Lesson 3: Likelihood-based inference for POMP models

Qianying (Ruby) Lin   Spencer J. Fox   Zian (Larry) Zhuang

Objectives

Students completing this lesson will:

1. Gain an understanding of the nature of the problem of likelihood computation for POMP models.
2. Be able to explain the simplest particle filter algorithm.
3. Gain experience in the visualization and exploration of likelihood surfaces.
4. Be able to explain the tools of likelihood-based statistical inference that become available given numerical accessibility of the likelihood function.

Overview I

A general framework of epidemiological inference includes three layers:

▶ The input: a model of interest and the given data
▶ A method for inference
▶ Inferences include estimation, uncertainty, prediction and forecast, and model selection.
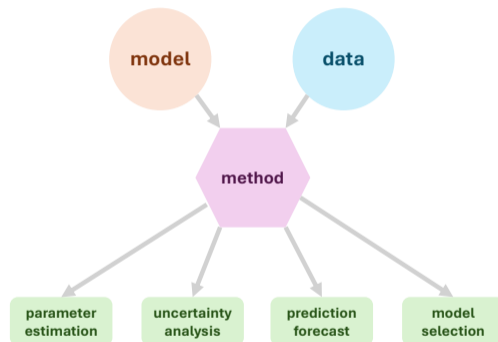
Overview II

Methods for inference can be categorized into three groups:

▶ Optimization-based: minimize a cost function (e.g., SSE, MSE, MAE) that measures the difference between observed data and model predictions

▶ Likelihood-based: maximize a likelihood function, which represents the probability of observing the given data given the parameters

▶ Summary Statistics-based: use a set of features of the data instead of the full set of data

In this lesson, we focus on the likelihood-based method because

▶ it fits for stochastic models and

▶ it incorporates all data (i.e., full-information).

The likelihood I

▶ The basis for modern frequentist, Bayesian, and information-theoretic inference.
▶ Method of maximum likelihood introduced by Fisher (1922).
▶ The likelihood function itself is a representation of the what the data have to say about the parameters.
▶ A good general reference on likelihood is by Pawitan (2001).

The likelihood II

▶ Goal: fit the model to the data and conduct statistical inferences, such as parameter estimation.

▶ The likelihood, thus, can be considered as a metric to assess the *goodness* of the proposed parameters.

▶ By exploring the space of parameters, we can eventually obtain the maximum likelihood estimator (MLE).
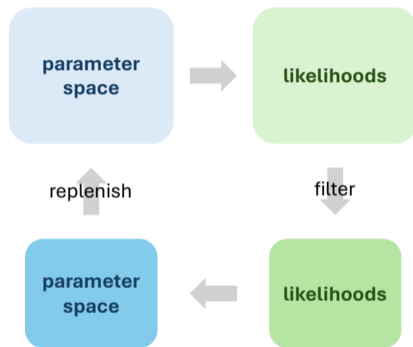
Thus, the objective of this lesson is to discuss how we compute the likelihood given a model of interest with a proposed set of parameters in both theory and in pomp.

Definition of the likelihood function

▶ Notations:
  ▶ $y_{1:N}^*$: the data, a sequence of N observations
  ▶ $f_{Y_{1:N}}(y_{1:N}; \theta)$: the statistical model, a probability distribution for each value of a parameter vector θ
  ▶ $Y_{1:N} \sim f_{Y_{1:N}}(y_{1:N}; \theta)$: a random variable drawn from distribution $f_{Y_{1:N}}(y_{1:N}; \theta)$

▶ The likelihood function is

$$\mathcal{L}(\theta) = f_{Y_{1:N}}(y_{1:N}^*; \theta),$$

the density function evaluated at the data.

▶ It is often convenient to work with the log-likelihood function,

$$\ell(\theta) = \log \mathcal{L}(\theta) = \log f_{Y_{1:N}}(y_{1:N}^*; \theta).$$

A simulator is implicitly a statistical model

▶ $f_{Y_{1:N}}(y_{1:N};\theta)$ is simple and with an explicit expression:
  ▶ the simulation of $Y_{1:N}$ is direct, e.g., $Y_k \sim N(0,1)$ for $k = 1, ..., N$
  ▶ the likelihood function is explicit

▶ $f_{Y_{1:N}}(y_{1:N};\theta)$ is complex or even without an explicit expression:
  ▶ the simulation of $Y_{1:N}$, given the underlying dynamical model, is a bit more complex but convenient
  ▶ the likelihood function exists with a complicated expression or even without an explicit expression

Thus, we can develop numerical methods to compute the complex or implicit likelihood functions!

