Collective Inference on Markov Models for Modeling Bird Migration

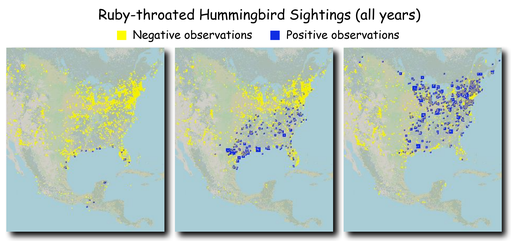
Cornell University

Daniel Sheldon M. A. Saleh Elmohamed Dexter Kozen

1. Summary

eBird: Citizen Science

More than 10 million bird observations reported voluntarily by general public, since 2002. Migration is evident:



Goal: automatically infer migration paths

Family of inference problems:

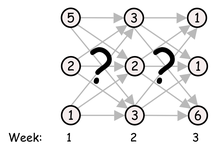
- Given partial information about many independent samples from Markov chain, reconstruct most probable paths matching observations.
- Like Hidden Markov Models (HMMs), but many paths at once.
- Techniques: network flows in the trellis graph.

Demonstration:

Visualization of Ruby-throated Hummingbird migration.

2. Problem Setup

Input: Weekly distributions of hummingbirds across locations (say, map grid cells), estimated from eBird.



Example with 3 locations. How to "connect the dots"?

(n) : Number of birds at location

Probabilistic model:

Assume a population-wide Markov model (given) governs migration. States are locations, each bird is an independent sample.

Inference Framework:

Given partial observations about many independent sample paths from a Markov chain, find most probable paths matching observations.

Here, observations are a distribution over states (locations) at each time step. We also consider a generalization, where states may be hidden, and we only observe (distributions of) state-dependent output symbols. This gives a true generalization of HMMs.

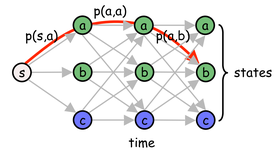
Problems:

- Multiple Path Problem: find exactly M paths.
- Fractional Path Problem: find a distribution of paths.

3. Illustrating the Problems

Trellis Graph

States replicated for each time step; edges forward in time, labeled with transition probability.



●●:Colors indicate output symbol. Assume WLOG that each state outputs a unique symbol.

Probability of red path is p(s,a) x p(a,a) x p(a,b).

Summaries

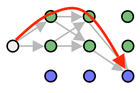
Single Path (classical HMM)	Multiple Path	Fractional Path
Observe symbols. Find best path.	Observe multisets of M symbols. Find best set of M paths.	Observe distributions of symbols. Find best distribution of paths.
● ● ●	2 x ● 1 x ● 3 x ● 0 x ● 1 x ● 2 x ●	.6 x ● .4 x ● .9 x ● .1 x ● .2 x ● .8 x ●

Example observations.

4. Flow Techniques

Background: Viterbi Algorithm ⇒ Shortest Path:

For the single path problem, Viterbi finds the highest probability path matching a single sequence of observed symbols. Transforming probabilities to costs c(u,v) = -log p(u,v), Viterbi finds a shortest path in the trellis.



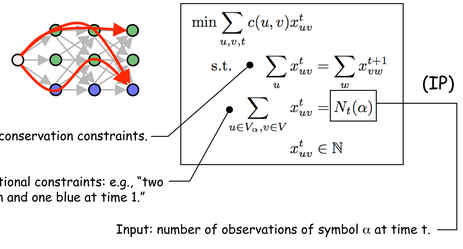
E.g., suppose observations are ● ● ●.

Delete edges to nodes of wrong color, then find shortest path.

Cost of red path is c(s,a) + c(a,a) + c(a,c).

Multiple Path ⇒ Minimum Cost Flow:

Most probable set of M paths is the minimum cost flow of M units through the trellis. Force paths to match observations by adding additional constraints to integer program formulation of the flow problem.



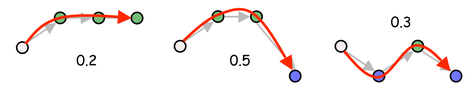
Multiple Path Problem is NP-hard.

- By reduction from SET COVER.
- Can solve IP as an integer program without polynomial-time guarantees.
- Or seek more efficient problem variants.

