Learning Optimal Test Statistics

in Presence of Nuisance Parameters

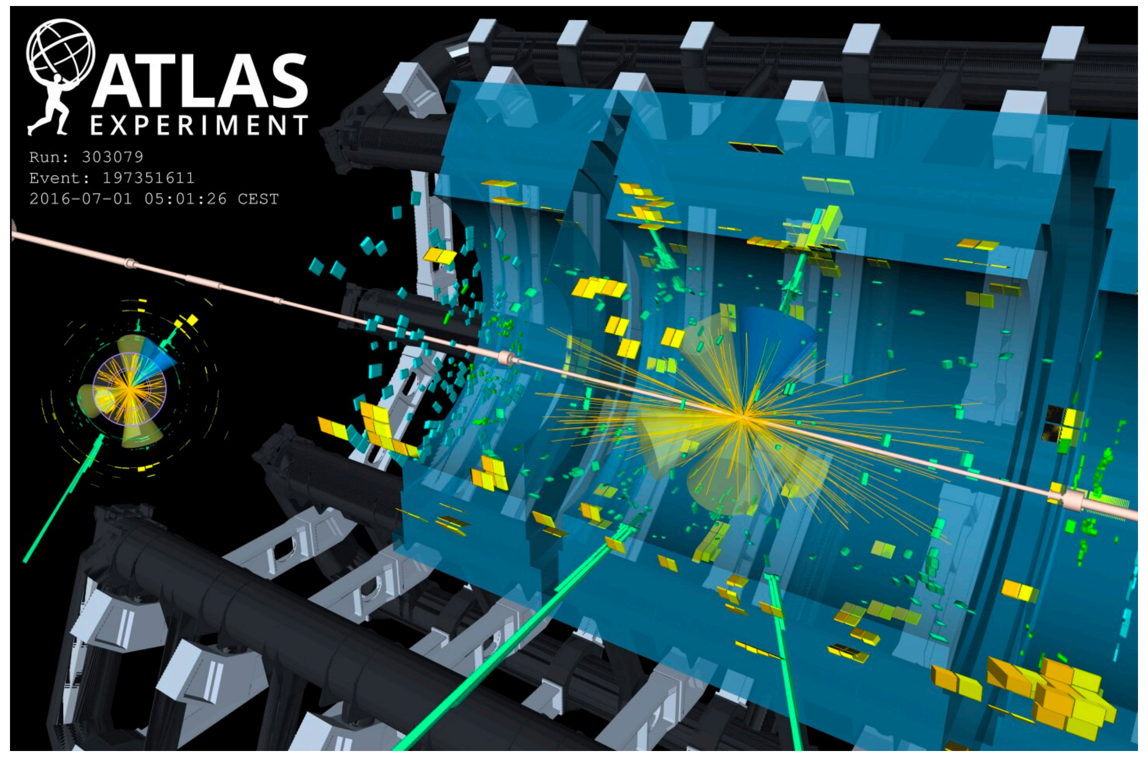
Lukas Heinrich, TUM

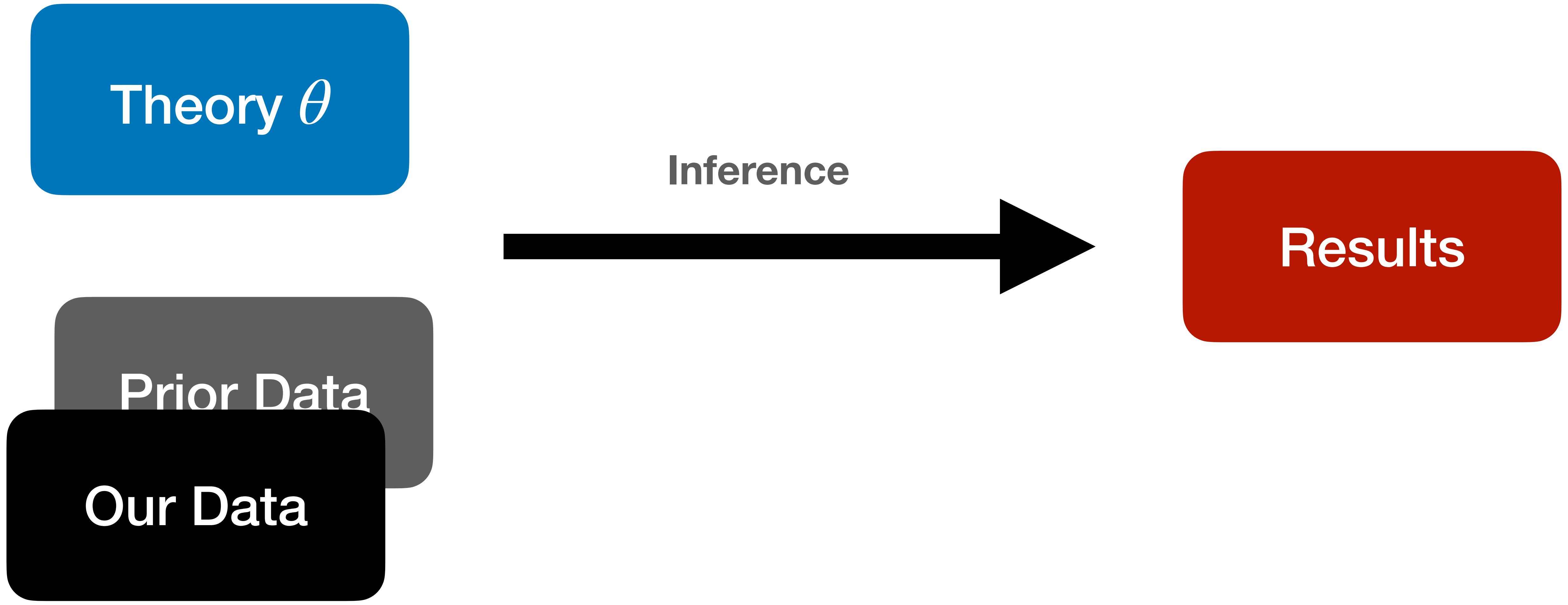
arXiv:2203.13079

Introduction

Our main goal as physicists is to make inferences about nature in light of the data we collect

$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + i\bar{\psi}\not{D}\psi + h.c. + \bar{\psi}_i y_{ij} \psi_j \phi + h.c. + |D_\mu \phi|^2 - V(\phi)$$





Likelihoods

The way we do this usually is through statistical inference by formulating a statistical model $p(x|\theta)$ that describes our data-generating process

When we say $p(x|\theta)$ (or "likelihood") we actually mean two things:

- **ability to generate data for a given theory:** $x \sim p(x|\theta)$
- **ability to evaluate the probability under a given theory:** $L(\theta) = p(x|\theta)$

