knowledge about $\bar{\boldsymbol{M}}_p$. In this figure, we also display the MSE of the MMSE estimator derived in [4]. The total number of snapshots is $N = 20$ and $\nu = 20$. As can be observed, for a given number of snapshots, the BB is smaller when $L_k = 1$ than when $K = 1$, which confirms the previous observations made on the CRB. Also, as could be expected, the BB decreases as μ increases, i.e., as the prior is more and more informative. Finally, we note that the MMSE estimator has a MSE close to the BB only for large values of μ.

V. Concluding Remarks

This correspondence derived lower bounds on the estimation of a covariance matrix $\boldsymbol{M}_p$ using heterogeneous samples $\boldsymbol{Z}_k$, $k = 1, \ldots, K$, which have covariance matrices $\boldsymbol{M}_k$ different from $\boldsymbol{M}_p$. When $\boldsymbol{M}_p$ is deterministic, we showed that consistent estimation of $\boldsymbol{M}_p$ is not feasible, when all samples share the same covariance matrix, i.e., when $K = 1$. Indeed, the CRB does not converge to zero as the number of training samples increases. In contrast, if all snapshots have different covariance matrices, randomly distributed around $\boldsymbol{M}_p$ (i.e., $L_k = 1$, for $k = 1, \ldots, K$), the CRB goes to zero when the number of training samples increases. The correspondence also derived the Bayesian bound associated to a random covariance matrix $\boldsymbol{M}_p$. The bounds derived herein enable one to quantify the degradation induced by heterogeneity, and can serve as references for any estimator of the covariance matrix $\boldsymbol{M}_p$.

Acknowledgment

The authors would like to thank Prof. G. Letac for enthusiastically sharing his expert knowledge on multivariate Wishart and beta distributions.

References

[1] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation and Time Series Analysis*. Reading, MA: Addison-Wesley, 1991.
[2] W. L. Melvin, "Space-time adaptive radar performance in heterogeneous clutter," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 36, no. 2, pp. 621–633, Apr. 2000.
[3] W. L. Melvin, "A STAP overview," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 19, no. 1 , pt. 2, pp. 19–35, Jan. 2004.
[4] O. Besson, S. Bidon, and J.-Y. Tourneret, "Covariance matrix estimation with heterogeneous samples," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 909–920, Mar. 2008.
[5] S. Bidon, O. Besson, and J.-Y. Tourneret, "A Bayesian approach to adaptive detection in non-homogeneous environments," *IEEE Trans. Signal Process.*, vol. 56, no. 1, pp. 205–217, Jan. 2008.
[6] J. A. Tague and C. I. Caldwell, "Expectations of useful complex Wishart forms," *Multidimen. Syst. Signal Process.*, vol. 5, pp. 263–279, 1994.
[7] C. G. Khatri and C. R. Rao, "Effects of estimated noise covariance matrix in optimal signal detection," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 5, pp. 671–679, May 1987.
[8] H. L. V. Trees, *Optimum Array Processing*. New York: Wiley, 2002.
[9] H. Lütkepohl, *Handbook of Matrices*. Chichester, U.K.: Wiley, 1996.
[10] C. G. Khatri, "Classical statistical analysis based on a certain multivariate complex Gaussian distribution," *Ann. Math. Stat.*, vol. 36, no. 1, pp. 98–114, Feb. 1965.
[11] M. Capitaine and M. Casalis, "Asymptotic freeness by generalized moments for Gaussian and Wishart matrices," *Indiana Univ. Math. J.*, vol. 53, no. 2, pp. 397–431, 2004.
[12] M. Capitaine and M. Casalis, "Cumulants for random matrices as convolutions on the symmetric group," *Probab. Theory Relat. Fields*, vol. 136, no. 1, pp. 19–36, Sep. 2006.
[13] G. Letac, Expectation of $Z_{ij}Z_{kl}$ for a Matrix Beta Law 2007, private communication.

Cross Entropy Approximation of Structured Gaussian Covariance Matrices

Cheng-Yuan Liou and Bruce R. Musicus

***Abstract*—We apply two variations of the principle of minimum cross entropy (the Kullback information measure) to fit parameterized probability density models to observed data densities. For an array beamforming problem with P incident narrowband point sources, $N > P$ sensors, and colored noise, both approaches yield eigenvector fitting methods similar to that of the MUSIC algorithm and of the oblique transformation in factor analysis. Furthermore, the corresponding cross entropies (CE) are related to the MDL model order selection criterion .**

***Index Terms*—Array beamforming, eigenvector methods, factor analysis, generalized principle component analysis, Kullback information measure, minimum cross entropy (CE), oblique transformation, stochastic estimation, structured covariance.**

I. Introduction

Many existing high resolution methods for spectral analysis and for optimal beamforming utilize covariance matrices estimated from observed data. Often, an underlying structure for the covariance matrix is known in advance, and our goal is to estimate the covariance matrix with this structure which best fits the observed data. Previous literature has suggested a variety of methods of optimally estimating structured covariance matrices from data [1]–[5]. In this correspondence, we will apply the minimum cross entropy (CE) and minimum reverse cross-entropy (RCE) [6] principles to estimate the covariance matrix. These principles have proved to be quite powerful in a wide variety of signal processing applications, such as complex independent component analysis [7], [8], encoding mechanism [9]. They have been justified as being "optimal" under suitable assumptions. In Section II, we apply the CE and RCE procedures to the problem of estimating structured covariance matrices, and in Section III we demonstrate the utility of the idea for a beamforming application.

