Correcting for the look-elsewhere effect: why, when and how

Sara Algeri

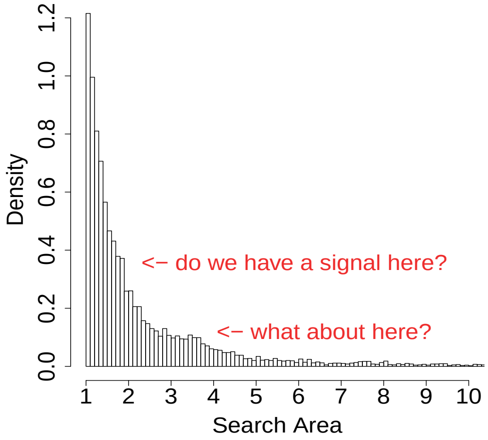
School of Statistics, University of Minnesota

PHYSTAT Dark Matter 2019,
Stockholm University.

July 31, 2019.

The physics problem

We would like to detect the signal of a new particle/astronomical source/astrophysical phenomenon BUT we do not known its location.



Let's consider the simplest possible scenario

- **Underlying assumptions:** (i) we search for at most one signal, (ii) the only unknown parameters are the intensity and the location, (iii) we know both the background and signal distributions (up to some free parameters).
- **Model:**

$$(1-\eta)\underbrace{f(y,\phi)}_{\text{background}} + \underbrace{\eta}_{\text{signal relative intensity}} \underbrace{g(y,\overbrace{\theta}^{\text{signal location}})}_{\text{bump}} \qquad 0 \leq \eta \leq 1 \tag{1}$$

- **Test**

$$H_0 : \eta = 0 \quad \text{versus} \quad H_1 : \eta > 0$$

- **Test statistics:**

$$LRT = -2\log[\underbrace{L(0,\hat{\phi}_0,\text{-})}_{\text{Likelihood under } H_0} - \underbrace{L(\hat{\eta},\hat{\phi},\theta)}_{\text{Likelihood under } H_1}] \tag{2}$$

Why can't we just use Wilks or similar results?

- **If θ is fixed** $\Rightarrow$ under H_0, $LRT(\theta) \xrightarrow[n\to\infty]{d} \frac{1}{2}\chi^2_1 + \frac{1}{2}\delta(0)$

  (Chernoff, 1954, Self and Liang, 1987).

- **If θ is not fixed** $\Rightarrow$ under H_0, $LRT(\theta) \not\xrightarrow[n\to\infty]{d}$ **???**

  (**non-identifiability**, i.e., θ can take any value when $\eta = 0$).

- **If we define a grid of possible locations** $\Theta_R = \{\theta_1, \ldots, \theta_R\}$, for all $\theta_r \in \Theta_R$ we calculate R local p-values p_r

(**multiple comparisons problem**, i.e., our p-values must be corrected for the fact that many tests are conducted simultaneously).

How can we tackle this problem?

- Approach 1: **Multiple hypotesis testing** $\Rightarrow$ that's how the
  e.g., Bonferroni's correction. **Look-Elsewhere Effect** (**LEE**) problem was originally formulated.

- Approach 2: **Simulations/resampling**
  e.g., Monte Carlo, Bootstrap methods.

- Approach 3: **Extreme value theory/Random fields theory**
  e.g., Gross and Vitells (2010) $\Rightarrow$ nowadays this paper is essentially synonym of **LEE**.

Multiple hypothesis testing

In his Class Notes for Statistics 411 at Princeton University in 1976, John Tukey introduced the problem of multiple comparisons by means of a story:

A young psychologist administers many hypothesis tests as part of a research project, and finds that, out of 250 tests 11 were significant at the 5% level. The young researcher feels very proud of this fact and is ready to make a big deal about it, until a senior researcher suggests that one would expect 12.5 significant tests even in the purely null case, merely by chance. In that sense, finding only 11 significant results is actually somewhat disappointing!

Local p-values and type I error

- We have an ensemble of R local p-values $p_1, \ldots, p_r, \ldots, p_R$.
- The smallest, names p_{L} is then compared with the target probability of type I error α_{L}.
- But what is α_{L} if we want to claim a discovery at 5σ?

Global and local probability of false detection

$\alpha_{\mathrm{L}} =$ specific probability of false detection for each of the R tests.

$\neq$

$\alpha_{\mathrm{G}} = 1 - \Phi(5) = 2.87 \cdot 10^{-7} =$ probability of having at least one false detection over the whole ensemble of R tests.