Bayesian and Frequentist Inference

With a likelihood in hand, we can follow inference procedures

Bayesians: $p(x\,|\,\theta)$ **into** $p(\theta\,|\,x)$

$$p(\theta\,|\,x) = \frac{p(x\,|\,\theta)p(\theta)}{p(x)}$$

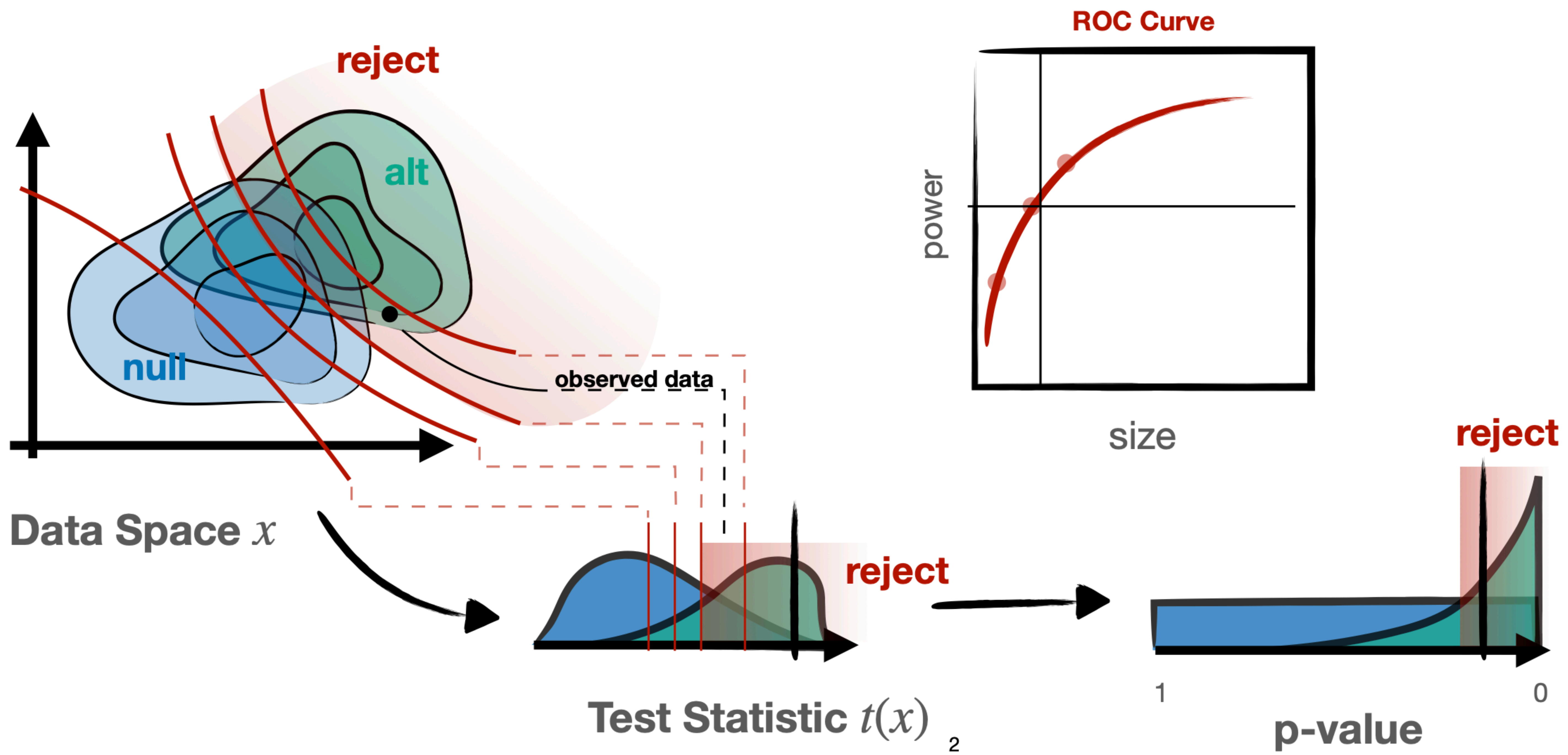
Note 1: requires ability to compute $p(x\,|\,\theta)$

Note 2: subjective choice on your priors $p(\theta)$

Bayesian and Frequentist Inference

Frequentists: let's look at the data distribution!

- **ideally in a way that accentuates the difference between theories**



Note 1: in principle only requires ability to sample $x \sim p(x \,|\, \theta)$ and compute $t(x)$

Note 2: subjective choice of which test statistic $t(x)$ to use

Optimal Test Statistics

In reality, we often to want evaluate the likelihood $p(x\,|\,\theta)$ also for frequentist inference as we can show that the **optimal test statistics** are related to it

Neyman-Pearson Lemma (no nuisance parameters)

$$t(x) = -2\log\frac{p(x|H_0)}{p(x|H_1)}$$

most powerful test at fixed size

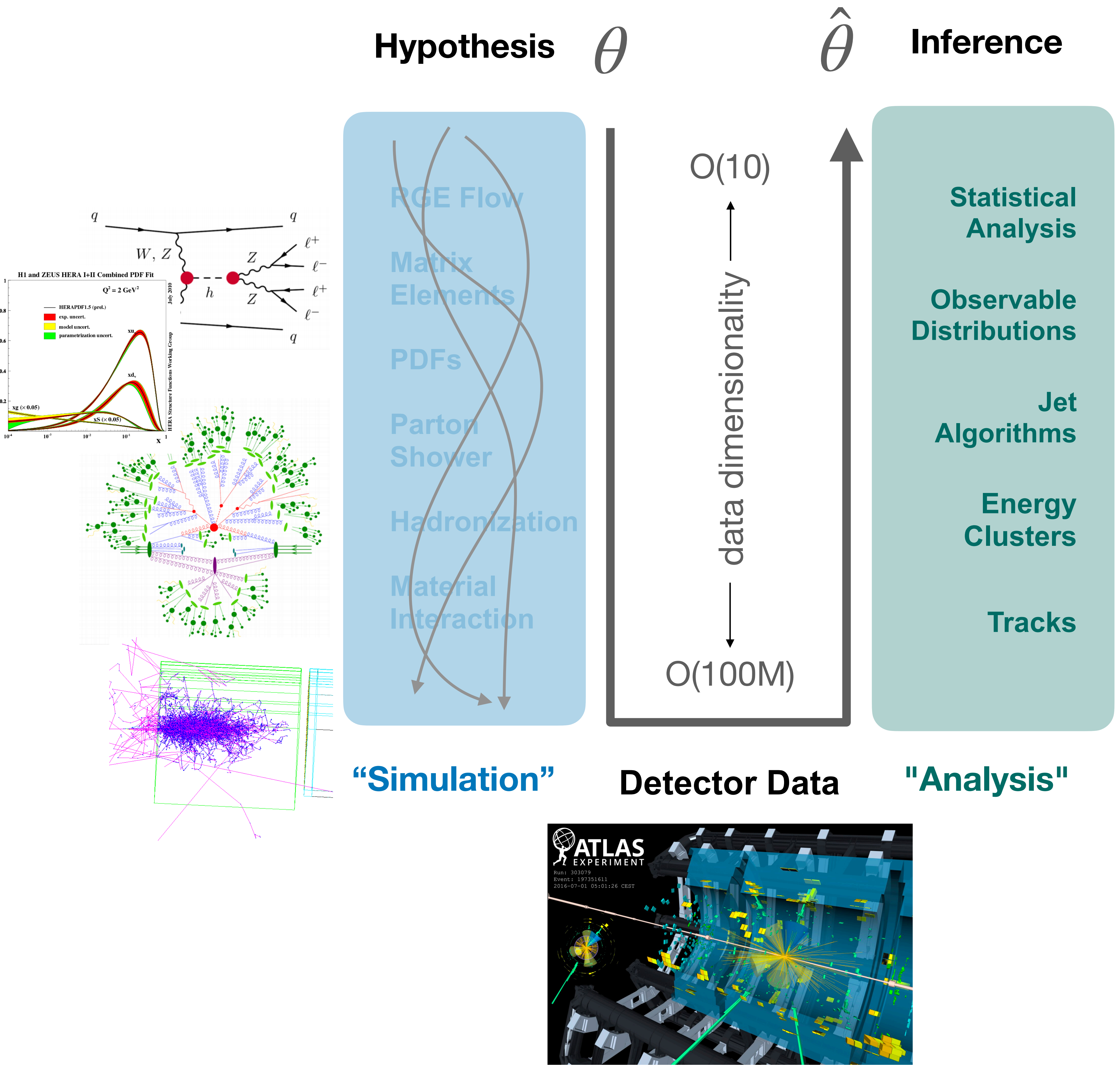
Wald's Profile Likelihood Ratio (with nuisance parameters)

