Null Hypothesis Test for Anomaly Detection

Theoretical Physics Seminar @ Fermilab

Manuel Szewc, University of Cincinnati

University of CINCINNATI

This talk

Based on arxiv:2210.02226 by J.F. Kamenik and M.S.

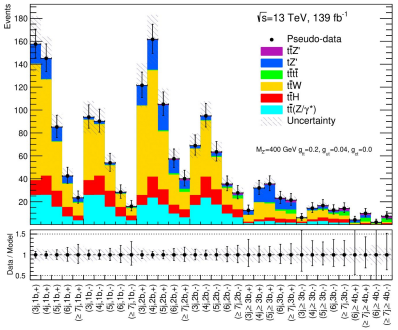
We present a method for translating **anomaly detection** into **p-values for background rejection** that does not rely on selection cuts nor background interpolation.

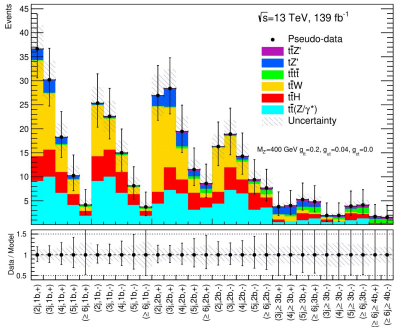
We show that we are robust to the absence of signal while still picking up small S/B

All code can be found at
ManuelSzewc/Null_Hypothesis_Test_for_Anomaly_Detection

"Traditional" Searches for New Physics in colliders

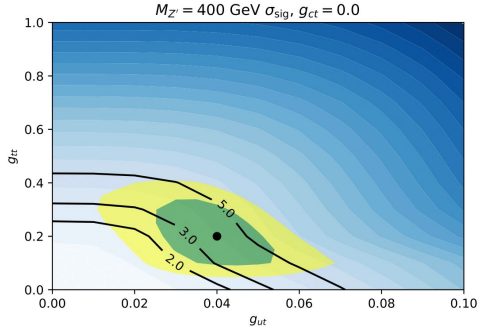
BSM model + specific observables → statistical expected discovery and/or exclusion significances

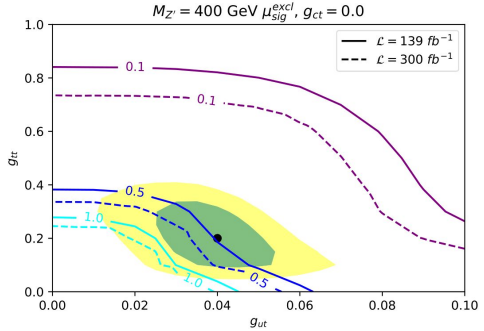




$$\frac{\mathcal{L}(\mathcal{D}|H_1)}{\mathcal{L}(\mathcal{D}|H_0)}$$

Example taken from a Z' pheno study by E. Alvarez et al, arxiv:2011.06514.





Anomaly detection

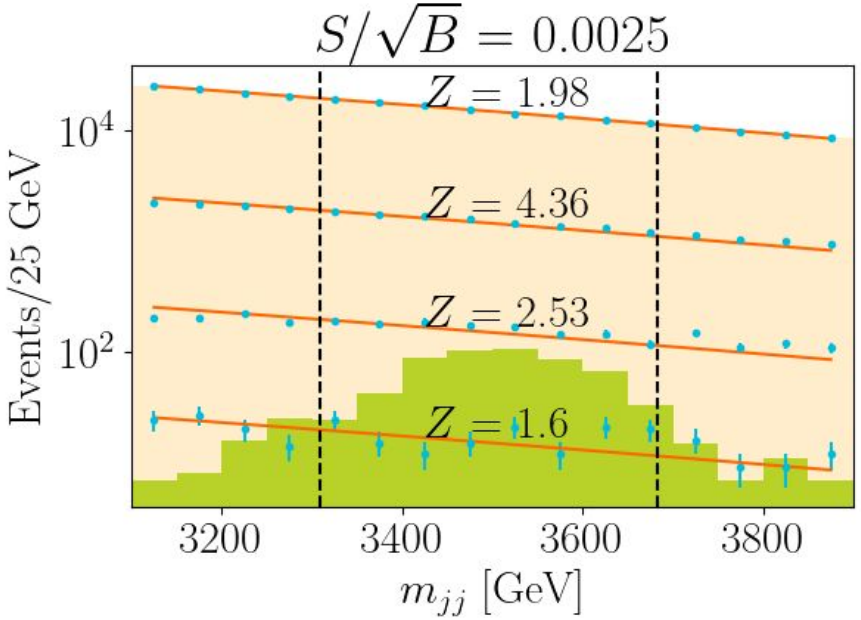
This result was obtained with the LHC Olympics (arXiv:2101.08320) datasets, more on this later!

Implementing **unsupervised and/or weakly supervised** algorithms to search for **small signals** among **large backgrounds**.

Less specific at a cost: sensitive to a large variety of signals but loss of statistical power in comparison to dedicated searches.

