New Technique for Finding Needles in Haystacks: Geometric Approach to Distinguishing between a New Source and Random Fluctuations

Ramani S. Pilla,[1,*] Catherine Loader,[1] and Cyrus C. Taylor[2]

[1]*Department of Statistics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106, USA*
[2] *Department of Physics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106, USA*

(Received 26 May 2005; published 1 December 2005)

We propose a new test statistic based on a score process for determining the statistical significance of a putative signal that may be a small perturbation to a noisy experimental background. We derive the reference distribution for this score test statistic; it has an elegant geometrical interpretation as well as broad applicability. We illustrate the technique in the context of a model problem from high-energy particle physics. Monte Carlo experimental results confirm that the score test results in a significantly improved rate of signal detection.

DOI: 10.1103/PhysRevLett.95.230202 PACS numbers: 02.50.Sk, 02.50.Tt, 07.05.Kf

One of the fundamental problems in the analysis of experimental data is determining the statistical significance of a putative signal. Such a problem can be cast in terms of classical "hypothesis testing," where a null hypothesis $\mathcal{H}_0$ describes the background and an alternative hypothesis $\mathcal{H}_1$ characterizes the signal together with the background. A test statistic (a function of the data) is used to decide whether to reject $\mathcal{H}_0$ and conclude that a signal is present.

The hypothesis test concludes that a signal is present whenever the test statistic falls in a critical region W. One is interested in the probability that a signal is found under two scenarios. First, when the null hypothesis $\mathcal{H}_0$ is true, the *significance level* α is the probability of incorrectly concluding that a signal is present. Second, when the alternative $\mathcal{H}_1$ is true, the *power* of the test is the probability that the signal is found. The goal is to construct a test statistic whose asymptotic distribution (reference distribution under $\mathcal{H}_0$ for large sample size n) can be calibrated accurately and that the associated test has high power at a fixed α, such as $\alpha = 0.01$.

When the two hypotheses are distinct, a powerful technique based on the likelihood ratio test (LRT) is often used. Suppose $p(x;\boldsymbol{\theta})$ is a probability density function for a measurement x with a parameter vector $\boldsymbol{\theta} \in \Theta \subset \mathcal{R}^d$. The joint probability density function evaluated with n measurements $\mathbf{X}$ for an unknown $\boldsymbol{\theta}$ is the likelihood function [1] $L(\boldsymbol{\theta}|\mathbf{X})$. An effective approach to the problem of choosing between $\mathcal{H}_0$ [corresponding likelihood $L(\boldsymbol{\theta}_0|\mathbf{X})$] and $\mathcal{H}_1$ [with a likelihood $L(\boldsymbol{\theta}_1|\mathbf{X})$] for explaining the data is to consider the LRT statistic

$$\Lambda = \frac{L(\hat{\boldsymbol{\theta}}_1|\mathbf{X})}{L(\hat{\boldsymbol{\theta}}_0|\mathbf{X})},$$

where $\hat{\boldsymbol{\theta}}$ is the value of $\boldsymbol{\theta}$ that maximizes $L(\boldsymbol{\theta}|\mathbf{X})$ [1–3]. To employ the LRT, the parsimonious model under $\mathcal{H}_0$ (with s_0 parameters) must be nested within the more complicated alternative model under $\mathcal{H}_1$ (with s_1 parameters). For simple models, under regularity conditions, $2\log(\Lambda)$ has an asymptotic (i.e., $n \to \infty$) χ^2 distribution with $(s_1 - s_0)$ degrees of freedom under $\mathcal{H}_0$ [1].

