Received March 26, 2013, accepted May 24, 2013, date of publication August 15, 2013, date of current version August 26, 2013.

Digital Object Identifier 10.1109/ACCESS.2013.2277930

An Informative Interpretation of Decision Theory: The Information Theoretic Basis for Signal-to-Noise Ratio and Log Likelihood Ratio

JOHN POLCARI
Center for Engineering Science Advanced Research, Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA
(polcarij@ornl.gov)

Support for this work was provided by the Office of Naval Research Maritime Sensing Program, Code 321MS.

ABSTRACT The signal processing concept of signal-to-noise ratio (SNR), in its role as a performance measure, is recast within the more general context of information theory, leading to a series of useful insights. Establishing generalized SNR (GSNR) as a rigorous information theoretic measure inherent in any set of observations significantly strengthens its quantitative performance pedigree while simultaneously providing a specific definition under general conditions. In turn, this directly leads to consideration of the log likelihood ratio (LLR): first, as the simplest possible information-preserving transformation (i.e., signal processing algorithm) and subsequently, as an absolute, comparable measure of information for any specific observation exemplar. The information accounting methodology that results permits practical use of both GSNR and LLR as diagnostic scalar performance measurements, directly comparable across alternative system/algorithm designs, applicable at any tap point within any processing string, in a form that is also comparable with the inherent performance bounds due to information conservation.

INDEX TERMS Data compression, decision theory, detection algorithms, information measures, information theory, Kullback–Leibler divergence, log likelihood ratio, performance evaluation, performance measures, self-scaling property, signal processing algorithms, signal to noise ratio, statistical analysis.

I. INTRODUCTION

Optimal detection theory is traditionally developed from the underlying perspective of either Bayesian cost minimization or Neyman-Pearson hypothesis testing. By contrast, one purpose of this paper (and anticipated future papers) is to demonstrate that optimal detection theory may be at least equally well founded on the fundamental principles of information theory, upon recognition of generalized signal-to-noise ratio (GSNR) as a measure of relative entropy, as first proposed for discrete random variables by Shannon and later extended to continuous random variables by Kullback and Leibler.

Fundamentally, the current work contrasts with traditional signal processing and detection literature in its underlying roadmap of concept and consequence. While precise generalizations of alternative conceptual structures are difficult, the author takes a definitive text such as [1] as representative of the consolidated line of thought in this area. Figure 1 provides a conceptual block diagram for a generic binary decision process. The conceptual focus of traditional theory is the careful formulation of the right-hand block. For this particular problem, the log likelihood ratio (LLR) and the associated threshold are derived as the two sides of the quantitative thresholding process which optimizes either of two useful (but heuristic) objective functions, namely average cost (or risk) as defined in a Bayesian sense, and probability of detection P_D for any specific constrained probability of false alarm P_F (the Neyman-Pearson criteria). Since Bayesian cost factors are often difficult to specify on a practical basis (and thus typically treated as a thought experiment for the purpose of justifying the optimal decision process), performance is then measured by assembly of (P_D, P_F) pairs into two dimensional receiver operating characteristic (ROC) curves.

As many of the underlying data models involve various forms of signal in additive noise, the concept of signal-to-noise ratio (SNR), specifying the relative size of the two contributions, is subsequently introduced as an ad-hoc measure related to performance in an intuitively obvious but quantitatively ambiguous way. For the specific case of known signal in Gaussian noise, the mathematics is simple enough to actually derive the relationship and subsequently prove that, for the

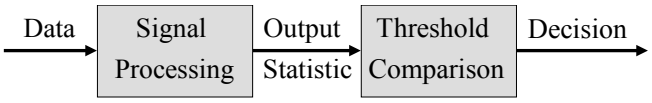


FIGURE 1. Block diagram of generic binary decision process.

specific case, it is the only parameter required to uniquely evaluate the associated ROC curve.

For all intents and purposes, this traditional conceptual roadmap does not consider potential existence of the left-hand signal processing block. While those carefully schooled in such matters might (correctly) argue that the traditional line of thought implies that the optimal solution is to use the data as the output statistic (completely eliminating existence of the left-hand block), this implication is generally left unstated, let alone being explored and emphasized. Further, no special pedigree is imputed to the LLR; indeed, the full LLR is often considered too tedious to compute, leading to partial calculation of only the data dependent terms.

The traditional line of thought was established long before the formulation of information theoretic concepts, the initiation of which is commonly attributed to Shannon [2], and subsequent extension to continuous random variables to Kullback and Leibler [3]. As shown here and in subsequent papers, consideration of information-based performance concepts allows one to explicitly address both the left and right hand blocks in a unified manner. This modified line of thought starts with the recognition of SNR (in a generalized form) as a measure of the mean information available, and as a consequence then introduces the LLR as the scalar statistic literally quantifying that information for any particular exemplar of the data.

Multiple considerations motivate the author's interest in documenting the information theoretic basis for optimal detection theory. First, it provides an alternate way of thinking about signal processing at a very fundamental level; the author is unaware of any existing orderly exposition of either the train of logic or the theoretical and practical implications that result. Second, it imparts impeccable information measurement pedigrees to the traditional signal processing constructs of SNR and LLR, significantly strengthening both their meaningfulness and their quantitative utility. Third, it vastly simplifies (both conceptually and practically) the relationship between signal processing application and performance, yielding rigorous *scalar* performance measures to replace the (P_D, P_F) pair that comprises the centerpiece of traditional detector performance assessment.

The last item deserves further elucidation, particularly from a practical perspective. The careful empirical measurement required to obtain a high-quality ROC curve often entails a task of such magnitude that it can easily dwarf system development in both cost and schedule. Moreover, the real world presents such a range of uncontrollable variables that most field tests can provide only a gross average result, with large, inherent uncertainty when applied to any one particular operating condition. In the sonar arena (the author's experience base), this all too often leads to discounting the utility of quantitative performance prediction; comparative performance assessments involving side-by-side ''beauty contests;'' and a distinct inability to precisely ''prove'' (and thus *improve*) system performance. The author suspects that similar (although perhaps less extreme) situations exist in other applications. By contrast, the approach outlined here supports quantitative comparisons on a case-by-case basis. This massively simplifies any required test regimen in a manner that is tolerant of real-world variability.

This paper is intended to be the first in a series, specifically focusing on the interpretation and practical application of quantitative information measurement to the performance characterization of arbitrary signal processing algorithms (i.e., the left-hand block of Figure 1). Developments are couched specifically in terms of joint density functions of random vectors (the practical reality of signal processing in the digital age), with all associated multi-dimensional functions assumed to possess well-defined Jacobian matrices so the resulting random variable transformations may be handled using traditional density stretching techniques. As such, some more esoteric mathematical concerns are handled less rigorously than might otherwise be desired; in particular, probability space concepts are avoided, solely in favor of the practical consideration of extending the audience for this paper.

In the preparation of this paper, a concerted (but certainly not exhaustive) literature search was conducted. Surprisingly little existing work could be found along these lines, and what material was found is in the information theory literature rather than the signal processing literature. In particular, the closest recent work appears to the cross-entropy analyses of Shore and Johnson, circa 1980, that form the basis for the (indirectly) related subject of least informative priors ([4], [5], [6]); and a series of more recent papers, beginning with Guo, et.al. [7], relating mutual information with minimum mean-square estimation error. The latter primarily address Gaussian background statistics, with the exception of a very recent extension to Poisson channels [8], and make no claim to consider the general case. That said, particularly given the extensive base of potentially relevant literature across multiple fields, the author is extremely hesitant to label any particular result presented here as new.

As an outgrowth of the review process, the author has been asked to comment on why the results found here should be considered novel and not obvious. Indeed, several of the theoretical results in Sections 2 and 3 could quite reasonably described as ''intuitively obvious''. The unique aspect of the theory as developed here is rather the demonstration of its rigorous mathematical underpinnings, elevating these concepts beyond conjecture and into the realm of the ''laws of information''. The self-scaling property of LLRs is an ideal example of this, a property so obvious that its critical role as proof of any ''true'' information measure has been previously overlooked. By contrast, the true novelty of the current work resides in the future application of these newly minted laws to

practical real-world signal processing design and assessment, which even a cursory consideration of Sections 4, 5, and 6 shows is not currently the case.

II. A RIGOROUS GENERALIZED DEFINITION OF SNR

The term ''signal-to-noise ratio'' (or SNR) is loosely defined as a ratio of a size or strength measure (typically energy or power) of the signal (or ''interesting'') component to an equivalent measure of the noise (or ''uninteresting'') component against which it is competing in some set of observations [9]. While a detailed history of the term is beyond the scope of this paper, over time it has clearly acquired a second distinct (but related) connotation as a scalar measure of signal processing performance. The source of the duality of nomenclature may be seen by considering the specific case of detecting a known signal $\underline{s}$ in additive, jointly Gaussian random noise that is zero mean with covariance $\underline{\overline{R}}_n$. For this case SNR, specifically defined for real data as [10]

$$SNR = \underline{s}^{+}\underline{\overline{R}}_n^{-1}\underline{s} \tag{1}$$

is demonstrably a rigorous performance measure, since, for threshold η, the binary detector ROC curve depends only upon SNR, i.e., [11]

$$P_D = \frac{1}{2}\mathrm{erfc}\left(\frac{\ln\eta}{\sqrt{2\,SNR}} - \frac{1}{2}\sqrt{\frac{SNR}{2}}\right) \quad \text{and}$$
$$P_F = \frac{1}{2}\mathrm{erfc}\left(\frac{\ln\eta}{\sqrt{2\,SNR}} + \frac{1}{2}\sqrt{\frac{SNR}{2}}\right). \tag{2}$$

For complex data, an additional factor of two must be included in (1).

Unfortunately, for more general cases, this duality breaks down. While the equivalent non-dimensional ratio can be (and typically is) defined in multiple ways, demonstrability as a measure of performance is never addressed, leaving a performance accounting regimen that, short of the laborious (and often ambiguous) process of explicit ROC curve evaluation, is less than fully rigorous.

By comparison, information theory provides a direct, rigorous measure of the average information on the detection decision that is carried by the observations. This measure is the relative entropy, (also called the Kullback-Leibler (KL) divergence or distance), defined in symmetric form as [12]

$$\begin{aligned} J\left(\underline{\delta}\right) &= I_{1:0}\left(\underline{\delta}\right) + I_{0:1}\left(\underline{\delta}\right) \\ &= \int d\underline{\delta}\left[\ln\left(\frac{p_1\left(\underline{\delta}\right)}{p_0\left(\underline{\delta}\right)}\right)p_1\left(\underline{\delta}\right) + \ln\left(\frac{p_0\left(\underline{\delta}\right)}{p_1\left(\underline{\delta}\right)}\right)p_0\left(\underline{\delta}\right)\right] \\ &= \int d\underline{\delta}\ln\left(\frac{p_1\left(\underline{\delta}\right)}{p_0\left(\underline{\delta}\right)}\right)\left[p_1\left(\underline{\delta}\right) - p_0\left(\underline{\delta}\right)\right] \end{aligned} \tag{3}$$

where the lack of stated bounds implies integration over the fully allowable, multi-dimensional range of the observations $\underline{\delta}$. It is well recognized [13], [14] that the KL divergence, in either the symmetric form $J\left(\underline{\delta}\right)$ or the two asymmetric components $I_{1:0}\left(\underline{\delta}\right)$ and $I_{0:1}\left(\underline{\delta}\right)$, is the direct generalization of Shannon's fundamental measure of information (relative entropy) for the case of continuous random variables. As such, it precisely quantifies the mean available information, once the two data densities $p_{1/0}\left(\underline{\delta}\right)$ (conditional *only* upon the pending decision) have been specified. For the previous case of known signal in additive Gaussian noise, the data model for target present is

$$\begin{aligned} H_1 : \underline{\delta} &= \underline{s} + \underline{n} \\ \Rightarrow\quad p_1\left(\underline{\delta}\right) &= \frac{1}{(2\pi)^{N/2}\sqrt{\det\left(\underline{\overline{R}}_n\right)}}e^{-\frac{1}{2}\left(\underline{\delta}-\underline{s}\right)^{+}\underline{\overline{R}}_n^{-1}\left(\underline{\delta}-\underline{s}\right)} \end{aligned} \tag{4}$$

(where N is the number of elements in $\underline{\delta}$), while that for target absent is

$$H_0 : \underline{\delta} = \underline{n} \;\Rightarrow p_0\left(\underline{\delta}\right) = \frac{1}{(2\pi)^{N/2}\sqrt{\det\left(\underline{\overline{R}}_n\right)}}e^{-\frac{1}{2}\underline{\delta}^{+}\underline{\overline{R}}_n^{-1}\underline{\delta}}. \tag{5}$$

Evaluation of the intermediate algebra then leads to the conclusion that

$$J\left(\underline{\delta}\right) = \underline{s}^{+}\underline{\overline{R}}_n^{-1}\underline{s} = SNR \tag{6}$$

strongly suggesting the author's fundamental proposition; namely, that the symmetric KL divergence is the appropriate generalization of SNR when used in the role of a measure of performance.

