Optimally Distinguishable Distributions: a New Approach to Composite Hypothesis Testing With Applications to the Classical Linear Model

Seyed Alireza Razavi, *Student Member, IEEE*, and Ciprian Doru Giurcăneanu, *Member, IEEE*

***Abstract*—The newest approach to composite hypothesis testing proposed by Rissanen relies on the concept of optimally distinguishable distributions (ODD). The method is promising, but so far it has only been applied to a few simple examples. We derive the ODD detector for the classical linear model. In this framework, we provide answers to the following problems that have not been previously investigated in the literature: i) the relationship between ODD and the widely used Generalized Likelihood Ratio Test (GLRT); ii) the connection between ODD and the information theoretic criteria applied in model selection. We point out the strengths and the weaknesses of the ODD method in detecting subspace signals in broadband noise. Effects of the subspace interference are also evaluated.**

***Index Terms*—Generalized likelihood ratio test, information theoretic criteria, linear model, minimum description length, optimally distinguishable distributions.**

I. Introduction and Preliminaries

THE most recent developments in methods of inference based on the minimum description length (MDL) principle [1], [2] emerge from a happy union between algorithmic complexity theory (ACT) [3] and coding theory. Because the central notions from ACT, namely *Kolmogorov complexity* (KC), *universal distribution* and the *Kolmogorov structure function* (KSF) are noncomputable, their use in practical applications poses troubles. To circumvent such difficulties, Rissanen extends all the notions from ACT to statistical models by replacing the *set of programs* with classes of parametric models $\{f(\mathbf{x};\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, where $\mathbf{x} = [x_0, \ldots, x_{N-1}]^\top$ is the vector of observations and Θ is a bounded closed subset of $\Re^k$ [1]. The symbol $^\top$ is used for transposition. With the understanding that each model class is a likelihood function, the role of the *universal model* is played by the normalized maximum likelihood (NML) density function [4]

$$\tilde{f}(\mathbf{x}) = \frac{f(\mathbf{x};\hat{\boldsymbol{\theta}}(\mathbf{x}))}{\int\limits_{\mathbf{y}:\hat{\boldsymbol{\theta}}(\mathbf{y})\in\Theta} f(\mathbf{y};\hat{\boldsymbol{\theta}}(\mathbf{y}))\mathrm{d}\mathbf{y}} \tag{1}$$

Manuscript received June 27, 2008; accepted February 02, 2009. First published March 16, 2009; current version published June 17, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Pramod K. Varshney. This work was supported by the Academy of Finland, Project No. 113572, 118355, and 213462. The article extends the results of the paper "Composite Hypothesis Testing by Optimally Distinguishable Distributions," authored by S. A. Razavi and C. D. Giurcăneanu, which was presented at International Conference on Acoustics, Speech and Signal Processing (ICASSP), Las Vegas, NV, 2008

The authors are with the Department of Signal Processing, Tampere University of Technology, FIN-33101, Tampere, Finland (e-mail: alireza.razavi@tut.fi; ciprian.giurcaneanu@tut.fi).

Digital Object Identifier 10.1109/TSP.2009.2017568

where $\hat{\boldsymbol{\theta}}(\mathbf{x})$ denotes the maximum likelihood (ML) estimate. Whenever it is clear from the context which measurements are used for estimation, the simpler notation $\hat{\boldsymbol{\theta}}$ will be preferred to $\hat{\boldsymbol{\theta}}(\mathbf{x})$. Our interest is confined to models for which $f(\mathbf{x};\boldsymbol{\theta})$ can be factored as [4]

$$f(\mathbf{x};\boldsymbol{\theta}) = f(\mathbf{x} \mid \hat{\boldsymbol{\theta}})g(\hat{\boldsymbol{\theta}};\boldsymbol{\theta}) \tag{2}$$

where $g(\hat{\boldsymbol{\theta}};\boldsymbol{\theta})$ is the marginal density of $\hat{\boldsymbol{\theta}}$. The conditional density $f(\mathbf{x} \mid \hat{\boldsymbol{\theta}})$ does not depend on the unknown parameter vector $\boldsymbol{\theta}$. Furthermore, KC is replaced by stochastic complexity (SC), whose expression is given by $\ln(1/\tilde{f}(\mathbf{x}))$.

The definition of the KSF involves a partition of the parameter space into rectangles such that the Kullback–Leibler (KL) divergence between any two adjacent models is constant [1]. For the detection problems discussed in this study, the partition of Θ associated with the KSF is significantly more important than the expression of the KSF itself. This motivates us to emphasize the main steps of the construction as they are outlined in [1]. Let $\mathbf{J}_N(\boldsymbol{\theta}) = -\frac{1}{N}E\left[\frac{\partial^2 \ln f(\mathbf{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right]$ be the Fisher information matrix (FIM), and $\mathbf{J}_\infty(\boldsymbol{\theta}) = \lim_{N\to\infty}\mathbf{J}_N(\boldsymbol{\theta})$. The limit is finite for most of the models in signal processing, but not for all of them; for example, the limit is not finite in the case of a sinusoidal regression model with unknown frequency [5]. In the following derivations, we prefer to use $\mathbf{J}_N(\boldsymbol{\theta})$, with the supplementary assumption that none of its singular points are included in Θ. For an arbitrary $\overline{\boldsymbol{\theta}} \in \Theta$, consider the hyper-ellipsoid

$$(\boldsymbol{\theta}-\overline{\boldsymbol{\theta}})^\top\mathbf{J}_N(\overline{\boldsymbol{\theta}})(\boldsymbol{\theta}-\overline{\boldsymbol{\theta}}) = \frac{d}{N} \tag{3}$$

where d is a parameter whose optimal value we will find next. We take the largest rectangle within this hyper-ellipsoid, and then we continue defining a complete set of $\mathsf{N}_{d/N}$ disjoint rectangles whose union is the entire parameter space Θ. The procedure is complicated if the entries of $\mathbf{J}_N(\overline{\boldsymbol{\theta}})$ depend on $\overline{\boldsymbol{\theta}}$ [1]. In [6], it is described how the partition can be obtained in the general case (see also [2, Ch. 10]). For the problem addressed in this study, $\mathbf{J}_N(\overline{\boldsymbol{\theta}})$ is the same for all $\overline{\boldsymbol{\theta}}$, which simplifies significantly the construction of the partition, as we will see in the following sections. Remark that $\mathsf{N}_{d/N}$ decreases when d grows from zero to a value where a single rectangle covers the entire parameter space Θ [7]. With the conventions from [1], we let $B_{d/N}(j)$ denote the jth rectangle within this set, and we denote its center as $\boldsymbol{\theta}^j$. For all $j \in \{0, \ldots, \mathsf{N}_{d/N}-1\}$, the probability density $\hat{f}(\mathbf{x} \mid \boldsymbol{\theta}^j)$ is defined by

$$\hat{f}(\mathbf{x} \mid \boldsymbol{\theta}^j) = \begin{cases} \frac{f(\mathbf{x};\hat{\boldsymbol{\theta}}(\mathbf{x}))}{Q_{d/N}(j)}, & \hat{\boldsymbol{\theta}}(\mathbf{x}) \in B_{d/N}(j) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

1053-587X/$25.00 © 2009 IEEE

Authorized licensed use limited to: Tampereen Teknillinen Korkeakoulu. Downloaded on June 17, 2009 at 02:17 from IEEE Xplore. Restrictions apply.

with

$$Q_{d/N}(j) = \int_{\mathbf{x}:\hat{\boldsymbol{\theta}}(\mathbf{x})\in B_{d/N}(j)} f(\mathbf{x};\hat{\boldsymbol{\theta}}(\mathbf{x}))\mathrm{d}\mathbf{x}$$

$$= \int_{\hat{\boldsymbol{\theta}}\in B_{d/N}(j)} \left[\int_{\mathbf{x}:\hat{\boldsymbol{\theta}}(\mathbf{x})=\hat{\boldsymbol{\theta}}} f(\mathbf{x}\mid\hat{\boldsymbol{\theta}})\mathrm{d}\mathbf{x}\right] g(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\theta}})\mathrm{d}\hat{\boldsymbol{\theta}} \quad (5)$$

$$= \int_{\hat{\boldsymbol{\theta}}\in B_{d/N}(j)} g(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\theta}})\mathrm{d}\hat{\boldsymbol{\theta}}. \quad (6)$$

The calculations above used (2) together with the fact that the inner integral in (5) gives unity [1].

The key point is that the distributions in (4) are perfectly distinguishable. The idea of distinguishability is borrowed from [8]: whenever $\boldsymbol{\theta}'$ is located close to a point $\boldsymbol{\theta}$ in the parameter space, it is difficult to decide if the measurements $\mathbf{x}$ are outcomes of the model $f(\mathbf{x};\boldsymbol{\theta}')$ or $f(\mathbf{x};\boldsymbol{\theta})$. By contrast, when the distance between $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}$ is large, it is easy to make a decision based on the sample $\mathbf{x}$, and consequently $f(\mathbf{x};\boldsymbol{\theta}')$ and $f(\mathbf{x};\boldsymbol{\theta})$ are deemed to be distinguishable. Relying on this property, Balasubramanian [8] collapses all the models whose parameters are within the hyper-ellipsoid (3) to a single probability distribution that it is conventionally assigned to $\bar{\boldsymbol{\theta}}$. Note that the hyper-ellipsoid in (3) shrinks when the sample size N increases. Because it is not possible to construct a partition of Θ with hyper-ellipsoids, Rissanen uses the largest rectangle within the hyper-ellipsoid (3) instead, as already explained above. Then, the probability distribution $\hat{f}(\mathbf{x}\mid\boldsymbol{\theta}^j)$ is assigned to the center of the jth rectangle, or equivalently, to the jth equivalence class.

Note that $\hat{f}(\mathbf{x}\mid\boldsymbol{\theta}^i)$ and $\hat{f}(\mathbf{x}\mid\boldsymbol{\theta}^j)$ are distinguishable for $i\neq j$ because their supports are disjoint. In [1], [9], it is shown that $\hat{f}(\mathbf{x}\mid\boldsymbol{\theta}^j)$, $j\in\{0,\ldots,\mathsf{N}_{d/N}-1\}$, are optimally distinguishable distributions (ODD). The proof is technical and involves a carefully defined index of separation. Additionally, $\hat{f}(\mathbf{x}\mid\boldsymbol{\theta}^j)$ is almost constant for all the estimates $\hat{\boldsymbol{\theta}}(\mathbf{x})$ within $B_{d/N}(j)$. Since the probability distributions in (4) have desirable properties, we want to minimize the KL divergence $D(\hat{f}(\mathbf{x}\mid\boldsymbol{\theta}^j)\parallel f(\mathbf{x};\boldsymbol{\theta}^j))$ between the "artificial" model $\hat{f}(\mathbf{x}\mid\boldsymbol{\theta}^j)$ and the "natural" model $f(\mathbf{x};\boldsymbol{\theta}^j)$ for all $j\in\{0,\ldots,\mathsf{N}_{d/N}-1\}$. If the Central Limit Theorem holds, then there exists a unique $\hat{d}$ that minimizes the KL divergence, and asymptotically, $\hat{d}=3k$ [1]. Moreover, Rissanen shows that the number of distinguishable distributions obtained when $\hat{d}=3k$ agrees with the number of distinguishable distributions given in [8].

These findings can be applied almost straightforwardly to composite hypothesis testing, defining a totally new framework for this problem. We briefly explain the ODD test between the hypotheses specified by

$$\mathcal{M}_0=\{f(\mathbf{x};\boldsymbol{\theta}):\boldsymbol{\theta}=\boldsymbol{\theta}^0\}$$
$$\mathcal{M}_1=\{f(\mathbf{x};\boldsymbol{\theta}):\boldsymbol{\theta}\neq\boldsymbol{\theta}^0\}.$$

For partitioning the parameter space in this case, we first demarcate the rectangle centered at the point $\boldsymbol{\theta}^0$ denoted $B_{\hat{d}/N}(0)$, then fix the centers of its neighbors, and finally continue the construction until the complete set of rectangles is settled. The ODD criterion selects the model class $\mathcal{M}_0$ whenever $\hat{\boldsymbol{\theta}}(\mathbf{x})\in B_{\hat{d}/N}(0)$, where $\hat{\boldsymbol{\theta}}(\mathbf{x})$ is the ML estimate for the model class $\mathcal{M}_1$ [1].

