Profile likelihood

Patrick Breheny

November 11, 2024

Introduction

- We have now derived all the core results of likelihood theory, developed inferential tools using them, and implemented them computationally
- In this third and final part of the course, we are going to explore challenges to likelihood
- Specifically, we will focus on cases where "ordinary" likelihood may be problematic, but modifications or extensions of the likelihood idea prove beneficial

Nuisance parameters

- The primary challenge that likelihood-based inference faces is the problem of nuisance parameters
- Nuisance parameters are unavoidable in practice – with the exception of highly controlled experimental settings, there are almost always additional factors and dependencies that we want to adjust for
- Even if we are genuinely interested in every single parameter, we must almost always think about them in isolation at some point, and when doing so, the other parameters become nuisance parameters
- Furthermore, this problem doesn't go away – the more data we have, the more complex of a model we may try to fit

Bayesian approach

- It is worth noting that this is not really an issue in the Bayesian paradigm
- In Bayesian statistics, the way to handle nuisance parameters is obvious, universal, and works very well (computational challenges aside) – we simply integrate them out:

$$p(\theta_j|\mathbf{x}) = \int p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta}_{-j}$$

- Unfortunately, this approach is not possible with likelihood; likelihoods are not probability distributions, so integrating them is not meaningful

Whither likelihood?

- Indeed, there is no single technique that is ideal in all situations
- For this reason, we will spend time covering a variety of methods for modifying the likelihood
- We won't necessarily be able to cover each one in full detail, but hopefully we can learn the main ideas of how each modification works

Profile likelihood: Definition

- The first modification we will discuss is called the "profile" likelihood
- Given the joint likelihood $L(\boldsymbol{\theta}, \boldsymbol{\eta})$, the *profile likelihood* of $\boldsymbol{\theta}$ is

$$L_p(\boldsymbol{\theta}) = \max_{\boldsymbol{\eta}} L(\boldsymbol{\theta}, \boldsymbol{\eta})$$

- Note that this is equivalent to

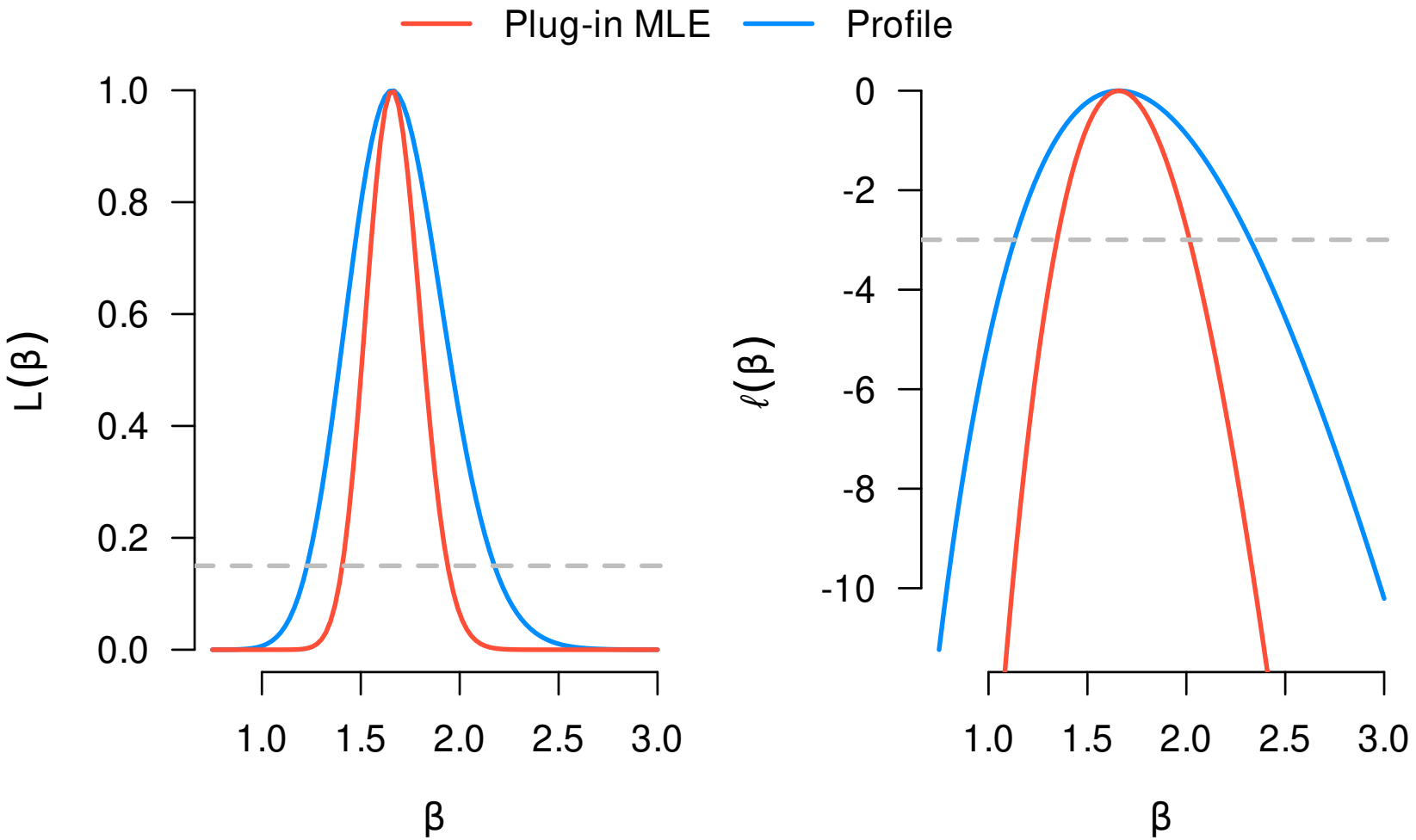
$$L_p(\boldsymbol{\theta}) = L(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}(\boldsymbol{\theta}));$$

