The Bayes factor surface

[2401.11710]

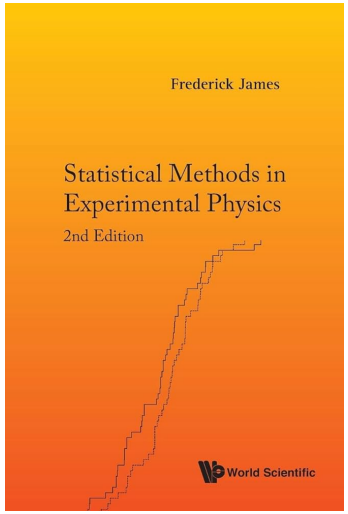
Andrew Fowlie

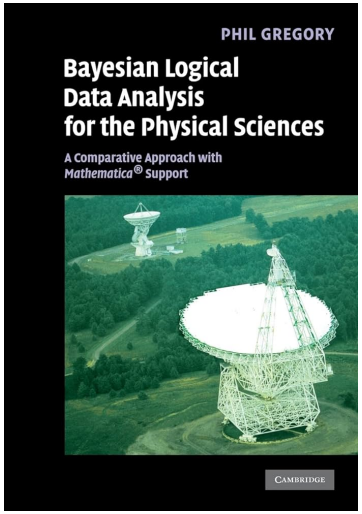
Annual Symposium for Young Mathematicians
in the Suzhou Region | Soochow University

