Pitfalls in likelihood land

Andrew Fowlie (speaker) & Anders Kvellestad

18 February 2021

Nanjing Normal University

Likelihood land

The likelihood (see e.g., Cousins 2020)

$$\mathcal{L}(\boldsymbol{\Theta}) = p\left(D \mid M, \boldsymbol{\Theta}\right)$$

tells us the probability (density) of the observed data, D, given a particular model, M, and choice of parameters.

This is a function of the model's parameters, $\boldsymbol{\Theta}$, for fixed, observed data.

Likelihood land

Key part of Bayesian analysis

$$\text{Likelihood} \times \text{Prior density} = \text{Evidence} \times \text{Posterior density}$$

in which we marginalise over unknown parameters by multi-dimensional *integration* over the likelihood function.

Frequentist analysis, on the other hand, usually involves finding the best-fit parameters by *finding the maximum* of the likelihood function.

General pitfalls

Pitfall — the likelihood is not enough?

- Frequentist analysis violates likelihood principle (Berger and Wolpert 1988) and in fact requires whole sampling distribution

$$p\left(D \,\middle|\, M, \boldsymbol{\Theta}\right)$$

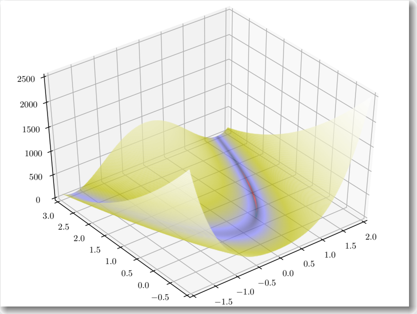
  as a distribution in the data, D.
- Often we make asymptotic approximations (Wilks 1938; Chernoff 1954) for the sampling distribution that only require the likelihood, e.g., $2 \times$ log likelihood ratio follows a χ^2 distribution. See e.g. Cowan et al. 2011.
- But what if those assumptions don't apply? See e.g., Algeri et al. 2019. Should we be talking about public sampling distributions?[1]

[1] See also Prosper's talk on Monday. Perhaps this is a matter of terminology and by public likelihoods we really mean public statistical models. Not clear to me at least.

Common pitfalls

Shape

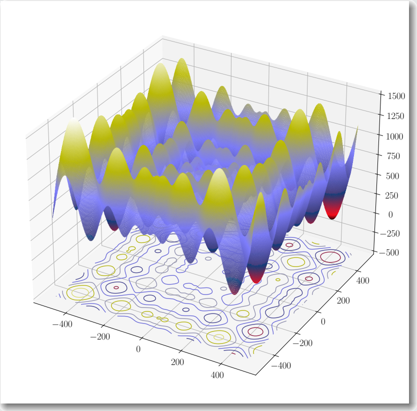
Likelihood could contain non-convex features, e.g., the classic Rosenbrock banana shape. Challenging for traditional optimisation and MCMC algorithms



Common pitfalls

Multi-modal

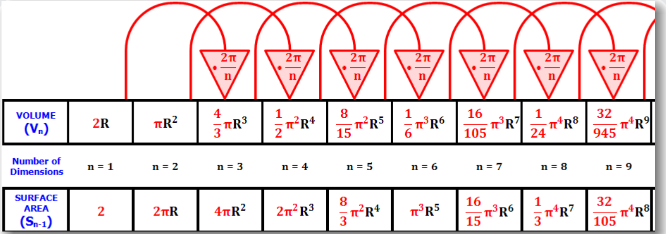
Likelihood could contain several distinct modes — how to find them all and the best one? (figure from Balázs et al. 2021)



