TRANSIENT DETECTION USING A HOMOGENEITY TEST

Biao Chen, Peter Willett

U-157, University of Connecticut
Storrs, CT 06269-3157

Roy Streit

Code 202, Naval Undersea Warfare Center
Newport, RI 02841

Abstract

A simple yet effective statistic is proposed for detecting transient buried in partially unknown ambient noise. The transient model is the frequency scattered increased variance observations. We pose the transient detection problem as homogeneity test and the statistic is derived as the (generalized) likelihood ratio test of overdispersion when the underlying observation sequence follows a double exponential distribution. Numerical testing focuses on the comparison of this scheme with the CFAR power-law detector.

1. INTRODUCTION

A recently proposed power-law detector [1] has attracted much attention in addressing transient detection especially for passive underwater acoustic application. The problem setting is to assume that the frequency domain observation sequence follows a unity-exponential distribution under the noise only hypothesis, while if a transient is present, some observations have an increased mean value. Posed as a standard hypothesis testing problem, we have

$$\begin{aligned} \mathbf{H}: \quad f(\mathbf{x}) &= \prod_{i=0}^{n-1} \frac{1}{\mu_0} e^{-x_i/\mu_0} \\ \mathbf{K}: \quad f(\mathbf{x}) &= \prod_{i \notin \mathcal{S}} \frac{1}{\mu_0} e^{-x_i/\mu_0} \prod_{i \in \mathcal{S}} \frac{1}{\mu_1} e^{-x_i/\mu_1} \quad (1) \end{aligned}$$

in which $\mathcal{S}$ denotes that subset of $\{0, 1, \ldots, n-1\}$ which correspond to "signal-containing" observations and $\mu_1 = \mu_0(1 + snr/M)$ where snr is the aggregate SNR of the transient signal. Such a model arises naturally when we take the magnitude-square FFT on time sequences which assume a normal distribution. Assuming appropriate normalization, the unity-exponential bins correspond to those signal absent frequency band while those with increased scale, which correspond to those observations with increased variance, suggest the presence of transient signal.

Nuttall's power-law statistic is defined as

$$T_{pl}(\mathbf{x}) = \sum_{i=0}^{n-1} x_i^{\nu} \quad (2)$$

in which x_i denotes the magnitude-square of the i^{th} FFT bin, and ν is an adjustable exponent. Assuming the cardinality of $\mathcal{S}$ in (1) is M and further that the M signal containing bins are equally likely distributed among the total FFT bins, Nuttall derived his power-law statistic as an approximation to the optimal likelihood ratio test which is otherwise computationally infeasible. A remarkable feature is that the performance of power-law detector does not depend on M very much. In fact, it is found through extensive numerical work that for a wide range of M, power-law detector with $2 < \nu < 3$ has near-optimal performance and is hard to improve upon. However, its reliance on the known μ_0 is too demanding in many applications. As a remedy, Nuttall in [2] has proposed a constant false-alarm rate (CFAR) version of (2):

$$T_{cpl}(\mathbf{x}) = \frac{\left(\sum_{i=0}^{n-1} x_i^{\nu}\right)}{\left(\sum_{i=0}^{n-1} x_i\right)^{\nu}} \quad (3)$$

Our task here is to develop a new statistic that is competitive with, or even better than, the CFAR power-law. The approach we take is totally different from that of Nuttal. The model in (1) suggests that under H all data follow an *i.i.d.* exponential distribution, while under K part of the observations follow a different exponential distribution. Therefore we could consider the transient detection problem as test of homogeneity: whether or not the whole data set comes from a single exponential distribution. This formulation of the problem is in fact more consistent with the practical situation we are likely to encounter. The transient signal, if present, does not necessarily have equal strength among all the frequency bins it occupies (consider, for example, the case that more than one transient appears at the same time).

This research was supported by ONR through NUWC Newport under contracts N66604-97-M-3139, N00014-98-1-0049, and N00014-98-WX30115.