When the alternative hypothesis corresponds to a signal which is a perturbation of the background, regularity conditions required for this large-sample asymptotic theory are violated, since (a) some of the parameters under $\mathcal{H}_0$ are on the boundaries of their region of support and (b) different parameter values give rise to the same null model. As a result, the LRT has lacked an analytically tractable reference distribution required to calibrate a test statistic. Such a difficulty occurs in many practical applications, for example, when testing for a new particle resonance of an unknown production cross section, since the signal strength must be nonnegative. Hence, the LRT must be employed cautiously; however, it has been employed in several problems of practical importance where certain required regularity conditions are violated [2]. Misapplication of the LRT statistics can lead to incorrect scientific conclusions [4,5].

Because of above difficulties with the LRT, a χ^2 goodness-of-fit test is commonly employed. However, it typically has less power than might be hoped for as it does not take into account information about the anticipated form of the signal. We propose a new test statistic (closely related to the LRT for sufficiently large n) based on a *score process* to detect the presence of a signal and present its reference distribution.

Consider the model

$$p(x;\eta,\boldsymbol{\theta}) = (1-\eta)f(x) + \eta\psi(x;\boldsymbol{\theta}),$$

where $f(x)$ is a specified *null density* and $\psi(x,\boldsymbol{\theta})$ is a *perturbation density*. The parameter vector $\boldsymbol{\theta}$ is the "location" of the perturbation, and $\eta \in [0,1]$ measures the "strength" of the perturbation. The null hypothesis of no signal ($\mathcal{H}_0$: $\eta = 0$) implies η is on the

boundary; scenario (a) applies. Under $\mathcal{H}_0$, $p(x;0,\boldsymbol{\theta}) = f(x)$ for all x independently of $\boldsymbol{\theta}$; scenario (b) is also applicable.

Consider a search for a new particle resonance. One may measure the frequency of events as a function of energy e, modeling it by $p(e;\eta,E_0)$, where $f(e)$ characterizes the background density and

$$\psi(e;E_0) = \frac{\Gamma}{2\pi}\frac{1}{[(e-E_0)^2 + (\Gamma/2)^2]}$$

is the Cauchy (Breit-Wigner) density describing a resonance centered on E_0 with full width at half maximum Γ. We consider the two-dimensional $\boldsymbol{\theta} = (E_0,\Gamma)$ and for a fixed Γ, a one-dimensional $\boldsymbol{\theta} = E_0$.

A key obstacle to detecting the signal is finding the tail probability for test statistics, enabling valid statistical inference. We provide an asymptotic solution to this problem via a geometric formula [see Eq. (3)]. The relative improvement of the score test over the χ^2 goodness-of-fit test is particularly salient when the signal is hard to detect (Fig. 4). The development of the reference distribution and a flexible computational method will enable making probabilistic statements to solving some of the fundamental problems arising in many experimental physics.

Pilla and Loader [6] have developed a general theory and a computationally flexible method to determine the asymptotic reference distribution of a test statistic under $\mathcal{H}_0$. Their method is based on the "score process," indexed by the parameter vector $\boldsymbol{\theta}$ and defined as

$$\mathrm{S}(\boldsymbol{\theta}) := \frac{\partial}{\partial\eta}\log\left[\prod_{i=1}^{n} p(E_i;\eta,\boldsymbol{\theta})\right]\Bigg|_{\eta=0}$$

for a given data $\mathbf{E} = (E_1,\ldots,E_n)$. Under $\mathcal{H}_0$, the expectation of $\mathrm{S}(\boldsymbol{\theta})$ is 0 for all $\boldsymbol{\theta}$, while under $\mathcal{H}_1$ it has a peak at the true value of $\boldsymbol{\theta}$. Hence, the statistic $\mathrm{S}(\boldsymbol{\theta})$ is sensitive to the signal of interest. The random variability of $\mathrm{S}(\boldsymbol{\theta})$ can exhibit significant dependence on the parameter vector $\boldsymbol{\theta}$; hence, we consider the *normalized score process* defined as

$$\mathrm{S}^{\star}(\boldsymbol{\theta}) := \frac{\mathrm{S}(\boldsymbol{\theta})}{\sqrt{nC(\boldsymbol{\theta},\boldsymbol{\theta})}} \quad \text{for } \boldsymbol{\theta}\in\Theta\subset\mathcal{R}^d, \tag{1}$$

