LECTORIAL ON MAXIMUM LIKELIHOOD ESTIMATION

Oliver Maclaren
Department of Engineering Science
The University of Auckland, New Zealand
oliver.maclaren@auckland.ac.nz

WORKSHEET!

Worksheet plus recommended reading probably the most useful part of this lectorial!

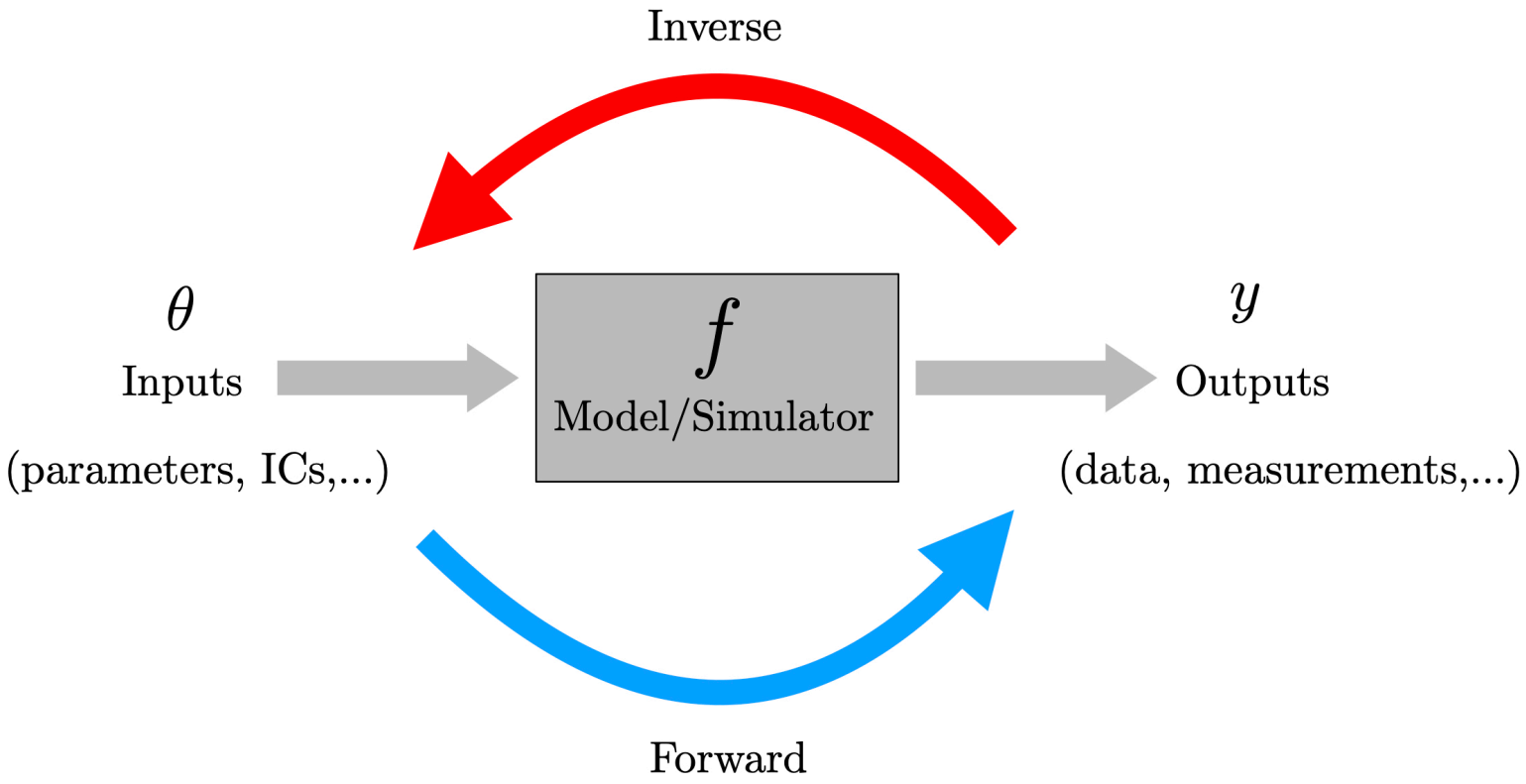
MODELS AND DATA

There are many types of *model*, and many types of *data*.

In general, a model is a *theoretical picture of the world* that implies some potentially *observable data*.