$$t_\mu(x) = -2\log\frac{p(x|\mu,\hat{\hat{\nu}})}{p(x|\hat{\mu},\hat{\nu})}$$

highest average power for nested hypotheses

A slight problem

Unfortunately in HEP we cannot evaluate $p(x\,|\,\theta)$ - it's likelihood-free!



often try to at least build an approximate likelihood using smart dim. reduction e.g. reconstruction & analysis

$$p(x\,|\,\theta) \approx p(\boxed{f_{\text{ana}}(x)}\,|\,\theta)$$

proceed using standard techniques e.g. via pyhf-based models

derived $t(x)$ (e.g. approx. LR) may not be optimal, but inference will never be wrong

A simple Idea

There is a path to likelihood-free frequentist inference by exploiting the optimality properties of the test statistic we seek

If we're using $t(x)$ (e.g. likelihood ratio or profile likelihood ratio) because it is optimal….

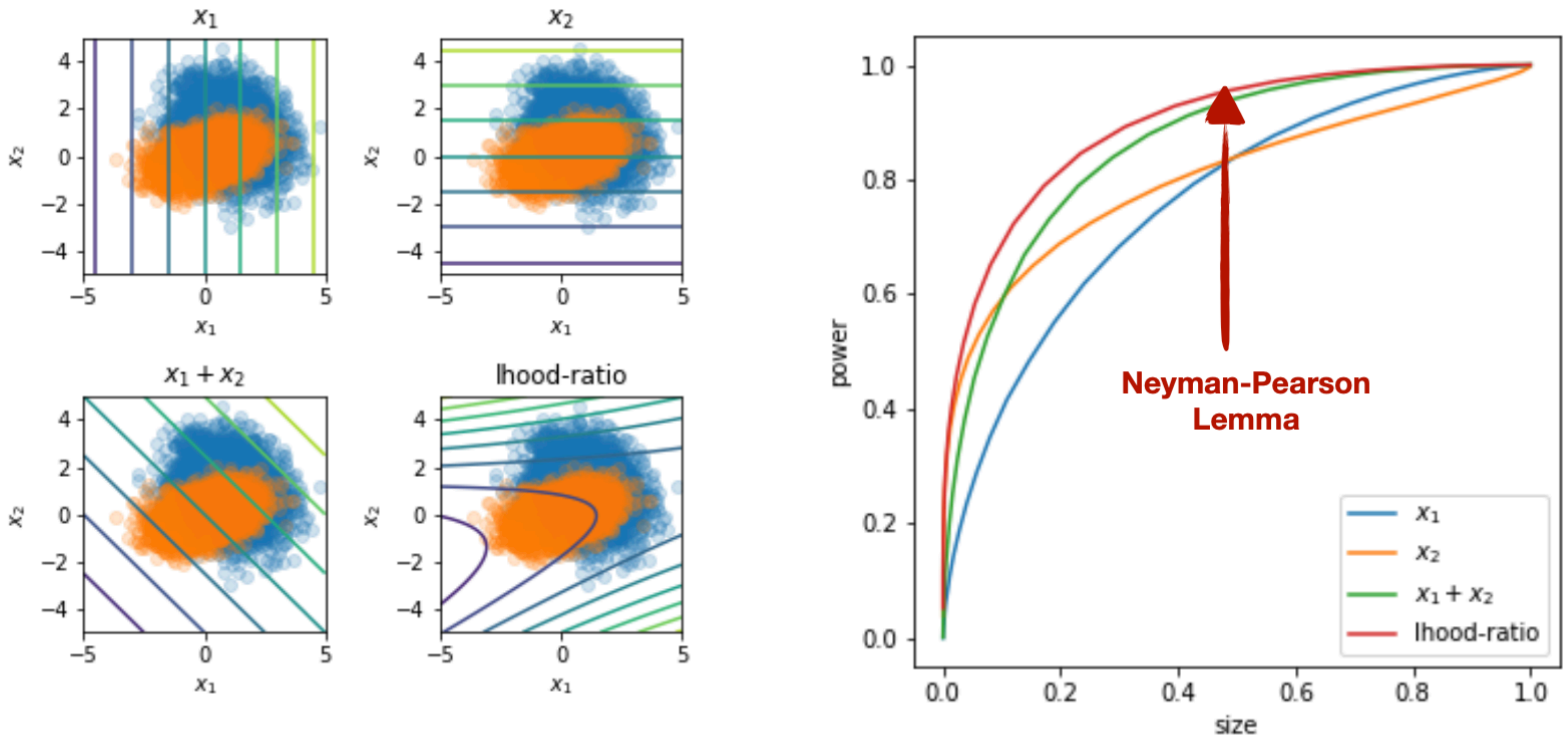
… that just means that **we can find $t(x)$ through optimization in function space, a.k.a. Machine Learning**

- **just requires samples from $p(x \mid \theta)$, not the likelihood**

Likelihood Ratio Trick

For the non-nuisance case this is the “likelihood ratio trick”

Training a Neural Network to discriminate H_0 v. H_1 will converged to a function $f(x)$ that is 1↔1 to Neyman-Pearson’s Likelihood Ratio



With Nuisance Parameters

Can we extend this to the case with nuisance parameters?

i.e. train a neural network $f(x, \mu)$ such that it converges to the profile likelihood (or a function that is 1↔1 with it)

data features

$$f_\phi(x, \mu)$$

parameters of interest

neural network weights

train against loss $L(\phi, x)$

$$t_\mu(x) = -2 \log \frac{p(x|\mu, \hat{\hat{\nu}})}{p(x|\hat{\mu}, \hat{\nu})}$$

What is the appropriate training procedure?

