Quiz 2, STATS/DATASCI 531/631 W25

In class on 4/16, 2:30pm to 3:00pm

Name: **UMID:**

Instructions. You have a time allowance of 30 minutes. The quiz may be ended early if everyone is done. The quiz is closed book, and you are not allowed access to any notes. Any electronic devices in your possession must be turned off and remain in a bag on the floor.

For each question, circle one letter answer and provide supporting reasoning. If you run out of space, you may continue on the back of the page, but please indicate to the reader that you are doing so.

Q1. Foundations of POMP models

Consider a model $Y_{1:N}$ for data $y^*_{1:N}$, with a latent variable $X_{0:N}$, and a statistical model defined by a joint density $f_{X_{0:N},Y_{1:N}}(x_{0:N},y_{1:N};\theta)$. The likelihood function is

$$L(\theta) = f_{Y_{1:N}}(y^*_{1:N};\theta).$$

Are the following identities (A) true for all statistical models; (B) true for general POMP models but not all models; (C) true for linear Gaussian POMP models but not general POMP models; (D) generally false? Give a letter answer for each identity.

$$L(\theta) = \int f_{Y_{1:N}|X_{0:N}}(y^*_{1:N}|x_{0:N};\theta)\, f_{X_{0:N}}(x_{0:N};\theta)\, dx_{0:N} \tag{1}$$

$$L(\theta) = \prod_{n=1}^{N} f_{Y_n|Y_{1:n-1}}(y^*_n|y^*_{1:n-1};\theta) \tag{2}$$

$$\mathrm{Var}\{X_{n+1}\,|\,Y_{1:n}\} = E\big[\mathrm{Var}\{X_{n+1}\,|\,X_n\}\,\big|\,Y_{1:n}\big] + \mathrm{Var}\big\{E[X_{n+1}\,|\,X_n]\,\big|\,Y_{1:n}\big\} \tag{3}$$

$$L(\theta) = \int \left[\prod_{n=1}^{N} f_{Y_n|X_n}(y^*_n|x_n;\theta)\right] f_{X_{0:N}}(x_{0:N};\theta)\, dx_{0:N} \tag{4}$$

Supporting reasoning:

Q2. Likelihood evaluation; the particle filter

Effective sample size (ESS) is one of the main tools for diagnosing the success of a particle filter. If you plot an object of class `pfilterd_pomp` (created by applying `pfilter` to a pomp object), the ESS is displayed. 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.

Supporting reasoning:

Q3. Likelihood maximization; iterated filtering

People sometimes confuse likelihood profiles with likelihood slices. When you read a report claiming to have computed a profile it can be worth checking whether it is actually computed as a slice. 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.

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 (made available to you by the authors as an Rmarkdown file) 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.

Supporting reasoning:

Q4. Data analysis: epidemiological models

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).

Supporting reasoning:

Q5. Data analysis: financial models

The Heston model for volatility, V_n, is a stochastic volatility (SV) model with

$$V_n = (1-\phi)\theta + \phi V_{n-1} + \sqrt{V_{n-1}}\,\omega_n,$$

for $\omega_n \sim N[0, \sigma_\omega^2]$. The log return is $Y_n \sim N[0, V_n]$, conditional on V_n. A previous 531 project (W22, #14) fitted the Heston model to investment in Ethereum, a crypto currency. They obtained a log-likelihood of 34975.3, compared to 28587.4 for GARCH and 28977 for the SV model with leverage presented in class. Their iterated filtering convergence diagnostics are shown in figure 1. What is the best conclusion from this information?