$\Rightarrow$ we must correct p_{L} accordingly

The simplest corrections for the local p-values

- **If the R tests were independent**

$$\alpha_{\mathrm{G}} = 1 - (1-\alpha_{\mathrm{L}})^R \quad \Rightarrow \quad \boxed{p_{\mathrm{G}} = 1 - (1-p_{\mathrm{L}})^R \quad \text{Sidak's correction}} \qquad (3)$$

E.g.: Suppose we are conducting $R = 50$ simultaneous test, each of them at 5σ

$$\alpha_{\mathrm{L}} = 1 - \Phi(5) \quad \Rightarrow \text{ by (3):} \quad \alpha_{\mathrm{G}} = 1 - \Phi(4.18)$$

i.e., $\frac{\alpha_{\mathrm{G}}}{\alpha_{\mathrm{L}}} \approx 50$.

- **If the R tests were dependent** (which is generally the case)

$$\alpha_{\mathrm{G}} \leq R\alpha_{\mathrm{L}} \quad \Rightarrow \quad \boxed{p_{\mathrm{B}F} = Rp_{\mathrm{L}} \quad \text{Bonferroni's correction}} \qquad (4)$$

Pros: Easy to implement. **Cons:** May be overly conservative.

The LEE as a non-identifiability problem

- **If θ is not fixed** $\Rightarrow$ under H_0, $LRT(\theta) \underset{n\to\infty}{\overset{d}{\not\longrightarrow}}$ **???**

What about simulating the distribution of $LRT(\theta)$?

Pros: Easy to implement. **Cons:** May be computationally intensive.

What about looking at $LRT(\theta)$ as a stochastic process?

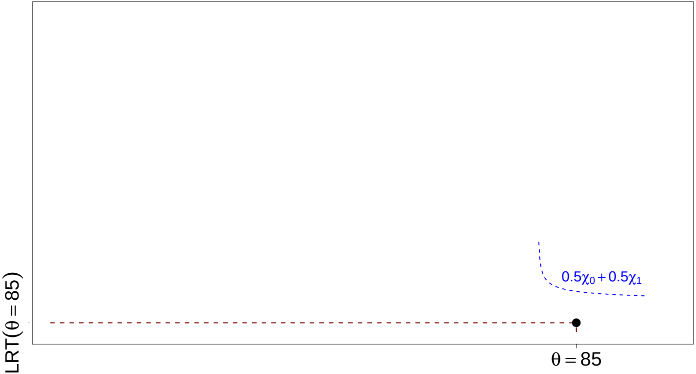
$\Rightarrow$ Extreme value theory $\Rightarrow$ **Gross and Vitells, 2010**.

Let's see what Gross and Vitells, 2010 is all about...

The rationale (1)

If θ is fixed, under H_0

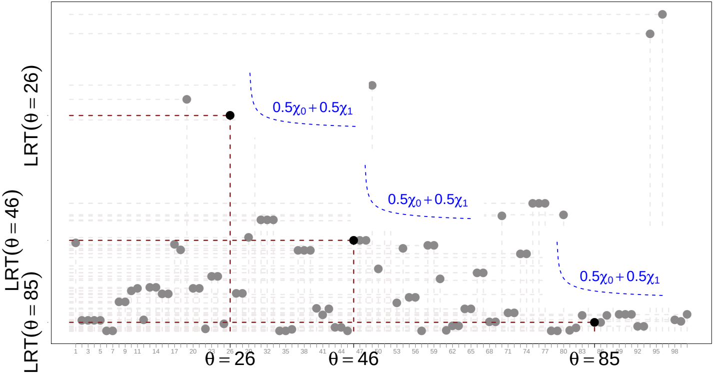
$$LRT(\theta) \xrightarrow[n\to\infty]{d} \frac{1}{2}\chi^2_1 + \frac{1}{2}\delta(0)$$



The rationale (2)

If θ is fixed, under H_0,

$$LRT(\theta) \xrightarrow[n\to\infty]{d} \frac{1}{2}\chi^2_1 + \frac{1}{2}\delta(0)$$

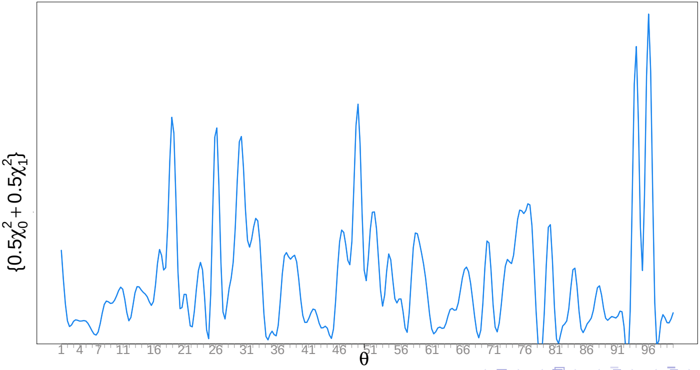


The rationale (3)

If we let θ vary, it can be shown (e.g., Ghosh and Sen, 1985) that under suitable regularity conditions, if H_0 is true,

$$\{LRT(\theta), \theta \in \Theta\} \xrightarrow[n\to\infty]{d} \left\{\frac{1}{2}\chi^2_1 + \frac{1}{2}\delta(0)\right\} \tag{5}$$

(i.e., the LRT process is asymptotically, under H_0, a $\bar{\chi}^2_{01}$).



