A DATA VERSION OF THE GAUSS-MARKOV THEOREM AND ITS APPLICATION TO ADAPTIVE SUBSPACE SPLITTING

Louis L. Scharf and John K. Thomas
Dept. Of Electrical and Computer Engineering
University of Colorado
Boulder, Co-80309
USA

ABSTRACT

How does one adaptively split a measurement subspace into signal and orthogonal subspaces of reduced rank so that detectors, estimators, and quantizers may be adaptively designed from experimental data? We provide some answers to this question by decomposing experimental correlations into their Wishart distributed Schur complements and showing how these distributions may be used to identify subspaces.

1.0 INTRODUCTION

In [1], the authors outlined a number of problems where low-rank approximations reduce the sum of bias-squared plus variance to minimize mean-squared error. These ideas were further extended in [2]. All of the solutions advocated in these two papers required knowledge of the second-order information for their implementation. Since then, there has been a considerable amount of work done on subspace identification as a preliminary step to rank reduction [3] - [5]. In [3] the authors derived hypothesis tests to determine signal rank in the presence of AWGN, assuming that a useful decomposition of subspaces can be made. The work of [4] and [5] has clarified the probability that a useful subspace decomposition can be made.

In this paper we pose a sequence of problems which generalize the problems in [1] - [5] and illuminate the study of subspace splitting from random data. These problems and the connections between them are illustrated in Figure 1. Problem 1 is a standard linear minimum mean squared error (LMMSE) estimation problem of estimating a random vector x from a random vector y when second-order information is known. When snapshots of x and y, assembled in matrices X and Y, are available instead of the second-order information, a least squares (LS) approach is used to estimate the matrix X from Y. This is Problem 2 in Figure 1. The performance of the LS estimator can be characterized in terms of estimation error covariances from Problem 1, and this forms the connection between the two problems. In Problems 3 and 4, we specialize the definitions of X and Y. In Problem 3, X and Y are resolutions of a data matrix onto subspaces generated by known second-order information. In Problem 4, X and Y are projections onto estimated subspaces generated by the data matrix itself. First-order perturbation models for Problem 3 may be used to analyze Problem 4, and this forms the connection between the two problems.

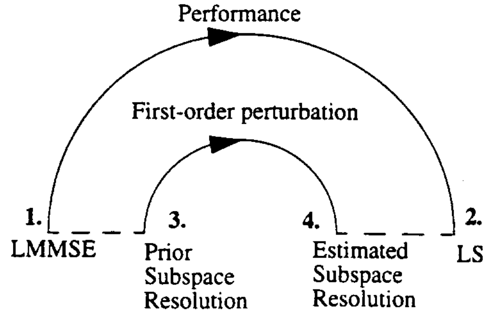

FIGURE 1. Framework of Problems

2.0 Problem #1: LMMSE Estimation

In Problem #1, , we are given a single snapshot of N-dimensional data, $z^* = [x^* \;\; y^*]$, drawn from a distribution with known covariance R_{zz}, and asked to estimate x from y. The covariance matrix exhibits the following structure and Schur decomposition:

$$R_{zz} = Ezz^* = \begin{bmatrix} R_{xx} & R_{xy} \\ R_{yx} & R_{yy} \end{bmatrix}$$
$$= \begin{bmatrix} I & T \\ 0 & I \end{bmatrix} \begin{bmatrix} Q & 0 \\ 0 & R_{yy} \end{bmatrix} \begin{bmatrix} I & 0 \\ T^* & I \end{bmatrix} = \begin{bmatrix} I & T \\ 0 & I \end{bmatrix} E\left[\begin{bmatrix} e_x \\ y \end{bmatrix} \begin{bmatrix} e_x^* & y^* \end{bmatrix}\right] \begin{bmatrix} I & 0 \\ T^* & I \end{bmatrix} \quad (1)$$

In this decomposition, T, $\hat{x}$, e_x, and Q are defined as follows:

$$\begin{aligned} T &= R_{xy}R_{yy}^{-1} \text{ : LMMSE filter} \\ \hat{x} &= Ty \text{ : LMMSE estimate of x} \\ e_x &= x - Ty \text{ : estimator error} \\ Q &= E[e_x e_x^*] = R_{xx} - R_{xy}R_{yy}^{-1}R_{yx} \text{: error covariance} \end{aligned} \quad (2)$$

This is a well understood problem with all its connections to the Wiener-Hopf equations, conditional mean estimators, and the Statistician's Pythagorean Theorem [6]. Analogous results for estimating y from x can be obtained.