This proposed definition of generalized SNR (labeled here as GSNR, to denote its intended role as a performance measure) is mathematically self-consistent, as Gibbs' Inequality guarantees that $I_{1:0}$, $I_{0:1}$, and J all be non-negative. Several important advantages derive, including

- A precise, quantitative definition of SNR under general conditions.
- Rigorous validation of the role of SNR as a *scalar* performance measure of Shannon information (versus the rather ad hoc pedigree it currently possesses).
- Related general (but mathematically demonstrable) insights within the arena of optimal detection theory.

In the next section, a few of the most critical insights are explored for the purpose of establishing the information-theoretic basis of the log likelihood ratio. The remainder of this paper then addresses the practical implications of the second item, with further conceptual expansions left as the subject for future articles.

III. THE LOG LIKELIHOOD RATIO (LLR) FROM THE PERSPECTIVE OF INFORMATION THEORY

A critical insight gleaned from the proposed definition in Section 2 is that, rather than resulting from the signal processing applied to the observations, GSNR is a property inherent in the observations themselves. Indeed, the following theorem (proven in Kullback and Leibler's original paper, but recast here in the parlance of signal processing and traditional probability densities [15]), being the statistical equivalent of the Second Law of Thermodynamics, represents a death knell to notions of ''SNR improvement.''

Theorem 1 (Information Limit Theorem): Let the mean information available from an observation vector $\underline{\boldsymbol{\delta}}$, with associated densities $p_{1/0}\left(\underline{\boldsymbol{\delta}}\right)$, be denoted by

$$GSNR\left(\underline{\boldsymbol{\delta}}\right) = J\left(\underline{\boldsymbol{\delta}}\right). \tag{7}$$

No well-defined function $\underline{\boldsymbol{F}}$ transforming $\underline{\boldsymbol{\delta}}$ to a vector $\underline{\boldsymbol{\varepsilon}} = \underline{\boldsymbol{F}}\left(\underline{\boldsymbol{\delta}}\right)$ can provide additional information beyond that originally available; that is, under all circumstances

$$GSNR\left(\underline{\boldsymbol{\varepsilon}}\right) \leq GSNR\left(\underline{\boldsymbol{\delta}}\right). \tag{8}$$

Proof: Theorem 4.1 [16] guarantees that $I_{1:0}\left(\underline{\boldsymbol{\varepsilon}}\right) \leq I_{1:0}\left(\underline{\boldsymbol{\delta}}\right)$ and $I_{0:1}\left(\underline{\boldsymbol{\varepsilon}}\right) \leq I_{0:1}\left(\underline{\boldsymbol{\delta}}\right)$, which, in turn, implies the inequality in (8). While the underlying functional transformation is strictly required only to be measurable [17], the words ''well-defined function'' are used herein to mildly further restrict consideration to those random variable transformations for which the joint probability densities $p_{1/0}\left(\underline{\boldsymbol{\varepsilon}}\right)$ may be computed using traditional density stretching techniques. Because of difficulty in identifying specific documentation elsewhere, Appendix A provides the general form of density stretching for random vectors, i.e., evaluating the density function of an $M\,x\,1$ multivariate random variable $\underline{\boldsymbol{y}}$ arising from the general, non-linear transformation of an $N\,x\,1$ multivariate random variable $\underline{\boldsymbol{x}}$ $(M \leq N)$.

This result immediately forces the question of ''GSNR preservation'' to the forefront, a natural outgrowth of which is the introduction of the same LLR statistic that is normally developed in the context of optimal detection theory.

Corollary 1 (Information Preservation Corollary): A well-defined function $\underline{\boldsymbol{F}}$ preserves the original mean information, i.e., $GSNR\left(\underline{\boldsymbol{\varepsilon}}\right) = GSNR\left(\underline{\boldsymbol{\delta}}\right)$, if and only if

$$\ln\left(\frac{p_{1\,\underline{\boldsymbol{\varepsilon}}}\left(\underline{\boldsymbol{\varepsilon}}\right)}{p_{0\,\underline{\boldsymbol{\varepsilon}}}\left(\underline{\boldsymbol{\varepsilon}}\right)}\right) \equiv \ln\left(\frac{p_{1\,\underline{\boldsymbol{\delta}}}\left(\underline{\boldsymbol{\delta}}\right)}{p_{0\,\underline{\boldsymbol{\delta}}}\left(\underline{\boldsymbol{\delta}}\right)}\right). \tag{9}$$

Proof: The second part of the proof of Theorem 4.1 [18] provides the necessary and sufficient requirements for preservation of GSNR.

Taking this result all the way to a simple scalar form leads directly to the LLR.

Definition 1 (LLR Definition): For an observation vector $\underline{\boldsymbol{\delta}}$, with densities $p_{1/0}\left(\underline{\boldsymbol{\delta}}\right)$, the LLR associated with $\underline{\boldsymbol{\delta}}$ is

$$\lambda = L\left(\underline{\boldsymbol{\delta}}\right) = \ln\left(\frac{p_1\left(\underline{\boldsymbol{\delta}}\right)}{p_0\left(\underline{\boldsymbol{\delta}}\right)}\right). \tag{10}$$

Throughout this paper, use of the form $L\left(\underline{\boldsymbol{\delta}}\right)$ is restricted to cases where it is useful to make the functional dependence on $\underline{\boldsymbol{\delta}}$ explicit.

Lemma 1: The associated LLR preserves the mean information available in the original observation vector $\underline{\boldsymbol{\delta}}$.

Proof: It suffices to demonstrate that (10) also guarantees that

$$\ln\left(\frac{p_{1\,\lambda}\left(\lambda\right)}{p_{0\,\lambda}\left(\lambda\right)}\right) \equiv \ln\left(\frac{p_{1\,\underline{\boldsymbol{\delta}}}\left(\underline{\boldsymbol{\delta}}\right)}{p_{0\,\underline{\boldsymbol{\delta}}}\left(\underline{\boldsymbol{\delta}}\right)}\right) \tag{11}$$

so that information preservation is guaranteed. Considering the scalar transformation of random variables

$$L\left(\underline{\boldsymbol{\delta}}\right) = \ln\left(\frac{p_{1\,\underline{\boldsymbol{\delta}}}\left(\underline{\boldsymbol{\delta}}\right)}{p_{0\,\underline{\boldsymbol{\delta}}}\left(\underline{\boldsymbol{\delta}}\right)}\right). \tag{12}$$

the scalar form from Appendix A yields

$$p_{1/0\,\lambda}\left(\lambda\right) = \int\limits_{\{\underline{\boldsymbol{\delta}}|\lambda=F\left(\underline{\boldsymbol{\delta}}\right)\}} d\underline{\boldsymbol{\delta}}\,\frac{p_{1/0\,\underline{\boldsymbol{\delta}}}\left(\underline{\boldsymbol{\delta}}\right)}{\left|\overrightarrow{\nabla}_{\underline{\boldsymbol{\delta}}}L\left(\underline{\boldsymbol{\delta}}\right)\right|}. \tag{13}$$

Noting that (10) implies that $p_{1\,\underline{\boldsymbol{\delta}}}\left(\underline{\boldsymbol{\delta}}\right) = e^{\lambda}p_{0\,\underline{\boldsymbol{\delta}}}\left(\underline{\boldsymbol{\delta}}\right)$

$$\begin{aligned}\ln\left(\frac{p_{1\,\lambda}\left(\lambda\right)}{p_{0\,\lambda}\left(\lambda\right)}\right) &= \ln\left(\frac{\int\limits_{\{\underline{\boldsymbol{\delta}}|\lambda=F\left(\underline{\boldsymbol{\delta}}\right)\}} d\underline{\boldsymbol{\delta}}\,\frac{p_{1\,\underline{\boldsymbol{\delta}}}\left(\underline{\boldsymbol{\delta}}\right)}{\left|\overrightarrow{\nabla}_{\underline{\boldsymbol{\delta}}}L\left(\underline{\boldsymbol{\delta}}\right)\right|}}{\int\limits_{\{\underline{\boldsymbol{\delta}}|\lambda=F\left(\underline{\boldsymbol{\delta}}\right)\}} d\underline{\boldsymbol{\delta}}\,\frac{p_{0\,\underline{\boldsymbol{\delta}}}\left(\underline{\boldsymbol{\delta}}\right)}{\left|\overrightarrow{\nabla}_{\underline{\boldsymbol{\delta}}}L\left(\underline{\boldsymbol{\delta}}\right)\right|}}\right) \\ &= \ln\left(\frac{\int\limits_{\{\underline{\boldsymbol{\delta}}|\lambda=F\left(\underline{\boldsymbol{\delta}}\right)\}} d\underline{\boldsymbol{\delta}}\,e^{\lambda}\,\frac{p_{0\,\underline{\boldsymbol{\delta}}}\left(\underline{\boldsymbol{\delta}}\right)}{\left|\overrightarrow{\nabla}_{\underline{\boldsymbol{\delta}}}L\left(\underline{\boldsymbol{\delta}}\right)\right|}}{\int\limits_{\{\underline{\boldsymbol{\delta}}|\lambda=F\left(\underline{\boldsymbol{\delta}}\right)\}} d\underline{\boldsymbol{\delta}}\,\frac{p_{0\,\underline{\boldsymbol{\delta}}}\left(\underline{\boldsymbol{\delta}}\right)}{\left|\overrightarrow{\nabla}_{\underline{\boldsymbol{\delta}}}L\left(\underline{\boldsymbol{\delta}}\right)\right|}}\right) \\ &= \lambda = \ln\left(\frac{p_{1\,\underline{\boldsymbol{\delta}}}\left(\underline{\boldsymbol{\delta}}\right)}{p_{0\,\underline{\boldsymbol{\delta}}}\left(\underline{\boldsymbol{\delta}}\right)}\right)\end{aligned} \tag{14}$$

where e^{λ} is constant over the region of integration. Since this result is true for all possible values of λ, it confirms (11).

Implied in the above result is a powerful but often overlooked characteristic, the self-scaling property, differentiating LLRs from all other statistics, and ultimately casting the LLR as the instantaneous analog of GSNR. That is, while GSNR measures differential information in an average sense over the statistical ensemble of observations, the LLR provides an absolute measure of the information contained in any specific exemplar.

Theorem 2 (Self-Scaling Property): A scalar statistic λ, with densities $p_{1/0}\left(\lambda\right)$, is an LLR if and only if

$$\lambda \equiv \ln\left(\frac{p_1\left(\lambda\right)}{p_0\left(\lambda\right)}\right). \tag{15}$$

Proof: If λ is the LLR of some other random variable $\underline{\boldsymbol{\delta}}$, then (14) must hold. Conversely, if (15) holds, then λ is, at a minimum, the LLR of itself.

