2017

Some Contributions to Radar Detection Theory

Zhenghan Zhu
University of Rhode Island, zhu.zhenghan1987@gmail.com

SOME CONTRIBUTIONS TO RADAR DETECTION THEORY

BY

ZHENGHAN ZHU

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

ELECTRICAL ENGINEERING

UNIVERSITY OF RHODE ISLAND

2017

DOCTOR OF PHILOSOPHY DISSERTATION

OF

ZHENGHAN ZHU

APPROVED:

Dissertation Committee:

Major Professor Steven Kay

Ramdas Kumaresan

Mustafa Kulenovic

Nasser H. Zawia

DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND

2017

ABSTRACT

This dissertation focuses on statistical signal processing theory and its applications to radar, complex-valued signal processing and model selection.

The transmit signal critically affects a radar system's performance. Its design is an important task and is an active research area. We provide an optimal design for detecting extended targets in colored noise based on the locally most powerful detector. We also establish a fundamental relationship between the Kullback-Leibler divergence, signal-to-noise ratio, and mutual information, all of which have been used as waveform design metrics the literature. The relationship explains the role of each metric.

In space-time adaptive processing (STAP), the nonstationarity of the data samples causes a mismatch between the estimated covariance matrix and the true one, and consequently leads to the degradation of STAP performance. We propose an asymptotically optimal detector for testing the non-stationarity via the generalized likelihood ratio test and an alternative Rao test with lower computational cost

The Rao test is a very useful method in signal processing. A complex parameter Rao test is proposed and serves as a new method for complex-valued parameter testing. Different from the traditional way, it reformulates the calculations with respect to the complex-valued quantities directly and often leads to more intuitive, and more computationally efficient test statistics. Applying the complex parameter Rao test to the bandedness of the Cholesky factor of the inverse of a

complex-valued covariance matrix is an example of its application.

Model order selection is another fundamental but important task that arises in many areas. We propose a new Bayesian model order selection method by employing the exponentially embedded family (EEF) technique. In addition to the established important properties of EEF, the new Bayesian model selection method can use vague proper priors and improper non-informative priors without the criticisms of Lindley's paradox and the Information paradox. The penalty term of the Bayesian EEF is shown to have a very intuitive meaning as the sum of the model parameter dimension and the estimated mutual information between the parameter and observed data. The EEF is also used to estimated the degree of noncircularity of a complex random vector and is shown to have good performance.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor Prof. Steven Kay. I feel lucky and honored to be his student. He is the best Ph.D. mentor I could imagine to have. His invaluable advice through words and deeds on research and life will benefit me for a life-long time. He has set such a great role model of being an excellent researcher, respectful educator and great person. Without his inspiration and support, this dissertation would have been impossible.

I would also like to thank the rest of my dissertation committee members: Prof. Ramdas Kumaresan, Prof. Mustafa Kulenovic, Prof. Tom Bella for their time and effort in my comprehensive exams and dissertation defense.

Thanks are due to many faculty members, staff, graduate students, undergraduates and friends, too many to name individually, in the Department of Electrical, Computer and Biomedical Engineering at URI. They offered help and suggestions in several ways.

My family and friends have supported me along my journey towards this degree. The encouragement and support from Fanglin Wu, Meilan Zhu, Meiyun Gu and Yuru Xiao are especially appreciated.

This dissertation is dedicated to my father, Qingbao Zhu, whom I will miss forever. Words cannot express how grateful I am for all of the love he gave and sacrifices he made on my behalf.

DEDICATION

in memory of my father, Qingbao Zhu

PREFACE

This dissertation is organized in the manuscript format consisting of seven manuscripts as following:

- Manuscript 1:

  Z. Zhu, S. Kay and R.S. Raghavan, "Information-theoretic optimal radar waveform design," in part published in *IEEE Signal Processing Letters*, vol. 24, no. 3, pp.274–278, Mar. 2017.

- Manuscript 2:

  Z. Zhu, S. Kay, F. Cogun and R.S. Raghavan, "On detection of nonstationarity of the Covariance Matrix in radar signal processing", to be submitted to *IEEE Transactions on Aerospace and Electronic Systems*, in part published *in the proceedings of 2016 IEEE Radar Conference*, pp.1–4, Philadelphia, PA, May. 2016.

- Manuscript 3:

  S. Kay and Z. Zhu, "The complex parameter Rao test," *IEEE Transactions on Signal Processing*, vol. 64, no. 24, pp.6580–6588, Dec. 2016.

- Manuscript 4:

  Z. Zhu and S. Kay, "The Rao test for testing bandedness of complex-valued covariance matrix", *in the proceedings of the* 41^{st} *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.3960–3963, Shanghai, Mar. 2016.

- Manuscript 5:

  Z. Zhu and S. Kay, "On the Bayesian exponentially embedded family for model order selection," submitted to *IEEE Transactions on Signal Processing.*

- Manuscript 6:

  Z. Zhu and S. Kay, "The penalty term of the exponentially embedded family is estimated mutual information," *in the proceedings of* 42^{nd} *IEEE International Conference on Acoustics, Speech and Signal Processing.* pp. 4149–4152, New Orleans, LA, Mar. 2017.

- Manuscript 7:

  Z. Zhu and S. Kay, "Estimate the degree of noncircularity of complex-valued vectors via exponentially embedded family,"*in the proceedings of 2017 IEEE Radar Conference*, to appear.

TABLE OF CONTENTS

Page

Page

Page

LIST OF FIGURES

LIST OF TABLES

Table **Page**

MANUSCRIPT 1

Information-Theoretic Optimal Radar Waveform Design

by

Zhenghan Zhu, Steven Kay and R. S. Raghavan

in part published in

IEEE Signal Processing Letters, vol. 24, no. 3, pp.274–278, Mar. 2017.

Abstract

In this paper we address the problem of designing the optimal radar waveform for the detection of an extended target in a colored noise environment. Specifically, the impulse response of the target is assumed to be linear and time invariant, which is modeled by a wide sense stationary Gaussian random process. The noise is also assumed to be a wide sense stationary Gaussian random process with known power spectral density. We derive the locally most powerful detector and the corresponding optimal waveform based on maximizing the detector's performance under a small signal assumption. The performance is evaluated analytically, and numerically compared to that of the existing information-theoretic method, i.e., maximizing the mutual information between the received data and target response. The locally most powerful detection metric is shown to be the Kullback-Leibler divergence. Thus, use of the latter measure for waveform optimization shows that a substantial performance improvement is achieved by adopting the proposed waveform design approach instead of mutual information. Moreover, an interesting relationship among the three waveform design metrics, namely, the output signal-to-noise ratio, the Kullback-Leibler divergence, and the mutual information, is derived. This provides an important relationship that explains the tradeoffs of the various metrics currently used for radar waveform design.

1.1 Introduction

Transmit signal waveform design is one of critical factors affecting radar detection performance [3]. This problem has attracted much attention in the last few

decades (see [5]–[8] and references therein). We focus on the problem of optimal signal waveform design for detecting an extended target in a wide-sense stationary (WSS) noise environment. The target is modeled as a linear time invariant (LTI) system, whose impulse response is assumed to be a WSS random process.

The first question for signal waveform design is the choice of an optimization metric. Many metrics have been proposed such as maximizing the mutual information (MI), minimizing the mean square error (MMSE) and maximizing the output-signal-to-noise ratio (SNR) [5], [8], [15], [10]. Clearly, the choice of a metric should be done *to maximize detection performance*. In particular, the information-theoretic waveform design method of maximizing the MI [6], is a widely used one [5], [8]. It maximizes the MI between the target impulse response and the received data. However, the waveform design based on the MI criteria does not guarantee detection optimality, as Bell has remarked in [7]: " It is not clear under what circumstances these approaches lead to optimal or near-optimal results".

In this paper we hope to answer this question. To do so we derive an optimal radar signal waveform for the detection of an extended target in colored noise. Although the performance of the Neyman-Pearson (NP) detector is a natural choice to start with [2][17], determining its performance analytically is nontrivial due to the difficulty of obtaining the probabilities of detection and false alarm analytically [2]. On the other hand, the analytical performance of the locally most powerful (LMP) detector, which is equivalent to the NP detector under the small signal assumption, is easily derived. Furthermore, the performance of the LMP detector is

a function of the Kullback-Leibler divergence (KLD), denoted by $D(p_1(\mathbf{x})||p_0(\mathbf{x}))$, where $p_1(\mathbf{x})$ is the PDF of the received data $\mathbf{x}$ under the alternative (i.e. the signal present) hypothesis $\mathcal{H}_1$ and $p_0(\mathbf{x})$ is the PDF under the null hypothesis $\mathcal{H}_0$. For simplicity we will at times use $D(p_1||p_0)$ to denote the KLD. Hence, it would seem that the KLD measure in the small signal case would be more appropriate for radar waveform design. Our results demonstrate by computer simulation that this is indeed true.

The main contributions of this paper are:

1) The LMP detector and its performance are derived.

2) A waveform design solution based on maximizing KLD-LMP is introduced.

3) The detection performances resulting from using the KLD-based and MI-based waveform are studied and compared.

4) An interesting relationship among KLD, MI and SNR is derived, which provides some insights into the role of each criterion.

The content of the paper is as follows. We first discuss the problem of choosing an optimality criteria regarding detection versus estimation in Section 1.2. Next, the radar waveform design problem is formulated in Section 1.3. We then derive the LMP detector and its performance in Section 1.4. The optimal waveform design based on the LMP detector follows in Section 1.5, while that based on mutual information is given in Section 1.6. An important relationship between KLD, MI and SNR is derived in Section 1.7. In Section 1.8, the performance of the derived waveform design method is evaluated and compared with that of the existing MI-

based design via computer simulation. Finally, some discussion and conclusions are given in Section 1.9.

Notation: Throughout the paper, $I(x;s)$ denotes the MI between the random variable x and s, $I_x(f)$ denotes the periodogram of the data $\mathbf{x}$, $I(\theta)$ denotes the Fisher information of the parameter θ, $P_x(f)$ denotes the power spectral density of a random signal $x[n]$, $|S(f)|^2$ denotes the energy spectral density of a deterministic signal $s[n]$, $E[\cdot]$ denotes taking expectation, and $p(x|t)$ denotes the PDF of x conditioned on t.

1.2 Detection versus Estimation - Optimality Criteria

The link between the problems of detection and parameter estimation is not always a strong one. Hence, one should be wary of applying the solution for one of problem to the other. To illustrate a breakdown of the sometimes accepted linkage and to set the stage for the waveform design problem, which is the topic of this paper, consider the following detection problem. Although somewhat contrived, it conveys our assertion with clarity. Assume we wish to decide among the binary hypotheses

$$\begin{aligned} \mathcal{H}_0 &: x \sim \mathcal{N}(0, 2+2\rho_0) \\ \mathcal{H}_1 &: x \sim \mathcal{N}(0, 2+2\rho_1) \end{aligned}$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian probability density function (PDF) with mean μ and variance σ^2, and ρ_i is a parameter taking on values $|\rho_i| < 1$. Note that as ρ_1 becomes closer to ρ_0, it will be exceedingly difficult to decide between the

nearly identical hypotheses. In this case, we can expect $P_D \approx P_{FA}$, where P_{FA} is the probability of false alarm and P_D is the probability of detection, and hence, an exceedingly poor detector. Now, under $\mathcal{H}_1$ assume that $x = s + w$, where s represents a zero-mean random signal and w is a noise, and instead consider the estimation problem. That is to say we wish to estimate s based on the observation x, assuming that the joint PDF under $\mathcal{H}_1$ is

$$\begin{bmatrix} s \\ w \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{bmatrix} \right).$$

Note that under this joint PDF assumption $x = s + w$ has the given PDF under $\mathcal{H}_1$. Choosing the Bayesian mean square error (BMSE) to minimize, it is well known [1] that the MMSE estimator is $\hat{s} = E[s|x] = (\mathrm{cov}(x, s)/\mathrm{var}(x))x$. This is evaluated as

$$\begin{aligned} \hat{s} &= \frac{E[(s+w)s]}{E[(s+w)^2]}x \\ &= \frac{1+\rho_1}{2+2\rho_1}x \\ &= \frac{1}{2}x. \end{aligned} \tag{1.1}$$

The minimum BMSE is well known to be

$$\mathrm{BMSE}_{\min} = \mathrm{var}(s)(1 - \rho_{x,s}^2)$$

where $\rho_{x,s}$ is the correlation coefficient between x and s. But

$$\begin{aligned}\rho_{x,s}^2 &= \frac{\text{cov}^2(x,s)}{\text{var}(x)\text{var}(s)} \\ &= \frac{E^2[xs]}{E[x^2]E[s^2]} \\ &= \frac{(1+\rho_1)^2}{(2+2\rho_1)1} \\ &= \frac{1+\rho_1}{2}. \end{aligned} \tag{1.2}$$

Hence, as $\rho_1 \to 1$, the BMSE goes to zero, and the estimation of s using (1.1) is without error. Hence, the two goals of detection and estimation are not coupled in that good performance in one problem does not guarantee good performance in another.

To illustrate this further we note that the KLD, which measures detectability (at least in an asymptotic sense), is zero for the detection problem since

$$\begin{aligned} D(p_1||p_0) &= \int p_1(x) \ln \frac{p_1(x)}{p_0(x)} dx \\ &= \frac{1}{2}\left(\frac{2+2\rho_1}{2+2\rho_0}\right) - \frac{1}{2}\ln\left(\frac{2+2\rho_1}{2+2\rho_0}\right) - \frac{1}{2} \end{aligned}$$

as $\rho_1 \to \rho_0$. However, the MI between the random variables x and s under $\mathcal{H}_1$ is known to be

$$\begin{aligned} I(x;s) &= \int\int p_1(x,s) \ln \frac{p_1(x,s)}{p_1(x)p(s)} ds dx \\ &= \frac{1}{2}\ln\frac{1}{1-\rho_{x,s}^2}. \end{aligned}$$

As $\rho_1 \to 1$, we have from (1.2) that $\rho_{x,s}^2 \to 1$ and hence, $I(x;s) \to \infty$, indicating perfect knowledge of s given the observation x. Thus, *the detectability is zero even though the MI can be made arbitrarily large.* In summary, this simple example

illustrates the possible pitfalls in using mutual information as a design metric for a detection problem. We continue this discussion in Section 1.7 where an explicit relationship between KLD and MI is established.

1.3 Problem Formulation

We now consider the detection of an extended target in colored wide sense stationary Gaussian noise. This is the problem originally posed by Bell [6] that has led to the use of MI as a waveform design criterion. To simplify the discussion we assume sampled real data. For complex data we will state the obvious extensions without proof. To begin assume we have the detection problem shown in Figure 1.1, where $\mathbf{s} = [s[0]\, s[1] \ldots s[N-1]]^T$ is the transmitted deterministic signal, $\mathbf{h} = [h[0]\, h[1] \ldots h[N-1]]^T$ is the impulse response of the target, and $\mathbf{w} = [w[0]\, w[1] \ldots w[N-1]]^T$ is the observation noise. We model the impulse re-

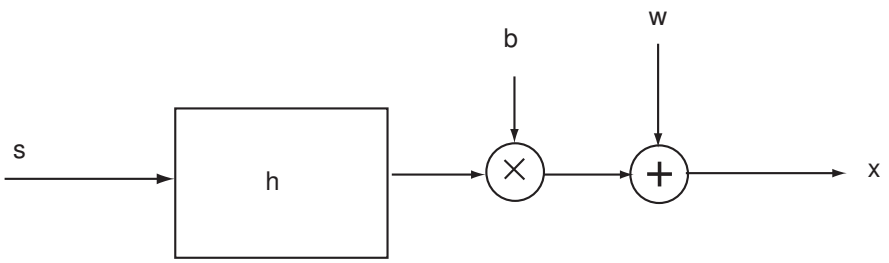

Figure 1.1. Model for the received data for an extended target in noise.

sponse as a Gaussian WSS random process with zero mean and power spectral density (PSD) $P_h(f)$. Likewise, the observation noise is modeled as a Gaussian WSS random process with zero mean and PSD $P_w(f)$. The received data is assumed to be $x[n] = w[n]$ under $\mathcal{H}_0$ and $x[n] = \sqrt{\theta} s[n] \star h[n] + w[n]$ under $\mathcal{H}_1$,

where $b = \sqrt{\theta}$ denotes a small positive number and $\star$ indicates convolution. The parameter θ accounts for channel attenuation and is assumed known. Finally, we will assume that N is large so that the detection problem may equivalently be posed in the frequency domain [2] as a hypothesis test on the PSD $P_x(f)$ of $x[n]$ as follows

$$\begin{aligned} \mathcal{H}_0 &: P_x(f) = P_w(f) \\ \mathcal{H}_1 &: P_x(f) = \theta P_h(f)|S(f)|^2 + P_w(f) \end{aligned}$$

where the transmit signal energy is constrained to be $\int_{-\frac{1}{2}}^{\frac{1}{2}} |S(f)|^2 df \leq \mathcal{E}$, and $|S(f)|^2$ is the energy spectral density (ESD). We wish to choose $|S(f)|^2$ to maximize P_D subject to a constraint on P_{FA} as per the NP approach to detection. To do so we first need to determine P_D and P_{FA}. The problem posed is that of detection of a Gaussian signal in Gaussian noise. Much is known about this problem with the definitive analysis contained in [3]. With a large data record assumption (i.e. as $N \to \infty$) the NP detector test statistic can be shown to be given by [2]

$$T_{NP}(\mathbf{x}) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{I_x(f)}{P_w(f)} \frac{\theta P_h(f)|S(f)|^2}{\theta P_h(f)|S(f)|^2 + P_w(f)} df$$

where $I_x(f)$ is the periodogram of the received data. The NP test statistic can be further approximated for large N in discrete-time as

$$T_{NP}(\mathbf{x}) = \frac{1}{N} \sum_{k=0}^{N-1} \frac{I_x(f_k)}{P_w(f_k)} \frac{\theta P_h(f_k)|S(f_k)|^2}{\theta P_h(f_k)|S(f_k)|^2 + P_w(f_k)} \tag{1.3}$$

for $f_k = k/N$, where we have changed the frequency interval to the equivalent one of $[0, 1]$. As the $I_x(f_k)$'s are mutually independent and their PDFs are asymptotically

(for large N, which is denoted by the "a") [2]

$$\frac{I_x(f_k)}{P_w(f_k)/2} \overset{a}{\sim} \chi_2^2 \quad \text{under } \mathcal{H}_0$$
$$\frac{I_x(f_k)}{[\theta P_h(f_k)|S(f_k)|^2 + P_w(f_k)]/2} \overset{a}{\sim} \chi_2^2 \quad \text{under } \mathcal{H}_1$$

the test statistic $T_{NP}(\mathbf{x})$ can be viewed as a sum of weighted independent and identically distributed (IID) random variables, which are chi-squared distributed. The exact closed-form PDF of $T_{NP}(\mathbf{x})$ is not trivial to derive. A number of attempts to determine the exact PDF of the weighted sum of IID chi-squared random variables is summarized in [11]. Unfortunately, the determination of these probabilities is exceedingly difficult, even with the large sample-frequency domain approach we have taken. The optimization of the detector with respect to the transmit signal is even more formidable. Probably this may have motivated the MI approach to signal design. We propose to solve this problem by using an approximation to the NP detector, known as the LMP approach. By simulation it will later be shown to have performance nearly identical to the NP detector under small signal assumption. By using the LMP formulation, an analytical result is obtained for the performance, and this then leads to a simple criterion for signal design. The optimal signal obtained is shown, again via computer simulation, to be superior to the use of the MI waveform design criterion.

1.4 The LMP Detector and its Performance

As shown in Appendix A the LMP detector decides a signal is present if

$$T_{\text{LMP}}(\mathbf{x}) = \frac{\frac{N}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{P_h(f)|S(f)|^2}{P_w(f)} \left(\frac{I_x(f)-P_w(f)}{P_w(f)}\right) df}{\sqrt{\frac{N}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\frac{P_h(f)|S(f)|^2}{P_w(f)}\right)^2 df}} > \gamma. \tag{1.4}$$

This may be interpreted as a correlation in frequency between the true normalized signal PSD i.e., $P_h(f)|S(f)|^2/P_w(f)$, and the data normalized signal PSD, i.e., $(I_x(f) - P_w(f))/P_w(f)$, in the numerator, which is normalized to yield unit variance. The detection performance of the LMP detector is approximated for large data records as [2]

$$T_{\text{LMP}} \overset{a}{\sim} \begin{cases} \mathcal{N}(0,1) & \text{under } \mathcal{H}_0 \\ \mathcal{N}(\sqrt{I(0)}\theta, 1) & \text{under } \mathcal{H}_1. \end{cases} \tag{1.5}$$

This is the standard Gauss-Gauss detection problem for which P_D is monotonically increasing with the deflection coefficient defined as

$$d^2_{\text{LMP}} = \theta^2 I(0).$$

As a result, to maximize the detection performance with respect to the transmitted signal, we need only maximize d^2_{LMP}, which is given as (see Appendix A)

$$d^2_{\text{LMP}} = \frac{N}{2}\theta^2 \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\frac{P_h(f)|S(f)|^2}{P_w(f)}\right)^2 df. \tag{1.6}$$

It is important to note that *the deflection coefficient for this case of a small signal is identical to the twice the KLD* [14]. *Thus, optimal waveform design in the small signal case, which is usually of primary importance, should be approached by maximizing the KLD.* In doing so, we will maximize the detection performance.

1.5 Optimal Waveform Design

The deflection coefficient of (1.6) may be expressed for large data records using $\Delta f = 1/N$ and $f_k = k/N$ as the Riemann sum over the equivalent frequency interval of $[0, 1]$ as

$$\begin{aligned} d^2_{\text{LMP}} &= \frac{N}{2}\theta^2 \sum_{k=0}^{N-1} \left(\frac{P_h(f_k)|S(f_k)|^2}{P_w(f_k)} \right)^2 \Delta f \\ &= \frac{1}{2}\theta^2 \sum_{k=0}^{N-1} \left(\frac{P_h(f_k)|S(f_k)|^2}{P_w(f_k)} \right)^2 \end{aligned}$$

and letting $S_k = |S(f_k)|^2$, $w_k = P_h(f_k)/P_w(f_k)$, we have

$$d^2_{\text{LMP}} = \frac{1}{2}\theta^2 \sum_{k=0}^{N-1} (w_k S_k)^2 \,. \tag{1.7}$$

To maximize this with respect to S_k we note that with the usual energy constraint of $\sum_{k=0}^{N-1} |S(f_k)|^2 \Delta f \leq \mathcal{E}$, we have that $\sum_{k=0}^{N-1} S_k \leq \mathcal{E}/\Delta f$. Thus, we wish to maximize (1.7) over S_k for $k = 0, 1, \ldots, N-1$, where $S_k \geq 0$ and $\sum_{k=0}^{N-1} S_k \leq \mathcal{E}/\Delta f$. This domain for S_k is a simplex and hence a convex set. The function to be maximized is a convex function since it is a quadratic function. Hence, the problem reduces to maximizing a convex function over a convex set, the solution of which is well known to be among the extreme points [4]. The extreme points of the simplex are those for which $S_k = \mathcal{E}/\Delta f$ for a given $k = k_0$ and $S_k = 0$ otherwise. Thus, the solution is to let $S_k = (\mathcal{E}/\Delta f)\delta_{k,k_0}$, where δ_{ij} is the Kronecker delta. Substituting this into (1.7) produces

$$\begin{aligned} d^2_{\text{LMP}} &= \frac{1}{2}\theta^2 \left(w_{k_0} S_{k_0}\right)^2 \\ &= \frac{1}{2}\theta^2 w^2_{k_0} \left(\frac{\mathcal{E}}{\Delta f} \right)^2 \end{aligned}$$

which is maximized for the value of $k = k_0$ that maximizes w_k. Hence, *the optimal signal places all its energy at the frequency where* $w_k = P_h(f_k)/P_w(f_k)$ *is maximum.* Equivalently, we place all the energy at which $P_w(f_k)/P_h(f_k)$ is minimized. Interestingly, this is the same result as for detection of a *deterministic signal* in colored noise [2]. Hence, we have finally that

$$d^2_{\text{LMP}} = \frac{(N\theta\mathcal{E})^2}{2}\left[\max_{k=0,1,\ldots,N-1}\left(\frac{P_h(f_k)}{P_w(f_k)}\right)\right]^2. \tag{1.8}$$

Since we have assumed real data and hence a symmetric PSD, the energy is split between the positive and negative frequency bins. For the case of complex data the same result is valid if the symmetry condition on the PSD is not imposed. Hence, one needs only concentrate all the signal energy in the bin that achieves the maximum value of $w_k = P_h(f_k)/P_w(f_k)$.

1.6 Waveform Design Based on Maximizing Mutual Information

In this section, we consider a signal waveform design solution based on maximizing the MI between the received data and the target ensemble [6], which has been widely used as a metric in the literature. It has also been proved that minimizing the mean square error of estimating the target yields the same signal waveform design solution as the MI-based method in white Gaussian noise [8]. The MI between the received data and target ensemble under $\mathcal{H}_1$ can be shown to be (see Appendix D)

$$I(\mathbf{x};\mathbf{t}) \;=\; \frac{1}{2}\sum_{k=0}^{N-1}\ln\left(1+\frac{\theta P_h(f_k)|S(f_k)|^2}{P_w(f_k)}\right) \tag{1.9}$$

where $\mathbf{t} = \sqrt{\theta}\mathbf{s} \star \mathbf{h}$. It is well known that the MI-based waveform design is the "water-filling" solution [6], [12], [8] given as

$$\theta |S(f_k)|^2 = \max\left[0, \lambda - \frac{P_w(f_k)}{P_h(f_k)}\right] \tag{1.10}$$

where λ, termed the water level, is a constant determined by the energy constraint and which is usually found numerically from the solution of

$$\sum_{k=0}^{N-1} \max\left[0, \lambda - \frac{P_w(f_k)}{P_h(f_k)}\right] = N\theta\mathcal{E}.$$

1.7 A Relationship Between KLD, MI and SNR

Not only have the KLD and the MI served as metrics of radar waveform design, but output SNR has as well [10], [15]. It is felt that a relationship among all the three terms/metrics and its discussion will be beneficial in shedding further light on the waveform design problem.

Recall that $\mathbf{t} = \sqrt{\theta}\mathbf{s} \star \mathbf{h}$, where $\mathbf{s}$ is the transmitted signal sequence and $\mathbf{h}$ is the target impulse response sequence. Then the detection problem can be written as follows.

$$\begin{aligned} \mathcal{H}_0 &: \mathbf{x} = \mathbf{w} \\ \mathcal{H}_1 &: \mathbf{x} = \mathbf{t} + \mathbf{w} \end{aligned}$$

A general relationship between the KLD, the output SNR, and MI is derived in Appendix C. A related result has been used to compute MI in order to obtain the channel capacity per unit cost [13]. For our problem the relationship is best

expressed as

$$D(p_1(\mathbf{x})||p_0(\mathbf{x})) = E_{\mathbf{t}}\left[D(p_1(\mathbf{x}|\mathbf{t})||p_0(\mathbf{x}))\right] - I(\mathbf{x};\mathbf{t}) \tag{1.11}$$

where $p_1(\mathbf{x}|\mathbf{t})$ is the PDF of $\mathbf{x}$ conditioned on the target response $\mathbf{t}$ under $\mathcal{H}_1$ and the first term on the right-hand-side of the above equation is defined as

$$E_{\mathbf{t}}\left[D(p_1(\mathbf{x}|\mathbf{t})||p_0(\mathbf{x}))\right] = \int_t p(\mathbf{t}) D\big(p_1(\mathbf{x}|\mathbf{t})||p_0(\mathbf{x})\big) d\mathbf{t} \tag{1.12}$$

and can be interpreted as an SNR, as shown next.

Also, $p_1(\mathbf{x})$ can be written as an averaged conditional PDF by averaging $p_1(\mathbf{x}|\mathbf{t})$ over $\mathbf{t}$, as

$$p_1(\mathbf{x}) = \int_{\mathbf{t}} p_1(\mathbf{x},\mathbf{t}) d\mathbf{t} = \int_{\mathbf{t}} p_1(\mathbf{x}|\mathbf{t}) p(\mathbf{t}) d\mathbf{t} \tag{1.13}$$

Thus the term $D(p_1(\mathbf{x})||p_0(\mathbf{x}))$ is the KLD of the *averaged conditional* PDF $p_1(\mathbf{x})$ from the PDF $p_0(\mathbf{x})$.

Moreover, the MI $I\big(\mathbf{x};\mathbf{t}\big)$ is also an averaged KLD obtained by averaging KLD of the conditional PDF $p_1(\mathbf{x}|\mathbf{t})$ from the unconditional PDF $p_1(\mathbf{x})$, $D(p_1(\mathbf{x}|\mathbf{t})||p_1(\mathbf{x}))$, over all possible target signals $\mathbf{t}$ as suggested by

$$\begin{aligned} I(\mathbf{x};\mathbf{t}) &= \int_{\mathbf{t}}\int_{\mathbf{x}} p_1(\mathbf{x},\mathbf{t}) \ln \frac{p_1(\mathbf{x},\mathbf{t})}{p_1(\mathbf{x})p(\mathbf{t})} d\mathbf{x} d\mathbf{t} \\ &= \int_{\mathbf{t}}\int_{\mathbf{x}} p(\mathbf{t}) p_1(\mathbf{x}|\mathbf{t}) \ln \frac{p_1(\mathbf{x}|\mathbf{t})}{p_1(\mathbf{x})} d\mathbf{x} d\mathbf{t} \\ &= \int_{\mathbf{t}} p(\mathbf{t}) D(p_1(\mathbf{x}|\mathbf{t})||p_1(\mathbf{x})) d\mathbf{t}. \end{aligned} \tag{1.14}$$

Therefore, all the three terms of the decomposition (1.11) can be interpreted respectively as distance measurements in the KLD sense. Alternatively, we can

write the relationship as

$$\underbrace{D(p_1(\mathbf{x})||p_0(\mathbf{x}))}_{\text{KLD}} = \underbrace{E_t[D(p_1(\mathbf{x}|\mathbf{t})||p_0(\mathbf{x}))]}_{\text{SNR}} - \underbrace{I(\mathbf{x};\mathbf{t})}_{\text{MI}}. \tag{1.15}$$

Specifically, for the problem at hand the terms may be easily evaluated (see Appendix D) to yield

$$\begin{aligned} E_{\mathbf{t}}\left[D(p_1(\mathbf{x}|\mathbf{t})||p_0(\mathbf{x}))\right] &= \frac{1}{2}\sum_{k=0}^{N-1}\frac{\theta|S(f_k)|^2P_h(f_k)}{P_w(f_k)} \\ I(\mathbf{x};\mathbf{t}) &= \frac{1}{2}\sum_{k=0}^{N-1}\ln\left(1+\frac{\theta P_h(f_k)|S(f_k)|^2}{P_w(f_k)}\right) \end{aligned}$$

so that the KLD is given from (6.5) as

$$D(p_1(\mathbf{x})||p_0(\mathbf{x})) = \frac{1}{2}\sum_{k=0}^{N-1}\frac{\theta|S(f_k)|^2P_h(f_k)}{P_w(f_k)} - \frac{1}{2}\sum_{k=0}^{N-1}\ln\left(1+\frac{\theta P_h(f_k)|S(f_k)|^2}{P_w(f_k)}\right). \tag{1.16}$$

The negative of the second term is the MI and is seen to be concave, leading to the usual maximization of the MI, subject to the energy constraint. Of course, the appropriate criterion for maximization must also take into account the term, which is the SNR. Finally, the result in (1.16) agrees with the usual asymptotic KLD between two multivariate Gaussian PDFs with PSDs given as [9]

$$D(p_1(\mathbf{x})||p_0(\mathbf{x})) = \frac{N}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}}\left(\frac{P_1(f)}{P_0(f)} - \ln\frac{P_1(f)}{P_0(f)} - 1\right)df$$

with $P_1(f) = \theta P_h(f)|S(f)|^2 + P_w(f)$ and $P_0(f) = P_w(f)$. Also, note that for this case of a small signal, i.e., asymptotically, it is shown in Appendix B that the KLD is symmetric in that $D(p_1(\mathbf{x})||p_0(\mathbf{x})) = D(p_0(\mathbf{x})||p_1(\mathbf{x}))$.

This reveals an interesting relationship among the three terms KLD, MI and

SNR, all of which have been used as design metrics. It is

$$\text{KLD} = \text{SNR} - \text{MI}.$$

As shown, the important metric for detection performance is KLD. The SNR term is the "distance" between the PDFs when the signal $\mathbf{t}$ is known as in a matched filter detector when averaged over the possible signals. The MI term is a degradation factor which accounts for the fact that the signal, i.e., target response, is actually unknown. Since the target response is unknown, it has been modeled as the outcome of a random process, for which the optimal detector is an estimator-correlator. The loss in performance is the MI, which measures how much the unknown target response is reflected in the PDF of the received data. In summary, the MI can be viewed as the loss in detection performance between a matched filter and an estimator-correlator. Note that for the known signal case, in which the target $\mathbf{t}$ is known and not random, we have that MI = 0. Hence, it follows that KLD = SNR, which is 1/2 the deflection coefficient.

1.8 Computer Simulations and Analysis

In this section the performance of the proposed LMP-based signal design solution is evaluated through computer simulations and compared with that of the MI-based signal design solution. We consider a signal $\mathbf{s} = \begin{bmatrix} s[0] & s[1] & \cdots & s[N-1] \end{bmatrix}^T$ with length $N = 64$ for a case when the signal energy $\theta\mathcal{E} = 5.12$ and the ratio of the noise PSD $P_w(f)$ and target PSD $P_h(f)$, termed the noise-target power ratio at each frequency bin, $\frac{P_w(f_k)}{P_h(f_k)}$ for $k = 0, 1, \cdots, N-1$, is shown in Figure 1.2a. As

both PSDs are symmetric, the noise-target power ratio is also symmetric. Recall that the signal and data were assumed real.

The MI-based waveform design is found by the water-filling technique according to (1.10) and is depicted in Figure 1.2b. However, the LMP-based waveform design is to split all the signal energy into two symmetric frequency bins f_{13} and f_{51} (centered about $f = 1/2$), where the ratio $\frac{P_w(f_k)}{P_h(f_k)}$ is the minimum among all frequency bins, as shown in Figure 1.2c.

We next use the two waveforms in the NP detector in (1.3) as well as the LMP detector in (1.4). The resulting receiver operating characteristic (ROC) are given in Figure 1.3. The LMP-based waveform design outperforms the MI-based design substantially, and the LMP detector has a performance close to that of the NP detector for both waveform designs.

To further explore the performance of the LMP-based waveform under different signal energy constraints, the signal energy $\theta\mathcal{E}$ is varied from small values to larger ones. The rest of the simulation parameters remains unaltered. The LMP detector detection performance is nearly identical to that of the NP detector under small signal cases, so we only compare the performances of the two waveform designs using the NP detector. In Figure 1.4 we compare the NP detector P_D's by using the two waveform design methods for different signal energy constraints, with the P_{FA} being fixed. It is seen that once again the LMP-based waveform produces better detection performance than the MI-based waveform. This is seen to be true even when the small signal assumption is no longer valid.

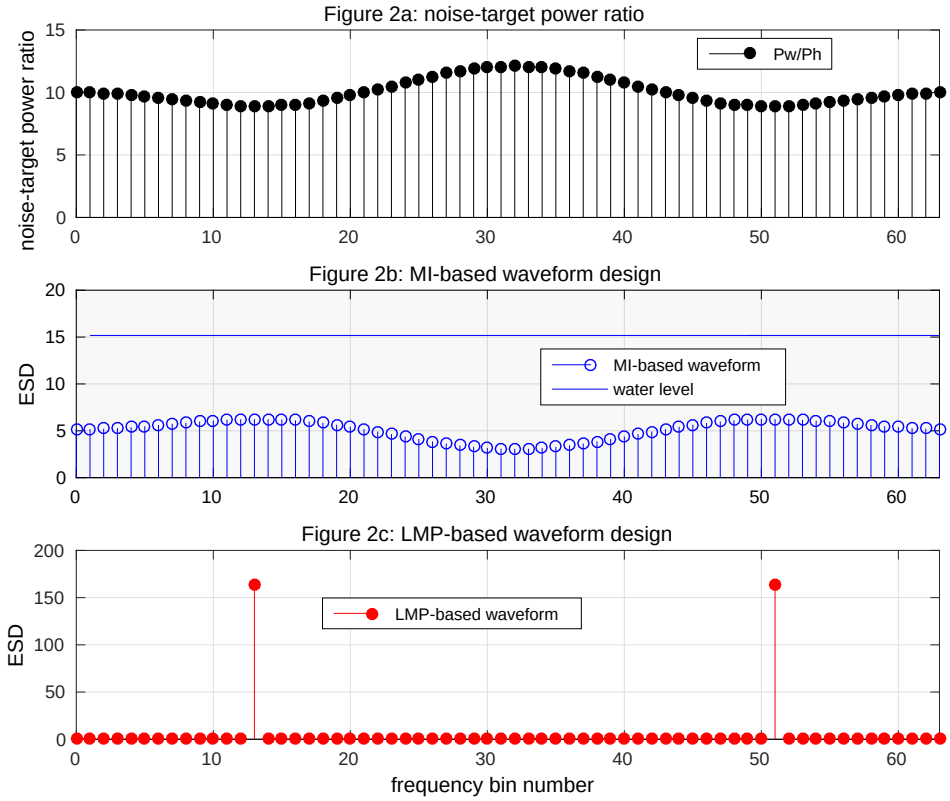

Figure 1.2. Simulation setting up and the two waveform design solutions. Top: the samples of noise-target ratio $\frac{P_w(f_k)}{P_h(f_k)}$. Middle: the MI-based waveform design solution and the water filling level λ. Bottom: the LMP-based waveform design solution

.

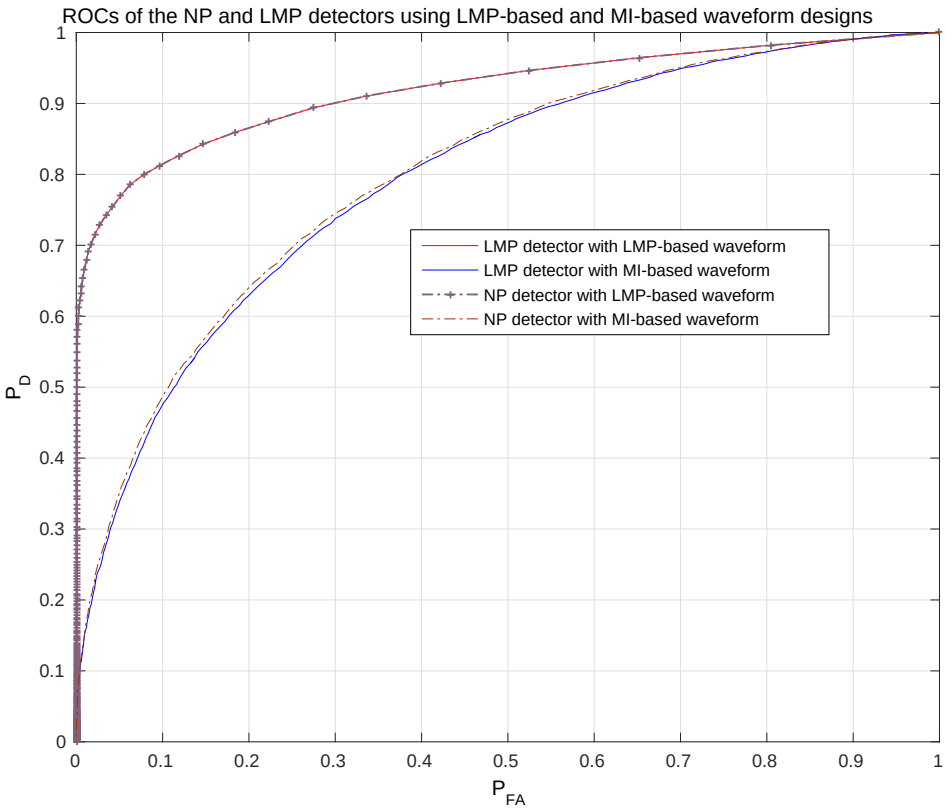

Figure 1.3. The ROCs of the LMP and NP detectors using LMP-based and MI-based waveform designs with $\theta\mathcal{E} = 5.12$

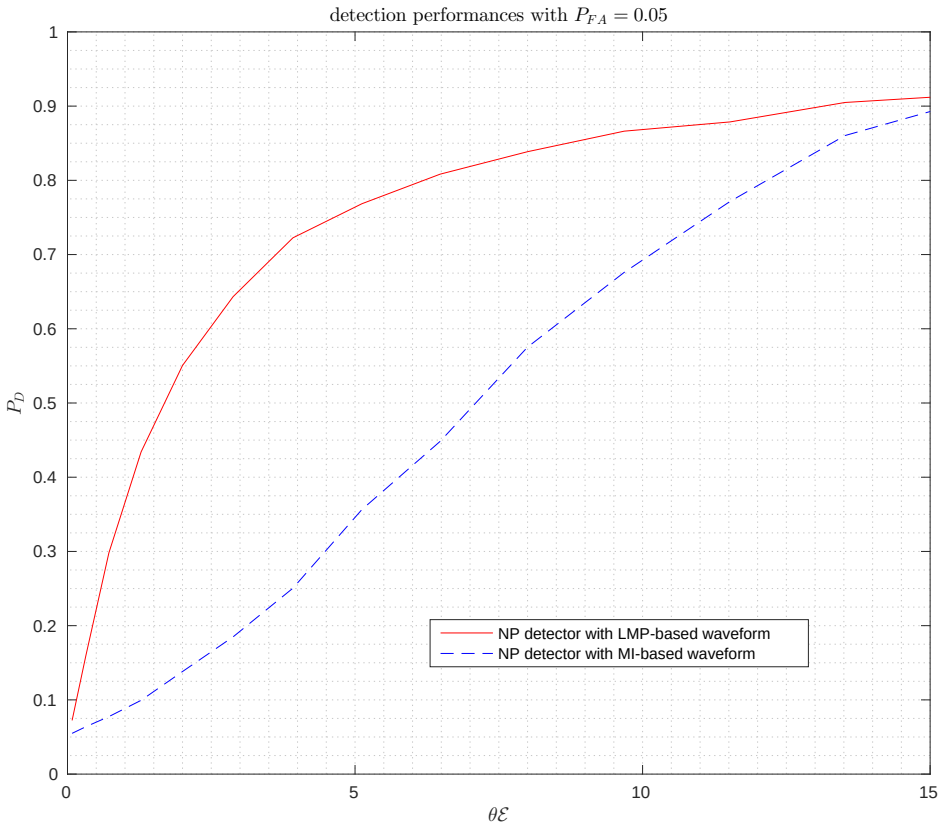

Figure 1.4. The detection performance versus signal energy.

The improvement of the LMP-designed waveform over the MI-designed waveform depends critically on the dependence of the noise-target frequency dependence. To illustrate this dependence we use the same signal as before, having a length $N = 64$ and an energy $\theta\mathcal{E} = 5.12$. However, the noise-target power ratio is now given in Figure 1.5a. Note that the noise-target ratio changes more drastically from one frequency bin to another as compared to the previous example. The corresponding MI-based waveform and LMP-based waveform are given in Figure 1.5b

and Figure 1.5c respectively. Compared with Figure 1.2b, the MI-based waveform for this example becomes more "concentrated" since it allocates zero energy for many of the frequency bins. The LMP-based waveform still splits the total energy into two symmetric frequency bins where the noise-target power ratio is minimum. The resulting ROCs for the NP detector using the two waveforms are shown in Figure 1.6. It is seen that the performance differential of the two waveforms is now decreased, although the LMP-based waveform still outperforms MI-based waveform in the low P_{FA} region. As an extreme case, the two waveform designs may lead to the same solution, splitting signal energy into the two symmetric bins where the noise-target power ratio is minimum. This occurs when the noise-target power ratios of all other bins reach/exceed the water-level of MI-based solutions.

1.9 Conclusions

We have shown that the link between the problems of detection and parameter estimation is not always a strong one, and illustrates possible pitfalls in using MI as a design metric for a detection problem. The LMP detector and its asymptotic performance for the case of an extended target in WSS colored noise are derived. The asymptotic results are fundamental in nature against which performance for signal sets with finite number of samples can be compared. To maximize detection performance it has been shown that the KLD is the appropriate measure to be maximized. The very simple result that the signal should place all its energy at the minimum of the noise PSD (normalized by $P_h(f)$) is both satisfying and intuitively appealing. Similar results are well known for signal design in the case of an assumed

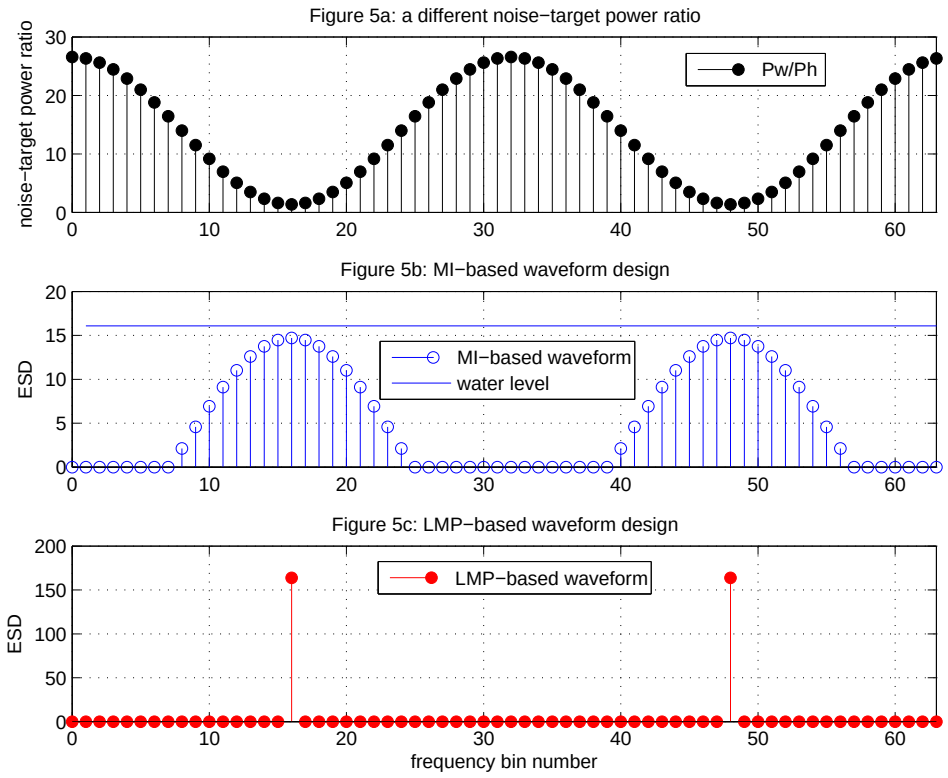

Figure 1.5. A second noise-target power ratio and the corresponding waveform designs.