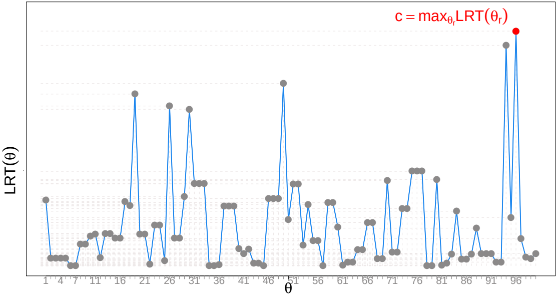
The extreme value theory part (1)

- Define a grid $\boldsymbol{\Theta}_R$ of R values θ_r over the search region $\boldsymbol{\Theta}$.
- For all $\theta_r \in \boldsymbol{\Theta}_R$ calculate $LRT(\theta_r)$.

Combine the R values $LRT(\theta_r)$ in a unique test statistic

$$c = \max_{\theta_r \in \boldsymbol{\Theta}_R}\{LRT(\theta_r)\}$$

This is essentially the maximum of the stochastic process $\{LRT(\theta_r)\}$!



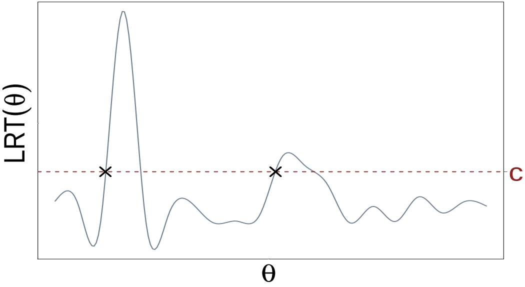
The extreme value theory part (2)

The p-value...

The **p-value** of our test $H_0 : \eta = 0$ versus $H_a : \eta > 0$ is in the form

$$P(\sup_{\theta \in \Theta}\{LRT(\theta)\} > c) \tag{6}$$

...which we must calculate/approximate somehow!



To do so we consider the **number of upcrossings** (N_c) of c by the process $\{LRT(\theta)\}$ when H_0 is true.

E.g. In this image $N_c = 2$.

Approximation of $P(\sup_{\theta\in\Theta}\{LRT(\theta)\} > c)$

- From **Davies, 1987** we have that as $c \to +\infty$ and additional regularity conditions:

$$P(\sup_{\theta\in\Theta}\{LRT(\theta)\} > c) \approx \frac{P(\chi^2_1 > c)}{2} + \frac{e^{-\frac{c}{2}}}{\sqrt{2\pi}} \int_L^U \kappa(\theta) d\theta \quad \text{Expected \# of upcrossings over } c \text{ of the LRT process under } H_0 \qquad (7)$$

- if $c \not\to +\infty \Rightarrow$ we have an upper bound for $P(\sup LRT(\theta) > c)$.
- $\kappa(\theta)$ is complicated $\Rightarrow$ (7) proposed in **Gross and Vitells, 2010**

$$P(\sup_{\theta\in\Theta}\{LRT(\theta)\} > c) \approx \frac{P(\chi^2_1 > c)}{2} + \underbrace{e^{-\frac{c-c_0}{2}} E[N_{c_0}|H_0]}_{=E[N_c|H_0]} \quad \text{Expected \# of upcrossings over } c_0 \text{ of the LRT process under } H_0 \qquad (8)$$

- where $c_0 << c$ and $E[N_{c_0}|H_0]$ is estimated using (few) Bootstrap simulations.

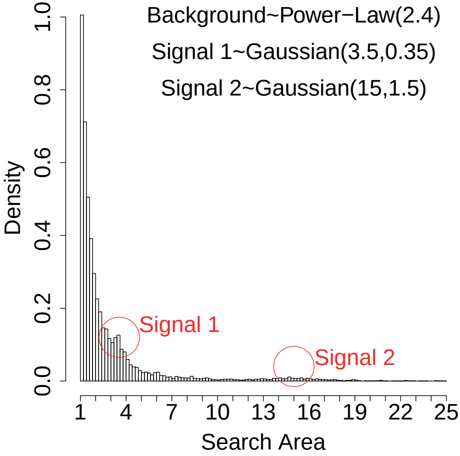
Pros: Efficient. **Cons:** Relies on "unspoken" regularity conditions.