Go back to Wald 1943

To find appropriate training procedure to optimize $f(x, \mu) \to t_\mu(x)$ we need to recall in what sense the profile likelihood is optimal

TESTS OF STATISTICAL HYPOTHESES CONCERNING SEVERAL PARAMETERS WHEN THE NUMBER OF OBSERVATIONS IS LARGE[1]

BY
ABRAHAM WALD

TABLE OF CONTENTS

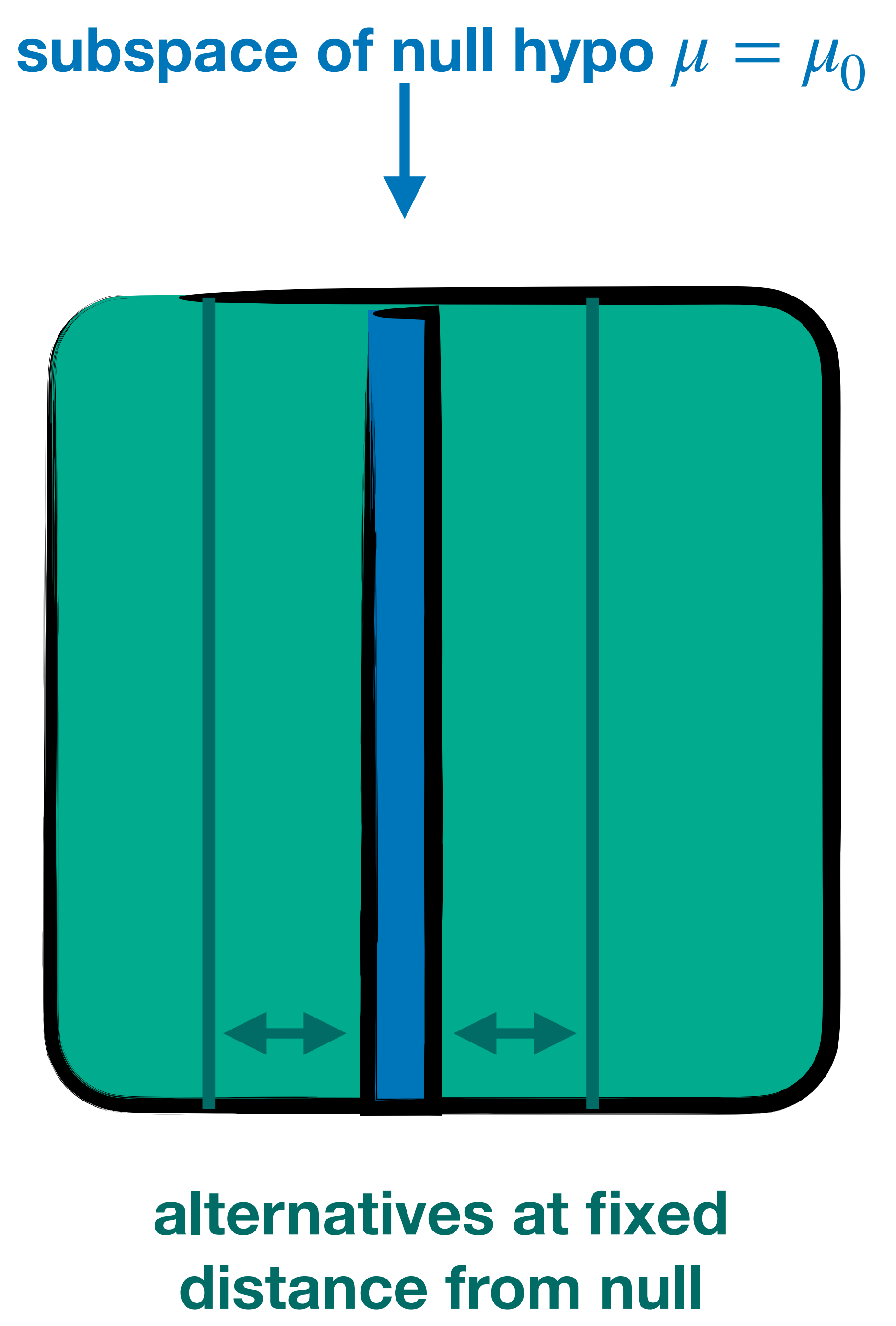
1. **Introduction.** In this paper we shall deal with the following general problem: Let $f(x^1, x^2, \cdots, x^r, \theta^1, \cdots, \theta^k)$ be the joint probability density function of the variates (chance variables) $x^1, \cdots, x^r$ involving k unknown parameters $\theta^1, \cdots, \theta^k$. Any set of k values $\theta^1, \cdots, \theta^k$ can be represented by a point θ in the k-dimensional Cartesian space with the coordinates $\theta^1, \cdots, \theta^k$. We shall denote the set of all possible parameter points by Ω. The set Ω is called parameter space. The parameter space Ω may be the whole k-dimensional Cartesian space, or a subset of it. For any subset ω of Ω, we shall denote by H_ω the hypothesis that the parameter point lies in ω. If ω consists of a single point, H_ω is called a simple hypothesis, otherwise H_ω is called a composite hypothesis. In this paper we shall discuss the question of an appropriate test of the hypothesis H_ω based on a large number of independent observations on $x^1, \cdots, x^r$.

For simplicity we shall introduce the following notations: The letter θ or θ_i for any subscript i will denote a point in the parameter space Ω. The letter x

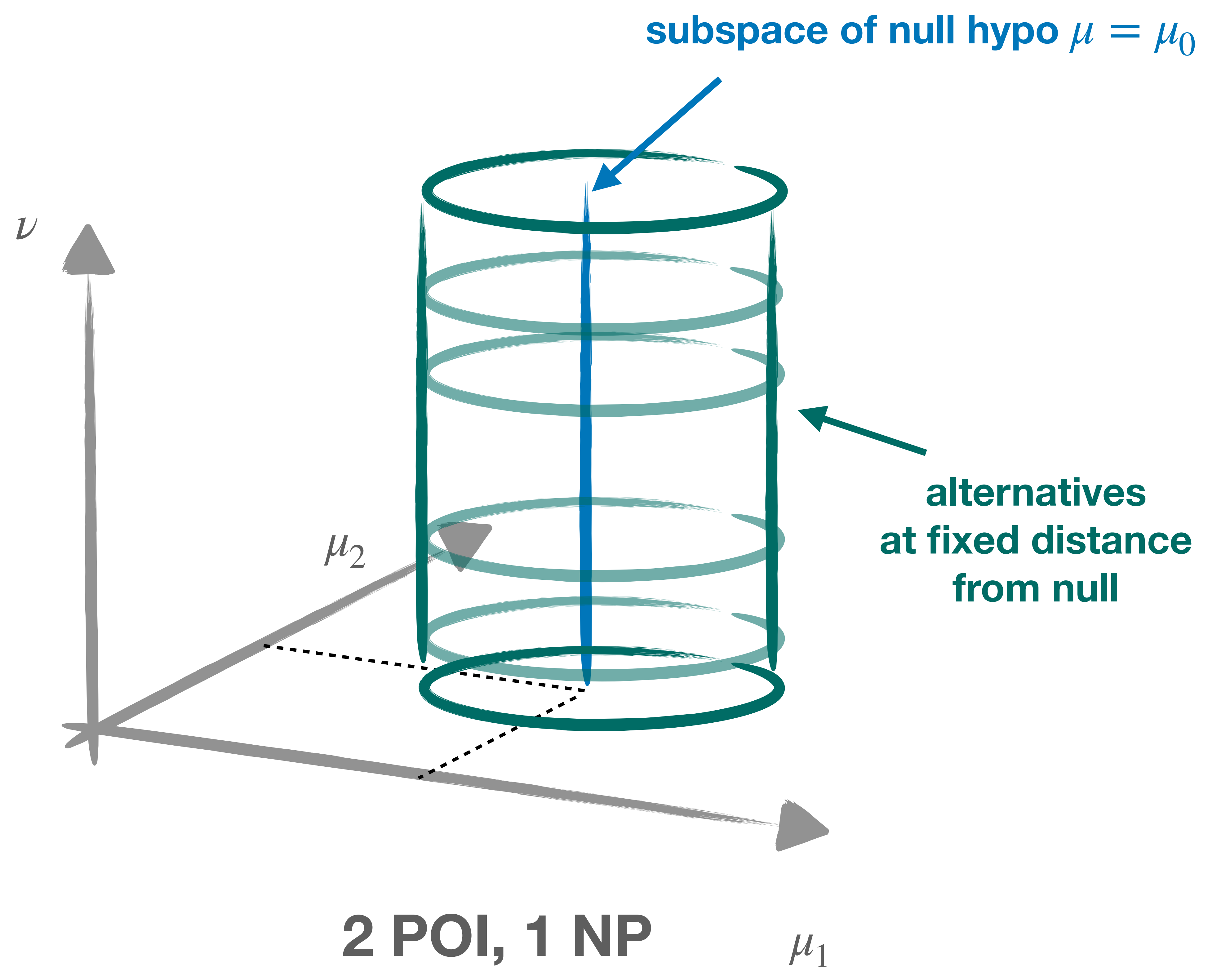
Some of the results contained in this paper were presented to the Society, February 22, 1941 and September 2, 1941; received by the editors March 31, 1943.
(1) Research under a grant-in-aid from the Carnegie Corporation of New York.