10 November 2024

Xi'an Jiaotong Liverpool University
西交利物浦大学

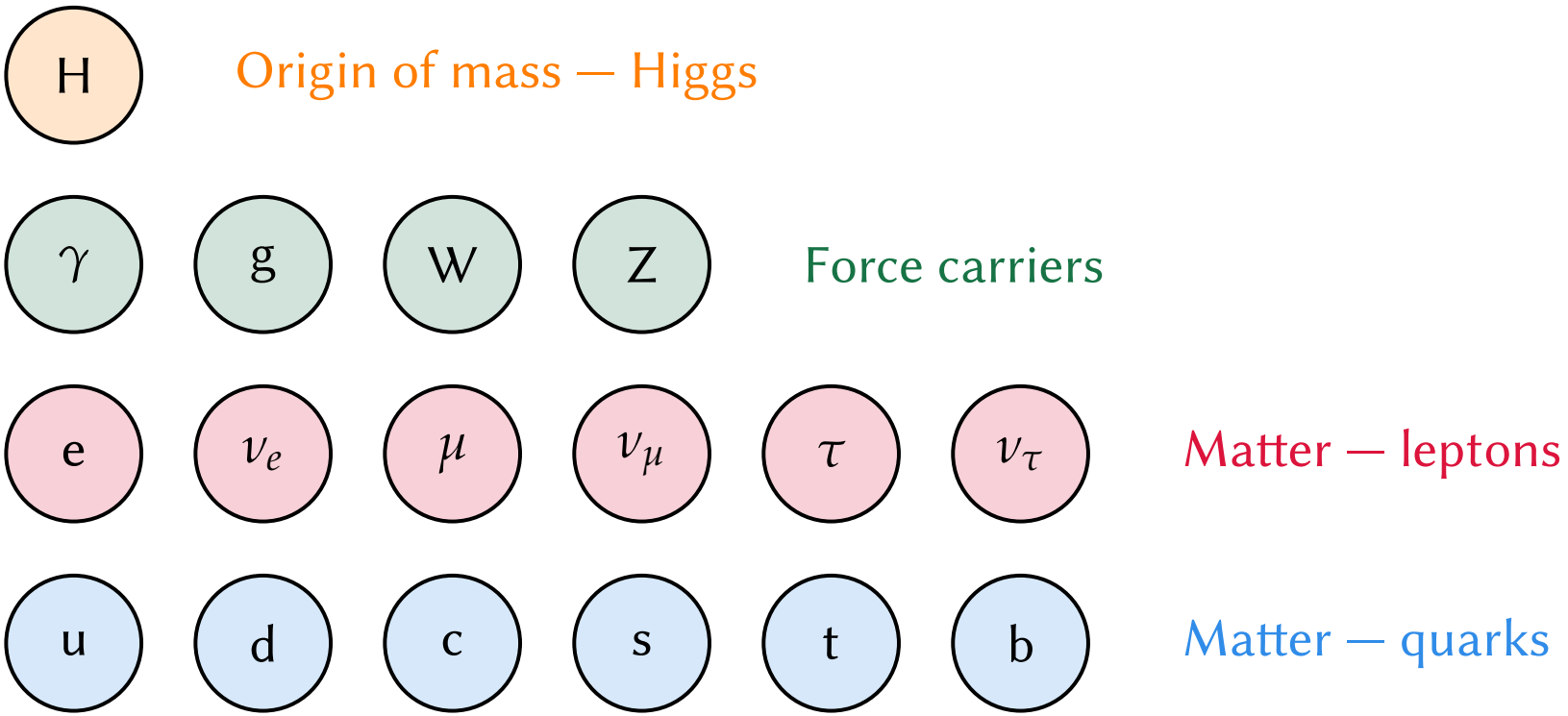
Context



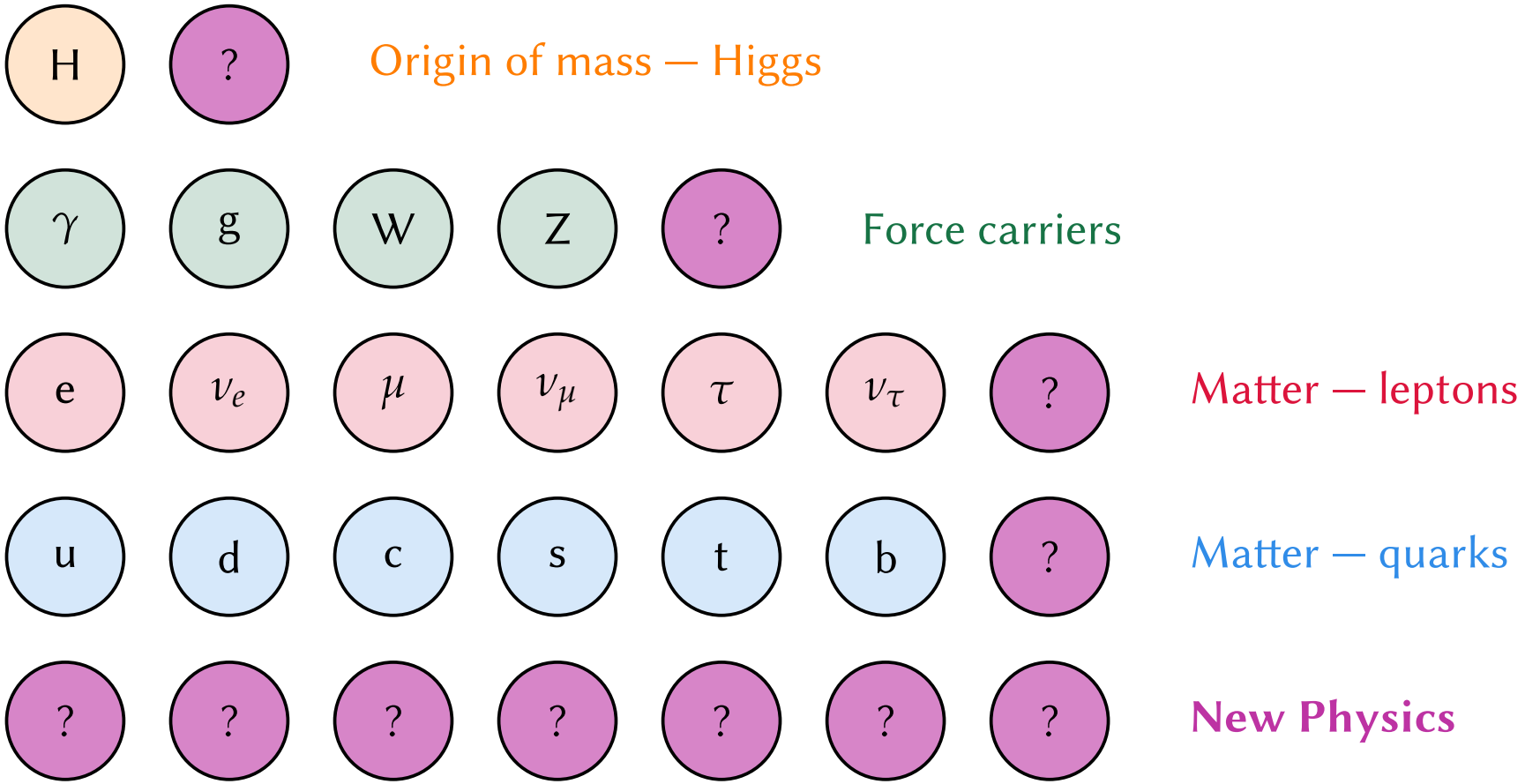


- I am a particle physicist phenomenologist
- I began my Ph.D. around start of Large Hadron Collider (LHC) in 2009
- Focused on statistical analyses of supersymmetric models in light of first LHC data and direct searches for dark matter (Fowlie, 2013)
- Early in Ph.D., studied statistical methods in physics from these two books