II. Problem Statement

Let $\underline{x}$ be an N-dimensional real or complex random vector. Assume that a Gaussian probability density for $\underline{x}$ is either known *a priori* or has been estimated by some procedure from observed data

$$p(\underline{x}) = N(\underline{m}, R) \tag{1}$$

where $\underline{m}$ is the expected value of $\underline{x}$, and R is the covariance matrix, $R = E[(\underline{x} - \underline{m})(\underline{x} - \underline{m})^H]$, and where $\underline{x}^H$ is the Hermitian (complex conjugate transpose) of $\underline{x}$. Suppose we wish to approximate this $p(\underline{x})$ with a parameterized probability density function (pdf)

$$q_\theta(\underline{x}) = N(\underline{m}_\theta, R_\theta) \tag{2}$$

where $\underline{\theta}$ denotes the unknown parameters in the model $q_\theta(\underline{x})$ which are to be estimated. Conceptually, we wish to choose $\underline{\theta}$ to make $q_\theta(\underline{x})$ op-

Manuscript received March 25, 2007; revised January 8, 2008. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sven Nordebo.

C.-Y. Liou is with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, R.O.C. (e-mail: cyliou@csie.ntu.edu.tw).

B. R. Musicus resides in Boston, MA 02421 USA (e-mail: bmusicus@rcn.com).

Digital Object Identifier 10.1109/TSP.2008.917878

timally match $p(\underline{x})$. An appropriate objective function is the Kullback information measure [6], otherwise known as the minimum CE principle [10]. Because this measure is asymmetric, we can apply it in two different ways to this problem. Note that there exist symmetric measures, such as the J-divergence [11]. Following [7], [8], [10], we call these the "Cross-Entropy" and "Reverse Cross-Entropy" methods

$$\mathrm{CE}: \hat{q}_\theta \leftarrow \min_{\underline{\theta}} H(q_\theta, p) \tag{3}$$

$$\mathrm{RCE}: \hat{q}_\theta \leftarrow \min_{\underline{\theta}} H(p, q_\theta) \tag{4}$$

where

$$H(p_1,p_2) = \int p_1(\underline{x}) \log \frac{p_1(\underline{x})}{p_2(\underline{x})} d\underline{x}. \tag{5}$$

Kullback [6] has argued that $H(p_1,p_2)$ measures the mean amount of information for discriminating in favor of the hypothesis that p_1 is the correct density of $\underline{x}$ rather than p_2. The measure $H(p_1,p_2)$ has several pleasing mathematical properties: it is convex in p_1, and convex in p_2, and attains its minimum value of zero when $p_1(\underline{x}) = p_2(\underline{x})$ almost everywhere. Another useful property is that estimating $\underline{\theta}$ from either (3) or (4) is straightforward. Substitute (1) and (2) into the CE and RCE formulas to obtain

$$\begin{aligned}\mathrm{CE}: H(q_\theta,p) =& \xi\{\mathrm{tr}(R^{-1}R_\theta) - N - \log|R^{-1}R_\theta| \\ &+ (\underline{m}_\theta - \underline{m})^H R^{-1}(\underline{m}_\theta - \underline{m})\} \\ \mathrm{RCE}: H(p,q_\theta) =& \xi\{\mathrm{tr}(R_\theta^{-1}R) - N - \log|R_\theta^{-1}R| \\ &+ (\underline{m}_\theta - \underline{m})^H R_\theta^{-1}(\underline{m}_\theta - \underline{m})\}\end{aligned}$$

where $\xi = 1/2$ when $\underline{x}$ is real and $\xi = 1$ when $\underline{x}$ is complex.

To simplify the remainder of the discussion, assume that the mean is known, $\underline{m}_\theta = \underline{m}$, so that we can focus on the estimation of the covariance matrix and compare the results with those by Burg, Luenberger, Wenger [4] and Gray, Anderson, Sim [5] . The two estimation problems reduce to minimizing

$$\mathrm{CE}: H(q_\theta,p) = \xi\{\mathrm{tr}(R^{-1}R_\theta) - N - \log|R^{-1}R_\theta|\} \tag{6}$$

$$\mathrm{RCE}: H(p,q_\theta) = \xi\{\mathrm{tr}(R_\theta^{-1}R) - N - \log|R_\theta^{-1}R|\} \tag{7}$$

Setting the gradients of the above two objective functions with respect to $\underline{\theta}$ to zero, we obtain the necessary conditions that $\underline{\widehat{\theta}}$ be the optimal solution

$$\mathrm{CE}: \mathrm{tr}\left\{\left(R^{-1} - R_\theta^{-1}\right)\frac{\partial R_\theta}{\partial \theta_i}\right\}\Bigg|_{\underline{\theta}=\underline{\widehat{\theta}}} = 0 \tag{8}$$

$$\mathrm{RCE}: \mathrm{tr}\left\{(R - R_\theta)\frac{\partial R_\theta^{-1}}{\partial \theta_i}\right\}\Bigg|_{\underline{\theta}=\underline{\widehat{\theta}}} = 0 \tag{9}$$

for all i, where θ_i is the ith element of $\underline{\theta}$. When R_θ is invertible and differentiable in $\underline{\theta}$

$$\frac{\partial R_\theta^{-1}}{\partial \theta_i} = -R_\theta^{-1}\frac{\partial R_\theta}{\partial \theta_i}R_\theta^{-1}. \tag{10}$$

Substituting this into the RCE formula gives an alternate set of necessary conditions for the optimal RCE solution

$$\mathrm{RCE}: \mathrm{tr}\left\{\left(R_\theta^{-1}RR_\theta^{-1} - R_\theta^{-1}\right)\frac{\partial R_\theta}{\partial \theta_i}\right\}\Bigg|_{\underline{\theta}=\underline{\widehat{\theta}}} = 0. \tag{11}$$

III. Application to Array Beamforming

In this section, we will apply the CE and RCE methods to fitting a low rank plus noise covariance matrix to data. Such problems arise in a variety of contexts, including narrowband sensor array processing, harmonic retrieval, and factor analysis. We focus on the array problem. Let $\underline{x}[n] = (x_1[n], \ldots, x_N[n])^T$ be a vector of sensor measurements at time n, where N is the total number of sensors in the array. Assume that the signal is narrowband (perhaps because the sensor data has been preprocessed through a fast Fourier transform of each sensor's data). Let our initial pdf estimate for the data be given by $p(\underline{x}[n]) = N(\underline{0}, R)$, where R is any non-parameterized estimate of the signal covariance, such as $R = (1/K)\sum_{k=1}^K \underline{x}[k]\underline{x}^H[k]$ where K snapshots of array data are used.

