COMPOSITE HYPOTHESIS TESTING BY OPTIMALLY DISTINGUISHABLE DISTRIBUTIONS

Seyed Alireza Razavi and Ciprian Doru Giurcăneanu

Tampere University of Technology, Institute of Signal Processing
P.O. Box 553, FIN-33101 Tampere, Finland
alireza.razavi@tut.fi, ciprian.giurcaneanu@tut.fi

ABSTRACT

Relying on optimally distinguishable distributions (ODD), it was defined very recently a new framework for the composite hypothesis testing. We resort to the linear model to investigate the performances of the ODD detector and to compare it with the widely used Generalized Likelihood Ratio Test (GLRT). As the ODD concept is very new, its application to models with nuisance parameters was not discussed in the previous literature. The present study attempts to fill the gap by proposing a modified ODD criterion to accommodate the practical case of unknown noise variance.

Index Terms— Composite hypothesis testing, optimally distinguishable distributions, linear model, Kolmogorov structure function, Generalized Likelihood Ratio Test.

1. INTRODUCTION AND PRELIMINARIES

The most recent developments in methods of inference based on the Minimum Description Length (MDL) principle emerge from a happy union between the algorithmic complexity theory and the coding theory [1],[2]. As the central notions from the algorithmic complexity theory [3], namely *Kolmogorov complexity*, *universal distribution* and the *structure function* are non-computable, their use in practical applications poses troubles. To circumvent such difficulties, Rissanen extends all these notions to statistical models by replacing the set of programs from the algebraic theory of complexity 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 $\Re^k$ [2]. 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]:

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

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 $\hat{\boldsymbol{\theta}}$ is 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]

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

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 the *Kolmogorov complexity* is replaced by the stochastic complexity (SC) whose expression is given by $\ln(1/\tilde{f}(\mathbf{x}))$.

To construct the *Kolmogorov structure function*, the parameter space Θ is partitioned into rectangles such that the Kullback-Leibler (KL) distance between any two adjacent models is constant [2]. We outline the steps of the construction as they are given in [2]. 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 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 $(\boldsymbol{\theta}-\overline{\boldsymbol{\theta}})^\top\mathbf{J}_N(\overline{\boldsymbol{\theta}})(\boldsymbol{\theta}-\overline{\boldsymbol{\theta}}) = d/N$, 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 the procedure until defining a complete set of $\mathfrak{N}_{d/N}$ disjoint rectangles whose reunion is the entire parameter space Θ. With the conventions from [2], we dub $B_{d/N}(j)$ the j-th rectangle within this set, and we denote $\boldsymbol{\theta}^j$ its center. For all $j \in \{0,\ldots,\mathfrak{N}_{d/N}-1\}$, the probability distribution $\hat{f}(\mathbf{x} \mid \boldsymbol{\theta}^j)$ is defined by

$$\hat{f}(\mathbf{x} \mid \boldsymbol{\theta}^j) \quad = \quad \begin{cases} f(\mathbf{x};\hat{\boldsymbol{\theta}}(\mathbf{x}))/Q_{d/N}(j), & \hat{\boldsymbol{\theta}}(\mathbf{x}) \in B_{d/N}(j) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

$$Q_{d/N}(i) \quad = \quad \int_{\hat{\boldsymbol{\theta}}\in B_{d/N}(i)} g(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\theta}})\mathrm{d}\hat{\boldsymbol{\theta}}, \tag{4}$$

where $g(\hat{\boldsymbol{\theta}};\boldsymbol{\theta})$ is the same as in (2). The key point is that the newly introduced distributions are perfectly distinguishable, and here the sense of distinguishability is borrowed from the differential geometry [6]. In [7], Rissanen proposes an index to measure the distinguishability, and this allows to prove that $\hat{f}(\mathbf{x} \mid \boldsymbol{\theta}^j)$ are *optimally distinguishable distributions* (ODD).

The KL distance $D(\hat{f}(\mathbf{x} \mid \boldsymbol{\theta}^j) \parallel f(\mathbf{x};\boldsymbol{\theta}^j))$ between the "artificial" model $\hat{f}(\mathbf{x} \mid \boldsymbol{\theta}^j)$ and the "natural" model $f(\mathbf{x};\boldsymbol{\theta}^j)$ depends on the parameter d for all $j \in \{0,\ldots,\mathfrak{N}_{d/N}-1\}$. If the Central Limit Theorem is verified, then there exists a unique $\hat{d}$ that minimizes this distance, and asymptotically $\hat{d} = 3k$ [2].

