Review questions

Edward L. Ionides, Aaron A. King, Qianying Lin

June 28, 2021

Questions on Lesson 2: Simulation of stochastic dynamic models

Question 2.1. Scientifically, our conclusions should not depend on the units we choose, but we must get the details right. Suppose our data are two years of weekly aggregated case reports of a disease and we have a continuous time model solved numerically by an Euler timestep of size dt. Which of the following is a correct explanation of our options for properly implementing this in a `pomp` object called `po`?

(A) The measurement times, `time(po)`, should be in units of weeks, such as $1, 2, \ldots, 104$. The latent process can be modeled using arbitrary time units, say days or weeks or years. The units of dt should match the time units of the **latent** process.

(B) The measurement times, `time(po)`, should be in units of weeks, such as $1, 2, \ldots, 104$. The latent process can be modeled using arbitrary time units, say days or weeks or years. The units of dt should be in weeks (in practice, usually a fraction of a week) to match the units of the **measurement** times.

(C) The measurement times do not have to be in units of weeks. For example, we could use `time(po)`$=1/52$, $2/52, \ldots, 2$. The latent process and dt should use the same units of time as the measurement times.

(D) The measurement times do not have to be in units of weeks. For example, we could use `time(po)`$=1/52$, $2/52, \ldots, 2$. The latent process can also use arbitrary units of time, which do not necessarily match the units of the measurement times. The units of dt should match the units used for the **latent** process.

(E) The measurement times do not have to be in units of weeks. For example, we could use `time(po)`$=1/52$, $2/52, \ldots, 2$. The latent process can also use arbitrary units of time, which do not necessarily match the units of the measurement times. The units of dt should match the units used for the **measurement** times.

Question 2.2. Suppose you obtain the following error message when you build your pomp model using Csnippets.

```
Error: error in building shared-object library from C snippets: in ‘Cbuilder’: compilation error:
cannot compile shared-object library ‘/var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_be9007eb030e47cb34264e3e779b6da9.so’:
status = 1
compiler messages:
clang -mmacosx-version-min=10.13 -I"/Library/Frameworks/R.framework/Resources/include" -DNDEBUG
-I’/Users/ionides/Library/R/x86_64/4.1/library/pomp/include’
-I’/Users/ionides/sbied/questions’  -I/usr/local/include   -fPIC  -Wall -g -O2
-c /var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_be9007eb030e47cb34264e3e779b6da9.c
-o /var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_be9007eb030e47cb34264e3e779b6da9.o
/var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
```

```
5324/pomp_be9007eb030e47cb34264e3e779b6da9.c:39:5:
error: called object type 'int' is not a function or function pointer
    W = 0;
    ^
/var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/5
In addition: Warning message:
In system2(command = R.home("bin/R"), args = c("CMD", "SHLIB", "-c",  :
running command 'PKG_CPPFLAGS="-I'/Users/ionides/Library/R/x86_64/4.1/library/pomp/include'
-I'/Users/ionides/sbied/questions'" '/Library/Frameworks/R.framework/Resources/bin/R'
CMD SHLIB -c -o /var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_be9007eb030e47cb34264e3e779b6da9.so
/var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_be9007eb030e47cb34264e3e779b6da9.c
2>&1' had status 1
```

Which of the following is a plausible cause for this error?

(A) Using R syntax within a C function that has the same name as an R function.

(B) A parameter is missing from the `paramnames` argument to `pomp`.

(C) Indexing past the end of an array because C labels indices starting at 0.

(D) Using `beta` as a parameter name when it is a declared C function.

(E) A missing semicolon at the end of a line.

Question 2.3. Suppose you obtain the following error message when you build your pomp model using Csnippets.

```
Error: error in building shared-object library from C snippets: in ‘Cbuilder’: compilation error:
cannot compile shared-object library ‘/var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_b675d99e691eda865610f570058ea3be.so’: status = 1
compiler messages:
clang -mmacosx-version-min=10.13 -I"/Library/Frameworks/R.framework/Resources/include" -DNDEBUG
-I'/Users/ionides/Library/R/x86_64/4.1/library/pomp/include' -I'/Users/ionides/sbied/questions'
-I/usr/local/include   -fPIC  -Wall -g -O2
-c /var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_b675d99e691eda865610f570058ea3be.c
-o /var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_b675d99e691eda865610f570058ea3be.o
/var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_b675d99e691eda865610f570058ea3be.c:33:16:
error: use of undeclared identifier 'pop'; did you mean 'pow'?
    double m = pop/(S_0+I_0+R_0);
               ^~~
               pow
/Applications/
In addition: Warning message:
In system2(command = R.home("bin/R"), args = c("CMD", "SHLIB", "-c",  :
  running command 'PKG_CPPFLAGS="-I'/Users/ionides/Library/R/x86_64/4.1/library/pomp/include'
-I'/Users/ionides/sbied/questions'" '/Library/Frameworks/R.framework/Resources/bin/R' CMD SHLIB
-c -o /var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_b675d99e691eda865610f570058ea3be.so
/var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_b675d99e691eda865610f570058ea3be.c 2>&1' had status 1
```

Which of the following is a plausible cause for this error?