"Typical" AD:

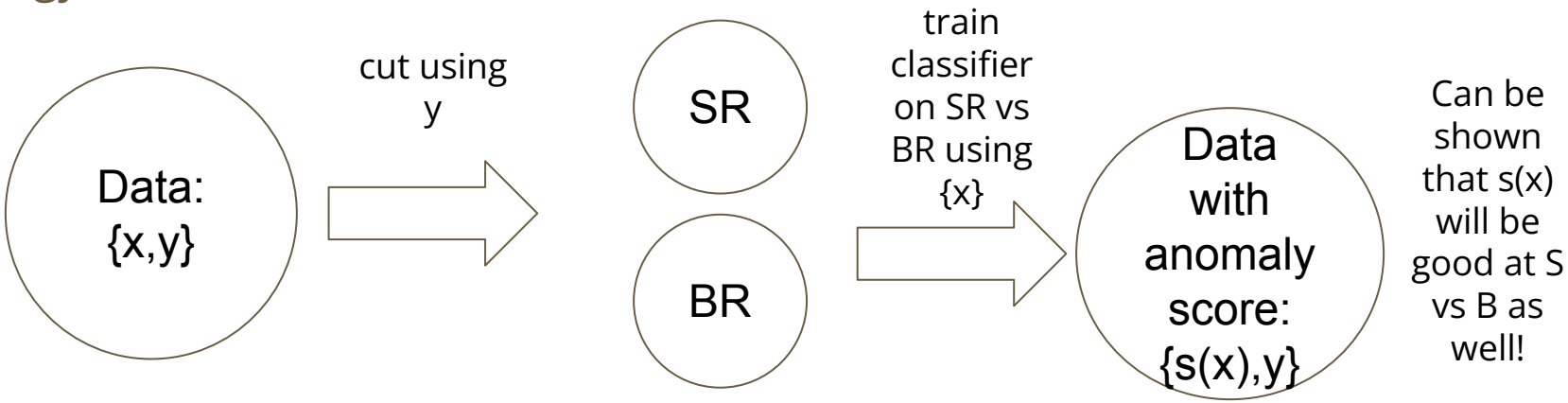
- SR and BR with **different amounts of S and B → Learn Anomaly Score from SR and BR distr.**
- **Cuts on Anomaly Score → Interesting events.**
- **Interesting events + extrapolation from BR → Significance**



An example: Classification Without Labels

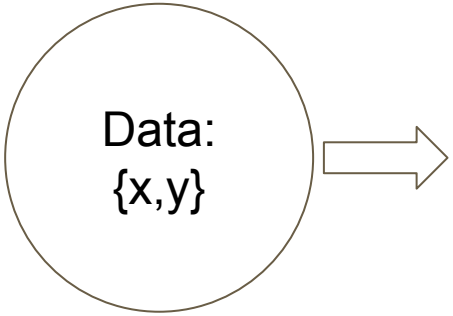
One of the first and most widely used Anomaly Detection techniques, introduced in arXiv:1708.02949 by E. M. Metodiev, B. Nachman, and J. Thaler.

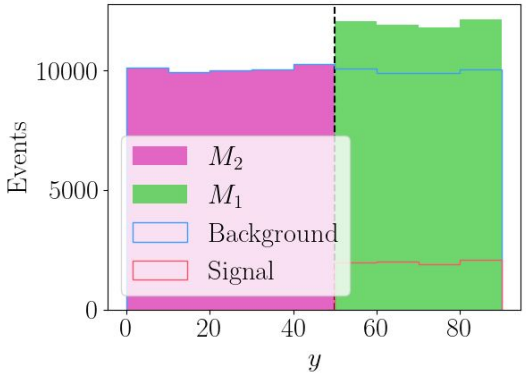
Already considered by experimental searches as a CWoLa + bump-hunt strategy.

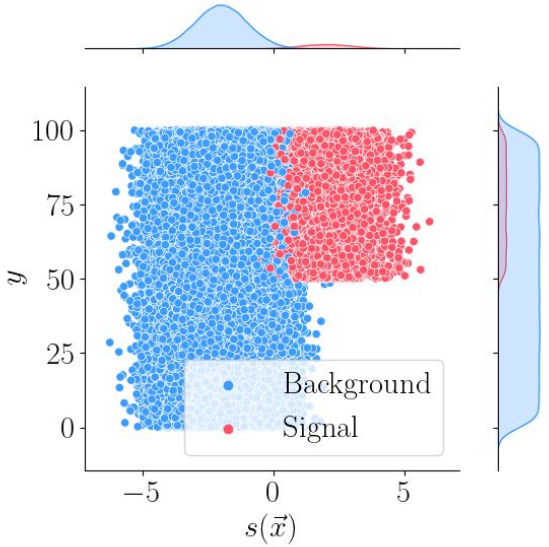


An example: Classification Without Labels

The anomaly score can then be used to select events and get a p-value from the $p(y|s>s_{cut})$!





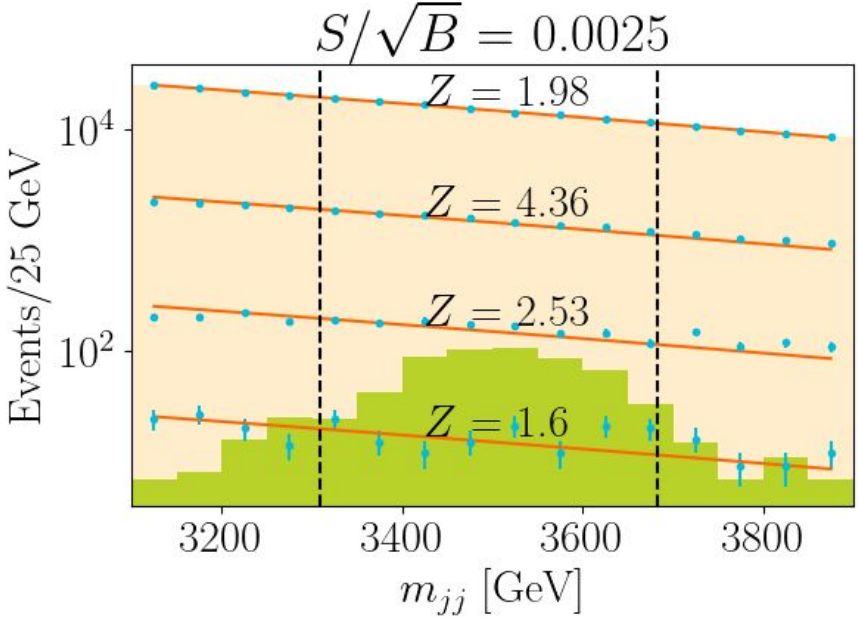


Toy dataset to show how CWoLa works

Back to Anomaly detection

Still a lot of open questions. Sensitivity and measure problem:

- Anomaly cuts are **not robust** to data representation choices.
- They **cannot be optimised** and may cost precious signal events.
- BR **extrapolations** may introduce **biases**.



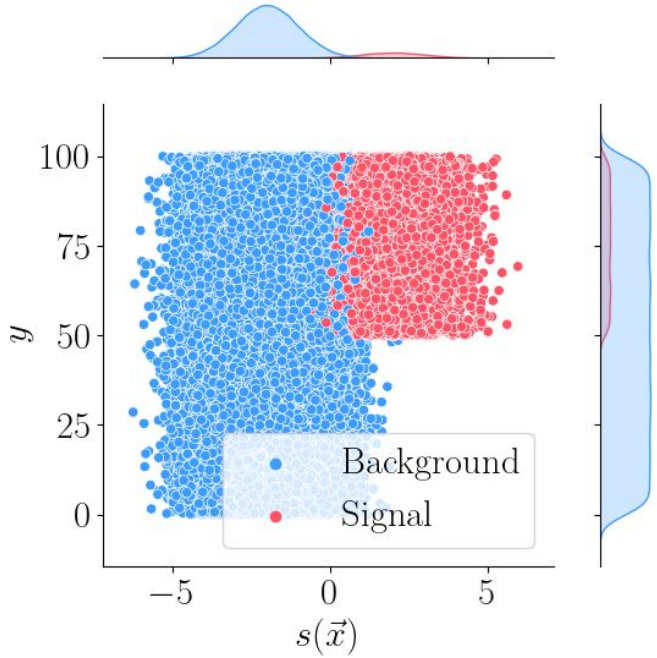
Anomaly detection

One possible path: since anomaly detection techniques have a null-hypothesis baked in:

- Can we design a statistical test with known (asymptotic) distributions to exclude it?
- Can we match the null-hypothesis to the background-only hypothesis?

If yes to both, we can design **hypothesis tests to exclude background-only without fixed anomaly score cuts nor background model extrapolations**

CWoLa: underlying model



CWoLa is at its heart a simple mixture model of **conditionally independent variables**.

$$p(s(\vec{x}), y|\pi) = (1-\pi)\, p(s(\vec{x})|B)p(y|B) + \pi\, p(s(\vec{x})|S)p(y|S)$$

If more than one process, p(s(x),y) **does not factorize**. The null hypothesis is

$$p(s(\vec{x}), y|\pi = 0) = p(s(\vec{x})|B)p(y|B)$$

Testing for independence

Testing for independence using a finite dataset of measured $\{s(x_i),y_i\}$

If $\{s(x_i),y_i\}$ are **assumed to be conditionally independent**, ruling out independence rules out a unique process. **Null hypothesis ↔ Background-only hypothesis**

If independence cannot be ruled out, **clear statement** about CWoLa incapable of stating whether differences between M_1 and M_2 are merely statistical fluctuations or signs of two underlying processes.

Mutual Information

MI encodes exactly what we want: the difference between the joint distribution and the marginals. It quantifies it in terms of the relative entropy

$$I(s,y) = D_{\mathrm{KL}}(p(s,y)||p(s)p(y))$$

$$I(s,y) = \int ds\, dy\; p(s,y) \log \frac{p(s,y)}{p(s)p(y)}$$

The null hypothesis

The conditional independence hypothesis becomes

$$I(s,y|z) = \int ds\, dy\; p(s,y|z) \log \frac{p(s,y|z)}{p(s|z)p(y|z)} = 0$$

and because

$$I(s,y) \geq 0$$

The null hypothesis can be phrased as

$$I(s,y|\pi = 0) = 0$$

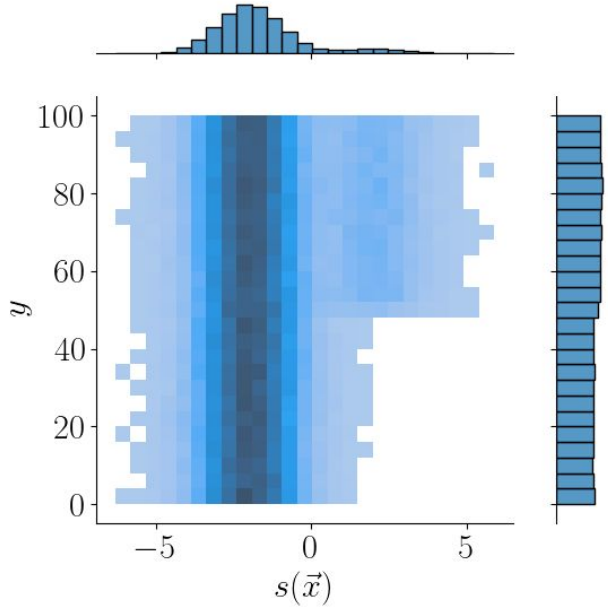
Mutual Information

One possible choice among many (e.g. Hoeffding's D independence test and distance correlation).

We focus on MI because it's cheap to estimate for large datasets by binning on {s,y} and the estimator has well behaved **asymptotic properties** in the limit of small MI and large sample size.

$$p(\hat{I}|I=0) = \Gamma\left(\frac{(d_s-1)(d_y-1)}{2}, N\right)$$

Estimating Mutual Information



$$\hat{p}(s(\vec{x}))$$
$$\hat{p}(y)$$
$$\hat{p}(s(\vec{x}), y)$$

$$\hat{I}(s, y)$$

SA-CWoLa

All this assumes **conditional independence** which may not be exactly true.

Correlation between features addressed by K. Benkendorfer, L. L. Pottier, and B. Nachman in arXiv:2009.02205: **the Simulation Assisted CWoLa or SA-CWoLa**

A simulation dataset is introduced and the loss function is modified to

$$\mathcal{L}_{\text{SA-CWoLa}}[s] = -\left(\sum_{\vec{x}_n \in M_1^{\text{data}}} \log s(\vec{x}_n) + \sum_{\vec{x}_n \in M_2^{\text{data}}} \log\left(1 - s(\vec{x}_n)\right)\right)$$
$$-\lambda\left(\sum_{\vec{x}_n \in M_1^{\text{sim.}}} \log\left(1 - s(\vec{x}_n)\right) + \sum_{\vec{x}_n \in M_2^{\text{sim.}}} \log s(\vec{x}_n)\right)$$

SA-CWoLa

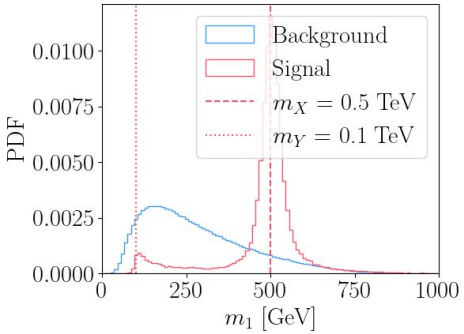
With imperfect simulations, you train your classifier to be agnostic to correlations between features in the background process.

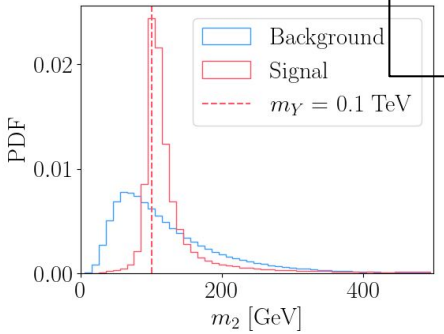
This ensures that the **null-hypothesis and the background-only hypothesis coincide** at the expense of a new **hyperparameter λ**.

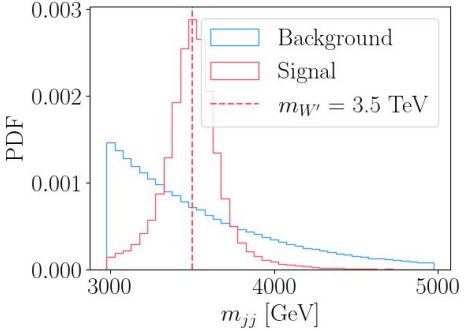
We verify that it works by computing the AUC on the simulated samples after training. We need AUC = 0.5.

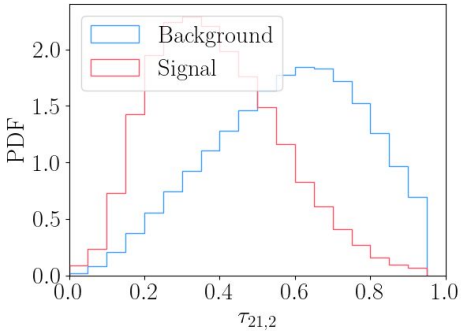
A test case: LHC Olympics

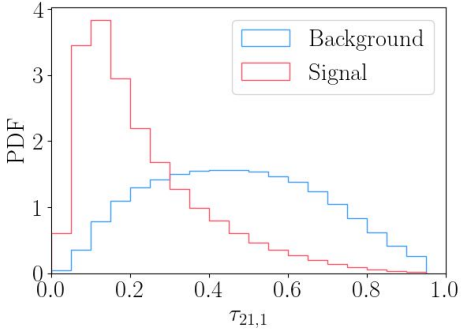
We consider the LHC Olympics (arXiv:2101.08320) R&D dataset with the LHC Olympics BB1 Background as simulation.