Now suppose we wish to model the data $\underline{x}[n]$ as

$$\underline{x}[n] = \sum_{i=1}^P s_i[n]\underline{u}_i + \sigma\underline{w}[n] \tag{12}$$

where $s_1[n], \ldots, s_P[n]$ are P source signals, $P < N$, arriving from unknown directions $\underline{u}_1, \ldots, \underline{u}_P$, with additive noise $\underline{w}[n]$ with gain σ. Assume the number P is known. Suppose that signals $s_i[n]$ are statistically independent, real or complex zero mean Gaussian random variables with covariance $\Lambda_i > 0$, and that the noise samples $\underline{w}[n]$ are statistically independent, real or complex zero mean Gaussian random variables with covariance W

$$p(s_i[n]) = N(0, \Lambda_i) \tag{13}$$

$$p(\underline{w}[n]) = N(\underline{0}, W). \tag{14}$$

Thus, the parameterized model pdf of $\underline{x}[n]$ is Gaussian

$$q_\theta(\underline{x}[n]) = N(\underline{0}, R_\theta) \tag{15}$$

where

$$R_\theta = \sum_{i=1}^P \Lambda_i \underline{u}_i\underline{u}_i^H + \sigma^2 W. \tag{16}$$

We will assume that the noise covariance W is known, but that all the other parameters $\underline{\theta} = (\Lambda_1, \ldots, \Lambda_P, \underline{u}_1, \ldots \underline{u}_P, \sigma)^T$ must be estimated. For convenience, define

$$R_\theta = U\Lambda U^H + \sigma^2 W \tag{17}$$

where

$$U = [\underline{u}_1 \quad \underline{u}_2 \quad \cdots \quad \underline{u}_P]$$

and

$$\Lambda = \begin{bmatrix}\Lambda_1 & & 0 \\ & \ddots & \\ 0 & & \Lambda_P\end{bmatrix}. \tag{18}$$

Suppose there are no *a priori* constraints on the matrix U, and that the only constraints on Λ are that $\Lambda_i > 0$. This would typically be true if the array were uncalibrated, or subject to heavy unknown multipath distortion. (Note that because we assume an uncalibrated array, we will not be able to directly derive information about the direction of arrival.) Appendices A and B apply the CE and RCE criteria to this model. They show that the solution to these two problems are quite similar, and can be found by the following algorithm.

CE and RCE BEAMFORMING ALGORITHMS

1) Find the generalized eigenvector $\underline{u}_i$ and eigenvalue λ_i solutions to

$$\lambda_i R^{-1}\underline{u}_i = W^{-1}\underline{u}_i \tag{19}$$

with normalization constraint $\underline{u}_i^H W^{-1}\underline{u}_j = \delta_{i,j}$.

2) Sort the eigenvectors and eigenvalues so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$. Then the optimal structured covariance matrix approximation $\hat{R}_\theta$ to R is

$$\hat{R}_\theta = (\underline{u}_1 \quad \underline{u}_2 \quad \cdots \quad \underline{u}_P) \times \begin{pmatrix} \lambda_1 - \hat{\sigma}^2 & & 0 \\ & \ddots & \\ 0 & & \lambda_P - \hat{\sigma}^2 \end{pmatrix} \times \begin{pmatrix} \underline{u}_1^H \\ \underline{u}_2^H \\ \cdot \\ \cdot \\ \cdot \\ \underline{u}_P^H \end{pmatrix} + \hat{\sigma}^2 W \tag{20}$$

where

$$\begin{cases} \frac{1}{\hat{\sigma}^2} = \frac{1}{N-P}\sum_{i=P+1}^N \frac{1}{\lambda_i} \text{ for CE} \\ \hat{\sigma}^2 = \frac{1}{N-P}\sum_{i=P+1}^N \lambda_i \text{ for RCE} \end{cases} \tag{21}$$

3) The CE for the optimal model is

$$\text{CE}: H(\hat{q}_\theta, p) = \xi \sum_{i=P+1}^{N} \log\left(\frac{\lambda_i}{\hat{\sigma}^2}\right) \tag{22}$$

$$\text{RCE}: H(p, \hat{q}_\theta) = \xi \sum_{i=P+1}^{N} \log\left(\frac{\hat{\sigma}^2}{\lambda_i}\right). \tag{23}$$

The estimates of $\hat{R}_\theta$ and $\hat{\sigma}^2$ will be unique if and only if $\lambda_P > \lambda_{P+1}$ (The estimate of U will not be unique.)

An interesting alternative form for the CE formulas can be found by substituting the value of $\hat{\sigma}^2$ from (21) into (22)

$$\text{CE}: H(\hat{q}_\theta, p) = \xi(N-P)\log\left(\frac{\left[\frac{1}{\lambda_{P+1}}, \ldots, \frac{1}{\lambda_N}\right]_{\text{avg}}}{\left[\frac{1}{\lambda_{P+1}}, \ldots, \frac{1}{\lambda_N}\right]_{geo}}\right) \tag{24}$$

$$\text{RCE}: H(p, \hat{q}_\theta) = \xi(N-P)\log\left(\frac{[\lambda_{P+1}, \ldots, \lambda_N]_{\text{avg}}}{[\lambda_{P+1}, \ldots, \lambda_N]_{geo}}\right) \tag{25}$$

where

$$[\beta_{P+1}, \ldots, \beta_N]_{\text{avg}} = \frac{1}{N-P}\sum_{i=P+1}^{N} \beta_i \tag{26}$$

$$[\beta_{P+1}, \ldots, \beta_N]_{\text{geo}} = (\beta_{P+1}\beta_{P+2}\cdots\beta_N)^{1/(N-P)}. \tag{27}$$

The CEs are proportional to the log of the ratio of the arithmetic mean to the geometric mean of the eigenvalues (or their inverses) that are not used in building U. The CEs will, therefore, be positive, and will attain their minimum value of zero only if the geometric average of $\lambda_{P+1}, \ldots, \lambda_N$ (or their inverses) equals their arithmetic mean. This will only occur if these $N-P$ smallest generalized eigenvalues are all equal. Note that (21), (24), and (25) have either CE or RCE meaning and are useful in identifying the number P.

