X-ray Dark Sources Detection

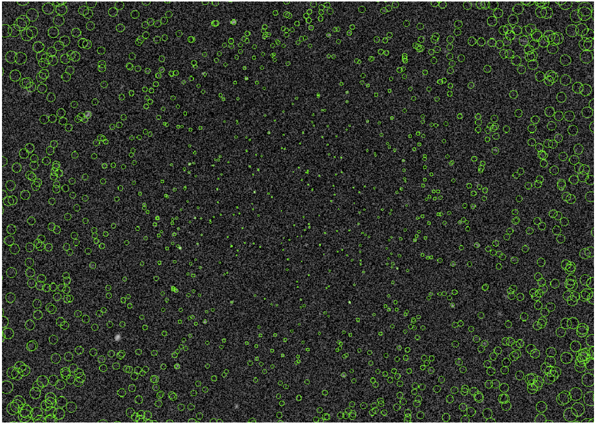
Lazhi Wang

Department of Statistics, Harvard University

Nov. 5th, 2013

Data

- Y_i, background contaminated photon counts in a source exposure over $T = 48984.8$ seconds (13.6 hours),
- X, photon counts in the exposure of pure background over T seconds.



Goals of the Project

1. To develop a fully Bayesian model to infer the distribution of the intensities of all the sources in a population.

2. To identify the existence of dark sources in the population.

Outline

1. The basic hierarchical Bayesian model
2. Extensions of the basic model
3. Extensive simulation studies:
    - Robustness of the model
    - Non-informativeness of the prior
4. Identifying the existence of dark sources via hypothesis testing:
    - Calculation of test-statistic and posterior predictive p-value
    - Simulation study
5. Real Data Application
6. One Difficult Problem and Discussion

Basic Hierarchical Bayesian Model

- Level I:

$$
\begin{array}{rcl}
Y_i & = & \mathcal{S}_i + \mathcal{B}_i \\
\mathcal{S}_i \big| \lambda_i & \sim & \text{Poisson}(r_i e_i T \lambda_i) \\
\mathcal{B}_i \big| \xi & \sim & \text{Poisson}(a_i T \xi) \\
X \big| \xi & \sim & \text{Poisson}(A T \xi)
\end{array}
$$

  - $\mathcal{S}_i$ (counts): number of photons from source i in the source region,
  - $\mathcal{B}_i$ (counts): number of photons from the background in the source region,
  - λ_i (counts/s/cm^2): the intensity of source i,
  - ξ (counts/s/pixels): the intensity of background,
  - T (seconds): exposure time, $T = 48984.8$,
  - e_i (cm^2): the telescope effective area,
  - r_i:proportion of photons from source i expected to fall in source region,
  - a_i (pixels): the size of source region i,
  - A (pixels): the size of background region.

$\mathcal{S}_i$, $\mathcal{B}_i$, λ_i, ξ are all unobserved/latent, T, e_i, r_i, a_i, A are all known constant. Y_i, X are observed data.

Basic Hierarchical Bayesian Model

- Level II:

$$\xi \sim \text{Gamma}(\alpha_0, \beta_0)$$

$$\lambda_i \big| \alpha, \beta, \pi_d \begin{cases} = 0 & \text{with probability } \pi_d, \\ \sim \text{Gamma}(\alpha, \beta) & \text{with probability } 1 - \pi_d. \end{cases}$$

- Level III: Prior on the hyper-parameters $\pi_d, \mu = \frac{\alpha}{\beta}, \theta = \frac{\alpha}{\beta^2}$

$$\pi_d \sim Unif(0, 1)$$

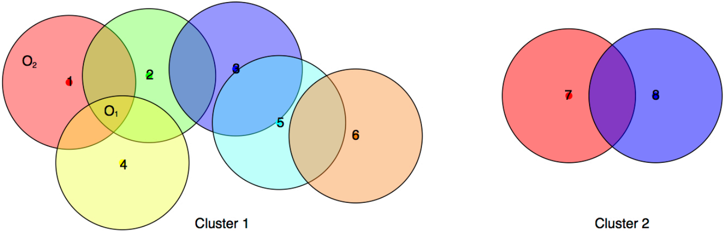
$$P(\mu, \theta) \propto \frac{1}{c_1^2 + (\mu - c_2)^2} \frac{1}{c_3^2 + (\theta - c_4)^2} I_{\mu>0, \theta>0},$$

