Statistica Sinica Preprint No: SS-2018-0027	
Title	Testing One Hypothesis Multiple Times
Manuscript ID	SS-2018-0027
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0027
Complete List of Authors	Sara Algeri and David A. van Dyk
Corresponding Author	Sara Algeri
E-mail	salgeri@umn.edu
Notice: Accepted version subject to English editing.	

Testing One Hypothesis Multiple Times

Sara Algeri*,† and David A. van Dyk†

*University of Minnesota and † Imperial College London

Abstract: In applied settings, tests of hypothesis where a nuisance parameter is only identifiable under the alternative often reduces into one of *Testing One Hypothesis Multiple times* (TOHM). Specifically, a fine discretization of the space of the non-identifiable parameter is specified, and the null hypothesis is tested against a set of *sub-alternative hypothesis*, one for each point of the discretization. The resulting *sub-test statistics* are then combined to obtain a *global p-value*. In this paper, we discuss a computationally efficient inferential tool to perform TOHM under stringent significance requirements, such as those typically required in the physical sciences, (e.g., p-value $< 10^{-7}$). The resulting procedure leads to a generalized approach to perform inference under non-standard conditions, including non-nested models comparisons.

Key words and phrases: Multiple hypothesis testing, bump hunting, non-identifiabily in hypothesis testing, non-nested models comparison.

1. Introduction

{intro}

A fundamental statistical challenge in scientific discoveries is the so called "bump-hunting" problem (Choudalakis, 2011), where researchers aim to

distinguish peaks due to a signal of interest (the new discovery) from peaks due to random fluctuations of the background. In the framework of hypothesis testing, the null model specified by H_0 is typically the background-only model, and a signal bump is added in the alternative model specified by H_1. Consider for example a dark matter search where we aim to distinguish events associated with a power-law (Pareto type I) distributed background from the signal of a dark matter source modeled as a narrow Gaussian bump with unknown location over the search area $\Theta \equiv [\mathcal{L}, \mathcal{U}] \subset \mathbb{R}$. We can specify the model of interest using a mixture model

$$(1-\eta)\frac{1}{k_\phi y^{\phi+1}} + \frac{\eta}{k_\theta}\exp\left\{-\frac{(y-\theta)^2}{0.02\theta^2}\right\} \quad \text{for } y \geq 1, \qquad (1.1) \quad \{\texttt{ex1}\}$$

where k_ϕ and k_θ are normalizing constants, $y \geq 1$, $\phi > 0$, and $\theta \geq 1$. Notice that the parameter θ characterizes both the location of the signal over the search region and its standard deviation. Specifically, the bump becomes wider the further its position is in the tail of the background distribution. The model in (1.1) is a toy example which simplifies the models involved in the context of searches for γ-ray emissions ina cluster of galaxies (Anderson et al., 2016); where for example the width of the signal may be a more complex function of its location. Despite its simplicity, the model in (1.1)

introduces the key statistical issues arising in the context of dark matter searches, as described below.

In order to assess the evidence in favor of the signal, we test

$$H_0 : \eta = 0 \quad \text{versus} \quad H_1 : \eta > 0. \tag{1.2}$$ {testex1}

where η is the proportion of events due to the dark matter emission, and typically $0 \leq \eta \leq 1$. Despite its straightforward formulation, testing (1.2) is non-trivial. Difficulties arise because θ is not defined under H_0. Consequently, classical asymptotic properties of, e.g., Maximum Lixelihood Estimates (MLE) and the Likelihood Ratio Test (LRT), fail. Analogously, complications may arise when using resampling techniques, such as bootstrapping (Efron and Tibshirani, 1994), to derive the null distribution of the test statistic, in the presence of stringent significance requirements. For searches in high energy physics for instance, the significance level necessary to claim a discovery can be in the order of 10^{-7} (see Lyons, 2013, Table 1). Hence, a large (e.g., $O(10^8)$) simulation may be infeasible when dealing with complex models. This is a key motivation for a computationally efficient inferential solution.

To address these difficulties, in this paper, we consider the bump-

hunting problem as a special case of what is known in statistical literature as *"testing statistical hypotheses when a nuisance parameter is present only under the alternative"*. In addition to bump-hunting, classical examples may include regression models where structural changes, such as break-points and threshold-effects, occur (Andrews, 1993; Hansen, 1992b, 1999; Davies, 2002).

The general problem has long been studied, starting at least from the seminal work of Hotelling (1939) and Davies (1977, 1987), and further investigated in the econometrics literature by several authors including Andrews and Ploberger (1994) and Hansen (1991, 1992a, 1996). In their practical implementation, these methods reduce the problem of testing with unidentifiable parameters under H_0 into one of *Testing One Hypothesis Multiple times* (TOHM), where a single null hypothesis H_0 is tested against different *sub-alternative hypotheses* of the form $H_1(\theta)$, one for each fixed θ in Θ, and a corresponding ensemble of *sub-test statistics* indexed by θ, namely $W(\theta)$, is specified. The goal is to provide a global p-value as the standard of evidence for comparing H_0 and the global alternative hypothesis H_1, of which each $H_1(\theta)$ is a special case. Unfortunately, existing methods often require case-by-case mathematical computations (e.g., Davies, 1977), estimating the covariance structure (e.g., Hansen, 1991), choosing weighting

functions (e.g., Andrews and Ploberger, 1994), or full simulations of the empirical process (e.g., Hansen, 1992a, 1996).

In this paper we discuss a computationally efficient method to perform TOHM which overcomes these limitations. Specifically, as in Davies (1977, 1987) we consider a stochastic process, $\{W(\theta)\}$, indexed by $\theta \in \Theta \equiv [\mathcal{L}, \mathcal{U}]$, and with covariance function $\rho(\theta, \theta^{\dagger})$. We consider the global p-value

$$P\left(\sup_{\theta\in\Theta}\{W(\theta)\} > c\right), \tag{1.3}$$ {pval}

where c is the observed value of the *global test statistic*, $\sup_{\theta\in\Theta}\{W(\theta)\}$. The central difficulty of this approach is to derive or approximate (1.3). One possible way forward is to consider the Extreme Value Theory (EVT) argument developed by Cramér and Leadbetter (2013, p. 272), where a bound for (1.3) is obtained considering the upcrossings of c by $\{W(\theta)\}$ (see Figure **??**). Specifically, $\{W(\theta)\}$ has an *upcrossing* of a threshold $c \in \mathbb{R}$ at $\theta_0 \in \Theta$ if, for some $\epsilon > 0$, $W(\theta) \leq c$ in the interval $(\theta_0 - \epsilon, \theta_0)$ and $W(\theta) \geq c$ in the interval $[\theta_0, \theta_0 + \epsilon)$ (Adler, 2000). Let N_c be the number of upcrossings of c by $\{W(\theta)\}$. Using Markov's inequality, Cramér and Leadbetter (2013,

p. 272) show that (1.3) can be bounded as in (1.4),

$$P\left(\sup_{\theta\in\Theta}\{W(\theta)\} > c\right) \leq P(W(\mathcal{L}) > c) + E[N_c] \qquad (1.4) \quad \texttt{\{general\_bound\}}$$

where $P(W(\mathcal{L}) > c)$ is typically known. Davies (1977, 1987) consider the cases where $\{W(\theta)\}$ is a Gaussian or a χ^2-process, estimate $E[N_c]$ via total variation, and show that (1.4) becomes sharp, as $c \to \infty$ (under long-range independence, i.e., if $\rho(\theta, \theta^\dagger) \to 0$ as $|\theta - \theta^\dagger| \to \infty$). Unfortunately, Hansen (1991) points out that situations exist where the total variation diverges.

