Detecting new signals under background mismodelling

Sara Algeri

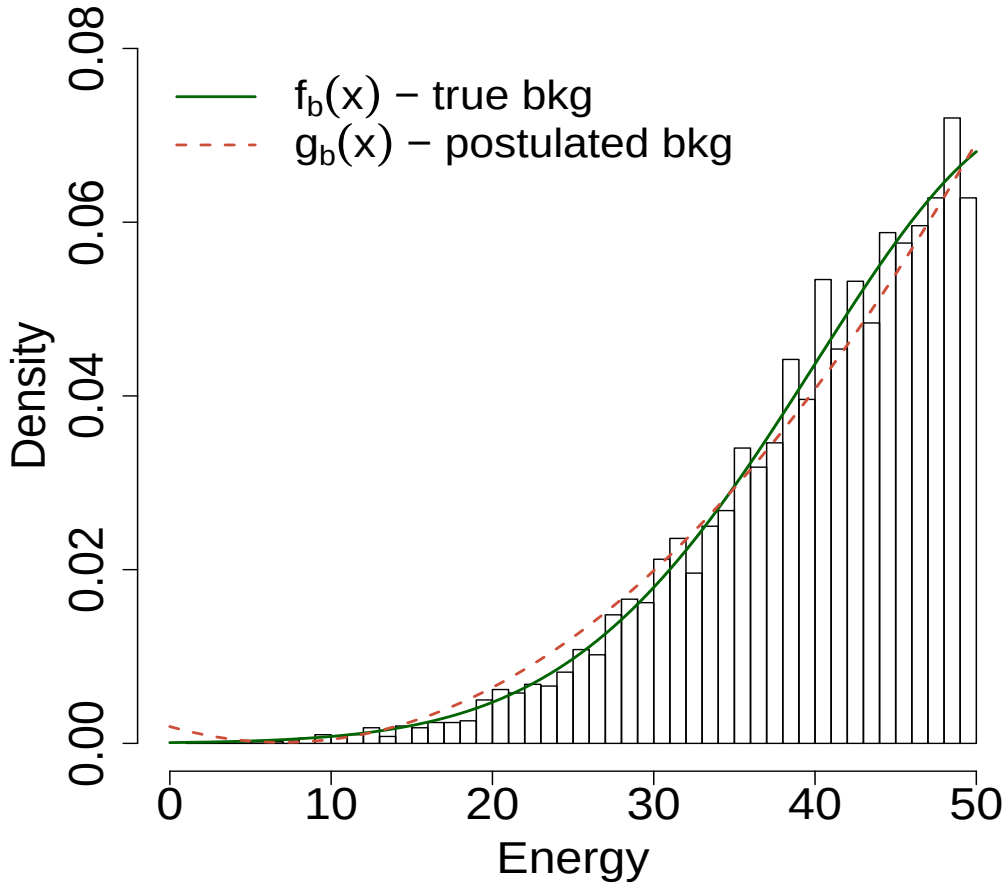
School of Statistics, University of Minnesota

PHYSTAT Dark Matter 2019,
Stockholm University.

August 1, 2019.

The physics problem

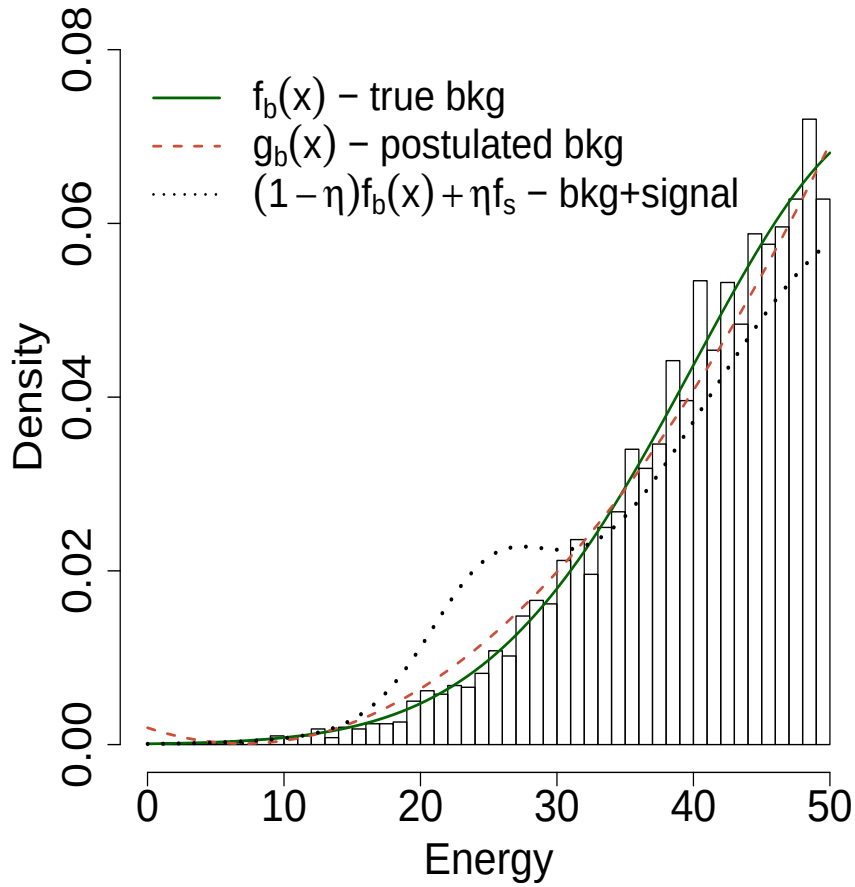
PHYSICIST(S): I would like to detect the signal of a new particle/astronomical source/astrophysical phenomenon BUT I am not sure if I am modeling the background correctly.



Problems associated with background mismodeling

- If the background is **overestimated** in the signal region
  $\Rightarrow$ it is harder to detect the signal if present.

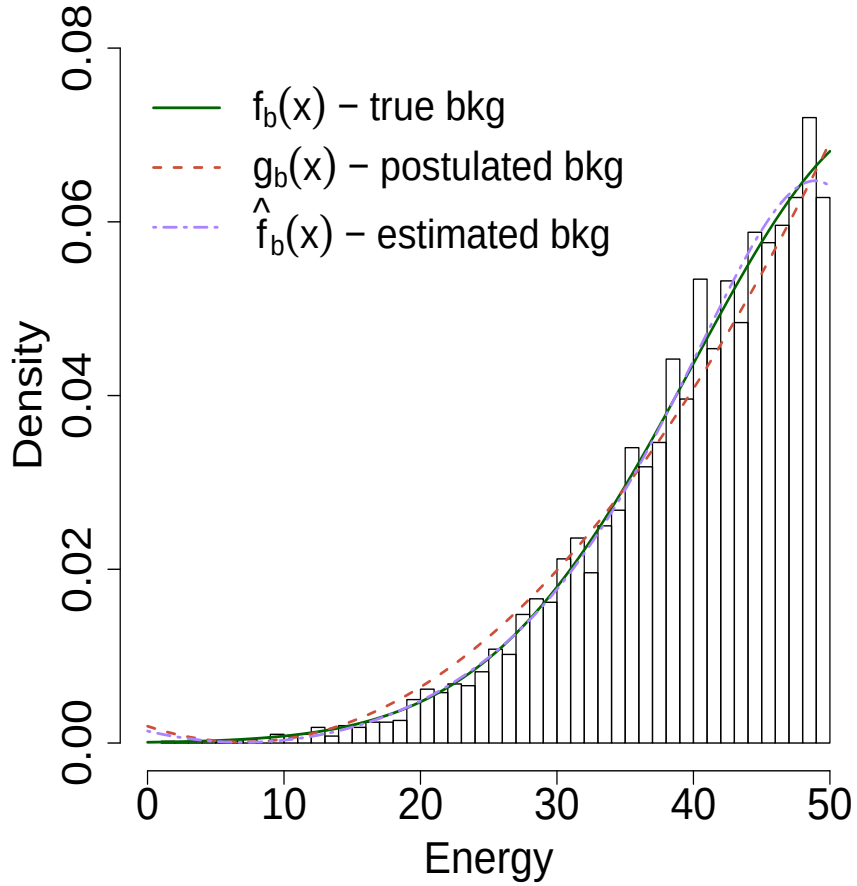
- If the background is **underestimated**
  $\Rightarrow$ the sensitivity is artificially enhanced (we may claim that a signal is there when it isn't).



The first statistical question (Q1)

STATISTICIAN: Can we provide a "data-updated" version of the pdf postulated by the physicists?

- INPUT 1: the postulated background distribution $g_b(x)$.
- INPUT 2: the data $x_1, \ldots, x_n$.
- OUTPUT: an estimate for the true background distribution $\hat{f}_b(x)$ constructed using both $g_b(x)$ and the data.



The second statistical question (Q2)

STATISTICIAN: Can we use the "data-updated" version of the pdf postulated by the physicists to perform inference?

When we say **inference** we mean:

- Inference to assess if the postulated background requires an "update".
- Inference to assess if the signal the physicist is looking for is there or not (and characterize/identify it).
- Inference to check if there is something else in addition to the signal the physicist is looking for (and characterize/identify it).

Summary and the last statistical questions (Q3-Q4)

- **Q1**: Can we provide a "data-updated" version of the pdf postulated by the physicists? (Modelling)
- **Q2**: Can we use the "data-updated" version of the pdf postulated by the physicists to perform inference? (Inference)
- **Q3**: Can we address all of the above using just one plot? (Graphics)

Q4: Can we design a unified algorithm for data(+science)-driven discoveries by integrating all of the above? (Data Science)

Answer(s):**YES**.

...and guess what, we do not even need Bayes!

The key element of the solution

The skew-G density model

(Mukhopadhyay and Parzen, 2014, Mukhopadhyay, 2017)

Given a random variable X, let F and f be its cdf and pdf (or pmf) respectively. Let G a suitable cdf and let g be the respective pdf (or pmf). Then,

$$f(x) = g(x)\, d(G(x); G, F) \tag{1}$$

where

$$d(G(x); G, F) = \frac{f(x)}{g(x)} \qquad \text{is called } \textbf{comparison density} \tag{2}$$

Note:

$d(G(x); G, F)$ it is not a density w.r.t. x but it is with respect to $G(x)$.

$\Rightarrow$ Set $u = G(x)$, $d(u; G, F)$ is a density w.r.t. u.

The LP representation

The approach of Mukhopadhyay and Parzen (2014) provides a representation of $d(u; G, F)$ by means of a orthonormal basis of shifted **L**egendre **P**olynomials, $Leg_j(u)$, i.e.,

$$d(u; G, F) = 1 + \sum_{j>0} LP_j \, Leg_j(u) \tag{3}$$

where the $LP_j = \int_0^1 Leg_j(u) d(u; G, F) du$.

The first few LP bases are:

$$Leg_0(u) = 1$$
$$Leg_1(u) = \sqrt{12}(u - 0.5)$$
$$Leg_2(u) = \sqrt{5}(6u^2 - 6u + 1) \quad \text{etc.}$$

The LP skew density estimator

An estimate of $d(u; G, F)$ can be obtained via the **LP skew density estimator** (Mukhopadhyay and Parzen, 2014), i.e.,

$$\widehat{d}(u; G, F) = 1 + \sum_{j=1}^{M} \widehat{LP}_j \, Leg_j(u) \tag{4}$$

where M is typically small, and given $u_1, \ldots, u_n$, i.i.d,

$$\widehat{LP}_j = \frac{1}{n} \sum_{i=1}^{n} Leg_j(u_i)$$

.

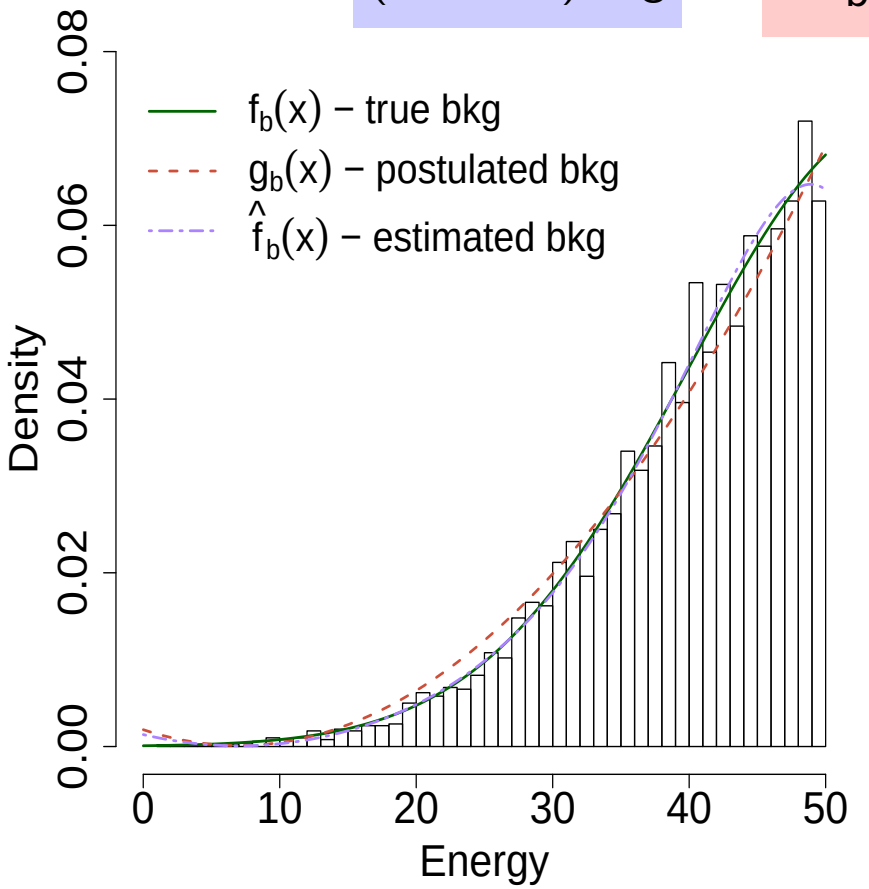
Hence, an estimate of $f(x)$ is given by

$$\hat{f}(x) = g(x) \, \widehat{d}(G(x); G, F) \tag{5}$$

In the background problem...