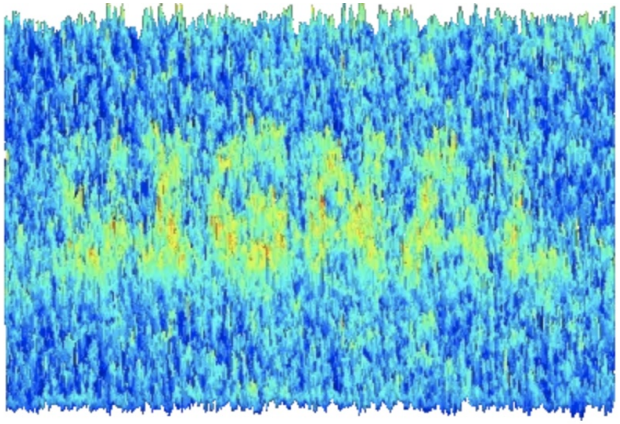
The Standard Model



Beyond the Standard Model



Searches for new physics



- In the hunt for new physics, we search for a signal on top of a noisy background
- Signal model usually parameterised by an effect size, e.g., a cross section for the production of a new particle
- Background model often equivalent to effect size = 0
- There could be nuisance parameters describing systematic uncertainties
- Signal model could have other unknown parameters, e.g. the unknown mass of a new particle

Statistical practices in searches for new physics

In particle physics, experimental results of searches for new particles are shown by

- Confidence intervals for effect size (e.g. cross section) and other parameters of interest
- P-values for discovery of new particles

This is done through

- Statistical recipes using Wilks' theorem; see Cowan et al., 2011. These recipes are popular — about 500 cites/year
- Exact prescriptions and conventions still debated — see PHYSTAT-DM 2019, about 50 participant workshop for conventions for dark matter searches
- Led to recommendations paper on confidence intervals (Baxter et al., 2021)

Confidence interval

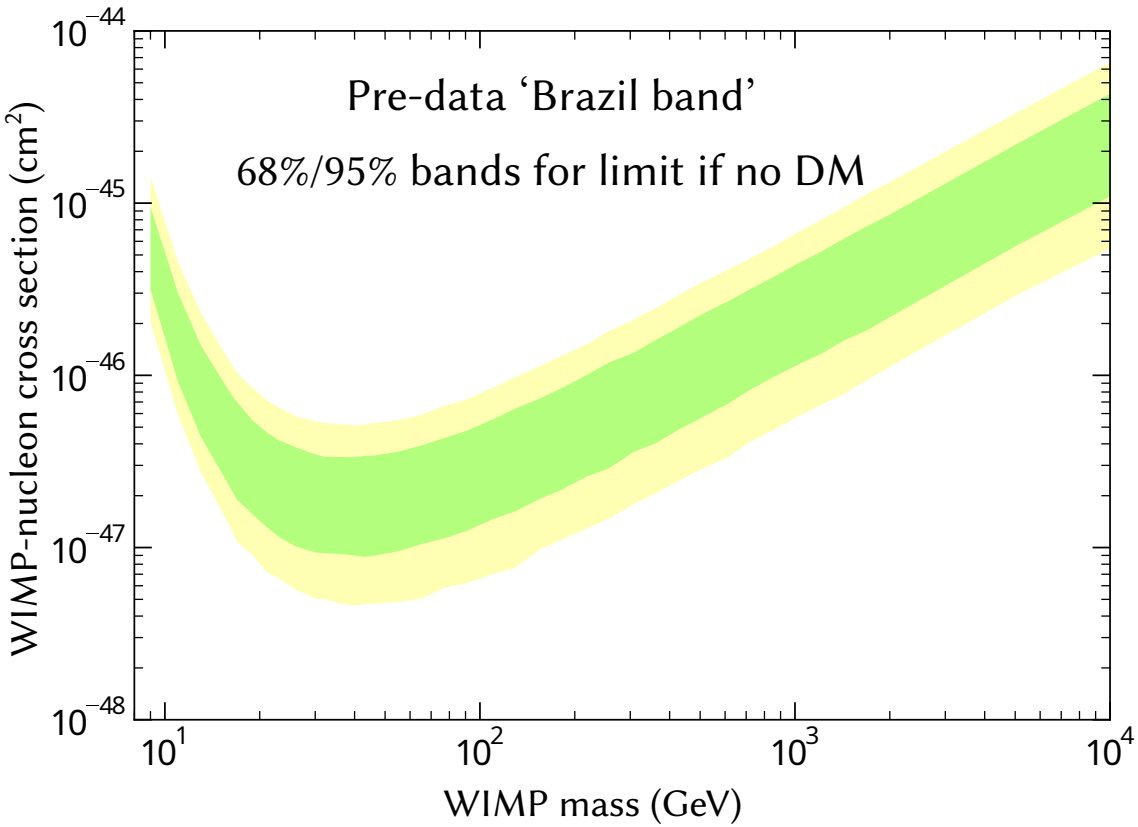
Definition

An interval found from a procedure that, under repeated sampling, would include the true value of the unknown parameter at a guaranteed rate (Neyman, 1937).