Remark that it is not necessary to resort to the maximization of the probability of detection (P_D) for a given probability of false alarm (P_{FA}), as in traditional Neyman-Pearson (NP) methodology [10]. However, the performance of the ODD procedure can be assessed by calculating indexes $E1=1-P_{0|0}$ and $E2=P_{0|j}$ for $j\neq 0$ [1], [9]. For an arbitrary pair (i,j), $P_{i|j}$ denotes the probability mass of $B_{\hat{d}/N}(i)$ induced by the model $f(\mathbf{x};\boldsymbol{\theta}^j)$. $E1$ is intended as a confidence measure for being wrong in accepting the null hypothesis. Similarly, $E2$ is a confidence measure for being wrong in rejecting the null hypothesis. $E1$ is interpreted by Rissanen as something very different from P_{FA}, even if $E1$ is equal to P_{FA} by definition. The probabilistic interpretation of $E1$ and $E2$ is difficult, and the interested reader can find more details in [1], [9]. Here, we take $E1$ and $E2$ to be confidence measures.

ODD testing is promising, but so far it has only been applied in the following examples [1], [7], [9]: (i) for the model class $\mathcal{M}_0$, the observed random variable X is Gaussian with mean 0 and variance 1, whereas for $\mathcal{M}_1$, X is Gaussian with nonzero mean and nonunitary variance; (ii) $X\in\{0,1\}$ is Bernoulli distributed for both $\mathcal{M}_0$ and $\mathcal{M}_1$, and under the null hypothesis, $X=0$ with probability $1/3$. The relation between P_D and P_{FA} when the parameter space partition is constructed with $\hat{d}$ has not been previously investigated. In this study, we provide answers to unsolved problems connected with ODD testing by considering the linear model (LM), which has many applications in signal processing [10].

The rest of this paper is focused on the detection of a deterministic signal with unknown linear parameters in zero-mean Gaussian noise. More precisely, the signal obeys the linear subspace model $\mathbf{H}\boldsymbol{\theta}$, where $\mathbf{H}\in\Re^{N\times k}$ $(N>k)$ is a full-rank matrix and $\boldsymbol{\theta}\in\Re^{k\times 1}$ is the vector of the unknown parameters. The detection problem reduces to deciding if the measurements $\mathbf{x}=[x_0,\ldots,x_{N-1}]^\top$ are outcomes from $\mathcal{N}_N(\mathbf{0},\tau\mathbf{I})$ or from $\mathcal{N}_N(\mathbf{H}\boldsymbol{\theta},\tau\mathbf{I})$ [10], where $\tau>0$ and $\mathcal{N}_N(\boldsymbol{\varrho},\mathbf{R})$ denotes the multivariate Gaussian distribution with mean $\boldsymbol{\varrho}\in\Re^{N\times 1}$ and covariance matrix $\mathbf{R}\in\Re^{N\times N}$. We adopt the convention that $\mathbf{0}$ is a null vector/matrix of appropriate dimension. Similarly, $\mathbf{I}$ is employed for the identity matrix with appropriate dimension.

Section II derives the ODD detector and evaluates its performance when both the matrix $\mathbf{H}$ and the noise variance τ are known. The most important result within Section II is Theorem II.1, which was included without a proof in [11]. Here, we give a rigorous proof of the theorem, and we extend it to colored noise with known covariance matrix. Additionally, we investigate the connection between ODD and model selection, which was not treated in [11].

The analysis continues in Section III by decomposing $\mathbf{H}\boldsymbol{\theta}=\mathbf{H}_r\boldsymbol{\theta}_r+\mathbf{H}_s\boldsymbol{\theta}_s$ [12], where the first component bears information on the signal and the second one models interference. The signal component lies in $\langle\mathbf{H}_r\rangle$, the subspace spanned by the columns of $\mathbf{H}_r$, whereas the interference lies in $\langle\mathbf{H}_s\rangle$. Assuming that $\mathbf{H}$ and τ are known, and the subspaces $\langle\mathbf{H}_r\rangle$ and $\langle\mathbf{H}_s\rangle$

Authorized licensed use limited to: Tampereen Teknillinen Korkeakoulu. Downloaded on June 17, 2009 at 02:17 from IEEE Xplore. Restrictions apply.

are linearly independent, we elaborate on the ODD rule to test $\mathbf{x} \sim \mathcal{N}_N(\mathbf{H}_s\boldsymbol{\theta}_s, \tau\mathbf{I})$ versus $\mathbf{x} \sim \mathcal{N}_N(\mathbf{H}_r\boldsymbol{\theta}_r + \mathbf{H}_s\boldsymbol{\theta}_s, \tau\mathbf{I})$. We note that the results within Section III were not published in the conference paper [11].

II. Subspace Signal in Gaussian Noise of Known Level

A. Main Results

The definitions from the previous section lead to the following theorem. For writing the equations within the theorem more compactly, we notate the Gaussian right-tail probability as $\mathbb{Q}(x) = \int_x^\infty 1/\sqrt{2\pi}\exp\left(-1/2y^2\right)\mathrm{d}y$ for an arbitrary $x \in \Re$ [10]. We also use the notation $\lfloor x \rfloor$ for the largest integer less than or equal to the real-valued argument x.

Theorem II.1: For the data sequence $\mathbf{x} = [x_0, \ldots, x_{N-1}]^\top$, we consider the Gaussian density function with zero mean and known variance τ,

$$f(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{(2\pi\tau)^{N/2}}\exp\left(-\frac{1}{2\tau}\left\|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\right\|^2\right) \tag{7}$$

where $\mathbf{H} \in \Re^{N\times k}$ is a known matrix of rank k, $\boldsymbol{\theta} \in \Re^{k\times 1}$ is the vector of parameters ($N > k$), and $\|\cdot\|$ denotes the Euclidean norm. For ODD testing between the hypotheses specified by the model classes

$$\mathcal{M}_0 = \{f(\mathbf{x};\boldsymbol{\theta}) : \boldsymbol{\theta} = \mathbf{0}\}$$
$$\mathcal{M}_1 = \{f(\mathbf{x};\boldsymbol{\theta}) : \boldsymbol{\theta} \neq \mathbf{0}\}$$

we have the following results.

a) For $\boldsymbol{\theta}^0 = \mathbf{0}$, $D(\hat{f}(\mathbf{x} \mid \boldsymbol{\theta}^0) \,\|\, f(\mathbf{x};\boldsymbol{\theta}^0))$ is a convex function that attains its minimum $(k/2)\ln(\pi\exp(1)/6)$ for $\hat{d} = 3k$.

b) After observing $\mathbf{x}$, select $\mathcal{M}_0$ if

$$\max\left(|z_1|, \ldots, |z_k|\right) < \sqrt{3} \tag{8}$$

where $z_j = \left(\mathbf{v}_j^\top\mathbf{H}^\top\mathbf{x}/\sqrt{\ell_j}\right)/\sqrt{\tau}\ \forall j \in \{1, \ldots, k\}$, with the convention that $\ell_1, \ldots, \ell_k$ are the eigenvalues of the matrix $\mathbf{H}^\top\mathbf{H}$, and $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are the corresponding eigenvectors.

c) When condition (8) is satisfied,

$$E1 = 1 - \left(1 - 2\mathbb{Q}(\sqrt{3})\right)^k \approx 1 - 0.917^k. \tag{9}$$

Otherwise,

$$E2 = \prod_{j=1}^{k}\left[\mathbb{Q}\left((2m_j - 1)\sqrt{3}\right) - \mathbb{Q}\left((2m_j + 1)\sqrt{3}\right)\right] \tag{10}$$

where $m_j = \left\lfloor (z_j + \sqrt{3})/(2\sqrt{3}) \right\rfloor\ \forall j \in \{1, \ldots, k\}$.

The proof is deferred to Appendix A. □

B. Discussion

Theorem II.1 and its proof can be easier understood via Fig. 1, which depicts the particular case $k = 2$. The ODD detector selects $\mathcal{M}_1$ whenever $\hat{\boldsymbol{\theta}} \notin B_{\hat{d}/N}(0)$. For testing the condition, we calculate the coordinates of $\hat{\boldsymbol{\theta}}$ in the cartesian system determined by the principal axes of the hyper-ellipsoid. Then we decide $\mathcal{M}_1$ if there exists $j \in \{1, \ldots, k\}$ such that the magnitude of the jth coordinate is larger than one half of the jth side length of the rectangle $B_{\hat{d}/N}(0)$. Fig. 1 illustrates the situation when $\hat{\boldsymbol{\theta}} \notin B_{\hat{d}/N}(0)$, but $\hat{\boldsymbol{\theta}}$ lies in the interior of the hyper-ellipsoid $(\boldsymbol{\theta} - \boldsymbol{\theta}^0)^\top\mathbf{J}_N(\boldsymbol{\theta} - \boldsymbol{\theta}^0) = \hat{d}/N$. Like in Appendix A, we have $\boldsymbol{\theta}^0 = \mathbf{0}$, and $\mathbf{J}_N$ is taken to be constant in the parameter space.

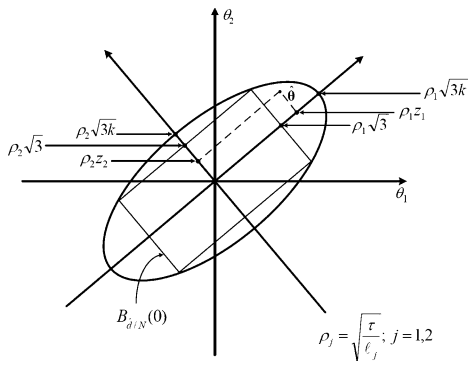

Fig. 1. The ellipse $(\boldsymbol{\theta} - \boldsymbol{\theta}^0)^\top\mathbf{J}_N(\boldsymbol{\theta} - \boldsymbol{\theta}^0) = \hat{d}/N$ and the rectangle $B_{\hat{d}/N}(0)$ when $k = 2$. Note that $\boldsymbol{\theta}^0 = \mathbf{0}$ and $\hat{d} = 6$.

An equivalent form of the condition in (8) is obtained via the singular value decomposition (SVD) of the matrix $\mathbf{H}$. Let

$$\mathbf{H} = \mathbf{U}\mathbf{D}^{1/2}\mathbf{V}^\top \tag{11}$$

where $\mathbf{U} = [\mathbf{u}_1 \ldots \mathbf{u}_k]$ is the matrix formed by the eigenvectors of $\mathbf{H}\mathbf{H}^\top$ that correspond to nonzero eigenvalues. With the notations from Appendix A, we have $\mathbf{V} = [\mathbf{v}_1 \ldots \mathbf{v}_k]$, and $\mathbf{D}$ is the diagonal matrix whose nonzero entries are $\ell_1, \ldots, \ell_k$. Simple calculations lead to the identity $z_j = (\mathbf{u}_j^\top\mathbf{x})/\sqrt{\tau}$ for all $j \in \{1, \ldots, k\}$.

Theorem II.1 can be extended to the case of zero-mean Gaussian noise with known covariance matrix $\mathbf{R}$ that it is not necessarily diagonal. If $\mathbf{R}$ is nonsingular, then its inverse can be factored as $\mathbf{R}^{-1} = \mathbf{L}^\top\mathbf{L}$, where $\mathbf{L}$ is an invertible matrix [13]. Detection in colored noise reduces to replacing, in the results above, $\mathbf{x}$ with $\mathbf{L}\mathbf{x}$, $\mathbf{H}$ with $\mathbf{L}\mathbf{H}$ and τ with one. When $\mathbf{R}$ is singular, the problem can be also solved by discarding the entries of $\mathbf{x}$ that are linearly dependent on the retained entries (see the discussion in [14]).