5. Fractional Paths

Linear Program

- Simple idea: relax IP to get a linear program (LP).
- Result: flow paths may split into fractional parts.
- Normalize to one unit of flow:



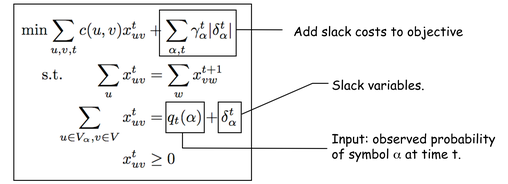
Interpretation:

LP solution is a distribution over paths. This is the "most probable" distribution that matches the observations, i.e., gives correct distribution of symbols at each step.

Formally: Let P(Y) be the probability that path Y is generated by the Markov model. The fractional paths LP finds a distribution π^* (over paths) maximizing $E_{Y\sim\pi}[\log P(Y)]$ over all distributions π whose marginal distributions over symbols at each time step match the observations.

Slack:

It is useful to incorporate slack into the LP constraints. Observations might not be exact, and poorly estimated observations can cause "jitter" in the form of unnecessary transitions.



- Also add constraints so $q_t(\alpha) + \delta_\alpha{}^t$ is probability distribution.
- Slack costs can be tailored to individual input values: charge more to deviate from confident estimates.

6. Demo: Ruby-throated Hummingbird

Markov Model

We use a simple model favoring short flights over long ones. Transition probabilities are Gaussian:

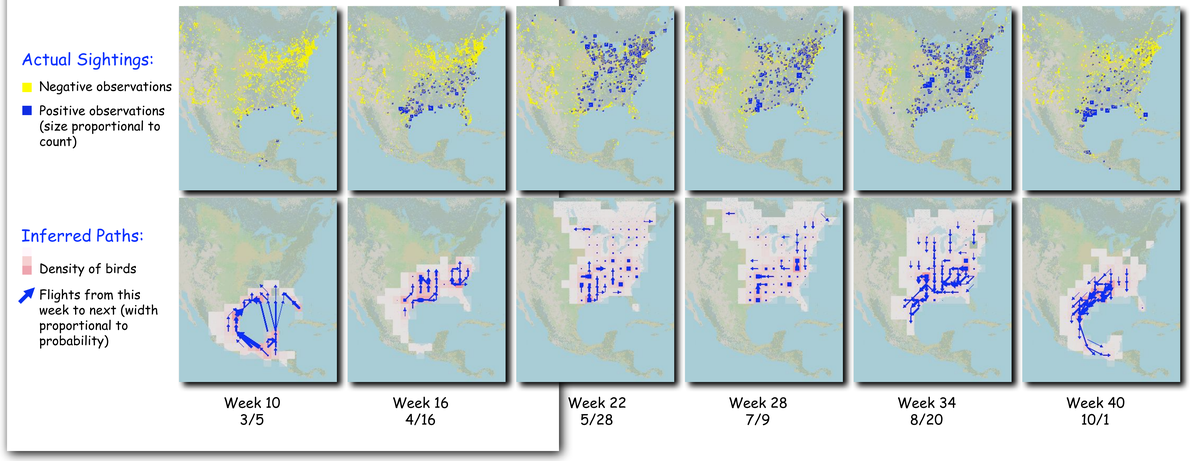
$$p(x,y) \sim \exp(-||x-y||^2),$$

truncated so p(x,y) = 0 for ||x-y|| > 1350 km.

Weekly Distributions

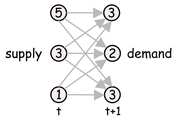
From eBird, we need estimates for the weekly distributions $q_t(u)$, and slack costs γ^t_u. In our paper, we describe a simple technique to derive these values from eBird observations, using a spatial smoothing technique based on harmonic energy minimization to infer values for locations with little data.

In general, producing these estimates is an important and challenging component of the overall problem that could benefit from more attention.



Aside: Efficient Special Case

In our application, we directly observe distributions of states (instead of symbols). Ideally, the problem simplifies in this case: decisions at each level of the trellis are independent, and each subproblem can be solved directly as a network flow problem.



Match supply at time t with demand at time t+1 to minimize cost.

Decisions are independent of how supply got there.

However, this simplification is incompatible with slack variables, so our application uses the complete fractional paths LP.

Acknowledgements: We are grateful to Daniel Fink, Wesley Hochachka and Steve Kelling at the Cornell Lab of Ornithology for useful discussions.