An alternative solution can overcome this problem and has had significant impact in physics (Gross and Vitells, 2010). Consider a set of observations $y_1, \ldots, y_n$, and let $T_n(\theta)$ the LRT statistics used to test (1.2) and evaluated on $y_1, \ldots, y_n$ when θ is fixed. We denote the LRT-process indexed by different values of θ with $\{T_n(\theta)\}$. Under H_0 and suitable uniformity conditions (Hansen, 1991), $\{T_n(\theta)\} \xrightarrow[n\to\infty]{d} \{W_\chi(\theta)\}$, where $W_\chi(\theta)$ is a χ^2-process with components $W_\chi(\theta) \sim \chi^2_s$, for each $\theta \in [\mathcal{L}, \mathcal{U}]$ fixed. Let $E[N_c^\chi]$ be the expected number of upcrossings of c by $\{W_\chi(\theta)\}$ over Θ. One possible way to compute (1.4) is to estimate $E[N_c^\chi]$ via Monte Carlo simulations. However, when dealing with stringent significance requirements, the corresponding significance threshold c is typically very large. Hence,

upcrossings of c are expected to occur infrequently when simulating under H_0, and thus a massive simulation is required to estimate $E[N_c^\chi]$ directly. Gross and Vitells (2010) exploit the χ^2 distribution of $\{W_\chi(\theta)\}$, and rewrite $E[N_c^\chi]$ as a function of $E[N_{c_0}^\chi]$, see (1.5), for some $c_0 << c$,

$$P\left(\sup_{\theta\in\Theta}\{W_\chi(\theta)\} > c\right) \leq P(W_\chi(\mathcal{L}) > c) + \left(\frac{c}{c_0}\right)^{\frac{s-1}{2}} e^{-\frac{c-c_0}{2}} E[N_{c_0}^\chi]. \quad (1.5) \quad \{\texttt{gv\_bound}\}$$

where $E[N_c^\chi] = \left(\frac{c}{c_0}\right)^{\frac{s-1}{2}} e^{-\frac{c-c_0}{2}} E[N_{c_0}^\chi]$. This allows a drastic reduction in the computational effort needed to compute $E[N_c^\chi]$. Specifically upcrossings of $c_0 << c$ are expected to occur often, and thus $E[N_{c_0}^\chi]$ can be estimated accurately with a small Monte Carlo simulation.

Gross and Vitells (2010) do not formally justify (1.5). In Section 2, we derive (1.5), we generalized it to any process $\{W(\theta)\}$, and we clarify the conditions under which (1.5) and its generalization hold. Efficient choices of c_0 are discussed in Section 3 and a simple graphical tool is proposed to validate the adequacy of the number of sub-tests conducted.

The resulting procedure leads to a generalized approach to perform inference under non-standard regularity conditions including, as discussed in Section 3, comparisons of non-nested models. This can be done by specifying a comprehensive model that includes the two (non-nested) models

under comparison as special cases. Two tests of hypothesis where a nuisance parameter is present only under the alternative are then performed to select among the two models (Algeri et al., 2016).

In principle, the problem of testing in presence of a nuisance parameter which is present only under the alternative can be formulated as a multiple hypothesis testing (MHT) problem, where several tests are conducted over a grid of possible values of θ, and corrected using Bonferroni's correction (Bonferroni, 1935, 1936) or similar methods to control for the probability of type I error. Although the Bonferroni correction is easy to implement, it is often dismissed by practitioners both because of its stringent control of the overall false detection rate and its artificial dependence on the number of tests conducted. In Section 4 we compare TOHM and Bonferroni's correction via a suite of numerical studies and data applications; we also discuss how the tools introduced in this manuscript can be used to identify situations where, by virtue of its relationship with TOHM, Bonferroni can be used without worry about obtaining an overly conservative result.

The remainder of the paper is organized as follows. In Section 2, we define the framework for TOHM, and we derive a computable upper bound for (1.3) by generalizing (1.5). In Section 3, we illustrate how TOHM can be used to distinguish among non-nested models, we validate our results with

simulation studies and we discuss graphical tools to select the necessary quantities involved in the computation of the bound proposed in Section 2. In Section 4 we investigate the relationship between TOHM and the classical Bonferroni correction, and we apply both methods on several realistic data sets. A summary and a discussion of our findings appear in Section 5. Additional figures, data and proofs are collected in the Supplementary Material.

2. TOHM via EVT

{sec3}

2.1 Definition and formalization

{GV}

In this section, we generalize the testing procedure of Gross and Vitells (2010) beyond the LRT and the χ^2 case and formalize it in statistical terms. This allows us to establish a general theoretical framework to efficiently bound/approximate the global p-value in (1.3).

Recall that $\{W(\theta)\}$ is a generic stochastic process indexed by $\theta \in \Theta \equiv [\mathcal{L};\mathcal{U}]$ with covariance function $\rho(\theta, \theta^\dagger)$. Following Davies (1987) we stipulate

{cond31}

Condition 1. *$\{W(\theta)\}$ has continuous sample paths; $\{W(\theta)\}$ has continuous first derivative, except possibly for a finite number of jumps; and its components $W(\theta)$ are identically distributed for all $\theta \in \Theta$.*

To exploit (1.4), we aim to conveniently estimate $E[N_c]$ and bound or

approximate (1.3). Results 2 and 3 allow this.

{theo1}

Result 2. *Let $c \in \mathbb{R}$ be an arbitrary threshold, $a(c)$ be a function which depends on c but not on θ, and $b(\Theta)$ be a function which does not depend on c, and to be calculated over Θ. Under Condition 1, if $E[N_c]$ can be decomposed as*

$$E[N_c] = a(c)b(\Theta) \tag{2.6}$$ {decompose}

then,

$$E[N_c] = \frac{a(c)}{a(c_0)} E[N_{c_0}] \qquad \forall c_0 \leq c, c_0 \in \mathbb{R}. \tag{2.7}$$ {expect}

The function $b(\Theta)$ typically involves integration over the interval Θ, and should not be confused with a function of θ. Deriving a closed-form expression of $b(\Theta)$ in (2.6) may be challenging, and may require knowledge of $\rho(\theta, \theta^{\dagger})$. Conversely, the form of $a(c)$ typically depends on the marginal distribution of the components $W(\theta)$ of $\{W(\theta)\}$, hence the requirement of identical distribution in Condition 1. The continuity assumptions on $\{W(\theta)\}$ and its first derivative prevent $E[N_c]$ from diverging.

Equation (2.7) offers a simple way to compute $E[N_c]$, provided that, as discussed below, $E[N_{c_0}]$ can be estimated accurately. Result 3 follows from (1.4), (2.6), and (2.7).

{coroll1}

Result 3. *Under Condition 1, if* (2.6) *holds,* (1.3) *can be bounded by*

$$P\left(\sup_{\theta\in\Theta}\{W(\theta)\} > c\right) \leq P(W(\mathcal{L}) > c) + \frac{a(c)}{a(c_0)}E[N_{c_0}] \qquad (2.8) \quad \{\texttt{bound2}\}$$

for all $c_0 \leq c, c_0 \in \mathbb{R}$. *If additionally,* $\rho(\theta, \theta^\dagger) \to 0$ *as* $|\theta - \theta^\dagger| \to \infty$, *the difference between the left and the right hand side of* (2.8) *approaches zero as* $c \to \infty$.