where n is the total number of events observed, and

$$C(\boldsymbol{\theta},\boldsymbol{\theta}^{\dagger}) = \int_{-\infty}^{\infty}\frac{\psi(x;\boldsymbol{\theta})\psi(x;\boldsymbol{\theta}^{\dagger})}{f(x)}dx - 1 \tag{2}$$

is the covariance function of $\mathrm{S}(\boldsymbol{\theta})$ for $\boldsymbol{\theta}\in\Theta$.

For exposition, we assume that $f(e)$ is completely specified. In practice, it often contains unknown parameters. In this scenario, the covariance function $C(\boldsymbol{\theta},\boldsymbol{\theta}^{\dagger})$ in Eq. (2) for $\mathrm{S}(\boldsymbol{\theta})$ needs modification. Pilla and Loader [6] derive an appropriate $C(\boldsymbol{\theta},\boldsymbol{\theta}^{\dagger})$ under estimated parameters.

For testing the hypotheses $\mathcal{H}_0$: $\eta = 0$ (no signal) versus $\mathcal{H}_1$: $\eta > 0$ (signal is present) consider the test statistic

$$\mathbb{T} := \sup_{\boldsymbol{\theta}\in\Theta}\mathrm{S}^{\star}(\boldsymbol{\theta}).$$

We conclude that a signal is present if $\mathbb{T}$ exceeds a critical value $c\in\mathcal{R}$ corresponding to a specified significance level α. In order to determine c, we need to find an approximation to the null distribution of $\mathbb{T}$.

Under $\mathcal{H}_0$, $\mathrm{S}^{\star}(\boldsymbol{\theta})$ converges in distribution to a Gaussian process $Z(\boldsymbol{\theta})$ with mean 0 and covariance function

$$\frac{C(\boldsymbol{\theta},\boldsymbol{\theta}^{\dagger})}{\sqrt{C(\boldsymbol{\theta},\boldsymbol{\theta})C(\boldsymbol{\theta}^{\dagger},\boldsymbol{\theta}^{\dagger})}}$$

as $n\to\infty$ [6]. The null distribution of $\mathbb{T}$ converges to that of $\sup_{\boldsymbol{\theta}} Z(\boldsymbol{\theta})$ as $n\to\infty$. Except in special cases, this distribution cannot be expressed analytically. However, a good asymptotic solution to the tail probability $P(\sup_{\boldsymbol{\theta}} Z(\boldsymbol{\theta})\geq c)$, where $c\in\mathcal{R}$ is large, can be obtained via the *volume-of-tube* formula [7–9]. This formula provides an elegant geometric approach for solving problems in simultaneous inference [10] by reducing the evaluation of tail probabilities to that of finding the $(J-1)$-dimensional volume of the set of points lying within a distance r of the curve $(d=1)$ or *manifold* $(d\geq 2)$, with boundaries, on the surface of the unit sphere $S^{(J-1)}$ embedded in $\mathcal{R}^J$ for some integer J (see Fig. 1).

We assume that the covariance function $C(\boldsymbol{\theta},\boldsymbol{\theta}^{\dagger})$ is twice differentiable. The results of Pilla and Loader [6] provide an expansion of the distribution of $\sup_{\boldsymbol{\theta}} Z(\boldsymbol{\theta})$ in terms of the χ^2 probabilities:

$$P(\sup_{\boldsymbol{\theta}\in\Theta} Z(\boldsymbol{\theta})\geq c) = \sum_{k=0}^{d}\frac{\zeta_k}{A_k A_{d+1-k}}P(\chi^2_{d+1-k}\geq c^2) + o(c^{-1}e^{-c^2/2}) \quad \text{as } c\to\infty, \tag{3}$$

where $A_0 = 1$ and $A_k = 2\pi^{k/2}/\Gamma(k/2)$ for $k\geq 1$.

