A Modified Generalized Likelihood Test for a Class of Detection Problems

Boaz Porat
Technion-Israel Institute of Technology
Haifa 3200, Israel

Benjamin Friedlander
Dept. Elec. & Comp. Eng.
University of California
Davis, CA 95616
USA

In this paper we examine a class of detection problems in which the classical GLRT analysis does not apply. This class is characterized by two sets of parameters, where the model is linear in one set and nonlinear in the other. We show, both by theoretical arguments and by simulations, that the GLRT for this model fails to be chi-square under $H_0$. Moreover, the distribution under $H_0$ may be very 'bad', requiring large threshold values. In practice, this means that one has to set the threshold experimentally, without being able to rely on theory and without being able to predict the actual false-alarm and detection probabilities. We then propose a modification of the GLRT, which restores its asymptotic chi-square behavior under both $H_0$ and $H_1$. Since the distribution of the modified GLRT is chi-square under $H_0$, the threshold can be easily computed according to the permitted probability of false-alarm. The performance of the modified GLRT is analyzed theoretically, and illustrated by an example.

2. Problem Statement

Assume we are given a parametric family of discrete-time vector signals $s(n, \theta)$. Here $n$ is the integer time index and $\theta$ is the parameter vector, which is assumed to belong to an open subset $\Theta$ of the $M$-dimensional Euclidean space $\mathbb{R}^M$. For fixed $n$ and $\theta$, $s(n, \theta)$ is a $K$-dimensional real vector, having components $s_k(n, \theta)$, $1 \leq k \leq K$. The $s_k(n, \theta)$, regarded as functions of $\theta$, are assumed to be twice differentiable for all $\theta$ and for all $k, n$.

The noise-free received signal $x(n, \theta)$ is assumed to be a scalar linear combination of the components of $s(n, \theta)$, i.e., $x(n, \theta) = \sum_{k=1}^{K} s_k(n, \theta)$. The noisy signal is $y(n) = x(n, \theta) + w(n)$, where $w(n)$ is a discrete-time zero mean white Gaussian noise.

Assume we are given $N$ measurements $\{y(n), 0 \leq n \leq N-1\}$. Collecting these measurements in a column vector $y$ and defining $S(\theta)$, $a$ and $w$ in an obvious manner, we can write

$$y = S(\theta)a + w.$$  \hspace{1cm} (1)

The matrix $S(\theta)$ is assumed to have full rank for all $\theta \in \Theta$. 

---

This work was supported by the Army Research Office under contract no. DAAJ03-91-C-0022, sponsored by U.S. Army Communic. Electronics Command, Center for Signals Warfare.
Consider the binary detection problem:

\[ H_0: \text{The signal is absent, i.e., } y = w. \]

The variance of the noise is assumed to be known and equal to \( \sigma^2 \).

\[ H_1: \text{The signal is present and obeys the model (1), but neither } \theta \text{ nor } \alpha \text{ is known (the noise variance is still assumed to be known and equal to } \sigma^2). \]

As a concrete example, consider the rational discrete time causal and stable transfer function

\[ g(z) = \sum_{k=1}^{K} a_k z^{-k}. \]

Assume that the noise-free signal consists of the first \( N \) values of the impulse response of \( g(z) \). Define

\[ s_1(n, \theta) = \begin{cases} 0, & n < k - 1 \\ 1, & n \geq k - 1 \end{cases} \]

and

\[ s_k(n, \theta) = \begin{cases} 0, & n < k - 1 \\ s_1(n - k + 1, \theta), & n \geq k - 1 \end{cases} \]

Then the problem can obviously be described by the model (1).

A common approach to detection problems of this kind is the generalized likelihood ratio test (GLRT) [1].

Let

\[ t = \max_{\theta, \alpha} \{2 \log f_1(y) - 2 \log f_0(y)\}, \]

where \( f_0(y) \) and \( f_1(y) \) are the probability densities under \( H_0 \) and \( H_1 \) respectively. The statistic \( t \) is compared to a threshold and \( H_1 \) is accepted if it exceeds that threshold. The threshold is usually determined so as to make the probability of false-alarm fixed and equal to a pre-chosen value \( P_f \).