The statistic for homogeneity testing is derived using the concept of the double exponential family, also called the overdispersed model. A special property about the exponential distribution is that its mean and variance are functionally related to each other. In fact, this is true for a whole class of distributions called "one-parameter exponential families". Other examples include Poisson and Binomial. From a statistical standpoint, this functional relationship between mean and variance incurs some restriction in data modeling; when the data set exhibits some heterogeneity, a one-parameter exponentially family does not provide enough flexibility. A common remedy is to use mixture models; but in this case the implementation is usually restricted to numerical approaches, such as the EM algorithm [3, 4].

Recognizing such a lack of flexibility in addressing heterogeneity, there has been considerable development of the so-called double exponential families (DEF). DEF generalize the one-parameter exponential family by introducing a second parameter that controls its variance independent of its mean – the overdispersion parameter, which leads to another name, the so-called overdispersed model, i.e., it is the overdispersed version of the corresponding one-parameter exponential family. An advantage of using DEF over the mixture model is that it remains a member of the exponential family for which the inference tools are both rich and mature.

The original motivation for double exponential families was to use them as constituent distributions in generalized linear regressions for increased modeling flexibility. Its application to transient detection is exploited in this paper. The idea, as implied above, is to consider the signal-absent observations as the "no overdispersion" case since they all come from the same exponential distribution; while for the signal-present observations, a mixture of two or more exponential distributions causes nonhomogeneity in the data set, which corresponds to the overdispersed case. Thus we have transformed the transient detection problem to one of testing the overdispersion (or from the opposite point of view, testing the homogeneity) of the data set.

2. THE DERIVED TEST STATISTIC

2.1. Overdispersed Model

Supposing the original one-parameter exponential family is of the form $f_\mu(x)$, then its double exponential correspondent has the form [5]

$$\tilde{f}_{\mu,\theta}(x) = c(\mu,\theta)\theta^{1/2}\{f_\mu(x)\}^\theta\{f_x(x)\}^{1-\theta} \tag{4}$$

where θ is the so-called overdispersion parameter. The notation $f_x(x)$ refers to the density function with μ replaced by x. The normalizing constant $c(\mu,\theta)$ is introduced to make the integration equal unity so that $\tilde{f}$ is a legitimate density function. Setting $\theta = 1$ reduces the DEF to its one-parameter exponential counterpart; while $\theta < 1$ corresponds to the "overdispersed" case. It is shown in Efron's paper the variance of the DEF equals σ^2/θ where σ^2 is the variance of the original one-parameter exponential family, meaning the observation exhibits heterogeneity (larger variance) when $\theta < 1$.

A nice property about this double-exponential family is that the normalizing constant is close to unity. For example, when the original one-parameter exponential family is the $exp(\mu)$, it was shown using the Edgeworth expansion that $c(\mu,\theta)$ can be approximated by

$$c(\mu,\theta) = 1 - \frac{1}{12}\frac{1-\theta}{\theta}$$

Thus for non-extreme value of θ, e.g., $\theta > 1/2$, $c(\mu,\theta)$ can be reasonably approximated by unity. This allows us to drop out this normalizing term when doing inference; otherwise, the lack of explicit form of $c(\mu,\theta)$ would exacerbate the difficulty in the likelihood based inference.

