STAT 153 - Introduction to Time Series
Homework Three

Fall 2022, UC Berkeley

Due by 11:59 pm on 11 October 2022
Total Points = 65

1. A noisy measurement device is being examined for understanding the distribution of the errors that are being produced by it. Suppose that ten measurements led to the following observations on the errors made by the device:

   `-0.69, -4.26, 0.14, -0.86, 0.42, 24.21, 0.51, -1.23, 2.30, 4.15`

   Consider the following three models for the distribution of these errors:

   - **Model 1**: $\epsilon_1, \ldots, \epsilon_n \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$.
   - **Model 2**: $\epsilon_1, \ldots, \epsilon_n \overset{\text{i.i.d}}{\sim} \text{Lap}(0, \sigma)$. Recall that the $\text{Lap}(0, \sigma)$ density is given by $x \mapsto \frac{1}{2\sigma} \exp\left(-\frac{|x|}{\sigma}\right)$.
   - **Model 3**: $\epsilon_1, \ldots, \epsilon_n \overset{\text{i.i.d}}{\sim} \text{Cauchy}(0, \sigma)$. Recall that the $\text{Cauchy}(0, \sigma)$ density is given by $x \mapsto \frac{1}{\pi} \frac{\sigma}{x^2+\sigma^2}$.

   a) Under the prior $\log \sigma \sim \text{Unif}(-15, 15)$, calculate the evidences of each of the above three models, given the observed data. Use a numerical approximation method (by gridding the set of possible σ values) for evaluating the integral over σ. (**6 points**)

   b) Normalize the three evidences to obtain posterior probabilities for the three models. Which model has the highest posterior probability? Comment on whether your results seem intutively sensible. (**2 points**).

2. In the file "HW3Data153Fall2022.csv", you will find data on two variables $(x_1, y_1), \ldots, (x_n, y_n)$. Consider the following four models for this dataset:

   - **Model 1**: $Y_i = \beta_0 + \beta_1 x_i + Z_i$ with $Z_i \overset{\text{i.i.d}}{\sim} C(0, \sigma)$ ($C(0, \sigma)$ has the density $x \mapsto \frac{1}{\pi} \frac{\sigma}{x^2+\sigma^2}$).
   - **Model 2**: $Y_i = \beta_0 + \beta_1 x_i + Z_i$ with $Z_i \overset{\text{i.i.d}}{\sim} \text{Lap}(0, \sigma)$ ($\text{Lap}(0, \sigma)$ has the density $x \mapsto \frac{1}{2\sigma} \exp\left(-\frac{|x|}{\sigma}\right)$).
   - **Model 3**: $Y_i = \beta_0 + Z_i$ with $Z_i \overset{\text{i.i.d}}{\sim} C(0, \sigma)$.
   - **Model 4**: $Y_i = \beta_0 + Z_i$ with $Z_i \overset{\text{i.i.d}}{\sim} \text{Lap}(0, \sigma)$.

a) Numerically calculate the Evidence for each of these models given the observed data. As priors, assume that β_0, σ and β_1 (if β_1 exists in the model) are independent with

$$\beta_0 \sim \text{unif}\{-10, -9.9, -9.8, \ldots, 9.8, 9.9, 10\}$$
$$\beta_1 \sim \text{unif}\{-10, -9.9, -9.8, \ldots, 9.8, 9.9, 10\}$$
$$\log \sigma \sim \text{unif}\{-10, -9.9, -9.8, \ldots, 9.8, 9.9, 10\}$$

Report the normalized evidences. Which model has the highest evidence and what is the value of this highest evidence? (**6 points**)

b) For your chosen model, describe your best estimates of β_0, σ and β_1 (if β_1 exists in the model) along with appropriate uncertainty quantification. Plot your best estimate of $\beta_0 + \beta_1 x$ on a plot of the data. (**4 points**)

3. `R` has an inbuilt dataset called `state` which gives some data on 50 states in America from the 1970s. You can access this dataset via (see `help(state)` for information about the data)

```
data(state); dt = data.frame(state.x77, row.names = state.abb)
```

We want to fit a linear model to this dataset with life expectancy as the response variable and some subset of the remaining seven explanatory variables (including the intercept term). Your goal is to figure out which subset of the explanatory variables provides the best explanation for the response variable in a linear model.

a) Use the Evidence-based Bayesian model selection method from Lecture 11 to calculate the evidences for each of the $2^7 = 128$ models (obtaining by taking all possible subsets of the explanatory variables along with the intercept term). How many of the 128 models get nontrivial Evidences (after normalization so that the Evidences sum to one)? Describe the models getting high evidences? Do the results of your analysis seem sensible? (**6 points**)

b) Calculate the best model (among the 128 possible models) using the BIC. Compare this model with the models obtaining high evidence from part (a). (**3 points**)

c) Calculate the best model (among the 128 possible models) using the AIC. Compare this model with the models obtaining high evidence from part (a). (**3 points**)