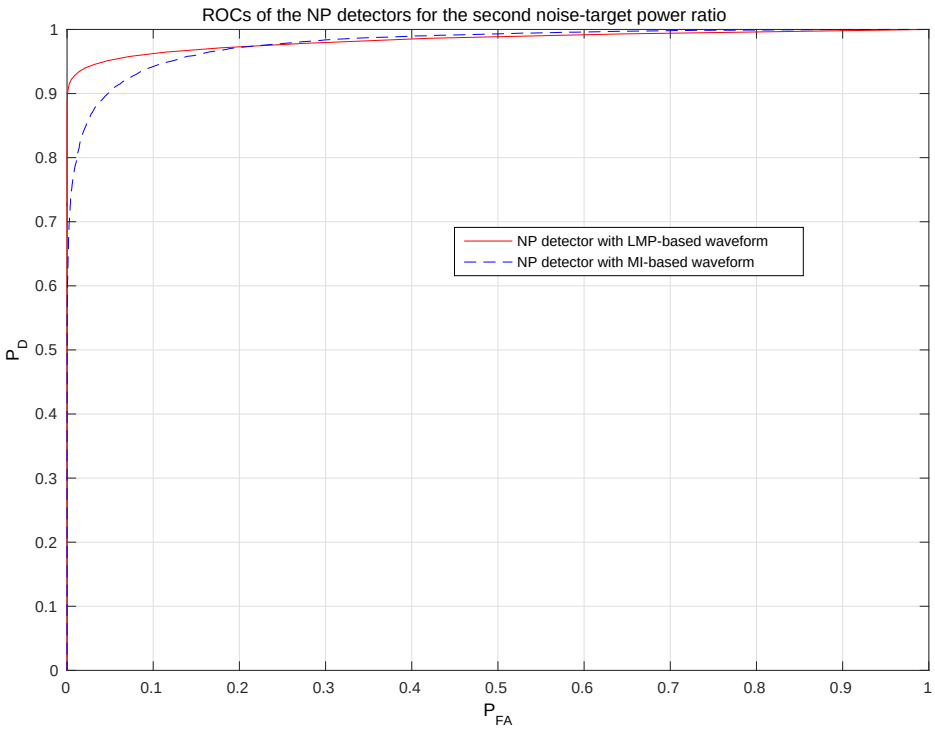

Figure 1.6. ROCs for NP detectors for the second noise-target ratio power ratio example.

known signal at the receiver. Even in the case of an NP detector implementation signal design based on KLD yields superior performance to signal design based on MI, as shown via computer simulation. Furthermore, the important relationship KLD = SNR − MI connects the three design metrics, namely KLD, SNR, and MI, in extended target waveform design and suggests that MI can be viewed as a performance loss between a matched filter and an estimator-correlator.

List of References

[1] Kay, S., *Fundamentals of Statistical Signal Processing: Estimation*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[2] Kay, S., *Fundamentals of Statistical Signal Processing: Detection*, Prentice-Hall, Englewood Cliffs, NJ, 1998.

[3] Van Trees, H.L., *Detection, Estimation, and Modulation Theory, Vol. III*, New York: J. Wiley, 1971.

[4] Rockafellar, R.T., *Convex Analysis*, Princeton: Princeton University Press, 1970

[5] Gini, F., A. De Maio, and L. Patton, *Waveform Design and Diversity for Advanced Radar Systems*. London, U.K.: IET Radar and Sonar Navigation, 2011

[6] Bell, M.R., "Information Theory and Radar Waveform Design," *IEEE Trans. Inf. Theory* , vol. 39, no. 5, pp. 1578–1597, Sept. 1993.

[7] Bell, M.R., "Information Theory and Radar Waveform Design," in *Waveform Design and Diversity for Advanced Radar Systems*. London, U.K.: IET Radar and Sonar Navigation, 2011

[8] Yang, Y., and R.S. Blum, "MIMO Radar Waveform Design Based on Mutual Information and Minimum Mean-square Error Estimation," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 43, pp. 330–343, Jan. 2007.

[9] Kazakos, D. and P. Papantoni-Kazakos, *Detection and Estimation*, New York: Computer Science Press,1990.

[10] Kay, S., " Waveform Design for Multistatic Radar Detection," *IEEE Trans. Aerosp. Electron. Syst.,* vol. 45, no.4, pp. 1153–1166. Jul. 2009.

[11] Mathai, A.M., and S. B. Provost, *Quadratic Forms in Random Variables*, New York: Marcel Dekker, 1992.

[12] Romero, R., and N. A. Goodman, "Improved Waveform Design for Target Recognition with Multiple Transmissions", in *Proceedings of the International Waveform Diversity and Design Conference (WDD '09)*, pp. 26–30, Fla, USA, Feb. 2009.

[13] Verdu, S., "On Channel Capacity per Unit Cost," *IEEE Trans. Inform. Theory,* vol. 36, pp. 1019–1030, Sept. 1990.

[14] Kullback, S., "Information Theory and Statistics", John Wiley & Sons, 1959.

[15] Romero, R.A., J. Bae, and N. A. Goodman, "Theory and Application of SNR and Mutual Information Matched Illumination Waveforms," *IEEE Trans. Aerosp. Electron. Syst.,*, vol. 47, no.2, pp. 912–927, April 2011.

[16] Kass, R.E. and P. W. Voss, *Geometrical Foundations of Asymptotic Inference*, J. Wiley, NY, 1997.

[17] Kay, S., "Optimal Signal Design for Detection of Gaussian Point Targets in Stationary Gaussian Clutter/Reverberation," *IEEE Journal of Selected Topics in Signal Processing,* vol.1, no.1, pp.31–41, June 2007.

Appendix A - Derivation of LMP **Detector**

The PDF of $\mathbf{x} = [x[0]\, x[1] \ldots x[N-1]]^T$, where $x[n]$ is a Gaussian WSS random process is given for large N by [1]

$$\ln p(\mathbf{x}) = -\frac{N}{2}\ln 2\pi - \frac{N}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\ln P_x(f) + \frac{I_x(f)}{P_x(f)}\right) df$$

where $P_x(f)$ is the PSD of the received data $\mathbf{x}$. Now under $\mathcal{H}_1$ we have upon replacing $P_x(f)$ by $\theta P_h(f)|S(f)|^2 + P_w(f)$

$$\begin{aligned}\ln p(\mathbf{x};\theta) &= -\frac{N}{2}\ln 2\pi \\ &\quad -\frac{N}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\ln\left[\theta P_h(f)|S(f)|^2 + P_w(f)\right] + \frac{I_x(f)}{\theta P_h(f)|S(f)|^2 + P_w(f)}\right) df.\end{aligned}$$

The LMP detector decides $\mathcal{H}_1$ if [2]

$$\frac{\left.\frac{\partial \ln p(\mathbf{x};\theta)}{\partial\theta}\right|_{\theta=\theta_0}}{\sqrt{I(\theta_0)}} > \gamma$$

where $I(\theta_0)$ is the Fisher information evaluated at θ_0. For our problem we have $\theta_0 = 0$. Now we have

$$\begin{aligned}\frac{\partial \ln p(\mathbf{x};\theta)}{\partial\theta} &= -\frac{N}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\frac{P_h(f)|S(f)|^2}{\theta P_h(f)|S(f)|^2 + P_w(f)}\right. \\ &\quad \left. -\frac{P_h(f)|S(f)|^2 I_x(f)}{(\theta P_h(f)|S(f)|^2 + P_w(f))^2}\right) df\end{aligned}$$

and evaluating this at $\theta = 0$ produces

$$\left.\frac{\partial \ln p(\mathbf{x};\theta)}{\partial\theta}\right|_{\theta=0} = -\frac{N}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\frac{P_h(f)|S(f)|^2}{P_w(f)} - \frac{P_h(f)|S(f)|^2 I_x(f)}{P_w^2(f)}\right) df.$$

The Fisher information is found as

$$\begin{aligned}\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial\theta^2} &= -\frac{N}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} \left(-\frac{P_h^2(f)|S(f)|^4}{(\theta P_h(f)|S(f)|^2 + P_w(f))^2}\right. \\ &\quad \left. +2\frac{P_h^2(f)|S(f)|^4 I_x(f)}{(\theta P_h(f)|S(f)|^2 + P_w(f))^3}\right) df\end{aligned}$$

and taking the expected value and noting that $E[I_x(f)] \approx P_x(f) = \theta P_h(f)|S(f)|^2 + P_w(f)$ for large data records, we have

$$\begin{aligned} E\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial\theta^2}\right] &= -\frac{N}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} \left(-\frac{P_h^2(f)|S(f)|^4}{(\theta P_h(f)|S(f)|^2 + P_w(f))^2} \right. \\ &\quad \left. +2\frac{P_h^2(f)|S(f)|^4}{(\theta P_h(f)|S(f)|^2 + P_w(f))^2}\right) df \\ &= -\frac{N}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{P_h^2(f)|S(f)|^4}{(\theta P_h(f)|S(f)|^2 + P_w(f))^2} df. \end{aligned}$$

Finally, we have for the Fisher information

$$\begin{aligned} I(\theta) &= -E\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial\theta^2}\right] \\ &= \frac{N}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{P_h^2(f)|S(f)|^4}{(\theta P_h(f)|S(f)|^2 + P_w(f))^2} df \end{aligned}$$

which when evaluated at $\theta = 0$ is

$$I(0) = \frac{N}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\frac{P_h(f)|S(f)|^2}{P_w(f)}\right)^2 df.$$

As a result the LMP test statistic is

$$T_{\text{LMP}}(\mathbf{x}) = \frac{\frac{N}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\frac{P_h(f)|S(f)|^2 I_x(f)}{P_w^2(f)} - \frac{P_h(f)|S(f)|^2}{P_w(f)}\right) df}{\sqrt{\frac{N}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\frac{P_h(f)|S(f)|^2}{P_w(f)}\right)^2 df}}$$

where we have used (1.4).

Appendix B - Symmetry of KL Measure for Small Signals

From Appendix A, we have for a large data record and using a discrete approximation to the integrals involved

$$\ln p_0(\mathbf{x}) = -\frac{N}{2}\ln 2\pi - \frac{1}{2}\sum_{k=0}^{N-1}\left(\ln P_w(f_k) + \frac{I_x(f_k)}{P_w(f_k)}\right)$$

and

$$\ln p_1(\mathbf{x}) = -\frac{N}{2}\ln 2\pi - \frac{1}{2}\sum_{k=0}^{N-1}\left(\ln\left(\theta P_h(f_k)|S(f_k)|^2 + P_w(f_k)\right) + \frac{I_x(f_k)}{\theta P_h(f_k)|S(f_k)|^2 + P_w(f_k)}\right.$$

bigg).

Then,

$$\ln p_0(\mathbf{x}) - \ln p_1(\mathbf{x}) = \frac{1}{2}\sum_{k=0}^{N-1}\left[\ln\frac{\theta P_h(f_k)|S(f_k)|^2 + P_w(f_k)}{P_w(f_k)} + I_x(f_k)\left(\frac{1}{\theta P_h(f_k)|S(f_k)|^2 + P_w(f_k)} - \frac{1}{P_w(f_k)}\right)\right]$$

The KLD $D(p_0||p_1)$ can be evaluated to be

$$\begin{aligned} D(p_0||p_1) &= E_0(\ln p_0 - \ln p_1) \\ &= \frac{1}{2}\sum_{k=0}^{N-1}\left[\ln\frac{\theta P_h(f_k)|S(f_k)|^2 + P_w(f_k)}{P_w(f_k)} + \frac{P_w(f_k)}{\theta P_h(f_k)|S(f_k)|^2 + P_w(f_k)} - \frac{P_w(f_k)}{P_w(f_k)}\right], \end{aligned} \tag{1.17}$$

where $E_0(\cdot)$ denotes taking expectation under $p_0(\mathbf{x})$. Furthermore, using the following Taylor expansions

$$\sum_{k=0}^{N-1}\frac{P_w(f_k)}{\theta P_h(f_k)|S(f_k)|^2 + P_w(f_k)} \approx N - \sum_{k=0}^{N-1}\frac{\theta P_h(f_k)|S(f_k)|^2}{P_w(f_k)} + \sum_{k=0}^{N-1}\left(\frac{\theta P_h(f_k)|S(f_k)|^2}{P_w(f_k)}\right)^2 \tag{1.18}$$

and

$$\sum_{k=0}^{N-1}\ln\frac{\theta P_h(f_k)|S(f_k)|^2 + P_w(f_k)}{P_w(f_k)} \approx \sum_{k=0}^{N-1}\frac{\theta P_h(f_k)|S(f_k)|^2}{P_w(f_k)} - \frac{1}{2}\sum_{k=0}^{N-1}\left(\frac{\theta P_h(f_k)|S(f_k)|^2}{P_w(f_k)}\right)^2, \tag{1.19}$$

we have in the small signal case,

$$D(p_0||p_1) \approx \frac{1}{4}\sum_{k=0}^{N-1}\left(\frac{\theta P_h(f_k)|S(f_k)|^2}{P_w(f_k)}\right)^2. \tag{1.20}$$

Moreover, the KLD $D(p_1||p_0)$ can also be approximated in the small signal case as

$$\begin{aligned} D(p_1||p_0) &= E_1(\ln\frac{p_1}{p_0}) \\ &= -E_1(\ln p_0 - \ln p_1) \\ &= -\frac{1}{2}\sum_{k=0}^{N-1}\Bigg[\ln\frac{\theta P_h(f_k)|S(f_k)|^2 + P_w(f_k)}{P_w(f_k)} \\ &\quad + \frac{\theta P_h(f_k)|S(f_k)|^2 + P_w(f_k)}{\theta P_h(f_k)|S(f_k)|^2 + P_w(f_k)} - \frac{\theta P_h(f_k)|S(f_k)|^2 + P_w(f_k)}{P_w(f_k)}\Bigg] \\ &\approx \frac{1}{4}\sum_{k=0}^{N-1}\left(\frac{\theta P_h(f_k)|S(f_k)|^2}{P_w(f_k)}\right)^2, \end{aligned} \tag{1.21}$$

where $E_1(\cdot)$ denotes taking expectation under $p_1(\mathbf{x})$. This proves the asymptotic symmetry of $D(p_1||p_0)$ and $D(p_0||p_1)$ in the small signal case. Both KLD measures are locally equal to $\frac{1}{2}\theta^2 I(0)$, which is $\frac{1}{2}d^2_{\text{LMP}}$ in the small signal case. This result may also be obtained more generally by noting that the KL measure is a Riemannian metric when we consider the Riemannian space of PDFs [16].

Appendix C - Derivation of Relationship Between KLD, SNR, and MI

In this appendix we prove that for $\mathbf{x}$ and $\mathbf{t}$ jointly distributed random variables that

$$D(p_1(\mathbf{x})||p_0(\mathbf{x})) = E_{\mathbf{t}}[D(p_1(\mathbf{x}|\mathbf{t})||p_0(\mathbf{x}))] - I(\mathbf{x};\mathbf{t}) \tag{1.22}$$

where $\pi(\mathbf{t})$ is the PDF of $\mathbf{t}$. The derivation is straightforward.

$$\begin{aligned} \ln\frac{p_1(\mathbf{x})}{p_0(\mathbf{x})} &= \ln\frac{p_1(\mathbf{x}|\mathbf{t})}{p_0(\mathbf{x})} - \ln\frac{p_1(\mathbf{x}|\mathbf{t})}{p_1(\mathbf{x})} \\ &= \ln\frac{p_1(\mathbf{x}|\mathbf{t})}{p_0(\mathbf{x})} - \ln\frac{p_1(\mathbf{x},\mathbf{t})}{p_1(\mathbf{x})\pi(\mathbf{t})} \end{aligned}$$

and taking the expected value with respect to $p_1(\mathbf{x}, \mathbf{t})$ produces

$$E_{\mathbf{x},\mathbf{t}}\left[\ln \frac{p_1(\mathbf{x})}{p_0(\mathbf{x})}\right] = E_{\mathbf{t}} E_{\mathbf{x}|\mathbf{t}}\left[\ln \frac{p_1(\mathbf{x}|\mathbf{t})}{p_0(\mathbf{x})}\right] - E_{\mathbf{x},\mathbf{t}}\left[\ln \frac{p_1(\mathbf{x},\mathbf{t})}{p_1(\mathbf{x})\pi(\mathbf{t})}\right]$$

to yield (1.11).

Appendix D - Evaluation of KLD, SNR, and MI Terms

The output SNR is now shown to be

$$E_{\mathbf{t}}\big[D(p_1(\mathbf{x}|\mathbf{t})||p_0(\mathbf{x}))\big] = \frac{1}{2}\sum_{k=0}^{N-1} \frac{\theta P_h(f_k)|S(f_k)|^2}{P_w(f_k)}$$

Firstly, we have for the asymptotic conditional PDF for a given $\mathbf{t}$, using a discrete approximation to the integral

$$\ln p_1(\mathbf{x}|\mathbf{t}) = -\frac{N}{2}\ln 2\pi - \frac{1}{2}\sum_{k=0}^{N-1} \ln P_w(f_k) - \frac{1}{2}\sum_{k=0}^{N-1} \frac{I_{x|t}(f_k)}{P_w(f_k)}$$

where

$$I_{x|t}(f_k) = \frac{1}{N}\left|\sum_{n=0}^{N-1}(x[n] - t[n])\exp(-j2\pi f_k n)\right|^2$$

is the periodogram of the received data $\mathbf{x}$ under the assumption of a fixed and known target $\mathbf{t}$, and the asymptotic PDF of the received data $\mathbf{x}$ under $\mathcal{H}_0$

$$\ln p_0(\mathbf{x}) \;=\; -\frac{N}{2}\ln 2\pi - \frac{1}{2}\sum_{k=0}^{N-1} \ln P_w(f_k) - \frac{1}{2}\sum_{k=0}^{N-1} \frac{I_x(f_k)}{P_w(f_k)}. \tag{1.23}$$

Then,

$$\ln p_1(\mathbf{x}|\mathbf{t}) - \ln p_0(\mathbf{x}) = \frac{1}{2}\sum_{k=0}^{N-1} \frac{I_x(f_k) - I_{x|t}(f_k)}{P_w(f_k)} \tag{1.24}$$

Therefore,

$$\begin{aligned} E_{\mathbf{t}}\left[D(p_1(\mathbf{x}|\mathbf{t})||p_0(\mathbf{x}))\right] &= \int_{\mathbf{t}}\int_{\mathbf{x}} p(\mathbf{t}) p_1(\mathbf{x}|\mathbf{t})(\ln p_1(\mathbf{x}|\mathbf{t})(\mathbf{x}) - \ln p_0(\mathbf{x}))d\mathbf{x}d\mathbf{t} \\ &= \frac{1}{2}\int_{\mathbf{x}} p_1(\mathbf{x}) \sum_{k=0}^{N-1} \frac{I_x(f_k) - P_w(f_k)}{P_w(f_k)} d\mathbf{x} \end{aligned}$$

where we have used $E_{\mathbf{t}}[I_{x|t}(f_k)] = P_w(f_k)$. Continuing we have

$$\begin{aligned} E_{\mathbf{t}}\left[D(p_1(\mathbf{x}|\mathbf{t})||p_0(\mathbf{x}))\right] &= \frac{1}{2}\sum_{k=0}^{N-1}\frac{P_t(f_k)}{P_w(f_k)} \\ &= \frac{1}{2}\sum_{k=0}^{N-1}\frac{\theta|S(f_k)|^2 P_h(f_k)}{P_w(f_k)} \end{aligned} \tag{1.25}$$

where we use the assumption that the target $\mathbf{t}$ is independent of the noise $\mathbf{w}$ under $\mathcal{H}_1$ and $P_x(f_k) = P_t(f_k) + P_w(f_k) = \theta P_h(f_k)|S(f_k)|^2 + P_w(f_k)$.

To evaluate $I(\mathbf{x};\mathbf{t})$ we proceed as follows.

$$\begin{aligned} I(\mathbf{x};\mathbf{t}) &= \int_{\mathbf{x}}\int_{\mathbf{t}} p_1(\mathbf{x},\mathbf{t})\ln\frac{p_1(\mathbf{x},\mathbf{t})}{p_1(\mathbf{x})p(\mathbf{t})}d\mathbf{x}d\mathbf{t} \\ &= E_{\mathbf{x},\mathbf{t}}[\ln p_1(\mathbf{x}|\mathbf{t}) - p_1(\mathbf{x})] \\ &= E_{\mathbf{x},\mathbf{t}}\Bigg[-(N/2)\ln 2\pi - \frac{1}{2}\sum_{k=0}^{N-1}\ln P_w(f_k) - \frac{1}{2}\sum_{k=0}^{N-1}\frac{I_{x|t}(f_k)}{P_w(f_k)} \\ &\quad +(N/2)\ln 2\pi + \frac{1}{2}\sum_{k=0}^{N-1}\ln[\theta P_h(f_k)|S(f_k)|^2 + P_w(f_k)] \\ &\quad +\frac{1}{2}\sum_{k=0}^{N-1}\frac{I_x(f_k)}{\theta P_h(f_k)|S(f_k)|^2 + P_w(f_k)}\Bigg] \\ &= \frac{1}{2}\sum_{k=0}^{N-1}\ln\frac{\theta P_h(f_k)|S(f_k)|^2 + P_w(f_k)}{P_w(f_k)} - \frac{1}{2}\sum_{k=0}^{N-1}\frac{P_w(f_k)}{P_w(f_k)} \\ &\quad +\frac{1}{2}\sum_{k=0}^{N-1}\frac{\theta P_h(f_k)|S(f_k)|^2 + P_w(f_k)}{\theta P_h(f_k)|S(f_k)|^2 + P_w(f_k)} \\ &= \frac{1}{2}\sum_{k=0}^{N-1}\ln\left(1 + \frac{\theta P_h(f_k)|S(f_k)|^2}{P_w(f_k)}\right). \end{aligned}$$

MANUSCRIPT 2

Detection of Nonstationarity of the Covariance Matrix in Radar Signal Processing

by

Zhenghan Zhu, Steven Kay, Fuat Cogun and R.S. Raghavan

in part published

in Proc. of 2016 IEEE Radar Conference, pp.1–4, Philadelphia, May. 2016.

to be submitted to

IEEE Transactions on Aerospace and Electronic Systems

Abstract

Space-time adaptive processing (STAP) has become a leading technique in airborne radar signal processing. One of its key steps is using the interference-plus-noise covariance matrix to form an adaptive filter. The optimality of the STAP assumes the stationarity of the covariance matrices. In practice, however, the covariance matrices may be nonstationary. If such nonstationarity is not detected and not well treated, the STAP system's performance decreases substantially. In this paper, we present two detectors for detecting the covariance matrix nonstationarity. We form the first detector based on the generalized lkelihood ratio test, which inherits the property of asymptotically optimal detection performance. A second detector employs the Rao test and requires significantly less computation than the first detector, which can be the favorable choice when computational load is of concern to the signal processing system. Numerical simulations are run to test the performances of the detectors. The proposed detectors may be used as a pre-processing step in STAP to choose applicable training data, and therefore to improve STAP system's performance.

2.1 Introduction

To indicate ground/airborne moving targets, the signal processing of a modern airborne radar system commonly applies space-time adaptive processing (STAP) [1]. The STAP technique was first introduced by Brennan and Reed in the 1970s [2]. STAP techniques provide significant improvement in radar system output

signal-to-interference-plus-noise-ratio (SINR) and better detection performance in moving target indicator (MTI) system through its two-dimensional, space-time or angle-Doppler, adaptive filtering [3]. Usually, the array of antenna elements carried on an airborne platform provides the spatial frequencies, while a Fourier transform of the slow-time voltage collected from pulsed-transmissions at certain clutter range translates into a Doppler-frequency [4].

The statistical features of the interference and noise environment, in terms of interference-plus-noise covariance matrices, play an essential role in adaptive filtering. The covariance matrices are used to form the optimal weights for adaptive filters. The well-known formula of the optimal weight $\mathbf{w}$ maximizing the SINR is [5]

$$\mathbf{w} = p\mathbf{C}^{-1}\mathbf{s},$$

where p is an arbitrary constant, $\mathbf{C}$ is the interference-plus-noise covariance matrix, and $\mathbf{s}$ is the data vector under test. However, the covariance matrix $\mathbf{C}$ is usually unknown in practice. Therefore, target-free training data, or so-called secondary data, are collected from reference ranges close to the cell under test (CUT) to estimate the covariance matrix in STAP [5]. A commonly used estimator of the matrix is [6]

$$\hat{\mathbf{C}} = \frac{1}{K}\sum_{k=0}^{K-1}\mathbf{x}_k\mathbf{x}_k^H,$$

where H denotes the Hermitian operator, K is the number of training data vectors used in estimation, and $\mathbf{x}_k$ for $k = 0, 1, \cdots, K-1$ are the training data vectors assuming no target present from adjacent ranges of the CUT. Such an estimator

assumes the training data vectors are independently identically distributed (IID) or homogeneous in order to be accurate. In other words, the interference-plus-noise covariance matrices for all training data are assumed to be stationary and representative for that of the testing data. Nevertheless, such an assumption may not hold in real scenarios. The reasons for such nonstationarity being present include: variation in clutter amplitude or spectral spread due to a mixture of clutter types, abrupt edge characteristics of clutter interfaces and so on [6].

Such nonstationarity causes a mismatch between the estimated and actual covariance matrices of the data under test, and furthermore, leads to substantial loss in SINR and detection performance of the STAP. The performance loss due to the nonstationary covariance matrices has been studied in [7], [8], and [9]. Melvin [9] reported that in nonstationary scenarios, STAP could lose SINR by an amount ranging from a few tenths of a decibel to greater than 16 dB for specific cases. Armstrong, et al. reported even great loss as in their analysis [8]. In fact, the operation of STAP in the non-stationary, heterogeneous interference environment is one of the current challenges and open problems [10], [11]. Many efforts have been made to improve STAP algorithms for detection in heterogeneous environments: [12] used reduced dimension/rank algorithms; [13] proposes estimation strategies via structured interferences. Another way to deal with the nonstationarity is the careful selection of the secondary data by discarding heterogeneous samples according to certain criterion, e.g., power considerations or more complex metrics such as nonhomogeneity statistics.

It is therefore critical to detect the nonstationarity so that mismatching covariance matrices are not used to produce adaptive filtering weights. In this paper, two well-behaved detectors to detect the nonstationarity of the covariance matrices are proposed. First, we consider a detector based on generalized lkelihood ratio test (GLRT), which has an asymptotic optimality property [14]. In STAP, the computational load is often a necessary consideration [11]. To ease the computational load of the system, we also derive another nonstationarity detector based on the Rao test. The Rao test has asymptotically equivalent performance to the GLRT when the degree of the nonstationarity is small, yet it requires noticeably lower computation cost as it only needs the maximum lkelihood estimate under null hypothesis [14]. Several computer simulations are carried out to test the performances of both detectors.

The remainder of this paper is organized as follows: Section 2.2 presents an interference-plus-noise covariance matrix model and formulates the detection problem, Section 2.3 derives the GLRT detector and Rao test, Computer simulations and the detectors' performances are illustrated in Section 2.4; Finally, conclusions are drawn in Section 2.5.

2.2 Problem Formulation

The radar system under consideration is a pulsed Doppler radar residing on an airborne platform. The radar antenna is a uniformly spaced linear array antennas of S elements. The radar transmits a burst of T pulses in a coherent processing interval (CPI) and samples from N range rings are collected to cover the range

interval. A radar data cube consists of $S \times T \times N$ complex-valued data. Let the x_{stn} be the complex sample from the s^{th} sensor element, t^{th} pulse, at the n^{th} range gate, and $\mathbf{x}_{t,n}$ be the $S \times 1$ vector of antenna element outputs, or a spatial snapshot, at t^{th} pulse and the n^{th} range gate. Then a $ST \times 1$ vector, $\mathbf{x}_n = \text{vec}(\mathbf{x}_{1,n}\ \mathbf{x}_{2,n}\ \cdots\ \mathbf{x}_{T,n})$, is termed the *space-time snapshot*. Now assume that we have observed data from N snapshots, $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_N]$ and each $\mathbf{x}_n$ is a $M \times 1$ complex vector with $M = ST$, which obeys a zero-mean multivariate complex Gaussian distribution $\mathbf{x}_n \sim \mathcal{CN}(\mathbf{0}, \mathbf{C}_n)$ for $n = 1, 2, \cdots, N$.

The airborne radar data mainly contains of three types of interference and noise: clutter, jamming, and thermal noise, and these three unwanted components are assumed to be mutually uncorrelated [6]. Clutter is the most complicated because it is distributed in both angle and range and is spread in Doppler frequency due to platform motion. The interference-plus-noise covariance matrices $\mathbf{C}_n$ for $n = 1, 2, \cdots, N$ are modelled (with jamming neglected) as follows [7]

$$\mathbf{C}_n = \sigma^2 \mathbf{I} + \sigma_n^2 \mathbf{R}; n = 1, 2, \cdots, N \tag{2.1}$$

where σ^2 is the power of the thermal noise, and σ_n^2 represents the range-dependent clutter power. We do not assume prior knowledge of these power parameters. They are considered as unknown real parameters in our model, which is often the case in practice. $\mathbf{I}$ is an $M \times M$ identity matrix representing the normalized thermal noise covariance matrix. $\mathbf{R}$ is a normalized clutter covariance matrix for the snapshots from all N considered ranges, which is not dependent on the range and is assumed to be known. Typically, $\mathbf{R}$ is an Toeplitz-block-Toeplitz matrix,

consisting of $T \times T$ block matrices, where each block is an $S \times S$ cross-covariance of the spatial snapshots from two pulse repetition intervals (PRI) [6].

If the N range-dependent clutter power parameters σ_n^2's are all equal to each other for all $n = 1, 2, \cdots, N$, then the interference-plus-noise covariance matrices for these N ranges data are stationary. Otherwise, there is a non-stationarity/non-homogeneity existing. The objective is to detect such a nonstationarity if exists. With the covariance matrix model shown in (2.1), the detection problem is equivalent to choose between the following hypotheses

$$\begin{aligned} \mathcal{H}_0 : &\quad \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_N^2 = \sigma_0^2, \sigma^2 = \sigma_{h0}^2; \\ \mathcal{H}_1 : &\quad \sigma_1^2, \sigma_2^2, \cdots, \sigma_N^2 \text{ are not all equal}, \sigma^2 = \sigma_{h1}^2, \end{aligned}$$

where, σ_{h0}^2 and σ_{h1}^2 are the thermal noise power under $\mathcal{H}_0$ and $\mathcal{H}_1$ respectively. Note that σ^2's are nuisance parameters for this hypothesis testing problem.

2.3 GLRT and Rao test for detecting the nonstationarity

In this section, we present two detectors for the aforementioned detection problem. The first detector is formed by the GLRT, which is widely used because of its asymptotic optimality property for large data records, and other favorable properties such as consistency and unbiasedness [14]. The second detector employs a Rao test which attains the same asymptotic (as $N \to \infty$) detection performance as the GLRT. For finite data records (finite N), the GLRT usually outperforms Rao test. However, the Rao test only requires an MLE under the null hypothesis $\mathcal{H}_0$, so its computational cost can be remarkably less than the GLRT. This can be a

desirable property in real-time STAP where the volume of radar data processed is huge. Trade-offs can be made between detection performance and computational complexity when choosing an appropriate detector.

2.3.1 GLRT for detecting the nonstationarity

Next, GLRT detector for detecting the nonstationarity of the interference-plus-noise covariance matrices is derived. It can be readily written

$$L_G(\mathbf{X}) = \frac{p(\mathbf{X}; \hat{\sigma}_1^2, \hat{\sigma}_2^2, \cdots, \hat{\sigma}_N^2, \hat{\sigma}_{h1}^2, \mathcal{H}_1)}{p(\mathbf{X}; \hat{\sigma}_0^2, \hat{\sigma}_{h0}^2, \mathcal{H}_0)} > \gamma \tag{2.2}$$

where $\hat{\sigma}_n^2$ are the MLEs of the unknown parameters σ_n^2 under $\mathcal{H}_1$ for $n = 1, 2, \cdots, N$; $\hat{\sigma}_0^2$ is the MLE of σ_0^2 under $\mathcal{H}_0$; $\hat{\sigma}_{h0}^2$ and $\hat{\sigma}_{h1}^2$ are the MLEs of σ^2 under $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively. Let

$$\begin{aligned} \hat{\mathbf{C}}_0 &= \hat{\sigma}_0^2 \mathbf{R} + \hat{\sigma}_{h0}^2 \mathbf{I} \\ \hat{\mathbf{C}}_n &= \hat{\sigma}_n^2 \mathbf{R} + \hat{\sigma}_{h1}^2 \mathbf{I}, \ n = 1, 2, \cdots, N \end{aligned}$$

Recall the assumption that $\mathbf{x}_n \sim \mathcal{CN}(\mathbf{0}, \mathbf{C}_n)$ for $n = 1, 2, \cdots, N$. Then we have

$$p(\mathbf{X}; \hat{\sigma}_1^2, \hat{\sigma}_2^2, \cdots, \hat{\sigma}_N^2, \hat{\sigma}_{h1}^2, \mathcal{H}_1) = \prod_{n=1}^{N} \frac{1}{\pi^M |\hat{\mathbf{C}}_n|} \exp\left(-\mathbf{x}_n^H \hat{\mathbf{C}}_n^{-1} \mathbf{x}_n\right)$$

and

$$p(\mathbf{X}; \hat{\sigma}_0^2, \hat{\sigma}_{h0}^2, \mathcal{H}_0) = \prod_{n=1}^{N} \frac{1}{\pi^M |\hat{\mathbf{C}}_0|} \exp\left(-\mathbf{x}_n^H \hat{\mathbf{C}}_0^{-1} \mathbf{x}_n\right),$$

where $|\cdot|$ denotes determinant. Then, the GLRT test statistic becomes

$$\begin{aligned} T_G(\mathbf{X}) &= 2 \ln L_G(\mathbf{X}) \\ &= 2 \sum_{n=1}^{N} \left[\ln\left(\frac{|\hat{\mathbf{C}}_0|}{|\hat{\mathbf{C}}_n|}\right) + \mathbf{x}_n^H \left(\hat{\mathbf{C}}_n^{-1} - \hat{\mathbf{C}}_0^{-1}\right) \mathbf{x}_n \right] > \gamma', \end{aligned} \tag{2.3}$$

Note that, due to the complexity, approximate MLEs $\hat{\sigma}_0^2, \hat{\sigma}_1^2, \cdots, \hat{\sigma}_N^2, \hat{\sigma}_{h0}^2$ and $\hat{\sigma}_{h1}^2$ must be used and are to be found in Appendix 2.5 instead of exact MLEs.

2.3.2 Rao test for detecting the nonstationarity

As shown in (2.3), the GLRT requires MLEs to be computed under both hypotheses. When the computational cost becomes a priority, it is necessary to have an alternative detector with lower computational cost and reasonably good performance. The Rao test can serve well as such a detector. The following presents the Rao test detector for detecting the nonstationarity. In order to form the Rao test, the following parameter transformation is made first. Let

$$\begin{aligned} \theta_{s1} &= \sigma_1^2 \\ \theta_1 &= \sigma_2^2 - \sigma_1^2 \\ \theta_2 &= \sigma_3^2 - \sigma_1^2 \\ &\vdots \\ \theta_{N-1} &= \sigma_N^2 - \sigma_1^2 \\ \theta_{s2} &= \sigma^2 \end{aligned} \tag{2.4}$$

Denote $\boldsymbol{\theta}_r = [\theta_1 \ \theta_2 \ \cdots \ \theta_{N-1}]^T$ which is an $(N-1) \times 1$ parameter vector for the testing problem, $\boldsymbol{\theta}_s = [\theta_{s1} \ \theta_{s2}]^T$ which is a 2×1 nuisance parameter vector, and let $\boldsymbol{\theta} = [\boldsymbol{\theta}_r^T \ \boldsymbol{\theta}_s^T]^T$ which is an $(N+1) \times 1$ vector containing all unknown parameters for the testing problem. With these notations, the testing problem is equivalent

to choosing between the following two hypotheses

$$\mathcal{H}_0 : \boldsymbol{\theta}_r = \mathbf{0}, \boldsymbol{\theta}_s$$

$$\mathcal{H}_1 : \boldsymbol{\theta}_r \neq \mathbf{0}, \boldsymbol{\theta}_s$$

The Rao test can be determined to be

$$\begin{aligned} T_R(\mathbf{X}) &= \frac{\sum_{n=2}^{N}\left[\mathrm{tr}(\mathbf{x}_n\mathbf{x}_n^H\hat{\mathbf{C}}_0^{-1}\mathbf{R}\hat{\mathbf{C}}_0^{-1}) - \mathrm{tr}(\mathbf{R}\hat{\mathbf{C}}_0^{-1})\right]^2}{\mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2)} \\ &\quad + \frac{\left[\sum_{n=2}^{N}\left(\mathrm{tr}(\mathbf{x}_n\mathbf{x}_n^H\hat{\mathbf{C}}_0^{-1}\mathbf{R}\hat{\mathbf{C}}_0^{-1}) - \mathrm{tr}(\mathbf{R}\hat{\mathbf{C}}_0^{-1})\right)\right]^2}{\mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2)} \end{aligned}$$

where, $\hat{\mathbf{C}}_0 = \hat{\sigma}_0^2\mathbf{R} + \hat{\sigma}_{h0}^2\mathbf{I}$, and $\hat{\sigma}_0^2, \hat{\sigma}_{h0}^2$ are the MLEs of σ_0^2, σ^2 under $\mathcal{H}_0$ respectively. The latters are derived in Appendix 2.5. The derivation of the Rao test is given in Appendix 2.5. Note that the Rao test shown in (2.5) only requires MLEs under $\mathcal{H}_0$.

2.4 Computer Simulations

Several computer simulations are carried out to evaluate the performances of the proposed GLRT and Rao test detectors for the nonstationarity. To have meaningful results relevant to a real scenario, we choose a normalized clutter covariance matrix $\mathbf{R}$ likely to be of practical interest. The details of choosing the $\mathbf{R}$ can be found in Appendix 2.5.

Some important parameter settings remaining the same for all simulations are listed in Table 2.1, where we set the PRF to be 4KHz, the antenna

platform speed to be 100 m/s, the radar operating wavelength to be 0.1 m, and the sensor space to be 0.1 m. Since the number of array elements is $S = 5$ and the number of pulses is $T = 10$, the corresponding clutter covariance matrix $\mathbf{R}$ is a 50×50 matrix. The modulus of the designed $\mathbf{R}$ is given in Figure 2.4.

Moreover, the thermal noise power is $\sigma^2 = 1$ under both hypotheses, the clutter power under $\mathcal{H}_0$ is $\sigma_0^2 = 5$ for all N ranges, and under $\mathcal{H}_1$ is $\sigma_n^2 = 5\alpha^{(n-1)}$ with $0 < \alpha < 1$ so the clutter power decays with range bin number n.The nonstationarity modeled is the clutter power loss with range [6]. The decaying ratio α can be viewed as an indicator of the degree of nonstationarity of the covariance matrices. Two different simulation set-ups are considered by changing the decaying ratio α parameter for comparison. In simulation 1, we let $\alpha = 0.9$. The receiver operating characteristic (ROC) curves, giving the relationship between detector's probability of detection (P_d) and probability of false alarm (P_{fa}), for both GLRT and Rao detector are shown in Figure 2.1. Both the GLRT and Rao test detectors yield "perfect" detection performance in this simulation setting. In simulation 2, for $\alpha = 0.95$, the degree of the nonstationarity of the covariance matrices becomes smaller than that of simulation 1. The ROCs for simulation 2 are given in Figure 2.2. Comparing with the ROCs in simulation 1, both the GLRT and Rao test's performances drop by certain level, with the GLRT slightly outperforming the Rao test.

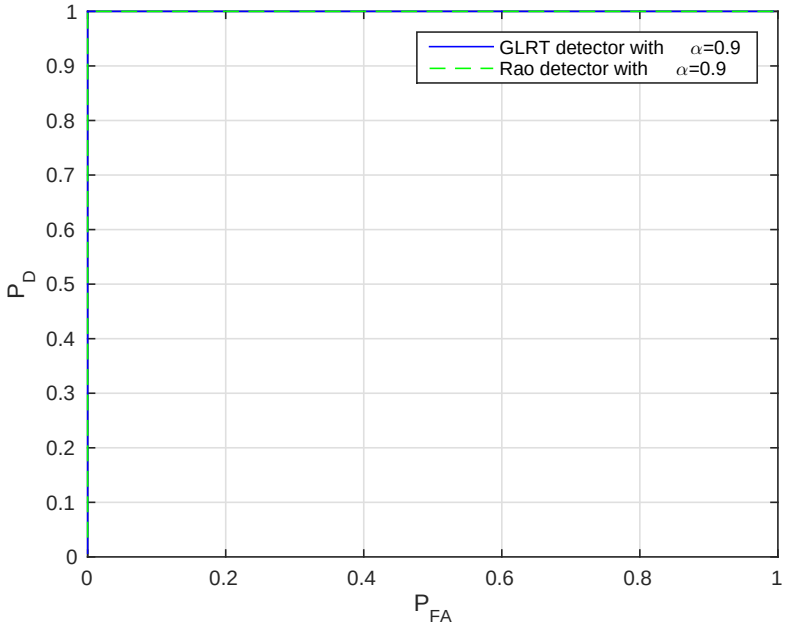

Figure 2.1. ROC curves for GLRT and Rao test detectors with $\alpha = 0.9$ in Simulation 1

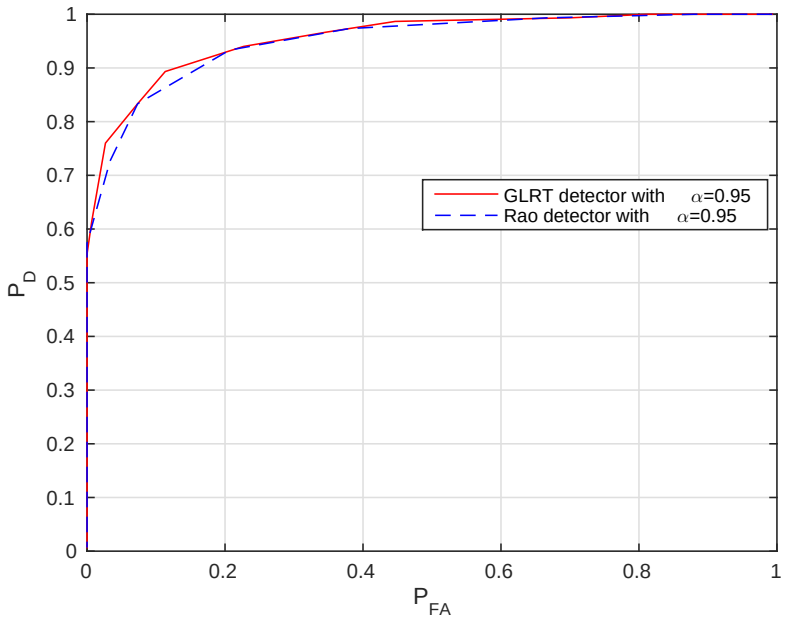

Figure 2.2. ROC curves for GLRT and Rao test detectors with $\alpha = 0.95$ in Simulation 2

Table 2.1. Simulations Parameter Setting

Parameters	Values
number of array element S	5
number of pulses per CPI T	10
number of ranges N	15
PRF f_r	4KHz
receive platform speed v	100m/s
radar operating wavelength λ	0.1m
inter-sensor spacing d	0.1m

2.5 Conclusions

In STAP, the possible heterogeneity of the training data may result in a degradation in performance and should be taken into consideration. In this paper, we presented two detectors to detect the non-homogeneity of the clutter covariance matrices, which further can be utilized to choose suitable training datasets for STAP in airborne radar signal processing scenarios. The first detector employs the GLRT; therefore, it inherits the asymptotic optimality property. The second detector is based on the Rao test with its intrinsic property of only requiring MLEs under the null hypothesis, so it can be used in a situation where the computational cost of the STAP system is of high consideration. Simulations attempting to model a practical situation are performed to test the performances of both detectors. The proposed nonstationarity detectors in this paper can be used as a pre-processing stage in STAP and by employing them, improved output SINR of STAP can be obtained.

List of References

[1] Melvin, W.L., "A STAP Overview", *IEEE AES Systems Magazine Special Tutorials Issue*, Vol. 19, No.1, Jan. 2004 , pp. 19–35

[2] L. E. Brennan and I. S. Reed, "Theory of adaptive radar", *IEEE Trans. Aero. Elec. Syst.*, Vol. AES-9, no. 2, pp. 237–252, 1973

[3] Fa, R., de Lamare, R.C., Wang, L., "Reduced-Rank STAP Schemes for Airborne Radar Based on Switched Joint Interpolation, Decimation and Filtering Algorithm" ,*IEEE Transactions on signal processing* Vol.58, no.8, 2010, pp. 4182–4194

[4] R. Chellappa and S. Theodoridis, *Academic Press Library in Signal Processing*, Vol.2, Communications and Radar Signal Processing, Academic Press, 2014, pp. 595–665.

[5] R. Klemm, *Principle of space-time adaptive processing*, IEE Press, Bodmin, UK, 2002

[6] J. Ward, "Space-time adaptive processing for airborne radar", *Tech. Rep. 1015, MIT Lincoln lab.*, Lexington, MA, Dec. 1994

[7] Nitzberg, R.,"An effect of range-heterogeneous clutter on adaptive Doppler filters", *IEEE Transactions on Aerospace and Electronic Systems*, Vol.26, No.3, May 1990, pp.475-480.

[8] Armstrong, B. C., Griffiths, H. D., Baker, C. J., and White, R. G., "Performance of adaptive optimal Doppler processors in heterogeneous clutter",*IEE Proceedings on Radar, Sonar, Navigation*, 142, 4 (Aug.1995), 179–190.

[9] Melvin, W.L.,"Space-Time Adaptive Radar Performance in Heterogeneous Clutter", *IEEE Transaction on Aerospace and Electronic Systems*, Vol.36, No.2, April 2000, pp.621–633.

[10] M. Rangaswamy, F.C. Lin and K.R. Gerlach, "Robust adaptive signal processing methods for heterogeneous radar clutter scenarios, *Proceedings of the 2003 IEEE Radar conference*, Huntsville, AL, May 2003

[11] M. Rangaswamy, "An Overview of Space-Time Adaptive Processing for Radar",*Radar Conference, 2003. Proceedings of the International*, 45–50. 2003.

