6th Workshop of (Re)Interpretation Forum

Harrison B. Prosper
Florida State University

15 February 2021

OPEN SCIENCE NEEDS OPEN LIKELIHOODS

Outline

- Open Science
- Likelihood Functions
- Science 2061
- Summary

OPEN SCIENCE

From Closed Science…

In 1609, Galileo turned his telescope to Saturn and encrypted his discovery in an anagram:

s m a i s m r m i l m e p o e t a l e u m i b u n e n u g t t a u i r a s

which when unscrambled read:

Altissimum planetam tergeminum observavi

“I have observed the highest planet to have a triple form.”

…to Open Science

CERN Convention (1953):

"…the results of its experimental and theoretical work shall be published or otherwise made generally available"

We have had open access to *published results* for decades.

In recent years, the statement in the CERN Convention has been interpreted to imply open access to *data* also.

CERN announces new open data policy in support of open science

A new open data policy for scientific experiments at the Large Hadron Collider (LHC) will make scientific research more reproducible, accessible, and collaborative

11 DECEMBER, 2020

English (United States)

Why Open Science?

1. More brainpower means, potentially, more creative science.

   According to the World Bank, in the age group 15 – 64 years, there are **5.0 billion brains**. Compare this with the **2500** permanent brains at CERN + its **8000** visiting brains.

2. Data acquired today (at great cost) could yield novel science tomorrow.

3. Our future may depend on it!

LIKELIHOOD FUNCTIONS

Likelihood Functions: 1

Ideally, we should distinguish two distinct but related concepts:

1. A **probability model** (of the data generation mechanism)

   $p(x, \theta) = p(x|\theta)\, p(\theta)$, where x denotes data that are possible *a priori* and θ are model parameters.

2. A **likelihood function**

   $p(D|\theta)$, which is the conditional model $p(x|\theta)$ into which the *observed* data D have been inserted.

Likelihood Functions: 2

Sometimes in order to highlight the difference between $p(x|\theta)$ and $p(D|\theta)$, the latter is written as

$$L(\theta) \equiv p(D|\theta).$$

This emphasizes the fact that the data D are *known constants* and the likelihood is a function of the *unknown* parameters θ.

However, $p(D|\theta)$ is preferred so that theorems such as $p(\theta|D) = p(D|\theta)p(\theta)/p(D)$ look correct.

Likelihood Functions: 3

The parameters θ can be partitioned into at least four classes:

1. Parameters of interest
   Example: A cross section
2. Parameters that are internal to an analysis
   Example: The parameters of a background model
3. Parameters that are internal to an experiment
   Example: The jet energy scale
4. Parameters that are external to an experiment
   Example: The parameters of MadGraph or CT14nlo

Handling these four classes in full generality is challenging:
Reinterpretation of LHC results for new physics: status and recommendations after run 2, The LHC BSM Reinterpretation Forum, SciPost Phys. 9, 022 (2020).

Likelihood Functions: 4

Example: Consider binned data. The mean count per bin can be written in a model-agnostic way as

$$n = \sigma\ell + b$$

where σ, ℓ, and b are unknown *parameters*; the former, the the parameter of interest, is the effective cross section, ℓ is the integrated luminosity and b is the total background.

The connection to a given physics model is through the effective cross section σ, which depends on the parameters α of the model through some conditional density $f(\hat{\sigma}|\sigma, \alpha)$, where $\hat{\sigma}$ is the predicted cross section and f encodes whatever theoretical uncertainties are judged to be relevant.

Likelihood Functions: 5

The integrated luminosity is constrained either by a likelihood function one of whose parameters is ℓ or by a *prior density* for ℓ.

Likewise, each background component is constrained by a probabilistic function, which itself is constrained by multiple probabilistic functions that encode uncertainties* that affect the background estimate.

*There is uncertainty in the jet energy scale and jet energy resolution; uncertainty due to the finite size of a data sample used in a data-driven background estimate; uncertainties in various scale factors, and uncertainties in the models used in simulations of some backgrounds.

Why Are Open Likelihoods Useful?

By an open likelihood I mean simply a function in which both the data and the functional form are public.

Likelihoods are useful (in principle) because:

1. They can model non-Gaussian features.
2. They permit the optimal combination of results.
3. They permit the use of different inference methods.
4. They permit global analyses, e.g., of astrophysical and accelerator data.

Note, however, in order to benefit from the above, it is important to make public a precise definition of the analysis pipeline that led to the data in the likelihood.

A Modest Proposal (from 2011!)

1. Release all CDF and Dzero internal notes to INSPIRE

2. Make *physics* data associated with publications open.

   - Release the likelihood functions $p(D \mid \theta, \omega)$ of all analyses in RooStats workspaces.

   - Have CMS and ATLAS Collaborations officially endorse *open-source* fast simulations of their respective detectors. An excellent candidate is Delphes.

SCIENCE 2061

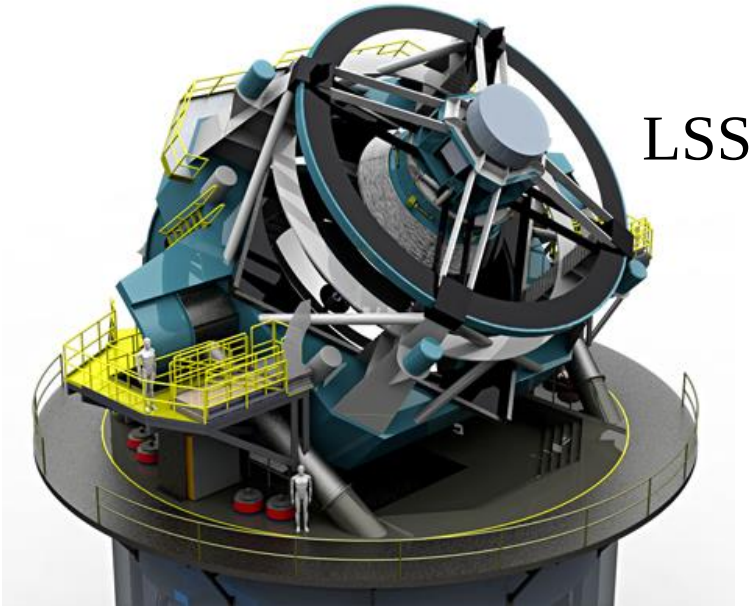
Exploring The Data Landscape

Hanford, Washington

Livingston, Louisiana

LIGO
Virgo

LHC



LSST

Science 2061

- Many members of the generation who may wish to re-analyze the LHC Run 3 data in 2061 have yet to be born.

- That generation may wish to re-analyze these data simultaneously with old LSST and LIGO data using BSM models that they invent from which (with the help of their AI assistants) they are able to compute testable predictions for LHC, LSST, and LIGO data.

- Now ask yourself: What are we doing wrong, or what are we forgetting, that risks thwarting such heroic efforts forty year hence?

Summary

- Open Data is now a welcome reality in particle physics that will further the move towards Open Science.

- The publication of full likelihoods is also a reality. However, work is needed to tame the explosion in the number of nuisance parameters, while preserving the ability to account for dependences across analyses whether they be within an experiment or across experiments.

- But we can already go a long way towards making the data of the LHC even more valuable by adhering to some field-wide recommendations: SciPost Phys. 9, 022 (2020).