JOINT INSTITUTE
交大密西根学院

VE414 • Bayesian Analysis • Project

Author: Chenmin Hou, Yanbo Song, Zhanpeng Zhou
Instructor: Jing Liu

Problem Statement

Jiuling trees, a kind of invisible plant, have been in the Forbidden Forest of Hogward for many years, which can produce a visible fruits, taye. With the help of an old spell, the positions of tayes in some trips can be recorded. Our project is to identify the number and locations of the trees given the information recorded by the travellers.

Main Assumption

1. Observed data are independent of obersevers
2. The location and number of the tayes remain the same at different time periods

Visualization of information

Our group plotted the trajectory of all the travellers from all trips in one figure. In the figure, the darker the color, the more tayes observed near that particular data point.

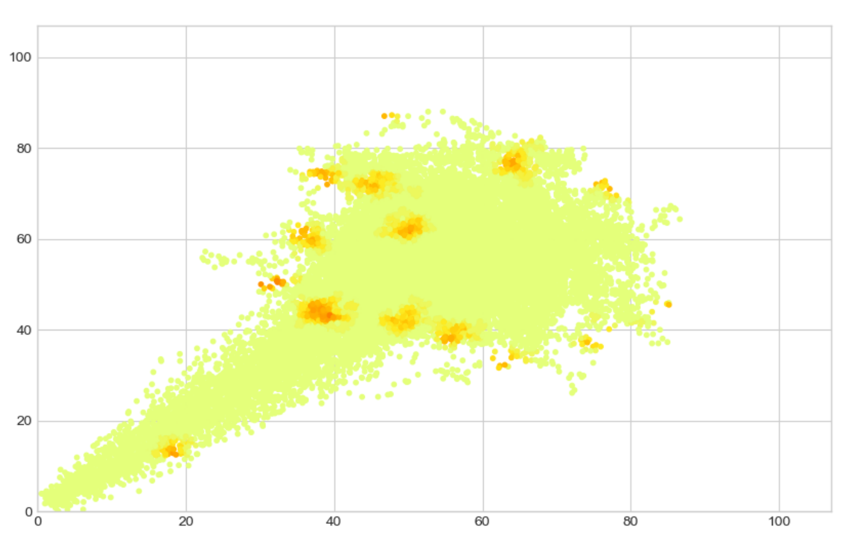

Fig. 1: Raw Data Visualization

Data Preprocessing

- Due to the inaccuracy of the spell, ignore all the data that tayes are not too far away

Data preprocessing is done to estimate the location of the tayes given the number of neary tayes information.

- Delete all the points on the map which have 0 tayes close to them.
- Divide the forest into square grids.
- For each grid, pick out the points p that lies in this grid, which forms a set of points P.
- Define the centroid position of this grid as the weighted average of the locations of points $p \in P$.
- Generate a number n of random samples in the circle centered at the centroid, where n is defined as the average number of tayes close to the point $p \in P$.
- Treat these generated points as estimates of the tayes' position, shown in the following figure.

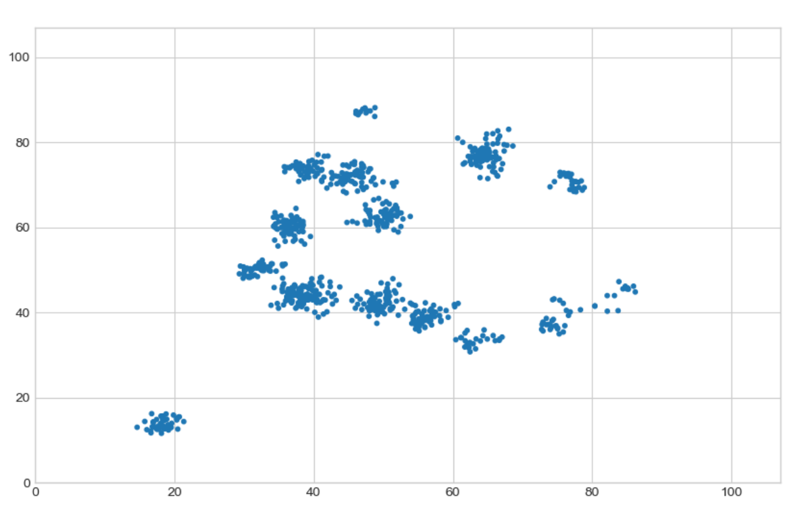

Fig. 2: Generated Tayes' location

Modeling by GMM and EM

Assumption

- Tayes location of Jiuling centered at $\vec{c} \sim$ Bivariate Gaussian Distribution$(\vec{c}, \Sigma)$
- The covariance matrix of the Bivariate Gaussian Distribution is $\begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$

We use GMM and EM to fit the generated samples data of tayes' positions. Given n samples and assume k clusters, we should find $\mu_1, \cdots \mu_k$ and $\sigma_1, \cdots, \sigma_k$ to maximize $\mathcal{L}_W = \Pi_{i=1}^{n}(\Sigma_{j=1}^{k} W_{i,j} P(X_i|\mu_j, \sigma_j))$, where $P(X_i|\mu_j, \sigma_j)$ follows Gaussian Distribution and W is a $n \times k$ matrix. Then we should repeat the following step

- Expectation: to udpate W, where

$$W_{i,j} = \frac{\pi_j P(X_i|\mu_j, var_j)}{\Sigma_{m=1}^{k} \pi_m P(X_i|\mu_j, var_m)}$$

$$\pi_j = \frac{\Sigma_{i=1}^{n} W_{i,j}}{\Sigma_{j=1}^{k} \Sigma_{i=1}^{n} W_{i,j}}$$

- Maximization: to update μ, σ until the log likelihood converges, where

$$\mu_{j,k} = \frac{\Sigma_{i=1}^{n} W_{i,j} X_{i,k}}{\Sigma_{i=1}^{n} W_{i,j}}$$

$$var_{j,k} = \frac{\Sigma_{i=1}^{n} W_{i,j}(X_{i,k} - \mu_{j,k})^2}{\Sigma_{i=1}^{n} W_{i,j}}$$

To find the optimal number of clusters of GMM, we compare the values of BIC of different results using different number of clusters.