Corollary 2 (Mean LLR Property): For any LLR statistic λ,

$$\langle\lambda\rangle_1 \geq 0, \quad \langle\lambda\rangle_0 \leq 0, \quad \text{and} \quad GSNR\left(\lambda\right) = \langle\lambda\rangle_1 - \langle\lambda\rangle_0 \geq 0. \tag{16}$$

Proof: The asymmetric forms of the KL divergence are [19]

$$I_{1:0}\left(\underline{\boldsymbol{\delta}}\right) = \int d\underline{\boldsymbol{\delta}}\ln\left(\frac{p_{1\,\underline{\boldsymbol{\delta}}}\left(\underline{\boldsymbol{\delta}}\right)}{p_{0\,\underline{\boldsymbol{\delta}}}\left(\underline{\boldsymbol{\delta}}\right)}\right)p_{1\,\underline{\boldsymbol{\delta}}}\left(\underline{\boldsymbol{\delta}}\right) = \int d\lambda\,\lambda\,p_{1\,\lambda}\left(\lambda\right) = \langle\lambda\rangle_1 \tag{17}$$

and

$$I_{0:1}\left(\underline{\delta}\right) = \int d\underline{\delta} \ln\left(\frac{p_{0\underline{\delta}}\left(\underline{\delta}\right)}{p_{1\underline{\delta}}\left(\underline{\delta}\right)}\right) p_{0\underline{\delta}}\left(\underline{\delta}\right) = -\int d\lambda\, \lambda\, p_{0\lambda}\left(\lambda\right) = -\left\langle\lambda\right\rangle_0 \tag{18}$$

with GSNR being the sum. Gibb's Inequality then applies to both components.

A specific benefit of the LLR is that, although it is an instantaneous measure, it remains quantitatively comparable.

Theorem 3 (LLR Ordering Theorem): Let $\underline{\boldsymbol{F}}$ be any well-defined function transforming observation vector $\underline{\delta}$ to vector $\underline{\varepsilon} = \underline{\boldsymbol{F}}\left(\underline{\delta}\right)$, with respective LLRs

$$\lambda_{\underline{\delta}} = L_{\underline{\delta}}\left(\underline{\delta}\right) = \ln\left(\frac{p_{1\underline{\delta}}\left(\underline{\delta}\right)}{p_{0\underline{\delta}}\left(\underline{\delta}\right)}\right) \quad \text{and}$$

$$\lambda_{\underline{\varepsilon}} = L_{\underline{\varepsilon}}\left(\underline{\varepsilon}\right) = \ln\left(\frac{p_{1\underline{\varepsilon}}\left(\underline{\varepsilon}\right)}{p_{0\underline{\varepsilon}}\left(\underline{\varepsilon}\right)}\right). \tag{19}$$

Then, for any choice of $\underline{\varepsilon}$ in the range of $\underline{\boldsymbol{F}}$

$$\min_{\{\underline{\delta}|\lambda_{\underline{\varepsilon}}=L_{\underline{\varepsilon}}(\underline{\boldsymbol{F}}(\underline{\delta}))\}} \lambda_{\underline{\delta}} \le \left\langle\lambda_{\underline{\delta}}\right\rangle_{q_0} \le \lambda_{\underline{\varepsilon}} \le \left\langle\lambda_{\underline{\delta}}\right\rangle_{q_1} \le \max_{\{\underline{\delta}|\lambda_{\underline{\varepsilon}}=L_{\underline{\varepsilon}}(\underline{\boldsymbol{F}}(\underline{\delta}))\}} \lambda_{\underline{\delta}} \tag{20}$$

where the densities $q_{1/0}\left(\underline{\delta}\,|\lambda_{\underline{\varepsilon}}\right)$ are defined only over the limited support space $\left\{\underline{\delta}\,|\lambda_{\underline{\varepsilon}} = L_{\underline{\varepsilon}}\left(\underline{\boldsymbol{F}}\left(\underline{\delta}\right)\right)\right\}$ (which is the pre-image region of $\lambda_{\underline{\varepsilon}}$ in $\underline{\delta}$) as

$$q_{1/0}\left(\underline{\delta}\,|\lambda_{\underline{\varepsilon}}\right) = \frac{p_{1/0\underline{\delta}}\left(\underline{\delta}\right)}{\int\limits_{\{\underline{\delta}|\lambda_{\underline{\varepsilon}}=L_{\underline{\varepsilon}}(\underline{\boldsymbol{F}}(\underline{\delta}))\}} d\underline{\delta}\, p_{1/0\underline{\delta}}\left(\underline{\delta}\right)}. \tag{21}$$

Proof: As they are non-negative across the support space and integrate to unity, the functions defined in (21) are clearly probability densities. Consider the expectation of the LLR computed over $q_1\left(\underline{\delta}\,|\lambda_{\underline{\varepsilon}}\right)$

$$\left\langle\lambda_{\underline{\delta}}\right\rangle_{q_1} = \int d\underline{\delta}\, L_{\underline{\delta}}\left(\underline{\delta}\right) q_1\left(\underline{\delta}\,|\lambda_{\underline{\varepsilon}}\right) \tag{22}$$

where the functional form has been used to denote the LLR. Using (21)

$$L_{\underline{\delta}}\left(\underline{\delta}\right) = \ln\left(\frac{p_{1\underline{\delta}}\left(\underline{\delta}\right)}{p_{0\underline{\delta}}\left(\underline{\delta}\right)}\right) = \ln\left(\frac{\int\limits_{\{\underline{\delta}|\lambda_{\underline{\varepsilon}}=L_{\underline{\varepsilon}}(\underline{\boldsymbol{F}}(\underline{\delta}))\}} d\underline{\delta}\, p_{1\underline{\delta}}\left(\underline{\delta}\right)}{\int\limits_{\{\underline{\delta}|\lambda_{\underline{\varepsilon}}=L_{\underline{\varepsilon}}(\underline{\boldsymbol{F}}(\underline{\delta}))\}} d\underline{\delta}\, p_{0\underline{\delta}}\left(\underline{\delta}\right)}\right) + \ln\left(\frac{q_1\left(\underline{\delta}\,|\lambda_{\underline{\varepsilon}}\right)}{q_0\left(\underline{\delta}\,|\lambda_{\underline{\varepsilon}}\right)}\right). \tag{23}$$

From Appendix A (Equation (80)), the integral may be written in terms of $\underline{\varepsilon}$ as

$$\int\limits_{\{\underline{\delta}|\lambda_{\underline{\varepsilon}}=L_{\underline{\varepsilon}}(\underline{\boldsymbol{F}}(\underline{\delta}))\}} d\underline{\delta}\, p_{1/0\underline{\delta}}\left(\underline{\delta}\right)$$

$$= \int\limits_{\{\underline{\varepsilon}|\lambda_{\underline{\varepsilon}}=L_{\underline{\varepsilon}}(\underline{\varepsilon})\}} d\underline{\varepsilon} \int\limits_{\{\underline{\delta}|\underline{\varepsilon}=\underline{\boldsymbol{F}}(\underline{\delta})\}} d\underline{\delta}\, \frac{p_{1/0\underline{\delta}}\left(\underline{\delta}\right)}{\sqrt{\det\left(\left(\frac{\partial \underline{\boldsymbol{F}}}{\partial \underline{\delta}}\right)\left(\frac{\partial \underline{\boldsymbol{F}}}{\partial \underline{\delta}}\right)^T\right)}}$$

$$= \int\limits_{\{\underline{\varepsilon}|\lambda_{\underline{\varepsilon}}=L_{\underline{\varepsilon}}(\underline{\varepsilon})\}} d\underline{\varepsilon}\, p_{1/0\underline{\varepsilon}}\left(\underline{\varepsilon}\right) \tag{24}$$

so that

$$L_{\underline{\delta}}\left(\underline{\delta}\right) = \ln\left(\frac{\int\limits_{\{\underline{\varepsilon}|\lambda_{\underline{\varepsilon}}=L_{\underline{\varepsilon}}(\underline{\varepsilon})\}} d\underline{\varepsilon}\, p_{1\underline{\varepsilon}}\left(\underline{\varepsilon}\right)}{\int\limits_{\{\underline{\varepsilon}|\lambda_{\underline{\varepsilon}}=L_{\underline{\varepsilon}}(\underline{\varepsilon})\}} d\underline{\varepsilon}\, p_{0\underline{\varepsilon}}\left(\underline{\varepsilon}\right)}\right) + \ln\left(\frac{q_1\left(\underline{\delta}\,|\lambda_{\underline{\varepsilon}}\right)}{q_0\left(\underline{\delta}\,|\lambda_{\underline{\varepsilon}}\right)}\right). \tag{25}$$

However, $p_{1\underline{\varepsilon}}\left(\underline{\varepsilon}\right) = e^{\lambda_{\underline{\varepsilon}}} p_{0\underline{\varepsilon}}\left(\underline{\varepsilon}\right)$, with $e^{\lambda_{\underline{\varepsilon}}}$ constant over the range of integration; hence,

$$L_{\underline{\delta}}\left(\underline{\delta}\right) = \lambda_{\underline{\varepsilon}} + \ln\left(\frac{q_1\left(\underline{\delta}\,|\lambda_{\underline{\varepsilon}}\right)}{q_0\left(\underline{\delta}\,|\lambda_{\underline{\varepsilon}}\right)}\right) \tag{26}$$

and

$$\left\langle\lambda_{\underline{\delta}}\right\rangle_{q_1} = \lambda_{\underline{\varepsilon}} + \int d\underline{\delta} \ln\left(\frac{q_1\left(\underline{\delta}\,|\lambda_{\underline{\varepsilon}}\right)}{q_0\left(\underline{\delta}\,|\lambda_{\underline{\varepsilon}}\right)}\right) q_1\left(\underline{\delta}\,|\lambda_{\underline{\varepsilon}}\right). \tag{27}$$

The last term is non-negative by Gibbs' Inequality, providing the upper bounds in (20). Analogous consideration of $\left\langle\lambda_{\underline{\delta}}\right\rangle_{q_0}$ demonstrates the lower bounds.

The central insight provided by the LLR Ordering Theorem is that, subject to multi-valued branching effects, instantaneous information (as measured by LLRs) may be meaningfully compared in a direct manner, without consideration of the statistical expectations inherent in average information measures such as entropy and GSNR. As might then be expected, LLR may be clearly differentiated from other scalar statistics as a direct quantitative distillation of such instantaneous information.

Corollary 3 (LLR Uniqueness): Any scalar statistic ζ of $\underline{\delta}$ is information-preserving (but possibly not self-scaling) if and only if it maps one-to-one with the LLR.

Proof: Consider the scalar-to-scalar transformation

$$\lambda_\zeta = L_\zeta\left(\zeta\right) = \ln\left(\frac{p_{1\zeta}\left(\zeta\right)}{p_{0\zeta}\left(\zeta\right)}\right). \tag{28}$$

From Appendix A,

$$p_{1/0\lambda_\zeta}\left(\lambda_\zeta\right) = \sum_i \frac{p_{1/0\zeta}\left(\zeta_i\right)}{\left|dL_\zeta/d\zeta\right|\big|_{\zeta_i}} \tag{29}$$

so that the Self-Scaling Property requires

$$\lambda_\zeta \equiv \ln\left(\frac{p_{1\lambda_\zeta}\left(\lambda_\zeta\right)}{p_{0\lambda_\zeta}\left(\lambda_\zeta\right)}\right) \equiv \ln\left(\frac{\sum\limits_i \frac{p_{1\zeta}(\zeta_i)}{|dL_\zeta/d\zeta||_{\zeta_i}}}{\sum\limits_i \frac{p_{0\zeta}(\zeta_i)}{|dL_\zeta/d\zeta||_{\zeta_i}}}\right) \equiv ln\left(\frac{p_{1\zeta}\left(\zeta\right)}{p_{0\zeta}\left(\zeta\right)}\right). \tag{30}$$

which can only be true if, at all points, i never exceeds one, i.e., the mapping between ζ and λ_ζ must be one-to-one.