in other words, we have encountered this basic idea already, when deriving score and likelihood ratio confidence intervals

Is the profile likelihood a likelihood?

- The difference now is that we're treating the quantity $L_p(\boldsymbol{\theta})$ as a likelihood itself
- But is it?
- No; there is no probability model $p(x|\boldsymbol{\theta})$ such that $L_p(\boldsymbol{\theta}|x) = p(x|\boldsymbol{\theta})$
- Nevertheless, in many ways the profile likelihood behaves as a likelihood does, but one that reflects any uncertainty concerning the nuisance parameters

Profile likelihood: Illustration (Gamma distribution)



Profile likelihood theory

- So, in what ways does the profile likelihood have the same properties as a regular likelihood?
- First, and most obviously, the MLE of $L_p(\boldsymbol{\theta})$ is equal to $\hat{\boldsymbol{\theta}}$, the first component of the MLE for $L(\boldsymbol{\theta}, \boldsymbol{\eta})$
- Also, as we have already proved, the profile likelihood ratio test is equivalent to the full likelihood ratio test in the presence of nuisance parameters
- What about the Score and Wald tests?

Total derivatives

- First, a brief review of the concept of a *total derivative*
- For a differentiable function parameterized as $f(\mathbf{x}, \mathbf{y}(\mathbf{x}))$, we have

$$\nabla f(\mathbf{x}) = \frac{\partial f}{\partial \mathbf{x}} + \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \frac{\partial f}{\partial \mathbf{y}}$$

- For, example, suppose $y = x^2$ and $f(x, y) = xy$:
  - Since $f(x) = x^3$, we have $f'(x) = 3x^2$
  - Alternatively, $f'(x) = y + (2x)(x) = 3x^2$

