Random Interval Graphs for Birdwatching and other Chronological Sampling Activities

Edgar Jaramillo-Rodriguez

UC Davis

May 16th, 2022

Joint Work with: J. De Loera, D. Oliveros, and A. Torres-Hernández [JDLTH]

Motivating Problem: Bird Watching

Figure: Clockwise from top left: Canadian Geese, Whitehead Sparrow, Mourning Doves, Scrubjay

Motivating Problem: Bird Watching

Assume bird sightings come as a sequence of i.i.d. random variables $Y_1, Y_2, \ldots$ taking values in $[m] = \{1, \ldots, m\}$, where m is the number of bird species.

Motivating Problem: Bird Watching

Assume bird sightings come as a sequence of i.i.d. random variables $Y_1, Y_2, \ldots$ taking values in $[m] = \{1, \ldots, m\}$, where m is the number of bird species.

Question: how many sightings will it take before we have observed every type of bird?

The Coupon Collector's Problem (a brief history)

Problem: *"If each box of a brand of cereals contains a coupon, and there are* **m** *different types of coupons, what is the probability that more than* **n** *boxes need to be bought to collect all* **m** *coupons?"*

The Coupon Collector's Problem (a brief history)

Problem: *"If each box of a brand of cereals contains a coupon, and there are* **m** *different types of coupons, what is the probability that more than* **n** *boxes need to be bought to collect all* **m** *coupons?"*

- Euler and Laplace proved that when the coupons are equally likely, the expected number of boxes needed grows as $\mathcal{O}(m\log(m))$.

The Coupon Collector's Problem (a brief history)

Problem: *"If each box of a brand of cereals contains a coupon, and there are* **m** *different types of coupons, what is the probability that more than* **n** *boxes need to be bought to collect all* **m** *coupons?"*

- Euler and Laplace proved that when the coupons are equally likely, the expected number of boxes needed grows as $\mathcal{O}(m\log(m))$.
- In 1954 Hermann Von Schelling obtained the expected number of boxes when the coupons are not equally likely. In this case the expected waiting time is $\sum_{k=0}^{m-1}(-1)^{m-1-k}\sum_{|J|=k}\frac{1}{1-p_J}$, where J is a subset of $[m]$ and p_J denotes the probability of getting any coupon from J. [Sch54]

More Questions

Question 1: how many sightings will it take before we have observed every type of bird? ✓

More Questions

Question 1: how many sightings will it take before we have observed every type of bird? ✓

Question 2: what is the likelihood that a pair of species passed through the area at the same time?

More Questions

Question 1: how many sightings will it take before we have observed every type of bird? ✓

Question 2: what is the likelihood that a pair of species passed through the area at the same time?

Question 3: what is the best time to go bird watching? Is there a time where we are most likely to see the greatest number of species?

Random Interval Graph Model

Random Interval Graph Model

- Recall, in the coupon collector problem we assume the sightings come as a sequence of i.i.d. random variables $Y_1, Y_2, \ldots$, however we want a model that can account for seasonal changes in distributions.

Random Interval Graph Model

- Recall, in the coupon collector problem we assume the sightings come as a sequence of i.i.d. random variables $Y_1, Y_2, \ldots$, however we want a model that can account for seasonal changes in distributions.
- Therefore, our first modeling choice is that our observations are samples from a *stochastic process* Y with indexing set $[0, T] \subset \mathbb{R}$ and state space $[m]$.
- When we conduct an observation at some time $t_0 \in [0, T]$, we are taking a sample of the random variable Y_{t_0}.

Random Interval Graph Model

For each $i \in [m]$, the probabilities that $Y_t = i$ give us a function from $[0, T] \to [0, 1]$, which we call the *rate function* of Y corresponding to i.

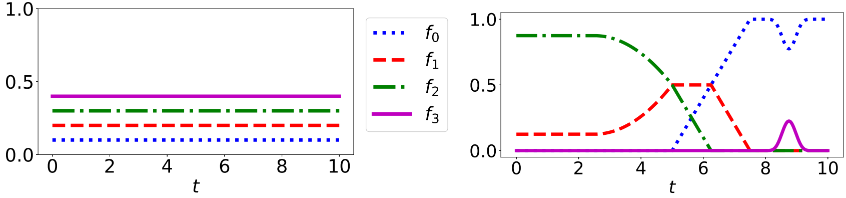
Definition (Rate function)

Let $Y = \{Y_t : t \in [0, T]\}$ be a stochastic process with indexing set $I = [0, T]$ and state space $S = [m]$. The *rate function* corresponding to label $i \in S$ in this process is the function $f_i : I \to [0, 1]$ given by

$$f_i(t) = P(Y_t = i) = P(\{\omega : Y(t, \omega) = i\}).$$

If f_i is constant for all $i \in [m]$, we say the process Y is *stationary*.

Random Interval Graph Model



(a) Stationary (b) Non-Stationary

Figure: Two examples of hypothetical rate functions.

Observe that at a fixed time t_0, the values $f_i(t_0)$ sum to 1 and thus determine the probability density function of Y_{t_0}.

Random Interval Graph Model

- Note that the set of times where species i might be present is exactly the *support* of the rate function f_i. Therefore, **our key problem is to estimate the support of the rate functions from finitely many samples**.

Random Interval Graph Model

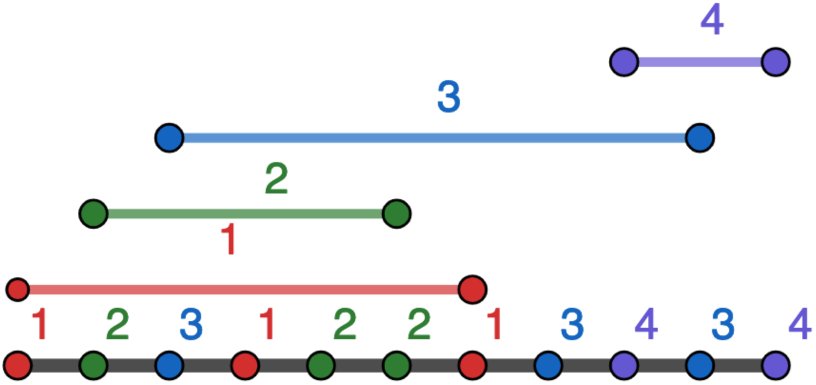
- Note that the set of times where species i might be present is exactly the *support* of the rate function f_i. Therefore, **our key problem is to estimate the support of the rate functions from finitely many samples**.
- This brings us to our next modeling choice: *we assume the rate functions f_i have connected support for all $i \in [m]$*.

Random Interval Graph Model

- Note that the set of times where species i might be present is exactly the *support* of the rate function f_i. Therefore, **our key problem is to estimate the support of the rate functions from finitely many samples**.
- This brings us to our next modeling choice: *we assume the rate functions f_i have connected support for all $i \in [m]$.*
- Now $\text{supp}(f_i)$ is a sub-interval of $[0, T]$. This fact provides a natural way of approximating the support of f_i: given a sequence of observations $Y_{t_1}, Y_{t_2}, \ldots, Y_{t_n}$ with $0 \le t_1 < t_2 < \ldots < t_n \le T$, let $I_n(i)$ denote the sub-interval of $[0, T]$ whose endpoints are the first and last times t_k for which $Y_{t_k} = i$. Explicitly,

$$I_n(i) = \text{Conv}(t_j : Y_j = i, j \le n).$$

Random Interval Graph Model



Random Interval Graph Model

- We call the interval $I_n(i)$ the *empirical support* of f_i, as it is an approximation of $\text{supp}(f_i)$ taken from a random sample. Note that it is possible for $I_n(i)$ to be empty or a singleton.

Random Interval Graph Model

- We call the interval $I_n(i)$ the *empirical support* of f_i, as it is an approximation of $\operatorname{supp}(f_i)$ taken from a random sample. Note that it is possible for $I_n(i)$ to be empty or a singleton.
- Now, the birdwatching questions from earlier may be expressed in terms of the empirical supports as follows:

1. *How many observations are required before we can expect all the empirical supports are non-empty?*
2. *What are the chances that a particular pair of empirical supports $I_n(i)$ and $I_n(j)$ intersect?*
3. *What is the greatest number of empirical supports that mutually intersect?*

Random Interval Graph Model

To make these questions even easier to analyze, we will present a combinatorial object: an *interval graph* that records the intersections of the intervals $I_n(i)$ in its edge set.

Definition

Given a finite collection of m intervals on the real line, its corresponding *interval graph*, $G(V, E)$, is the simple graph with m vertices, each associated to an interval, such that an edge $\{i, j\}$ is in E if and only if the associated intervals have a nonempty intersection, i.e., they overlap.

Random Interval Graph Model

To make these questions even easier to analyze, we will present a combinatorial object: an *interval graph* that records the intersections of the intervals $I_n(i)$ in its edge set.

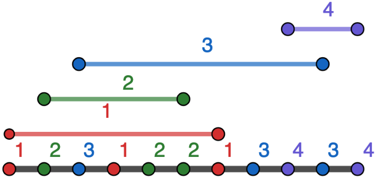
Definition

Given a finite collection of m intervals on the real line, its corresponding *interval graph*, $G(V,E)$, is the simple graph with m vertices, each associated to an interval, such that an edge $\{i,j\}$ is in E if and only if the associated intervals have a nonempty intersection, i.e., they overlap.

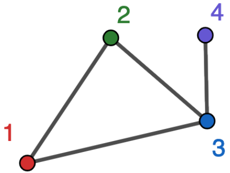
Definition

Let $\mathcal{F} = \{F_1, \ldots, F_m\}$ be a family of convex sets in $\mathbb{R}^d$. The *nerve complex* $\mathcal{N}(\mathcal{F})$ is the abstract simplicial complex whose k-facets are the $(k+1)$-subsets $I \subset [m]$ such that $\bigcap_{i \in I} F_i \neq \emptyset$.

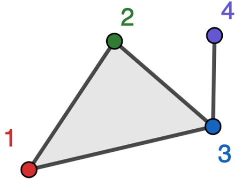
Random Interval Graph Model



(a) Labeled observations and induced intervals (empirical supports)



(b) Interval Graph



(c) Nerve Complex (Empirical Nerve)

Figure: Example observations with their corresponding graph and nerve.

Random Interval Graph Model

- By construction, the interval graph G is exactly the 1-skeleton of the nerve complex $\mathcal{N}$ generated by the intervals.

Random Interval Graph Model

- By construction, the interval graph G is exactly the 1-skeleton of the nerve complex $\mathcal{N}$ generated by the intervals.
- Recall **Helly's Theorem** [Bar02] in dimension 1: if a collection of convex sets in $\mathbb{R}$ all intersect pairwise, then the whole collection has a non-empty intersection.

Random Interval Graph Model

- By construction, the interval graph G is exactly the 1-skeleton of the nerve complex $\mathcal{N}$ generated by the intervals.
- Recall **Helly's Theorem** [Bar02] in dimension 1: if a collection of convex sets in $\mathbb{R}$ all intersect pairwise, then the whole collection has a non-empty intersection.

Lemma

Higher dimensional faces of $\mathcal{N}_n(Y)$ are exactly cliques in the 1-skeleton (interval graph). Hence the empirical nerve is completely determined by the interval graph and vice-versa.

Note: this *only* holds in dimension 1.

Random Interval Graph Model, Summary

1. We let $Y = \{Y_t : t \in [0, T]\}$ be a stochastic process as above and let $\mathcal{P} = \{t_1, t_2, ..., t_n\}$ be a set of n distinct observation times or sample points in $[0, T]$ with $t_1 < t_2 < \ldots < t_n$.
2. Let $Y = (Y_1, Y_2, \ldots, Y_n)$ be a random vector whose components Y_i are samples from Y where $Y_i = Y_{t_i}$, so each Y_i takes values $\{1, \ldots, m\}$.
3. For each label i we define the (possibly empty) interval $I_n(i) = \text{Conv}(\{t_j \in \mathcal{P} : Y_j = i\})$, which we refer to as the *empirical support* of label i.
4. Furthermore, because it comes from the n observations or samples, we call the nerve complex, $\mathcal{N}(\{I_n(i) : i = 1, \ldots m\})$, the *empirical nerve* of Y and denote it $\mathcal{N}_n(Y)$.

Most General Results

Theorem

Let $(t)_{n\in\mathbb{N}}$ *be a dense sequence in* $[0, T]$. *If* $\mu(\mathrm{supp}(f_i) \cap \mathrm{supp}(f_j)) > 0$, *then as* $n \to \infty$, $P(I_n(i) \cap I_n(j) \neq \emptyset) \to 1$.

Most General Results

Theorem

Let $(t)_{n\in\mathbb{N}}$ *be a dense sequence in* $[0, T]$*. If* $\mu(\text{supp}(f_i) \cap \text{supp}(f_j)) > 0$*, then as* $n \to \infty$*,* $P(I_n(i) \cap I_n(j) \neq \emptyset) \to 1$*.*

- You can give more specific bounds by making additional assumptions on Y.
- Our paper [JDLTH] pays special attention to the stationary case, where all the rate functions are constant. This situation is like the classical coupon collector problem, but asks more nuanced questions about the sequence of coupons.

Some Highlights for the Stationary Case

Theorem

Let $Y = (Y_1, \ldots, Y_n)$ be a random vector whose components are i.i.d. random variables such that $P(Y_j = i) = p_i > 0$ for all $i \in [m]$. Let $\mathcal{N}_n = \mathcal{N}_n([n], Y)$ denote the empirical nerve of the random coloring induced by Y, and let ω be a random variable equal to the clique number of $\mathcal{N}_n$, i.e., the size of the largest clique in $\mathcal{N}_n$. Then

$$\mathbb{E}\,\omega \geq \sum_{i=1}^{m} \left(1 - [(1-p_i)^{\lceil \frac{n}{2} \rceil} + (1-p_i)^{n - \lceil \frac{n}{2} \rceil + 1} - (1-p_i)^n]\right).$$

Proof Sketch:

Some Highlights for the Stationary Case

Theorem

Let $Y = (Y_1, \ldots, Y_n)$ be a random vector whose components are i.i.d. random variables such that $P(Y_j = i) = p_i > 0$ for all $i \in [m]$. Let $\mathcal{N}_n = \mathcal{N}_n([n], Y)$ denote the empirical nerve of the random coloring induced by Y, and let ω be a random variable equal to the clique number of $\mathcal{N}_n$, i.e., the size of the largest clique in $\mathcal{N}_n$. Then

$$\mathbb{E}\,\omega \geq \sum_{i=1}^{m} \left(1 - [(1-p_i)^{\lceil \frac{n}{2} \rceil} + (1-p_i)^{n - \lceil \frac{n}{2} \rceil + 1} - (1-p_i)^n] \right).$$

Proof Sketch:

- Let $f(x) = 1 - [(1-p_i)^x + (1-p_i)^{n-x+1} - (1-p_i)^n]$ and note $f(x) = P(x \in I_n(i))$. Observe $f(x)$ is maximized over $[n]$ at $x^* = \lceil \frac{n}{2} \rceil$.

Some Highlights for the Stationary Case

Theorem

Let $Y = (Y_1, \ldots, Y_n)$ be a random vector whose components are i.i.d. random variables such that $P(Y_j = i) = p_i > 0$ for all $i \in [m]$. Let $\mathcal{N}_n = \mathcal{N}_n([n], Y)$ denote the empirical nerve of the random coloring induced by Y, and let ω be a random variable equal to the clique number of $\mathcal{N}_n$, i.e., the size of the largest clique in $\mathcal{N}_n$. Then

$$\mathbb{E}\,\omega \geq \sum_{i=1}^{m} \left(1 - [(1-p_i)^{\lceil \frac{n}{2} \rceil} + (1-p_i)^{n - \lceil \frac{n}{2} \rceil + 1} - (1-p_i)^n]\right).$$

Proof Sketch:

- Let $f(x) = 1 - [(1-p_i)^x + (1-p_i)^{n-x+1} - (1-p_i)^n]$ and note $f(x) = P(x \in I_n(i))$. Observe $f(x)$ is maximized over $[n]$ at $x^* = \lceil \frac{n}{2} \rceil$.
- Let $X_i = 1_{\{x^* \in I_n(i)\}}$ for $i \in [m]$. Then,

$$\omega \geq \sum_{i=1}^{m} X_i \implies \mathbb{E}\omega \geq \sum_{i=1}^{m} \mathbb{E}X_i$$

Some Highlights for the Stationary Case

Corollary

The probability that all intervals intersect, tends to 1 as the number of samples n tends to infinity.

Some Highlights for the Stationary Case

Corollary

The probability that all intervals intersect, tends to 1 as the number of samples n tends to infinity.

Theorem

Let $Y = Y_1, Y_2, \ldots$ *be a sequence i.i.d. random variables such that* $P(Y_j = i) = p_i > 0$ *for all* $i \in [m]$*. Let* $\mathcal{N}_n = \mathcal{N}_n([n], Y)$ *denote the empirical nerve of the random coloring induced by the first n terms. Now let* X *be the random variable for the waiting time until* $\mathcal{N}_n = \Delta_{m-1}$*. Then,*

$$\mathbb{E}X \leq 2\int_0^\infty \left(1 - \prod_{i=1}^{m}(1 - e^{-p_i x})\right) dx.$$

Moreover, in the uniform case where $P(Y_j = i) = \frac{1}{m}$ *for all* $i \in [m]$*, we have that* $\mathbb{E}X \leq 2m\sum_{i=1}^{m}\frac{1}{i} = O(mlog(m))$*.*

Connections to Convex Geometry

- Tverberg's theorem states that if finite set of point $S \subset \mathbb{R}^d$ has cardinality $|S| \geq (m-1)(d+1)+1$, for a positive integer m, then S can be partitioned into m sets $S_1, \ldots, S_m$ in such a way that $\cap_{i=1}^{m} \mathrm{Conv}(S_i) \neq \emptyset$ [Tve81]; such a partition is called a *Tverberg partition*.

Connections to Convex Geometry

- Tverberg's theorem states that if finite set of point $S \subset \mathbb{R}^d$ has cardinality $|S| \geq (m-1)(d+1)+1$, for a positive integer m, then S can be partitioned into m sets $S_1, \ldots, S_m$ in such a way that $\cap_{i=1}^{m} \text{Conv}(S_i) \neq \emptyset$ [Tve81]; such a partition is called a *Tverberg partition*.
- Cover's theorem [Cov65] states that if n data points $x_i \in \mathbb{R}^d$ are partitioned into two classes independently at random with equal probability then the probability that the resulting partition is Tverberg goes to 1 or 0 depending on the asymptotic behavior of the ratio $\frac{n}{d}$.

Connections to Convex Geometry

- Tverberg's theorem states that if finite set of point $S \subset \mathbb{R}^d$ has cardinality $|S| \geq (m-1)(d+1)+1$, for a positive integer m, then S can be partitioned into m sets $S_1, \ldots, S_m$ in such a way that $\cap_{i=1}^{m} \text{Conv}(S_i) \neq \emptyset$ [Tve81]; such a partition is called a *Tverberg partition*.
- Cover's theorem [Cov65] states that if n data points $x_i \in \mathbb{R}^d$ are partitioned into two classes independently at random with equal probability then the probability that the resulting partition is Tverberg goes to 1 or 0 depending on the asymptotic behavior of the ratio $\frac{n}{d}$.
- J.A. De Loera and T. Hogan proved a lower bound on the likelihood that a uniformly random m-partition of n points in $\mathbb{R}^d$ is Tverberg [DLH20]. Their bound requires $O(m \log(m) \log(\log(m)))$ points.
- **Our bounds work for non-uniform colorings and improve De Loera and Hogan's bound in the uniform case, from $O(m \log(m) \log(\log(m))$ to just $O(m \log(m)$.**

Further Directions

In future work, we plan to study the following:

- Which results from our first paper can we extend to other special cases? For example, if the rate functions are assumed to be piecewise linear or Gaussians?
- Which results can be extended to higher dimensions? Note the interval graph and nerve complex are no longer equivalent.
- Note that many sequences of observations can lead to the same nerve complex, can we count them, i.e., for a given nerve complex how many distinct sequences of observations can produce that nerve?*

Acknowledegement and References I

I gratefully acknowledge partial support from NSF DMS-grant 1818969 from the NSF-AGEP supplement.

A. Barvinok, *A course in convexity*, vol. 54, American Mathematical Society, 2002.

T. M. Cover, *Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition*, IEEE Transactions on Electronic Computers **EC-14** (1965), no. 3, 326–334.

J.A. De Loera and T. Hogan, *Stochastic tverberg theorems with applications in multiclass logistic regression, separability, and centerpoints of data*, SIAM Journal on Mathematics of Data Science **2** (2020), 1151–1166.

D. Oliveros J. De Loera, E. Jaramillo-Rodriguez and A. Torres-Hernandez, *A model for birdwatching and other chronological sampling activities*.

H. Von Schelling, *Coupon collecting for unequal probabilities*, The American Mathematical Monthly **61** (1954), no. 5, 306–311.

H. Tverberg, *A generalization of Radon's theorem. II*, Bull. Austral. Math. Soc. **24** (1981), no. 3, 321–325. MR 647358 (83d:52009)