C. ODD and Other Detectors

To gain more insight, we relate the ODD criterion (8) with the widely used Generalized Likelihood Ratio Test (GLRT). Assuming the hypotheses from Theorem II.1, the GLRT as well as the Rao and Wald test decide $\mathcal{M}_1$ if $\hat{\boldsymbol{\theta}}^\top\mathbf{H}^\top\mathbf{H}\hat{\boldsymbol{\theta}}/\tau > \gamma$, where $\hat{\boldsymbol{\theta}}$ is the ML estimate of $\boldsymbol{\theta}$ for the model class $\mathcal{M}_1$, and the threshold γ is selected based on the desired P_{FA} [10]. Since it is easy to confirm that $\hat{\boldsymbol{\theta}}^\top\mathbf{H}^\top\mathbf{H}\hat{\boldsymbol{\theta}}/\tau = \sum_{j=1}^k z_j^2$, we have:

Proposition II.1:

a) For $k = 1$, the ODD detector is equivalent to the GLRT with $\gamma = 3$.

b) For $k > 1$, there is no γ such that the ODD detector is equivalent to the GLRT. Supplementarily, GLRT with $\gamma = 3$ will select $\mathcal{M}_1$ whenever ODD detector selects $\mathcal{M}_1$.

D. ODD and Model Selection

We now relate the ODD detector with information theoretic criteria (ITC) applied in model selection. For ease of presen-

Authorized licensed use limited to: Tampereen Teknillinen Korkeakoulu. Downloaded on June 17, 2009 at 02:17 from IEEE Xplore. Restrictions apply.

tation, we consider only the particular case when the matrix $\mathbf{H}^\top\mathbf{H}$ is diagonal. Model structure estimation is equivalent to finding a subset of the variables $\theta_1,\ldots,\theta_k$ that minimize the ITC. With the notations used in the Theorem II.1 and its proof, the most popular selection rules can be written in the general form [15]–[17]

$$\begin{aligned}\mathrm{ITC}(\mathbf{x};k) &= \frac{1}{2\tau}\left(\|\mathbf{x}-\mathbf{H}\hat{\boldsymbol{\theta}}\|^2 + k\tau c\right)\\ &= \frac{\|\mathbf{x}\|^2}{2\tau} + \frac{1}{2\tau}\left(-\|\mathbf{H}\hat{\boldsymbol{\theta}}\|^2 + k\tau c\right)\\ &= \frac{\|\mathbf{x}\|^2}{2\tau} - \frac{1}{2}\sum_{j=1}^{k}\left(z_j^2 - c\right)\\ &= \frac{\|\mathbf{x}\|^2}{2\tau} - \frac{1}{2}\sum_{j=1}^{k}\left(\frac{\ell_j}{\tau}\hat{\theta}_j^2 - c\right)\end{aligned}$$

where c is a fixed threshold. For example, $c=2$ for the Akaike Information criterion (AIC) [18], and $c=\ln N$ for the Bayesian information criterion (BIC) [19]. It is evident that the minimum condition for the ITC will retain only the indexes j for which $|\hat{\theta}_j| > (c\tau/\ell_j)^{1/2}$. In terms of Theorem II.1, this is equivalent to selecting the model class $\mathcal{M}_1$ whenever $\max_{1\le j\le k}|z_j| > \sqrt{c}$, which shows clearly the connection between the ODD detector and the ITC. Remark that the ODD detector is more similar to AIC than to BIC, in the sense that c does not depend on the sample size N. This is surprising because ODD was derived from the MDL principle, and a two-part code criterion equivalent to BIC was also obtained by resorting to the same principle [20], whereas AIC has different grounds [18]. The interested reader can find more details on the relationship between ITC and the GLRT in [21] and [22].

E. Confidence Indexes $E1$, $E2$ and the Probabilities P_{FA}, P_D

It is customary to asses the performance of a detector by evaluating P_{FA} and P_D. For the decision rule (8), we get:

$$\begin{aligned}P_{\mathrm{FA}} &= E1,\\ P_D &= 1-\prod_{j=1}^{k}\left(\mathbb{Q}(-\sqrt{3}-\zeta_j)-\mathbb{Q}(\sqrt{3}-\zeta_j)\right)\end{aligned} \quad (12)$$

where $\zeta_j = \mathbf{v}_j^\top\underline{\boldsymbol{\theta}}\sqrt{\ell_j}/\sqrt{\tau}\ \forall j\in\{1,\ldots,k\}$ and $\boldsymbol{\theta}=\underline{\boldsymbol{\theta}}$ for the model class $\mathcal{M}_1$. Notice the major difference between evaluating performance in terms of $E1$ and $E2$ instead of P_{FA} and P_D. The calculation of P_D assumes that data was generated by $\mathcal{M}_1$ with a particular parameter vector $\underline{\boldsymbol{\theta}}$. Such an assumption is not necessary when computing $E1$ and $E2$ because they depend only on the ML estimate $\hat{\boldsymbol{\theta}}$. More precisely, if $\hat{\boldsymbol{\theta}}\in B_{\hat{d}/N}(j)$, then $E1$ and $E2$ depend on the equivalence class defined by the rectangle $B_{\hat{d}/N}(j)$. Therefore, for each rectangle we have a different confidence index whose value is calculated with the $E1$ formula when $\hat{\boldsymbol{\theta}}$ falls into $B_{\hat{d}/N}(0)$, and with the $E2$ formula for all other rectangles.

For illustration, Fig. 2 considers the LM with $k=2$ parameters. We draw, in the (z_1,z_2) plane, the squares obtained from the original rectangles within the (θ_1,θ_2) plane after applying the rotation and the scaling required by the condition in (8). Thus, there exists a bijection from the original $B_{\hat{d}/N}(0)$ to the central square in Fig. 2, where the value of $E1$, the confidence measure for being wrong in accepting $\mathcal{M}_0$ when $\hat{\boldsymbol{\theta}}\in B_{\hat{d}/N}(0)$ is written. Note that $E1$ approaches 1 when k, the number of parameters, is large. Similar bijections also exist for the squares around the central one and for which we indicate the value of $E2$. Observe that the greater the distance is from the center of the square to the null hypothesis, the smaller $E2$ is; hence, the confidence in rejecting $\mathcal{M}_0$ is greater. We mention that $E2$ is smaller than 10^{-7} for all the squares that are situated far away from the null hypothesis, which are not drawn in Fig. 2.

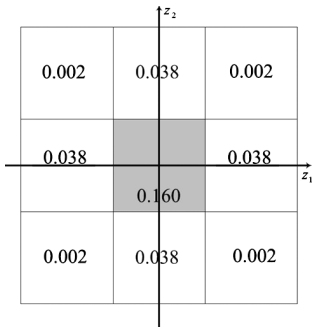

Fig. 2. Linear model with $k=2$ parameters: values of $E1$ (central square) and $E2$ (all other squares). The edge length of each square is $2\sqrt{3}$.

For comparison with the performance of the GLRT, we extend the results of Theorem II.1 by replacing (8) with the modified ODD condition

$$\max\left(|z_1|,\ldots,|z_k|\right) < \sqrt{\frac{d}{k}} \quad (13)$$

and d is chosen such that P_{FA} takes a predefined value. More precisely

$$d = k\left[\mathbb{Q}^{-1}\left(\frac{1-(1-P_{\mathrm{FA}})^{1/k}}{2}\right)\right]^2. \quad (14)$$

We readily obtain

$$P_D = 1-\prod_{j=1}^{k}\left(\mathbb{Q}\left(-\sqrt{\frac{d}{k}}-\zeta_j\right)-\mathbb{Q}\left(\sqrt{\frac{d}{k}}-\zeta_j\right)\right) \quad (15)$$

where the ζ_j are the same as in (12) for all $j\in\{1,\ldots,k\}$.

F. Example: Sinusoidal Detection

One constraint in using the ODD criterion is that the FIM must be nonsingular for the parameters that correspond to the null hypothesis. The condition is not satisfied in sinusoidal detection if the value of the frequency is not known *a priori*. This difficulty was also noticed in connection with the Rao detector [10]. A solution for such cases is the detection method based on the NML of the competing models [23].

Here we consider

$$\mathbf{H} = \begin{bmatrix} 1 & 0\\ \cos(\omega) & \sin(\omega)\\ \vdots & \vdots\\ \cos(\omega(N-1)) & \sin(\omega(N-1))\end{bmatrix}$$

Authorized licensed use limited to: Tampereen Teknillinen Korkeakoulu. Downloaded on June 17, 2009 at 02:17 from IEEE Xplore. Restrictions apply.

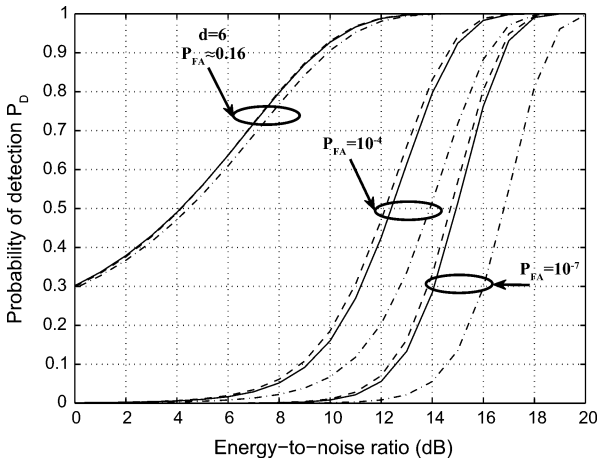

Fig. 3. Performance for sinusoidal detection: GLRT (solid line), best results of ODD (dashed line), worst results of ODD (dashed–dotted line).

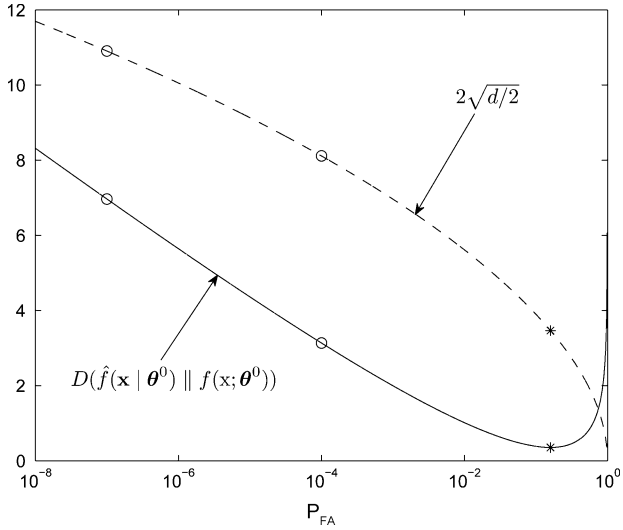

Fig. 4. Sinusoidal detection: For each P_{FA}, the parameter d is computed with formula (14), and the edge length $(2\sqrt{d/2})$ of the central square within the (z_1, z_2) plane is plotted; see the modified detection condition (13). For each P_{FA}, the KL divergence between the "artificial model" (4) assigned to $B_{d/N}(0)$ and the "natural" model (7) evaluated under the null hypothesis is also plotted. The star symbol marks the points that correspond to the optimum $\hat{d} = 6$ given by Theorem II.1 (a). The circles indicate the two cases, $P_{\mathrm{FA}} = 10^{-4}$ and $P_{\mathrm{FA}} = 10^{-7}$, that are compared with the optimal ODD criterion in Fig. 3.

where the frequency $\omega \in (0, \pi)$ is *known*, and we use the asymptotic approximation $\mathbf{H}^\top \mathbf{H} \approx (N/2)\mathbf{I}$ [10]. For the model class $\mathcal{M}_0$, we have $\boldsymbol{\theta} = \mathbf{0}$. For $\mathcal{M}_1$, the parameters are $\theta_1 = A\cos\phi$ and $\theta_2 = -A\sin\phi$, with the convention that $A > 0$ is the unknown amplitude and $\phi \in [-\pi, \pi)$ is the unknown phase. With the notation from (15), $\zeta_1 = \sqrt{\eta}\cos\phi$ and $\zeta_2 = -\sqrt{\eta}\sin\phi$, where $\eta = NA^2/(2\tau)$ is the signal energy-to-noise ratio (ENR) [10]. For a given P_{FA}, the P_D of the ODD detector depends on both η and ϕ. To clarify the dependence on ϕ, we remark that for fixed η and P_{FA}, P_D is maximized when $|\phi| \in \{0, \pi/2, \pi\}$ and it is minimized when $|\phi| \in \{\pi/4, 3\pi/4\}$. In other words, P_D is maximized when either $\zeta_1^2 = \eta$ or $\zeta_2^2 = \eta$, and it is minimized when ENR is equally "distributed" between ζ_1 and ζ_2. For better understanding this result, we relate it to Theorem II.1, where the ODD test amounts to comparing both $|z_1|$ and $|z_2|$ with the threshold $\sqrt{3}$. Hence, the decision does not depend, as in GLRT case, only on the estimated ENR given by $z_1^2 + z_2^2$; it also depends on how the energy of the signal is "distributed" between z_1 and z_2.