Note the similarity of the RCE formula to the MDL order determination algorithm suggested by Wax and Kailath [2]. The RCE criterion is also strongly related to the Maximum Likelihood problem of estimating the structured covariance matrix given independent observations $\underline{x}_1, \ldots, \underline{x}_K$

$$\hat{R}_\theta \leftarrow \max_\theta \log p(\underline{x}_1, \ldots, \underline{x}_K|\underline{\theta}) \tag{28}$$

where

$$p(\underline{x}_1, \ldots, \underline{x}_K|\underline{\theta}) = \prod_{i=1}^{K} p(\underline{x}_i|\underline{\theta}) \tag{29}$$

and

$$p(\underline{x}_i|\underline{\theta}) = N(\underline{0}, R). \tag{30}$$

This is because

$$H(p, q_\theta) = \frac{1}{K}\log p(\underline{x}_1, \ldots, \underline{x}_K|\underline{\theta}) - \xi(N + \log|R|) \tag{31}$$

Since the second term in (31) does not depend on $\underline{\theta}$, the RCE estimate of R_θ will be identical to the ML estimate.

For the special case when the background noise is white Gaussian noise $W = I$, the $\underline{u}_i$ must satisfy

$$R\underline{u}_i = \lambda_i \underline{u}_i \tag{32}$$

and, thus, the $\underline{u}_i$ are the eigenvectors of the observed data correlation matrix R. This special case is thus quite similar to that used in the MUSIC algorithm [1] and other similar beamforming algorithms.

If subroutines for computing generalized eigenvectors are not available, we can use subroutines for computing eigenvectors of symmetric positive definite matrices as follows. Factor $W = W^{1/2}W^{H/2}$ where $W^{1/2}$ is any square root of W and $W^{H/2}$ is its Hermitian. Then to compute the $\underline{u}_i$

1) From the whitened data correlation matrix

$$\tilde{R} = W^{-1/2}RW^{-H/2} \tag{33}$$

where $W^{-1/2}$ is the inverse of $W^{1/2}$. Note that $\tilde{R}$ is symmetric and positive definite.

2) Solve for the eigenvectors $\underline{t}_i$ and corresponding eigenvalues λ_i of $\tilde{R}$.

$$\tilde{R}\underline{t}_i = \lambda_i \underline{t}_i \tag{34}$$

where $\underline{t}_j^T \underline{t}_i = \delta_{i.j}$. Sort these so that the eigenvalues are in descending order.

3) Then

$$\underline{u}_i = W^{1/2}\underline{t}_i \tag{35}$$

It is also interesting to consider the effect of using the structured covariance matrix estimate when forming either a classical or optimal beamformer. Let $\underline{w}_0$ be the ideal array response for a signal in a particular direction. The classical beamformer estimates the signal $s[n]$ from the array data as $s[n] = \underline{w}_0^T \underline{x}[n]$. The expected received power from this direction is then $E[s^2[n]] = \underline{w}_0^T R_\theta \underline{w}_0$. Now suppose that $\underline{w}_o$ is in the space spanned by the columns of $R^{-1}U$, i.e., $\underline{w}_0 = R^{-1}U\underline{\alpha}$ for some vector $\underline{\alpha}$. It is shown in Appendix A that $R_\theta^{-1}U = R^{-1}U$. Therefore

$$\begin{aligned} \underline{w}_0^H R_\theta \underline{w}_0 &= \underline{\alpha}^H U^H R^{-H} R_\theta R^{-1} U \underline{\alpha} \\ &= \underline{\alpha}^H U^H R^{-H} U \underline{\alpha} \\ &= \underline{w}_0^H R \underline{w}_0 \end{aligned} \tag{36}$$

In this case, replacing R with the structured covariance estimate R_θ in the classical beamformer makes no difference. However, if $\underline{w}_0$ is not in the subspace spanned by $R^{-1}\underline{u}_1, \ldots, R^{-1}\underline{u}_P$, then $R_\theta^{-1}\underline{w}_0 \neq$

$R^{-1}\underline{w}_0$, and using the structured covariance estimate in the classical beamformer will yield a different beam pattern.

A similar statement holds for the optimum minimum variance beamformer, $s[n] = \underline{w}^T \underline{x}[n]$, which uses a window $\underline{w}$ designed such that the expected response energy $\underline{w}^T R_\theta \underline{w}$ is minimized subject to the constraint that the response to a plane wave from the direction of interest is unity, $\underline{w}^T \underline{w}_0 = 1$. The solution is $\underline{w} = (\underline{w}_0^T R_\theta^{-1} \underline{w}_0)^{-1} R_\theta^{-1} \underline{w}_0$. Note that if $\underline{w}_0$ is in the subspace spanned by the columns of U, then there exists some vector $\underline{\alpha}$ such that $\underline{w}_0 = U\underline{\alpha}$. Since $R_\theta^{-1}U = R^{-1}U$,

$$\begin{aligned} R_\theta^{-1}\underline{w}_0 &= R_\theta^{-1}U\underline{\alpha} \\ &= R^{-1}U\underline{\alpha} \\ &= R^{-1}\underline{w}_0 \end{aligned} \tag{37}$$

which in turn implies

$$\begin{aligned} \underline{w} &= \left(\underline{w}_0^T R_\theta^{-1} \underline{w}_0\right)^{-1} R_\theta^{-1}\underline{w}_0 \\ &= \left(\underline{w}_0^T R^{-1} \underline{w}_0\right)^{-1} R^{-1}\underline{w}_0. \end{aligned} \tag{38}$$

In this case, replacing R with the structured covariance estimate R_θ in the optimal beamformer makes no difference. However, if $\underline{w}_0$ is not in the subspace spanned by the columns of U, then $R_\theta^{-1}\underline{w}_0 \neq R^{-1}\underline{w}_0$, and using the structured covariance estimate in the optimal beamformer will yield a different beam pattern. These results are contrary to the suggestion implied in [5] that replacing R with R_θ in an optimal beamformer should make no difference.