- The desired rate is known as the confidence level, and 90% and 95% are common choices
- Procedures that guarantee that the rate is exactly the confidence level are known as exact
- Those that guarantee that the rate at least as great as the confidence level are known as valid

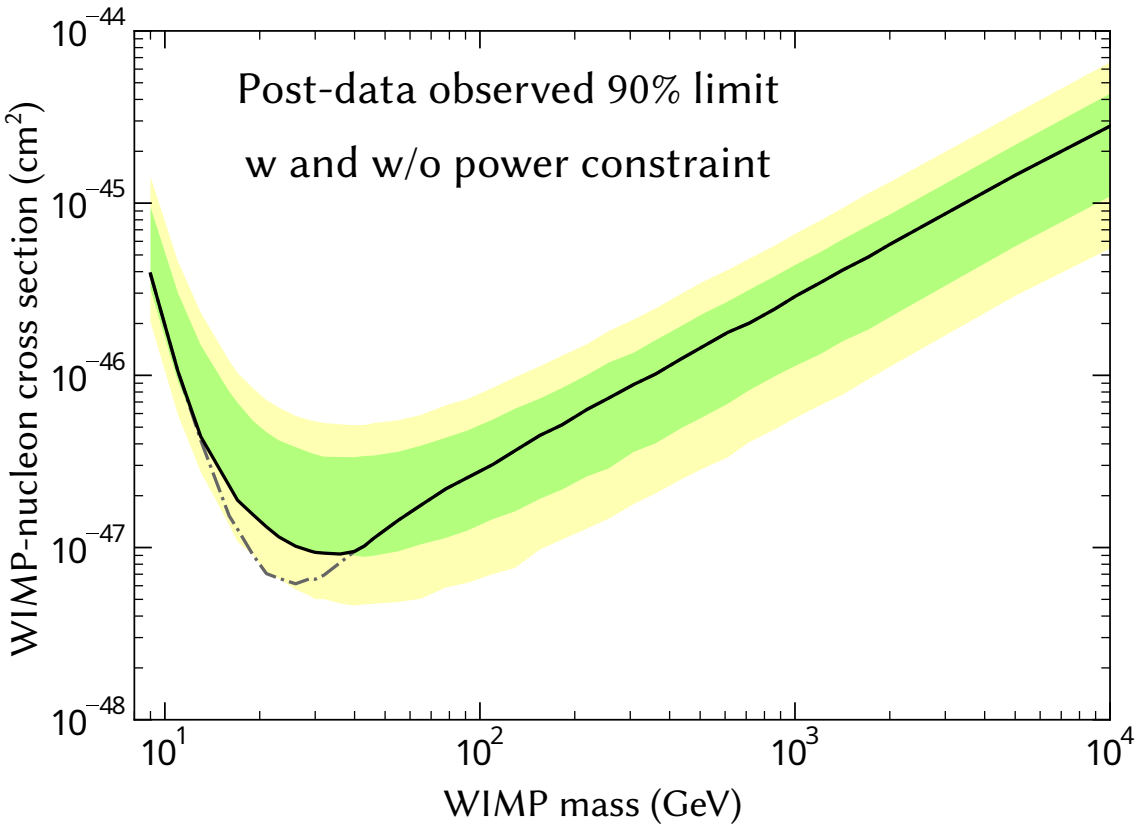
Search for dark matter

Result from LZ shown as a one-sided confidence interval (upper limit) on cross section (an effect size) as function of mass (Aalbers et al., 2023)