These findings can be applied almost straightforward for composite hypothesis testing and, more importantly, they define a totally new framework for this problem. We explain briefly the ODD testing between the hypotheses specified by $\mathcal{M}_0 = \{f(\mathbf{x};\boldsymbol{\theta}) : \boldsymbol{\theta} = \boldsymbol{\theta}^0\}$ and $\mathcal{M}_1 = \{f(\mathbf{x};\boldsymbol{\theta}) : \boldsymbol{\theta} \neq \boldsymbol{\theta}^0\}$. It is evident in this case that, for partitioning the parameter space, we first demarcate the rectangle centered at the point $\boldsymbol{\theta}^0$ and denoted $B_{\hat{d}/N}(0)$, second we fix the centers of its neighbors, and then we continue the construction until the complete set of rectangles is settled. The ODD criterion selects

This work was supported by the Academy of Finland, project No. 113572, 118355 and 213462.

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 $\mathcal{M}_1$ [2].

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 it is done in the traditional Neyman-Pearson methodology [8]. However, the performances of the ODD procedure are assessed by calculating two indices [7]: $E1 = 1 - P_{0|0}$ and $E2 = P_{0|j}$ for $j \neq 0$. We note that, for an arbitrary pair (i, j), $P_{i|j}$ is the probability mass of $B_{\hat{d}/N}(i)$ induced by the model $f(\mathbf{x}; \boldsymbol{\theta}^j)$. The significance of $E1$ and $E2$ will be clarified in the next Section.

The approach based on the ODD testing is very promising, but so far it was applied only in the following examples [1],[2],[7]: (i) for the model class $\mathcal{M}_0$, the observed variable X is Gaussian with mean 0 and variance 1, whereas for $\mathcal{M}_1$, X is Gaussian with non-zero mean and non-unitary 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 $\frac{1}{3}$. Moreover, the ODD criterion for models with nuisance parameters was not introduced in the previous literature. Also it was not yet investigated the relation between P_D and P_{FA} when the parameter space partition is constructed with $\hat{d}$.

In this study, we provide answers to the unsolved problems connected with the ODD testing by considering the linear model (LM). The motivation of our choice is twofold: (i) the most important results obtained in [2] by resorting to asymptotic approximations turns out to be non-asymptotic for the LM; (ii) LM has many applications in signal processing [8].

The rest of the paper is focused on the detection of a deterministic signal with unknown linear parameters in zero-mean Gaussian noise. In Section 2 we derive the ODD detector and evaluate its performances assuming the noise is white with known variance. The case of the unknown noise variance is treated in Section 3.

2. LINEAR MODEL:WHITE GAUSSIAN NOISE WITH *KNOWN* VARIANCE

The main results and definitions from the previous Section lead to the Theorem below. For writing the equations within the Theorem more compactly, we resort to the formula of the right-tail probability: $\mathbb{Q}(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right) \mathrm{d}y$ for an arbitrary $x \in \Re$ [8]. We also use the notation $\lfloor x \rfloor$ for the largest integer less than or equal to the real-valued argument x.

Theorem 2.1. *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,*

$$f(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{(2\pi\tau)^{N/2}} \exp\left(-\frac{1}{2\tau}\|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\|^2\right), \tag{5}$$

where $\mathbf{H}$ is a known $N \times k$ matrix of rank k, $\boldsymbol{\theta}$ is a $k \times 1$ vector of parameters ($N > k + 1$), and $\|\cdot\|$ denotes the Euclidean norm. For the ODD testing between the hypotheses specified by the model classes $\mathcal{M}_0 = \{f(\mathbf{x}; \boldsymbol{\theta}) : \boldsymbol{\theta} = \mathbf{0}\}$ and $\mathcal{M}_1 = \{f(\mathbf{x}; \boldsymbol{\theta}) : \boldsymbol{\theta} \neq \mathbf{0}\}$, we have the following results:
a) For $\boldsymbol{\theta}^0 = \mathbf{0}$, $D(\hat{f}(\mathbf{x} \mid \boldsymbol{\theta}^0) \parallel f(\mathbf{x}; \boldsymbol{\theta}^0))$ is a convex function that attains its minimum $\frac{k}{2} \ln \frac{\pi \exp(1)}{6}$ for $\hat{d} = 3k$.
b) After observing $\mathbf{x}$, select $\mathcal{M}_1$ if

