Bird embedding

Li-Ping Liu

1 Introduction

The eBird data consists of checklists of bird observations. The figure below shows some sites having checklist submissions on Manhattan island. eBird project even includes a webpage that shows real-time submissions.

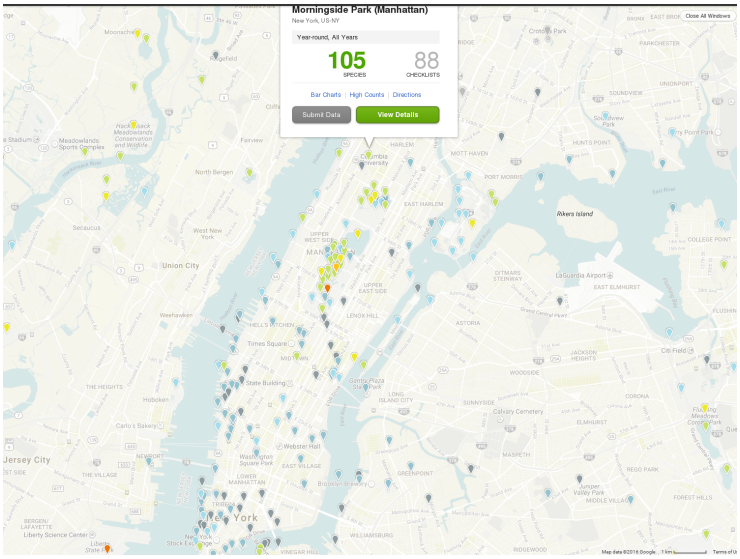

Figure 1: eBird checklist submissions on Manhattan island. (I believe) Checklists are aggregated by a number of sites for better plot. The location at CU is mistakenly labeled as Morningside Park. Zoom in for better view.

Each checklist contains its location (latitude & longitude), time point, and a list of bird counts for 953 species. Two sets of covariates are associated to each checklist. The first set (≈ 12) of covariates describe the environmental information, such as elevation, temperature, and vegetation coverage, and explain why the bird is there. The second set (≈ 4) of covariates are about the observation process and states how the observation is made, such as the type of observation (staitionary, traveling, area survey, etc.) and duration of observation.

There are about 6 million checklists during the last 5 years (from 2010 to 2014).

In this work, we would like to discover bird relations by bird embedding. The idea of bird embedding originates from word embedding in NLP, where various relations among words are discovered[citation of word embedding]. In this work, we would like to discover relations among birds by applying the embedding technique to the eBird data.

Dispite its similarity with word embedding, there are new problems to consider for bird embedding.

- Large quantity of data, which has 6 millions observations.
- Rich information about checklists. In another word, checklists are not from the same distribution. We should construct a model in which environmental covariates expain presence/absence of species, and bird embedding captures interactions among species.
- Locations near each other should have similar distribution of birds, not only because they share similar covariates, but also because birds fly around. How to smooth the distribution?
- How can we explain the embedding result? Making recommendation may not be a good idea here.

2 Bird Embedding

The first model to consider is the combination of exponential family embedding and the exposure model.

In the embedding model, we essentially need to define the conditional distribution of the bird count of a species in a checklist given its *context*. Since we are interested in relationships among bird species, the *context* of a bird count is the vector of bird counts of other species in the checklist. Due to the generality of context, we can define the context of a bird count flexibly, for example, as bird counts of other species averaged over checklists within some radius of the current observation. The average may give more stable results, but we will consider this later.

The exposure model can be used to describe the observation process. Mathematically, it plays the role of down-weighting zero entries in the observation matrix. In another word, a species is not observed either because no such bird lives there or because it is not detected by that observation. If the model choose the second explanation, then model would use little strength to fit the zero value. We will see this after we have defined the model formally.

Let's define the model. Suppose there are N checklists, and let i, $1 \leq i \leq N$, index checklists. In the data there are $J = 953$ species, each of which is indexed by j, $1 \leq j \leq J$. In each checklist i, y_{ij} birds are observed for species j.

