Quiz 1, STATS/DATASCI 531/631 W25

In class on 2/17, 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. Stationarity and unit roots.

Q1-01.

Suppose that a dataset $y^*_{1:N}$ is well described by the statistical model

$$Y_n = a + bn + \epsilon_n,$$

where ϵ_n is a Gaussian ARMA process and $b \neq 0$. Which of the following is the best approach to time series modeling of $y^*_{1:N}$?

A. The data are best modeled as non-stationary, so we should take differences. The differenced data are well described by a stationary ARMA model.

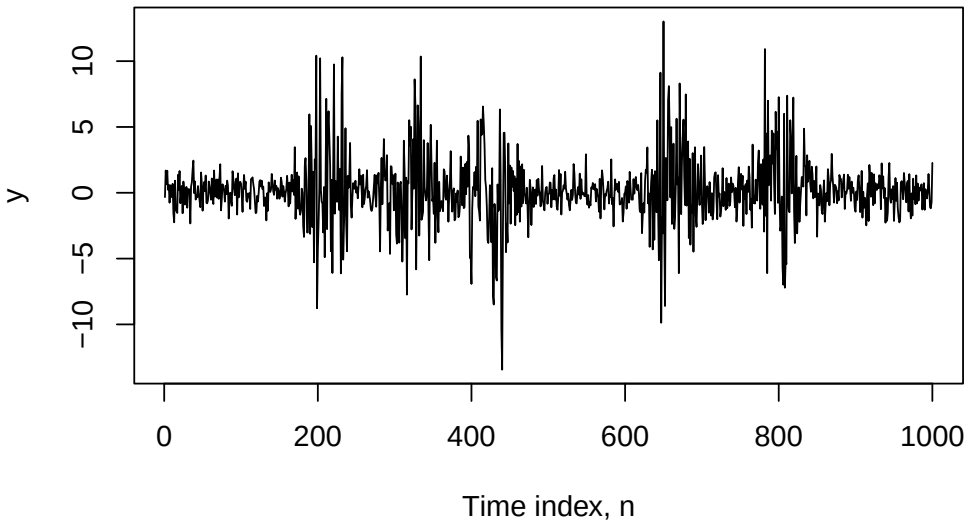
B. The data are best modeled as non-stationary, and we should use a trend plus ARMA noise model.

C. The data are best modeled as non-stationary. It does not matter if we difference or model as trend plus ARMA noise since these are both linear time series models which become equivalent when we estimate their parameters from the data.

D. We should be cautious about doing any of A, B or C because the data may have nonstationary sample variance in which case it may require a transformation before it is appropriate to fit any ARMA model.

Solution. B.
It does matter whether we take differences. For example, the differenced model is non-causal (has an MA root on the unit circle) so cannot be fitted by usual ARMA methods. D is not relevant since we are told that the data are well described by a model that rules out this possibility.

Q1-02.



Consider the time series plotted above. Which of the below is the most accurate statement about stationarity?

A. The plot shows that the data are clearly non-stationary. We could make a formal hypothesis test to confirm that, but it would not be insightful. To describe the data using a statistical model, we will need to develop a model with non-constant variance.

B. The sample variance is evidently different in different time intervals. However, we should not conclude that the underlying data generating mechanism is non-stationary before making a formal statistical test of equality of variances between the time regions that have lower sample variance and the regions that have higher sample variance. Visual impressions without a formal hypothesis test can be deceptive.

C. A model with randomly changing variance looks appropriate for these data. Since the variance for such a model is time-varying, the model must be non-stationary.

D. A model with randomly changing variance looks appropriate for these data. Despite the variance for such a model being time-varying, the model is stationary.

E. The sample variance is evidently different in different time intervals. An appropriate next step to investigate stationary would be to plot the sample autocorrelation function for different intervals to see if the dependence between time points is also time-varying.

Solution. D.
This is a subtle question, so let's discuss each option. The plotted time series is a realization of a stationary model:

```
N <- 1000
sd1 <- rep(1,N)
events <- runif(N) < 5/N
sigma <- 20
amplitude <- 10
sd2 <- sd1 + filter(events,
  dnorm(seq(from=-2.5*sigma,to=2.5*sigma,length=5*sigma),sd=sigma)*sigma*amplitude,
  circular=T)
Y <- rnorm(n=N,mean=0,sd=sd2)
```

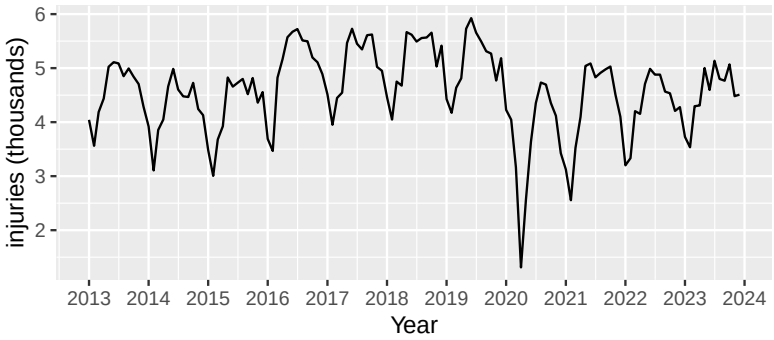
Hopefully, this suggests that it should not be clearly non-stationarity, ruling out A.

C and D contain a value judgement, "looks appropriate" which is hard to quantify but is (in this case) correct! "Randomly changing variance" is an informal description of a model with stochastic conditional variance. The sample variance estimates the variance conditional on the realization of the conditional variance. The actual variance is an expectation over posible values of the conditional variance. So, between C and D, only

D can be correct.

B and E acknowledge the variation in sample variance but do not provide useful ways to assess whether this variable sample variance comes about via a stationary stochastic conditional variance model or via a non-stationary model. In particular, if you follow the advice in B you would conclude that an appropriate model should be non-stationary, which would be incorrect in this case. In financial applications, it is common to fit stationary models to time series of financial returns that often resemble this model. For this particular case, E would not show significant autocorrelation in any time interval, but the same reasoning applies.

Q1-03.



Above are monthly injuries from motor vehicle collisions in New York City. An augmented Dickey-Fuller test, `tseries::adf.test(injuries)`, gives a p-value of 0.01. Which is the best way to proceed:

A: The time plot indicates a non-constant mean function describing a major dip due to the COVID-19 pandemic and an increasing trend at other times. The ADF test does not support or refute that model.

B: The ADF test suggests the series is stationary, supporting a decision to fit a SARMA model.

C: The ADF test suggests the series is non-stationary; it should be differenced before fitting a SARMA.

D: The ADF test indicates that the series is non-stationary, supporting the use of a non-constant mean function to describe a major dip due to the COVID-19 pandemic and an increasing trend at other times.

Solution. A.
The ADF test has a null hypothesis of a unit root linear model and an alternative of a stationary linear model, so neither of these describes a nonlinear trend. Here, the role of the COVID-19 pandemic is large enough that it does not make much sense to build a model that omits it, or describes it as a large perturbation of a stationary process. The conventional interpretation of the ADF test is option B (we reject the unit root hypothesis, and so we are invited to fit a stationary model). Here, a nonlinear trend (which is neither a unit root nor a stationary model) makes sense.

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.

Solution. C.
A test on simulated data, when the truth is known, can help pin down an optimization problem. All the issues other than C can cause inference problems, but likely would cause similar problems on simulated data.

When there is a reproducible and stable phenomenon of decreasing likelihood, it generally indicates that the unperturbed model is a worse fit to the data than the perturbed model. Recall that the likelihood calculated by iterated filtering at each iteration corresponds to the model with perturbed parameters rather than the actual postulated model with fixed parameters. If the perturbed model has higher likelihood, it may mean that the data are asking to have time-varying parameters. It may also be a signature of any other weakness in the model that can be somewhat accommodated by perturbing the parameters.

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.

Solution. B.
If the researchers calculate a sliced likelihood through the MLE and tell you it is a profile, but you are concerned they might have constructed a slice by mistake, it is hard to know without looking at the code. (A) is the proper construction of a profile if the points are maximizations over the remaining parameters for a range of fixed values of rho. However, if the code does not involve maximization over other parameters at each value of ρ, it cannot be a proper profile. It could be a slice accidentally explained to be a profile, and with a confidence interval constructed as if it were a profile.

If there is only one unknown parameter then a slice and a profile are the same thing, and no maximization is required. This is an unusual situation; there is usually more than one unknown parameter.