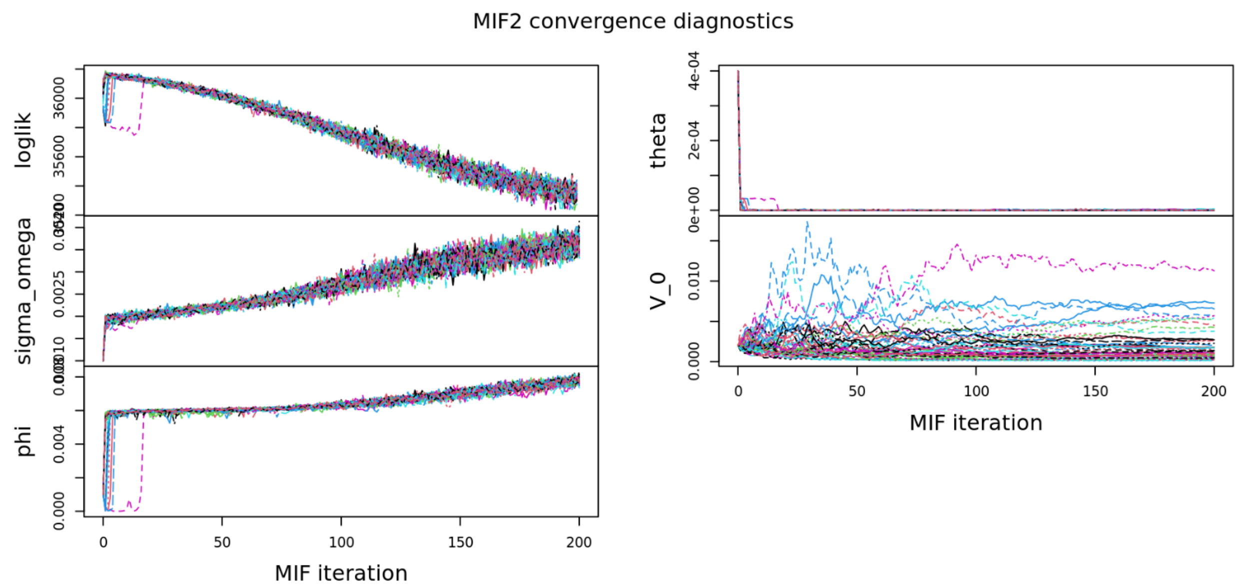

Figure 1: Diagnostic plot for fitting the Heston model

A. The high likelihood shows this is a promising model despite the convergence problems identified in the figure. Attention to the diagnostics may lead to additional improvements.

B. The most important diagnostic feature is the observation that the log-likelihood trace plot peaks and then declines. From the y-axis scale we see the decline is of order 1000 log units. This is evidence of substantial model misspecification which should be addressed.

C. The most important dignostic feature is that the `theta` traces all drop quickly to zero. Since that is not a scientifically plausible value for the parameter, we can deduce that the model is unsuccessful despite its high likelihood.

D. The most important diagnostic feature is that `phi` is close to zero and well identified. This shows that the volatility is close to constant, and is supported by the high likelihood.

E. The decreasing likelihood and other convergence diagnostics problems show there is a problem with the model. Likely, there is a bug and the high likelihood obtained is simply an error.

Supporting reasoning:

Q6. Computing with POMP models

Suppose you obtain the following error message when you build your pomp model using C snippets.

```
##
## Error: error in building shared-object library from C snippets: in 'Cbuilder': compilation error:
##  cannot compile shared-object library
##  '/tmp/RtmpFkkeCQ/24104/pomp_068eedfcaf62b1e391363bbdd99fbe8c.so': status = 1
## compiler messages:
## gcc -I"/usr/local/apps/R/ubuntu_20.04/4.2.1/lib64/R/include" -DNDEBUG
## -I'/home/kingaa/R/x86_64-pc-linux-gnu-library/4.2/pomp/include' -I'/home/kingaa/teach/sbied'
##   -I/usr/local/include   -fpic  -g -O2 -Wall -pedantic
## -c /tmp/RtmpFkkeCQ/24104/pomp_068eedfcaf62b1e391363bbdd99fbe8c.c
## -o /tmp/RtmpFkkeCQ/24104/pomp_068eedfcaf62b1e391363bbdd99fbe8c.o
## /tmp/RtmpFkkeCQ/24104/pomp_068eedfcaf62b1e391363bbdd99fbe8c.c:
## In function '__pomp_rinit':
## /tmp/RtmpFkkeCQ/24104/pomp_068eedfcaf62b1e391363bbdd99fbe8c.c:38:13:
##  error: called object is not a function or function pointer
##    38 |     cases = 0
##       |             ^
## make: *** [/usr/local/apps/R/ubuntu_20.04/4.2.1/lib64/R/etc/Makeconf:168:
##  /tmp/RtmpFkkeCQ/24104/pomp_068eedfcaf62b1e391363bbdd99fbe8c.o] Error 1
## In addition: Warning message:
## In system2(command = R.home("bin/R"), args = c("CMD", "SHLIB", "-c",  :
##   running command 'PKG_CPPFLAGS="-I'/home/kingaa/R/x86_64-pc-linux-gnu-library/4.2/pomp/include'
##  -I'/home/kingaa/teach/sbied'" '/usr/local/apps/R/ubuntu_20.04/4.2.1/lib64/R/bin/R' CMD SHLIB -c
##  -o /tmp/RtmpFkkeCQ/24104/pomp_068eedfcaf62b1e391363bbdd99fbe8c.so
##  /tmp/RtmpFkkeCQ/24104/pomp_068eedfcaf62b1e391363bbdd99fbe8c.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.

Supporting reasoning:
