PPAML Challenge Problem: Bird Migration

Authors: Tom Dietterich (tgd@cs.orst.edu) and Shahed Sorower (sorower@eecs.oregonstate.edu) .
Version 5 May 13, 2014.

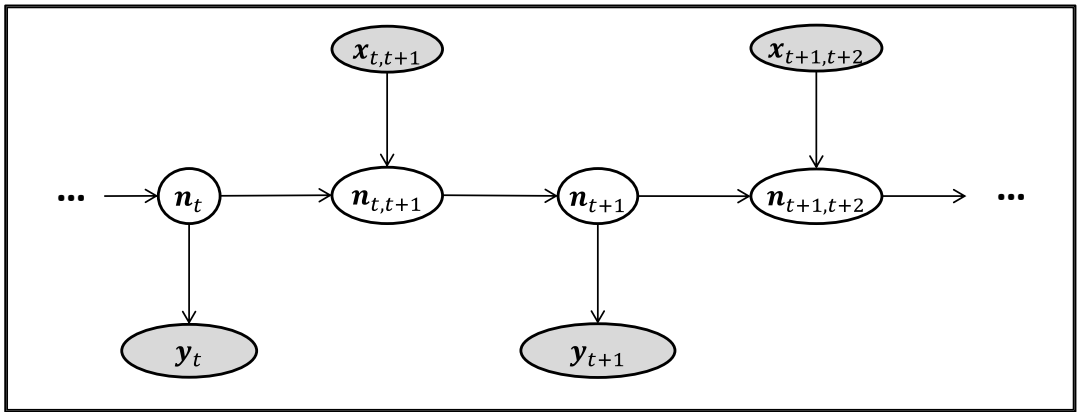
Credits: Simulator developed by Tao Sun (UMass, Amherst) and Liping Liu (Oregon State University)

Background: On peak nights during migration season, billions of birds take to the air across the US. However, because migration proceeds over vast temporal and spatial scales, and because it is difficult to observe directly, it is poorly understood. Scientists would like answers to questions such as (a) do birds wait for favorable winds before migrating (or are they on a fixed schedule)? (b) what factors influence a bird's decision to stop at some location? (c) what factors influence a bird's decision to resume migration? and (d) how do these factors vary from one species to another? Answering these questions requires constructing a model of the dynamics of bird migration. A team of researchers from UMass, Cornell, and Oregon State University is in the second year of an NSF grant to develop such a model (see http://birdcast.info). We have designed a challenge problem based on this project.

Modeling approach: The Eastern half of the US is divided into a rectangular grid containing a total of J cells. We will assume that all migration takes place at night (typically starting just after sundown). Each evening at sunset, a set of features (e.g., wind speed, direction, temperature, relative humidity) is measured in each cell. Each day, bird watchers participating in Project eBird make observations at locations of their choosing and upload a checklist to the web site http://ebird.org. There are three goals for the analysis:

1. **Reconstruction.** Given data for a series of years, estimate the number of birds $n_{t,t+1}(i,j)$ flying from cell i to cell j during the night separating day t from day $t+1$ (for all i,j,t). The absolute number is not identifiable from eBird observations, so we will assume a known total population size.
2. **Prediction.** Given data for a series of training years and data for the current year up through day t, predict the number of birds $n_{t,t+1}(i,j)$ flying from cell i to cell j during the night separating day t from day $t+1$. Also predict $n_{t+1,t+2}(i,j)$.
3. **Estimation.** Fit a log linear model to estimate $\log P(j|i,\boldsymbol{x}) \propto \beta^{\top}\boldsymbol{x}$, the probability of a bird in cell i flying to cell j under conditions $\boldsymbol{x}$ and output your estimated parameter vector $\hat{\beta}$.