Kendall and Stuart [2] treated a general detection problem, which includes the above as a special case. They considered a density function which depends on a 'primary parameter' \( \theta \) and a 'nuisance parameter' \( \alpha \). Under \( H_0 \), the primary parameter has a known value \( \theta_0 \), while under \( H_1 \) its value is unknown. The value of the nuisance parameter is unknown under both \( H_0 \) and \( H_1 \). The problem (1) can be considered as a special case, with \( \alpha \) taking the role of the primary parameter, and \( \theta \) the nuisance parameter. Under \( H_0 \), the primary parameter is \( \alpha = 0 \).

The analysis in [2] assumes that \( \theta \) is estimable under both \( H_0 \) and \( H_1 \), and that \( \alpha \) is estimable under \( H_1 \). By 'estimable' we mean that their values can be estimated under the corresponding hypothesis, and the maximum likelihood (ML) estimate is asymptotically normal, with mean equal to the true value and covariance equal to the inverse of the Fisher information matrix. If these assumptions hold, the asymptotic distribution of the statistic \( t \) is chi-square with \( r \) degrees of freedom (d.o.f.), \( r \) being the dimension of \( \theta_r \). The distribution is central under \( H_0 \) and noncentral under \( H_1 \). The noncentrality under \( H_1 \) is

\[ \rho = \theta - \theta_0 \]

where \( I(\theta) \) is the Fisher information of \( \theta \) under \( H_1 \). It is interesting to note that the asymptotic result is independent of the dimension of the nuisance parameter.

When the above assumptions are violated, the analysis does not apply and the distribution of the GLRT is not necessarily as stated above. In particular, in our model (1), the parameter \( \theta \) is not estimable under \( H_0 \). Therefore, even though our problem is a special case of the one in [2], the analysis there does not apply, and the distribution under \( H_0 \) is not necessarily chi-square with \( K \) d.o.f.

In the next section we continue to explore the structure and the properties of the GLRT, and exhibit the existence of pathological cases, where the distribution under \( H_0 \) is very 'bad' compared to the \( K \)-d.o.f. chi-square.

3. The GLRT Under \( H_0 \)

To derive the GLRT for the model (1), write the probability density function of \( y \) under \( H_1 \),

\[ f_1(y) = \frac{1}{(2\pi)^{N/2} \sigma^N} \exp\left(-\frac{y^T S^T \sigma^{-2} S y}{2}\right). \]

where \( S \) is the covariance matrix of the noise. The probability density under \( H_0 \) is

\[ f_0(y) = \frac{1}{(2\pi)^{N/2} \sigma^N} \exp\left(-\frac{y^T \sigma^{-2} y}{2}\right). \]

Hence,

\[ \hat{\theta}_{ML} = \arg\min_{\theta} \left\{ y^T S (\hat{\theta}_0) - \frac{1}{2} \right\} \]

where \( \hat{\theta}_0 \) is the projection on the column space of \( S(\theta) \).

The probability density under \( H_0 \) is

\[ f_0(y) = \frac{1}{(2\pi)^{N/2} \sigma^N} \exp\left(-\frac{y^T \sigma^{-2} y}{2}\right). \]

In summary, the GLRT statistic is given by

\[ t(y) = \frac{\sigma^{-2} y^T \sigma^{-2} y}{\sigma^{-2} y^T \theta}. \]
In particular, under $H_0$,

$$t(w) = \sigma^{-2} \max_{\theta \in \Theta} \{w^T P(\theta) w\}. \quad (12)$$

While $t_0(w)$ is chi-square with $K$ d.o.f. for any specific value of $\theta$, the maximum of (12) is not chi-square in general. Moreover, in most cases the probability distribution of $t(w)$ is intractable.

The distribution of $t(w)$ is important in two respects. First, it is needed for determining the threshold to achieve a pre-assigned false-alarm rate. Second, we will derive bounds on the probability that $t(w)$ exceeds a given value $\alpha$. We will denote by $F_K(\alpha)$ the probability distribution function of a central chi-square random variable with $K$ degrees of freedom.