If ζ maps one-to-one with $\lambda_{\underline{\delta}}$, clearly λ_ζ must also. The LLR Ordering Theorem then requires that $\lambda_\zeta \equiv \lambda_{\underline{\delta}}$, implying

$$GSNR(\zeta) = GSNR(\lambda_\zeta) = GSNR(\lambda_{\underline{\delta}}) = GSNR(\underline{\delta}) \quad (31)$$

so that ζ must be information-preserving.

Conversely, if ζ is an information-preserving transformation, then by Corollary 1

$$\lambda_\zeta \equiv \ln\left(\frac{p_{1\zeta}(\zeta)}{p_{0\zeta}(\zeta)}\right) \equiv \ln\left(\frac{p_{1\underline{\delta}}(\underline{\delta})}{p_{0\underline{\delta}}(\underline{\delta})}\right) \equiv \lambda_{\underline{\delta}} \quad (32)$$

so that ζ must map one-to-one with $\lambda_{\underline{\delta}}$.

However, in either case, ζ itself cannot be self-scaling unless $\zeta \equiv \lambda_{\underline{\delta}}$.

Closely related to the issue of LLR uniqueness is the concept of the null LLR.

Corollary 4: An LLR statistic $\lambda_\emptyset$ possesses $GSNR = 0$ if and only if

$$p_1(\lambda_\emptyset) \equiv p_0(\lambda_\emptyset) \equiv \delta(\lambda_\emptyset) \quad (33)$$

where $\delta(\cdot)$ is a Dirac delta function.

Proof: If $GSNR = 0$, the KL divergence must be zero, so that $p_{1\lambda_\emptyset}(\lambda_\emptyset) \equiv p_{0\lambda_\emptyset}(\lambda_\emptyset)$. Self-scaling then implies $\lambda_\emptyset \equiv 0$, so the conditional densities can only have support at the origin. Proof of the converse is a trivial application of self-scaling and Corollary 2.

In summary, the above results paint the following conceptual picture:

- There is a finite amount of information available in any set of observations.
- GSNR quantifies the average information available, in a Shannon sense, so that it is universally comparable across the spectrum of potential observations.
- Any transformation of the observations (i.e., processing) either preserves or generates average information loss, but can never increase average information.
- The most compact (i.e., single number) informational compression of the observations in a lossless manner is the LLR.
- As GSNR quantifies average information in scalar form, the LLR distills the Shannon information contained in any specific exemplar of the observations.
- An authentic LLR may be differentiated from any other arbitrary statistic by recourse to the self-scaling property.
- As it is permissible to universally compare average information content using GSNR, it is also permissible to compare instantaneous information content using LLR (subject to multi-valued branching effects).
- Any other method for lossless compression of the observations to a scalar value *must* possess a one-to-one mapping to the LLR.

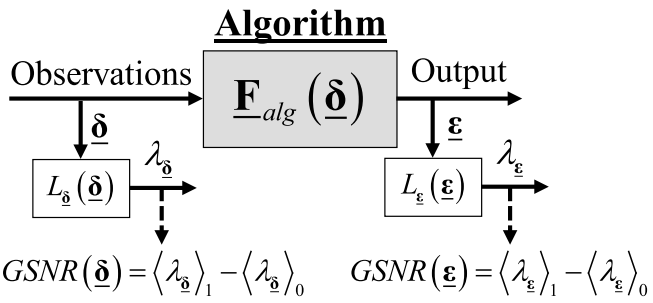


FIGURE 2. Block diagram of processing algorithm with performance characterization.

IV. PRACTICAL CHARACTERIZATIONS OF PROCESSING PERFORMANCE USING SNR

The essential methodology of using GSNR for real-world performance characterization is introduced here by considering the canonical problem of attaching a performance value to an arbitrary signal processing algorithm. Such an algorithm is illustrated as a block diagram in Figure 2. The algorithmic description may be cast mathematically as a multi-dimensional (presumably non-linear) transformation $\underline{\boldsymbol{F}}_{alg}$ from an $N \times 1$ vector of inputs $\underline{\delta}$ to an $M \times 1$ vector of outputs $\underline{\varepsilon}$,

$$\underline{\varepsilon} = \underline{\boldsymbol{F}}_{alg}(\underline{\delta}). \quad (34)$$

Some assumptions implicit in this statement are worth stating explicitly. First, the probability densities $p_{1/0\underline{\delta}}(\underline{\delta})$ must be fully defined. An inability or unwillingness to do so in an unconditional manner (for anything other than the hypothetical outcomes) *must* yield an ill-posed problem, as the original amount of exploitable information cannot be quantified. Second, the output $\underline{\varepsilon}$ must be describable as a vector of joint random variables; that is, $p_{1/0\underline{\varepsilon}}(\underline{\varepsilon})$ must also exist, which, in turn, constrains both $\underline{\boldsymbol{F}}_{alg}$ and $\underline{\varepsilon}$. While choices of $\underline{\boldsymbol{F}}_{alg}$ in this paper are intentionally limited in this regard to avoid mathematical digression, the limitation on $\underline{\varepsilon}$ is more easily stated; namely, individual random variables which are duplicative (in the sense of being strictly linearly dependent on the remaining entries) cannot be included. Here, the prototypical example is the random vector

$$\underline{y} = \begin{bmatrix} x \\ x \end{bmatrix}. \quad (35)$$

which, for a Gaussian random variable x, implies a singular covariance matrix. However, this is clearly only a matter of the specific definition of $\underline{\varepsilon}$, since any removed entries may always be subsequently recreated (implying that they can carry no additional information); an obvious implication of this constraint is that M cannot exceed N.

Performance characterization of $\underline{\boldsymbol{F}}_{alg}$ is then straightforward. Given $p_{1/0\underline{\delta}}(\underline{\delta})$, the input LLR may be computed and GSNR then evaluated from the Mean LLR Property. Then, based upon $p_{1/0\underline{\delta}}(\underline{\delta})$, $p_{1/0\underline{\varepsilon}}(\underline{\varepsilon})$ may be evaluated. This may be accomplished using any convenient method ranging from fully analytic approaches (per Appendix A) to fully numerical approaches employing Monte-Carlo techniques. Once in

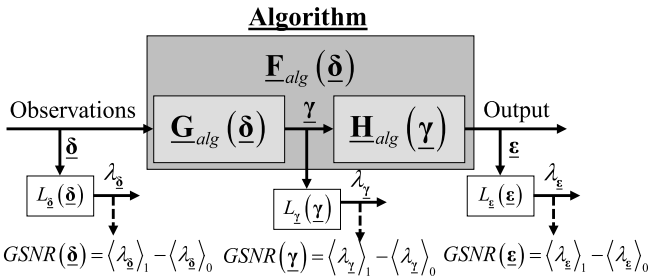


FIGURE 3. LLRs used as performance probes for a processing algorithm.

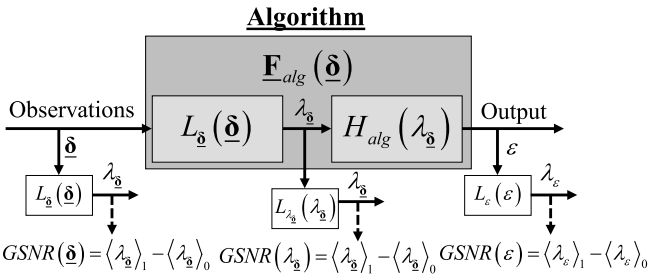


FIGURE 4. LLRs used as performance probes on the original observation LLR algorithm.

possession of $p_{1/0\,\underline{\varepsilon}}(\underline{\varepsilon})$, the output LLR and GSNR may be computed in analogous fashion. Both LLRs and GSNRs may be meaningfully compared; differences provide a direct measure of the amount of information lost due to the construction of the algorithm; if equal, the algorithm provides lossless conversion and may be considered "optimal." While it is not absolutely necessary to compute LLRs as an intermediate step in obtaining the GSNRs, doing so adds significant value by allowing instantaneous, single input comparisons to be made. If properly measured, output LLR/GSNR cannot exceed input LLR/GSNR, since the LLR Ordering and Information Limit Theorems apply.

As shown in Figure 3, formulation of the LLR statistic may be used diagnostically to provide a fully calibrated "performance probe" at any selected point in a processing string, with the assurance that equality of LLRs at different taps guarantees that optimal performance is being obtained between those taps.

Because of the self-scaling property, LLRs serve a unique dual role as both algorithm and statistic, resulting in two derivative implications: (1) LLR statistics are self-calibrating/self-repairing; and (2) information loss can only occur at points in a processing stream where irreversible transformation occurs. To illustrate this, consider the block diagram in Figure 4, where the LLR of the original observations is now the first part of the algorithm under consideration. By the self-scaling property, the LLR at the middle tap point must be $\lambda_{\underline{\delta}}$. Then, if $H_{alg}(\lambda_{\underline{\delta}})$ is information preserving,

$$\lambda_{\varepsilon} \equiv \lambda_{\underline{\delta}} \Rightarrow \; L_{\varepsilon}\left(H_{alg}\left(\lambda_{\underline{\delta}}\right)\right) \equiv \lambda_{\underline{\delta}} \Rightarrow \; L_{\varepsilon}(\varepsilon) = H_{alg}^{-1}(\varepsilon) \tag{36}$$

so that the LLR operation at the final tap point simply undoes the final algorithmic transformation. Conversely, the only way that $H_{alg}(\lambda_{\underline{\delta}})$ cannot be information preserving is for it to be irreversible (such as a square or absolute value operator), so $\lambda_{\underline{\delta}}$ cannot be recovered. While the specific example involves a scalar tap point, the train of logic readily generalizes multi-dimensional situations.

A similar approach permits the more challenging issue of "density mismatch" to be addressed. Suppose an LLR operator is formulated based upon one set of observational densities (the "presumed" densities $p_{1/0p}(\underline{\delta})$), but then employed on observations generated by random draw from a

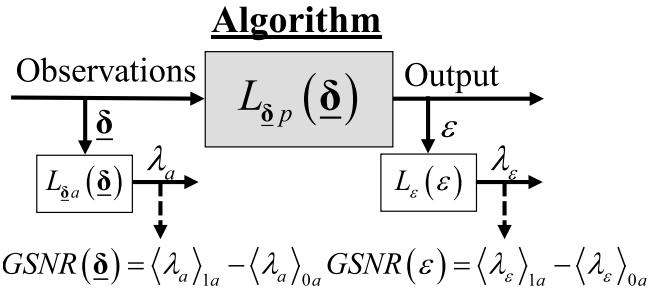


FIGURE 5. Performance measurement of a statistically mismatched LLR detector.

potentially different set of densities (differentiated by labeling them as the "actual" densities $p_{1/0a}(\underline{\delta})$). This situation is shown in Figure 5. Here, the output of the presumed LLR operator is labeled as ε, since it is not a true LLR (i.e., self-scaling does not hold – if it does, then the presumed and actual densities must match, and there is no performance loss). Any differences between the LLRs (GSNRs) are direct instantaneous (average) measures of the information lost due to the mismatch in statistical assumptions. This approach can be daisy-chained as many times as necessary; any sequence of such operations will cease changing only when a true LLR is reached, a very useful form of idempotence.