You may assume that the birds in the population behave in an iid fashion. The decision to move from cell i to cell j on any night is made independently by each bird. Under this assumption, the formalism of Collective Graphical Models (Sheldon & Dietterich, 2011) can be applied. The graphical model is the following:



In this figure $\boldsymbol{n}_t$ is a vector whose ith entry indicates the number of birds in cell i on day t. $\boldsymbol{n}_{t,t+1}$ is a matrix whose (i,j) entry specifies the number of birds that flew from cell i to cell j on the night between day t and day $t+1$. $\boldsymbol{y}_t$ is a vector whose ith entry specifies the number of birds observed in cell i on day t. (We assume that the same amount of observation effort is made in each cell.) Finally, $\boldsymbol{x}_{t,t+1}$ is a set of features that determine the transition probabilities between cells. Specifically, entry (i,j) is a vector of four features that determine the probability that birds in cell i will fly to cell j during the night from t to $t+1$. These features determine a multinomial $P(j|i,\boldsymbol{x})$ over the destination states j conditioned on the source state i. Each bird then makes its migration decision independently by drawing from this multinomial. We can see that this is a special kind of Input-Output HMM. The main challenge is that the number of possible states is very large (consisting of all possible ways of assigning the population of N birds to the J cells).

To promote efficient inference, we set an upper limit of $d_{max} = 3\sqrt{2} \approx 4.2426$ units of distance as the maximum that a bird can fly in a single night.

The generative model can be written as follows:

- $y_t(i) \sim \text{Poisson}(n_t)$
- $\phi_{t,t+1}(i,j) = \beta^\top x_{t,t+1}(i,j)$
- $\theta_{t,t+1}(i,j) = \frac{\exp \phi_{t,t+1}(i,j)}{\sum_{j'} \exp \phi_{t,t+1}(i,j')}$ However, this is zero if the distance from i to j exceeds d_{max}, and the sum in the denominator is restricted only to those cells j' that are within d_{max} of i.
- $n_{t,t+1}(i,j) \sim \text{Multinomial}\left(n_t(i);\ \theta_{t,t+1}(i,1), \ldots, \theta_{t,t+1}(i,J)\right)$. This is $n_t(i)$ draws from the multinomial defined by the θs.
- $n_{t+1}(j) = \sum_i n_{t,t+1}(i,j)$. This is deterministic.

Data: For phase 1, we are providing three simulated data sets. Our intention is that the first data set can be used for basic debugging, especially for integrating the log-linear model into the transitions and the Poisson model into the observations. The second data set involves a "small" population of 1000 birds, which may be small enough to permit certain brute-force inference methods to succeed. The third data

set involves a much larger population of one million birds, which presumably will require more clever inference.

The data are organized into folders as follows:

```
data
      input    # read input features and observations from here
            dataset1
            dataset2
            dataset3
      output   # your answers should be written into these folders
            dataset1
            dataset2
            dataset3
      ground   # ground truth is stored here
            dataset1
            dataset2
            dataset3
```

Data set 1: "One Bird". The data consist of observations of a single bird traversing a 4x4 grid over a period of 30 years. Each year, the bird starts in the lower left cell and attempts to migrate to the upper right cell. In effect, we are observing 30 state sequences, so this is a fully-observed IOHMM. Two data files are provided:

onebird-observations.csv:

- Columns:
  - Year {1, 2, ..., 30}
  - Day {1, 2, ..., 20}
  - Cell1: Number of birds observed in cell 1 for that year and day. The observations are distributed as a Poisson random variable that depends on the true number of birds in the cell according to $P(O|N) = Poisson(N)$. That is, the intensity parameter is equal to the true number of birds in the cell (0 or 1 in this data set).
  - Cell2, ..., Cell16. Cells are indexed in Matlab order (column major)

onebird-features.csv:

- Columns:
  - Year {1, 2, ..., 30}
  - Day {1, 2, ..., 20}
  - From.cell {1, 2, ..., 16}
  - To.cell {1, 2, ..., 16}
  - f1 (float): encodes distance from the from.cell to the to.cell with noise
  - f2 (float): encodes the difference between the vector from from.cell to to.cell and the desired destination (which is the upper right corner)
  - f3 (float): encodes wind direction
  - f4 (float): encodes whether from.cell == to.cell

Our intent is that you will use these four features directly as the potentials in a log-linear model.