**Proposition 1.**

$$P\{t(w) > \alpha\} \geq 1 - F_K(\alpha), \quad \forall \alpha. \quad (13)$$

Moreover, suppose that $\Theta$ contains at least two points $\theta_1$ and $\theta_2$ such that $P(\theta_1) \neq P(\theta_2)$. Then the inequality (13) is strict.

**Proof.** Fix a point $\theta_0$ in $\Theta$. Then,

$$P\{t(w) > \alpha\} \geq P\{t_{\theta_0}(w) > \alpha\} = 1 - F_K(\alpha).$$

To prove the second part, let us show first that $t_{\theta_1}(w)$ and $t_{\theta_2}(w)$ are different with probability one. Indeed, equality holds if and only if $w^T [P(\theta_1) - P(\theta_2)] w = 0$, i.e., if and only if $w$ is in the null space of $P(\theta_1) - P(\theta_2)$. But, by the assumption, the rank of this matrix is at least one, so the dimension of the null space is no larger than $N - 1$. Since the multivariate Gaussian distribution is absolutely continuous, it follows that $P\{t_{\theta_1}(w) = t_{\theta_2}(w)\} = 0$.

Next note that, since $P(\theta_1)$ and $P(\theta_2)$ have the same rank, their difference is necessarily indefinite (i.e., it has both positive and negative eigenvalues). Therefore $P\{t_{\theta_1}(w) \leq \alpha\} \cap \{t_{\theta_2}(w) > \alpha\} > 0$ for all $\alpha > 0$. Hence

$$P\{t(w) > \alpha\} \geq P\{\max\{t_{\theta_1}(w), t_{\theta_2}(w)\} > \alpha\} = P\{t_{\theta_1}(w) > \alpha\} + P\{t_{\theta_2}(w) \leq \alpha\} \cap \{t_{\theta_2}(w) > \alpha\} > 1 - F_K(\alpha).$$

**Proposition 2.**

$$P\{t(w) > \alpha\} \leq 1 - F_N(\alpha), \quad \forall \alpha. \quad (14)$$

Furthermore, this bound is tight, i.e., there exists a family of $P(\theta)$ for which equality holds.

**Proof.** Obviously, $t_{\theta}(w) \leq \sigma^{-2} w^T w$ for all $\theta$, hence $t(w) \leq \sigma^{-2} w^T w$. Therefore,

$$P\{t(w) > \alpha\} \leq P\{\sigma^{-2} w^T w > \alpha\} = 1 - F_N(\alpha).$$

As is well known, both the interval $[0, 1]$ and the hypersphere $\{v \in \mathbb{R}^N : v^T v = 1\}$ have the cardinal number $N$, so there exists a one-to-one function from the former set to the latter. Let $\Theta = [0, 1]$, $K = 1$ and $P(\theta) = \nu(\theta) v^T$, where $\nu(\theta)$ is the above function. Then, for every $w$, there exists $\theta$ such that $\nu(\theta)$ is collinear with $w$. For this $\theta$ we have $t_{\theta}(w) = \sigma^{-2} w^T w$, so necessarily $t(w) = \sigma^{-2} w^T w$. Therefore the inequality (14) becomes an equality in this case.

It is interesting to note that (12) is the probability distribution of the energy detector $\sigma^{-2} y^T y$ under $H_0$. Therefore, the GLRT is guaranteed to perform no worse than the energy detector.

The example in Proposition 2 is admittedly unrealistic. As will be illustrated by the next example, there are realistic problems in which the probability $P\{t(w) > \alpha\}$ is much larger than the chi-square function $1 - F_K(\alpha)$.