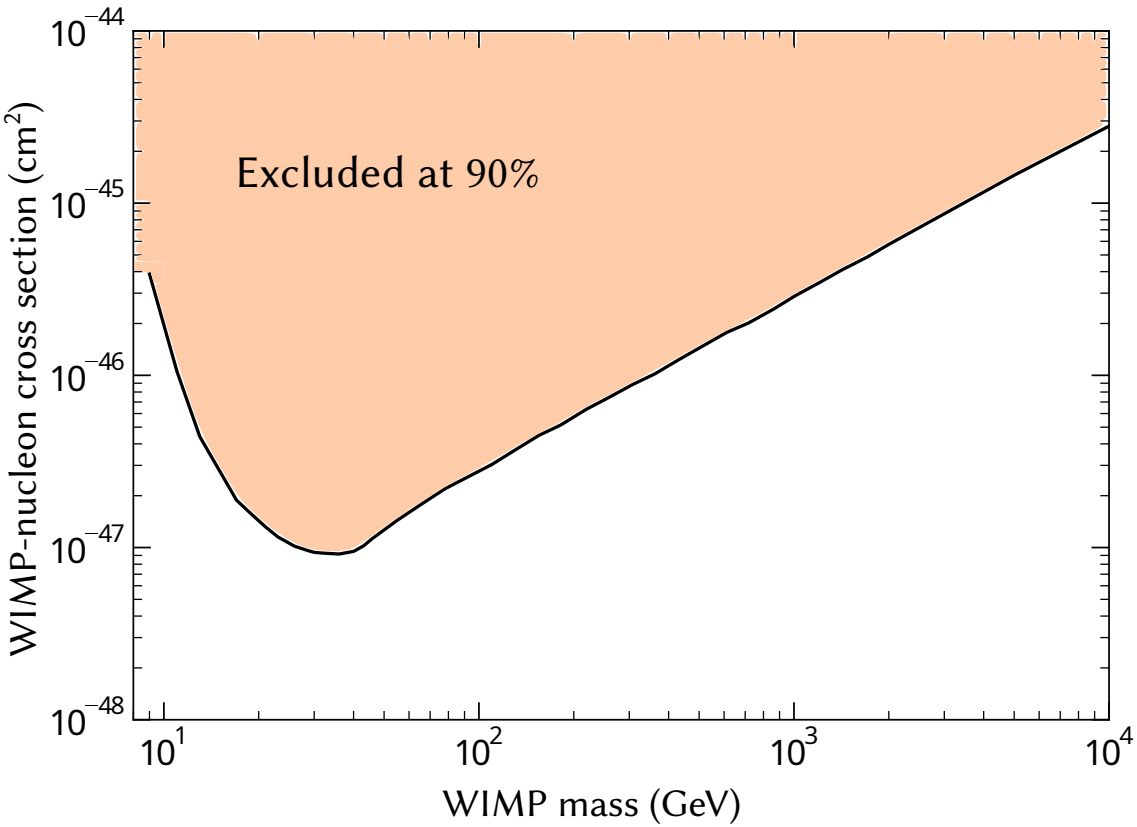
Search for dark matter

Result from LZ shown as a one-sided confidence interval (upper limit) on cross section (an effect size) as function of mass (Aalbers et al., 2023)



Search for dark matter

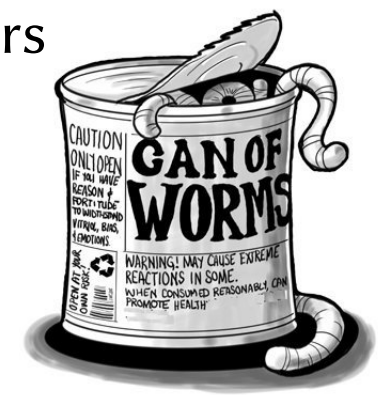
Result from LZ shown as a one-sided confidence interval (upper limit) on cross section (an effect size) as function of mass (Aalbers et al., 2023)



Can of worms

Properties of confidence intervals are not always satisfactory, even to adherents:

- Flip-flopping — coverage spoilt by choosing between one- or two-tailed limits after seeing data
- False exclusion — confidence intervals exclude effect sizes at rate of e.g., 5%. Including arbitrarily tiny effect sizes to which the experiment had no power
  - This led to creation of CL_s
- Systematics — hard to handle nuisance parameters describing systematic uncertainties



Can of worms

On more general and philosophical grounds, in my opinion,

- Whole frequentist edifice is full of thinking traps
- Prone to misinterpretation — hard to interpret or communicate a confidence interval (Morey et al., 2016)
- Informal interpretation — that effect sizes outside CI are ruled out or implausible — hard to justify

Something else

Easy to criticise confidence intervals. Can we do better?

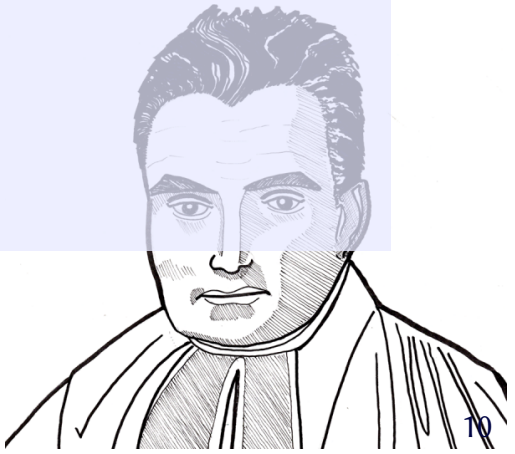
- Obviously, I wanted to be Bayesian
- The Bayesian analogue of the confidence interval — the credible region (Jaynes, 1989a)!