2.2 TOHM bounds for Gaussian-related processes {bounds}

The bound in (1.5) and the analogous bounds for Gaussian and related processes such as F and t-processes, can be derived using results of random fields theory as discussed in Algeri and van Dyk (2018). In this setting, it can be shown that, under mild smoothness conditions (see Taylor and Adler (2003, p. 547)), $E[N_c]$ enjoys the decomposition in (2.6), where $a(c)$ only depends on the distribution of the marginals of $\{W(\theta)\}$, whereas $b(\Theta)$ corresponds to the so-called Lipschitz-Killing curvature of first order (e.g., Adler and Taylor, 2009) and is typically difficult to compute. Here, we report explicit forms of the right hand side of (2.8) for Gaussian, F and t processes which can be obtained on the basis of these results (see Taylor and Worsley, 2008; Adler and Taylor, 2009; Algeri and van Dyk, 2018, for

more details).

Gaussian process. Let $\{Z(\theta)\}$ be a mean zero and variance one Gaussian process, such that $Z(\theta) \sim N(0,1)$ for all $\theta \in \Theta$, and let N_c^Z be the process of upcrossings of c_0 by $\{Z(\theta)\}$ over $\Theta \equiv [\mathcal{L},\mathcal{U}]$. The TOHM bound in equation (2.8) takes the form

$$P\left(\sup_{\theta\in\Theta}\{Z(\theta)\} \geq c\right) \leq \Phi(-c) + e^{-\frac{c^2-c_0^2}{2}} E[N_{c_0}^Z]. \qquad (2.9) \quad \texttt{\{z\_bound\}}$$

where $\Phi(-c)$ is the cumulative density function of a standard normal random variable evaluated at $-c$ and the ratio $\frac{a(c)}{a(c_0)}$ is givan by $e^{-\frac{c^2-c_0^2}{2}}$. For the stationary case, the same result can be obtained by expressing $E[N_c^Z]$ via Rice's formula (Rice, 1944) i.e.,

$$E[N_c^Z] = \frac{|\mathcal{L}-\mathcal{U}|}{2\pi}\sqrt{\rho''(\theta,\theta)}e^{-\frac{c^2}{2}}$$

where $\rho''(\theta,\theta) = \frac{\partial\theta}{\partial\theta\partial\theta^\dagger}\rho(\theta,\theta^\dagger)\big|_{\theta^\dagger=\theta}$ is the second spectral moment of $\{Z(\theta)\}$ and is assumed to be finite, and $|\mathcal{L}-\mathcal{U}|$ is the length of Θ. As discussed in Davies (1987), for a two-sided test, the excursion probability of interest is $P(\sup_{\theta\in\Theta}|\{Z(\theta)\}| \geq c)$; the bound of which is twice the right hand side of

(2.9).

The rate of convergence of the difference between the right and left hand side of (1.5) and (2.9) are discussed in Section S.1 of the Supplementary Material. We further study the sharpness of the bounds in (1.5) and (2.9), as $c \to \infty$ in Section 3 via a suite of simulation studies.

F-process. Consider an F-process $\{F(\theta)\}$ with s and v degrees of freedom such that $F(\theta) \sim F_{s,v}$ for all $\theta \in \Theta$. Let $E[N^F_{c_0}]$ be the expected number of upcrossings of c_0 by $\{F(\theta)\}$, then the TOHM bound in equation (2.8) takes the form

$$P\left(\sup_{\theta\in\Theta}\{F(\theta)\} \geq c\right) \leq P(F(\mathcal{L}) \geq c) + \left(\frac{c}{c_0}\right)^{\frac{s-1}{2}} \left(\frac{v + s\cdot c}{v + s\cdot c_0}\right)^{-\frac{s+v-2}{2}} E[N^F_{c_0}] \tag{2.10}$$

{F_bound}

for all $c_0 \leq c, c_0 \in \mathbb{R}$, and with $a(c) = c^{\frac{s-1}{2}}(v + s\cdot c)^{-\frac{s+v-2}{2}}$.

t-process. Consider a t-process $\{V(\theta)\}$ with s degrees of freedom such that $V(\theta) \sim t_s$. Let $E[N^V_{c_0}]$ be the expected number of upcrossings of c_0 by

$\{V(\theta)\}$, then the TOHM bound in equation (2.8) takes the form

$$P\left(\sup_{\theta\in\Theta}\{V(\theta)\}\geq c\right)\leq P(V(\mathcal{L})\geq c)+\left(\frac{1+c^2}{1+c_0^2}\right)^{-\frac{s-1}{2}}E[N_{c_0}^V] \qquad (2.11) \quad \{\texttt{t\_bound}\}$$

for all $c_0\leq c, c_0\in\mathbb{R}$, and with $a(c)=(1+c^2)^{-\frac{s-1}{2}}$.

2.3 Testing one hypothesis multiple times in practice {Ncrdef}

In practice, we evaluate $\{W(\theta)\}$ on a fine grid of points, namely $\Theta_R=\{\theta_1,\dots\theta_R\}\subseteq\Theta$, with R being the typically large number of grid points. Let $\{W(\theta_r)\}$ be the random sequence which coincides with $\{W(\theta)\}$ at each $\theta_r\in\Theta_R$ and $\{w(\theta_r)\}$ be its observed value. We approximate $\sup_{\theta\in\Theta}\{W(\theta)\}$ with its discrete counterpart $\max_{\theta_r\in\Theta_R}\{W(\theta_r)\}$, the observed value of which is given by

$$c_R=\max_{\theta_r\in\Theta_R}\{w(\theta_r)\}. \qquad (2.12) \quad \{\texttt{cR}\}$$

Let the process of upcrossings of c_R by $\{W(\theta_r)\}$, namely $\tilde{N}_{c_R}$, be events of the type $\{W(\theta_{r-1})\leq c_R, W(\theta_r)>c_R\}$. We assume that Θ_R is sufficiently dense, so that the right hand side of (2.8) can be approximated by (2.13), as $R\to\infty$,

$$P(W(\mathcal{L})>c_R)+\frac{a(c_R)}{a(c_0)}E[\tilde{N}_{c_0}] \qquad \forall c_0\leq c_R, c_0\in\mathbb{R} \qquad (2.13) \quad \{\texttt{real\_bound}\}$$

where $E[\tilde{N}_{c_0}]$ can be replaced by its Monte Carlo estimate, namely $\widehat{E[\tilde{N}_{c_0}]}$.

Notice that the null hypothesis, H_0, is tested versus an ensable of alternative hypotheses H_{1r}, one for each value of θ_r fixed. The observed *sub-test statistics* $\{w(\theta_1), \ldots, w(\theta_R)\}$, realizations of $\{W(\theta)\}$, are combined into the global test statistic c_R and an approximated bound for the global p-value is computed via (2.13). Thus, the problem of testing (1.2) is reduced to testing H_0 versus the R *sub-alternative hypotheses* H_{1r}, i.e., *Testing One Hypothesis Multiple Times*.

Cramér and Leadbetter (2013, p. 63 and 195) discuss adequate choices of Θ_R for which c, N_c and $\sup_{\theta\in\Theta}\{W(\theta)\}$ are well approximated by c_R, $\tilde{N}_{c_R}$ and $\max_{\theta_r\in\Theta_R}\{W(\theta_r)\}$, respectively. However, since in practice Θ_R may be determined by the experiment, in Section 3 we discuss graphical tools to assess whether these approximations hold.