We can exploit the skew-G density model and write

$$\underbrace{f_b(x)}_{\substack{\text{true}\\ \text{(unknown) bkg}}} = \underbrace{g_b(x)}_{\substack{\text{postulated}\\ \text{bkg}}} \underbrace{d(G_b(x); G_b, F_b)}_{\substack{\text{comparison}\\ \text{density}}} \tag{6}$$



We then obtain the LP skew density estimator $\hat{f}_b(x)$ of $f_b(x)$ as discussed in the previous slide.

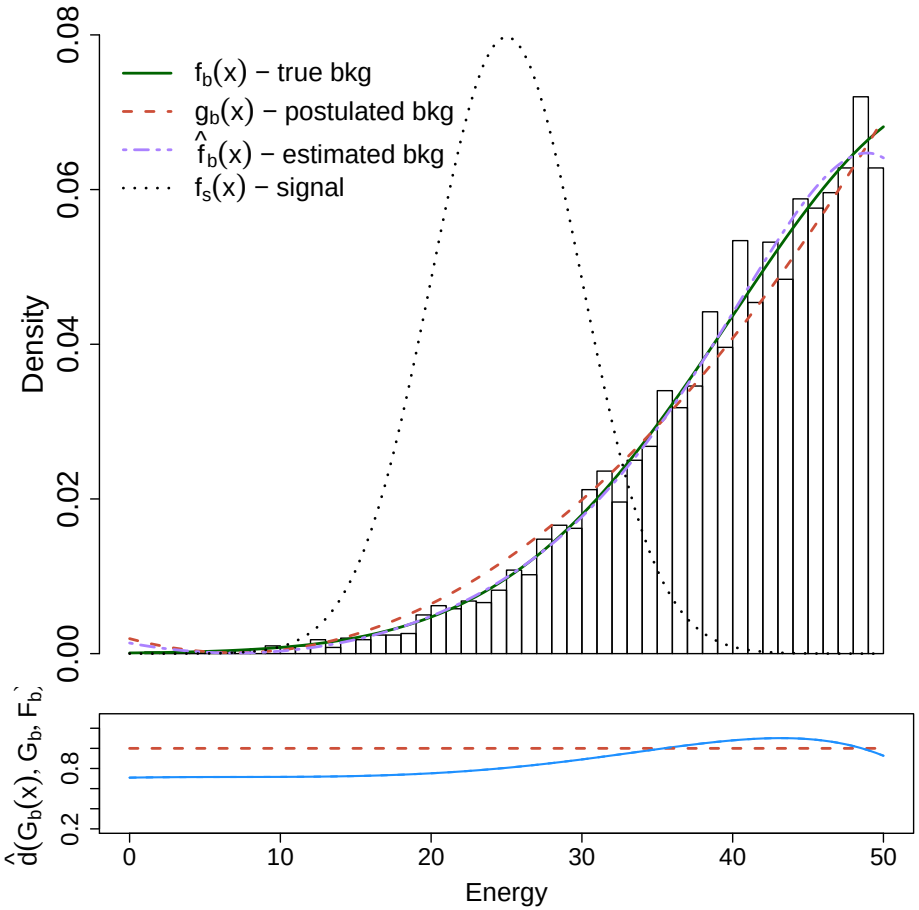
$\Rightarrow$**Q1** (Modelling) ✓

The beauty of a comparison density approach

The comparison density provides us with important insights on the physics problem...

...it tells us if the departure of $g_b(x)$ from $f_b(x)$ occurs in the signal region or not!

- If $\hat{d}(G_b(x); G_b, F_b)$ is above one outside the signal region $\Rightarrow$ using $\hat{f}_b(x)$ instead of $g_b(x)$ we prevent false discoveries.
- If $\hat{d}(G_b(x); G_b, F_b)$ is below one in the signal region $\Rightarrow$ using $\hat{f}_b(x)$ instead of $g_b(x)$ we increase our chances of discovering the new signal.



What about the inference?

A measure of departure from uniformity is given by the deviance, i.e.,

$$D = \sum_{j=1}^{M} \widehat{LP}_j^2 \qquad \text{(Mukhopadhyay 2016)}$$

Furthermore under $H_0 : F \equiv G$, and as $n \to \infty$

$$\sqrt{n}\widehat{LP}_j \xrightarrow{d} N(0,1) \quad \text{hence,} \quad \boxed{n \cdot D \sim \chi^2_M \quad \text{(Deviance test statistic)}}. \tag{7}$$

Confidence bands under H_0 can be constructed by means of tube formulae, i.e.,

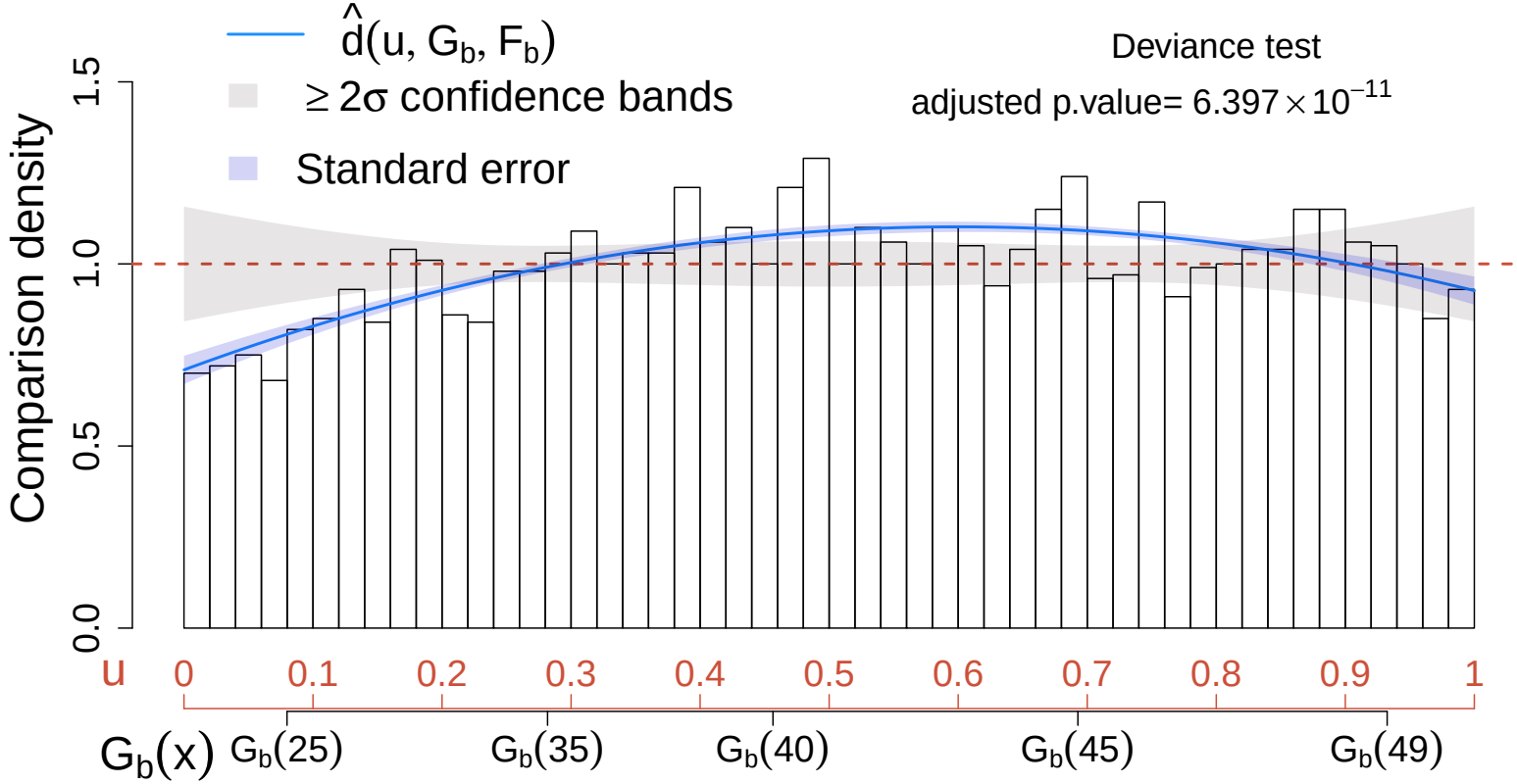
$$\text{CI}_{(1-\alpha)\%}: \left[1 - c_\alpha \sqrt{\textstyle\sum_{j=1}^{M} \frac{1}{n} Leg_j^2(u)}, 1 + c_\alpha \sqrt{\textstyle\sum_{j=1}^{M} \frac{1}{n} Leg_j^2(u)}\right], \tag{8}$$