Model Extension I: Overlapping Sources

- Notation:
  - $O = \{i_1, \cdots, i_k\}$ indicates the region formed by the overlap of source $i_1, \cdots, i_k$. For example, $O_1 = \{1, 2, 4\}$, $O_2 = \{1\}$.
  - $\mathcal{O}$: the collection of all such regions.
- Level I model:

$$Y_o = \mathcal{S}_o + \mathcal{B}_o = \sum_{j \in O} \mathcal{S}_{oj} + \mathcal{B}_o,$$

$$\begin{aligned} \mathcal{S}_{oj} \big| \lambda_j &\sim \text{Poisson}(r_{oj} e_o T \lambda_j) \\ \mathcal{B}_o \big| \xi &\sim \text{Poisson}(a_o T \xi) \end{aligned}$$



Model Extension II: Different Background Intensities

- In our data, the background intensity has an increasing trend as the projected angle (in arcmin) on the sky from the center of the field of view increases from 0 to 16.

Projected Angle	Counts (counts)	Region (pixels)	Intensity (counts/pixels)
0-6	219962	22029408	0.0010
6-8	146332	14093856	0.0104
8-16	285300	26448800	0.0108
overall 0-16	651891	62572560	0.0104

Model Extension II: Different Background Intensities

- Notation:
  - X_k (counts): number of photons collected in background region k over T seconds
  - ξ_k (counts/s/pixels): the background intensity in regions k
  - A_k (pixels): the size of background region k
  - $\mathcal{O}_k$: the collection of source regions in the background region k
- Model:
  - For counts from the pure background:

$$X_k \big| \xi_k \sim \text{Poisson}(A_k T \xi_k)$$

  - For counts from the source region $O \in \mathcal{O}_k$:

$$B_o \big| \xi_k \sim \text{Poisson}(a_o T \xi_k)$$

Simulation Study: The Robustness of the Model

$$Y_i \sim \text{Poisson}(r_i e_i T \lambda_i + 5), \text{ for } i = 1, \cdots, 1000, \quad X = 2.5 \times 10^5,$$

$$r_i e_i T \lambda_i \begin{cases} = 0 & \text{with probability } \pi_d, \\ \sim \text{Gamma}[\mu^* = 15, \theta^*] & \text{with probability } 1 - \pi_d. \end{cases}$$

	π_d									
θ^*	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
50	0.002 (0,0.01)	0.111 (0.09,0.14)	0.209 (0.17,0.24)	0.281 (0.26,0.33)	0.422 (0.37,0.44)	0.506 (0.48,0.55)	0.58 (0.53,0.61)	0.696 (0.68,0.75)	0.795 (0.76,0.82)	0.866 (0.86,0.91)
100	0.009 (0,0.03)	0.102 (0.07,0.13)	0.226 (0.18,0.27)	0.255 (0.22,0.31)	0.367 (0.33,0.42)	0.525 (0.48,0.57)	0.589 (0.52,0.62)	0.702 (0.64,0.73)	0.795 (0.77,0.85)	0.838 (0.78,0.93)
200	0.021 (0,0.05)	0.117 (0.06,0.17)	0.159 (0.11,0.24)	0.32 (0.24,0.37)	0.366 (0.29,0.44)	0.509 (0.41,0.55)	0.54 (0.49,0.62)	0.703 (0.62,0.76)	0.76 (0.68,0.83)	0.791 (0.47,0.95)
300	0.007 (0,0.06)	0.134 (0.03,0.18)	0.231 (0.13,0.3)	0.31 (0.27,0.43)	0.329 (0.18,0.44)	0.447 (0.21,0.54)	0.637 (0.53,0.69)	0.733 (0.65,0.77)	0.816 (0.75,0.88)	0.931 (0.87,0.95)
500	0.005 (0,0.08)	0.067 (0,0.22)	0.266 (0.12,0.39)	0.262 (0.03,0.35)	0.505 (0.41,0.58)	0.561 (0.51,0.68)	0.564 (0.14,0.67)	0.606 (0.52,0.84)	0.789 (0.5,0.9)	0.931 (0.73,0.97)
1000	0.16 (0.02,0.33)	0.296 (0,0.4)	0.176 (0,0.36)	0.415 (0.07,0.54)	0.418 (0.08,0.61)	0.568 (0.05,0.64)	0.594 (0.11,0.74)	0.544 (0.04,0.75)	0.829 (0.23,0.9)	0.921 (0.73,0.98)

