Semiparametric signal detection under unknown background

Aritra Banerjee

School of Statistics
University of Minnesota, Twin Cities

Joint work with Sara Algeri (University of Minnesota, Twin Cities), Lydia Brenner (Dutch National Institute for Subatomic Physics) and Oliver Reiger (Dutch National Institute for Subatomic Physics)

Agenda

Introduction

Discovery of new physics is formulated as testing for the presence of a specific signal of interest in the data observed from an experiment.

However, the data collected from the experiments also contains signals from a range of nuisance sources that constitutes the background.

Therefore, the density of the data generating process can be looked as a convex combination of the signal density f_s and a background density f_b

Introduction

The density of the data generating process f can be written as

$$f(x;\eta) = \eta \cdot f_s(x) + (1-\eta) \cdot f_b(x); \;\; x \in [\mathcal{L}, \mathcal{U}]; \;\; \eta \in [0,1)$$

where $[\mathcal{L}, \mathcal{U}]$ is the search region, η is the signal proportion or signal strength.
Note that, 1 is outside the range of η as $\eta = 1$ suggests a data with only signal which is unrealistic

The signal search is formulated as the following hypothesis test,

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

- In our framework, we assume that we do not have access to a background only sample and the background density f_b is unknown
- We carry out estimation and test for η using only the physics sample (data generated from the experiments that may or may not contain the signal)

Mathematical Framework

- We start with a proposal background density g_b which acts as a proxy for the unknown density f_b
- Let $\{1, S_1, T_1, T_2, \cdots\}$ be an orthonormal basis on $L_2(G_b)$ where,

$$S(x) = \frac{f_s(x)}{g_b(x)} - 1$$

and $S_1(x) = S(x)/\|S\|_{G_b}$. S captures the deviation from g_b in the direction of the signal

Mathematical Framework

- We can express f_b as,

$$\frac{f_b(x)}{g_b(x)} = 1 + \sum_{j=1}^{\infty} \beta_j T_j(x) + \delta \cdot S_1(x)$$

where δ is departure from the background in the direction of the signal f_s

- Plugging above in $f = \eta \cdot f_s + (1-\eta) \cdot f_b$ we get,

$$\frac{f(x)}{g_b(x)} = 1 + \sum_{j=1}^{\infty} \tau_j T_j(x) + \theta \cdot S_1(x)$$

where,

$$\tau_j = (1-\eta) \cdot \beta_j;$$

$$\theta = \eta \cdot \|S\|_{G_b} + (1-\eta) \cdot \delta$$

Estimation

We have,

$$\theta = \left\langle \frac{f}{g_b}, S_1 \right\rangle_{G_b} = \int S_1(x)dF(x)$$

$$\implies \quad \hat{\theta} = \int S_1(x)d\mathbb{F}(x) = \frac{1}{n}\sum_{i=1}^{n} S_1(X_i)$$

$$\tau_j = \left\langle \frac{f}{g_b}, T_j \right\rangle_{G_b} = \int T_j(x)dF(x)$$

$$\implies \quad \hat{\tau}_j = \int T_j(x)d\mathbb{F}(x) = \frac{1}{n}\sum_{i=1}^{n} T_j(X_i)$$

where $\mathbb{F}$ is the empirical estimate of the mixture CDF F.

Empirical estimators have a tractable asymptotic distribution.

However,

Estimation

$$\theta = \eta \cdot \|S\|_{G_b} + (1-\eta)\cdot \delta \implies \eta = \frac{\theta - \delta}{\|S\|_{G_b} - \delta}$$

δ is not estimable since,

$$\delta = \left\langle \frac{f_b}{g_b}, S_1 \right\rangle_{G_b} = \int S_1(x)d\,F_b(x)$$

Alternative: We simply ignore δ

$$\tilde{\eta} = \frac{\theta}{\|S\|_{G_b}} \implies \widehat{\tilde{\eta}} = \frac{\hat{\theta}}{\|S\|_{G_b}}$$

It is easy to show that $\delta \leq 0 \implies \tilde{\eta} \leq \eta$ i.e. $\widehat{\tilde{\eta}}$ is a conservative estimate for η

