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.

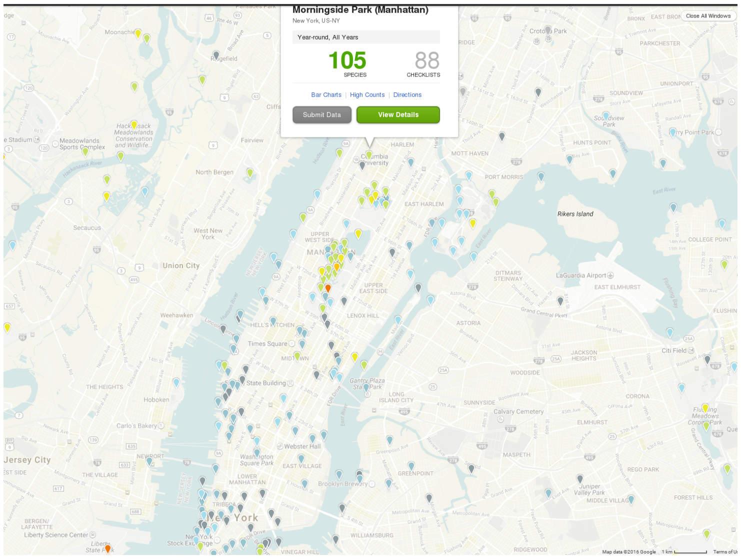

Figure 1: eBird checklist submissions on Manhattan island. (I believe) Checklists are aggregated by a number of sites for better plot. The location at CU is mistakenly labeled as Morningside Park. Zoom in for better view.

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 $J = 214$ 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 $u_{ij} = \mathrm{logistic}(\boldsymbol{\beta}_j^\top \mathbf{x}_i + \bar{\beta}_j)$, where $\boldsymbol{\beta}_j$ is a parameter, and $\bar{\beta}_j$ is the intercept term. The indicator b_{ij} of observing species j at checklist i is sampled from Bernoulli distribution with probability u_{ij}.

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

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

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

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.

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

The vector $\boldsymbol{\rho}_j$ is the weight vector shared by checklists of 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 the product from $\mathbb{R}$ to $\mathbb{R}^+$. It can be defined as the exponential function or the softplus function ($f(x) = \log(1 + \exp(x))$). Different with traditional embedding, we also include an intercept term $\bar{\lambda}_j$ as the base rate.

Gaussian priors are put on the parameters α_j, ρ_j, and β_j.

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

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. If the embedding is to capture more about birds' interactions, the base rate $\bar{\lambda}_j$ needs to explain more about environmental factors by including environmental covariates.

2.1 Settings

The model is defined to be flexible. Four different configurations are listed here.

downweighting zeros: There are three settings for downweighting zeros. The first one is not downweighting zeros, the probability u_{ij} for each (i, j) being set to 1 instead of calculated by the logistic function. The second setting fits the probability u_{ij} by the term $\bar{\beta}$ only, then the distribution of counts is essentially a zero-inflated Poisson distribution. The third setting fits the probability u_{ij} by both observation covariates and $\bar{\beta}$, so zeros a further explained by covariates.

scaling context: The counts in context exhibit large varince, and counts of some species are much larger than those of others. To prevent the embedding being dominated by species with large counts, counts are normalized by $r(\cdot)$ function. One method is to divide each count by the 95% quantile of the counts of that species. If the setting is on, the counts are scaled.

intercept term: $\bar{\lambda}_j = 0$ gives the basic embedding, or $\bar{\lambda}_j$ is fit by the data.

link function: $f(x) = \exp(x)$, or $f(x) = \log(1 + \exp(x))$.

3 Inference

In this section, we develop a varitional inference method to infer parameters of the model. The inference method below is general enough to include all settings above.

3.1 E-step: calculating posterior distribution of observation variable

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

$$q_{ij}^0 = p(b_{ij} = 0|y_{ij} > 0, \lambda_{ij}, u_{ij}) = 0 \tag{8}$$

$$q_{ij}^1 = p(b_{ij} = 1|y_{ij} > 0, \lambda_{ij}, u_{ij}) = 1 \tag{9}$$

$$q_{ij}^0 = p(b_{ij} = 0|y_{ij} = 0, \lambda_{ij}, u_{ij}) = \frac{1 - u_{ij}}{1 - u_{ij} + u_{ij}\exp(-\lambda_{ij})} \tag{10}$$

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

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

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

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

Take derivatives with respect to the parameters.

$$\nabla_{\boldsymbol{\rho}_j} LL = \sum_i q_{ij}^1 \left(y_{ij}/\lambda_{ij} - 1\right) \nabla_{\boldsymbol{\rho}_j} \lambda_{ij} \quad (12)$$

$$\nabla_{\boldsymbol{\alpha}_j} LL = \sum_i \sum_{j' \in C_{ij}} q_{ij'}^1 \left(y_{ij'}/\lambda_{ij'} - 1\right) \nabla_{\boldsymbol{\alpha}_j} \lambda_{ij'} \quad (13)$$

$$\nabla_{\boldsymbol{\beta}_j} LL = \sum_i -q_{ij}^0 \text{logistic}(\boldsymbol{\beta}_j^\top \mathbf{x}_{ij}) \mathbf{x}_{ij} + q_{ij}^1 (1 - \text{logistic}(\boldsymbol{\beta}_j^\top \mathbf{x}_{ij})) \mathbf{x}_{ij} \quad (14)$$

With the softplus function,

$$\nabla_{\boldsymbol{\rho}_j} \lambda_{ij} = \frac{\exp(h_{ij})}{1 + \exp(h_{ij}/\delta)} \sum_{j' \in C_{ij}} r(y_{ij'}) \boldsymbol{\alpha}_{j'}, \quad (15)$$

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

where $h_{ij} = f^{-1}(\lambda_{ij})$.

