PPAML Challenge Problem: Bird Migration

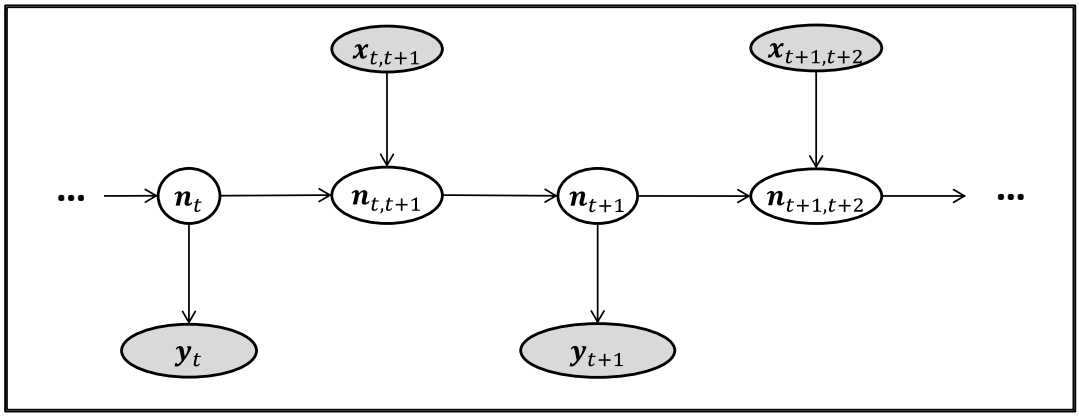
Version 6 – August 30, 2014
Authors: Tom Dietterich (tgd@cs.orst.edu) and Shahed Sorower (sorower@eecs.oregonstate.edu)
Credits: Simulator developed by Tao Sun (UMass, Amherst) and Liping Liu (Oregon State University)

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.

Teams may also wish to consider Gaussian approximations to the Collective Graphical Model (Liu et al. 2014). Liu et al. showed that, when the population is large, Gaussian approximation (GCGM) is able to preserve the dependency structure of the Collective Graphical Model (CGM), and is able to make efficient and accurate approximate inference.

Data: The data package for this challenge problem contains three data sets. The first data set '1-bird' 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 '1000-birds' 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 '1M-birds' involves a much larger population of one million birds, which presumably will require more clever inference.

For each of these datasets, we provide two CSV files that the teams can use as input to their probabilistic programs. The observation data file contains the number of birds observed in each cell for each day and year. The feature data file contains the features values observed for each pair of cells for each day and year. Notice that the features on the last day of each year is not specified because that is irrelevant for the problems here.

Data set 1: 1-Bird. The data consist of observations of a single bird traversing a 4x4 grid (Figure 1) 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.

4	8	12	16
3	7	11	15
2	6	10	14
1	5	9	13

Figure 1: A 4x4 grid with cell numbers

Two data files are provided:

4x4x1-train-observations.csv:

Column	Column Data
Year	Year number, {1, 2, ..., 30}
Day	Day number, {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	As Cell1. Cells are indexed in Matlab order (column major)

4x4x1-train-features.csv:

Column	Column Data
Year	Year number, {1, 2, ..., 30}
Day	Day number, {1, 2, ..., 19}
FromCell	Cell number, {1, 2, ..., 16}
ToCell	Cell number, {1, 2, ..., 16}

Column	Column Data
f1	Float value. Encodes distance from the FromCell to the ToCell with noise (the distance between two cells is passed through a lognormal distribution (μ=1, σ=1) to obtain a desirability score for flying that distance)
f2	Float value. Encodes the difference between the vector from FromCell to ToCell and the desired destination, which is the upper right corner (cosine distance between the two vectors)
f3	Float value. Encodes wind direction (cosine distance between the vector from FromCell to ToCell and the wind direction)
f4	Float value. Encodes whether FromCell == ToCell. Contains "2" if true.

Our intent is that you will use these four features directly as the potentials in a log-linear model. Note that all feature values are set to zero if the distance between a pair of cells is greater than d_{max}.

The task for the '1-bird' data set is the following.

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

   Your program will output a comma-separated-value file in the following format.

   File name: 4x4x1-estimated-parameters.csv
   Column Headers (first row of the csv file): b1,b2,b3,b4

   For example, your CSV should have only the following table (2 rows including the header):

b1	b2	b3	b4
2.099555	4.561738	2.966087	3.723159

   The numeric values in this table should be your estimated parameters.

   The evaluation metric will be the squared difference between the predicted and actual values of the parameters.

Data set 2: 1000-birds. 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:

Column	Column Data
Year	Year number, {1, 2, 3}
Day	Day number, {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…Cell100	As Cell1. Cells are indexed in Matlab order (column major)

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

Column	Column Data
Year	Year number, {1, 2, 3}
Day	Day number, {1, 2, ..., 19}
FromCell	Cell number, {1, 2, ..., 100}
ToCell	Cell number, {1, 2, ..., 100}
f1	Float value. Encodes distance from the FromCell to the ToCell with noise (the distance between two cells is passed through a lognormal distribution (μ=1, σ=1) to obtain a desirability score for flying that distance)
f2	Float value. Encodes the difference between the vector from FromCell to ToCell and the desired destination, which is the upper right corner (cosine distance between the two vectors)
f3	Float value. Encodes wind direction (cosine distance between the vector from FromCell to ToCell and the wind direction)
f4	Float value. Encodes whether FromCell == ToCell. Contains “2” if true.