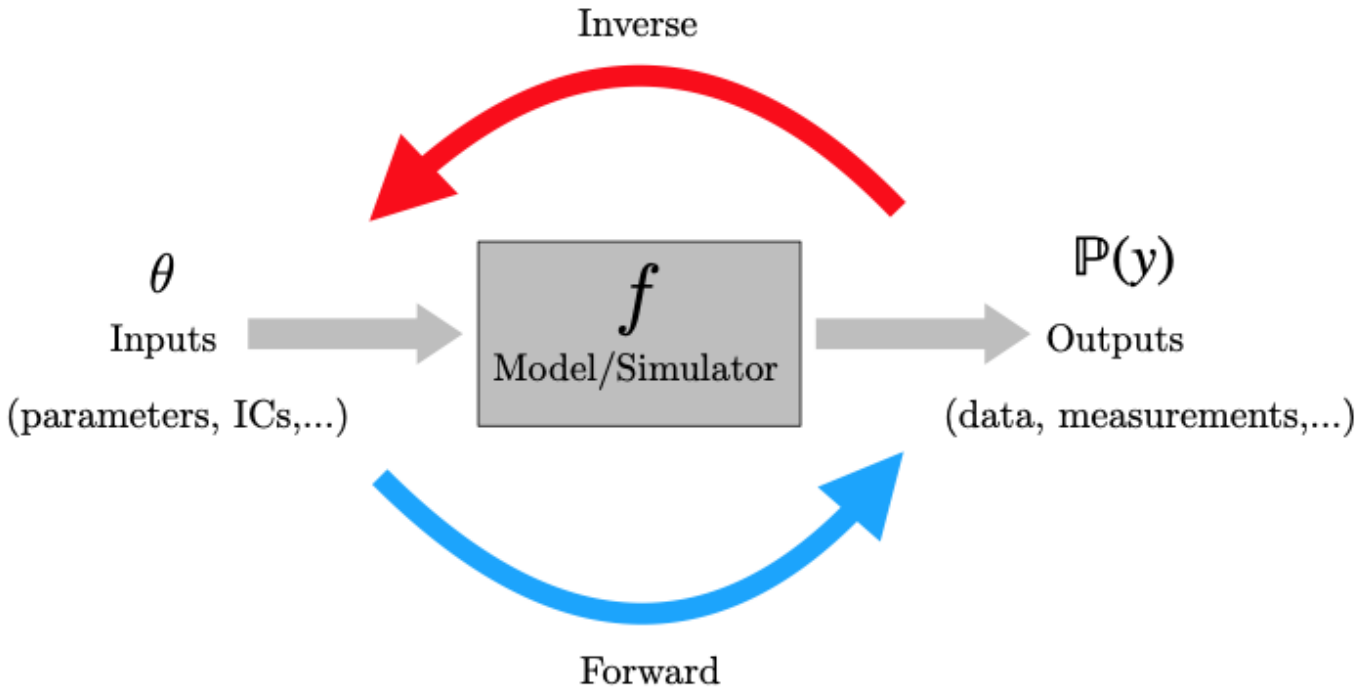
There are two sorts of problems involved in relating models and data: the *forward* problem and the *inverse* problem.

MODELS AND DATA



STATISTICS

Statistics is a general framework for thinking about this connection. We need to slightly modify the previous picture:



STATISTICS

In particular, here a model is a theoretical picture of the world that *implies probability distributions* over observable data.

$$\mathbb{P}(y;\theta) : \theta \mapsto \mathbb{P}(y)$$

Here θ represents everything about the theory required to imply (probabilistic) predictions.

NOTE ON VECTORS, FUNCTIONS ETC

Note: in this approach both the parameter and random variable can be *vector-valued* or even *infinite-dimensional*.

I will still denote these with ordinary non-bold, often lowercase symbols (though occasionally using a capital letter when emphasising something is a random variable).

NOTE ON PROBABILITY DENSITIES

Here we will often work with probability *density* functions. Informally, we take

$$\mathbb{P}(Y = y) = p(Y = y)dy$$

where $p(Y = y) = p(y)$ is the probability density of Y at y.

NOTE ON INTERPRETATION OF PROBABILITY

'The Bayesian' approach directly models both *epistemic* (personal knowledge) and *aleatoric* (natural variability or 'out there') uncertainty with probability.

'The frequentist' view tends to restrict probability models to just capturing *aleatoric* 'variability in nature' rather than personal uncertainty about parameters.

There are many subtleties and complications to this 'rule' though.