Fig. 3 plots the maximum and the minimum of the P_D computed for the ODD criterion when d is chosen to be optimal, namely $d = \hat{d} = 6$ as in Theorem II.1 (a). In this case, we have $P_{\mathrm{FA}} \approx 0.16$, and this is used to compute the P_D of the GLRT detector for various ENR values. The evaluation of the GLRT performance relies on the results from [10]; we emphasize that the P_D of the GLRT is independent of ϕ. For $d = 6$, Fig. 3 shows that ϕ has a marginal influence on the P_D of ODD, and the ODD and the GLRT detectors perform similarly. The main drawback is that the P_{FA} has a value that may be considered too large in most of practical applications. To investigate the cases with lower P_{FA}, we apply (14), which leads to $d \approx 32.90$ when $P_{\mathrm{FA}} = 10^{-4}$ and $d \approx 59.43$ when $P_{\mathrm{FA}} = 10^{-7}$. For both cases, we plot the performance of the ODD and the GLRT in Fig. 3. Note that ϕ has an important influence on P_D of ODD, and this makes the maximum P_D of the ODD superior to the GLRT, but the minimum P_D of the ODD is clearly inferior to GLRT.

The results can be better understood by noting that the KL divergence between the "artificial" and the "natural" models in the ODD settings is about 0.35 for the optimum $\hat{d} = 6$, but it becomes as large as 3.13 and 6.97 for the values of d that correspond to $P_{\mathrm{FA}} = 10^{-4}$ and $P_{\mathrm{FA}} = 10^{-7}$, respectively. It is evident that the constraints on P_{FA} are not in agreement with the ODD methodology; Fig. 4 shows that the central rectangle $B_{d/N}(0)$ must be enlarged by increasing the value of d to ensure a small P_{FA}. This makes the "artificial" model $\hat{f}(\mathbf{x} \mid \boldsymbol{\theta}^0)$ a poor approximation of the "natural" model $f(\mathbf{x}; \theta^0)$. In contrast, the ODD strategy selects $d = \hat{d}$ such that $\hat{f}(\mathbf{x} \mid \boldsymbol{\theta}^0)$ is the best possible approximation of $f(\mathbf{x}; \theta^0)$ in the KL sense, which leads to a slightly large P_{FA}.

III. Subspace Signal in Subspace Interference and Gaussian Noise

A. Main Results

We assume the measurements $x_0, \ldots, x_{N-1}$ to be distributed according to (7), and we write the vector of parameters as

$$\boldsymbol{\theta} = [\boldsymbol{\theta}_r^\top \ \boldsymbol{\theta}_s^\top]^\top$$

where $\boldsymbol{\theta}_r \in \Re^{r\times 1}$ and $\boldsymbol{\theta}_s \in \Re^{s\times 1}$. Moreover, $r + s = k$ and $N > k$. To simplify the calculations, we partition the full-rank matrix $\mathbf{H}$ into two blocks:

$$\mathbf{H} = [\mathbf{H}_r \ \mathbf{H}_s].$$

$\mathbf{H}_r$ contains the first r columns of $\mathbf{H}$, and $\mathbf{H}_s$ is formed by the rest of the columns. Next we define

$$\tilde{\mathbf{H}}_r = \mathbf{P}_{\mathbf{H}_s}^{\perp}\mathbf{H}_r = (\mathbf{I} - \mathbf{P}_{\mathbf{H}_s})\mathbf{H}_r \tag{16}$$

with the convention that $\mathbf{P}_{\mathbf{H}_s} = \mathbf{H}_s\mathbf{H}_s^{\#} = \mathbf{H}_s(\mathbf{H}_s^\top\mathbf{H}_s)^{-1}\mathbf{H}_s^\top$ is the orthogonal projection onto the linear subspace $\langle \mathbf{H}_s \rangle$, and the symbol $^{\#}$ is used for the Moore-Penrose pseudoinverse. The range of $\mathbf{P}_{\mathbf{H}_s}^{\perp} = \mathbf{I} - \mathbf{P}_{\mathbf{H}_s}$ is $\langle \mathbf{H}_s \rangle^{\perp}$, the orthogonal complement

Authorized licensed use limited to: Tampereen Teknillinen Korkeakoulu. Downloaded on June 17, 2009 at 02:17 from IEEE Xplore. Restrictions apply.

of $\langle \mathbf{H}_s \rangle$. More details on the geometry of the linear transformations involved can be found in [12], [14]. The matrix $\hat{\mathbf{H}}_r^\top \hat{\mathbf{H}}_r$ is positive definite [17]. The subspaces $\langle \mathbf{H}_r \rangle$ and $\langle \mathbf{H}_s \rangle$ are linearly independent, but they are not constrained to be orthogonal. Proposition III.1 below shows how the ODD methodology can be applied to detect the signal that lies in the subspace $\langle \mathbf{H}_r \rangle$ when the interference lies in the subspace $\langle \mathbf{H}_s \rangle$ and the additive noise is Gaussian with known variance.

Proposition III.1: For the data sequence $\mathbf{x} = [x_0, \ldots, x_{N-1}]^\top$, consider the Gaussian density function (7) with zero mean and known variance τ. Additionally, $\mathbf{H} \in \Re^{N \times k}$ is a known matrix of rank k. The results on the ODD testing between the hypotheses specified by the model classes

$$\mathcal{M}_0 = \{f(\mathbf{x};\boldsymbol{\theta}) : \boldsymbol{\theta}_r = \mathbf{0}, \boldsymbol{\theta}_s \in \Theta_s\}$$
$$\mathcal{M}_1 = \{f(\mathbf{x};\boldsymbol{\theta}) : \boldsymbol{\theta}_r \neq \mathbf{0}, \boldsymbol{\theta}_s \in \Theta_s\}$$

are obtained by replacing $\mathbf{H}$ by $\tilde{\mathbf{H}}_r$ and k by r in the Theorem II.1.

In the above proposition, Θ_s is a bounded closed subset of $\Re^s$. See Appendix B for the proof. □

B. Discussion

For the detection problem in Proposition III.1, it was shown in [12] that the GLRT remains unchanged when the vector of measurements $\mathbf{x}$ is transformed to $\mathcal{T}(\mathbf{x}) = \mathbf{T}\mathbf{x} + \mathbf{w}$, where $\mathbf{T} = \mathbf{U}_{\tilde{\mathbf{H}}_r} \mathbf{Q} \mathbf{U}_{\tilde{\mathbf{H}}_r}^\top + \mathbf{P}_{\tilde{\mathbf{H}}_r}^\perp$ and $\mathbf{w}$ is an arbitrary vector from $\langle \tilde{\mathbf{H}}_r \rangle^\perp$. Here $\mathbf{U}_{\tilde{\mathbf{H}}_r}$ is an $N \times r$ matrix whose columns form an orthonormal basis for $\langle \tilde{\mathbf{H}}_r \rangle$, $\mathbf{Q} \in \Re^{r \times r}$ is an arbitrary orthogonal matrix, and $\mathbf{P}_{\tilde{\mathbf{H}}_r}^\perp$ is the orthogonal projection onto $\langle \tilde{\mathbf{H}}_r \rangle^\perp$. To better understand the effect of the transformation defined above, consider the decomposition $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$, where $\mathbf{x}_1 \in \langle \tilde{\mathbf{H}}_r \rangle$ and $\mathbf{x}_2 \in \langle \tilde{\mathbf{H}}_r \rangle^\perp$. It is clear that $\mathcal{T}(\cdot)$ rotates the component $\mathbf{x}_1$ around $\langle \tilde{\mathbf{H}}_r \rangle^\perp$, retains the component $\mathbf{x}_2$, and adds the bias component $\mathbf{w}$ in $\langle \tilde{\mathbf{H}}_r \rangle^\perp$ [12]. The invariance property of the GLRT can be easily verified using the formula in (41) from Appendix B. In the detection literature [12], such a property is considered desirable because it makes all the signals of constant energy in $\langle \hat{\mathbf{H}}_r \rangle$ "equally detectable," and it also makes the detector invariant to the components of the signal that are orthogonal to $\langle \tilde{\mathbf{H}}_r \rangle$.

Our concern is the influence of the transformation $\mathcal{T}(\cdot)$ on the ODD decision. Like in (11), we consider the SVD $\tilde{\mathbf{H}}_r = \tilde{\mathbf{U}}_r \tilde{\mathbf{D}}_r^{1/2} \tilde{\mathbf{V}}_r^\top$, where $\tilde{\mathbf{U}}_r \in \Re^{N \times r}$ is the matrix formed by the eigenvectors of $\tilde{\mathbf{H}}_r \tilde{\mathbf{H}}_r^\top$ that correspond to nonzero eigenvalues, $\tilde{\mathbf{D}}_r \in \Re^{r \times r}$ is a diagonal matrix, and $\tilde{\mathbf{V}}_r \in \Re^{r \times r}$ satisfies $\tilde{\mathbf{V}}_r^{-1} = \tilde{\mathbf{V}}_r^\top$. We take $\mathcal{T}(\mathbf{x}) = \left(\tilde{\mathbf{U}}_r \mathbf{Q} \tilde{\mathbf{U}}_r^\top + \mathbf{P}_{\tilde{\mathbf{H}}_r}^\perp\right)\mathbf{x} + \mathbf{w}$, and we define the vectors $\tilde{\mathbf{z}}_{\mathbf{x}} = \tilde{\mathbf{U}}_r^\top \mathbf{x}/\sqrt{\tau}$ and $\tilde{\mathbf{z}}_{\mathcal{T}(\mathbf{x})} = \tilde{\mathbf{U}}_r^\top \mathcal{T}(\mathbf{x})/\sqrt{\tau} = \mathbf{Q}\tilde{\mathbf{z}}_{\mathbf{x}}/\sqrt{\tau}$. Based on the original data vector $\mathbf{x}$, the ODD detector selects $\mathcal{M}_0$ whenever $\max(|\tilde{\mathbf{z}}_{\mathbf{x}}|) < \sqrt{3}$, where $\max(|\cdot|)$ denotes the maximum magnitude for the entries of the vector in the argument. Similarly, for the transformed data vector $\mathcal{T}(\mathbf{x})$, the ODD detector selects $\mathcal{M}_0$ whenever $\max\left(\left|\tilde{\mathbf{z}}_{\mathcal{T}(\mathbf{x})}\right|\right) < \sqrt{3}$. In general, $\max(|\tilde{\mathbf{z}}_{\mathbf{x}}|) < \sqrt{3}$ does not imply $\max\left(\left|\tilde{\mathbf{z}}_{\mathcal{T}(\mathbf{x})}\right|\right) < \sqrt{3}$. This can be easily verified for the following simple example: $\tilde{\mathbf{z}}_{\mathbf{x}} = \sqrt{2}\begin{bmatrix}1\\1\end{bmatrix}$ and $\mathbf{Q} = \left(\sqrt{2}/2\right)\begin{bmatrix}1 & 1\\ -1 & 1\end{bmatrix}$. We conclude that the ODD detector is not invariant to $\mathcal{T}$-transformations like the GLRT.