For each checklist, the feature vector $\mathbf{x}_i$ brings some information of the observation process. The probability u_{ij} of observing each species j is calculated as $u_{ij} = \text{logistic}(\boldsymbol{\beta}_j^\top \mathbf{x}_i)$, where $\boldsymbol{\beta}_j$ is a parameter. The indicator b_{ij} of observing species j at checklist i is sampled from Bernoulli distribution with probability u_{ij}.

$$b_{ij} \sim \text{Bernoulli}(u_{ij}). \tag{1}$$

The observed count y_{ij} is from Poisson distribution defined as follows.

$$y_{ij} \sim \text{Poisson}(b_{ij}\lambda_{ij}), \tag{2}$$

where the rate λ_{ij} is the rate calculated from the embedding.

To define the embedding, we first define the *context* of y_{ij}, which consists of species with positive observations.

$$C_{ij} = \{j' : y_{ij'} > 0, j' \neq j\} \tag{3}$$

The embedding of species j is $\boldsymbol{\alpha}_j$. The rate is defined as follows.

$$\lambda_{ij} = f\left(\boldsymbol{\rho}_j^\top \sum_{j' \in C_i} r(y_{ij'})\boldsymbol{\alpha}_{j'}\right) \tag{4}$$

where the vector $\boldsymbol{\rho}_j$ is the weight vector shared by species j. The function $r(\cdot)$ maps counts to a value in $[0, 1]$ to avoid that a large count dominate the embedding. The function $f(\cdot)$ maps a value in $\mathcal{R}$ to $\mathcal{R}^+$. It can be defined as the exponential function or the softplus function.

The parameters α_j, ρ_j, and β_j are given Gaussian priors.

$$\boldsymbol{\alpha}_j \sim \text{Normal}(\mathbf{0}, \sigma_1^2 I), \tag{5}$$
$$\boldsymbol{\rho}_j \sim \text{Normal}(\mathbf{0}, \sigma_2^2 I), \tag{6}$$
$$\boldsymbol{\beta}_j \sim \text{Normal}(\mathbf{0}, \sigma_3^2 I), \tag{7}$$

where σ_1^2, σ_2^2, and σ_3^2 are hyper-parameters, $\mathbf{0}$ represents a zero vector with proper length N, and I represents the identity matrix of with proper size.

The parameter $\boldsymbol{\rho}$ and and $\boldsymbol{\alpha}$ explains the correlation among bird species. The correlation comes from either shared environmental factors or birds' interactions. As an extension of the model, we can explain bird distributions with environmental covariates first and the apply the embedding model to the residual to capture more information about bird interactions. We can define the Poisson rate λ as follows.

$$\lambda'_{ij} = \lambda_{ij} + g_j(\mathbf{v}_i) \tag{8}$$

where $g_j(\cdot)$ is a function that maps the covariates $\mathbf{v}_i$ at location i to a fitness value for species j. If $g_j(\mathbf{v}_i)$ can perfectly model species distributions for single species, then the embedding model will capture interactions among species.

However, it is more likely that there are still some factors not included in $\mathbf{v}$. Nevertheless, the embedding model will contain both the marginalized correlation and interactions among birds. Experts may be able to identify factors missing from $\mathbf{v}$ or manually discover more bird interactions from the embedding model.

The function $g_j(\mathbf{v}_i)$ should be learned first, so the species distribution is explained by environmental covariates first. The embedding model is used to explain the part not explained by f. Compared with joint learning of f and θ, there are two benefits from this considerations: a) the embedding part can focus more on interactions among bird species, and b) learning is computationally more efficient.

3 Inference

In this second, we develop a varitional inference method to infer parameters of the model.

3.1 E-step: calculating posterior distribution of observation variable

In this subsection, we calculate the posterior distribution of a_{ij}.

$$q_{ij}^0 = p(a_{ij} = 0 | y_{ij} > 0, \lambda_{ij}, u_{ij}) = 0 \tag{9}$$

$$q_{ij}^1 = p(a_{ij} = 1 | y_{ij} > 0, \lambda_{ij}, u_{ij}) = 1 \tag{10}$$

$$q_{ij}^0 = p(a_{ij} = 0 | y_{ij} = 0, \lambda_{ij}, u_{ij}) = \frac{1 - u_{ij}}{1 - u_{ij} + u_{ij} \exp(-\lambda_{ij})} \tag{11}$$

$$q_{ij}^1 = p(a_{ij} = 1 | y_{ij} = 0, \lambda_{ij}, u_{ij}) = \frac{u_{ij} \exp(-\lambda_{ij})}{1 - u_{ij} + u_{ij} \exp(-\lambda_{ij})} \tag{12}$$