Simulation Study: The Robustness of the Model

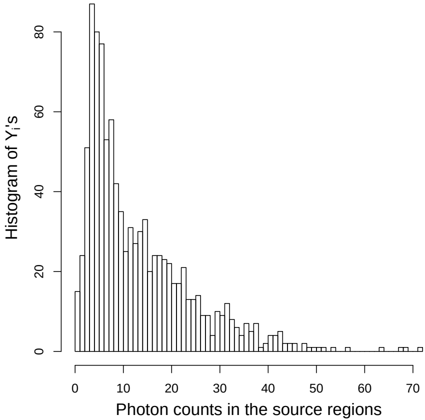
$$Y_i \sim \text{Poisson}(r_i e_i T\lambda_i + 10), \text{ for } i = 1, \cdots, 1000, \quad X = 2.5 \times 10^5,$$

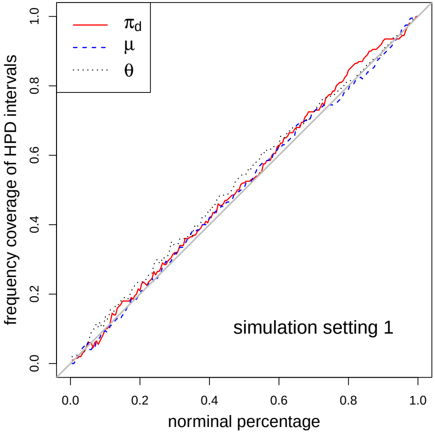
$$r_i e_i T\lambda_i \begin{cases} = 0 & \text{with probability } \pi_d, \\ \sim \text{Gamma}[\mu^* = 15, \theta^*] & \text{with probability } 1 - \pi_d. \end{cases}$$

	π_d									
θ^*	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
50	0.006 (0,0.02)	0.133 (0.09,0.15)	0.21 (0.17,0.24)	0.312 (0.26,0.34)	0.399 (0.34,0.43)	0.498 (0.46,0.54)	0.588 (0.52,0.62)	0.727 (0.69,0.76)	0.745 (0.74,0.82)	0.889 (0.85,0.91)
100	0.003 (0,0.03)	0.06 (0.03,0.11)	0.257 (0.18,0.28)	0.257 (0.19,0.3)	0.377 (0.35,0.45)	0.581 (0.5,0.6)	0.56 (0.5,0.64)	0.719 (0.67,0.76)	0.816 (0.78,0.87)	0.911 (0.85,0.94)
200	0.028 (0,0.1)	0.188 (0.09,0.22)	0.221 (0.12,0.27)	0.291 (0.24,0.4)	0.331 (0.24,0.47)	0.537 (0.44,0.6)	0.523 (0.45,0.62)	0.736 (0.64,0.79)	0.785 (0.61,0.82)	0.903 (0.69,0.95)
300	0.02 (0,0.1)	0.034 (0,0.15)	0.193 (0.07,0.31)	0.375 (0.24,0.45)	0.417 (0.31,0.51)	0.437 (0.21,0.57)	0.604 (0.5,0.71)	0.745 (0.64,0.81)	0.818 (0.58,0.88)	0.951 (0.73,0.96)
500	0.004 (0,0.09)	0.274 (0,0.26)	0.188 (0,0.31)	0.095 (0.06,0.4)	0.497 (0.3,0.57)	0.521 (0.24,0.65)	0.713 (0.5,0.76)	0.769 (0.32,0.85)	0.642 (0.19,0.9)	0.935 (0.54,0.97)
1000	0.106 (0,0.27)	0.268 (0,0.38)	0.082 (0,0.37)	0.339 (0.07,0.59)	0.327 (0.06,0.61)	0.542 (0.13,0.73)	0.633 (0.04,0.69)	0.476 (0.05,0.79)	0.812 (0.48,0.93)	0.959 (0.82,0.98)