4. Download the google trends time series dataset for the query *yahoo*. This should be a monthly time series dataset that indicates the search popularity of this query from January 2004 to August 2022. The goal of this exercise is to use model selection to figure out the best polynomial trend model

$$Y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_k x_i^k + Z_i \qquad \text{with } Z_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$$

among the values $k = 1, 2, 3, 4, 5, 6, 7, 8$. To prevent numerical instability issues, take x_i to be some scaled version of time (for example, take $x_i = i/n$).

a) Use the Evidence-based Bayesian model selection method from Lecture 11 to calculate the evidences for each of the above 8 models. Report the normalized evidences of each of the 8 models. Which model has the highese evidence? Does this model selection method select models that seemings overfit? (**6 points**)

b) Calculate the best model using the BIC. Compare this model with the models obtaining high evidence from part (a). (**3 points**)

c) Calculate the best model using the AIC. Compare this model with the models obtaining high evidence from part (a). (**3 points**)

5. Download the FRED dataset on "Retail Sales: Beer, Wine, and Liquor Stores" from `https://fred.stlouisfed.org/series/MRTSSM4453USN`. This is a monthly dataset (the units are millions of dollars) and is not seasonally adjusted. To this data, we want to fit one of the models $M_{k,l}$ where k ranges in $0, 1, 2, 3, 4, 5$ and l ranges in $0, 1, 2, 3, 4, 5$. The model M_{kl} is given by:

$$Y_t = \beta_0 + \sum_{j=1}^{k} \beta_j \left(\frac{t}{n}\right)^j + \sum_{u=1}^{l} \left\{\alpha_{1u} \cos\left(\frac{2\pi u t}{12}\right) + \alpha_{2u} \sin\left(\frac{2\pi u t}{12}\right)\right\} + Z_t \qquad \text{with } Z_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$$

for $t = 1, \ldots, n$. Note that $k = 0$ means that the term $\sum_{j=1}^{k} \beta_j \left(\frac{t}{n}\right)^j$ is just missing (similarly for $l = 0$). For example, M_{00} corresponds to just fitting the constant β_0. The total number of models considered is $6^2 = 36$ corresponding to all pairs of choices $(k, l) \in \{0, 1, 2, 3, 4, 5\} \times \{0, 1, 2, 3, 4, 5\}$.

a) Use the Evidence-based Bayesian model selection method from Lecture 11 to calculate the evidences for each of the above 36 models. Report the normalized evidences of each of the 36 models. Which models have high evidence? Does this model selection method favors models that seemingly overfit? (**6 points**)

b) Calculate the best model using the BIC. Compare this model with the models obtaining high evidence from part (a). (**3 points**)

c) Calculate the best model using the AIC. Compare this model with the models obtaining high evidence from part (a). (**3 points**)

6. A classic time series dataset is the Box and Jenkins airline passenger data (can be accessed in R via `data(AirPassengers)`). This gives monthly totals of international airline passengers from 1949 to 1960. There are $n = 144$ observations in total corresponding to 12 years. To this dataset, consider fitting the models:

$$Y_t = a + bt + \sum_{f \in S} \left[\beta_{1f} \cos(2\pi f t) + \beta_{2f} \sin(2\pi f t)\right] + Z_t \qquad \text{with } Z_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$$

for some subset

$$S \subseteq \left\{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{18}{n}\right\}$$

There are 2^{18} models here in total and we shall denote them by M_S as S ranges over all subsets of $\{1/n, \ldots, 18/n\}$. In addition to these 2^{18} models, also consider the models:

$$\log Y_t = a + bt + \sum_{f \in S} \left[\beta_{1f} \cos(2\pi f t) + \beta_{2f} \sin(2\pi f t)\right] + Z_t \qquad \text{with } Z_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$$

again as S ranges over all subsets of $\{1/n, \ldots, 18/n\}$ (note that the response variable above is $\log Y_t$ as opposed to Y_t). Let us denote these models by LM_S (L standing for logarithm). Consider all these $2^{18} + 2^{18} = 2^{19}$ models together.

a) Use the Evidence-based Bayesian model selection method from Lecture 11 to calculate the evidences for each of the above 2^{19} models. Which models have high evidences? Do the frequencies k/n appearing in the high evidence models have any intuitive meaning? (**8 points**)

b) For fixed S, which of the two models M_S and LM_S generally has higher evidence? (**3 points**)