Mathematical Contest in Modeling with R

Jiang Qi

Department of Statistics
Renmin University of China

June 5, 2010

Brief Introduction

- MCM, short for Mathematical Contest in Modeling held by COMAP, is an Annual online competition for teams of undergraduates to use mathematical modeling to solve real world problems.

Brief Introduction

- MCM, short for Mathematical Contest in Modeling held by COMAP, is an Annual online competition for teams of undergraduates to use mathematical modeling to solve real world problems.
- Problem B: Peter Sutcliffe, a notorious series murder. The contesters should provided a plan to aid the police agency on locating Sutcliffe and predicting the next possible crime site.

Modeling

- My part of work is to locate the criminal's 'anchor point' (a criminology term referred to the base of murder). This process is named Geographical Profiling.

Modeling

- My part of work is to locate the criminal's 'anchor point' (a criminology term referred to the base of murder). This process is named Geographical Profiling.
- I assume $f(s, z)$ is the probability density for the criminal to commit a crime at site s from the anchor point z, which has the following form:

$$f(s, z) \propto \frac{1}{\sqrt{2\pi}\sigma} \left\{ -\frac{1}{2\sigma^2} \left(d(s, z) - d \right)^2 \right\}$$

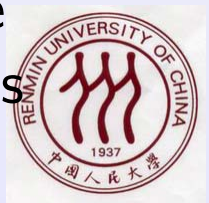
where $d(s, z)$ denotes the distance between crime site s and anchor point z, σ, d are parameters in the model.

Modeling

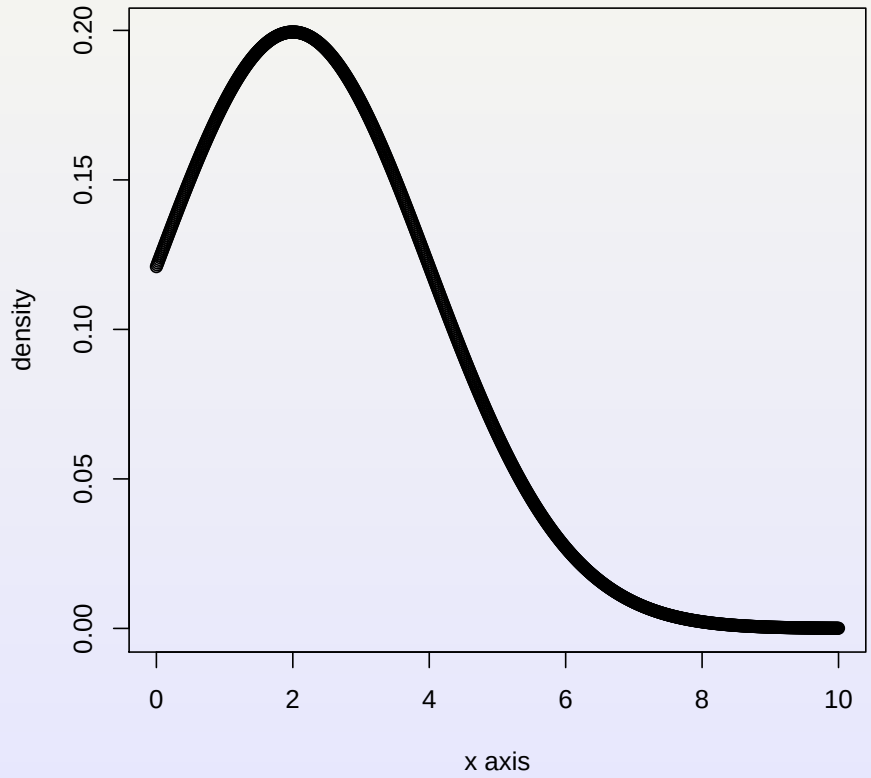
- My part of work is to locate the criminal's 'anchor point' (a criminology term referred to the base of murder). This process is named Geographical Profiling.
- I assume $f(s, z)$ is the probability density for the criminal to commit a crime at site s from the anchor point z, which has the following form:

$$f(s, z) \propto \frac{1}{\sqrt{2\pi}\sigma} \left\{ -\frac{1}{2\sigma^2} \left(d(s, z) - d \right)^2 \right\}$$

  where $d(s, z)$ denotes the distance between crime site s and anchor point z, σ, d are parameters in the model.
- Note that the $f(s, z)$ is a 'cut-off' density, not the complete normal density (d varies from 0 to ∞), so a notation of $\propto$ is used.



```
> x = seq(0, 10, 0.01)
> y = dnorm(x, 2, 2)
> plot(x, y, xlim = c(0, 10), ylab = "density", xlab = "x axis")
```