Definition

A credible region, $\mathcal{R}$, contains a specified percentile of posterior probability, e.g., 95%,

$$\int_{\mathcal{R}} p(\theta \mid x, M)\, \mathrm{d}\theta = 0.95$$

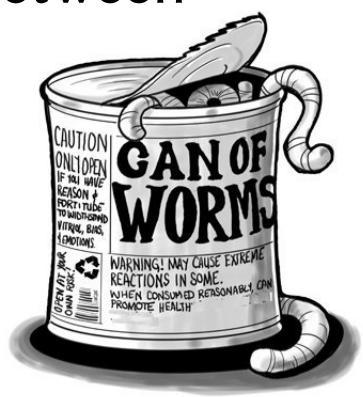
for parameter θ, data x and model M.



More worms!

In this context, credible regions are not much better

- Depend on ordering rule — which region $\mathcal{R}$?
- This induces a dependence on parametrization
- Sensitive to arbitrary aspects of the prior
- Cannot tell us whether a model predicting particular mass and cross section any better than background-only model
- This is connected to fact that there is no duality between testing and measurement



What is the question?

For a while I was stumped.

- If you don't like the answer, perhaps you asked the wrong question (Jaynes, 1989b)
- What question do we want to ask?

We don't like the credible region, because it answers the wrong question. We don't want to know the plausible parameters in a dummy model of (mass, cross section).

The question is: how do choices of (mass, cross section) compare to the background only model?

Bayes factor surface

Definition

For parameters θ and data x, compute the Bayes factor

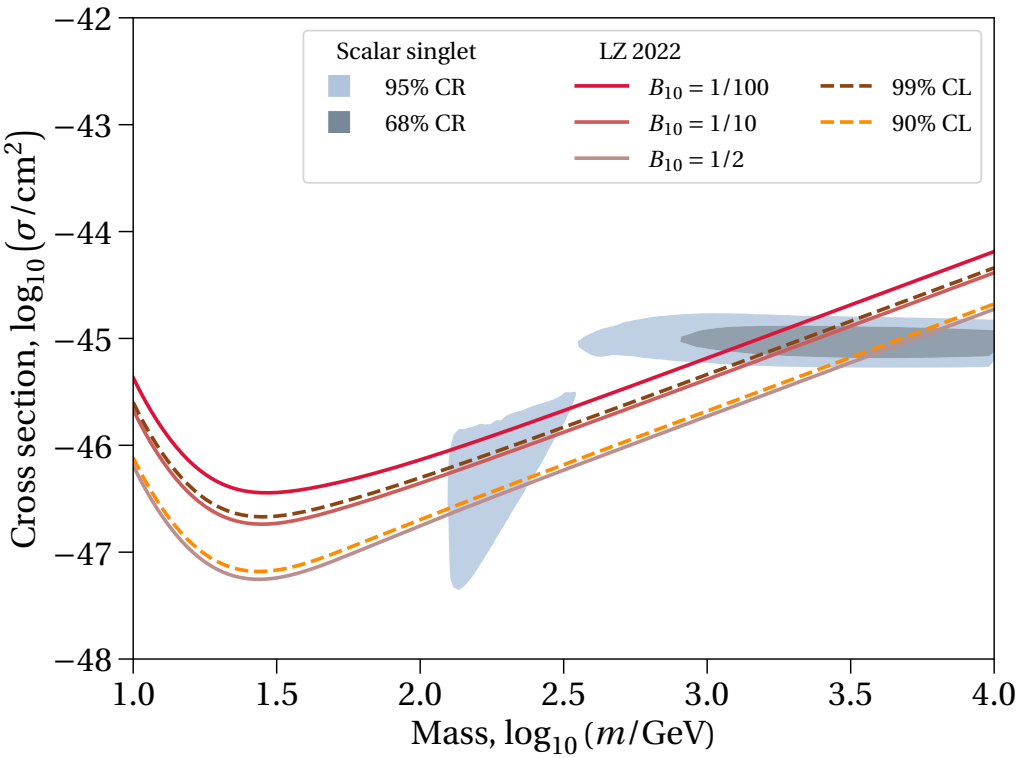
$$B_{10}(\theta) \equiv \frac{p(x \mid \theta, H_1)}{p(x \mid H_0)},$$

Show results of search by contours of $B_{10}(\theta)$. E.g., the contour of θ for which $B_{10}(\theta) = 1/10$.

- Does not depend a choice of prior, parametrization or ordering rule
- Directly tells us change in plausibility of model predicting particular parameters θ relative to background only model