$$BIC = k \times ln(n) - 2ln(L)$$

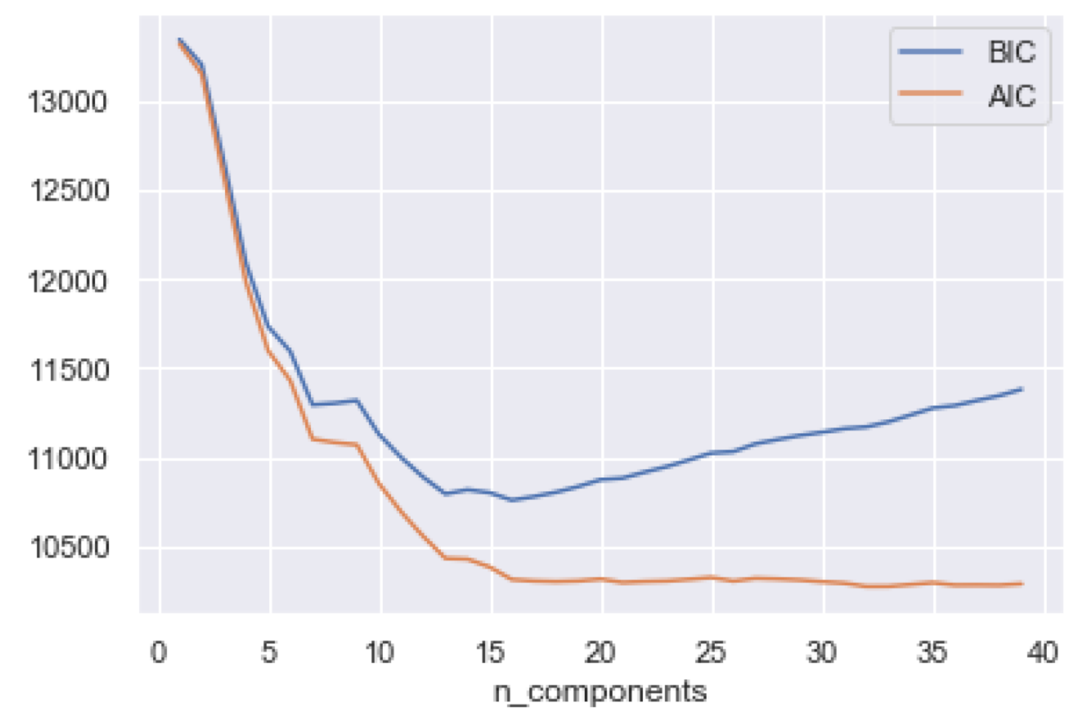

Fig. 3: BIC changes with k

Then we decide to use 15 as the number of cluster which gives the minimum of BIC and is just the number of Jiuling in observed area.

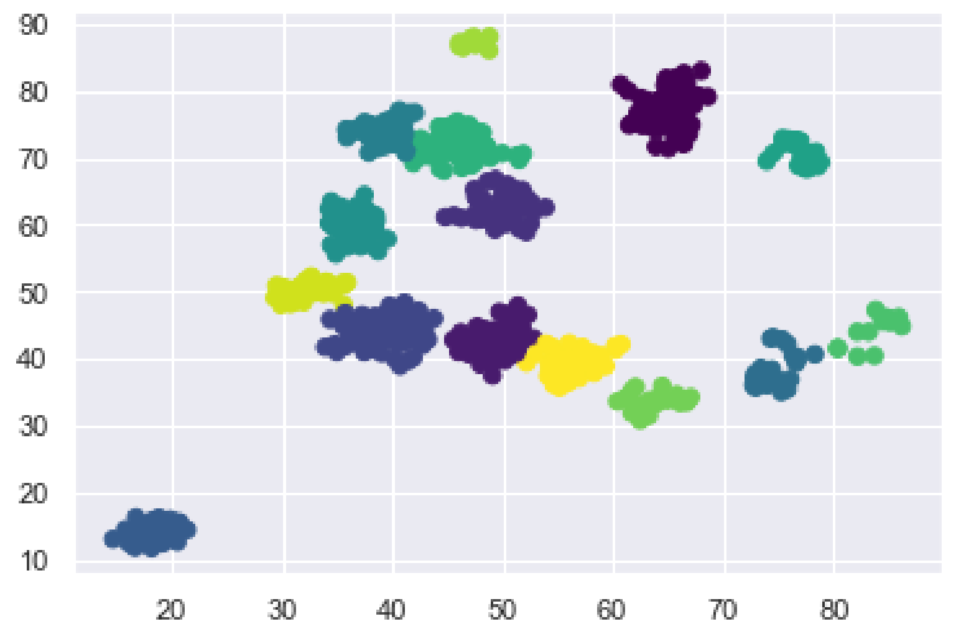

Fig. 4: GMM Result with 15 clusters.

Conclusion

Assumption

- the total number of Jiuling follows Possion distribution with $\lambda = p \cdot Total\ Area$

After we get the number of Jiulings by observed data, we can approximate the probability of finding a jiuling in the forest as $p = \frac{Number\ of\ Clusters}{Observed\ Area}$.