3. Practical matters

{practice}

3.1 Case studies: description

{examples}

Here we illustrate the implementation of TOHM in the context of three case studies, i.e., the "bump hunting" problem introduced in Section 1, a non-nested models comparison, and a logistic model with a break point. Hereafter, we refer to these as Examples 1, 2 and 3, respectively. Data

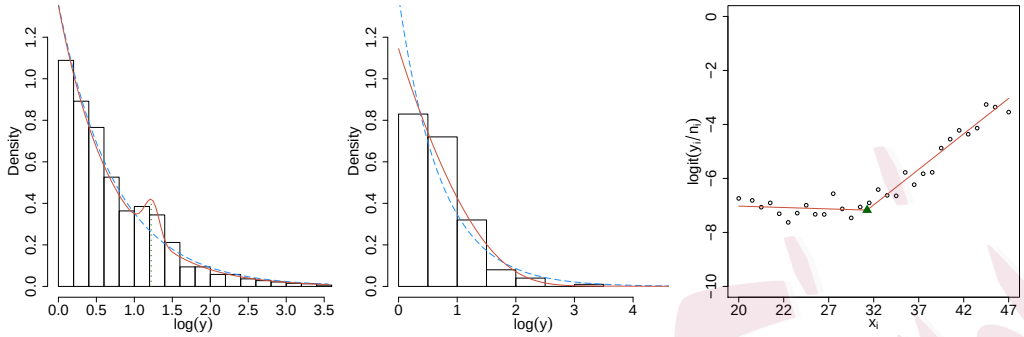

Figure 1: Data and fitted models. Left panel: histogram of the Fermi-LAT realistic data simulation for Example 1 (on log-scale), null model (blue dashed curve) fitted under the assuption of background only counts ($\hat{\phi} = 1.350$), and fitted alternative model (red solid curve) with $\hat{\eta} = 0.045$, $\hat{\phi} = 1.406$. The green dotted vertical line indicates the location of the observed Gaussian bump, i.e., $\hat{\theta} = 3.404$. Central panel: histogram of the Fermi-LAT realistic data simulation for Example 2 (log-scale), the null model when testing (1.2) *is fitted as a power-law distributed cosmic source with $\hat{\phi} = 1.395$ (blue dashed curve). The null model when testing* (3.16) *is the dark matter model in* (3.14) *with $\hat{\theta} = 27.89$ obtained via MLE (red solid curve) . Right panel: Down syndrome data and fitted regression model (red piecewise-linear solid lines), with break-point (green triangle) at $\hat{\theta} = 31.266$.* {real_plots}

for Examples 1 and 2 were generated using simulations of the Fermi Large Area Telescope (LAT) obtained with the *gtobssim* package and include representations of detector effects and systematic errors. The Fermi-LAT is a γ-ray telescope on the orbiting Fermi satellite (Atwood et al., 2009).

In Example 1, our data analysis aims to properly distinguish between γ-ray signals induced by dark matter annihilations and those induced by the astrophysical background. As in (1.1), dark matter events are modeled as a Gaussian bump with mean energy θ and standard deviation varying with

http://fermi.gsfc.nasa.gov/ssc/data/analysis/software

θ. The astrophysical background is power-law (Pareto type I) distributed with index ϕ. In our simulation, we set $\theta = 3.5$GeV (where GeV denotes Giga electron-volt), $\phi = 1.4$, $\eta = 0.02$, and we consider the energy band $y \in [1; 35]$. This setup resulted in 64 dark matter events and 2274 background events. For more physics details, see Algeri et al. (2016).

In Example 2, the non-nested models to be compared are a dark matter emission with probability density given by

$$g(y, \theta) \propto y^{-1.5} \exp\left\{-7.8\frac{y}{\theta}\right\}, \qquad (3.14) \quad \texttt{\{DMmodel\}}$$

with $y \geq 1$, $\phi > 0$ and $\theta \geq 1$ (see Bergström et al., 1998) and a power-law distributed cosmic source with density $f(y, \phi) \propto \frac{1}{k_\phi y^{\phi+1}}$. In our simulation we set the putative dark matter emission to occur at $\theta = 35$GeV, and the power-law index to $\phi = 1.4$. In this way, we obtained 200 dark matter events over the energy band $y \in [1; 100]$.

Since the models $f(y, \phi)$ and $g(y, \theta)$ are non-nested, the classical asymptotic properties of the MLE and LRT fail. However, as shown in Algeri et al. (2016), the framework of Section 2 can be extended to compare non-nested models by reformulating this comparison as a test in which a nuisance parameter is identified only under H_1. Specifically, following Cox (1962) and

Atkinson (1970), we specify a comprehensive model that embes two non-nested models, i.e.,

$$(1-\eta)f(y,\phi)+\eta g(y,\theta) \qquad 0\leq\eta\leq 1. \tag{3.15}$$ {comprehensive}

This reduces the problem to a nested models comparison and we test (1.2). However, in contrast to the bump-hunting example in (1.1), here η has no physical interpretation. Rather, as in Quandt (1974), η is an auxiliary parameter which allows us to exploit the normality of its MLE to apply well-know asymptotic results. In addition to (1.2), the hypotheses

$$H_0:\eta=1 \quad \text{versus} \quad H_1:\eta<1 \tag{3.16}$$ {flipping}

should also be tested in order to exclude intermediate situations (e.g., Cox, 1962, 2013). I.e., we want to avoid treating (3.15) as a mixture and focus on comparing the two models. Testing both (1.2) and (3.16) is particularly suited to particle physics searches where researchers typically assign different degrees of belief to the models being tested. Specifically, as described in van Dyk (2014), the most stringent significance requirements (e.g., Lyons, 2013, Table 1) are typically used only in the *detection* stage, i.e., when testing (1.2) to assess the presence of a new signal. Conversely, in the *exclusion*

stage, i.e., when testing (3.16) to exclude the hypothesis of a signal being present, a significance level of 0.05 is typically sufficient. The Fermi-LAT datasets for Examples 1 and 2 are plotted in the first two panels of Figure 1. Both simulations are downloadable among the Supplementary Materials.

Finally, in Example 3 we consider the *Down Syndrome dataset* available in the R package `segmented` (Muggeo et al., 2008). The dataset records whether babies born to 354,880 women are affected by Down Syndrome. We use (3.17) to model the probability, π_i, that a woman of age x_i has a baby with down syndrome, where $x_i \in [17; 47]$, and we let $\theta \in [20; 44]$. The logit of the ratio between the number of down syndrome cases and number of births by age group is plotted in the right panel of Figure 1.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \phi_1 + \phi_2 x_i + \xi(x_i - \theta)\mathbb{1}_{\{x_i \geq \theta\}} \quad \forall i = 1, \ldots, n, \qquad (3.17) \quad \{\texttt{ex3}\}$$

where $\theta \in \mathbb{R}$ is the location of the unknown break-point. In this case, we test $H_0 : \xi = 0$ versus $H_1 : \xi \neq 0$.

In Example 1 and 2 we use the LRT, $T_n(\theta)$, as the sub-test statistic. Since both tests are of the form in (1.2), the test is on the boundary of the parameter space and for each θ fixed the asymptotic distribution under H_0 is a mixture of χ^2_1 and zero (Chernoff, 1954; Self and Liang, 1987), also known

as $\bar{\chi}$-distribution and which we dentote with $\bar{\chi}^2_{01}$. It can be shown (Algeri and van Dyk, 2018) that in this setting the bound in (2.8) has the same form as in the χ^2_1 case, i.e., it is given by (1.5) with $s = 1$. In Example 3, we use the signed-root of the LRT $Q_n(\theta) = \text{sign}(\hat{\eta}_\theta - \eta_0)\sqrt{T_n(\theta)}$, hence the sub-tests statistics are asymptotically normally distributed under H_0 (e.g., Davies, 1977).