The evaluation metric for this data set is to estimate the parameters of the log linear model for the transition probabilities: $P(i|j) \propto \exp[\beta_1 f_1(i,j) + \beta_2 f_2(i,j) + \beta_3 f_3(i,j) + \beta_4 f_4(i,j)]$.

Data set 2: "10x10x1000". The data consist of observations of a population of 1000 birds traversing a 10x10 grid from the lower left corner to the upper right region. Data are provided for 3 years, 20 days per year. For this problem, we provide separate train and test data sets (of the same size).

10x10x1000-train-observations.csv, 10x10x1000-test-observations.csv:

Columns:

Year {1, 2, 3}

Day {1, 2, ..., 20}

Cell1: Number of birds observed in cell 1 for that year and day. The observations are distributed as Poisson(N), where N is the true number of birds in the cell.

Cell2, ..., Cell100. Cells are indexed in Matlab order (column major)

10x10x1000-train-features.csv, 10x10x1000-test-features.csv:

Columns:

Year {1, 2, 3}

Day {1, 2, ..., 20}

From.cell {1, 2, ..., 100}

To.cell {1, 2, ..., 100}

f1 (float): same as in onebird

f2 (float): same as in onebird

f3 (float): same as in onebird

f4 (float): same as in onebird

There are three tasks to perform:

1. **Reconstruction:** Your program will output a comma-separated-value file in the following format

   10x10x1000-train-reconstruction.csv: year, day, from.cell, to.cell, number.of.birds

   This is your estimate for each year and day in the training data of the number of birds that flew from the from.cell to the to.cell. The evaluation metric will be the squared difference between the actual and the reconstructed number of birds, summed over all years, days, and cell pairs.

2. **Prediction:** In the prediction task, you should use the 3 training years to fit your model and then make predictions on the 3 testing years. When predicting night $t+1$, you should use the observations $\boldsymbol{y}_{1:t}$ and the weather covariates $\boldsymbol{x}_{1:t+1}$. For predicting night $t+2$, you should still use observations $\boldsymbol{y}_{1:t}$ but you can use the weather covariates $\boldsymbol{x}_{1:t+2}$. In effect, you will have perfect weather forecasts for nights $t+1$ and $t+2$. Your program will output a comma-separated-value file in the following format:

10x10x1000-test-prediction.csv: year, day, from.cell, to.cell, number.birds, number.birds2

number.birds is the number of birds on this night. number.birds2 is the number of birds on the next night (i.e., predicted 48 hours in advance).

The metric will be the squared difference between the actual and the predicted number of birds, summed over the three test years and all days and cell pairs.

3. **Estimation:** The metric will be the squared difference between the predicted and actual values of the parameters. You should use only the training data to make this estimate.

Data set 3: "10x10x1000000". The structure of this data set is the same as for the 10x10x1000 data set except that the population now consists of one million birds. The files are named as follows:

10x10x1000000-train-observations.csv, 10x10x1000000-test-observations.csv:
10x10x1000000-train-features.csv, 10x10x1000000-test-features.csv:

The same tasks should be solved as for Data set 2.

References:

The Birdcast team has published the following papers that are relevant to this problem:

Sheldon, D., Elmohamed, M. A. S., & Kozen, D. (2007). Collective inference on Markov models for modeling bird migration. *Advances in Neural Information Processing Systems*, *20*, 1321–1328.

Sheldon, D., & Dietterich, T. G. (2011). Collective Graphical Models. In *NIPS 2011*.

Sheldon, D., Sun, T., Kumar, A., & Dietterich, T. G. (2013). Approximate Inference in Collective Graphical Models. In *Proceedings of ICML 2013*.

Liu, L-P., Sheldon, D., Dietterich, T. G. (2014). Gaussian Approximation of Collective Graphical Models. In *Proceedings of ICML 2014.*