C. More on the Relation Between Theorem II.1 and Proposition III.1

Let us consider the SVD

$$\mathbf{H}_s = [\mathbf{U}_s \ \mathbf{U}_{s_0}] \left[\mathbf{D}_s^\top \ \mathbf{0}^\top\right]^\top \mathbf{V}_s^\top \tag{17}$$

where the matrix $[\mathbf{U}_s \ \mathbf{U}_{s_0}]$ has orthonormal columns, $\mathbf{U}_s \in \Re^{N \times s}$, and $\mathbf{U}_{s_0} \in \Re^{N \times (N-s)}$. As usual, $\mathbf{D}_s \in \Re^{s \times s}$ is a diagonal matrix and $\mathbf{V}_s \in \Re^{s \times s}$ satisfies $\mathbf{V}_s^{-1} = \mathbf{V}_s^\top$. After applying the transformation

$$\bar{\mathbf{x}} = \mathbf{U}_{s_0}^\top \mathbf{x} \tag{18}$$

the decision as to whether the measurements $\mathbf{x}$ are outcomes from $\mathcal{N}_N(\mathbf{H}_s\boldsymbol{\theta}_s, \tau\mathbf{I})$ or from $\mathcal{N}_N(\mathbf{H}_r\boldsymbol{\theta}_r + \mathbf{H}_s\boldsymbol{\theta}_s, \tau\mathbf{I})$ reduces to the decision as to whether $\bar{\mathbf{x}}$ are outcomes from $\mathcal{N}_{N-s}(\mathbf{0}, \tau\mathbf{I})$ or from $\mathcal{N}_{N-s}(\bar{\mathbf{H}}_r\boldsymbol{\theta}_r, \tau\mathbf{I})$, with the convention that

$$\bar{\mathbf{H}}_r = \mathbf{U}_{s_0}^\top \mathbf{H}_r. \tag{19}$$

Therefore, Theorem II.1 can be applied after replacing the triplet $(\mathbf{x}, k, \mathbf{H})$ by $(\bar{\mathbf{x}}, r, \bar{\mathbf{H}}_r)$, and the model class $\mathcal{M}_0$ is selected whenever $\max_{1 \le j \le r}\left[\left(|\bar{\mathbf{v}}_j^\top \bar{\mathbf{H}}_r^\top \bar{\mathbf{x}}| \Big/ \sqrt{\bar{\ell}_j}\right)/\sqrt{\tau}\right] < \sqrt{3}$, where $\bar{\ell}_1, \ldots, \bar{\ell}_r$ are the eigenvalues of $\bar{\mathbf{H}}_r^\top \bar{\mathbf{H}}_r$ and $\bar{\mathbf{v}}_1, \ldots, \bar{\mathbf{v}}_r$ are the corresponding eigenvectors. Because

$$\mathbf{P}_{\bar{\mathbf{H}}_s}^\perp = \mathbf{U}_{s_0} \mathbf{U}_{s_0}^\top, \tag{20}$$

we have $\tilde{\mathbf{H}}_r = \mathbf{U}_{s_0} \bar{\mathbf{H}}_r$ from (16) and (19). Consequently, $\tilde{\mathbf{H}}_r^\top \tilde{\mathbf{H}}_r = \bar{\mathbf{H}}_r^\top \bar{\mathbf{H}}_r$ and $\left(|\bar{\mathbf{v}}_j^\top \tilde{\mathbf{H}}_r^\top \mathbf{x}| \Big/ \sqrt{\bar{\ell}_j}\right)/\sqrt{\tau} = \left(|\bar{\mathbf{v}}_j^\top \bar{\mathbf{H}}_r^\top \bar{\mathbf{x}}| \Big/ \sqrt{\bar{\ell}_j}\right)/\sqrt{\tau}$, which shows the equivalence between the detection strategy in Proposition III.1 and the approach based on the transformation in (18). The key observation in (18) is that $\mathbf{U}_{s_0}^\top$ nulls everything in the interference subspace $\langle \mathbf{H}_s \rangle$, while the distribution of the noise remains unmodified. The price to be paid is a degradation of $\mathbf{H}_r$ that is transformed to $\bar{\mathbf{H}}_r$.

We quantify the effects of this degradation via the ENR, calculated as $\eta(\mathbf{H}_r, \mathbf{H}_s, \boldsymbol{\theta}_r) = \|\tilde{\mathbf{H}}_r\boldsymbol{\theta}_r\|^2/\tau = \|\bar{\mathbf{H}}_r\boldsymbol{\theta}_r\|^2/\tau$. Conventionally, we denote $\eta(\mathbf{H}_r, \mathbf{0}, \boldsymbol{\theta}_r) = \|\mathbf{H}_r\boldsymbol{\theta}_r\|^2/\tau$ as the ENR in the absence of interference. The ENR reduction is analyzed below in connection with how close the signal and the interference subspaces are. For a rigorous measure of "closeness", we employ the definition of the *principle angles* [24], [25] between the subspaces $\langle \mathbf{H}_r \rangle$ and $\langle \mathbf{H}_s \rangle$. Because the definition involves the SVD of both $\mathbf{H}_r$ and $\mathbf{H}_s$, similar to (17), we write

$$\mathbf{H}_r = [\mathbf{U}_r \ \mathbf{U}_{r_0}] \left[\mathbf{D}_r^\top \ \mathbf{0}^\top\right]^\top \mathbf{V}_r^\top \tag{21}$$

where $\mathbf{U}_r \in \Re^{N \times r}$, $\mathbf{U}_{r_0} \in \Re^{N \times (N-r)}$, $\mathbf{D}_r \in \Re^{r \times r}$, and $\mathbf{V}_r \in \Re^{r \times r}$. The matrix $[\mathbf{U}_r \ \mathbf{U}_{r_0}]$ has orthonormal columns, the diagonal matrix $\mathbf{D}_r$ is invertible, and $\mathbf{V}_r^{-1} = \mathbf{V}_r^\top$. Let

Authorized licensed use limited to: Tampereen Teknillinen Korkeakoulu. Downloaded on June 17, 2009 at 02:17 from IEEE Xplore. Restrictions apply.

$\sigma_1 \geq \cdots \geq \sigma_p \geq 0$ be the singular values of the matrix $\mathbf{U}_r^\top \mathbf{U}_s$, where $p = \min(r, s)$. Then $\alpha_i = \arccos(\sigma_i)$, $i \in \{1, \ldots, p\}$, are the principle angles between $\langle \mathbf{H}_r \rangle$ and $\langle \mathbf{H}_s \rangle$ [24], [25]. Because we assume that $\langle \mathbf{H}_r \rangle$ and $\langle \mathbf{H}_s \rangle$ are disjoint, we have $\min_{1 \leq i \leq p} \alpha_i > 0$; hence, $\alpha_i \in (0, \pi/2]$ $\forall i \in \{1, \ldots, p\}$.

In the corollary below, we take $\boldsymbol{\theta}_r$ to be arbitrary from $\Re^{r \times 1} \setminus \{\mathbf{0}\}$, because the ODD methodology does not need any prior knowledge on the parameter vector. We prove that removing the interference by the transformation in (18) decreases the ENR, except in the case when $\mathbf{H}_r \boldsymbol{\theta}_r \in \langle \mathbf{H}_s \rangle^\perp$. Moreover, it is natural to expect that the impact on the ENR is more important when the subspace $\langle \mathbf{H}_s \rangle$ is closer to $\langle \mathbf{H}_r \rangle$. To clarify this aspect, Corollary III.1 gives necessary and sufficient conditions for the dependence between the ENR reduction and the geometry of the two subspaces. The result appears to be novel, and we formalize it as follows.

Corollary III.1:

a) For $\mathbf{H}_r, \mathbf{H}_s$ like in Proposition III.1 and $\boldsymbol{\theta}_r \in \Re^{r \times 1} \setminus \{\mathbf{0}\}$, we have

$$\eta(\mathbf{H}_r, \mathbf{H}_s, \boldsymbol{\theta}_r) \leq \eta(\mathbf{H}_r, \mathbf{0}, \boldsymbol{\theta}_r) \tag{22}$$

with equality if and only if $\mathbf{H}_r \boldsymbol{\theta}_r \in \langle \mathbf{H}_s \rangle^\perp$.

b) Let $1 \leq r \leq s$ and $\mathbf{H}_r \in \Re^{N \times r}$, $\mathbf{H}_s^{(1)} \in \Re^{N \times s}$, $\mathbf{H}_s^{(2)} \in \Re^{N \times s}$ such that the matrices $\left[\mathbf{H}_r \ \mathbf{H}_s^{(1)}\right]$ and $\left[\mathbf{H}_r \ \mathbf{H}_s^{(2)}\right]$ are full-rank, where $N > r + s$. For $i \in \{1, 2\}$, the principle angles between the subspaces $\langle \mathbf{H}_r \rangle$ and $\langle \mathbf{H}_s^{(i)} \rangle$ are $\alpha_1^{(i)} \leq \cdots \leq \alpha_r^{(i)}$. We have:

b1) If the inequality

$$\eta(\mathbf{H}_r, \mathbf{H}_s^{(1)}, \boldsymbol{\theta}_r) \leq \eta(\mathbf{H}_r, \mathbf{H}_s^{(2)}, \boldsymbol{\theta}_r) \tag{23}$$

holds for all $\boldsymbol{\theta}_r \in \Re^{r \times 1} \setminus \{\mathbf{0}\}$, then $\alpha_j^{(1)} \leq \alpha_j^{(2)}$ $\forall j \in \{1, \ldots, r\}$.

b2) If $\alpha_r^{(1)} \leq \alpha_1^{(2)}$, then the inequality in (23) is verified for all $\boldsymbol{\theta}_r \in \Re^{r \times 1} \setminus \{\mathbf{0}\}$.

The proof is deferred to Appendix C. □

D. Example: Detection in Sinusoidal Interference

We assume that the amplitude and the phase of the interference are unknown, but the frequency is known [12]. Also assume that the signal is known except for its amplitude [10]. Hence, $r = 1$ and $s = 2$. Because $r = 1$, we obtain immediately from Proposition II.1 and Proposition III.1 that, except the value of the threshold used in the test, the ODD detector is equivalent to the GLRT, which is analyzed in Example 7.6 from [10].

IV. Conclusion

We investigated the use of the ODD detector for the LM by emphasizing the strengths and the weaknesses of the method. The confidence indexes provided by ODD without assuming knowledge of the true parameter values are an advantage. For the GLRT, the complement set of the critical region is a solid hyper-ellipsoid. The ODD decision does not involve an hyper-ellipsoid, but the largest rectangle within it, and this can reduce P_D for a given P_{FA}, as was apparent from the comparisons with the GLRT. Moreover, the GLRT is invariant to a "natural" class of transformations, whereas the ODD detector does not share the same invariances. The performance of ODD testing can be potentially improved by applying results from lattice theory [26]: for example, in the two-dimensional case, the rectangles can be replaced by hexagons. This would make the theoretical analysis more difficult than that outlined in this paper.

Appendix A
Proof of Theorem II.1

The proof contains three important parts. First, we construct the partition of the parameter space, and then we obtain a closed-form expression for the KL divergence between the most distinguishable models and the real ones. After these preliminaries, the main results of the Theorem II.1 are proven.

Partition of the Parameter Space: The FIM for the model class $\mathcal{M}_1$ is given by $\mathbf{J}_N(\boldsymbol{\theta}) = \mathbf{H}^\top \mathbf{H}/(N\tau)$ [13], and does not depend on the values of the parameters $\boldsymbol{\theta}$. To emphasize this property, we use the notation $\mathbf{J}_N$ instead of $\mathbf{J}_N(\boldsymbol{\theta})$. Consider the hyper-ellipsoid centered at $\boldsymbol{\theta}^0 = \mathbf{0}$ and defined by $(\boldsymbol{\theta} - \boldsymbol{\theta}^0)^\top \mathbf{J}_N (\boldsymbol{\theta} - \boldsymbol{\theta}^0) = d/N$, where d is a parameter whose optimal value we will find next. Furthermore, let $B_{d/N}(0)$ be the largest rectangle within this hyper-ellipsoid. Its volume is $|B_{d/N}(0)| = 2^k \prod_{j=1}^k \mu_j$, where $\mu_j = [d/(Nk\lambda_j)]^{1/2}$ and λ_j is the jth eigenvalue of the matrix $\mathbf{J}_N$ [1]. The procedure continues until a complete set of disjoint rectangles whose union is the entire parameter space is defined.

