Dark Sources Detection

Lazhi Wang

Department of Statistics, Harvard University

April 16, 2013
Introduction: Data and Project Goal

Data:
- $Y_i$, observed photon counts, contaminated with background in a source exposure.
- $X$, observed photon counts in the exposure of pure background.

Goals of the Project:
- To develop a fully Bayesian model to infer the distribution of the intensities of all the sources in a population
- To identify the existence of dark sources in the population
A Brief Review: Bayesian Model

- **Level 1 model:**

  \[
  X | \xi \sim \text{Pois}(\xi),
  \]

  \[
  Y_i = Y_{iB} + Y_{iS}, \text{ where } Y_{iB} | \xi \sim \text{Pois}(a_i \xi),
  \]

  \[
  Y_{iS} | \lambda_i \sim \text{Pois}(b_i \lambda_i) \sim \begin{cases} 
  \delta_0, & \text{if } \lambda_i = 0; \\
  \text{Pois}(b_i \lambda_i), & \text{if } \lambda_i \neq 0.
  \end{cases}
  \]

- \(\xi\) is the background intensity,
- \(\lambda_i\) is the intensity of source \(i\),
- \(a_i\) is ratio of source area to background area (known constant),
- \(b_i\) is the telescope effective area (known constant).
A Brief Review: Bayesian Model

- Level II model:

  \[ \lambda_i | \alpha, \beta, \pi \begin{cases} = 0, & \text{with probability } 1 - \pi; \\ \sim \text{Gamma}(\alpha, \beta), & \text{with probability } \pi. \end{cases} \]

- Level III model:

  \[ P(\alpha, \beta, \pi) \propto P(\alpha, \beta). \]
The prior distribution of $\alpha, \beta$ needs to be proper

We do not want the proper prior to be very informative

Let $\mu = \frac{\alpha}{\beta}, \theta = \frac{\alpha}{\beta^2}$ be the mean and variance parameters of the Gamma distribution.

Weakly informative prior on $\mu, \theta$:

$$P(\mu) \propto \frac{1}{1 + \left(\frac{\mu - 20}{20}\right)^2}, \quad P(\theta) \propto \frac{1}{1 + \left(\frac{\theta - 1000}{1000}\right)^2}$$
Weakly Informative Prior on $\alpha, \beta$

Prior distribution of $\mu$

Prior distribution of $\theta$
Frequency Coverage: Simulation Setting 1

Observed data $Y$

Underlying intensity $\lambda$

Background data $Y_B$

Photons from the source $Y_S$
Frequency Coverage: Simulation Setting 1

Frequency coverage of interval for $\alpha/\beta$

Frequency coverage of interval for $\alpha/\beta^2$

Frequency coverage of interval for $1 - \pi$

![Graphs showing frequency coverage for different intervals](image-url)
Frequency Coverage: Simulation Setting 2

- **Observed data** $Y$
- **Underlying intensity** $\lambda$
- **Background data** $Y_B$
- **Photons from the source** $Y_S$
Frequency Coverage: Simulation Setting 2
Identifying the Existence of Dark Sources

- Hypothesis Testing:

  \[ H_0 : 1 - \pi = 0, \quad H_a : 1 - \pi > 0. \]

- \( H_0 \) corresponds to \( M_0 \) with the second level

  \[ \lambda_i|\alpha, \beta \sim \text{Gamma}(\alpha, \beta) \]

- \( H_a \) corresponds to \( M_a \) with the second level

  \[ \lambda_i|\alpha, \beta, \pi \begin{cases} = 0, & \text{with probability } 1 - \pi; \\ \sim \text{Gamma}(\alpha, \beta), & \text{with probability } \pi. \end{cases} \]
Hypothesis Testing

- Likelihood Ratio Test Statistics:

\[ R = \frac{L_a(\hat{\alpha}_{MLE}, \hat{\beta}_{MLE}, \hat{\pi}_{MLE}|Y)}{L_0(\hat{\alpha}_{MLE}, \hat{\beta}_{MLE}|Y)} \]

- What’s the distribution of \( R \) or \( \log(R) \) under \( H_0 \)?

- p-value is used to measure how likely we are to see a value of the test statistics as extreme as the observed value under \( H_0 \).

\[ \text{p-value} = P(R \geq R^{obs}|H_0) \]
The Distribution of $R$ under $H_0$

- Simulate $N$ data sets $Y^{rep}$ under $H_0$ and compute $R^{rep}$ for each of the $N$ data sets.

- P-value can be approximated by

$$ p\text{-value} \approx \frac{\#\{i : R_i^{rep} \geq R^{obs}\}}{N} $$

- However, we can not simulate data sets under $H_0$ because $\alpha$ and $\beta$ are unknown.

- Instead, we simulate $Y^{rep} \sim M_0$ with $\alpha, \beta \sim P_0(\alpha, \beta| Y^{obs})$. So the resulted “p-value” is the posterior predictive p-value under the $M_0$. 