SIMPLE EXAMPLES

Estimating a constant

$$y = \theta + e$$

where y is (random) observable data, θ is a fixed, unknown scalar parameter to estimate, e is a random error variable with (here) known distribution, e.g.

$$e \sim \mathcal{N}(0, \sigma^2)$$

for known σ.

SIMPLE EXAMPLES

Fitting a linear (in parameters) regression

$$y = ax + bx^2 + e$$

where y and e as before, here $\theta = (a, b)$ is a fixed, unknown vector parameter to estimate. The x are usually assumed known and held fixed in regression.

SIMPLE EXAMPLES

Fitting an ODE

$$y = Az + e$$

$$\frac{dz}{dt} = f(z; \theta)$$

where y and e as before, A is an observation operator extracting the solution z at measurement locations, θ is a fixed, unknown vector parameter for the ODE system.

CONSTANT EXAMPLE: ERROR FORM

$$p_Y(y;\theta) = p_E(y-\theta).$$

Here we know everything about this distribution

except the (fixed but unknown) parameter θ.

Note can obtain above via a heuristic change of variables derivation:

$$p_Y(y;\theta) = \int p_{Y|E}(y|e;\theta)p_E(e;\theta)de$$

$$= \int \delta(y-\theta-e)p_E(e)de$$

$$= p_E(y-\theta)$$

MULTIPLE MEASUREMENTS: IID DATA

For *independent, identically distributed* data we can write

$$p(y;\theta) = \prod_{i=1}^{n} p(y_i;\theta)$$

where $p(y;\theta)$ is the probability density for the vector $y = (y_1, ..., y_n)$ and $p(y_i;\theta)$ for $i = 1, ..., n$ are the single observation densities.

CONSTANT EXAMPLE CONT'D

Given n iid observations $y = (y_1, \ldots, y_n)$ then:

$$p(y_1, \ldots, y_n; \theta) =$$

$$\prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(\frac{-(y_i - \theta)^2}{2\sigma^2}\right)$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(\sum_{i=1}^{n} \frac{-(y_i - \theta)^2}{2\sigma^2}\right)$$

LIKELIHOOD FUNCTION

Given observed data, the likelihood function is defined as proportional *probability of the observed data as a function of the parameter* :

$$\mathcal{L}(\theta; y) \propto p(y; \theta)dy$$

As indicated, the likelihood is often taken as defined only up to a constant. Hence the dy and other terms can be dropped.

WHO CARES?

Intuitively, the likelihood function tells you *how well each parameter fits the data*.

As we will see, *maximising the likelihood function* can be thought of as a natural generalisation of ideas like *minimising the sum of squared errors*.

The full likelihood function also has many interesting theoretical statistical properties. Think of it as representing *the information the data provides about the parameter*.

LIKELIHOOD OR MAXIMUM LIKELIHOOD?

…it is difficult to convey to a statistical audience the vital distinction between likelihood regarded as a basis for a theory of inference, and likelihood regarded as a commodity to be maximised in a method of point estimation….

LIKELIHOOD OR MAXIMUM LIKELIHOOD?

…At one recent interational conference at which I laboured for three-quarters of an hour to make clear the advantages of likelihood inference, the chairman thanked me for my lecture on the Method of Maximum Likelihood.

A. W. F. EDWARDS

Likelihood

EXPANDED EDITION

LEADS US BACK TO: MANY APPROACHES TO STATISTICS

Frequentists, Bayesians, Likelihoodists, ...

A NEO-FISHERIAN APPROACH

- Use *whole likelihood function* to form point estimates, interval estimates, represent info in data etc
- *Calibrate* these estimation procedures in a frequentist manner
  - i.e. check how the perform under *(hypothetical) repeated sampling*

A NEO-FISHERIAN APPROACH

- *'Automatic' recipe* for deriving estimators that *often* have good *frequentist* properties
- Parameterisation *invariant*
- Fairly intuitive 'case at hand' or *'evidential'* interpretation as well (pursued by 'pure likelihood' school)
- Reasonably compatible with a *Bayesian* approach if you decide to go that way

RELATIVE LIKELIHOOD FUNCTION

In likelihood-based frequentist and pure likelihood inference, it is convenient to fix the arbitrary proportionality factor by working with the *relative likelihood*. We will do this from now, taking

$$\mathcal{L}(\theta; y) = \frac{p(y;\theta)}{\sup_\theta p(y;\theta)} = \frac{p(y;\theta)}{p(y;\hat{\theta})}$$

This is also our first proper encounter with the *maximum likelihood estimate* $\hat{\theta}$!