$$\max\left(|z_1|, \ldots, |z_k|\right) > \sqrt{3}, \tag{6}$$

where $z_j = \frac{(\mathbf{v}_j^\top \mathbf{H}^\top \mathbf{x})/\sqrt{\ell_j}}{\sqrt{\tau}}$ $\forall j \in \{1, \ldots, k\}$, 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 the condition (6) is verified, we are wrong in accepting the null hypothesis with probability $E1 = 1 - \left(1 - 2\mathbb{Q}(\sqrt{3})\right)^k \approx 1 - 0.917^k$, otherwise, we are wrong in rejecting the null hypothesis with probability $E2 = \prod_{j=1}^k \left(\mathbb{Q}\left((2m_j - 1)\sqrt{3}\right) - \mathbb{Q}\left((2m_j + 1)\sqrt{3}\right)\right)$, where $m_j = \left\lfloor \frac{z_j + \sqrt{3}}{2\sqrt{3}} \right\rfloor$ $\forall j \in \{1, \ldots, k\}$.

Sketch of the proof. Note for the model class $\mathcal{M}_1$ that $\mathbf{J}_N(\boldsymbol{\theta}) = \frac{1}{N\tau}\mathbf{H}^\top\mathbf{H}$ [9], hence FIM does not depend on the values of the parameters $\boldsymbol{\theta}$. We denote it $\mathbf{J}_N$. The proof for a) is similar with the one from [2], with the remarkable difference that we do not use asymptotic approximations. Condition (6) is readily obtained with a chain of equivalent inequalities. Because FIM does not depend on $\boldsymbol{\theta}$, the parameter space is partitioned into congruent rectangles. Additionally, we have for all j: $P_{j|j} = P_{0|0}$ and $P_{0|j} = P_{j|0}$. The identities are instrumental in the proof of c). □

To gain more insight, we show below the relation between the ODD criterion and the widely used Generalized Likelihood Ratio Test (GLRT). Assuming the hypotheses from Theorem 2.1, GLRT as well as Rao and Wald test decide $\mathcal{M}_1$ if $T(\mathbf{x}) = \dfrac{\hat{\boldsymbol{\theta}}^\top \mathbf{H}^\top \mathbf{H} \hat{\boldsymbol{\theta}}}{\tau} > \gamma$, 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 P_{FA} [8]. As it is easy to check that $T(\mathbf{x}) = \sum_{j=1}^k z_j^2$, we have:

Proposition 2.1. *a) For $k = 1$, the ODD detector is equivalent with the GLRT for which $\gamma = 3$.*
b) For $k > 1$, it does not exist any γ such that the ODD detector is equivalent with the GLRT. Supplementarily, GLRT with $\gamma = 3$ will select $\mathcal{M}_1$ whenever ODD detector selects $\mathcal{M}_1$.

It is customary to asses the performances of a detector by evaluating P_{FA} and P_D. For the detection rule (6), we get immediately: $P_{FA} = E1$ and $P_D = 1 - \prod_{j=1}^k \left(\mathbb{Q}(-\sqrt{3} - \zeta_j) - \mathbb{Q}(\sqrt{3} - \zeta_j)\right)$, where $\zeta_j = \frac{\mathbf{v}_j^\top \underline{\boldsymbol{\theta}} \sqrt{\ell_j}}{\sqrt{\tau}} \forall j \in \{1, \ldots, k\}$ and $\boldsymbol{\theta} = \underline{\boldsymbol{\theta}}$ for the model class $\mathcal{M}_1$. Now we can notice the major difference between evaluating the performances 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 that it is calculated with the $E1$ formula when $\hat{\boldsymbol{\theta}}$ falls into $B_{\hat{d}/N}(0)$, and with $E2$ formula for all other rectangles.

