How good is my model?

1. After fitting multiple SARIMA models to some process X_t, a few natural questions arise:

   - How good is a given model?
   - How do I choose between two models?

   Let's focus on the first question: criteria for testing model fitness. Usually, given a model $(\hat{\phi}(B), d, \hat{\theta}(B))$, the prediction at time t, $\hat{X}_t$, is the best linear predictor given $(X_1, ..., X_{t-1})$ under the model. [1] In testing model fit, one approach is to study the standardized residuals:

   $$\tilde{e}_t := \frac{e_t}{\sqrt{\mathrm{Var}(e_t)}}$$

   where $e_t = X_t - \hat{X}_t$ denotes the residual at time t. For simplicity, assume (X_t) is a zero-mean process.

   (a) What is $\mathbb{E}\, e_t$?

   (b) What is the correlation between e_s and e_t for $s \neq t$?

[1] This can be quickly calculated via the Durbin-Levinson algorithm as covered in Chapter 3.4 of Shumway and Stoffer.

Thus, if the model is approximately correct, then the standardized residuals should be like a white noise process with variance 1, and under some mild conditions their correlations are asymptotically iid. [2] One way to check for independence for a sequence of data $y_1, ..., y_n$ is the Ljung-Box test, which relies on the Q-statistic:

$$Q_H = n(n+2)\sum_{h=1}^{H}\frac{\hat{\rho}_y^2(h)}{n-h}; \qquad \hat{\rho}_y^2(h) \text{ is the sample correlation at lag } h \text{ of the } y_t's.$$

Under the null assumption of the sample autocorrelations $\hat{\rho}_y(h)$ being iid Normal$(0, \frac{1}{n})$, (in this case $y_t = \tilde{e}_t$), then the Q-statistic has a χ^2 distribution and we reject for sufficiently large Q_H for a fixed lag H.[3]

Therefore, the above discussion suggests three criteria for checking model fit:

- **Plot the standardized residuals**. If the model is approximately true, then the $\tilde{e}_t$ should look like a white noise process with variance 1.
- **Plot the ACF of the standardized residuals**. If the model is approximately true, then asymptotic distribution of the sample autocorrelations of $\tilde{e}_t$, a white noise process, is Normal$(0, \mathbf{I}_n)$.
- **Perform a Ljung-Box test.** If the model is approximately true, then the sample autocorrelations of $\tilde{e}_t$ should be approximately iid Normal$(0, \frac{1}{n})$, which we check for using the Ljung-Box test. Note that the degree of freedom for the associated χ^2 distribution is $H - p - q$, accounting for the $p + q$ degrees of freedom lost in the model fitting.

Next, let's focus on second question: **criteria for model selection**. When trying to select between multiple candidate models, two good criteria to compare them on is the fit between data and model as well as trying to reduce model overfitting. There are two similar quantities that try to capture this.

- **Akaike's Information Criterion (AIC)** $= -2L_{model} + 2k$.
- **Bayesian Information Criterion (BIC)** $= -2L_{model} + k \log n$.

In both formulas above, L_{model} denotes the the maximum of the log-likelihood for the data under the considered model, k is the number of parameters, and n is how much data we have. The first term measures the fit between the data and the model, while the second term penalizes models with a large number of parameters. The BIC has a larger parameter penalty, so it values simpler models more than the AIC. [4]

[2] The sample variance is consistent (converges in probability to true variance).

[3] Usually, the null assumption is independence of the data y_t. However, the distribution of the Q-statistic really just depends on the distribution of the sample autocorrelations, which is why we can use it even though the $\tilde{e}_t$ are not necessarily independent.

[4] The AIC/BIC are only useful when comparing models. It's a relative value and is meaningless by itself.

Data analysis: airline passenger count

2. We'll be looking at the popular `AirPassengers` dataset in **R**, which consists of the number of US airline passengers per month from 1949 to 1960. We'll be needing the `astsa` and `forecast` packages too.

   (a) Load the `AirPassengers` dataset in **R**, as well as the above packages.

   (b) Recall that a SARIMA process is stationary after appropriate differencing. Therefore, before fitting a SARIMA model, transform the data so that it's stationary. This implies that the mean AND the variance have to be constant over time.

   (c) Fit an appropriate SARIMA model to the data using the function `sarima`. You need to specify the order of both the ARIMA and seasonal polynomials (AR, differencing, MA), as well as the period. Here are some useful criteria:

   - PACF/ACF for determining if process is pure (S)AR/(S)MA/mixed and the order.
   - `sarima` provides various diagnostics of model fit and model selection criteria:[5]
     - p-values of the coefficients of the model. Can be accessed using `$ttable`.
     - AIC, BIC, AICc, the model selection criteria. Can be accessed using `$AIC`, `$BIC`, `$AICc`.
     - Plot the standardized residuals, their ACF, and a table of Box-Ljung tests at multiple lags.

   (d) Perform cross validation of your given model using the following steps:

   (i) Save the first ≈ 80% of your transformed, stationary dataset as a separate object. This will be the training set.

   (ii) Using your parameters from 2(b), use `sarima.for` to predict 20 steps into the future and save it as an object. The `sarima.for` function will, given your input parameters, estimate the weights of the model and generate forecasts. [6]

   (e) Next, plot both your predictions and your actual data. You can access the predictions by using `$pred`. Also plot the residuals. How does it look?

   (f) Finally, a helpful function for guiding your estimation of the order of the SARIMA polynomials is `auto.arima`. Try it out on your transformed data. Is it similar to the model you found?

   While `auto.arima` is a great tool for helping you, be advised that in this class (including on the homework assignments and midterm) you still need to justify each step in your data analysis and verify that the model is good using the diagnostics. All else equal, a well-justified model is preferable to some black box model. Finally, be sure to check out `https://www.stat.pitt.edu/stoffer/tsa4/Rissues.htm` for a few puzzling issues you may encounter when doing time series analysis in **R**!

[5] We check for marginal normality of the residuals using the Q-Q plot to generate the prediction intervals. The AICc is just the AIC plus some extra parameter penalty term. Finally, since `sarima` gives a table of Ljung-Box statistics, be cognizant of the possibility of multiple testing, depending on how the Q-statistics are jointly distributed.

[6] Note that this model might differ from the previous model trained on the full data. However, in practice they should be similar since the training subset is most of the data.