence of the signal contained in E (i.e., when $E = 0$). This is, for instance, the case in ESPRIT [19]. In other cases, a square root R can be directly computed from the structure of the equations (see, e.g., [9]). Another remark concerns the numerical computation of the SVD of MR^{-1}. An explicit formation of the inverse R^{-1} followed by the explicit calculation of the product MR^{-1} may result in dramatic loss of numerical accuracy in the data. This can be avoided by using the quotient singular value of decomposition (QSVD) (called the generalized SVD in [13], see [7]), which delivers the required factorizations without forming quotients and products (see, e.g., [9] for a detailed explanation). In [5], we describe a framework based upon the SVD and the QSVD to analyze the geometrical structure of signals with respect to disturbances (oriented energy of vector signals and oriented signal-to-signal ratios).

III. Minimum Variance Estimation

What is the best estimate of E that could be obtained by making linear combinations of the noisy data in the matrix M? This problem can be formulated in the following minimum variance estimation framework.

Given the $p \times q$ matrix M as in (1), with E satisfying (2), find the matrix X that minimizes:

$$\min_{X \in \mathbb{R}^{q\times q}} \|MX - E\|_F^2.$$

If E would be known, then setting to zero the derivatives of the object function

$$\|MX - E\|_F^2 = \mathrm{tr}[X^tM^tMX + E^tE - 2X^tM^tE]$$

with respect to the elements of X, results in $X = (M^tM)^{-1}M^tE$. Hence, the minimum variance estimate of E is given by $MX = M(M^tM)^{-1}M^tE$. The geometrical interpretation is that the minimum variance estimate of E given M is the orthogonal projection of E onto the column space of M. Observe that $\mathrm{rank}(MX) = \mathrm{rank}(E)$. Since E is unknown, this solution cannot be computed from the noisy data matrix M alone unless additional assumptions are made. Under the assumptions of Section II-C ($N^tN = \sigma^2I_q$, $N^tE = 0$), we find that an expression for the minimum variance estimate is given by

$$\begin{aligned} MX &= M(M^tM)^{-1}M^tE = (U_{m1} \quad U_{m2})\begin{pmatrix} U_{m1}^t \\ U_{m2}^t \end{pmatrix} \\ &\quad \cdot (U_{e1} \quad U_{e2})\begin{pmatrix} S_{e1} & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} V_{e1}^t \\ V_{e2}^t \end{pmatrix} \\ &= (U_{m1} \quad U_{m2}) \\ &\quad \cdot \begin{pmatrix} (S_{e1}^2 + \sigma^2I_r)^{-1/2}(S_{e1}^tU_{e1}^t + V_{e1}^tN^t)(U_{e1} \quad U_{e2}) \\ \sigma^{-1}V_{e2}^tN^t(U_{e1} \quad U_{e2}) \end{pmatrix} \\ &\quad \cdot \begin{pmatrix} S_{e1} & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} V_{e1}^t \\ V_{e2}^t \end{pmatrix} \\ &= (U_{m1} \quad U_{m2})\begin{pmatrix} (S_{e1}^2 + \sigma^2I_r)^{-1/2}S_{e1} & * \\ 0 & * \end{pmatrix} \\ &\quad \cdot \begin{pmatrix} S_{e1} & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} V_{e1}^t \\ V_{e2}^t \end{pmatrix} \\ &= [U_{m1}][S_{e1}^2(S_{e1}^2 + \sigma^2I_r)^{-1/2}][V_{m1}^t] \end{aligned} \tag{18}$$

which is a singular value decomposition. Despite the fact that we do not know the original matrix E, it is possible to find the minimum variance estimate from the SVD of M if all the geometrical assumptions from Sections II-B and C are satisfied (ie., $N^tN = \sigma^2I_q$ and $N^tE = 0$). From expression (18), it can be seen that we do not obtain a consistent estimate of the long space since $U_{m1} \neq U_{e1}$. The singular values of the minimum variance estimate (18) are given by

$$S_{e1}^2(S_{e1}^2 + \sigma^2I_r)^{-1/2} = (S_{e1}S_{m1}^{-1})S_{e1} = CS_{e1} \tag{19}$$

where C are the cosines of the canonical angles as defined in (16). The relation of this expression to the classical geometry of minimum variance estimation can best be il-

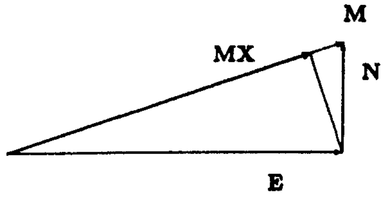


Fig. 3. The large arrow is the minimum variance estimate MX of E given M.

lustrated via the same simple vector example of Section II-C:

Example: Let $M \in \mathbb{R}^{p \times 1}$ be the orthogonal sum of two vectors E and N as in Fig. 1. The minimum variance estimate of E, given M, is obtained from the solution of the optimization problem $\min_{\xi \in \mathbb{R}} \|M\xi - E\|^2$. The solution ξ is easily found to be $\xi = M^tE/M^tM$ which can be rewritten as

$$\xi = \frac{E^tE}{M^tM} = \frac{E^tE}{E^tE + N^tN} = \frac{\|E\|^2}{\|E\|^2 + \|N\|^2}.$$

Hence, the minimum variance estimate of E given M is $M\xi = M/\|M\| \; \|E\| \cos \alpha$ where (see Fig. 3):

$$\cos \alpha = \sqrt{\frac{E^tE}{M^tM}} = \frac{\|E\|}{\|M\|}.$$

Expression (18) is the multivariate generalization of this observation.

The minimum variance estimate (18) is of course not the only possible estimate of the original signals. Another estimate is obtained when we approximate M by a matrix of rank r in least squares sense as

$$\min_{E_{LS} \in \mathbb{R}^{p \times q},\ \mathrm{rank}(E_{LS}) = r < q} \|M - E_{LS}\|_F^2.$$

The solution is a classical result (originally due to Eckart and Young, see, e.g., [13]) and follows from the SVD (10) of M as

$$E_{LS} = U_{m1} S_{m1} V_{m1}^t = [(U_{e1}S_{e1} + NV_{e1})(S_{e1}^2 + \sigma^2 I_r)^{-1/2}] \cdot [(S_{e1}^2 + \sigma^2 I_r)^{1/2}] V_{e1}^t.$$

Observe that the left and right singular vectors of this least squares estimate are the same as those of the minimum variance estimate (18) but the singular values are different.

IV. A Matrix Ball of Signal Reconstructions

In this section, we derive the multidimensional generalization of the ambiguity illustrated in Fig. 2(b). Even in the two-dimensional situation illustrated there, the original vector E could not be reconstructed uniquely from the given noisy data vector M, even if the noise variance σ_N were known. The situation is similar in the multidimensional case ($p > 2$). Despite the fact that we can compute the minimum variance estimate as derived in the previous section, it is not possible to estimate consistently the "long space" of the exact matrix E from the SVD of M. However, from (19) we can compute the canonical angles between the column spaces of M and E, without knowing E if only the assumptions $N^tN = \sigma^2 I_q$ and $N^tE = 0$ are satisfied. We will now show how we can obtain a whole set of "equivalent" reconstructions of the original signals. In other words, since we cannot reconstruct uniquely the exact data matrix E from the noisy version M, we will describe a whole set of data matrices E that could have generated the observed data matrix M under the assumptions that $N^tN = \sigma^2 I_q$ and $N^tE = 0$.

First, we define a matrix ball.

Definition 1: A matrix ball $\mathcal{B}$ with center matrix $B_c \in \mathbb{R}^{p \times q}$, left radius matrix $B_l \in \mathbb{R}^{p \times l}$ and right radius matrix $B_r \in \mathbb{R}^{m \times q}$ is a set of matrices of the form

$$\mathcal{B} = \{B \mid B = B_c + B_l X B_r,\ X \in \mathbb{R}^{l \times m} \text{ is orthogonal}\}.$$

Recall the SVD of M as in (10). Consider a set $\mathcal{B}$ of matrices with the following properties:

1) The rank is r and the singular values are S_{e1}.
2) The right singular vectors are the columns of V_{e1}.
3) The cosines of the canonical angles between the left principal singular subspace and $R(U_{m1})$ are given by the diagonal elements of C as in (16).
4) The original, exact matrix E is an element of $\mathcal{B}$.

The main purpose of this section is to show that the set of matrices $\mathcal{B}$ as determined by these four conditions is a matrix ball. Let us first concentrate on the third constraint. All orthogonal matrices $Y_1 \in \mathbb{R}^{p \times r}$ that satisfy this constraint are given by

$$Y_1 = U_{m1} Z_1 C W_1^t + U_{m2} Z_2 (\sqrt{I_r - C^tC}) W_1^t \quad (20)$$

where Z_1, Z_2, and W_1 are arbitrary orthogonal matrices of appropriate dimensions. This can be seen by computing $U_{m1}^t Y_1 = Z_1 C W_1^t$ which has singular values C. Hence, the cosines of the canonical angles between $R(Y_1)$ and $R(U_{m1})$ are precisely the diagonal elements of C. That Y_1 is orthogonal can be verified from

$$Y_1^t Y_1 = W_1 C^t C W_1^t + W_1 (I_r - C^tC) W_1 = I_r.$$

The fourth constraint implies that the original matrix U_{e1} belongs to the set of orthogonal matrices described by (20) for a specific choice of Z_1, Z_2 and W_1:

$$U_{e1} = Y_1 = U_{m1} Z_1 C W_1^t + U_{m2} Z_2 \sqrt{I_r - C^tC}\, W_1^t$$
$$\Rightarrow Z_1 C W_1^t = U_{m1}^t U_{e1} = C \Rightarrow Z_1 = W_1 = I_r$$
$$\Rightarrow Z_2 \sqrt{I_r - C^tC} = U_{m2}^t U_{e1}$$
$$\Rightarrow Z_2 = U_{m2}^t U_{e1} (\sqrt{I_r - C^tC})^{-1}.$$

(In order to avoid technical complications, we assume that all diagonal elements of C are distinct, which is the generic situation anyway. Also note that $\sqrt{I_r - C^tC}$ is a diagonal matrix with the sines of the canonical angles between the column space E and the column space generated by U_{m1}.) While Z_1 and W_1 can apparently be determined, we cannot determine Z_2 because U_{e1} itself is unknown. Hence, the best we can do for the long space of the original matrix E is to require that U_{e1} belongs to the set of orthogonal matrices

$$\mathcal{Y}_1 = \{Y_1 \mid Y_1 = U_{m1} C + U_{m2} Z \sqrt{I_r - C^t C}, \quad Z \in \mathbb{R}^{(p-r)\times r} \text{ is orthogonal}\}. \quad (21)$$

If we now add the first and second constraint, we find that the set of matrices $\mathcal{B}$ is a matrix ball given by

$$\mathcal{B} = \{B \mid B = B_c + B_l Z B_r, \quad Z \in \mathbb{R}^{(p-r)\times r} \text{ is orthogonal}\}. \quad (22)$$

where the center matrix is the minimum variance estimate (18) $B_c = U_{m1} C S_{e1} V_{e1}^t$, the left radius matrix is the orthogonal complement of U_{m1}: $B_l = U_{m2}$ and the right radius matrix is

$$B_r = \sqrt{I - C^t C} S_{e1} V_{e1}^t.$$

V. Statistical Arguments

The analysis so far has concentrated on the algebraic and geometrical assumptions on E and N that are sufficient to recover model information about E from the SVD of M. In Section V-A, we show that asymptotically, as the number of measurements $p \to \infty$, these conditions are satisfied if the noise is zero mean and has bounded fourth-order moments. In Section V-B, we investigate how the exact singular values of E and the angles that characterize the bias in the long space, can be estimated from the measured data. In Section V-C, we concentrate on the case where the noise in N is Gaussian.

A. Asymptotic Behavior as $p \to \infty$

In principle, in all that follows, we could indicate the dependence of the matrices M, E and N on p, but for clarity we will not do so. Note that $\forall p$:

$$M^t M = E^t E + N^t N + E^t N + N^t E. \quad (23)$$

Let us first look at the cross terms $E^t N$ and $N^t E$ in (23) under the following assumption.

Assumption 1: The elements of N have zero mean.