[12] Guerci, J. R., Goldstein, J. S., and Reed, I. S. Optimal and adaptive reduced-rank STAP. *IEEE Transactions on Aerospace and Electronic Systems*, 36, 2 (Apr. 2000), 647–663.

[13] Fuhrmann, D. R., Application of Toeplitz estimation to adaptive beamforming and detection. *IEEE Transactions on Signal Processing*, 39, 10 (Oct.1991), 2194–2198.

[14] S. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1998.

[15] Wong,J.L., Reed,I.S., and Kaprielian,Z.A. "A Model for the Radar Echo from a Random Collection of Rotating Dipole Scatterers", *IEEE Tran. on Aero. and Elec. Sys.*, AES-3, 2, Mar. 1967, 171–178.

[16] Barlow, E.J., "Doppler radar", *Proceedings of the IRE*,37, 1949

Appendix A–Computing the MLEs $\hat{\sigma}_n^2$, $\hat{\sigma}_0^2$,$\hat{\sigma}_{h0}^2$ and $\hat{\sigma}_{h1}^2$

This section derives the approximate MLEs $\hat{\sigma}_n^2$'s for $n = 1, 2, \cdots, N$ and $\hat{\sigma}_{h1}^2$ under $\mathcal{H}_1$, and $\hat{\sigma}_0^2$, and the exact MLE $\hat{\sigma}_{h0}^2$ under $\mathcal{H}_0$.

Exact MLEs under $\mathcal{H}_0$

Under $\mathcal{H}_0$, σ_0^2 and $\sigma^2 = \sigma_{h0}^2$ are the unknown parameters. For $n = 1, 2, .., N$, we have $\mathbf{x}_n \sim \mathcal{CN}(\mathbf{0}, \mathbf{C}_0)$,where

$$\mathbf{C}_0 = \sigma_0^2 \mathbf{R} + \sigma_{h0}^2 \mathbf{I}.$$

The joint probability distribution function (pdf) can be expressed as following:

$$\begin{aligned} p(\mathbf{X}; \sigma_0^2, \sigma^2) &= \prod_{n=1}^{N} p(\mathbf{x}_n; \sigma_0^2, \sigma^2) = \prod_{n=1}^{N} \frac{1}{\pi^M |\mathbf{C}|} \exp\left[-\mathbf{x}_n^H \mathbf{C}^{-1} \mathbf{x}_n\right] \\ &= \frac{1}{\pi^{MN} |\sigma_0^2 \mathbf{R} + \sigma^2 \mathbf{I}|^N} \exp\left[-\sum_{n=1}^{N} \mathbf{x}_n^H (\sigma_0^2 \mathbf{R} + \sigma^2 \mathbf{I})^{-1} \mathbf{x}_n\right] \quad (2.5) \end{aligned}$$

Maximizing (2.5) with respect to σ_0^2 and σ_{h0}^2 is equivalent to maximizing $\ln p(\mathbf{X}; \sigma_0^2, \sigma_{h0}^2)$

$$\ln p(\mathbf{X}; \sigma_0^2, \sigma^2) = -MN \ln \pi - N \ln |\sigma_0^2 \mathbf{R} + \sigma^2 \mathbf{I}| - \sum_{n=1}^{N} \mathbf{x}_n^H (\sigma_0^2 \mathbf{R} + \sigma^2 \mathbf{I})^{-1} \mathbf{x}_n.$$

Furthermore, it is equivalent to minimizing the following function $J(\sigma_0^2, \sigma^2)$ over σ_0^2 and σ^2

$$J(\sigma_0^2, \sigma^2) = N \ln |\sigma_0^2 \mathbf{R} + \sigma^2 \mathbf{I}| + \sum_{n=1}^{N} \mathbf{x}_n^H (\sigma_0^2 \mathbf{R} + \sigma^2 \mathbf{I})^{-1} \mathbf{x}_n$$

Unfortunately, an analytical solution is not available. One solution is to carrying out a 2-dimensional grid-search over σ_0^2 and σ^2 to find the values that minimizes $J(\sigma_0^2, \sigma^2)$. As an alternative we can reduce the computational cost by employing

the transformation $\alpha = \sigma_0^2/\sigma^2$ and $\beta = \sigma^2$. Then it becomes equivalent to a grid search of $\hat{\alpha}$ which minimizes

$$J'(\alpha) = MN \ln \left(\frac{1}{NM} \sum_{n=1}^{N} \mathbf{x}_n^H (\alpha \mathbf{R} + \mathbf{I})^{-1} \mathbf{x}_n \right) + N \ln |\alpha \mathbf{R} + \mathbf{I}|$$

and

$$\hat{\beta} = \hat{\sigma}^2 = \frac{1}{NM} \sum_{n=1}^{N} \mathbf{x}_n^H (\hat{\alpha} \mathbf{R} + \mathbf{I})^{-1} \mathbf{x}_n$$

Then we have $\hat{\sigma}^2 = \hat{\beta}$ and $\hat{\sigma}_0^2 = \hat{\alpha}\hat{\beta}$. To this end, the approximate MLEs under $\mathcal{H}_0$ are found.

Approximate MLEs under $\mathcal{H}_1$

Next, the approximate MLEs $\hat{\sigma}_n^2$'s for $n = 1, 2, \cdots, N$ and $\hat{\sigma}_{h1}^2$ under $\mathcal{H}_1$ are derived. Under $\mathcal{H}_1$, $\mathbf{x}_n \sim \mathcal{CN}(\mathbf{0}, \mathbf{C}_n)$ for $n = 1, 2, \cdots, N$, where

$$\mathbf{C}_n = \sigma_n^2 \mathbf{R} + \sigma^2 \mathbf{I}.$$

Under the assumption that the $\mathbf{x}_n$'s are mutually independent, the PDF is

$$p(\mathbf{X}; \sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2, \sigma^2) = \prod_{n=1}^{N} p(\mathbf{x}_n; \sigma_n^2, \sigma^2) = \prod_{n=1}^{N} \frac{1}{\pi^M |\mathbf{C}_n|} \exp(-\mathbf{x}_n^H \mathbf{C}_n^{-1} \mathbf{x}_n)$$

Using the fact that $\mathbf{C}_n$ is Hermitian and positive definite for $n = 1, 2, \cdots, N$, we can diagonalize $\mathbf{C}_n$ by a unitary matrix $\mathbf{V}_n$ such that $\mathbf{V}_n^H \mathbf{V}_n = \mathbf{V}_n \mathbf{V}_n^H = \mathbf{I}$ and $\mathbf{V}_n^H \mathbf{C}_n \mathbf{V}_n$ is a diagonal matrix whose diagonal elements are real. Then, we have

$$\begin{aligned} \mathbf{V}_n^H \mathbf{C}_n \mathbf{V}_n &= \mathbf{V}_n^H (\sigma^2 \mathbf{I} + \sigma_n^2 \mathbf{R}) \mathbf{V}_n \\ &= \sigma^2 \mathbf{V}_n^H \mathbf{V}_n + \sigma_n^2 \mathbf{V}_n^H \mathbf{R} \mathbf{V}_n \\ &= \sigma^2 \mathbf{I} + \sigma_n^2 \mathbf{V}_n^H \mathbf{R} \mathbf{V}_n \end{aligned}$$

Since $\mathbf{V}_n^H\mathbf{C}_n\mathbf{V}_n$ and $\sigma^2\mathbf{I}$ are diagonal, $\mathbf{V}_n^H\mathbf{R}\mathbf{V}_n$ is also a diagonal matrix. Let $\mathbf{\Lambda} = \mathbf{V}_n^H\mathbf{R}\mathbf{V}_n$ which is defined as $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$. We use the linear transformation

$$\mathbf{y}_n = \mathbf{V}_n^H\mathbf{x}_n, \qquad \text{for } n = 1, 2, \cdots, N \tag{2.6}$$

Thus, each $\mathbf{y}_n$ is distributed according to

$$\begin{aligned} \mathbf{y}_n &\sim \mathcal{CN}(\mathbf{0}, \boldsymbol{V}_n^H\mathbf{C}_n\boldsymbol{V}_n) \\ &\sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I} + \sigma_n^2\mathbf{\Lambda}) \end{aligned}$$

Now, by letting $\mathbf{Y} = [\mathbf{y}_1\ \mathbf{y}_2\ \cdots\ \mathbf{y}_N]$, we can write

$$\begin{aligned} p(\mathbf{Y}; \sigma_1^2, \sigma_2^2, \cdots, \sigma_N^2, \sigma^2) &= \prod_{n=1}^{N} p(\mathbf{y}_n; \sigma_n^2, \sigma^2) \\ &= \prod_{n=1}^{N} \left(\frac{1}{\pi^M |\mathbf{V}_n^H\mathbf{C}_n\mathbf{V}_n|} \exp\left[-\mathbf{y}_n^H(\mathbf{V}_n^H\mathbf{C}_n\mathbf{V}_n)^{-1}\mathbf{y}_n\right] \right) \\ &= \prod_{n=1}^{N} \left(\frac{1}{\pi^M |\sigma^2\mathbf{I} + \sigma_n^2\mathbf{\Lambda}|} \exp\left[-\mathbf{y}_n^H(\sigma^2\mathbf{I} + \sigma_n^2\mathbf{\Lambda})^{-1}\mathbf{y}_n\right] \right) \\ &= \prod_{n=1}^{N} \left(\frac{1}{\pi^M \prod_{k=1}^{M}(\sigma^2 + \sigma_n^2\lambda_k)} \exp\left[-\sum_{k=1}^{M} \frac{|[\mathbf{y}_n]_k|^2}{\sigma^2 + \sigma_n^2\lambda_k} \right] \right) \\ &= \frac{1}{\pi^{MN} \prod_{k=1}^{M}\prod_{n=1}^{N}(\sigma^2 + \sigma_n^2\lambda_k)} \exp\left[-\sum_{k=1}^{M}\sum_{n=1}^{N} \frac{\sum_{n=1}^{N} |[\mathbf{y}_n]_k|^2}{\sigma^2 + \sigma_n^2\lambda_k} \right] \end{aligned}$$

Taking the $\ln(\cdot)$ of both sides:

$$\ln p(\mathbf{Y}; \sigma_1^2, \sigma_2^2, \cdots, \sigma_N^2, \sigma^2) = -MN\ln\pi - \sum_{n=1}^{N}\sum_{k=1}^{M} \ln(\sigma^2 + \sigma_n^2\lambda_k) - \sum_{n=1}^{N}\sum_{k=1}^{M} \frac{|[\mathbf{y}_n]_k|^2}{\sigma^2 + \sigma_n^2\lambda_k}$$

Maximizing the $\ln p(\mathbf{Y}; \sigma_1^2, \sigma_2^2, \cdots, \sigma_N^2, \sigma^2)$ is equivalent to minimizing the following

$J''(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2, \sigma^2)$:

$$J''(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2, \sigma^2) = \sum_{n=1}^{N}\sum_{k=1}^{M} \ln(\sigma^2 + \sigma_n^2\lambda_k) + \sum_{n=1}^{N}\sum_{k=1}^{M} \frac{|[\mathbf{y}_n]_k|^2}{\sigma^2 + \sigma_n^2\lambda_k} \tag{2.7}$$

To find the values of unknown parameters $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2, \sigma^2$ that minimize J'', the derivatives $\frac{\partial J''}{\partial \sigma_n^2}$ for $n = 1, 2, \cdots, N$ are formed as

$$\begin{aligned}
\frac{\partial J''}{\partial \sigma_n^2} &= \sum_{k=1}^{M}\left[\frac{\lambda_k}{\sigma^2 + \sigma_n^2\lambda_k} - \frac{\lambda_k|[\mathbf{y}_n]_k|^2}{(\sigma^2 + \sigma_n^2\lambda_k)^2}\right] \\
&= \sum_{k=1}^{M} \frac{\lambda_k(\sigma^2 + \sigma_n^2\lambda_k) - \lambda_k|[\mathbf{y}_n]_k|^2}{(\sigma^2 + \sigma_n^2\lambda_k)^2}
\end{aligned}$$

The approximation $(\sigma^2 + \sigma_n^2\lambda_k)^2 \approx (\sigma_n^2\lambda_k)^2$ for $\sigma_n^2\lambda_k >> \sigma^2$,$n = 1, 2, \cdots, N$ (i.e. high clutter-to-noise ratio) are made in order to proceed on an analytical computation. By doing so, we have the following simplification

$$\begin{aligned}
\frac{\partial J''}{\partial \sigma_n^2} &\approx \frac{1}{\sigma_n^4}\sum_{k=1}^{M} \frac{\lambda_k(\sigma^2 + \sigma_n^2\lambda_k) - \lambda_k|[\mathbf{y}_n]_k|^2}{\lambda_k^2} \\
&= \frac{1}{\sigma_n^4}\sum_{k=1}^{M}\left(\frac{\sigma^2 + \sigma_n^2\lambda_k}{\lambda_k} - \frac{|[\mathbf{y}_n]_k|^2}{\lambda_k}\right) \\
&= \frac{1}{\sigma_n^4}\sum_{k=1}^{M}\left(\sigma_n^2 - \frac{|[\mathbf{y}_n]_k|^2 - \sigma^2}{\lambda_k}\right)
\end{aligned}$$

Setting the derivative to be zero, we obtain the approximate MLEs $\hat{\sigma}_n^2$ for $n = 1, 2, \cdots, N$

$$\begin{aligned}
\frac{\partial J''}{\partial \sigma_n^2} &= \sum_{k=1}^{M}\left(\hat{\sigma}_n^2 - \frac{|[\mathbf{y}_n]_k|^2 - \sigma^2}{\lambda_k}\right) \\
&= M\hat{\sigma}_n^2 - \sum_{k=1}^{M}\frac{|[\mathbf{y}_n]_k|^2 - \sigma^2}{\lambda_k} \\
&= 0
\end{aligned}$$

Solving the equation, produce

$$\hat{\sigma}_n^2 = \frac{1}{M}\sum_{k=1}^{M}\frac{|[\mathbf{y}_n]_k|^2}{\lambda_k} - \frac{1}{M}\sum_{k=1}^{M}\frac{\sigma^2}{\lambda_k}. \tag{2.8}$$

Recalling that $\mathbf{y}_n = \mathbf{V}_n^H\mathbf{x}_n$, we have

$$\begin{aligned}\hat{\sigma}_n^2 &= \frac{1}{M}\sum_{k=1}^{M}\frac{|[\mathbf{V}_n^H\mathbf{x}_n]_k|^2}{\lambda_k} - \frac{1}{M}\sum_{k=1}^{M}\frac{\sigma^2}{\lambda_k} \\ &= \frac{1}{M}\mathbf{x}_n^H\mathbf{V}_n\mathbf{\Lambda}^{-1}\mathbf{V}_n^H\mathbf{x}_n - \frac{\sigma^2}{M}\mathrm{tr}(\mathbf{\Lambda}^{-1}),\end{aligned}$$

and with $\mathbf{\Lambda} = \mathbf{V}_n^H\mathbf{R}\mathbf{V}_n$, we have

$$\hat{\sigma}_n^2 = \frac{1}{M}\mathbf{x}_n^H\mathbf{R}^{-1}\mathbf{x}_n - \frac{\sigma^2}{M}\mathrm{tr}(\mathbf{R}^{-1}) \tag{2.9}$$

We now have the approximate MLEs $\hat{\sigma}_n^2$ for $n = 1, 2, \cdots, N$. The next step is plugging these approximate MLEs into the PDF of data set $\mathbf{X}$, so that the PDF only depends on σ^2 as follows:

$$\begin{aligned}p(\mathbf{X}; \hat{\sigma}_1^2, \hat{\sigma}_2^2, \cdots, \hat{\sigma}_N^2, \sigma^2) &= \prod_{n=1}^{N} p(\mathbf{x}_n; \hat{\sigma}_n^2, \sigma^2) \\ &= \prod_{n=1}^{N}\frac{1}{\pi^M|\hat{\mathbf{C}}_n|}\exp(-\mathbf{x}_n^H\hat{\mathbf{C}}_n^{-1}\mathbf{x}_n) \\ &= \frac{1}{\pi^{MN}}\prod_{n=1}^{N}\frac{1}{|\hat{\sigma}_n^2\mathbf{R}+\sigma^2\mathbf{I}|}\exp\left[-\mathbf{x}_n^H(\hat{\sigma}_n^2\mathbf{R}+\sigma^2\mathbf{I})^{-1}\mathbf{x}_n\right]\end{aligned}$$

Then,

$$\ln p(\mathbf{X}; \hat{\sigma}_1^2, \hat{\sigma}_2^2, \cdots, \hat{\sigma}_N^2, \sigma^2) = -MN\ln\pi - \sum_{n=1}^{N}\ln|\hat{\sigma}_n^2\mathbf{R}+\sigma^2\mathbf{I}| - \sum_{n=1}^{N}\mathbf{x}_n^H(\hat{\sigma}_n^2\mathbf{R}+\sigma^2\mathbf{I})^{-1}\mathbf{x}_n.$$

Maximizing the log-lkelihood function above over σ^2 is equivalent to minimizing the following function over σ^2

$$J_2(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \cdots, \hat{\sigma}_N^2, \sigma^2) = \sum_{n=1}^{N}\ln|\hat{\sigma}_n^2\mathbf{R}+\sigma^2\mathbf{I}| + \sum_{n=1}^{N}\mathbf{x}_n^H(\hat{\sigma}_n^2\mathbf{R}+\sigma^2\mathbf{I})^{-1}\mathbf{x}_n$$

Using (2.9), it reduces to minimizing

$$\begin{aligned} J_2'(\sigma^2) &= J(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \cdots, \hat{\sigma}_N^2, \sigma^2) \\ &= \sum_{n=1}^{N} \ln \left| \left[\frac{1}{M} \mathbf{x}_n^H \mathbf{R}^{-1} \mathbf{x}_n - \frac{\sigma^2}{M} \mathrm{tr}(\mathbf{R}^{-1}) \right] \mathbf{R} + \sigma^2 \mathbf{I} \right| \\ &+ \sum_{n=1}^{N} \mathbf{x}_n^H \left(\left[\frac{1}{M} \mathbf{x}_n^H \mathbf{R}^{-1} \mathbf{x}_n - \frac{\sigma^2}{M} \mathrm{tr}(\mathbf{R}^{-1}) \right] \mathbf{R} + \sigma^2 \mathbf{I} \right)^{-1} \mathbf{x}_n \end{aligned}$$

over σ^2. We resort it to grid search to find the value of σ^2 minimizing $J_2'(\sigma^2)$, and the minimizing value is $\hat{\sigma}^2$ under $\mathcal{H}_1$. Then substituting $\hat{\sigma}^2$ of (2.9) produces the remaining numerical MLEs $\hat{\sigma}_n^2$. Thus, all the approximate MLEs under $\mathcal{H}_1$ are found.

Appendix B–Derivation of the Rao test Detector

This section derives the Rao test. Let

$$\begin{aligned} \theta_{s1} &= \sigma_1^2 \\ \theta_1 &= \sigma_2^2 - \sigma_1^2 \\ \theta_2 &= \sigma_3^2 - \sigma_1^2 \\ &\vdots \\ \theta_{N-1} &= \sigma_N^2 - \sigma_1^2 \\ \theta_{s2} &= \sigma^2 \end{aligned} \tag{2.10}$$

Denote $\boldsymbol{\theta}_r = [\theta_1 \ \theta_2 \ \cdots \ \theta_{N-1}]^T$ which is the parameter vector for the testing problem, $\boldsymbol{\theta}_s = [\theta_{s1} \ \theta_{s2}]^T$ which is the 2×1 nuisance parameter vector, and let $\boldsymbol{\theta} = [\boldsymbol{\theta}_r^T \ \boldsymbol{\theta}_s^T]^T$, which is a $(N+1) \times 1$ vector contains all unknown parameters for the testing problem. With these notations, the testing problem is equivalent to

testing between the following two hypotheses

$$\mathcal{H}_0 : \boldsymbol{\theta}_r = \mathbf{0}, \boldsymbol{\theta}_s$$

$$\mathcal{H}_1 : \boldsymbol{\theta}_r \neq \mathbf{0}, \boldsymbol{\theta}_s$$

First,

$$\ln p(\mathbf{X}; \boldsymbol{\sigma}^2_{ns'}) = \ln p(\mathbf{X}; \boldsymbol{\theta}) = \ln \frac{1}{\pi^{NM}} - \sum_{n=1}^{N} \left[\mathbf{x}_n^H \mathbf{C}_n^{-1} \mathbf{x}_n + \ln |\mathbf{C}_n| \right], \tag{2.11}$$

where

$$\begin{aligned} \mathbf{C}_1 &= \sigma_1^2 \mathbf{R} + \sigma^2 \mathbf{I} = \theta_{s1} \mathbf{R} + \theta_{s2} \mathbf{I}, \\ \mathbf{C}_n &= \sigma_n^2 \mathbf{R} + \sigma^2 \mathbf{I} = (\theta_{n-1} + \theta_{s1}) \mathbf{R} + \theta_{s2} \mathbf{I}, n = 2, 3, \cdots, N. \end{aligned}$$

Then,

$$\begin{aligned} \frac{\partial \ln p(\mathbf{X}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \begin{bmatrix} -\sum_{n=1}^{N} \frac{\partial \left[\mathbf{x}_n^H \mathbf{C}_n^{-1} \mathbf{x}_n + \ln |\mathbf{C}_n| \right]}{\partial \theta_1} \\ -\sum_{n=1}^{N} \frac{\partial \left[\mathbf{x}_n^H \mathbf{C}_n^{-1} \mathbf{x}_n + \ln |\mathbf{C}_n| \right]}{\partial \theta_2} \\ \vdots \\ -\sum_{n=1}^{N} \frac{\partial \left[\mathbf{x}_n^H \mathbf{C}_n^{-1} \mathbf{x}_n + \ln |\mathbf{C}_n| \right]}{\partial \theta_{N-1}} \\ -\sum_{n=1}^{N} \frac{\partial \left[\mathbf{x}_n^H \mathbf{C}_n^{-1} \mathbf{x}_n + \ln |\mathbf{C}_n| \right]}{\partial \theta_{s1}} \\ -\sum_{n=1}^{N} \frac{\partial \left[\mathbf{x}_n^H \mathbf{C}_n^{-1} \mathbf{x}_n + \ln |\mathbf{C}_n| \right]}{\partial \theta_{s2}} \end{bmatrix} \\ &= \begin{bmatrix} -\frac{\partial \left[\mathbf{x}_2^H \mathbf{C}_2^{-2} \mathbf{x}_2 + \ln |\mathbf{C}_2| \right]}{\partial \theta_1} \\ -\frac{\partial \left[\mathbf{x}_3^H \mathbf{C}_3^{-3} \mathbf{x}_3 + \ln |\mathbf{C}_3| \right]}{\partial \theta_2} \\ \vdots \\ -\frac{\partial \left[\mathbf{x}_N^H \mathbf{C}_N^{-1} \mathbf{x}_N + \ln |\mathbf{C}_N| \right]}{\partial \theta_{N-1}} \\ -\sum_{n=1}^{N} \frac{\partial \left[\mathbf{x}_n^H \mathbf{C}_n^{-1} \mathbf{x}_n + \ln |\mathbf{C}_n| \right]}{\partial \theta_{s1}} \\ -\sum_{n=1}^{N} \frac{\partial \left[\mathbf{x}_n^H \mathbf{C}_n^{-1} \mathbf{x}_n + \ln |\mathbf{C}_n| \right]}{\partial \theta_{s2}} \end{bmatrix} \end{aligned} \tag{2.12}$$

Furthermore, we have for $n = 2, 3, \cdots, N$

$$\frac{\partial \ln |\mathbf{C}_n|}{\partial \theta_{n-1}} = \frac{\partial \ln |\mathbf{C}_n|}{\partial \sigma_n^2} = \operatorname{tr}\left(\mathbf{R} (\sigma_n^2 \mathbf{R} + \sigma^2 \mathbf{I})^{-1} \right) = \operatorname{tr}\left(\mathbf{R} \mathbf{C}_n^{-1} \right),$$

$$\frac{\partial \mathbf{x}_n^H \mathbf{C}_n^{-1} \mathbf{x}_n}{\partial \theta_{n-1}} = \frac{\partial \mathbf{x}_n^H \mathbf{C}_n^{-1} \mathbf{x}_n}{\partial \sigma_n^2} \quad = \mathrm{tr}(\mathbf{x}_n \mathbf{x}_n^H \frac{\partial \mathbf{C}_n^{-1}}{\partial \sigma_n^2}) = -\mathrm{tr}(\mathbf{x}_n \mathbf{x}_n^H \mathbf{C}_n^{-1} \mathbf{R} \mathbf{C}_n^{-1});$$

And also for $n = 1, 2, \cdots, N$

$$\frac{\partial \ln |\mathbf{C}_n|}{\partial \theta_{s1}} = \mathrm{tr}\left(\mathbf{R}\mathbf{C}_n^{-1}\right),$$

$$\frac{\partial \mathbf{x}_n^H \mathbf{C}_n^{-1} \mathbf{x}_n}{\partial \theta_{s1}} = \mathrm{tr}(\mathbf{x}_n \mathbf{x}_n^H \frac{\partial \mathbf{C}_n^{-1}}{\partial \theta_{s1}}) \quad = -\mathrm{tr}(\mathbf{x}_n \mathbf{x}_n^H \mathbf{C}_n^{-1} \mathbf{R} \mathbf{C}_n^{-1});$$

$$\frac{\partial \ln |\mathbf{C}_n|}{\partial \theta_{s2}} = \mathrm{tr}\left(\mathbf{C}_n^{-1}\right),$$

$$\frac{\partial \mathbf{x}_n^H \mathbf{C}_n^{-1} \mathbf{x}_n}{\partial \theta_{s2}} = \mathrm{tr}(\mathbf{x}_n \mathbf{x}_n^H \frac{\partial \mathbf{C}_n^{-1}}{\partial \theta_{s2}}) \quad = -\mathrm{tr}(\mathbf{x}_n \mathbf{x}_n^H \mathbf{C}_n^{-1} \mathbf{C}_n^{-1}).$$

Therefore,

$$\frac{\partial \ln p(\mathbf{X}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \mathrm{tr}(\mathbf{x}_2 \mathbf{x}_2^H \mathbf{C}_2^{-1} \mathbf{R} \mathbf{C}_2^{-1}) - \mathrm{tr}\left(\mathbf{R}\mathbf{C}_2^{-1}\right) \\ \mathrm{tr}(\mathbf{x}_3 \mathbf{x}_3^H \mathbf{C}_3^{-1} \mathbf{R} \mathbf{C}_3^{-1}) - \mathrm{tr}\left(\mathbf{R}\mathbf{C}_3^{-1}\right) \\ \vdots \\ \mathrm{tr}(\mathbf{x}_N \mathbf{x}_N^H \mathbf{C}_N^{-1} \mathbf{R} \mathbf{C}_N^{-1}) - \mathrm{tr}\left(\mathbf{R}\mathbf{C}_N^{-1}\right) \\ \sum_{n=1}^{N} \left[\mathrm{tr}(\mathbf{x}_n \mathbf{x}_n^H \mathbf{C}_n^{-1} \mathbf{R} \mathbf{C}_n^{-1}) - \mathrm{tr}\left(\mathbf{R}\mathbf{C}_n^{-1}\right)\right] \\ \sum_{n=1}^{N} \left[\mathrm{tr}(\mathbf{x}_n \mathbf{x}_n^H \mathbf{C}_n^{-1} \mathbf{C}_n^{-1}) - \mathrm{tr}\left(\mathbf{C}_n^{-1}\right)\right] \end{bmatrix} \tag{2.13}$$

Next, we are to compute the Fisher Information Matrix (FIM) $\mathbf{I}(\boldsymbol{\theta})$.

$$\begin{aligned} \mathbf{I}(\boldsymbol{\theta}) &= -E\left(\frac{\partial^2 \ln p(\mathbf{X}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right) \\ &= \begin{bmatrix} \mathbf{I}_{rr} & \mathbf{I}_{rs} \\ \mathbf{I}_{sr} & \mathbf{I}_{ss} \end{bmatrix} \\ &= \begin{bmatrix} (N-1) \times (N-1) & (N-1) \times 2 \\ 2 \times (N-1) & 2 \times 2 \end{bmatrix} \end{aligned} \tag{2.14}$$

Notice that the block $\mathbf{I}_{rr}$ is a diagonal matrix, given that $\mathbf{C}_n$ is not dependent on θ_{m-1} when $n \neq m$ for $n, m = 2, 3, \cdots, N$.

And, for $n = 2, 3, \cdots, N$ we have

$$\begin{aligned}
[\mathbf{I}_{rr}]_{n-1,n-1} &= -E\left(\frac{\partial\left[\mathrm{tr}(\mathbf{x}_n\mathbf{x}_n^H\mathbf{C}_n^{-1}\mathbf{R}\mathbf{C}_n^{-1}) - \mathrm{tr}(\mathbf{R}\mathbf{C}_n^{-1})\right]}{\partial\theta_{n-1}}\right) \\
&= 2E\left(\mathrm{tr}(\mathbf{x}_n\mathbf{x}_n^H\mathbf{C}_n^{-1}\mathbf{R}\mathbf{C}_n^{-1}\mathbf{R}\mathbf{C}_n^{-1})\right) - \mathrm{tr}\left(\mathbf{R}\mathbf{C}_n^{-1}\mathbf{R}\mathbf{C}_n^{-1}\right) \\
&= 2\mathrm{tr}(\mathbf{C}_n\mathbf{C}_n^{-1}\mathbf{R}\mathbf{C}_n^{-1}\mathbf{R}\mathbf{C}_n^{-1}) - \mathrm{tr}\left(\mathbf{R}\mathbf{C}_n^{-1}\mathbf{R}\mathbf{C}_n^{-1}\right) \\
&= \mathrm{tr}\left(\mathbf{R}\mathbf{C}_n^{-1}\mathbf{R}\mathbf{C}_n^{-1}\right),
\end{aligned} \tag{2.15}$$

and

$$\begin{aligned}
[\mathbf{I}_{rs}]_{n-1,1} &= -E\left(\frac{\partial\left[\mathrm{tr}(\mathbf{x}_n\mathbf{x}_n^H\mathbf{C}_n^{-1}\mathbf{R}\mathbf{C}_n^{-1}) - \mathrm{tr}(\mathbf{R}\mathbf{C}_n^{-1})\right]}{\partial\theta_{s1}}\right) \\
&= 2\mathrm{tr}(\mathbf{C}_n\mathbf{C}_n^{-1}\mathbf{R}\mathbf{C}_n^{-1}\mathbf{R}\mathbf{C}_n^{-1}) - \mathrm{tr}\left(\mathbf{R}\mathbf{C}_n^{-1}\mathbf{R}\mathbf{C}_n^{-1}\right) \\
&= \mathrm{tr}\left(\mathbf{R}\mathbf{C}_n^{-1}\mathbf{R}\mathbf{C}_n^{-1}\right),
\end{aligned} \tag{2.16}$$

$$\begin{aligned}
[\mathbf{I}_{rs}]_{n-1,2} &= -E\left(\frac{\partial\left[\mathrm{tr}(\mathbf{x}_n\mathbf{x}_n^H\mathbf{C}_n^{-1}\mathbf{R}\mathbf{C}_n^{-1}) - \mathrm{tr}(\mathbf{R}\mathbf{C}_n^{-1})\right]}{\partial\theta_{s2}}\right) \\
&= 2\mathrm{tr}(\mathbf{C}_n^{-1}\mathbf{R}\mathbf{C}_n^{-1}) - \mathrm{tr}\left(\mathbf{R}\mathbf{C}_n^{-1}\mathbf{C}_n^{-1}\right) \\
&= \mathrm{tr}\left(\mathbf{C}_n^{-1}\mathbf{R}\mathbf{C}_n^{-1}\right).
\end{aligned} \tag{2.17}$$

With a similar calculation procedure, we have

$$[\mathbf{I}_{ss}]_{1,1} = \sum_{n=1}^{N}\mathrm{tr}\left(\mathbf{R}\mathbf{C}_n^{-1}\mathbf{R}\mathbf{C}_n^{-1}\right) \tag{2.18}$$

$$[\mathbf{I}_{ss}]_{1,2} = \sum_{n=1}^{N}\mathrm{tr}\left(\mathbf{R}\mathbf{C}_n^{-1}\mathbf{C}_n^{-1}\right) \tag{2.19}$$

and

$$[\mathbf{I}_{ss}]_{2,2} = \sum_{n=1}^{N}\mathrm{tr}\left(\mathbf{C}_n^{-1}\mathbf{C}_n^{-1}\right) \tag{2.20}$$

With the property that FIM is a symmetric matrix, we already have it as:

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{bmatrix} \mathrm{tr}\left((\mathbf{R}\mathbf{C}_2^{-1})^2\right) & 0 & \cdots & \mathrm{tr}\left((\mathbf{R}\mathbf{C}_2^{-1})^2\right) & \mathrm{tr}\left(\mathbf{R}\mathbf{C}_2^{-2}\right) \\ 0 & \mathrm{tr}\left((\mathbf{R}\mathbf{C}_3^{-1})^2\right) & \cdots & \mathrm{tr}\left((\mathbf{R}\mathbf{C}_3^{-1})^2\right) & \mathrm{tr}\left(\mathbf{R}\mathbf{C}_3^{-2}\right) \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \mathrm{tr}\left((\mathbf{R}\mathbf{C}_2^{-1})^2\right) & \mathrm{tr}\left((\mathbf{R}\mathbf{C}_3^{-1})^2\right) & \cdots & \sum_{n=1}^{N}\mathrm{tr}\left((\mathbf{R}\mathbf{C}_n^{-1})^2\right) & \sum_{n=1}^{N}\mathrm{tr}\left(\mathbf{R}\mathbf{C}_n^{-2}\right) \\ \mathrm{tr}\left(\mathbf{R}\mathbf{C}_2^{-2}\right) & \mathrm{tr}\left(\mathbf{R}\mathbf{C}_3^{-2}\right) & \cdots & \sum_{n=1}^{N}\mathrm{tr}\left(\mathbf{R}\mathbf{C}_n^{-2}\right) & \sum_{n=1}^{N}\mathrm{tr}\left(\mathbf{C}_n^{-2}\right) \end{bmatrix} \tag{2.21}$$

Under $\mathcal{H}_0$, where $\mathbf{C}_1 = \mathbf{C}_2 = \cdots = \mathbf{C}_N = \hat{\mathbf{C}}_0 = \hat{\theta}_{s1}\mathbf{R} + \hat{\theta}_{s2}\mathbf{I} = \hat{\sigma}_0^2\mathbf{R} + \hat{\sigma}_{h0}^2\mathbf{I}$, with $\hat{\sigma}_0^2$ and $\hat{\sigma}_{h0}^2$ denoting the MLEs of σ_0^2 and σ^2 under $\mathcal{H}_0$ respectively, the FIM reduces to

$$\mathbf{I}(\boldsymbol{\theta})|_{\mathcal{H}_0} = \begin{bmatrix} \mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2) & 0 & \cdots & \mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2) & \mathrm{tr}(\mathbf{R}\hat{\mathbf{C}}_0^{-2}) \\ 0 & \mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2) & \cdots & \mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2) & \mathrm{tr}(\mathbf{R}\hat{\mathbf{C}}_0^{-2}) \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2) & \mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2) & \cdots & N\mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2) & N\mathrm{tr}(\mathbf{R}\hat{\mathbf{C}}_0^{-2}) \\ \mathrm{tr}(\mathbf{R}\hat{\mathbf{C}}_0^{-2}) & \mathrm{tr}(\mathbf{R}\hat{\mathbf{C}}_0^{-2}) & \cdots & N\mathrm{tr}(\mathbf{R}\hat{\mathbf{C}}_0^{-2}) & N\mathrm{tr}(\hat{\mathbf{C}}_0^{-2}) \end{bmatrix} \tag{2.22}$$

$$= \begin{bmatrix} \mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2)\mathbf{I}_{(N-1)\times(N-1)} & \mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2)\mathbf{1}_{(N-1)\times 1} & \mathrm{tr}(\mathbf{R}\hat{\mathbf{C}}_0^{-2})\mathbf{1}_{(N-1)\times 1} \\ \mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2)\mathbf{1}_{1\times(N-1)} & N\mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2) & N\mathrm{tr}(\mathbf{R}\hat{\mathbf{C}}_0^{-2}) \\ \mathrm{tr}(\mathbf{R}\hat{\mathbf{C}}_0^{-2})\mathbf{1}_{1\times(N-1)} & N\mathrm{tr}(\mathbf{R}\hat{\mathbf{C}}_0^{-2}) & N\mathrm{tr}(\hat{\mathbf{C}}_0^{-2}) \end{bmatrix} \tag{2.23}$$

In the Rao test, we treat $\boldsymbol{\theta}_s$ as a nuisance parameter, so we calculate the FIM

of the testing parameter vector $\boldsymbol{\theta}_r$ as follows.

$$
\begin{aligned}
[\mathbf{I}(\boldsymbol{\theta})]_{\boldsymbol{\theta}_r\boldsymbol{\theta}_r}\Big|_{\mathcal{H}_0} &= \left[\mathbf{I}_{rr} - \mathbf{I}_{rs}\mathbf{I}_{ss}^{-1}\mathbf{I}_{sr}\right]\big|_{\mathcal{H}_0} \qquad (2.24)\\
&= \mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2)\mathbf{I}_{(N-1)\times(N-1)}\\
&\quad - \begin{bmatrix} \mathrm{tr}\left((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2\right) & \mathrm{tr}\left(\mathbf{R}\hat{\mathbf{C}}_0^{-2}\right) \\ \mathrm{tr}\left((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2\right) & \mathrm{tr}\left(\mathbf{R}\hat{\mathbf{C}}_0^{-2}\right) \\ \vdots & \\ \mathrm{tr}\left((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2\right) & \mathrm{tr}\left(\mathbf{R}\hat{\mathbf{C}}_0^{-2}\right) \end{bmatrix} \cdot \frac{1}{N} \begin{bmatrix} \mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2) & \mathrm{tr}(\mathbf{R}\hat{\mathbf{C}}_0^{-2}) \\ \mathrm{tr}(\mathbf{R}\hat{\mathbf{C}}_0^{-2}) & \mathrm{tr}(\hat{\mathbf{C}}_0^{-2}) \end{bmatrix}^{-1}\\
&\quad \cdot \begin{bmatrix} \mathrm{tr}\left((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2\right) & \mathrm{tr}\left(\mathbf{R}\hat{\mathbf{C}}_0^{-2}\right) \\ \mathrm{tr}\left((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2\right) & \mathrm{tr}\left(\mathbf{R}\hat{\mathbf{C}}_0^{-2}\right) \\ \vdots & \\ \mathrm{tr}\left((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2\right) & \mathrm{tr}\left(\mathbf{R}\hat{\mathbf{C}}_0^{-2}\right) \end{bmatrix}^{T}
\end{aligned}
$$

To compute the inverse of FIM $[\mathbf{I}^{-1}(\boldsymbol{\theta})]_{\boldsymbol{\theta}_r\boldsymbol{\theta}_r}\Big|_{\mathcal{H}_0}$ we apply Woodbury's identity

$$
\left(\mathbf{I}_{rr} - \mathbf{I}_{rs}\mathbf{I}_{ss}^{-1}\mathbf{I}_{sr}\right)^{-1}\Big|_{\mathcal{H}_0} = \left[\mathbf{I}_{rr}^{-1} + \mathbf{I}_{rr}^{-1}\mathbf{I}_{rs}(\mathbf{I}_{ss} - \mathbf{I}_{sr}\mathbf{I}_{rr}^{-1}\mathbf{I}_{rs})^{-1}\mathbf{I}_{sr}\mathbf{I}_{rr}^{-1}\right]\big|_{\mathcal{H}_0}
$$

It can be shown that

$$
(\mathbf{I}_{ss} - \mathbf{I}_{sr}\mathbf{I}_{rr}^{-1}\mathbf{I}_{rs})\big|_{\mathcal{H}_0} = \begin{bmatrix} \mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2) & \mathrm{tr}(\mathbf{R}\hat{\mathbf{C}}_0^{-2}) \\ \mathrm{tr}(\mathbf{R}\hat{\mathbf{C}}_0^{-2}) & \frac{N\mathrm{tr}(\hat{\mathbf{C}}_0^{-2})\mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2) - (N-1)\mathrm{tr}^2(\mathbf{R}\hat{\mathbf{C}}_0^{-2})}{\mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2)} \end{bmatrix}
$$

and then

$$
\begin{aligned}
(\mathbf{I}_{ss} - \mathbf{I}_{sr}\mathbf{I}_{rr}^{-1}\mathbf{I}_{rs})^{-1}\big|_{\mathcal{H}_0} &= \frac{1}{N\mathrm{tr}(\hat{\mathbf{C}}_0^{-2})\mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2) - N\mathrm{tr}^2(\mathbf{R}\hat{\mathbf{C}}_0^{-2})}\\
&\quad \cdot \begin{bmatrix} \frac{N\mathrm{tr}(\hat{\mathbf{C}}_0^{-2})\mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2) - (N-1)\mathrm{tr}^2(\mathbf{R}\hat{\mathbf{C}}_0^{-2})}{\mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2)} & -\mathrm{tr}(\mathbf{R}\hat{\mathbf{C}}_0^{-2}) \\ -\mathrm{tr}(\mathbf{R}\hat{\mathbf{C}}_0^{-2}) & \mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2) \end{bmatrix}
\end{aligned}
$$

It can be shown that

$$
\mathbf{I}_{rs}(\mathbf{I}_{ss} - \mathbf{I}_{sr}\mathbf{I}_{rr}^{-1}\mathbf{I}_{rs})^{-1}\mathbf{I}_{sr}\big|_{\mathcal{H}_0} = \mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2) \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \cdots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}_{(N-1)\times(N-1)}
$$

$$\begin{aligned}\left(\mathbf{I}_{rr}-\mathbf{I}_{rs}\mathbf{I}_{ss}^{-1}\mathbf{I}_{sr}\right)^{-1}\Big|_{\mathcal{H}_0} &= \left[\mathbf{I}_{rr}^{-1}+\mathbf{I}_{rr}^{-1}\mathbf{I}_{rs}(\mathbf{I}_{ss}-\mathbf{I}_{sr}\mathbf{I}_{rr}^{-1}\mathbf{I}_{rs})^{-1}\mathbf{I}_{sr}\mathbf{I}_{rr}^{-1}\right]\Big|_{\mathcal{H}_0} \\ &= \frac{1}{\mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2)}\left[\begin{bmatrix}1&0&\cdots&0\\0&1&\cdots&0\\\vdots&\vdots&\cdots&\vdots\\0&0&\cdots&1\end{bmatrix}+\begin{bmatrix}1&1&\cdots&1\\1&1&\cdots&1\\\vdots&\vdots&\cdots&\vdots\\1&1&\cdots&1\end{bmatrix}\right]\end{aligned} \tag{2.25}$$

Finally, the Rao test can be formulated as follows

$$T_R(\mathbf{X}) = \left[\frac{\partial \ln p(\mathbf{X};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_r}^T \left[\mathbf{I}^{-1}(\boldsymbol{\theta})\right]_{\boldsymbol{\theta}_r\boldsymbol{\theta}_r} \frac{\partial \ln p(\mathbf{X};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_r}\right]\Big|_{\mathcal{H}_0} \tag{2.26}$$

$$\begin{aligned}&= \begin{bmatrix}\mathrm{tr}(\mathbf{x}_2\mathbf{x}_2^H\hat{\mathbf{C}}_0^{-1}\mathbf{R}\hat{\mathbf{C}}_0^{-1})-\mathrm{tr}\left(\mathbf{R}\hat{\mathbf{C}}_0^{-1}\right)\\ \mathrm{tr}(\mathbf{x}_3\mathbf{x}_3^H\hat{\mathbf{C}}_0^{-1}\mathbf{R}\hat{\mathbf{C}}_0^{-1})-\mathrm{tr}\left(\mathbf{R}\hat{\mathbf{C}}_0^{-1}\right)\\ \vdots\\ \mathrm{tr}(\mathbf{x}_N\mathbf{x}_N^H\hat{\mathbf{C}}_0^{-1}\mathbf{R}\hat{\mathbf{C}}_0^{-1})-\mathrm{tr}\left(\mathbf{R}\hat{\mathbf{C}}_0^{-1}\right)\end{bmatrix}^T \\ &\quad \cdot\frac{1}{\mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2)}\left[\begin{bmatrix}1&0&\cdots&0\\0&1&\cdots&0\\\vdots&\vdots&\cdots&\vdots\\0&0&\cdots&1\end{bmatrix}+\begin{bmatrix}1&1&\cdots&1\\1&1&\cdots&1\\\vdots&\vdots&\cdots&\vdots\\1&1&\cdots&1\end{bmatrix}\right] \\ &\quad \cdot\begin{bmatrix}\mathrm{tr}(\mathbf{x}_2\mathbf{x}_2^H\hat{\mathbf{C}}_0^{-1}\mathbf{R}\hat{\mathbf{C}}_0^{-1})-\mathrm{tr}\left(\mathbf{R}\hat{\mathbf{C}}_0^{-1}\right)\\ \mathrm{tr}(\mathbf{x}_3\mathbf{x}_3^H\hat{\mathbf{C}}_0^{-1}\mathbf{R}\hat{\mathbf{C}}_0^{-1})-\mathrm{tr}\left(\mathbf{R}\hat{\mathbf{C}}_0^{-1}\right)\\ \vdots\\ \mathrm{tr}(\mathbf{x}_N\mathbf{x}_N^H\hat{\mathbf{C}}_0^{-1}\mathbf{R}\hat{\mathbf{C}}_0^{-1})-\mathrm{tr}\left(\mathbf{R}\hat{\mathbf{C}}_0^{-1}\right)\end{bmatrix} \\ &= \frac{\sum_{n=2}^N\left[\mathrm{tr}(\mathbf{x}_n\mathbf{x}_n^H\hat{\mathbf{C}}_0^{-1}\mathbf{R}\hat{\mathbf{C}}_0^{-1})-\mathrm{tr}(\mathbf{R}\hat{\mathbf{C}}_0^{-1})\right]^2}{\mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2)} \\ &\quad +\frac{\left[\sum_{n=2}^N\left(\mathrm{tr}(\mathbf{x}_n\mathbf{x}_n^H\hat{\mathbf{C}}_0^{-1}\mathbf{R}\hat{\mathbf{C}}_0^{-1})-\mathrm{tr}(\mathbf{R}\hat{\mathbf{C}}_0^{-1})\right)\right]^2}{\mathrm{tr}((\mathbf{R}\hat{\mathbf{C}}_0^{-1})^2)}\end{aligned} \tag{2.27}$$

where, $\hat{\mathbf{C}}_0 = \hat{\sigma}_0^2\mathbf{R} + \hat{\sigma}_{h0}^2\mathbf{I}$, and $\hat{\sigma}_0^2, \hat{\sigma}_{h0}^2$ are MLE of the σ_0^2, σ^2 under $\mathcal{H}_0$ respectively, which are derived in Appendix A.