Q10-03.

The iterated filtering convergence diagnostics in figure 1 come from a student project investigating the market value of Gamestop. 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, 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.

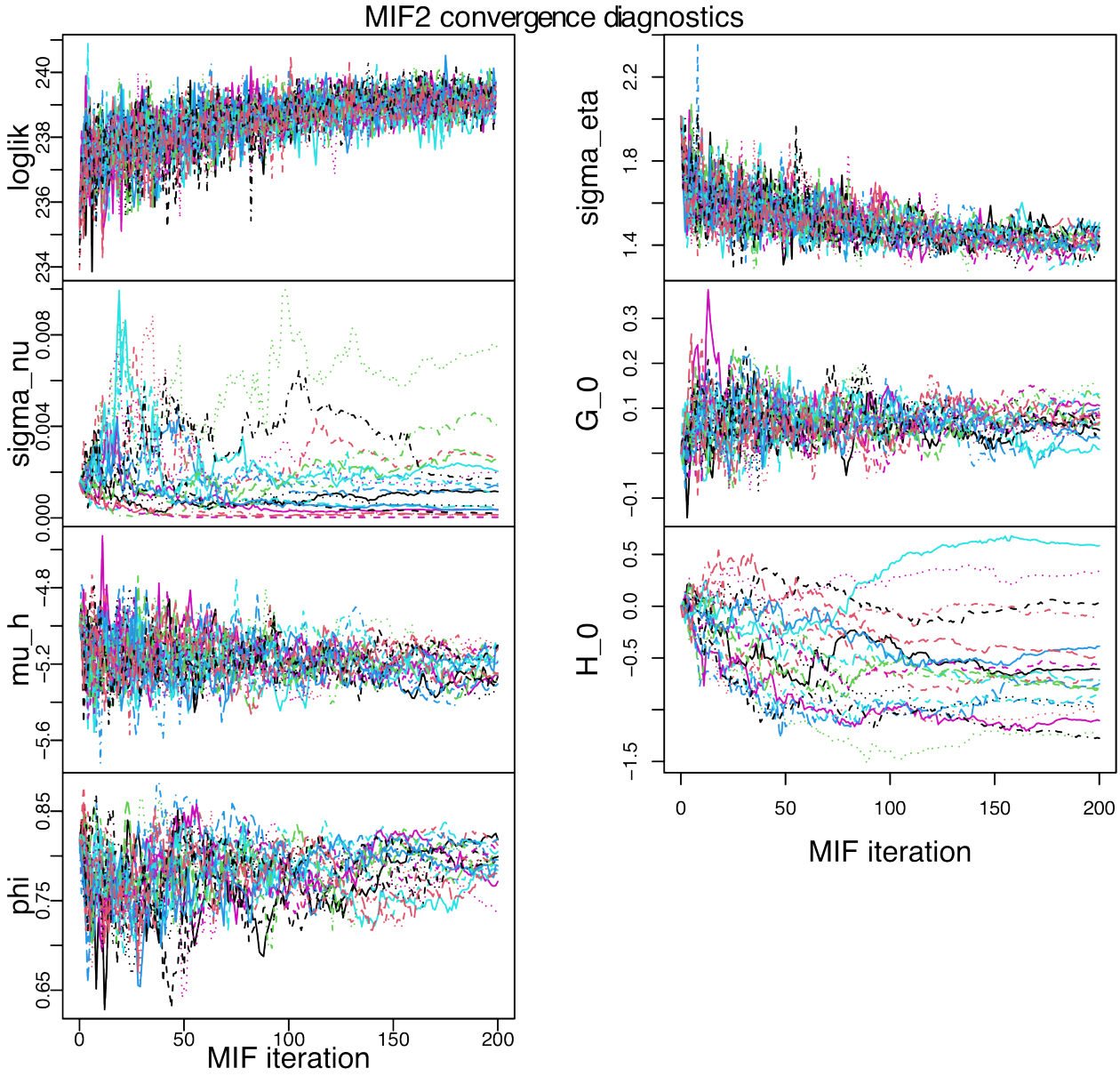


Figure 1: Iterated filtering diagnostic plot

Solution. A.
All searches are finding parameters with consistent likelihood. The discrepancies of a few log likelihood units put the parameter values within statistical uncertainty according to Wilks's Theorem. Therefore, the spread in the parameter estimates reflects uncertainty about the parameter given the data, rather than a lack of convergence.

That perspective suggests that the goal of the Monte Carlo optimizer is to get close to the MLE, measured by likelihood, rather than to obtain it exactly. Independent Mont Carlo searches can be combined via a profile likelihood to get a more exact point estimate and a confidence interval.

Wide confidence intervals, also called weak identifability, are not necessarily a problem for the scientific investigation. Some parameters may be imprecisely estimable, while others can be obtained more precisely, and part of the analysis is to find which is in each category. It may also be of interest to investigate what extra precision can be obtained on one parameter by making assumptions about the value of another, as in D, but this is not mandatory for proper inference.

Overall, the convergence plots here look good. The plots show that the seaches are all started from a single high likelihood starting point. Now this has been done successfully, a natural next step would be to start some searches from more diverse starting points to look for any global features missed by this local search.

Q10-04.

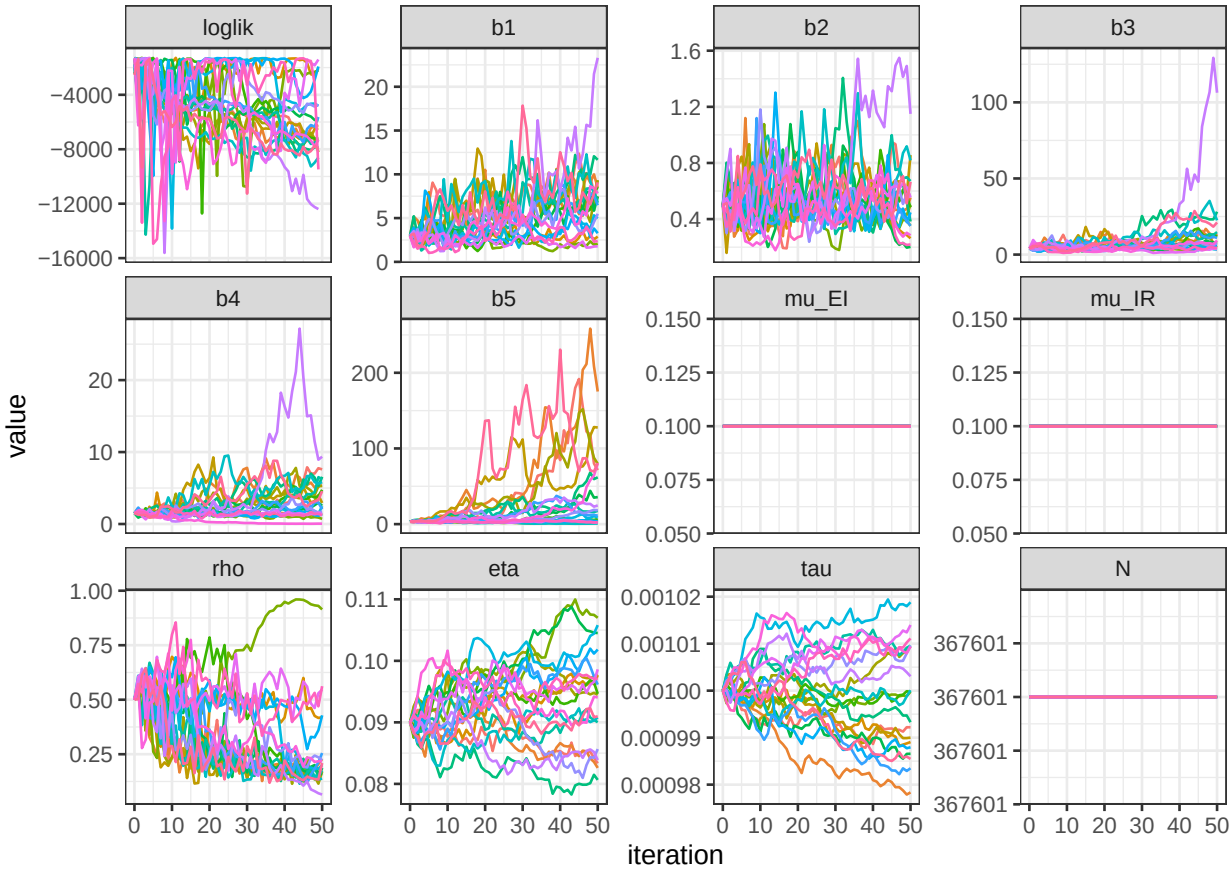


Figure 2: Diagnostic plot for a COVID-19 model

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.