3.2 The choices of c_0 and R

{choosingR}

One way to select an appropriate thresholds c_0 is to perform a sensitivity analysis based on few Monte Carlo simulations of the traces of the underlying processes under H_0. As discussed in Section 2, under suitable regularity conditions and when H_0 is true, the LRT and signed-root LRT processes $\{T_n(\theta)\}$ and $\{Q_n(\theta)\}$ converge uniformly to $\{W_\chi(\theta)\}$ and $\{Z(\theta)\}$, respectively, as $n \to +\infty$. More generally, given a test statistics $W_n(\theta)$ to be evaluated on the data $y_1, \ldots, y_n$ for each θ fixed, we write $\{W_n(\theta)\} \xrightarrow{d} \{W(\theta)\}$. Consequently, for each sample generated under H_0, we compute $\{W_n(\theta)\}$ over a fine grid of values of θ and which approximates $\{W(\theta)\}$ when n is large. In all our simulations, the nuisance parameters under the null model have been estimated via MLE and each simulated sample under H_0 is obtained via parametric bootstrap (Efron and Tibshirani, 1994). We plot the results of our simulation in order to visualize the traces of $\{W_n(\theta)\}$ as

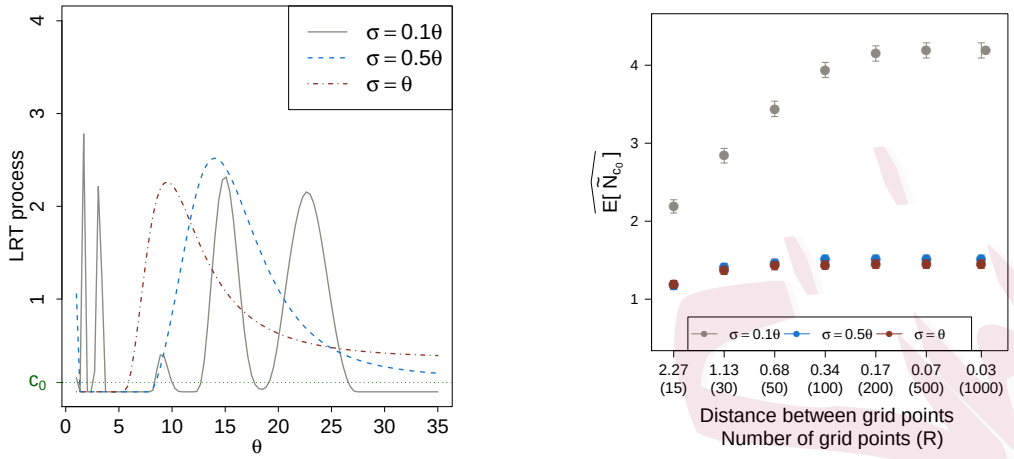

Figure 2: Left panel: simulated sample paths of the LRT process, $\{T_n(\theta)\}$, under H_0 in Example 1. Both plots consider different widths of the Gaussian bump. Right panel: upcrossings plot showing Monte Carlo estimates of $E[\tilde{N}_{c_0}]$ and standard errors (whiskers), under H_0, for Example 1, and evaluated over grids of $R = 15, 30, 50, 100, 200, 500$ points, and for three choices of the Gaussian width, namely $\sigma = 0.1\theta, \sigma = 0.5\theta$ and $\sigma = \theta$. {upc_Gauss}

shown in Figure 2 for Example 1. (The analogous plots for Examples 2 and 3 appear in Figure **??**.) In order to calculate (2.8), it is important to provide an accurate estimate of $E[N_{c_0}]$. Hence, we choose c_0 to be at a level (on the y-axis) around which the process $\{W_n(\theta)\}$ oscillates often, and thus, with respect to which the upcrossings occur with high frequency. For Examples 1, 2 and 3, this leads to values c_0 equal to $0.1, 0.3$ and 0, respectively. Inspecting the smoothness of the trace plots also allows us to qualitatively assess Condition 1 and verify the goodness of the approximation of $E[N_{c_0}]$ by $E[\tilde{N}_{c_0}]$, necessary for the validity of the results of Section 2.

As discussed in Section 1, the implementation of our procedure requires the specification of a grid Θ_R over $\Theta \equiv [\mathcal{L};\mathcal{U}]$, where R is the number of times H_0 is tested versus the ensemble of sub-alternatives $H_{11}, \ldots, H_{1R}$. In practice, R must either be chosen arbitrarily by the researcher or determined by the nature of the experiment. In either case, R must be sufficiently large to guarantee robustness of the results, yet small enough to ensure computational efficiency when calculating (2.13). One possibility is to choose R large enough so that, for a given c_0, $E[\tilde{N}_{c_0}]$ converges to a finite limit, which we expect, for sufficiently dense Θ_R, to correspond to $E[N_{c_0}]$. This strategy requires us to set c_0 before setting R.

In order to identify the value of R that best negotiates the trade-off between accuracy and computational efficiency, one can consider different values of R and for each of them compute an estimate of $E[N_{c_0}]$ by means of a small Monte Carlo simulation. The results can then be summarized in an *upcrossing plot* where the values for R considered are reported on the x-axis and the respective $\widehat{E[\tilde{N}_{c_0}]}$ estimates of $E[N_{c_0}]$ are reported on the y-axis. The upcrossing plot in the right panel of Figure 2 displays Monte Carlo estimates $\widehat{E[\tilde{N}_{c_0}]}$ for the LRT in Example 1, under H_0, as a function of R (with $R = 15, 30, 50, 100, 200, 500, 1000$). For each value of R considered, the grid points have been chosen to be equally spaced over

Θ. Analogous plots for Examples 2 and 3 appear in Figure **??**. For each R considered we computed 100 Monte Carlo simulations, each of size 1000. In all our examples, 100 simulations are sufficient to achieve small Monte Carlo errors.

As a rule of thumb, if the number of upcrossings increases with R but does not converge, it means that the resolution is not sufficiently high to catch all the crossings or, the underlying process is not sufficiently smooth to guarantee $E[N_{c_0}] < \infty$. Conversely, if the number of upcrossings converges, as in the well-known scree-plot used for Principal Component Analysis (PCA) (e.g., James et al., 2013, p. 383), we look for an "elbow" in the plot of $\widehat{E[\tilde{N}_{c_0}]}$. The value of R corresponding to the elbow is the smallest value for which $\widehat{E[\tilde{N}_{c_0}]}$ converges to its limit, $E[N_{c_0}]$, up to Monte Carlo error. In physics terms, this corresponds to the minimal value of R for which $\widehat{E[\tilde{N}_{c_0}]}$ well approximates the number of upcrossings of the underlying continuous time process.

