DATA6810 - Probabilistic Models for Complex Data

Assignment 2

Due: Monday 18 April 2022, 23:59AEST

1 General Description

This assignment covers topics 2, 3 and 4 of DATA6810, evaluating knowledge in the topics of Generalised Linear Models and Graphical Models. It consists of programming and technical exercises which are closely related to the topics covered in the respective tutorials. The expected completion time is 4 hours.

1.1 Deliverables

Submission will be through Microsoft Teams. Please complete the python notebooks / r markdown files for each section and upload them to the Teams assignment environment in the "your work" section.

Marking criteria involves a total of 100 points, with 50 points assigned to the content of each section. A penalty of MINUS 5 points per each day after the due date will be applied.

2 Generalised Linear Models (50%)

The datasets 'woodlark.csv' contains the number of migrating woodlarks sighted from Sept 2nd to November 11th for three years, 2007, 2008 and 2009.

For students whose surname begins with A-H, use 2007, for those whose surnames begins with I-P use 2008 and for those whose surnames begins with Q-Z use 2009.

Let $\mathbf{y} = (y_1, \ldots, y_n)$ be the vector of the number of observed migrating woodlarks for a chosen year. A common model for these sightings is to assume that on any given day t, the distribution of y_t has one of two forms, either

$$y_t \sim \text{Poisson}(\lambda_t) \tag{1}$$

or

$$y_t \sim \text{Negative Binomial}(\lambda_t, \theta)\ . \tag{2}$$

For both distributions the mean, λ_t has the following form

$$\log(\lambda_t) = \mathbf{x}_t \boldsymbol{\beta} \tag{3}$$

with $\mathbf{x}_t = [1, t, t^2]$, for $t = 1, \ldots, 71$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$. Assuming the Poisson distribution adequately models the data do the following

1. Compute the MLE estimates of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_{MLE}$, and the variance/covariance matrix of these estimates, $\boldsymbol{V}_{\hat{\beta}}$.

2. Construct a MCMC scheme using a Metropolis-Hastings transition kernel where the proposal distribution, $q(\cdot)$, is $\mathcal{N}\left(\hat{\boldsymbol{\beta}}_{MLE}, \boldsymbol{V}_{\hat{\beta}}\right)$.

3. Using the output from this sampling scheme:

   (a) Plot the posterior mean number of sightings, as a function of time with 95% credible intervals.

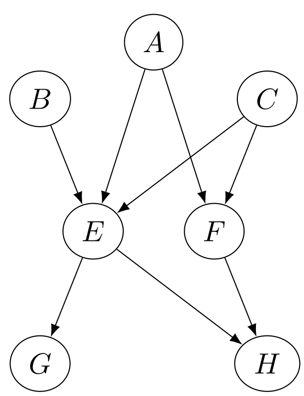
   (b) Compute the posterior distribution of the day on which the maximum number of sightings is predicted. To do this you will need to find the day on which the maximum number of expected sightings occurs at each iteration in the sampling scheme.

   (c) Using the sampling scheme comput the posterior probability that the number of sightings, on the day on which the maximum occurs, will exceed 40.

4. Is the Poisson distribution a suitable model for this data? Explain.

3 Graphical Models (50%)

Given the following DAG,



Answer the following

1. What is the Markov blanket of E?

2. Show the expression for the joint distribution of the DAG.

3. Assume that G is observed, is B and F conditionally independent given G? i.e.

$$B \perp\!\!\!\perp F \mid G \quad ? \tag{4}$$

4. Assume that A and C are observed, is B and F conditionally independent given A and C? i.e.

$$B \perp\!\!\!\perp F \mid A, C \quad ? \tag{5}$$

5. Explain the "Explaining Away" effect and provide an example for it.