An equivalent yet more useful form of (4) using Hoeffding's representation [5] is

$$\tilde{f}_{\mu,\theta}(x) = c(\mu,\theta)\theta^{1/2} f_x(x)\exp[-\theta I(x,\mu)] \tag{5}$$

where $I(y,\mu)$ is the usual Kullback-Leibler distance with parameters y and μ respectively, i.e.,

$$I(\mu_1,\mu_2) = E_{\mu_1}\{\log[f_{\mu_1}(x)/f_{\mu_2}(x)]\} \tag{6}$$

From (4) and (5), dropping the normalizing constant, it is straightforward to derive the MLEs from n observations $x_1,\cdots,x_n$, all assumed to follow the distribution (4), as

$$\hat{\mu} = \bar{x} \tag{7}$$

$$\hat{\theta} = \frac{n}{2\sum_{i=1}^n I(x_i,\bar{x})} \tag{8}$$

2.2. The New Statsitic

As mentioned earlier, we are interested in the test of whether the observations come from a single exponential distribution $exp(\mu)$ where μ is the unknown ambient noise level; or whether they come from the overdispersed model $\tilde{f}_{\mu,\theta}(x)$ with $\theta < 1$. Equivalently, using the fact that $exp(\mu)$ is a special case of the overdispersed model with $\theta = 1$, our goal is to test, given $x_i \sim \tilde{f}_{\mu,\theta}(x_i)$, the following hypothesis regarding θ

$$\begin{array}{ll}\mathbf{H}: & \theta = 1\\ \mathbf{K}: & \theta < 1\end{array} \tag{9}$$

The GLRT for this test utilizes the MLE for θ as obtained in (8). Thus the decision rule is

$$\hat{\theta} \quad \underset{H}{\overset{K}{\lessgtr}} \quad \tau \tag{10}$$

where τ is a constant smaller than one which satisfies $P_H(\hat{\theta} \geq \tau) \leq \alpha$ with α the prefixed false alarm rate. To obtain an explicit test structure with the DFT-square samples, we need to evaluate the Kullback-Leibler number for the exponential distribution. We get

$$I(\mu_1, \mu_2) = \log \frac{\mu_2}{\mu_1} - 1 + \frac{\mu_1}{\mu_2}$$

Substituting $\mu_1 = x_i$ and $\mu_2 = \bar{x}$ and inserting to (8), we obtain, after some cleaning-up,

$$\hat{\theta} = \frac{n}{2\sum_{i=1}^{n} \log[\bar{x}/x_i]} \tag{11}$$

The test (10) is thus equivalent to a threshold comparison of the statistic

$$T_{mvd} = \frac{1}{n}\sum_{i=1}^{n} I(y_i, \hat{\mu}) = \frac{1}{n}\sum_{i=1}^{n} \log[\bar{x}/x_i] \quad \underset{H}{\overset{K}{\lessgtr}} \quad h \tag{12}$$

That this be the generalized likelihood ratio test is in fact a direct result of the following observation: we recognize (5) as an exponential family (hence have monotone likelihood ratio) for θ with natural statistic $I(x, \mu)$, assuming μ is known. Thus the most powerful test amounts to the threshold testing of

$$\sum_{i=1}^{n} I(x_i, \mu) = \sum \log \frac{\mu}{x_i} - n + \frac{\sum_{i=1}^{n} x_i}{\mu}$$

Substituting μ by $\bar{x}$ results in the generalized likelihood ratio test of H vs K which is identical to that of (12).

Throughout this paper, we refer to the test statistic as one of "mean-value deviance"(MVD). It quantifies in a certain way the average deviation of observations from its mean value. Unusually large value of $T_{mvd}(\mathbf{x})$ implies such deviation is too great to be explained using a simple exponential distribution model.

2.3. Properties of the New Statistic

Property 1 $T_{mvd}(\mathbf{x})$ is nonnegative.

Proof: $-T_{mvd}(\mathbf{x}) = \frac{1}{n}\sum_{i=0}^{n-1} \log \frac{x_i}{\bar{x}} \leq \frac{1}{n}\sum_{i=0}^{n-1} (\frac{x_i}{\bar{x}} - 1) = 0$

Property 2 It is CFAR with respect to μ_0 in (1).
Proof: Recoganizing that μ is a scale parameter for $f(x_i)$, thus $x_i/\bar{x}$, hence $T_{mvd}(\mathbf{x})$, is an ancillary statistic, meaning its distribution is free of μ [6].