The likelihood for a POMP model I

Recall the following schematic diagram, showing dependence among variables in a POMP model.
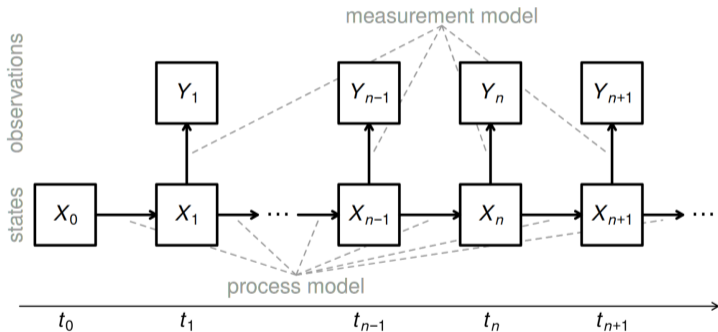
The likelihood for a POMP model II

Recall the following definitions and properties:

▶ **Measurements**: Y_n, at time t_n depend on the latent process, X_n, at that time.

▶ **The Markov property**: latent process variables depend on their value at the previous timestep.

1. The distribution of the state X_{n+1}, conditional on X_n, is independent of the values of X_k, $k < n$ and Y_k, $k \le n$.
2. The distribution of the measurement Y_n, conditional on X_n, is independent of all other variables.

▶ **The latent process**: $X(t)$, may be defined at all times, but we are particularly interested in its value at observation times. Therefore, we write

$$X_n = X(t_n).$$

▶ We write collections of random variables using the notation $X_{0:N} = (X_0, ..., X_N)$.

The likelihood for a POMP model III

▶ **The one-step transition density**: $f_{X_n|X_{n-1}}(x_n|x_{n-1}; \theta)$, together with the measurement density, $f_{Y_n|X_n}(y_n|x_n; \theta)$ and the initial density, $f_{X_0}(x_0; \theta)$, specify the entire joint density via

$$f_{X_{0:N}, Y_{1:N}}(x_{0:N}, y_{1:N}; \theta)$$
$$= f_{X_0}(x_0; \theta) \prod_{n=1}^{N} f_{X_n|X_{n-1}}(x_n|x_{n-1}; \theta) \, f_{Y_n|X_n}(y_n|x_n; \theta).$$

▶ **The marginal density for sequence of measurements**: $Y_{1:N}$, evaluated at the data, $y_{1:N}^*$, is

$$\mathcal{L}(\theta) = f_{Y_{1:N}}(y_{1:N}^*; \theta) = \int f_{X_{0:N}, Y_{1:N}}(x_{0:N}, y_{1:N}^*; \theta) \, dx_{0:N}.$$

Special case: deterministic latent process

▶ When the latent process is non-random, the log-likelihood for a POMP model closely resembles a nonlinear regression model.

▶ In this case, we can write $X_n = x_n(\theta)$, and the log-likelihood is

$$\ell(\theta) = \sum_{n=1}^{N} \log f_{Y_n|X_n}(y_n^*|x_n(\theta); \theta).$$

▶ If we have a Gaussian measurement model, where Y_n given $X_n = x_n(\theta)$ is conditionally normal with mean $\hat{y}_n(x_n(\theta))$ and constant variance σ^2, then the log-likelihood contains a sum of squares which is exactly the criterion that nonlinear least squares regression seeks to minimize.

▶ More details on deterministic latent process models are given as a supplement.

General case: stochastic unobserved state process

▶ For a POMP model, the likelihood takes the form of an integral:

$$
\begin{aligned}
\mathcal{L}(\theta) &= f_{Y_{1:N}}(y_{1:N}^*; \theta) \\
&= \int f_{X_0}(x_0; \theta) \prod_{n=1}^{N} f_{Y_n|X_n}(y_n^*|x_n; \theta) \, f_{X_n|X_{n-1}}(x_n|x_{n-1}; \theta) \, dx_{0:N}.
\end{aligned}
\tag{1}
$$

▶ This integral is high dimensional and, except for the simplest cases, can not be reduced analytically.

Monte Carlo likelihood: direct simulation I

Spoiler Alert: This section serves to introduce the concept of the **particle filter** and the approach of Monte Carlo integration by first proposing an intuitive and a simpler method. This simple method usually **does NOT work** on anything but **very short** time series.

1. Let's rewrite the likelihood integral using an equivalent factorization. As an exercise, you could check how the equivalence of Equation 1 and Equation 2 follows algebraically from the Markov property and the definition of conditional density.

$$
\begin{aligned}
\mathcal{L}(\theta) &= f_{Y_{1:N}}(y_{1:N}^*; \theta) \\
&= \int \left\{ \prod_{n=1}^{N} f_{Y_n|X_n}(y_n^*|x_n; \theta) \right\} f_{X_{0:N}}(x_{0:N}; \theta) \, dx_{0:N}.
\end{aligned}
\tag{2}
$$

Monte Carlo likelihood: direct simulation II

2. Notice, using the representation in Equation 2, that the likelihood can be written as an expectation,

$$\mathcal{L}(\theta) = \mathbb{E}\left[\prod_{n=1}^{N} f_{Y_n|X_n}(y_n^*|X_n;\theta)\right],$$

where the expectation is taken with $X_{0:N} \sim f_{X_{0:N}}(x_{0:N};\theta)$.