Common pitfalls

Curse of dimensionality

Performance of numerical algorithms deteriorate exponentially with dimension

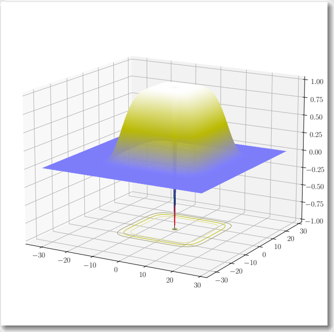


VOLUME (V_n)	$2R$	πR^2	$\frac{4}{3}\pi R^3$	$\frac{1}{2}\pi^2 R^4$	$\frac{8}{15}\pi^2 R^5$	$\frac{1}{6}\pi^3 R^6$	$\frac{16}{105}\pi^3 R^7$	$\frac{1}{24}\pi^4 R^8$	$\frac{32}{945}\pi^4 R^9$
Number of Dimensions	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n = 7$	$n = 8$	$n = 9$
SURFACE AREA (S_{n-1})	2	$2\pi R$	$4\pi R^2$	$2\pi^2 R^3$	$\frac{8}{3}\pi^2 R^4$	$\pi^3 R^5$	$\frac{16}{15}\pi^3 R^6$	$\frac{1}{3}\pi^4 R^7$	$\frac{32}{105}\pi^4 R^8$

Common pitfalls

Compression

If the best-fit is a flagpole in the Atlantic ocean, it could easily be missed (Balázs et al. 2021)

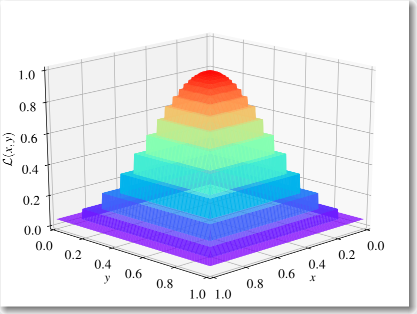


If sampling from the prior or if estimating the compression is important, convergence could be slow (Skilling 2006)

Common pitfalls

Plateaus

Likelihood functions involving plateaus may pose unexpected problems, see e.g., Schittenhelm and Wacker 2020; Fowlie, Handley, and Su 2020



Pitfall — incomplete information

What if we don't know the form of the likelihood? What if an experiment only reports an asymmetric error (see e.g., Barlow 2004 for further discussion)

$$x = 45^{+5}_{-8}$$

or a bound

$$x < 62 \text{ @ } 95\% \text{ CL}$$

or an interval etc. Actually, what if they only report a symmetric error

$$x = 45 \pm 10$$

It's a Gaussian, $\mathcal{N}(45, 10^2)$, e.g. Wald 1943, right? or is it?

The prior isn't the only source of subjectivity!

Experimentalists know more, and they should tell us — solved by public likelihoods and this workshop!

Pitfall — intractable likelihood

What if we just can't work out the likelihood — it's intractable?

Fine — if we can draw pseudo data from the sampling distribution we can use likelihood-free inference methods, e.g., ABC (Diggle and Gratton 1984).

Public sampling distributions?

Pitfall — noisy estimate of likelihood

Consider case of LHC likelihoods. The likelihood might be a Poisson involving the expected number of events.

The expected number of events, though, can often only be computed through forward simulations of the whole experiment, involving MC simulations in e.g., `Pythia` and detector simulations

So we end up with a noisy estimate of the likelihood, $\hat{\mathcal{L}}$.

Pitfall — noisy estimate of likelihood

Careful! — if we run an optimiser, it will favour points with upward fluctuations in estimate of likelihood. Fitting noise.

If we run MCMC, we might get correct inferences if we have an unbiased estimator of the likelihood, $\langle\hat{\mathcal{L}}\rangle = \mathcal{L}$ — this is pseudo-marginal MCMC (Andrieu and Roberts 2009).

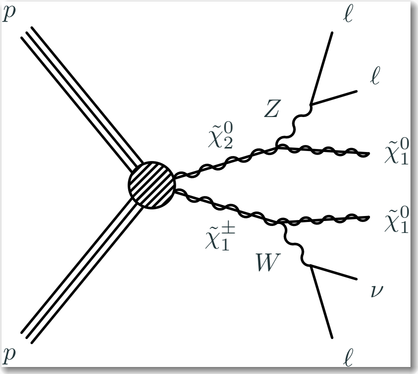
However, common MLE estimator of $\mathcal{L}$ is biased.

Interplay between noise and incomplete information

GAMBIT study of electroweakinos

GAMBIT studied constraints on neutralino and chargino supersymmetric particles (Athron et al. 2019)

Compiled relevant constraints from LHC and LEP and constructed joint likelihood function.