Ok, but what if we are dealing with more complex situations?

	Multiple Hypothesis Testing	**Simulation Methods (MC/Bootstrap)**	**Random Fields or Extreme Value Theory**
Multidimensional searches	✓ May become even more conservative	✓ May become unfeasible	✓ Use Vitells and Gross (2011) or Algeri and van Dyk (2018) instead
Multiple signals	✓ Either you know how many or they do not overlap	✓ Either you know how many or they do not overlap	✓ If you know how many or they do not overlap use upper limits instead of p-values
Using a different test statistics than the LRT	✓	✓	✓ Use Algeri and van Dyk (2017-2018) or Pilla et al. (2005) instead
Bkg and/or signal models are unknown	X	X	X

I will cover this in my talk tomorrow at 11AM (the solution proposed automatically deals with the multiple signals setting as well).

Example: detecting non-overlapping signals



Upper limits construction

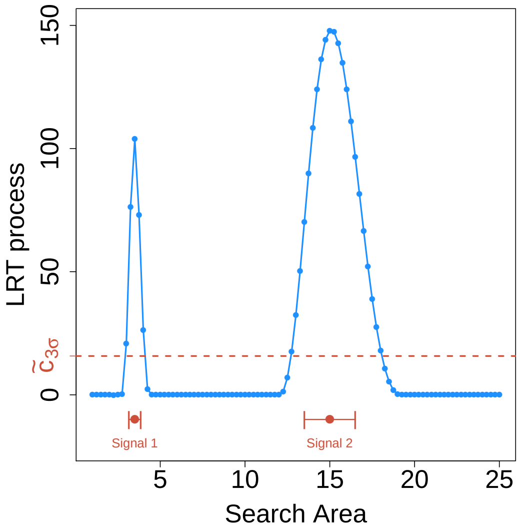
- **Model:** $(1-\eta)\frac{e^{-2.4}}{k_1} + \eta\frac{e^{-\frac{(x-\theta)^2}{0.2\theta^2}}}{k_2}$, with k_1, k_2 normalizing constants.
- From **Gross and Vitells, 2010** we have that, for large c_α (and small α)

$$P(\sup_{\theta\in\Theta}\{LRT(\theta)\} \leq c_\alpha) \approx \frac{P(\chi^2_1 > c_\alpha)}{2} + e^{-\frac{c_\alpha - c_0}{2}} E[N_{c_0}|H_0]$$

- **Approximated α-quantile:** Find $\tilde{c}_\alpha$ which satisfies

$$\alpha \approx \frac{P(\chi^2_1 > \tilde{c}_\alpha)}{2} + e^{-\frac{\tilde{c}_\alpha - c_0}{2}} E[N_{c_0}|H_0]$$

3σ upper limit on the LRT process



Summary

- Given their simple implementation and wide applicability, multiple hypothesis testing correction methods such as Bonferroni should always be implemented.
- If we can afford a large simulation, then you should proceed with that as we can easily control the error of the approximation (MC error) and no regularity conditions are required.
- If we cannot find anything with multiple hypothesis testing, and if a simulation would be too expensive, then we can use Gross and Vitells (2010) or similar approaches, but we must keep in mind that these procedure rely on regularities conditions which should be assessed.
- For all the methods discussed here, **we must know the background and signal models!**

References

- Algeri S., van Dyk D.A., Conrad J.and Brandon, A. *Looking for a Needle in a Haystack? Look Elsewhere! A statistical comparison of approximate global p-values.* Journal of Instrumentation, 2016.
- Algeri S. and van Dyk D.A. *Testing one hypothesis multiple times.* Under third review, 2017.
- Algeri S. and van Dyk D.A. *Testing one hypothesis multiple times: the multidimensional case.* Under third review, 2018.
- H. Chernoff. *On the Distribution of the Likelihood Ratio*. The Annals of Mathematical Statistics, 1954.
- R. B. Davies. *Hypothesis Testing when a Nuisance Parameter is Present Only Under the Alternatives*. Biometrika, 1987.
- J.K. Ghosh and P.K. Sen. *On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results*. Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, 1985.
- E. Gross and O. Vitells. *Trial factors for the look elsewhere effect in high energy physics*. The European Physical Journal C, 2010.
- R.S. Pilla et al. *New Technique for Finding Needles in Haystacks: Geometric Approach to Distinguishing between a New Source and Random Fluctuations*. Physical Review Letters, 2005.
- S.G. Self and K.-Y. Liang. *Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions*. Journal of the American Statistical Association, 1987.
- O. Vitells and E. Gross. *Estimating the significance of a signal in a multi-dimensional search*. Astroparticle Physics, 2011.