Solution. E.
This project team were able to show evidence of adequate global maximization for their model, but their maximized likelihood was 47 log units lower than ARMA model. The wide spread in likelihood, thousands of log units, shown in this convergence plot suggests that the numerics are not working smoothly. This could mean that more particles are needed: 10^3 particles is relatively low for a particle filter. However, if the model fit is not great (as revealed by comparison against a benchmark) this makes the filtering harder as well as less scientifically satisfactory. If the model is fitting substantially below ARMA benchmarks, it is worth considering some extra time on model development. Identifying time points with low effective sample size can help to identify which parts of the data are problemtic for the model to explain.

In this case, the clearest clue happens to come from the benchmark ARMA comparison. The model would have fitted better with overdispersion on the latent process. If the model has a substantial flaw, this can make filtering hard but it is unproductive to bandaid the problem by using massive computational effort. It is better to fix the model.

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).

Solution. D.
Why not A? Likelihoods of different models for the same data can be compared. Likelihood ratio tests using Wilks's theorem specifically require nested models, but in other contexts (such as AIC and the Neyman-Pearson lemma) the models being compared by likelihood do not need to have any particular relationship.

Why not B? Likelihood ratios have statistical meaning, which corresponds to differences of log likelihoods. The likelihood is a dimensional quantity, whereas the likelihood ratio is dimensionless. The units used correspond to a scientifically arbitrary additive constant to the log likelihood, which disappears after taking differences.

Why not C? If our only goal were to find a predictive model, then (C) could be a reasonable position. Usually, we want to find a model that also has interpretable structure, leading to understanding of the system or estimating the effect of interventions. A simple regression model cannot do those things, even if it fits a bit better. If the mechanistic model fits much worse than simple alternatives, it is not providing a reasonable explanation of the data, suggesting that there may be important things missing from the model specification.

Quite likely, with some persistence, a mechanistic specification will beat a simple off-the-shelf statistical model.

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.

Solution. B.
All the possibilities are worth consideration. However, adding noise in rates to give flexibility in mean-variance relationships is commonly an important part of developing a stochastic model. The simple compartment model interpretation of a ODE as a Markov chain is determined by the rates and therefore does not have free parameters to describe variance. There is some variance inherent in the Markov chain (demographic stochasticity) but additional variability may be needed. It will be hard to investigate the other possibilities if the model has not been given enough stochasticity to explain the variability in the data, so including overdispersion should be an early step. Note that overdispersion can be included in both the process model and the measurement model.

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.

Solution. A.
Transferring parameter estimates between scales is hard. An example is the difficulty of reconciling micro and macro economics. It is generally not possible to guarantee that a parameter means exactly the same thing in models at different scales. (A) acknowledges this. The other answers, in various ways, assume that there should be a single parameter value that describes the system at all scales. There is some merit also to (D), since it is reasonable to try to gain biological understanding by investigating why the fitted model is successful at explaining the data. However, this is an observational study and so we should be cautious of making a causal interpretation of models fitted to data due to the possiblity of confounding. Only (A) addresses this concern.

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.

Solution. B.
It would be surprising if substantial sample autocorrelation appeared in the residuals of a GARCH model, at least for a highly traded financial instrument. This would violate the efficient market hypothesis. But that observation is just as true for an i.i.d. white noise model. The reason to prefer GARCH over an i.i.d. white noise model is to explain the conditional heteroskedasticity, but the ACF does not reveal whether or not that is successful.

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.

Solution. E.

```
?tseries::logLik.garch
```

reveals that

> 'logLik' returns the log-likelihood value of the GARCH(p, q) model
> represented by 'object' evaluated at the estimated coefficients.
> It is assumed that first max(p, q) values are fixed.

Therefore, the log-likelihood for fiting GARCH(p,q) corresponds only to $y^*_{(\max(p,q)+1):N}$. The violations of nesting occur because different amounts of data are used for differeng values of $\max(p, q)$. Therefore, we cannot easily compare likelihoods or AIC values.

Q12-03.

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 $\omega_n \sim N[0, \sigma^2_\omega]$. 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 3. What is the best conclusion from this information?

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.

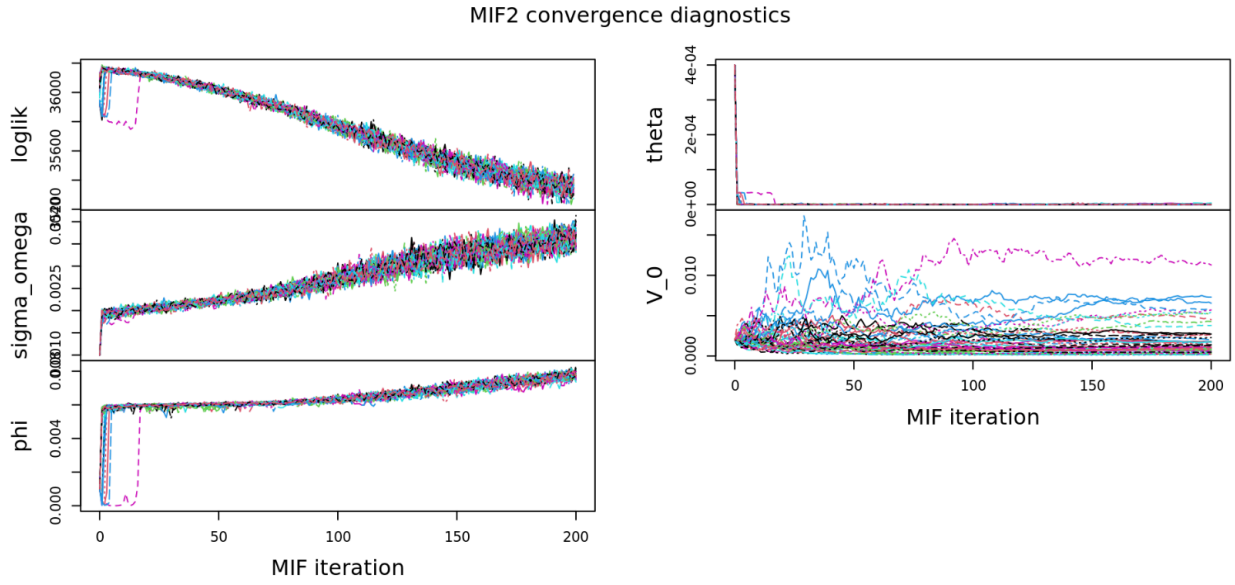


Figure 3: Diagnostic plot for fitting the Heston model

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.

Solution. A.
The structure of the particle filter makes it hard to obtain an artificially high likelihood by cheating. Mathematically, the best expected log-likelihood is obtained by a one-step forecast distribution matching the true prediction distribution, assuming the model is correct. (This is just another way of stating the property that the expected log-likelihood is highest under the true model.)

If `dmeasure` is not in fact a density then artificially high likelihoods are possible, but in most models (including this one) the measurement model is a call to a basic R function known to be a density (i.e., integrating to 1).

Inspection of the source code, available online, reveals that the authors made a mistake in implementing `rprocess`. Specifically, the `rprocess` line

```
V = theta*(1 - phi) + phi*sqrt(V) + sqrt(V)*omega;
```

should be

```
V = theta*(1 - phi) + phi*V + sqrt(V)*omega;
```

Thus, their model is not exactly the model they thought they were implementing, leading to incorrect interpretations of their results. Nevertheless, this error turns out to give rise to a model which fits the data very well. This happy accident suggests that a key to modeling the data may be to use a longer-tailed distribution than normal for the returns.

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.

Solution. A.
The code producing the error is below. Within C snippets, the C versions of R distribution functions are available but they have slightly different syntax from their more familiar R children. A complete reference guide to R's C interface is available as part of R's documentation. In particular, the C form of R's distribution functions is useful for writing C snippets.

```
sir4 <- simulate(
  sir1,
  statenames=c("S","I","R","cases","W"),
  paramnames=c(
    "gamma","mu","iota",
    "beta1","beta_sd","pop","rho",
    "S_0","I_0","R_0"
  ),
  rmeasure=Csnippet("
    double mean, sd;
    double rep;
    mean = cases*rho;
    sd = sqrt(cases*rho*(1-rho));
    rep = nearbyint(rnorm(1,mean,sd));
    reports = (rep > 0) ? rep : 0;"
  )
)
```

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.