3. Now, using a law of large numbers, we can approximate an expectation by the average of a Monte Carlo sample. Thus,

$$\mathcal{L}(\theta) \approx \frac{1}{J}\sum_{j=1}^{J}\prod_{n=1}^{N} f_{Y_n|X_n}(y_n^*|X_n^j;\theta),$$

where $\{X_{0:N}^j, j = 1, \dots, J\}$ is a Monte Carlo sample of size J drawn from $f_{X_{0:N}}(x_{0:N};\theta)$.

Monte Carlo likelihood: direct simulation III

In conclusion, we can generate trajectories by simulation and all we need to do to get a Monte Carlo estimate of the likelihood is to evaluate the measurement density of the data at each trajectory and average. In the context of the **plug-and-play** framework, our algorithm depends on rprocess for simulation but does not require dprocess for evaluation. However, this naive approach scales poorly with dimension:

▶ it requires a Monte Carlo effort that scales exponentially with the length of the time series, and so is infeasible on anything but a short data set;

▶ due to stochasticity, once a simulated trajectory diverges from the data, it will seldom come back;

▶ simulations that lose track and deviate from the data are harmful for likelihood estimation;

▶ when simulating a long time series, almost all the simulated trajectories will eventually lose track of the data.

▶ measles outbreak example: supplementary material.

Sequential Monte Carlo: The particle filter I

Fortunately, we can compute the likelihood for a POMP model by a much more efficient algorithm than direct Monte Carlo integration:

1. We proceed by factorizing the likelihood in a different way:

$$\mathcal{L}(\theta) = f_{Y_{1:N}}(y^*_{1:N}; \theta) = \prod_{n=1}^{N} f_{Y_n|Y_{1:n-1}}(y^*_n|y^*_{1:n-1}; \theta)$$

$$= \prod_{n=1}^{N} \int f_{Y_n|X_n}(y^*_n|x_n; \theta) \, f_{X_n|Y_{1:n-1}}(x_n|y^*_{1:n-1}; \theta) \, dx_n,$$

with the understanding that $f_{X_1|Y_{1:0}} = f_{X_1}$.

Sequential Monte Carlo: The particle filter II

2. The Markov property leads to the **prediction formula:**

$$f_{X_n|Y_{1:n-1}}(x_n|y_{1:n-1}^*;\theta)$$
$$= \int f_{X_n|X_{n-1}}(x_n|x_{n-1};\theta) \, f_{X_{n-1}|Y_{1:n-1}}(x_{n-1}|y_{1:n-1}^*;\theta) \, dx_{n-1}.$$

3. Bayes' theorem gives the **filtering formula:**

$$f_{X_n|Y_{1:n}}(x_n|y_{1:n}^*;\theta)$$
$$= f_{X_n|Y_n,Y_{1:n-1}}(x_n|y_n^*, y_{1:n-1}^*;\theta)$$
$$= \frac{f_{Y_n|X_n}(y_n^*|x_n;\theta) \, f_{X_n|Y_{1:n-1}}(x_n|y_{1:n-1}^*;\theta)}{\int f_{Y_n|X_n}(y_n^*|u_n;\theta) \, f_{X_n|Y_{1:n-1}}(u_n|y_{1:n-1}^*;\theta) \, du_n}.$$

Sequential Monte Carlo: The particle filter III

▶ This suggests that we keep track of two key distributions at each time t_n,
  ▶ The **prediction distribution** is $f_{X_n|Y_{1:n-1}}(x_n|y_{1:n-1}^*)$.
  ▶ The **filtering distribution** is $f_{X_n|Y_{1:n}}(x_n|y_{1:n}^*)$.

▶ The prediction and filtering formulas give us a two-step recursion:
  ▶ The prediction formula gives the prediction distribution at time t_n using the filtering distribution at time t_{n-1}.
  ▶ The filtering formula gives the filtering distribution at time t_n using the prediction distribution at time t_n.

▶ The **particle filter** use Monte Carlo techniques to sequentially estimate the integrals in the prediction and filtering recursions. Hence, the alternative name of **sequential Monte Carlo (SMC)**.

Sequential Monte Carlo: The particle filter IV

A basic particle filter is described as follows:

1. Suppose $X_{n-1,j}^F$, $j = 1, \ldots, J$ is a set of J points drawn from the filtering distribution at time t_{n-1}.

2. We obtain a sample $X_{n,j}^P$ of points drawn from the prediction distribution at time t_n by simply simulating the process model:

$$X_{n,j}^P \sim \text{process}(X_{n-1,j}^F, \theta), \qquad j = 1, \ldots, J.$$

3. Having obtained $x_{n,j}^P$, we obtain a sample of points from the filtering distribution at time t_n by *resampling* from $\left\{ X_{n,j}^P, j \in 1 : J \right\}$ with weights

$$w_{n,j} = f_{Y_n | X_n}(y_n^* | X_{n,j}^P; \theta).$$

Sequential Monte Carlo: The particle filter V

4. The Monte Carlo principle tells us that the conditional likelihood

$$\mathcal{L}_n(\theta) = f_{Y_n|Y_{1:n-1}}(y_n^*|y_{1:n-1}^*;\theta)$$
$$= \int f_{Y_n|X_n}(y_n^*|x_n;\theta)\, f_{X_n|Y_{1:n-1}}(x_n|y_{1:n-1}^*;\theta)\, dx_n$$

is approximated by

$$\hat{\mathcal{L}}_n(\theta) \approx \frac{1}{J} \sum_j f_{Y_n|X_n}(y_n^*|X_{n,j}^P;\theta)$$

since $X_{n,j}^P$ is approximately a draw from $f_{X_n|Y_{1:n-1}}(x_n|y_{1:n-1}^*;\theta)$.

5. We can iterate this procedure through the data, one step at a time, alternately simulating and resampling, until we reach $n = N$.

Sequential Monte Carlo: The particle filter VI

6. The full log-likelihood then has approximation

$$\ell(\theta) = \log \mathcal{L}(\theta) = \sum_n \log \mathcal{L}_n(\theta) \approx \sum_n \log \hat{\mathcal{L}}_n(\theta).$$

▶ References on the particle filter include Kitagawa (1987), Arulampalam et al. (2002), Doucet, Freitas, and Gordon (2001), King, Nguyen, and Ionides (2016).

▶ It can be shown that the particle filter provides an unbiased estimate of the likelihood. This implies a consistent but biased estimate of the log-likelihood.

A block diagram representation of a particle filter
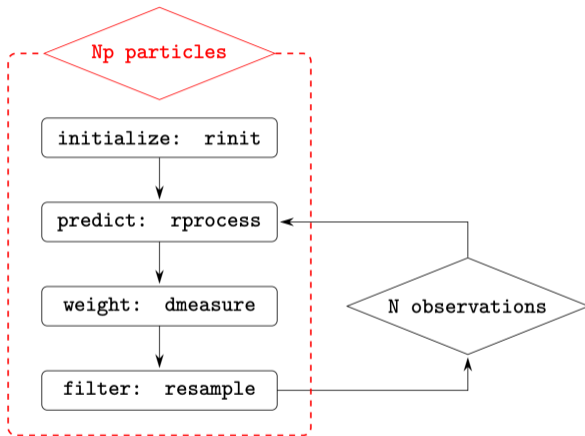
Parallel computing

It will be helpful to parallelize most of the computations. Most machines nowadays have multiple cores and using this computational capacity is as simple as:

1. letting R know you plan to use multiple processors;
2. using the parallel for loop provided by the foreach package; and
3. paying proper attention to the use of parallel random number generators (RNG).

For example:

```
library(foreach)            # load foreach
library(doFuture)           # load doFuture (and future)
plan(multisession)          # using multiple R sessions
```

The second line tells foreach that we will use the doFuture backend. By default, R will attempt to determine how many cores are available and will run an appropriate number of concurrent R processes.

Particle filtering in pomp I

Recall the measles-outbreak example and the stochastic SIR model that we construct in the previous lesson, we can using the `pfilter` function to compute the likelihood using particle filtering method, given the parameters chosen by looking at simulations. R code to build the model is available here. We can execute this code by sourcing the file and check the parameters:

```
source("model_measSIR.R")
measSIR@params
```

```
   Beta    Gamma      Rho        k      Eta        N
1.5e+01  5.0e-01  5.0e-01  1.0e+01  6.0e-02  3.8e+04
```

Particle filtering in pomp II

In pomp, we can compute the likelihood using the particle filtering method, implemented by function `pfilter`. The argument `Np` assigns the number of particles used:

```
library(pomp)
pf <- measSIR |> pfilter(Np=5000)
logLik(pf)
```

```
[1] -131.0779
```

Particle filtering in pomp III

The particle filtering method relies heavily on the state process and the measurement model. Therefore, it is necessary to make sure that the basic particle filter is working.

1. Check the `rprocess` and the `rmeasure` by simulation, as shown in Lesson 2:

```
measSIR |>
  simulate(nsim=20,format="data.frame",include.data=TRUE) |>
  ggplot(aes(x=week,y=reports,group=.id,color=.id=="data")) +
  geom_line() + guides(color="none")
```
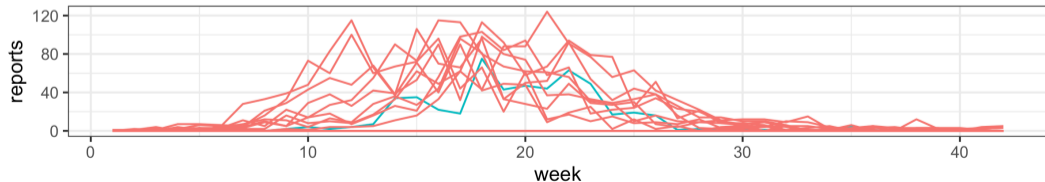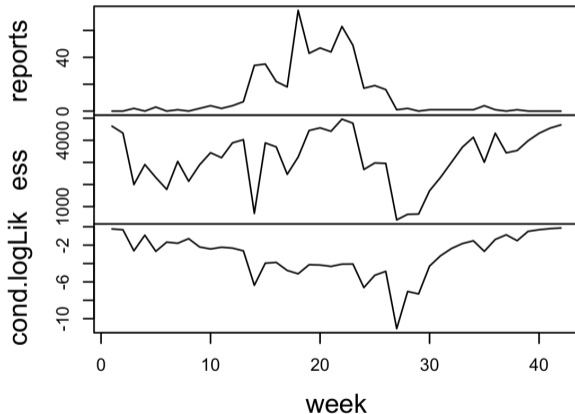
Particle filtering in pomp IV

2. A diagnostic plot to check the `rprocess` and the `dmeasure`:

`plot(pf)`

▶ The data, `reports`;

▶ The **effective sample size** of the particle filter, `ess`;

▶ The log-likelihood of each observation conditioned on the preceding ones, `cond.logLik`.