Best Average Power

Wald defines optimality as a test having best average power against alternatives "equally distant" from subspace defined by the parameters of interest



1 POI, 1 NP



2 POI, 1 NP

Best Average Power

Gives us a clear recipe on what loss to train our network on.

- LR trick: binary cross entropy optimizes for best power for fixed alternative
- Wald: Profile LR will emerge from optimized for best average power
  - **optimize on best average BXE by sampling fixed-distance alternatives and average over them. Then watch** $f(x, \mu) \to t_\mu(x)$

Algorithm 1 Training a Test Statistic with Best Average Power

---

Require: η: learning rate
Require: ϕ_0: initial parameters
Require: $\theta \sim p(\theta),\ \theta \sim p(\theta, S_c|\theta_0)$: sampling routines
while not converged **do**
  $\theta_0 = (\mu_0, \nu_0) \sim p(\theta)$ ▷ sample null
  $\theta_i = (\mu_i, \nu_i) \sim p(\theta, S_c|\theta_0)$ ▷ sample alternatives
  $(x_i, y_i) \sim p(x|\theta_0), p(x|\theta_i)$ ▷ null: $y_i = 0$, all alternatives have $y_i = 1$
  $p_i \leftarrow s_\phi(x_i; \mu_0)$
  $L = \sum_{\text{null,alts}} L_{\text{BXE}}(y_i, p_i)$
  $\phi_{i+1} \leftarrow \phi_i - \eta \nabla_\phi L$
end while
return ϕ_N

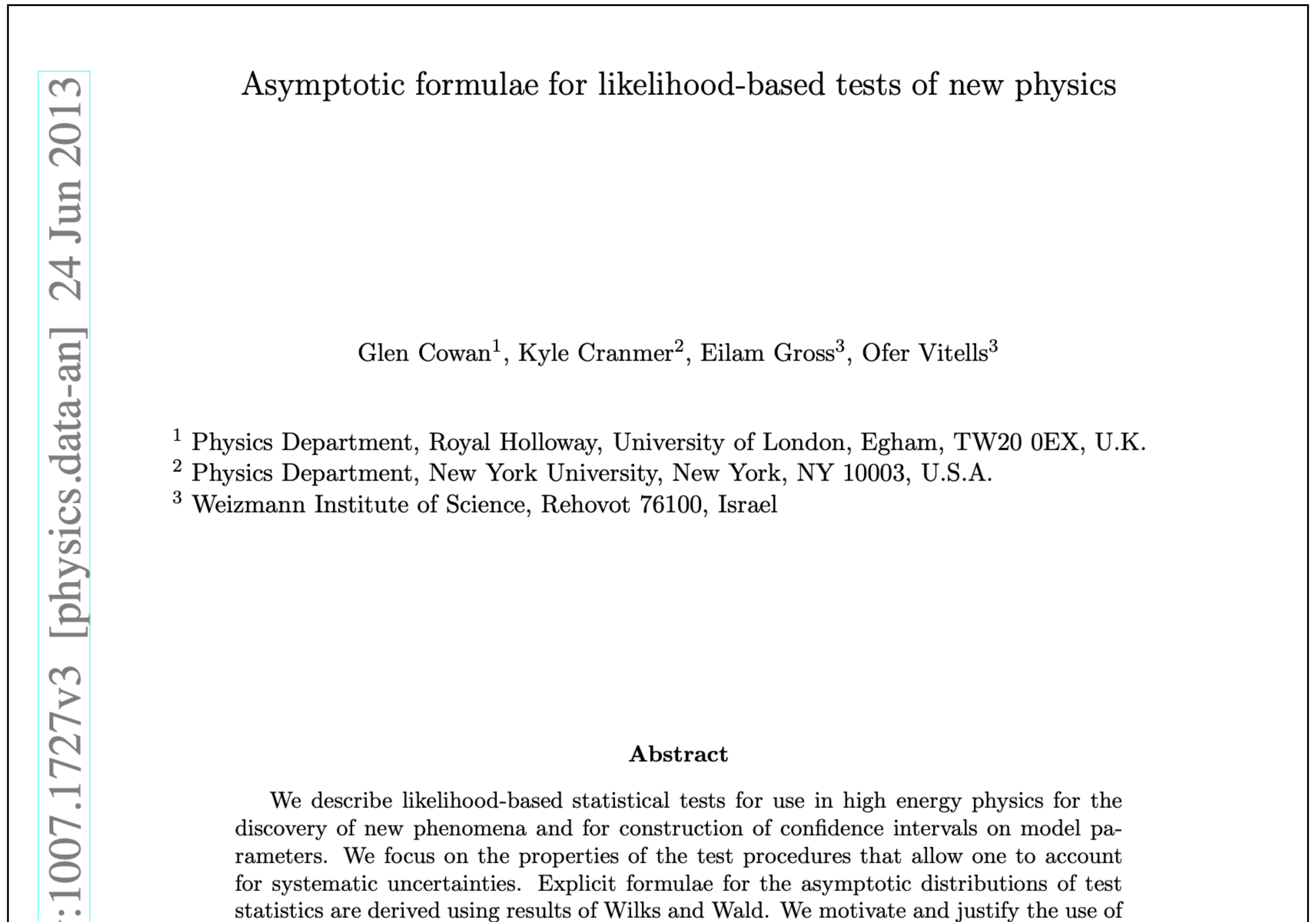
Does this work?