For the ‘1000-bird’ dataset there are three tasks to perform:

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

   File name: 10x10x1000-train-reconstruction.csv:
   Column headers: Year, Day, FromCell, ToCell, NumberOfBirds

   This is your estimate for each year and day in the training data of the number of birds that flew from the FromCell to the ToCell. For example, your CSV should have only the following table (570001 rows including the header):

Year	Day	FromCell	ToCell	NumberOfBirds
1	1	1	1	888.5977
1	1	1	2	0.129334

Year: 1,2,3
Day: 1,2,…,19
FromCell: 1,2…,100
ToCell: 1,2,…,100

The evaluation metric will be the squared difference between the actual and the reconstructed number of birds, summed over all the 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.

   File name: 10x10x1000-test-prediction.csv
   Column headers: Year, Day, FromCell, ToCell, NumberOfBirds, NumberOfBirds2

   NumberOfBirds is the number of birds on this night. NumberOfBirds2 is the number of birds on the next night (i.e., predicted 48 hours in advance).

   For example, your CSV should have only the following table (570001 rows including the header row):

Year	Day	FromCell	ToCell	NumberOfBirds	NumberOfBirds2
1	1	1	1	45.35418	2.094707
1	1	1	2	1.722693	6.116287

   Year: 1,2,3
   Day: 1,2,…,19
   FromCell: 1,2…,100
   ToCell: 1,2,…,100

   Because our observation is limited to 20 days only, you do not have to predict 'NumberOfBirds2' on day 19. You can leave those cells empty or can print as -1.

   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:** Similar to the '1-bird' data set, estimate the parameters of the log linear model for the transition probabilities: $P(i|j) \propto \exp[\beta_1 f_1(i,j) + \beta_2 f_2(i,j) + \beta_3 f_3(i,j) + \beta_4 f_4(i,j)]$.

   Your program will output a comma-separated-value file in the following format.

File name: 10x10x1000-estimated-parameters.csv
Column Headers (first row of the csv file): b1,b2,b3,b4

For example, your CSV should have only the following table (2 rows including the header):

b1	b2	b3	b4
2.099555	4.561738	2.966087	3.723159

The numeric values in this table should be your estimated parameters.

The evaluation 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: 1M-birds. The structure of this data set is the same as for the '1000-birds' 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 the '1000-birds' data set, and therefore, your program should output the following files.

10x10x1000000-train-reconstruction.csv
10x10x1000000-test-prediction.csv:
10x10x1000000-estimated-parameters.csv

The data formatting of these CSV output files should be same as specified in '1000-birds' dataset.

Data Input and Output

The data package is organized into folders of input and ground truth data as follows.

```
data
    input    # read input features and observations from here; this will be provided to you
        4x4x1-train-observations.csv
        4x4x1-train-features.csv
        10x10x1000-train-observations.csv
        10x10x1000-train-features.csv
        10x10x1000-test-observations.csv
        10x10x1000-test-features.csv
```

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

`ground` # ground truth is stored here; this will be used for evaluation
4x4x1-ground-parameters.csv
10x10x1000-ground-parameters.csv
10x10x1000-reconstruction-ground.csv
10x10x1000-prediction-ground.csv
10x10x1000000-ground-parameters.csv
10x10x1000000-reconstruction-ground.csv
10x10x1000000-prediction-ground.csv

The output files of your challenge problem solution should be written to a single output folder. For example, if this folder is named 'output', we would expect the following solution output files:

`output` # your answers should be written into a folder; this is what you need to submit
4x4x1-estimated-parameters.csv
10x10x1000-estimated-parameters.csv
10x10x1000-train-reconstruction.csv
10x10x1000-test-prediction.csv
10x10x1000000-estimated-parameters.csv
10x10x1000000-train-reconstruction.csv
10x10x1000000-test-prediction.csv

Frequently Asked Questions

Covariate Specifications

1. What are the exact specifications of the four covariates?
   Ans: Here is the snippet of matlab code that computes them. The four coefficients are a1, a2, a3, and a4, and they multiply the four features:

```
% Phi(i,j,t) = ( a1*(4*lognpdf(|v_{ij}|,mu,sigma)
%             + a2*cos(v_{ij}, U_i)
%             + a3*cos(v_{ij}, W) )*(i ~= j)
%             +( a4*selfT )*(i == j);
% v_{ij} : unit vector from cell i to j
% U_i    : birds' perferred transition direction, always points to the
%          top-right cell.
% W      : random wind direction, unit vector
% selfT = 2 is the self-transition parameter
```

Interpretation of Data Sets

1. It appears that the only feature between a pair of cells which changes over time is the wind direction – That is, the value of f1, f2 and f4 are the same between cell i and cell j, regardless of the day or year. Is this correct, and if so, will it be the case for the evaluation data as well?

   Answer: Yes, this is correct. The features will be the same on the evaluation data as they are with the data you have.

Range of Values for Parameters

1. Is there a range of values a scientist would consider reasonable for the beta parameters in the bird migration problem?

   Answer: Since this is a logistic regression, the parameters are on a log-odds scale. A reasonable range is [-20,20]. The usual prior for logistic regression is that the parameters are drawn from zero-mean gaussians (e.g., with standard deviation of 10). In this case, the features have been coded such that we would expect all of the parameter values to be non-negative. So you might use the absolute value of the Gaussian as a prior.

Interpretation of Observation Counts

1. How should we interpret the observation counts?

   Answer: The way to think about the observation count data is the following:

   Suppose we have a single bird watcher in each cell. Every day, if the true number of birds in the cell is N, the birdwatcher reports Poisson(N) number of birds. So the reported number can be 0 even when N is non-zero. It can also be greater than N. But its expected value will be N.

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.*