3. The Monte Carlo variability of the likelihood:

```
plan(multisession)
foreach (
  i=1:10, .combine=c, .options.future=list(seed=652643293)
) %dofuture% {
    measSIR |> pfilter(Np=5000)
} -> pf
logLik(pf) -> ll
logmeanexp(ll,se=TRUE)
```

```
        est              se
-130.037782      1.253115
```

Note that we set the parallel RNG seed in the `foreach` call.

Review of likelihood-based inference

For now, let us suppose that software exists to evaluate and maximize the likelihood function, up to a tolerable numerical error, for the dynamic models of interest. Our immediate task is to think about how to use that capability.

▶ Likelihood-based inference (meaning statistical tools based on the likelihood function) provides tools for parameter estimation, standard errors, hypothesis tests and diagnosing model misspecification.

▶ Likelihood-based inference often (but not always) has favorable theoretical properties. Here, we are not especially concerned with the underlying theory of likelihood-based inference. On any practical problem, we can check the properties of a statistical procedure by simulation experiments.

The maximum likelihood estimate (MLE)

▶ A maximum likelihood estimate (MLE) is

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \, \ell(\theta),$$

where $\operatorname{argmax}_\theta g(\theta)$ means a value of argument θ at which the maximum of the function g is attained, so $g\left(\operatorname{argmax}_\theta g(\theta)\right) = \max_\theta g(\theta)$.

▶ If there are many values of θ giving the same maximum value of the likelihood, then an MLE still exists but is not unique.

▶ Note that $\operatorname{argmax}_\theta \mathcal{L}(\theta)$ and $\operatorname{argmax}_\theta \ell(\theta)$ are the same. Why?

Standard errors for the MLE

▶ Parameter estimates are not very useful without some measure of their uncertainty.

▶ Usually, this means obtaining a confidence interval, or in practice an interval close to a true confidence interval which should formally be called an approximate confidence interval. In practice, the word "approximate" is often dropped!

There are three main approaches to estimating the statistical uncertainty in an MLE.

1. The Fisher information.
2. Profile likelihood estimation.
3. A simulation study, also known as a bootstrap.

Fisher information

▶ A computationally quick approach when one has access to satisfactory numerical second derivatives of the log-likelihood.

▶ The approximation is satisfactory only when $\hat{\theta}$ is well approximated by a normal distribution.

▶ Neither of the two requirements above are typically met for POMP models.

▶ A review of standard errors via Fisher information is provided as a supplement.

Profile likelihood estimation

This approach is generally preferable to the Fisher information for POMP models.
We will explain this method below and put it into practice in the next lesson.

The bootstrap

▶ If done carefully and well, this can be the best approach.

▶ A confidence interval is a claim about reproducibility. You claim, so far as your model is correct, that on 95% of realizations from the model, a 95% confidence interval you have constructed will cover the true value of the parameter.

▶ A simulation study can check this claim fairly directly, but requires the most effort.

▶ The simulation study takes time for you to develop and debug, time for you to explain, and time for the reader to understand and check what you have done. We usually carry out simulation studies to check our main conclusions only.

▶ Further discussion of bootstrap methods for POMP models is provided as a supplement.

Confidence intervals via the profile likelihood I

▶ Let's consider the problem of obtaining a confidence interval for the first component of θ. We'll write

$$\theta = (\phi, \psi).$$

▶ The **profile log-likelihood function** of ϕ is defined to be

$$\ell^{\mathrm{profile}}(\phi) = \max_{\psi} \ell(\phi, \psi).$$

In general, the profile likelihood of one parameter is constructed by maximizing the likelihood function over all other parameters.

▶ Note that, $\max_{\phi} \ell^{\mathrm{profile}}(\phi) = \max_{\theta} \ell(\theta)$ and that maximizing the profile likelihood $\ell^{\mathrm{profile}}(\phi)$ gives the MLE, $\hat{\theta}$. Why?

▶ An approximate 95% confidence interval for ϕ is given by

$$\{\phi : \ell(\hat{\theta}) - \ell^{\mathrm{profile}}(\phi) < 1.92\}.$$

Confidence intervals via the profile likelihood II

▶ This is known as a profile likelihood confidence interval. The cutoff 1.92 is derived using Wilks' theorem, which we will discuss in more detail when we develop likelihood ratio tests.