3.3 Maximization with Stochastic Gradient

Combining the E-step with the M-step by appying q values to the gradients above gives the gradient of the log-likelihood with respect to model parameters. Instead of calculating the exact gradient with all training instances, we only calculate a noisy gradient by using only one instance randomly drawn from the training set. Then we use AdaGrad to update model parameters with noisy graidents. Practically, the algorithm converges within 100,000 iterations.

4 Experiment

In the experiment, a subset of checklists are taken from a rectangular area that mostly overlps with Pennsylvania and the period from day 180 to day 210 in 2014. The subset has overall 5488 checklists (trips) and 214 bird species (items).

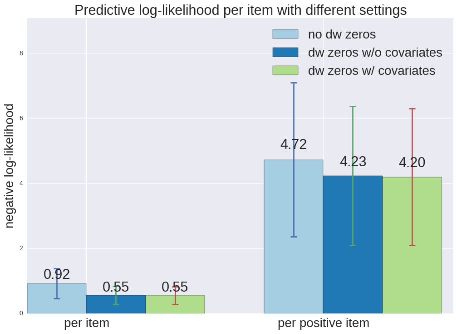

Figure 2: Negative predicted log-likelihood with different measures of downweighting zeros

The dataset is randomly split into two thirds as the training set and one third as the test set. The model is trained on the training set and tested on the test set for ten times on 10 different random splits. In each training run, the model is optimized on nine tenth of the training set and validated on the rest one tenth. For each random split, the negative predictive log-likelihood is recorded. In two randomly splits, there are some outliers that greatly affect the performance of the model, so these two random splits are removed. We report the average negative predictive log-likelihood on 8 random splits.

The basic setting is as follows. $K = 10, \sigma_1 = \sigma_2 = \sigma 3 = 100, \bar{\lambda}_j = 0, f(\cdot) = \exp(\cdot)$. Counts in context are scaled by 95% quantile value of each species. Observation covariates and $\bar{\beta}$ are used to fit u_{ij}. In the following result, only the setting to be examined is varied.

Downweighting zeros: See results in Fig. 2. This result shows that the average performance with different measures of downweighting zeros. When both covariates and $\bar{\beta}$ are used, the mean value of the performance is smallest.

Scaling context: See results in Fig. 3. Scaling counts in contexts helps to stablize the algorithm. Without scaling the context counts, the log-likelihood of validation set stops increasing before much less iterations than that of the setting with scaling counts.

Intercept term: See results in Fig. 4. Intercept terms $(\bar{\lambda}_j)_j$ helps to explain the data.

: See results in Fig. 5. Two link functions give very similar results.

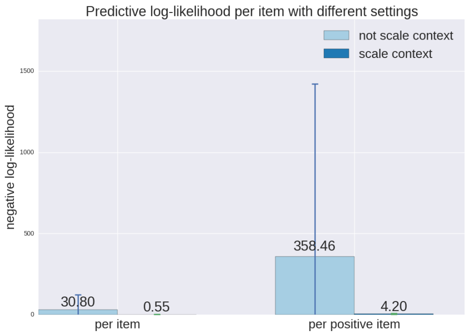

Figure 3: Negative predicted log-likelihood with and without scaling the context

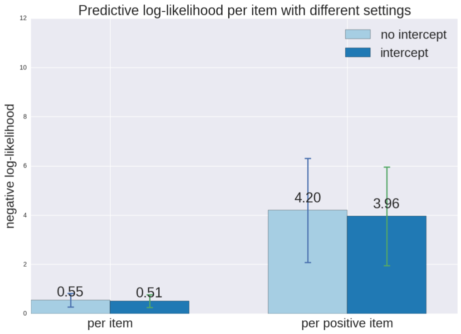

Figure 4: Negative predicted log-likelihood with and without the intercept term

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 predict the presence/absence of species?

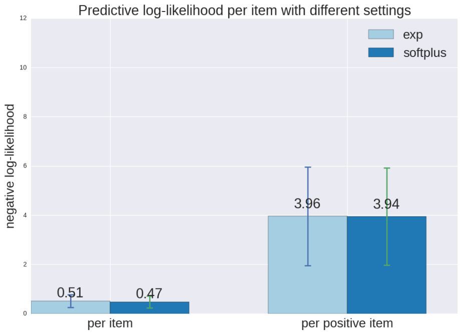

Figure 5: Negative predicted log-likelihood with exp and softmax as the link function

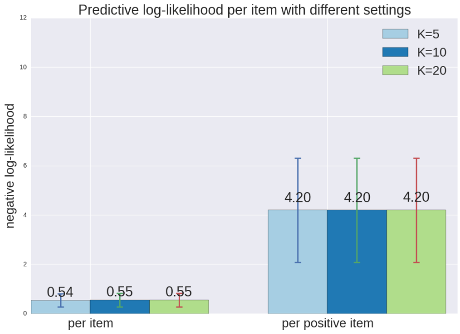

Figure 6: Negative predicted log-likelihood with different dimensionalities