IV. Conclusion

In this paper, we have derived the optimal solution for correlation matrix estimation by the CE and RCE principles. The two methods give identical directions $\underline{u}_i$ of each source, differing only in the value of the noise level estimate. Strikingly, the closed-form solution in the ALGORITHMS is similar to the oblique rotation in factor analysis in psychometrics where various intelligence traits are obtained from such analysis. To our knowledge, the closed-form solution cannot be reached or derived from any other existing approaches. The RCE method gives the same results as the maximum likelihood approach, and when the noise is white, both methods are similar to MUSIC. It is interesting that the CE approach thus provides a framework for deriving spectral estimation algorithm including Bartlett, MLM [8], MEM [10], and now MUSIC. The results of the approach can be connected, via the Eckart-Young-Mirsky theorem [12], to the approximation of covariance matrices in norms 2 and Frobenius.

Appendix A
Derivation of CE Beamforming Algorithm

In this Appendix we derive the optimal structured covariance estimate using the CE principle. First, to simplify the effort, let us define $V = U\Lambda^{1/2}$, where $\Lambda^{1/2} = \mathrm{diag}(\Lambda_1^{1/2}, \ldots, \Lambda_N^{1/2})$. Then

$$R_\theta = VV^H + \sigma^2 W \tag{39}$$

Substitute this into the CE entropy expression (6), and set the derivatives with respect to the real and imaginary part of every element of the V matrix, and with respect to σ^2, to zero. Arranging these derivatives in complex matrix fom gives

$$\left(R^{-1} - R_\theta^{-1}\right)V = 0 \tag{40}$$

$$\mathrm{tr}\left\{\left(R^{-1} - R_\theta^{-1}\right)W\right\} = 0. \tag{41}$$

Using the Woodward lemma

$$\begin{aligned} R_\theta^{-1} = \frac{1}{\sigma^2}W^{-1} &- \frac{1}{\sigma^2}W^{-1}V \\ &\times \left[V^H \frac{1}{\sigma^2}W^{-1}V + I\right]^{-1} V^H \frac{1}{\sigma^2}W^{-1}. \end{aligned} \tag{42}$$

Substituting into (40) and simplifying gives

$$R^{-1}V = \frac{1}{\sigma^2}W^{-1}V\left[V^H \frac{1}{\sigma^2}W^{-1}V + I\right]^{-1}. \tag{43}$$

This equation has many possible solutions. Let V refer to any one of these. Then let $\Psi = V^H W^{-1} V$. Diagonalize Ψ by factoring it $\Psi = Q\Phi Q^H$, where Φ is diagonal and Q is orthonormal, $Q^H Q = I$. Define $\tilde{V} = VQ$. Note that $\tilde{V}$ is also a solution to (43). In fact

$$R^{-1}\tilde{V} = \frac{1}{\sigma^2}W^{-1}\tilde{V}\left[\frac{1}{\sigma^2}\Phi + I\right]^{-1} \tag{44}$$

and

$$\tilde{V}^H W^{-1} \tilde{V} = \Phi. \tag{45}$$

Let the P columns of $\tilde{V}$ be $\underline{\tilde{v}}_1, \ldots, \underline{\tilde{v}}_P$, and let the P diagonal elements of Φ be $\phi_1, \ldots, \phi_P$. Then

$$\lambda_i R^{-1}\underline{\tilde{v}}_i = W^{-1}\underline{\tilde{v}}_i \tag{46}$$

where

$$\lambda_i = \phi_i + \sigma^2. \tag{47}$$

The columns of $\tilde{V}$ must therefore either be zero, or else must be generalized eigenvector solutions to (46). Because R and W are conjugate symmetric and positive definite, there are always N linearly independent generalized eigenvector solutions $\underline{\tilde{v}}_1, \ldots, \underline{\tilde{v}}_N$ to (46), with corresponding generalized eigenvalues $\lambda_1, \ldots, \lambda_N$ which are positive. Assume without loss of generality that the first P_0 columns of $\tilde{V}$ are nonzero, where $P_0 \le P$. These first P_0 columns must be selected from among the N possible generalized eigenvectors, in a manner we will determine later. Also note that it is not necessary to estimate Q or V directly, since we can construct R_θ directly from $\tilde{V}$

$$\begin{aligned} R_\theta &= VV^H + \sigma^2 W \\ &= VQQ^H V^H + \sigma^2 W \\ &= \tilde{V}\tilde{V}^H + \sigma^2 W. \end{aligned} \tag{48}$$

Now to solve for σ^2. Substitute (42) into (41), and simplify by exploiting the facts that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ and $\mathrm{tr}(C+D) = \mathrm{tr}(C) + \mathrm{tr}(D)$ and $\mathrm{tr}(\alpha C) = \alpha\mathrm{tr}(C)$ where A, B are matrices, C, D are square matrices, and α is a scalar.

$$\begin{aligned} 0 &= \mathrm{tr}\left\{\left(R_\theta^{-1} - R^{-1}\right)W\right\} \\ &= \mathrm{tr}\left\{\left(\frac{1}{\sigma^2}W^{-1} - \frac{1}{\sigma^2}W^{-1}\tilde{V}\left[\tilde{V}^H\frac{1}{\sigma^2}W^{-1}\tilde{V} + I\right]^{-1}\right.\right. \\ &\qquad \left.\left.\times\ \tilde{V}^H\frac{1}{\sigma^2}W^{-1} - R^{-1}\right)W\right\} \\ &= \mathrm{tr}\left\{\frac{1}{\sigma^2}I\right\} - \frac{1}{\sigma^2}\mathrm{tr}\left\{\left[\tilde{V}^H\frac{1}{\sigma^2}W^{-1}\tilde{V} + I\right]^{-1}\right. \\ &\qquad \left.\times\left[\tilde{V}^H\frac{1}{\sigma^2}W^{-1}\tilde{V}\right]\right\} - \mathrm{tr}\{R^{-1}W\} \\ &= \frac{N}{\sigma^2} - \frac{1}{\sigma^2}\sum_{i=1}^{P}\frac{\phi_i}{\phi_i + \sigma^2} - \mathrm{tr}\{R^{-1}W\} \\ &= \frac{N - P_0}{\sigma^2} + \sum_{i=1}^{P_0}\frac{1}{\lambda_i} - \mathrm{tr}\{WR^{-1}\} \end{aligned} \tag{49}$$

where we used (45) in the fourth line, and (47) in the fifth. This can be further simplified by noticing that if $\underline{\tilde{v}}_i$ is any generalized eigenvector solution to (46), then

$$WR^{-1}\underline{\tilde{v}}_i = W\left(\frac{1}{\lambda_i}W^{-1}\underline{\tilde{v}}_i\right) = \frac{1}{\lambda_i}\underline{\tilde{v}}_i. \quad (50)$$

Therefore, the $\underline{\tilde{v}}_i$ are eigenvectors of WR^{-1} with eigenvalues $1/\lambda_i$. Thus

$$\mathrm{tr}\{WR^{-1}\} = \sum_{i=1}^{N}\frac{1}{\lambda_i}. \quad (51)$$

Substituting back into (49), then solving for σ^2 gives

$$\sigma^2 = \frac{N-P_0}{\sum_{i=P_0+1}^{N}\frac{1}{\lambda_i}}. \quad (52)$$

Now substitute the solution for $\tilde{V}$ and for σ^2 into (48), and then substitute this back into the formula (6) for the CE. The algebra is simplified by noting that if $\underline{\tilde{v}}_i$ is any generalized eigenvector solution to (46), then

$$\begin{aligned} R_\theta R^{-1}\underline{\tilde{v}}_i &= (\tilde{V}\tilde{V}^H + \sigma^2 W)\left(\frac{1}{\lambda_i}W^{-1}\underline{\tilde{v}}_i\right) \\ &= \frac{1}{\lambda_i}(\tilde{V}\tilde{V}^H W^{-1}\underline{\tilde{v}}_i + \sigma^2\underline{\tilde{v}}_i) \\ &= \begin{cases} \frac{1}{\lambda_i}(\phi_i+\sigma^2)\underline{\tilde{v}}_i & \text{for } i=1,\ldots,P_0 \\ \frac{1}{\lambda_i}\sigma^2\underline{\tilde{v}}_i & \text{for } i=P_0+1,\ldots,N \end{cases} \\ &= \begin{cases} \underline{\tilde{v}}_i & \text{for } i=1,\ldots,P_0 \\ \frac{\sigma^2}{\lambda_i}\underline{\tilde{v}}_i & \text{for } i=P_0+1,\ldots,N. \end{cases} \end{aligned} \quad (53)$$

Therefore, the $\underline{\tilde{v}}_i$ are all eigenvectors of $R_\theta R^{-1}$. The first P_0 eigenvalues are equal to 1, and the remainder are equal to $\sigma^2/\lambda_{P_0+1},\ldots,\sigma^2/\lambda_N$. Putting all this together, the CE at this solution has the value

$$\begin{aligned} H(q_\theta,p) &= \xi\{\mathrm{tr}\{R_\theta R^{-1}\} - N - \log|R_\theta R^{-1}|\} \\ &= \xi\left\{P_0 + \sigma^2\sum_{i=P_0+1}^{N}\frac{1}{\lambda_i} - N - \log\prod_{i=P_0+1}^{N}\frac{\sigma^2}{\lambda_i}\right\} \\ &= \xi\sum_{i=P_0+1}^{N}\log\left(\frac{\lambda_i}{\sigma^2}\right). \end{aligned} \quad (54)$$

Substituting the value of σ^2 from (52) gives the alternate form

$$H(q_\theta,p) = \xi(N-P_0)\log\left[\frac{\frac{1}{N-P_0}\sum_{i=P_0+1}^{N}\frac{1}{\lambda_i}}{\left(\prod_{i=P_0+1}^{N}\frac{1}{\lambda_i}\right)^{1/(N-P_0)}}\right]. \quad (55)$$

Now to return to the issue of which of the N possible generalized eigenvector solutions should be used for the P_0 nonzero columns of $\tilde{V}$. Let us call the selected P_0 eigenvectors $\underline{\tilde{v}}_1,\ldots,\underline{\tilde{v}}_{P_0}$ the "signal eigenvectors," and let us call the remainder the "noise eigenvectors." The signal eigenvectors satisfy $\underline{\tilde{v}}_i \neq 0$; since $W^{-1} > 0$, then $\phi_i = \underline{\tilde{v}}_i^H W^{-1}\underline{\tilde{v}}_i > 0$ and thus $\lambda_i = \phi_i + \sigma^2 > \sigma^2$ for $i = 1,\ldots,P_0$. We show that these signal eigenvalues must be the largest eigenvalue solutions to (46). Suppose this were not true, so that the global optimum solution corresponded to an R_θ such that one of the signal eigenvalues, say λ_{P_0}, was smaller than the largest of the noise eigenvalues, say λ_{P_0+1}. Thus, $\sigma^2 < \lambda_{P_0} < \lambda_{P_0+1}$. But then, as we will see, swapping these eigensolutions, making $\underline{\tilde{v}}_{P_0+1}$ a signal eigenvector and $\underline{\tilde{v}}_{P_0}$ a noise eigenvector will further decrease the CE, contradicting our assumption of global optimality. To show this, let $H(\lambda_{P_0+1},\lambda_{P_0+2},\ldots,\lambda_N)$ represent the CE with a model R_θ built using nonzero solutions $\underline{\tilde{v}}_1,\ldots,\underline{\tilde{v}}_{P_0-1},\underline{\tilde{v}}_{P_0}$, and let $H(\lambda_{P_0},\lambda_{P_0+2},\ldots,\lambda_N)$ represent the CE with a model R_θ built using nonzero solutions $\underline{\tilde{v}}_1,\ldots,\underline{\tilde{v}}_{P_0-1},\underline{\tilde{v}}_{P_0+1}$. Then, because the CE formula (55) is an analytic function of the λ_i, by the mean value theorem