THE LOG (RELATIVE) LIKELIHOOD FUNCTION

It's also convenient to define the log (relative) likelihood function:

$$l(\theta; y) = \log p(y; \theta) - \log p(y; \hat{\theta})$$

CONSTANT EXAMPLE CONT'D

For fixed, known σ we get

$$\mathcal{L}(\theta; y) = \frac{\exp\left(-\frac{1}{2\sigma^2}\sum_i (y_i - \theta)^2\right)}{\sup_\theta \exp\left(-\frac{1}{2\sigma^2}\sum_i (y_i - \theta)^2\right)},$$

where the common factor $\left(2\pi\sigma^2\right)^{-\frac{n}{2}}$ cancels.

CONSTANT EXAMPLE CONT'D

The log relative likelihood is then

$$l(\theta; y) =$$

$$-\frac{1}{2\sigma^2}\sum_i (y_i - \theta)^2 - \left(-\frac{1}{2\sigma^2}\sum_i (y_i - \hat{\theta})^2\right),$$

using that log and max commute, i.e. $\log \sup f = \sup \log f$.

CONSTANT EXAMPLE CONT'D

This is

$$l(\theta; y) = \frac{1}{2\sigma^2} \sum_i (y_i - \hat{\theta})^2 - \frac{1}{2\sigma^2} \sum_i (y_i - \theta)^2$$

IMPLICATION

Here we really just need to work with the unnormalised form:

$$\frac{1}{2\sigma^2}\sum_i (y_i - \theta)^2$$

and compute the ‘normalising’ factor for the relative likelihood as the minimum (maximum of negative) of this. We can take exponentials etc after if we want.

Also, as can be seen…

IMPLICATION

In the case of *additive Gaussian errors with known variance, the log relative likelihood is basically just a plot of the sum of squared errors relative to the least squares solution*

Furthermore, the *maximum likelihood estimate is then the least squares estimate*.