Appendix C–Design of normalized covariance matrix R

This section presents in detail the choice of the normalized covariance matrix **R** used in the simulations. The clutter covariance matrix **R** can be modeled as Toeplitz-block-Toeplitz structure assuming that the clutter patches are mutually uncorrelated and the antenna is an ideal uniform linear array [6]. It can be briefly summarized as follows. The clutter component of a certain space-time snapshot from n^{th} range cell can be expressed as [6]

$$\mathbf{x}_c = \sum_{k=1}^{N_c} \beta_k \mathbf{s}(\bar{w}_k, \bar{v}_k)$$

where β_k is the random amplitude from the k^{th} clutter patch, N_c is the number of independent clutter patches evenly distributed in a certain range, and $\mathbf{s}(\bar{w}_k, \bar{v}_k)$ is the spatial-temporal steering vector at k^{th} clutter patch

$$\mathbf{s}(\bar{w}_k, \bar{v}_k) = \mathbf{b}(\bar{w}_k) \otimes \mathbf{a}(\bar{v}_k)$$

where

$$\mathbf{b}(\bar{w}_k) = [1 \; e^{j2\pi\bar{w}_k} \; \cdots \; e^{j(S-1)\pi\bar{w}_k}]^T$$

is the $S \times 1$ temporal steering vector with normalized Doppler frequency $\bar{w}_k$ and

$$\mathbf{a}(\bar{v}_k) = [1 \; e^{j2\pi\bar{v}_k} \; \cdots \; e^{j(T-1)\pi\bar{v}_k}]^T$$

represents $T \times 1$ the spatial steering vector at normalized spatial frequency $\bar{v}_k$, and the operator $\otimes$ is the Kronecker product. Also

$$\bar{w}_k = \frac{2v}{f_r \lambda} \sin\theta_k \cos\phi_k$$

$$\bar{v}_k = \frac{d}{\lambda} \sin\theta_k \cos\phi_k$$

where, f_r is the PRF, λ is the radar operating wavelength, and d is the inter-sensor distance, the values used can be found in Table 2.1, also θ_k and ϕ_k are the azimuth and elevation of k^{th} clutter patch respectively. Then the clutter covariance can be expressed as

$$\begin{aligned} \mathbf{R} &= E\{\mathbf{x}_c \mathbf{x}_c^H\} \\ &= \sigma^{2'} \sum_{k=1}^{N_c} \epsilon_k [\mathbf{b}(\bar{w}_k)\mathbf{b}(\bar{w}_k)^H] \otimes [\mathbf{a}(\bar{v}_k)\mathbf{a}(\bar{v}_k)^H], \end{aligned} \tag{2.28}$$

where $\sigma^{2'}$ and ϵ_k are constants. Thus, the clutter covariance matrix $\mathbf{R}$ is an $S \times S$ block matrix, and each block is a $T \times T$ cross-covariance of the spatial snapshots from two pulsess, which is Toeplitz. Thus, the $\mathbf{R}$ is of Toeplitz-block-Toeplitz strucure.

In our simulation, the $\mathbf{R}$ is formed by the inverse of a two-dimensional Fourier transform of an angle-Doppler power spectral density (PSD). In general, the *clutter ridge*,which is the locus of the PSD distribution, may span a portion of the Doppler space, or the whole Doppler space, depending on the platform velocity, the operating wavelength, and the radar pulse repetition frequency (PRF) [6]. Also, the Doppler spectrum of the ground clutter can be modeled as Gaussian model, as reported in [15] and [16]. Therefore, the angle-Doppler PSD of the clutter is modelled as shown in Figure 2.3., which is a two dimensional Gaussian distribution along the clutter ridge.

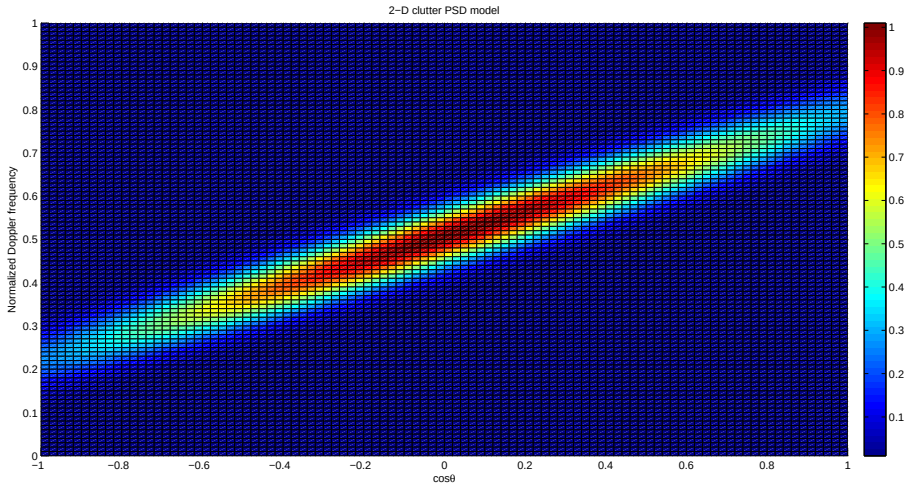

Figure 2.3. Angle-Doppler Power Spectral Density of the normalized clutter covariance $\mathbf{R}$

With the parameter settings in Table 2.1, we have $\beta = 0.25$, the slope of the clutter line, representing the number of half-interelement spacings traversed by the platform during one PRI. The two-dimensional inverse Fourier transform of the clutter angle-Doppler PSD results in the the clutter covariance matrix $\mathbf{R}$, whose modulus is plotted in Figure 2.4.

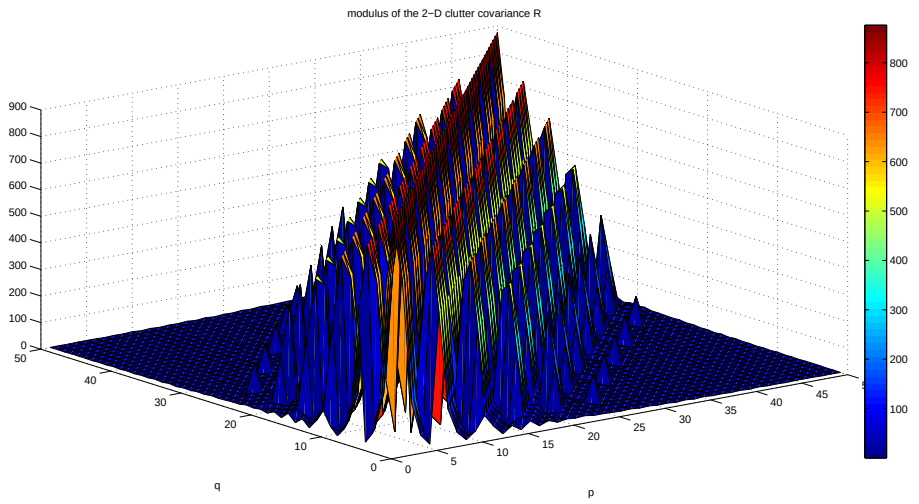

Figure 2.4. Modulus of the normalized Clutter Covariance Matrix **R**

MANUSCRIPT 3

The Complex Parameter Rao Test

by

Steven Kay and Zhenghan Zhu

published in

IEEE Transactions on Signal Processing, vol. 64, no. 24, pp.6580–6588, Dec. 2016.

Abstract

The Rao test is an important method in signal detection in the presence of unknown parameters. The traditional approach to the problem when the unknown parameters are complex-valued is to form a corresponding real-valued parameter vector and then to use the real Rao test. Alternatively, we present a complex parameter Rao test by reformulating the calculations with respect to the complex-valued quantities directly. Two important examples of the application of the complex parameter Rao test are given to illustrate the procedure.

3.1 Introduction

The Rao test is an important and useful tool in signal detection. It is asymptotically equivalent to the generalized likelihood ratio test (GLRT). However, it does not require the maximum likelihood estimates (MLEs) of testing parameters and only needs the MLEs of nuisance parameters (if present) under the null hypothesis [1, pp. 209-217]. Therefore, it often has lower computational complexity and is much easier to use in practice. These properties of the Rao test are advantageous and desirable especially in applications when computational cost is of concern. As an example, compared with GLRT, the Rao test reduces computational cost substantially, but maintains good performance in detecting the nonstationarity of radar signal in space-time adaptive processing [4]. Moreover, the Rao test is employed to form an adaptive detector for testing the presence of a deterministic signal with unknown parameter in noise of unknown autoregressive parameterized spectra [5]. It has been shown that under several detection problems commonly

encountered in practice such as the detection of subspace signals in the presence of subspace interference and Gaussian noise with unknown covariance, the Rao test is statistically equivalent to the GLRT and Wald test [6]. When the unknown parameters are real-valued, the Rao test is well-known and can be found in [1, pp. 209-217].

However, in many applications, such as radar [18] and sonar [19], the data and unknown parameters are both complex-valued. Traditionally, to find the Rao test for such problems requires one to form real-valued vectors and substitute them into the Rao test. The procedure of this straightforward approach can be found in several literature with practical applications. For example, it is used in distributed target detection in compound-Gaussian noise [14], in testing a target in partially homogeneous environment [15] and in testing a signal in homogeneous environment when the covariance matrix is unknown [16]. As an alternative and hopefully more insightful procedure, a more natural method is presented that formulates the Rao test for complex-valued data and parameters, termed the *complex parameter Rao test*. A similar attempt can be found in [3]; however, the extension of the Rao test to complex parameters given there is only valid (if ignoring a factor of two) under the special condition of the real Fisher information matrix (FIM) of the unknown parameter having a special form. Note that even under the special condition, the Rao test statistic given in [3] is incorrect by a factor of two. This can cause problems if the usual asymptotic statistics for the Rao test [1, pp. 473-526] are used to compute the probabilities of false alarm and detection.

The complex parameter Rao test can be applied to many applications. For instance, the Rao test for complex parameters derived in [3] have been applied to detect a subspace signal in colored noise with unknown covariance matrix in [7],to detect a distributed target in interference and noise with unknown covariance matrix in [8]. In addition, the complex parameter Rao test derived in this paper has been used to test the bandedness of a complex-valued covariance matrix [17].

The paper is organized as follows. Section 3.2 describes the standard real parameter and real data Rao test. The extension to the complex data and complex parameters is given in Section 3.3. A simplified version, which is valid when the FIM satisfies certain conditions is derived in Section 3.4. In Section 3.5, some important practical examples, such as complex linear model and autoregressive model, are used to illustrate the theorems. Finally, conclusions are given in Section 3.6.

Notation: Scalar quantities are denoted by lower-case symbols. Vectors are denoted by boldface lowercase symbols. The matrices are denoted by boldface uppercase symbols (except $\tilde{\boldsymbol{\Theta}}$, which is a vector). All complex-valued quantities are labeled with "tilde" while real-valued quantities are not. The symbols T, H and $*$ denote transpose, Hermitian and complex conjugate respectively. A lower-case letter with footnote such as b_i and b_{ij} denote the i^{th} element of a vector $\mathbf{b}$ and ij^{th} element of a matrix $\mathbf{B}$ respectively. The symbol "$|\cdot|$" represents the modulus of a complex scalar. The symbol $E(\cdot)$ denotes expectation of a random quantity. vec$(\cdot)$ represents the vectorization of a matrix. $j = \sqrt{-1}$. Lastly, $\otimes$

denotes Kronecker product.

3.2 Real Vector Parameter Rao Test

We first set up some notations that will be useful in the formulation of the complex parameter Rao test. Then, we summarize the usual real vector approach to implementing the test. Suppose we observe a complex data vector $\tilde{\mathbf{x}} = \mathbf{u} + j\mathbf{v} \in \mathbb{C}^{N\times 1}$. We form the real data vector as $\mathbf{x} = [\mathbf{u}^T \ \mathbf{v}^T]^T \in \mathbb{R}^{2N\times 1}$. Similarly, assume the probability density function (PDF) depends upon the unknown parameters $\tilde{\boldsymbol{\theta}} = \boldsymbol{\alpha} + j\boldsymbol{\beta}$, where $\boldsymbol{\alpha} \in \mathbb{R}^{p\times 1}$, $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ and $\tilde{\boldsymbol{\theta}} \in \mathbb{C}^{p\times 1}$. We denote $\boldsymbol{\xi} = [\boldsymbol{\alpha}^T \boldsymbol{\beta}^T]^T$, so $\boldsymbol{\xi} \in \mathbb{R}^{2p\times 1}$. Note that we can represent the PDF as $p_{\mathbf{x}}(\mathbf{x};\boldsymbol{\xi})$ or equivalently as $p_{\tilde{\mathbf{x}},\tilde{\mathbf{x}}^*}(\tilde{\mathbf{x}},\tilde{\mathbf{x}}^*;\tilde{\boldsymbol{\theta}},\tilde{\boldsymbol{\theta}}^*)$. For the simplicity of notation, we will write it as $p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})$.

We focus on discussing complex Rao test when no nuisance parameter is present in this paper, and the case of nuisance parameters will be treated in a follow-up work. The real Rao test statistic for deciding between the hypotheses $\mathcal{H}_0 : \boldsymbol{\xi} = \boldsymbol{\xi}_0$ versus $\mathcal{H}_1 : \boldsymbol{\xi} \neq \boldsymbol{\xi}_0$ (without nuisance parameters) is [1, pp. 221-230]

$$T_R(\mathbf{x}) = \left.\frac{\partial \ln p_{\mathbf{x}}(\mathbf{x};\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}\right|_{\boldsymbol{\xi}_0}^{T} \mathbf{I}^{-1}(\boldsymbol{\xi}_0) \left.\frac{\partial \ln p_{\mathbf{x}}(\mathbf{x};\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}\right|_{\boldsymbol{\xi}_0}, \tag{3.1}$$

where $\mathbf{I}(\boldsymbol{\xi})$ is the FIM of $\boldsymbol{\xi}$ and can be partitioned as

$$\mathbf{I}(\boldsymbol{\xi}) = \begin{bmatrix} \mathbf{I}_{\alpha\alpha} & \mathbf{I}_{\alpha\beta} \\ \mathbf{I}_{\beta\alpha} & \mathbf{I}_{\beta\beta} \end{bmatrix} \tag{3.2}$$

where $\mathbf{I}_{\alpha\alpha}, \mathbf{I}_{\beta\beta}, \mathbf{I}_{\beta\alpha}, \mathbf{I}_{\alpha\beta} \in \mathbb{R}^{p\times p}$, $\mathbf{I}_{\alpha\alpha}^T = \mathbf{I}_{\alpha\alpha}$, $\mathbf{I}_{\beta\beta}^T = \mathbf{I}_{\beta\beta}$, and $\mathbf{I}_{\alpha\beta}^T = \mathbf{I}_{\beta\alpha}$.

Our objective is to replace the real parameter vector $\boldsymbol{\xi}$ by the complex parameter vector $\tilde{\boldsymbol{\theta}}$. We will derive the complex parameter Rao test statistic for the unknown complex parameter $\tilde{\boldsymbol{\theta}}$ by carrying out the mathematical operations

with respect to the complex-valued quantities instead of using the real Rao test statistic of (3.1). Note that in doing so the PDF will then be expressed in terms of its original complex data $\tilde{\mathbf{x}}$ and complex parameter vector $\tilde{\boldsymbol{\theta}}$ as $p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})$. This allows us to differentiate the log-likelihood function with respect to the complex parameters while also maintaining the complex nature of the data, thus leading to a more intuitive and direct means of constructing the Rao test.

3.3 Complex Parameter Rao Test

Observe that $p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})$ is a real function of $\tilde{\boldsymbol{\theta}}$ and thus, must depend on $\tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}^*$. To make this apparent, we will denote the PDF $p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})$ at times as $p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\Theta}}) \equiv p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}},\tilde{\boldsymbol{\theta}}^*)$, where $\tilde{\boldsymbol{\Theta}} = \left[\tilde{\boldsymbol{\theta}}^T \ \tilde{\boldsymbol{\theta}}^H\right]^T$ and $\tilde{\boldsymbol{\Theta}} \in \mathbb{C}^{2p\times 1}$ [9]. Also, for $\tilde{z} = x + jy$ we have $2x = \tilde{z} + \tilde{z}^*$ and $2jy = \tilde{z} - \tilde{z}^*$, and the complex partial derivatives of a real scalar function $g(\tilde{z},\tilde{z}^*) \equiv f(x,y)$ are given by [9]

$$\frac{\partial g(\tilde{z},\tilde{z}^*)}{\partial \tilde{z}} = \frac{1}{2}\left(\frac{\partial f}{\partial x} - j\frac{\partial f}{\partial y}\right) \tag{3.3}$$

and

$$\frac{\partial g(\tilde{z},\tilde{z}^*)}{\partial \tilde{z}^*} = \frac{1}{2}\left(\frac{\partial f}{\partial x} + j\frac{\partial f}{\partial y}\right). \tag{3.4}$$

Finally, for a real function g of complex vectors $\tilde{\mathbf{z}}$ and $\tilde{\mathbf{z}}^*$, the complex gradient is given by

$$\left[\frac{\partial g(\tilde{\mathbf{z}},\tilde{\mathbf{z}}^*)}{\partial \tilde{\mathbf{z}}}\right]_i = \frac{\partial g(\tilde{\mathbf{z}},\tilde{\mathbf{z}}^*)}{\partial \tilde{z}_i} \tag{3.5}$$

$$\left[\frac{\partial g(\tilde{\mathbf{z}},\tilde{\mathbf{z}}^*)}{\partial \tilde{\mathbf{z}}^*}\right]_i = \frac{\partial g(\tilde{\mathbf{z}},\tilde{\mathbf{z}}^*)}{\partial \tilde{z}_i^*}. \tag{3.6}$$

With these definitions, we are able to formulate the complex parameter Rao test.

Theorem 1 (Complex parameter Rao test)**.** *The complex parameter Rao test statistic for testing the hypotheses* $\mathcal{H}_0 : \tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}_0$ *versus* $\mathcal{H}_1 : \tilde{\boldsymbol{\theta}} \neq \tilde{\boldsymbol{\theta}}_0$ *is given as*

$$T_{\tilde{R}}(\tilde{\mathbf{x}}) = \left.\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\Theta}})}{\partial \tilde{\boldsymbol{\Theta}}^*}\right|_{\tilde{\boldsymbol{\Theta}}_0}^{H} \underline{\tilde{\mathbf{I}}}^{-1}(\tilde{\boldsymbol{\Theta}}_0) \left.\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\Theta}})}{\partial \tilde{\boldsymbol{\Theta}}^*}\right|_{\tilde{\boldsymbol{\Theta}}_0} \tag{3.7}$$

where

$$\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\Theta}})}{\partial \tilde{\boldsymbol{\Theta}}^*} = \begin{bmatrix} \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\Theta}})}{\partial \tilde{\boldsymbol{\theta}}^*} \\ \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\Theta}})}{\partial \tilde{\boldsymbol{\theta}}} \end{bmatrix}, \tag{3.8}$$

and the $2p \times 2p$ *Fisher information matrix is*

$$\begin{aligned} \underline{\tilde{\mathbf{I}}}(\tilde{\boldsymbol{\Theta}}) &= E\left(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\Theta}})}{\partial \tilde{\boldsymbol{\Theta}}^*} \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\Theta}})}{\partial \tilde{\boldsymbol{\Theta}}^*}^H\right) & (3.9) \\ &= \begin{bmatrix} \tilde{\mathbf{I}}(\tilde{\boldsymbol{\theta}}) & \tilde{\mathbf{J}}(\tilde{\boldsymbol{\theta}}) \\ \tilde{\mathbf{J}}^*(\tilde{\boldsymbol{\theta}}) & \tilde{\mathbf{I}}^*(\tilde{\boldsymbol{\theta}}) \end{bmatrix}. & (3.10) \end{aligned}$$

Each block in the FIM has dimension $p \times p$ *and is defined as*

$$\tilde{\mathbf{I}}(\tilde{\boldsymbol{\theta}}) = E\left(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^*} \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^*}^H\right) \tag{3.11}$$

and

$$\tilde{\mathbf{J}}(\tilde{\boldsymbol{\theta}}) = E\left(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^*} \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}}^H\right). \tag{3.12}$$

The elements of the variance matrices are more conveniently evaluated by using second derivatives as

$$[\,\tilde{\boldsymbol{I}}(\tilde{\boldsymbol{\theta}})]_{kl} = -E\left(\frac{\partial^2 \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\theta}_k^* \partial \tilde{\theta}_l}\right) \tag{3.13}$$

and

$$[\,\tilde{\boldsymbol{J}}(\tilde{\boldsymbol{\theta}})]_{kl} = -E\left(\frac{\partial^2 \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\theta}_k^* \partial \tilde{\theta}_l^*}\right) \tag{3.14}$$

for $k = 1, 2, \ldots, p$ *and* $l = 1, 2, \ldots, p$.

The complex parameter form of the Rao test statistic as given by (3.7) is numerically identical to the Rao test statistic as given by (3.1) so that $T_{\tilde{R}}(\tilde{\mathbf{x}}) = T_R(\mathbf{x})$.

Note that $\underline{\tilde{\mathbf{I}}}(\tilde{\boldsymbol{\Theta}})$ *is a* $2p \times 2p$ *complex hermitian matrix, and hence,* $T_{\tilde{R}}(\tilde{\mathbf{x}})$ *is real.*

The reader should observe that no assumptions have been made on the form of $\mathbf{I}(\boldsymbol{\xi})$. For a proof of this theorem see Appendices A and C.

3.4 Complex Parameter Rao Test - Special Fisher Information Matrix

Next we consider a special form of the complex FIM, which is common in practice.

Theorem 2 (Special Form of Fisher Information Matrix)**.** *Assume the real FIM as given by (3.2) has the special form*

$$\mathbf{I}(\boldsymbol{\xi}) = 2 \begin{bmatrix} \mathbf{E} & -\mathbf{F} \\ \mathbf{F} & \mathbf{E} \end{bmatrix} \tag{3.15}$$

where $\mathbf{E} \in \mathbb{R}^{p \times p}$, $\mathbf{F} \in \mathbb{R}^{p \times p}$ *so that* $\mathbf{I}(\boldsymbol{\xi}) \in \mathbb{R}^{2p \times 2p}$, $\mathbf{E}^T = \mathbf{E}$, *and* $\mathbf{F}^T = -\mathbf{F}$. *Then,* $T_R(\mathbf{x})$ *can be equivalently expressed as*

$$T_{\tilde{R}}(\tilde{\mathbf{x}}) = 2 \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}}; \tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^*} \bigg|_{\tilde{\boldsymbol{\theta}}_0}^H \tilde{\mathbf{I}}^{-1}(\tilde{\boldsymbol{\theta}}_0) \, \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}}; \tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^*} \bigg|_{\tilde{\boldsymbol{\theta}}_0} \tag{3.16}$$

where

$$\tilde{\mathbf{I}}(\tilde{\boldsymbol{\theta}}) = E \left(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}}; \tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^*} \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}}; \tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^*}^H \right). \tag{3.17}$$

Note that $\tilde{\mathbf{I}}(\tilde{\boldsymbol{\theta}})$ *is a* $p \times p$ *complex Hermitian matrix.*

For a proof see Appendix B. In this case it is shown in Appendix B that $\tilde{\boldsymbol{J}}(\tilde{\boldsymbol{\theta}}) = \mathbf{0}$ and therefore the complex FIM of (3.9) is block diagonal, leading to the simplification of the complex parameter Rao test.

3.5 Some Examples

In this section we apply the theorems to two important problems. The first case is the general complex classical linear model [2, pp. 529-531], in which the Fisher information matrix takes on the special form. The second case is the testing of complex covariance matrix parameters in a complex Gaussian distribution, in which no special form applies in general.

3.5.1 Complex Classical Linear Model

Assume the data is modeled as follows [2, pp. 529-531]

$$\tilde{\mathbf{x}} = \tilde{\mathbf{H}}\tilde{\boldsymbol{\theta}} + \tilde{\mathbf{w}}, \tag{3.18}$$

where $\tilde{\mathbf{H}} \in \mathbb{C}^{N \times p}$ is a known matrix with $N > p$ and full rank, $\tilde{\boldsymbol{\theta}}$ is an unknown complex $p \times 1$ parameter vector, and $\tilde{\mathbf{w}}$ is a complex $N \times 1$ random vector with PDF $\tilde{\mathbf{w}} \sim \mathcal{CN}(\mathbf{0}, \tilde{\mathbf{C}})$, with $\tilde{\mathbf{C}} \in \mathbb{C}^{N \times N}$. Then, by the properties of the complex Gaussian PDF

$$\tilde{\mathbf{x}} \sim \mathcal{CN}(\tilde{\mathbf{H}}\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{C}}) \tag{3.19}$$

with $\tilde{\mathbf{C}}$ not dependent on $\tilde{\boldsymbol{\theta}}$ or $\tilde{\boldsymbol{\theta}}^*$. The PDF is (omitting the $\tilde{\boldsymbol{\theta}}^*$ dependence)

$$p(\tilde{\mathbf{x}}; \tilde{\boldsymbol{\theta}}) = \frac{1}{\pi^N \det(\tilde{\mathbf{C}})} \exp\left[-(\tilde{\mathbf{x}} - \tilde{\mathbf{H}}\tilde{\boldsymbol{\theta}})^H \tilde{\mathbf{C}}^{-1} (\tilde{\mathbf{x}} - \tilde{\mathbf{H}}\tilde{\boldsymbol{\theta}})\right]. \tag{3.20}$$

For this example it has been shown that the real FIM $\mathbf{I}(\boldsymbol{\xi})$ has the special form required [2, pp.529-531]. Hence, we can use Theorem 2 in formulating the

Rao test. We use some previously derived results which are [1, pp.484-485]

$$\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^*} = \tilde{\mathbf{H}}^H \tilde{\mathbf{C}}^{-1}(\tilde{\mathbf{x}} - \tilde{\mathbf{H}}\tilde{\boldsymbol{\theta}}), \tag{3.21}$$

and

$$\tilde{\boldsymbol{I}}(\tilde{\boldsymbol{\theta}}) = \tilde{\mathbf{H}}^H \tilde{\mathbf{C}}^{-1}\tilde{\mathbf{H}}. \tag{3.22}$$

Inserting these expressions into (3.16), we have that

$$T_{\tilde{\mathrm{R}}}(\tilde{\mathbf{x}}) = 2(\tilde{\mathbf{x}} - \tilde{\mathbf{H}}\tilde{\boldsymbol{\theta}}_0)^H \tilde{\mathbf{C}}^{-1}\tilde{\mathbf{H}}(\tilde{\mathbf{H}}^H \tilde{\mathbf{C}}^{-1}\tilde{\mathbf{H}})^{-1}\tilde{\mathbf{H}}^H \tilde{\mathbf{C}}^{-1}(\tilde{\mathbf{x}} - \tilde{\mathbf{H}}\tilde{\boldsymbol{\theta}}_0). \tag{3.23}$$

For the particular case in which $\tilde{\boldsymbol{\theta}}_0 = \mathbf{0}$ and $\tilde{\mathbf{C}} = \mathbf{I}$, this reduces to

$$T_{\tilde{\mathrm{R}}}(\tilde{\mathbf{x}}) = \frac{\tilde{\mathbf{x}}^H \tilde{\mathbf{H}}(\tilde{\mathbf{H}}^H \tilde{\mathbf{H}})^{-1}\tilde{\mathbf{H}}^H \tilde{\mathbf{x}}}{\sigma^2/2}, \tag{3.24}$$

which agrees with the generalized likelihood ratio test (GLRT) previously derived results in [1, pp. 484-485]. In fact, the result that the GLRT, the Rao test, and the Wald test are identical for the real linear model is also true for the complex linear model. Another example follows.

3.5.2 Complex autoregressive filter parameter

To illustrate that the special form of the real FIM does not always hold, we consider the problem of testing the complex parameter of a complex autoregressive (AR) random process. Even this simple case can involve some difficult calculations so that we restrict our example to a data set with $N = 3$.

In addition, it shows that the special FIM form does not hold in general for the unknown parameters of the covariance matrix of a multivariate complex Gaussian

PDF, with proof given in Appendix D and calculations of the AR example are consistent with the general results.

Suppose we have a complex Gaussian AR(1) random process sampled at $N = 3$ times to form the complex data vector $\tilde{\mathbf{x}} = [\tilde{x}_1 \ \tilde{x}_2 \ \tilde{x}_3]^T$. Assuming the process is zero mean and has parameters $\tilde{a}$, the filter parameter, and $\sigma_{\tilde{u}}^2 = 1$, the excitation noise variance. That is, for $n = 2, 3, \cdots, N$,

$$\tilde{x}_n = -\tilde{a}\tilde{x}_{n-1} + \tilde{u}_n \tag{3.25}$$

where $\tilde{u}_n$ is a complex white Gaussian noise with variance $\sigma_{\tilde{u}}^2 = 1$. Then, it has the following PDF

$$p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}}; \tilde{a}) = \frac{1}{\pi^3 \det(\tilde{\mathbf{C}})} \exp\left[-\tilde{\mathbf{x}}^H \tilde{\mathbf{C}}^{-1} \tilde{\mathbf{x}}\right], \tag{3.26}$$

where the covariance matrix is [13, pp.114-119]

$$\tilde{\mathbf{C}} = \frac{1}{1 - |\tilde{a}|^2} \begin{bmatrix} 1 & -\tilde{a}^* & (-\tilde{a}^*)^2 \\ -\tilde{a} & 1 & -\tilde{a}^* \\ (-\tilde{a})^2 & -\tilde{a} & 1 \end{bmatrix}. \tag{3.27}$$

The inverse covariance matrix is

$$\tilde{\mathbf{C}}^{-1} = \begin{bmatrix} 1 & \tilde{a}^* & 0 \\ \tilde{a} & 1 + \tilde{a}\tilde{a}^* & \tilde{a}^* \\ 0 & \tilde{a} & 1 \end{bmatrix} \tag{3.28}$$

Upon differentiating, we have

$$\begin{aligned} \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}}; \tilde{a})}{\partial \tilde{a}} &= -\frac{\partial \ln \det(\tilde{\mathbf{C}})}{\partial \tilde{a}} - \frac{\partial \tilde{\mathbf{x}}^H \tilde{\mathbf{C}}^{-1} \tilde{\mathbf{x}}}{\partial \tilde{a}} \\ &= -\frac{\tilde{a}^*}{1 - |\tilde{a}|^2} - (\tilde{x_2}^* \tilde{x_1} + \tilde{a}^* \tilde{x_2}^* \tilde{x_2} + \tilde{x_3}^* \tilde{x_2}). \end{aligned} \tag{3.29}$$

Now to determine whether $\tilde{\mathbf{J}}^*(\tilde{\boldsymbol{\theta}}) = \tilde{J}^*(\tilde{\theta}) = 0$, we have from (3.14) with $\tilde{\boldsymbol{\theta}} = \tilde{a}$ and $\tilde{\boldsymbol{\Theta}} = [\tilde{a}\, \tilde{a}^*]^T$

$$\begin{aligned} E\left(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{a})}{\partial \tilde{a}} \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{a})}{\partial \tilde{a}}\right) &= -E\left(\frac{\partial^2 \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{a})}{\partial \tilde{a} \partial \tilde{a}}\right) \\ &= \frac{(\tilde{a}^*)^2}{(1-|\tilde{a}|^2)^2} \\ &\neq 0, \end{aligned} \tag{3.30}$$

showing that the special form does not hold for $\tilde{a} \neq 0$. Hence, the Rao test must make use of Theorem 1. Then we have from (3.29) that

$$\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{a})}{\partial \tilde{a}^*} = -\frac{\tilde{a}}{1-|\tilde{a}|^2} - \tilde{x_2} \tilde{x_1}^* - \tilde{a}|\tilde{x_2}|^2 - \tilde{x_3} \tilde{x_2}^*. \tag{3.31}$$

Thus,

$$\begin{bmatrix} \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{a})}{\partial \tilde{a}^*} \\ \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{a})}{\partial \tilde{a}} \end{bmatrix} = - \begin{bmatrix} \frac{\tilde{a}}{1-|\tilde{a}|^2} + \tilde{a}|\tilde{x}_2|^2 + \tilde{x_2}\tilde{x_1}^* + \tilde{x_3}\tilde{x_2}^* \\ \frac{\tilde{a}^*}{1-|\tilde{a}|^2} + \tilde{a}^*|\tilde{x}_2|^2 + \tilde{x_2}^*\tilde{x_1} + \tilde{x_3}^*\tilde{x_2} \end{bmatrix}. \tag{3.32}$$

To find the FIM we use the second derivative form from (3.13) and (3.14)

$$\underline{\tilde{\boldsymbol{I}}}\left(\begin{bmatrix} \tilde{a} \\ \tilde{a}^* \end{bmatrix}\right) = \begin{bmatrix} -E\left[\frac{\partial^2 \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{a})}{\partial \tilde{a}^* \partial \tilde{a}}\right] & -E\left[\frac{\partial^2 \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{a})}{\partial \tilde{a}^* \partial \tilde{a}^*}\right] \\ -E\left[\frac{\partial^2 \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{a})}{\partial \tilde{a} \partial \tilde{a}}\right] & -E\left[\frac{\partial^2 \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{a})}{\partial \tilde{a} \partial \tilde{a}}\right] \end{bmatrix}. \tag{3.33}$$

From (3.29) and (3.31) we have that

$$-\frac{\partial^2 \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{a})}{\partial \tilde{a}^* \partial \tilde{a}} = |\tilde{x}_2|^2 + \frac{1}{(1-|\tilde{a}|^2)^2}, \tag{3.34}$$

$$-\frac{\partial^2 \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{a})}{\partial \tilde{a}^* \partial \tilde{a}^*} = \frac{\tilde{a}^2}{(1-|\tilde{a}|^2)^2}, \tag{3.35}$$

and therefore since

$$E[|\tilde{x}_2|^2] = \frac{1}{1-|\tilde{a}|^2}, \tag{3.36}$$

we have that

$$\underline{\tilde{\boldsymbol{I}}}(\tilde{\boldsymbol{\Theta}}) = \tilde{\boldsymbol{I}}\left(\begin{bmatrix} \tilde{a} \\ \tilde{a}^* \end{bmatrix}\right) = \begin{bmatrix} \frac{2-|\tilde{a}|^2}{(1-|\tilde{a}|^2)^2} & \frac{\tilde{a}^2}{(1-|\tilde{a}|^2)^2} \\ \frac{\tilde{a}^{*2}}{(1-|\tilde{a}|^2)^2} & \frac{2-|\tilde{a}|^2}{(1-|\tilde{a}|^2)^2} \end{bmatrix}. \tag{3.37}$$

Note that for $\tilde{a} \neq 0$, the FIM is not diagonal (see Appendix B) and hence $\mathbf{I}(\boldsymbol{\xi})$ will not have the special form of (3.15). Using the FIM and also

$$\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}}; \tilde{\boldsymbol{\Theta}})}{\partial \tilde{\boldsymbol{\Theta}}^*} = \begin{bmatrix} \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}}; \tilde{a})}{\partial \tilde{a}^*} \\ \\ \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}}; \tilde{a})}{\partial \tilde{a}} \end{bmatrix} \tag{3.38}$$

from (3.32) in Theorem 1 produces the complex parameter Rao test statistic. As a special case, if $\tilde{a}_0 = 0$, then we see that $\underline{\tilde{\boldsymbol{I}}}(\tilde{\boldsymbol{\Theta}}_0) = 2\mathbf{I}_2$ and also that

$$\left.\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}}; \tilde{\boldsymbol{\Theta}})}{\partial \tilde{\boldsymbol{\Theta}}^*}\right|_{\tilde{\boldsymbol{\Theta}}_0} = - \begin{bmatrix} \tilde{x_2} \tilde{x_1}^* + \tilde{x_3} \tilde{x_2}^* \\ \\ \tilde{x_2}^* \tilde{x_1} + \tilde{x_3}^* \tilde{x_2} \end{bmatrix}, \tag{3.39}$$

and therefore we have from (3.7) that

$$T_{\tilde{R}}(\tilde{\mathbf{x}}) = |\tilde{x}_1^* \tilde{x}_2 + \tilde{x}_2^* \tilde{x}_3|^2 \tag{3.40}$$

or in general we would have

$$T_{\tilde{R}}(\tilde{\mathbf{x}}) = \frac{2}{N-1} \left| \sum_{n=1}^{N-1} \tilde{x}_n^* \tilde{x}_{n+1} \right|^2. \tag{3.41}$$

The Rao test is then just a test of the estimated autocorrelation sequence (ACS) at lag one, after it is magnitude-squared. Clearly, for noise only the theoretical ACS at lag one would be zero while for the case of a signal present it would be

$$\begin{aligned} \left| \sum_{n=1}^{N-1} \tilde{x}_n^* \tilde{x}_{n+1} \right|^2 &\approx |(N-1)\tilde{r}_{\tilde{x}}[1]|^2 \\ &= \left| (N-1) \frac{-\tilde{a}}{1-|\tilde{a}|^2} \right|^2 \\ &= \frac{(N-1)^2 |\tilde{a}|^2}{(1-|\tilde{a}|^2)^2}. \end{aligned} \tag{3.42}$$

Note also that for this parameter value, i.e., $\tilde{a}_0 = 0$, the FIM has the special form and so (3.16) could have been used instead.

In fact, the special form of FIM does not hold in general for the unknown parameters of the covariance matrix of a multivariate complex Gaussian PDF. Therefore, the Theorem 2 can not be applied in this case. Instead, it must employ the Theorem 1. Suppose we observe a $N \times 1$ complex-valued vector $\tilde{\mathbf{x}}$, which is multivariate complex Gaussian distributed. Its mean is zero, and its covariance matrix is $\tilde{\mathbf{C}} = \tilde{\mathbf{C}}(\tilde{\boldsymbol{\Theta}}) = \tilde{\mathbf{C}}(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}^*)$, parameterized by an unknown $p \times 1$ complex parameter vector $\tilde{\boldsymbol{\theta}}$.That is,

$$\tilde{\mathbf{x}} \sim \mathcal{CN}(\mathbf{0}, \tilde{\mathbf{C}}(\tilde{\boldsymbol{\Theta}})), \tag{3.43}$$

and

$$p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}}; \tilde{\boldsymbol{\Theta}}) = \frac{1}{\pi^N \det\left(\tilde{\mathbf{C}}(\tilde{\boldsymbol{\Theta}})\right)} \exp\left[-\tilde{\mathbf{x}}^H \tilde{\mathbf{C}}^{-1}(\tilde{\boldsymbol{\Theta}})\tilde{\mathbf{x}}\right]. \tag{3.44}$$

It is derived in Appendix D that the $\tilde{\mathbf{J}}^*(\tilde{\boldsymbol{\theta}})$ element of the FIM $\underline{\tilde{\boldsymbol{I}}}(\tilde{\boldsymbol{\Theta}})$ is found to be nonzero in general,

$$\begin{aligned} \tilde{\mathbf{J}}^*(\tilde{\boldsymbol{\theta}}) &= E\left\{ \frac{\partial \ln p(\tilde{\mathbf{x}}; \tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}} \frac{\partial \ln p(\tilde{\mathbf{x}}; \tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^*}^H \right\} \\ &= \left(\mathcal{D}_{\tilde{\boldsymbol{\theta}}} \tilde{\mathbf{C}}\right)^T \left(\tilde{\mathbf{C}}^{-1} \otimes \tilde{\mathbf{C}}^{-T}\right) \mathbf{K}^N \left(\mathcal{D}_{\tilde{\boldsymbol{\theta}}} \tilde{\mathbf{C}}\right), \end{aligned} \tag{3.45}$$

where $\mathcal{D}_{\tilde{\boldsymbol{\theta}}} \tilde{\mathbf{C}} = \frac{\partial \operatorname{vec}(\tilde{\mathbf{C}})}{\partial \tilde{\boldsymbol{\theta}}^T}$ is the Jacobian matrix of $\tilde{\mathbf{C}}$ with respect to $\tilde{\boldsymbol{\theta}}$ and $\mathbf{K}^N$ is a $N^2 \times N^2$ commutation matrix. Therefore, Theorem 2 can not be applied to this case.

As an example, we show that this general result (3.45) applies to the AR

example. First,

$$\begin{aligned}
&\mathcal{D}_{\tilde{a}}\tilde{\mathbf{C}} \\
&\quad = \frac{\partial \operatorname{vec}(\tilde{\mathbf{C}})}{\partial \tilde{a}} \\
&\quad = \operatorname{vec}\left(\frac{1}{(1-\tilde{a}\tilde{a}^*)^2} \begin{bmatrix} \tilde{a}^* & -(\tilde{a}^*)^2 & (\tilde{a}^*)^3 \\ -1 & \tilde{a}^* & -(\tilde{a}^*)^2 \\ 2\tilde{a}-\tilde{a}^2\tilde{a}^* & -1 & \tilde{a}^* \end{bmatrix} \right) \\
&\quad = \operatorname{vec}\left(\frac{\partial \tilde{\mathbf{C}}}{\partial \tilde{a}} \right). \qquad (3.46)
\end{aligned}$$

Then, with (3.45), we have that

$$\begin{aligned}
\tilde{J}^*(\tilde{a}) &= (\mathcal{D}_{\tilde{a}}\tilde{\mathbf{C}})^T(\tilde{\mathbf{C}}^{-1}\otimes\tilde{\mathbf{C}}^{-T})\mathbf{K}^3(\mathcal{D}_{\tilde{a}}\tilde{\mathbf{C}}) \\
&= (\mathcal{D}_{\tilde{a}}\tilde{\mathbf{C}})^T(\tilde{\mathbf{C}}^{-1}\otimes\tilde{\mathbf{C}}^{-T}) \operatorname{vec}\left(\left(\frac{\partial \tilde{\mathbf{C}}}{\partial \tilde{a}} \right)^T \right) \qquad (3.47) \\
&= (\mathcal{D}_{\tilde{a}}\tilde{\mathbf{C}})^T \operatorname{vec}\left(\tilde{\mathbf{C}}^{-T} \left(\frac{\partial \tilde{\mathbf{C}}}{\partial \tilde{a}} \right)^T \tilde{\mathbf{C}}^{-T} \right).
\end{aligned}$$

While,

$$\tilde{\mathbf{C}}^{-T} \left(\frac{\partial \tilde{\mathbf{C}}}{\partial \tilde{a}} \right)^T \tilde{\mathbf{C}}^{-T} = \begin{bmatrix} 0 & -1 & 0 \\ 0 & -\tilde{a}^* & -1 \\ 0 & 0 & 0 \end{bmatrix}. \qquad (3.48)$$

Plugging equations (3.46) and(3.48) into (3.47), we have

$$\tilde{J}^*(\tilde{a}) = \frac{\tilde{a}^*\tilde{a}^*}{(1-\tilde{a}\tilde{a}^*)^2}, \qquad (3.49)$$

which is the same result with (3.30).

3.6 Conclusions

We have derived the Rao test for complex data and complex parameters. It can be used as an alternative to representing the data and parameters as concatenated real vectors. The alternative approach described computes the Rao test

statistic directly from the complex data with complex parameters rather than the approach of concatenating the complex parameter vector into a real vector. When the Fisher information matrix of the real parameters satisfies some special conditions, then the complex parameter Rao test resembles the real parameter Rao test, except for a factor of two. This result clarifies some supposedly but incorrect extensions of the real Rao test to the complex case. The important example of the complex linear model, in which the special conditions are satisfied has been given. Furthermore, the problem of testing of a complex covariance matrix parameter indicates a case in which the special conditions are not satisfied and hence, the slightly more complicated form of the complex parameter Rao test is required.

Acknowledgment

The authors would like to thank Dr. R. S. Raghavan for helpful discussions and insights. He is with the Air Force Research Laboratory, Wright-Patterson Air Force Base, Dayton, OH.

List of References

[1] S. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1998.

[2] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[3] W. Liu, Y. Wang and W. Xie, "Fisher information matrix, Rao test, and Wald test for complex-valued signals and their applications," *Signal Processing*, vol. 94, pp.1–5, 2014.

[4] Z. Zhu, S. Kay, F. Cogun and R.S. Raghavan. "On detection of nonstationarity in radar signal processing", *Proc of the 2016 IEEE Radar Conf.,* May 2–6, 2016, Philadelphia, pp. 1–4.

[5] P. M. Baggenstoss and S. Kay, "An adaptive detector for deterministic signals in noise of unknown spectra using the Rao test," *IEEE Trans. Signal Process.*, vol. 40, no. 6, pp. 1460–1468, Jun. 1992

[6] A. De Maio, S. Kay and A. Farina, "On the invariance, coincidence, and statistical equivalence of the GLRT, Rao test, and Wald test," *IEEE Trans. Signal Process.,* vol. 58, no. 4, pp. 1967–1979, 2010

[7] J. Liu, W. Liu, B. Chen, H. Liu, H. Li and C. Hao, "Modified rao test for multichannel adaptive signal detection, *IEEE Transactions on Signal Processing,* vol. 64, no. 3, pp. 714–725, Feb. 2016.

[8] W. Liu, J. Liu , L. Huang, D. Zou and Y. Wang, "Rao tests for distributed target detection in interference and noise, *Signal Processing,* vol. 117, pp. 333–342, Dec. 2015.

[9] V. Nagesha and S. Kay, "Cramer–Rao lower bounds for complex parameters," Unpublished, 1993, Available: http://www.ele.uri.edu/faculty/kay/New%20web/Books.html

[10] A. Hjorungnes and D. Gesbert, "Complex-valued matrix differentiation: techniques and key results," *IEEE Tran. on Signal Processing*, vol. 55, no. 6, pp. 2740–2746, June 2007.

[11] J.R. Magnus and H. Neudecker, "The commutation matrix: some properties and applications," *Annals of Statistics*, vol. 7, no. 2, pp. 381–394, 1979.

[12] S.A. Sultan and D.S. Tracy, "Moments of the complex multivariate normal distribution," *Linear Algebra and Its Applications*, vol. 237–238, pp. 191– 204, 1996.

[13] S. Kay, *Modern Spectral Estimation: Theory and Application*, Prentice–Hall, Englewood Cliffs, NJ, 1988.

[14] E. Conte and A. De Maio, "Distributed target detection in compound–Gaussian noise with Rao and Wald tests," *IEEE Trans. Aerosp. Electron. Syst.,* vol. 39, pp. 568–582, Apr. 2003.

[15] A. De Maio and S. Iommelli, "Coincidence of the Rao test, Wald test and GLRT in partially homogeneous environment," *IEEE Signal Process. Lett.,* vol. 15, pp. 385–388, 2008.