where c_α is the solutions of

$$2(1 - \Phi(c_\alpha)) + \frac{k_0}{\pi} e^{-0.5c_\alpha^2} = \alpha \quad \text{with} \quad k_0 = \sqrt{\sum_{j=1}^{M} [\frac{\partial}{\partial u} Leg_j(u)]^2} \tag{9}$$

$\Rightarrow$ **Q3** (Inference) ✓.

What about the graph?



$\Rightarrow$ **Q4** (Graphics) ✓

Note: The term "adjusted" highlights the fact that, when M is selected among a pool of possible values, the inference must be adjusted accordingly (see Algeri S., 2019 for more details).

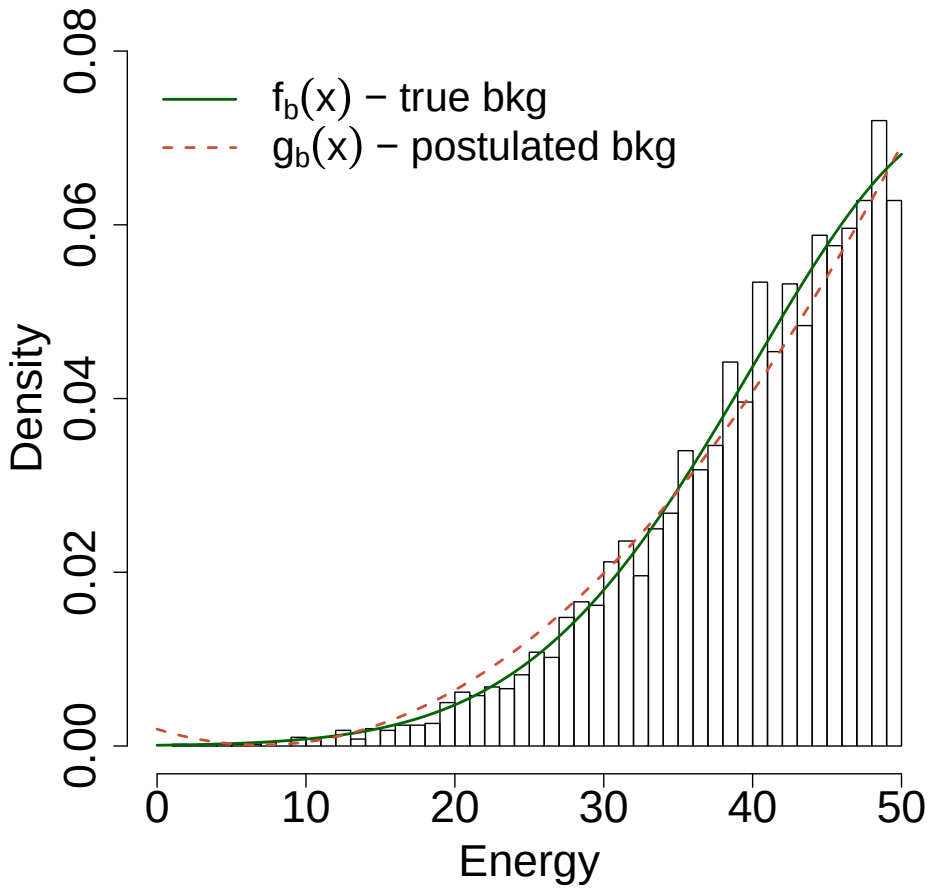
What about the algorithm?

Let's go through the main steps considering a running example...