Because LLR and GSNR are inherent characteristics of the observations/outputs on which they are based, they must be universally meaningful within the context of the hypothesis under consideration. This is illustrated in Figure 6. Consider first the simpler case where the algorithms being compared operate on the same observations, so $\underline{\delta}_1 = \underline{\delta}_2 = \underline{\delta}$. Since both output LLRs/GSNRs are comparable to the equivalent input values, they may then be meaningfully compared among themselves, with the larger providing better preservation of available information. Note, however, that this requires a single set of statistical assumptions on the inputs that are then self-consistently carried through to both sets of outputs. For algorithms processing different observations (such as comparing a sonar sensor with a radar sensor), the LLR/GSNR of a composite set of observations is readily calculable; for statistically independent observations, it is just the sum of the LLRs/GSNRs of the components. Thus, all true LLRs/GSNRs must be universally comparable.

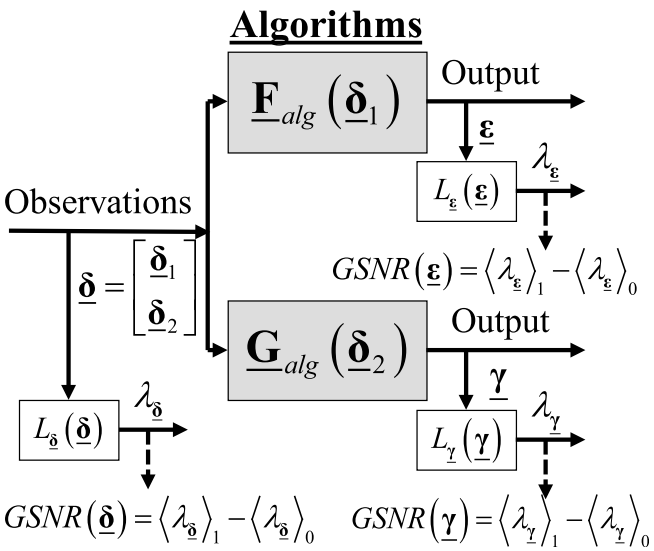


FIGURE 6. Comparing the performance of different processing algorithms.

V. SOME SPECIFIC EXAMPLES

Several examples of specific results are now provided to demonstrate both the immediacy and the breadth of utility of these insights. First, the results for the very classic problem of a known signal in additive Gaussian noise are provided; these results are well understood, but requoting them permits validation that they fit within this more general framework. The data model for the known signal in Gaussian noise detector is

$$\begin{aligned} H_1:&\ \underline{\delta} = \underline{s} + \underline{n} \quad \underline{s} \text{ known} \\ H_0:&\ \underline{\delta} = \underline{n} \qquad \underline{n} \sim N\left(\underline{0},\ \overline{\underline{R}}_n\right) \end{aligned} \tag{37}$$

where it is assumed that the observations are complex numbers. For this data model,

$$\begin{aligned} \lambda &= \underline{s}^{+}\overline{\underline{R}}_n^{-1}\underline{\delta} + \underline{\delta}^{+}\overline{\underline{R}}_n^{-1}\underline{s} - \underline{s}^{+}\overline{\underline{R}}_n^{-1}\underline{s} \\ &\Rightarrow \quad \langle\lambda\rangle_{1/0} = \pm\underline{s}^{+}\overline{\underline{R}}_n^{-1}\underline{s} \\ &\Rightarrow \quad GSNR = 2\underline{s}^{+}\overline{\underline{R}}_n^{-1}\underline{s}. \end{aligned} \tag{38}$$

As mentioned earlier, for cases where a composite observation may be divided into statistically independent components for which the data model applies individually, all of these results may be simply summed over the components.

Next, the results for the equally classic problem of Gaussian signal in additive Gaussian noise are provided, primarily because most traditional methods of evaluating the performance of the underlying energy (or quadratic) operation are laced with approximations, generated primarily by a presumed need to refer performance to the detector input (i.e., the input of the non-linear squaring operation). By contrast, the GSNR performance measurement is fully capable of handling the effect of non-linear operations exactly. The data model for the Gaussian signal in Gaussian noise detector is

$$\begin{aligned} H_1:&\ \underline{\delta} = \alpha\underline{e} + \underline{n} \quad \alpha \sim N\left(\underline{0},\ \sigma_s^2\right),\ \underline{e} \text{ known} \\ H_0:&\ \underline{\delta} = \underline{n} \qquad \underline{n} \sim N\left(\underline{0},\ \overline{\underline{R}}_n\right). \end{aligned} \tag{39}$$

with the further proviso that α and $\underline{n}$ be statistically independent. Again assuming complex observations,

$$\begin{aligned} \lambda &= \left(\frac{\sigma_s^2\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{e}}{1+\sigma_s^2\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{e}}\right)\left(\frac{\left|\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{\delta}\right|^2}{\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{e}}\right) - \ln\left(1+\sigma_s^2\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{e}\right) \\ \Rightarrow \quad \langle\lambda\rangle_1 &= \sigma_s^2\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{e} - \ln\left(1+\sigma_s^2\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{e}\right) \\ \langle\lambda\rangle_0 &= \left(\frac{\sigma_s^2\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{e}}{1+\sigma_s^2\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{e}}\right) - \ln\left(1+\sigma_s^2\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{e}\right) \\ \Rightarrow \quad GSNR &= \frac{\left(\sigma_s^2\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{e}\right)^2}{1+\sigma_s^2\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{e}} \end{aligned} \tag{40}$$

with independent components again summable. This exact result reduces to

$$\begin{aligned} GSNR_{LS} &= \lim_{\sigma_s^2\to\infty} GSNR \sim \sigma_s^2\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{e} \quad \text{and} \\ GSNR_{SS} &= \lim_{\sigma_s^2\to 0} GSNR \sim \left(\sigma_s^2\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{e}\right)^2 \end{aligned} \tag{41}$$

in the large and small signal limits, respectively.

Classical performance measurement approaches struggle when applied to more challenging densities, such as those with ill-conditioned means and variances. As an example of the true generality of this approach, the GSNR associated with scalar Cauchy statistics (for which neither mean, variance, nor any higher order moment exists) [20], is evaluated. Here, the data model is

$$\begin{aligned} H_1:&\quad \delta = \kappa_1 \qquad p(\kappa_1) = \frac{1}{\pi}\left(\frac{\gamma_1}{(\kappa_1 - x_1)^2 + \gamma_1^2}\right) \\ H_0:&\quad \delta = \kappa_0 \qquad p(\kappa_0) = \frac{1}{\pi}\left(\frac{\gamma_0}{(\kappa_0 - x_0)^2 + \gamma_0^2}\right) \end{aligned} \tag{42}$$

where $x_{1/0}$ and $\gamma_{1/0}$ are, respectively, the $H_{1/0}$ location and scale parameters, so [21]

$$\begin{aligned} \lambda &= 7\ln\left(\frac{(\delta - x_0)^2 + \gamma_0^2}{(\delta - x_1)^2 + \gamma_1^2}\right) + \ln\left(\frac{\gamma_1}{\gamma_0}\right) \\ &\Rightarrow \langle\lambda\rangle_{1/0} = \pm\ln\left(\frac{(x_1 - x_0)^2 + (\gamma_1 + \gamma_0)^2}{4\gamma_1\gamma_0}\right) \\ &\Rightarrow GSNR = 2\ \ln\left(\frac{(x_1 - x_0)^2 + (\gamma_1 + \gamma_0)^2}{4\gamma_1\gamma_0}\right). \end{aligned} \tag{43}$$

The above examples demonstrate the use of the theory in Section 2 and 3 to evaluate specific GSNRs. The next two examples demonstrate its use in characterizing information loss caused by suboptimal processing algorithms. The first considers a Generalized Likelihood Ratio Test (GLRT) processing formulation; GLRTs are regularly used for convenience but often lack meaningful performance characterization, a situation easily rectified using information theory constructs.

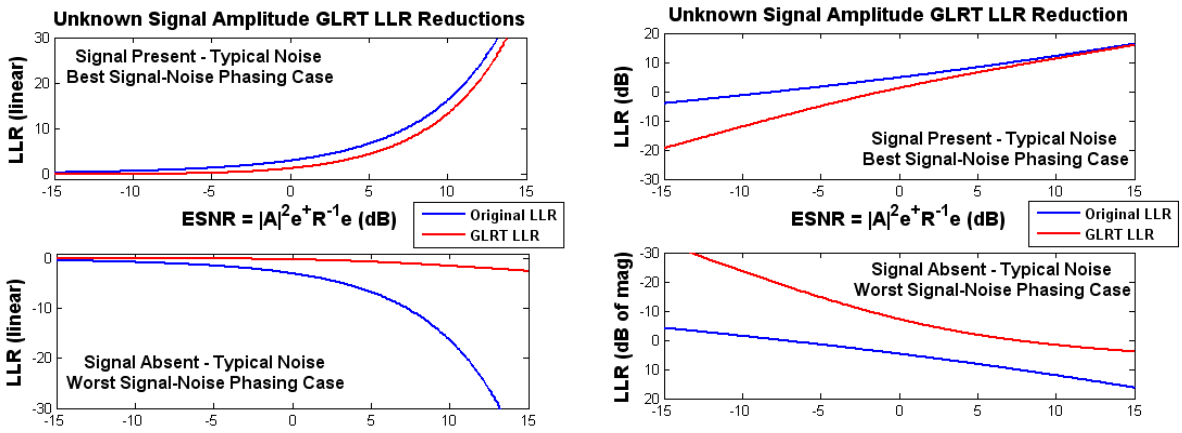


FIGURE 7. LLR reductions generated by unknown signal amplitude GLRT formulation (linear [left] and logarithmic [right] vertical scales).

Consider one of the simplest GLRT formulations, involving unknown signal amplitude. The underlying data model is the same as that presented in (37), with the exception that the signal vector (still presumed to be fully known) is written in terms of a shape vector and a complex scalar amplitude, or equivalently, a real scalar amplitude and a scalar phase, i.e.,

$$\underline{s} = A\,\underline{e} = |A|\,e^{j\phi}\,\underline{e} \tag{44}$$

where, to make the decomposition unique, it is required that $\underline{e}^{+}\underline{e} = 1$. Then, for any actual complex amplitude A_a, the information available within the observations is that of (38)

$$GSNR\left(\underline{\delta}\right) = 2\,|A_a|^2\,\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{e} \tag{45}$$

However, the appropriate LLR formulation (also from (38)) is

$$\lambda = A^{*}\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{\delta} + A\,\underline{\delta}^{+}\overline{\underline{R}}_n^{-1}\underline{e} - |A|^2\,\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{e}. \tag{46}$$

In the all too common situation where the underlying data model is chosen for convenience rather than realism, this formulation inconveniently requires prior knowledge of signal amplitude and phase that is, in reality, unknown. The unknown amplitude GLRT attempts to repair the situation by using the amplitude which maximizes the target present likelihood of the observations (which is identical to maximizing the LLR in (46)) so that

$$\begin{aligned}\hat{A} &= \frac{\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{\delta}}{\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{e}} \quad\Rightarrow\quad \varepsilon_{GLRT} = \lambda_{max} \\ &= \hat{A}^{*}\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{\delta} + \hat{A}\,\underline{\delta}^{+}\overline{\underline{R}}_n^{-1}\underline{e} - \left|\hat{A}\right|^2\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{e} \\ &= \frac{\left|\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{\delta}\right|^2}{\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{e}}\end{aligned} \tag{47}$$

The notation ε_{GLRT} is used to emphasize that the GLRT does not yield a true LLR, since

$$\langle\varepsilon_{GLRT}\rangle_1 = 1 + |A_a|^2\,\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{e} \text{ and } \langle\varepsilon_{GLRT}\rangle_0 = 1, \tag{48}$$

which violates the mean LLR property (see (16)). The resulting information loss may be quantified by evaluating $GSNR\,(\varepsilon_{GLRT})$ and comparing it to that available in the observations. By inspection of (47), all the elements of $\underline{\delta}$ combine linearly, so ε_{GLRT} is a chi-square random variable with one complex degree of freedom

$$\begin{aligned}H_1:&\quad p_1\left(\varepsilon_{GLRT}\right) = \left(\frac{1}{\langle\varepsilon_{GLRT}\rangle_1}\right)e^{-\varepsilon_{GLRT}/\langle\varepsilon_{GLRT}\rangle_1}u\left(\varepsilon_{GLRT}\right) \\ H_0:&\quad p_0\left(\varepsilon_{GLRT}\right) = e^{-\varepsilon_{GLRT}}u\left(\varepsilon_{GLRT}\right).\end{aligned} \tag{49}$$

Hence

$$\begin{aligned}\lambda_\varepsilon &= \left(\frac{|A_a|^2\,\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{e}}{1+|A_a|^2\,\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{e}}\right)\varepsilon_{GLRT} - \ln\left(1+|A_a|^2\,\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{e}\right) \\ \Rightarrow\quad GSNR\left(\varepsilon_{GLRT}\right) &= \frac{\left(|A_a|^2\,\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{e}\right)^2}{1+|A_a|^2\,\underline{e}^{+}\overline{\underline{R}}_n^{-1}\underline{e}}.\end{aligned} \tag{50}$$

To exemplify the LLR ordering theorem, λ_ε is compared with the original LLR from (46) in Figure 7, for specific cases with signal present and signal absent, using both linear and logarithmic (dB) scales. Note that the bottom right panel of Figure 7 depicts negative values on a logarithmic scale. In Figure 8, the GLRT GSNR is compared with the originally available GSNR from (45).