The constants $\zeta_0,\ldots,\zeta_d$ depend on the geometry of a manifold as described next. The Gaussian random field $Z(\boldsymbol{\theta})$ is represented via the Karhunen-Loève expansion [11] as

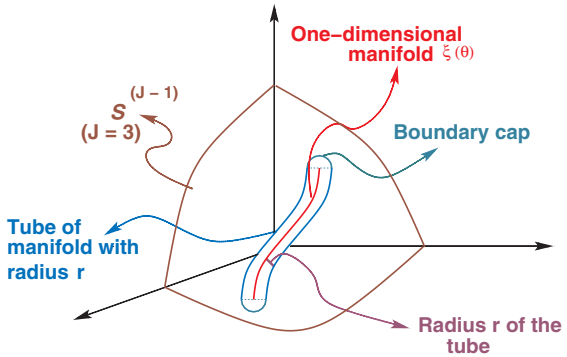

FIG. 1 (color). Tube around a one-dimensional manifold $\boldsymbol{\xi}(\boldsymbol{\theta})$, with boundaries, embedded in $S^2\subset\mathcal{R}^3$.

$$Z(\boldsymbol{\theta}) = \sum_{k=1}^{\infty} \upsilon_k \xi_k(\boldsymbol{\theta}) = \langle \boldsymbol{\nu}, \boldsymbol{\xi}(\boldsymbol{\theta}) \rangle,$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product, $\boldsymbol{\nu}$ and $\xi(\boldsymbol{\theta})$ are vectors, and $\upsilon_k \sim N(0, 1)$. The vector function $\xi(\boldsymbol{\theta})$ defines a curve (or surface, depending on the dimensionality of $\boldsymbol{\theta}$) on $S^{(J-1)}$. The coefficient ζ_0 is the length ($d = 1$), area ($d = 2$), or, in general, volume of the manifold $\xi(\boldsymbol{\theta})$ which is found via

$$\zeta_0 = \int_{\boldsymbol{\theta} \in \Theta} [C(\boldsymbol{\theta}, \boldsymbol{\theta})]^{-(d+1)/2} D(\boldsymbol{\theta}, \boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{4}$$

where $D(\boldsymbol{\theta}, \boldsymbol{\theta})$ is defined as

$$\left| \det \begin{pmatrix} C(\boldsymbol{\theta}, \boldsymbol{\theta}^\dagger) & \nabla_1 C(\boldsymbol{\theta}, \boldsymbol{\theta}^\dagger) \\ \nabla_2 C(\boldsymbol{\theta}, \boldsymbol{\theta}^\dagger) & \nabla_1 \nabla_2 C(\boldsymbol{\theta}, \boldsymbol{\theta}^\dagger) \end{pmatrix} \right|^{1/2}_{\boldsymbol{\theta}^\dagger = \boldsymbol{\theta}}$$

with ∇_1 and ∇_2 as the partial derivative operators with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^\dagger$, respectively. The constant ζ_1 measures the size of the boundaries of the manifold, and for $d = 1$ just counts the end points. The remaining constants involve curvature of the manifold and its boundaries, and become progressively more complex. In practice, the first few terms will suffice and an implementation of the first four terms is described in Ref. [12]. When the null distribution can be approximated by a χ^2 distribution, a tabulated value can be employed to calibrate the test statistic. However, the geometric constants appearing in Eq. (3) depend on the problem at hand.