▶ Although the asymptotic justification of Wilks' theorem is the same limit that justifies the Fisher information standard errors, profile likelihood confidence intervals tend to work better than Fisher information confidence intervals when N is not so large—particularly when the log-likelihood function is not close to quadratic near its maximum.

The likelihood surface

▶ It is extremely useful to visualize the geometric surface defined by the likelihood function.

▶ If Θ is two-dimensional, then the surface $\ell(\theta)$ has features like a landscape.

▶ Local maxima of $\ell(\theta)$ are peaks.

▶ Local minima are valleys.

▶ Peaks may be separated by a valley or may be joined by a ridge. If you go along the ridge, you may be able to go from one peak to the other without losing much elevation. Narrow ridges can be easy to fall off, and hard to get back on to.

▶ In higher dimensions, one can still think of peaks and valleys and ridges. However, as the dimension increases it quickly becomes hard to imagine the surface.

Exploring the likelihood surface: slices

▶ To get an idea of what the likelihood surface looks like in the neighborhood of a point in parameter space, we can construct some likelihood *slices*.

▶ A likelihood slice is a cross-section through the likelihood surface.

▶ We'll make slices for our Consett measles POMP model, in the β and μ_{IR} directions.

▶ Both slices will pass through our current candidate parameter vector, stored in the pomp model object.

Questions:

1. What is the difference between a likelihood slice and a profile?
2. What is the consequence of this difference for the statistical interpretation of these plots?
3. How should you decide whether to compute a profile or a slice?

Worked solution to the Exercise

Slicing the measles SIR likelihood I

▶ We first construct a data frame to explore the parameter slice, with $40 \times 3 + 40 \times 3 = 240$ rows:

```
slice_design(
  center = coef(measSIR),
  Beta = rep(seq(from=5,to=30,length=40),each=3),
  Gamma = rep(seq(from=0.2,to=2,length=40),each=3)
) -> param_slice

dim(param_slice)
```

```
[1] 240   7
```

Slicing the measles SIR likelihood II

▶ We compute the likelihoods 3 times for each combination (i.e., row) in
`param_slice`:

```
library(iterators)
plan(multisession)
foreach (theta=iter(param_slice,"row"), .combine=rbind,
  .options.future=list(seed=108028909)) %dofuture% {
  measSIR |> pfilter(params=theta,Np=5000) -> pf
  theta$loglik <- logLik(pf)
  theta
} -> lik_slice
```

Slicing the measles SIR likelihood III
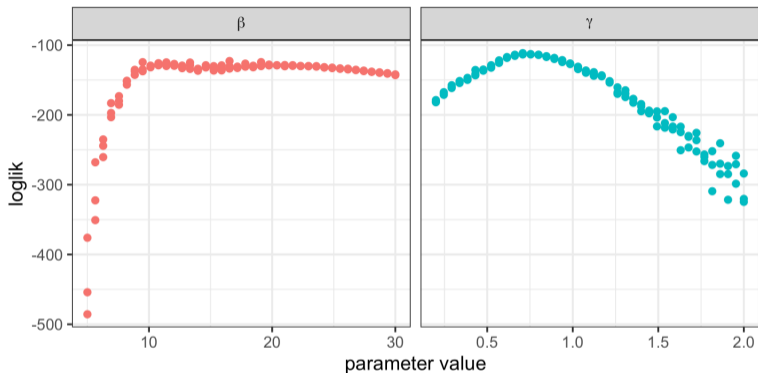
▶ Slices offer a very limited perspective on the geometry of the likelihood surface.
▶ When there are only one or two unknown parameters, we can evaluate the likelihood at a grid of points and visualize the surface directly.

Two-dimensional likelihood slice I

▶ We first construct the parameters grid data frame `param_grid`, with
$40 \times 3 \times 40 \times 3 = 14,400$ rows:

```
expand.grid(
  Beta = rep(seq(from=10,to=30,length=40), each=3),
  Gamma = rep(seq(from=0.4,to=1.5,length=40), each=3),
  Rho = 0.5, k=10, Eta=0.06, N=38000
) -> param_grid

dim(param_grid)
```
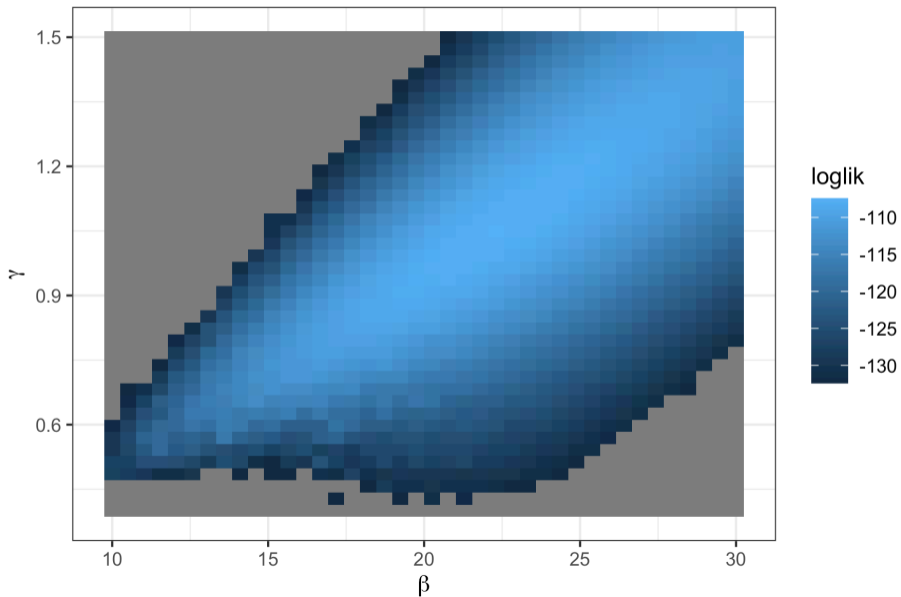
```
[1] 14400     6
```