For illustration, we consider in Figure 1 the LM with $k = 2$ parameters, and we draw in the (z_1, z_2) plan the squares obtained from the original rectangles within the $(\hat{\theta}_1, \hat{\theta}_2)$ plan after applying the rotation and the scaling required by the condition (6). Thus there exists a bijection from the original $B_{\hat{d}/N}(0)$ to the central square in Figure 1, where it is written the value of $E1$, the error probability of accepting $\mathcal{M}_0$ when $\hat{\boldsymbol{\theta}} \in B_{\hat{d}/N}(0)$. Note that $E1$ approaches 1 when k, the number of parameters, is large. Similar bijections exist also for the squares around the central one and for which we indicate the value of $E2$. Observe that the larger is the distance from the center of the square to the null hypothesis, the smaller is $E2$, hence the greater is the confidence in rejecting $\mathcal{M}_0$. We mention that $E2$ is smaller than 10^{-7} for all the squares that are situated faraway from the null hypothesis, and which are not drawn in Figure 1.

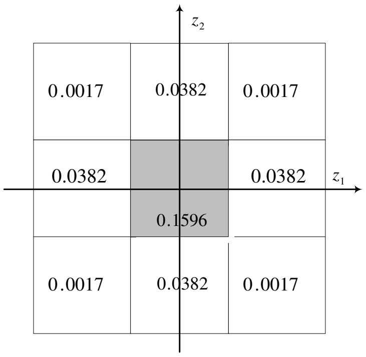

Fig. 1. LM with 2 parameters: values of $E1$ (central square) and $E2$ (all other squares). The edge length of each square is $2\sqrt{3}$.

Example: sinusoidal detection We consider as usual $\mathbf{H} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ \cos(\omega(N-1)) & \sin(\omega(N-1)) \end{bmatrix}$, where the frequency $\omega \in (0, \pi)$ is known, and we utilize the asymptotic approximation $\mathbf{H}^\top \mathbf{H} \approx (N/2)\mathbf{I}_2$. 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. Obviously $\boldsymbol{\theta} = \mathbf{0}$ for $\mathcal{M}_0$. For sake of comparison with the performances of the GLRT, we extend the results of Theorem 2.1 for an arbitrary d: the condition (6) becomes $\max(|z_1|, \ldots, |z_k|) > \sqrt{d/2}$. With the modified ODD condition, we can choose $d = 2\left(\mathbb{Q}^{-1}\left(\frac{1-\sqrt{1-P_{FA}}}{2}\right)\right)^2$ such that to satisfy the constraint on P_{FA}. Once d is found, we calculate $P_D = 1 - \prod_{j=1}^{2}\left(\mathbb{Q}(-\sqrt{d/2} - \xi_j) - \mathbb{Q}(\sqrt{d/2} - \xi_j)\right)$ for a fixed pair (A, ϕ), where $\xi_1 = \sqrt{\eta}\cos\phi$, $\xi_2 = -\sqrt{\eta}\sin\phi$ and $\eta = \frac{NA^2}{2\tau}$. Observe that η is the signal energy-to-noise ratio (ENR) [8]. For a given P_{FA}, P_D of the ODD detector depends on both η and ϕ. To clarify the dependence on ϕ, we remark for fixed η and P_{FA} that P_D is maximized when $|\phi| \in \{0, \frac{\pi}{2}, \pi\}$ and it is minimized when $|\phi| \in \{\frac{\pi}{4}, \frac{3\pi}{4}\}$.

In Figure 2, we plot the maximum and the minimum of P_D computed for the ODD criterion when d is chosen to be optimal, namely $d = 6$ as in Theorem 2.1 a). In this case, we have $P_{FA} \approx 0.16$, and this is used to get P_D of the GLRT detector for various ENR values. The evaluation of the GLRT performances rely on the results from [8], and we emphasize that P_D of GLRT is independent of ϕ. For $d = 6$, it is easy to note from Figure 2 that ϕ has a marginal influence on the P_D of ODD, and the performances of ODD and GLRT detectors are very similar. The main drawback is that P_{FA} has a value considered to be too large in most of the practical applications. To investigate the case of low P_{FA}, we find $d \approx 32.90$ when $P_{FA} = 10^{-4}$ and $d \approx 59.43$ when $P_{FA} = 10^{-7}$. For both cases we plot the performances of ODD and GLRT in Figure 2. Remark that ϕ has an important influence on the P_D of ODD, and this makes the maximum P_D of ODD to be superior to GLRT, but the minimum P_D of ODD is clearly inferior to GLRT. This outcome can be better understood if we mention additionally that the KL distance between the "artificial" and the "natural" models in the ODD settings is about 0.35 for the optimum $d = 6$, but becomes as large as 3.13 and 6.97 for the values of d that correspond to $P_{FA} = 10^{-4}$ and $P_{FA} = 10^{-7}$, respectively. It is evident that the constraints on P_{FA} are not in agreement with the ODD methodology.