In order to derive Eq. (3), we employ results of Hotelling-Weyl-Naiman [7–9]. If the Karhunen-Loève expansion is terminated after J terms, it follows that [6]

$$P(\sup_{\boldsymbol{\theta} \in \Theta} Z(\boldsymbol{\theta}) \geq c) = \int_{c^2}^{\infty} P(\sup_{\boldsymbol{\theta} \in \Theta} \langle \boldsymbol{U}, \boldsymbol{\xi}(\boldsymbol{\theta}) \rangle \geq w) \, h_J(y) dy,$$

where

$$\boldsymbol{U} = \left(U_1 = \frac{\upsilon_1}{\| \boldsymbol{\nu} \|}, \ldots, U_J = \frac{\upsilon_J}{\| \boldsymbol{\nu} \|} \right)$$

is uniformly distributed on $S^{(J-1)}$, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_J)$, $w = c/\sqrt{y}$, and $h_J(y)$ is a χ^2 density with J degrees of freedom. The probability in the integrand measures the chance that the random point $\boldsymbol{U}$ is sufficiently close to the curve $\xi(\boldsymbol{\theta})$. For small radii, this is determined by the volume-of-tube formula [7,8], extended by Ref. [9] to manifolds with boundaries. Performing the integration yields Eq. (3).

In order to apply Eq. (3), one has to evaluate ζ_k. First, one must find the covariance function in Eq. (2) by numerical integration. The constant ζ_0 is then found using a numerical integration to evaluate Eq. (4).

In many applications, including the one considered in this Letter, one is interested in the probabilities of rare events (i.e., $c \to \infty$). Therefore, the terms in Eq. (3) are of descending size, and the error term is asymptotically negligible.

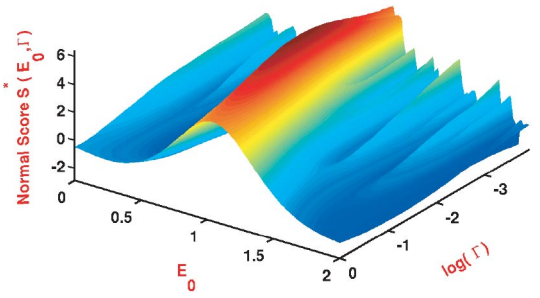

FIG. 2 (color). Surface of the process $S^\star(\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta} = (E_0, \Gamma)$.

We demonstrate the power of the score test with a Monte Carlo simulation experiment drawn from high-energy physics. In our simulations, we consider measurements of energy in a region $e \in [0, 2]$ in which the background (null) density is modeled as linear, with a specific form $f(e) = (1/2.6)(1 + 0.3e)$. The resonance is modeled by a Breit-Wigner density function. The parameters for this problem are modeled following an example in Roe [13].

To examine the effectiveness of $\mathbb{T}$ in detecting a signal, we perform Monte Carlo analyses of 10 000 samples each with a size of $n = 1000$ events spread over 50 bins at the values of $\Gamma = 0.2$ and $E_0 = 1$. For a single simulated data set, Fig. 2 shows the normalized score surface as a function of $\boldsymbol{\theta} = (E_0, \Gamma)$; the maximum is achieved near $E_0 = 1$ irrespective of the value of Γ.

Since the method is robust with respect to Γ, one can obtain significant computational gains by fixing Γ and optimizing over $\boldsymbol{\theta} = E_0$ alone. Figure 3 shows histograms over 10 000 samples under $\mathcal{H}_0$: $\eta = 0$ and $\mathcal{H}_1$: $\eta = 0.1$. We fixed $\Gamma = 0.2$, and optimized $S^\star(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta} = E_0$. The former histogram confirms that, about 5% of the time, the hypothesis of no signal is rejected. For this case, $\zeta_0 = 8.877$ and $\zeta_1 = 2.0$. The asymptotic null density

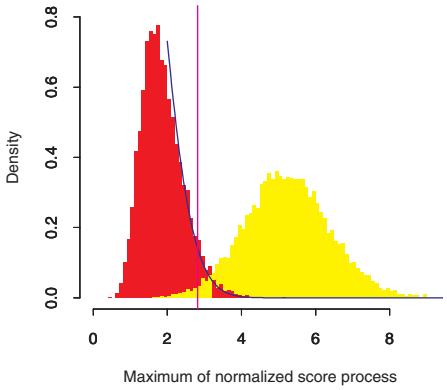

FIG. 3 (color). Histograms of the simulated null ($\eta = 0$) density (red) and alternative ($\eta = 0.1$) density (yellow) of the test statistic $\mathbb{T}$ with a superimposed (blue) asymptotic null density [derivative of Eq. (3)] for a fixed $\Gamma = 0.2$. The purple vertical bar is the cut off for the test statistic $\mathbb{T}$ at the 5% false positive rate calculated via the volume-of-tube formula [Eq. (3) with $d = 1$].