Score and information

- With this is mind, we can express the score and information of the profile likelihood in terms of the score and information of the ordinary joint likelihood
- **Theorem:** Suppose the log-likelihood function $\ell(\boldsymbol{\theta}, \boldsymbol{\eta})$ is twice differentiable, with $\mathbf{u}_1$ referring to the portion of the score corresponding to $\boldsymbol{\theta}$, and so on. Then

$$\begin{aligned} \nabla \ell_p(\boldsymbol{\theta}) &= \mathbf{u}_1(\boldsymbol{\theta}) \\ -\nabla^2 \ell_p(\boldsymbol{\theta}) &= \boldsymbol{\mathcal{I}}_{11}(\boldsymbol{\theta}) - \boldsymbol{\mathcal{I}}_{12}(\boldsymbol{\theta}) \boldsymbol{\mathcal{I}}_{22}^{-1}(\boldsymbol{\theta}) \boldsymbol{\mathcal{I}}_{21}(\boldsymbol{\theta}) \end{aligned}$$

- Note here that $\mathbf{u}_1$ and $\boldsymbol{\mathcal{I}}$ are evaluated at $(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}(\boldsymbol{\theta}))$; the usual score and information are functions of all parameters, but are written here as functions of $\boldsymbol{\theta}$ alone since $\boldsymbol{\eta}$ is determined by $\boldsymbol{\theta}$

Score test and Wald test

- In other words, the profile likelihood information correctly reflects the loss of information caused by uncertainty regarding the nuisance parameters
- Thus, the Wald test based on the profile likelihood is equivalent to the earlier test we derived for the Wald test in the presence of nuisance parameters
- Furthermore, because the profile likelihood score is $\mathbf{u}_1$ and the result on the previous slide holds everywhere (not just at $\hat{\boldsymbol{\theta}}$), the profile likelihood score test is also the same as the score test in the presence of nuisance parameters that we derived earlier

Example: Normal variance

- At this point, you may be inclined to think that the profile likelihood has all the properties of a regular likelihood
- This is not true, however
- For example, consider the profile likelihood of σ^2 for the normal distribution $X_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$:

$$\ell_p(\sigma^2) \propto -\tfrac{n}{2} \log \sigma^2 - \tfrac{1}{2\sigma^2} \sum_i (x_i - \bar{x})^2;$$

the corresponding score equation is

$$u_p(\sigma^2) = -\frac{n}{2\sigma^2} + \frac{\sum_i (x_i - \bar{x})^2}{2\sigma^4}$$

Example: Normal variance (cont'd)

- Since the expected value of $\sum_i (X_i - \bar{X})^2$ is $(n-1)\sigma^2$, the expected value of the score equation evaluated at the true value of σ^2 is $-1/(2\sigma^2)$
- In other words, this does not satisfy the usual property of zero expectation for a score statistic
- Note that the full likelihood has a μ instead of $\bar{x}$; in a sense, the profile likelihood has replaced μ with its estimate without really accounting for the bias this introduces