We also investigate the relationship between the width of the signal in the bump-hunting example, and the grid resolution. In particular, we replicate the simulation for three choices of the Gaussian width, namely $\sigma = 0.1\theta, \sigma = 0.5\theta$ and $\sigma = \theta$. (In our actual analysis $\sigma = 0.1\theta$.) As expected, wider signals correspond to smoother underlying processes (Figure 2, left

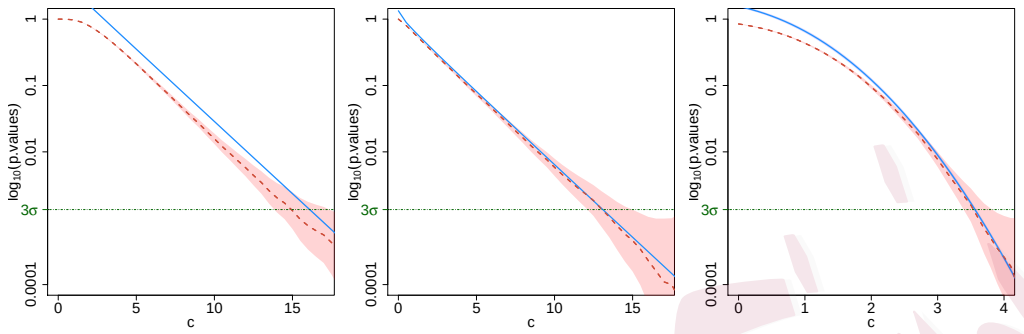

Figure 3: Estimated bound/approximation in (2.13) *(blue solid line), simulated global p-values (on* $\log_{10}$*-scale), Monte Carlo estimates of* $P(\sup_{\theta\in\Theta}\{W(\theta)\} > c)$ *(red dashed line), and Monte Carlo Errors (pink areas) for increasing values of the threshold c, for Example 1 (left panel), Example 2 (central panel) and Example 3 (right panel). Monte Carlo errors associated with* $\widehat{E[\tilde{N}_{c_0}]}$ *on the bound in* (2.13) *are plotted in grey, but are too small to be visible.*

{assess}

panel) and $\widehat{E[\tilde{N}_{c_0}]}$ converges (Figure 2, right panel) at lower grid resolution.

In general, R impacts the upper bound/approximation for the global p-value in (2.8), as well as the observed value of the test statistics, c_R, which we assume converges to c, as $R \to \infty$. Specifically, if the gap between θ_r and θ_{r+1} is wider than the signal width, c_R may underestimate c, and the signal may be missed. Thus, if the signal is suspected to be localized over a small region of the search interval, a higher resolution is required to accurately estimate (2.13) and avoid false negatives, which would in turn adversely affect the power of the test.

Conversely, in Examples 2 and 3, the signal is spread either over the whole parameter space or over a large portion of it. In these cases the

choice of R should be based on the desired level of accuracy of both c_R as an estimate for the maximum of the underlying process and of the value of θ at which the maximum occurs, i.e.,

$$\tilde{\theta} = \text{argmax}_{\theta_r \in \Theta_R}\{W(\theta_r)\}. \qquad (3.18) \quad \{\texttt{thetatilde}\}$$

Finally, based on the elbow in the upcrossing plots in Figures 2 and **??**, the values of R we select are $R = 100$ in Example 1 (with $\sigma = 0.1\theta$ as in (1.1)), $R = 50$ in Example 2, and $R = 30$ in Example 3. In order to guarantee accuracy of at least 0.5 for the identified location, $\tilde{\theta}$, of the break-point, however, we set $R = 50$ in Example 3. For each of the models considered, we computed (2.13) using the R and c_0 selected above. The results obtained are compared in Figure 3 with the Monte Carlo estimates of $P(\sup_{\theta\in\Theta}\{W(\theta)\} > c)$ for increasing values of c, obtained using 100,000 simulations, each of size 10,000. The pink areas correspond to the respective Monte Carlo errors. The Monte Carlo errors associated to the estimate $\widehat{E[\tilde{N}_{c_0}]}$ for $E[\tilde{N}_{c_0}]$ in (2.13) (and displayed on a lower scale in the upcrossing plots) are also incorporated in Figure 3, but they are too small to be visible. As expected, the estimated TOHM bounds approach the "truth"

as $c \to \infty$. Convergence appears to be slower for Example 1. The plots, however, are presented on $\log_{10}$-scale, and thus in all cases we obtain a good approximations of the global p-values.

4. Comparing TOHM and Bonferroni's bounds {analysis}

In fields such as high energy physics and astrophysics, experiments are often characterized by the search of one signal over a wide pool of possibilities. The simplest possible way to tackle this problem using classical Multiple Hypothesis Testing (MHT) is by means of Bonferroni correction (Bonferroni, 1935, 1936). The Bonferroni bound for the global p-value is

$$p_{\mathrm{BF}} = R \cdot \min_{\theta_r \in \Theta_R} P(W(\mathcal{L}) \geq w(\theta_r)) = R \cdot P(W(\mathcal{L}) \geq c_R). \qquad (4.19) \quad \{\texttt{bonfcorr1}\}$$

The standard Bonferoni correction, p_{BF}, used to bound statistical significance in multiple testing also yields a bound on $P(\max_{\theta_r \in \Theta_R}\{W(\theta_r)\} \geq c_R)$. Specifically,

$$P\left(\max_{\theta_r \in \Theta_R}\{W(\theta_r)\} \geq c_R\right) = P\left(\cup_{\theta_r \in \Theta_R}\{W(\theta_r) > c_R\}\right) \leq \sum_{\theta_r \in \Theta_R} P(W(\theta_r) > c_R)$$

$$= R \cdot P(W(\mathcal{L}) > c_R) = p_{BF}.$$

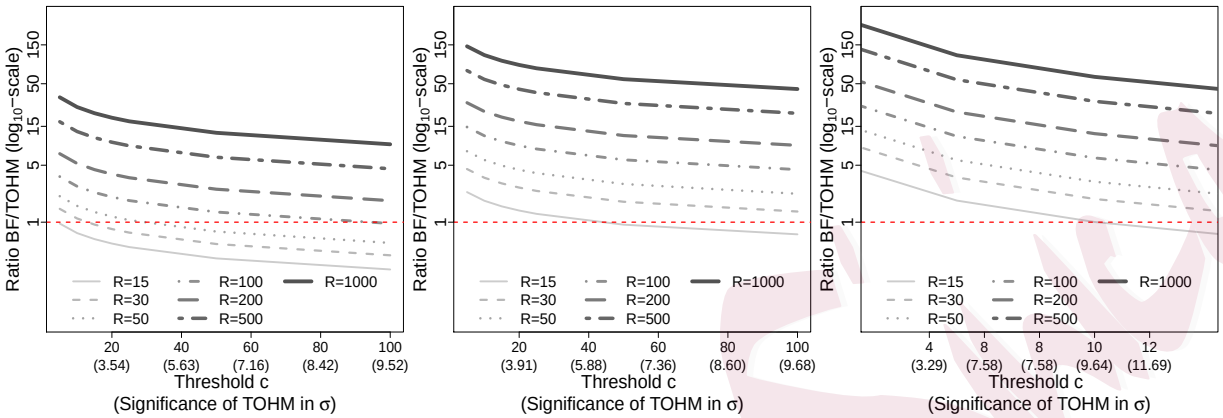

Figure 4: Ratio of Bonferroni and TOHM's bounds at increasing values of c (and corresponding significance for TOHM), and considering different resolutions (grey curves). The left, central and right panels correspond to Example 1, 2 and 3, respectively.

{EVTBonf}

In this section, we investigate the relationship between the TOHM and Bonferroni bounds using simple constructs from EVT in order to individuate situations where the latter can be used without leading to overly conservative results.