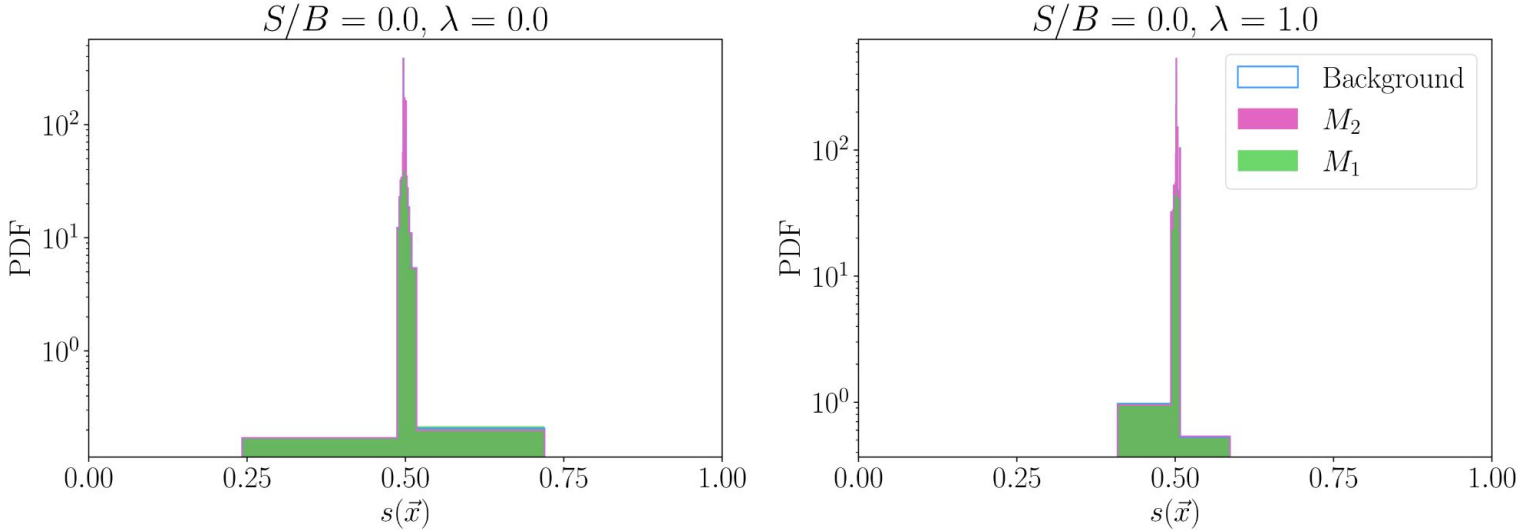




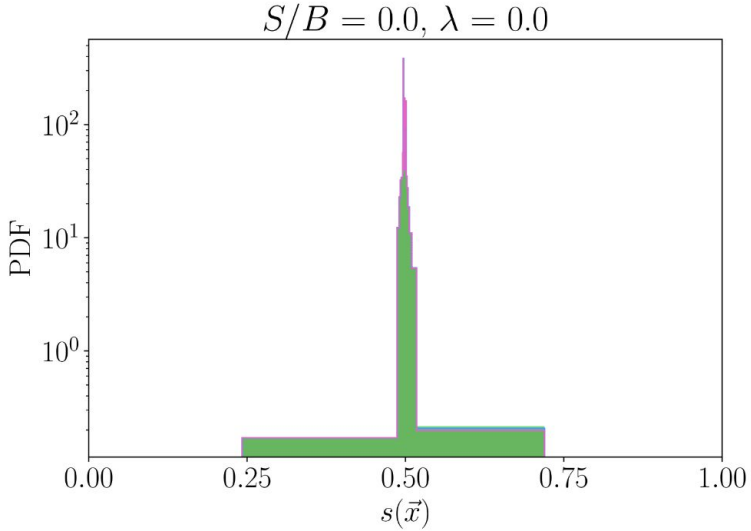
M_1 and M_2 are defined with the invariant mass of the event, with M_1={3.3,3.7} TeV and M_2={3.1,3.3}+{3.7,3.9} TeV
B = 250k, variable S/B

No signal? No problem

To interpret the effect of λ, we look at the learned variable's PDF
Beware, non-uniform binning to ensure ≤ 1% stat. uncertainty per bin



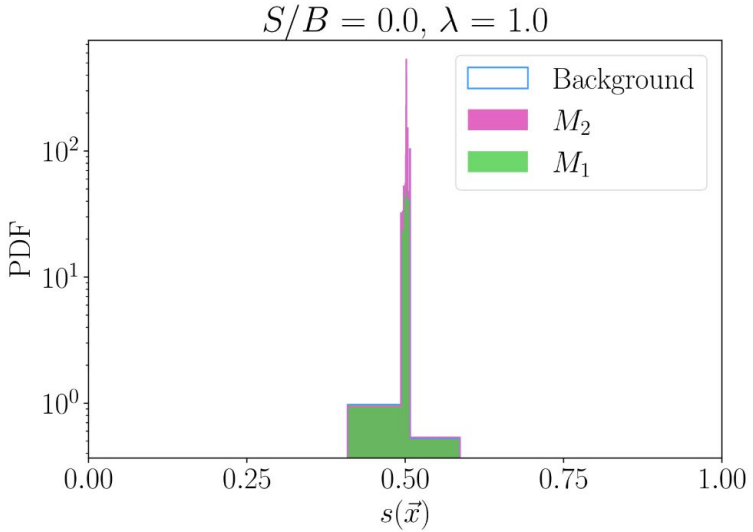
No signal? No problem



With no regularization, slight correlations sculpt the distribution (M_1 and M_2 are distinguishable)

No signal? No problem

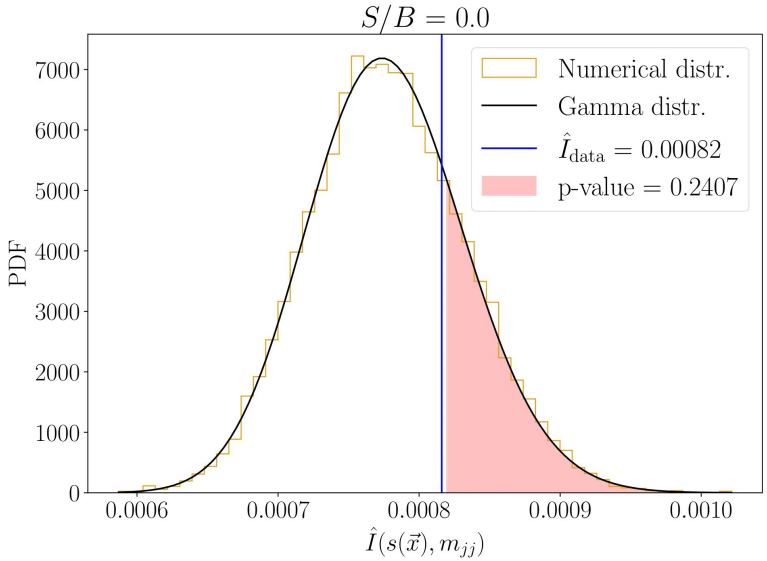
Regularization clusters the PDF around 0.5 and renders M_1 and M_2 indistinguishable for S/B = 0