$$\begin{aligned} &H(\lambda_{P_0+1},\lambda_{P_0+2},\ldots,\lambda_N) - H(\lambda_{P_0},\lambda_{P_0+2},\ldots,\lambda_N) \\ &\quad = \frac{\partial H}{\partial\lambda}(\lambda,\lambda_{P_0+2},\ldots,\lambda_N)|_{\lambda=\bar{\lambda}}(\lambda_{P_0+1}-\lambda_{P_0}) \end{aligned} \quad (56)$$

where $\bar{\lambda}$ is some value in the range $\lambda_{P_0} < \bar{\lambda} < \lambda_{P_0+1}$. But

$$\begin{aligned} \frac{\partial H}{\partial\lambda} &= \xi\frac{1}{\lambda^2}\left(\lambda - \frac{N-P_0}{\frac{1}{\lambda}+\sum_{i=P_0+2}^{N}\frac{1}{\lambda_i}}\right) \\ &> 0 \end{aligned} \quad (57)$$

for all $\lambda_{P_0} < \lambda < \lambda_{P_0+1}$, where the last line is true because

$$\begin{aligned} \lambda &> \lambda_{P_0} \\ &> \sigma^2 \\ &= \frac{N-P_0}{\sum_{i=P_0+1}^{N}\frac{1}{\lambda_i}} \\ &> \frac{N-P_0}{\frac{1}{\lambda}+\sum_{i=P_0+2}^{N}\frac{1}{\lambda_i}}. \end{aligned} \quad (58)$$

Since $\lambda_{P_0+1} - \lambda_{P_0} > 0$, the change in (56) must be positive. Therefore, swapping $\underline{\tilde{v}}_{P_0}$ and $\underline{\tilde{v}}_{P_0+1}$ reduces the CE, and our assumed global optimum solution cannot be globally optimum. The P_0 signal eigenvalues must therefore be the largest eigenvalue solutions to (46), and the nonzero P_0 columns of $\tilde{V}$ must be the corresponding general eigenvectors.

Finally, we must show that we should always choose $P_0 = P$ eigenvectors. Without loss of generality, let us sort all the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$. Let H_i represent the minimum CE with i nonzero columns in $\tilde{V}$. Then using (55)

$$\begin{aligned} H_{P_0} - H_{P_0+1} &= \xi(N-P_0) \\ &\quad\times\log\left[\frac{\frac{1}{(N-P_0)}\frac{1}{\lambda_{P_0+1}} + \left(1-\frac{1}{(N-P_0)}\right)(1/\bar{\lambda})}{\left(\frac{1}{\lambda_{P_0+1}}\right)^{\frac{1}{(N-P_0)}}(1/\bar{\lambda})^{1-\frac{1}{(N-P_0)}}}\right] \\ &\geq 0 \end{aligned} \quad (59)$$

where $(1/\bar{\lambda}) = [(1)/(N-P_0-1)]\sum_{i=P_0+2}^{N}(1)/(\lambda_i)$ and where we used the inequality $\rho\alpha + (1-\rho)\beta \geq \alpha^\rho\beta^{(1-\rho)}$ for any $0 \leq \rho \leq 1$ in the last line. Thus, the CE decreases as P_0 varies from 0 to P, so the best choice for P_0 must be $P_0 = P$.

The proof that R_θ is unique when $\lambda_P > \lambda_{P+1}$ is messy but straightforward. The key issue is that the space spanned by the signal eigenvectors is uniquely determined. If there are multiple signal eigenvalues, then the eigenvectors themselves may not be uniquely determined, and, thus, $\tilde{V}$ may not be uniquely determined. We get the formulas in the text by defining $U = \tilde{V}\Phi^{-1/2}$.

APPENDIX B
DERIVATION OF RCE ALGORITHM

In this Appendix we give the solution to the RCE problem. The derivation is quite similar to that for the CE problem, and, therefore, we present this quickly. With our Gaussian models, the RCE cross-entropy has the value

$$\mathrm{RCE}: H(p,q_\theta) = \xi\left\{\mathrm{tr}\left(R_\theta^{-1}R\right) - N - \log\left|R_\theta^{-1}R\right|\right\}. \quad (60)$$

Differentiating with respect to the real and imaginary parts of V and setting these to zero, as before, gives

$$\left(R_\theta^{-1}RR_\theta^{-1} - R_\theta^{-1}\right)V = 0. \tag{61}$$

Multiplying both sides by $R^{-1}R_\theta$ gives

$$\left(R_\theta^{-1} - R^{-1}\right)V = 0 \tag{62}$$

which is exactly the same equations which the solution for V in the CE problem must satisfy, (40). Therefore, we can construct R_θ from (48), where the columns of $\tilde{V}$ must be solutions to the generalized eigenvector problem (46).