Signal Search

The signal search is formulated as the following test

$$H_0 : \eta = 0 \;\text{ vs }\; H_1 : \eta > 0$$

But we propose the test

$$H_0 : \tilde{\eta} = 0 \;\text{ vs }\; H_1 : \tilde{\eta} > 0$$

Which is again a conservative test given we have $\delta \leq 0$ It is equivalent to testing,

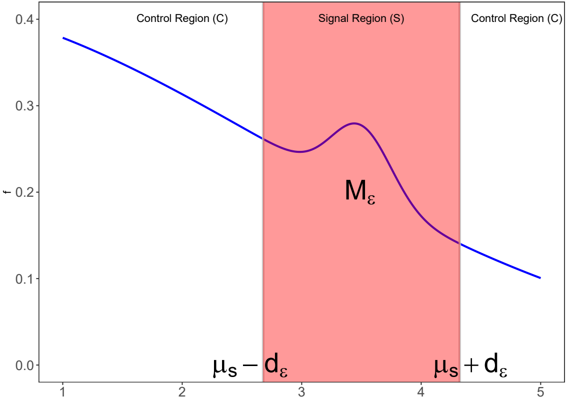
$$H_0 : \theta = 0 \;\text{ vs }\; H_1 : \theta > 0$$

The obvious test statistic would be

$$\mathcal{Z} = \frac{\hat{\theta}}{s.e.(\hat{\theta})} \xrightarrow{H_0} \mathcal{N}(0,1) \quad \text{where} \quad s.e.(\hat{\theta}) = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n} S_1^2(X_i) - \hat{\theta}^2}{n}}$$

How to ensure $\delta \leq 0$?

Consider a narrow enough interval around the signal location μ_s say, $M_\epsilon = (\mu_s - d_\epsilon, \mu_s + d_\epsilon)$ such that $F_s(M_\epsilon) = 1 - \epsilon$ for small enough ϵ



How to ensure $\delta \leq 0$?

Now, we can show

$$\sup_{M_\epsilon^c} \frac{f_b(x)}{g_b(x)} \cdot \epsilon + \sup_{M_\epsilon} \frac{f_b(x)}{g_b(x)} \cdot (1-\epsilon) \leq 1 \implies \delta \cdot \|S\|_{G_b} \leq 0$$

Therefore, we need:

- small ϵ
- $\sup_{M_\epsilon} \frac{f_b}{g_b}$ is preferably below 1 : g_b should dominate f_b in majority of M_ϵ (if not completely)
- $\sup_{M_\epsilon^c} \frac{f_b}{g_b}$ is not too large : M_ϵ should be narrow enough (which works for signals with localized peak)

How to ensure $\delta \leq 0$?

- One way to ensure $\delta \leq 0$ is to make sure that g_b dominates f_b in majority of M_ϵ
- One way to achieve that is to give g_b a wide bump throughout the signal region M_ϵ and make sure it cuts through the bump in f

How to ensure $\delta \leq 0$?

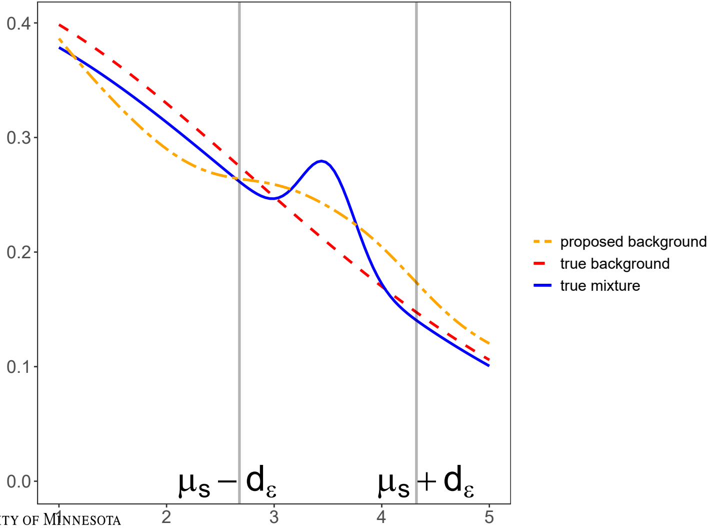
One possible way to construct g_b:

$$g_b(x) = \lambda \cdot \left[fs(x, \mu_1, \sigma) + fs(x, \mu_2, \sigma) \right] + (1 - 2\lambda) q_b(x); \quad 0 < \lambda < \frac{1}{2}$$