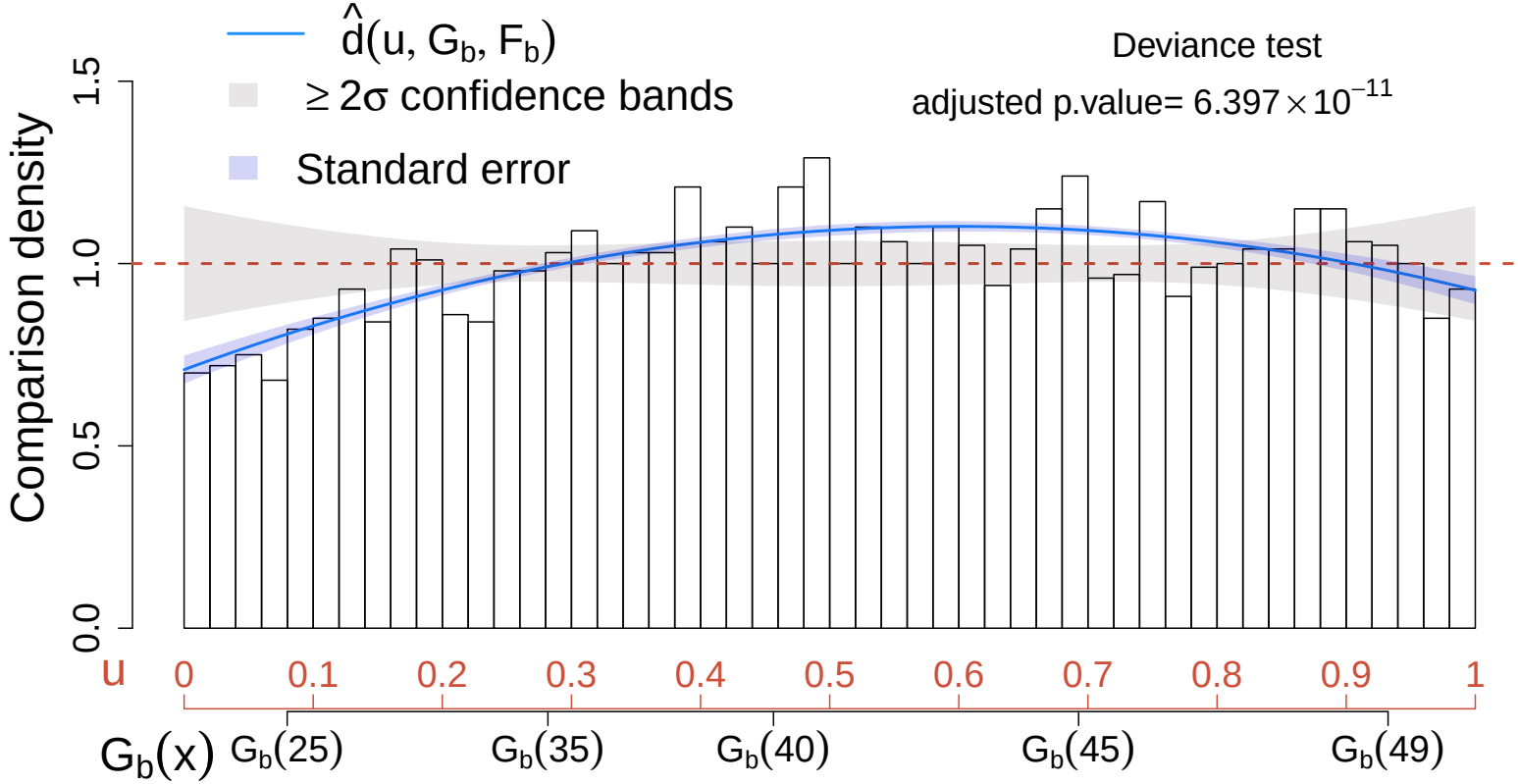
Phase A: Background calibration

First of all our initial **INPUTS** are:

- A source-free sample $\tilde{\boldsymbol{x}}_b = \tilde{x}_1, \ldots, \tilde{x}_{5000}$ from a $N(55, 15^2)$ truncated over $[0; 50]$.
- A postulated background distribution $g_b(\tilde{x})$ given by the best fit of a quadratic polynomial.

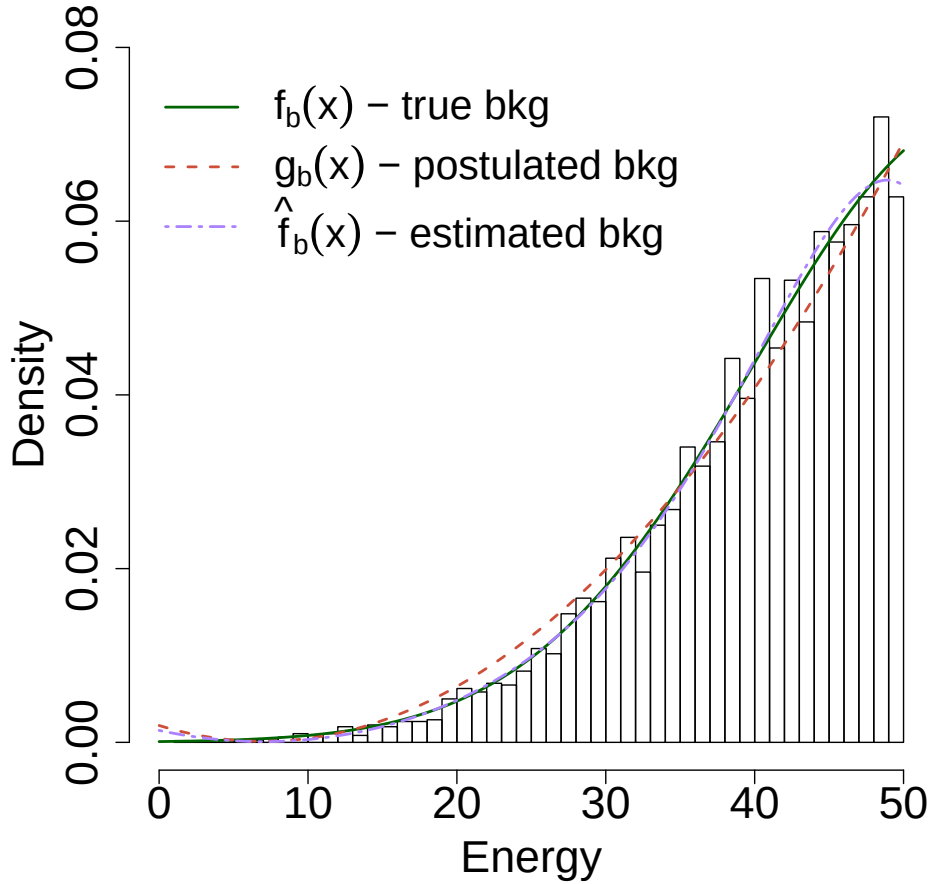


Step 1: estimate/test $d(\tilde{u}; G_b, F_b)$



Step 2: set $\hat{f}_b(\tilde{x})$

- If $F_b \not\equiv G_b$, set
  $\hat{f}_b(\tilde{x}) = g_b(\tilde{x})\hat{d}(\tilde{u}; G_b, F_b)$.
- Set $\hat{f}_b(\tilde{x}) = g_b(\tilde{x})$ otherwise.



Phase B: signal detection/characterization

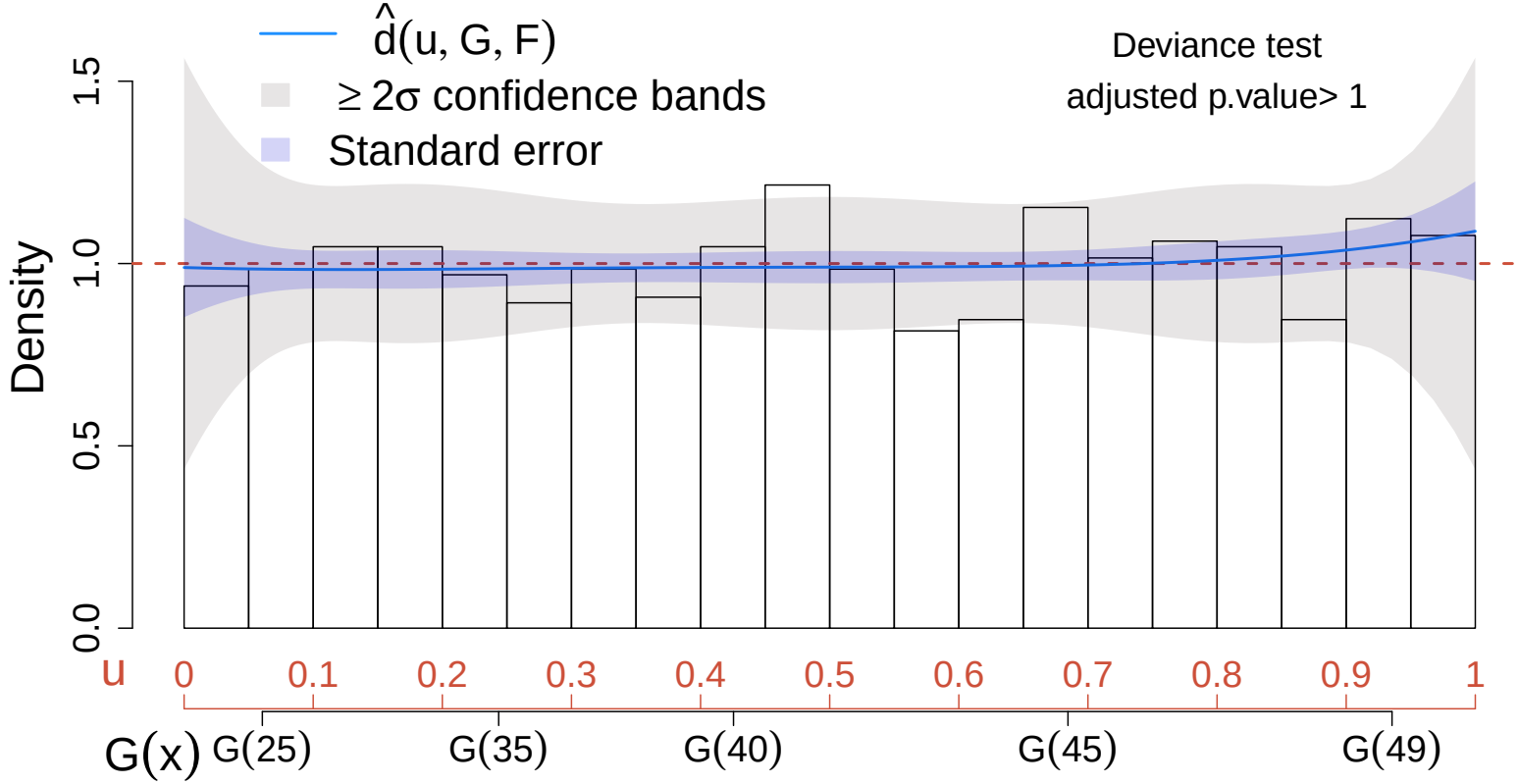
We can now consider our second **INPUT**

Physics sample $\boldsymbol{x} = x_1, \ldots, x_{1300}$.
We consider three possible scenarios:

- **Background only:** We have no signal in the data.
- **Background + Signal:** The signal we are looking for is present.
- **Background + Signal + Something else:** The signal we are looking for is present but we also have the signal of another source that we didn't expect.

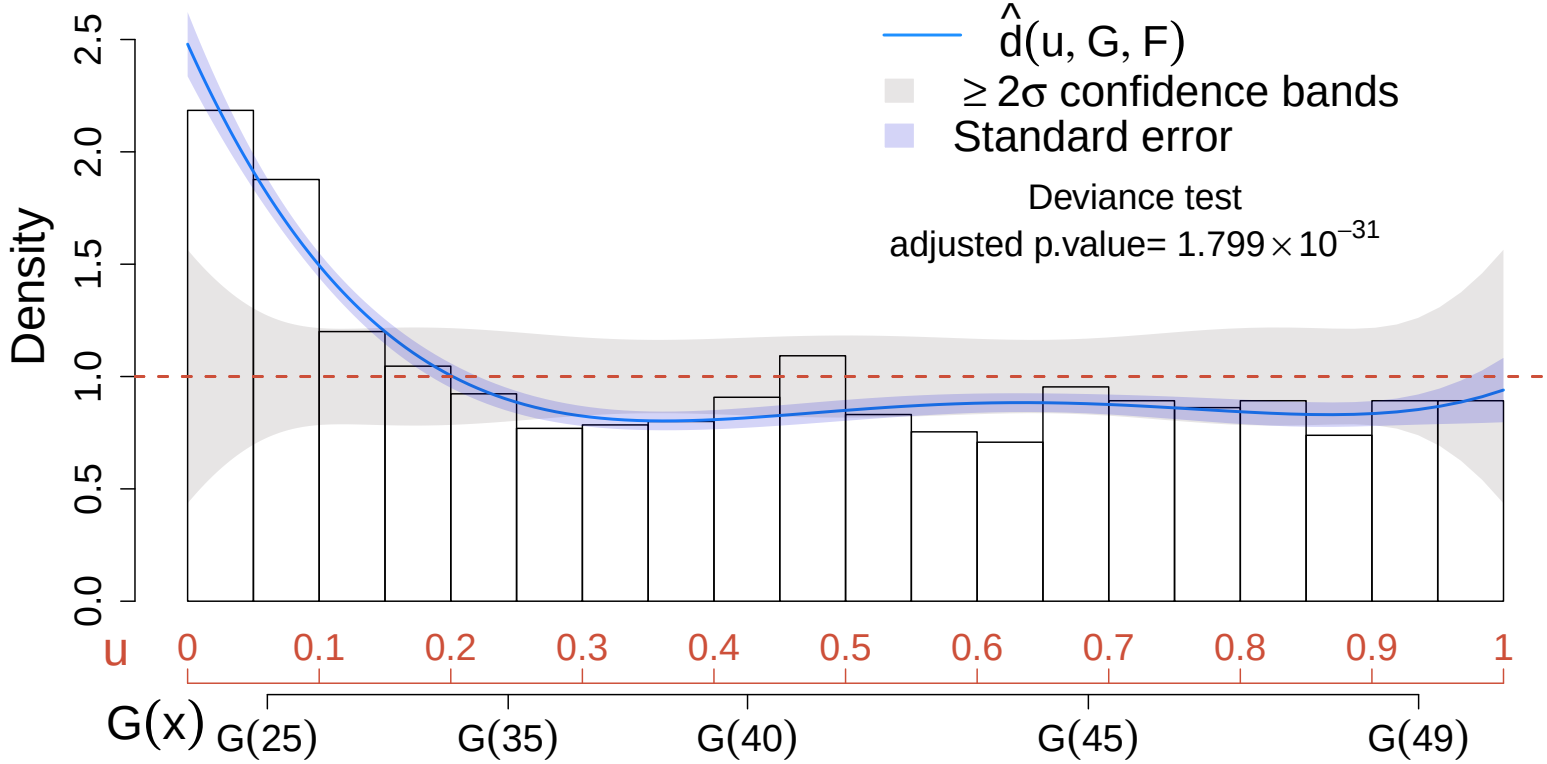
Step 3: set $g \equiv \hat{f}_b$ and estimate/test $\hat{d}(u; G, F)$ (Background only case)

