VE414 Presentation

Group 17

August 2020

Main Assumptions

- Observed data are independent of observers
- The location and number of the Tayes remain the same at different time periods
- Tayes location of Jiuling centered at $\vec{c} \sim$ Bivariate Gaussian Distribution$(\vec{c}, \Sigma)$
- The total number of Jiuling follows Possion distribution with $\lambda = p \cdot Total\ Area$

Data Preprocessing

Data Preprocessing: Method 1

- Consider all recordings made by all three people in all 49 trips together.
- Keep only the recordings that has at least 1 close (within $1m$) Tayes.
- Apply EM + GMM algorithm directly to these points

Data Preprocessing: Method 1

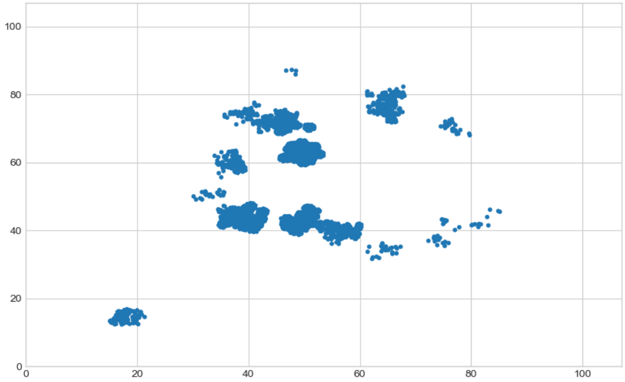

Figure: Points fed into GMM by Method 1

Data Preprocessing: Method 2

- Keep only the recordings that has at least 1 close-by (within $1m$) Tayes.
- Divide the whole 107×107 region to grids. Each with same height and width $n \times n$.
- Group recording points by the grid that they belong to.
- For each grid, denote the set of all points belong to this grid P. Compute the centroid as,

$$x = \frac{\sum_{p \in P} x_i}{\sum_{p \in P} w_i} \quad y = \frac{\sum_{p \in P} y_i}{\sum_{p \in P} w_i},$$

  where x_i and y_i are the x, y coordinate of the point p, and w_i is the number of close-by Tayes of this point.
- Apply EM + GMM algorithm directly to these centroids.

Data Preprocessing: Method 2

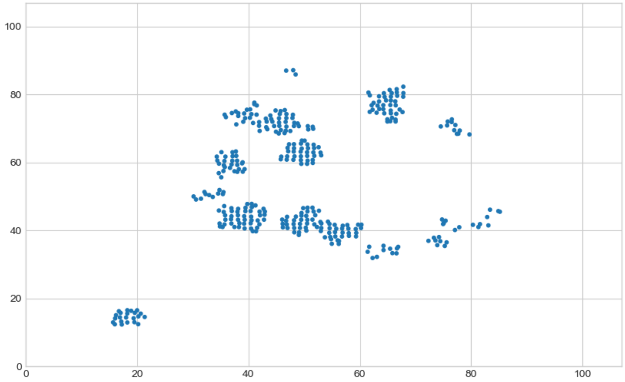

Figure: Points fed into GMM by Method 2

Data Preprocessing: Method 3

- Compute the centroids of each grid like Method 2.
- Generate m ramdom samples in the unit circle centered at this grid, m equals to the average number of close-by Tayes in this grid. They are estimates of the Tayes' location.
- Apply EM + GMM to the generated samples.

Data Preprocessing: Method 3

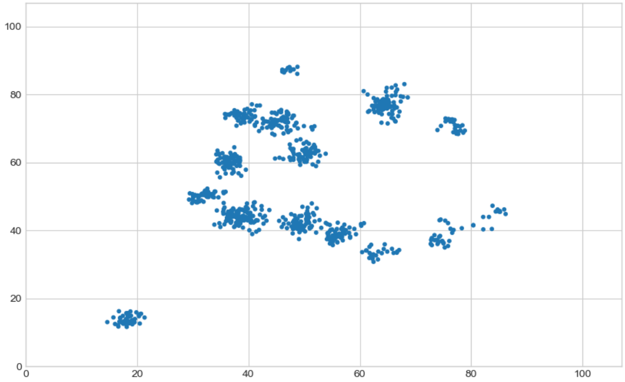

Figure: Points fed into GMM by Method 3

Data Preprocessing: Method 4

- Apply EM + GMM to the Tayes' location (assume we know it)
- This method is a benchmark. Theoretically, we cannot do better than this method.

Definition of the metric

- We need to define a metric that measures quantitatively how "close" our result is to the real result.
- Let the set of real Jiuling locations as T_1, the set of estimated Jiuling locations as T_2.
- For each $t_i \in T_1$, compute the closest distance in the set T_2, denote it as d,

$$d_i = \min_{t' \in T_2} \|t_i - t'\|_2$$

- We define D as the metric operator,

$$D : T_1 \times T_2 \to \mathbb{R}$$

by computing d_i, for each $t_i \in T_1$, then take the median of all the d_i.

Comparison of the Data Preprocessing methods

- Randomly initialize the Jiulings' locations.
- Generate Tayes' locations based on our assumptions.
- Do this 100 times for each pair of grid size n, and gaussian variance σ^2
- Evaluate the result provided by the 4 methods

	X1	X2	X3	X4
Min.	:1.219	:0.5609	:0.4150	:0.4865
1st Qu.	:1.225	:0.7741	:0.5504	:0.4899
Median	:1.225	:0.7887	:0.6349	:0.5569
Mean	:1.248	:0.7923	:0.6335	:0.5436
3rd Qu.	:1.226	:0.8311	:0.6967	:0.5782
Max.	:1.590	:0.9552	:1.0015	:0.6705

- Among the first three methods, Method 3 is the best in terms of the metric we deined. Its performance is very close to the theoretically optimal one produced by Method 4.

Analysis using GMM

GMM with Expectation-Maximization algorithm (EM)

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 $\mathcal{L}_W = \prod_{i=1}^{n}(\sum_{j=1}^{k} W_{i,j} P(X_i|\mu_j, \sigma_j))$, 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

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

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

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

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

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

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

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

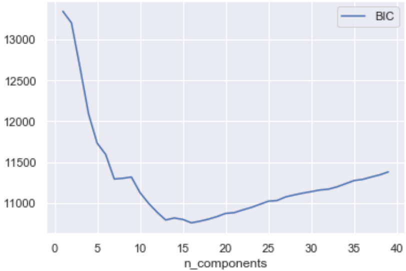

Figure: BIC changes with k

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.

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

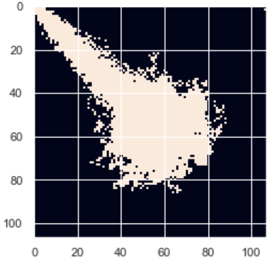

Figure: The Estimated Observed Area by Discretizing 2D Space

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

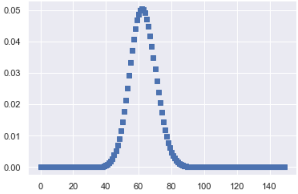

Figure: Possion Distribution with λ

The Position of Jiuling

We can easily get the position of Jiuling in the observed area. But we can not predict the position of Jiuling in unobserved area.

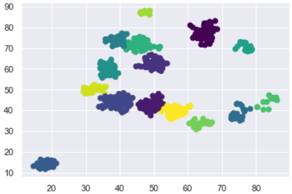

Figure: Position of Jiuling in oberved area

Question 3: Propose what we will need in order to address the main task if Jiuling can actually move!
Solution: The extra information to address this problem is the time when the spell was executed.

Thank You!