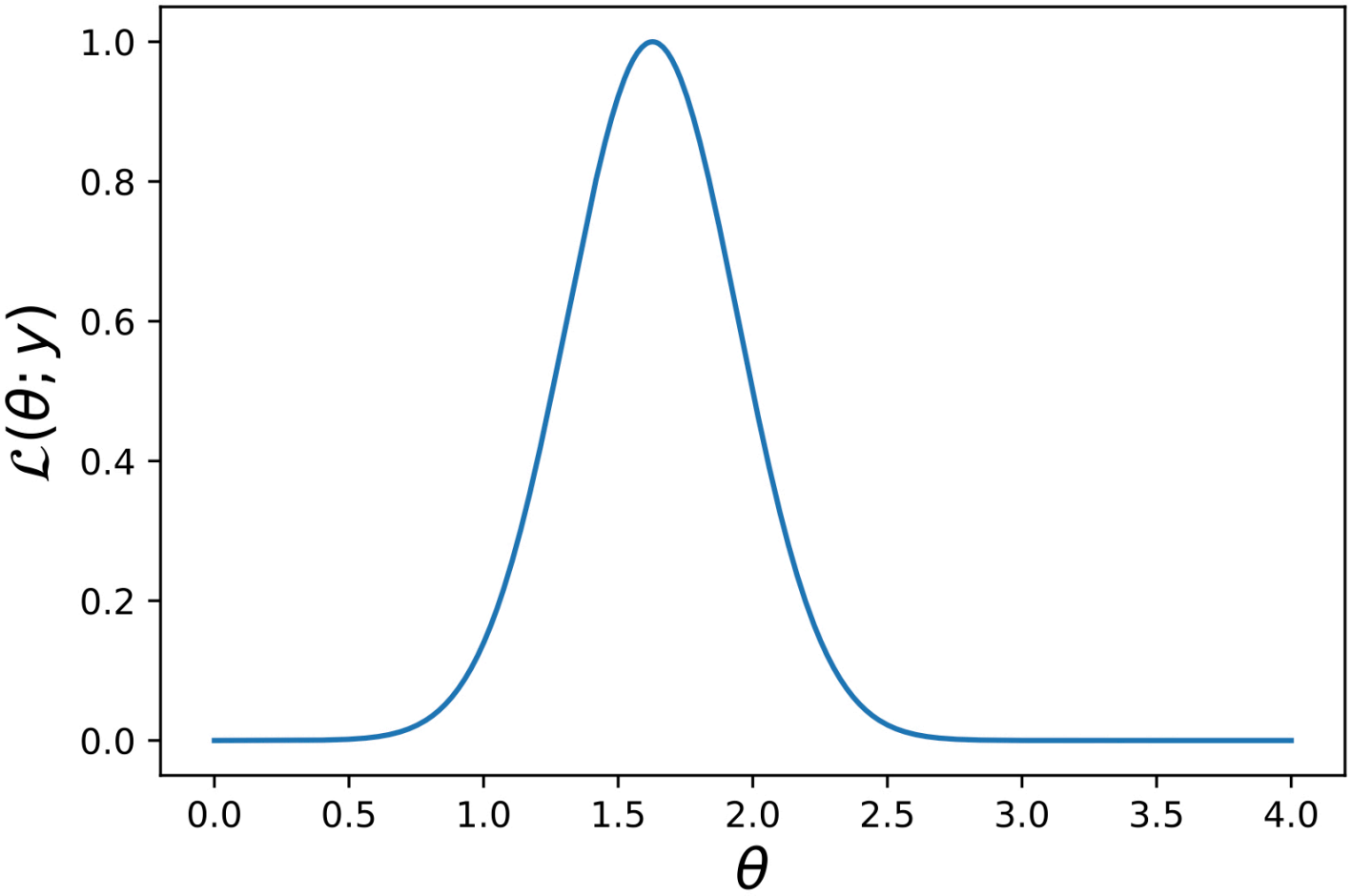
COMPUTING

Here we simply need to be able to *directly evaluate the unnormalised likelihood* over a grid of scalar parameter values.

The *maximum* can then be found analytically (here), numerically from this grid or (in general) using a *numerical optimisation tool* (either generic or optimised for least-squares problems).

I use e.g. *scipy.optimize* in Python (similar tools in Matlab, R, Julia etc)

PLOTTING SIMPLE EXAMPLE



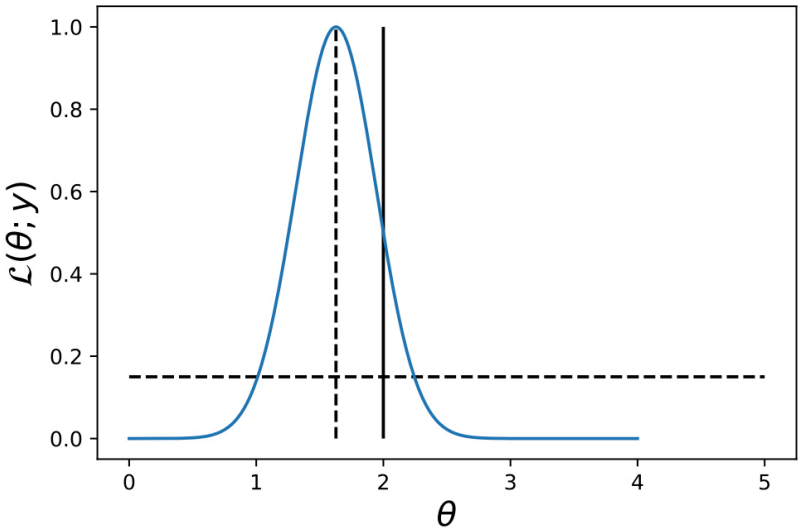
NOW WHAT?

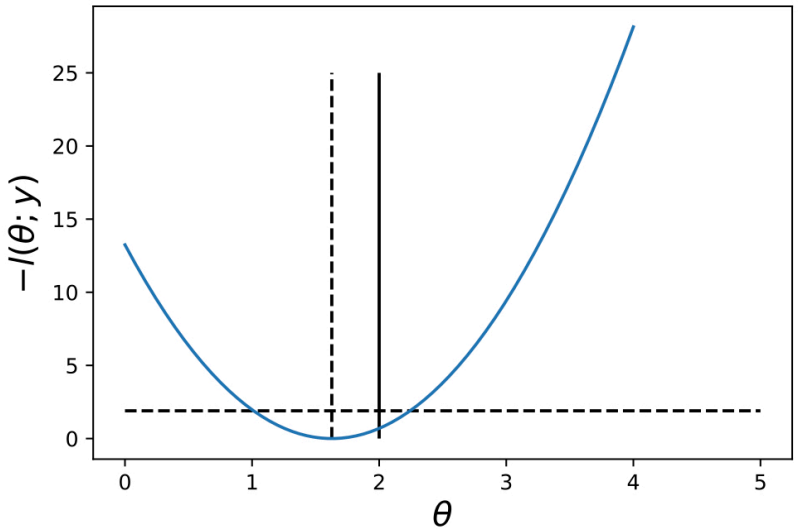
We want to use the likelihood function to *estimate* our unknown parameters.