Check on a well-known example from HEP stats: **the on-off problem**

$$p(x_1, x_2|\mu, \nu) = \mathrm{Pois}(x_1|\mu s + \nu b)\mathrm{Pois}(x_2|\nu\tau b),$$

In this case we can solve for the true profile likelihood analytically

$$\begin{aligned}
\hat{\mu} &= \frac{n - m/\tau}{s}\,, \\
\hat{b} &= \frac{m}{\tau}\,, \\
\hat{\hat{b}} &= \frac{n + m - (1+\tau)\mu s}{2(1+\tau)} + \left[\frac{(n + m - (1+\tau)\mu s)^2 + 4(1+\tau)m\mu s}{4(1+\tau)^2}\right]^{1/2}\,.
\end{aligned}$$


:1007.1727v3 [physics.data-an] 24 Jun 2013

Asymptotic formulae for likelihood-based tests of new physics

Glen Cowan[1], Kyle Cranmer[2], Eilam Gross[3], Ofer Vitells[3]

[1] Physics Department, Royal Holloway, University of London, Egham, TW20 0EX, U.K.
[2] Physics Department, New York University, New York, NY 10003, U.S.A.
[3] Weizmann Institute of Science, Rehovot 76100, Israel

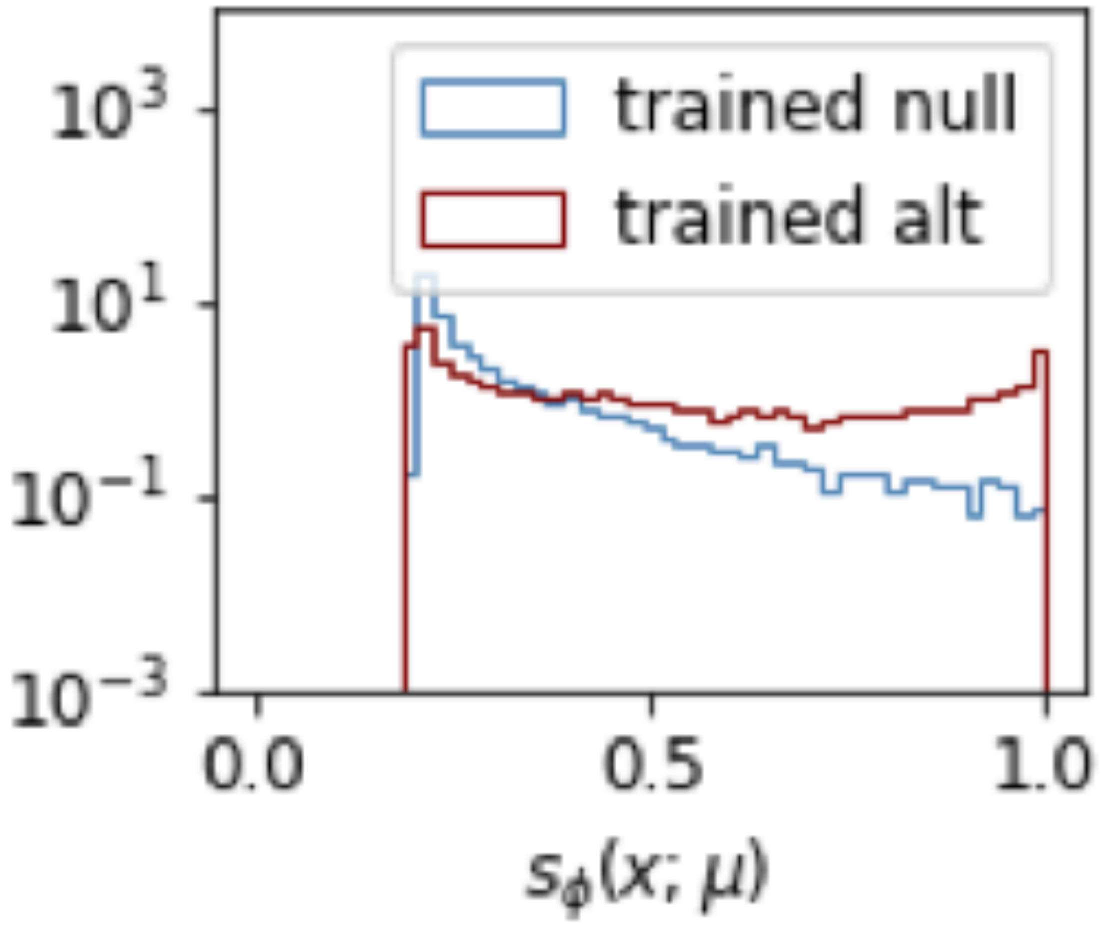
Abstract

We describe likelihood-based statistical tests for use in high energy physics for the discovery of new phenomena and for construction of confidence intervals on model parameters. We focus on the properties of the test procedures that allow one to account for systematic uncertainties. Explicit formulae for the asymptotic distributions of test statistics are derived using results of Wilks and Wald. We motivate and justify the use of

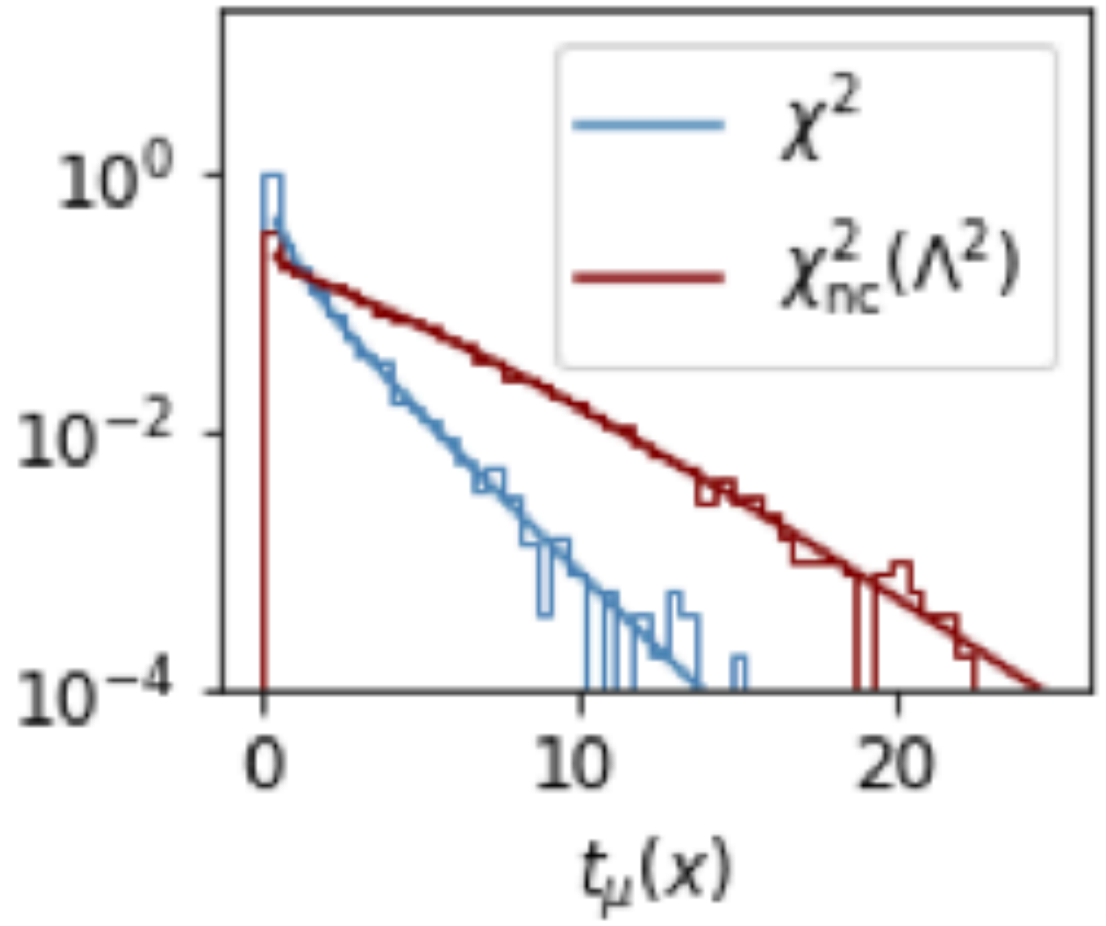
Examples

Does this work? Check on a well-known example from HEP Stats

$$p(x_1, x_2|\mu, \nu) = \mathrm{Pois}(x_1|\mu s + \nu b)\mathrm{Pois}(x_2|\nu\tau b),$$