Solution. E.
The error message was produced by the code below. `pomp` passes on the C compiler error message for you to inspect. Note the missing semicolon in the next-to-last line.

```
sir1 <- sir()
sir2 <- pomp(
  sir1,
  statenames=c("S","I","R","cases","W"),
  paramnames=c(
    "gamma","mu","iota",
    "beta1","beta_sd","pop","rho",
    "S_0","I_0","R_0"
  ),
  rinit=Csnippet("
  double m = pop/(S_0+I_0+R_0);
  S = nearbyint(m*S_0);
  I = nearbyint(m*I_0);
  R = nearbyint(m*R_0);
  cases = 0
  W = 0;"
  )
)
```

Q13-03.

A useful way to check statistical methodology is to apply an inference method to a collection of simulated datasets from the fitted model with the estimated parameter values (say, the maximum likelihood estimate,

MLE). This is sometimes called a "parametric bootstrap". Suppose that we carry out this check for a POMP data analysis, using plug-and-play inference methodology such as iterated filtering, and we find that the re-estimated parameters from inference on the simulated data are close to the MLE. What can we infer about the correctness of our inference.

A. This is a strong check that both the model and the methodology are correctly implemented. Except for some rare special cases, an error in either one of these will lead the check to fail.

B. This checks the implementation of the inference methodology but not the model. Even if the model is implemented wrongly, the check will still show us whether the inference methodology is correct.

C. This checks the implementation of the model but not the inference methodology. As long as the model is implemented correctly, any reasonable inference methodology should pass the check successfully.

D. This is not a strong check of either the model or the methodology. It shows self-consistency but that is different from showing accuracy.

Solution. B.

This is a useful property to bear in mind when debugging statistical analysis carried out using plug-and-play methodology. By definition, the inference methodology defines the model via a simulator, and presumably the same simulator is used for inference as for the simulation used to test the inference. Thus, the parametric bootstrap exercise tests the inference methodology but not the correctness of the model implementation; errors in the latter will apply in the same way to both the simulation and the inference, so cannot show up as a mismatch between inferred parameters and re-estimated parameters.

Errors in `rprocess` for a POMP model are hard to debug for this reason. Best practice is to present the Csnippet right next to the math representation, and use the same notation for both, so that the visual match is evident.

It is a good idea to carry out a parametric bootstrap despite this limitation.

Q2. Calculations for ARMA models

Q2-01.

Let $Y_n = \phi Y_{n-1} + \epsilon_n$ for $n = 1, 2, \ldots$ with $\epsilon_n \sim \text{iid}N[0, \sigma^2]$ and $Y_0 = 0$. The covariance of Y_n with Y_{n+k} for $k \geq 0$ is

A. $\sigma^2\phi^k/(1-\phi^2)$

B. $\sigma^2\phi^{2k}/(1-\phi^2)$

C. $\sigma^2\phi^k/(1-\phi)$

D. $\sigma^2\phi^{2k}/(1-\phi)$

E. None of the above.

Solution. E.
This model is not started in its stationary distribution, leading to a covariance that is not shift invariant. The exact calculation is not needed, but it is as follows.

$$\begin{aligned}
\text{Cov}(Y_n, Y_k) &= \text{Cov}\left(\textstyle\sum_{i=1}^{n} \phi^{n-i}\epsilon_i, \sum_{j=1}^{n+k} \phi^{n+k-j}\epsilon_j\right) \\
&= \textstyle\sum_{i=1}^{n}\sum_{j=1}^{n+k} \phi^{2n+k-(i+j)}\text{Cov}(\epsilon_i, \epsilon_j) \\
&= \sigma^2\phi^k \textstyle\sum_{i=1}^{n} \phi^{2(n-i)} \\
&= \sigma^2\phi^k \textstyle\sum_{i=0}^{n-1} \phi^{2i} \\
&= \sigma^2\phi^k(1-\phi^{2n})\big/(1-\phi^2)
\end{aligned}$$

If you didn't see this, you may feel you were tricked. However, it is a common mistake in practical data analysis to pay insufficient attention to initial conditions, so it is worth bringing this to your attention.

Q2-02.

Let Y_n be an ARMA model solving the difference equation

$$Y_n = (1/4)Y_{n-2} + \epsilon_n + (1/2)\epsilon_{n-1}.$$

This is equivalent to which of the following:

A. $Y_n = (1/2)Y_{n-1} + \epsilon_n$

B. $Y_n = -(1/2)Y_{n-1} + \epsilon_n$

C. $Y_n = (1/2)Y_{n-2} - (1/16)Y_{n-4} + \epsilon_n + \epsilon_{n-1} + (1/4)\epsilon_{n-2}$

D. $Y_n = -(1/2)Y_{n-2} - (1/16)Y_{n-4} + \epsilon_n + \epsilon_{n-1} + (1/4)\epsilon_{n-2}$

E. None of the above

Solution. A.
Writing the model in terms of the lag operator, L, we get

$$(1 - (1/2)L)(1 + (1/2)L)Y_n = (1 + (1/2)L)\epsilon_n.$$

Canceling out a factor of $(1 + (1/2)L)$, we obtain

$$(1 - (1/2)L)Y_n = \epsilon_n.$$

Q2-03.

Is it possible for an $AR(2)$ model to have a finite moving average representation, so that it is equivalent to some $MA(q)$ model for $q < \infty$?

A. No. Any moving average representation of any $AR(2)$ model is $MA(\infty)$

B. Yes. Although it is not true for any $AR(2)$ process, it is possible to find particular choices of the autoregressive coefficients, p_1 and p_2, that lead to a finite $MA(q)$ representation.

C. It is not possible for any real-valued p_1 and p_2, but it is possible if you permit p_1 and p_2 to be complex-valued.

Solution. A.
For any $AR(p)$ model with $p \geq 1$, an $MA(q)$ representation always has $q = \infty$. One way to see this is to argue by contradiction, by supposing there is a value of $q < \infty$. Then, setting $\phi(x)$ and $\psi(x)$ as the AR and MA polynomials, we can write

$$\frac{1}{\phi(B)} = \psi(B)$$

Thus, $\phi(x)\psi(x) = 1$. But $\phi(x)\psi(x)$ has a nonzero x^{p+q} coefficient so it cannot equal 1. This argument applies for either real-valued or complex-valued coefficients.

Q3. Likelihood-based inference for ARMA models

Q3-01.

The following table of AIC values results from fitting ARMA(p,q) models to a time series $y_{1:415}$ where y_n is the time, in miliseconds, between the nth and $(n+1)$th firing event for a monkey neuron. The experimental details are irrelevant here. You are asked to check how many adjacent pairs of AIC values in this table are inconsistent, such that they could mathematically arise only from a numerical error? Adjacent pairs of models are those directly above or below or left or right of each other in the table.

	MA0	MA1	MA2	MA3
AR0	3966.0	3961.5	3962.7	3964.7
AR1	3961.1	3962.6	3964.6	3966.6
AR2	3962.7	3960.5	3959.8	3961.7
AR3	3964.6	3965.5	3962.6	3968.4

A: 0, so the table is mathematically plausible.

B: 1

C: 2

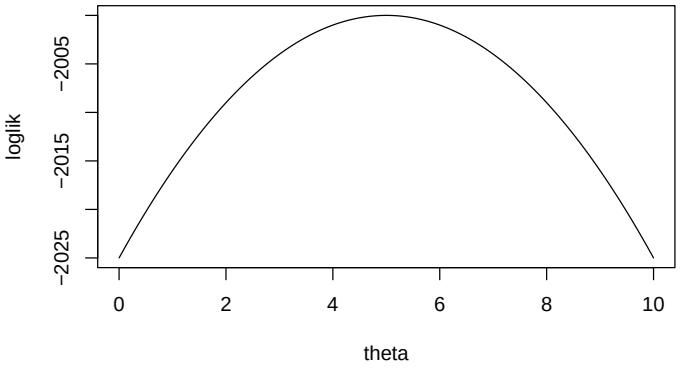
D: 3

E: 4 or more

Solution. E.
Adding one parameter in a nested model cannot decrease the maximized log-likelihood, so it can increase the AIC by at most 2 units. Adjacent pairs $\{(p,q),(p',q')\}$ inconsistent with this are $\{(3,2),(3,3)\}$, $\{(2,3),(3,3)\}$, $\{(2,1),(3,1)\}$, $\{(2,2),(3,2)\}$.

Q3-02.

The R function `arima()` provides standard errors calculated using observed Fisher information. This question tests your understanding of what that means. Suppose a parametric model has a single parameter, θ, and the log-likelihood function when fitting this model to dataset is as follows:



What is the observed Fisher information (I_{obs}) for θ?

Hint 1. The observed Fisher information is accumulated over the whole dataset, not calculated per observation, so we don't have to know the number of observations, N.

Hint 2. Observations in time series models are usually not independent. Thus, the log-likelihood is not the sum of the log-likelihood for each observation. Its calculation will involve consideration of the dependence, and usually the job of calculating the log-likelihood is left to a computer.

Hint 3. The usual variance estimate for the maximum likelihood estimate, $\hat{\theta}$, is $\mathrm{Var}(\hat{\theta}) \approx 1/I_{obs}$.

A: $I_{obs} = 2$

B: $I_{obs} = 1$

C: $I_{obs} = 1/2$

D: $I_{obs} = 1/4$

E: None of the above

Solution. A.
The log-likelihood here is a quadratic function. We can see by inspection that this quadratic is given by

$$\ell(\theta) = -2000 - (\theta - 5)^2.$$

The observed Fisher information is the negative of the second derivative of the log-likelihood at the MLE, so $I_{obs} = 2$. Thus, the standard error is $1/\sqrt{2} = 0.707$

Q3-03.

```
##
## Call:
## arima(x = huron_level, order = c(2, 0, 1))
##
## Coefficients:
##          ar1     ar2     ma1  intercept
##       0.3388  0.4092  0.6320   176.4821
## s.e.  0.4646  0.4132  0.4262     0.1039
##
## sigma^2 estimated as 0.04479:  log likelihood = 21.42,  aic = -32.84
```

```
##
## Call:
## arima(x = huron_level, order = c(2, 0, 2))
##
## Coefficients:
##           ar1     ar2     ma1     ma2  intercept
##       -0.1223  0.7646  1.1310  0.1310   176.4815
## s.e.   0.0682  0.0550  0.1084  0.1004     0.1004
##
## sigma^2 estimated as 0.04364:  log likelihood = 22.64,  aic = -33.28
```

The R output above uses `stats::arima` to fit ARMA(2,1) and ARMA(2,2) models to the January level (in meters above sea level) of Lake Huron from 1860 to 2024. Residual diagnostics (not shown) show no major violation of model assumptions. We aim to choose one of these as a null hypothesis of no trend for later comparison with models including a trend.

Which is the best conclusion from the available evidence:

A: The ARMA(2,2) model has a lower AIC so it should be preferred.

B: We cannot reject the null hypothesis of ARMA(2,1) since the ARMA(2,2) model has a likelihood less than 1.92 log units higher than ARMA(2,1). Since there is not sufficient evidence to the contrary, it is better to select the simpler ARMA(2,1) model.

C: Since the comparison of AIC values and the likelihood ratio test come to different conclusions in this case, it is more-or-less equally reasonable to use either model.

D: When the results are borderline, numerical errors in the `stats::arima` optimization may become relevant. We should check using optimization searches from multiple starting points in parameter space, for example, using `arima2::arima`.

Solution. D.
All the answers are fairly reasonable here! Perhaps the most unreasonable thing would be to be sure there's only one reasonable answer.

However, a more careful optimization using `arima2::arima` shows us that ARMA(2,1) actually has a higher AIC than ARMA(2,2) so all lines of evidence suggest ARMA(2,1) is a better choice. The differences are small, so the choice is unlikely to be highly consequential.

```
##
## Call:
## arima2::arima(x = huron_level, order = c(2, 0, 1), max_iters = 200, max_repeats = 20)
##
## Coefficients:
##          ar1     ar2     ma1  intercept
##       -0.0674  0.7781  1.0000   176.4860
## s.e.   0.0517  0.0518  0.0541     0.1095
##
## sigma^2 estimated as 0.04401:  log likelihood = 21.82,  aic = -33.63
```

```
##
## Call:
## arima2::arima(x = huron_level, order = c(2, 0, 2), max_iters = 200, max_repeats = 20)
##
## Coefficients:
##          ar1     ar2     ma1     ma2  intercept
##       -0.1223  0.7646  1.1310  0.1310   176.4815
## s.e.   0.0682  0.0550  0.1084  0.1004     0.1004
##
## sigma^2 estimated as 0.04364:  log likelihood = 22.64,  aic = -33.28
```

Sometimes the multiple starts used by `arima2::arima` make a difference, sometimes the results from `stats::arima` are unchanged. In this case, it happens to make a difference.

Q3-04.

Suppose model M_0 is nested within a larger model M_1 which has one additional parameter. Suppose that the AIC for M_1 is 0.5 units lower than the AIC for M_0. Which of the following is a correct expression for the p-value of a likelihood ratio test for M_1 against the null hypothesis M_0, supposing that a Wilks approximation is accurate? Here, χ^2_1 is a chi-square random variable on 1 degree of freedom.

A: $P(\chi^2_1 > 0.5)$

B: $P(\chi^2_1 > 1)$

C: $P(\chi^2_1 > 1.5)$

D: $P(\chi^2_1 > 2)$

E: $P(\chi^2_1 > 2.5)$

F: $P(\chi^2_1 > 3)$

Solution. E.
The AIC for each model $k \in \{0, 1\}$ is $AIC_k = -2\ell_k + 2_{D_k}$, where ℓ_k is the log-likelihood for M_k and D_k is the number of parameters. Thus, $AIC_0 - AIC_1 = 2(\ell_1 - \ell_0) - 2 = 0.5$, and so $ 2(_1- _0)=2.5$. Under M_0, according to Wilks' approximation,

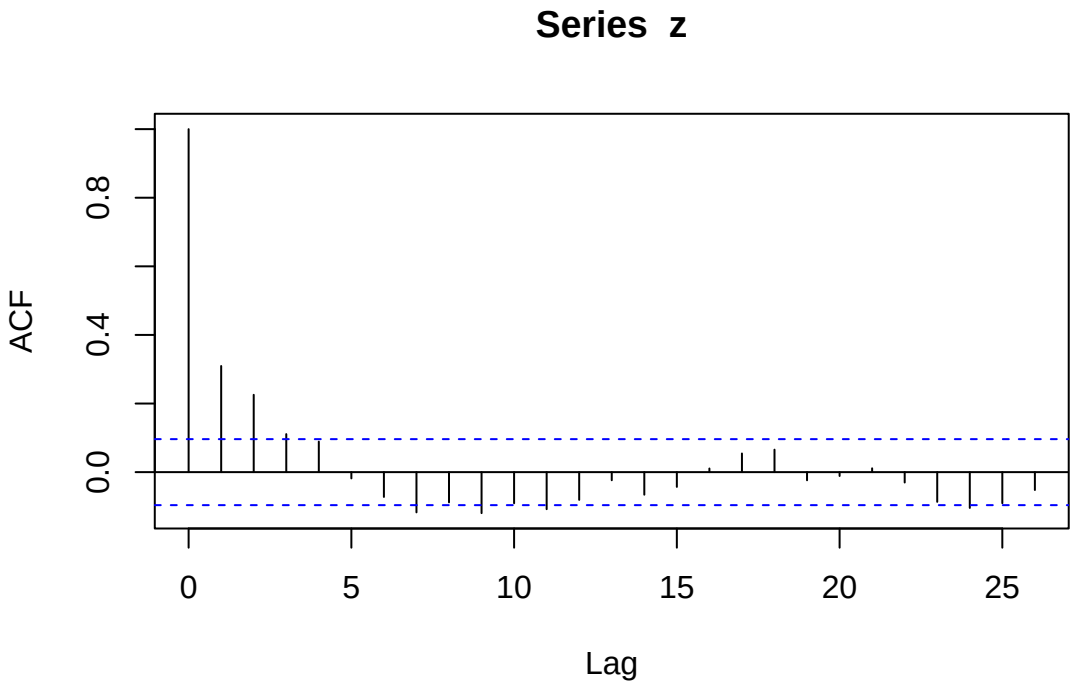
$$2(\ell_1 - \ell_0) \sim \chi^2_1.$$

Therefore, the p-value is $P(\chi^2_1 > 2.5) = 0.11$.

Q4. Interpreting diagnostics

Q4-01.

We consider data $y_{1:415}$ where y_n is the time, in miliseconds, between the nth and $(n + 1)$th firing event for a monkey neuron. Let $z_n = \log(y_n)$, with log being the natural logarithm. The sample autocorrelation function of $z_{1:415}$ is shown below.



We are interested about whether it is appropriate to model the time series as a stationary causal ARMA process. Which of the following is the best interpretation of the evidence from these plots:

A. There is clear evidence of a violation of stationarity. We should consider fitting a time series model, such as ARMA, and see if the residuals become stationary.

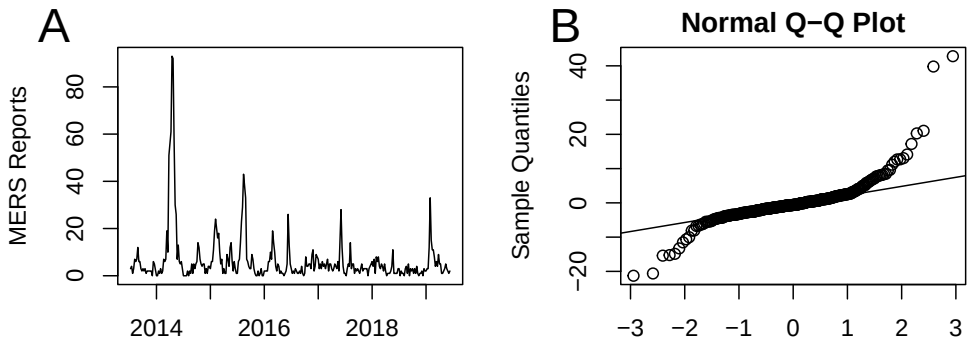
B. This plot suggests there would be no benefit from detrending or differencing the time series before fitting a stationary ARMA model. It does not rule out a sample covariance that varies with time, which is incompatible with ARMA.

C. This plot is enough evidence to demonstrate that a stationary model is reasonable. We should proceed to check for normality, and if the data are also not far from normally distributed then it is reasonable to fit an ARMA model by Gaussian maximum likelihood.

Solution. B.
The usual interpretation of the sample ACF assumes that variance and covariance depend only on lag (i.e., the data are well modeled by a shift-invariant autocovariance) but the plot does not check this.

There is clear evidence that the process is not well modeled by white noise. There is also clear evidence against a trend in the mean, which would show up as a slowly decaying sample ACF.

Q4-02.

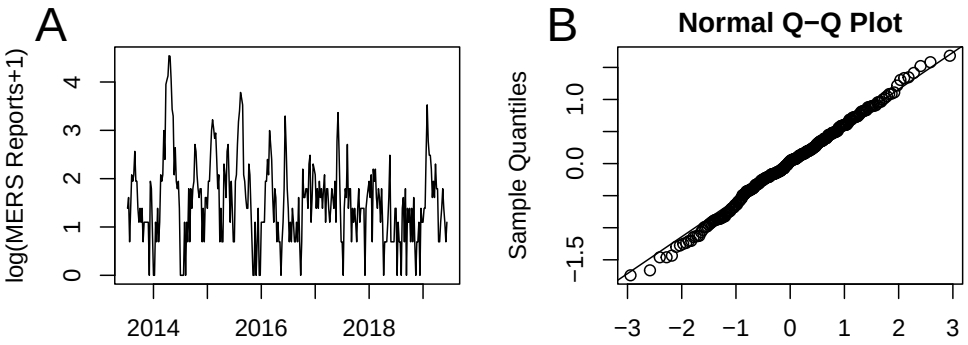


(A) Weekly cases of Middle East Respiratory Syndome (MERS) in Saudi Arabia. (B) a normal quantile

plot of the residuals from fitting an ARMA(2,2) model to these data using `arima()`. What is the best interpretation of (B)?

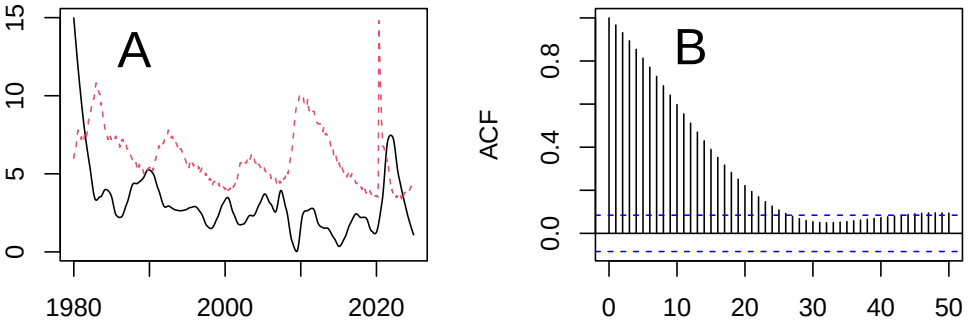
A: We should consider fitting a long-tailed error distribution, such as the t distribution.
B: The model is missing seasonality, which could be critical in this situation.
C: For using ARMA methods, these data should be log-transformed to make a linear Gaussian approximation more appropriate.
D: The normal quantile plot shows a long-tailed distribution, but this is not a major problem. We have over 300 data points, so the central limit theorem should hold for parameter estimates.
E: The normal quantile plot shows long tails, but with the right tail noticeably longer than the left tail. We should consider an asymmetric error distribution.
F: We should not interpret (B) before testing for stationarity. First run `adf.test()` and, if the null hypothesis is not rejected, recalculate (B) when fitting to the differenced data.

Solution. C.
Here's what happens when we take a `log(x+1)` transform, fitting ARMA(2,2) and checking the residuals as before.



It is often a good idea to log-transform non-negative quantities, and failure to do this can show up as long tailed residuals. Fitting long-tailed ARMA models is possible, but non-standard and not necessary here. There is seasonality, but an ARMA(2,2) model can already explain some periodicity so including a seasonal term in the model is not critical. There may be some non-stationarity here, but nothing that resembles the null hypothesis of the Augmented Dickey-Fuller test, so that is not relevant here.

Q4-03.



(A) Inflation (black) and unemployment (red) for the USA, 1980-2024. (B) Sample autocorrelation function of the residuals from a least square regression, 'lm(inflation~unemployment)', with estimated coefficients below. Which is the best interpretation of these graphs and fitted model?

```
##  (Intercept) unemployment
##   2.87056052   0.04543759
```

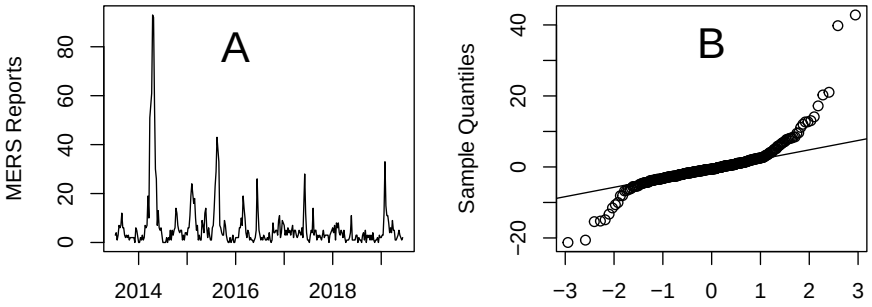
A: 0.05 is a reasonable estimate for the additional unemployment caused by one percentage point of additional inflation. We should not trust the uncertainty estimate (not shown), since our model does not allow for autocorrelation of the residuals.

B: 0.05 is a reasonable estimate for the association between inflation and unemployment. We should not assume there is a causal relationship. We should not trust the uncertainty estimate (not shown), since our model does not allow for autocorrelation of the residuals.

C: 0.05 is a reasonable estimate for the association between inflation and unemployment. We should make an additional assumption that there are no confounding variables, and then we can interpret this association to be causal. We should not trust the uncertainty estimate (not shown), since our model does not allow for autocorrelation of the residuals.

Solution. B.
Association is not causation. Inflation and unemployment are two aspects of a complex system. They may have some direct effect on each other, but they are also both driven by other aspects of the economy such as interest rates, consumer spending, and international trade. Mathematically, we could wish away these confounding variables (as proposed in answer C) but for practical purposes it does not make sense to do so.

Q4-04.