It is interesting to note that, if one accepts the conceptual equivalence of σ_s^2 and $|A_a|^2$, the information retained by the unknown amplitude GLRT exactly matches that available in the Gauss-Gauss problem (Equation (40)). This is suggestive of the ability of these constructs to adjudicate both the value provided and the accuracies required when attempting to exploit prior information in signal processing applications.

As a final example, consider the following question: What is the performance loss caused by errors in the knowledge of the covariance matrix? To the author, this appears to be a critical question that has been addressed at best peripherally in the mass of adaptive processing analysis conducted to date. Adopting the Gauss-Gauss data model of (39),

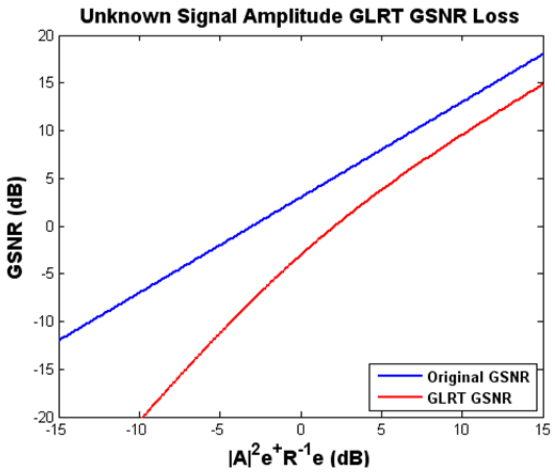


FIGURE 8. GSNR loss generated by unknown signal amplitude GLRT formulation.

the information available in the original observations is, given actual noise covariance matrix $\underline{\overline{\boldsymbol{R}}}_{na}$,

$$GSNR\left(\underline{\delta}\right) = \frac{\left(\sigma_s^2 \underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{na}^{-1} \underline{e}\right)^2}{1 + \sigma_s^2 \underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{na}^{-1} \underline{e}}. \tag{51}$$

However, the mismatched Gauss-Gauss LLR formulation for an estimated or otherwise mismatched (i.e., presumed) noise covariance matrix $\underline{\overline{\boldsymbol{R}}}_{np}$ is

$$\varepsilon_{mm} = \left(\frac{\sigma_s^2 \underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{e}}{1 + \sigma_s^2 \underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{e}}\right)\left(\frac{\left|\underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{\delta}\right|^2}{\underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{e}}\right) - \ln\left(1 + \sigma_s^2 \underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{e}\right) \tag{52}$$

where the notation ε_{mm} is again used to avoid confusion with a true LLR. As a practical matter, it is convenient to introduce the modified statistic

$$\gamma_{mm} = \left(\frac{1 + \sigma_s^2 \underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{e}}{\sigma_s^2 \underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{e}}\right)\left(\varepsilon_{mm} + \ln\left(1 + \sigma_s^2 \underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{e}\right)\right) = \left(\frac{\left|\underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{\delta}\right|^2}{\underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{e}}\right). \tag{53}$$

Since the transformation is linear, and thus reversible, the information contained in the two statistics is guaranteed to be the same. Now γ_{mm} is once more a chi-square random variable with one complex degree of freedom, with means

$$\langle \gamma_{mm} \rangle_1 = \frac{\underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{\overline{\boldsymbol{R}}}_{na} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{e} + \sigma_s^2 \left(\underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{e}\right)^2}{\underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{e}} \quad \text{and}$$

$$\langle \gamma_{mm} \rangle_0 = \frac{\underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{\overline{\boldsymbol{R}}}_{na} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{e}}{\underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{e}}, \tag{54}$$

implying densities

$$H_1 : p_1\left(\gamma_{mm}\right) = \left(\frac{1}{\langle \gamma_{mm} \rangle_1}\right) e^{-\gamma_{mm}/\langle \gamma_{mm} \rangle_1} u\left(\gamma_{mm}\right)$$
$$H_0 : p_0\left(\gamma_{mm}\right) = \left(\frac{1}{\langle \gamma_{mm} \rangle_0}\right) e^{-\gamma_{mm}/\langle \gamma_{mm} \rangle_0} u\left(\gamma_{mm}\right) \tag{55}$$

LLR

$$\lambda_\gamma = \left(\frac{1}{\langle \gamma_{mm} \rangle_0} - \frac{1}{\langle \gamma_{mm} \rangle_1}\right) \gamma_{mm} + \ln\left(\langle \gamma_{mm} \rangle_1\right) - \ln\left(\langle \gamma_{mm} \rangle_0\right) \tag{56}$$

and GSNR

$$GSNR\left(\varepsilon_{mm}\right) = \frac{\left(\langle \gamma_{mm} \rangle_1 - \langle \gamma_{mm} \rangle_0\right)^2}{\langle \gamma_{mm} \rangle_1 \langle \gamma_{mm} \rangle_0} = \frac{\left(\sigma_s^2 \left(\underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{e}\right)^2 \Big/ \underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{\overline{\boldsymbol{R}}}_{na} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{e}\right)^2}{1 + \sigma_s^2 \left(\underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{e}\right)^2 \Big/ \underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{\overline{\boldsymbol{R}}}_{na} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{e}}. \tag{57}$$

This exact result reduces to

$$GSNR_{LS}\left(\varepsilon_{mm}\right) = \lim_{\sigma_s^2 \to \infty} GSNR\left(\varepsilon_{mm}\right) \sim \sigma_s^2 \left(\underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{e}\right)^2 \Big/ \underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{\overline{\boldsymbol{R}}}_{na} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{e}$$

and

$$GSNR_{SS}\left(\varepsilon_{mm}\right) = \lim_{\sigma_s^2 \to 0} GSNR\left(\varepsilon_{mm}\right) \sim \left(\sigma_s^2 \left(\underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{e}\right)^2 \Big/ \underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{\overline{\boldsymbol{R}}}_{na} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{e}\right)^2 \tag{58}$$

in the large and small signal limits, respectively. Since the mismatched information must be smaller than the original information, this implies

$$\left(\underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{e}\right)^2 \Big/ \left[\left(\underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{\overline{\boldsymbol{R}}}_{na} \underline{\overline{\boldsymbol{R}}}_{np}^{-1} \underline{e}\right)\left(\underline{e}^{+} \underline{\overline{\boldsymbol{R}}}_{na}^{-1} \underline{e}\right)\right] \leq 1. \tag{59}$$

which is easily demonstrable from the properties of positive definite matrices [22]. The critical observation is that, in the small signal limit, GSNR drops as the *square* of the fractional mismatch in (59). This makes the effectiveness of small signal incoherent integration extremely sensitive to covariance errors, an effect elsewhere given the name Incoherent Gain Limit (IGL).

VI. SOME CONCEPTUAL CONSEQUENCES

A firm understanding of the information measurement role that GSNR and LLR play leads to a few more fundamental observations (perhaps bordering on the philosophic) that deserve some discussion.

GSNR measures the average information available from the probabilistic ensemble in a relative sense, that is, as a difference which is guaranteed to be positive. LLR measures the instantaneous information provided by any specific draw from the probabilistic ensemble in a manner that captures

both information strength (magnitude) and information content (sign). Retention of a signed scale by LLR suggests much greater import to knowing the ''zero point'' of that scale than might be traditionally perceived. Indeed, although individual mean LLRs provide a more precise alternative to describing average information content, they (for the most part) are unused in current practice.

In retrospect, the whole of information accounting in the signal processing context hinges upon the specification of the data model; that is, the pair of joint probability densities for the original observations. Since both GSNR and LLR depend upon this specification, the implication is that the very concept of information is inherently probabilistic; unless one can and does properly specify the statistical situation, the amount of information available as well as the rules for its proper extraction remain indeterminate, and it is exceedingly difficult to consider any resulting methodology as fundamentally well posed.

This proscription extends to any form of ''loose'' constant, such as the ''deterministic but unknown'' parameters that ultimately lurk at the bottom of GLRT formulations. The available information is only fully specified for specific densities with known parameters. To make this idea explicit, consider a density on $\underline{\boldsymbol{x}}$ possessing parameter $\underline{\boldsymbol{\mu}}$ (be it mean, covariance, higher moment, or other arbitrary parameter). It is always acceptable to write this density as conditional on $\underline{\boldsymbol{\mu}}$, i.e., $p_{\underline{\boldsymbol{x}}}\left(\underline{\boldsymbol{x}}\middle|\,\underline{\boldsymbol{\mu}}\right)$; however, computation of the unconditional density needed to specify information content requires that $\underline{\boldsymbol{\mu}}$ then be treated as a random variable with a specified secondary density $p_{\underline{\boldsymbol{\mu}}}\left(\underline{\boldsymbol{\mu}}\right)$, so that the unconditional density may then be computed by marginalizing over $\underline{\boldsymbol{\mu}}$

$$p_{\underline{\boldsymbol{x}}}\left(\underline{\boldsymbol{x}}\right) = \int d\underline{\boldsymbol{\mu}}\, p_{\underline{\boldsymbol{x}}}\left(\underline{\boldsymbol{x}}\middle|\,\underline{\boldsymbol{\mu}}\right) p_{\underline{\boldsymbol{\mu}}}\left(\underline{\boldsymbol{\mu}}\right). \tag{60}$$

Under the special case that the value of $\underline{\boldsymbol{\mu}}$ is known to have value $\underline{\boldsymbol{\mu}}_0$, it may be assigned a delta function density located at $\underline{\boldsymbol{\mu}}_0$, in which case (60) reduces to

$$p_{\underline{\boldsymbol{x}}}\left(\underline{\boldsymbol{x}}\right) = p_{\underline{\boldsymbol{x}}}\left(\underline{\boldsymbol{x}}\middle|\,\underline{\boldsymbol{\mu}}\right)\Big|_{\underline{\boldsymbol{\mu}}=\underline{\boldsymbol{\mu}}_0}. \tag{61}$$

However, an inability or unwillingness to specify $\underline{\boldsymbol{\mu}}_0$ still yields an incomplete definition of $p_{\underline{\boldsymbol{x}}}\left(\underline{\boldsymbol{x}}\right)$, and, ultimately, indeterminate specification of the available information. This is why GLRT formulations cannot, in general, lead to true LLR statistics. The author believes that the precept of *complete* prior specification of the available information is closely related to many of the deeper existential issues that complicate the procedural landscapes of detection and estimation theory.