First, we introduce the distinction between upcrossings and *exceedances* of $\{W(\theta_r)\}$. Specifically, an exceedance of c_R by $\{W(\theta_r)\}$ occurs at θ_r if $\{W(\theta_r) > c_R\}$. An illustration of the difference between upcrossings and exceedances is given in Figure **??**. We denote by $\tilde{N}_{c_R}$, the process of exceedances of c_R by $\{W(\theta_r)\}$, and let $\dot{N}_{c_R}$ be the process of upcrossings as

defined in 2.3. Notice that

$$E[\dot{N}_{c_R}] = \sum_{r=1}^{R} P\Big(W(\theta_r) \geq c_R\Big) = \sum_{r=1}^{R} P\Big(W(\theta_r) \geq \max_{\theta_{r'} \in \Theta_R} \{w(\theta_{r'})\}\Big) \quad (4.20) \quad \{\texttt{derivationBF1}\}$$

$$= R \min_{\theta_r \in \Theta_R} P(W(\mathcal{L}) \geq w(\theta_r)) = p_{\mathrm{BF}} \quad (4.21) \quad \{\texttt{derivationBF2}\}$$

Because each upcrossing requires at least one exceedance, $E[\dot{N}_{c_R}] \geq E[\tilde{N}_{c_R}]$. Moreover, we expect that the clusters of exceedances corresponding to each upcrossing to be smaller, and consequently $E[\dot{N}_{c_R}]$ to approach $E[\tilde{N}_{c_R}]$ as c_R increases. $E[\dot{N}_{c_R}]$ can be easily computed using p_{BF} in (4.20)-(4.21); whereas, when $\{W(\theta)\}$ satisfies Condition 1, $E[\tilde{N}_{c_R}]$ is approximately equal to the second term in (2.13), for large R. Further, $E[\tilde{N}_{c_R}]$ dominates the first term in (2.13), as $c_R \to \infty$. Thus, it is natural to consider if there are situations where (2.13) and p_{BF} are approximately equivalent bounds on $P(\max_{\theta_r \in \Theta_R}\{W(\theta_r)\} \geq c_R)$, i.e,

$$P(W(\mathcal{L}) > c_R) + \frac{a(c_R)}{a(c_0)} E[\tilde{N}_{c_0}] \approx p_{\mathrm{BF}}, \quad (4.22) \quad \{\texttt{approximation}\}$$

for $c_0 \leq c_R$, $c_R \to +\infty$ and $R \to +\infty$. Unfortunately, simultaneously quantifying the rates at which c_R and R must increase for (4.22) to hold is not an easy task; hence, we investigate the approximation in (4.22) by

means of a numerical simulation where we compare the performance of Bonferroni and the TOHM bounds with respect to the number of tests considered and the level of significance for Examples 1, 2 and 3.

The results are reported in Figure 4, where we plot the ratio of the two bounds for increasing values of c, using different grid sizes, R. Because the signed-root LRT, $\{Q_n(\theta)\}$, is used in Example 3 rather than the LRT, smaller values of c correspond to equally significant results. In the horizontal axes, the statistical significance is reported in terms of σ-significance, i.e., the number of standard deviations from the mean of a standard normal distribution that corresponds to the tail probability expressed by the one-sided p-value, i.e.,

$$\#\sigma = \Phi^{-1}(1 - \text{p-value}),$$

where Φ is the standard normal cumulative function.

In Examples 2 and 3, Bonferroni is always more conservative than the TOHM bound when at least 30 tests are performed. For $R = 15$, Bonferroni becomes less conservative only when the level of significance achieved is of the order of 6σ and 11σ, respectively.

A more interesting situation is observed for Example 1. Here, equiv-

alence of p_{TOHM} and p_{BF} occurs for values of c much smaller than those for which the same limit is achieved in Examples 2 and 3. Further, when $R \leq 50$, Bonferroni quickly becomes less conservative than the TOHM bound as c increases. For $R = 50$ for instance, Bonferroni performs better than TOHM when $c > 30$ ($\sim 4.5\sigma$ significance).

Finally, all the plots in Figure 4 suggest that the TOHM bound is preferable to Bonferroni with very high resolutions, i.e. $R \geq 500$, for all the significance levels considered (up to $\sim 10\sigma$).

It is important to point out that the value of R selected via the upcrossing plots discussed in Section 3.2 is the minimum number of grid points (among those considered) for which $\widehat{E[\tilde{N}_{c_0}]}$ converges to its limit. As R increases beyond this point, the estimated TOHM bound remains constant, whereas Bonferroni's continues to increase. This implies that, when the number of tests to be conducted can be selected arbitrarly, Bonferroni will not be overly conservative if the "elbow" in the upcrossings plot appears at a relatively small value of R and the observed value of c is large. However, practitioners should keep in mind that when attempting to identify the signal location, $\tilde{\theta}$, a higher resolution is typically required and thus TOHM is preferable.

Example	Test	Method	R	c_R	$\tilde{\theta}$	p-value (Significance)
Example 1	$H_0 : \eta = 0$ $H_1 : \eta > 0$	Bonferroni TOHM	100	38.326	3.404	$2.99 \cdot 10^{-8}$ (5.42σ) $2.11 \cdot 10^{-8}$ (5.48σ)
Example 2	$H_0 : \eta = 0$ $H_1 : \eta > 0$ $H_0 : \eta = 1$ $H_1 : \eta < 1$	Bonferroni TOHM Bonferroni TOHM	50 50 	21.021 0.606 	27.265 27.890 	$1.14 \cdot 10^{-4}$ (3.69σ) $2.51 \cdot 10^{-5}$ (4.06σ) > 1 $(0.00\sigma$ $)$ $7.201 \cdot 10^{-1}$ (0.58σ)
Example 3	$H_0 : \xi = 0$ $H_1 : \xi \neq 0$	Bonferroni TOHM	50	11.826	31.266	$1.43 \cdot 10^{-30}$ (11.43σ) $5.06 \cdot 10^{-31}$ (11.52σ)

Table 1: Summary of the results of TOHM and MHT via Bonferroni on real data for Examples 1, 2 and 3.

{real_table}

4.1 Data analyses

{application}

In this section we compare the TOHM and Bonferroni bounds for Examples 1, 2 and 3. The results are summarized in Table 1. In the dark matter search problem of Example 1, we obtain a significance in favour of the presence of a dark matter emission of about 5.4σ using both TOHM and MHT. This result is not surprising since $c_R = 38.326$ and as shown in the central panel of Figure 4, at $c \approx 40$ the gray line associated with $R = 100$ is very close the red dashed line. The signal location selected is close to the truth (3.5GeV), and the estimated model is plotted as a solid red line in the left panel of Figure 1; the signal location selected, $\tilde{\theta} = 3.404$, is indicated by the green

dotted vertical line.

In Example 2 both TOHM and Bonferroni reject the hypothesis that the observed emission is due to a power-law distributed cosmic source at 4.06σ and 3.69σ respectively. Because this example involves a non-nested models comparison, we invert the null of the hypotheses in order to avoid meaningless results (see Section 3.1 for more details). In the inverted test, the power-law model cannot be rejected. Both the fitted dark matter model and the fitted power-law cosmic source model are displayed in the central panel of Figure 1. In Example 2, when testing (1.2), the value of θ (i.e., the signal annihilation of the dark matter model) selected by TOHM is $\tilde{\theta} = 27.265\text{GeV}$. This is somewhat off from the true value used to simulate the data ($\theta = 35\text{GeV}$), perhaps because our analysis does not account for instrumental errors. Our analysis also only uses the spectral energy of the γ-ray signals, whereas in practice the directions of the γ-ray would also be used, thus increasing the statistical power.

Finally, for the break-point regression model in Example 3, both TOHM and MHT give similar inferences (11.52σ and 11.43σ respectively) when rejecting the hypothesis of a linear model with no break-point. The equivalence among the two procedure is likely due to the very high statistical significance, and the only moderately large number of tests conducted

($R = 50$). The fitted model is displayed in Figure 1 where the green triangle corresponds to the optimal break-point location, i.e., the maximum of the signed-root LRT process occurs at a mother's age of 31.266 years.

5. Discussion

{discussion}

In this paper we discuss a highly generalizable method to efficiently conduct statistical tests under non-standard conditions, including bump-hunting, structural change detection and non-nested models comparison.