Let $\mathbf{F}$ be the unitary DFT matrix, i.e., $\mathbf{F}_{k\ell} = N^{-1/2} e^{-j\ell k \pi / N}$, $0 \leq k, \ell \leq N - 1$. Separate $\mathbf{F}$ into its real and imaginary components $\mathbf{F} = \mathbf{F}_R + j \mathbf{F}_I$. Let

$$\tilde{\mathbf{F}} = \begin{bmatrix} \mathbf{F}_R & \mathbf{F}_I \\ -\mathbf{F}_I & \mathbf{F}_R \end{bmatrix}.$$  

(15)

Note that $\tilde{\mathbf{F}}$ is a real orthonormal matrix. Let $s(\theta)$ be the $N$-dimensional column vector whose $k$th entry is $\sin(k\theta)$ and $c(\theta)$ the $N$-dimensional vector whose $k$th entry is $\cos(k\theta)$. Also let

$$\tilde{\mathbf{S}}(\theta) = \frac{1}{\sqrt{N}} \begin{bmatrix} c(\theta) & s(\theta) \\ -s(\theta) & c(\theta) \end{bmatrix}; \quad \mathbf{S}(\theta) = \tilde{\mathbf{F}}^T \tilde{\mathbf{S}}(\theta). \quad (16)$$

Note that $\tilde{\mathbf{S}}(\theta)$ satisfies $\tilde{\mathbf{S}}^T(\theta) \mathbf{S}(\theta) = I$, so $\mathbf{S}(\theta) \tilde{\mathbf{S}}^T(\theta)$ is the projection on the column space of $\mathbf{S}(\theta)$. The two columns of $\tilde{\mathbf{S}}$ are 'generic' columns of $\tilde{\mathbf{F}}$. In particular, if we define $\theta_i = 2\pi i / N$, $1 \leq i \leq N - 1$, then $\mathbf{S}(\theta_i) = [e_i, e_{N+i}]$, where $e_i$ denotes a $2N$-dimensional unit vector with 1 in the $i$th position and zeros elsewhere.

Consider the detection problem with $K = 2$ and $\mathbf{S}(\theta)$ as above, and take $\sigma^2 = 1$ for simplicity. Let $w$ be a $2N$-dimensional Gaussian vector. Then

$$t_\theta(w) = u_{\theta}^w + w_{N+i}, \quad (17)$$

where $u_{\theta}^w$ is a random variable with a $\mathcal{N}(0, \theta)$ distribution, and $w_{N+i}$ is a standard normal random variable.

438
Theorem 3.2 \[ P(t(w) > \alpha) \geq P(\max_{i} t_i(w) > \alpha) = 1 - F_{2N}(\alpha)^N. \] 

As we see, while \( P(t(w) > \alpha) \) is not as high as \( 1 - F_Z(\alpha) \), it is still much higher than \( 1 - F_Z(\alpha) \) if \( N \) is large. We should note that using \( F \) in the definition of \( S(\theta) \) was only for simplifying the derivation of (18). The same result would have been obtained with \( S(\theta) \).

4. Modification of the GLRT

A partial remedy to the shortcomings of the GLRT is obtained by the detector discussed below. The main idea is to split the measurements into two sets. One set is used to estimate \( \theta \) and \( a \) by maximum likelihood, and the other for the test statistic, using the estimate from the first set. We have chosen the two sets as the even- and odd-numbered measurements respectively. The test statistic is essentially a linearized GLRT, where the linearization of the model is about the ML estimates \( \hat{\theta} \) and \( \hat{a} \).

Let \( y_1 \) be the vector consisting of the measurements at odd time points and \( y_2 \) the vector consisting of the measurements at even time points. Then we can write, similarly to (1),

\[ y_1 = S_1(\theta)a + w_1 \quad ; \quad y_2 = S_2(\theta)a + w_2. \] \hspace{1cm} (19)

Let \( \hat{\theta} \) and \( \hat{a} \) be the ML estimates of \( \theta \) and \( a \) using the data \( y_2 \). Define

\[ b_m = \frac{\partial S(\theta)}{\partial \theta_m}, \quad B(\theta) = [b_1, b_2, \ldots, b_M] \] \hspace{1cm} (20)

\[ C(\theta) = [S(\theta), B(\theta)], \quad x = \begin{bmatrix} a - \hat{a} \\ \theta - \hat{\theta} \end{bmatrix} \] \hspace{1cm} (21)

The notations \( B_1(\theta, a), B_2(\theta, a), C_1(\theta, a) \) and \( C_2(\theta, a) \) are defined in a similar manner. All the C matrices are assumed to have full rank almost everywhere (i.e., except on a set of \( \{ \theta, a \} \) of measure zero).