Back to direct detection of dark matter

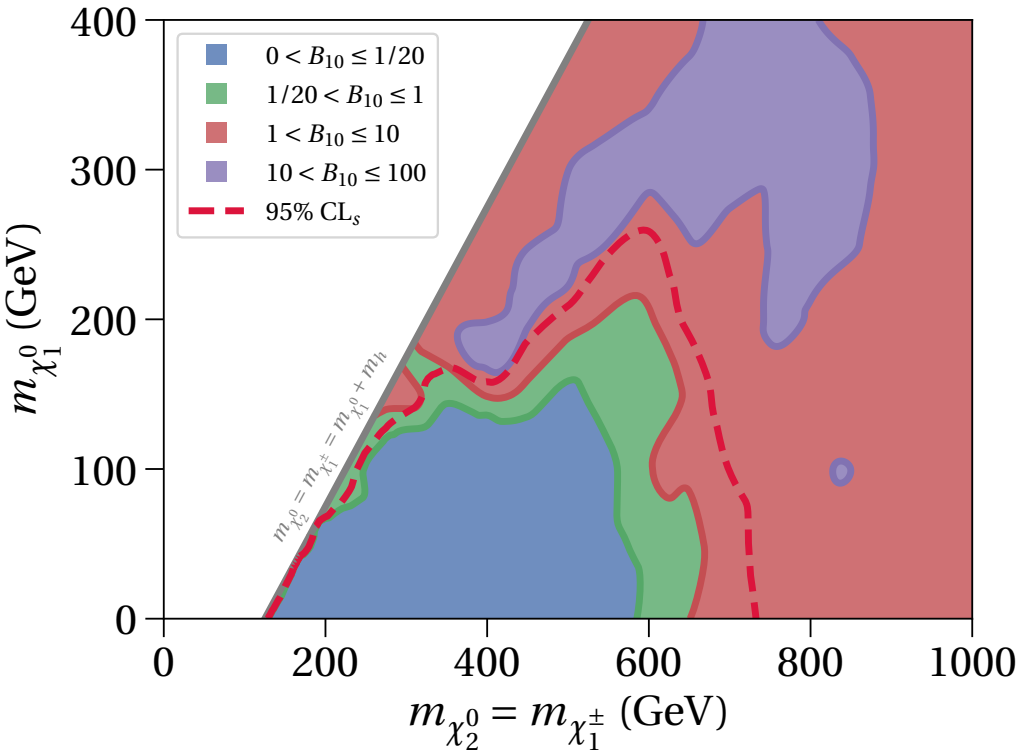
We compute the Bayes factor surface, as well as confidence limits



We can tell by eye change in plausibility of models making particular predictions

Searches for supersymmetry

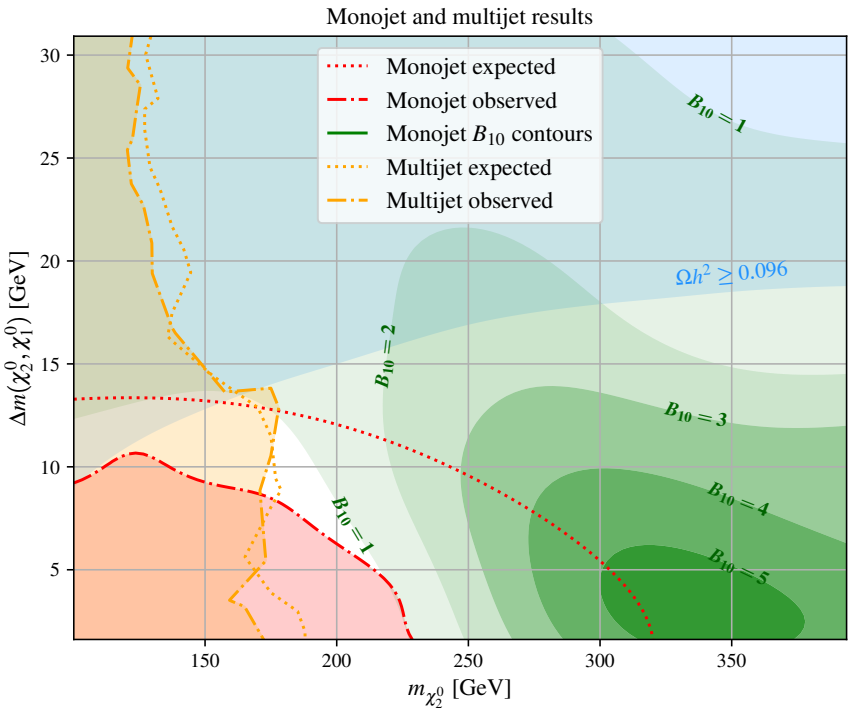
Because the idea is simple and general, it works in many other cases. E.g., an LHC search for supersymmetric particles



Do any fundamental theories make predictions in the purple region?

In the Wild

Used by Goodsell, 2024; Fuks, Goodsell, and Murphy, 2024. E.g.,



This plot from Fuks, Goodsell, and Murphy, 2024 the impact of a monojet search in a model of dark matter with a compressed spectrum

Computational methods

There are often nuisance parameters, such that we cannot just directly compute the Bayes factor surface; we need to marginalise

In searches for new physics, models are often nested.