The main advantages of the method proposed are its easy implementation and its efficiency in providing accurate inference, while controlling for very small Type I errors rates. Following Davies (1987) and Gross and Vitells (2010) we combine the theoretical framework of EVT with the practical simplicity of Monte Carlo simulations and we generalize their results beyond the LRT and χ^2. Using a suite of simulation studies we show that as few as 100 Monte Carlo simulations are often sufficient to achieve a high level of accuracy. Although we do not investigate the power of TOHM here, readers interested in power are directed to Davies (1977) for a formal derivation of lower and upper bounds of the power function in the normal case, or the simulation studies conducted in Algeri et al. (2016) and Algeri et al. (2016) for the $\bar{\chi}^2_{01}$ case.

From a more practical perspective, we propose simple graphical tools

to select the threshold c_0 and to specify an appropriate number of sub-tests R to guarantee robustness of the resulting inference. Finally, we investigate the relationship between the TOHM and Bonferroni bounds and we implement both procedures on our running examples. Extensions of our results to the case where the nuisance parameter specified only under the alternative, $\boldsymbol{\theta}$, is multi-dimensional are the subject of a forthcoming paper (Algeri and van Dyk, 2018).

It is important to point out that the stringent significance requirements play a critical role in both the theory discussed in Section 2 and practical applications. Specifically, this setup is particularly well suited for searches in high energy physics, where the significance level necessary to claim a discovery is of at least 5σ. However, in light of the recent "p-value crisis", culminated with the *Journal Basic and Applied Social Psychology* banning the use of the p-value in future submissions (Wasserstein and Lazar, 2016; Leek and Peng, 2015), stringent significance criteria may become more popular in other scientific communities.

Supplementary Materials In Section S.1 we discuss the error rate of (2.8) for Gaussian, χ^2 and $\bar{\chi}^2_{01}$ processes. Proofs of Result 2 and Result 3 are collected in Section S.2. Additionally figures are reported in Section S.3. Data used in Examples 1 and 2 are also downloadable among the

Supplementary Materials.

Acknowledgements The authors thank two anonymous referees and the associate editor for their constructive feedback. SA and DvD also thank Jan Conrad for the valuable discussion of the physics problems which motivated this work, and Brandon Anderson who provided the Fermi-LAT datasets used in the analyses. Finally the authors acknowledge support from Marie- Skodowska-Curie RISE (H2020-MSCA-RISE-2015-691164) Grant provided by the European Commission.

References

Adler, R. (2000). On excursion sets, tube formulas and maxima of random fields. *Annals of Applied Probability*, 1–74.

Adler, R. and J. Taylor (2009). *Random fields and geometry*. Springer Science & Business Media.

Algeri, S. et al. (2016). On methods for correcting for the look-elsewhere effect in searches for new physics. *Journal of Instrumentation 11*(12), P12010.

Algeri, S., J. Conrad, and D. van Dyk (2016). A method for comparing non-nested models with application to astrophysical searches for new physics.

Monthly Notices of the Royal Astronomical Society: Letters 458(1), L84–L88.

Algeri, S. and D. van Dyk (2018). Testing one hypothesis multiple times: The multidimensional case. *arXiv:1803.03858*.

Anderson, B. et al. (2016). Search for gamma-ray lines towards galaxy clusters with the fermi-lat. *Journal of Cosmology and Astroparticle Physics 2016*(02), 026.

Andrews, D. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica: Journal of the Econometric Society*, 821–856.

Andrews, D. and W. Ploberger (1994). Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica: Journal of the Econometric Society*, 1383–1414.

Atkinson, A. (1970). A method for discriminating between models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 323–353.

Atwood, W. B. et al. (2009). The large area telescope on the fermi gamma-ray space telescope mission. *The Astrophysical Journal 697*(2), 1071.

Bergström, L., P. Ullio, and J. Buckley (1998). Observability of γ rays from

dark matter neutralino annihilations in the milky way halo. *Astroparticle Physics 9*(2), 137–162.

Bonferroni, C. (1935). Il calcolo delle assicurazioni su gruppi di teste. In *Studi in Onore del Professore Salvatore Ortu Carboni*, pp. 13–60. Rome.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8*, 3–62.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics*, 573–578.

Choudalakis, G. (2011). On hypothesis testing, trials factor, hypertests and the BumpHunter. In *Proceedings, PHYSTAT 2011. ArXiv:1101.0390.*

Cox, D. (1962). Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society. Series B (Methodological)*, 406–424.

Cox, D. (2013). A return to an old paper:'tests of separate families of hypotheses'. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(2), 207–215.

Cramér, H. and M. Leadbetter (2013). *Stationary and related stochastic*

processes: Sample function properties and their applications. Courier Corporation.

Davies, R. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika 64*(2), 247–254.

Davies, R. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika 74*(1), 33–43.

Davies, R. (2002). Hypothesis testing when a nuisance parameter is present only under the alternative: linear model case. *Biometrika*, 484–489.

Efron, B. and R. Tibshirani (1994). *An introduction to the bootstrap*. CRC press.

Gross, E. and O. Vitells (2010). Trial factors for the look elsewhere effect in high energy physics. *The European Physical Journal C 70*(1-2), 525–530.

Hansen, B. (1991). Inference when a nuisance parameter is not identified under the null hypothesis. *Rochester Center for Economic Research Working Paper No. 296*.

Hansen, B. (1992a). The likelihood ratio test under nonstandard conditions: testing the markov switching model of gnp. *Journal of applied Econometrics 7*(S1).

Hansen, B. (1992b). Testing for parameter instability in linear models. *Journal of policy Modeling 14*(4), 517–533.

Hansen, B. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica: Journal of the econometric society*, 413–430.

Hansen, B. (1999). Threshold effects in non-dynamic panels: Estimation, testing, and inference. *Journal of econometrics 93*(2), 345–368.

Hotelling, H. (1939). Tubes and spheres in n-spaces, and a class of statistical problems. *American Journal of Mathematics 61*(2), 440–460.

James, G. et al. (2013). *An introduction to statistical learning*, Volume 112. Springer.

Leek, J. and R. Peng (2015). Statistics: P values are just the tip of the iceberg. *Nature 520*(7549), 612.

Lyons, L. (2013). Discovering the significance of 5 sigma. *arXiv preprint arXiv:1310.1284*.

Muggeo, V. et al. (2008). Segmented: an r package to fit regression models with broken-line relationships. *R news 8*(1), 20–25.

Quandt, R. (1974). A comparison of methods for testing nonnested hypotheses. *The Review of Economics and Statistics*, 92–99.

Rice, S. (1944). Mathematical analysis of random noise. *Bell Labs Technical Journal 23*(3), 282–332.

Self, S. and K.-Y. Liang (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association 82*(398), 605–610.

Taylor, J. and R. Adler (2003). Euler characteristics for gaussian fields on manifolds. *Annals of Probability*, 533–563.

Taylor, J. and K. Worsley (2008). Random fields of multivariate test statistics, with applications to shape analysis. *The Annals of Statistics*, 1–27.

van Dyk, D. (2014). The role of statistics in the discovery of a higgs boson. *Annual Review of Statistics and Its Application 1*, 41–59.

Wasserstein, R. and N. Lazar (2016). The asa's statement on p-values: context, process, and purpose.

Sara Algeri

School of Statistics

University of Minnesota

Minneapolis, MN, 55455

E-mail: salgeri@umn.edu

David van Dyk

Statistics Section

Dept of Mathematics

Imperial College London

London, UK SW7 2AZ

E-mail: d.van-dyk@imperial.ac.uk