- f_s : density from the same family as the signal
- $q_b(x)$: guess for the smooth background density or the available description of f_b
- μ_1 : Slightly above $\mu_s - d$; μ_2 : Slightly below $\mu_s + d$;
- σ : scale parameter larger than that of the signal
- λ : mixing proportion

How to ensure $\delta \leq 0$?

We want something like this



What if the data is binned?

- The methodology described thus far applies to unbinned data
- To deal with the binned data $(n_1, x_1), (n_2, x_2), \cdots, (n_k, x_k)$ we need to modify our estimator $\hat{\theta}$
- We propose the estimators

$$\hat{\theta}_{(b)} = \frac{1}{N}\sum_{i=1}^{k} S_1(x_i) n_i \quad \text{and} \quad \widehat{\tilde{\eta}}_{(b)} = \frac{\hat{\theta}_{(b)}}{\|S\|_{G_b}}$$

- As long as we are estimating $\tilde{\eta}$ and testing $\theta = 0$, we are still performing a conservative inference given $\delta < 0$

Signal Search for Binned Data

- One can show that $\hat{\theta}_{(b)}$ is consistent for θ
- We can decompose $\hat{\theta}_{(b)}$

$$\hat{\theta}_{(b)} = \frac{T}{N} \cdot \frac{k}{T} \cdot \frac{1}{k} \sum_{i=1}^{k} n_i S_1(x_i) = \frac{T}{N} \cdot \frac{k}{T} \cdot \check{\theta}$$

- To test $H_0 : \theta = 0$ against $H_1 : \theta > 0$ we can use the test statistic

$$\mathcal{Z}_{(b)} = \frac{k \cdot \check{\theta}}{\sqrt{\sum_{i=1}^{k} n_i S_1^2(x_i)}} \xrightarrow{H_0} \mathcal{N}(0, 1)$$

Safeguard can be anticonservative

- The *safeguard method* by Priel et al. (2017) and the *spurious signal method* by Aad et al. (2014) are two similar methods used at the ATLAS collaboration for particle discovery
- Both of them try to obtain a conservative estimate of the background distribution by accounting for signal fluctuations into them

SAFEGUARD CAN BE ANTICONSERVATIVE

- In the safeguard method, first the background is estimated by fitting the following model on a signal free calibration dataset

$$g_b^{(sf)}(x) = \epsilon \cdot f_s(x) + (1-\epsilon) \cdot f_b^{(model)}(x)$$

- In the context of the safeguard method as well, we have the δ

$$\frac{f_b(x)}{g_b^{(sf)}(x)} = 1 + \sum_{j=1}^{\infty} \beta_j T_j(x) + \delta^{(sf)} S_1^{(sf)}(x)$$

where

$$S^{(sf)}(x) = \frac{f_s(x)}{g_b^{(sf)}(x)} - 1 \text{ ; and } S_1^{(sf)}(x) = \frac{S^{(sf)}(x)}{\|S^{(sf)}\|_{G_b^{(sf)}}}$$

Safeguard can be anticonservative

- Then, the final model that is fit on the physics data using maximum likelihood estimation to obtain $\hat{\eta}_{MLE}$

$$\tilde{f}(x) = \eta \cdot f_s(x) + (1 - \eta) \cdot g_b^{(sf)}(x)$$

- One can show that if $\delta^{(sf)} > 0$ the asymptotic limit of $\hat{\eta}_{MLE}$ i.e. η^* is strictly positive even when there is no signal

Numerical Results: Simulation

- For demonstration purposes we have chosen a scenario where $\delta^{(sf)} > 0$
- The data generation process is