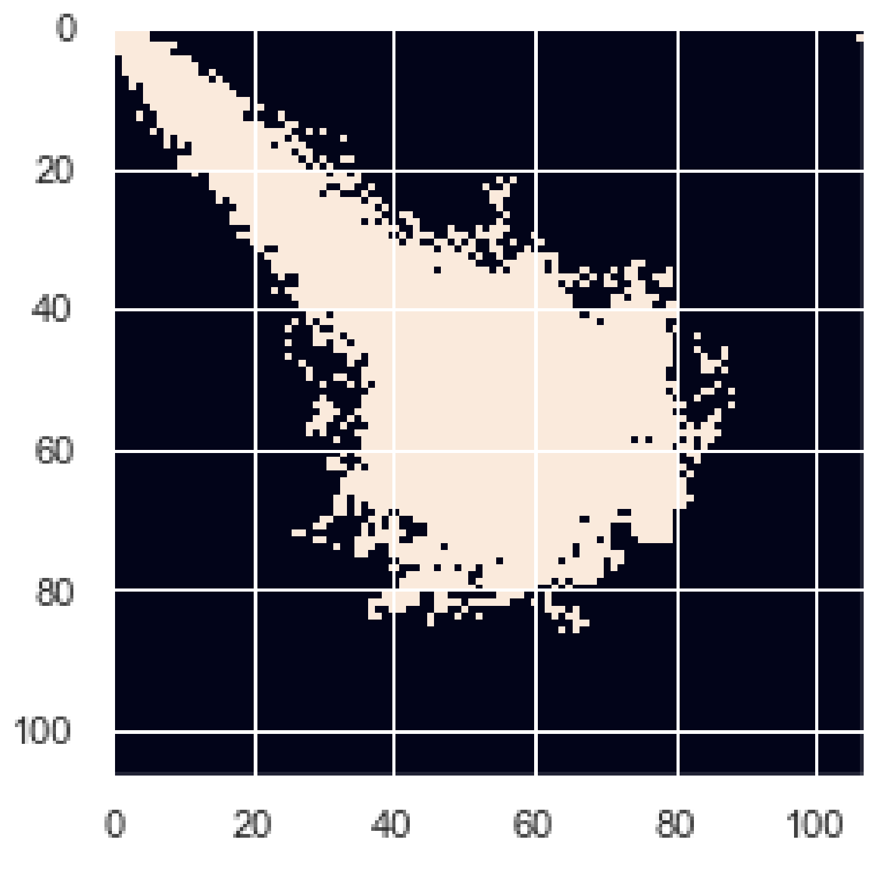

Fig. 5: The Estimated Observed Area by Discretizing 2D Space

Thus, we can a point estimate of total number of Jiulings in the forest bt taking the Possion model, which is 62.

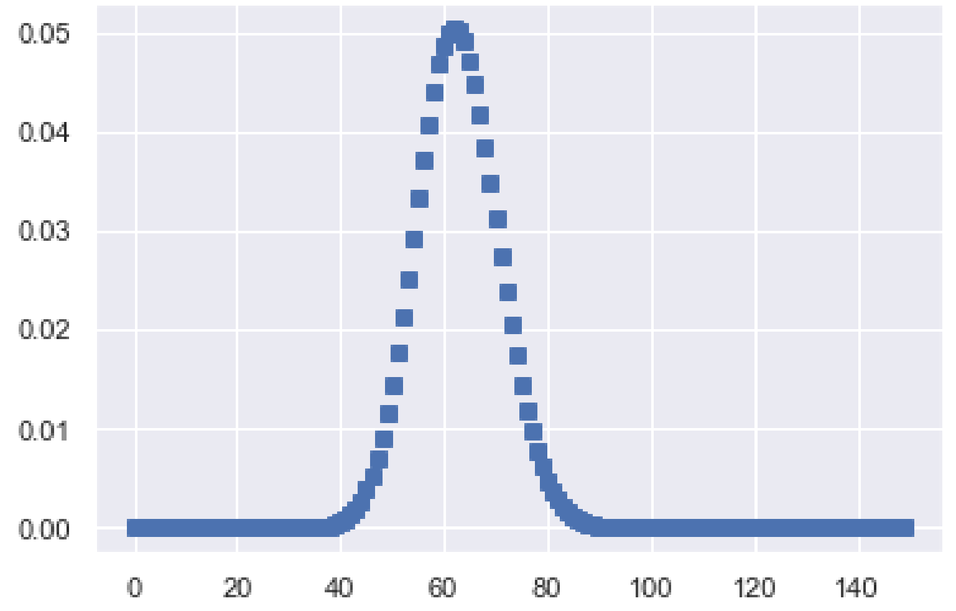

Fig. 6: Possion Distribution with λ

University of Michigan - Shanghai Jiao Tong University Joint Institute