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

Let V_n be a Markov process and let $W_n = h(V_n)$ for some function h. Let (X_n, Y_n) be a POMP with latent process X_n and observed process Y_n. 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 $V_n = (X_n, Y_n)$ and $h(X_n, Y_n) = X_n$ then W_n is a Markov process.

v. If $V_n = (X_n, Y_n)$ 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. Some combination other than those listed above

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

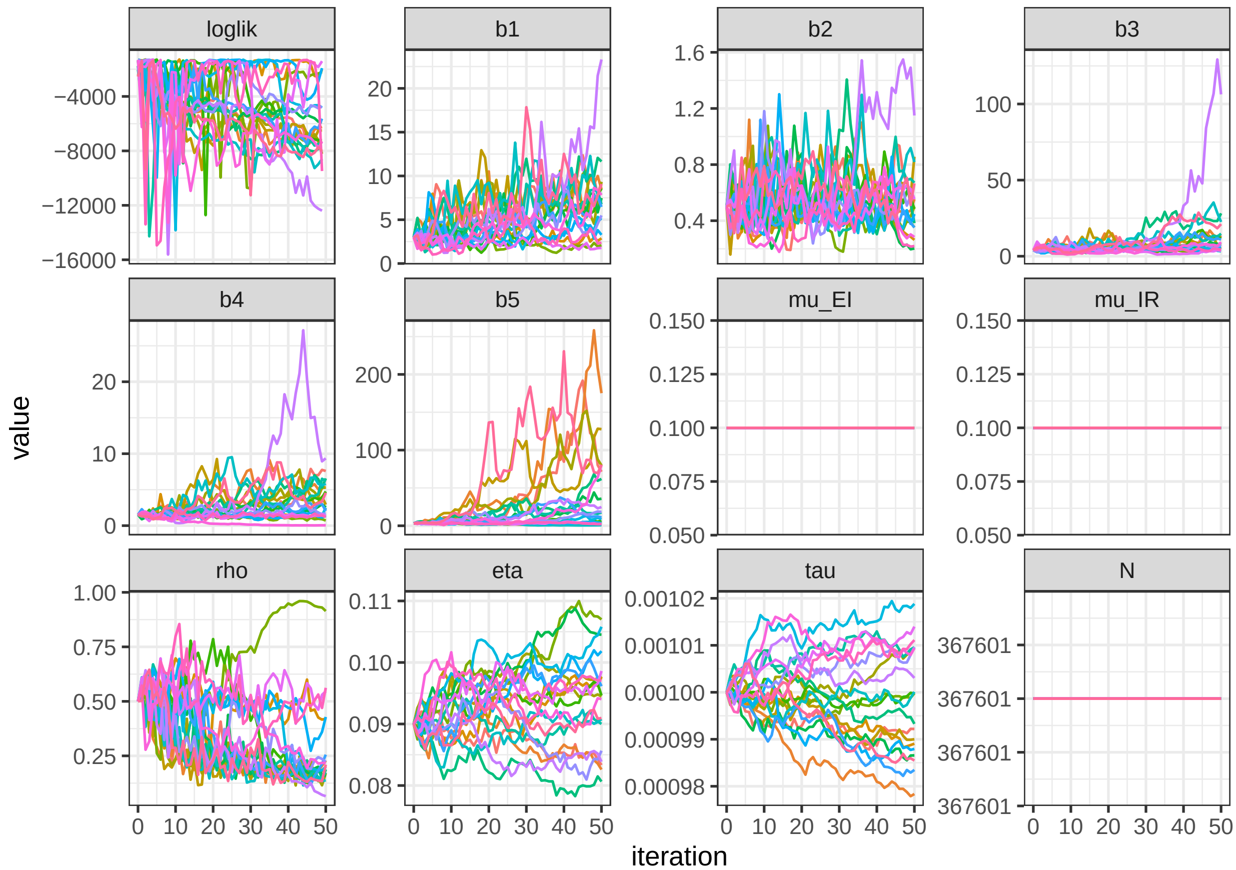

Figure 1: Diagnostic plot for a COVID-19 model

The iterated filtering convergence diagnostics plot in figure 1 comes from a 2021 student project investigating COVID-19. The calculation used 10^3 particles. What is the best interpretation?