Missing correlation information

CMS and ATLAS perform many searches. Some searches use several distinct search regions for which the SM backgrounds (b) are only known with appreciable and correlated uncertainties,

$$p(b_1, b_2, \ldots) \neq \prod_{i=1}^{n} p(b_i)$$

In some cases, only information about the marginals, $p(b_i)$ were publicly available at the time, i.e., information of the form

$$b_1 = 46 \pm 10$$

without any correlation information, e.g., a covariance matrix.

A way forward?

Decided that for each parameter point, we would approximate the likelihood by a single search region, since we know the individual likelihoods. But which one?

- We didn't want to tamper with the asymptotic sampling distributions of any test-statistics (needn't be a consideration in Bayesian approach)
- So couldn't make the choice based on the observed numbers of events
- Used region with greatest expected sensitivity i.e., the region most likely to result in rejection of the signal model when the background model was true.

Pitfall — noisy sensitivity estimates

Since we use MC simulations, only have noisy estimates of likelihood and sensitivity.

- At a boundary, two or more searches could have very similar or equal sensitivities — with only noisy estimates available, which one gets selected is a coin flip!
- Even though the sensitivities are similar, the likelihood could be very different!
- The likelihood flips between the very different likelihoods of the two regions with very similar sensitivity — which one you get depends on noisy estimate of sensitivity.

Toy example

Consider two signal regions (SR), each targeting a specific signal

SR1

$n_1 = 65$ observed events, $b_1 = 50$ expected background events
$\rightarrow$ excess of 15 events

SR2

$n_2 = 35$, $b_2 = b_1 = 50 \rightarrow$ deficit of 15 events

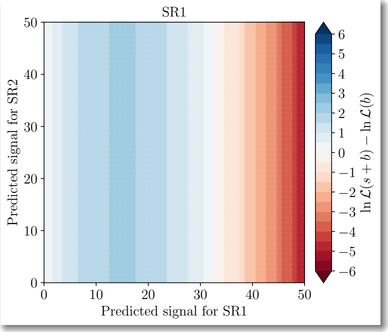
Uncertainties on backgrounds not important for this example

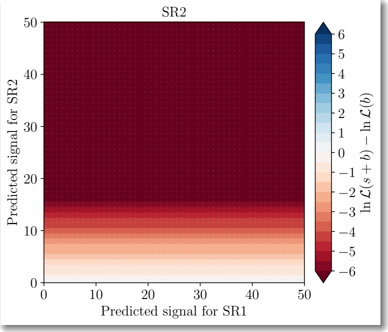
Now perform a parameter scan of some large theory that can produce both signals

We'll consider this in the plane of the two signal predictions directly

Toy example

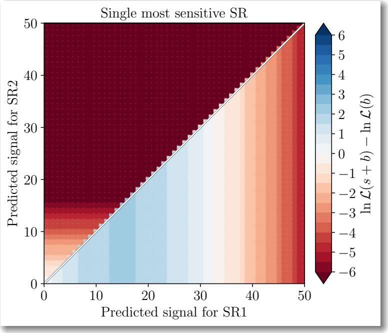
True log-likelihood functions for SR1 and SR2

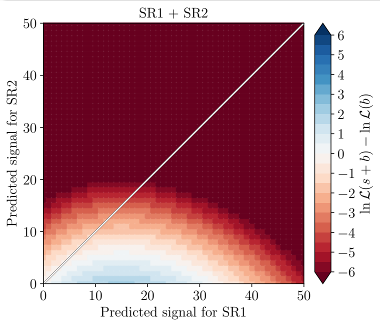




Toy example

True log-likelihood functions for most sensitive region and sum of regions



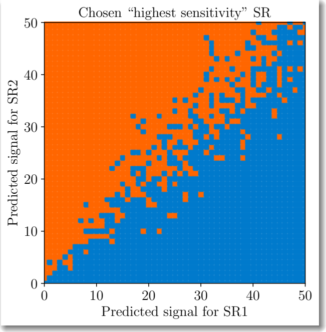


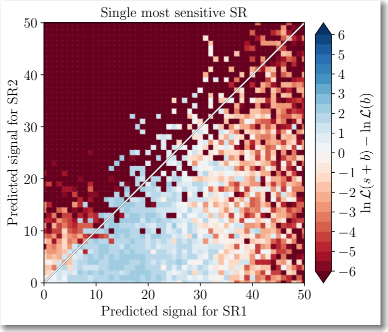
Sensitives of two SRs equal along diagonal

When using most sensitive SR, chosen SR will be a coin toss for points near the diagonal.

Toy example

With MC uncertainty on the estimates of the sensitivities, we often select *wrong* SR and obtain greater likelihood, especially above the diagonal



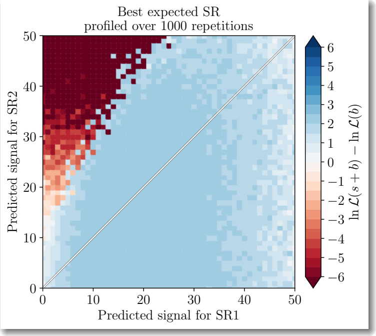


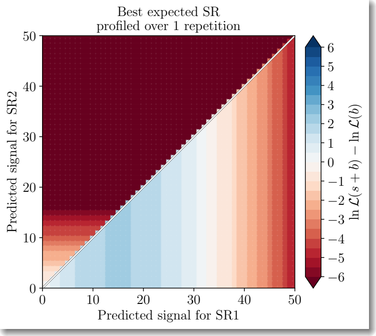
Toy example

- Suppose we were performing a high-dimensional parameter scan.
- In each cell on these plots, we could have 1000 or so samples.
- If we then computed the profile likelihood, we would select sample with the greatest likelihood of the 1000
- This effect would favour wrongly selected SRs with greater likelihoods

Toy example

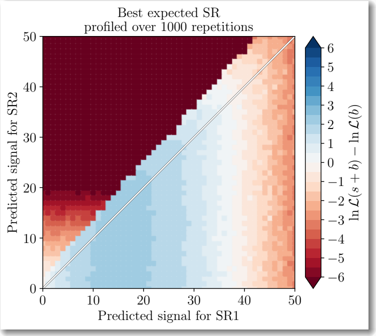
To illustrate this, now perform 1000 repetitions per cell and show the one with the greatest likelihood from the most sensitive SR

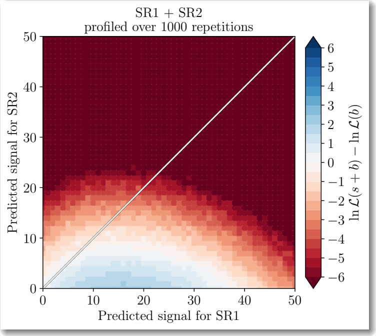




Toy example

Increase number of simulations to reduce uncertainties on sensitivities by factor 4 — 16 times more MC





Still large likelihoods above diagonal on LHS. Harder to fix problems caused by fluctuations on most sensitive SR (LHS) than sum of log-likelihoods (RHS).

Solution

Temporary solution

For important regions of parameter space, ramp up number of events in MC simulations.

Reduce uncertainty on estimated sensitivities.

This is CPU intensive — 64 million events per parameter point.

The ultimate solution is public likelihoods — the problem becomes smaller for every ATLAS and CMS analysis that is released with simplified/full likelihood information.

Summary

- Likelihood key component of statistical analysis (though not always enough)
- Likelihoods may be pathological — multimodal, flagpoles in the middle of an ocean, etc
- Moreover, our estimates of likelihoods may be pathological because of
    1. Incomplete information (e.g., missing correlations)
    2. Noisy estimator of likelihood (e.g., MC simulations)
- In Athron et al. 2019, these two problems conspired in a particularly problematic way in likelihoods for searches for supersymmetric particles

Backup

Bayesian land

Here, we don't care about spoiling or knowing sampling behaviour. We can just build the best approximation to the likelihood. This problem needn't occur.