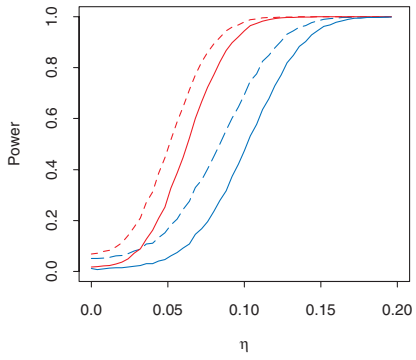

FIG. 4 (color). Power comparison of the χ^2 goodness-of-fit test (blue) and normalized score test $\mathbb{T}$ (red) for $d = 2$ at $\alpha = 0.05$ (dashed lines) and $\alpha = 0.01$ (solid lines), calculated via the volume-of-tube formula, based on 10 000 simulations for binned data.

[derivative of Eq. (3) with $d = 1$] agrees with the simulated null distribution as expected.

Figure 4 displays power curves for $\boldsymbol{\theta} = (E_0, \Gamma)$. In this case, $\zeta_0 = 13.51$, $\zeta_1 = 35.3$, and $\zeta_2 = -7.23$. The results demonstrate an increase in power as the signal strength η increases; $\mathbb{T}$ is significantly more powerful than the χ^2 goodness-of-fit test in detecting the signal. The geometric asymptotic tail probability result [Eq. (3)] is elegant, simple and powerful in distinguishing the signal and the random fluctuations in data.

Financial support from the U.S. National Science Foundation, Division of Mathematical Sciences [DMS 02-39053 (R. S. P.) and DMS 03-06202 (C. L.)] and the Office of Naval Research, Probability and Statistics Program [N00014-02-1-0316 (R. S. P.) and N00014-04-1-0481 (R. S. P. and C. L.)] is gratefully acknowledged.

*Email address: pilla@case.edu
Corresponding author.
Electronic address: http//stat.case.edu/~pillar/

[1] S. S. Wilks, *Mathematical Statistics* (Princeton University Press, New Jersey, 1944).
[2] W. T. Eadie *et al.*, *Statistical Methods in Experimental Physics* (North-Holland, New York, 1971).
[3] K. S. Cranmer, in *Proceedings of PHYSTAT2003, SLAC, Stanford, California, 2003*, edited by L. Lyons, R. Mount, and R. Reitmeyer, eConf C030908, 211 (2003); physics/031011002.
[4] P. E. Freeman *et al.*, Astrophys. J. **524**, 753 (1999).
[5] R. Protassov *et al.*, Astrophys. J. **571**, 545 (2002).
[6] R. S. Pilla, and C. Loader, math.ST/0511503.
[7] H. Hotelling, Am. J. Math. **61**, 440 (1939).
[8] H. Weyl, Am. J. Math. **61**, 461 (1939).
[9] D. Q. Naiman, Ann. Stat. **18**, 685 (1990).
[10] M. Knowles, and D. Siegmund, Int. Stat. Rev. **57**, 205 (1989).
[11] R. J. Adler, *An introduction to Continuity, Extrema and Related Topics for General Gaussian Processes* (Institute of Mathematical Statistics, Hayward, California, 1990).
[12] C. Loader, math.ST/0511502.
[13] B. P. Roe, *Probability and Statistics in Experimental Physics* (Springer, New York, 1992).