0-7803-2431-5/95 $4.00 © 1995 IEEE

3.0 Problem #2: Least Squares Estimation

When second-order information is unavailable, as in adaptive filtering, we use snapshots of x and y, assembled into data matrices X and Y, to approximate the LMMSE filter T with the LS filter $\hat{T}$. Assuming that the snapshots are independent Gaussian random vectors, we derive the performance of the LS filter and connect it with error covariances from Problem 1. We then derive a reduced-rank approximation to the LS filter.

3.1 Filter Characterization

Let Z represent an NxM data matrix (M>N), which can be partitioned into X (rxM) and Y (N-rxM):

$$Z = \begin{bmatrix} z_1 & \dots & z_M \end{bmatrix} = \begin{bmatrix} x_1 & \dots & x_M \\ y_1 & \dots & y_M \end{bmatrix} = \begin{bmatrix} X \\ Y \end{bmatrix} \quad (3)$$

The correlation matrix for Z is a crude estimate of R_{zz}:

$$\hat{R}_{zz} = ZZ^* = \begin{bmatrix} XX^* & XY^* \\ YX^* & YY^* \end{bmatrix} \quad (4)$$

The Schur decomposition of ZZ^* exhibits the following structure:

$$ZZ^* = \begin{bmatrix} I & \hat{T} \\ 0 & I \end{bmatrix} \begin{bmatrix} \hat{Q} & 0 \\ 0 & YY^* \end{bmatrix} \begin{bmatrix} I & 0 \\ \hat{T}^* & I \end{bmatrix} \quad (5)$$

In this decomposition $\hat{T}$, $\hat{X}$, $\hat{E}_x$, and $\hat{Q}$ are defined as follows:

$$\begin{aligned} \hat{T} &= XY^*(YY^*)^{-1} : \text{LS filter} \\ \hat{X} &= \hat{T}Y = XY^*(YY^*)^{-1}Y = XP_{Y*} : \text{LS estimate of } X \\ \hat{E}_X &= X - \hat{X} = X - XP_{Y*} = XP_{Y*}^{\perp} : \text{estimator error} \\ \hat{Q} &= \hat{E}_X\hat{E}_X^* = XX^* - XY^*(YY^*)^{-1}YX^* \\ &= XP_{Y*}^{\perp}X^* : \text{error covariance} \end{aligned} \quad (6)$$

The error matrix $\hat{E}_X$ is orthogonal to the estimator $\hat{X}$ and the input matrix Y:

$$\hat{X}\hat{E}_X^* = XP_{Y*}P_{Y*}^{\perp}X^* = 0 \quad (7)$$

$$Y\hat{E}_X^* = YP_{Y*}^{\perp}X^* = 0 \quad (8)$$

These formulas mimic the structure and behavior of the Gauss-Markov formulas of Problem 1, with exact covariances replaced by estimates. Furthermore, $\hat{X}$ and $\hat{E}_X$ orthogonally decompose X:

$$\begin{aligned} X &= X\left(P_{Y*} + P_{Y*}^{\perp}\right) = \hat{X} + \hat{E}_X \\ XX^* &= \hat{X}\hat{X}^* + \hat{E}_X\hat{E}_X^* \end{aligned} \quad (9)$$

Equation (9) leads to the geometric interpretation illustrated in Figure 2.