Calculation of $R$: Maximum likelihood under $M_0$

- Likelihood under $M_0$:

\[
L_0(\alpha, \beta | Y^{rep}) = \int P(Y^{rep}, \lambda | \alpha, \beta) d\lambda \\
\propto \left( \frac{\beta^\alpha}{\Gamma(\alpha)} \right)^n \prod_{i=1}^{n} \int e^{-(b_i + \beta)\lambda_i \frac{(a_i \xi + b_i \lambda_i) Y_i^{rep}}{Y_i^{rep}}} \lambda_i^{\alpha - 1} d\lambda_i
\]

- EM algorithm (\(\lambda\)'s are treated as missing data).
- In the E-step, we need to find

\[
T_1^{(t)} = E_t(\sum_{i=1}^{n} \lambda_i | Y^{rep}) \quad \text{and} \quad T_2^{(t)} = E_t(\sum_{i=1}^{n} \log(\lambda_i) | Y^{rep})
\]

- Simulation to estimate $T_1^{(t)}$ and $T_2^{(t)}$:

  Gibbs sampling: $\lambda_i^{(t)} \sim P(\lambda_i | \alpha^{(t)}, \beta^{(t)}, Y^{rep}), i = 1, \ldots, n$
Calculation of $R$: Maximum likelihood under $M_0$

$\alpha^{(t)}$ in the EM algorithm

$\beta^{(t)}$ in the EM algorithm

observed log lik

$\alpha^{(t)}$ in the EM algorithm

$\beta^{(t)}$ in the EM algorithm

observed log lik

$\alpha^{(t)}$ in the EM algorithm

$\beta^{(t)}$ in the EM algorithm

observed log lik

$\alpha^{(t)}$ in the EM algorithm

$\beta^{(t)}$ in the EM algorithm

observed log lik

$\alpha^{(t)}$ in the EM algorithm

$\beta^{(t)}$ in the EM algorithm

observed log lik

$\alpha^{(t)}$ in the EM algorithm

$\beta^{(t)}$ in the EM algorithm

observed log lik
Calculation of $R$: Maximum likelihood under $M_a$

- **EM** algorithm ($\lambda$’s are treated as missing data)
- Gibbs sampling: $\lambda_i^{(t)} \sim P(\lambda_i|\alpha^{(t)}, \beta^{(t)}, \pi^{(t)}, Y^{rep})$
- However,
  - Each step in the EM algorithm is very slow
  - EM algorithm converges very slowly
Calculation of $R$: Maximum likelihood under $M_a$
A More Efficient Method to Calculate the Maximum likelihood under $M_a$

- Observation: for a fixed $\pi$, the EM converges very fast.
- A more efficient algorithm:
  1. Explore the space of $\pi$: fix $\pi$ at a range of values $\pi_1, \pi_2, \cdots, \pi_K$ and compute the
     \[ L_k = L_a(\hat{\alpha}_k, \hat{\beta}_k, \pi_k | Y^{rep}) \]
  2. Choose $k^*$ such that
     \[ k^* = \arg \max_k L_a(\hat{\alpha}_k, \hat{\beta}_k, \pi_k | Y^{rep}) \]
  3. Doing the complete EM algorithm with starting points
     \[ \pi^{(0)} = \pi_{k^*}, \alpha^{(0)} = \hat{\alpha}_{k^*}, \beta^{(0)} = \hat{\beta}_{k^*} \]
A More Efficient Method to Calculate the Maximum likelihood under $M_a$
Posterior Predictive P-value

Histogram of $l_a(\hat{\alpha}, \hat{\beta}, \hat{\pi}|y^{rep}) - l_0(\hat{\alpha}, \hat{\beta}|y^{rep})$

$log(R) = 0.877$ for the real data

Posterior predictive p-value $= P(log(R^{rep}) \geq log(R^{obs})) = 0.105$
\[ \frac{\hat{\alpha}_{MLE}}{\hat{\beta}_{MLE}} = 20.22, \quad \frac{\hat{\alpha}_{MLE}}{\hat{\beta}_{MLE}^2} = 1451.2, \quad 1 - \hat{\pi}_{MLE} = 0.624. \]
Model for dealing with Overlapping Sources

\[ Y_{i,j} = Y_{i,j,1}^{(s)} + Y_{i,j,2}^{(s)} + Y_{i,j}^{(b)} \]

\[ Y_{i,j,k}^{(s)} \sim \text{Pois}(b_{i,j,k} \lambda_{i,k}), \]

where \( b_{i,j,k} = b_{i,k} c_{i,j,k} \), \( b_{i,k} \) is the effective area and \( c_{i,j,k} \) is the expected proportion of photons from source \( k \) counted in \( Y_{i,j} \).
Model for dealing with Overlapping Sources

Level I Model:

\[ Y_{i,j} = Y_{i,j}^{(s)} + Y_{i,j}^{(b)}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n_i, \]

\[ Y_{i,j}^{(b)} | \xi \sim \text{Pois}(a_i \xi), \]

\[ Y_{i,j}^{(s)} = \sum_{k=1}^{n_i} Y_{i,j,k}^{(s)} \]

\[ Y_{i,j,k}^{(s)} | \lambda_{i,k} \sim \text{Pois}(b_{i,j,k} \lambda_{i,k}), \quad k = 1, \ldots, n_i, \]
Simulation Results

35% $n_i$’s are 1, 5% $n_i$’s are 2 and 65% $n_i$’s are 3.
Maximum Likelihood under $M_a$ for the Real Data

- $\alpha^{(v)}$ with $0.2816$
- $\beta^{(v)}$ with $0.01393$
- $\pi^{(v)}$ with $0.876$
- $\|\alpha^{(v)},\beta^{(v)},\pi^{(v)}\|$ with $6875.48$
Posterior Distribution under $M_0$

\[ \hat{\alpha}_{MLE} \frac{MLE}{\hat{\beta}} = 7.87, \quad \hat{\alpha}_{MLE} \frac{MLE}{\hat{\beta}^2} = 787, \]