No signal? No problem

Let’s look at the estimated mutual information

We can check how likely it is under the null hypothesis both numerically (with bootstrapping) and with asymptotic results

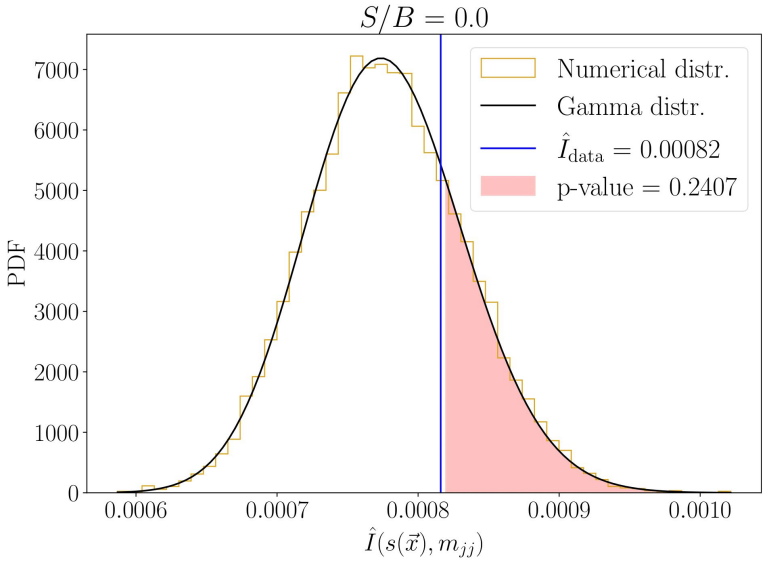


No signal? No problem

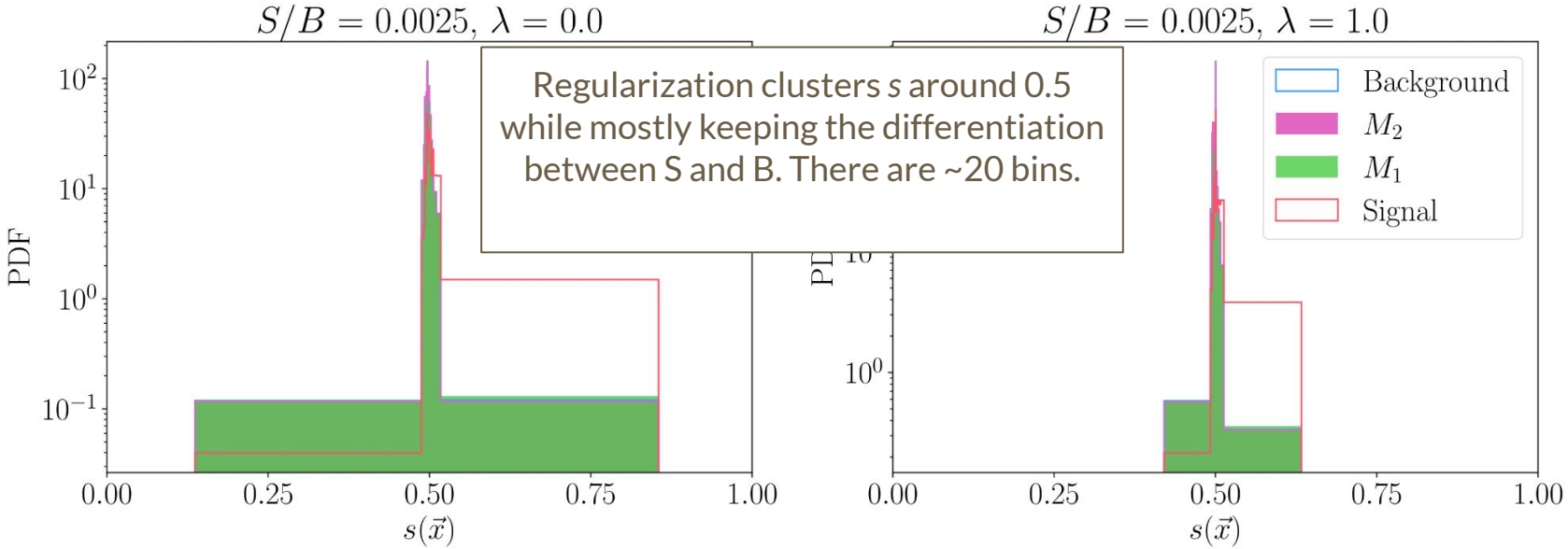
We only show the regularized case, because without regularization CWoLa picks-up the very slight correlations

The relevant result is that **we are not able to exclude the null-hypothesis.**

This implies we are robust against the absence of signal.



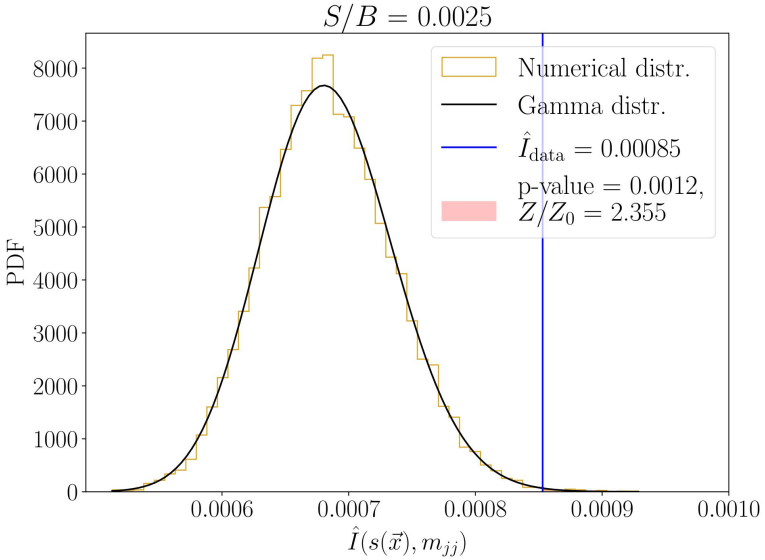
In the presence of small S/B, we still find effects



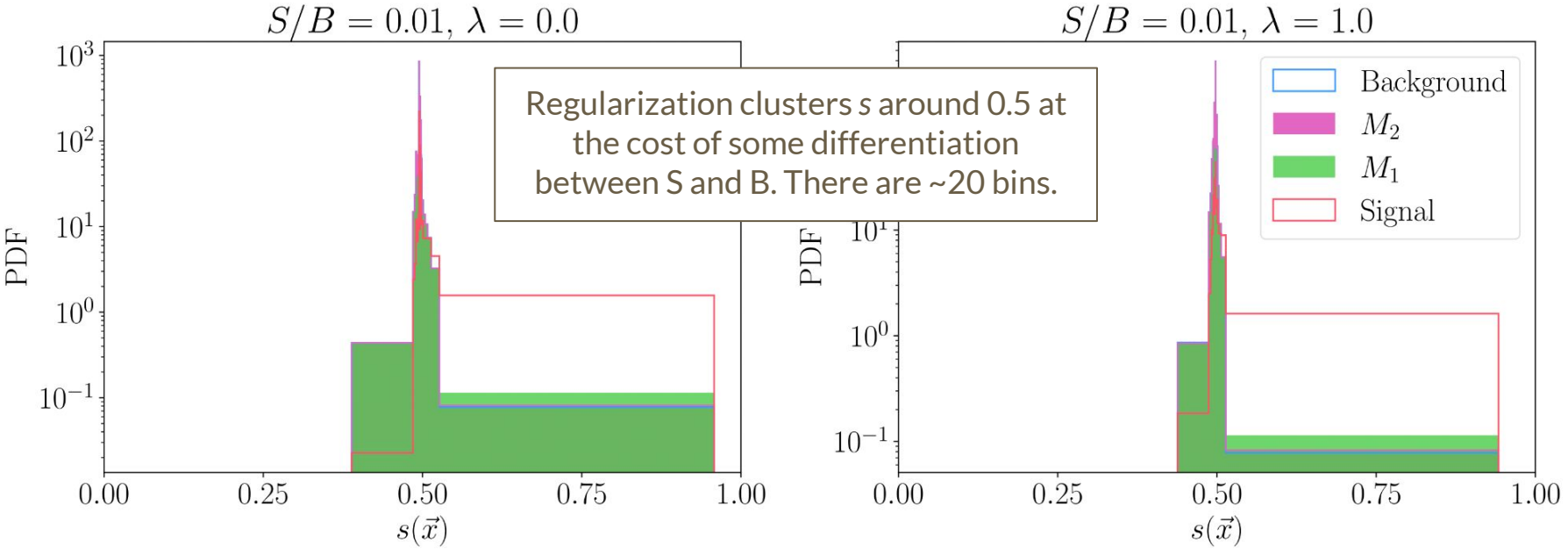
In the presence of small S/B, we still find effects

In the presence of small S/B, our method picks-up the absence of statistical independence.

There is a gain of significance with respect to the naive count S/√B. We are picking differential information.



For larger S/B, decorrelation reduces performance

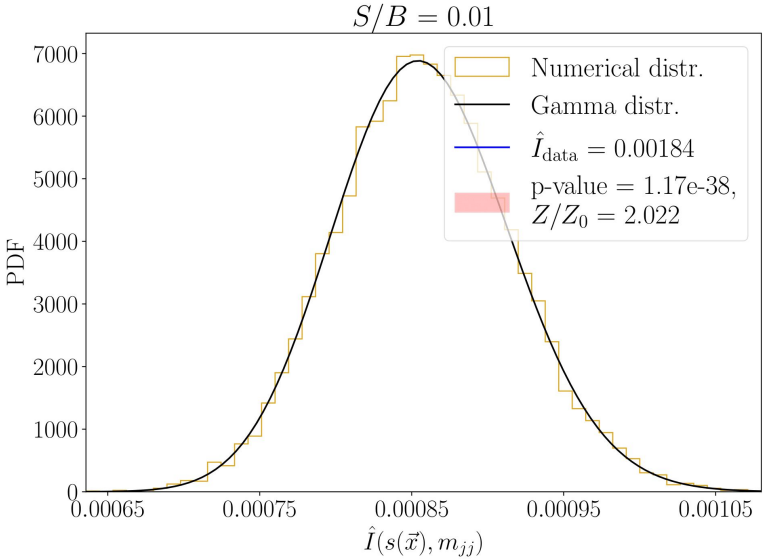


Regularization clusters *s* around 0.5 at the cost of some differentiation between S and B. There are ~20 bins.