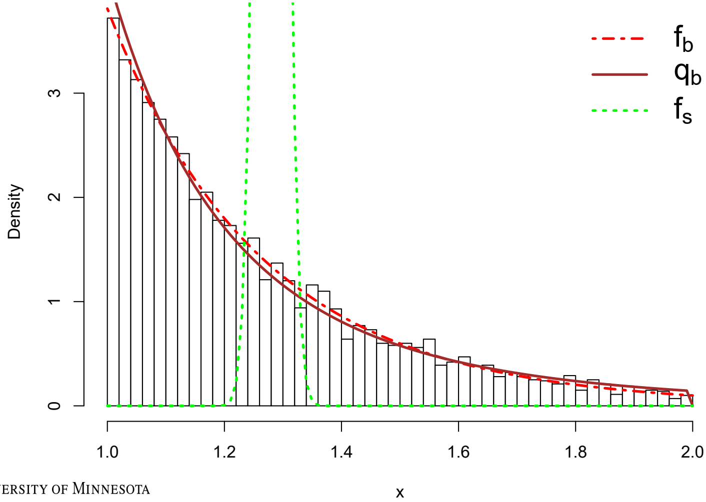
$$f(x;\eta) = \frac{\eta}{c_1} exp\left\{-\frac{(x-1.28)^2}{2\cdot(0.02)^2}\right\} + \frac{(1-\eta)}{c_2}\frac{e^{-3.3x}}{\sqrt{x}};\;\; 1 \leq x \leq 2$$

- While fitting the model, we have used a common guess for the smooth background

$$q_b(x) = f_b^{(model)}(x) = C\cdot x^{-3.87-1};\;\; 1 \leq x \leq 2$$

- We implemented our method using g_b with $\lambda = 0.002, 0.005, 0.007$ and 0.01.

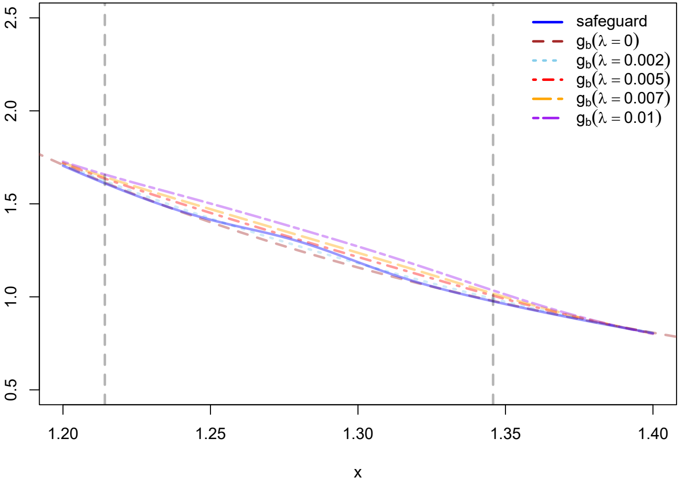
Numerical Results: Simulation



NUMERICAL RESULTS: SIMULATION

Method	δ	**Type I error**	**Power** ($\eta = 0.01$)
safeguard	0.013	0.242	0.923
Unbinned Method			
$g_b(\lambda = 0.002)$	0.016	0.291	0.942
$g_b(\lambda = 0.005)$	0.008	0.128	0.840
$g_b(\lambda = 0.007)$	0.002	0.064	0.732
$g_b(\lambda = 0.01)$	-0.006	0.017	0.520

Numerical Results: Simulation



Numerical Results: VBF Analysis

- From our ATLAS collaborators, we obtained data (binned) on Higgs to dimuon decay ($pp \to H \to \mu\mu$) via the Vector Boson Fusion (VBF) process
- Data generating process:

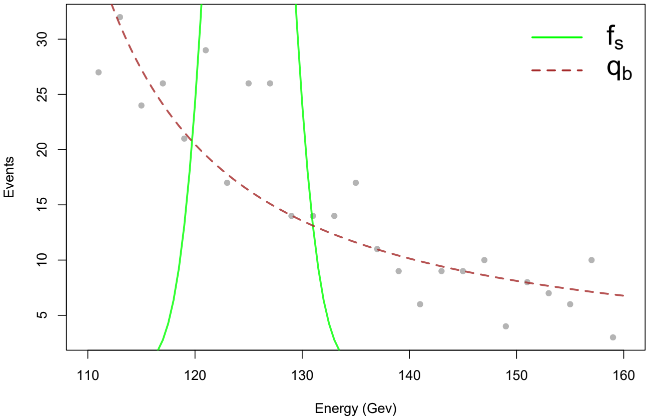
$$f(x;\eta) = \frac{\eta}{c_1} exp\left\{-\frac{(x-125)^2}{2\cdot(3)^2}\right\} + \frac{(1-\eta)}{c_2} f_b(x); \quad 110 \leq x \leq 160$$

- Benchmark background density:

$$q_b(x) \propto \frac{1}{(x-91.2)^2+\left(\frac{2.49}{2}\right)^2} + x^{-1.55}; \quad 110 \leq x \leq 160$$

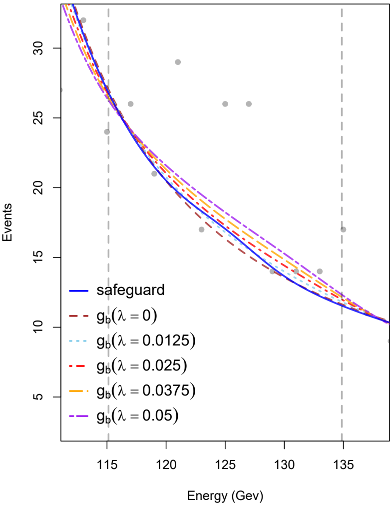
- We had 4 categories of VBF data. We found significant amount of signal in the *VBF - medium* category.

Numerical Results: VBF Analysis



Numerical Results: VBF Analysis

λ in g_b	$\hat{\eta}$	p-value
0	0.096	0.0017
0.0125	0.088	0.0037
0.025	0.080	0.0077
0.0375	0.072	0.0150
0.05	0.064	0.0278
safeguard	0.089	0.0020



Discussion

- We have identified the central role of δ in determining validity of inference for signal detection under unknown background
- Safeguard and spurious signal method try to be conservative but can still end up with a positive δ
- We propose a heuristic approach to construct the proposal background via a sensitivity analysis to ensure $\delta < 0$ and perform a conservative inference without any prior knowledge about the true background density
- The better is the guess q_b for the background, the more reliable is our sensitivity analysis
- For cases where the large sample assumption is not met, we can approximate the null distribution using smoothed bootstrap

Acknowledgment

I would like to thank the CHASC Collaboration for inviting me to present my work and the DSMMA (Data Science in Multi-Messenger Astrophysics) program at UMN for supporting this visit.

Thank You!!

Numerical Results: Simulation

Method	δ	Type I error	Power $\eta = 0.005$	Power $\eta = 0.01$	Power $\eta = 0.02$
safeguard	0.01340076	0.2417	0.6487	0.9226	0.99975
Unbinned Method					
$g_b(\lambda = 0.002)$	0.01624906	0.29077	0.70299	0.94227	0.99986
$g_b(\lambda = 0.005)$	0.007890316	0.12836	0.48119	0.84015	0.99875
$g_b(\lambda = 0.007)$	0.002424086	0.06422	0.33311	0.73232	0.99582
$g_b(\lambda = 0.01)$	-0.005624272	0.01748	0.15643	0.52019	0.98037
Binned Method					
$g_b(\lambda = 0.002)$	0.01624906	0.29379	0.70093	0.94087	0.99981
$g_b(\lambda = 0.005)$	0.007890316	0.12911	0.47856	0.8380	0.99862
$g_b(\lambda = 0.007)$	0.002424086	0.06385	0.33187	0.72815	0.99541
$g_b(\lambda = 0.01)$	-0.005624272	0.01738	0.1550	0.51609	0.97929

Tail in the bump