One way to get a *point* (single best) estimate is to maximise the likelihood.

We can also get an *interval estimate* by including all parameters with sufficiently high likelihood.

POINT AND INTERVAL ESTIMATION WITH THE LIKELIHOOD FUNCTION





Vertical dashed = max likelihood

Horizontal dashed = confidence interval thresholds

INTERVAL ESTIMATION?

Likelihood-based confidence intervals can be formed by taking all parameters within a set *cutoff* of the maximum likelihood value:

$$\{\theta \ : \ \frac{\mathcal{L}(\theta)}{\mathcal{L}(\hat{\theta})} \geq c\}$$

Magical frequentist (repeated sampling) derivations show that, typically, we can choose this based on the *chi-squared* distribution…

INTERVAL ESTIMATION WITH THE LIKELIHOOD FUNCTION

For a *scalar* parameter of interest choosing

$$\{\theta \ : \ \frac{\mathcal{L}(\theta)}{\mathcal{L}(\hat{\theta})} \geq 0.15\}$$

gives, approximately, a *95% confidence interval*.

Vector parameters have a different calibration. E.g. for a two-dimesional parameter we can choose $c = 0.05$ (see Pawitan for details).

INTEREST PARAMETERS AND NUISANCE PARAMETERS

When dealing with vector parameters we often partition it into an *interest* parameter ψ and *nuisance* parameter λ.

$$\theta = (\psi, \lambda)$$

We then profile (maximise) out our nuisance parameter.

PROFILE LIKELIHOOD FOR INTEREST PARAMETER

For *each ψ value*, define

$$\mathcal{L}_p(\psi; y) = \sup_{\lambda} \mathcal{L}(\psi, \lambda; y)$$

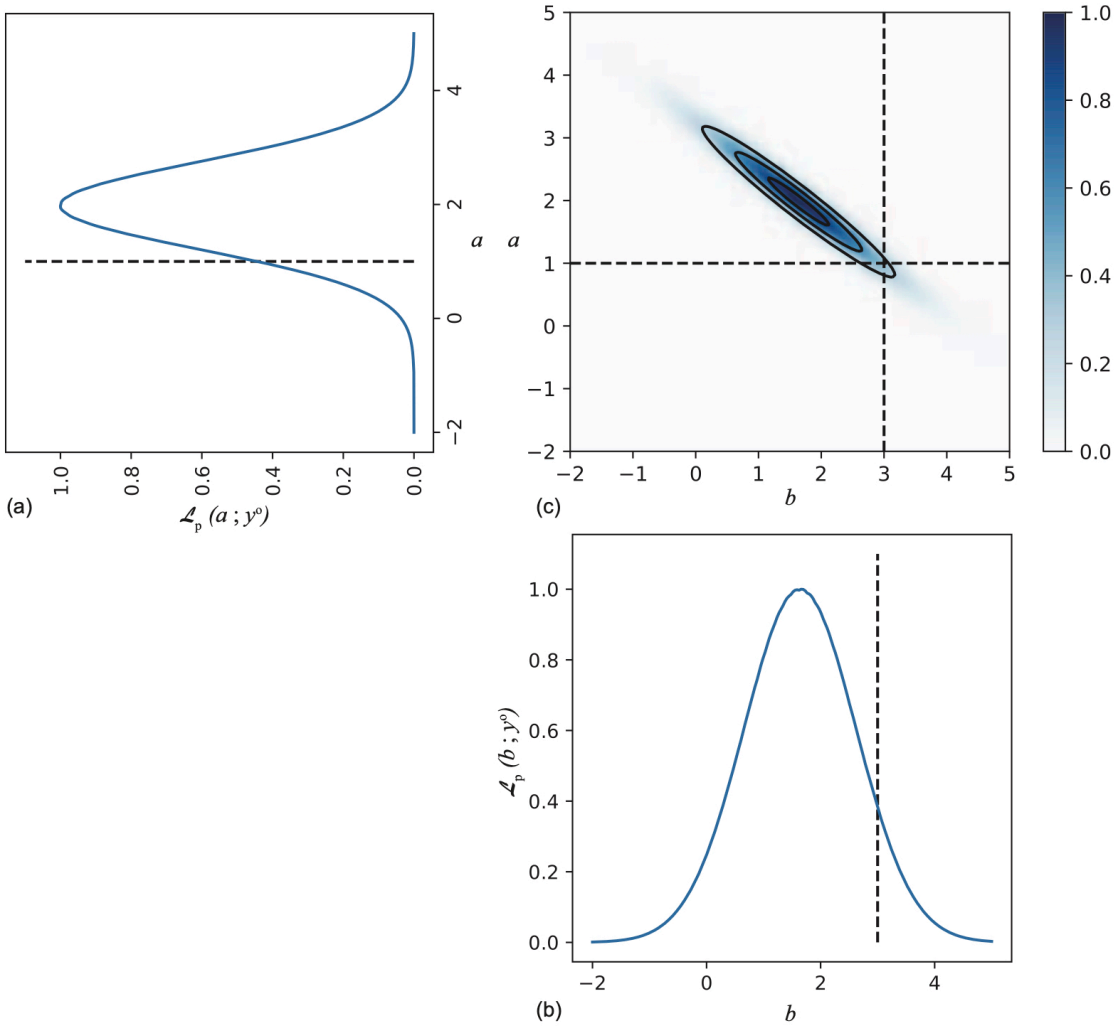
Note: gives curve of optimal $\hat{\lambda}(\psi)$ for each ψ (re-optimise each time).