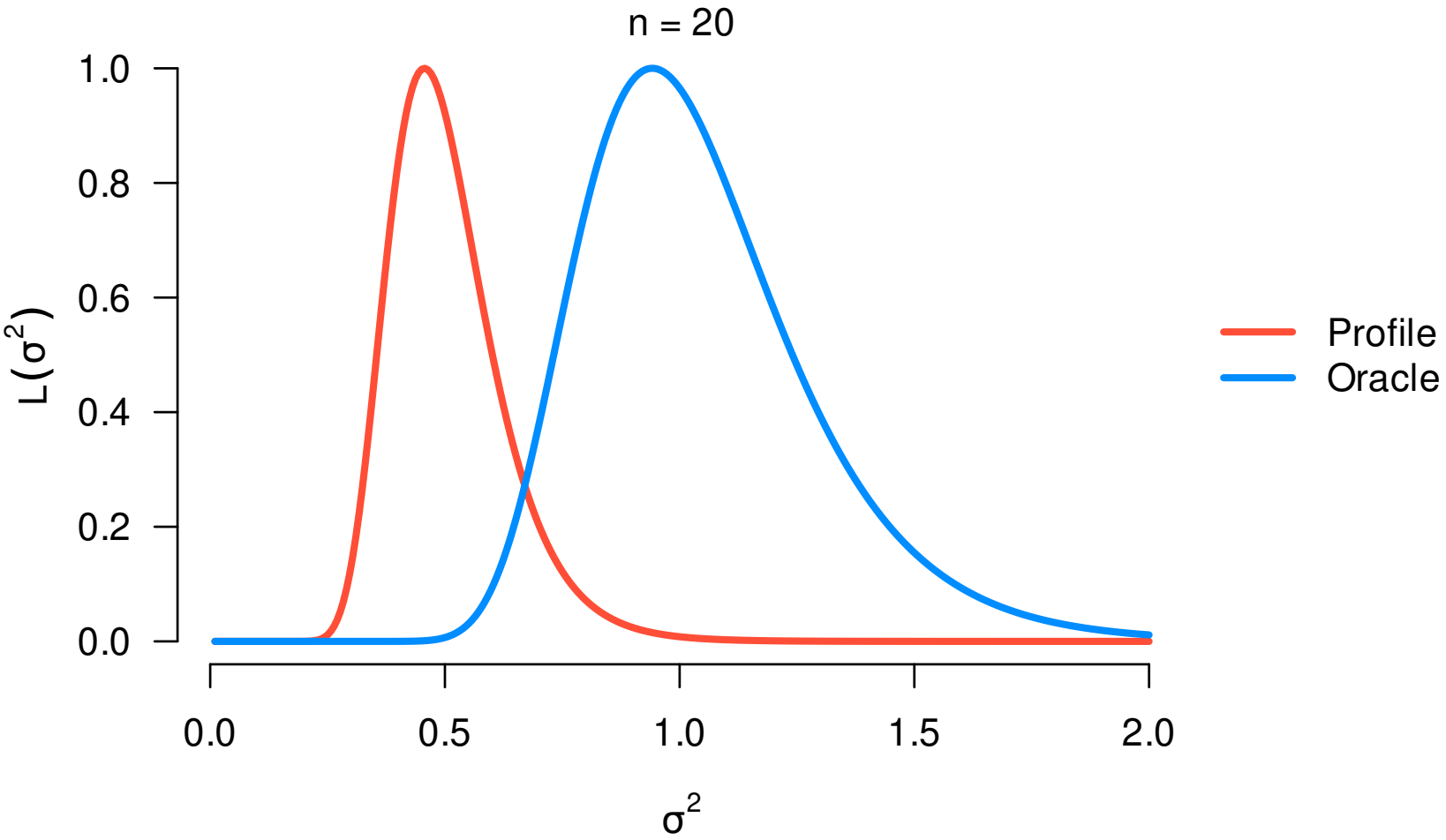
Profile likelihood and bias

- As a consequence, the MLE for σ^2 is also biased (downwards)
- As $n \to \infty$, this is irrelevant since the MLE is consistent and the difference between $\bar{x}$ and μ^* goes to zero
- For small samples – or more accurately, when the number of nuisance parameters is large with respect to the sample size – this bias can be significant
- For example, in linear regression with normal errors, the (profile) MLE of σ^2 is RSS/n while the unbiased estimator is $\mathrm{RSS}/(n-p)$; how large a problem this is depends on the ratio of n and p

Neyman-Scott problem

- In the extreme case, as we have already seen, this can cause the MLE to be inconsistent
- For example, recall the Neyman-Scott problem from an earlier assignment, where y_{i1} and y_{i2} are iid samples from a $\mathrm{N}(\mu_i, \sigma^2)$ distribution
- In this case, the number of nuisance parameters was going to infinity, creating a problem that could never be overcome
- To get a sense of how this relates to the profile likelihood, let's plot the profile and "oracle" likelihoods (using the unknown true values of μ_i in the likelihood)

Neyman-Scott illustration



Final remarks

- None of this means that profile likelihood is bad or invalid – without question, it remains the most widely used and readily applicable method for dealing with nuisance parameters
- Nevertheless, hopefully this example provides a motivation for developing other extensions of the likelihood that perhaps improve upon profile likelihood, at least in certain settings