Property 3 It is sensitive to small outliers, that is, observations with extremely small values.

This is so as the contribution to the test statistic from each single observation x_i is $\log(\bar{x}/x_i)$. This differs from the CFAR power-law statistic that is more sensitive to large outliers.

3. COMPARISON WITH CFAR POWER-LAW DETECTOR

In this section we give examples to illustrate the advantages and disadvantages of the MVD statistic. Specifically, our statistic is compared with the CFAR power-law detector with different ν's. The aggregate $SNR = 50$ and the FFT bin number is 128.

- CFAR power-law detector.

  The results are plotted in figure 1. Clearly for $M = 5$, CFAR power-law detectors with $\nu = 2, 3$ are superior to the new detector; while for $M = 40$ and $M = 100$, the new detector exhibits significant performance gain over the CFAR power-law detector. This is consistent with the property 3 of the MVD statistic. For small M, the power-law detector with reasonably large ν can easily pick up these large outliers On the other hand, with M large, most observations tend to have a large value and the new detector is more responsive to the small fraction (corresponding to noise only bins) of observations which have relatively smaller value.

- CFAR power-law detector with non-Gaussian noise.

  In figure 2 we repeat the above experiments for non-Gaussian noise. Correspondingly the magnitude-square DFT outputs no longer have an exponential distribution; in fact we assume they are actually squares of ones that follow an exponential distribution. That is

  $$f_\mu(x_i) \quad = \quad \frac{1}{2\mu\sqrt{x_i}} e^{-\sqrt{x_i}/\mu} \tag{13}$$

  meaning that the underlying bin-level random variables are "heavy-tailed". It is apparent that the new detector based on T_{mvd} is significantly more robust to such observations than is the power-law statistic. The reason that the T_{mvd} is more robust to heavy-tailed ambient is again a direct result of the property 3, that is, it is not as sensitive to large value outliers as the (CFAR) power-law statistic.

4. CONCLUSION

In this paper we propose a statistic for testing homogeneity of a data set which assumes exponential distribution under the null hypothesis. We derive this as the generalized likelihood ratio test of the "overdispersion" in a double exponentially distributed data set. Some properties of this statistic relevant to the transient detection task are summarized. Its performance evaluation is done numerically and is compared to that of the power-law detector.

5. REFERENCES

[1] A. Nuttall, "Detection Performance of Power-Law Processors for Random Signals of Unknown Location, Structure, Extent, and Strength", *NUWC-NPT Technical Report 10,751*, September, 1994.

[2] A. Nuttall, "Performance of Power-Law Processors with Normalization for Random Signals of Unknown Structure", *NUWC-NPT Technical Report 10,760*, May, 1997.

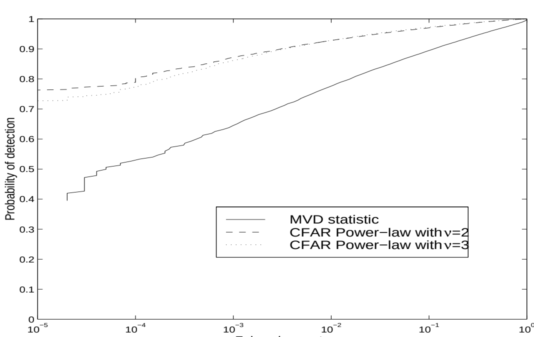
[3] R. Redner, H. Walker "Mixture Densities, Maximun Likelihood, and the EM Algorithm", *SIAM Review*, Vol. 26, No. 2, pp195-239, 1984.

[4] R. L. Streit, P. Willett, "Hybrid Bayes/GLRT Detection of Transient Signals", *Proceedings of the CISS'97*, March 1997.