Simulation Study: Non-informativeness of the Prior

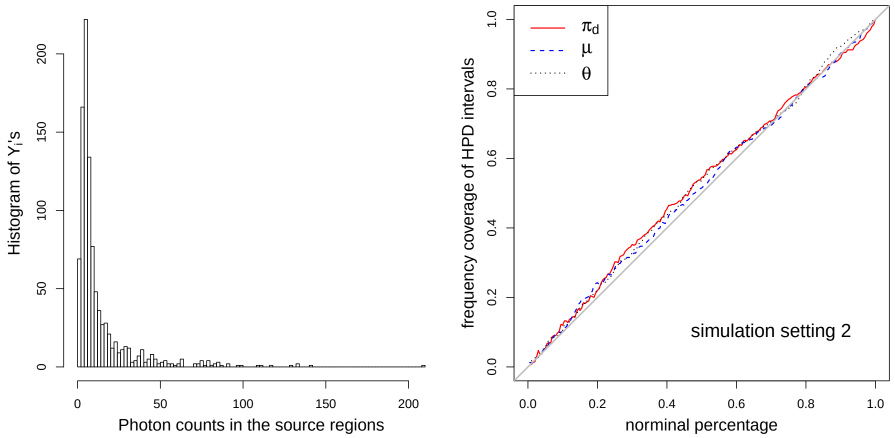
$$\mathcal{B}_i \sim \text{Poisson}(5), \quad \pi_d = 0.4, \quad \mu^* = 15, \quad \theta^* = 100$$





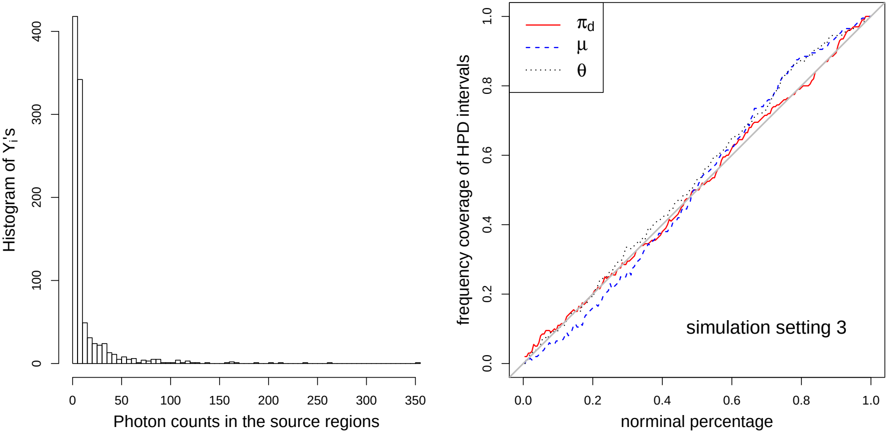
Simulation Study: Non-informativeness of the Prior

$$\mathcal{B}_i \sim \text{Poisson}(5), \quad \pi_d = 0.4, \quad \mu^* = 15, \quad \theta^* = 500$$



Simulation Study: Non-informativeness of the Prior

$$\mathcal{B}_i \sim \text{Poisson}(5), \quad \pi_d = 0.4, \quad \mu^* = 15, \quad \theta^* = 1000$$



Hypothesis Testing for Existence of Dark Sources

- Hypothesis Testing:

$$H_0 : \pi_d = 0, \quad H_a : \pi_d > 0.$$

- Reject H_0 if the p-value is low,

$$\text{p-value } = P(T(\mathbb{D}) \geqslant T^{obs} \big| H_0),$$

where $\mathbb{D} \sim H_0$ and $T(\mathbb{D})$ is a test statistic.

Hypothesis Testing for Existence of Dark Sources

- Hypothesis Testing:

$$H_0 : \pi_d = 0, \quad H_a : \pi_d > 0.$$

- Reject H_0 if the p-value is low,

$$\text{p-value } = P(T(\mathbb{D}) \geqslant T^{obs} \big| H_0),$$

where $\mathbb{D} \sim H_0$ and $T(\mathbb{D})$ is a test statistic.

- However, $\mathbb{D} \big| H_0$ is unknown because α and β are unknown:

$$\lambda_i | \alpha, \beta \sim Gamma(\alpha, \beta)$$

- Posterior predictive p-value (*ppp*):

$$ppp = P_0(T(\mathbb{D}) \geqslant T^{obs} \big| \mathbb{D}^{obs}),$$

where $\mathbb{D} \sim \mathbb{D} \big| H_0$ with $(\alpha, \beta) \sim \alpha, \beta \big| \mathbb{D}^{obs}, H_0$.