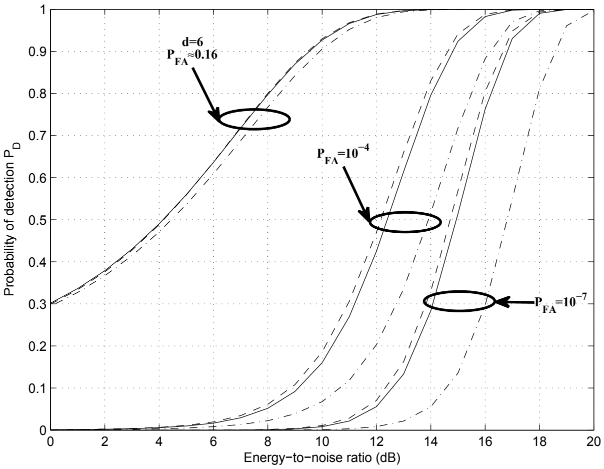

Fig. 2. Performances for sinusoidal detection: GLRT (solid line), best results of ODD (dashed line), worst results of ODD (dashdot line).

3. LINEAR MODEL:WHITE GAUSSIAN NOISE WITH *UNKNOWN* VARIANCE

Without loss of generality we make the following assumption on the noise variance: $\tau_1 < \tau < \tau_2$, where τ_1 and τ_2 are arbitrary, and they do not have any influence on the ODD decision. We use the notation $\mathbb{Q}_{t_\nu}(x)$ for the right-tail probability of a Student t-distribution with ν degrees of freedom [10]. Analogously $\mathbb{Q}_{t'_\nu(\delta)}(x)$ is the right-tail probability of a noncentral Student t-distribution with ν degrees of freedom and noncentrality parameter δ. The next result extends Theorem 2.1:

Theorem 3.1. *For the data sequence* $\mathbf{x} = [x_0, \ldots, x_{N-1}]^\top$*, consider the ODD testing between the hypotheses specified by the model classes* $\mathcal{M}_0 = \{f(\mathbf{x}; \boldsymbol{\theta}, \tau) : \boldsymbol{\theta} = \mathbf{0}, \tau_1 < \tau < \tau_2\}$ *and* $\mathcal{M}_1 = \{f(\mathbf{x}; \boldsymbol{\theta}, \tau) : \boldsymbol{\theta} \neq \mathbf{0}, \tau_1 < \tau < \tau_2\}$*. The normal density function* $f(\mathbf{x}; \boldsymbol{\theta}, \tau)$ *is given in (5), where* $\mathbf{H}$ *is a known* $N \times k$ *matrix of rank* k *and* $\boldsymbol{\theta}$ *is a* $k \times 1$ *vector of parameters* ($N > k+2$)*. Then we have:*
a) The optimum value of the d *parameter is* $\hat{d} = 3k$*.*
b) Select $\mathcal{M}_1$ *if*