It is obvious that the use of the LLR statistic is completely justifiable solely as an information and performance measurement tool, independent of any other role it might play. This observation turns out to be the justification of the author's long-held belief that any optimal processing technique is inherently ''aware'' of its own performance (expected and actual). The LLR plays a unique, dual role in the exploitation of information; it provides the rules for the precise distillation of that information into a single value as well as measuring the amount of information so distilled. In this light, the self-scaling property is simply a statement that these two items are really just one and the same. This elegant arrangement then yields a universal capability for comparative performance measurement, even to the point of addressing mismatch effects through recursive self-calibration and self-correction.

VII. SUMMARY AND FUTURE EXTENSIONS

In this paper, a quantitative, universal information accounting methodology built around GSNR and the LLR has been developed. This methodology has the advantage of providing performance metrics that are scalar in nature, simplifying evaluation, comparison, and optimization. The validity of the methodology has been proven theoretically, and its practical application to issues of signal processing algorithm performance prediction, measurement, and comparison demonstrated.

In toto, these results raise the question of why any algorithmic approaches other than true LLR formulation and evaluation should even be contemplated, at least for binary detection applications. It is the considered opinion of the author that if, across the signal processing arena, the performance measurement methodology presented here were well understood and careful quantitative performance assessment and comparison of processing techniques were routinely mandated, then the use of algorithms other than those formally defensible as true LLR statistics would largely dry up, simply to minimize associated performance characterization effort. Among other benefits, this would force a much stronger emphasis on the explicit identification and validation of the underlying data models, which appears to be precisely the correct point of technical focus. That this state of affairs is not currently the case is, perhaps, the best empirical evidence that the author can provide for the importance of this material.

It is anticipated that future extensions of this work will

- Expand the application of information-based scalar performance measurements through completion of the decision-making step, demonstrating how the traditional precepts of optimal detection theory may then be cast as optimizations of such measurements; and
- Delineate in much greater depth the wealth of useful mathematical properties possessed by LLRs when considered as a specific class of scalar statistics.

APPENDIX A

In this appendix, the general form for the density function of an $M \times 1$ multivariate random variable $\underline{\boldsymbol{y}}$ arising as the multi-dimensional transformation $\underline{\boldsymbol{F}}$ of an $N \times 1$ multivariate random variable $\underline{\boldsymbol{x}}$ ($M \leq N$) is developed. This specific development assumes that the Jacobian of $\underline{\boldsymbol{F}}$ exists; situations where the Jacobian may not exist are beyond the scope of this paper. Development is cast in terms of real random variables, with the implication that it may also be applied to complex

forms by explicit separation into real and imaginary components.

Theorem (Densities of Generalized Functions of Random Variables): For the $N\ x\ 1$ real multivariate random variable

$$\underline{\boldsymbol{x}} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \tag{62}$$

with joint density function $p\left(\underline{\boldsymbol{x}}\right)$, and the transformation

$$\underline{\boldsymbol{y}} = \begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} f_1\left(\underline{\boldsymbol{x}}\right) \\ \vdots \\ f_M\left(\underline{\boldsymbol{x}}\right) \end{bmatrix} = \underline{\boldsymbol{F}}\left(\underline{\boldsymbol{x}}\right) \tag{63}$$

mapping $\underline{\boldsymbol{x}}$ to an $M\ x\ 1$ real multivariate random variable $\underline{\boldsymbol{y}}$ ($M \leq N$), with all components of $\underline{\boldsymbol{y}}$ linearly independent, the probability density function of $\underline{\boldsymbol{y}}$ is

$$p\left(\underline{\boldsymbol{y}}\right) = \int\limits_{\left\{\underline{\boldsymbol{x}}\middle|F\left(\underline{\boldsymbol{x}}\right)=\underline{\boldsymbol{y}}\right\}} d\underline{\boldsymbol{x}} \frac{p\left(\underline{\boldsymbol{x}}\right)}{\sqrt{\det\left(\left(\partial\underline{\boldsymbol{y}}\middle/\partial\underline{\boldsymbol{x}}\right)\left(\partial\underline{\boldsymbol{y}}\middle/\partial\underline{\boldsymbol{x}}\right)^T\right)}} \tag{64}$$

where $\partial\underline{\boldsymbol{y}}/\partial\underline{\boldsymbol{x}}$ is the $M\ x\ N$ Jacobian matrix of the transformation $\underline{\boldsymbol{F}}\left(\underline{\boldsymbol{x}}\right)$, i.e.,

$$\partial\underline{\boldsymbol{y}}\middle/\partial\underline{\boldsymbol{x}} = \begin{bmatrix} \partial f_1\middle/\partial x_1 & \cdots & \partial f_1\middle/\partial x_N \\ \vdots & \ddots & \vdots \\ \partial f_M\middle/\partial x_1 & \cdots & \partial f_M\middle/\partial x_N \end{bmatrix}. \tag{65}$$

Proof: One seeks to extend the transformation to full rank by appending the $(N-M)\ x\ 1$ vector $\underline{\boldsymbol{y}}_*$ to $\underline{\boldsymbol{y}}$ so that

$$\begin{aligned} \underline{\boldsymbol{y}}_{ex} = \begin{bmatrix} \underline{\boldsymbol{y}} \\ \underline{\boldsymbol{y}}_* \end{bmatrix} &= \begin{bmatrix} y_1 \\ \vdots \\ y_M \\ y_{*M+1} \\ \vdots \\ y_{*N} \end{bmatrix} = \begin{bmatrix} f_1\left(\underline{\boldsymbol{x}}\right) \\ \vdots \\ f_M\left(\underline{\boldsymbol{x}}\right) \\ f_{*M+1}\left(\underline{\boldsymbol{x}}\right) \\ \vdots \\ f_{*N}\left(\underline{\boldsymbol{x}}\right) \end{bmatrix} \\ &= \begin{bmatrix} \underline{\boldsymbol{F}}\left(\underline{\boldsymbol{x}}\right) \\ \underline{\boldsymbol{F}}_*\left(\underline{\boldsymbol{x}}\right) \end{bmatrix} = \underline{\boldsymbol{F}}_{ex}\left(\underline{\boldsymbol{x}}\right). \end{aligned} \tag{66}$$

Then the determinant of the Jacobian of the transformation $\det\left(\partial\underline{\boldsymbol{y}}_{ex}/\partial\underline{\boldsymbol{x}}\right)$ is well defined, and the density of $\underline{\boldsymbol{y}}_{ex}$ is [23]

$$p\left(\underline{\boldsymbol{y}}_{ex}\right) = \int\limits_{\left\{\underline{\boldsymbol{x}}\middle|F_{ex}\left(\underline{\boldsymbol{x}}\right)=\underline{\boldsymbol{y}}_{ex}\right\}} d\underline{\boldsymbol{x}} \frac{p\left(\underline{\boldsymbol{x}}\right)}{\left|\det\left(\partial\underline{\boldsymbol{y}}_{ex}\middle/\partial\underline{\boldsymbol{x}}\right)\right|}. \tag{67}$$

The desired density may be recovered by integrating out the extended random variables

$$\begin{aligned} p\left(\underline{\boldsymbol{y}}\right) &= \int d\underline{\boldsymbol{y}}_* p\left(\underline{\boldsymbol{y}}, \underline{\boldsymbol{y}}_*\right) = \int d\underline{\boldsymbol{y}}_* p\left(\underline{\boldsymbol{y}}_{ex}\right) \\ &= \int d\underline{\boldsymbol{y}}_* \int\limits_{\left\{\underline{\boldsymbol{x}}\middle|F_{ex}\left(\underline{\boldsymbol{x}}\right)=\underline{\boldsymbol{y}}_{ex}\right\}} d\underline{\boldsymbol{x}} \frac{p\left(\underline{\boldsymbol{x}}\right)}{\left|\det\left(\partial\underline{\boldsymbol{y}}_{ex}\middle/\partial\underline{\boldsymbol{x}}\right)\right|}. \end{aligned} \tag{68}$$

To this end, the Jacobian of $\underline{\boldsymbol{F}}$ may WLOG be rewritten in terms of its SVD

$$\partial\underline{\boldsymbol{y}}\middle/\partial\underline{\boldsymbol{x}} = \overline{\underline{\boldsymbol{U}}}\,\overline{\underline{\boldsymbol{\Lambda}}}\,\overline{\underline{\boldsymbol{V}}}_p^T \tag{69}$$

where $\overline{\underline{\boldsymbol{U}}}$ is an $M\ x\ M$ orthogonal matrix, $\overline{\underline{\boldsymbol{\Lambda}}}$ is an $M\ x\ M$ non-negative diagonal matrix, and $\overline{\underline{\boldsymbol{V}}}_p$ is an $N\ x\ M$ partial orthogonal matrix, the columns of which represent the first M components of a full $N\ x\ N$ orthogonal basis set. Now, from (69)

$$\begin{aligned} \left\{\left(\partial\underline{\boldsymbol{y}}\middle/\partial\underline{\boldsymbol{x}}\right)\left(\partial\underline{\boldsymbol{y}}\middle/\partial\underline{\boldsymbol{x}}\right)^T\right\}^{1/2} &= \left\{\left(\overline{\underline{\boldsymbol{U}}}\,\overline{\underline{\boldsymbol{\Lambda}}}\,\overline{\underline{\boldsymbol{V}}}_p^T\right)\left(\overline{\underline{\boldsymbol{V}}}_p\overline{\underline{\boldsymbol{\Lambda}}}\,\overline{\underline{\boldsymbol{U}}}^T\right)\right\}^{1/2} \\ &= \left\{\overline{\underline{\boldsymbol{U}}}\,\overline{\underline{\boldsymbol{\Lambda}}}^2\overline{\underline{\boldsymbol{U}}}^T\right\}^{1/2} = \overline{\underline{\boldsymbol{U}}}\,\overline{\underline{\boldsymbol{\Lambda}}}\overline{\underline{\boldsymbol{U}}}^T \end{aligned} \tag{70}$$

where the matrix square root is unique if the symmetric form is (WLOG) used. Hence,

$$\overline{\underline{\boldsymbol{U}}}\,\overline{\underline{\boldsymbol{\Lambda}}} = \left\{\left(\partial\underline{\boldsymbol{y}}\middle/\partial\underline{\boldsymbol{x}}\right)\left(\partial\underline{\boldsymbol{y}}\middle/\partial\underline{\boldsymbol{x}}\right)^T\right\}^{1/2}\overline{\underline{\boldsymbol{U}}}. \tag{71}$$

and (69) may be conveniently rewritten as

$$\begin{aligned} \partial\underline{\boldsymbol{y}}\middle/\partial\underline{\boldsymbol{x}} &= \left\{\left(\partial\underline{\boldsymbol{y}}\middle/\partial\underline{\boldsymbol{x}}\right)\left(\partial\underline{\boldsymbol{y}}\middle/\partial\underline{\boldsymbol{x}}\right)^T\right\}^{1/2}\overline{\underline{\boldsymbol{U}}}\,\overline{\underline{\boldsymbol{V}}}_p^T \\ &= \left\{\left(\partial\underline{\boldsymbol{y}}\middle/\partial\underline{\boldsymbol{x}}\right)\left(\partial\underline{\boldsymbol{y}}\middle/\partial\underline{\boldsymbol{x}}\right)^T\right\}^{1/2}\overline{\underline{\boldsymbol{W}}}_p^T \end{aligned} \tag{72}$$

where $\overline{\underline{\boldsymbol{W}}}_p$ is now a modified $N\ x\ M$ partial orthogonal matrix.