PROFILE LIKELIHOOD FOR INTEREST PARAMETER

It can be shown that profile likelihood *usually has similar properties to 'standard' likelihood functions*.

E.g. we can use them to form *confidence intervals for interest parameters* by treating them 'as if' they were standard likelihood functions for the interest parameters (same approximate calibration).

PREVIEW



OUR OTHER EXAMPLES

These are not too different to our constant example!

Try! See worksheet.

Important note: same ideas apply for *non-Gaussian models!*

FURTHER TOPICS

MULTIPLE MEASUREMENTS: DATA SYNTHESIS

For independent data sources representing information on the same parameter we can write

$$p(y, z; \theta) = p(y; \theta)p(z; \theta).$$

Note: doesn't require the same probability model in each case, just same parameter. Gives a (non-Bayesian) way of *capturing stochastic prior information*.

LATENT RANDOM VARIABLES

Often, e.g. with stochastic process models, we might know the form of $p(y, z; \theta)$ but not $p(y; \theta)$. If we only observe y, i.e. z is *latent*, then our likelihood is based on

$$p(y; \theta) = \int p(y, z; \theta) dz$$

This integral is often very difficult. Various approximate methods e.g. Monte-Carlo or Laplace available, but active research area.

LIKELIHOOD-FREE INFERENCE?

Often we have a ‘recipe’ for simulating realisations of data given the parameter, but can’t write down $p(y; \theta)$. Similar to above but might be fully black box, e.g. no $p(y, z; \theta)$ etc either.

This is sometimes called *likelihood-free* inference. A possibly better name is *simulation-based* inference as a likelihood might still ‘exist’ even if we don’t know it….

LIKELIHOOD-FREE INFERENCE?

A general approach is to construct a *surrogate* or *approximate* likelihood and then use in usual way. Or, construct as you go.

Active area of research for both Bayesians and frequentists, though arguably Bayesians most active in this area (they essentially *need* a likelihood as a key ingredient; frequentists don't necessarily).

See Chris Drovandi's talk!

QUASI-PSEUDO….

Related: use one of various ‘likelihood-like’ inference functions. A whole zoo:

More popular in frequentist world (don’t care if ‘true’ likelihood).

READING RECOMMENDATIONS: NEO-FISHERIAN

- In All Likelihood by Y. Pawitan
- Statistical Inference in Science by D.A. Sprott
- Principles of Statistical Inference by D.R. Cox
- Principles of Statistical Inference from a Neo-Fisherian Perspective by L. Pace and A. Salvan

READING RECOMMENDATIONS: PURE LIKELIHOOD

- Likelihood by A.W.F. Edwards
- Statistical Evidence: A Likelihood Paradigm by R. Royall
- Logic of Statistical Inference by I. Hacking

YOUR TURN!

THANKS!