Hypothesis Testing for Existence of Dark Sources

- Estimation of *ppp*:

  1. Draw $(\alpha^{(t)}, \beta^{(t)})$ from $(\alpha, \beta)\big|\mathcal{D}^{obs}$ for $t = 1, 2, \cdots, m$,
  2. For each pair $(\alpha^{(t)}, \beta^{(t)})$, simulate $\mathcal{D}^{(t)}$ from the null model and calculate $T^{(t)} = T(\mathcal{D}^{(t)})$,
  3. Estimate *ppp* by

$$ppp \approx \frac{1}{m} \sum_{t=1}^{m} I\left(T^{(t)} \geqslant T^{obs}\right).$$

Hypothesis Testing for Existence of Dark Sources

- Estimation of *ppp*:
  1. Draw $(\alpha^{(t)}, \beta^{(t)})$ from $(\alpha, \beta)\big|\mathcal{D}^{obs}$ for $t = 1, 2, \cdots, m$,
  2. For each pair $(\alpha^{(t)}, \beta^{(t)})$, simulate $\mathcal{D}^{(t)}$ from the null model and calculate $T^{(t)} = T(\mathcal{D}^{(t)})$,
  3. Estimate *ppp* by

$$ppp \approx \frac{1}{m} \sum_{t=1}^{m} I\left(T^{(t)} \geqslant T^{obs}\right).$$

- Likelihood Ratio Test Statistics:

$$R(\mathbb{D}) = \frac{\sup_{\alpha,\beta,\pi_d} L_a(\alpha, \beta, \pi_d \big| \mathbb{D})}{\sup_{\alpha,\beta} L_0(\alpha, \beta \big| \mathbb{D})},$$

We use $T(\mathbb{D}) = log(R(\mathbb{D}))$ as the test statistic.

Calculation of Test Statistics

- One simplification: $\xi = X$
- $L_0(\alpha, \beta \big| \mathbb{Y})$:

$$P_0(\mathbb{Y}\big|\alpha,\beta) = \int P(\mathbb{Y}\big|\boldsymbol{\lambda})P_0(\boldsymbol{\lambda}\big|\alpha,\beta)d\boldsymbol{\lambda}$$

$$= C\frac{\beta^\alpha}{\Gamma(\alpha)}\prod_{i=1}^{N}\left[\sum_{j=1}^{Y_i} c_i^j \binom{Y_i}{j} \frac{\Gamma(Y_i - j + \alpha)}{(\beta + r_i e_i T)^{Y_i - j + \alpha}}\right].$$

- $L_a(\alpha, \beta, \pi_d \big| \mathbb{Y})$:

$$P_a(\mathbb{Y}\big|\alpha,\beta,\pi_d) = \int P(\mathbb{Y}\big|\boldsymbol{\lambda})P_a(\boldsymbol{\lambda}\big|\alpha,\beta,\pi_d)d\boldsymbol{\lambda}$$

$$= C\prod_{i=1}^{N}\left[\pi_d c_i^{Y_i} + (1-\pi_d)\frac{\beta^\alpha}{\Gamma(\alpha)}\sum_{j=1}^{Y_i} c_i^j \binom{Y_i}{j} \frac{\Gamma(Y_i - j + \alpha)}{(\beta + r_i e_i T)^{Y_i - j + \alpha}}\right].$$

Simulation Study

$$Y_i \sim \text{Poisson}(r_i e_i T \lambda_i + 5), \;\text{for } i = 1, \cdots, 1000, \quad X = 2.5 \times 10^5,$$

$$r_i e_i T \lambda_i \begin{cases} = 0 & \text{with probability } \pi_d, \\ \sim \text{Gamma}[\mu^* = 15, \theta^*] & \text{with probability } 1 - \pi_d. \end{cases}$$

	π_d									
θ^*	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
50	1	1	0	0	0	0	0	0	0	0
100	0.179	0	0	0	0	0	0	0	0	0.001
200	0.332	0	0	0	0	0	0	0	0	0.197
300	1	0.01	0	0	0.002	0.003	0	0	0	0.001
500	1	0.232	0.001	0.064	0	0	0.058	0.01	0.035	0.039
1000	0.074	0.211	0.226	0.051	0.118	0.152	0.147	1	0.334	0.03

Simulation Study

$$Y_i \sim \text{Poisson}(r_i e_i T\lambda_i + 10), \text{ for } i = 1, \cdots, 1000, \quad X = 2.5 \times 10^5,$$

$$r_i e_i T\lambda_i \begin{cases} = 0 & \text{with probability } \pi_d, \\ \sim \text{Gamma}[\mu^* = 15, \theta^*] & \text{with probability } 1 - \pi_d. \end{cases}$$