(A) Using R syntax within a C function that has the same name as an R function.

(B) A parameter is missing from the `paramnames` argument to `pomp`.

(C) Indexing past the end of an array because C labels indices starting at 0.

(D) Using `beta` as a parameter name when it is a declared C function.

(E) A missing semicolon at the end of a line.

Question 2.4. Suppose you obtain the following error message when you build your pomp model using Csnippets.

```
Error: error in building shared-object library from C snippets: in ‘Cbuilder’: compilation error:
cannot compile shared-object library ‘/var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_77886fb66d95b4b9904440d86a4425b3.so’: status = 1
compiler messages:
clang -mmacosx-version-min=10.13 -I"/Library/Frameworks/R.framework/Resources/include" -DNDEBUG
-I’/Users/ionides/Library/R/x86_64/4.1/library/pomp/include’
-I’/Users/ionides/sbied/questions’  -I/usr/local/include   -fPIC  -Wall -g -O2
-c /var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_77886fb66d95b4b9904440d86a4425b3.c
-o /var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_77886fb66d95b4b9904440d86a4425b3.o
/var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_77886fb66d95b4b9904440d86a4425b3.c:39:36:
error: too many arguments to function call, expected 2, have 3
      rep = nearbyint(rnorm(1,mean,sd));
                      ~~~~~        ^~
/Librar
In addition: Warning message:
In system2(command = R.home("bin/R"), args = c("CMD", "SHLIB", "-c",  :
running command ’PKG_CPPFLAGS="-I’/Users/ionides/Library/R/x86_64/4.1/library/pomp/include’
-I’/Users/ionides/sbied/questions’" ’/Library/Frameworks/R.framework/Resources/bin/R’
CMD SHLIB -c -o /var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_77886fb66d95b4b9904440d86a4425b3.so
/var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_77886fb66d95b4b9904440d86a4425b3.c 2>&1’ had status 1
```

Which of the following is a plausible cause for this error?

(A) Using R syntax within a C function that has the same name as an R function.

(B) A parameter is missing from the `paramnames` argument to `pomp`.

(C) Indexing past the end of an array because C labels indices starting at 0.

(D) Using `beta` as a parameter name when it is a declared C function.

(E) A missing semicolon at the end of a line.

Question 2.5. Let V_n be a Markov process and let $W_n = h(V_n)$ for some function h. Which of the following statements are true?
i) W_n is a Markov process for all choices of h.
ii) W_n is a Markov process for some choices of h.

iii) W_n is not a Markov process for any choice of h.
iv) If $V_n = (X_n, Y_n)$ where X_n and Y_n are a POMP model, and $h(X_n, Y_n) = X_n$ then W_n is a Markov process.
v) If $V_n = (X_n, Y_n)$ where X_n and Y_n are a POMP model, and $h(X_n, Y_n) = Y_n$ then W_n is a Markov process.

(A) i,iv,v

(B) ii,iv

(C) ii,v

(D) iii

(E) None of the above

Questions on Chapter 3: Likelihood for POMPs

Question 3.1. Suppose that 10 replications of a particle filter, each using 10^3 particles, runs in 15 minutes with no parallelization. To look for a more precise likelihood evaulation, you consider running 20 replicates, each with 10^4 particles. How many minutes will this take, if you distribute the calculation across 4 cores?

(A) 50

(B) 60

(C) 75

(D) 120

(E) 300

Question 3.2. A particle filter is repeated 5 times to evaluate the likelihood at a proposed maximum likelihood estimate, each time with 10^4 particles. Suppose the log likelihood estimates are -2446.0, -2444.0, -2443.0, -2442.0, -2440.0. Which of the following is an appropriate estimate for the log likelihood at this parameter value and its standard error.

(A) Estimate = -2443.0, with standard error 1.0

(B) Estimate = -2443.0, with standard error 2.2

(C) Estimate = -2443.0, with standard error 5.0

(D) Estimate = -2441.4, with standard error 2.2

(E) Estimate = -2441.4, with standard error 1.0

Question 3.3. What is the log likelihood (to the nearest unit) of the Dacca cholera data for the POMP model constructed in pomp via

```
d <- dacca(deltaI=0.08)
```

with cholera mortality rate 8% and other parameters fixed at the default values.

(A) -3764

(B) -3765

(C) -3766

(D) -3767

(E) -3768

Question 3.4. Effective sample size (ESS) is one of the main tools for diagnosing the success of a particle filter. It is shown if you plot an object of class `pfilterd_pomp` which is created by applying `pfilter` to a `pomp` object. Suppose one or more time points have low ESS (say, less than 10) even when using a fairly large number of particles (say, 10^4). What is the proper interpretation?

(A) There is a problem with data, perhaps an error recording an observation.

(B) There is a problem with the model which means that it cannot explain something in the data.