$$\max(|t_1|, \ldots, |t_k|) > \sqrt{3}, \tag{7}$$

where $t_j = \frac{(\mathbf{v}_j^\top \mathbf{H}^\top \mathbf{x})/\sqrt{\ell_j}}{\sqrt{\hat{\mathfrak{v}}}}$ $\forall j \in \{1, \ldots, k\}$, $\hat{\mathfrak{v}} = \frac{\mathbf{x}^\top(\mathbf{I}_k - \mathbf{H}(\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top)\mathbf{x}}{N-k}$, $\ell_1, \ldots, \ell_k$ *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 the condition (7) is verified, we are wrong in accepting the null hypothesis with probability $E1 = 1 - \left(1 - 2\mathbb{Q}_{t_{N-k}}(\sqrt{3})\right)^k$*, otherwise, we are wrong in rejecting the null hypothesis with probability* $E2 = \prod_{j=1}^{k}\left(\mathbb{Q}_{t'_{N-k}(2m_j\sqrt{3})}\left(-\sqrt{3}\right) - \mathbb{Q}_{t'_{N-k}(2m_j\sqrt{3})}\left(\sqrt{3}\right)\right)$*, where* $m_j = \left\lfloor \frac{t_j + \sqrt{3}}{2\sqrt{3}} \right\rfloor$ $\forall j \in \{1, \ldots, k\}$*.*

Proof. We revisit briefly some well-known results [10],[11]. For the model class $\mathcal{M}_1$, the ML estimates are: $\hat{\boldsymbol{\theta}} = (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\mathbf{x}$ and $\hat{\tau} = \frac{1}{N}\|\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}}\|^2$. Similarly $\hat{\tau}_0 = \frac{1}{N}\|\mathbf{x}\|^2$ for $\mathcal{M}_0$. The function $g(\cdot;\cdot)$, involved in the factorization based on sufficient statistics (2), has the expression $g(\hat{\boldsymbol{\theta}}, \hat{\tau}; \boldsymbol{\theta}, \tau) = g_1(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta})g_2(\hat{\tau}; \tau)$, where $g_1(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}) = \frac{|\mathbf{H}^\top\mathbf{H}|^{1/2}}{(2\pi\tau)^{k/2}}\exp\left(-\frac{1}{2\tau}\|\mathbf{H}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\|^2\right)$ and $g_2(\hat{\tau}; \tau) =$

$\frac{(N/2)^{(N-k)/2}}{\Gamma((N-k)/2)}\left(\frac{\hat{\tau}}{\tau}\right)^{(N-k)/2}\frac{1}{\hat{\tau}}\exp\left(-\frac{N}{2}\frac{\hat{\tau}}{\tau}\right)$. $\Gamma(\cdot)$ denotes the usual *Gamma* function.

In the proof of Theorem 2.1, it was straightforward to consider as "natural" model the unique density function that belongs to $\mathcal{M}_0$. Here we take as "natural" model for $\mathcal{M}_0$ the universal model given by the NML function (1): $\tilde{f}_0(\mathbf{x}) = f(\mathbf{x};\boldsymbol{\theta}^0,\hat{\tau}_0)/C_0$, where $\boldsymbol{\theta}^0 = \mathbf{0}$. The interested reader can find in [12] details on the computation of $C_0 = \frac{(N/2)^{N/2}\exp(-N/2)}{\Gamma(N/2)}\ln\frac{\tau_2}{\tau_1}$.

Like in the proof of Theorem 2.1, we 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 the parameter whose optimal value we will find next. Let $B_{d/N}(0)$ be the largest rectangle within this hyper-ellipsoid. Its volume is $|B_{d/N}(0)| = 2^k\prod_{j=1}^k \mu_j$ [2], where $\mu_j = \left(\frac{d}{Nk\lambda_j}\right)^{1/2}$ and $\lambda_j = \ell_j/(N\tau)$ is the j-th eigenvalue of the matrix $\mathbf{J}_N$. Because FIM for the class $\mathcal{M}_1$ is $\begin{bmatrix}\mathbf{J}_N & \mathbf{0}\\ \mathbf{0} & 1/(2\tau^2)\end{bmatrix}$, and the constraints for τ are the same for both $\mathcal{M}_0$ and $\mathcal{M}_1$, we choose $B^\tau_{d/N}(0) = \big\{(\boldsymbol{\theta},\tau) : \boldsymbol{\theta}\in B_{d/N}(0), \tau\in(\tau_1,\tau_2)\big\}$ to be the rectangle associated to the model class $\mathcal{M}_0$. Therefore the density function $\hat{f}(\mathbf{x}\mid\boldsymbol{\theta}^0)$ defined in (3) is zero outside $B^\tau_{d/N}(0)$, and inside $B^\tau_{d/N}(0)$ equals $\hat{f}(\mathbf{x}\mid\boldsymbol{\theta}^0) = f(\mathbf{x};\hat{\boldsymbol{\theta}},\hat{\tau})/Q^\tau_{d/N}(0)$. The expression of the normalization factor (4) is obtained after some calculations: $Q^\tau_{d/N}(0) = d^{k/2}\left(\frac{2}{k\pi}\right)^{k/2}\frac{(N/2)^{(N-k)/2}\exp(-N/2)}{\Gamma((N-k)/2)}\ln\frac{\tau_2}{\tau_1}$.