Then, obviously $E(E^t N) = 0$ because E is now considered as a fixed deterministic matrix. For the sake of simplicity, we assume that there is no structure in the matrices (such as (block-)Hankel, (block-)Toeplitz, etc.). In the case where there is structure, the analysis below can be adapted without much difficulty. In any case, we assume that the noise covariance matrix $E(N^t N)$ can be computed or estimated up to within a scalar multiple. In particular, we will assume that

Assumption 2: The elements of N are independently and identically distributed with possibly unknown variance ν^2.

If, for instance, the elements of N are independently and identically distributed with zero mean and (possibly unknown) elementwise variance ν^2, we have $E(N^t N) = p\nu^2 I_q$. Hence, in this case[3]

$$E(M^t M) = E^t E + E(N^t N) = E^t E + p\nu^2 I_q. \quad (24)$$

This suggests that when we would average over several experiments (recall that $E(\cdot)$ is the ensemble average) with identical E matrix but with different realizations for N, the three "geometrical" conditions of Section II-B are satisfied. Indeed, we have

1) $E(E^t N) = 0$.
2) $V_{e1}^t E(N^t N) V_{e2} = 0$.
3) $\lambda_i E(M^t M) = \lambda_i(E^t E) + p\nu^2$.

In practical experimental situations, however, one cannot repeat the same experiment over and over again with the same identical matrix E. Hence, in practice, it is impossible to rely on (24) to arrive at the geometrical conditions of Section II-B.

In one experiment, however, it is very well possible to take a large number of measurements, especially in signal processing applications. We will now show that, under some mild conditions, the geometrical and algebraic requirements of Section II-B are achieved asymptotically, as $p \to \infty$. We start again from expression (23). The variance of the elements of $E^t N$ as a function of p is given by $E(E^t N)_{ij}^2 = \nu^2 \Sigma_{k=1}^p e_{ki}^2$. This variance increases with p according to the energy (the sum of squares) of the elements in the columns of E. In order to keep it finite as $p \to \infty$, we will divide expression (23) by some function of p, so that for $p \to \infty$, the influence of the cross terms vanishes. In this way, we will achieve the first geometrical condition (8) of Section II-B. The precise function of p that will do the job depends on the properties of the exact signals in E. Quite often, these signals satisfy the following.

Assumption 3: The exact signals are quasi-stationary, i.e.,

$$\lim_{p\to\infty} \frac{\sum_{k=1}^{p} e_{ki}^2}{p} = \text{finite}$$

$$\lim_{p\to\infty} \frac{\sum_{k=1}^{p} e_{ki}^2}{p^2} = 0$$

In this case, the variance of the inner products between the columns of E and N grows at most linearly with p. Let's now divide (23) by p:

$$M^t M/p = E^t E/p + N^t N/p + (E^t N + N^t E)/p. \quad (25)$$

[3]Note that the eigenvalues of $E(M^t M)$ will be larger than the eigenvalues of $E^t E$, even if $E(N^t N)$ is not a multiple of the identity matrix. Indeed, if a Hermitian matrix is perturbed by a positive definite matrix, its eigenvalues must increase [20, p. 203].

We find for the variances in the cross terms:

$$E(E^tN/p)_{ij}^2 = \nu^2 \frac{\sum_{k=1}^{p} e_{ki}^2}{p^2} \to 0 \qquad \text{as } p \to \infty. \tag{26}$$

Hence, as p grows, the variances go to 0 while the mean value is also 0. This implies that, even if we do only one experiment (i.e., simply omit the expected value operator in (26)), the cross terms in (25) will go to zero as $p \to \infty$.

Let us now investigate the term N^tN/p in (25). We find for the off-diagonal elements of N^tN in (23) that $E(N^tN)_{ij,i\neq j} = 0$ and $E(N^tN)_{ij,i\neq j}^2 = p\nu^4$. This last equation follows from

$$E(N^tN)_{ij,i\neq j}^2 = E\left[\left(\sum_{k=1}^{p} N_{ki}N_{kj}\right)^2\right]$$
$$= E\left[\sum_{k=1}^{p} N_{ki}^2 N_{kj}^2\right] + E\left[2\sum_{k=1}^{p}\sum_{l=1,k<l}^{p} N_{ki}N_{kj}N_{li}N_{lj}\right].$$

The second term is 0 because of the independence of different elements of N. We see that the variance of the off-diagonal elements grows linearly with p. When scaled by p as in (25) we find $E(N^tN/p)_{ij,i\neq j}^2 = \nu^4/p \to 0$ as $p \to \infty$. Hence, the variance of the scaled off-diagonal elements goes to zero for increasing p so that N^tN becomes more and more a diagonal matrix. For the diagonal elements, we find $E(N^tN)_{ii} = p\nu^2$ and for their variance

$$E((N^tN)_{ii} - p\nu^2)^2 = E[(N^tN)_{ii}^2 + p^2\nu^4 - 2p\nu^2(N^tN)_{ii}]$$
$$= pE(N_{ii}^4) + 2E\left[\sum_{k=1}^{p}\sum_{l=1,k<l}^{p} N_{ik}^2 N_{il}^2\right] + p^2\nu^4 - 2p\nu^2 E[(N^tN)_{ii}]$$
$$= pE[N_{ii}^4] + 2[(p-1)(p-2)\cdots 2.1]\nu^4 + p^2\nu^4 - 2p^2\nu^4$$
$$= pE(N_{ii}^4) - p\nu^4. \tag{27}$$

For the scaled matrix N^tN/p, we then find for the mean and variance of the diagonal elements $E(N^tN/p)_{ii} = \nu^2$ and $E((N^tN/p)_{ii} - \nu^2)^2 = E(N_{ii}^4)/p - \nu^4$. Obviously, we need the additional assumption that

Assumption 4: The fourth-order moments of the noise are bounded.

Hence, even for one experiment (i.e., when we simply omit the expected value operator), the matrix N^tN/p approaches the identity matrix $\nu^2 I_q$ as $p \to \infty$, because the variance of the diagonal elements goes to zero, so that asymptotically, the second and third geometrical condition of Section II-B are satisfied. Summarizing, we find that

1) $E^tN/p \to 0$ as $p \to \infty$.
2) $N^tN/p \to \nu^2 I_q$ which implies that $V_{e1}^t(N^tN/p)V_{e2} \to 0$ as $p \to \infty$.
3) $\lambda_i(M^tM/p) \to \lambda_i(E^tE/p) + \nu^2$ as $p \to \infty$.