[16] A. De Maio, "Rao test for adaptive detection in Gaussian interference With unknown covariance matrix," *IEEE Trans. on Signal Processing*, vol. 55, no. 7, July, 2007.

[17] Z. Zhu and S. Kay,"The Rao test for testing the bandedness of complex-valued covariance matrix," *Proc. of IEEE International Conf. on Acoustic, Speech, and Signal Processing*, Mar 20–25, 2016, Shanghai, China, pp. 3960–3963.

[18] N. Levanon and E. Mozeson, *Radar Signals*, J. Wiley, NY, 2004.

[19] W. Knight, R. Pridham and S. Kay, "Digital signal processing for sonar," *Proc. of the IEEE*, vol. 69, no.11, pp.1451–1506, Nov. 1981.

Appendix A - Proof of Theorem 1

We use some results from [9]. With $\tilde{\boldsymbol{\Theta}} = [\tilde{\boldsymbol{\theta}}^T \ \tilde{\boldsymbol{\theta}}^H]^T$ and since $\tilde{\boldsymbol{\theta}} = \boldsymbol{\alpha} + j\boldsymbol{\beta}$, then by the definition of the complex gradient, we have

$$\begin{aligned}
\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\Theta}})}{\partial \tilde{\boldsymbol{\Theta}}^*} &= \begin{bmatrix} \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\Theta}})}{\partial \tilde{\boldsymbol{\theta}}^*} \\ \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\Theta}})}{\partial \tilde{\boldsymbol{\theta}}} \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{2}\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\Theta}})}{\partial \boldsymbol{\alpha}} + \frac{j}{2}\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\Theta}})}{\partial \boldsymbol{\beta}} \\ \frac{1}{2}\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\Theta}})}{\partial \boldsymbol{\alpha}} - \frac{j}{2}\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\Theta}})}{\partial \boldsymbol{\beta}} \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{2}\mathbf{I}_p & \frac{j}{2}\mathbf{I}_p \\ \frac{1}{2}\mathbf{I}_p & -\frac{j}{2}\mathbf{I}_p \end{bmatrix} \begin{bmatrix} \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\Theta}})}{\partial \boldsymbol{\alpha}} \\ \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\Theta}})}{\partial \boldsymbol{\beta}} \end{bmatrix} \\
&= \mathbf{T}\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\Theta}})}{\partial \boldsymbol{\xi}},
\end{aligned} \tag{3.50}$$

where $\mathbf{I}_p$ is a $p \times p$ identity matrix, and $\boldsymbol{\xi} = [\boldsymbol{\alpha}^T \boldsymbol{\beta}^T]^T$. Therefore,

$$\begin{aligned}
\underline{\tilde{\mathbf{I}}}(\tilde{\boldsymbol{\Theta}}) &= E\left(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\Theta}})}{\partial \tilde{\boldsymbol{\Theta}}^*}\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\Theta}})}{\partial \tilde{\boldsymbol{\Theta}}^*}^H\right) \\
&= E\left(\mathbf{T}\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\Theta}})}{\partial \boldsymbol{\xi}}\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\Theta}})}{\partial \boldsymbol{\xi}}^H \mathbf{T}^H\right) \\
&= \mathbf{T}\left[E\left(\frac{\partial \ln p_{\mathbf{x}}(\mathbf{x};\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}\frac{\partial \ln p_{\mathbf{x}}(\mathbf{x};\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}^T\right)\right]\mathbf{T}^H \\
&= \mathbf{T}\ \mathbf{I}(\boldsymbol{\xi})\ \mathbf{T}^H.
\end{aligned} \tag{3.51}$$

Note that $\tilde{\mathbf{x}} = \mathbf{u} + j\mathbf{v} \in \mathbb{C}^{N\times 1}$ while $\mathbf{x} = [\mathbf{u}^T \mathbf{v}^T]^T \in \mathbb{R}^{2N\times 1}$. Hence, we have

$$\begin{aligned} T_{\tilde{R}}(\tilde{\mathbf{x}}) &= \left(\mathbf{T} \frac{\partial \ln p_{\mathbf{x}}(\mathbf{x};\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}\bigg|_{\boldsymbol{\xi}_0} \right)^H \left(\mathbf{T}\mathbf{I}(\boldsymbol{\xi}_0)\mathbf{T}^H\right)^{-1} \\ &\quad \cdot \mathbf{T} \frac{\partial \ln p_{\mathbf{x}}(\mathbf{x};\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}\bigg|_{\boldsymbol{\xi}_0} \\ &= \frac{\partial \ln p_{\mathbf{x}}(\mathbf{x};\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}\bigg|_{\boldsymbol{\xi}_0}^T \mathbf{T}^H (\mathbf{T}^H)^{-1}\, \mathbf{I}^{-1}(\boldsymbol{\xi}_0)\, \mathbf{T}^{-1} \\ &\quad \cdot \mathbf{T}\, \frac{\partial \ln p_{\mathbf{x}}(\mathbf{x};\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}\bigg|_{\boldsymbol{\xi}_0} \\ &= \frac{\partial \ln p_{\mathbf{x}}(\mathbf{x};\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}\bigg|_{\boldsymbol{\xi}_0}^T \mathbf{I}^{-1}(\boldsymbol{\xi}_0)\, \frac{\partial \ln p_{\mathbf{x}}(\mathbf{x};\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}\bigg|_{\boldsymbol{\xi}_0} \\ &= T_R(\mathbf{x}). \end{aligned} \tag{3.52}$$

Appendix B - Proof of Theorem 2

We can write $\underline{\tilde{\mathbf{I}}}(\tilde{\boldsymbol{\Theta}})$, where $\underline{\tilde{\mathbf{I}}}(\tilde{\boldsymbol{\Theta}}) \in \mathbb{C}^{2p\times 2p}$, in block form from (3.9) as

$$\begin{aligned} \underline{\tilde{\mathbf{I}}}(\tilde{\boldsymbol{\Theta}}) &= E\left[\begin{bmatrix} \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^*} \\ \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}} \end{bmatrix} \begin{bmatrix} \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^*} \\ \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}} \end{bmatrix}^H \right] \\ &= \begin{bmatrix} \tilde{\mathbf{I}}(\tilde{\boldsymbol{\theta}}) & \tilde{\mathbf{J}}(\tilde{\boldsymbol{\theta}}) \\ \tilde{\mathbf{J}}^*(\tilde{\boldsymbol{\theta}}) & \tilde{\mathbf{I}}^*(\tilde{\boldsymbol{\theta}}) \end{bmatrix}, \end{aligned} \tag{3.53}$$

where

$$\tilde{\mathbf{I}}(\tilde{\boldsymbol{\theta}}) = E\left(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^*} \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^*}^H \right), \tag{3.54}$$

$$\tilde{\mathbf{J}}(\tilde{\boldsymbol{\theta}}) = E\left(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^*} \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}}^H \right), \tag{3.55}$$

$$\tilde{\mathbf{J}}^*(\tilde{\boldsymbol{\theta}}) = E\left(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}} \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^*}^H \right), \tag{3.56}$$

and

$$\tilde{\mathbf{I}}^*(\tilde{\boldsymbol{\theta}}) = E\left(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}} \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}}^H \right). \tag{3.57}$$

Note $\tilde{\mathbf{I}}(\tilde{\boldsymbol{\theta}}), \tilde{\mathbf{J}}(\tilde{\boldsymbol{\theta}}), \tilde{\mathbf{I}}^*(\tilde{\boldsymbol{\theta}}), \tilde{\mathbf{J}}^*(\tilde{\boldsymbol{\theta}}) \in \mathbb{C}^{p\times p}$. According to the definition of complex partials, we have

$$\begin{aligned}\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^*} &= \frac{1}{2}\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\alpha}} + \frac{j}{2}\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\beta}} \\ &= \frac{1}{2}\frac{\partial \ln p_{\mathbf{x}}(\mathbf{x};\boldsymbol{\xi})}{\partial \boldsymbol{\alpha}} + \frac{j}{2}\frac{\partial \ln p_{\mathbf{x}}(\mathbf{x};\boldsymbol{\xi})}{\partial \boldsymbol{\beta}}.\end{aligned} \tag{3.58}$$

Then, given with the special form of real FIM

$$\mathbf{I}(\boldsymbol{\xi}) = 2\begin{bmatrix} \mathbf{E} & -\mathbf{F} \\ \mathbf{F} & \mathbf{E} \end{bmatrix}, \tag{3.59}$$

we have

$$\begin{aligned}&\tilde{\mathbf{J}}(\tilde{\boldsymbol{\theta}}) \\ &\quad= E\left(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^*}\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}}^H\right) \\ &\quad= \frac{1}{4}(\mathbf{I}_{\alpha\alpha} + j\mathbf{I}_{\alpha\beta} + j\mathbf{I}_{\beta\alpha} - \mathbf{I}_{\beta\beta}) \\ &\quad= \frac{1}{4}(2\mathbf{E} - j2\mathbf{F} + j2\mathbf{F} - 2\mathbf{E}) \\ &\quad= \mathbf{0}.\end{aligned} \tag{3.60}$$

Thus,

$$\underline{\tilde{\mathbf{I}}}(\tilde{\boldsymbol{\Theta}}) = \begin{bmatrix} \tilde{\mathbf{I}}(\tilde{\boldsymbol{\theta}}) & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{I}}^*(\tilde{\boldsymbol{\theta}}) \end{bmatrix}. \tag{3.61}$$

Now, from (3.7)

$$\begin{aligned}T_{\tilde{R}}(\tilde{\mathbf{x}}) &= \begin{bmatrix} \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}} \\ \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^*} \end{bmatrix}\Bigg|_{\tilde{\boldsymbol{\theta}}_0}^T \begin{bmatrix} \tilde{\mathbf{I}}^{-1}(\tilde{\boldsymbol{\theta}}_0) & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{I}}^{-*}(\tilde{\boldsymbol{\theta}}_0) \end{bmatrix} \\ &\quad \cdot \begin{bmatrix} \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^*} \\ \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}} \end{bmatrix}\Bigg|_{\tilde{\boldsymbol{\theta}}_0} \\ &= 2\,\mathrm{Re}\left\{\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}}\Bigg|_{\tilde{\boldsymbol{\theta}}_0}^T \tilde{\mathbf{I}}^{-1}(\tilde{\boldsymbol{\theta}}_0)\,\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^*}\Bigg|_{\tilde{\boldsymbol{\theta}}_0}\right\},\end{aligned} \tag{3.62}$$

where, $\tilde{\mathbf{I}}^{-*}(\tilde{\boldsymbol{\theta}})$ denotes the inverse of $\tilde{\mathbf{I}}^{*}(\tilde{\boldsymbol{\theta}})$. Also, note that $\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}}\Big|_{\tilde{\boldsymbol{\theta}}_0}^{T} \tilde{\mathbf{I}}^{-1}(\tilde{\boldsymbol{\theta}}_0) \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^{*}}\Big|_{\tilde{\boldsymbol{\theta}}_0}$ is real. Therefore,

$$\begin{aligned} T_{\tilde{R}}(\tilde{\mathbf{x}}) &= 2 \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}}\Big|_{\tilde{\boldsymbol{\theta}}_0}^{T} \tilde{\mathbf{I}}^{-1}(\tilde{\boldsymbol{\theta}}_{\mathbf{0}}) \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^{*}}\Big|_{\tilde{\boldsymbol{\theta}}_0} \\ &= T_R(\mathbf{x}). \end{aligned} \tag{3.63}$$

Appendix C - Expression of complex Fisher information matrix with second derivatives of PDF

This section is to express complex Fisher information matrix in terms of second derivatives of PDF. We follow the lead of [9] to do so. First we consider

$$\tilde{\mathbf{I}}(\tilde{\boldsymbol{\theta}}) = E\left(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^{*}} \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^{*}}^{H} \right) \tag{3.64}$$

where $\tilde{\mathbf{I}}(\tilde{\boldsymbol{\theta}}) \in \mathbb{C}^{p\times p}$. Then, we will expand the result to the case

$$\underline{\tilde{\mathbf{I}}}(\tilde{\boldsymbol{\Theta}}) = \begin{bmatrix} \tilde{\mathbf{I}}(\tilde{\boldsymbol{\theta}}) & \tilde{\mathbf{J}}(\tilde{\boldsymbol{\theta}}) \\ \tilde{\mathbf{J}}^{*}(\tilde{\boldsymbol{\theta}}) & \tilde{\mathbf{I}}^{*}(\tilde{\boldsymbol{\theta}}) \end{bmatrix}. \tag{3.65}$$

Recall

$$\tilde{\mathbf{J}}(\tilde{\boldsymbol{\theta}}) = E\left(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^{*}} \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}}^{H} \right), \tag{3.66}$$

and $\tilde{\boldsymbol{\theta}} = \boldsymbol{\alpha} + j\boldsymbol{\beta}$, $\boldsymbol{\xi} = [\boldsymbol{\alpha}^T \ \boldsymbol{\beta}^T]^T$. Then, if the following regularity condition is satisfied

$$E\left(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}} \right) = \mathbf{0}, \tag{3.67}$$

then, according to (3),(4),(5), and (6), we have

$$E\left(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} \right) = \mathbf{0}. \tag{3.68}$$

Then, we have for $1 \le m, n \le 2p$ [2],

$$E\left(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\boldsymbol{\xi})}{\partial \xi_m} \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\boldsymbol{\xi})}{\partial \xi_n} \right) = -E\left(\frac{\partial^2 \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\boldsymbol{\xi})}{\partial \xi_m \partial \xi_n} \right) \tag{3.69}$$

With all these quantities, we have for $1 \leq k, l \leq p$,

$$
\begin{aligned}
\left[\tilde{\mathbf{I}}(\tilde{\boldsymbol{\theta}})\right]_{kl} &= E\left(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\theta}_k^*}\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\theta}_l}\right) \\
&= E\left[\frac{1}{2}\left(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \alpha_k} + j\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \beta_k}\right)\cdot\right. \\
&\qquad \left.\frac{1}{2}\left(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \alpha_l} - j\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \beta_l}\right)\right] \\
&= \frac{1}{4}E\Bigg(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \alpha_k}\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \alpha_l} \\
&\qquad + j\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \beta_k}\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \alpha_l} \\
&\qquad - j\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \alpha_k}\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \beta_l} \\
&\qquad + \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \beta_k}\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \beta_l}\Bigg) \\
&= -\frac{1}{4}E\Bigg(\frac{\partial^2 \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \alpha_k \partial \alpha_l} + j\frac{\partial^2 \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \beta_k \partial \alpha_l} \\
&\qquad - j\frac{\partial^2 \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \alpha_k \partial \beta_l} + \frac{\partial^2 \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \beta_k \partial \beta_l}\Bigg) \\
&= -\frac{1}{4}E\Bigg[\frac{\partial}{\partial \alpha_k}\left(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \alpha_l} - j\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \beta_l}\right) \\
&\qquad + j\frac{\partial}{\partial \beta_k}\left(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \alpha_l} - j\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \beta_l}\right)\Bigg] \\
&= -E\left[\frac{\partial}{\partial \tilde{\theta}_k^*}\left(\frac{1}{2}\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \alpha_l} - \frac{j}{2}\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \beta_l}\right)\right] \\
&= -E\left[\frac{\partial^2 \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\theta}_k^* \partial \tilde{\theta}_l}\right] \qquad (3.70)
\end{aligned}
$$

This shows that $\tilde{\mathbf{I}}(\tilde{\boldsymbol{\theta}})$ can be computed via second derivatives of PDFs element-wisely. Now by (3.65), it reduces to find $\tilde{\mathbf{J}}(\tilde{\boldsymbol{\theta}})$ to express $\underline{\tilde{\mathbf{I}}}(\tilde{\boldsymbol{\Theta}})$ in terms of second

derivatives of PDFs. Similarly, we have for $1 \leq k, l \leq p$,

$$\begin{aligned}
\left[\tilde{\mathbf{J}}(\tilde{\boldsymbol{\theta}})\right]_{kl} &= E\left(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\theta}_k^*}\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\theta}_l^*}\right) \\
&= E\left[\frac{1}{2}\left(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \alpha_k} + j\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \beta_k}\right)\cdot\right. \\
&\qquad \left.\frac{1}{2}\left(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \alpha_l} + j\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \beta_l}\right)\right] \\
&= \frac{1}{4}E\Bigg(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \alpha_k}\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \alpha_l} \\
&\qquad + j\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \beta_k}\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \alpha_l} \\
&\qquad + j\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \alpha_k}\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \beta_l} \\
&\qquad - \frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \beta_k}\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \beta_l}\Bigg)
\end{aligned}$$

Continuing the computation, we have

$$\begin{aligned}
\left[\tilde{\mathbf{J}}(\tilde{\boldsymbol{\theta}})\right]_{kl} &= -\frac{1}{4}E\Bigg(\frac{\partial^2 \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \alpha_k \partial \alpha_l} + j\frac{\partial^2 \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \beta_k \partial \alpha_l} \\
&\qquad + j\frac{\partial^2 \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \alpha_k \partial \beta_l} - \frac{\partial^2 \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \beta_k \partial \beta_l}\Bigg) \\
&= -\frac{1}{4}E\Bigg[\frac{\partial}{\partial \alpha_k}\left(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \alpha_l} + j\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \beta_l}\right) \\
&\qquad + j\frac{\partial}{\partial \beta_k}\left(\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \alpha_l} + j\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \beta_l}\right)\Bigg] \\
&= -E\left[\frac{\partial}{\partial \tilde{\theta}_k^*}\left(\frac{1}{2}\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \alpha_l} + \frac{j}{2}\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \beta_l}\right)\right] \\
&= -E\left[\frac{\partial^2 \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\theta}_k^* \partial \tilde{\theta}_l^*}\right]
\end{aligned} \tag{3.71}$$

This completes the expression of $\underline{\tilde{\mathbf{I}}}(\tilde{\boldsymbol{\Theta}})$ with second derivatives of PDF.

Appendix D - Discussion on the Fisher Information Matrix Form of Covariance Parameters

This example discuss the case when the unknown parameters parameterize covariance matrix. It shows that the FIM of covariance parameters does not have the special FIM form in general and it can not employ Theorem 2 in this case. Suppose we observe a $N \times 1$ complex-valued vector $\tilde{\mathbf{x}}$, which is multivariate complex Gaussian distributed with its mean being zero and its covariance matrix being $\tilde{\mathbf{C}} = \tilde{\mathbf{C}}(\tilde{\boldsymbol{\Theta}})$. Then,

$$p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}}; \tilde{\boldsymbol{\Theta}}) = \frac{1}{\pi^N \det\left(\tilde{\mathbf{C}}(\tilde{\boldsymbol{\Theta}})\right)} \exp\left[-\tilde{\mathbf{x}}^H \tilde{\mathbf{C}}^{-1}(\tilde{\boldsymbol{\Theta}})\tilde{\mathbf{x}}\right], \tag{3.72}$$

and

$$\frac{\partial \ln p_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}}; \tilde{\boldsymbol{\Theta}})}{\partial \tilde{\boldsymbol{\theta}}} = -\frac{\partial \ln \det \tilde{\mathbf{C}}(\tilde{\boldsymbol{\Theta}})}{\partial \tilde{\boldsymbol{\theta}}} - \frac{\partial \tilde{\mathbf{x}}^H \tilde{\mathbf{C}}^{-1}(\tilde{\boldsymbol{\Theta}})\tilde{\mathbf{x}}}{\partial \tilde{\boldsymbol{\theta}}}. \tag{3.73}$$

Our notations in the following part are based on [10]. Also, from [10] the Jacobian matrix $\mathcal{D}_{\tilde{\boldsymbol{Z}}}\tilde{\mathbf{F}}(\tilde{\mathbf{Z}}, \tilde{\mathbf{Z}}^*)$ of $\tilde{\mathbf{F}}$ with respect to $\tilde{\mathbf{Z}}$ can be computed as

$$\mathcal{D}_{\tilde{\boldsymbol{Z}}}\tilde{\mathbf{F}}(\tilde{\mathbf{Z}}, \tilde{\mathbf{Z}}^*) = \frac{\partial \operatorname{vec}(\tilde{\mathbf{F}}(\tilde{\mathbf{Z}}, \tilde{\mathbf{Z}}^*))}{\partial \operatorname{vec}^T(\tilde{\mathbf{Z}})}. \tag{3.74}$$

Then, we have

$$\mathcal{D}_{\tilde{\boldsymbol{\theta}}}\left(\ln \det \tilde{\mathbf{C}}(\tilde{\boldsymbol{\Theta}})\right) = \left(\frac{\partial \ln \det \tilde{\mathbf{C}}(\tilde{\boldsymbol{\Theta}})}{\partial \tilde{\boldsymbol{\theta}}}\right)^T, \tag{3.75}$$

and

$$\mathcal{D}_{\tilde{\boldsymbol{\theta}}}\left(\tilde{\mathbf{x}}^H \tilde{\mathbf{C}}^{-1}(\tilde{\boldsymbol{\Theta}})\tilde{\mathbf{x}}\right) = \left(\frac{\partial\left(\tilde{\mathbf{x}}^H \tilde{\mathbf{C}}^{-1}(\tilde{\boldsymbol{\Theta}})\tilde{\mathbf{x}}\right)}{\partial \tilde{\boldsymbol{\theta}}}\right)^T. \tag{3.76}$$

By the Chain rule[10], we have

$$\begin{aligned}\mathcal{D}_{\tilde{\boldsymbol{\theta}}}(\ln\det\tilde{\mathbf{C}}) &= \left[\mathcal{D}_{\tilde{\mathbf{C}}}(\ln\det\tilde{\mathbf{C}})\right]\mathcal{D}_{\tilde{\boldsymbol{\theta}}}\tilde{\mathbf{C}} \\ &= \mathrm{vec}^T\left(\tilde{\mathbf{C}}^{-T}\right)\mathcal{D}_{\tilde{\boldsymbol{\theta}}}\tilde{\mathbf{C}}, \end{aligned} \tag{3.77}$$

where, we have applied that $\mathcal{D}_{\tilde{\mathbf{C}}^*}(\ln\det\tilde{\mathbf{C}}) = \mathbf{0}$. Similarly,

$$\begin{aligned}\mathcal{D}_{\tilde{\boldsymbol{\theta}}}(\tilde{\mathbf{x}}^H\tilde{\mathbf{C}}^{-1}\tilde{\mathbf{x}}) &= \mathcal{D}_{\tilde{\mathbf{C}}^{-1}}\left(\tilde{\mathbf{x}}^H\tilde{\mathbf{C}}^{-1}\tilde{\mathbf{x}}\right)\mathcal{D}_{\tilde{\mathbf{C}}}\left(\tilde{\mathbf{C}}^{-1}\right)\mathcal{D}_{\tilde{\boldsymbol{\theta}}}\tilde{\mathbf{C}} \\ &= \mathrm{vec}^T(\tilde{\mathbf{x}}^*\tilde{\mathbf{x}}^T)\left(-\tilde{\mathbf{C}}^{-T}\otimes\tilde{\mathbf{C}}^{-1}\right)\mathcal{D}_{\tilde{\boldsymbol{\theta}}}\tilde{\mathbf{C}}, \end{aligned} \tag{3.78}$$

where we have applied $\mathcal{D}_{\tilde{\mathbf{C}}^{-*}}\left(\tilde{\mathbf{x}}^H\tilde{\mathbf{C}}^{-1}\tilde{\mathbf{x}}\right) = \mathbf{0}$ and $\mathcal{D}_{\tilde{\mathbf{C}}^*}\tilde{\mathbf{C}}^{-1} = \mathbf{0}$. Thus,

$$\begin{aligned}\frac{\partial\ln p(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial\tilde{\boldsymbol{\theta}}} &= -\left[\mathcal{D}_{\tilde{\boldsymbol{\theta}}}\left(\ln\det\tilde{\mathbf{C}}\right)\right]^T - \left[\mathcal{D}_{\tilde{\boldsymbol{\theta}}}\left(\tilde{\mathbf{x}}^H\tilde{\mathbf{C}}^{-1}\tilde{\mathbf{x}}\right)\right]^T \\ &= \left[\mathrm{vec}^T(\tilde{\mathbf{x}}^*\tilde{\mathbf{x}}^T)\left(\tilde{\mathbf{C}}^{-T}\otimes\tilde{\mathbf{C}}^{-1}\right)\mathcal{D}_{\tilde{\boldsymbol{\theta}}}\tilde{\mathbf{C}}\right]^T \\ &\quad -\left[\mathrm{vec}^T(\tilde{\mathbf{C}}^{-T})\mathcal{D}_{\tilde{\boldsymbol{\theta}}}\tilde{\mathbf{C}}\right]^T \\ &= (\mathcal{D}_{\tilde{\boldsymbol{\theta}}}\tilde{\mathbf{C}})^T\left(\tilde{\mathbf{C}}^{-1}\otimes\tilde{\mathbf{C}}^{-T}\right)\mathrm{vec}(\tilde{\mathbf{x}}^*\tilde{\mathbf{x}}^T) \\ &\quad -\left(\mathcal{D}_{\tilde{\boldsymbol{\theta}}}\tilde{\mathbf{C}}\right)^T\mathrm{vec}(\tilde{\mathbf{C}}^{-T}) \\ &= (\mathcal{D}_{\tilde{\boldsymbol{\theta}}}\tilde{\mathbf{C}})^T\,\mathrm{vec}\left(\tilde{\mathbf{C}}^{-T}(\tilde{\mathbf{x}}^*\tilde{\mathbf{x}}^T)\tilde{\mathbf{C}}^{-T}\right) \\ &\quad -(\mathcal{D}_{\tilde{\boldsymbol{\theta}}}\tilde{\mathbf{C}})^T\,\mathrm{vec}(\tilde{\mathbf{C}}^{-T}). \end{aligned} \tag{3.79}$$

The result

$$\left(\tilde{\mathbf{C}}^{-1}\otimes\tilde{\mathbf{C}}^{-T}\right)\mathrm{vec}\left(\tilde{\mathbf{x}}^*\tilde{\mathbf{x}}^T\right) = \mathrm{vec}\left(\tilde{\mathbf{C}}^{-T}\left(\tilde{\mathbf{x}}^*\tilde{\mathbf{x}}^T\right)\tilde{\mathbf{C}}^{-T}\right) \tag{3.80}$$

has been used above. Also, the regularity condition is met as shown as below.

$$\begin{aligned} E\left\{\frac{\partial \ln p(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}}\right\} &= (\mathcal{D}_{\tilde{\boldsymbol{\theta}}}\tilde{\mathbf{C}})^T \operatorname{vec}\left(\tilde{\mathbf{C}}^{-T}\mathbf{E}\{\tilde{\mathbf{x}}^*\tilde{\mathbf{x}}^T\}\tilde{\mathbf{C}}^{-T}\right) \\ &\quad -(\mathcal{D}_{\tilde{\boldsymbol{\theta}}}\tilde{\mathbf{C}})^T \operatorname{vec}(\tilde{\mathbf{C}}^{-T}) \\ &= (\mathcal{D}_{\tilde{\boldsymbol{\theta}}}\tilde{\mathbf{C}})^T \operatorname{vec}\left(\tilde{\mathbf{C}}^{-T}\tilde{\mathbf{C}}^{T}\tilde{\mathbf{C}}^{-T}\right) \\ &\quad -(\mathcal{D}_{\tilde{\boldsymbol{\theta}}}\tilde{\mathbf{C}})^T \operatorname{vec}(\tilde{\mathbf{C}}^{-T}) \\ &= \mathbf{0}. \end{aligned} \tag{3.81}$$

Next, we explore to see if in this case the FIM of $\tilde{\boldsymbol{\theta}}$ has the special form as shown in Theorem 2. That reduces to see if the following equation is met.

$$E\left\{\frac{\partial \ln p(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}}\frac{\partial \ln p(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^T}\right\} = \mathbf{0}. \tag{3.82}$$

For the purpose of simplifying notations, we denote $\mathcal{D}_{\tilde{\boldsymbol{\theta}}}\tilde{\mathbf{C}} = \tilde{\mathbf{G}}$, then $\tilde{\mathbf{G}} \in \mathbb{C}^{N^2\times p}$, and $\operatorname{vec}\left(\tilde{\mathbf{C}}^{-T}(\tilde{\mathbf{x}}^*\tilde{\mathbf{x}}^T)\tilde{\mathbf{C}}^{-T}\right) = \tilde{\mathbf{f}}$, then $\tilde{\mathbf{f}} \in \mathbb{C}^{N^2\times 1}$, and $\operatorname{vec}(\tilde{\mathbf{C}}^{-T}) = \tilde{\mathbf{h}}$, $\tilde{\mathbf{h}} \in \mathbb{C}^{N^2\times 1}$. Note the relationship $E\{\tilde{\mathbf{f}}\} = \tilde{\mathbf{h}}$. Let $\tilde{\mathbf{B}} = E\left\{\frac{\partial \ln p(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}}\frac{\partial \ln p(\tilde{\mathbf{x}};\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^T}\right\}$. Then,

$$\begin{aligned} \tilde{\mathbf{B}} &= E\{\tilde{\mathbf{G}}^T(\tilde{\mathbf{f}}\tilde{\mathbf{f}}^T - \tilde{\mathbf{f}}\tilde{\mathbf{h}}^T - \tilde{\mathbf{h}}\tilde{\mathbf{f}}^T + \tilde{\mathbf{h}}\tilde{\mathbf{h}}^T)\tilde{\mathbf{G}}\} \\ &= \tilde{\mathbf{G}}^T\left[E\{\tilde{\mathbf{f}}\tilde{\mathbf{f}}^T\} - \tilde{\mathbf{h}}\tilde{\mathbf{h}}^T\right]\tilde{\mathbf{G}} \end{aligned} \tag{3.83}$$

and

$$\begin{aligned} E\{\tilde{\mathbf{f}}\tilde{\mathbf{f}}^T\} &= E\Big\{\left(\tilde{\mathbf{C}}^{-1}\otimes\tilde{\mathbf{C}}^{-T}\right)\operatorname{vec}(\tilde{\mathbf{x}}^*\tilde{\mathbf{x}}^T) \\ &\quad \operatorname{vec}^T(\tilde{\mathbf{x}}^*\tilde{\mathbf{x}}^T)\left(\tilde{\mathbf{C}}^{-T}\otimes\tilde{\mathbf{C}}^{-1}\right)\Big\} \\ &= \left(\tilde{\mathbf{C}}^{-1}\otimes\tilde{\mathbf{C}}^{-T}\right)E\{\operatorname{vec}(\tilde{\mathbf{x}}^*\tilde{\mathbf{x}}^T) \\ &\quad \operatorname{vec}^T(\tilde{\mathbf{x}}^*\tilde{\mathbf{x}}^T)\}\left(\tilde{\mathbf{C}}^{-T}\otimes\tilde{\mathbf{C}}^{-1}\right). \end{aligned} \tag{3.84}$$

It reduces to find $E\{\text{vec}(\tilde{\mathbf{x}}^*\tilde{\mathbf{x}}^T)\,\text{vec}^T(\tilde{\mathbf{x}}^*\tilde{\mathbf{x}}^T)\}$, which is denoted as $\widetilde{\mathbf{W}}$.Observe that $\text{vec}(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^H) = \tilde{\mathbf{x}}^* \otimes \tilde{\mathbf{x}}$ and $\text{vec}(\tilde{\mathbf{x}}^*\tilde{\mathbf{x}}^T) = \mathbf{K}^{N,N}\,\text{vec}(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^H)$, where $\mathbf{K}^{N,N}$ is $N^2 \times N^2$ commutation matrix, denoted as $\mathbf{K}^N$ for simplicity, and it has special property $\mathbf{K}^N = (\mathbf{K}^N)^T$ [11]. Then,

$$\begin{aligned}\widetilde{\mathbf{W}} &= E\{\mathbf{K}^N\,\text{vec}(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^H)\,\text{vec}^T(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^H)(\mathbf{K}^N)^T\} \\ &= \mathbf{K}^N E\{\text{vec}(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^H)\,\text{vec}^T(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^H)\}\mathbf{K}^N. \end{aligned} \tag{3.85}$$

Let $\tilde{\mathbf{V}} = \text{vec}(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^H)\,\text{vec}^T(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^H)$, then $E\{\text{vec}(\tilde{\mathbf{V}})\}$ is the fourth moment of $\tilde{\mathbf{x}}$ [12].

$$\begin{aligned}E\{\text{vec}(\tilde{\mathbf{V}})\} &= E\{\text{vec}(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^H) \otimes \text{vec}(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^H)\} \\ &= E\{(\tilde{\mathbf{x}}^* \otimes \tilde{\mathbf{x}}) \otimes (\tilde{\mathbf{x}}^* \otimes \tilde{\mathbf{x}})\} \\ &= \text{vec}[\text{vec}(\tilde{\mathbf{C}})\,\text{vec}^T(\tilde{\mathbf{C}})] + \text{vec}((\tilde{\mathbf{C}}^T \otimes \tilde{\mathbf{C}})\mathbf{K}^N). \end{aligned} \tag{3.86}$$

Therefore, we have

$$E\{\tilde{\mathbf{V}}\} = \text{vec}(\tilde{\mathbf{C}})\,\text{vec}^T(\tilde{\mathbf{C}}) + (\tilde{\mathbf{C}}^T \otimes \tilde{\mathbf{C}})\mathbf{K}^N. \tag{3.87}$$

Substituting (3.87) in (3.85)

$$\begin{aligned}\widetilde{\mathbf{W}} &= \mathbf{K}^N \left[\text{vec}(\tilde{\mathbf{C}})\,\text{vec}^T(\tilde{\mathbf{C}}) + (\tilde{\mathbf{C}}^T \otimes \tilde{\mathbf{C}})\mathbf{K}^N\right]\mathbf{K}^N \\ &= \text{vec}(\tilde{\mathbf{C}}^T)\,\text{vec}^T(\tilde{\mathbf{C}}^T) + \mathbf{K}^N(\tilde{\mathbf{C}}^T \otimes \tilde{\mathbf{C}}). \end{aligned} \tag{3.88}$$

The properties of the commutation matrix $\mathbf{K}^N$ that $\mathbf{K}^N\,\text{vec}(\tilde{\mathbf{C}}) = \text{vec}(\tilde{\mathbf{C}}^T)$ and $\mathbf{K}^N\mathbf{K}^N = \mathbf{I}$ have been used above. Substituting (3.88) in (3.85) and (3.84) and

denoting $\tilde{\mathbf{P}} = \tilde{\mathbf{C}}^{-1} \otimes \tilde{\mathbf{C}}^{-T}$, then

$$\begin{aligned} E\{\tilde{\mathbf{f}}\tilde{\mathbf{f}}^T\} &= \tilde{\mathbf{P}} \left[\mathrm{vec}(\tilde{\mathbf{C}}^T)\, \mathrm{vec}^T(\tilde{\mathbf{C}}^T) + \mathbf{K}^N(\tilde{\mathbf{C}}^T \otimes \tilde{\mathbf{C}}) \right] \tilde{\mathbf{P}}^T \\ &= \mathrm{vec}(\tilde{\mathbf{C}}^{-T})\, \mathrm{vec}^T(\tilde{\mathbf{C}}^{-T}) + \tilde{\mathbf{P}}\mathbf{K}. \end{aligned} \tag{3.89}$$

Using (3.89) in (3.83), we have

$$\begin{aligned} \tilde{\mathbf{B}} &= E\left\{ \frac{\partial \ln p(\tilde{\mathbf{x}}; \tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}} \frac{\partial \ln p(\tilde{\mathbf{x}}; \tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}^T} \right\} \\ &= \tilde{\mathbf{J}}^*(\tilde{\boldsymbol{\theta}}) \\ &= \tilde{\mathbf{G}}^T \tilde{\mathbf{P}} \mathbf{K}^N \tilde{\mathbf{G}} \\ &= \left(\mathcal{D}_{\tilde{\boldsymbol{\theta}}} \tilde{\mathbf{C}} \right)^T \left(\tilde{\mathbf{C}}^{-1} \otimes \tilde{\mathbf{C}}^{-T} \right) \mathbf{K}^N \left(\mathcal{D}_{\tilde{\boldsymbol{\theta}}} \tilde{\mathbf{C}} \right) \end{aligned} \tag{3.90}$$

which is not zero in general. Thus, Theorem 2 cannot be applied to this case.

MANUSCRIPT 4

The Rao Test for Testing Bandedness of Complex-Valued Covariance Matrix

by

Zhenghan Zhu and Steven Kay

published in

in Proc. of the 41^{st} *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.3960–3963, Shanghai, Mar. 2016.

Abstract

Banding the inverse of covariance matrix has become a popular technique to estimate a high dimensional covariance matrix from limited number of samples. However, little work has been done in providing a criterion to determine when a matrix is bandable. In this paper, we present a detector to test the bandedness of a Cholesky factor matrix. The test statistic is formed based on the Rao test, which does not require the maximum likelihood estimates under the alternative hypothesis. In many fields, such as radar signal processing, the covariance matrix and its unknown parameters are often complex-valued. We focus on dealing with complex-valued cases by utilizing the complex parameter Rao test, instead of the traditional real Rao test. This leads to a more intuitive and efficient test statistic. Examples and computer simulations are given to investigate the derived detector performance.

4.1 Introduction

In statistical signal processing, such as used in a radar signal processing system, the sample covariance matrix plays an essential role. [1]. It is often estimated from N adjacent sample data vectors $[\mathbf{x}_0\ \mathbf{x}_1\ \cdots, \mathbf{x}_{N-1}]$, where $\mathbf{x}_n$'s are assumed to be $L \times 1$ identical and independent distributed (IID) complex-valued data vectors, with the general maximum likelihood covariance matrix estimate $\hat{\mathbf{C}} = \frac{1}{N}\sum_{n=0}^{N-1} \mathbf{x}_n \mathbf{x}_n^H$ [3], where H denotes hermitian. A good covariance matrix estimate usually requires N to be large. For example, it requires $N \geq 2L$ in space-time adaptive processing (STAP) to have a good clutter covariance matrix

estimate [2] . In practice, however, this is not valid due to the nonstationary environment. For example, the data for a STAP system is often nonstationary due to the heterogeneous clutter [1]. The number of data sufficiently IID (homogeneous) can be relatively small $N \leq L$ [2].

A popular solution to the problem is adopting banding/tapering techniques. Wu and et al. proposed to estimate the covariance matrix by banding the cholesky factor matrix and applying kernel smoothing estimation [4]. Bickel demonstrated that within the bandable class of covariance matrices, the estimator $\hat{\mathbf{C}}^{-1}$ obtained by banding the cholesky factor matrix of the covariance matrix's inverse is consistent [5]. However, little work is available to provide a guideline/criterion on deciding if a covariance matrix or the cholesky factor matrix of its inverse is bandable. Such a criterion is important and useful to decide if the banding technique is a suitable strategy. Other covariance estimation methods, such as modeling the covariance matrix as a time-varying autoregressive moving average (ARMA) model [8] also requires testing to decide if the model is a good fit. Some recent hypothesis tests for bandedness can be found in [6].

In this paper, a new test based on the Rao test is presented to test the bandedness of a Cholesky factor matrix. The Rao test has an asymptotic optimality property for large data records, yet it requires noticeably lower computation cost than some other detectors, ie., generalized likelihood ratio test (GLRT), as it only needs the maximum likelihood estimates (MLE) under the null hypotheses

[9]. This property in computational cost of the Rao test can be an advantage in high-dimensional multivariate signal processing. We consider a complex-valued covariance matrix and unknown parameters in this paper. We adopt the complex parameter Rao test, which offers a more intuitive detector than the traditional real Rao test for testing complex-valued parameters [7]. It should be pointed out, however, that the concept of utilizing the Rao test for testing the bandedness of a matrix can also be easily applied to the real-valued covariance matrix case via the real Rao test.

The paper is organized as follows: Section 4.2 formulates the problem; Section 4.3 derives the Rao test detector for testing the bandedness of the cholesky factor matrix; Examples and computer simulations for evaluating the detector's performance are given in Section 4.4; Finally, conclusions are drawn in Section 4.5.

4.2 Problem Formulation

Assume that we have N IID observed data vectors, $\mathbf{X} = [\mathbf{x}_0^T \ \mathbf{x}_1^T \ \cdots \ \mathbf{x}_{N-1}^T]^T$, where T denotes transpose and each $\mathbf{x}_n$ is an $L \times 1$ complex-valued data vector, which obeys a zero-mean multivariate complex Gaussian distribution $\mathbf{x}_n \sim \mathcal{CN}(\mathbf{0}, \mathbf{C})$ for $n = 0, 1, \cdots, N-1$, and the $\mathbf{x}_n$'s are mutually independent. We assume the $N \leq L$ limitation. The $L \times L$ covariance matrices $\mathbf{C}$ is a Hermitian matrix, so its inverse can be decomposed via the Cholesky decomposition as

$$\mathbf{C}^{-1} = \mathbf{D}^H \mathbf{D},$$

where $\mathbf{D}$ is a lower triangular $L \times L$ matrix with a testing model as follows.

$$\mathbf{D} = \mathbf{D}_B + \sum_{k=1}^{M} b_k \mathbf{\Phi}_k$$

$\mathbf{D}_B$ is a known banded lower triangular matrix, with the bandwidth to be m, the b_k's are unknown complex-valued parameters, and the $\mathbf{\Phi}_k$'s are known basis matrices. Specifically,

$$\begin{aligned}
b_1 &= [\mathbf{D}]_{m+2,1}, & \mathbf{\Phi}_1 &= \mathbf{e}_{m+2}\mathbf{e}_1^T \\
b_2 &= [\mathbf{D}]_{m+3,2}, & \mathbf{\Phi}_2 &= \mathbf{e}_{m+3}\mathbf{e}_2^T \\
&\vdots & &\vdots \\
b_{L-m-1} &= [\mathbf{D}]_{L,L-m-1}, & \mathbf{\Phi}_{L-m-1} &= \mathbf{e}_L\mathbf{e}_{L-m-1}^T \\
b_{L-m} &= [\mathbf{D}]_{m+3,1}, & \mathbf{\Phi}_{L-m} &= \mathbf{e}_{m+3}\mathbf{e}_1^T \\
b_{L-m+1} &= [\mathbf{D}]_{m+4,2}, & \mathbf{\Phi}_{L-m+1} &= \mathbf{e}_{m+4}\mathbf{e}_2^T \\
&\vdots & &\vdots \\
b_M &= [\mathbf{D}]_{L,1}, & \mathbf{\Phi}_M &= \mathbf{e}_L\mathbf{e}_1^T
\end{aligned}$$

where $M = \frac{(L-m-1)(L-m)}{2}$ and each $\mathbf{e}_k$ is an $L \times 1$ vector with k^{th} element being one and the rest being all zeros. The objective is to test if the lower triangular Cholesky factor matrix $\mathbf{D}$ is equal to the banded lower triangular matrix $\mathbf{D}_B$. Let $\mathbf{b} = [b_1\ b_2\ ...\ b_M]^T$. The detection problem is equivalent to choosing between the

following hypotheses:

$$\mathcal{H}_0 : \mathbf{b} = \mathbf{0};$$

$$\mathcal{H}_1 : \mathbf{b} \neq \mathbf{0};$$

4.3 The Rao test for testing the bandedness

In this section, we derive the complex parameter Rao test for the aforementioned detection problem. The Rao test attains the asymptotic (as $N \to \infty$) performance as the GLRT but avoids requiring MLEs under the alternative hypothesis $\mathcal{H}_1$, so its computation cost is often substantially less than the GLRT. This can be a desirable property in high-dimensional signal processing, such as real-time STAP. The derivation of the Rao test statistics follows. Let $\mathbf{b}^* = [b_1^* \; b_2^* \; ... \; b_M^*]^T$, where $*$ denotes conjugate, and $\underline{\mathbf{b}} = [\mathbf{b}^T \; \mathbf{b}^H]^T$, which is an $2M \times 1$ complex-valued parameter vector. The complex parameter Rao test detector can be formed [7]

$$T_R(\mathbf{X}) = \left.\frac{\partial \ln p(\mathbf{X};\underline{\mathbf{b}})}{\partial \underline{\mathbf{b}}^*}\right|_{\mathbf{b}=\mathbf{0}}^{H} \mathbf{I}^{-1}(\underline{\mathbf{b}})\big|_{\mathbf{b}=\mathbf{0}} \left.\frac{\partial \ln p(\mathbf{X};\underline{\mathbf{b}})}{\partial \underline{\mathbf{b}}^*}\right|_{\mathbf{b}=\mathbf{0}} \tag{4.1}$$

where,

$$\frac{\partial \ln p(\mathbf{X};\underline{\mathbf{b}})}{\partial \underline{\mathbf{b}}} = \left[\frac{\partial \ln p(\mathbf{X};\underline{\mathbf{b}})}{\partial \mathbf{b}}^T \frac{\partial \ln p(\mathbf{X};\underline{\mathbf{b}})}{\partial \mathbf{b}^*}^T\right]^T,$$

$$\frac{\partial \ln p(\mathbf{X};\underline{\mathbf{b}})}{\partial \mathbf{b}} = \left[\frac{\partial \ln p(\mathbf{X};\underline{\mathbf{b}})}{\partial b_1} \;\; \frac{\partial \ln p(\mathbf{X};\underline{\mathbf{b}})}{\partial b_2} \;\; \cdots \;\; \frac{\partial \ln p(\mathbf{X};\underline{\mathbf{b}})}{\partial b_M}\right]^T,$$

$$\frac{\partial \ln p(\mathbf{X};\underline{\mathbf{b}})}{\partial \mathbf{b}^*} = \left[\frac{\partial \ln p(\mathbf{X};\underline{\mathbf{b}})}{\partial b_1^*} \;\; \frac{\partial \ln p(\mathbf{X};\underline{\mathbf{b}})}{\partial b_2^*} \;\; ... \;\; \frac{\partial \ln p(\mathbf{X};\underline{\mathbf{b}})}{\partial b_M^*}\right]^T,$$

are based on Wirtinger derivatives. We next find each element $\frac{\partial \ln p(\mathbf{X};\mathbf{b})}{\partial b_k}$ as follows.