3.2 M-step: maximizing the log-likelihood with respect to model parameter

In the M-step, we want to maximize the following objective.

$$\begin{aligned} LL(\boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \sum_{ij} E_{a_{ij}} \Big[\log p\left(y_{ij} | a_{ij} \lambda_{ij}\right) + \log p\left(a_{ij} | \boldsymbol{\beta}_j^\top \mathbf{x}_{ij}\right) \Big] \\ &= \sum_{ij} -q_{ij}^0 \log(1 + \exp(\boldsymbol{\beta}_j^\top \mathbf{x}_i)) + q_{ij}^1 \Big(y_{ij} \log(\lambda_{ij}) - \lambda_{ij} - \log\left(1 + \exp(-\boldsymbol{\beta}_j^\top \mathbf{x}_i)\right) \Big) \end{aligned}$$

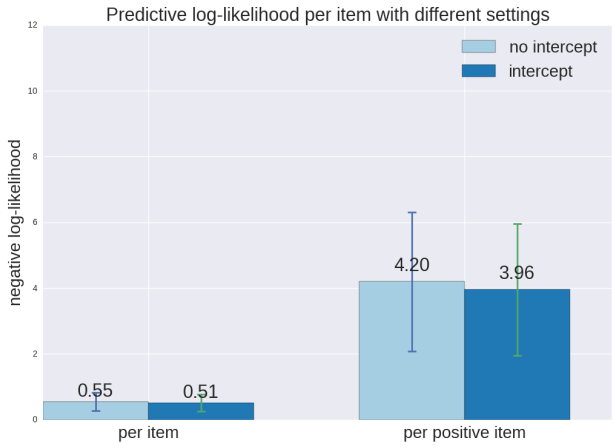

Figure 2: Negative predicted log-likelihood with and without the intercept term

Take derivatives with respect to the parameters.

$$\nabla_{\boldsymbol{\rho}_j} LL = \sum_i q^1_{ij} \left(y_{ij}/\lambda_{ij} - 1\right) \nabla_{\boldsymbol{\rho}_j} \lambda_{ij} \tag{13}$$

$$\nabla_{\boldsymbol{\alpha}_j} LL = \sum_i \sum_{j' \in C_{ij}} q^1_{ij'} \left(y_{ij'}/\lambda_{ij'} - 1\right) \nabla_{\boldsymbol{\alpha}_j} \lambda_{ij'} \tag{14}$$

$$\nabla_{\boldsymbol{\beta}_j} LL = \sum_i -q^0_{ij} \text{logistic}(\boldsymbol{\beta}_j^\top \mathbf{x}_{ij}) \mathbf{x}_{ij} + q^1_{ij} (1 - \text{logistic}(\boldsymbol{\beta}_j^\top \mathbf{x}_{ij})) \mathbf{x}_{ij} \tag{15}$$

With the softplus function,

$$\nabla_{\boldsymbol{\rho}_j} \lambda_{ij} = \frac{\exp(h_{ij}/\delta)}{1 + \exp(h_{ij}/\delta)} \sum_{j' \in C_{ij}} r(y_{ij'}) \boldsymbol{\alpha}_{j'} \tag{16}$$

$$\nabla_{\boldsymbol{\alpha}_j} \lambda_{ij'} = \frac{\exp(h_{ij'}/\delta)}{1 + \exp(h_{ij'}/\delta)} r(y_{ij}) \boldsymbol{\rho}_{j'} \tag{17}$$

4 Experiment

Intercept term:

5 Research Directions

Assumption: embeddings at similar time-locations are similar but different. Can we use a $\boldsymbol{\rho}_{ij}$ for each checklist i and species j and put a GP prior over $\boldsymbol{\rho}_{ij}$-s to

encourage strong correlation among $\boldsymbol{\rho}_{ij}$-s at neighboring locations?

Can we do inference over 6 million checklists?

Can we predict the presence/absence of species?

6 TODO list

- [Done]Derive the inference method
- [Done]Run a small dataset with the P-EMB model
- [Mostly done]Code the algorithm
- Setup the experiment protocol
- ...