MATH 640: Exam 1, Computing

Instructions

1. You have until Thursday, February 28, at 6:30pm to finish the computing portion.
2. Please submit your answers electronically in PDF or DOC form to the assignment on Canvas before 6:30pm (i.e. class time). Your report must be typed (including any formulas or derivations).
3. Late submissions will be accepted up to 24 hours after the deadline, however they will be penalized: 4 points off for every six hours the submission is late.
4. Do not consult with each other, e-mail or meet with me or the TA if you have questions.
5. The derivations of the likelihood, prior, posterior, and any mathematical derivations should be shown for new models. If we have previously used the model, simply state the likelihood, prior, and resulting posterior.
6. Code must be included, either in line or in a Code Appendix at the end.
7. The score for each question will be based on the derivation/statement of the model (6 points), the code to generate posterior samples (4 points), and the generation and interpretation of the results (10 points).

Questions

1. Wind speed measurements were taken from New York's LaGuardia Airport over the course of 153 days starting in May and ending in September. We are interested in building a Bayesian model to examine attributes of the distribution of wind speeds. One common way to model wind speeds is using the Rayleigh distribution which is parameterized by its mode, $\theta > 0$. The density of the Rayleigh has the form

$$f(w_i) = \frac{w_i}{\theta^2} \exp\left(-\frac{w_i^2}{2\theta^2}\right).$$

With the distribution of the mode, we can empirically estimate the distribution of the mean $\left(\text{calculated as } \theta\sqrt{\pi/2}\right)$ as well as the median $\left(\text{calculated as } \theta\sqrt{2\log(2)}\right)$.

Assuming the sample is iid, first determine the likelihood and then find a conjugate prior for θ^2, making sure to state the resulting posterior. For your hyper-parameters, select them to make the prior non-informative—justify your choice.

Using the data in the file `wind.txt`[1] (wind speeds in mph), find the posterior distribution of the mode, mean, and median wind speeds at LaGuardia. Calculate summary statistics for each measures along with credible intervals. Finally, determine the posterior predictive probability that the wind speeds on one day between May and September of a future year will exceed 15 mph. Generate $B = 20000$ samples and set the seed to 17. (Hint 1: the posterior mode, mean, median, and predictive probability may be determined empirically using samples from the posterior of θ^2. Hint 2: the `rrayleigh()` function in the package `VGAM` may be useful.)

[1] The assumption of independence in time series data is usually suspect, however, the autocorrelation among the samples is quite low, even for successive days and drops off quickly suggesting that we can feel OK with this assumption.

2. Data from the Capital Bikeshare system was collected from 2011 and 2012 along with a set of predictors with the aim of building a model to assess the impact of different phenomenon, weather and workday related, on daily system usage. The dataset is in the file `day.txt` and contains the following variables:

   (a) `casual`: count of casual users on a day (outcome # 1)

   (b) `registered`: count of registered users on a day (outcome # 2)

   (c) `yr`: year (coded 0 for 2011, 1 for 2012)

   (d) `holiday`: whether day is a holiday or not

   (e) `workingday`: 1 if day is neither weekend nor holiday, 0 otherwise

   (f) `temp`: normalized temperature in Celsius

   (g) `atemp`: normalized "feeling" temperature in Celsius

   (h) `hum`: normalized humidity, the values were divided by 100

   (i) `windspeed`: normalized wind speed, the values were divided by 67

   The rows of the dataset correspond to days.

   Build two regression models, one with casual users as the outcome, the other with registered users. For each model, determine the "best" set of predictors supporting your choice with Bayesian inference results. Before modeling, check each outcome variable to ensure normality is a reasonable assumption, transform accordingly if not. Take $B = 20000$ samples for each, setting the seed to 13 for the casual users model and 235 for the registered users. Compare and contrast the final models discussing any differences, if any, and suggest possible reasons for what you see. Also discuss any similarities you see between the models. Provide relevant posterior summaries for all model parameters.

   Finally, Leap Day 2012 (February 29, 2012) was withheld from the dataset. Using its covariates, predict both the number of casual and registered users: `yr` = 1, `holiday` = 0, `workingday` = 1, `temp` = 0.344348, `atemp` = 0.34847, `hum` = 0.804783, and `windspeed` = 0.179117. Be sure to provide credible intervals for both of your predicted values. Set the seed to 2011 for your casual-model prediction and 2012 for your registered-model prediction. (Hint 1: you are free to choose the formulation of the regression model you prefer, from class or homework 3. Hint 2: you may end up fitting the two models on different scales, so comparisons should be based on Bayesian inference instead of the interpretation of coefficients.)