A. Everything seems to be working fine. There is a clear consensus from the different searches concerning the highest likelihood that can be found. Therefore, the search is doing a good job of maximization. Occasional searches get lost, such as the purple line with a low likelihood, but that is not a problem.

B. The searches obtain likelihood values spread over thousands of log units. We would like to see consistent convergence within a few log units. We should use more particles and/or more iterations to achieve this.

C. The searches obtain likelihood values spread over thousands of log units. We would like to see consistent convergence within a few log units. We should compare the best likelihoods obtained with simple statistical models, such as an auto-regressive moving average model, to look for evidence of model misspecification.

D. The searches obtain likelihood values spread over thousands of log units. We would like to see consistent convergence within a few log units. We should look at the effective sample size plot for the best fit we have found yet, to see whether there are problems with the particle filtering.

E. All of B, C, and D.

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

A generalized autoregressive conditional heteroskedasticity (GARCH) model has $Y_n = \sigma_n Z_n$ where $Z_n \sim$ i.i.d.$N(0,1)$ and

$$\sigma_n^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{n-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{n-j}^2.$$

For data $y^*_{1:N}$, residuals may be defined by $r_n = Y_n/\hat{\sigma}_n$ where $\hat{\sigma}_n$ is an estimate of σ_n. Suppose that we fit a GARCH model to the log-returns of a financial time series, and we find that the sample ACF of $r_{1:N}$ is consistent with white noise (e.g., 531W24 final project #7). What is the best inference from the residual ACF about the success of the GARCH model for these data?

A. This supports the use of GARCH over ARMA. That is not especially surprising, since it is true for essentially all financial time series, but it is good to check.
B. A fitted ARMA model is also anticipated to have a residual ACF consistent with white noise. The problem with the ARMA model for financial data is not residual autocorrelation.
C. We should also make a normal quantile plot of the residuals. If the residuals are approximately normal then the ACF plot becomes more trustworthy as a test for lack of correlation. If the residuals are far from normal, we should not draw conclusions from the sample ACF.
D. GARCH aims to fix the problem of conditional heteroskedasticity in financial data that ARMA cannot explain. However, fixing this might break the negligible autocorrelation that is critical for the efficient market hypothesis. It is good to see that we can fix conditional heteroskedasticity while remaining compatible with the efficient market hypothesis.

Supporting reasoning:

Q6. Computing with POMP models

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

```
##
## Error: in ‘simulate’: error in building shared-object library from C snippets: in ‘Cbuilder’:
##  compilation error: cannot compile shared-object library
##  ‘/tmp/RtmpFkkeCQ/24104/pomp_4fc43714a7a9ebddf896bbc51635d211.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_4fc43714a7a9ebddf896bbc51635d211.c
## -o /tmp/RtmpFkkeCQ/24104/pomp_4fc43714a7a9ebddf896bbc51635d211.o
## In file included from /home/kingaa/R/x86_64-pc-linux-gnu-library/4.2/pomp/include/pomp.h:9,
##                  from /tmp/RtmpFkkeCQ/24104/pomp_4fc43714a7a9ebddf896bbc51635d211.c:5:
## /tmp/RtmpFkkeCQ/24104/pomp_4fc43714a7a9ebddf896bbc51635d211.c: In function ‘__pomp_rmeasure’:
## /usr/local/apps/R/ubuntu_20.04/4.2.1/lib64/R/include/Rmath.h:333:16: error:
##  too many arguments to function ‘Rf_rnorm
## 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_4fc43714a7a9ebddf896bbc51635d211.so
##  /tmp/RtmpFkkeCQ/24104/pomp_4fc43714a7a9ebddf896bbc51635d211.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: