Quiz 2, STATS/DATASCI 531/631 W25

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

This document produces different random quizzes each time the source code generating it is run. The actual quiz will be a realization generated by this random process, or something similar.

This version lists all the questions currently in the quiz generator. The actual quiz will have one question sampled from each of the 6 question categories.

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 some supporting reasoning.

Q1. Foundations of POMP models

Q8-01.

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.

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

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

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

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

Q8-02.

Which of the following linear Gaussian POMP model have an observable variable Y_n with distribution matching an ARMA(1,1) model? Here, ϵ_n and η_n are Gaussian white noise. X_n is 1-dimensional in (1) and 2-dimensional in (2) and (3).

(A) Only (3)
(B) (1) and (2) but not (3)
(C) (2) and (3) but not (1)

(D) (1) and (3) but not (2)
(E) (1), (2) and (3)

$$\left.\begin{array}{rcl} X_n &=& aX_{n-1}+\epsilon_n \\ Y_n &=& X_n+\eta_n \end{array}\right\} \tag{5}$$

$$\left.\begin{array}{rcl} X_n &=& \left(\begin{array}{cc} a & 1 \\ 0 & 0 \end{array}\right) X_{n-1} + \left(\begin{array}{c} 0 \\ 1 \end{array}\right)\epsilon_n \\ Y_n &=& (1,0)\, X_n + \eta_n \end{array}\right\} \tag{6}$$

$$\left.\begin{array}{rcl} X_n &=& \left(\begin{array}{cc} a & 1 \\ 0 & 0 \end{array}\right) X_{n-1} + \left(\begin{array}{c} 1 \\ b \end{array}\right)\epsilon_n \\ Y_n &=& (1,0)\, X_n \end{array}\right\} \tag{7}$$

Q8-03.

Scientifically, our conclusions should not depend on the units of measurement we use, but we can make errors if we don't 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, ..., 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, ..., 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, ..., 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.

Q8-04.

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

Q2. Likelihood evaluation; the particle filter

Q9-01.

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. Approximately how many minutes will this take, if you distribute the calculation across 4 cores?

A: 50
B: 60
C: 75
D: 120
E: 300

Q9-02.

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 1.4

Q9-03.

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.

Q9-04.

In a particle filter, a particle that is resampled k times is said to be the parent of these k children, The complete ancestry graph of all the particles is an evolutionary tree for the population of particles. Each filtering iteration corresponds to a generation of this population process. Darwinian evolution on populations occurs when individuals with the fittest phenotype reproduce more offspring for the next generation, and these children inherit the parent's genotype subject to mutation.

Particle filter	Darwinian evolution
(1) Prediction step: simulation	(a) Genotype
(2) Filtering step: resampling	(b) Fitness
(3) State value, $X_{n,j}$	(c) Mutation
(4) Measurement density, $f_{Y_n\|X_n}(y_n^*\|X_{n,j})$	(d) Reproduction

What is the pairing between the particle filter concepts (numbers 1–4) and the analogous evolutionary concepts (letters a–d).

A: (1a) (2b) (3c) (4d)
B: (1b) (2a) (3c) (4d)
C: (1c) (2d) (3a) (4b)
D: (1d) (2c) (3a) (4b)
E: A combination not listed above.

Q3. Likelihood maximization; iterated filtering

Q10-01.

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 deviations 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.

Q10-02.

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.

Q10-03.

The iterated filtering convergence diagnostics plot shown above come from a student project. What is the best interpretation?

A: Everything seems to be working fine. The likelihood is climbing. The replicated searches are giving consistent runs. The spread of convergence points for σ_ν and H_0 indicates weak identifability, which is a statistical fact worth noticing but not a weakness of the model.

B: The consistently climbing likelihood is promising, but the failure of σ_ν and H_0 to converge needs attention. Additional searching is needed, experimenting with **larger** values of the random walk perturbation standard deviation for these parameters to make sure the parameter space is properly searched.

C: The consistently climbing likelihood is promising, but the failure of σ_ν and H_0 to converge needs attention. Additional searching is needed, experimenting with **smaller** values of the random walk perturbation standard deviation for these parameters to make sure the parameter space is properly searched.

D: The consistently climbing likelihood is promising, but the failure of σ_ν and H_0 to converge needs attention. This indicates weak identifiability which cannot be solved by improving the searching algorithm. Instead,

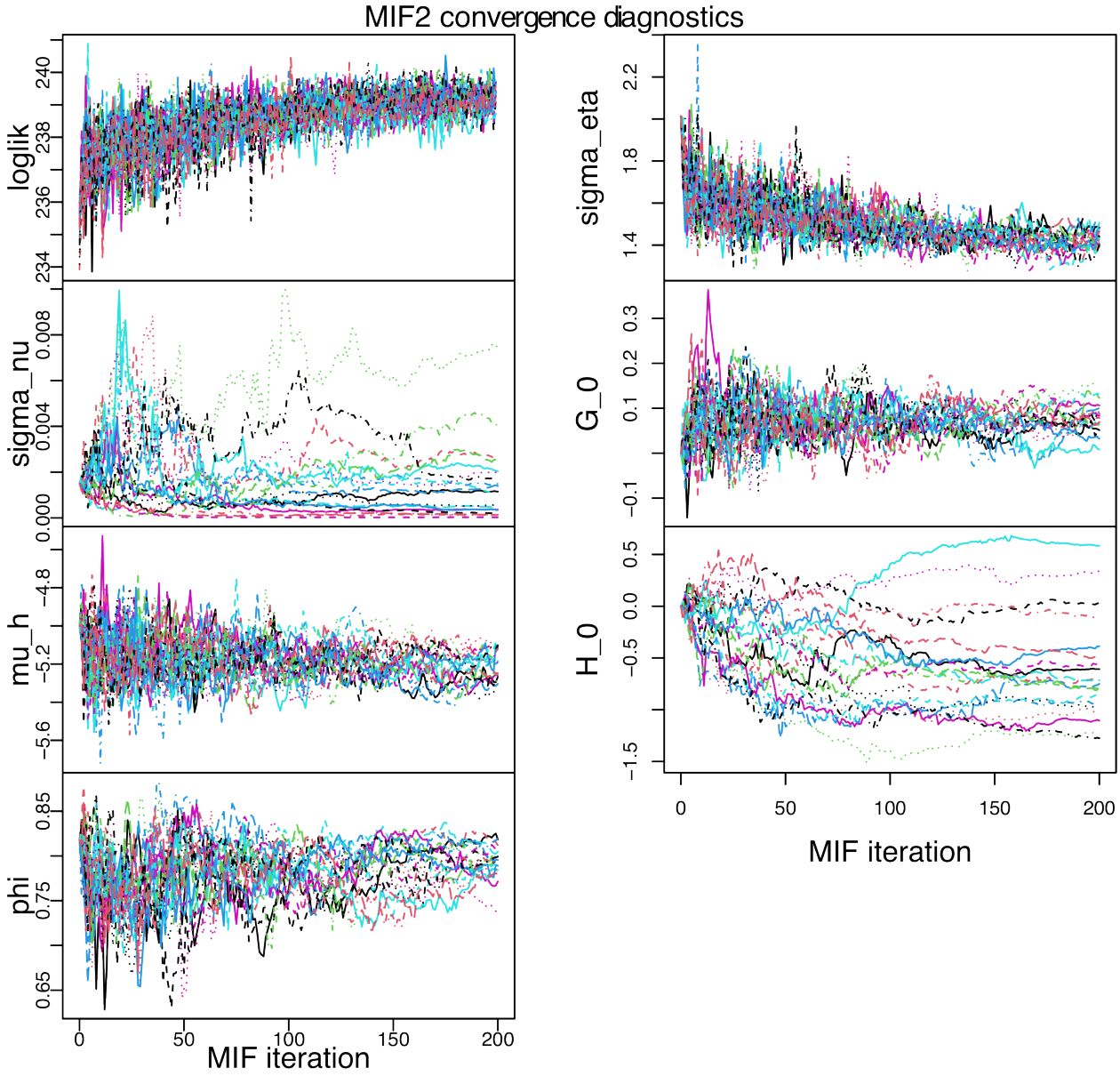

Figure 1: Diagnostic plot

we should change the model, or fix one or more parameters at scientifically plausible values, to resolve the identifiability issue before proceeding.

E: Although the log likelihood seems to be climbing during the search, until the convergence problems with σ_ν and H_0 have been addressed we should not be confident about the successful optimization of the likelihood function or the other parameter estimates.

Q10-04.

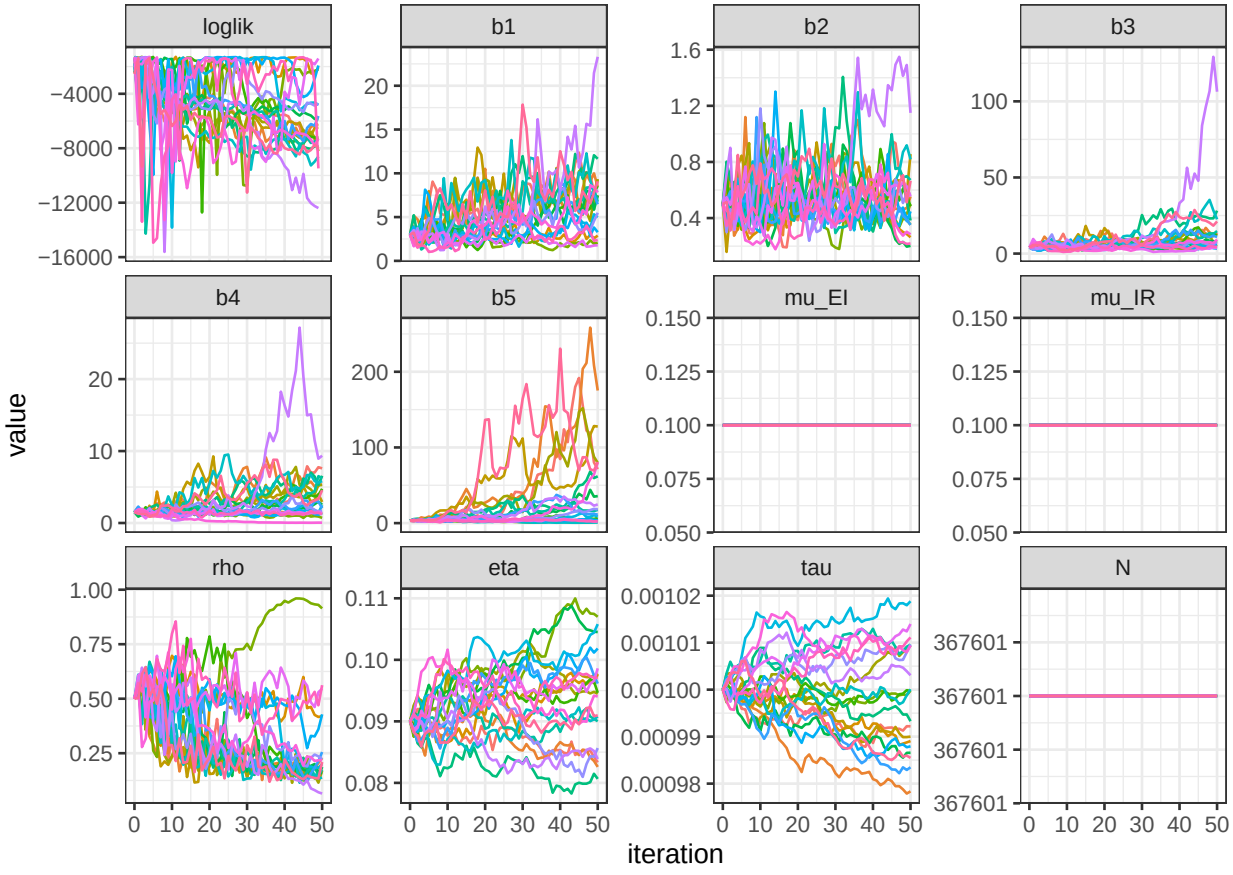

Figure 2: Diagnostic plot for a COVID-19 model

The iterated filtering convergence diagnostics plot in figure 2 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.

Q4. Data analysis: epidemiological models

Q11-01.

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).

Q11-02.

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.

Q11-03. You fit an SEIR model to case reports of an immunizing disease from 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.

B: The discrepancy shows that something is substantially wrong with the model. Extra biological detail must 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.

Q5. Data analysis: financial models

Q12-01.

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.

Q12-02.

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 $\sigma_n^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{n-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{n-j}^2$. There are many extensions to GARCH implemented by various R packages. When comparing models by likelihood or AIC, care is required since packages do not always use standard definitions. What is the most reasonable interpretation of this table?

```
for (i in 1:p) {
  for (j in 1:q) {
    fit_garch <- tseries::garch(log_returns, order = c(i, j))
    garch_table[i, j] <- tseries:::logLik.garch(fit_garch)
  }
}
```

	q1	q2	q3	q4
p1	2646.277	2642.919	2620.280	2616.151
p2	2644.417	2625.417	2622.460	2616.427
p3	2641.804	2637.538	2625.953	2625.740
p4	2639.728	2629.869	2629.969	2628.345

A. The positive values of the log-likelihood are implausible. Perhaps the software actually reports the

negative log-likelihood since many optimizers are designed to minimize rather than maximize.
B. The models are nested and so a larger model should mathematically have a larger likelihood. In this table, the larger model usually has lower likelihood, so optimization is problematic.
C. This table would make more sense if `logLik` in fact returns an AIC value. The preferred model is $(p, q) = (1, 4)$.
D. The preferred model is $(p, q) = (1, 1)$ since it is both the simplest model and the one with the highest log-likelihood.
E. `tseries::garch` produces something that is not the likelihood of $y_{1:N}$ or the AIC, and so we cannot readily compare it between models.

Q6. Computing with POMP models

Q13-01.

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.

Q13-02. 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.