For larger S/B, decorrelation reduces performance

For larger S/B, the method works but the decorrelation may hinder the performance with respect to the optimal classifier.

Not a problem! We prefer **robustness over sheer sensitivity** for Anomaly Detection as we look for small signals.



Comparison with other methods

We compare our method with the use of different anomaly cuts and with anomaly cuts + background modelling.

Mutual Info combines **robustness against no signal** and **good performance for small S/B** with no efficiency selection (no tuning!)

Significance	$S/B = 0.0$	$S/B = 0.0025$	$S/B = 0.005$	$S/B = 0.01$
$S/\sqrt{B}$	0.0	1.29	2.55	6.40
Mutual Info $\lambda = 0.0$	6.40	7.04	7.58	14.1
Mutual Info $\lambda = 1.0$	0.70	3.03	5.33	13.0
Anomaly cuts $\epsilon_2 = 0.1$, $\lambda = 0.0$	3.35	4.78	6.27	11.6
Anomaly cuts $\epsilon_2 = 0.1$, $\lambda = 1.0$	2.48	2.26	4.49	10.0
Anomaly cuts $\epsilon_2 = 0.01$, $\lambda = 0.0$	2.26	4.62	10.1	27.0
Anomaly cuts $\epsilon_2 = 0.01$, $\lambda = 1.0$	0.55	1.66	10.7	27.1
Anomaly cuts $\epsilon_2 = 0.001$, $\lambda = 0.0$	1.39	10.3	17.9	34.2
Anomaly cuts $\epsilon_2 = 0.001$, $\lambda = 1.0$	0.	0.57	13.6	37.0
Bump Hunt	0.95	1.97	2.74	5.30
Bump Hunt $\epsilon_2 = 0.1$, $\lambda = 0.0$	6.41	9.26	10.92	19.5
Bump Hunt $\epsilon_2 = 0.1$, $\lambda = 1.0$	3.81	4.35	6.93	16.0
Bump Hunt $\epsilon_2 = 0.01$, $\lambda = 0.0$	4.77	6.96	14.2	34.7
Bump Hunt $\epsilon_2 = 0.01$, $\lambda = 1.0$	0.97	2.53	14.0	35.0
Bump Hunt $\epsilon_2 = 0.001$, $\lambda = 0.0$	2.98	12.3	20.0	35.8
Bump Hunt $\epsilon_2 = 0.001$, $\lambda = 1.0$	0.29	1.60	15.9	38.7

Conclusions

We have presented a **novel strategy to quantify the sensitivity** of a specific anomaly detection technique, CWoLa.

We have shown that **by testing for statistical independence** between the learned {*s*,*y*} pairs using the estimated Mutual Information we obtain a test statistic that is **robust against the absence of signal and the correlation between {*x*,*y*}**.

This test **dispenses of the need for selection cuts and background modelling**, and can be generalized to other measures of statistical dependence and other decorrelation techniques, which need to be thought of as dataset-dependent.

This method also opens the door for **applying CWoLa to non-resonant searches** as in E. Alvarez, F. Lamagna and MS arXiv:1911.09699 and T. Finke, M. Kramer, M. Lipp and A. Muck, arXiv:2204.11889.

Backup

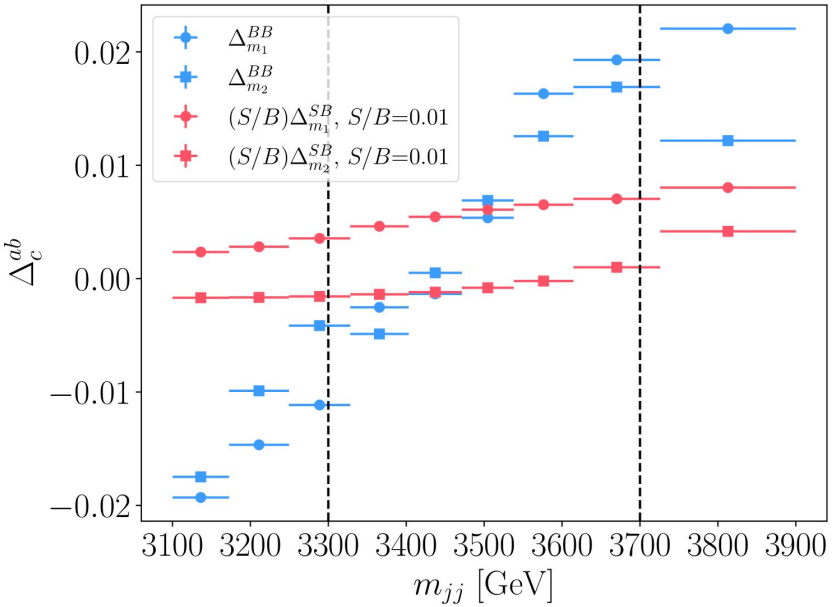
Is SA-CWoLa necessary?

A proxy for measuring the relative importance of x,y correlations:

$$\Delta_c^{ab}(m_{jj}^{\text{bin}}) = \frac{\mathbb{E}[m_c | m_{jj} \in m_{jj}^{\text{bin}}, a] - \mathbb{E}[m_c|b]}{\mathbb{E}[m_c|b]}$$

We compare the m_{jj} dependence for Background-Background and for Signal-Background.

$$\Delta_c^{S+B,B} \approx \Delta_c^{BB} + (S/B)\Delta_c^{SB}$$



The correlation between m_c and m_{jj} in the background is **more noticeable** than the difference between background and signal.