(C) The model and data have no major problems, but the model happens to be problematic for the particle filter algorithm.

(D) At least one of (A), (B) and (C).

(E) Either (A) or (B) or both, but not (C). If the model fits the data well, the particle filter is guaranteed to work well.

Questions on Chapter 4: Inference via iterated filtering

Question 4.1. When carrying out inference by iterated particle filtering, the likelihood increases for the first 10 iterations or so, and then steadily decreases. Testing the inference procedure on simulated data, this does not happen and the likelihood increases steadily toward convergence. Which of the following is the best explanation for this?

(A) One or more random walk standard deviation is too large.

(B) One or more random walk standard deviations is too small.

(C) The model is misspecified, so it does not fit the data adequately.

(D) A combination of the parameters is weakly identified, leading to a ridge in the likelihood surface.

(E) Too few particles are being used.

Question 4.2. People sometimes confuse likelihood profiles with likelihood slices. Suppose you read a figure which claims to construct a profile confidence interval for a parameter ρ in a POMP model with four unknown parameters. Which of the following confirms that the plot is, or is not, a properly constructed profile confidence interval. The code producing the plot is available to you as an Rmarkdown file.

(A) The CI is constructed by obtaining the interval of rho values whose log likelihood is within 1.92 of the maximum on a smoothed curve of likelihood values plotted against ρ.

(B) The code involves evaluation of the likelihood but not maximization.

(C) The points along the ρ axis are not equally spaced.

(D) The smoothed line shown in the plot is close to quadratic.

(E) (A) and (D) together.

Questions on Chapter 5: Measles

Question 5.1. Two models are fitted to case counts on an epidemic. Model 1 is an SIR POMP model with a negative binomial measurement model, and model 2 is a linear regression model estimating a cubic trend. The log likelihoods are $\ell_1 = -2037.91$ and $\ell_2 = -2031.28$ respectively. Which of the following do you agree with most?

(A) We should not compare the models using these likelihoods. They correspond to different model structures, so it is an apples-to-oranges comparison.

(B) We can compare them, but the difference is in the 4th significant figure, so the likelihoods are statistically indistinguishable.

(C) The linear model has a noticeably higher likelihood. Our mechanistic model needs to be updated to beat this benchmark before we can responsibly interpret the fitted model. If a simple regression model has higher likelihood than a more complex mechanistic model, one should prefer the simpler model.

(D) The linear model has a noticeably higher likelihood. The mechanistic model is somewhat validated by being not too far behind the simple regression model. We are justified in cautiously interpreting the mechanistic model, while continuing to look for further improvements.

(E) The log likelihoods cannot properly be compared as presented, but could be if we used a Gaussian measurement model for the POMP (or a negative binomial generalized linear model instead of least squares for the regression).

Question 5.2. A compartment model is first implemented as a system of ordinary differential equations (ODEs). This leads to qualitatively reasonable trajectories, but poor likelihood values. The researchers add stochasticity in an attempt to improve the fit of the model by interpreting the ODEs as rates of a Markov chain. The likelihood, maximized by iterated particle filtering, remains poor compared to ARMA benchmarks. In addition, the effective sample size for the particle filtering is low at many time points despite even using as many as 10^4 particles. Which of the following is the most promising next step.

(A) Increase to 10^5 particles, moving the computations to a cluster if necessary.

(B) Add noise to one or more rates to allow for overdispersion.

(C) Try adding extra features to the model to capture scientific details not present in the original model.

(D) Experiment with variations in the iterated filtering procedure; maybe more iterations, or a different cooling schedule.

(E) To address the possibility of reporting errors, see if the model fits better when the most problematic data points are removed.

Question 5.3. You fit an SEIR model to the case report data on an immunizing disease for a city. The resulting confidence interval for the mean latent period is 12-21 days, but clinical evidence points to a latent period averaging about 7 days. Which of the following is the most appropriate response to this discrepancy.

(A) The latent period may be confounded with some unmodeled aspect of the system, such as spatial or age structure. The model estimates an effective latent period at the population level, which may not perfectly match what is happening at the scale of individuals. One should be cautious of making a causal interpretation of models fitted to observational data.

(B) The discrepancy shows that something is substantially wrong with the model. Extra biological detail should be introduced with the goal of bringing the estimated parameter back in line with the known biology of the system.

(C) The discrepancy is problematic, but fortunately can readily be fixed. Since we know the clinical value of this parameter with reasonable accuracy, we should simply use this value in the model rather than estimating it.

(D) If the model fits the data statistically better than any known alternative model, then we have to take the estimated parameter at face value. It is certainly possible that the estimates in the literature correspond to some different population, or different strain, or have some other measurement bias such as corresponding to severe cases resulting in hospitalization. The discrepancy does not show that our model was wrong.

(E) This discrepancy suggests that we should take advantage of both C and D above by putting a Bayesian prior on the latent period. By quantifying the degree of our skepticism about the previously established clinical value of 7 days, we can optimally combine that uncertainty with the evidence from this dataset.