Now differentiating (60) with respect to σ^2 and setting it to zero gives

$$\mathrm{tr}\left\{(R_\theta^{-1}RR_\theta^{-1} - R_\theta^{-1})W\right\} = 0. \tag{63}$$

Combining this with (61) gives

$$\begin{aligned} 0 &= \mathrm{tr}\left\{\left(R_\theta^{-1}RR_\theta^{-1} - R_\theta^{-1}\right)\left(\tilde{V}\tilde{V}^H + \sigma^2 W\right)\right\} \\ &= \mathrm{tr}\left\{\left(R_\theta^{-1}RR_\theta^{-1} - R_\theta^{-1}\right)R_\theta\right\} \\ &= \mathrm{tr}\left\{R_\theta^{-1}R - I\right\} \end{aligned} \tag{64}$$

which implies that

$$\mathrm{tr}\left\{R_\theta^{-1}R\right\} = N. \tag{65}$$

But (53) implies that $R_\theta^{-1}R$ has P_0 eigenvalues equal to 1, and the rest have values $\lambda_{P_0+1}/\sigma^2, \ldots, \lambda_N/\sigma^2$. Since the trace of a matrix is just the sum of its eigenvalues

$$P_0 + \frac{1}{\sigma^2}\sum_{i=P_0+1}^{N}\lambda_i = N \tag{66}$$

which gives

$$\sigma^2 = \frac{1}{(N-P_0)}\sum_{i=P_0+1}^{N}\lambda_i. \tag{67}$$

Using the facts that the trace of a matrix is the sum of the eigenvalues, and the determinant is the product of the eigenvalues

$$\begin{aligned} H(p, q_\theta) &= \xi\left\{\mathrm{tr}\{R_\theta^{-1}R\} - N - \log\left|R_\theta^{-1}R\right|\right\} \\ &= \xi\sum_{i=P_0+1}^{N}\log\frac{\sigma^2}{\lambda_i}. \end{aligned} \tag{68}$$

The proofs that we must choose $\lambda_1, \ldots, \lambda_{P_0}$ to be the largest eigenvalues, that we should choose $P_0 = P$, and that the solution R_θ is unique if $\lambda_P > \lambda_{P+1}$, are similar to the proofs for the CE algorithm.

REFERENCES

[1] R. Schmidt, "Multiple emitter location and signal parameter estimation," in *Proc. RADC Spectral Estimation Workshop*, Rome, NY, 1978, pp. 243–258.

[2] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech Signal Process.*, vol. ASSP-33, pp. 387–392, 1985.

[3] Q. A. Nguyen, "On the uniqueness of the maximum likelihood estimate of structured covariance matrices," *IEEE Trans. Acoust., Speech Signal Process.*, vol. ASSP-32, pp. 1249–1251, 1984.

[4] J. Burg, D. Luenberger, and D. Wenger, "Estimation of structured covariance matrices," *Proc. IEEE*, vol. 70, pp. 963–974, 1982.

[5] D. Gray, B. Anderson, and P. Sim, "Estimation of structured covariances with application to array beamforming," *Circuits, Systems, and Signal Proc.*, vol. 6–4, pp. 421–447, 1987.

[6] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.

[7] B. Musicus, "Iterative algorithms for optimal signal reconstruction and parameter identification given noisy and incomplete data," Ph.D. dissertation, Dept. Elec. Eng. Comp. Sci., Mass. Inst. Technol., Cambridge, 1982.

[8] C.-Y. Liou and B. Musicus, "A separable cross-entropy approach to power spectral estimation," *IEEE Trans. Acoust., Speech Signal Process.*, vol. ASSP-38, no. 1, pp. 105–113, 1990.

[9] C.-Y. Liou and S.-L. Lin, "The other variant Boltzmann machine," in *Int. Joint Conf. Neural Netw., IJCNN*, Wash., DC, Jun. 18–22, 1989, pp. 449–454.

[10] J. Shore, "Minimum cross-entropy spectral analysis," *IEEE Trans. Acoust., Speech Signal Process.*, vol. ASSP-29, pp. 230–236, 1981.

[11] R. D. Palkki, A. D. Lanterman, and W. D. Blair, "Addressing track coalescence in sequential K-best multiple hypothesis tracking," in *Proc. 38th Southeastern Symp. System Theory*, Cookeville, TN, Mar. 5–7, 2006, Tennessee Technolog. Univ..

[12] S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem*. Philadelphia, PA: SIAM, 1991.

Comments on "Statistical Analysis of the Nonhomogeneity Detector for Non-Gaussian Interference Backgrounds"

Saralees Nadarajah

***Abstract*—Rangaswamy ["Statistical Analysis of the Nonhomogeneity Detector for Non-Gaussian Interference Backgrounds," *IEEE Trans. Signal Process.*, vol. 53, no. 6, pp. 2101–2111, Jun. 2005] presented a statistical analysis with respect to the nonhomogeneity detector (NHD) for non–Gaussian interference scenarios. However, except for one case, Rangaswamy did not provide explicit expressions for the statistical quantities of interest. In this Comment, exact and elementary expressions are derived for all of the quantities considered by the statistical analysis.**

***Index Terms*—Beta function, moments, nonhomogeneity detector, probability density function, probability of type I error.**

I. INTRODUCTION

Rangaswamy [1] derived the nonhomogeneity detector (NHD) for non–Gaussian interference scenarios and presented a statistical analysis of the method. The statistical analysis focused on the following quantities:

1) the probability density function (pdf) of the NHD test statistic Λ_{eq} given by

$$f_{\Lambda_{\mathrm{eq}}}(r) = \frac{L}{B(L+1, M-1)}\int_0^1 \frac{\gamma^L(1-\gamma)^{M-1}}{\{1+(1-\gamma)r\}^{L+1}}d\gamma \tag{1}$$

[see (10)], where $B(a,b)$ denotes the beta function defined by

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt;$$

Manuscript received December 17, 2006; revised December 17, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Steven M. Kay.

The author is with the University of Manchester, Manchester, M60 1QD, U.K. (e-mail: saralees.nadarajah@manchester.ac.uk).

Digital Object Identifier 10.1109/TSP.2008.917361