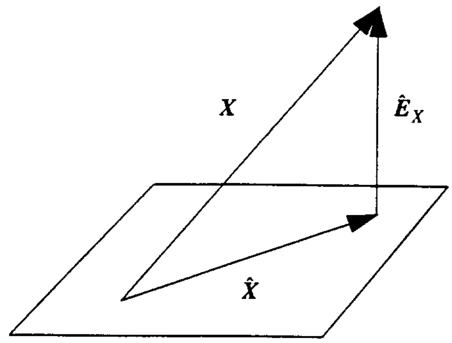

FIGURE 2. Geometric Interpretation

3.2 Performance

Suppose the columns of the data matrix Z are Gaussian, independent, and identically distributed with zero mean and covariance matrix R_{zz}. Then the conditional distributions of $\hat{X}$, $\hat{E}_X$, and X can be characterized in terms of the *Vec(.)* operator:

$$\begin{aligned} Vec(\hat{X}|Y) &: N[(P_{Y*} \otimes I)m, P_{Y*} \otimes Q] \\ Vec(\hat{E}_X|Y) &: N\left[\left(P_{Y*}^{\perp} \otimes I\right)m, P_{Y*}^{\perp} \otimes Q\right] \\ Vec(X|Y) &: N[m, I \otimes Q] \\ m &= Vec(R_{xy}R_{yy}^{-1}Y) \end{aligned} \quad (10)$$

The means and covariances of $\hat{X}$ and $\hat{E}_X$ add to give the mean and covariance of X. This is a statement of the Statistician's Pythagorean Theorem [6].

The random matrix XX^* is Wishart distributed with parameterization $W_r(M, R_{XX}, 0)$, where $W_r(M, Q, \Theta)$ denotes a Wishart distribution with M degrees of freedom, covariance Q, and non-centrality parameter Θ for an rxr matrix ($\Theta = 0$ is a central Wishart). The conditional distributions of $\hat{X}\hat{X}^*$, and $\hat{E}_X\hat{E}_X^*$, given Y, are Wishart with the following parameterizations:

$$\begin{aligned} \hat{X}\hat{X}^*|Y &: W_r(M - r, Q, \Theta) \; ; \; \Theta = R_{xy}R_{yy}^{-1}YY^*R_{yy}^{-1}R_{yx} \\ \hat{E}_X\hat{E}_X^*|Y &: W_r(M - N + r, Q, 0) \end{aligned} \quad (11)$$

Since the conditional distribution of $\hat{E}_X\hat{E}_X^*$ is independent of Y, the unconditional distribution remains Wishart. $\hat{E}_X$ and $\hat{X}$ are conditionally Gaussian and orthogonal, meaning $\hat{X}\hat{X}^*$ and

$\hat{E}_X\hat{E}_X^*$ are conditionally independent. The conditional joint density of $\hat{X}\hat{X}^*$ and $\hat{E}_X\hat{E}_X^*$ can be factored into its conditional marginals:

$$\begin{aligned} f\left(\hat{X}\hat{X}^*, \hat{E}_X\hat{E}_X^*|Y\right) &= f(\hat{X}\hat{X}^*|Y)\, f\left(\hat{E}_X\hat{E}_X^*|Y\right) \\ &= f(\hat{X}\hat{X}^*|Y)\, f\left(\hat{E}_X\hat{E}_X^*\right) \end{aligned} \quad (12)$$

Taking the expectation over Y, shows $\hat{X}\hat{X}^*$ and $\hat{E}_X\hat{E}_X^*$ are unconditionally independent:

$$f\left(\hat{X}\hat{X}^*, \hat{E}_X\hat{E}_X^*\right) = f(\hat{X}\hat{X}^*)\, f\left(\hat{E}_X\hat{E}_X^*\right) \quad (13)$$

XX^* is the sum of two independent random matrices, one of which $(\hat{E}_X\hat{E}_X^*)$ is $W_r(M-N+r, Q, 0)$. This allows us to write the characteristic function of $\hat{X}\hat{X}^*$ as the ratio of the characteristic function of XX^* and $\hat{E}_X\hat{E}_X^*$:

$$\Phi_{\hat{X}\hat{X}^*}(j\Omega) = \frac{|I + j\Omega R_{XX}|^{-M}}{|I + j\Omega Q|^{-(M-N+r)}} \quad (14)$$

This is a generalized Wishart that completely characterizes the performance of the estimator $\hat{X}$.

3.3 Applications

In this section we approximate the LS filter $\hat{T}$ with a low rank filter $\hat{T}_k$. Figure 3 illustrates the synthesis and analysis filters for going back and forth between the data X and the error $\hat{E}_X$.