$X \sim N_{[0,50]}(55, 15^2)$



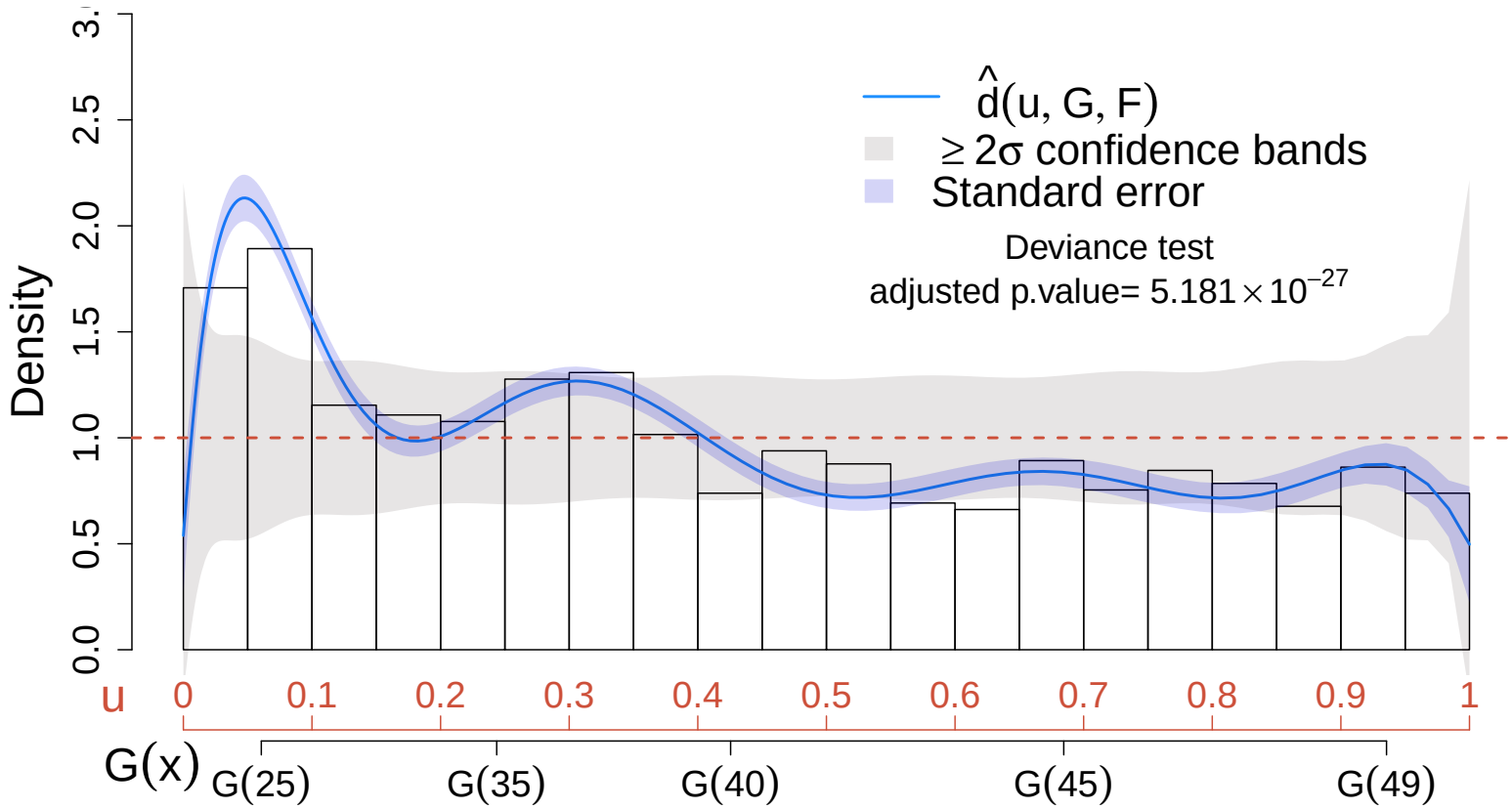
Step 3: set $g \equiv \hat{f}_b$ and estimate/test $\hat{d}(u; \hat{F}_b, F)$ (Background+Signal case)

$$X \sim (1 - 0.15) N_{[0,50]}(55, 15^2) + 0.15 N_{[0,50]}(25, 4.5^2)$$



Step 3: set $g \equiv \hat{f}_b$ and estimate/test $\hat{d}(u; \hat{F}_b, F)$ (Background+Signal+ Something else case)

$$X \sim (1 - 0.15 - 0.1)N_{[0,50]}(55, 15^2) + 0.15N_{[0,50]}(25, 4.5^2) + 0.1N_{[0,50]}(37, 1.8^2)$$



Step 4: Set $\hat{f}(x)$ or move to semiparametric stage

- If $\hat{F}_b \not\equiv F$, go to Step 5 (semiparametric signal characterization).
- Set $\hat{f}(x) = \hat{f}_b(x)$ otherwise and stop.

Step 5: ML estimation
(Semiparametric stage)

If $f_s(x)$ is unknown, one can explore the theories available and assess their validity fitting the data parametrically to obtain a "good" candidate model for our data e.g.,

$$g_{bs}(x) = (1-\eta)\hat{f}_b(x) + \eta f_s(x), \qquad 0 \leq \eta \leq 1 \tag{10}$$

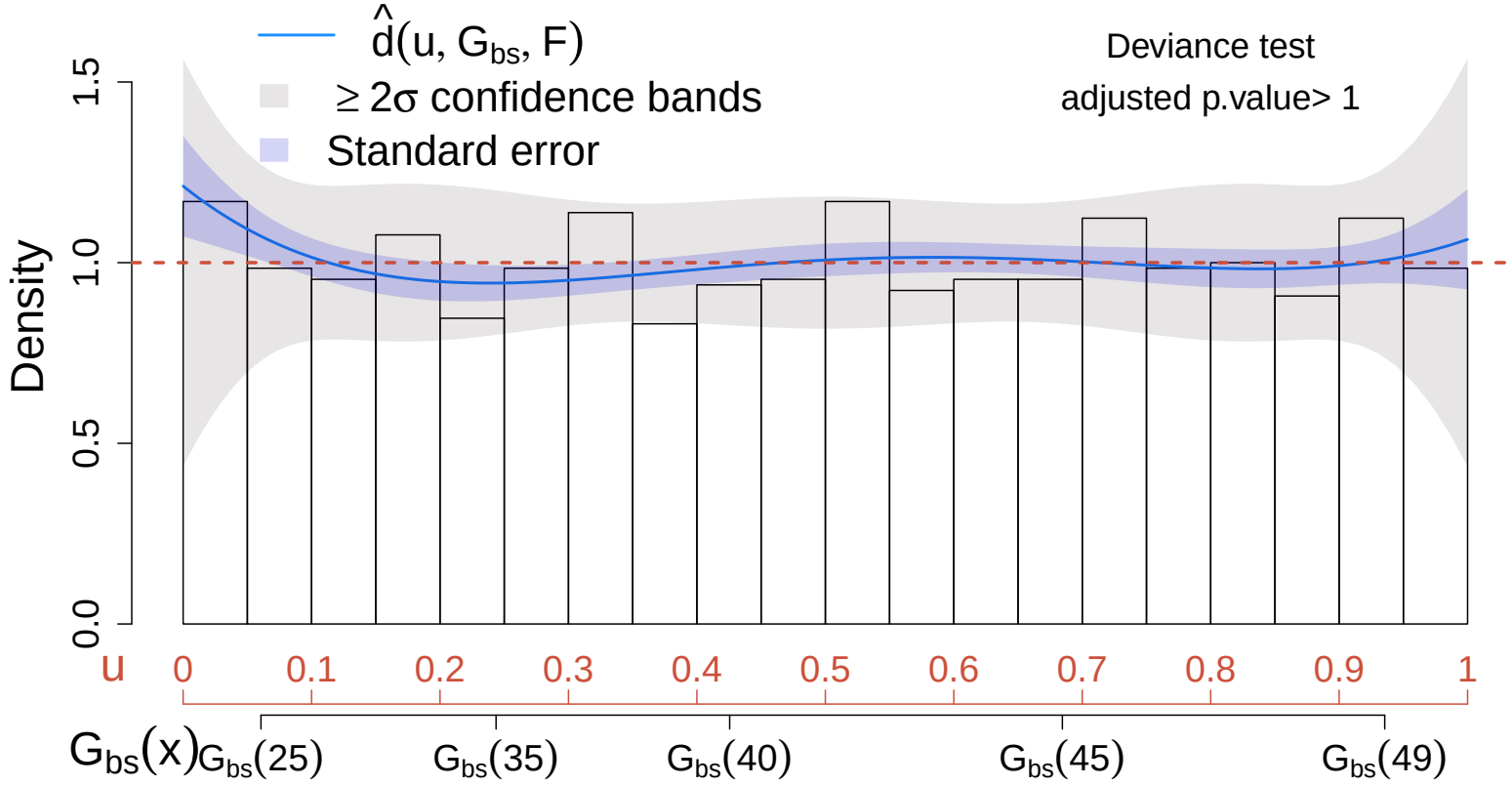
If $f_s(x)$ is known (up to some free parameters), we can test directly if (10) provides a good fit for the data.

