Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
January 20, 2025

Problem n.1

A *Wildlife Conservation Team* is monitoring the behavior of a rare bird species in a remote forest. The birds alternate between Foraging for Food (F) and Resting in Nest (R). To track their behavior, the team measures the level of sound produced by the birds, recorded as discrete values: 1 (Low), 2 (Moderate), or 3 (High). Some conservationists hypothesize:

- When the bird is Foraging for Food (F), there is a 0.1 probability of producing a Low sound level (1), a 0.2 probability of Moderate sound level (2), and a 0.7 probability of High sound level (3). The bird has a 0.4 probability of transitioning to the Resting in Nest (R) state.
- When the bird is Resting in Nest (R), there is a 0.6 probability of producing a Low sound level (1), a 0.3 probability of Moderate sound level (2), and a 0.1 probability of High sound level (3). The bird has a 0.3 probability of transitioning back to the Foraging for Food (F) state.

By modeling this situation as a Hidden Markov Model, answer the following questions:

a) What are the hidden states of the model, and why?

Between 6:00 am and 7:00 am on May 3rd, 2024, the team recorded the following sequence of sound levels:
3, 3, 1, 1, 3, 3, 3, 2, 1, 3, 2, 1, 3, 3, 1, 1, 1, 2, 3, 1, 1, 1

b) Estimate Emission matrix and Transition matrix from the observed sequence, assuming that the initial distribution is uniform on the states (default).

Report the estimated probability of producing a High sound level (3) when the bird is Resting in Nest (R).

What is the most probable path for generating the observed sequence, given the estimated matrices?

Other conservationists decide to approach the issue differently and analyze the spatial distribution of *nest density* for a rare bird species, by collecting a dataset `birds.txt` in May 2024 containing data collected from various locations within the study area. It includes the UTM geographical coordinates s of the sampling locations, `vegetation_type` indicating the type of vegetation (dense forest = D or open forest = O or wetland = W), the `elevation` of the site [in kilometers], and the nest density $y(s)$ [nests per hectare]. Consider the following model:

$$y(s) = b_{0,j} + b_1 \texttt{elevation}(s) + \delta(s)$$

where $\delta(s)$ represents a 2^{nd} order stationary residual with an exponential variogram *without* nugget effect, and $j = 0, 1$ corresponds to the grouping induced by the variable `vegetation_type` ($j = 0$ for dense forest D, $j = 1$ for open forest O, $j = 2$ for wetlands W). Notice that no constant intercept should be included.

c) Report a plot of the fitted variogram, initialising the variogram fit with the model `vgm(0.25, "Exp", 1000)`. Indicate the estimate of the range and the sill.

d) Using the model fitted in (d), estimate the nest density $y^*(s_0)$ at location $s_0 = (5140000, 427000)$, in the dense forest, at an altitude of 3200m (=... km).

e) According to the model, what is the expected difference of nest density between a dense and a open forest, all other parameters being equal?

Upload your solution https://forms.office.com/e/di8rHbudcX