WXML Final Report: Chinese Restaurant Process

Dr. Noah Forman, Gerandy Brito,
Alex Forney, Yiruey Chou, Chengning Li

Spring 2017

1 Introduction

The Chinese Restaurant Process (CRP) generates random partitions of the natural numbers and has applications in clustering algorithms in machine learning and in biological modeling. The first goal of this project is to learn about stochastic branching processes, the CRP, and the ordered CRP with reseating. The second goal is to program a simulation to generate images and an animation of a surprising visual representation of the ordered CRP with reseating.

2 Theory

2.1 Exchangeability

2.1.1 Polya's Urn

We begin our investigation by considering Polya's urn model. Imagine that we have an urn containing n colored balls, k of which are green and $n-k$ are blue. We can see that the probability of randomly selecting a green ball from the urn is k/n and the probability of randomly selecting a blue ball is $(n-k)/n$. After actually randomly choosing a ball, we observe its color, put the ball back into the urn, and also put another ball of the same color into the urn. This process is then repeated.

Consider an example in which we draw three green balls and then two blue balls from the urn (we will denote such a draw sequence as GGGBBB). It is easily shown that the probability of obtaining such a sequence is

$$\left(\frac{1}{2}\right)\left(\frac{2}{3}\right)\left(\frac{3}{4}\right)\left(\frac{1}{5}\right)\left(\frac{2}{6}\right) = \frac{1}{60}.$$

Now, consider another draw sequence with the same number of green balls and blue balls, GBGBG. The probability of obtaining this sequence is

$$\left(\frac{1}{2}\right)\left(\frac{1}{3}\right)\left(\frac{2}{4}\right)\left(\frac{2}{5}\right)\left(\frac{3}{6}\right) = \frac{1}{60}.$$

Surprisingly, these probabilities are the same. In fact, Polya's urn model is exchangeable. This means that the probability of any particular sequence with k green balls and $n - k$ blue balls is

$$\frac{k!\,(n-k)!}{(n+1)!}.$$

One can ask whether the percentage of green draws out of the total number of draws converges to some number. To answer this question, we call upon the following theorem:

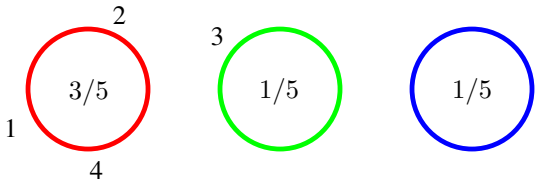
Theorem (de Finetti)**.** *The probability distribution of any infinite exchangeable sequence of Bernoulli random variables is a "mixture" of the probability distributions of independent and identically distributed sequences of Bernoulli random variables.*

A consequence of the theorem is that the limit of the percentage of green draws does indeed converge. In fact, this percentage is actually random with uniform $[0, 1]$ distribution.

2.1.2 Chinese Restaurant Process

We now consider a model much like Polya's urn, the Chinese Restaurant Process. To set the stage, imagine that we have a Chinese restaurant that has infinitely many tables that can each seat infinitely many customers. When a new customer, say the nth, enters the restaurant, the probability of them sitting at a table with m other customers is m/n, and the probability of this nth customer sitting alone at a new table is $1/n$. This seating rule can be visualized in the following example where $n = 5$:

Probabilities of customer 5 joining each table



It turns out that this process is also exchangeable. Specifically, the probability of having a particular seating configuration is not affected by seating order. One can see this demonstrated in the following example. Suppose that we have a sequence of 7 customers who sit at 5 tables in the order 1 1 1 2 2 3 4. The probability of obtaining this sequence is

$$(1)\left(\frac{1}{2}\right)\left(\frac{2}{3}\right)\left(\frac{1}{4}\right)\left(\frac{1}{5}\right)\left(\frac{1}{6}\right)\left(\frac{1}{7}\right) = \frac{1}{168}.$$

Alternatively, consider the same scenario with the sequence 1 2 3 4 1 2 1, which has the probability

$$(1)\left(\frac{1}{2}\right)\left(\frac{1}{3}\right)\left(\frac{1}{4}\right)\left(\frac{1}{5}\right)\left(\frac{1}{6}\right)\left(\frac{2}{7}\right) = \frac{1}{168}.$$

We can generalize this formulation of the CRP. Specifically, we can introduce a two parameter (α, θ)-Chinese Restaurant Process, for $\alpha, \theta \in \mathbb{R}$. These parameters, α and θ, are used in determining new seating rules. In this new formulation, the nth customer entering a restaurant with k non-empty tables will sit at a table with m customers with probability

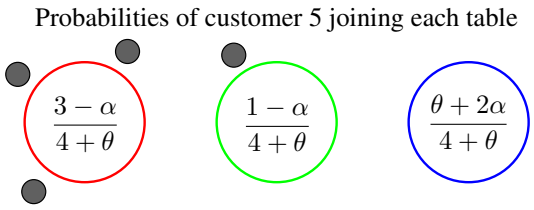
$$\frac{m - \alpha}{n - 1 + \theta}$$

and will sit alone with probability

$$\frac{\theta + k\alpha}{n - 1 + \theta}.$$

We will make the restrictions that $\alpha \in [0, 1)$ and $\theta > -\alpha$ in order to avoid negative probabilities and to keep our simulations interesting (cf. Section 3).

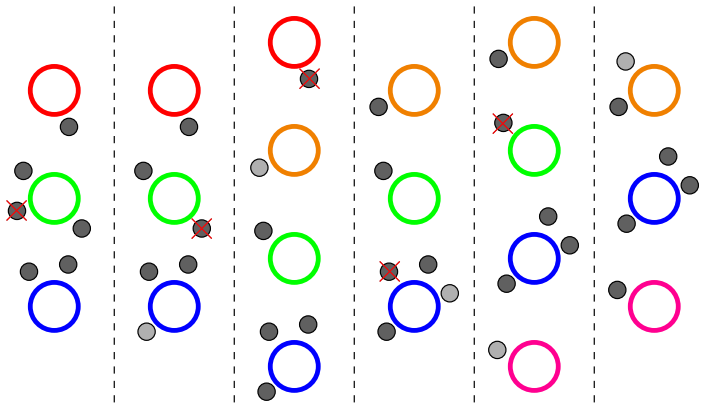
The above seating rule can be visualized in the following example where $n = 5$ and $k = 2$:



Notice that the initial formulation of the CRP had $\alpha = 0$ and $\theta = 1$.

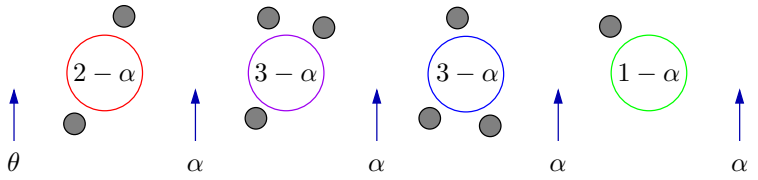
2.2 Chinese Restaurant Process with Reseating

As a further extension of the Chinese Restaurant Process, we now introduce *reseating*. As the name suggests, for a fixed number of customers in the restaurant, reseating involves a uniformly random customer leaving the restaurant, reentering, and being reseated according to the (α, θ)-CRP seating rule. Specifically, each customer leaves the restaurant with rate 1.



Notice that reseating allows for tables to "die off." That is, it is possible for every customer sitting at a given table to leave (the red table above demonstrates this phenomenon). In fact, given a restaurant with n customers, its population will decrease with exponential rate n and increase with exponential rate $n + \theta$. The individual table populations will evolve, too. A table in this restaurant with m people will also lose customers with exponential rate $m - \alpha$ and gain customers with exponential rate m.

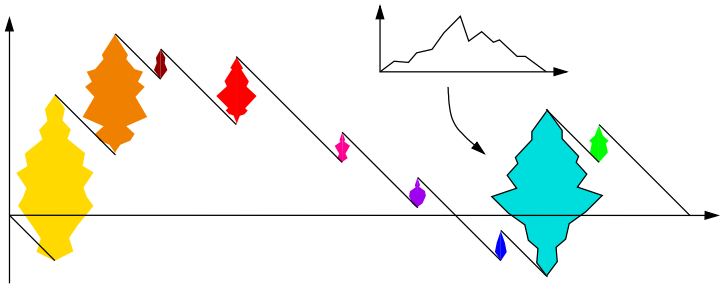
In order to observe this birth-and-death cycle of tables, we want to keep track of each table. Hence, we will randomly order the tables from left to right. Additionally, rather than only allowing new tables to be born on the right side, we will allow for customers to sit at new tables at any point between two other tables, as shown in the following illustration:



In fact, The limiting proportions of customers in the (α, θ)-CRP, listed in ranked order, is a random monotone decreasing sequence of positive real numbers that sum to 1. We say it has Poisson-Dirichlet distribution with parameters (α, θ).

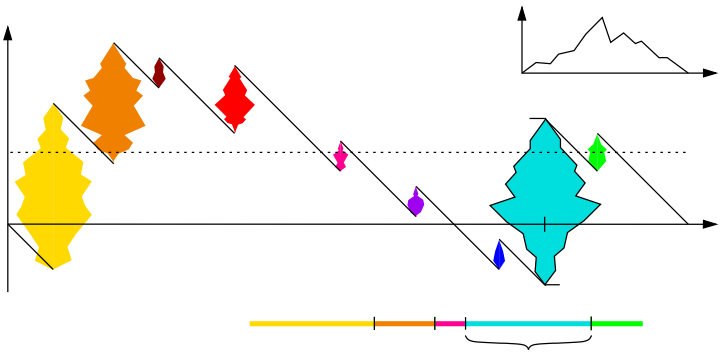
In 2009, Leonid Petrov [1] described the limit of evolution of the ranked sequences of table sizes in the (α, θ)-CRP with reseating. In 1981, Ethier and Kurtz [2] did this for $\alpha = 0$.

As we mentioned previously, one of the main goals of this project is to create a graphical representation of the CRP. This representation arises through what we will call *scaffolding* and *spindles*.



The scaffolding (black lines above) is a real-valued cádlág process (spectrally positive $Stable(3/2)$) of negative drift (or compensation) and positive jumps, where the jump times are exponentially distributed with rate 1. The heights of the jumps are given by the lifespans of tables. The leftmost jump corresponds to the leftmost table, and so on. The spindles (colored portions above) are continuous, $\mathbb{R}_+$-valued excursions with lifetime equal to the jump height. This means that the vertical component corresponds to the lifespan of a given table in the restaurant, while the horizontal component corresponds to the population of the table at a given time.

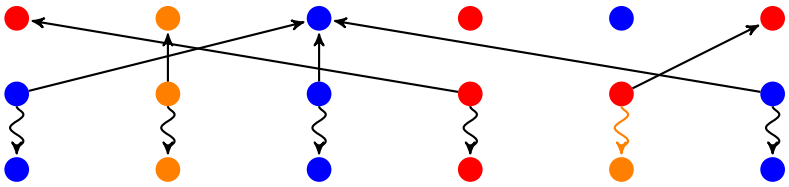
This graphical representation also gives us information about the distribution of the population of the restaurant, which is to say the table populations, at any given time. This comes through what we will call the *skewer* process.



For some $y \in \mathbb{R}$, we get the level y skewer by drawing a line across the picture at y. Then, from left to right, we collect the cross-section of the spindles and slide these values together, as if on a skewer, to remove gaps. As the skewer moves up, the interval partition evolves continuously. The cross-section sizes shrink and grow, which corresponds to the table populations shrinking and growing.

2.2.1 Moran Model

The Moran model is a stochastic process used in biological modeling. It describes how alleles change in some gene pool from generation to generation. In this model, the children choose their parent i.i.d.. The child will then inherit their parent's alleles. However, each child has some independent probability of mutating and adopting the alleles of some other parent that currently exists in the gene pool. This is to say that the child cannot mutate to a brand new genotype. In this model, we see genetic drift, where alleles become more and less common at random. There is a tendency, though, for common genes to stay common and uncommon genes to stay uncommon.

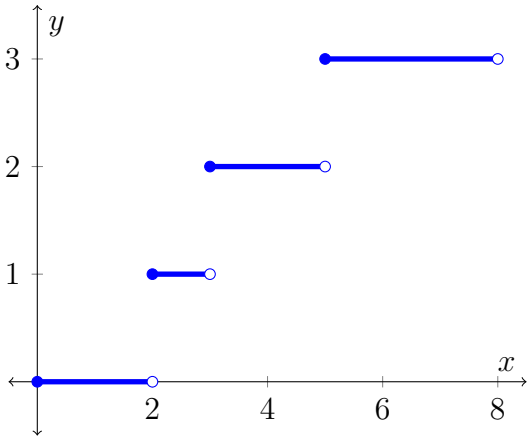


The Moran Model is similar to the CRP with reseating. For example, the children in the Moran model select types like new customers in the CRP randomly select tables. Also, in both models we see the property that common types self-perpetuate. Finally, the limit of the CRP with $\alpha = 0$ is close to the Miran model

with infinitely many types. One difference between the two, though, is that people choose tables one-by-one in the CRP, whereas all children choose their parents at once in the Moran model. Although mutation in the Moran model is similar to sitting alone in the CRP, there is always a chance to introduce a new table in the CRP, but the Moran model does not allow for the introduction of new genotypes.

2.3 Lévy Processes

We now turn out attention to Poisson processes, which model a number of "events" happening over time. Consider the process $(N(t), t \geq 0)$ shown below:

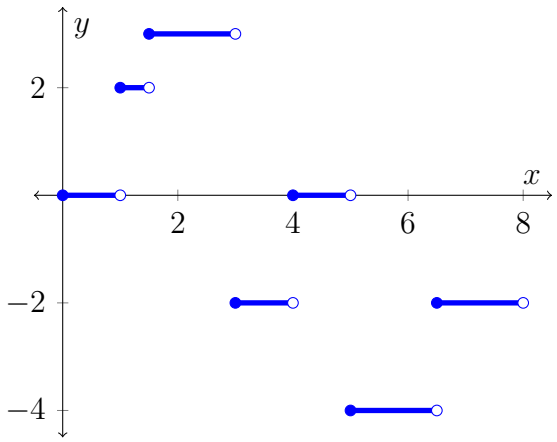


Here, $N(t)$ has Poisson distribution with rate $t\lambda$ for some number λ. Then,

$$\mathbb{P}\{N(t) = k\} = e^{-t\lambda}\frac{(t\lambda)^k}{k!}.$$

The interarrival times in this process are i.i.d., specifically exponentially distributed with rate λ.

The Poisson process can be further generalized. As before, we have "arrivals" or "hits" with rate λ, but instead of jumping by 1, we have i.i.d. jump sizes according to some probability distribution Λ. The generalized Poisson process can be visualized below:



We can also add Poisson processes. If $(N(t), t \geq 0)$ is a Poisson process with rate λ and $(M(t), t \geq 0)$ is a Poisson process with rate μ, then $(N(t) + M(t), t \geq 0)$ is a Poisson process with rate $\lambda + \mu$. We have a slightly different result for generalized Poisson processes. If $(N_1(t), t \geq 0)$ is a generalized Poisson process with rate λ_1 and jump distribution Λ_1 and $(N_2(t), t \geq 0)$ is a generalized Poisson process with rate λ_2 and jump distribution Λ_2, then $(N_1(t) + N_2(t), t \geq 0)$ is a generalized Poisson process with rate $\lambda_1 + \lambda_2$ and jump distribution

$$\frac{\lambda_1 \Lambda_2 + \lambda_2 \Lambda_1}{\lambda_1 + \lambda_2}.$$

Now, consider an infinite series of generalized Poisson processes where jumps come faster and faster but jump sizes get really small. This gives rise to a Lévy process, which is a stochastic process on $\mathbb{R}^n$ (for our purposes, just $\mathbb{R}$) with stationary, independent, increments. This process can be thought of as a continuum analogue to random walks. In fact, there are only three behaviors of this process:

1. Poisson jumps
2. Brownian motion
3. (Deterministic) drift (and "compensation")

The aforementioned scaffolding process is a certain Lévy process with downward drift (or compensation) and upward jumps.

2.4 Branching Processes

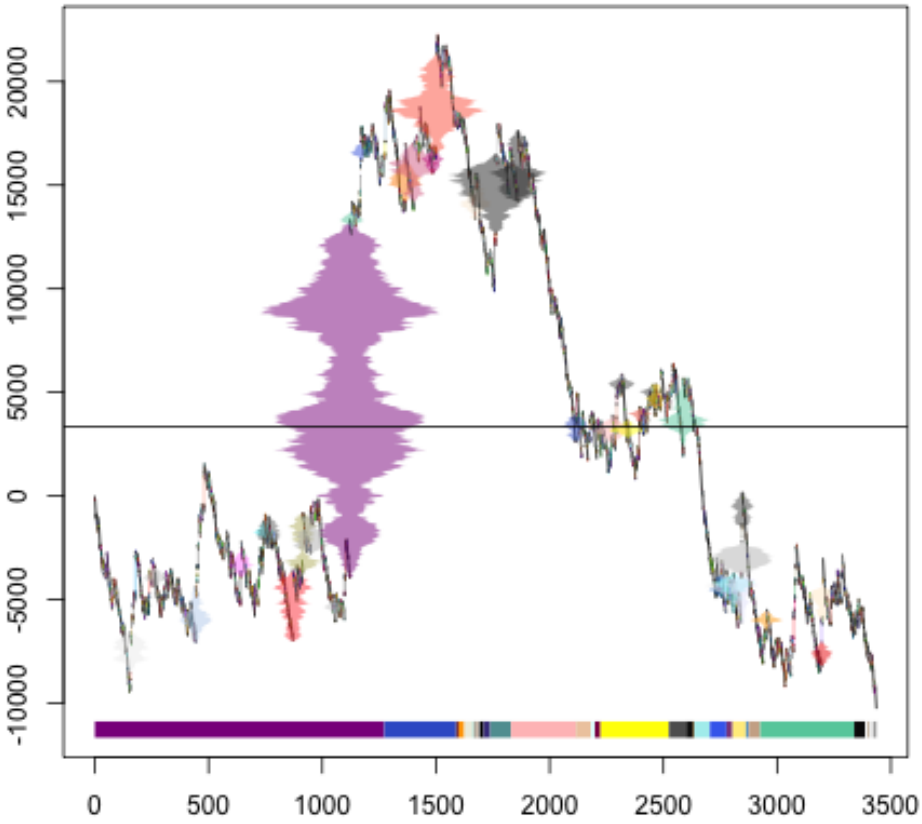
As we mentioned in Section 2.2, the lifetime of a table in the (α, θ)-CRP gives rise to a sort of "birth-and-death chain." It turns out that this chain actually converges to the squared Bessel process with dimension -1, or $BESQ(-1)$. Hence, our spindles are actually squared-Bessel excursions.

We finally turn our attention to the Galton-Watson process. The Galton-Watson process with offspring distribution μ, where μ is some probability distribution on nonnegative integers, serves as a way to model the extinction of certain populations (Francis Galton used it to model the death of family names). We start with some population, then at each time step, each individual dies and is replaced by a random number of offspring with distribution μ i.i.d..

The Galton-Watson process is significant to us because $BESQ(-1)$ can be approximated by the Galton-Watson process with emigration rate 1. Since the graphical representation of our model is reliant on generating spindles, the Galton-Watson process gives us an easy way to simulate the birth-and-death chain of the tables in our Chinese Restaurant.

3 Simulation

As we have mentioned, the ultimate goal of this project is to simulate the ordered CRP with reseating. So, we begin by attempting to simulate the three major components of our graphical representation: the scaffolding, the spindles, and the skewer. The end result of this simulation is an image such as the one below. (We also produced animated GIFs, but these could not be included in this report.)



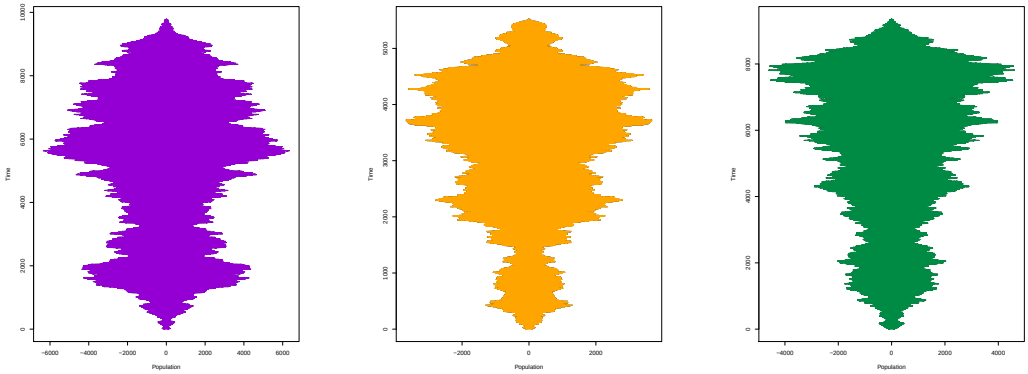
3.1 Spindles

Creating spindles amounts to simulating the Galton-Watson process until our initial population reaches 0. As discussed in Section 2.4, how we actually simulate the Galton-Watson process is important because we need our simulation to converge to $BESQ(-1)$. So, for the distribution μ, we sample from the binomial distribution $Bm(n, 1/2)$, where n is the current table population. For the emigration with rate 1, we sample from the Bernoulli distribution $Ber(1/2)$. The new table population is then $2 \cdot Bm(n, 1/2) - Ber(1/2)$, which gives us the desired

convergence to $BESQ(-1)$.

There is still an issue of an initial population, however. The initial table population should be 1, but this leads to an issue where the picture is not adequate because there is 50% chance that the table dies off in one time step. So, we choose an initial population of 100. This value is large enough to allow for large spindles to be created but also small enough that the spindles in the image appear to be generated from an initial population of 1.

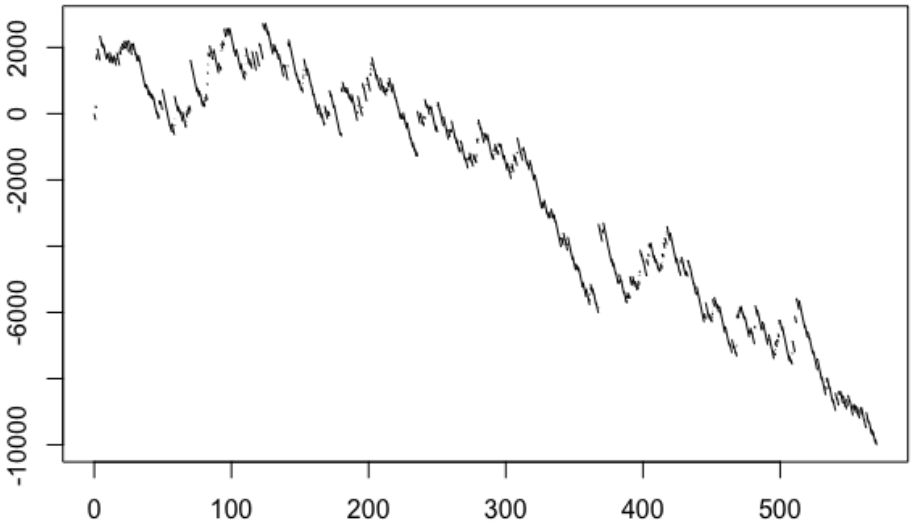
Some examples of carrying out this simulation can be seen below. The spindle colors are not significant; the colors are randomly chosen for the sake of making a better picture.



3.2 Scaffolding

In our preliminary simulations of spindles, we found that the average lifespan of a table was double its initial population. More specifically, a table with initial population 100 typically lasted for 200 time steps. Since the scaffolding is negative drift (or compensation) and positive jumps, we choose the slope of the negative drift to be -200 in order to "cancel out" the expected table lifespan.

As mentioned in Section 2.2, the x-coordinates of the jumps are exponentially distributed with rate 1. In order to simulate the scaffolding, then, we start at the origin and sample from the exponential distribution $Exp(1)$. We drift downward until reaching the random x-coordinate and then jump and repeat the drift. This process results in something looking like the image below.

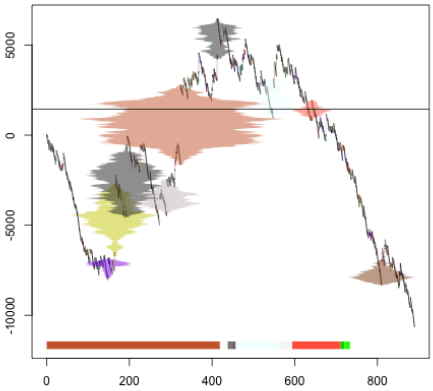


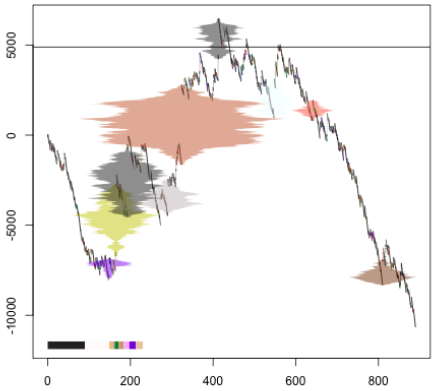
3.3 Skewer

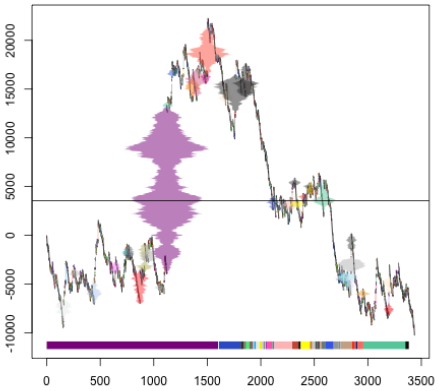
The actual calculation of the skewer is rather simple. When we simulate a spindle, we store the beginning and ending y-values of the spindle as well as the table populations at each time step. So, to create the skewer, we fix some nonnegative y-value and check if it was between the start and end points of each spindle. If it is, then we calculate where the fixed y-value lands inside the spindle itself. We then access the table population at that time step and plot double this value (since the spindles are reflected horizontally over the center) on the skewer. We then repeat this process for each spindle that contains the y-value.

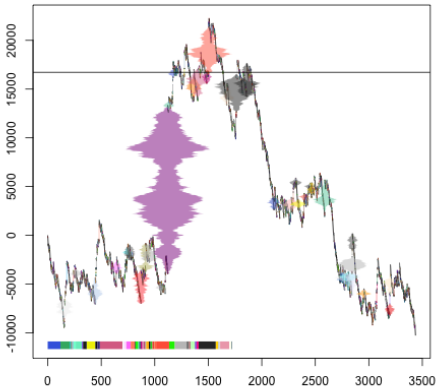
3.4 Pictures

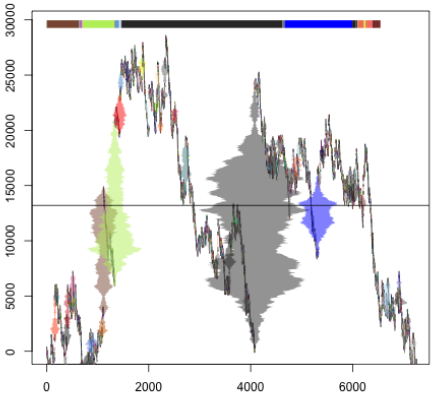
Combining all of these pieces, we were able to create nice visual representations of the ordered Chinese Restaurant Process with reseating.

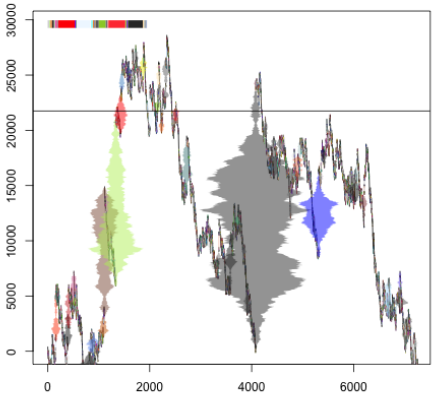

5000
0
-5000
-10000
0
200
400
600
800


5000
0
-5000
-10000
0
200
400
600
800


20000
15000
10000
5000
0
-5000
-10000
0
500
1000
1500
2000
2500
3000
3500


20000
15000
10000
5000
0
-5000
-10000
0
500
1000
1500
2000
2500
3000
3500


30000
25000
20000
15000
10000
5000
0
0
2000
4000
6000


30000
25000
20000
15000
10000
5000
0
0
2000
4000
6000

References

[1] L. Petrov. Two-parameter family of infinite-dimensional diffusions on the kingman simplex. *Functional Analysis and Its Applications*, 43(4):279–296, 2009.

[2] S. Ethier and T. Kurtz. The infinitely-many-neutral-alleles diffusion model. *Advances in Applied Probability*, 13(3):429–452, 1981.