B. Estimates of the Noise Variance and the Canonical Angles

The preceding analysis shows that, as $p \to \infty$, we gradually approach the geometric situation of Section II-C, where $E^tN = 0$ and N^tN was a multiple of the identity matrix. Since singular values are perfectly well conditioned (see, e.g., [20]), it follows from the perturbation theory for singular values, that with this additive noise, the noise threshold will become more and more pronounced for increasing p. This allows us to estimate the noise variance as follows: Let $\mu_1, \mu_2, \cdots, \mu_r, \mu_{r+1}, \cdots, \mu_q$ be the singular values of M. Then, an estimate ν_{est}^2 of the variance ν^2 of the elements of N can be obtained (from (11) with $\sigma^2 = p\nu^2$) as

$$\nu_{est}^2 = \frac{\mu_{r+1}^2 + \cdots + \mu_q^2}{p(q-r)}. \tag{28}$$

An estimate of the exact singular values σ_i, $i = 1, \cdots, r$ of E can be obtained (from (12) with $\sigma^2 = p\nu^2$) as

$$\sigma_i = \sqrt{\mu_i^2 - p\nu_{est}^2} \qquad i = 1, \cdots, r. \tag{29}$$

Together, expressions (28) and (29) allow us to obtain an estimate for the signal-to-noise ratio SNR (13) and (14). Using (28) and (29) the canonical angles that characterize the bias of the long space can be estimated from (16).

As a matter of fact, the strong consistency of estimates of the singular values and of quantities that are associated with the short space only depends on the convergence of the sample covariance matrix (which is M^tM/p) with probability 1 to its expected value. That this is the case when the fourth-order moments are bounded and the exact signal is quasi-stationary, has been observed before by several authors [2], [17]. However, here we have also described the asymptotic behavior of the long space.

C. Gaussian Noise and Signals

It is often assumed that the elements of the matrix N have zero mean and are identically and independently normally distributed. This *a priori* assumption of normality is not only mathematically convenient, but via the central limit theorem (see, e.g., [18]), it is often a good engineering approximation of the real circumstances. When ξ is normally distributed with zero mean and variance ν^2, then $E(\xi^4) = 3\nu^4$ so that (27) becomes $E[(N^tN)_{ii} = p\nu^2]^2 = 2p\nu^4$. In the case of Gaussian noise, the fact that $N^tN/p \to \sigma^2 I_q$ as $p \to \infty$ can also be interpreted as follows: If we consider the rows of N as vectors in a q-dimensional space, then all directions in $\mathbb{R}^q$ are sampled with equal probability, i.e., the row vectors of N lie equally dense in all directions of $\mathbb{R}^q$. In other words, the probability density function of the row vectors normalized to have norm 1, is uniform on the unit sphere in $\mathbb{R}^q$. The reverse statement is less trivial (see, e.g., [15]): If the elements of $x \in \mathbb{R}^q$ are independent zero-mean random variables such that $x/\|x\|$ is uniformly distributed on the unit sphere on $\mathbb{R}^q$, then the elements of x are normally distributed if $q \geq 3$.

If $N'N/p$ is not a multiple of the identity matrix, the transformation R^{-1} discussed in Section II-E, where $E(N'N) = RR'$, reduces the noise situation to the isotropic case.

In order to make the statistical analysis more tractable, it is often assumed that also the rows of E are independently and identically normally distributed, i.e., also for the exact signal a statistical model is assumed. In this case, the asymptotic distribution (i.e., as $p \to \infty$) of the eigenvectors of the sample covariance matrix $M'M/p$ is Gaussian. Let v_{mk} be the eigenvector of $M'M/p$ associated with its kth eigenvalue (i.e., v_{mk} is the kth right singular vector of $M/\sqrt{p}$). Let v_{ek} be the kth right singular vector of $E/\sqrt{p}$ and let σ_i, $i = 1, 2, \cdots, r$ be its singular values. Define

$$\lambda_i = \sigma_i^2 + \nu^2 \qquad i = 1, \cdots, r$$

$$\lambda_i = \nu^2 \qquad i = r + 1, \cdots, q.$$

Then (see, e.g., [1]):

$$E[v_{mk}] = v_{ek} + 0(p^{-1}) \tag{30}$$

$$E[(v_{mk} - E[v_{mk}])(v_{mk} - E[v_{mk}])'] = \frac{\lambda_k}{p} \sum_{i=1, i\neq k}^{q} \frac{\lambda_i}{(\lambda_i - \lambda_k)^2} v_{ei} v_{ei}' + o(p^{-1}) \tag{31}$$

$$E[(v_{mk} - E[v_{mk}])(v_{ml} - E[v_{ml}])'] = -\frac{\lambda_k \lambda_l}{p(\lambda_k - \lambda_l)^2} v_{el} v_{ek}' + 0(p^{-1}) \qquad k \neq l \tag{32}$$

In the case of Gaussian E and N, formula (28) gives the maximum likelihood estimate of the noise variance [1, p. 130].

The following remarkable observation was made in [17, p. 134]: Consider the projection of the dominant eigenvectors V_{m1} of M onto the exact null space of E: $u_k = V_{e2} V_{e2}' v_{mk}$, $k = 1, 2, \cdots, r$. These projections are also Gaussian distributed with mean and covariances that can be easily computed from (30)–(32). For $k = 1, \cdots, r$, we have

$$E[u_k] = 0 + O(p^{-2})$$

$$E[u_k u_l'] = 0 + o(p^{-1}) \qquad k \neq l$$

$$E[u_k u_k'] = V_{e2} \left[\frac{\lambda_k}{p} \sum_{i=r+1}^{p} \frac{\lambda_i}{(\lambda_i - \lambda_k)^2} \right] V_{e2}' + o(p^{-1})$$

$$= V_{e2} \frac{\nu^2}{p} \frac{\lambda_k}{(\sigma_k^2)^2} I_{q-r} V_{e2}' + o(p^{-1}).$$

Observe that the scalars λ_k/σ_k^4 in the last expression are the squares of the diagonal elements of $S_{m1}/S_{e1}^2 = (CS_{e1})^{-1}$. We have already encountered the matrix CS_{e1} as the singular values of the minimum variance estimate (18).

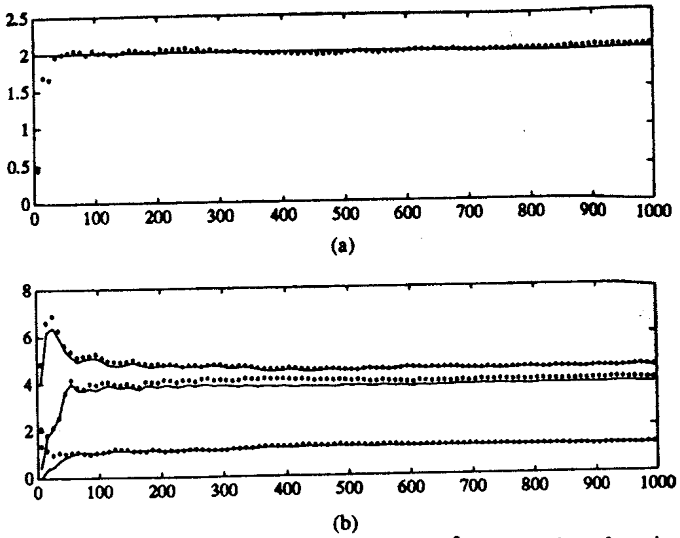


Fig. 4. (a) Estimate ν_{est}^2 (* * *) of the variance ν^2 as a function of p using (28). (b) Estimates (* * *) of the exact singular values (full line) as a function of p using (29).

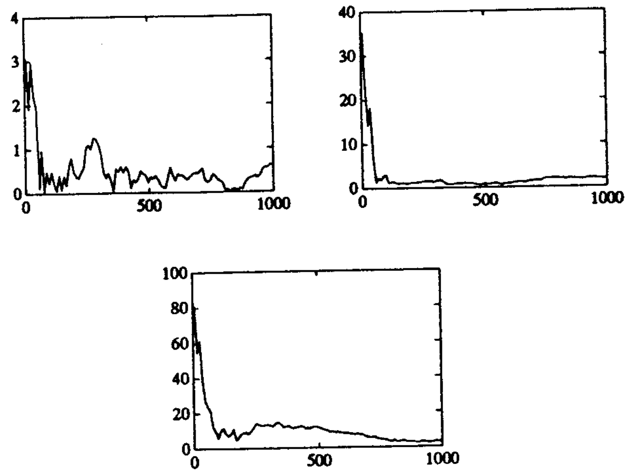


Fig. 5. The three canonical angles between $R(V_{e1})$ and $R(V_{m1})$ as a function of p, illustrating the fact that the estimated short space converges to the exact short space.

The diagonal matrix $(CS_{e1})^{-2}$ is the covariance matrix of the coordinates of the projections of the singular vectors that generate the short space (the signal subspace) onto the exact null space. This explains why $W = (CS_{e1})^2$ is the optimal weighting matrix in the so-called weighted subspace approach. Hence a maximum likelihood estimate of the r dimensional short space can be obtained from the solution of a quadratic optimization problem

$$\min_{\text{over} A, T \in \mathbb{R}^{r\times r}} \| V_{m1} W^{1/2} - AT \|_F^2$$

subject to some structural constraints on A (which maybe depends on some parameter vector θ (as in ESPRIT)) or has some structure (e.g., a shift structure in system identification). In this criterion V_{m1} are the r dominant singular vectors of the data matrix in the short space and W is the optimal weighting as described before. For more details refer to [17]. Even if the signal vectors in E are not

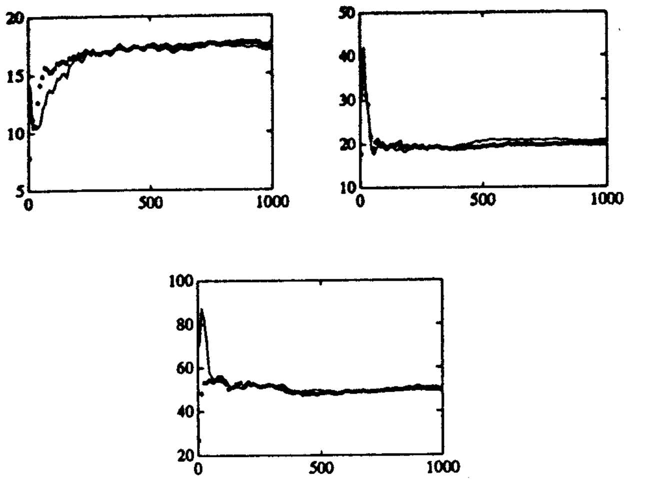


Fig. 6. Canonical angles between $R(U_{m1})$ and $R(U_{e1})$ (full line) and their estimates (* * *), using (28), (29), and (16).

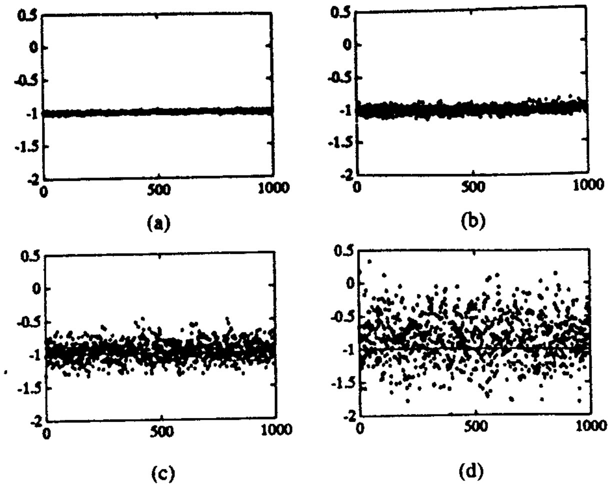


Fig. 8. Minimum variance reconstruction of third column of E, using (18) with $p = 996$ for four different elementwise noise variances: (a) $\nu^2 = 0.001$; (b) $\nu^2 = 0.01$; (c) $\nu^2 = 0.1$; (d) $\nu^2 = 1$. It is clearly visible how the "uncertainty" grows for increasing noise variance.

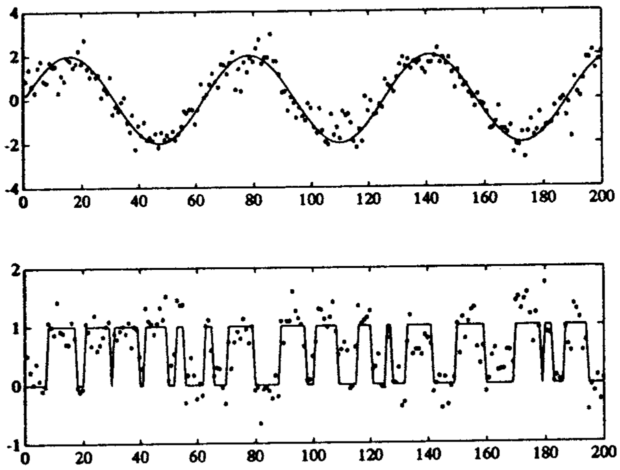


Fig. 7. Minimum variance reconstruction (* * *) of the first and second column of E (exact: full line) using (18) for $p = 996$. The noise variance is $\nu^2 = 2$.

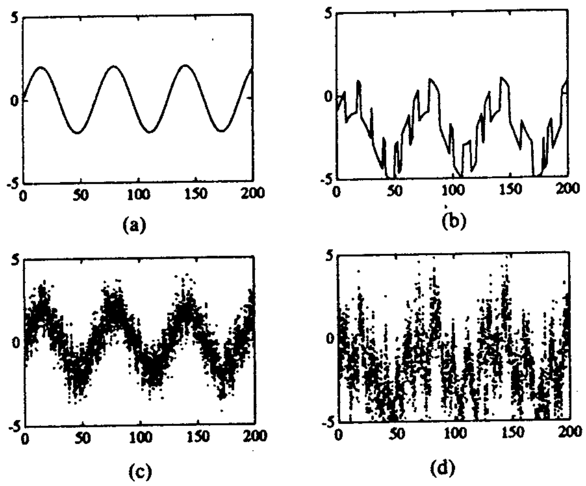


Fig. 9. (a) First column of E; (b) Fourth column of E; (c) and (d). Matrix ball of solutions around the minimum variance estimate, generated from (22) by plotting the results for 20 random choices of the orthonormal matrix Z in (21). The noise variance is $\nu^2 = 2$.

Gaussian but the rows of N still are, the asymptotic distribution of the eigenvectors of the sample covariance matrix $M^t M/p$ is also normal. The explicit distribution is given in, e.g., [4] [17, p. 207]. This distribution even applies when E is deterministic but quasi-stationary.

VI. A Numerical Example

In order to illustrate the main points, we now present a numerical example which was simulated using Matlab. A matrix E was constructed as follows: Its first column contains a sinusoid $2 * \sin(0.1 * k)$ where k is the row index. Its second column is pseudorandom binary noise, the switching moments of which are integer values between 0 and 10 that are all equally probable. The third column contains a constant signal with amplitude -1. Columns 4 to 6 are obtained by postmultiplying columns 1 to 3 with the matrix

$$\begin{pmatrix} 1 & -2 & 1 \\ -2 & 3 & 4 \\ 1 & 2 & -1 \end{pmatrix}.$$

The number of rows of E is p, where p increases from 6 to 1000. Obviously, for all values of p, the rank of E is 3. We also generated a 1000×6 noise matrix N, the elements of which were normally distributed with mean zero and variance $\nu^2 = 2$. Some results are presented in Figs. 4–9.

VII. Applications

We invite the reader to apply the insights of this paper to the following four subspace techniques. We only provide a brief summary of each of them together with some relevant characteristics. For more details, appropriate references are given.

A. Total Linear Least Squares

The total linear least squares approach is an alternative for the classical linear least squares scheme for solving overdetermined sets of linear equations, in the case that all data are corrupted by additive noise. Consider an inconsistent overdetermined set of linear equations $Ax \approx b$ where $A \in \mathbb{R}^{p \times (q-1)}$ and for simplicity we take rank $(A) = q - 1 << p$. A geometric interpretation of the least squares solution $x = (A^tA)^{-1}A^tb$ is that, first the right-hand side b is projected orthogonally onto $R(A)$, followed by a solution of the resulting consistent set of linear equations. Obviously, only the right-hand side b is modified to obtain a solution. The total linear least squares approach tries to modify all data in both A and b, with minimal effort, so that we obtain a consistent set of equations. This can be formulated as the following optimization problem

$$\min_{x, B \in \mathbb{R}^{p \times q}} \|[A \quad b] - B\|_F^2$$

subject to

$$B \begin{pmatrix} x \\ -1 \end{pmatrix} = 0.$$

It can be shown that the solution x can be obtained from the SVD of the concatenated matrix $[A \quad b]$ by scaling the right singular vector corresponding to the smallest singular value, so that its last component is -1. The solution x corresponds then to the first $(q - 1)$ components of this scaled right singular vector. Some references are [8], [12], [13], [22].

Using the notation of this paper, let us call $M = [A \quad b]$. The problem reduces to the least squares approximation of a matrix of observations $M = E + N$ by a rank deficient matrix, in precisely the same way as in Section II-D. The null space of E contains the linear relations and the rank r of E determines the number of linearly independent linear relations among its columns, which is $q - r$. Hence, if the elements of the noise N are zero mean and have bounded fourth-order moments, we can recover the original linear relations asymptotically as $p \to \infty$ from the SVD of the data matrix M. The long space of E, which contains the noiseless signals, cannot be estimated consistently. A least squares estimate can be obtained from the SVD of M as described in Section II-D. These consistency results (i.e., the asymptotic unbiasedeness) for the total least squares problem in particular were derived in [10], [11] together with an expression for the asymptotic error covariance matrix of the total least squares solution (which goes to zero as the number of equations goes to infinity).

B. Direction-of-Arrival Estimation ESPRIT

The basic model in the ESPRIT approach (see, e.g., [19]) is

$$M = E + N = TS + N \in \mathbb{R}^{p \times q}, \qquad q >> p$$

where $S \in \mathbb{R}^{r \times q}$ contains the emitted (small-band) signals. Note that the long space here is the row space and the short space is the column space. $T \in \mathbb{R}^{p \times r}$ is a static transfer matrix, which has some additional structure which is imposed by the array manifold. The rank r of E determines the number of sources while the direction-of-arrival of each source can be determined from the short space of E via the solution of an eigenvalue problem. When the elements of N are zero mean and have bounded fourth-order moments, a consistent estimate of the number of sources r and the directions of arrival is obtained. It is impossible from the SVD of M to obtain a consistent estimate of the emitted signals which are the rows of S (the so-called signal copy problem). These signals belong to the long space of E and only a least squares or minimum variance estimate can be obtained. In this context, it was shown recently [17, p. 133] that the weighted subspace fitting method to estimate the angles of arrival, asymptotically achieves the lowest estimation error variance on the directions of arrival. It is remarkable that the optimal weighting matrix is precisely the square of the diagonal matrix (19), which contains the singular values of the minimum variance estimate (18).

C. Separation of the Fetal ECG From the Maternal ECG

The basic subspace model in this application is given by

$$M = E + N = (T_1 \quad T_2)\begin{pmatrix} S_1 \\ S_2 \end{pmatrix} + N \qquad \in \mathbb{R}^{p \times q}, \quad q >> p.$$

Here M is a matrix that contains cutaneous measurements. Typically three electrodes are placed on the abdomen and three others are placed on the thorax of the mother which results in $p = 6$ measurement channels. The mother's heartbeat is represented by $S_1 \in \mathbb{R}^{3 \times p}$, which can typically be decomposed in three orthogonal signals. T_1 represents the transfer, which is assumed to be static, from the mother heart to the measurement electrodes. The fetal ECG is represented by $S_2 \in \mathbb{R}^{2 \times p}$ which is two dimensional. T_2 is the static transfer from fetal ECG to electrodes. Typically the $5 \times p$ matrix

$$\begin{pmatrix} S_1 \\ S_2 \end{pmatrix}$$

is orthogonal and the smallest singular value of S_1 is larger than the largest of S_2. By an appropriate placing of the electrodes, one can ensure that also $(T_1 \quad T_2)$ is orthogonal. This implies that for this application, the SVD is most naturally reflected in the model itself! Extensive experiments have demonstrated the applicability of this approach together with the plausibility of some of the assumptions, such as, e.g., the ones concerning the noise N (see, e.g., [3], [21]). It can be concluded from the results in this paper that one cannot recover the original signals contained in the matrices S_1 and S_2. This is because they belong to the long space. One can, however, recover asymptotically the linear combinations that should be made of the columns of the exact matrix E, so that E would be decomposed into S_1 (the mother heartbeat) and S_2 (the fetal ECG). Despite the fact that we have the exact linear combinations, we can only apply them to the noisy measurement matrix M. This always results in inconsistent estimates. Equations (21) and (22) could be used to generate an uncertainty band around the minimum variance solution obtained from the SVD of M. Because physicians are typically interested in the signals themselves and not in the linear relations, this conclusion is a little bit disappointing. Extensive simulations and experiments however have demonstrated that the least squares estimate of the signals as described in Section II-D, still provide very useful medical diagnostic information (see, e.g., [3], [21]). The minimum variance estimate (18) suggested in this paper might even perform better, but this has not yet been verified experimentally.

D. Identification of State Space Models from Noisy Input-Output Data

In [6], [16] we have proved the following result. Assume that the vectors $w_k^t = (u_k^t \;\; y_k^t)^t$ contain the inputs $u_k \in \mathbb{R}^m$ and outputs $y_k \in \mathbb{R}^l$ of a multivariate linear time-invariant discrete-time system of order n. Construct a block Hankel matrix W with the vectors w_k, which has much more columns than rows and divide it into two parts of equal size: an upper part, which we call W_{past} and a lower part called W_{future}. Then, there is an n-dimensional nontrivial intersection between the row spaces of W_{past} and W_{future}. This intersection is nothing more than a state sequence of the linear dynamical system. The state space model can be elegantly computed from the short space of W via its SVD (for details see [6], [16]). The relevant elements for this subspace technique are: In the case of input-output data that are corrupted by *white*[4] noise, the short space can still be estimated consistently as the number of columns goes to infinity. This implies that we obtain consistent estimates of the state space matrices.

It is impossible to obtain a consistent estimate of the state sequence, since it belongs to the long space of the block Hankel matrix with input-output data. This result is apparently new and perhaps provides an answer to a long-standing open problem posed by Kalman in his path-breaking paper [14], namely, whether it is possible to obtain (consistent) estimates of the state from noise corrupted input-output data. Our approach suggests that the answer is no. If the noise is white and Gaussian, one can obtain consistent estimates of the state space model and then use a Kalman filter to obtain a least squares estimate of the state. The error covariance matrix of the estimate of the state is, however, not zero asymptotically.

[4]This is an additional assumption, which reflects a dynamical property of the noise. This assumption is enforced by the block Hankel structure of the data matrix and the requirement that N^tN must be a multiple of the identity matrix asymptotically.

VIII. Conclusions

In this paper, we have given a geometrical description and statistical description of a fundamental asymmetry in the consistency properties of estimates of the long and short space of a matrix. While the short space can be estimated consistently under certain orthogonality assumptions, the estimate of the long space suffers from an asymptotic bias, which can be characterized in terms of certain canonical angles. These can be estimated and allow us to describe a matrix ball of equivalent signal reconstruction, which describes the uncertainty in trying to reconstruct the long space. The central solution is the minimum variance estimate, which can be constructed from the SVD of the data matrix. We have shown that the geometric and algebraic conditions are satisfied for noisy matrices, if the noise elements are zero mean and have bounded fourth-order moments. The results were illustrated by a numerical example and four different signal processing applications were enumerated.

It is true that in this paper we have only analyzed the asymptotic behavior of the SVD of noisy matrices. It has been observed by many users (and it can be verified experimentally) that SVD-based algorithms remain robust with noisy data matrices that are only slightly overdetermined (short data records). The analysis in this case is much more involved; the main reason for the difficulty in the finite case is the lack of orthogonality between the column spaces of E and N, which corresponds to a certain (but unknown) amount of correlation between the exact signals and the noise. One could analyze the finite case statistically. However, there do not yet exist fully satisfactory approaches, although some analysis of the finite case can be found in [17].