	π_d									
θ^*	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
50	0.18	0	0	0	0	0	0	0	0	0
100	1	0.002	0	0	0	0	0	0	0	0
200	0.034	0	0	0	0	0	0	0	0	0.018
300	1	1	0.002	0	0	0.002	0	0	0.006	0.02
500	1	0.087	0.11	0.025	0	0.015	0	0.072	0.207	0.149
1000	0.426	0.46	0.392	0.086	0.146	0.05	0.451	1	0.05	0.016

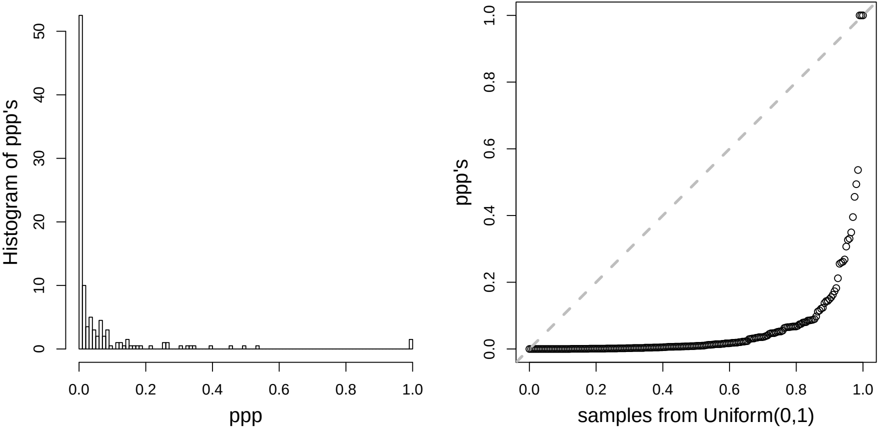
Simulation Study: Distribution of *ppp*

$$\mathcal{B}_i \sim \text{Poisson}(5), \quad \pi_d = 0.4, \quad \mu^* = 15, \quad \theta^* = 100$$

All the *ppp*'s are 0.

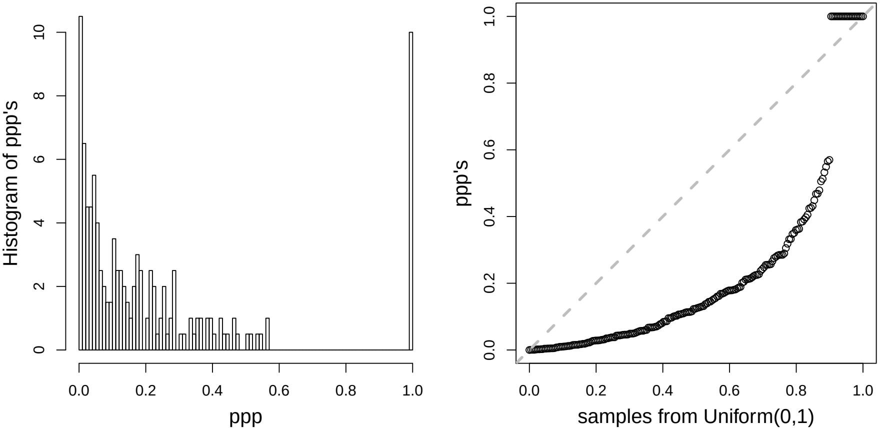
Simulation Study: Distribution of *ppp*

$$\mathcal{B}_i \sim \text{Poisson}(5), \quad \pi_d = 0.4, \quad \mu^* = 15, \quad \theta^* = 500$$



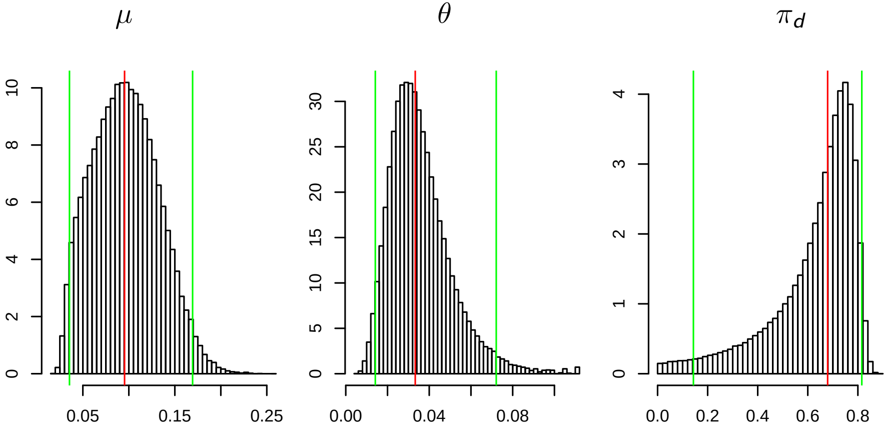
Simulation Study: Distribution of *ppp*

$$\mathcal{B}_i \sim \text{Poisson}(5), \quad \pi_d = 0.4, \quad \mu^* = 15, \quad \theta^* = 1000$$



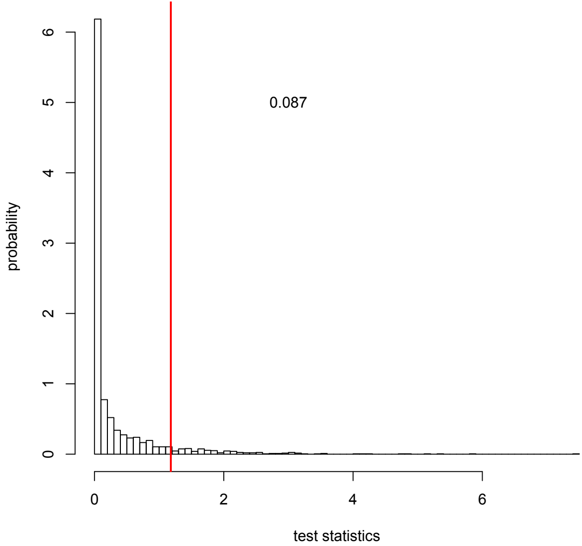
Real Data Analysis: No overlap sources, arcmin $\leqslant 6$

- Posterior distribution of the hyper-parameters



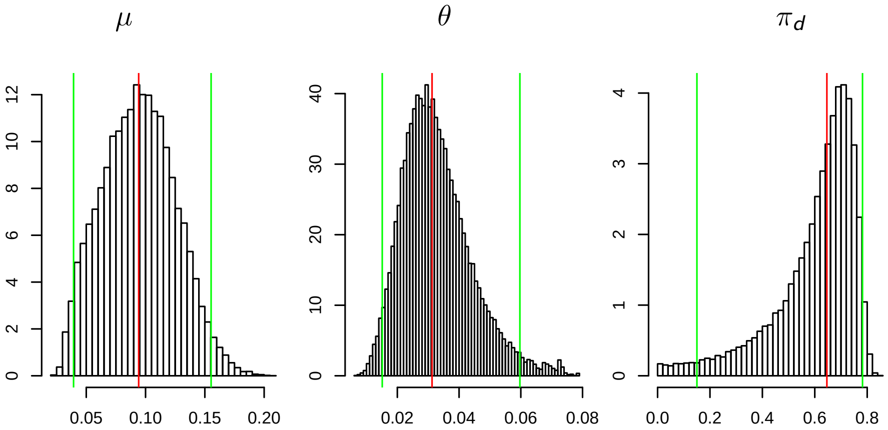
Real Data Analysis: No overlap sources, arcmin $\leqslant 6$

- Histogram of the test statistics: $ppp \approx 0.087$.

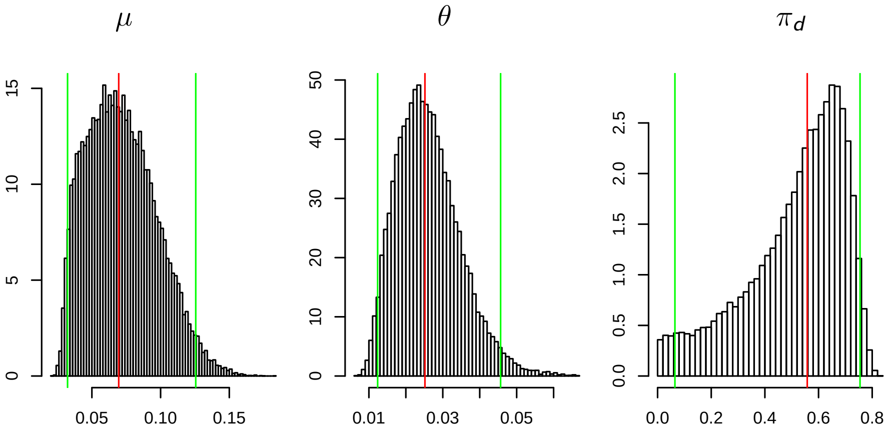


Real Data Analysis: all the overlap sources, arcmin $\leqslant 6$

- Posterior distribution of the hyper-parameters



- Posterior distribution of the hyper-parameters (two background intensities).