Neural Network Training



Analytic Result well inside the asymptotic regime

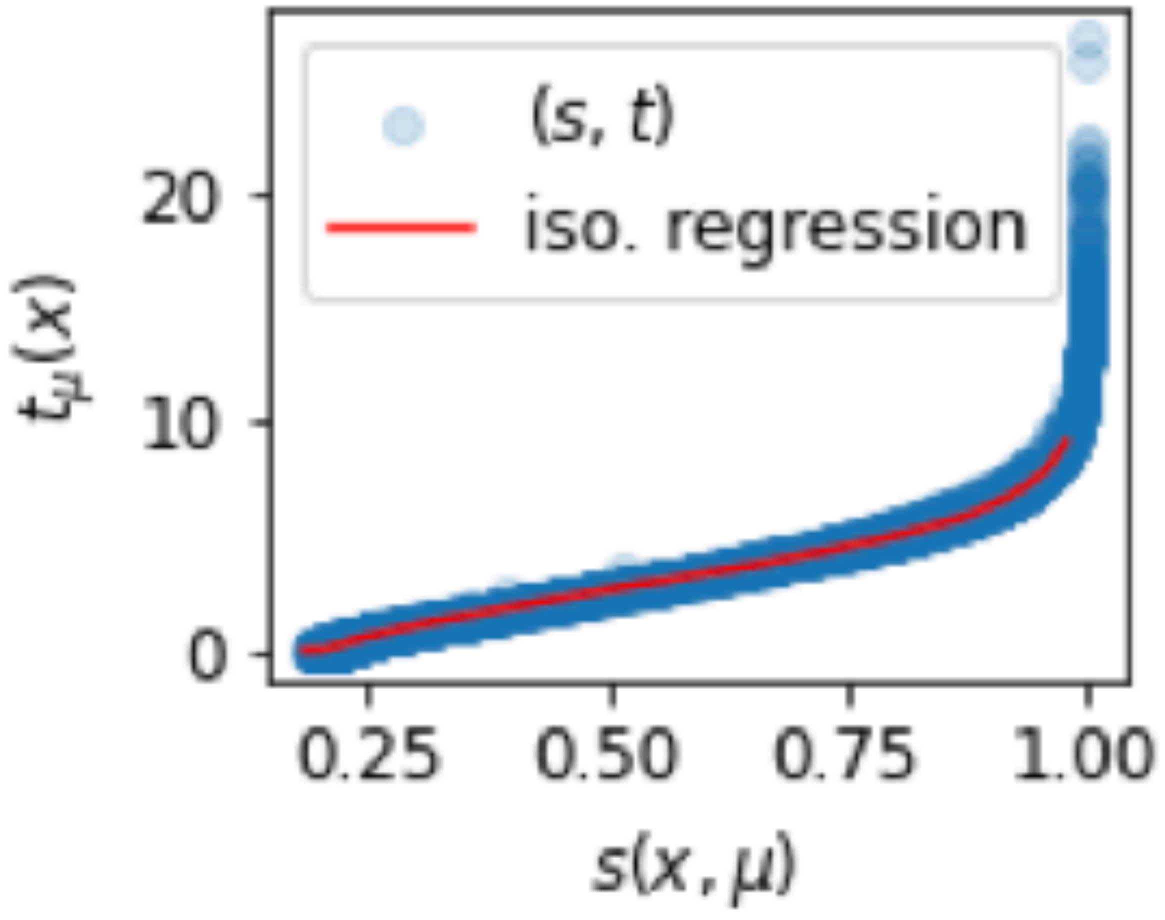
Examples

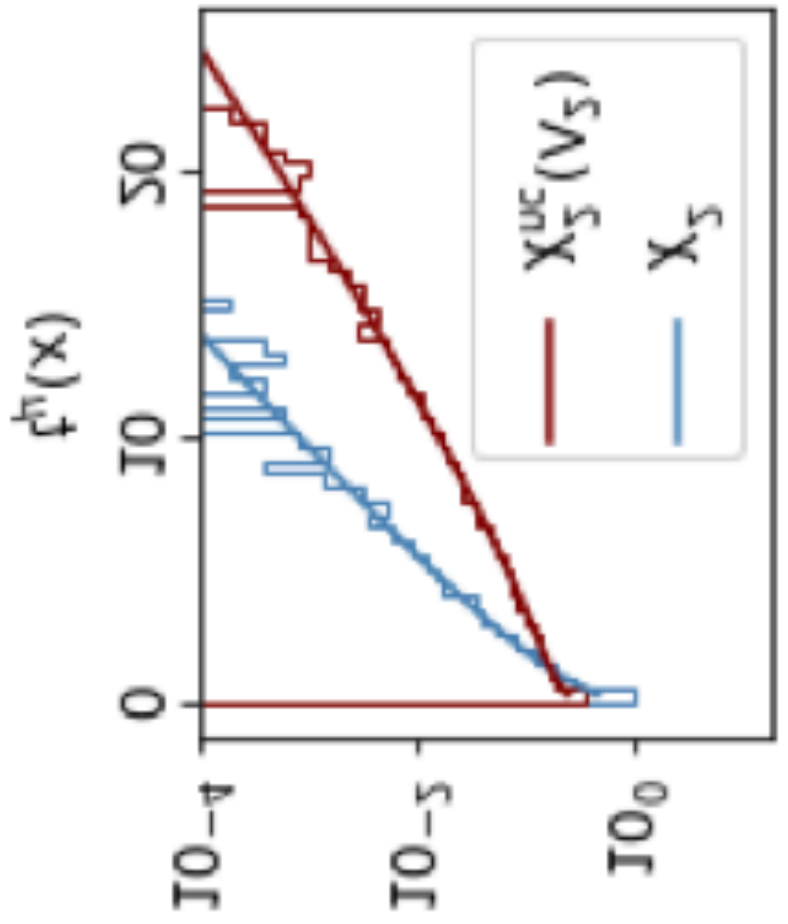
Are these two test statistics related?