Computation of the KL Divergence: For model class $\mathcal{M}_1$, the ML estimate is given by [13]

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{H}^{\#} \mathbf{x}.$$

The symbol $^{\#}$ is used for the Moore–Penrose pseudoinverse, hence $\mathbf{H}^{\#} = (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top$. The function $g(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta})$ from (2) takes the particular form [16], [17]

$$g(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}) = \frac{|\mathbf{H}^\top \mathbf{H}|^{1/2}}{(2\pi\tau)^{k/2}} \exp\left(-\frac{1}{2\tau} \|\mathbf{H}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\|^2\right). \tag{24}$$

We next compute the KL divergence between the "artificial" model $\hat{f}(\mathbf{x} \mid \boldsymbol{\theta}^0)$ assigned to $B_{d/N}(0)$ and the "natural" model (7) evaluated under the null hypothesis. With the supplementary notation $\mathcal{X}_0 = \{\mathbf{x} : \hat{\boldsymbol{\theta}}(\mathbf{x}) \in B_{d/N}(0)\}$, and applying the definition in (4), we readily obtain

$$D(\hat{f}(\mathbf{x} \mid \boldsymbol{\theta}^0) \parallel f(\mathbf{x}; \boldsymbol{\theta}^0)) = -\ln Q_{d/N}(0) + \frac{1}{Q_{d/N}(0)} \int_{\mathcal{X}_0} f(\mathbf{x}; \hat{\boldsymbol{\theta}}) \ln \frac{f(\mathbf{x}; \hat{\boldsymbol{\theta}})}{f(\mathbf{x}; \boldsymbol{\theta}^0)} \mathrm{d}\mathbf{x}. \tag{25}$$

Because (6) together with (24) leads to

$$Q_{d/N}(0) = \frac{|\mathbf{H}^T \mathbf{H}|^{1/2}}{(2\pi\tau)^{k/2}} |B_{d/N}(0)| \tag{26}$$

$$= \left(\frac{2d}{k\pi}\right)^{k/2}, \tag{27}$$

Authorized licensed use limited to: Tampereen Teknillinen Korkeakoulu. Downloaded on June 17, 2009 at 02:17 from IEEE Xplore. Restrictions apply.

all that remains is to calculate the integral. For an arbitrary $\hat{\boldsymbol{\theta}}$, we define the set $\mathcal{X}_{\hat{\boldsymbol{\theta}}} = \{\mathbf{x} : \hat{\boldsymbol{\theta}}(\mathbf{x}) = \hat{\boldsymbol{\theta}}\}$:

$$\begin{aligned}
&\int_{\mathcal{X}_0} f(\mathbf{x};\hat{\boldsymbol{\theta}}) \ln \frac{f(\mathbf{x};\hat{\boldsymbol{\theta}})}{f(\mathbf{x};\boldsymbol{\theta}^0)} \mathrm{d}\mathbf{x} \\
&\quad = \int_{B_{d/N}(0)} \left[\int_{\mathcal{X}_{\hat{\boldsymbol{\theta}}}} f(\mathbf{x} \mid \hat{\boldsymbol{\theta}}) \mathrm{d}\mathbf{x} \right] g(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\theta}}) \frac{\|\mathbf{H}\hat{\boldsymbol{\theta}}\|^2}{2\tau} \mathrm{d}\hat{\boldsymbol{\theta}} \quad (28) \\
&\quad = \frac{|\mathbf{H}^\top\mathbf{H}|^{1/2}}{(2\pi\tau)^{k/2}} \frac{N}{2} \int_{B_{d/N}(0)} \hat{\boldsymbol{\theta}}^\top \mathbf{J}_N \hat{\boldsymbol{\theta}} \mathrm{d}\hat{\boldsymbol{\theta}} \quad (29) \\
&\quad = \frac{|\mathbf{H}^\top\mathbf{H}|^{1/2}}{(2\pi\tau)^{k/2}} \frac{N}{2} \frac{d}{3N} |B_{d/N}(0)| \quad (30) \\
&\quad = \frac{d}{6} Q_{d/N}(0). \quad (31)
\end{aligned}$$

Note that (28) is obtained by applying the sufficiency factorization (2) and the well-known identity $\|\mathbf{H}\hat{\boldsymbol{\theta}}\|^2 = -\|\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}}\|^2 + \|\mathbf{x}\|^2$ [16]. Since the inner integral in (28) gives unity for a fixed $\hat{\boldsymbol{\theta}}$ [1], (24) and the definition of $\mathbf{J}_N$ yield the equality in (29). Rotation of the coordinates for the integral in (29) and some simple manipulations similar to those from of [1, Ch. 7] lead to (30). The result in (31) is an immediate consequence of (26) and (30). From (25), (27) and (31), we conclude

$$\begin{aligned}
D(\hat{f}(\mathbf{x} \mid \boldsymbol{\theta}^0) \parallel f(\mathbf{x};\boldsymbol{\theta}^0)) &= -\ln Q_{d/N}(0) + \frac{d}{6} \\
&= \frac{k}{2} \ln \frac{k\pi}{2d} + \frac{d}{6}. \quad (32)
\end{aligned}$$

Main Results:

a) From (32), $D(\hat{f}(\mathbf{x} \mid \boldsymbol{\theta}^0) \parallel f(\mathbf{x};\boldsymbol{\theta}^0))$ is a convex function that attains its minimum $(k/2)\ln(\pi\exp(1)/6)$ for $d = 3k$. Therefore, the condition of minimizing the KL divergence between the artificial models and the real ones leads to the optimum value $\hat{d} = 3k$. The proof for $a)$ is similar to that from [1], with the remarkable difference that we do not use asymptotic approximations.

b) $\mathbf{J}_N$ and $\mathbf{H}^\top\mathbf{H}$ have the same set of eigenvectors, and $\lambda_j = \ell_j/(N\tau)\ \forall j \in \{1,\ldots,k\}$. Denote $\mathbf{V} = [\mathbf{v}_1 \ldots \mathbf{v}_k]$, and let $\mathbf{D}$ be the diagonal matrix with the entries $\ell_1,\ldots,\ell_k$ on the main diagonal. According to the ODD testing procedure, we select $\mathcal{M}_0$ if and only if $\hat{\boldsymbol{\theta}} \in B_{\hat{d}/N}(0)$. The condition is equivalent to $|\mathbf{v}_j^\top\hat{\boldsymbol{\theta}}| < \mu_j$ for all $j \in \{1,\ldots,k\}$. Using $d = \hat{d}$ in the definition of μ_j, we obtain the chain of equivalent inequalities

$$\begin{aligned}
|\mathbf{v}_j^\top\hat{\boldsymbol{\theta}}| &< \mu_j, \\
\left|\mathbf{v}_j^\top(\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\mathbf{x}\right| &< \left(\frac{3k}{Nk\lambda_j}\right)^{1/2}, \\
\left|\mathbf{v}_j^\top\mathbf{V}\mathbf{D}^{-1}\mathbf{V}^\top\mathbf{H}^\top\mathbf{x}\right| &< \left(\frac{3}{\frac{N\ell_j}{N\tau}}\right)^{1/2}, \\
\frac{|\mathbf{v}_j^\top\mathbf{H}^\top\mathbf{x}|}{\ell_j} &< \left(\frac{3\tau}{\ell_j}\right)^{1/2},
\end{aligned}$$

which leads to the condition in (8).

c) Denote $\mathbf{z} = [z_1 \ldots z_k]^\top$, and let $\boldsymbol{\xi}$ be a column vector of length k for which all the entries are equal to $\sqrt{3}$. Note from the proof of point b) that

$$\mathbf{z} = \frac{1}{\sqrt{\tau}} \mathbf{D}^{1/2}\mathbf{V}^\top\hat{\boldsymbol{\theta}} \quad (33)$$

which implies $\hat{\boldsymbol{\theta}} \in B_{\hat{d}/N}(0)$ if and only if $-\boldsymbol{\xi} < \mathbf{z} < \boldsymbol{\xi}$. Since $\hat{\boldsymbol{\theta}} \sim \mathcal{N}_k(\boldsymbol{\theta}, \tau(\mathbf{H}^\top\mathbf{H})^{-1})$ [10], it is easy to check that $\mathbf{z} \sim \mathcal{N}_k(\mathbf{0},\mathbf{I})$ under the hypothesis $\boldsymbol{\theta} = \mathbf{0}$. Then

$$\begin{aligned}
E1 &= 1 - P_{0|0} \\
&= 1 - \mathrm{Prob}\{-\boldsymbol{\xi} < \mathbf{z} < \boldsymbol{\xi}; \boldsymbol{\theta} = \mathbf{0}\} \\
&= 1 - 2^k \prod_{j=1}^{k} \left[\frac{1}{2} - \mathbb{Q}(\sqrt{3})\right] \\
&= 1 - \left[1 - 2\mathbb{Q}(\sqrt{3})\right]^k \\
&\approx 1 - 0.917^k
\end{aligned}$$

and we get (9).

Before computing $E2$, remark that the parameter space Θ is partitioned into congruent rectangles because the matrix $\mathbf{J}_N$ does not depend on $\boldsymbol{\theta}$. Consequently, the $\mathbf{z}$-space is partitioned into congruent hypercubes whose sides have length $2\sqrt{3}$. The hypercube centered at $\mathbf{z} = \mathbf{0}$ is the one associated with model $\mathcal{M}_0$.

Assume, without loss of generality, that the ML estimate $\hat{\boldsymbol{\theta}}$ falls within $B_{\hat{d}/N}(i)$, where $i \neq 0$. Based on (33), $\mathbf{z}$ is located inside the hypercube centered at $2\mathbf{M}\boldsymbol{\xi}$, where $\mathbf{M}$ is a diagonal matrix with the entries $m_1,\ldots,m_k$ on the main diagonal, and $m_j = \lfloor (z_j + \sqrt{3})/(2\sqrt{3}) \rfloor\ \forall j \in \{1,\ldots,k\}$. Moreover, $\boldsymbol{\theta}^i = \sqrt{\tau}\mathbf{V}\mathbf{D}^{-1/2}(2\mathbf{M}\boldsymbol{\xi})$. The evaluation of E_2 is as follows:

$$\begin{aligned}
E2 &= P_{0|i} \\
&= P\left(B_{\hat{d}/N}(0) \mid \boldsymbol{\theta} = \boldsymbol{\theta}^i\right) \\
&= \mathrm{Prob}\left\{-\boldsymbol{\xi} < \frac{1}{\sqrt{\tau}}\mathbf{D}^{1/2}\mathbf{V}^\top\mathbf{H}^{\#}\mathbf{x} < \boldsymbol{\xi}; \boldsymbol{\theta} = \boldsymbol{\theta}^i\right\} \\
&= \prod_{j=1}^{k} \Big[\mathbb{Q}\left(-(2m_j+1)\sqrt{3}\right) \\
&\qquad - \mathbb{Q}\left(-(2m_j-1)\sqrt{3}\right)\Big] \quad (34) \\
&= \prod_{j=1}^{k} \left[\mathbb{Q}\left((2m_j-1)\sqrt{3}\right) - \mathbb{Q}\left((2m_j+1)\sqrt{3}\right)\right] \quad (35) \\
&= \mathrm{Prob}\left\{(2\mathbf{M}-\mathbf{I})\boldsymbol{\xi} < \mathbf{z} < (2\mathbf{M}+\mathbf{I})\boldsymbol{\xi}; \boldsymbol{\theta} = \mathbf{0}\right\} \\
&= P\left(B_{\hat{d}/N}(i) \mid \boldsymbol{\theta} = \mathbf{0}\right) \\
&= P_{i|0}.
\end{aligned}$$

The key observation for proving (34) is that $(\mathbf{D}^{1/2}\mathbf{V}^\top\mathbf{H}^{\#}\mathbf{x})/\sqrt{\tau} \sim \mathcal{N}_k(2\mathbf{M}\boldsymbol{\xi}, \mathbf{I})$ when $\boldsymbol{\theta} = \boldsymbol{\theta}^i$. Equation (35), which coincides with (10), is obtained by resorting to the properties of the right-tail probability $\mathbb{Q}(\cdot)$. The above calculations verify that $P_{0|i} = P_{i|0}$ for an arbitrary index $i \neq 0$. It is easy to extend the results by observing that $P_{i|i} = P_{0|0}$ for all i. □

Authorized licensed use limited to: Tampereen Teknillinen Korkeakoulu. Downloaded on June 17, 2009 at 02:17 from IEEE Xplore. Restrictions apply.

APPENDIX B
PROOF OF PROPOSITION III.1

Partition of the Parameter Space: For the model class $\mathcal{M}_1$, let $\check{\boldsymbol{\theta}} = [\check{\boldsymbol{\theta}}_r^\top \ \check{\boldsymbol{\theta}}_s^\top]^\top$ be an estimate of $\boldsymbol{\theta}$ distributed as $\mathcal{N}_k(\boldsymbol{\theta}, \mathbf{C})$. The Cramér–Rao bound guarantees that

$$\mathbf{C} \geq (N\mathbf{J}_N)^{-1} = \tau \begin{bmatrix} \mathbf{H}_r^\top \mathbf{H}_r & \mathbf{H}_r^\top \mathbf{H}_s \\ \mathbf{H}_s^\top \mathbf{H}_r & \mathbf{H}_s^\top \mathbf{H}_s \end{bmatrix}^{-1} = \tau \begin{bmatrix} \left(\mathbf{H}_r^\top (\mathbf{I} - \mathbf{P}_{\mathbf{H}_s}) \mathbf{H}_r\right)^{-1} & * \\ * & * \end{bmatrix}$$

[14]. It is straightforward to write the $r \times r$ northwest block as $(\tilde{\mathbf{H}}_r^\top \tilde{\mathbf{H}}_r)^{-1}$. We do not give the closed-form expressions for the other three blocks of the matrix $\mathbf{J}_N^{-1}$ because they are not important in our detection problem.

It can be easily checked that $E[(\check{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r)(\check{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r)^\top] \geq \tau(\tilde{\mathbf{H}}_r^\top \tilde{\mathbf{H}}_r)^{-1} \geq \tau(\mathbf{H}_r^\top \mathbf{H}_r)^{-1}$ [27]. Based on these results, we choose $B^r_{d/N}(0)$ to be the largest rectangle within the hyper-ellipsoid $(\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^0)^\top \tilde{\mathbf{J}}_N^r (\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^0) = d/N$, where $\tilde{\mathbf{J}}_N^r = \tilde{\mathbf{H}}_r^\top \tilde{\mathbf{H}}_r/(N\tau)$ and $\boldsymbol{\theta}_r^0 = \mathbf{0}$. We focus next on determining the optimal value of d.

Computation of the KL Divergence: For $\mathcal{M}_0$, we have $\hat{\boldsymbol{\theta}}_s^0 = \mathbf{H}_s^{\#}\mathbf{x}$, and for $\mathcal{M}_1$, the ML estimates are $[\hat{\boldsymbol{\theta}}_r^\top \ \hat{\boldsymbol{\theta}}_s^\top]^\top = [\mathbf{H}_r \ \mathbf{H}_s]^{\#}\mathbf{x}$. It can be easily shown that [24]

$$\hat{\boldsymbol{\theta}}_r = \tilde{\mathbf{H}}_r^{\#}\mathbf{x}. \tag{36}$$

The region of the parameter space associated with $\mathcal{M}_0$ is given by the cartesian product $B^r_{d/N}(0) \times \Theta_s$, and according to (4), the density function $f(\mathbf{x} \mid \boldsymbol{\theta}^0)$ is zero outside this region. Inside the $\mathcal{M}_0$ region, $f(\mathbf{x} \mid \boldsymbol{\theta}^0) = f(\mathbf{x}; \hat{\boldsymbol{\theta}}_r, \hat{\boldsymbol{\theta}}_s)/Q^r_{d/N}(0)$, where the normalization factor is

$$\begin{aligned} Q^r_{d/N}(0) &= \int_{\Theta_s} \int_{B^r_{d/N}(0)} g(\hat{\boldsymbol{\theta}}_r, \hat{\boldsymbol{\theta}}_s; \hat{\boldsymbol{\theta}}_r, \hat{\boldsymbol{\theta}}_r) \mathrm{d}\hat{\boldsymbol{\theta}}_r \mathrm{d}\hat{\boldsymbol{\theta}}_s \\ &= \frac{|\mathbf{H}^\top \mathbf{H}|^{1/2}}{(2\pi\tau)^{k/2}} |B^r_{d/N}(0)||\Theta_s| && (37) \\ &= \frac{|\tilde{\mathbf{H}}_r^\top \tilde{\mathbf{H}}_r|^{1/2} |\mathbf{H}_s^\top \mathbf{H}_s|^{1/2}}{(2\pi\tau)^{k/2}} |B^r_{d/N}(0)||\Theta_s| && (38) \\ &= \left(\frac{2d}{r\pi}\right)^{r/2} \frac{|\mathbf{H}_s^\top \mathbf{H}_s|^{1/2} |\Theta_s|}{(2\pi\tau)^{s/2}}. && (39) \end{aligned}$$

The equality in (37) is obtained immediately via (24). Equation (38) exploits the Cholesky factorization of $\mathbf{J}_N$ as given in [14]. Then, we get (39) by using the formula for $|B^r_{d/N}(0)|$.

Because $\mathcal{M}_0$ is not a singleton class, we proceed as in [23] by selecting the "natural" model for $\mathcal{M}_0$ to be the NML density function (1): $\tilde{f}_0(\mathbf{x}) = f\left(\mathbf{x}; \boldsymbol{\theta}_r^0, \hat{\boldsymbol{\theta}}_s^0\right) \Big/ C_0$, where the normalization factor is given by $C_0 = \int_{\mathbf{x}: \hat{\boldsymbol{\theta}}_s^0(\mathbf{x}) \in \Theta_s} f\left(\mathbf{x}; \boldsymbol{\theta}_r^0, \hat{\boldsymbol{\theta}}_s^0\right) \mathrm{d}\mathbf{x}$. For the computation of C_0, we refer to [1], [16]. Here, we do not need to calculate C_0; it is enough to assume $C_0 < \infty$. We refer to [2] (see page 406) for a more general discussion on choosing between the use of ML or NML in statistical inference.

With the notation $\mathcal{X}_0^r = \{\mathbf{x} : \hat{\boldsymbol{\theta}}_r(\mathbf{x}) \in B^r_{d/N}(0), \hat{\boldsymbol{\theta}}_s \in \Theta_s\}$, we have the following expression of the KL divergence

$$D(\hat{f}(\mathbf{x}|\boldsymbol{\theta}^0) \,\|\, \tilde{f}_0(\mathbf{x})) = \ln \frac{C_0}{Q^r_{d/N}(0)} + \frac{1}{Q^r_{d/N}(0)} \int_{\mathcal{X}_0^r} f(\mathbf{x}; \hat{\boldsymbol{\theta}}_r, \hat{\boldsymbol{\theta}}_s) \times \ln \frac{f(\mathbf{x}; \hat{\boldsymbol{\theta}}_r, \hat{\boldsymbol{\theta}}_s)}{f\left(\mathbf{x}; \boldsymbol{\theta}_r^0, \hat{\boldsymbol{\theta}}_s^0\right)} \mathrm{d}\mathbf{x} \tag{40}$$

and for its calculation, we first evaluate the likelihood ratio

$$\begin{aligned} \ln \frac{f(\mathbf{x}; \hat{\boldsymbol{\theta}}_r, \hat{\boldsymbol{\theta}}_s)}{f\left(\mathbf{x}; \boldsymbol{\theta}_r^0, \hat{\boldsymbol{\theta}}_s^0\right)} &= \frac{1}{2\tau} \mathbf{x}^\top \mathbf{P}_{\tilde{\mathbf{H}}_r} \mathbf{x} && (41) \\ &= \frac{1}{2\tau} \mathbf{x}^\top \mathbf{P}_{\tilde{\mathbf{H}}_r}^\top \mathbf{P}_{\tilde{\mathbf{H}}_r} \mathbf{x} && (42) \\ &= \frac{1}{2\tau} \left(\tilde{\mathbf{H}}_r^{\#}\mathbf{x}\right)^\top \left(\tilde{\mathbf{H}}_r^\top \tilde{\mathbf{H}}_r\right) \left(\tilde{\mathbf{H}}_r^{\#}\mathbf{x}\right) && (43) \\ &= \frac{1}{2\tau} \|\tilde{\mathbf{H}}_r \hat{\boldsymbol{\theta}}_r\|^2. && (44) \end{aligned}$$

The identity in (41) was obtained in [12]. To get (42), we use the fact that the projection matrix is symmetric and idempotent. Equation (43) is a straightforward application of the definition of the projection matrix, and (44) is a consequence of (36).

Then, we compute the integral

$$\begin{aligned} &\int_{\mathcal{X}_0^r} f(\mathbf{x}; \hat{\boldsymbol{\theta}}_r, \hat{\boldsymbol{\theta}}_s) \ln \frac{f(\mathbf{x}; \hat{\boldsymbol{\theta}}_r, \hat{\boldsymbol{\theta}}_s)}{f\left(\mathbf{x}; \boldsymbol{\theta}_r^0, \hat{\boldsymbol{\theta}}_s^0\right)} \mathrm{d}\mathbf{x} \\ &= \int_{\Theta_s} \int_{B^r_{d/N}(0)} \left[\int_{\mathcal{X}^r_{\hat{\boldsymbol{\theta}}}} f(\mathbf{x} \mid \hat{\boldsymbol{\theta}}_r, \hat{\boldsymbol{\theta}}_s) \mathrm{d}\mathbf{x} \right] \\ &\quad \times g(\hat{\boldsymbol{\theta}}_r, \hat{\boldsymbol{\theta}}_s; \hat{\boldsymbol{\theta}}_r, \hat{\boldsymbol{\theta}}_s) \frac{\|\tilde{\mathbf{H}}_r \hat{\boldsymbol{\theta}}_r\|^2}{2\tau} \mathrm{d}\hat{\boldsymbol{\theta}}_r \mathrm{d}\hat{\boldsymbol{\theta}}_s && (45) \\ &= \frac{|\mathbf{H}^\top \mathbf{H}|^{1/2}}{(2\pi\tau)^{k/2}} \frac{N}{2} \int_{\Theta_s} \int_{B_{d/N}(0)} \hat{\boldsymbol{\theta}}_r^\top \tilde{\mathbf{J}}_N^r \hat{\boldsymbol{\theta}}_r \mathrm{d}\hat{\boldsymbol{\theta}}_r \mathrm{d}\hat{\boldsymbol{\theta}}_s \\ &= \frac{|\mathbf{H}^\top \mathbf{H}|^{1/2}}{(2\pi\tau)^{k/2}} \frac{N}{2} \frac{d}{3N} |B^r_{d/N}(0)||\Theta_s| \\ &= \frac{d}{6} Q^r_{d/N}(0). && (46) \end{aligned}$$

For $\hat{\boldsymbol{\theta}} = [\hat{\boldsymbol{\theta}}_r^\top \ \hat{\boldsymbol{\theta}}_s^\top]^\top$ with $\hat{\boldsymbol{\theta}}_s \in \Theta_s$, we have used the notation $\mathcal{X}^r_{\hat{\boldsymbol{\theta}}} = \{\mathbf{x} : \hat{\boldsymbol{\theta}}(\mathbf{x}) = \hat{\boldsymbol{\theta}}\}$ in the calculations above. The innermost integral in (45) evaluates to one for a fixed $\hat{\boldsymbol{\theta}}$. To get (46), we have also used (2), (24) and (37), together with the same type of reasoning that earlier led from (29) to (30).

Main Results: The identities in (39), (40), and (46) lead to

$$D(\hat{f}(\mathbf{x} \mid \boldsymbol{\theta}^0) \,\|\, \tilde{f}_0(\mathbf{x})) = \ln \frac{C_0 (r\pi)^{r/2} (2\pi\tau)^{s/2}}{2^{r/2} |\mathbf{H}_s^\top \mathbf{H}_s|^{1/2} |\Theta_s|} - \frac{r}{2} \ln d + \frac{d}{6}$$

Authorized licensed use limited to: Tampereen Teknillinen Korkeakoulu. Downloaded on June 17, 2009 at 02:17 from IEEE Xplore. Restrictions apply.

which is minimized by $\hat{d} = 3r$. For selecting the optimum d, we do not need closed-form expressions for C_0 and $|\Theta_s|$; it is enough to assume that both of them are finite. The calculations above also show that selecting the "natural" model for $\mathcal{M}_0$ to be the ML function $f\left(\mathbf{x};\boldsymbol{\theta}_r^0,\hat{\boldsymbol{\theta}}_s^0\right)$ instead of the NML function $\tilde{f}_0(\mathbf{x}) = f\left(\mathbf{x};\boldsymbol{\theta}_r^0,\hat{\boldsymbol{\theta}}_s^0\right)\Big/C_0$ leads to the same optimum $\hat{d} = 3r$. The KL divergence between the "artificial" and the "natural" models will be different when choosing ML instead of the NML, but this is less important for our detection problem.

The remaining steps of the proof are similar to those from Appendix A, and we skip them for brevity. □

Appendix C
Proof of Corollary III.1

a) We employ the definition of $\eta(\mathbf{H}_r,\mathbf{H}_s,\boldsymbol{\theta}_r)$ and (16) and (17) to get

$$\begin{aligned}\eta(\mathbf{H}_r,\mathbf{H}_s,\boldsymbol{\theta}_r) &= \frac{\|\tilde{\mathbf{H}}_r\boldsymbol{\theta}_r\|^2}{\tau}\\ &= \frac{\boldsymbol{\theta}_r^\top\mathbf{H}_r^\top(\mathbf{I}-\mathbf{P}_{\mathbf{H}_s})\mathbf{H}_r\boldsymbol{\theta}_r}{\tau}\\ &= \eta(\mathbf{H}_r,\mathbf{0},\boldsymbol{\theta}_r) - \frac{\|\mathbf{U}_s^\top\mathbf{H}_r\boldsymbol{\theta}_r\|^2}{\tau}\end{aligned}$$

which proves the inequality in (22). The equality occurs if and only if $\|\mathbf{U}_s^\top\mathbf{H}_r\boldsymbol{\theta}_r\|^2 = 0$. This is equivalent to $\mathbf{H}_r\boldsymbol{\theta}_r \in \langle\mathbf{H}_s\rangle^\perp$ because the columns of $\mathbf{H}_r$ are linearly independent and the null space of $\mathbf{U}_s^\top$ coincides with the orthogonal complement of the range of $\mathbf{H}_s$.

b) From (20) and (21), we obtain

$$\begin{aligned}&\eta(\mathbf{H}_r,\mathbf{H}_s^{(2)},\boldsymbol{\theta}_r) - \eta(\mathbf{H}_r,\mathbf{H}_s^{(1)},\boldsymbol{\theta}_r)\\ &=\frac{1}{\tau}\boldsymbol{\theta}_r^\top\mathbf{H}_r^\top\left(\mathbf{P}^\perp_{\mathbf{H}_s^{(2)}} - \mathbf{P}^\perp_{\mathbf{H}_s^{(1)}}\right)\mathbf{H}_r\boldsymbol{\theta}_r\\ &=\frac{1}{\tau}\boldsymbol{\theta}_r^\top\mathbf{V}_r\mathbf{D}_r^\top\mathbf{U}_r^\top\left(\mathbf{U}_{s_0}^{(2)}\mathbf{U}_{s_0}^{(2)\top} - \mathbf{U}_{s_0}^{(1)}\mathbf{U}_{s_0}^{(1)\top}\right)\\ &\quad\times\mathbf{U}_r\mathbf{D}_r\mathbf{V}_r^\top\boldsymbol{\theta}_r\\ &=\frac{1}{\tau}\boldsymbol{\theta}_r^\top\mathbf{V}_r\mathbf{D}_r^\top\left(\mathbf{A}^{(2)} - \mathbf{A}^{(1)}\right)\mathbf{D}_r\mathbf{V}_r^\top\boldsymbol{\theta}_r\\ &=\frac{1}{\tau}\boldsymbol{\theta}_r^\top\mathbf{V}_r\mathbf{D}_r^\top\mathbf{B}\mathbf{D}_r\mathbf{V}_r^\top\boldsymbol{\theta}_r \qquad (47)\end{aligned}$$

where $\mathbf{B} = \mathbf{A}^{(2)} - \mathbf{A}^{(1)}$. For $i \in \{1,2\}$, $\mathbf{A}^{(i)} = \mathbf{U}_r^\top\mathbf{U}_{s_0}^{(i)}\mathbf{U}_{s_0}^{(i)\top}\mathbf{U}_r$, and let $\lambda_1(\mathbf{A}^{(i)}) \le \cdots \le \lambda_r(\mathbf{A}^{(i)})$ be the eigenvalues of $\mathbf{A}^{(i)}$ arranged in increasing order. Then, we have [24]

$$\lambda_j(\mathbf{A}^{(i)}) = \sin^2(\alpha_j^{(i)}), \quad 1 \le i \le 2,\ 1 \le j \le r. \qquad (48)$$

To complete the proof, we need the following result.

Theorem C.1: [25]If $\mathbf{E}$ and $\mathbf{E}+\mathbf{F}$ are $r \times r$ symmetric matrices, then

$$\begin{aligned}\lambda_j(\mathbf{E}) + \lambda_1(\mathbf{F}) &\le \lambda_j(\mathbf{E}+\mathbf{F})\\ &\le \lambda_j(\mathbf{E}) + \lambda_r(\mathbf{F}),\ j \in \{1,\ldots,r\} \qquad (49)\end{aligned}$$

where for an arbitrary symmetric matrix $\mathbf{S}$, the notation $\lambda_j(\mathbf{S})$ designates the jth smallest eigenvalue such that $\lambda_1(\mathbf{S}) \le \cdots \le \lambda_r(\mathbf{S})$.

b1) For an arbitrary $\mathbf{w} \in \Re^{r\times 1} \setminus \{\mathbf{0}\}$, we define $\boldsymbol{\theta}_r(\mathbf{w}) = \mathbf{V}_r\mathbf{D}_r^{-1}\mathbf{w}$. From the assumptions of the Corollary III.1 b1), we have $\eta(\mathbf{H}_r,\mathbf{H}_s^{(2)},\boldsymbol{\theta}_r(\mathbf{w})) - \eta(\mathbf{H}_r,\mathbf{H}_s^{(1)},\boldsymbol{\theta}_r(\mathbf{w})) \ge 0$, and using (47) we get $\mathbf{w}^\top\mathbf{B}\mathbf{w} \ge 0$. Hence, the matrix $\mathbf{B}$ is positive semidefinite, or equivalently, its minimum eigenvalue $\lambda_1(\mathbf{B})$ is nonnegative. By choosing $\mathbf{E} = \mathbf{A}^{(1)}$ and $\mathbf{F} = \mathbf{B}$ in (49), we obtain the inequality $\lambda_j(\mathbf{A}^{(1)}) + \lambda_1(\mathbf{B}) \le \lambda_j(\mathbf{A}^{(2)})$, $j \in \{1,\ldots,r\}$. This result, together with (48) and the fact that $\sin^2(\cdot)$ is monotonically increasing on $(0,\pi/2]$, leads to $\alpha_j^{(1)} \le \alpha_j^{(2)}\ \forall j \in \{1,\ldots,r\}$.

b2) We again apply (49) by choosing $\mathbf{E} = \mathbf{B}$, $\mathbf{F} = \mathbf{A}^{(1)}$ and $j = 1$, which leads to

$$\begin{aligned}\lambda_1(\mathbf{B}) &\ge \lambda_1(\mathbf{A}^{(2)}) - \lambda_r(\mathbf{A}^{(1)})\\ &= \sin^2(\alpha_1^{(2)}) - \sin^2(\alpha_r^{(1)}) \qquad (50)\\ &\ge 0 \qquad (51)\end{aligned}$$

where (50) is a straightforward application of (48), and (51) is due to the assumptions of Corollary III.1 b2). Therefore, the matrix $\mathbf{B}$ is positive semidefinite and the inequality in (23) is readily obtained using (47).

References

[1] J. Rissanen, *Information and Complexity in Statistical Modeling*. New York: Springer-Verlag, 2007.

[2] P. Grünwald, *The Minimum Description Length principle*. Cambridge, MA: MIT Press, 2007.

[3] M. Li and P. Vitanyi, *An Introduction to Kolmogorov Complexity and its Applications*. New York: Springer-Verlag, 1997.

[4] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Inf. Theory*, vol. 44, pp. 2743–2760, Oct. 1998.

[5] P. Stoica and Y. Selen, "A review of information criterion rules," *IEEE Signal. Process. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.

[6] G. Qian and H. Künsch, "Some notes on Rissanen's stochastic complexity," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 782–786, Mar. 1998.

[7] J. Rissanen, "The structure function and distinguishable models of data," *Comput. J.*, vol. 49, no. 6, pp. 657–664, 2006.

[8] V. Balasubramanian, "Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions," *Neural Comput.*, vol. 9, no. 2, pp. 349–368, 1997.

[9] J. Rissanen, *Optimally Distinguishable Distributions*, p. 8, Sep. 2007.

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

[11] S. Razavi and C. Giurcăneanu, "Composite hypothesis testing by optimally distinguishable distributions," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Las Vegas, NV, Mar. 4, 2008, pp. 3897–3900.

[12] L. Scharf and B. Friedlander, "Matched subspace detectors," *IEEE Trans. Signal. Process.*, vol. 42, no. 8, pp. 2146–2157, Aug. 1994.

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

[14] L. Scharf and L. McWhorter, "Geometry of the Cramer-Rao bound," *Signal Process.*, vol. 31, pp. 301–311, 1993.

[15] D. Foster and E. George, "The risk inflation criterion for multiple regression," *Ann. Stat.*, vol. 22, no. 4, pp. 1947–1975, 1994.

[16] E. Liski, "Normalized ML and the MDL principle for variable selection in linear regression," in *Festschrift for Tarmo Pukkila on His 60th Birthday*, E. Liski, J. Isotalo, J. Niemelä, S. Puntanen, and G. Styan, Eds. Tampere, Finland: Univ. of Tampere, 2006, pp. 159–172.

[17] G. Seber and A. Lee, *Linear Regression Analysis*. New York: Wiley-Interscience, 2003.

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

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

[20] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

Authorized licensed use limited to: Tampereen Teknillinen Korkeakoulu. Downloaded on June 17, 2009 at 02:17 from IEEE Xplore. Restrictions apply.

[21] T. Söderström, "On model structure testing in system identification," *Int. J. Control*, vol. 26, no. 1, pp. 1–18, 1977.
[22] P. Stoica, Y. Selen, and J. Li, "On information criteria and the generalized likelihood ratio test of model order selection," *IEEE Signal Process. Lett.*, vol. 11, pp. 794–797, 2004.
[23] J. Rissanen, "Hypothesis selection and testing by the MDL principle," *Comput. J.*, vol. 42, no. 4, pp. 260–269, 1999.
[24] R. Behrens and L. Scharf, "Signal processing applications of oblique projection operators," *IEEE Trans. Signal. Process.*, vol. 42, no. 6, pp. 1413–1424, Jun. 1994.
[25] G. Golub and C. van Loan, *Matrix Computations*. Baltimore, MD: The Johns Hopkins Univ. Press, 1996.
[26] J. Conway and N. Sloane, *Sphere Packings, Lattices and Groups*. New York: Springer-Verlag, 1988.
[27] T. McWhorter and L. Scharf, "Cramer-Rao bounds for deterministic modal analysis," *IEEE Trans. Signal. Process.*, vol. 41, no. 5, pp. 1847–1866, May 1993.

Seyed Alireza Razavi (S'08) was born in Birjand, Iran, in 1973. He received the B.S. and M.S. degrees both in electrical engineering from Amirkabir University of Technology, Tehran, Iran, in 1997 and 2000, respectively.

From 2000 to 2007, he was with the Faculty of Engineering, University of Birjand, Birjand, Iran, where he served as a Member of Academic Staff. Since 2007, he has been with the Department of Signal Processing, Tampere University of Technology, Tampere, Finland, where he is working towards the Ph.D. degree. His area of interest include statistical signal processing, information theory and wireless communications.

Ciprian Doru Giurcăneanu (S'98–M'02) received the Ph.D. degree (with honors) from the Department of Information Technology, Tampere University of Technology, Finland, in 2001.

From 1993 to 1997, he was a Junior Assistant at "Politehnica" University of Bucharest, and since 1997 he has been with Tampere University of Technology. He is currently a Research Fellow with the Academy of Finland. His research focuses on stochastic complexity and its applications.

Dr. Giurcăneanu has been the Chair of the IEEE Finland joint Signal Processing and Circuits and Systems Chapter since 2006.

Authorized licensed use limited to: Tampereen Teknillinen Korkeakoulu. Downloaded on June 17, 2009 at 02:17 from IEEE Xplore. Restrictions apply.