Since \( S_1(\theta)a \) is twice differentiable with respect to both \( \theta \) and \( a \), we get upon linearization of \( S_1(\theta)a \) around the ML estimates,

\[ y_1 = S_1(\hat{\theta})\hat{a} + C_1(\hat{\theta}, \hat{a})x + O(||x||^2) + w_1. \] \hspace{1cm} (22)

Neglecting the term \( O(||x||^2) \) and proceeding as in Section 3, we can easily derive the GLRT for the linearized model (22). The result is

\[ t(y) = \sigma^{-2}y^T P_{C_1}(\hat{\theta}, \hat{a})y_1, \] \hspace{1cm} (23)

where \( P_{C_1}(\hat{\theta}, \hat{a}) \) denotes the projection on the column space of \( C_1(\hat{\theta}, \hat{a}) \).

Consider now what happens under \( H_0 \). Since \( P_{C_1}(\hat{\theta}, \hat{a}) \) depends only on \( w_2 \), and since \( w_2 \) is independent of \( w_1 \), the conditional distribution of \( t(y) \) given \( w_2 \) is chi-square with \( K + M \) d.o.f., for all values of \( w_2 \) (we have used here the fact that the rank of this projection will be \( K + M \) with probability one). In particular, the 'garbage estimates' \( \hat{\theta} \) and \( \hat{a} \) have no effect on the conditional distribution of \( t(y) \). Therefore, the unconditional distribution of \( t(y) \) will be also chi-square with \( K + M \) d.o.f.

This technique of using two independent vectors, related to the same underlying parameters, is a special case of cross-validation, a known method in statistics. In essence, the projection of \( y_1 \) on the space constructed from the estimates based on \( y_2 \) attempts to validate that \( y_1 \) and \( y_2 \) are governed by the same parametric model. Since under \( H_0 \) they are not (being pure white noise), this cross-validation is likely to fail.

While the distribution of \( t(y) \) under \( H_0 \) is exactly chi-square, its exact form under \( H_1 \) is not computable in general, due to the nonlinear dependence of \( \theta \) and \( a \) on the data vector \( y_2 \). We will therefore derive an approximation which improves as the noise variance \( \sigma^2 \) approaches zero. We stress that the number of data points remains constant throughout. Therefore, the expression "asymptotic" means that \( \sigma^2 \to 0 \), not that \( N \to \infty \).

Under the assumptions that \( S_2(\theta) \) is twice differentiable and \( C_2(\theta, a) \) is nonsingular, it is easy to derive the result

\[ x = (C_2^T (\theta, a)C_2(\theta, a))^{-1}C_2^T (\theta, a)w_2 + O(||w_2||^2). \] \hspace{1cm} (24)

Now, \( w_2 = O_p(\sigma) \), hence \( x = O_p(\sigma) \), and \( O(||x||^2) = O_p(\sigma^2) \). Recall (22) and multiply it by \( P_{C_1}(\hat{\theta}, \hat{a}) \) to get

\[ P_{C_1}(\hat{\theta}, \hat{a})y_1 = P_{C_1}(\hat{\theta}, \hat{a})[S_1(\hat{\theta})\hat{a} + C_1(\hat{\theta}, \hat{a})x + O_p(\sigma^2) + w_1] = S_1(\hat{\theta})\hat{a} + C_1(\hat{\theta}, \hat{a})x + O_p(\sigma^2) + P_{C_1}(\hat{\theta}, \hat{a})w_1 = S_1(\theta)a + O_p(\sigma^2) + P_{C_1}(\hat{\theta}, \hat{a})w_1. \] \hspace{1cm} (25)

Also,

\[ P_{C_1}(\hat{\theta}, \hat{a})w_1 = P_{C_1}(\theta, a)w_1 + O_p(\sigma^2). \] \hspace{1cm} (26)