Modeling

- Notations: define $s_i = (x_i, y_i)$ as the coordinates of crime site, $z = (x_0, y_0)$ as the true location of anchor point. Function $d(s_i, z)$ is the distance from crime site to anchor point. d is a parameter to describe the psychological propensity that the murder is less willing to commit a crime either near his base, or very far from the base.

Modeling

- Notations: define $s_i = (x_i, y_i)$ as the coordinates of crime site, $z = (x_0, y_0)$ as the true location of anchor point. Function $d(s_i, z)$ is the distance from crime site to anchor point. d is a parameter to describe the psychological propensity that the murder is less willing to commit a crime either near his base, or very far from the base.
- Using Bayes Theorem, a Likelihood $\times$ Prior is given to deduce the posterior density for the parameter. The likelihood has the following form:

$$f(s_i, z|d, \sigma, x_0, y_0) \propto \prod_{i=1}^{n} \frac{1}{\sigma} exp\{-\frac{1}{2\sigma^2}\left(d(s_i, z) - d\right)^2)\}$$

Make Inference

- I assume the prior density for $(\sigma^2)^{-1}$, d, x_0, y_0

$$(\sigma^2)^{-1} \sim Gamma(c_1, d_1)$$

$$d \sim N(c_2, d_2)$$

$$x_0 \sim N(c_3, d_3)$$

$$y_0 \sim N(c_4, d_4)$$

Make Inference

- The posterior density for parameter σ, d, x_0, y_0:

$$(\sigma^2)^{-1}|d, x_0, y_0 \sim Gamma(\frac{n}{2} + c_1 - 1, d_1 + \sum_{i=1}^{n}(d(s_i, z) - d)^2$$

$$d|\sigma, x_0, y_0 \sim N\left(\frac{\sum_{i=1}^{n} d(s_i, z)d_2^2 + c_2\sigma^2}{nd_2^2 + \sigma^2}, \frac{\sigma^2 d_2^2}{nd_2^2 + \sigma^2}\right)$$

$$x_0|\sigma, d, y_0 \propto exp\left\{-\frac{\sum_{i=1}^{n}\left((x_i - x_0)^2 - 2dd(s_i, z)\right)}{\sigma^2} - \frac{1}{2\sigma_1^2(x_0 - c_3)^2}\right\}$$

$$y_0|\sigma, d, x_0 \propto exp\left\{-\frac{\sum_{i=1}^{n}\left((y_i - y_0)^2 - 2dd(s_i, z)\right)}{\sigma^2} - \frac{1}{2\sigma_1^2(y_0 - c_4)^2}\right\}$$

Make Inference

- A classical Gibbs Sampling approach is used to successively and repeatedly simulate from the posterior distribution.

Make Inference

- A classical Gibbs Sampling approach is used to successively and repeatedly simulate from the posterior distribution.
- The work flow is given as:
  Set $(x_0, y_0, \alpha, d) = (x_0^{(0)}, y_0^{(0)}, \alpha^{(0)}, d^{(0)})$.