Possible approaches:

- We know the marginals. We want write the joint distribution. A general approach is this problem is that of copulas (Nelsen 2006). Schematically,

$$\text{every joint distribution} = \text{marginal distributions} + \text{copula}$$

- Make approximation of likelihood using observed counts, e.g., crude bound

$$\mathcal{L} \leq \min_i P\left(\text{search}_i \,\middle|\, M, \boldsymbol{\Theta}\right)$$

Noisy estimates

As a proxy for sensitivity of each search region, we used

$$\frac{P\left(\text{observed} = \text{background} \,\middle|\, \text{Signal + background model}\right)}{P\left(\text{observed} = \text{background} \,\middle|\, \text{Background only model}\right)}$$

In fact, though, we are relying on MC simulations to estimate the numbers of expected events — only have noisy estimates of likelihood and sensitivity.

References

Algeri, Sara et al. (Nov. 2019). "Searching for new physics with profile likelihoods: Wilks and beyond." In: arXiv: 1911.10237 [physics.data-an].

Andrieu, Christophe and Gareth O. Roberts (Apr. 2009). "The pseudo-marginal approach for efficient Monte Carlo computations." In: *Ann. Statist.* 37.2, pp. 697–725. URL: https://doi.org/10.1214/07-AOS574.

Athron, Peter et al. (2019). "Combined collider constraints on neutralinos and charginos." In: *Eur. Phys. J. C* 79.5, p. 395. arXiv: 1809.02097 [hep-ph].

Balázs, Csaba et al. (Jan. 2021). "A comparison of optimisation algorithms for high-dimensional particle and astrophysics applications." In: arXiv: 2101.04525 [hep-ph].

Barlow, Roger (June 2004). "Asymmetric statistical errors." In: *Statistical Problems in Particle Physics, Astrophysics and Cosmology*, pp. 56–59. arXiv: physics/0406120.

Berger, J.O. and R.L. Wolpert (1988). *The Likelihood Principle.* Institute of Mathematical Statistics. Lecture notes : monographs series. Institute of Mathematical Statistics. ISBN: 9780940600133.

Chernoff, Herman (Sept. 1954). "On the Distribution of the Likelihood Ratio." In: *Ann. Math. Statist.* 25.3, pp. 573–578. URL: https://doi.org/10.1214/aoms/1177728725.

Cousins, Robert D. (Oct. 2020). "What is the likelihood function, and how is it used in particle physics?" In: arXiv: 2010.00356 [physics.data-an].

Cowan, Glen et al. (2011). "Asymptotic formulae for likelihood-based tests of new physics." In: *Eur. Phys. J. C* 71. [Erratum: Eur.Phys.J.C 73, 2501 (2013)], p. 1554. arXiv: 1007.1727 [physics.data-an].

Diggle, Peter J. and Richard J. Gratton (1984). "Monte Carlo Methods of Inference for Implicit Statistical Models." In: *Journal of the Royal Statistical Society. Series B (Methodological)* 46.2, pp. 193–227. ISSN: 00359246. URL: http://www.jstor.org/stable/2345504.

Fowlie, Andrew, Will Handley, and Liangliang Su (Oct. 2020). "Nested sampling with plateaus." In: arXiv: 2010.13884 [stat.CO].

Nelsen, R.B. (2006). *An Introduction to Copulas*. Springer Series in Statistics. Springer. ISBN: 9780387286594. URL: https://books.google.com/books?id=B3ONT5rBv0wC.

Schittenhelm, Doris and Philipp Wacker (May 2020). "Nested Sampling And Likelihood Plateaus." In: *arXiv e-prints*, arXiv:2005.08602, arXiv:2005.08602. arXiv: 2005.08602 [math.ST].

Skilling, John (Dec. 2006). "Nested sampling for general Bayesian computation." In: *Bayesian Anal.* 1.4, pp. 833–859. URL: https://doi.org/10.1214/06-BA127.

References v

Wald, Abraham (1943). "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large." In: *Transactions of the American Mathematical Society* 54.3, pp. 426–482. ISSN: 00029947. URL: http://www.jstor.org/stable/1990256.

Wilks, S. S. (1938). "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses." In: *Annals Math. Statist.* 9.1, pp. 60–62.