- This means that we can use a Savage-Dickey ratio (Dickey, 1971) to compute the Bayes factor surface
- Compute posterior usually conventional methods, e.g., Markov Chain Monte Carlo
- Use e.g., kernel density representation of posterior and identity

$$B_{10}(\theta) = \frac{p(x \mid \theta)}{p(x \mid \theta_0)} = \frac{p(\theta \mid x)}{p(\theta_0 \mid x)} \frac{\pi(\theta_0)}{\pi(\theta)}$$

where θ_0 correspond to background-only model

Connections

When I wrote the paper, I knew that

- Bayes factor surface had been proposed as a tool to study prior sensitivity — show the change in Bayes factor (Franck and Gramacy, 2020) when changing e.g., width of prior
- Bayes factor function had been proposed (Johnson, Pramanik, and Shudde, 2023) — one-dimensional and motivations somewhat unclear to me

Since writing the paper,

- I found the NANOGrav paper that uses a similar construction (Afzal et al., 2023)
- Statisticians contacted me to bring to my attention relevant works from the recent statistics literature (Wagenmakers et al., 2020; Pawel, Ly, and Wagenmakers, 2023)

A rose by any other name

- Bayes factor surface
- Bayes factor function
- Support interval
- K-ratio

Name will settle as idea matures

Summary

- Bayes factor surface — simple and intuitive way to summarise searches for new effects
- Shows change in plausibility of e.g. effect size relative to no effect model
- Wide-ranging applications across science, physics and beyond
- Idea popping up independently in various contexts
- Computational methods and some properties (e.g., frequentist properties) in their infancy

A new tool to present results of searches for new effects!

References I

- Aalbers, J. et al. (2023). “First Dark Matter Search Results from the LUX-ZEPLIN (LZ) Experiment”. *Phys. Rev. Lett.* 131.4, p. 041002. arXiv: 2207.03764 [hep-ex].
- Afzal, A. et al. (2023). “The NANOGrav 15 yr Data Set: Search for Signals from New Physics”. *Astrophys. J. Lett.* 951.1, p. L11. arXiv: 2306.16219 [astro-ph.HE].
- Baxter, D. et al. (2021). “Recommended conventions for reporting results from direct dark matter searches”. *Eur. Phys. J. C* 81.10, p. 907. arXiv: 2105.00599 [hep-ex].
- Cowan, G. et al. (2011). “Asymptotic formulae for likelihood-based tests of new physics”. *Eur. Phys. J. C* 71. arXiv: 1007.1727 [physics.data-an].
- Dickey, J. M. (1971). “The weighted likelihood ratio, linear hypotheses on normal location parameters”. *The Annals of Mathematical Statistics*, pp. 204–223.

References II

- Fowlie, A. (Dec. 2013). "Bayesian Approach to Investigating Supersymmetric Models". PhD thesis. Sheffield U.
- Franck, C. T. and R. B. Gramacy (2020). "Assessing Bayes Factor Surfaces Using Interactive Visualization and Computer Surrogate Modeling". *Am. Stat.* 74.4, pp. 359–369. arXiv: 1809.05580 [stat.ME]. URL: https://doi.org/10.1080/00031305.2019.1671219.
- Fuks, B., M. D. Goodsell, and T. Murphy (Sept. 2024). "Monojets from compressed weak frustrated dark matter". arXiv: 2409.03014 [hep-ph].
- Goodsell, M. D. (June 2024). "HackAnalysis 2: A powerful and hackable recasting tool". arXiv: 2406.10042 [hep-ph].
- Jaynes, E. T. (1989a). "Confidence Intervals vs Bayesian Intervals (1976)". *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*. Ed. by R. D. Rosenkrantz. Dordrecht: Springer Netherlands, pp. 149–209. ISBN: 978-94-009-6581-2.

References III

- Jaynes, E. T. (1989b). "What is the question?" URL: https://bayes.wustl.edu/etj/articles/what.question.pdf.
- Johnson, V. E., S. Pramanik, and R. Shudde (Feb. 2023). "Bayes factor functions for reporting outcomes of hypothesis tests". *Proceedings of the National Academy of Sciences* 120.8. ISSN: 1091-6490. arXiv: 2210.00049 [math.ST].
- Morey, R. D. et al. (2016). "The fallacy of placing confidence in confidence intervals". *Psychonomic Bulletin & Review* 23, pp. 103–123.
- Neyman, J. (Aug. 1937). "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability". *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* 236.767, pp. 333–380. ISSN: 2054-0272.

References IV

- Pawel, S., A. Ly, and E.-J. Wagenmakers (June 2023). "Evidential Calibration of Confidence Intervals". *Am. Stat.*, pp. 1–11. ISSN: 1537-2731. arXiv: 2206.12290 [stat.ME].
- Wagenmakers, E.-J. et al. (Feb. 2020). "The Support Interval". *Erkenntnis* 87.2, pp. 589–601. ISSN: 1572-8420.