E.g.,

- Suppose $f_s(x)$ is the pdf of a $N(25, 4.5^2)$ over $[0, 50]$.
- Estimate η by maximizing the likelihood of (10).
- Set $g_{bs}(x) = (1-\hat{\eta})\hat{f}_b(x) + \hat{\eta} f_s(x)$.

Step 6: Estimate/test $\hat{d}(u; G_{bs}, F)$ (Background+Signal case)

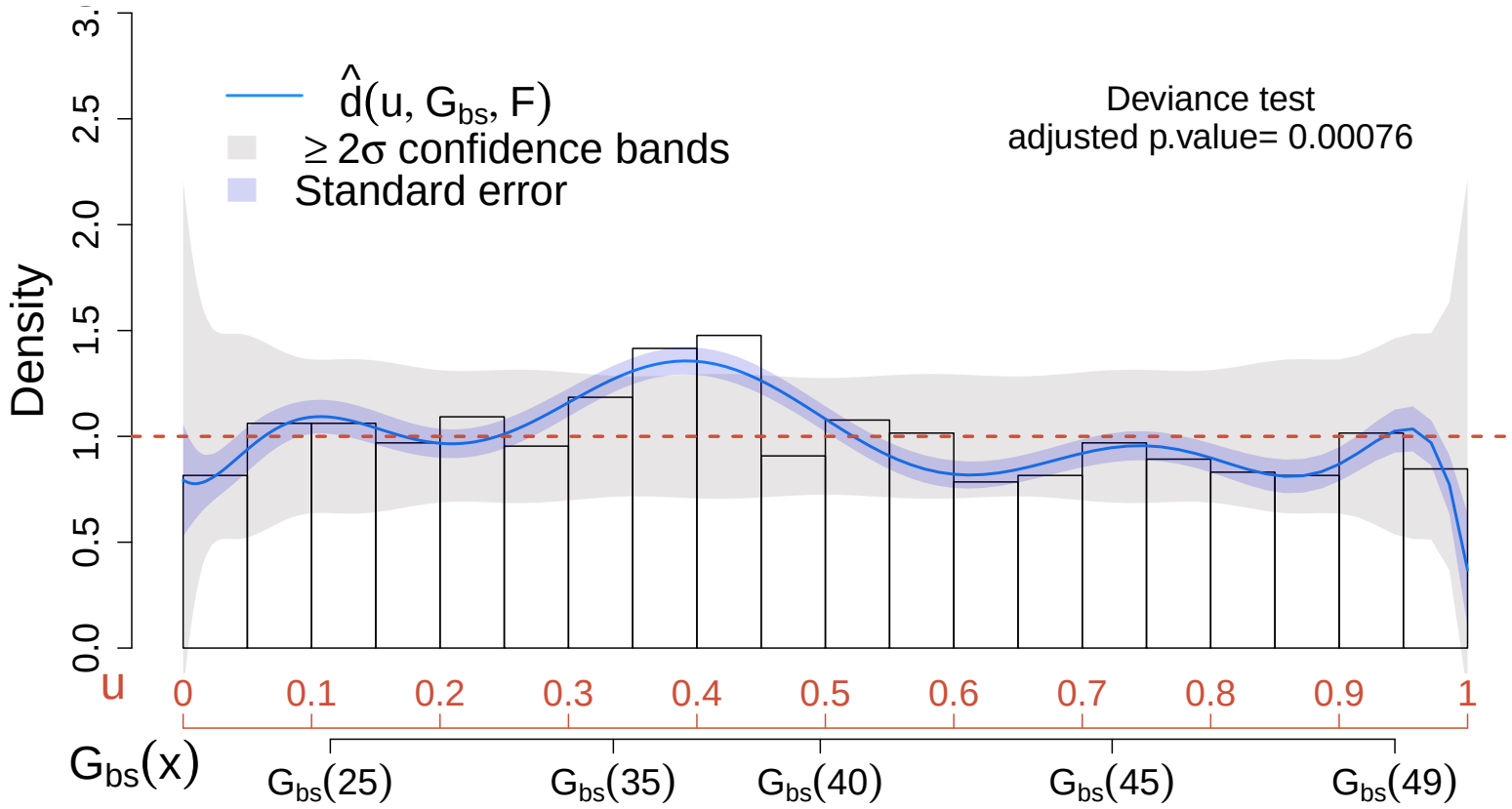
$$X \sim (1 - 0.15)N_{[0,50]}(55, 15^2) + 0.15N_{[0,50]}(25, 4.5^2)$$



Set $\hat{f}(x) = g_{bs}(x)$.

Step 6: Estimate/test $\hat{d}(u; G_{bs}, F)$ (Background+Signal+Something else case)

$$X \sim (1 - 0.15 - 0.1)N_{[0,50]}(55, 15^2) + 0.15N_{[0,50]}(25, 4.5^2) + 0.1N_{[0,50]}(37, 1.8^2)$$



Specify a different model for $f_s(x)$ which includes two bumps.

Summary

The LP approach to statistical modelling allows us to

(i) Assess the validity of existing models for the background.

(ii) "Update" the background distribution using the data.

(iii) Perform signal detection even if the signal distribution is not available.

(iv) Characterize the signal distribution.

(v) Detect additional signals of new unexpected sources.

References

- Algeri, S., 2019. Detecting new signals under background mismodelling. *Under review. arXiv:1906.06615.*
- Mukhopadhyay, S. and Parzen, E., 2014. LP approach to statistical modeling. *arXiv:1405.2601.*
- Mukhopadhyay, S., 2016. Large-scale signal detection: A unified perspective. *Biometrics*.
- Mukhopadhyay, S., 2017. Large-scale mode identification and data-driven sciences. *Electronic Journal of Statistics*.