Two-dimensional likelihood slice II

▶ We then compute likelihoods for each of the combinations:

```
plan(multisession)
foreach (theta=iter(param_grid,"row"), .combine=rbind,
  .options.future=list(seed=421776444)) %dofuture% {
  measSIR |> pfilter(params=theta,Np=5000) -> pf
  theta$loglik <- logLik(pf)
  theta
} -> lik_grid
```

Two-dimensional likelihood slice III

Cost of a particle-filter calculation

▶ How much computer processing time does a particle filter take?
▶ How does this scale with the number of particles?

Form a conjecture based upon your understanding of the algorithm. Test your conjecture by running a sequence of particle filter operations, with increasing numbers of particles (Np), measuring the time taken for each one using `system.time`. Plot and interpret your results.

Worked solution to the Exercise

log-likelihood estimation I

Here are some desiderata for a Monte Carlo log-likelihood approximation:

▶ It should have low Monte Carlo bias and variance.
▶ It should be presented together with estimates of the bias and variance so that we know the extent of Monte Carlo uncertainty in our results.
▶ It should be computed in a length of time appropriate for the circumstances.

Set up a likelihood evaluation for the measles model, choosing the numbers of particles and replications so that your evaluation takes approximately one minute on your machine.

▶ Provide a Monte Carlo standard error for your estimate.
▶ Comment on the bias of your estimate.
▶ Use doFuture to take advantage of multiple cores on your computer to improve your estimate.

Worked solution to the Exercise

Exercises

1. **One-dimensional likelihood slice**: Compute several likelihood slices in the η direction.
2. **Two-dimensional likelihood slice**: Compute a slice of the likelihood in the β-η plane.

Maximizing the particle filter likelihood

▶ Likelihood maximization is key to profile intervals, likelihood ratio tests and AIC as well as the computation of the MLE.

▶ An initial approach to likelihood maximization might be to stick the particle filter log-likelihood estimate into a standard numerical optimizer, such as the Nelder-Mead algorithm.

▶ In practice this approach is unsatisfactory on all but the smallest POMP models. Standard numerical optimizers are not designed to maximize noisy and computationally expensive Monte Carlo functions.

▶ Further investigation into this approach is available as a supplement.

▶ We'll present an *iterated filtering algorithm* for maximizing the likelihood in a way that takes advantage of the structure of POMP models and the particle filter.

▶ First, let's think a bit about some practical considerations in interpreting the MLE for a POMP.

Likelihood-based model selection and model diagnostics

▶ For nested hypotheses, we can carry out model selection by likelihood ratio tests.

▶ For non-nested hypotheses, likelihoods can be compared using Akaike's information criterion (AIC) or related methods.

Likelihood ratio tests for nested hypotheses I

▶ The whole parameter space on which the model is defined is $\Theta \subset \mathbb{R}^D$.

▶ Suppose we have two **nested** hypotheses

$$H^{\langle 0 \rangle} : \theta \in \Theta^{\langle 0 \rangle},$$
$$H^{\langle 1 \rangle} : \theta \in \Theta^{\langle 1 \rangle},$$

defined via two nested parameter subspaces, $\Theta^{\langle 0 \rangle} \subset \Theta^{\langle 1 \rangle}$, with respective dimensions $D^{\langle 0 \rangle} < D^{\langle 1 \rangle} \leq D$.

▶ We consider the log-likelihood maximized over each of the hypotheses,

$$\ell^{\langle 0 \rangle} = \sup_{\theta \in \Theta^{\langle 0 \rangle}} \ell(\theta),$$
$$\ell^{\langle 1 \rangle} = \sup_{\theta \in \Theta^{\langle 1 \rangle}} \ell(\theta).$$

Likelihood ratio tests for nested hypotheses II

▶ **Wilks approximation**: under the hypothesis $H^{\langle 0 \rangle}$,

$$\ell^{\langle 1 \rangle} - \ell^{\langle 0 \rangle} \approx \tfrac{1}{2}\, \chi^2_{D^{\langle 1 \rangle} - D^{\langle 0 \rangle}},$$

where χ^2_d is a chi-squared random variable on d degrees of freedom and $\approx$ means "is approximately distributed as".