Therefore

\[ \sigma^{-1}P_{C_1}(\hat{\theta}, \hat{a})y_1 = \sigma^{-1}[S_1(\theta)a + P_{C_1}(\theta, a)w_1] + O_p(\sigma). \] \hspace{1cm} (27)
Rewrite this as

\[ \xi_2 = \xi_1 + O_p(\sigma) \quad (28) \]

where the definition of \( \xi_1 \) and \( \xi_2 \) is clear from (27). The vector \( \xi_1 \) is Gaussian with mean \( \sigma^{-1}S_1(\theta)a \) and standard deviation \( O(1) \). The vector \( \xi_2 \) is non Gaussian, but its difference from \( \xi_1 \) is \( O_p(1) \). We cannot conclude that \( S \) converges to \( \xi_1 \) in distribution, since neither converges to anything (because the means go to infinity as \( \sigma \to 0 \)). Nevertheless, we can conclude from (28) that

\[ ||\xi_2|| - ||\xi_1|| = O_p(\sigma). \quad (29) \]

The statistic \( ||\xi_2|| \) is precisely the square-root of the modified GLRT, while \( ||\xi_1|| \) is exactly noncentral chi with \( K+M \) d.o.f. and noncentrality \( \sigma^{-1}||S_1(\theta)a|| \). As we see, while the mean of each increases in inverse proportion to \( \sigma \), the difference between the two approaches zero in probability as \( \sigma \to 0 \).

Summing up, the square-root of the modified GLRT is exactly central chi under \( H_0 \), and approximately noncentral chi under \( H_1 \). The error in the approximation decreases to zero in probability as \( \sigma \to 0 \). The number of d.o.f. is \( K+M \), which is higher than the 'ideal' number \( K \), corresponding to cases where the analysis of [2] applies. The effective signal to noise ratio is \( \sigma^{-2}||S_1(\theta)a||^2, \) which is about one half the SNR of the complete data vector \( y \). The loss in using the modified GLRT (compared to the ideal case) is therefore about 3dB of SNR and an increase in the number of d.o.f. from \( K \) to \( K+M \). The gain in using it should be clear from Section 2: without it there would be no control over the distribution under \( H_0 \), and in extreme cases it could be much worse than \( K \)- or \( (K+M) \)-d.o.f. chi-square.

A slight improvement of the detection performance can be obtained using the following heuristic procedure. Rename \( t(y) \) in (23) as \( t_1(y) \) and reverse the roles of \( y_1 \) and \( y_2 \) to get a second statistic

\[ t_2(y) = \sigma^{-2}y_2^TP_{C_2}(\hat{\theta}, \hat{a})y_2. \quad (30) \]

Then use

\[ t_3(y) = \min \{t_1(y), t_2(y)\} \quad (31) \]

as the final test statistic. While \( t_1(y) \) and \( t_2(y) \) are not independent in theory, they may be nearly independent in practice, i.e.,

\[ P\{\min \{t_1(y), t_2(y)\} > \alpha\} \approx P\{t_1(y) > \alpha\}P\{t_2(y) > \alpha\}. \quad (32) \]

If this approximation holds (a condition that can be checked in any particular application by Monte-Carlo simulation), one can use the distribution function \( 1 - (1 - F_{K+M}(\alpha))^2 \) for threshold determination. In the next section we give an example in which part of the 'lost 3dB' can be gained back this way.

A final remark concerns the choice of the projection matrix in (23). One might ask whether it would not be better to use \( P_{S_1}(\hat{\theta}, \hat{a}) \) in lieu of \( P_{C_2}(\hat{\theta}, \hat{a}) \). This would indeed decrease the number of d.o.f. from \( K+M \) to \( K \) under \( H_0 \), but may adversely affect the distribution under \( H_1 \). More precisely, the difference between \( \xi_1 \) and \( \xi_2 \) will be \( O(1) \), which does not improve when \( \sigma \to 0 \). Consequently, the difference between \( ||\xi_2|| \) and \( ||\xi_1|| \) will not go to zero as \( \sigma \to 0 \).

In [6] we present a numerical example which verifies the fact that the modified GLRT performs according to the theoretical analysis, and the use of the minimum statistic \( t_3(y) \) improves it as predicted. The overall difference in performance between the two detectors (as measured by the receiver operating characteristics) was marginal in this example. Nevertheless, the theoretical arguments given above indicate that the difference in performance may be greater in more complex problems.

References