While subject to the orthogonality requirements, the $y_{*i} = f_{*i}\left(\underline{\boldsymbol{x}}\right)$ may otherwise be chosen arbitrarily. Here, they are not defined explicitly, but rather implicitly through their Jacobian, with the understanding that specification of the gradient vector

$$\vec{\nabla}_{\underline{\boldsymbol{x}}} f_{*i}\left(\underline{\boldsymbol{x}}\right) = \sum_{k=1}^{N} \frac{\partial f_{*i}}{\partial x_k}\vec{\boldsymbol{u}}_k \tag{73}$$

is sufficient to fully define $f_{*i}\left(\underline{\boldsymbol{x}}\right)$ to within an integration constant (which may WLOG then be chosen arbitrarily). Thus, $\underline{\boldsymbol{y}}_*$ is defined by requiring

$$\partial\underline{\boldsymbol{y}}_*\middle/\partial\underline{\boldsymbol{x}} = \overline{\underline{\boldsymbol{W}}}_{res}^T \tag{74}$$

where $\overline{\underline{\boldsymbol{W}}}_{res}$ is an $N\ x\ (N-M)$ partial orthogonal matrix completing $\overline{\underline{\boldsymbol{W}}}_p$; that is, the composite $N\ x\ N$ matrix $\overline{\underline{\boldsymbol{W}}} = \left[\overline{\underline{\boldsymbol{W}}}_p\ \overline{\underline{\boldsymbol{W}}}_{res}\right]$ is orthogonal. Then the Jacobian of the extended transformation may be written as

$$\begin{aligned} \partial\underline{\boldsymbol{y}}_{ex}\middle/\partial\underline{\boldsymbol{x}} &= \begin{bmatrix} \partial\underline{\boldsymbol{y}}\middle/\partial\underline{\boldsymbol{x}} \\ \partial\underline{\boldsymbol{y}}_*\middle/\partial\underline{\boldsymbol{x}} \end{bmatrix} \\ &= \begin{bmatrix} \left\{\left(\partial\underline{\boldsymbol{y}}\middle/\partial\underline{\boldsymbol{x}}\right)\left(\partial\underline{\boldsymbol{y}}\middle/\partial\underline{\boldsymbol{x}}\right)^T\right\}^{1/2} & \underline{\boldsymbol{0}} \\ \underline{\boldsymbol{0}} & \underline{\boldsymbol{I}}_{N-M} \end{bmatrix} \begin{bmatrix} \overline{\underline{\boldsymbol{W}}}_p^T \\ \overline{\underline{\boldsymbol{W}}}_{res}^T \end{bmatrix} \\ &= \begin{bmatrix} \left\{\left(\partial\underline{\boldsymbol{y}}\middle/\partial\underline{\boldsymbol{x}}\right)\left(\partial\underline{\boldsymbol{y}}\middle/\partial\underline{\boldsymbol{x}}\right)^T\right\}^{1/2} & \underline{\boldsymbol{0}} \\ \underline{\boldsymbol{0}} & \underline{\boldsymbol{I}}_{N-M} \end{bmatrix} \overline{\underline{\boldsymbol{W}}}^T \end{aligned} \tag{75}$$

so that the determinant of that Jacobian must be

$$\det\left(\partial \underline{y}_{ex}/\partial \underline{x}\right) = \det\left(\left\{\left(\partial \underline{y}/\partial \underline{x}\right)\left(\partial \underline{y}/\partial \underline{x}\right)^T\right\}^{1/2}\right) = \sqrt{\det\left(\left(\partial \underline{y}/\partial \underline{x}\right)\left(\partial \underline{y}/\partial \underline{x}\right)^T\right)}. \quad (76)$$

Inserting this result into (68) yields

$$p\left(\underline{y}\right) = \int d\underline{y}_* \int\limits_{\left\{\underline{x}\left|F_{ex}\left(\underline{x}\right)=\underline{y}_{ex}\right.\right\}} d\underline{x} \frac{p\left(\underline{x}\right)}{\sqrt{\det\left(\left(\partial \underline{y}/\partial \underline{x}\right)\left(\partial \underline{y}/\partial \underline{x}\right)^T\right)}}. \quad (77)$$

Now, since

$$\left\{\underline{x}\left|\underline{F}_{ex}\left(\underline{x}\right)=\underline{y}_{ex}\right.\right\} = \left\{\underline{x}\left|\underline{F}\left(\underline{x}\right)=\underline{y} \text{ and } \underline{F}_*\left(\underline{x}\right)=\underline{y}_*\right.\right\}, \quad (78)$$

the order of integration may be reversed, and the integration over $\underline{y}_*$ evaluated by inspection (since the integrand does not depend upon $\underline{y}_*$)

$$p\left(\underline{y}\right) = \int\limits_{\left\{\underline{x}\left|F\left(\underline{x}\right)=\underline{y}\right.\right\}} d\underline{x} \frac{p\left(\underline{x}\right)}{\sqrt{\det\left(\left(\partial \underline{y}/\partial \underline{x}\right)\left(\partial \underline{y}/\partial \underline{x}\right)^T\right)}} \quad (79)$$

leading to the desired result.

Note that since

$$\begin{aligned} 1 &= \int d\underline{y}\, p\left(\underline{y}\right) \\ &= \int d\underline{y} \int\limits_{\left\{\underline{x}\left|F\left(\underline{x}\right)=\underline{y}\right.\right\}} d\underline{x} \frac{p\left(\underline{x}\right)}{\sqrt{\det\left(\left(\partial \underline{y}/\partial \underline{x}\right)\left(\partial \underline{y}/\partial \underline{x}\right)^T\right)}} \\ &= \int d\underline{x}\, p\left(\underline{x}\right), \end{aligned} \quad (80)$$

the double integral must be equivalent to integrating over the full range of $\underline{x}$ values.

In the case that y is a scalar, the result takes on a particularly simple form, as shown in the following corollary.

Corollary: For $\underline{x}$ as defined in (62) and the scalar transformation $y = F\left(\underline{x}\right)$, the probability density of y is

$$p\left(y\right) = \int\limits_{\left\{\underline{x}\left|F\left(\underline{x}\right)=y\right.\right\}} d\underline{x} \frac{p\left(\underline{x}\right)}{\left|\vec{\nabla}_{\underline{x}} F\left(\underline{x}\right)\right|} \quad (81)$$

for a gradient vector (written in physical vector notation rather than linear algebra vector notation) of

$$\vec{\nabla}_{\underline{x}} F\left(\underline{x}\right) = \sum_{k=1}^{N} \frac{\partial F}{\partial x_k} \vec{u}_k. \quad (82)$$

Proof: For scalar y,

$$\partial y/\partial \underline{x} = \left[\begin{array}{ccc} \partial F/\partial x_1 & \cdots & \partial F/\partial x_N \end{array}\right] \quad (83)$$

so that

$$\sqrt{\det\left(\left(\partial y/\partial \underline{x}\right)\left(\partial y/\partial \underline{x}\right)^T\right)} = \sqrt{\left(\sum_{k=1}^{N}\left(\partial F/\partial x_k\right)^2\right)} = \left|\vec{\nabla}_{\underline{x}} F\left(\underline{x}\right)\right|. \quad (84)$$

For scalar transformations of a scalar statistic, the integrals over the ambiguous regions of the inverse transformation typically reduce to a summation over discrete points [24]. That is, for $y = f\left(x\right)$,

$$p\left(y\right) = \sum_i \frac{p\left(x_i\right)}{\left|df/dx\right|_{x_i}} \quad (85)$$

where, for any particular value y_0, the sum is taken over all points x_i that yield $f\left(x_i\right) = y_0$. However, should continuous regions of the domain all map to the same y_0, the summation must be augmented by integrals covering those regions.

ACKNOWLEDGMENT

Oak Ridge National Laboratory is managed for the United States Department of Energy by Battelle LLC under contract DE-AC05-00OR22725. The author is indebted to Dr. Michael Traweek of ONR for his continued enthusiastic support (both intellectual and financial) of this type of fundamental conceptual investigation, which appears to be of decreasing interest within the larger DoD research community.

REFERENCES

[1] H. Van Trees, *Detection, Estimation, and Modulation Theory*, 1st ed. New York, NY, USA: Wiley, 1968, pp. 23–46.
[2] C. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 623–656, Oct. 1948.
[3] S. Kullback and R. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, no. 1, pp. 79–86, Mar. 1951.
[4] R. Johnson, "Axiomatic characterization of the directed divergences and their linear combinations," *IEEE Trans. Inf. Theory*, vol. 25, no. 6, pp. 709–716, Nov. 1979.
[5] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Inf. Theory*, vol. 26, no. 1, pp. 26–37, Jan. 1980.
[6] J. Shore and R. Johnson, "Properties of cross-entropy minimization," *IEEE Trans. Inf. Theory*, vol. 27, no. 4, pp. 472–482, Jul. 1981.
[7] D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1261–1282, Apr. 2005.
[8] R. Atar and T. Weissman, "Mutual information, relative entropy, and estimation in the Poisson channel," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1302–1318, Mar. 2012.
[9] Wikipedia. (2013). *Signal-to-Noise Ratio* [Online]. Available: http://en.wikipedia.org/wiki/Signal-to-noise_ratio
[10] H. Van Trees, *Detection, Estimation, and Modulation Theory*. New York, NY, USA: Wiley, 1968, p. 99.
[11] H. Van Trees, *Detection, Estimation, and Modulation Theory*. New York, NY, USA: Wiley, 1968, p. 37.
[12] S. Kullback and R. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, no. 1, pp. 79–86, Mar. 1951.
[13] O. Johnson, *Information Theory and The Central Limit Theorem*. London, U.K.: Imperial College Press, 2004, pp. 8–9.
[14] Wikipedia. (2013). *Kullback-Leibler Divergence* [Online]. Available: http://en.wikipedia.org/wiki/Kullback-Leibler_divergence

[15] Wikipedia. (2013). *Probability Space* [Online]. Available: http://en.wikipedia.org/wiki/Probability_space

[16] S. Kullback and R. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, no. 1, pp. 79–86, Mar. 1951.

[17] P. Halmos and L. Savage, "Application of the radon-nikodym theorem to the theory of sufficient statistics," *Ann. Math. Stat.*, vol. 20, no. 2, pp. 225–241, 1949.

[18] S. Kullback and R. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, no. 1, pp. 79–86, Mar. 1951.

[19] S. Kullback and R. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, no. 1, pp. 79–86, Mar. 1951.

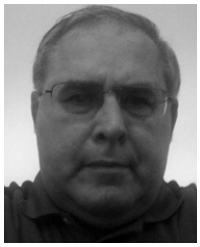
[20] Wikipedia. (2013). *Cauchy Distribution* [Online]. Available: http://en.wikipedia.org/wiki/Cauchy_distribution

[21] J. Polcari, "Closed form SNR for cauchy detection statistics," in *Proc. SAIC Working Paper*, 2012.

[22] J. Polcari, J. Gershone, and R. Perry, "Quadratic inequality notes," in *Proc. SAIC Working Paper*, 2012.

[23] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. New York, NY, USA: McGraw-Hill, 1965, pp. 201–202.

[24] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. New York, NY, USA: McGraw-Hill, 1965.



JOHN POLCARI received the B.S. degree in electrical engineering from the U.S. Naval Academy, Annapolis, MD, USA, in 1977. In 1980, he returned to post-graduate school at MIT, where he pursued master's and Engineers degrees through the Engineering Duty Naval Architecture Program, and the Sc.D. degree in oceanographic engineering through the MIT/WHOI Joint Program in 1986. His naval career spanned 23 years, including a tour at NUWC, New London, CT, USA, several scientific deployments to the Arctic, and multiple engineering and management tours in the Washington, DC, USA, area, where he was involved in developing advanced undersea and ASW systems. He was assigned to DARPA as a Program Manager prior to retirement from the U.S. Navy. Since then, he has been a Lead Scientist for AETC (subsequently acquired by SAIC) in the area of underwater acoustic signal processing. He has formulated, implemented, and tested multiple high performance detection and estimation algorithms. This paper (as well as several anticipated future follow-on papers) represents the natural outgrowth of the increasing generalization of optimal detection theory that guided those developments. He has recently joined the Staff of ORNL in the pursuit of an expanded breadth of applications.