Make Inference

- A classical Gibbs Sampling approach is used to successively and repeatedly simulate from the posterior distribution.
- The work flow is given as:
  Set $(x_0, y_0, \alpha, d) = (x_0^{(0)}, y_0^{(0)}, \alpha^{(0)}, d^{(0)})$.
- Simulate $\sigma^{(i)}$ from $p\left(\sigma^2 | x_0^{(i-1)}, y_0^{(i-1)}, d^{(i-1)}\right)$
  Simulate $d^{(i)}$ from $p\left(d | x_0^{(i-1)}, y_0^{(i-1)}, \sigma^{(\mathbf{i})}\right)$
  Simulate $x_0^{(i)}$ from $f(x_0 | y_0^{(i-1)}, \sigma^{(i)}, d^{(\mathbf{i})})$
  Simulate $y_0^{(i)}$ from $f(y_0 | x_0^{(\mathbf{i})}, \sigma^{(i)}, d^{(i)})$

Gibbs Sampling

- The posterior distribution of σ and d is straightforward, while for x_0 and y_0, it's not self-explanatory.

Gibbs Sampling

- The posterior distribution of σ and d is straightforward, while for x_0 and y_0, it's not self-explanatory.
- Use normal simulation function in R to simulate σ, d. And for x_0 and y_0, we only know the kernel of the density, so Metropolis Hasting Algorithm is recommended.

Metropolis Hasting Algorithm

- The objective of MH algorithm is to generate a Markov Chain which will converge to your designed distribution.

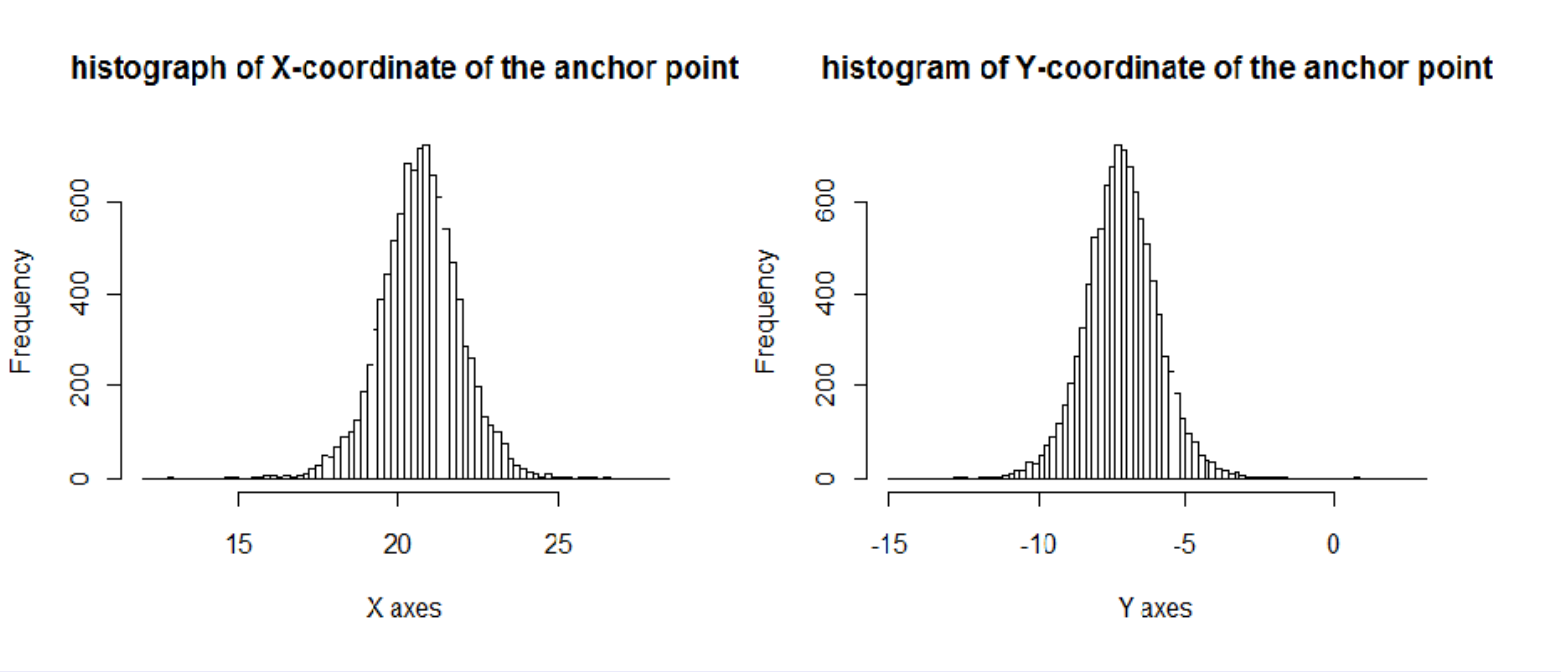
Metropolis Hasting Algorithm

Starting with $X^{(0)}$ iterate for K times.