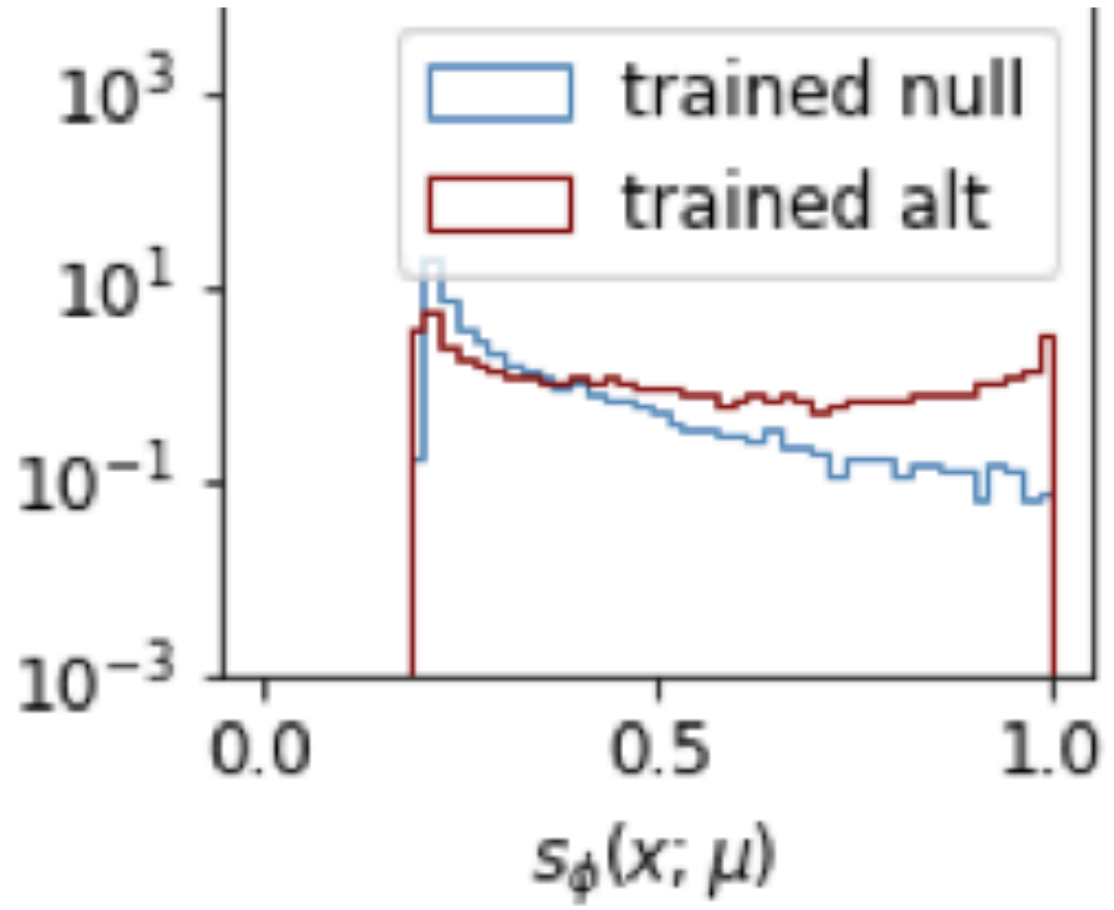
Yes: they're 1↔1

We can transform to standard χ^2-type units of just do inference in the learned test statistic

Both will produce the same results.

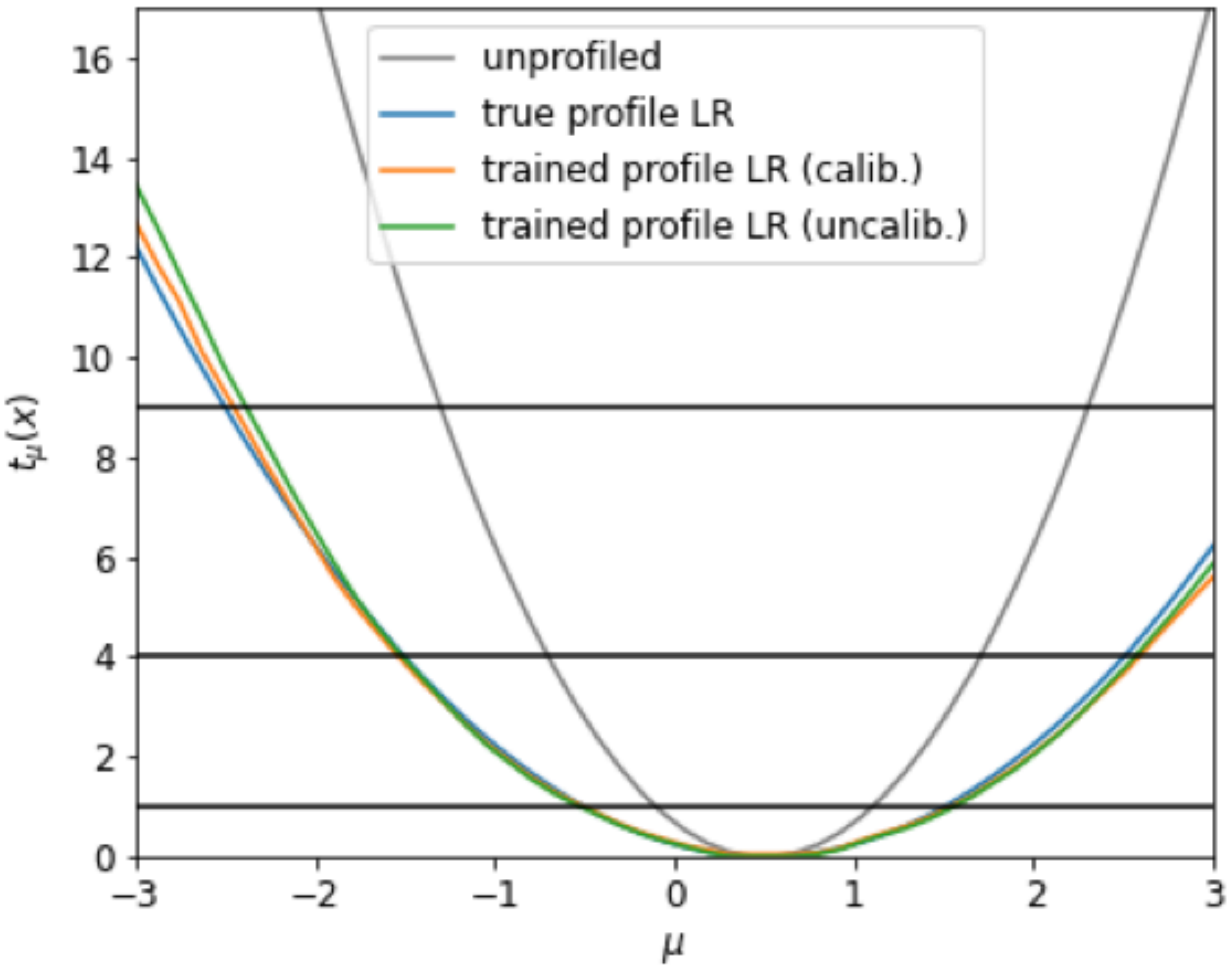






Examples

We can recover the "profile likelihood" in a fully likelihood free way



Summary

Described method to compute $$t_\mu(x) = -2\log\frac{p(x|\mu,\hat{\hat{\nu}})}{p(x|\hat{\mu},\hat{\nu})}$$

- without evaluating $p(x \mid \theta)$
- without running any fits to find $\hat{\mu}, \hat{\nu}, \hat{\hat{\nu}}$
- just using samples from $p(x \mid \theta)$

Taking Wald's optimality criterion seriously and using it as an optimization objective: extension of LR trick to case of nuisance parameters

Parting Note: the profile likelihood is only optimal asymptotically. In this case this method will find it, otherwise it'll find whatever is actually optimal.