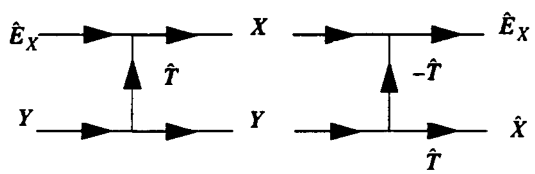

FIGURE 3. Filtering interpretation of Problem #2

Replace $\hat{T}$ in the figure with a low rank approximation, $\hat{T}_k$, to produces $\hat{X}_k$ and $\hat{E}_k$, which are the low rank approximations of $\hat{X}$ and $\hat{E}_X$:

$$\begin{bmatrix}\hat{E}_k \\ \hat{X}_k\end{bmatrix} = \begin{bmatrix} I & -\hat{T}_k \\ 0 & \hat{T}_k \end{bmatrix}\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} I & \hat{T} - \hat{T}_k \\ 0 & \hat{T}_k \end{bmatrix}\begin{bmatrix} \hat{E}_X \\ Y \end{bmatrix} \quad (15)$$

The error covariance of the reduced rank estimator is

$$\hat{E}_k\hat{E}_k^* = \hat{Q} + (\hat{T} - \hat{T}_k)\, YY^*\, (\hat{T} - \hat{T}_k)^* \quad (16)$$

If $\hat{T}(YY^*)^{1/2} = U\Sigma V^*$, where $\Sigma = diag\{\Sigma_k, \Sigma_{r-k}\}$, then minimization of the extra covariance $(\hat{T} - \hat{T}_k)\, YY^*\, (\hat{T} - \hat{T}_k)^*$ yields the solution

$$\hat{T}_k = U\begin{bmatrix} \Sigma_k & 0 \\ 0 & 0 \end{bmatrix} V^*(YY^*)^{-1/2} = U\begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix} U^*\hat{T} = P_{U_k}\hat{T} \quad (17)$$

Equation (17) characterizes the reduced rank LS filter $\hat{T}_k$, which is a data version of a reduced rank Wiener filter [6].

4.0 Problem 3: Resolution onto known prior subspaces

Suppose the matrices X and Y are projections of a data matrix, W, onto subspaces generated by the covariance matrix R_{ww}. Problems that fall into this category include vector quantizers, low-rank approximations, and threshold effects.

Define the matrix Z as a resolution of a data matrix W onto the modes of W:

$$\begin{aligned} Z &= \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} U_s^* \\ U_0^* \end{bmatrix} W \\ R_{ww} &= \left[U_s\ U_0\right] diag\{\Sigma_s, \Sigma_0\} \left[U_s\ U_0\right]^* \end{aligned} \quad (18)$$

If the columns of W are drawn from a Gaussian distribution, then the columns of Z are drawn from a Gaussian distribution with covariance R_{zz}:

$$R_{zz} = diag\{\Sigma_s, \Sigma_0\} \quad (19)$$

The least squares filter for estimating X from Y, and the corresponding error covariance, can be obtained by specializing the formulas in equations (6):

$$\begin{aligned} \hat{T}_s &= U^*_s WW^* U_0 (U^*_0 WW^* U_0)^{-1} : \text{LS filter} \\ \hat{E}_s\hat{E}_s^* &= U_s^* W P^{\perp}_{(U_0^* W)^*} W^* U_s : \text{error covariance} \end{aligned} \quad (20)$$

The statistical distributions of $\hat{E}_s\hat{E}_s^*$, $\hat{X}\hat{X}^*$ and XX^* for this problem are obtained by specializing the parameterization of the Wishart distributions in the previous section:

$$\hat{E}_s\hat{E}_s^*: W_r(M-N+r, \Sigma_s, 0)$$
$$XX^* = U_s^*ZZ^*U_s : W_r(M, \Sigma_s, 0) \quad (21)$$
$$\hat{X}\hat{X}^* : W_r(N-r, \Sigma_s, 0)$$

Similarly, the least squares filter, and the corresponding error covariance for estimating Y from X is

$$\hat{T}_0 = U^*_0WW^*U_s(U^*_sWW^*U_s)^{-1} : \text{LS filter}$$
$$\hat{E}_0\hat{E}_0^* = U^*_0WP^{\perp}_{(U^*_sW)^*}W^*U_0 : \text{error covariance} \quad (22)$$