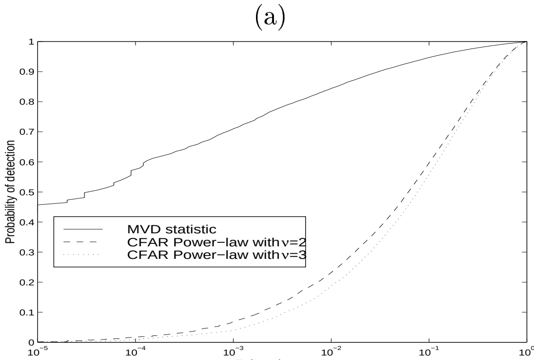
[5] B. Efron, "Double Exponential Families and Their Use in Generalized Linear Regression", *Journal of American Statistical Association*, Vol. 81, 1986.

[6] G. Casella, R. L. Berger, *Statistical Inference*, Duxbury Press, 1990.

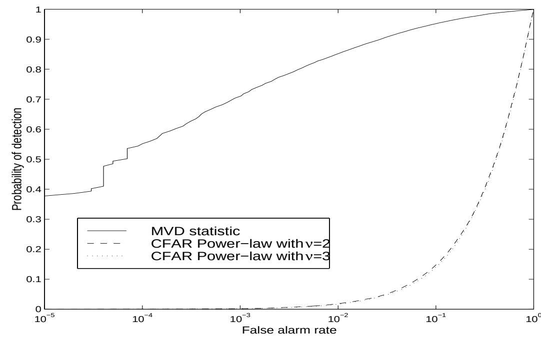
[7] E. L. Lehman, *Testing Statistical Hypotheses*, Wiley, 1986.



(a)

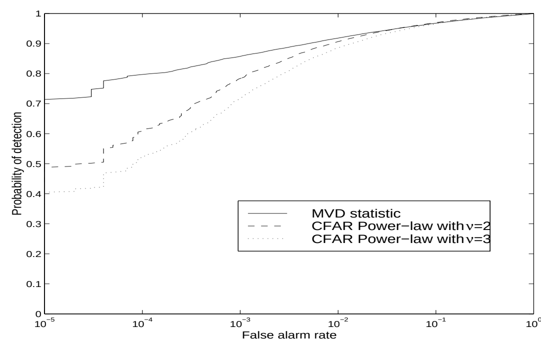


(b)

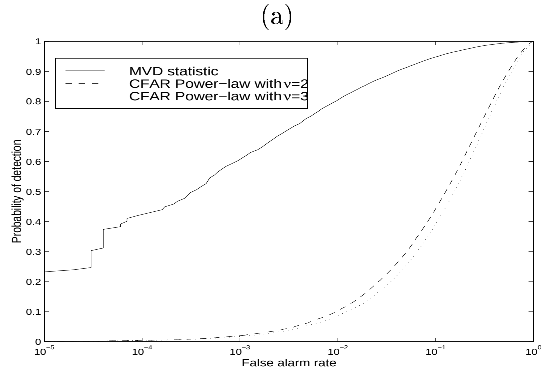


(c)

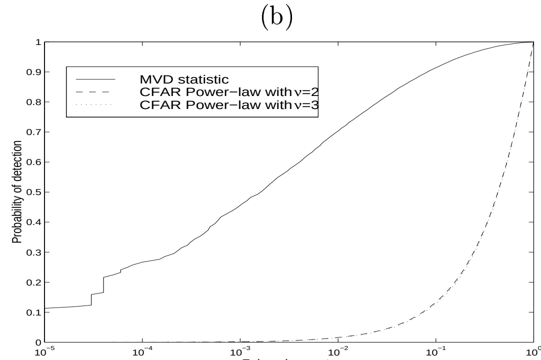
Figure 1: ROC curves of T_{mvd} and T_{cpl} with aggregate SNR=50. The number of signal containing bins is $M = 5, 40, 100$ respectively for plots (a), (b) and (c).



(a)



(b)



(c)

Figure 2: The same as in figure 1 except it is for the heavy-tailed model of (13).