(A) Weekly cases of Middle East Respiratory Syndome (MERS) in Saudi Arabia. (B) a normal quantile residual plot for ARMA(2,2). We can formally test for non-normality of these residuals by a Shapiro-Wilk test (p-value=4.8×10^{-21}). What best describes the value added by presenting the Shapiro-Wilk test here?

A. We should always be alert for the danger of seeing patterns in noise. The Shapiro-Wilk test is useful to confirm our assessment that the normal quantile plot shows long tails.
B. Presenting the Shapiro-Wilk test here is not very insightful here, since the long tails are obvious from the normal quantile plot. However, adding this test demonstrates technical competence so it is better to include it than to omit it.
C. The long tails are established from the normal quantile plot. We could consider a log transform, or a long-tailed model, or a bootstrap simulation study to investigate whether the conclusions are sensitive to non-normality. Adding a fairly uninformative test instead of investigating the consequences of the error distribution could be a distraction from good data analysis.
D. The Shapiro-Wilk test is useful, but has the problem that it only tells us about lack of normality, not whether the non-normality is due to skew or kurtosis. We should supplement with a Jarque-Bera test to assess those.

Solution. C.
The key step after plotting B is to notice that the residuals have a long tail, especially to the right. This may remind us to try a log transform, which is successful here.

It is rather uninteresting to test a null hypothesis that is no longer plausible after plotting the data. If the normal quantile plot shows little deviation from normal, then it may be interesting to test whether that is enough to reject a Gaussian null, though in that case we probably do not have to modify the model (recall that statistical significance and practical significance can differ).

If it is not obvious to you that the deviation shown in the normal quantile plot is entirely incompatible with a Gaussian model, train your intuition by plotting some simulated normal quantile plots for data generated using `rnorm`.

If making useless tests is harmless, it might not matter whether we make a formal test for normality here.

However, in practice, useless analysis distract from useful analysis.

The normal quantile plot has diagnostic value beyond simply rejecting the null of normality. For example, we can see whether there are outliers, and we can compare the left and right tails. Tests such as Shapiro-Wilk and Jarque Bera may be useful in some situations. However, they often add little to a normal quantile plot.

Q4-05.

	AIC MA0	AIC MA1	AIC MA2	AIC MA3	AIC MA4	LBT MA0	LBT MA1	LBT MA2	LBT MA3	LBT MA4
AR0	174.82	48.81	9.61	-14.62	-17.98	0.000	0.000	0.000	0.00165	0.0312
AR1	-33.05	-34.44	-32.68	-30.69	-30.31	0.429	0.903	0.890	0.88400	1.0000
AR2	-34.07	-33.63	-33.28	-31.28	-29.70	0.888	0.733	0.965	0.95800	0.9740
AR3	-32.67	-33.20	-31.28	-31.42	-28.22	0.886	0.968	0.963	0.92000	0.9960
AR4	-30.77	-31.36	-30.96	-31.17	-29.59	0.889	0.955	0.954	0.98000	0.9970

The Ljung-Box test (LBT) provides an alternative approach to comparison of AIC values for selecting ARMA models. Whereas the standard sample autocorrelation function (ACF) residual plot tests each ACF component $\hat{\rho}_k$ under a null hypothesis of white noise, LBT tests $\sum_{k=1}^{h} \hat{\rho}_k^2$. Here, we present an AIC table and an LBT table (for $h = 5$). This course have favored AIC, with visual inspection of ACF and checking whether residual patterns appear in the frequency domain. There may be reasons to prefer LBT. Which of the following are good reasons to use LBT?

(i). LBT provides a p-value which is more formal than the comparison of AIC values.
(ii). Numerical issues involved in fitting an ARMA model may cause problems for comparing AIC values.
(iii). The LBT gives insights into what model to investigate next if the null hypothesis is rejected.
(iv). The LBT is useful in conjunction with AIC and ACF, since it provides an alternative perspective.

A. (i) only
B. (i, ii, iv)
C. (i,iii, iv)
D. (ii, iii, iv)
E. None of the above

Solution. E.
"None of the above" is correct in two senses, since none of the reasons proposed are strong. Perhaps LBT had more value in an era of less computatoinal power, before maximum likelihood estimation became routine for ARMA models. LBT usually adds little or nothing to the methods covered in class. We consider each in turn.

AIC values are a formal quantitative measure, just like p-values. They measure different things, as explained in the notes. Using p-values for model diagnostics is actually informal, since a formal p-value should correspond to a hypothesis specified before examining the data. So, (i) is not a good answer.

It is correct that numerical optimization can be problematic for fitting ARMA models. However, both AIC and LBT assess the same fitted model and so they share any consequences of imperfect optimization. So, (ii) is not a good answer.

Null hypothesis tests have the feature that they are relatively weak at telling you what to do if the null hypothesis is rejected, especially when they test against a very general alternative as for LBT. AIC provides a ranking of the models under investigation (though there may be reasons not to proceed with the top ranked model according to AIC). By contrast, it is unclear from the LBT table which model to choose. We can see that AR(0) models are inappropriate, but that is also clear from AIC. So, (iii) is not a good answer.

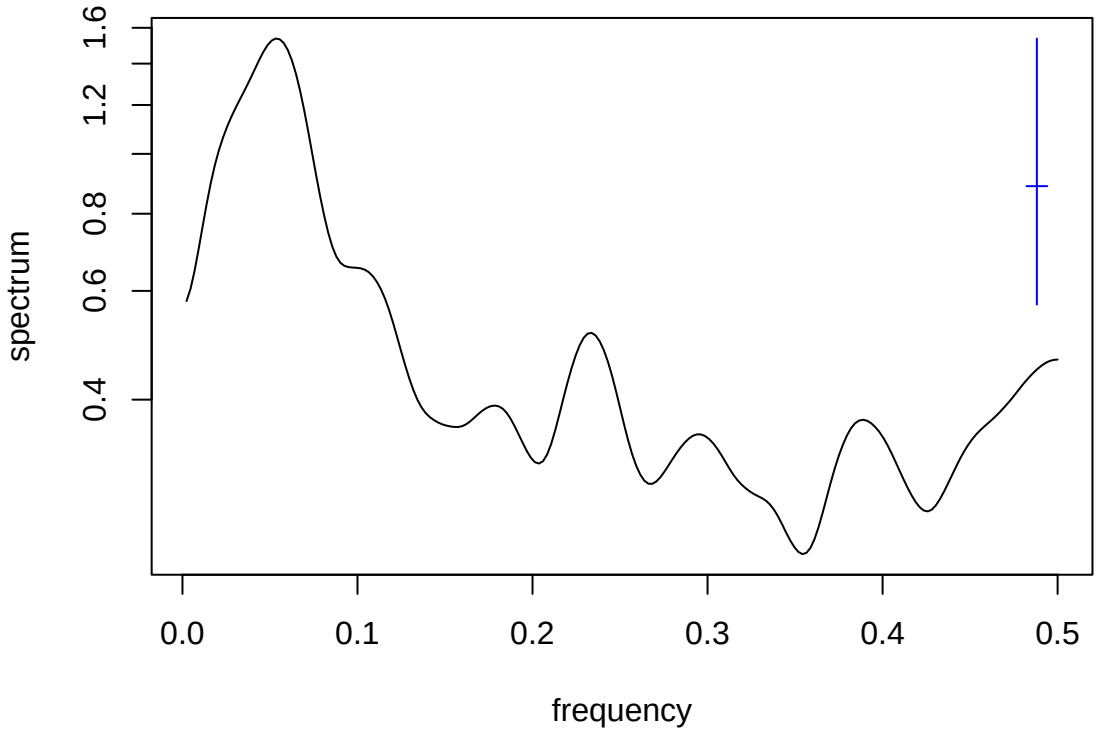
If LBT added anything substantial beyond AIC, it could be a useful extra component to the analysis. However, in this example, we see that every model with reasonable AIC values does not reject the LBT null.

This is commonly the case; if there is substantial autocorrelation in the residuals then a larger ARMA model will have better AIC. It can be useful to supplement AIC with a likelihood ratio test, since this provides a complementary perspective: AIC asks which model has better estimated predictive skill, whereas the null hypothesis test asks whether the simpler model is statistically plausible against the alternative of the more complex model. LBT does not address that. So, (iv) is not a good answer.

Q5. The frequency domain

Q5-01.