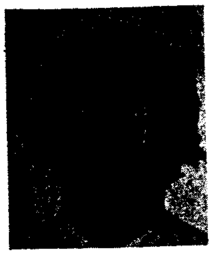
Acknowledgment

The author would like to thank D. Callaerts, B. Ottersten, R. Roy, J. Vandewalle, S. Van Huffel, P. Van Overschee, and M. Viberg for many suggestions which have greatly improved this paper's overall presentation. Also, the constructive critiques of the four reviewers were instrumental in obtaining this final version.

References

[1] T. W. Anderson, "Asymptotic theory for principal component analysis," *Ann. Math. Stat.*, vol. 34, pp. 122-148, 1963.
[2] T. W. Anderson, "Multivariate linear relations," in *Proc. Second Int. Tampere Conf. Stat.*, Tampere, Finland, 1987, pp. 9-36.
[3] D. Callaerts, B. De Moor, J. Vandewalle, and W. Sansen, "Comparison of SVD methods to extract the fetal electrocardiogram from cutaneous electrode signals," *Med. Biol. Eng. Computing*, vol. 28, pp. 217-224, May 1990.
[4] A. W. Davis, "Asymptotic theory for principal component analysis: Nonnormal case," *Aust. J. Stat.*, vol. 19, no. 3, pp. 206-212, 1977.

[5] B. De Moor, J. Staar, and J. Vandewalle, "Oriented energy and oriented signal-to-signal ratio concepts in the analysis of vector sequences and time series," in *SVD and Signal Processing*, E. Deprettere, Ed. New-Holland, 1988, pp. 209-232.
[6] B. De Moor, "Mathematical concepts and techniques for modelling of static and dynamic systems," Doctoral dissertation, Dept. Elect. Eng., Katholieke Universiteit Leuven, Belgium, June 1988.
[7] B. De Moor and G. H. Golub, "Generalized singular value decompositions: A proposal for a standardized nomenclature," ESAT-SISTA Rep. 1989-10, 15 pp., Apr. 1989, Dep. Elec. Eng., Katholieke Universiteit Leuven, Belgium; Numer. Anal. Proj. Manuscript NA-89-04, Dep. Comput. Sci., Stanford Univ., Stanford, CA.
[8] B. De Moor and J. Vandewalle, "A unifying theorem for linear and total linear least squares," *IEEE Trans. Automat. Contr.*, vol. 35, no. 5, pp. 563-566, May 1990.
[9] B. De Moor, J. Swevers, and H. Van Brussel, "Stepped sine system identification, errors-in-variables, and the quotient singular value decomposition," *Mechanical Syst. Signal Processing*, no. 5, vol. 6, Nov. 1991.
[10] W. A. Fuller, *Measurement Error Models*. New York: Wiley, 1987.
[11] L. J. Gleser, "Estimation in a multivariate 'errors in variables' regression model: Large sample results," *Ann. Stat.*, vol. 9, pp. 24-44, 1981.
[12] G. H. Golub and C. Van Loan, "An analysis of the total least squares problem," *SIAM J. Numer. Anal.*, vol. 17, pp. 883-893, 1980.
[13] G. Golub and C. Van Loan, *Matrix Computations*, second ed. Baltimore, MD: John Hopkins University Press, 1989.
[14] R. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Eng.*, vol. 82, pp. 35-45, Mar. 1960; also reprinted in *Linear Least-Squares Estimation*, Benchmark Papers in Electrical Engineering and Computer Science, 17, T. Kailath, Ed. Dowden, Hutchinson and Ross, Inc., 1977, pp. 254-264.
[15] G. Letac, "Isotropy and sphericity: Some characteristics of the normal distribution," *Ann. Stat.*, vol. 9, no. 2, pp. 408-417, 1981.
[16] M. Moonen, B. De Moor, L. Vandenberghe, and J. Vandewalle, "On- and off-line identification of linear state-space models," *Int. J. Contr.*, vol. 49, pp. 219-232, Jan. 1989.
[17] B. Ottersten, "Parametric subspace fitting methods for array signal processing," Ph.D. dissertation, Inform. Syst. Lab., Dep. Elec. Eng., Stanford Univ., Dec. 1989.
[18] A. Papoulis, *Probability Random Variables, and Stochastic Processes*. New York: McGraw-Hill, 1965.
[19] R. Roy and T. Kailath, "ESPRIT: Estimation of parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 7, pp. 984-995, July 1989.
[20] G. Stewart and J. Sun, *Matrix Perturbation Theory*. London: Academic, 1990.
[21] J. Vanderschoot, D. Callaerts, W. Sansen, J. Vandewalle, G. Vantrappen, and J. Janssens, "Two methods for optimal MECG elimination and FECG detection from skin electrode signals," *IEEE Trans. Biomed. Eng.*, vol. BME-34, no. 3, pp. 233-243, Mar. 1987.
[22] S. Van Huffel and J. Vandewalle, "Algebraic connections between the least squares and total least squares problems," *Numer. Math.*, vol. 55, no. 4, pp. 431-449, July 1989.



Bart De Moor (M'86) was born in Halle, Brabant, Belgium, on July 12, 1960. He received the degree of electromechanical engineering in control theory in 1983 and the doctoral degree in applied sciences in 1988, both from the Katholieke Universiteit Leuven, Belgium.

He was a Visiting Research Associate (1988-1989) at the Departments of Computer Science (Numerical Analysis) and Electrical Engineering (Information Systems Lab) of Stanford University, Stanford, CA. Since October 1989 he has been a Research Associate of the National Fund for Scientific Research of Belgium (NFWO) and Assistant Professor at the Department of Electrical Engineering of the Katholieke Universiteit Leuven. In 1991-1992 he was the Chief of Staff for the Belgian National Minister of Science, W. Demeester-DeMeyer. His research interests include systems and control theory and numerical linear algebra.

Dr. De Moor received an Advanced Research Fellowship in Science and Technology from the NATO Science Fellowships Programme (1988), the Leybold-Heraeus Prize (1986), the Leslie Fox Prize (1989), and the Guillemin-Cauer Best Paper Award from the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS (1990). In 1992 he was a Laureat of the Belgian Royal Academy of Sciences.