same background intensity

Difficulty

- Calculation of *ppp* in the presence of overlapping sources.
- We need to calculate the likelihood ratio test statistic:

$$R(\mathbb{Y}) = \frac{\sup_{\alpha,\beta,\pi_d} L_a(\alpha,\beta,\pi_d \big| \mathbb{Y})}{\sup_{\alpha,\beta} L_0(\alpha,\beta \big| \mathbb{Y})},$$

Difficulty

- Calculation of *ppp* in the presence of overlapping sources.
- We need to calculate the likelihood ratio test statistic:

$$R(\mathbb{Y}) = \frac{\sup_{\alpha,\beta,\pi_d} L_a(\alpha,\beta,\pi_d \big| \mathbb{Y})}{\sup_{\alpha,\beta} L_0(\alpha,\beta \big| \mathbb{Y})},$$

- For simplicity:
  - $N = 2$, the two sources overlap.
  - $\mathcal{O} = \{O_1 = \{1\}, O_2 = \{2\}, O_3 = \{1,2\}\}$

Difficulty

- Calculation of *ppp* in the presence of overlapping sources.
- We need to calculate the likelihood ratio test statistic:

$$R(\mathbb{Y}) = \frac{\sup_{\alpha,\beta,\pi_d} L_a(\alpha,\beta,\pi_d \big| \mathbb{Y})}{\sup_{\alpha,\beta} L_0(\alpha,\beta \big| \mathbb{Y})},$$

- For simplicity:
  - $N = 2$, the two sources overlap.
  - $\mathcal{O} = \{O_1 = \{1\}, O_2 = \{2\}, O_3 = \{1,2\}\}$

- The "complete" data likelihood under the null hypothesis is

$$P_0(\mathbb{Y}, \boldsymbol{\lambda}|\alpha,\beta) = P(Y_1 \big| \lambda_1) P(Y_2 \big| \lambda_2) P(Y_3 \big| \lambda_1, \lambda_2) P(\lambda_1, \lambda_2 \big| \alpha, \beta)$$

$$\propto \; e^{-c_1\lambda_1 - c_2\lambda_2} \lambda_1^{\alpha-1} \lambda_2^{\alpha-1} (1 + c_3\lambda_1)^{Y_1} (1 + c_3\lambda_2)^{Y_2} (1 + c_5\lambda_1 + c_6\lambda_2)^{Y_3},$$

where c_i's are some constants.

Difficulty

- Calculation of *ppp* in the presence of overlapping sources.
- We need to calculate the likelihood ratio test statistic:

$$R(\mathbb{Y}) = \frac{\sup_{\alpha,\beta,\pi_d} L_a(\alpha,\beta,\pi_d \big| \mathbb{Y})}{\sup_{\alpha,\beta} L_0(\alpha,\beta \big| \mathbb{Y})},$$

- For simplicity:
  - $N = 2$, the two sources overlap.
  - $\mathcal{O} = \{O_1 = \{1\}, O_2 = \{2\}, O_3 = \{1,2\}\}$
- The "complete" data likelihood under the null hypothesis is

$$P_0(\mathbb{Y}, \boldsymbol{\lambda}|\alpha,\beta) = P(Y_1 \big| \lambda_1) P(Y_2 \big| \lambda_2) P(Y_3 \big| \lambda_1, \lambda_2) P(\lambda_1, \lambda_2 \big| \alpha, \beta)$$

$$\propto \; e^{-c_1\lambda_1 - c_2\lambda_2} \lambda_1^{\alpha-1} \lambda_2^{\alpha-1} (1 + c_3\lambda_1)^{Y_1} (1 + c_3\lambda_2)^{Y_2} (1 + c_5\lambda_1 + c_6\lambda_2)^{Y_3},$$

  where c_i's are some constants.
- We need to integrate out λ_1 and λ_2 to get the likelihood $L_0(\alpha,\beta \big| \mathbb{Y})$.
- The calculation is "feasible" but very complicated when we have more overlaps and when N is large.

Real Data Analysis: all the overlap sources, arcmin $\leqslant 8$

- Posterior distribution of the hyper-parameters (same background intensities).