Firstly,

$$\begin{aligned}\ln p(\mathbf{X};\underline{\mathbf{b}}) &= \ln \prod_{n=0}^{N-1} p(\mathbf{x}_n;\underline{\mathbf{b}}) \\ &= \ln\left[\frac{1}{\pi^{NL}\prod_{n=0}^{N-1}\det(\mathbf{C})}\exp(-\sum_{n=0}^{N-1}\mathbf{x}_n^H\mathbf{C}^{-1}\mathbf{x}_n)\right] \\ &= \ln(\frac{1}{\pi^{NL}}) - \sum_{n=0}^{N-1}\mathbf{x}_n^H\mathbf{D}^H\mathbf{D}\mathbf{x}_n + N\ln\det(\mathbf{D}^H\mathbf{D}),\end{aligned} \tag{4.2}$$

and

$$\begin{aligned}\frac{\partial \ln p(\mathbf{X};\underline{\mathbf{b}})}{\partial b_k} &= N\frac{\partial \ln\det(\mathbf{D}^H\mathbf{D})}{\partial b_k} - \sum_{n=0}^{N-1}\frac{\partial \mathbf{x}_n^H\mathbf{D}^H\mathbf{D}\mathbf{x}_n}{\partial b_k} \\ &= N\frac{\partial \ln\det(\mathbf{D}^H\mathbf{D})}{\partial b_k} - \sum_{n=0}^{N-1}\frac{\partial \mathrm{tr}(\mathbf{D}\mathbf{x}_n\mathbf{x}_n^H\mathbf{D}^H)}{\partial b_k},\end{aligned} \tag{4.3}$$

for $k = 1, 2, \cdots, M$, where

$$\frac{\partial \ln\det(\mathbf{D}^H\mathbf{D})}{\partial b_k} = \mathrm{tr}(\mathbf{D}^{-1}\mathbf{\Phi}_k), \tag{4.4}$$

and

$$\frac{\partial \mathrm{tr}(\mathbf{D}\mathbf{x}_n\mathbf{x}_n^H\mathbf{D}^H)}{\partial b_k} = \mathrm{tr}(\mathbf{x}_n\mathbf{x}_n^H\mathbf{D}^H\mathbf{\Phi}_k). \tag{4.5}$$

Thus,

$$\frac{\partial \ln p(\mathbf{X};\underline{\mathbf{b}})}{\partial b_k} = N\mathrm{tr}(\mathbf{D}^{-1}\mathbf{\Phi}_k) - \sum_{n=0}^{N-1}\mathrm{tr}(\mathbf{x}_n\mathbf{x}_n^H\mathbf{D}^H\mathbf{\Phi}_k), \tag{4.6}$$

Under $\mathcal{H}_0$, where $\mathbf{b} = \mathbf{0}$,

$$\left.\frac{\partial \ln p(\mathbf{X};\underline{\mathbf{b}})}{\partial b_k}\right|_{\underline{\mathbf{b}}=\mathbf{0}} = N\mathrm{tr}(\mathbf{D}_B^{-1}\mathbf{\Phi}_k) - \sum_{n=0}^{N-1}\mathrm{tr}(\mathbf{\Phi}_k\mathbf{x}_n\mathbf{x}_n^H\mathbf{D}_B^H) \tag{4.7}$$

Also, we have

$$\frac{\partial \ln p(\mathbf{X};\underline{\mathbf{b}})}{\partial b_k^*} = N\mathrm{tr}(\mathbf{D}^{-H}\mathbf{\Phi}_k^H) - \sum_{n=0}^{N-1}\mathrm{tr}(\mathbf{D}\mathbf{x}_n\mathbf{x}_n^H\mathbf{\Phi}_k^H), \tag{4.8}$$

and its value under $\mathcal{H}_0$

$$\left.\frac{\partial \ln p(\mathbf{X};\underline{\mathbf{b}})}{\partial b_k^*}\right|_{\mathbf{b}=\mathbf{0}} = N\mathrm{tr}(\mathbf{D}_B^{-H}\mathbf{\Phi}_k^H) - \sum_{n=0}^{N-1}\mathrm{tr}(\mathbf{D}_B\mathbf{x}_n\mathbf{x}_n^H\mathbf{\Phi}_k^H) \tag{4.9}$$

We next compute $\mathbf{I}(\underline{\mathbf{b}})$.

$$\begin{aligned}\mathbf{I}(\underline{\mathbf{b}}) &= E\left(\frac{\partial \ln p(\mathbf{X};\underline{\mathbf{b}})}{\partial \underline{\mathbf{b}}^*}\frac{\partial \ln p(\mathbf{X};\underline{\mathbf{b}})^H}{\partial \underline{\mathbf{b}}^*}\right)\\ &= \begin{bmatrix}\mathbf{A} & \mathbf{B}^*\\ \mathbf{B} & \mathbf{A}^*\end{bmatrix}\\ &= \begin{bmatrix}M\times M & M\times M\\ M\times M & M\times M\end{bmatrix}\end{aligned} \tag{4.10}$$

where,

$$\begin{aligned}\mathbf{A} &= E\left(\frac{\partial \ln p(\mathbf{X};\underline{\mathbf{b}})}{\partial \mathbf{b}^*}\frac{\partial \ln p(\mathbf{X};\underline{\mathbf{b}})^H}{\partial \mathbf{b}^*}\right)\\ \mathbf{B} &= E\left(\frac{\partial \ln p(\mathbf{X};\underline{\mathbf{b}})}{\partial \mathbf{b}}\frac{\partial \ln p(\mathbf{X};\underline{\mathbf{b}})^T}{\partial \mathbf{b}}\right)\end{aligned}$$

For each element $[\mathbf{A}]_{k,l}$ and $[\mathbf{B}]_{k,l}$ for $1 \le k,l \le M$, we can compute as follows,

$$\begin{aligned}\mathbf{A}_{k,l} &= -E\left(\frac{\partial^2 \ln p(\mathbf{X};\underline{\mathbf{b}})}{\partial b_k^*\partial b_l}\right)\\ &= E\left(\sum_{n=0}^{N-1}\mathrm{tr}(\mathbf{\Phi}_l\mathbf{x}_n\mathbf{x}_n^H\mathbf{\Phi}_k^H)\right)\\ &= N\mathrm{tr}(\mathbf{\Phi}_l\mathbf{D}^{-1}\mathbf{D}^{-H}\mathbf{\Phi}_k^H)\end{aligned} \tag{4.11}$$

Under $\mathcal{H}_0$, where $\mathbf{b}=\mathbf{0}$, we have

$$\mathbf{A}_{k,l}|_{\mathbf{b}=\mathbf{0}} = N\mathrm{tr}(\mathbf{\Phi}_l\mathbf{D}_B^{-1}\mathbf{D}_B^{-H}\mathbf{\Phi}_k^H) \tag{4.12}$$

In a similar fashion, we have

$$\begin{aligned}\mathbf{B}_{k,l} &= -E\left(\frac{\partial^2 \ln p(\mathbf{X};\underline{\mathbf{b}})}{\partial b_k\partial b_l}\right)\\ &= N\mathrm{tr}(\mathbf{D}^{-1}\mathbf{\Phi}_l\mathbf{D}^{-1}\mathbf{\Phi}_k)\end{aligned} \tag{4.13}$$

and its value under $\mathcal{H}_0$

$$\mathbf{B}_{k,l}|_{\mathbf{b}=\mathbf{0}} = N\text{tr}(\mathbf{D}_B^{-1}\mathbf{\Phi}_l\mathbf{D}_B^{-1}\mathbf{\Phi}_k) \tag{4.14}$$

Using equations (4.7), (4.9), (4.10), (4.12), (4.14) and the complex parameter Rao test equation (4.1) will produce the Rao test statistic.

An explicit example is presented next to evaluate the performance of the detector.

4.4 Numerical Examples and Computer Simulations

Consider a simple example, where we only have the $N = 4$ observed data set $\mathbf{X} = [\mathbf{x}_0^T\ \mathbf{x}_1^T\ \mathbf{x}_2^T\ \mathbf{x}_3^T]^T$, each $\mathbf{x}_n$'s is a 4×1 complex-valued IID Gaussian vector, $\mathbf{x}_n \sim \mathcal{CN}(\mathbf{0}, \mathbf{C})$. Also, $\mathbf{C}^{-1} = \mathbf{D}^H\mathbf{D}$, and $\mathbf{D} = \mathbf{D}_B + b_1\mathbf{\Phi}_1$ with $\mathbf{\Phi}_1 = \mathbf{e}_4\mathbf{e}_1^T$ and

$$\mathbf{D}_B = \begin{bmatrix} 0.45 & 0 & 0 & 0 \\ -0.25+0.25j & 0.5 & 0 & 0 \\ -0.12+0.12j & -0.3+0.3j & 0.55 & 0 \\ 0 & -0.15-0.15j & 0.2-0.2j & 0.6 \end{bmatrix}$$

We are testing if the cholesky factor matrix $\mathbf{D}$ is banded and equal to the known $\mathbf{D}_B$. It is equivalent to testing if $b_1 = 0$ versus $b_1 \neq 0$. The Rao test for this example can be shown to be (4.15).

To evaluate the Rao test performance for this example, we consider three cases under the alternative hypothesis $\mathcal{H}_1$, $b_1 = 0.8 - j; b_1 = 0.5 + 0.5j; b_1 = -0.2 + 0.4j$ respectively. The receiver operating characteristic (ROC)s, showing the relationship of the probability of detection (P_d) versus the probability of false alarm (P_{fa}) of the derived Rao test is given in Figure 4.1.

The Rao test statistic under the null hypothesis $\mathcal{H}_0$ is chi-squared distributed

$$T_R(\mathbf{X}) = \frac{\mathrm{Re}\left\{\left(4\mathrm{tr}(\mathbf{D}_B^{-1}\mathbf{\Phi}_1) - \sum_{n=0}^{3}\mathrm{tr}(\mathbf{x}_n\mathbf{x}_n^H\mathbf{D}_B^H\mathbf{\Phi}_1)\right)^2 \mathrm{tr}^*(\mathbf{D}_B^{-1}\mathbf{\Phi}_1\mathbf{D}_B^{-H}\mathbf{\Phi}_1^H)\right\}}{2\left[|\mathrm{tr}(\mathbf{\Phi}_1\mathbf{D}_B^{-1}\mathbf{D}_B^{-H}\mathbf{\Phi}_1^H)|^2 - |\mathrm{tr}(\mathbf{D}_B^{-1}\mathbf{\Phi}_1\mathbf{D}_B^{-1}\mathbf{\Phi}_1)|^2\right]} - \frac{\mathrm{Re}\left\{\left(4\mathrm{tr}(\mathbf{D}_B^{-1}\mathbf{\Phi}_1) - \sum_{n=0}^{3}\mathrm{tr}(\mathbf{x}_n\mathbf{x}_n^H\mathbf{D}_B^H\mathbf{\Phi}_1)\right)^2 \mathrm{tr}(\mathbf{D}_B^{-1}\mathbf{\Phi}_1\mathbf{D}_B^{-1}\mathbf{\Phi}_1)\right\}}{2\left[|\mathrm{tr}(\mathbf{\Phi}_1\mathbf{D}_B^{-1}\mathbf{D}_B^{-H}\mathbf{\Phi}_1^H)|^2 - |\mathrm{tr}(\mathbf{D}_B^{-1}\mathbf{\Phi}_1\mathbf{D}_B^{-1}\mathbf{\Phi}_1)|^2\right]} \tag{4.15}$$

with one degree of freedom, $T_R(\mathbf{X}) \sim \chi_2^2$. The performance of the Rao test can be found asymptotically or as $N \to \infty$. An estimated probability density function (PDF), shown as a bar plot, and the theoretical PDF ($N \to \infty$) are shown in Figure 4.2.

4.5 Conclusions

The banding technique have become an important technique in high-dimensional covariance matrix estimation with a limited number of samples. However, before adopting the technique, it is important to test if the matrix is "bandable". We have introduced the Rao test of bandedness of Cholesky factor matrix of inverse of the covariance matrix in this paper.The Rao test's computational cost is relatively lower than other detectors such as GLRT, yet with reasonably good performance. A concise form of the Rao test for testing bandedness of a complex-valued covariance matrix with complex-valued unknown parameters is present. An example and a simulation are also given to evaluate the proposed detector. The method can be easily applied to the real-valued covariance matrix and parameters case. Moreover, the detector can be applied to test if any element is zero in a

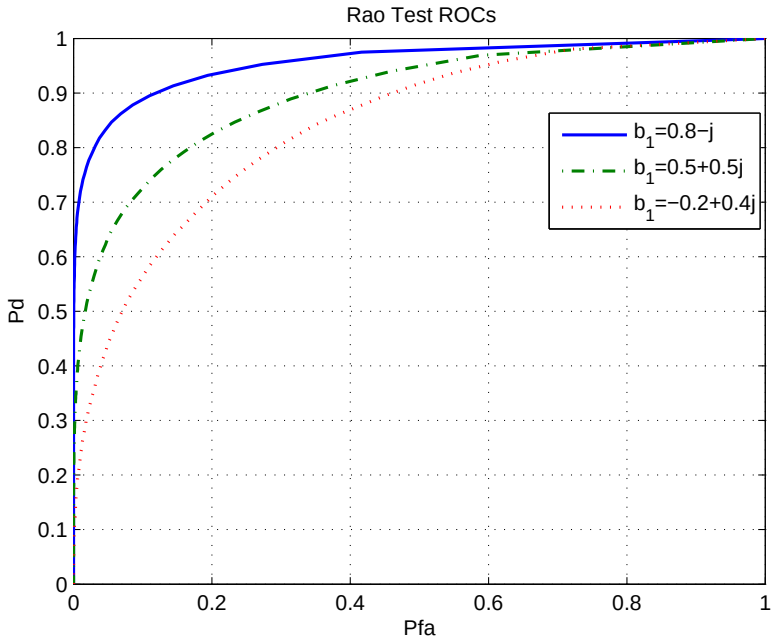

Figure 4.1. ROC curve of the Rao test detector with different b_1

matrix, by changing the basis matrix $\mathbf{\Phi}_k$ accordingly. The derived detector can be used as a pre-processing stage before adopting banding, or certain modeling method, such as ARMA modeling techniques in covariance matrix estimation.

List of References

[1] Melvin, W.L., "A STAP Overview", *IEEE AES Systems Magazine Special Tutorials Issue*, Vol. 19, No.1, Jan. 2004 , pp. 19-35

[2] Melvin, W. L. and Showman, G. A. "An approach to knowledge-aided covariance estimation". *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 42, No. 3, July 2006, pp. 1021-1042.

[3] T. Anderson, *An Introduction to Multivariate Statistical Analysis*. 3rd ed. Wiley, Hoboken, NJ.

[4] Wu, W. B. and Pourahmadi, M. "Nonparametric estimation of large covariance matrices of longitudinal data",*Biometrika* 2003,vol.90, pp.831-844.

[5] Bickel, P. J. and Levina, E. " Regularized estimation of large covariance matrices", *Ann. Statist.*2008, vol.36 pp.199-227.

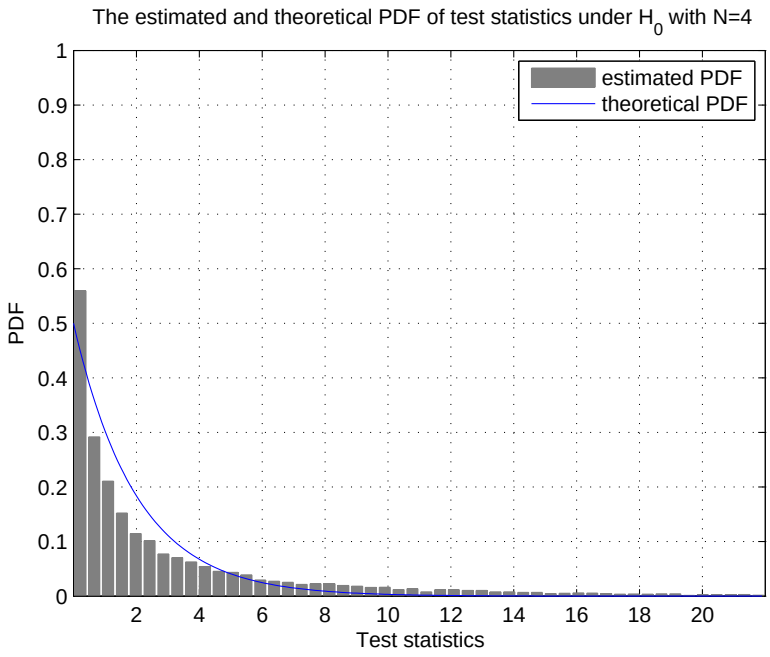

Figure 4.2. Estimated and theoretical PDF of the test statistic for the case $N = 4$

[6] Qiu, Y-M and Chen, S. X. "Test for bandedness of high dimensional covariance matrices with bandwidth estimation".*Ann. Stat.* 2012, vol.40, pp.1285-1314.

[7] S. Kay and Z. Zhu, "The complex parameter Rao Test", *IEEE Transactions on Signal Processing,* vol. 64, no. 24, pp. 6580–6588, Dec. 2016.

[8] A. Wiesel, O. Bibi, and A. Globerson, "Time varying autoregressive moving average models for covariance estimation", *IEEE Transactions on Signal Processing*, vol. 61, no. 11, pp. 2791-2801, 2013

[9] S. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1998.

MANUSCRIPT 5

On the Bayesian Exponentially Embedded Family Rule for Model Order Selection

by

Zhenghan Zhu and Steven Kay

submitted to

IEEE Transactions on Signal Processing

Abstract

In this paper, we derive a Bayesian model order selection rule by using the exponentially embedded family (EEF) method, termed *Bayesian EEF*. Unlike many other Bayesian model selection methods, the *Bayesian EEF* can use vague proper priors and improper non-informative priors to be objective in the elicitation of parameter priors. Moreover, the penalty term of the rule is shown to be the sum of half of the parameter dimension and the estimated mutual information between parameter and observed data. This helps to reveal the EEF mechanism in selecting model orders and may provide new insights into the open problem of choosing an optimal penalty term for model order selection and a good prior from information-theoretic viewpoints. The important example of linear model order selection is given to illustrate the algorithms and arguments. Lastly, the *Bayesian EEF* that uses Jeffreys' prior coincides with the EEF rule derived by frequentist strategies. This shows another interesting relationship between the frequentist and Bayesian philosophies for model selection.

5.1 Introduction

Model order selection is an important problem of active research in signal processing. It finds a wide range of applications. For example, determination of the number of sources in array signal processing [1] is essentially a model order selection problem. Overestimating the order fits the noise in the data; underestimating the order, on the other hand, fails to describe the data precisely [1]. Hence, a good model order selection rule is crucial for signal processing applications.

As a multiple hypotheses testing problem, model order selection lacks an optimal solution [9]. The generalized likelihood ratio test (GLRT) always favors the more complex model [7]. A typical model order selection algorithm introduces a penalty term to the GLRT, and it is the penalty term that makes one model order selection rule different from another. A model order selection rule derived from a Bayesian viewpoint typically tries to strike a balance between goodness of fit and model complexity [18].

Some leading algorithms, both frequentist and Bayesian, in literature [5] are Akaike's information criterion (AIC) [2], the minimum description length (MDL) [3], Bayesian information criterion (BIC) [4] and maximum a posteriori (MAP) [9]. For example, AIC and BIC rules are respectively

$$\ln p(\mathbf{x}|\hat{\boldsymbol{\theta}}) - k; \qquad \text{AIC}$$

$$\ln p(\mathbf{x}|\hat{\boldsymbol{\theta}}) - \tfrac{k}{2}\ln L; \quad \text{BIC}$$

where $\ln p(\mathbf{x}|\hat{\boldsymbol{\theta}})$ is the maximum log-likelihood under a certain model, k is the dimension of the model parameters, L is the data record length. As seen the AIC penalty is a constant k and BIC has a penalty $\frac{k}{2}\ln L$.

As an alternative, an EEF model order selection rule derived from a frequentist viewpoint is introduced in [8]. It is consistent and superior to others for several situations and has been adopted in many applications such as source enumeration, classification and sensor fusion [1],[12]-[14]. Different from [8], we derive in this paper the EEF rule from a Bayesian viewpoint, termed the *Bayesian EEF*, as a novel Bayesian model order selection rule. Using Bayesian strategies allows us the

possibilities to investigate the EEF mechanism in a new framework and from new viewpoints such as information theory and leads to the main contributions of this paper:

- A new Bayesian model order selection method, Bayesian EEF, is derived. It is proved that the Bayesian EEF can use both vague proper prior and improper non-informative prior for unknown parameters, both of which are usually forbidden for many Bayesian methods. The Bayesian EEF also does not have th Lindley's paradox or the Information paradox.

- An intuitive justification is given in interpreting the Bayesian EEF penalty term. The penalty term is a sum of half the model parameter dimension and the estimated mutual information between model parameters and observed data.

- It also shows that the Bayesian EEF using Jeffrey's prior coincides with the EEF derived from a frequentist viewpoint. This is another case of the interesting interaction between the frequentist and Bayesian philosophies and may provide useful insights into the discussion on the difference between the two.

The paper is organized as follows. In Section 5.2 we derive the Bayesian EEF order selection rule that uses a vague proper prior for linear model and discuss some desirable properties of the Bayesian EEF. In Section 5.3 we justify the Bayesian EEF penalty term. In Section 5.4 we derive the Bayesian EEF via improper non-

informative prior, Jeffreys' prior and discuss its interaction with frequentist EEF. Finally, some conclusions are given in Section 5.5.

5.2 Bayesian EEF rule for model selection via vague proper prior

Suppose there are M candidate models, $\mathcal{M}_0, \mathcal{M}_1, \cdots, \mathcal{M}_{M-1}$, where $\mathcal{M}_0$ is a null/reference model which has no unknown parameters and the model $\mathcal{M}_i$ (for $i = 1, \cdots, M-1$) has an unknown parameter vector $\boldsymbol{\theta}_i$ of dimension $k_i \times 1$. The probability density functions (PDF) of the observed data $\mathbf{x}$ of dimension $N \times 1$ for model $\mathcal{M}_i$ is denoted as $p_i(\mathbf{x})$. From the frequentist viewpoint, the unknown parameters are deterministic. The EEF model order selection rule proposed in [8] adopts this assumption and hence is termed *frequentist* EEF in this paper. On the other hand, a Bayesian model order selection method views the parameter vectors as random. The Bayesian EEF adopts this philosophy. If we know the the model parameter priors, we can compare marginal PDFs of $\mathbf{x}$ of different models or use a MAP rules to choose a model order. But in practice no prior information is available and the first question that arises for a Bayesian model order selection method is the specification of the prior distributions for the unknown parameter vector $\boldsymbol{\theta}_i$. Which prior to choose is a controversial and difficult task [17]. Ideally we want to use a prior with minimal influence on the Bayesian inference. Improper non-informative priors such as uniform distribution and vague prior distributions (a proper prior with large spread) seem to be natural choices because they are objective in that they do not favor one parameter value over another. However, they can, unfortunately, lead to non-sensible answers when used in many Bayesian

model selection methods. As shown next Bayesian EEF, on the other hand, can employ these two types of priors and still produce good results. This is a desirable property for a Bayesian model order selection algorithm. In this section, we derive the Bayesian EEF by assigning vague proper priors to unknown parameters. The resultant EEF is called the *reduced Bayesian EEF*. For illustration purposes, we focus on the normal linear model order selection problem. In Section 5.5, we give the Bayesian EEF that uses the improper non-informative Jeffreys' prior.

The vague proper prior adopted herein is constructed by letting the hyper-parameter of a g-prior goes to infinity. G-prior is widely used in Bayesian inference because of its conjugacy and computational efficiency in computing the marginal likelihoods and its simple, understandable interpretation [16][21]. The g-prior places less prior distribution mass in areas of the parameter space where the data is expected to be more informative about the unknown parameters. Assume we want to choose a model from the following linear model candidates

$$\mathcal{M}_i : \mathbf{x} = \mathbf{H}_i \boldsymbol{\theta}_i + \mathbf{w}, \ i = 1, \cdots, M-1.$$

where $\boldsymbol{\theta}_i$ is a $k_i \times 1$ unknown parameter vector, $\mathbf{H}_i$ is a $N \times k_i$ design matrix, and $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ is additive noise with $\mathbf{I}$ being a $N \times N$ identity matrix. There is also a null model $\mathcal{M}_0 : \mathbf{x} = \mathbf{w}$ which does not contain unknown parameters. Without loss of generality, we assume that $k_i \leq k_j$ for $i \leq j$.

We first assign $\boldsymbol{\theta}_i$ a vague proper prior, $\pi_i(\boldsymbol{\theta}_i)$, which is a g-prior with an

infinite hypeparameter g_i [21] as

$$\pi_i(\boldsymbol{\theta}_i) = \mathcal{N}\left(\mathbf{0}, g_i\sigma^2(\mathbf{H}_i^T\mathbf{H}_i)^{-1}\right) \text{ and } g_i \to \infty.$$

The marginal PDF $p_i(\mathbf{x})$ under the $\mathcal{M}_i$ model is then

$$\begin{aligned} p_i(\mathbf{x}) &= \int p_i(\mathbf{x}|\,\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)d\,\boldsymbol{\theta}_i \\ &= \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I} + g_i\sigma^2\mathbf{P}_i) \text{ and } \mathbf{P}_i = \mathbf{H}_i(\mathbf{H}_i^T\mathbf{H}_i)^{-1}\mathbf{H}_i^T \\ &= \mathcal{N}(\mathbf{0}, \mathbf{C}_i) \end{aligned} \tag{5.1}$$

where $p_i(\mathbf{x}|\,\boldsymbol{\theta}_i) = \mathcal{N}\left(\mathbf{H}_i\,\boldsymbol{\theta}_i, \sigma^2\mathbf{I}\right)$ is the conditional PDF of $\mathbf{x}$ on $\boldsymbol{\theta}_i$ under model $\mathcal{M}_i$ and the covariance matrix $\mathbf{C}_i = \sigma^2\mathbf{I} + g_i\sigma^2\mathbf{P}_i$. The PDF of $\mathbf{x}$ under the null model is

$$p_0(\mathbf{x}) = \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}) = \mathcal{N}(\mathbf{0}, \mathbf{C}_0), \tag{5.2}$$

where $\mathbf{C}_0 = \sigma^2\mathbf{I}$ is the covariance matrix of $p_0(\mathbf{x})$. Then for each $p_i(\mathbf{x})$, $i = 1, \cdots, M-1$, we can construct a new PDF, $p(\mathbf{x}; \eta_i)$ by exponentially embedding it with $p_0(\mathbf{x})$, which is parameterized by an embedding parameter η_i:

$$\begin{aligned} p(\mathbf{x}; \eta_i) &= \frac{p_i^{\eta_i}(\mathbf{x})p_0^{1-\eta_i}(\mathbf{x})}{\int p_i^{\eta_i}(\mathbf{x})p_0^{1-\eta_i}(\mathbf{x})d\mathbf{x}} \\ &= \exp(\eta_i T_i(\mathbf{x}) - K_0(\eta_i) + f_c(\mathbf{x})) \end{aligned} \tag{5.3}$$

with

$$\begin{aligned} &\text{sufficient statistic:} && T_i(\mathbf{x}) = \ln\frac{p_i(\mathbf{x})}{p_0(\mathbf{x})} \\ &\text{natural parameter:} && 0 \le \eta_i \le 1 \\ &\text{log-normalizer:} && K_0(\eta_i) = \ln\int p_i^{\eta_i}(\mathbf{x})p_0^{1-\eta_i}(\mathbf{x})d\mathbf{x} = \ln E_0\left(e^{\eta_i T_i(\mathbf{x})}\right) \\ &\text{carrier density:} && f_c(\mathbf{x}) = \ln p_0(\mathbf{x}) \end{aligned}$$

As shown the resulting PDF is an exponential family PDF and consequently inherits a multitude of mathematical and practical properties of the family. For example the statistic $T_i(\mathbf{x})$ is a minimal and complete sufficient statistic for η_i; its moments can be easily found and $K_0(\eta_i)$ is a convex function. The new PDF $p(\mathbf{x};\eta_i)$ is called the Bayesian EEF for the model $\mathcal{M}_i$ in that we employ both Bayesian philosophies and exponentially embedding to construct it. From the information-geometric viewpoints, the log-Bayesian EEF $\ln p(\mathbf{x};\eta_i)$ can be viewed as a point on the geodesic that connects $\ln p_i(\mathbf{x})$ and $\ln p_0(\mathbf{x})$ [8][10]. As seen from (5.3), the Bayesian EEF $p(\mathbf{x};\eta_i)$ reduces to $p_0(\mathbf{x})$ when $\eta_i = 0$ and $p_i(\mathbf{x})$ when $\eta_i = 1$.

Plugging $p_i(\mathbf{x})$ of (5.1) and $p_0(\mathbf{x})$ of (5.2) into (5.3) produces the reduced Bayesian EEF $p(\mathbf{x};\eta_i)$ for the linear model as follows.

$$\begin{aligned} p(\mathbf{x};\eta_i) &= \frac{p_i^{\eta_i}(\mathbf{x})p_0^{1-\eta_i}(\mathbf{x})}{\exp(K_0(\eta_i))} \\ &= \frac{\left[\frac{1}{\sqrt{|2\pi\mathbf{C}_i|}}\exp(-\frac{1}{2}\mathbf{x}^T\mathbf{C}_i^{-1}\mathbf{x})\right]^{\eta_i}\left[\frac{1}{\sqrt{|2\pi\mathbf{C}_0|}}\exp(-\frac{1}{2}\mathbf{x}^T\mathbf{C}_0^{-1}\mathbf{x})\right]^{1-\eta_i}}{\exp(K_0(\eta_i))} \\ &= c_1\exp\big[-\frac{1}{2}\mathbf{x}^T\underbrace{\left(\eta_i\mathbf{C}_i^{-1}+(1-\eta_i)\mathbf{C}_0^{-1}\right)}_{\mathbf{C}_{\eta_i}^{-1}}\mathbf{x}\big] \end{aligned}$$

where c_1 is a constant normalization term and $\mathbf{C}_{\eta_i} = \left(\eta_i\mathbf{C}_i^{-1}+(1-\eta_i)\mathbf{C}_0^{-1}\right)^{-1}$. It shows that the constructed EEF is also a zero mean normal distribution with a

covariance matrix $\mathbf{C}_{\eta_i}$ depending on η_i. Explicitly,

$$\begin{aligned}
\mathbf{C}_{\eta_i} &= \left[\eta_i(\sigma^2\mathbf{I} + g_i\sigma^2\mathbf{P}_i)^{-1} + (1-\eta_i)(\sigma^2\mathbf{I})^{-1}\right]^{-1} \\
&= \sigma^2\left[\eta_i(\mathbf{I} - \frac{g_i}{g_i+1}\mathbf{P}_i) + (1-\eta_i)\mathbf{I}\right]^{-1} \\
&= \sigma^2\left(\mathbf{I} - \frac{\eta_i g_i}{g_i+1}\mathbf{P}_i\right)^{-1} \\
&= \sigma^2\left(\mathbf{I} + \frac{\eta_i}{1-\eta_i+\frac{1}{g_i}}\mathbf{P}_i\right) \\
&\to \sigma^2\mathbf{I} + \frac{\eta_i}{1-\eta_i}\sigma^2\mathbf{P}_i \text{ as } g_i \to \infty
\end{aligned}$$

So the reduced Bayesian EEF for $\mathcal{M}_i$ is

$$p(\mathbf{x};\eta_i) = \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I} + \frac{\eta_i}{1-\eta_i}\sigma^2\mathbf{P}_i). \tag{5.4}$$

Then a model order selection algorithm based on the Bayesian EEF in (5.4) consists of two steps.

- Step1: Find the MLE of η_i, $0 \le \hat{\eta}_i \le 1$, which maximizes $p(\mathbf{x};\eta_i)$;

  For the linear model EEF in (5.4) we have

$$\hat{\eta}_i = \begin{cases} 0 & \text{if } \mathbf{x}^T\mathbf{P}_i\mathbf{x} < k_i\sigma^2 \\ \frac{\mathbf{x}^T\mathbf{P}_i\mathbf{x} - k_i\sigma^2}{\mathbf{x}^T\mathbf{P}_i\mathbf{x}} & \text{otherwise} \end{cases} \tag{5.5}$$

  where k_i is the dimension of $\boldsymbol{\theta}_i$.

- Step2: Compare the values of the $M-1$ maximized EEF $p(\mathbf{x};\hat{\eta}_i)$ or equivalently the log-likelihood ratio (LLR) $\ln\frac{p(\mathbf{x};\hat{\eta}_i)}{p_0(\mathbf{x})}$ and choose the model which is

associated with the maximum value.

For the linear model, plugging $\hat{\eta}_i$ into (5.4) produces the maximized LLR

$$\ln \frac{p(\mathbf{x};\hat{\eta}_i)}{p_0(\mathbf{x})} = \left(\frac{\mathbf{x}^T\mathbf{P}_i\mathbf{x}}{2\sigma^2} - \frac{k_i}{2} - \frac{k_i}{2} \ln \frac{\frac{\mathbf{x}^T\mathbf{P}_i\mathbf{x}}{\sigma^2}}{k_i} \right) u\left(\frac{\mathbf{x}^T\mathbf{P}_i\mathbf{x}}{2\sigma^2} - \frac{k_i}{2} \right).$$

where $u(\cdot)$ is a unit step function. In fact, the term $\frac{\mathbf{x}^T\mathbf{P}_i\mathbf{x}}{2\sigma^2}$ is the maximized LRT of the conditional PDF $p_i(\mathbf{x}|\hat{\boldsymbol{\theta}}_i)$ and $p_0(\mathbf{x})$, termed as l_{G_i}:

$$\begin{aligned} l_{G_i} &= \ln \frac{\max_{\boldsymbol{\theta}_i} p_i(\mathbf{x}|\boldsymbol{\theta}_i)}{p_0(\mathbf{x})} \\ &= \ln \frac{\max_{\boldsymbol{\theta}_i} \frac{1}{\sqrt{|2\pi\sigma^2\mathbf{I}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{H}_i\,\boldsymbol{\theta}_i)^T(\sigma^2\mathbf{I})^{-1}(\mathbf{x}-\mathbf{H}_i\,\boldsymbol{\theta}_i)\right)}{\frac{1}{\sqrt{|2\pi\sigma^2\mathbf{I}|}} \exp\left(-\frac{1}{2}\mathbf{x}^T(\sigma^2\mathbf{I})^{-1}\mathbf{x}\right)} \\ &= \ln \frac{p_i(\mathbf{x}|\hat{\boldsymbol{\theta}}_i)}{p_0(\mathbf{x})} \text{ with } \hat{\boldsymbol{\theta}}_i = (\mathbf{H}_i^T\mathbf{H}_i)^{-1}\mathbf{H}_i^T\mathbf{x} \\ &= \frac{\mathbf{x}^T\mathbf{P}_i\mathbf{x}}{2\sigma^2} \end{aligned}$$

In summary, we can write the linear model Bayesian EEF as

$$\ln \frac{p(\mathbf{x};\hat{\eta}_i)}{p_0(\mathbf{x})} = \left(l_{G_i} - \frac{k_i}{2} - \frac{k_i}{2} \ln \frac{l_{G_i}}{k_i/2} \right) u\left(l_{G_i} - \frac{k_i}{2} \right). \tag{5.6}$$

5.2.1 Rationale of Bayesian EEF model order selection algorithm

We now present the rationale for Bayesian EEF model order selection algorithm given above. First, when η_i is chosen as its MLE $\hat{\eta}_i$,

$$\frac{\partial \ln p(\mathbf{x};\eta_i)}{\partial \eta_i} = T_i(\mathbf{x}) - K_0'(\eta_i) = 0$$

follows from (5.3). That is $T_i(\mathbf{x}) = K_0'(\eta_i)$ evaluated at $\eta_i = \hat{\eta}_i$. Moreover, it holds in general $\int p(\mathbf{x};\eta_i)T_i(\mathbf{x})d\mathbf{x} = K_0'(\eta_i)$ for the exponential family [8]. Therefore

$$\left[\int p(\mathbf{x};\eta_i)T_i(\mathbf{x})d\mathbf{x} \right] \Big|_{\eta_i \to \hat{\eta}_i} = T_i(\mathbf{x}) \tag{5.7}$$

And consequently we have

$$\begin{aligned}
\mathrm{KL}(p(\mathbf{x};\hat{\eta}_i)||p_0(\mathbf{x})) &= \int p(\mathbf{x};\hat{\eta}_i)\ln\frac{p(\mathbf{x};\hat{\eta}_i)}{p_0(\mathbf{x})}d\mathbf{x} \\
&= \int p(\mathbf{x};\hat{\eta}_i)\left[\hat{\eta}_i T_i(\mathbf{x}) - K_0(\hat{\eta}_i)\right]d\mathbf{x} \\
&= \hat{\eta}_i T_i(\mathbf{x}) - K_0(\hat{\eta}_i) \\
&= \ln\frac{p(\mathbf{x};\hat{\eta}_i)}{p_0(\mathbf{x})} \qquad (5.8)
\end{aligned}$$

where $\mathrm{KL}(\cdot||\cdot)$ denotes Kullback Libler divergence (KLD).

Moreover, a Pythagorean-like relationship holds asymptotically for large data record among KLD quantities for EEF [8]

$$\mathrm{KL}(p_t(\mathbf{x})||p(\mathbf{x};\hat{\eta}_i)) = \mathrm{KL}(p_t(\mathbf{x})||p_0(\mathbf{x})) - \mathrm{KL}(p(\mathbf{x};\hat{\eta}_i)||p_0(\mathbf{x})),$$

where $p_t(\mathbf{x})$ denotes the true PDF of the data, which is unknown but fixed. The distance $\mathrm{KL}(p_t(\mathbf{x})||p_0(\mathbf{x}))$ is fixed, hence the model that maximizes the distance $\mathrm{KL}(p(\mathbf{x};\hat{\eta}_i)||p_0(\mathbf{x}))$ or equivalently $\ln\frac{p(\mathbf{x};\hat{\eta}_i)}{p_0(\mathbf{x})}$, among all models has the minimum $\mathrm{KL}(p_t(\mathbf{x})||p(\mathbf{x};\hat{\eta}_i))$-the "distance" from the true PDF $p_t(\mathbf{x})$. This is the reason why the Bayesian EEF model selection rule chooses the model with the maximum of the maximized EEF's of all models.

5.2.2 Discussion on paradoxes

The EEF model order selection algorithm has many desirable properties such as consistency [1] and better performances than many other algorithms in the low signal-to-noise ratio regime [8]. In addition to these properties, we now show that the newly derived Bayesian EEF has additional desirable properties-it does not

have *Lindley's paradox* nor the *Information paradox*. On the contrary, many other Bayesian model selection methods based on marginal bayes factor (BF) may suffer from these paradoxes [16]. *Lindley's paradox* can be understood as: "large spread of the prior induced by the non-informative choice of hyper-parameter has the unintended consequence of forcing the BF to favor the null model, the smallest model, regardless of the information in the data [16]". As shown in (5.6), the reduced Bayesian EEF does not necessarily favor the null model even if we let the hyper-parameter $g_i \to \infty$. This indicates that the reduced Bayesian EEF rule has no "Lindley's paradox". The *Information paradox* is "a paradox related to the limiting behavior of the BF. The BF yields a constant even when there is infinite amount of information supporting to choose a model [16]." For instance, the linear model BF resulted from assigning the parameter $\boldsymbol{\theta}_i$ a g-prior with a certain g_i is[16]

$$BF(\mathcal{M}_i : \mathcal{M}_0) = \frac{(1+g_i)^{(N-1-k_i)/2}}{(1+g_i(1-R_r^2))^{(N-1)/2}}$$

where R_r^2 is the ordinary coefficient of determination of the regression model $\mathcal{M}_i$. When there is overwhelming information supporting to choose $\mathcal{M}_i$ instead of $\mathcal{M}_0$, $R_r^2 \to 1$; however, the BF yields a constant $(1+g_i)^{(N-1-k_i)/2}$ instead of infinity. This information limiting behavior is called the information paradox. When $R_r^2 \to 1$ or equivalently $\mathbf{x}^T\mathbf{P}_i\mathbf{x} \gg k_i\sigma^2$ we have $\hat{\eta_i} \to 1$ from (5.5). In this case, the reduced Bayesian EEF $\ln \frac{p(\mathbf{x};\hat{\eta_i})}{p_0(\mathbf{x})}$ in (5.6) also goes to infinity. This shows that the Bayesian EEF has no information limiting behavior and hence no *Information paradox*. In fact, these two nice properties of the Bayesian EEF model selection rule are due to its mechanism of choosing the value of η_i. It uses the MLE $\hat{\eta}_i$ which is dependent

on data.

5.3 The Penalty Term of Reduced Bayesian EEF

The penalty term is the key term for a model order selection rule. Its function is to penalize the maximum log-likelihood with a measure of model complexity so that the model order selection rule can strike a tradeoff between goodness-of-fit and model complexity. In light of the general relationship KLD=SNR-MI [6], the reduced Bayesian EEF penalty term is found to possess a very intuitive interpretation. This not only helps further understanding EEF's mechanism in model selection but also provides new insights into the problem of choosing a good penalty term for model selection. As shown next, the EEF penalty term can be viewed as the sum of a term proportional to the parameter dimension, $\frac{k_i}{2}$, and estimated mutual information between the parameter and received data, $\frac{k_i}{2} \ln \frac{2l_{G_i}}{k_i}$.

First note that if assigning the unknown parameter $\boldsymbol{\theta}_i$ a prior that depends upon the embedding parameter η_i:

$$\pi'(\boldsymbol{\theta}_i; \eta_i) = \mathcal{N}(\mathbf{0}, \frac{\eta_i}{1-\eta_i}\sigma^2(\mathbf{H}_i^T\mathbf{H}_i)^{-1}),$$

the marginal PDF for model $\mathcal{M}_i$ becomes the reduced Bayesian EEF in (5.4)

$$\begin{aligned} p_i(\mathbf{x}) &= \int p_i(\mathbf{x}|\,\boldsymbol{\theta}_i)\pi'(\boldsymbol{\theta}_i;\eta_i)d\,\boldsymbol{\theta}_i \\ &= \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I} + \frac{\eta_i}{1-\eta_i}\sigma^2\mathbf{P}_i\right) \\ &= p(\mathbf{x};\eta_i). \end{aligned}$$

Note that this new $p_i(\mathbf{x})$ is in fact parameterized by η_i because $\pi'(\boldsymbol{\theta}_i;\eta_i)$ depends upon η_i. Hence we denote it as $p_i(\mathbf{x};\eta_i)$. Then we can write $p_{\eta_i}(\mathbf{x}) = p(\mathbf{x};\eta_i)$.

Together with the relationship in (5.8), we have the following decomposition [6] holds for $\eta_i = \hat{\eta}_i$

$$\begin{aligned}
\ln \frac{p(\mathbf{x};\eta_i)}{p_0(\mathbf{x})} &\approx \mathrm{KL}(p(\mathbf{x};\eta_i)||p_0(\mathbf{x})) \\
&= \mathrm{KL}(p_{\eta_i}(\mathbf{x})||p_0(\mathbf{x})) \\
&= \underbrace{\int\int p_{\eta_i}(\mathbf{x},\boldsymbol{\theta}_i) \ln \frac{p_{\eta_i}(\mathbf{x}|\,\boldsymbol{\theta}_i)}{p_0(\mathbf{x})} d\,\boldsymbol{\theta}_i\, d\mathbf{x}}_{\widehat{\mathrm{SNR}}} \\
&\quad \underbrace{- \int\int p_{\eta_i}(\mathbf{x},\boldsymbol{\theta}_i) \ln \frac{p_{\eta_i}(\mathbf{x}|\,\boldsymbol{\theta}_i)}{p_{\eta_i}(\mathbf{x})} d\mathbf{x} d\,\boldsymbol{\theta}_i}_{\widehat{\mathrm{MI}}}
\end{aligned} \tag{5.9}$$

Note $p_{\eta_i}(\mathbf{x},\boldsymbol{\theta}_i)$ denotes the joint PDF of $\mathbf{x}$ and $\boldsymbol{\theta}_i$ and $p_{\eta_i}(\mathbf{x}|\,\boldsymbol{\theta}_i) = \mathcal{N}(\mathbf{H}_i\,\boldsymbol{\theta}_i, \sigma^2\mathbf{I})$ is the conditional PDF. This says that the reduced EEF can be decomposed into two terms. As shown next, the first term is in fact an estimated SNR and hence is denoted as $\widehat{\mathrm{SNR}}$ and the second term is an estimated MI between parameter $\boldsymbol{\theta}_i$ and data $\mathbf{x}$, denoted as $\widehat{\mathrm{MI}}$. Note they are estimated terms in the sense that η_i is replaced by its MLE $\hat{\eta}_i$.