1. Draw X $\sim q(X|X^{(t-1)})$.
2. Compute acceptance probability
$\alpha = min\left\{1, \frac{f(X)q(X^{(t-1)}|X)}{f(X^{(t-1)})q(X|X^{(t-1)})}\right\}$.
3. With Probability α, set $X^{(t)} = X$, otherwise set $X^{(t)} = X^{(t-1)}$

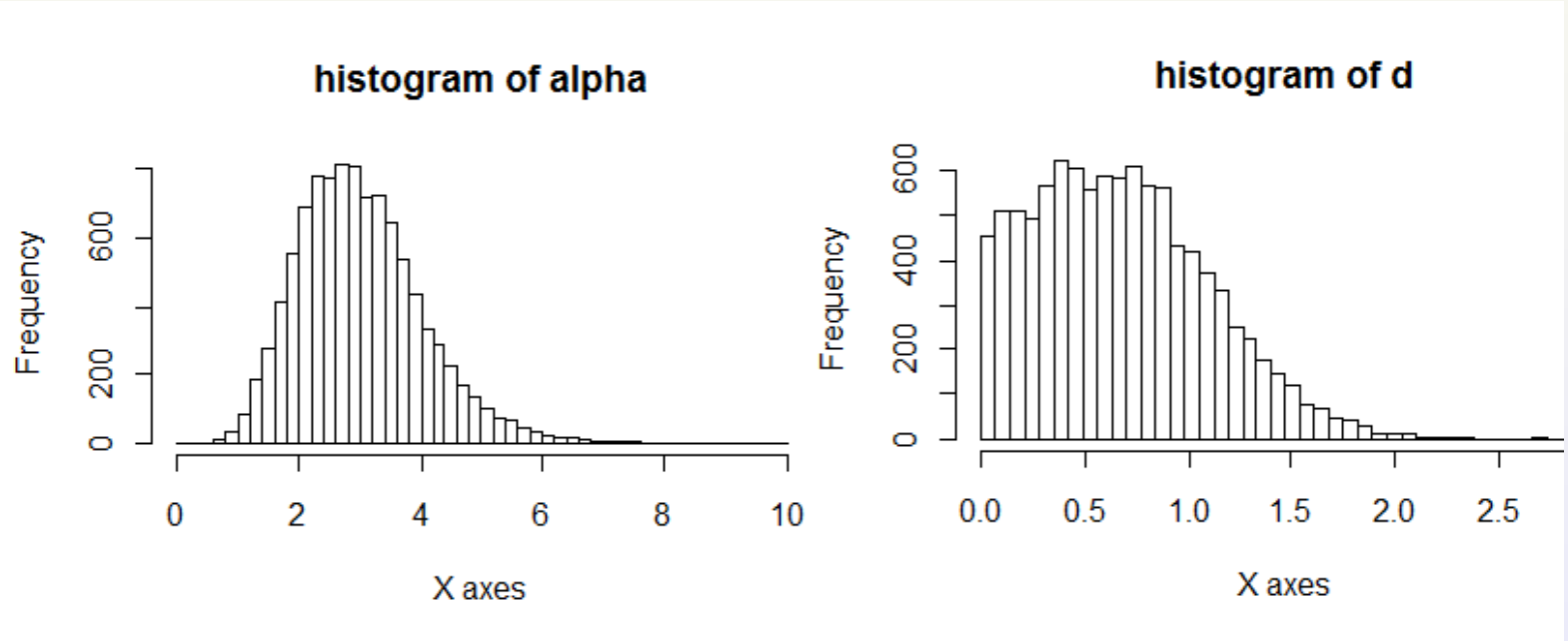
x_0, y_0

Figure: The histogram for x_0, y_0



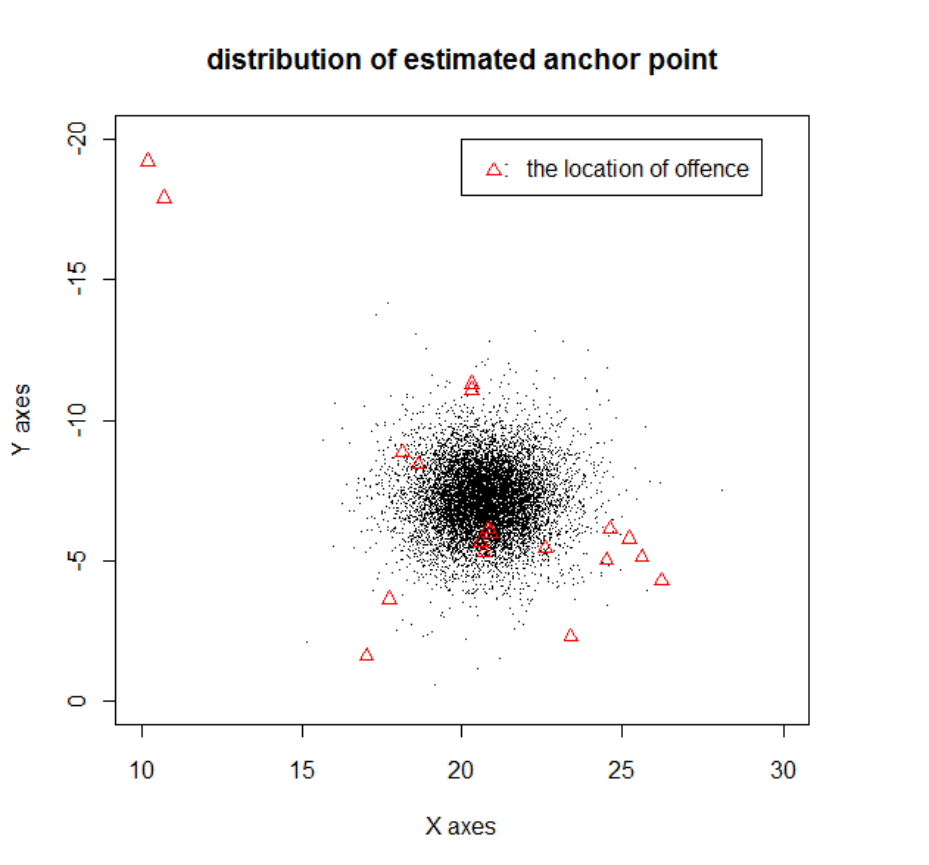
σ,d

Figure: The histogram for parameter σ, d



Z

Figure: The scatter plot of the estimated anchor point (the red triangles mark the location of offences)



- From the figure, we have estimates for parameters: $\sigma = 5.8$; $d = 2.3$. when transformed into a real length scale, $\sigma = 18.04km$; $d = 7.15km$. And also, we estimate the location of anchor point z is in southeast part of Bradford, West Yorkshine County, exactly the place Sutcliffe resides.

- From the figure, we have estimates for parameters: $\sigma = 5.8$; $d = 2.3$. when transformed into a real length scale, $\sigma = 18.04km$; $d = 7.15km$. And also, we estimate the location of anchor point z is in southeast part of Bradford, West Yorkshine County, exactly the place Sutcliffe resides.
- The weakness of the model, is no geographical information and time factor is incorporated into the estimation process. For further study, these features could be added into the likelihood function, and give more accurate estimate.

- From the figure, we have estimates for parameters: $\sigma = 5.8$; $d = 2.3$. when transformed into a real length scale, $\sigma = 18.04km$; $d = 7.15km$. And also, we estimate the location of anchor point z is in southeast part of Bradford, West Yorkshine County, exactly the place Sutcliffe resides.
- The weakness of the model, is no geographical information and time factor is incorporated into the estimation process. For further study, these features could be added into the likelihood function, and give more accurate estimate.
- Finally, we won Meritorious Award (6 percent in total) for our performance.

Thank you for your listening!