We consider data $y_{1:415}$ where y_n is the time interval, in milliseconds, between the nth and $(n+1)$th firing event for a monkey neuron. Let $z_n = \log(y_n)$, with log being the natural logarithm. A smoothed periodogram of $z_{1:415}$ is shown below. Units of frequency are the default value in R, i.e., cycles per unit observation. We see a peak at a frequency of approximately 0.07.



Which if the following is the best inference from this figure

A. Transitions between rapid neuron firing (short intervals between firing) and slow neuron firing (long intervals between firing) occur every $1/0.07 \approx 14$ firing events.

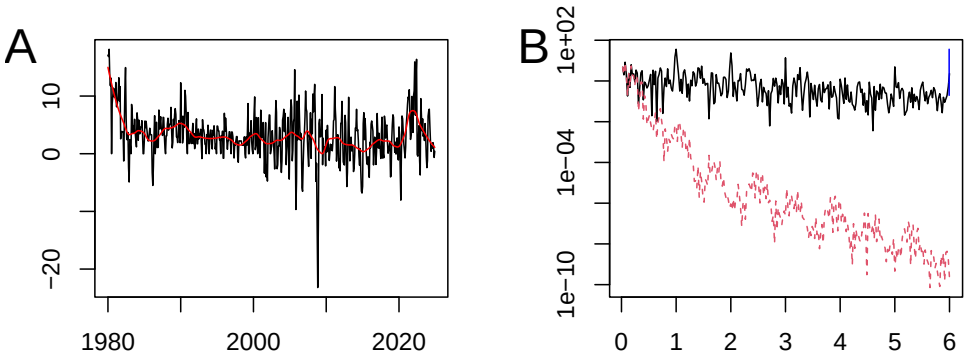
B. The neuron has a characteristic duration between firing events of $1/0.07 \approx 14$ milliseconds.

C. The neuron has a characteristic duration between firing events of $1/\exp(0.07) \approx 0.9$ milliseconds.

Solution. A.
In this example, the units of the "time" variable n are a dimensionless count of firing events. So, the units of frequency are cycles per firing event. The units of period are firing events per cycle. The peak in the smoothed periodogram corresonds to a cycle length of about $1/0.07 \approx 14$ observations. These cycles correspond to oscillations between low firing rate (i.e., long time intervals between firing) and high firing rate (short time intervals between firing).

The units of y_n are time, in milliseconds. It is unusual, and therefore somewhat counter-intuitive, to have a time series with units of time for the measured variable and a dimensionless "time" variable.

Q5-02.



The monthly US consumer price index (CPI) combines the price of a basket of products, such as eggs and bread and gasoline. (A) Annualized monthly percent inflation, i.e., the difference of log-CPI multiplied by 12×100 (black line); a smooth estimate via local linear regression (red line). (B) The periodogram of inflation and its smooth estimate. Which best characterizes the behavior of the smoother?

A: Cycles longer than 2 months are removed
B: Cycles shorter than 2 months are removed
C: Cycles longer than 2 year are removed
D: Cycles shorter than 2 year are removed
E: Cycles longer than (1/2) year are removed
F: Cycles shorter than (1/2) year are removed

Solution. D.
The units are in cycles per year; if we doubt this, we can tell because the largest peak at 1 corresponds to annual seasonality. At about (1/2) cycle per year, the power in the smooth fit (the red dashed line) rapidly falls several log units below the power in the raw monthly inflation data. (1/2) cycle per year corresponds to cycles of 2 year.

Q6. Scholarship for time series projects

Q6-01.

This question on citing references applies to any statistics report, but it is particularly relevant here since we are learning proper use of sources in order to write open-access midterm and final projects.

Suppose that the midterm project P1 cites a past project, P2, in the reference list. P1 refererences P2 at one point, mentioning that the projects have similarities. When you look at the source code and the writing, you find various points where P1 and P2 are almost identical, though at other points the projects are entirely different. What do you infer?

A: The authors of P1 have done enough to honestly disclose the relationship with P2. After all, there is sufficient information provided for any reader to track down the exact relationship.

B: The authors of P1 have misrepresented the relationship with P2 by appearing to take credit for some original work which was in fact heavily dependent on a source. This is a serious offence which should be reported to Rackham and/or the Associate Chair for Graduate Programs in Statistics as a violation of academic integrity.

C: There is not enough information to tell the actual story for certain. The authors of P2 may or may not have done something wrong, depending on information that is not available to us, but they did cite P2 so they should be given the benefit of the doubt and should not loose any scholarship points.

D: The authors of P1 have misrepresented the relationship with P2 by appearing to take credit for some original work which was in fact heavily dependent on a source. This is a moderately severe offence, partly offset by including P2 in the reference list. A substantial number of scholarship points should be subtracted.

E: P1 evidently has not shown perfect scholarship, but this is a small issue that could easily be an honest mistake given that the authors were not trying to hide the fact that they had studied P2. It is appropriate to substract, say, 1 point for scholarship for this mistake.

Solution. D.
If occurrences such as this were reported to the authorities, this would fill up too much time for the deans and chairs. This is a fairly serious issue, and a responsible student should not usually do this by mistake. It wastes everyone's time if proper credit is not clearly assigned to sources and if the grader has to track down the contribution of the authors.

Q6-02. Four people in a team collaborate on a project. After the project is submitted, a reader identifies that part of the project is adapted from an unreferenced source, i.e., it has been plagiarized. The team worked using git and cooperates on tracking down the issue, and the commit history clearly reveals who wrote the problematic part of the project. What is the most appropriate course of action:

A. The guilty coauthor should be penalized heavily for poor scholarship, and the other coauthors should have a minor penalty for failing to check their colleague's work.

B. All coauthors should share the same penalty, since this is a team project and all coauthors share equal responsibility for the submitted report.

C. The guilty coauthor should be penalized heavily for poor scholarship. The other coauthors have demonstrated strong scholarship by following good transparent working practices that enabled this issue to get quickly resolved, so they should not receive any penalty.

D. It is necessary to collect more information before coming to a decision. For example, the team may argue that the source is well known to all readers so did not have to be cited.

Solution. A.
Everyone should take responsibility for checking work submitted under their name. However, when it is possible to isolate the misconduct to one individual, and the rest of the team has demonstrated strong scholarship, they have some protection from the heavy scholarship consequences for the serious error.

One can always propose collecting new information, as in D, but in practice you often have to act on the information at hand. Parties can request a regrade if they think the decision is too far from justice.

Q6-03.

You discover that your team-mate is using Google Translate to carry out their share of the writing. The translation looks poorly done, similar in quality to ChatGPT, and does not use technical time series terminology correctly. What is the best course of action among the options below

A. Alert the instructor that you have a team mate adopting questionable scholarship strategies, in order to make sure you are not personally held responsible.

B. Ask ChatGPT to rewrite this problematic section to improve its quality

C. Help your team mate to rewrite the section in their own voice (shared with your voice).

Solution. C.
Different team mates bring different skills to the project, and perhaps you are more fluent at writing in English than one of your team mates. Major team problems can be discussed with the instructor, but this issue is best caught and corrrected early and solved within the team.

Q6-04.

Why is it helpful for a course such as DATASCI/STATS 531, that permits the use of internet resources including GenAI and past solutions, to require students to say explicitly say when they do not use sources?

A. Failure to give credit to sources is against the academic integrity rules of Rackham, the graduate school at University of Michigan.

B. It helps the GSI to grade the homework when they know exactly what sources have been used and for what question.

C. Students whose solution is more dependent on sources than they want to admit are reluctant to explicitly deny using sources.

D. The GSI has the task of evaluating whether the student has demonstrated thought about the homework task beyond collecting material from sources into a solution. This is not an easy task even when the sources are clearly listed and referenced at the point (or points) where they are used.

Solution. C.
The points made in A, B, D are all correct but are not directly relevant to the question. In an ideal world, the absence of a list of sources would be logically equivalent to an explicit statement saying that no sources were used. However, in practice that is not the case. We need a system that is robust against the natural tendency to hide information that could lead us to get a lower grade (because the grader would be able to see that our own contribution was smaller than it might otherwise appear). Most of us realize that explicitly saying we did not consult a source, when in fact we did, is a lie and amounts to academic misconduct. Failure to mention the source sounds like a milder misdemeanor. Therefore, to help the grader distinguish these things, it is important to be explicit about the lack of sources when indeed you did not need to consult any. It is appropriate for the GSI to award points for turning in solutions that are well-written and easier for the grader to evaluate.