5.3.1 The estimated SNR term

First, we have for $\eta_i = \hat{\eta}_i$

$$\begin{aligned}
\widehat{\mathrm{SNR}} &= \int\int p_{\eta_i}(\mathbf{x}, \boldsymbol{\theta}_i) \ln \frac{p_{\eta_i}(\mathbf{x}|\,\boldsymbol{\theta}_i)}{p_0(\mathbf{x})} d\,\boldsymbol{\theta}_i\, d\mathbf{x} && (5.10)\\
&= \int_{\boldsymbol{\theta}_i} \pi'(\boldsymbol{\theta}_i) \Big[\mathrm{KL}(p_{\eta_i}(\mathbf{x}|\,\boldsymbol{\theta}_i)||p_0(\mathbf{x})) \Big] d\,\boldsymbol{\theta}_i \\
&= \int_{\boldsymbol{\theta}_i} \pi'(\boldsymbol{\theta}_i) \Big[\mathrm{KL}\left(\mathcal{N}(\mathbf{H}_i\,\boldsymbol{\theta}_i, \sigma^2\mathbf{I}) || \,\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})\right) \Big] d\,\boldsymbol{\theta}_i \\
&= \int_{\boldsymbol{\theta}_i} \pi'(\boldsymbol{\theta}_i) \left[\frac{1}{2} \frac{\boldsymbol{\theta}_i^T\, \mathbf{H}_i^T \mathbf{H}_i\, \boldsymbol{\theta}_i}{\sigma^2} \right] d\,\boldsymbol{\theta}_i && (5.11)\\
&= \int_{\boldsymbol{\theta}_i} \frac{e^{-\frac{1}{2}\boldsymbol{\theta}_i^T \left[\frac{\eta_i}{1-\eta_i}\sigma^2(\mathbf{H}_i^T\mathbf{H}_i)^{-1})\right]^{-1}\boldsymbol{\theta}_i}}{\sqrt{\left|2\pi\frac{\eta_i}{1-\eta_i}\sigma^2(\mathbf{H}_i^T\mathbf{H}_i)^{-1}\right|}} \left[\frac{1}{2} \frac{\boldsymbol{\theta}_i^T\, \mathbf{H}_i^T \mathbf{H}_i\, \boldsymbol{\theta}_i}{\sigma^2} \right] d\,\boldsymbol{\theta}_i \,\Big|_{\eta_i=\hat{\eta}_i} && (5.12)\\
&= \frac{1}{2}\frac{\mathbf{x}^T\mathbf{P}_i\mathbf{x}}{\sigma^2} - \frac{k_i}{2} \\
&= l_{G_i} - \frac{k_i}{2} && (5.13)
\end{aligned}$$

where we have used the $\hat{\eta}_i$ in (5.5), treated as a constant, to replace η_i. The eqn (5.11) indicates that the first term is an average ratio of signal energy $||\mathbf{H}_i\,\boldsymbol{\theta}_i\,||^2$ and the noise power σ^2, and indeed is a measure of SNR; furthermore by (5.13) we see that $\widehat{\mathrm{SNR}}$ has introduced a penalty term $k_i/2$, which is proportional to the parameter dimension. In fact, (5.13) holds not only for linear model but in general. First, we can rewrite the $\widehat{\mathrm{SNR}}$ term as

$$\begin{aligned}
\widehat{\mathrm{SNR}} &= \int\int p_{\hat{\eta}_i}(\mathbf{x}, \boldsymbol{\theta}_i) \ln \frac{p_{\hat{\eta}_i}(\mathbf{x}|\,\boldsymbol{\theta}_i)}{p_0(\mathbf{x})} d\,\boldsymbol{\theta}_i\, d\mathbf{x} \\
&= \int_{\mathbf{x}} p_{\hat{\eta}_i}(\mathbf{x}) \int_{\boldsymbol{\theta}_i} \pi(\boldsymbol{\theta}_i\,|\mathbf{x}) \left[\ln \frac{p_{\hat{\eta}_i}(\mathbf{x}|\,\boldsymbol{\theta}_i)}{p_0(\mathbf{x})} \right] d\,\boldsymbol{\theta}_i\, d\mathbf{x}, && (5.14)
\end{aligned}$$

where $\pi(\boldsymbol{\theta}_i\,|\mathbf{x})$ is the posterior distribution of $\boldsymbol{\theta}_i$ after observing $\mathbf{x}$. For large data records we have approximately [15]

$$\pi(\boldsymbol{\theta}_i\,|\mathbf{x}) = \mathcal{N}(\hat{\boldsymbol{\theta}}_i, \mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}_i)),$$

where $\mathbf{I}(\hat{\boldsymbol{\theta}}_i)$ is the Fisher information matrix (FIM) of $\boldsymbol{\theta}_i$ evaluated at its MLE $\hat{\boldsymbol{\theta}}_i$. And using the Laplace approximation we have

$$\begin{aligned}\int_{\boldsymbol{\theta}_i} \pi(\boldsymbol{\theta}_i\,|\mathbf{x}) \ln \frac{p_{\hat{\eta}_i}(\mathbf{x}|\,\boldsymbol{\theta}_i)}{p_0(\mathbf{x})} d\,\boldsymbol{\theta}_i \;&\approx\; \int_{\boldsymbol{\theta}_i} \pi(\boldsymbol{\theta}_i\,|\mathbf{x}) \Big[\underbrace{\ln \frac{p_{\hat{\eta}_i}(\mathbf{x}|\hat{\boldsymbol{\theta}}_i)}{p_0(\mathbf{x})}}_{l_{G_i}} \\ &\qquad -\frac{1}{2}(\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)^T \mathbf{I}(\hat{\boldsymbol{\theta}}_i)(\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i) \Big] d\,\boldsymbol{\theta}_i \\ &= \; l_{G_i} - \frac{k_i}{2}.\end{aligned}$$

Therefore from (5.14)

$$\begin{aligned}\widehat{\mathrm{SNR}} \;&\approx\; \int p_{\hat{\eta}_i}(\mathbf{x}) \left[l_{G_i} - \frac{k_i}{2} \right] d\mathbf{x} \\ &= \; \int p(\mathbf{x}; \hat{\eta}_i)(T_i(\mathbf{x}) - \frac{k_i}{2}) d\mathbf{x} \\ &= \; T_i(\mathbf{x}) - \frac{k_i}{2} \\ &= \; l_{G_i} - \frac{k_i}{2} \end{aligned} \tag{5.15}$$

where we have used $\left[\int p(\mathbf{x}; \eta_i) T_i(\mathbf{x}) d\mathbf{x}\right]\big|_{\eta_i = \hat{\eta}_i} = T_i(\mathbf{x})$ in (5.7). This shows that the difference between l_{G_i} and the estimated SNR is asymptotically half of the parameter dimension.

5.3.2 The estimated mutual information term

We now consider the second term $\widehat{\mathrm{MI}}$ in the decomposition (5.9). For linear model it is $\frac{k_i}{2} \ln \frac{2 l_{G_i}}{k_i}$ as given in (5.6). It is shown next that in general it is the esti-

mated MI between $\boldsymbol{\theta}_i$ and $\mathbf{x}$. First we have from definition of mutual information,

$$
\begin{aligned}
\widehat{\mathrm{MI}} &= \int\int p_{\hat{\eta}_i}(\mathbf{x},\boldsymbol{\theta}_i)\ln\frac{p_{\hat{\eta}_i}(\mathbf{x}|\,\boldsymbol{\theta}_i)}{p_{\hat{\eta}_i}(\mathbf{x})}d\mathbf{x}d\,\boldsymbol{\theta}_i \qquad (5.16)\\
&= \int_{\boldsymbol{\theta}_i}\pi'(\boldsymbol{\theta}_i)|_{\eta_i=\hat{\eta}_i}\int_{\mathbf{x}}p_{\hat{\eta}_i}(\mathbf{x}|\,\boldsymbol{\theta}_i)\ln\frac{p_{\hat{\eta}_i}(\mathbf{x}|\,\boldsymbol{\theta}_i)}{p_{\hat{\eta}_i}(\mathbf{x})}d\mathbf{x}d\,\boldsymbol{\theta}_i\\
&= \int_{\boldsymbol{\theta}_i}\pi'(\boldsymbol{\theta}_i)|_{\eta_i=\hat{\eta}_i}\mathrm{KL}\Big(p_{\hat{\eta}_i}(\mathbf{x}|\,\boldsymbol{\theta}_i)||p_{\hat{\eta}_i}(\mathbf{x})\Big)d\,\boldsymbol{\theta}_i\\
&= \int_{\boldsymbol{\theta}_i}\pi'(\boldsymbol{\theta}_i)|_{\eta_i=\hat{\eta}_i}\mathrm{KL}\Big(\mathcal{N}(\mathbf{H}_i\,\boldsymbol{\theta}_i,\sigma^2\mathbf{I})||\,\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}+\frac{\hat{\eta}_i}{1-\hat{\eta}_i}\sigma^2\mathbf{P}_i)\Big)d\,\boldsymbol{\theta}_i\\
&= \int_{\boldsymbol{\theta}_i}\pi'(\boldsymbol{\theta}_i)|_{\eta_i=\hat{\eta}_i}\Big[\frac{1}{2}\ln\frac{|\sigma^2\mathbf{I}+\frac{\hat{\eta}_i}{1-\hat{\eta}_i}\sigma^2\mathbf{P_H}|}{|\sigma^2\mathbf{I}|}+\frac{1}{2}\mathrm{tr}\left(\sigma^2(\sigma^2\mathbf{I}+\frac{\hat{\eta}_i}{1-\hat{\eta}_i}\sigma^2\mathbf{P_H})^{-1}-\mathbf{I}\right)\\
&\quad+\frac{1}{2}(\mathbf{H}\,\boldsymbol{\theta})^T(\sigma^2\mathbf{I}+\frac{\hat{\eta}_i}{1-\hat{\eta}_i}\sigma^2\mathbf{P_H})^{-1}\mathbf{H}\,\boldsymbol{\theta}\Big]d\,\boldsymbol{\theta}_i\\
&= \frac{1}{2}\ln\frac{|\sigma^2\mathbf{I}+\frac{\hat{\eta}_i}{1-\hat{\eta}_i}\sigma^2\mathbf{P_H}|}{|\sigma^2\mathbf{I}|}+\frac{1}{2}\mathrm{tr}\left(\sigma^2(\sigma^2\mathbf{I}+\frac{\hat{\eta}_i}{1-\hat{\eta}_i}\sigma^2\mathbf{P_H})^{-1}-\mathbf{I}\right)\\
&\quad+\int_{\boldsymbol{\theta}_i}\left[\pi'(\boldsymbol{\theta}_i)|_{\eta_i=\hat{\eta}_i}\frac{1}{2}(\mathbf{H}\,\boldsymbol{\theta})^T(\sigma^2\mathbf{I}+\frac{\hat{\eta}_i}{1-\hat{\eta}_i}\sigma^2\mathbf{P_H})^{-1}\mathbf{H}\,\boldsymbol{\theta}\right]d\,\boldsymbol{\theta}_i\\
&= \frac{1}{2}\ln\frac{\left|\sigma^2\mathbf{I}+\frac{\hat{\eta}_i}{1-\hat{\eta}_i}\sigma^2\mathbf{P}_i\right|}{|\sigma^2\mathbf{I}|}\\
&= \frac{k_i}{2}\ln\left(\frac{1}{1-\hat{\eta}_i}\right) \qquad (5.17)\\
&= \frac{k_i}{2}\ln\left(\frac{\mathbf{x}^T\mathbf{P}_i\mathbf{x}}{k_i\sigma^2}\right) \qquad (5.18)\\
&= \frac{k_i}{2}\ln\frac{2l_{G_i}}{k_i} \qquad (5.19)
\end{aligned}
$$

This verifies that the term $\frac{k_i}{2}\ln\frac{2l_{G_i}}{k_i}$ of (5.6) is indeed an estimated mutual information term. As a measure of the statistical dependence of the parameter and observed data, the estimated MI is a reasonable measure of model complexity. First, the estimated MI can be viewed as averaged KLD distance between the $p_{\eta_i}(\mathbf{x}|\,\boldsymbol{\theta}_i)$ and $p_{\eta_i}(\mathbf{x})$, see (5.16), which assesses the "modeling potential" of the conditional distribution. Second, the estimated MI also measures the difference between the prior and posterior distributions of the unknown parameter and thus

relates to the "difficulty of estimation"[20]. From (5.17) we see that for linear model $\widehat{\mathrm{MI}}$ is monotonic with both the parameter dimension k_i and the embedding parameter $\hat{\eta}_i$. As $\hat{\eta}_i$ goes to zero, $\widehat{\mathrm{MI}} \to 0$. This is in agreement with the expectation from (5.3) in that when $\eta_i \to 0$, the Bayesian EEF $p(\mathbf{x}; \eta_i)$ reduces to the null model PDF $p_0(\mathbf{x})$. When $\hat{\eta}_i$ increases, the resulting Bayesian EEF $p(\mathbf{x}; \eta_i)$ moves closer towards $p_i(\mathbf{x})$ as shown in (5.3). The estimated MI simultaneously increases to reflect the increasing model complexity.

As shown, the Bayesian EEF penalty term takes into account three levels of model complexity, namely, parameter dimension, the prior of the unknown parameter $\pi_i'(\boldsymbol{\theta}_i)$ and the functional form on how the model is parameterized, the latter two of which contribute to the estimated MI. On the other hand, AIC only accounts for the dimension of unknown parameters k_i; BIC takes into consideration the parameter dimension k_i and the number of independently identical distributed (IID) data samples [2],[4] and [19].

5.3.3 An alternative interpretation of the estimated mutual information term

A closer look at the estimated mutual information term in (5.19) leads to an alternative intuition. Using the approximate relationship of $\widehat{\mathrm{SNR}}$ and l_{G_i} (5.15) in (5.19) we have

$$\begin{aligned} \widehat{\mathrm{MI}} &= \frac{k_i}{2} \ln \frac{2l_{G_i}}{k_i} \\ &= k_i \underbrace{\left[\frac{1}{2} \ln \left(1 + \frac{\widehat{\mathrm{SNR}}}{k_i/2} \right) \right]}_{\widehat{\mathrm{MI}} \text{ per dim}} \end{aligned}$$

The estimated mutual information term is the multiplicative result of parameter dimension k_i and the estimated MI per parameter dimension $\frac{1}{2}\ln\left(1+\frac{\hat{\text{SNR}}}{k_i/2}\right)$. As an example, for the normal linear model we have from (5.18) that $\widehat{\text{MI}} = \frac{k_i}{2}\ln\left(\frac{\mathbf{x}^T\mathbf{P}_i\mathbf{x}}{k_i\sigma^2}\right)$ and

$$\begin{aligned}
\mathbf{x}^T\mathbf{P}_i\mathbf{x} &= \mathbf{x}^T\mathbf{H}_i(\mathbf{H}_i^T\mathbf{H}_i)^{-1}\mathbf{H}_i^T\mathbf{x} \\
&= \mathbf{x}^T\mathbf{H}_i(\mathbf{H}_i^T\mathbf{H}_i)^{-\frac{1}{2}}(\mathbf{H}_i^T\mathbf{H}_i)^{-\frac{1}{2}}\mathbf{H}_i^T\mathbf{x} \\
&= ||\underbrace{(\mathbf{H}_i^T\mathbf{H}_i)^{-\frac{1}{2}}\mathbf{H}_i^T\mathbf{x}}_{\mathbf{y}}||^2 \\
&= ||(\mathbf{H}_i^T\mathbf{H}_i)^{-\frac{1}{2}}\mathbf{H}_i^T(\mathbf{H}_i\,\boldsymbol{\theta}_i+\mathbf{w})||^2 \\
&= ||\underbrace{(\mathbf{H}_i^T\mathbf{H}_i)^{-\frac{1}{2}}\mathbf{H}_i^T\mathbf{H}_i\,\boldsymbol{\theta}_i}_{\boldsymbol{\theta}_i'}+\underbrace{(\mathbf{H}_i^T\mathbf{H}_i)^{-\frac{1}{2}}\mathbf{H}_i^T\mathbf{w}}_{\mathbf{w}'}||^2
\end{aligned}$$

where we have denoted $\boldsymbol{\theta}_i' = (\mathbf{H}_i^T\mathbf{H}_i)^{-\frac{1}{2}}\mathbf{H}_i^T\mathbf{H}_i\,\boldsymbol{\theta}_i = (\mathbf{H}_i^T\mathbf{H}_i)^{\frac{1}{2}}\,\boldsymbol{\theta}_i$. It is of dimension $k_i\times 1$ and can be viewed as a signal coordinate vector. Also $\mathbf{w}' = (\mathbf{H}_i^T\mathbf{H}_i)^{-\frac{1}{2}}\mathbf{H}_i^T\mathbf{w}$ is of dimension $k_i\times 1$ and is a noise coordinate vector. Finally we denote $\mathbf{y} = \boldsymbol{\theta}_i' + \mathbf{w}'$, which is of dimension $k_i \times 1$.

With these notations, the estimated MI can be rewritten as

$$\widehat{\text{MI}} = \frac{k_i}{2}\ln\left(\frac{||\,\boldsymbol{\theta}_i'+\mathbf{w}'||^2}{k_i\sigma^2}\right) \tag{5.20}$$

$$= \frac{k_i}{2}\ln\left(\frac{\frac{1}{k_i}\sum_{j=1}^{k_i}(\theta_i'[j]+w'[j])^2}{\sigma^2}\right) \tag{5.21}$$

where $\theta_i'[j]$ and $w'[j]$ are the j^{th} elements of the vector $\boldsymbol{\theta}_i'$ and $\mathbf{w}'$ respectively.

Furthermore, we have the distributions of $\boldsymbol{\theta}_i'$ and $\mathbf{w}'$ based on the PDFs of $\boldsymbol{\theta}_i$ and $\mathbf{w}$, as

$$\boldsymbol{\theta}_i' \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{\boldsymbol{\theta}_i'})$$

with

$$\begin{aligned}\mathbf{C}_{\boldsymbol{\theta}'_i} &= (\mathbf{H}_i^T\mathbf{H}_i)^{\frac{1}{2}}\frac{\eta_i}{1-\eta_i}\sigma^2(\mathbf{H}_i^T\mathbf{H}_i)^{-1}(\mathbf{H}_i^T\mathbf{H}_i)^{\frac{1}{2}} \\ &= \underbrace{\frac{\eta_i}{1-\eta_i}\sigma^2}_{\sigma^2_{\theta_i}}\mathbf{I}_{k_i},\end{aligned}$$

where $\mathbf{I}_{k_i}$ denotes the identity matrix of dimension k_i and we have introduced $\sigma^2_{\theta_i} = \frac{\eta_i}{1-\eta_i}\sigma^2$ to simply the notation. This shows that by using the g-prior on $\boldsymbol{\theta}_i$, the coordinate vector $\boldsymbol{\theta}'_i$ has a scaled identity matrix as its covariance matrix; that is each element of the resulting vector $\boldsymbol{\theta}'_i$ is identically independently distributed (IID). The g-prior equalizes the distribution of each parameter of $\boldsymbol{\theta}_i$.

Similarly, we have the distribution of $\mathbf{w}'$ as

$$\mathbf{w}' \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{\mathbf{w}'})$$

with

$$\begin{aligned}\mathbf{C}_{\mathbf{w}'} &= (\mathbf{H}_i^T\mathbf{H}_i)^{-\frac{1}{2}}\mathbf{H}_i^T\sigma^2\mathbf{I}_N\mathbf{H}_i(\mathbf{H}_i^T\mathbf{H}_i)^{-\frac{1}{2}} \\ &= \sigma^2\mathbf{I}_{k_i}\end{aligned}$$

This shows that $\mathbf{w}'$ still has a zero mean normal distribution with a covariance matrix being $\sigma^2\mathbf{I}_{k_i}$. Then we have the PDF of $\mathbf{y} = \boldsymbol{\theta}'_i + \mathbf{w}'$, $p(\mathbf{y})$ as

$$\begin{aligned}p(\mathbf{y}) &= \mathcal{N}(\mathbf{0}, \mathbf{C}_{\boldsymbol{\theta}'_i} + \mathbf{C}_{\mathbf{w}'}) \\ &= \mathcal{N}\left(\mathbf{0}, (\sigma^2 + \sigma^2_{\theta_i})\mathbf{I}_{k_i}\right)\end{aligned}$$

In fact the term $\frac{1}{k_i}\sum_{j=1}^{k_i}(\theta'_i[j] + w'[j])^2$ in (5.21) is the estimate of $\sigma^2 + \sigma^2_{\theta_i}$ and the

hence (5.21) can be expressed alternatively as

$$\widehat{\mathrm{MI}} \;=\; k_i \underbrace{\frac{1}{2}\ln(\frac{\widehat{\sigma^2+\sigma^2_{\theta_i}}}{\sigma^2})}_{\widehat{\mathrm{MI}}\text{ per dim}}$$

The term $\widehat{\mathrm{MI}}$ per dim is the standard estimated mutual information for the case of Gaussian signal in additive Gaussian noise [22] for each signal component/parameter dimension. Since by employing the g-prior each element of the signal $\boldsymbol{\theta}'_i$ is IID, the total estimated MI is simply a multiplication of the $\widehat{\mathrm{MI}}$ per dim and the parameter dimension k_i. This provides another intuition on how the estimated MI depends on the parameter dimensions and the mechanism of the g-prior.

5.4 Bayesian EEF via Jeffreys' prior

Jeffreys' prior is another compelling non-informative prior [17] due to its property of invariance to reparameterization. In this section, we use the Jeffreys' prior in Bayesian EEF and derive the *asymptotic Bayesian EEF*. For each model $\mathcal{M}_i$ we assign a Jeffreys' prior $\pi_i(\boldsymbol{\theta}_i)$ to the unknown $\boldsymbol{\theta}_i$. The Jeffreys' prior PDF of $\boldsymbol{\theta}$ is proportional to the square root of the determinant of FIM of $\boldsymbol{\theta}_i$; that is, $\pi_i(\boldsymbol{\theta}_i) \propto \sqrt{|\mathbf{I}(\boldsymbol{\theta}_i)|}$. A motivation for the Jeffreys' prior is that Fisher information $\mathbf{I}(\boldsymbol{\theta}_i)$ is an indicator of the amount of information brought by the model/observations about unknown parameter $\boldsymbol{\theta}_i$. Favoring the values of $\boldsymbol{\theta}_i$ for which $\mathbf{I}(\boldsymbol{\theta}_i)$ is large, is equivalent to minimizing the influence of the prior [17]. By the Laplace approximation we have

$$p_i(\mathbf{x}|\,\boldsymbol{\theta}_i) \;\approx\; p_i(\mathbf{x}|\hat{\boldsymbol{\theta}}_i) e^{-\frac{1}{2}(\boldsymbol{\theta}_i-\hat{\boldsymbol{\theta}}_i)^T \mathbf{I}(\hat{\boldsymbol{\theta}}_i)(\boldsymbol{\theta}_i-\hat{\boldsymbol{\theta}}_i)}.$$

Moreover when assuming that $\pi_i(\boldsymbol{\theta}_i)$ is flat around $\hat{\boldsymbol{\theta}}_i$, which is valid for large data records, we have approximately

$$\begin{aligned} p_i(\mathbf{x}) &= \int_{\boldsymbol{\theta}_i} p_i(\mathbf{x}|\,\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)d\,\boldsymbol{\theta}_i \\ &\approx p_i(\mathbf{x}|\hat{\boldsymbol{\theta}}_i)\pi_i(\hat{\boldsymbol{\theta}}_i)\int e^{-\frac{1}{2}(\boldsymbol{\theta}_i-\hat{\boldsymbol{\theta}}_i)^T\mathbf{I}(\hat{\boldsymbol{\theta}}_i)(\boldsymbol{\theta}_i-\hat{\boldsymbol{\theta}}_i)}d\,\boldsymbol{\theta}_i \\ &= \frac{p_i(\mathbf{x}|\hat{\boldsymbol{\theta}}_i)\pi_i(\hat{\boldsymbol{\theta}}_i)}{(2\pi)^{-\frac{k_i}{2}}\sqrt{|\mathbf{I}(\hat{\boldsymbol{\theta}}_i)|}} \end{aligned}$$

Substituting this approximation into the EEF definition, we have

$$\begin{aligned} \ln\frac{p(\mathbf{x};\eta_i)}{p_0(\mathbf{x})} &= \eta_i \ln\frac{p_i(\mathbf{x})}{p_0(\mathbf{x})} - K_0(\eta_i) \\ &\approx \eta_i \ln\frac{\frac{p_i(\mathbf{x}|\hat{\boldsymbol{\theta}}_i)\pi_i(\hat{\boldsymbol{\theta}}_i)}{(2\pi)^{-\frac{k_i}{2}}\sqrt{|\mathbf{I}(\hat{\boldsymbol{\theta}}_i)|}}}{p_0(\mathbf{x})} \\ &\quad - \ln E_0 \exp\left(\eta_i \ln\frac{\frac{p_i(\mathbf{x}|\hat{\boldsymbol{\theta}}_i)\pi_i(\hat{\boldsymbol{\theta}}_i)}{(2\pi)^{-\frac{k_i}{2}}\sqrt{|\mathbf{I}(\hat{\boldsymbol{\theta}}_i)|}}}{p_0(\mathbf{x})}\right) \\ &= \eta_i \ln\frac{p_i(\mathbf{x}|\hat{\boldsymbol{\theta}}_i)}{p_0(\mathbf{x})} - \ln E_0 \exp\left(\eta_i \ln\frac{p_i(\mathbf{x}|\hat{\boldsymbol{\theta}}_i)}{p_0(\mathbf{x})}\right) \end{aligned}$$

Assigning $\boldsymbol{\theta}_i$ a Jeffreys' prior, the term $\frac{\pi_i(\hat{\boldsymbol{\theta}}_i)}{(2\pi)^{-\frac{k_i}{2}}\sqrt{|\mathbf{I}(\hat{\boldsymbol{\theta}}_i)|}}$ becomes a constant and thus the marginal PDF $p_i(\mathbf{x})$ becomes the multiplication of the maximized conditional PDF $p_i(\mathbf{x}|\hat{\boldsymbol{\theta}}_i)$ with the constant. From the derivation, it shows that by employing EEF mechanism, the resulting Bayesian model selection rule does not suffer from problems when $\int\sqrt{\mathbf{I}(\boldsymbol{\theta}_i)}d\,\boldsymbol{\theta}_i \to \infty$ as the FIM term is eliminated by the log-normalization term $K_0(\eta_i)$ using the Jeffreys' prior. This is one of many examples showing that the embedded family derives many of its useful properties from the use of the normalization term $K_0(\eta_i)$[8]. And it is this property that makes the *approximate Bayesian EEF* yield the same result as the frequentist EEF in [8].

For the normal linear model problem, the reduced Bayesian EEF, approximate Bayesian EEF methods and the reduced frequentist EEF all coincide with each other. This coincidence stems from the fact that the FIM for all $\boldsymbol{\theta}_i$ are the same under a certain model $\mathcal{M}_i$ in that $\mathbf{I}(\boldsymbol{\theta}_i) = \frac{\mathbf{H}_i^T \mathbf{H}_i}{\sigma^2}$. In this case the Jeffreys' prior, $\pi(\boldsymbol{\theta}_i) \propto \sqrt{|\mathbf{I}(\boldsymbol{\theta}_i)|}$, becomes an improper uniform distribution, $\pi(\boldsymbol{\theta}_i) = c > 0$, where c is a positive constant. This example also shows that Bayesian EEF can employ improper uniform prior without suffering from integration problems.

5.5 Conclusion

We have derived the Bayesian EEF, a new Bayesian model order selection rule, by using the EEF strategy in a Bayesian framework. The Bayesian EEF is shown to possess some desirable properties. To avoid introducing subjectivity in choosing parameter priors, the Bayesian EEF can utilize a vague proper prior as well as an improper non-informative prior, both of which are natural choices of non-informative priors but are usually forbidden by Bayesian model selection methods. It is also demonstrated that the EEF model order selection rule has a very intuitive penalty term as the sum of the parameter dimension and the estimated MI between parameter and received data. This interpretation not only helps in understanding the mechanisms at work in the EEF method but also provides new insights into the open question of designing an optimal penalty term for model selection. Some interesting interactions and coincidences between the EEF model order selection rules derived from Bayesian and frequentist viewpoints are also explained.

List of References

[1] C. Xu and S. Kay,"Source enumeration via the eef criterion,"*IEEE Signal Process. Lett.,* vol.15, pp.569–572,2008.

[2] H. Akaike,"A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol.19,pp.716–723, Dec.1974.

[3] J. Rissanen,"Modeling by shortest data description," *Automatica,* vol.14, no.5, pp.465–471,1978.

[4] G. Schwarz,"Estimating the dimension of a model," *Ann. Statist.,* vol.6, no.2, pp.461–464,1978.

[5] P. Stoica and Y. Selen, "Model-order selection: A review of information criterion rules," *IEEE Signal Process. Mag.,* vol.21, pp.36–47, Jul.2004.

[6] Z. Zhu, S. Kay, and R.S. Raghavan,"Information-theoretical optimal radar waveform design," *IEEE Signal Processing Letters*, vol. 24, no.3, pp.274-278, Mar. 2017.

[7] S. Kay, *Fundamentals of Statistical Signal Processing: Detection*, Prentice-Hall, Englewood Cliffs, NJ, 1998.

[8] S. Kay. "Exponentially embedded families-New approaches to model order estimation", *IEEE Trans. on Aerospace and Electronic Systems*, vol.41, no.1, pp.333–344, Jan. 2005.

[9] P. M. Djuric, "Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Processing,* vol.46, no.10, pp. 2726–2735, Oct. 1998.

[10] S. Amari and H. Nagaoka, "Methods of Information Geometry," New York: Oxford, 1993.

[11] J. Berger and L. Pericchi, "Objective Bayesian methods for model selection: Introduction and comparison," *in Model Selection, vol. 38 of IMS Lecture Notes–Monograph Series, (ed. P.Lahiri)*, pp.135-193, Institute of Mathematical Statistics, 2001.

[12] B. Tang, S. Kay, H. He and P.M. Baggenstoss, "EEF: exponentially embedded families with class-specific features for classification," *IEEE Signal Processing Letters,* vol. 23, no. 7, pp. 969–973, July 2016.

[13] B. Tang, H. He, Q. Ding, and S. Kay,"A parametric classification rule based on the exponentially embedded family," *IEEE Trans. Neural Netw. Learn. Syst.,* vol. 26, no. 2, pp. 367377, Feb. 2015.

[14] S. Kay and Q. Ding, "Exponentially embedded families for multimodal sensor processing," *in ICASSP,* 2010.

[15] D. J. Spiegelhalter, N. G. Best, B. P. Carlin and A. van der Linde, "Bayesian Measures of Model Complexity and Fit," *Journal of the Royal Statistical Society. Series B (Statistical Methodology),* Vol. 64, No. 4, pp. 583–639, 2002.

[16] Liang F., Paulo R., Molina G., Clyde M., Berger J. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103, 410.

[17] A. R. Syversveen, "Noninformative Bayesian priors. Interpretation and problems with construction and applications," *Preprint Statistics 3, Department of Mathematical Sciences, NTNU,Trondheim,* 1998.

[18] D. MacKay,"Bayesian interpolation," *Neural Computation*, vol. 4, no. 3, pp. 415–447, May. 1992.

[19] Myung, I. J., Balasubramanian, V. and Pitt, M. A.,"Counting probability distributions: Differential geometry and model selection," *Proceedings of the National Academy of Sciences USA,* vol. 97, pp.11170–11175, 2000.

[20] A. van der Linde, "A Bayesian view of model complexity," *Statistica Neerlandica* vol. 66, nr.3, pp.253–271, 2012.

[21] A. Zellner, "On assessing prior distributions and Bayesian regression analysis with g-prior distributions, *in Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti, eds. P. K,*1986.

[22] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

MANUSCRIPT 6

The Penalty Term of Exponentially Embedded Family is Estimated Mutual Information

by

Zhenghan Zhu and Steven Kay

published in

Proc. of 42^{nd} *IEEE International Conference on Acoustics, Speech and Signal Processing.* pp. 4149–4152, New Orleans, Mar. 2017.

Abstract

The penalty term plays an important role in model order selection. The Exponentially Embedded Families (EEF) has been proposed as an alternative approach for model order estimation. In this paper we show that part of the EEF penalty term is estimated mutual information (MI) between unknown parameters and received data. The finding is a result of an important relationship between Kullback-Leibler Divergence (KLD), signal-to-noise ratio (SNR) and MI in estimation/detection of random signals, which is also introduced.

6.1 Introduction

Model order selection is a fundamental problem in signal processing because observed data in practice usually is composed of an unknown number of signal components. For example, one may need to determine the number of sources in array signal processing [1]. Overestimating the order actually fits the noise in the data; underestimating the order, on the other hand, fails to describe the data precisely [1].

Model order selection problem, as a multiple hypotheses testing problem, lacks an optimal solution [11]. The traditional generalized likelihood ratio test (GLRT) tends to overestimate the order [7]. As a result, a typical model order selection algorithm introduces a penalty term to form a decision rule. Several popular algorithms are Akaike's information criterion (AIC) [2], the minimum description length (MDL) [3], Bayesian information criterion (BIC) [4] and maximum a posteriori (MAP) [11]. The reference [5] provides a review in this regard.

In addition to the aforementioned rules, EEF has been introduced in [8] as an alternative. It embeds two PDFs into a family of PDFs that are indexed by one or more parameters, and the new embedded family inherits many mathematical and optimality properties of the exponential family. It proves effective in model order selection and even superior under certain conditions. It has been shown to be consistent, i.e., as the data length $N \to \infty$, the probability of selecting the correct model goes to one [1]. The penalty term plays a central role in the EEF model order selection algorithm. In this paper we show that the EEF penalty term is actually the estimated mutual information between the unknown parameters and the received data. This hopefully can shed further light to understanding in choosing an optimal penalty term for model order selection. We limit the discussion in the context of linear normal model. A more general discussion will be our future work.

The paper is organized as follows. In Section 6.2 we introduce an useful relationship between KLD, SNR and MI, which holds in general in estimation/detection of random signals. In Section 6.3 a brief introduction is given to EEF. In Section 6.4 we discuss the EEF penalty term with an illustrative example. We then extend the discussion to the linear model in Section 6.5. Finally, some conclusions are drawn in Section 6.6.

6.2 An important relationship among KLD, SNR and MI

In signal processing, we often encounter problems of estimation/detection of random signals. Suppose we want to decide between the following hypotheses

$$\begin{aligned} \mathcal{H}_0 : \mathbf{x} &= \mathbf{w} \\ \mathcal{H}_1 : \mathbf{x} &= \mathbf{t} + \mathbf{w} \end{aligned}$$

where $\mathbf{w}$ is noise and $\mathbf{t}$ is a random signal. Denote $p_1(\mathbf{x})$ and $p_0(\mathbf{x})$ as the probability density function (PDF) of the received data $\mathbf{x}$ under $\mathcal{H}_1$ and $\mathcal{H}_0$ respectively, and $\pi(\mathbf{t})$ as the prior PDF of $\mathbf{t}$. An interesting and useful relationship is [6]

$$D(p_1(\mathbf{x})||p_0(\mathbf{x})) = E_{\mathbf{t}}[D(p_1(\mathbf{x}|\mathbf{t})||p_0(\mathbf{x}))] - I(\mathbf{x};\mathbf{t}), \tag{6.1}$$

where $D(p_1(\mathbf{x})||p_0(\mathbf{x}))$ is KLD, $E_{\mathbf{t}}(\cdot)$ denotes taking expectation according to $\mathbf{t}$, $p_1(\mathbf{x}|\mathbf{t})$ is the conditional PDF of $\mathbf{x}$ conditioned on $\mathbf{t}$ under $\mathcal{H}_1$ and $I(\mathbf{x};\mathbf{t})$ is the MI of $\mathbf{t}$ and $\mathbf{x}$ under $\mathcal{H}_1$. A related result has been used to compute MI in order to obtain the channel capacity per unit cost [10]. The derivation of (6.1) is straightforward

$$\begin{aligned} \ln \frac{p_1(\mathbf{x})}{p_0(\mathbf{x})} &= \ln \frac{p_1(\mathbf{x}|\mathbf{t})}{p_0(\mathbf{x})} - \ln \frac{p_1(\mathbf{x}|\mathbf{t})}{p_1(\mathbf{x})} \\ &= \ln \frac{p_1(\mathbf{x}|\mathbf{t})}{p_0(\mathbf{x})} - \ln \frac{p_1(\mathbf{x},\mathbf{t})}{p_1(\mathbf{x})\pi(\mathbf{t})} \end{aligned}$$

and taking the expected value with respect to $p_1(\mathbf{x},\mathbf{t})$ produces

$$E_{\mathbf{x},\mathbf{t}}\left[\ln \frac{p_1(\mathbf{x})}{p_0(\mathbf{x})}\right] = E_{\mathbf{t}} E_{\mathbf{x}|\mathbf{t}}\left[\ln \frac{p_1(\mathbf{x}|\mathbf{t})}{p_0(\mathbf{x})}\right] - E_{\mathbf{x},\mathbf{t}}\left[\ln \frac{p_1(\mathbf{x},\mathbf{t})}{p_1(\mathbf{x})\pi(\mathbf{t})}\right] \tag{6.2}$$

to yield (6.1). Also, $p_1(\mathbf{x})$ can be written as an averaged conditional PDF by averaging $p_1(\mathbf{x}|\mathbf{t})$ over $\mathbf{t}$, as

$$p_1(\mathbf{x}) = \int_{\mathbf{t}} p(\mathbf{x}, \mathbf{t}) d\mathbf{t} = \int_{\mathbf{t}} p_1(\mathbf{x}|\mathbf{t}) p(\mathbf{t}) d\mathbf{t} \tag{6.3}$$

Thus the term $D(p_1(\mathbf{x})||p_0(\mathbf{x}))$ is the KLD of the *averaged conditional* PDF $p_1(\mathbf{x})$ from the PDF $p_0(\mathbf{x})$.

Furthermore the MI $I(\mathbf{x};\mathbf{t})$ is also an averaged KLD obtained by averaging KLD of the conditional PDF $p_1(\mathbf{x}|\mathbf{t})$ from the unconditional PDF $p_1(\mathbf{x})$, $D(p_1(\mathbf{x}|\mathbf{t})||p_1(\mathbf{x}))$, over all possible signals $\mathbf{t}$

$$\begin{aligned} I(\mathbf{x};\mathbf{t}) &= \int_{\mathbf{t}} \int_{\mathbf{x}} p_1(\mathbf{x}, \mathbf{t}) \ln \frac{p_1(\mathbf{x}, \mathbf{t})}{p_1(\mathbf{x}) p(\mathbf{t})} d\mathbf{x} d\mathbf{t} \\ &= \int_{\mathbf{t}} \int_{\mathbf{x}} p(\mathbf{t}) p_1(\mathbf{x}|\mathbf{t}) \ln \frac{p_1(\mathbf{x}|\mathbf{t})}{p_1(\mathbf{x})} d\mathbf{x} d\mathbf{t} \\ &= \int_{\mathbf{t}} p(\mathbf{t}) D(p_1(\mathbf{x}|\mathbf{t})||p_1(\mathbf{x})) d\mathbf{t} \end{aligned} \tag{6.4}$$

Therefore, all three terms of the decomposition (6.1) can be interpreted respectively as a special distance measurement in the KLD sense. Alternatively, we can write the relationship as [6]

$$\underbrace{D(p_1(\mathbf{x})||p_0(\mathbf{x}))}_{\text{KLD}} = \underbrace{E_t[D(p_1(\mathbf{x}|\mathbf{t})||p_0(\mathbf{x}))]}_{\text{SNR}} - \underbrace{I(\mathbf{x};\mathbf{t})}_{\text{MI}}. \tag{6.5}$$

A simple example is next given to illustrate this important relationship. Assume $\mathbf{t}, \mathbf{w}$ are both independent $N \times 1$ random vectors and have distributions as $\mathbf{t} \sim N(\mathbf{0}, \sigma_t^2 \mathbf{I})$ and $\mathbf{w} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ respectively. Then we have

$$\begin{aligned} \mathbf{x} &\sim N\left(\mathbf{0}, \sigma^2 \mathbf{I}\right) \text{ under } \mathcal{H}_0 \\ \mathbf{x} &\sim N\left(\mathbf{0}, (\sigma^2 + \sigma_t^2) \mathbf{I}\right) \text{ under } \mathcal{H}_1 \end{aligned}$$

The KLD term is

$$\begin{aligned} D(p_1(\mathbf{x})||p_0(\mathbf{x})) &= \frac{1}{2}\ln\frac{|\sigma^2\mathbf{I}|}{|(\sigma^2+\sigma_t^2)\mathbf{I}|} \\ &\quad +\frac{1}{2}\mathrm{tr}\left[(\sigma^2+\sigma_t^2)\mathbf{I}(\sigma^2\mathbf{I})^{-1}-\mathbf{I}\right] \\ &= \frac{N}{2}\frac{\sigma_t^2}{\sigma^2}-\frac{N}{2}\ln\left(1+\frac{\sigma_t^2}{\sigma^2}\right). \end{aligned} \tag{6.6}$$

Next, for a given $\mathbf{t}$, the conditional PDF $p_1(\mathbf{x}|\mathbf{t})$ is a Gaussian distribution with mean $\mathbf{t}$ and variance $\sigma^2\mathbf{I}$, so

$$D(p_1(\mathbf{x}|\mathbf{t})||p_0(\mathbf{x})) = \frac{1}{2}\frac{\mathbf{t}^T\mathbf{t}}{\sigma^2}.$$

Thus, we have

$$\begin{aligned} E_{\mathbf{t}}[D(p_1(\mathbf{x}|\mathbf{t})||p_0(\mathbf{x}))] &= \int_{\mathbf{t}} p(\mathbf{t})\frac{1}{2}\frac{\mathbf{t}^T\mathbf{t}}{\sigma^2}d\mathbf{t} \\ &= \frac{N}{2}\frac{\sigma_t^2}{\sigma^2} \end{aligned}$$

which is indeed a measure of SNR. Lastly, it is easy to show that

$$I(\mathbf{x};\mathbf{t}) = \frac{N}{2}\ln\left(1+\frac{\sigma_t^2}{\sigma^2}\right).$$

Clearly, (6.5) applies to this simple example. This relationship (6.5) provides many insights into various problems. For instance, it suggests that MI measures the loss in detection performance between a matched filter, which is based on $\mathbf{t}$ known, and an estimator-correlator, which is based on an average $\mathbf{t}$ [6]. In this paper, however, we focus on using the relationship to justify the meaning of EEF penalty term. This hopefully will further the understanding of the problem of discrimination between normal linear models in [12].

6.3 Introduction of EEF

Assume that we have two distinct PDFs $p_1(\mathbf{x})$ and $p_0(\mathbf{x})$, and they model the data $\mathbf{x} = [x_0 \; x_1 \; \cdots \; x_{N-1}]^T$ under a general alternative model hypothesis $\mathcal{H}_1$ and a reference hypothesis $\mathcal{H}_0$. The EEF, denoted as $p(\mathbf{x}, \eta)$, is an exponential embedded PDF parameterized by an embedding parameter η, which takes on values $0 \leq \eta \leq 1$.

$$p(\mathbf{x}; \eta) = \frac{p_1^{\eta}(\mathbf{x}) p_0^{1-\eta}(\mathbf{x})}{\int p_1^{\eta}(\mathbf{x}) p_0^{1-\eta}(\mathbf{x}) d\mathbf{x}}. \tag{6.7}$$

Equivalently, the EEF is expressed as [8]

$$p(\mathbf{x}; \eta) = \exp\left[\eta T(\mathbf{x}) - K_0(\eta) + \ln p_0(\mathbf{x})\right]$$

where $T(\mathbf{x}) = \ln \frac{p_1(\mathbf{x})}{p_0(\mathbf{x})}$, $K_0(\eta) = \ln E_0(\exp(\eta T(\mathbf{x})))$, and $E_0(\cdot)$ denotes expectation under $\mathcal{H}_0$. If the PDF $p_1(\mathbf{x}) = p(\mathbf{x}; \boldsymbol{\theta})$ has unknown parameters $\boldsymbol{\theta}$, a $p \times 1$ vector and under $\mathcal{H}_0$, $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, then upon taking a reduced form and using an asymptotic approximation for the PDF, the EEF reduces to [8]

$$\text{EEF} = \max_{\eta} [\eta \ln \frac{1}{p_{T'}(T'(\mathbf{x}); \boldsymbol{\theta}_0)} - K_0(\eta)]$$

where $T'(\mathbf{x}) = \ln \frac{p(\mathbf{x}; \hat{\boldsymbol{\theta}})}{p(\mathbf{x}; \boldsymbol{\theta}_0)}$ and $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of the $\boldsymbol{\theta}$.

6.4 EEF penalty term-DC level in WGN

In this section we start the discussion of the penalty term of the EEF with a familiar example $\mathbf{x} = A\mathbf{1} + \mathbf{w}$, where A is assumed to be an unknown scalar, $\mathbf{w}$ is white Gaussian noise (WGN) with covariance $\sigma^2\mathbf{I}$, and $\mathbf{1} = [1 \; 1 \; \cdots \; 1]^T$ is a $N \times 1$ vector. The EEF, termed EEF_{d}, where the subscript "d" indicates that A

is assumed deterministic, is given in [8] as

$$\mathrm{EEF_d} = \max_{\eta} \left(\eta \frac{N\bar{x}^2}{2\sigma^2} + \frac{1}{2}\ln(1-\eta) \right),$$

where $\bar{x} = \frac{1}{N}\sum_{n=0}^{N-1} x_n$. With $\hat{\eta} = 1 - \frac{\sigma^2}{N\bar{x}^2}$ ($\hat{\eta} = 0$ if $N\bar{x}^2 < \sigma^2$), we have for $0 < \hat{\eta} < 1$

$$\mathrm{EEF_d} = \frac{1}{2}\left(\frac{N\bar{x}^2}{\sigma^2} - 1 \right) - \frac{1}{2}\ln\left(\frac{N\bar{x}^2}{\sigma^2} \right).$$

To verify the relationship between KLD, SNR and MI, we now assume the DC level A is a zero-mean Gaussian random variable with variance $k\frac{\sigma^2}{N}$ instead, and let $k \to \infty$. That is, we assign a vague proper prior to the unknown parameter in an attempt to assigning a non-informative prior. Then, we have

$$\begin{aligned} \mathcal{H}_0 : \mathbf{x} &\sim N(\mathbf{0}, \sigma^2\mathbf{I}) \\ \mathcal{H}_1 : \mathbf{x} &\sim N(\mathbf{0}, \sigma^2\mathbf{I} + k\frac{\sigma^2}{N}\mathbf{1}\mathbf{1}^T), \end{aligned}$$

and the resultant EEF PDF $p_\eta(\mathbf{x})$ can be shown to be

$$p_\eta(\mathbf{x}) = N(\mathbf{0}, \sigma^2\mathbf{I} + \frac{\eta}{1-\eta}\frac{\sigma^2}{N}\mathbf{1}\mathbf{1}^T). \tag{6.8}$$

Proof. $p_\eta(\mathbf{x})$ is an exponentially embedded PDF of two zero mean normal distributions PDFs with variance matrices being $\mathbf{C}_0 = \sigma^2\mathbf{I}$ and $\mathbf{C}_1 = \sigma^2\mathbf{I} + k\frac{\sigma^2}{N}\mathbf{1}\mathbf{1}^T$ respectively. According to (6.7), the resultant EEF $p_\eta(\mathbf{x})$ is also a zero mean

normal distribution with variance matrix $\mathbf{C}(\eta)$, depending on η, as [8]

$$\begin{aligned}
\mathbf{C}(\eta) &= \left(\eta\mathbf{C}_1^{-1} + (1-\eta)\mathbf{C}_0^{-1}\right)^{-1} \\
&= \left(\eta[\sigma^2\mathbf{I} + k\frac{\sigma^2}{N}\mathbf{1}\mathbf{1}^T]^{-1} + \frac{1-\eta}{\sigma^2}\mathbf{I}\right)^{-1} \\
&= \left(\frac{\eta}{\sigma^2}\left[\mathbf{I} - \frac{k}{k+1}\frac{1}{N}\mathbf{1}\mathbf{1}^T\right] + \frac{1-\eta}{\sigma^2}\mathbf{I}\right)^{-1} \\
&= \left(\frac{1}{\sigma^2}\left[\mathbf{I} - \frac{\eta k}{k+1}\frac{1}{N}\mathbf{1}\mathbf{1}^T\right]\right)^{-1} \\
&= \sigma^2\left(\mathbf{I} - \frac{-\frac{\eta k}{k+1}}{\frac{-\eta k}{k+1}+1}\frac{1}{N}\mathbf{1}\mathbf{1}^T\right) \\
&\overset{k\to\infty}{\to} \sigma^2\mathbf{I} + \frac{\eta}{1-\eta}\frac{\sigma^2}{N}\mathbf{1}\mathbf{1}^T
\end{aligned}$$

□

We denote $\mathbf{C}_\eta = \sigma^2\mathbf{I} + \frac{\eta}{1-\eta}\frac{\sigma^2}{N}\mathbf{1}\mathbf{1}^T$. Alternatively, we consider to assign a prior to A_η,

$$\pi(A_\eta) = N(0, \frac{\eta}{1-\eta}\frac{\sigma^2}{N})$$

for the following model

$$\mathbf{x}_\eta = A_\eta\mathbf{1} + \mathbf{w}$$

Then we have $p_\eta(\mathbf{x}) = p(\mathbf{x}_\eta)$; that is, the two are equivalent PDFs. This shows that EEF method can use vague proper prior and can find an equivalent PDF with a prior on unknown parameter relates to the embedding parameter η. This will be proved rigorously in an extended paper. On the other hand, it is generally a bad idea for many other Bayesian model selection methods to use vague proper prior[13].