We introduce two more notations: $\mathcal{X}^\tau_0 = \{\mathbf{x} : (\hat{\boldsymbol{\theta}}(\mathbf{x}),\hat{\tau}(\mathbf{x}))\in B^\tau_{d/N}(0)\}$ and $\mathfrak{M}_{d/N}(0) = [0,\mu_1]\times\cdots\times[0,\mu_k]$. With these preparations, we are ready to compute the KL distance,

$$D(\hat{f}(\mathbf{x}\mid\boldsymbol{\theta}^0)\parallel\tilde{f}_0(\mathbf{x}))$$
$$= \ln\frac{C_0}{Q^\tau_{d/N}(0)} + \frac{1}{Q^\tau_{d/N}(0)}\int_{\mathcal{X}^\tau_0} f(\mathbf{x};\hat{\boldsymbol{\theta}},\hat{\tau})\ln\frac{f(\mathbf{x};\hat{\boldsymbol{\theta}},\hat{\tau})}{f(\mathbf{x};\boldsymbol{\theta}^0,\hat{\tau}_0)}\mathrm{d}\mathbf{x}.$$

The change of variables $\hat{\boldsymbol{\rho}} = [\mathbf{v}_1\ldots\mathbf{v}_k]^\top\hat{\boldsymbol{\theta}}$ leads to:

$$\int_{\mathcal{X}^\tau_0} f(\mathbf{x};\hat{\boldsymbol{\theta}},\hat{\tau})\ln\frac{f(\mathbf{x};\hat{\boldsymbol{\theta}},\hat{\tau})}{f(\mathbf{x};\boldsymbol{\theta}^0,\hat{\tau}_0)}\mathrm{d}\mathbf{x} =$$
$$\frac{Nh_{\mathbf{H},N,k}}{2}\int\limits_{\tau_1}^{\tau_2}\frac{2^k}{\hat{\tau}^{k/2+1}}\left(\int\limits_{\mathfrak{M}_{d/N}(0)}\ln\left(1+\frac{\sum_{j=1}^k\hat{\rho}_j^2\ell_j}{N\hat{\tau}}\right)\mathrm{d}\hat{\boldsymbol{\rho}}\right)\mathrm{d}\hat{\tau},$$

where $h_{\mathbf{H},N,k} = \frac{|\mathbf{H}^\top\mathbf{H}|^{1/2}(N/2)^{(N-k)/2}\exp(-N/2)}{(2\pi)^{k/2}\Gamma((N-k)/2)}$. As $d \ll N$, we employ next the approximation $\ln(1+\varepsilon_N) = \varepsilon_N + O(\varepsilon_N^2)$, where $\varepsilon_N = \sum_{j=1}^k\frac{\hat{\rho}_j^2\ell_j}{N\hat{\tau}} < \sum_{j=1}^k\frac{\mu_j^2\ell_j}{N\hat{\tau}} = \frac{d}{N}$. Then we have $\int\limits_{\mathcal{X}^\tau_0} f(\mathbf{x};\hat{\boldsymbol{\theta}},\hat{\tau})\ln\frac{f(\mathbf{x};\hat{\boldsymbol{\theta}},\hat{\tau})}{f(\mathbf{x};\boldsymbol{\theta}^0,\hat{\tau}_0)}\mathrm{d}\mathbf{x} \approx \frac{d}{6}Q^\tau_{d/N}(0)$, and combining with the previous results we get $D(\hat{f}(\mathbf{x}\mid\boldsymbol{\theta}^0)\parallel\tilde{f}_0(\mathbf{x})) = \frac{k}{2}\ln\frac{k\pi}{2d}+\frac{d}{6}+\ln\left(\frac{(N/2)^{k/2}\Gamma((N-k)/2)}{\Gamma(N/2)}\right)$. It is clear that $D(\hat{f}(\mathbf{x}\mid\boldsymbol{\theta}^0)\parallel\tilde{f}_0(\mathbf{x}))$ is minimized for $\hat{d} = 3k$. Moreover, if we apply the Stirling approximation, the last term becomes
$\ln\left(\left(1+\frac{1}{a_{N,k}}\right)^{1/2}\left(\left(1+\frac{1}{a_{N,k}}\right)^{a_{N,k}}\right)^{-k/2}\exp\left(\frac{k}{2}\right)\right)$, where $a_{N,k} = (N-k)/k$. With this approximative formula, it is obvious that the last term is zero when $N\to\infty$ and k is fixed, hence the minimum of the KL distance between the "artificial" model and the "natural" model is asymptotically the same as in the case of *known* noise variance.