The distributions of $\hat{E}_0\hat{E}_0^*$, YY^*, and $\hat{Y}\hat{Y}^*$ are

$$\hat{E}_0\hat{E}_0^*: W_{N-r}(M-r, \Sigma_0, 0)$$
$$YY^* = U_0^*ZZ^*U_0 : W_{N-r}(M, \Sigma_0, 0) \quad (23)$$
$$\hat{Y}\hat{Y}^* : W_{N-r}(r, \Sigma_0, 0)$$

Sufficient conditions that a useful subspace decomposition can be made were identified in [4] and [5]. These conditions involve the energies resolved by projecting measurements onto known prior subspaces. In the notation in this paper, the probability of a successful decomposition is :

$$1 - P\left[\frac{trYY^*}{N-r} - \frac{trXX^*}{r} > 0\right] \quad (24)$$

From the distributions of XX^* and YY^* we may numerically compute this probability [5].

5.0 Problem 4: Resolution onto estimated subspaces

Finally, we project the data matrix W onto estimated subspaces U_1 and U_2 to produce X and Y. Order selection, universal coders, and low-rank approximations of random vectors are examples of problems that fall into this class. A first-order perturbation model for the estimated subspaces ties performance of solutions in this class of problems to error covariances in Problem 3.

Define the matrix Z to be

$$Z = \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} U_1^* \\ U_2^* \end{bmatrix} W \quad (25)$$

where U_1 and U_2 are subspaces generated by W:

$$W = \begin{bmatrix} U_1 & U_2 \end{bmatrix} diag\{\Sigma_1, \Sigma_2\} \begin{bmatrix} U_1 & U_2 \end{bmatrix}^* \quad (26)$$

The correlation matrix for Z is

$$ZZ^* = diag\{\Sigma_1, \Sigma_2\} \quad (27)$$

The least squares filters and corresponding covariances are

$$\hat{T}_1 = 0 : \hat{Q}_1 = \Sigma_1$$
$$\hat{T}_2 = 0 : \hat{Q}_2 = \Sigma_2 \quad (28)$$

In [3], the authors model U_2 in terms of U_s and U_0 under the assumption that a useful decomposition can be made. Their model is $U_2 = (U_0 + TU_s)V$, where V is an arbitrary rotation matrix. They solve for T by minimizing $\|U_0^*W + TU_s^*W\|_F$. But this minimization produces the LS estimation of Y from X with T described in equation (22). Therefore, we conclude that the model for U_2 is a rotated version of the error in estimating Y from X, and the energy resolved onto U_2 is approximated by $tr\left(\hat{E}_0\hat{E}_0^*\right)$, where $\hat{E}_0\hat{E}_0^*$ is described in equation [22]. Similarly, the energy in U_1 is approximated as $tr\left(\hat{E}_s\hat{E}_s^*\right)$.

6.0 REFERENCES

1. L. L. Scharf and D.W. Tufts, "Rank Reduction for Modeling Stationary Signals, *IEEE Trans ASSP* **ASSP-35**:3 (March 1987).
2. L.L. Scharf, "The SVD and Reduced-Rank Signal Processing", *Signal Processing* **24**, pp.111-130 (November 1991).
3. A. Shah and D.W. Tufts, "Model Order Determination", *Proc. 22nd Asilomar Conf on Signals, Systems, and Computers*, Pacific Grove, CA (November 1993).
4. D.W.Tufts, A.C.Kot, and R.J.Vaccaro, "The Threshold Effect in Signal Processing Algorithms which Use an Estimated Subspace", *SVD and Signal Processing II: Algorithms, Analysis, and Applications*, R.J.Vaccaro (ED.) (Elsevier Science Publishing Company, Inc., 1991).
5. J.K.Thomas, L.L.Scharf, and D.W.Tufts, "The Probability of a Subspace Swap in the SVD", to be published in *IEEE Trans Signal Proc* (1995).
6. L.L. Scharf, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*, (New York: Addison-Wesley Publishing Company, 1990).