▶ The Wilks approximation can be used to construct a hypothesis test of the null hypothesis $H^{\langle 0 \rangle}$ against the alternative $H^{\langle 1 \rangle}$.

▶ This is called a **likelihood ratio test** since a difference of log-likelihoods corresponds to a ratio of likelihoods.

▶ When the data are IID, $N \to \infty$, and the hypotheses satisfy suitable regularity conditions, this approximation can be derived mathematically and is known as **Wilks' theorem**.

Likelihood ratio tests for nested hypotheses III

▶ The chi-squared approximation to the likelihood ratio statistic may be useful, and can be assessed empirically by a simulation study, even in situations that do not formally satisfy any known theorem.

Wilks' theorem and profile likelihood I

▶ Suppose we have an MLE, written $\hat{\theta} = (\hat{\phi}, \hat{\psi})$, and a profile log-likelihood for ϕ, given by $\ell^{\text{profile}}(\phi)$.

▶ Consider the likelihood ratio test for the nested hypotheses

$$H^{\langle 0 \rangle} : \phi = \phi_0,$$
$$H^{\langle 1 \rangle} : \phi \text{ unconstrained.}$$

▶ We can compute the 95%-ile for a chi-squared distribution with one degree of freedom: `qchisq(0.95,df=1)` $= 3.841$.

▶ Wilks' theorem then gives us a hypothesis test with approximate size 5% that rejects $H^{\langle 0 \rangle}$ if $\ell^{\text{profile}}(\hat{\phi}) - \ell^{\text{profile}}(\phi_0) < 3.84/2$.

Wilks' theorem and profile likelihood II

▶ It follows that, with probability 95%, the true value of ϕ falls in the set

$$\{\phi : \ell^{\mathrm{profile}}(\hat{\phi}) - \ell^{\mathrm{profile}}(\phi) < 1.92\}.$$

So, we have constructed a profile likelihood confidence interval, consisting of the set of points on the profile likelihood within 1.92 log units of the maximum.

▶ This is an example of a general duality between confidence intervals and hypothesis tests.

Akaike's information criterion (AIC) I

▶ Likelihood ratio tests provide an approach to model selection for nested hypotheses, but what do we do when models are not nested?

▶ A more general approach is to compare likelihoods of different models by penalizing the likelihood of each model by a measure of its complexity.

▶ Akaike's information criterion **AIC** is given by

$$\text{AIC} = -2\,\ell(\hat{\theta}) + 2\,D$$

"Minus twice the maximized log-likelihood plus twice the number of parameters."

▶ We are invited to select the model with the lowest AIC score.

▶ AIC was derived as an approach to minimizing prediction error. Increasing the number of parameters leads to additional **overfitting** which can decrease predictive skill of the fitted model.

Akaike's information criterion (AIC) II

▶ Viewed as a hypothesis test, AIC may have weak statistical properties. It can be a mistake to interpret AIC by making a claim that the favored model has been shown to provide a superior explanation of the data. However, viewed as a way to select a model with reasonable predictive skill from a range of possibilities, it is often useful.

▶ AIC does not penalize model complexity beyond the consequence of reduced predictive skill due to overfitting. One can penalize complexity by incorporating a more severe penalty than the $2D$ term above, such as via BIC.

▶ A practical approach is to use AIC, while taking care to view it as a procedure to select a reasonable predictive model and not as a formal hypothesis test.

References

Arulampalam, M. S., S. Maskell, N. Gordon, and T. Clapp. 2002. "A Tutorial on Particle Filters for Online Nonlinear, Non-Gaussian Bayesian Tracking." *IEEE Trans Signal Process* 50: 174–88. https://doi.org/10.1109/78.978374.

Doucet, Arnaud, Nando de Freitas, and Neil Gordon, eds. 2001. *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag.

Fisher, R. A. 1922. "On the Mathematical Foundations of Theoretical Statistics." *Philos Trans R Soc London A* 222: 309–68. https://doi.org/10.1098/rsta.1922.0009.

King, Aaron A., Dao Nguyen, and Edward L. Ionides. 2016. "Statistical Inference for Partially Observed Markov Processes via the R Package Pomp." *J Stat Softw* 69 (12): 1–43. https://doi.org/10.18637/jss.v069.i12.

Kitagawa, Genshiro. 1987. "Non-Gaussian State-Space Modeling of Nonstationary Time Series." *J Am Stat Assoc* 82 (400): 1032–41. https://doi.org/10.1080/01621459.1987.10478534.

Pawitan, Yudi. 2001. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford: Clarendon Press.

License, acknowledgments, and links

▶ This lesson is prepared for the Simulation-based Inference for Epidemiological Dynamics module at the Summer Institute in Statistics and Modeling in Infectious Diseases, SISMID.

▶ The materials build on previous versions of this course and related courses.

▶ Licensed under the Creative Commons Attribution-NonCommercial license. Please share and remix non-commercially, mentioning its origin.

▶ Produced with R version 4.3.2 and pomp version 5.10.

▶ Compiled on 2024-07-24.

Back to Lesson
R code for this lesson