To prove the rest of the results, we apply similar techniques as in the proof of Theorem 2.1, with the main difference that τ is replaced with the unbiased estimate $\hat{\mathfrak{v}}$ [8]. □

Without difficulties, we can show for the detection rule (7) that $P_{FA} = E1$ and $P_D = 1-\prod_{j=1}^k\left(\mathbb{Q}_{t'_{N-k}(\zeta_j)}(-\sqrt{3}) - \mathbb{Q}_{t'_{N-k}(\zeta_j)}(\sqrt{3})\right)$, where $\zeta_j = \frac{\mathbf{v}_j^\top\underline{\boldsymbol{\theta}}\sqrt{\ell_j}}{\sqrt{\underline{\tau}}}\forall j\in\{1,\ldots,k\}$, and for the model class $\mathcal{M}_1$ we have: $\boldsymbol{\theta} = \underline{\boldsymbol{\theta}}$, $\tau = \underline{\tau}$.

We recall that the GLRT for testing if the j-th component of the parameter vector $\boldsymbol{\theta}$ is zero relies on the statistic $T(\mathbf{x}) = \frac{\hat{\theta}_j^2}{\hat{\mathfrak{v}}h^-_{jj}}$, where h^-_{jj} is the j-th diagonal entry of the matrix $(\mathbf{H}^\top\mathbf{H})^{-1}$ [8]. It is notorious that $T(\mathbf{x})$ is the square of the usual t-statistic [10]. When the matrix $\mathbf{H}^\top\mathbf{H}$ is diagonal, t_j becomes also identical with the usual t-statistic, and the ODD condition (7) reduces to compare with the threshold $\sqrt{3}$ the t-statistic computed for each component θ_j, and to select $\mathcal{M}_1$ if at least one component is found to be nonzero.

4. FINAL REMARKS

One of the constraints in utilizing the ODD criterion is the following: FIM must be non-singular for the parameters that correspond to the null hypothesis. The condition is not verified in sinusoidal detection if the value of the frequency is not known a priori. This difficulty was already noticed in connection with the Rao detector [8]. A solution for such cases is the detection method based on the NML of the competing models [12].

In the present study, we investigated the use of the ODD detector for the LM model by emphasizing the strengths and the weaknesses of the method. The confidence indices provided by ODD without assuming knowledge on the true parameter values are an advantage. For the GLRT, the complement set of the critical region is a solid hyper-ellipsoid. 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 it was apparent from the comparisons with the GLRT.

5. REFERENCES

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

[2] J. Rissanen, *Information and complexity in statistical modeling*, Springer Verlag, 2007.

[3] M. Li and P.M.B. Vitanyi, *An introduction to Kolmogorov complexity and its applications*, 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. Proces. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.

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

[7] J. Rissanen, "Optimally distinguishable distributions," Sep. 2007, (8 pages), personal communication.

[8] S.M. Kay, *Fundamentals of statistical signal processing: detection theory*, Prentice Hall, 1998.

[9] S.M. Kay, *Fundamentals of statistical signal processing: estimation theory*, Prentice Hall, 1993.

[10] G.A.F. Seber and A.J. Lee, *Linear regression analysis*, Wiley-Interscience, 2003.

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

[12] J. Rissanen, "Hypothesis selection and testing by the MDL principle," *Computer Journal*, vol. 42, no. 4, pp. 260–269, 1999.