Then, the EEF for this case, termed EEF_r, where the subscript "r" indicates that A is considered to be the outcome of a random variable, is KLD $D(p_{\hat{\eta}}(\mathbf{x})||p_0(\mathbf{x}))$. To compute it, we first should find the $\hat{\eta}$. It is also the value of η that maximizes the following likelihood ratio [8].

$$\begin{aligned} L_\eta(\mathbf{x}) &= 2\ln\frac{p_\eta(\mathbf{x})}{p_0(\mathbf{x})} \\ &= 2\ln\frac{\frac{1}{\sqrt{(2\pi)^N|\mathbf{C}_\eta|}}\exp(-\frac{1}{2}\mathbf{x}^T\mathbf{C}_\eta^{-1}\mathbf{x})}{\frac{1}{\sqrt{(2\pi)^N|\sigma^2\mathbf{I}|}}\exp(-\frac{1}{2}\mathbf{x}^T(\sigma^2\mathbf{I})^{-1}\mathbf{x})} \\ &= \mathbf{x}^T\left[(\sigma^2\mathbf{I})^{-1}-\mathbf{C}_\eta^{-1}\right]\mathbf{x}-\ln\frac{|\sigma^2\mathbf{I}+\frac{\eta}{1-\eta}\frac{\sigma^2}{N}\mathbf{1}\mathbf{1}^T|}{|\sigma^2\mathbf{I}|} \\ &= \frac{\eta}{N\sigma^2}\mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{x}-\ln\left|\mathbf{I}+\frac{\eta}{1-\eta}\frac{1}{N}\mathbf{1}\mathbf{1}^T\right| \\ &= \frac{\eta}{N\sigma^2}\mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{x}-\ln\left(1+\frac{\eta}{1-\eta}\right) \end{aligned} \tag{6.9}$$

Then the $\hat{\eta}$ is the value of η for which the derivative is equal to zero and hence, solves the equation

$$\frac{\partial L_\eta(\mathbf{x})}{\partial\eta} = \frac{1}{N\sigma^2}\mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{x}-\frac{1}{1-\eta} \tag{6.10}$$

Incorporating the definition of the embedding parameter $0\le\eta\le 1$, we have

$$\hat{\eta}=\begin{cases} 0 & \text{if } \mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{x}<N\sigma^2 \\ \frac{\mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{x}-N\sigma^2}{\mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{x}} & \text{otherwise} \end{cases}$$

When $\hat{\eta}=0$, the corresponding EEF_r penalty term is zero. We focus on the third case, $0<\hat{\eta}<1$, in the rest of the paper, which is of main interest. The resulting

EEF_r is

$$\begin{aligned}
\text{EEF}_r &= D(p_{\hat{\eta}}||p_0) \\
&= \frac{1}{2}\text{tr}\left[\frac{\hat{\eta}^2\sigma_A^2}{\sigma^2}\mathbf{1}\mathbf{1}^T\right] - \frac{1}{2}\ln\frac{|\hat{\eta}^2\sigma_A^2\mathbf{1}\mathbf{1}^T + \sigma^2\mathbf{I}|}{|\sigma^2\mathbf{I}|} \\
&= \frac{1}{2}\frac{\mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{x}}{N\sigma^2} - \frac{1}{2} - \frac{1}{2}\ln\left(\frac{\mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{x}}{N\sigma^2}\right) \\
&= \frac{1}{2}\left(\frac{N\bar{x}^2}{\sigma^2} - 1\right) - \frac{1}{2}\ln\left(\frac{N\bar{x}^2}{\sigma^2}\right). \qquad (6.11)
\end{aligned}$$

This shows that $\text{EEF}_\text{r} = \text{EEF}_\text{d}$, that is, the resulting EEFs for the two different problems of a deterministic A and a random A are the same. Note that when taking expectation according to $p_{\hat{\eta}}(\mathbf{x})$, $\hat{\eta}$ is considered as a constant parameter, not a function of $\mathbf{x}$.

It is easy to prove that the penalty term of EEF_r, $\frac{1}{2}\ln\left(\frac{\mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{x}}{N\sigma^2}\right)$ is indeed $I(\mathbf{x}_{\hat{\eta}}; A_{\hat{\eta}})$, the mutual information, since we have

$$\begin{aligned}
I(\mathbf{x}_{\hat{\eta}}; A_{\hat{\eta}}) &= E_{A_{\hat{\eta}}} D(p(\mathbf{x}_{\hat{\eta}}|A_{\hat{\eta}})||p(\mathbf{x}_{\hat{\eta}})) \\
&= \frac{1}{2}\ln\left(\frac{\mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{x}}{N\sigma^2}\right)
\end{aligned}$$

Strictly speaking, it is an estimated mutual information in that we only have the estimated PDF $p_{\hat{\eta}}(\mathbf{x})$, or equivalently $p_{\mathbf{x}_{\hat{\eta}}}$, instead of the true PDF.

This is a direct result of the equivalency of $p_\eta(\mathbf{x})$ and $p(\mathbf{x}_\eta)$ and the decomposition (6.5) when applied under the estimated PDF $p_{\hat{\eta}}(\mathbf{x})$ since the reduced EEF_r is an asymptotic KLD $D(p_{\hat{\eta}}||p_0)$. A modified version of decomposition (6.5) can be expressed as follows

$$\text{EEF}_r = \widehat{\text{SNR}} - \widehat{\text{MI}}$$

where $\widehat{\text{SNR}}, \widehat{\text{MI}}$ are the estimated SNR and estimated MI, respectively.

6.5 EEF penalty term of linear model

We now generalize the previous results to show that the EEF penalty term of model order selection for the linear model is the estimated MI. The linear model is an important one in practice and so a detailed analysis of this result is warranted. The linear model is $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$ where $\mathbf{H}$ is $N \times p$, $\boldsymbol{\theta}$ is a $p \times 1$ vector and $\mathbf{w} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Different models have different orders p and observation matrices $\mathbf{H}$. The model order selection problem is to decide the value of p to best model the data. It can be shown that assuming $\boldsymbol{\theta}$ is a deterministic unknown parameter yields the same EEF as assuming it is a random vector with a given prior PDF [8],[9]. We assume the latter by assigning to the unknown parameter $\boldsymbol{\theta}$ the prior PDF $N(\mathbf{0}, \frac{\xi^2}{p}\sigma^2(\mathbf{H}^T\mathbf{H})^{-1})$. When ξ^2 is assumed unknown, the EEF is proved to be equivalent to the model structure determination (MSD) [9]. If we reparameterize ξ^2 by letting

$$\frac{\xi^2}{p} = \frac{\eta}{1-\eta},$$

then a one-to-one transformation from ξ^2 to η $(0 < \eta < 1)$ is effected and finding $\hat{\eta}$ is equivalent to finding $\hat{\xi^2}$. With this setup, we have

$$\mathbf{x} \sim N(\mathbf{0}, \sigma^2\mathbf{I} + \frac{\xi^2}{p}\sigma^2\mathbf{P}_{\mathbf{H}}) \text{ under } p_\eta(\mathbf{x})$$

where $\mathbf{P}_{\mathbf{H}} = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$. It is shown in [9] that the EEF for model $\mathcal{M}_p$, i.e., with p unknown parameters, is

$$\text{EEF}_r(p) = \max_{\frac{\xi^2}{p}} \left[\frac{1}{2\sigma^2} \frac{\frac{\xi^2}{p}}{1+\frac{\xi^2}{p}} \mathbf{x}^T\mathbf{P}_{\mathbf{H}}\mathbf{x} - \frac{p}{2}\ln(1+\frac{\xi^2}{p}) \right].$$

The estimate $\frac{\hat{\xi^2}}{p}$, which maximizes $\mathrm{EEF}_r(p)$ is

$$\frac{\hat{\xi^2}}{p} = \frac{\mathbf{x}^T\mathbf{P_H}\mathbf{x}}{p\sigma^2} - 1$$

and hence, the maximized EEF is

$$\mathrm{EEF}_r(p) = \frac{1}{2}\left(\frac{\mathbf{x}^T\mathbf{P_H}\mathbf{x}}{\sigma^2} - p\right) - \frac{p}{2}\ln\frac{\frac{\mathbf{x}^T\mathbf{P_H}\mathbf{x}}{\sigma^2}}{p}. \tag{6.12}$$

Since $\frac{\mathbf{x}^T\mathbf{P_H}\mathbf{x}}{\sigma^2}$ obeys a χ^2_p distribution under the null hypothesis [7],[8], the term $\frac{1}{2}\left(\frac{\mathbf{x}^T\mathbf{P_H}\mathbf{x}}{\sigma^2} - p\right)$ subtracts out the mean p under $\mathcal{H}_0$, thereby producing $\widehat{\mathrm{SNR}}$. The term $\frac{p}{2}\ln\frac{\frac{\mathbf{x}^T\mathbf{P_H}\mathbf{x}}{\sigma^2}}{p}$ is the estimated $\widehat{\mathrm{MI}}$ as shown next. First, we have the following

$$\begin{aligned}
&D(p_\eta(\mathbf{x}|\,\boldsymbol{\theta})||p_\eta(\mathbf{x})) \\
&\qquad = \frac{1}{2}\ln\frac{|\sigma^2\mathbf{I} + \frac{\xi^2}{p}\sigma^2\mathbf{P_H}|}{|\sigma^2\mathbf{I}|} + \frac{1}{2}\mathrm{tr}\left(\sigma^2(\sigma^2\mathbf{I} + \frac{\xi^2}{p}\sigma^2\mathbf{P_H})^{-1} - \mathbf{I}\right) \\
&\qquad\quad + \frac{1}{2}(\mathbf{H}\,\boldsymbol{\theta})^T(\sigma^2\mathbf{I} + \frac{\xi^2}{p}\sigma^2\mathbf{P_H})^{-1}\mathbf{H}\,\boldsymbol{\theta} \\
&\qquad = \frac{1}{2}\ln\frac{|\sigma^2\mathbf{I} + \frac{\xi^2}{p}\sigma^2\mathbf{P_H}|}{|\sigma^2\mathbf{I}|} + \frac{1}{2}\mathrm{tr}\left(-\frac{\frac{\xi^2}{p}}{\frac{\xi^2}{p}+1}\mathbf{P_H}\right) \\
&\qquad\quad + \frac{1}{2\sigma^2}(\mathbf{H}\,\boldsymbol{\theta})^T(\mathbf{I} - \frac{\frac{\xi^2}{p}}{\frac{\xi^2}{p}+1}\mathbf{P_H})\mathbf{H}\,\boldsymbol{\theta}\,.
\end{aligned}$$

Then the computation of the estimated MI between $\mathbf{x}$ and $\boldsymbol{\theta}$ follows.

$$\begin{aligned}
I_{\hat{\eta}}(\mathbf{x};\boldsymbol{\theta}) &= E_{\boldsymbol{\theta}}D(p_{\hat{\eta}}(\mathbf{x}|\,\boldsymbol{\theta})||p_{\hat{\eta}}(\mathbf{x})) \\
&= \frac{1}{2}\ln\frac{|\sigma^2\mathbf{I} + \frac{\hat{\xi}^2}{p}\sigma^2\mathbf{P_H}|}{|\sigma^2\mathbf{I}|} + \frac{1}{2}\mathrm{tr}\left(-\frac{\frac{\hat{\xi}^2}{p}}{\frac{\hat{\xi}^2}{p}+1}\mathbf{P_H}\right) \\
&\qquad + E_{\boldsymbol{\theta}}\left[\frac{1}{2\sigma^2}\frac{1}{\frac{\hat{\xi}^2}{p}+1}(\mathbf{H}\,\boldsymbol{\theta})^T\mathbf{H}\,\boldsymbol{\theta}\right] \\
&= \frac{1}{2}\ln\frac{|\sigma^2\mathbf{I} + \frac{\hat{\xi}^2}{p}\sigma^2\mathbf{P_H}|}{|\sigma^2\mathbf{I}|} = \frac{p}{2}\ln\frac{\frac{\mathbf{x}^T\mathbf{P_H}\mathbf{x}}{\sigma^2}}{p}
\end{aligned} \tag{6.13}$$

where we have applied

$$E_{\boldsymbol{\theta}}\left[\frac{1}{2\sigma^2}\frac{1}{\frac{\hat{\xi}^2}{p}+1}(\mathbf{H}\,\boldsymbol{\theta})^T\mathbf{H}\,\boldsymbol{\theta}\right] \;=\; \frac{1}{2\sigma^2}\frac{1}{\frac{\hat{\xi}^2}{p}+1}\mathrm{tr}(\frac{\hat{\xi}^2}{p}\sigma^2\mathbf{P}_{\mathbf{H}}),$$

$(\mathbf{I}+\frac{\hat{\xi}^2}{p}\mathbf{P}_{\mathbf{H}})^{-1} = \mathbf{I}-\frac{\frac{\hat{\xi}^2}{p}}{\frac{\hat{\xi}^2}{p}+1}\mathbf{P}_{\mathbf{H}}$ and $\left|\mathbf{I}+\frac{\hat{\xi}^2}{p}\mathbf{P}_{\mathbf{H}}\right| = \left(1+\frac{\hat{\xi}^2}{p}\right)^p$. Thus, (6.13) proves that the EEF penalty term for the linear model in (6.12) is indeed the estimated MI. This is intuitively appealing in the sense that the model order selection rule should not take into account the information contributed by the distributional knowledge of the unknown parameters, which increases with its dimension [12]. As a special case, when $\mathbf{H}=\mathbf{1}$ and $\boldsymbol{\theta}=A$ then this example reduces to the DC level in WGN example for which $p=1$. Thus, the estimated MI term in (6.13) reduces to $\frac{1}{2}\ln\left(\frac{\mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{x}}{N\sigma^2}\right)$, which is the estimated MI in (6.11).

6.6 Conclusions

In this paper, we first introduce the important relationship KLD = SNR − MI. By employing this relationship, we have proved that the EEF penalty term in model order estimation is an estimated MI between the unknown parameters and the data. Intuitively, the estimated MI measures how much information of the data is contributed by the parameter $\boldsymbol{\theta}$. The EEF model order selection rule therefore subtracts it out so that the comparison among different models tends to be more fair.

List of References

[1] C. Xu and S. Kay,"Source enumeration via the eef criterion,"*IEEE Signal Process. Lett.,* vol.15, pp.569–572,2008.

[2] H. Akaike,"A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol.19,pp.716–723, Dec.1974.

[3] J. Rissanen,"Modeling by shortest data description," *Automatica,* vol.14, no.5, pp.465–471,1978.

[4] G. Schwarz,"Estimating the dimension of a model," *Ann. Statist.,* vol.6, no.2, pp.461–464,1978.

[5] P. Stoica and Y. Selen, "Model-order selection: A review of information criterion rules," *IEEE Signal Process. Mag.,* vol.21, pp.36–47, Jul.2004.

[6] Z. Zhu, S. Kay, and R.S. Raghavan,"Information-theoretical optimal radar waveform design," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp.274-278, Mar. 2017.

[7] S. Kay, *Fundamentals of Statistical Signal Processing: Detection*, Prentice-Hall, Englewood Cliffs, NJ, 1998.

[8] S. Kay. "Exponentially embedded families-New approaches to model order estimation", *IEEE Trans. on Aerospace and Electronic Systems*, vol.41, no.1, pp.333–344, Jan. 2005.

[9] S. Kay and Q. Ding, "Model estimation and classification via model structure determination", *IEEE Trans. on Signal Processing*,vol. 61, no.10, pp 2588-2597, 2013

[10] S. Verdu, "On channel capacity per unit cost," *IEEE Trans. Inform. Theory,* vol. 36, pp. 1019–1030, Sept. 1990.

[11] P. M. Djuric, "Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Processing,* vol.46, no.10, pp. 2726–2735, Oct. 1998.

[12] L.R. Pericchi, "An alternative to the standard Bayesian procedure for discrimination between normal linear models", *Biometrika*, vol.71, no.3, pp.575–586, Dec. 1984.

[13] J. Berger and L. Pericchi, "Objective Bayesian methods for model selection: Introduction and comparison," *in Model Selection, vol. 38 of IMS Lecture Notes–Monograph Series, (ed. P.Lahiri)*, pp.135-193, Institute of Mathematical Statistics, 2001.

MANUSCRIPT 7

Estimate the Degree of Noncircularity of Complex-valued Vectors via Exponentially Embedded Family

by

Zhenghan Zhu and Steven Kay

published in

the Proc. of IEEE Radar Conference 2017.

Abstract

Complex-valued signal processing is a fundamental task in many signal processing areas such as radar, sonar and communications. Modeling complex data as noncircular may provide better fitting of physical conditions. However, it requires more complicated signal processing algorithms and hence has more computations. Testing of noncircularity and estimating its degree are helpful in choosing a model. In this paper we focus on estimating the degree of noncircularity if the data is decided noncircular. It essentially a model order selection problem; therefore, we adopt the recently proposed exponentially embedded family (EEF) rule. Computer simulations are given to evaluate the EEF's performance and compare it with the minimum description length (MDL).

7.1 Introduction

In many areas, such as communication[1], radar[15] and sonar[16], a signal processing designer often deals with complex-valued data. One necessary consideration when designing algorithm is to model the complex data as noncircular or circular. A complex-valued random vector $\mathbf{x} \in \mathbb{C}^{N\times 1}$ is circular if its probability distribution is invariant to rotation in the complex plane, or equivalently, if its pseudocovariance matrix $\mathbf{P} = E(\mathbf{x}\mathbf{x}^T) = \mathbf{0}$, where T represents transpose. Conversely, it is noncircular if $\mathbf{P} \neq \mathbf{0}$ [1]. In many cases, circularity is traditionally assumed for simplification of computation, and this modeling is satisfyingly adequate[1]. On the other hand, there are cases where this simplified modeling fails and consequently produces very poor signal processing performance. Taking

the noncircularity of the data into account can achieve significant performance gains [1]. However, the noncircular modeling often requires more complicated signal processing algorithms and requires more computational resources. Therefore, an estimator of the degree of noncircularity is helpful in deciding which model to use.

A generalized likelihood ratio test (GLRT) has been proposed to test the noncircularity for Gaussian distributed data [2]. The probability distribution of the GLRT test statistic is shown to be a χ^2 distribution in [3]. It has then been extended to non-Gaussian cases [4]. In this paper we are concerned with the estimation of the degree of noncircularity, which is the number of nonzero noncircularity coefficients as defined in section 7.2. It is shown that this is essentially a model order selection problem. A GLRT-based sequential hypothesis test for estimating the degree of noncircularity is proposed in [5]. However, as viewed as a multiple hypotheses testing problem, the model order selection problem lacks an optimal solution [13]. Several important methods are Akaike's information criterion (AIC) [9], the minimum description length (MDL) [10], Bayesian information criterion (BIC) [11] and maximum a posteriori (MAP) [13]. A review in this regard can be found in [12].

A most recent alternative is the EEF model order selection method introduced in [6]. It embeds two PDFs into a family of PDFs that are indexed by one or more parameters, and the new embedded family inherits many mathematical and optimality properties of the exponential family. The rationale of the EEF for

model order estimator is: choosing the model order $\hat{k}$ that is associated with the maximum EEF makes the estimated PDF of the received data closest to its true PDF in Kullback-Leibler divergence (KLD) sense [6]. It proves effective in model order selection and superior under certain conditions such as in low signal-to-noise ratio cases. It has also been shown to be consistent, i.e., as the data length $N \to \infty$, the probability of selecting the correct model goes to one [14]. In this paper we employ the EEF rule to estimate the degree of noncircularity of the complex data. Computer simulations will be given to evaluate its performance and to compare it with the MDL since EEF rule has a similar computational load as MDL [14].

The paper is organized as follows. The problem under consideration is formulated in Section 7.2. Then, EEF rule for estimating the degree of noncircularity is derived in Section 7.3. Computer simulations and results are given to evaluate the performance of the proposed method 7.4. Finally, Section 7.5 draws some conclusions.

Notation: Throughout the paper, transpose of a vector/matrix will be denoted by T, H denotes conjugate transpose or hermitian, $*$ denotes conjugate, $E(\cdot)$ denotes expectation.

7.2 Problem Modeling

Assume we observe M independent identically distributed (IID) data vectors $\mathbf{x}_1,\ \mathbf{x}_2,\ \cdots,\ \mathbf{x}_M$ and each $\mathbf{x}_m$ for $m = 1, 2, \cdots, M$ is a $N \times 1$ complex-valued Gaussian random vector with its mean being zero, $\boldsymbol{\mu}(\mathbf{x}) = \mathbf{0}$. We will denote the received data as $\mathbf{X} = [\mathbf{x}_1 \mathbf{x}_2 \cdots \mathbf{x}_M]$. For a noncircular complex vector $\mathbf{x}$, the

conventional covariance matrix $\mathbf{C}$

$$\mathbf{C} = E(\mathbf{x}\mathbf{x}^H)$$

is not sufficient to fully describe its second-order properties. The pseudo-covariance matrix $\mathbf{P}$

$$\mathbf{P} = E(\mathbf{x}\mathbf{x}^T)$$

is required as complementary information. In fact, an argumented covariance matrix defined as the covariance matrix of the argumented random vector $\underline{\mathbf{x}} = [\mathbf{x}^T \mathbf{x}^H]^T$,

$$\underline{\mathbf{R}} = E(\underline{\mathbf{x}}\,\underline{\mathbf{x}}^H) = \begin{bmatrix} \mathbf{C} & \mathbf{P} \\ \mathbf{P}^* & \mathbf{C}^* \end{bmatrix}$$

is used to characterize the second-order properties of the noncircular complex-valued random variables [1]. As is shown, it is composed of $\mathbf{C}$ and $\mathbf{P}$. The probability density function (PDF) of the noncircular vector $\underline{\mathbf{x}}$ is then

$$p(\underline{\mathbf{x}} : \mathbf{C}, \mathbf{P}) = \frac{1}{\pi^N |\,\underline{\mathbf{R}}\,|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\,\underline{\mathbf{x}}^H\,\underline{\mathbf{R}}^{-1}\,\underline{\mathbf{x}}\right)$$

When the data is circular, then $\mathbf{P} = \mathbf{0}$, and the PDF reduces to a regular circularly complex Gaussian distribution.

Now, the circularity coefficients λ_k's for $k = 1, 2, \cdots, N$ are defined as singular values of the coherence matrix $\mathbf{C}^{-1/2}\mathbf{P}\mathbf{C}^{-\frac{T}{2}}$ [7]. Without loss of generality, let

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$$

Testing if the random vector is circular is equivalent to test if all of its circularity coefficients are equal to zero. On the other hand, if the data is noncircular, then

at least one of the circularity coefficients are nonzero. The test of noncircularity is in choosing between the following two hypotheses.

$$\mathcal{H}_0 : \qquad \lambda_1 = \lambda_2 = \cdots = \lambda_N = 0 \text{ (circular)}$$

$$\mathcal{H}_1 : \quad \text{at least one of the } \lambda_k \text{ is nonzero (noncircular)}$$

And the estimate of the degree of noncircularity is to estimate the number of nonzero λ's. In practice, the argumented covariance matrix of the Gaussian distribution from which the IID data is drawn is unknown and should be estimated. Thus maximum likelihood estimates (MLEs) $\hat{\mathbf{C}}$ and $\hat{\mathbf{P}}$ are used to replace the unknown $\mathbf{C}$ and $\mathbf{P}$ respectively,

$$\hat{\mathbf{C}} = \frac{1}{M}\sum_{k=1}^{M}(\mathbf{x}_k - \boldsymbol{\mu}_{\mathrm{x}})(\mathbf{x}_k - \boldsymbol{\mu}_{\mathrm{x}})^H$$

and

$$\hat{\mathbf{P}} = \frac{1}{M}\sum_{k=1}^{M}(\mathbf{x}_k - \boldsymbol{\mu}_{\mathrm{x}})(\mathbf{x}_k - \boldsymbol{\mu}_{\mathrm{x}})^T,$$

where $\boldsymbol{\mu}_{\mathrm{x}} = \frac{1}{M}\sum_{k=1}^{M}\mathbf{x}_k$ is the sample mean. Then the MLEs of the λ_k's, $\hat{\lambda}_k$'s are the ordered (from the largest to the smallest) singular values of the estimated coherence matrix

$$\hat{\mathbf{C}}^{-\frac{1}{2}}\hat{\mathbf{P}}\hat{\mathbf{C}}^{-\frac{T}{2}}$$

The goal is to estimate the number of nonzero circular coefficients, which is the degree of noncircularity, denoted as N_s. The problem is equivalent to choose

one of the following models.

$$\begin{aligned}
\mathcal{M}_1 : \lambda_1 &> \lambda_2 = \cdots = \lambda_N = 0 \\
&\vdots \\
\mathcal{M}_k : \lambda_1 &\geq \cdots \geq \lambda_k > \lambda_{k+1} = \cdots = \lambda_N = 0 \\
&\vdots \\
\mathcal{M}_N : \lambda_1 &\geq \lambda_2 \geq \cdots \geq \lambda_N > 0
\end{aligned} \tag{7.1}$$

Under the model $\mathcal{M}_k$, the degree of noncirculairty is k, and it has d_k unknown parameters, written as a vector $\boldsymbol{\theta}_k$. A good selection algorithm chooses the one is the true model or closest to the true model from all possible models $\mathcal{M}_1, \mathcal{M}_2, \cdots, \mathcal{M}_N$.

7.3 Estimate the degree of noncircularity via EEF

In this section, we derive the EEF rule for estimating the degree of noncircularity. The EEF rule chooses the k which maximizes the following [6]

$$\mathrm{EEF_k} = l_k(\mathbf{X}) - d_k \left[\ln \left(\frac{l_k(\mathbf{X})}{d_k} \right) + 1 \right] u\left(l_k(\mathbf{X}) - d_k \right),$$

where

$$l_k(\mathbf{X}) = 2 \ln \frac{p(\mathbf{X}; \hat{\boldsymbol{\theta}}_k)}{p(\mathbf{X}; \boldsymbol{\theta} = \mathbf{0})}$$

$\hat{\boldsymbol{\theta}}_k$ is the MLE of the unknown parameters and $u(\cdot)$ is the unit-step function. To evaluate the specific EEF rule for estimating the degree of noncircularity, we first find GLRT test statistic for each possible case corresponding to each degree of noncircularity as follows. We have

$$l_k(\mathbf{X}) = 2 \ln \frac{\max_{\mathbf{C}, \mathbf{P}_k} p(\mathbf{X}; \mathcal{H}_k)}{\max_{\mathbf{C}, \mathbf{0}} p(\mathbf{X}; \mathcal{H}_0)} = 2 \ln \frac{p(\mathbf{X}; \hat{\mathbf{C}}, \hat{\mathbf{P}}_k)}{p(\mathbf{X}; \hat{\mathbf{C}}, \mathbf{P} = \mathbf{0})},$$

where $\hat{\mathbf{P}}_k$ is the constrained MLE of the pseudo-covariance matrix $\mathbf{P}_k$, for which the degree of noncircularity is $N_s = k$, and $\hat{\mathbf{C}}$ is the unconstrained MLE of $\mathbf{C}$. First, let

$$p(\mathbf{X}; \mathcal{H}_N) = p(\mathbf{X}; \hat{\mathbf{C}}, \hat{\mathbf{P}})$$

It has been shown in [7] and [5] that

$$\max_{\mathbf{C},\mathbf{P}_k} p(\mathbf{X}; \mathbf{C}, \mathbf{P}_k) \quad = p(\mathbf{X}; \hat{\mathbf{C}}, \hat{\mathbf{P}}) \Pi_{i=k+1}^{N} (1 - \hat{\lambda}_i^2)^{\frac{M}{2}}$$

And specifically,

$$p_0(\mathbf{X}) = p(\mathbf{X}; \hat{\mathbf{C}}, \hat{\mathbf{P}}) \Pi_{i=1}^{N} (1 - \hat{\lambda}_i^2)^{\frac{M}{2}}$$

Then we have the log-likelihood ratio as

$$\begin{aligned} l_k(\mathbf{X}) &= 2 \ln \frac{\max_{\mathbf{C},\mathbf{P}_k} p(\mathbf{X}; \mathbf{C}, \mathbf{P}_k)}{p_0(\mathbf{X})} \\ &= -M \ln \left(\Pi_{i=1}^{k} (1 - \hat{\lambda}_i^2) \right) \end{aligned}$$

Note that under hypothesis $\mathcal{H}_k$ or equivalently the model $\mathcal{M}_k$, the number of unknown parameters is $d_k = k(2N - k + 1)$ more than that under $\mathcal{H}_0$ [8][5]. We now can form the EEF rule as follows

$$\begin{aligned} \mathrm{EEF_k} \quad = \quad & \Bigg\{ - M \ln \left(\Pi_{i=1}^{k} (1 - \hat{\lambda}_i^2) \right) \\ & - d_k \left[\ln \left(\frac{-M \ln \left(\Pi_{i=1}^{k} (1 - \hat{\lambda}_i^2) \right)}{d_k} \right) + 1 \right] \Bigg\} \\ & \cdot u \left(-M \ln \left(\Pi_{i=1}^{k} (1 - \hat{\lambda}_i^2) \right) - d_k \right), \end{aligned}$$

where $d_k = k(2N - k + 1)$.

The estimate of the degree of noncircularity, $\hat{k}$, is the k associated with EEF_k which is the maximum among all EEF's for $k = 1, \cdots, N$.

Note that the MDL rule for estimating the degree of noncircularity is [10][5]

$$\text{MDL}_\text{k} = -M \ln \left(\Pi_{i=1}^{k} (1 - \hat{\lambda}_i^2) \right) - d_k \ln M, \tag{7.2}$$

MDL rule chooses the k from $1, \cdots, N$ that maximizes (7.2).

7.4 Computer Simulations

In this section, a series of computer simulations with a similar setup as that in [5] are used to evaluate the performance of the EEF estimator for the degree of noncircularity.

For the first simulation, we generate $M = 100$ vectors for each trial which are drawn IID from a $N = 6$ variate CN distribution $\mathcal{CN}(\mathbf{0}, \mathbf{C}, \mathbf{P})$. Moreover, $\mathbf{C} = \mathbf{I}$ is an identity matrix and the pseudo-covariance matrix $\mathbf{P} = \mathbf{\Lambda}$ is a diagonal matrix with k nonzero diagonal elements, circularity coefficients, generated independently from the uniform distribution $U(0.05, 0.99)$ for each vector. In total, we run 1000 trials to calculate the correct estimate rate, i.e. the number of correct estimates when $\hat{k} =$ true k, over the number of trials. We repeat this for each $k = 1, \cdots, d$. In Figure 7.1, we list the probability of correct order of both EEF method and the MDL method for each true k. It is shown that the EEF generally has better performance over MDL. Note the GLRT-based sequential hypothesis testing method proposed in [5] requires setting a probability of false alarm (P_{FA}) which affects the

probability of correct order, thus we are not comparing with it directly.

For simulation 2, the number of vectors, M is increased from 100 to 500 without changing the other setup parameters. The probability of correct order corresponding to different true model orders of both EEF and MDL are displayed in Figure 7.2. It is no surprise that with longer observed data record, both methods achieve higher accuracy rate. However, EEF again outperforms the MDL in this case.

A third simulation investigates the estimation performances of EEF and MDL in a more difficult situation,i.e., the nonzero noncircularity coefficients on average are smaller, closer to zero, compared with those from the first two simulations. We keep $M = 500$ for this simulation but generate circularity coefficients by using a uniform distribution $U(0.05, 0.50)$ instead of the previous distribution $U(0.05, 0.99)$. The rest of setup remains unaltered. It is expected that both methods' performances will be degraded. The results, shown in Figure 7.3 agree with the theoretical expectation. It shows that the difference between the performances of the EEF and the MDL increases in this more "difficult" task.

To complete the performance evaluation, we modify the distribution from which the nonzero noncircularity coefficients are drawn from $U(0.05, 0.50)$ to $U(0.5, 0.99)$, and let the number of observation $M = 100$. The estimation performances of both methods are given in Figure 7.4, which should be compared with Figure 7.1. It is seen that both methods have very good performances, with the probabilities of correct order being close to one, since this is an easy scenario.

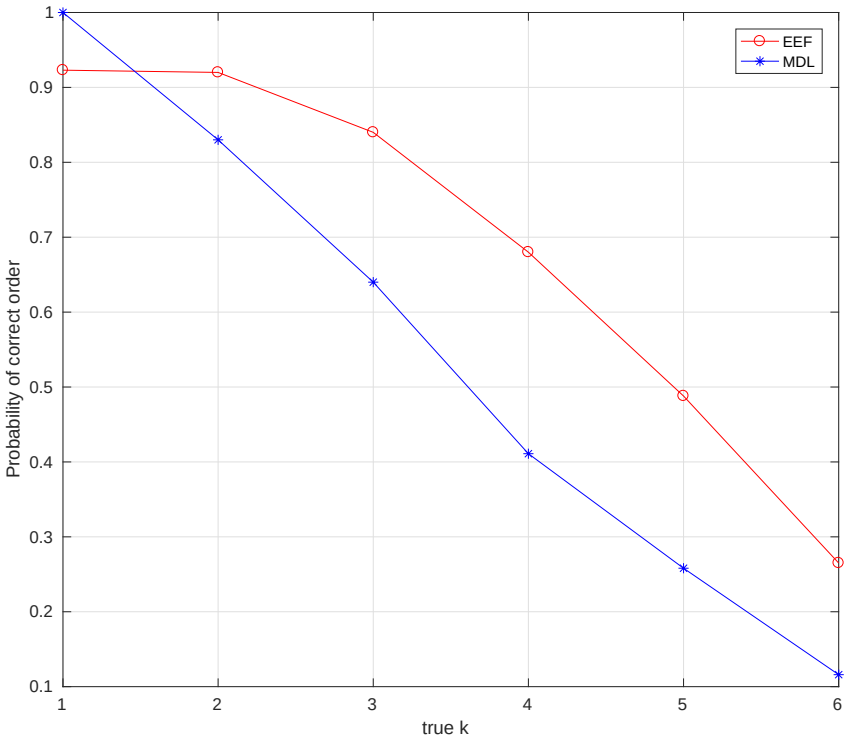

Figure 7.1. Performance Comparison between EEF and MDL of simulation 1 with 100 observations

In this case, the MDL method seems to have slightly better performance than the EEF.

7.5 Conclusions

In this paper, we have derived an EEF rule for estimating the degree of non-circularity of a complex-valued random vector. This estimator can be employed to decide whether to model complex random data as noncircular or circular with a trade-off between accuracy of modeling and algorithm complexity/computational cost when designing a signal processing algorithm. Computer simulations have shown that EEF method achieves good performance. It outperforms the MDL in

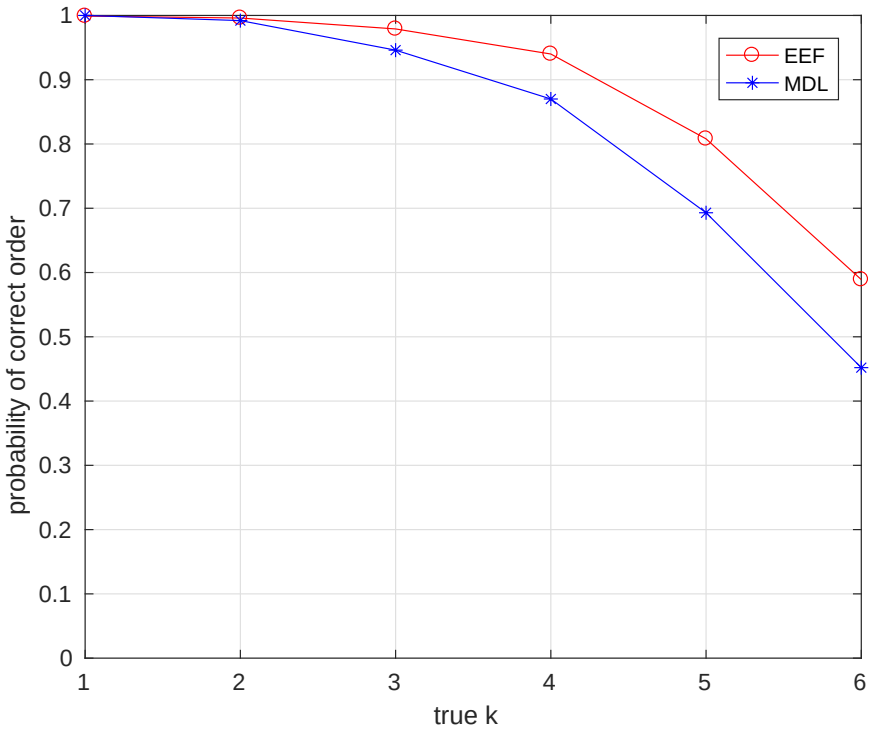

Figure 7.2. Performance Comparison between EEF and MDL of simulation 2 with 500 observations

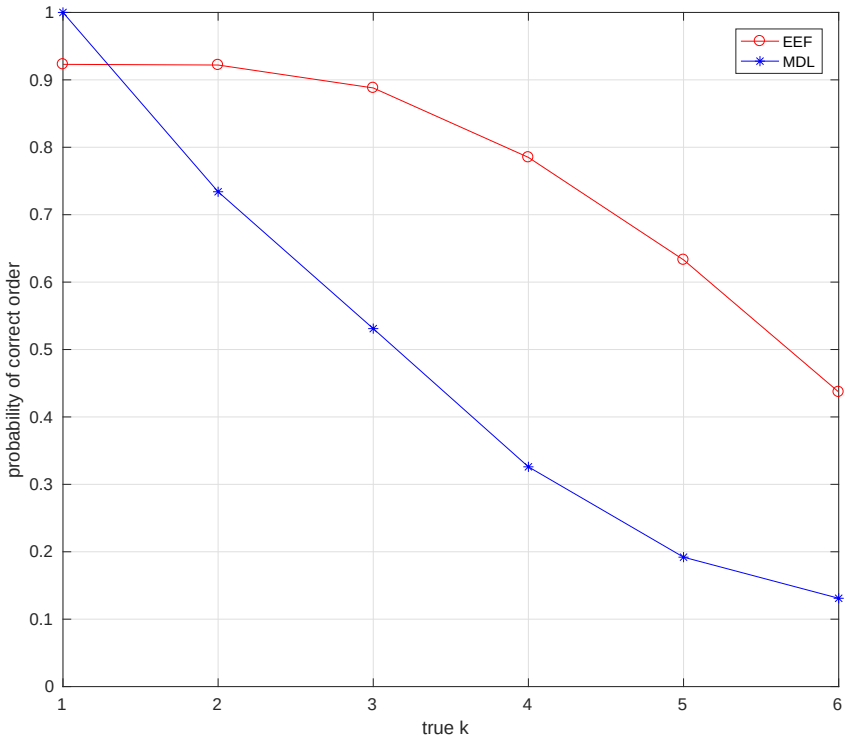

Figure 7.3. Performance Comparison between EEF and MDL for simulation 3 with smaller circularity coefficients and 500 observations

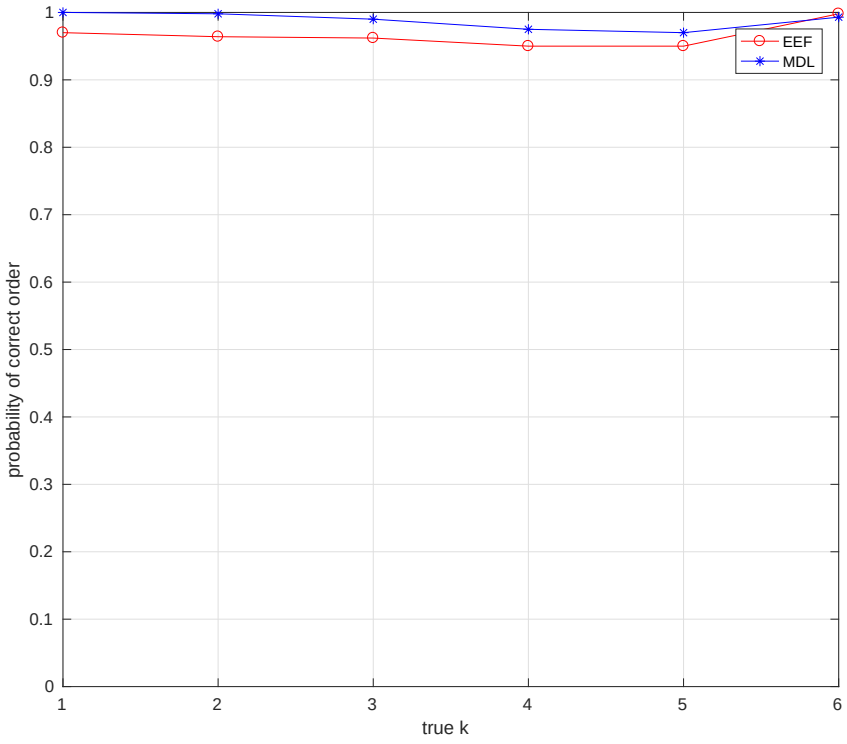

Figure 7.4. Performance Comparison between EEF and MDL for simulation 4 with larger circularity coefficients and 100 observations

general, especially in difficult situations.

List of References

[1] T. Adali, P. J. Schreier, and L. L. Scharf, "Complex-valued signal processing: The proper way to deal with impropriety," *IEEE Transactions on Signal Processing*, vol. 59, no. 11, pp. 5101–5125, 2011.

[2] P. Schreier, L. Scharf, and A. Hanssen, "A generalized likelihood ratio test for impropriety of complex signals," *IEEE Signal Processing Letters,* vol. 13, pp. 433–436, Jul. 2006.

[3] E. Ollila, "On the circularity of a complex random variable," *IEEE Signal Processing Letter,* vol. 15, pp. 841–844, 2008.

[4] E. Ollila and V. Koivunen, "Adjusting the generalized likelihood ratio test of circularity robust to non-normality," presented at the IEEE Int. Workshop Signal Process., Perugia, Italy, Jun. 2009.

[5] M. Novey, E. Ollila, and T. Adali, "On testing the extent of noncircularity," *IEEE Transactions on Signal Processing,* vol. 59, no. 11, pp. 5632–5637, 2011.

[6] S. Kay, "Exponentially embedded families-new approaches to model order estimation," *IEEE Transactions on Aerospace and Electronic Systems,* vol. 41, no. 1, pp. 333–345, 2005.

[7] J. Eriksson and V. Koivunen, "Complex random vectors and ICA models: Identifiability, uniqueness, and separability," *IEEE Transactions Information Theory,* vol. 52, no. 3, pp. 1017–1029, Mar. 2006.

[8] S. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[9] H. Akaike,"A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol.19,pp.716–723, Dec.1974.

[10] J. Rissanen,"Modeling by shortest data description," *Automatica,* vol.14,no.5, pp.465–471,1978.

[11] G. Schwarz,"Estimating the dimension of a model," *Ann. Statist.,* vol.6, no.2,pp.461–464,1978.

[12] P. Stoica and Y. Selen, "Model-order selection: A review of information criterion rules," *IEEE Signal Process. Mag.,* vol.21, pp.36–47, Jul.2004.

[13] P.M. Djuric,"Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Processing,* vol.46, no.10, pp. 2726–2735, Oct. 1998.

[14] C. Xu and S. Kay, "Source enumeration via the eef criterion," *IEEE Signal Process. Lett.,* vol.15, pp.569572,2008.

[15] N. Levanon, E. Mozeson, *Radar Signals*, J. Wiley, NY, 2004.

[16] W. Knight, R. Pridham, S. Kay, "Digital Signal Processing for Sonar", *IEEE Proceedings*, vol.69, no. 11, pp.1451-1506, Nov. 1981.

APPENDIX

Future Work

In this chapter, we discuss some future work that can be built upon the presenting work.

- The work in "Information-theoretic optimal radar waveform design" in Chapter 1 considers the optimal waveform design for detecting a stationary extended target. This research can be extended in many dimensions:
  - In practice, it is often of great interests to detect a moving target. Hence a future work will be devoted to investigate the problem of designing optimal transmit signals for detecting an extended target that is moving and fluctuating. This will require modeling the target impulse response as a two-dimensional function of both time and range. Mathematically, the response will be modeled as a linear time varying (LTV) random process. In frequency domain, the movement and fluctuation of the target lead to Doppler shift and Doppler spread of the target power spectral density (PSD). The target PSD is a 2-D function of both transmit signal frequency and Doppler frequency. Designing an optimal signal for detection of moving extended targets given the knowledge of the 2-D PSD is the first step of future research. Furthermore, in practice some parameters of the target PSD may be unknown, e.g., unknown Doppler shift. The lack of these information will require designing estimation

algorithms and the transmit signal at the same time. This will be the second research step.

  - It is of practical interests to pose some constraints on the waveform such as constant modulus and similarity constraint and the Peak-to-average ratio (PAR) constraint. The constant modulus and/or PAR constraints stem from the physical limitations of the radar amplifier and system. A similarity constraint is applicable when the signal waveform designer wants to use an existing waveform as a benchmark. The existing waveform is often known to have good properties, e.g. ambiguity. Hence designing optimal waveform with such constraints will make the solution easier to be implemented.

  - Design optimal signal waveform for multiple input multiple output (MIMO) radar. MIMO system transmits multiple probing signals and uses multiple receivers, hence provides extra degrees of freedom and has better detection and estimation performance. It has become a leading radar technique. It is thus useful to extend the waveform design for MIMO system.

- In the work of "On detection of nonstationarity of the Covariance Matrix in radar signal processing" in Chapter 2, we have assumed the complete knowledge of the normalized clutter covariance matrix $\mathbf{R}$. As a future work, we may extend the GLRT and Rao detectors of nonstationarity of the covariance matrix to the case where $\mathbf{R}$ is not completely known.

- In the work of " the complex parameter Rao test" in Chapter 3, it is assumed that no nuisance parameter is present in the hypothesis testing problem. In many practical problems, some nuisance parameters may appear in the hypothesis testing, it is hence useful to extend the complex parameter Rao test for the cases of nuisance parameters.

- In the work of "On the Bayesian exponentially embedded family for model order selection" in Chapter 5 we have focused on the linear model. Future work will consider model order selection for nonlinear models in which data depends on unknown parameter through nonlinear functions. As shown the penalty term of the EEF model order selection is the sum of the unknown parameter dimension and the estimated mutual information between the unknown parameter and received data. With these new insights, future work will also be devoted to the open question on how to design an optimal penalty term for the model order selection algorithm.