# A Latent Model To Detect Multiple Clusters of Varying Sizes

Minge Xie
Rutgers University
Joint Work with Qiankun Sun and Joe Naus
Research Partly Supported by NSA and ONR

## Clustering of Events

- Clustering of events in time and space
  - Unusually large number of events/patterns clumping within a small region of time, space or location in a sequence
    - Focus on a specific region, for example a region with heavy pollution;
    - Scan entire study area and seek to locate regions with unusually high likelihood of clustering
- Temporal Clusters: time intervals within which an incidence of interest is much more (or less) likely to happen than that outside these time intervals.

## Some Practical Examples

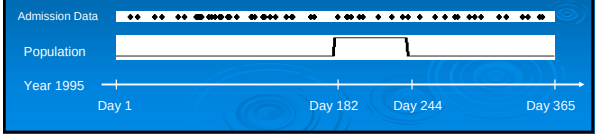
- Epidemiological study: Often required to study the data to obtain evidence of temporal or spatial clusters
  - Especially when the etiology of diseases has not yet been well established
- Surveillance for biological terrorism: Detecting relatively abrupt increase in incidence
  - Essential to provide early warnings of intentional releases of biological or nuclear agents
- Environmental study: Interest to detect and monitor population living near a factory generating pollution
  - Increased chance of certain diseases
- Many more ..

## Hospital Hemoptysis Admission Data

Days of hemoptysis admission at Nice University Hospital from January 1 to December 31, 1995 (Molinari, et al. 2001):

2 8 23 29 43 48 58 60 61 63
69 71 74 74 78 80 85 86 86 87
93 105 106 108 115 117 121 126 135 140
141 156 159 179 187 188 188 191 191 198
201 214 225 225 235 235 239 249 262 271
279 279 282 292 296 302 317 323 337 342
352 354

Nice, France (south)



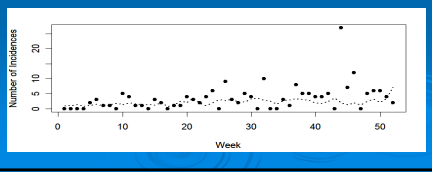

## 2004 Weekly Brucellosis Incidence Data (Collected by CDC)

2004 Data (number of incidences by week):

0 0 0 0 2 3 1 1 0 5 4 1 1 0 3 2 0 1 1 4 3 2 4 6 0 9 3
2 5 4 0 10 0 0 3 1 8 5 5 4 4 5 0 27 7 12 0 5 6 6 4 2

1997-2003 Average:

0.86 1.00 1.29 0.72 1.15 1.43 0.86 1.29 1.86 1.29 2.00
1.58 1.29 1.29 1.15 2.00 0.72 1.29 2.43 2.15 3.58 1.86
1.00 2.00 3.00 2.58 3.29 2.15 2.29 3.29 3.43 2.72 2.43
1.58 2.58 2.86 3.29 3.00 2.72 1.86 1.72 2.43 3.58 2.00
1.29 2.00 1.29 2.43 3.15 2.15 3.43 7.15



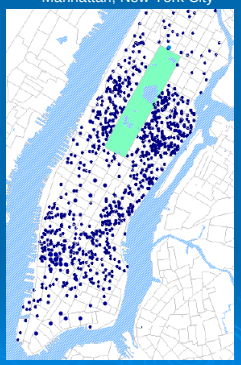

Bio-terror surveillance using Taxi Cabs
(On-going Project with DIMACS)

GPS tracking devise

+

Nuclear sensor devise

Manhattan, New York City



A simulation of taxi cab locations at morning rush hour

## SCAN Statistics

Traditional statistical method to detect a cluster of events is via *Scan Statistic*

- Most commonly used:

  S = Maximum number of cases in a fixed size moving window

  - Equivalent to a generalized likelihood ratio test for a uniform null ( ⎯⎯ ) against a pulse alternative ( ⎍⎍ )

- Variety of Scan Statistics:
  - S = Diameter of the smallest window that contains a fixed number of cases.
  - Other scan statistics
    - Generalized scan statistics
    - Likelihood based tests using a range of fixed window sizes or a range of fixed number of cases
    - Bayesian scan statistics
    - Etc.

*Four books:* Glaz and Balakrishnan '99, Glaz, Naus, and Wallenstein '01, Balakrishnan and Koutras ' 01, Fu and Lou '03

## Scan Statistics

- Very successful in detecting a single significant cluster, and some success in detecting multiple clusters of fixed sizes
  - Specificity in locating the unusual cluster
  - Can take into account the multiple comparisons
- Limited Success in detecting multiple clusters of varying sizes
  - Exist some technical difficulties

## Model Based Methods

- Stochastic process models
  - Localized estimates of likelihood of incidence in temporal spatial data (Diggle et al, 2005)
- "Disease mapping" Approaches
  - Examples: Clayton and Kaldor, 1987, Besag, York and Mollie, 1991, Waller et al, 1997, and Gangnon and Clayton, 2000
- The goal is to describe intensity functions instead of directly detecting and making inference on clusters

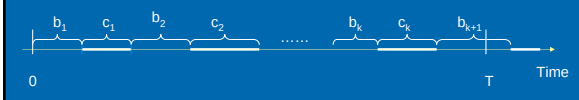
## Stepwise Regression Method

- Molinari et al (2001) proposed a stepwise regression method to detect multiple clusters of varying sizes in temporal data
  - Define responses as inter-arrival times (gaps) between events
  - Fit stepwise regression functions by least square method
  - Use AIC/BIC to determine number of clusters
  - Use Bootstrapped AIC/BIC to test significance of clusters

## Stepwise Regression Method

- The method can detect multiple clusters of varying sizes all together and is easy to program
- Bootstrap test is computationally expensive and not reliable
  - Demattei and Molinari (2007) suggested a new testing method using Beinstein's inequality
- Least square fit may not be efficient for often non-Gaussian responses

## Latent Clustering Model




**Model Assumption:**

Waiting time periods: $b_1, b_2, \ldots, b_{k+1} \sim \psi_b(t)$

Cluster interval lengths: $c_1, c_2, \ldots, c_k \sim \psi_c(t)$

Example: $b_1, b_2, \ldots, b_{k+1} \sim \exp(\lambda_b)$ & $c_1, c_2, \ldots, c_k \sim \exp(\lambda_c)$

*Figure 1.* An illustrative example of a latent model of multiple clusters.

## Data Model

- Observe data:
  - Time points $y_1$, $y_2$, …, $y_n$ when an incidence of interest occurs
  - A multiple step uniform distribution

$$f_\theta(y|\mathbf{b},\mathbf{c},k)=\begin{cases}\frac{\alpha_1}{T+\sum_{j=1}^{k}(\alpha_j-1)c_j}, & \text{if } y\in I_1;\\ \cdots & \\ \frac{\alpha_k}{T+\sum_{j=1}^{k}(\alpha_j-1)c_j}, & \text{if } y\in I_k;\\ \frac{1}{T+\sum_{j=1}^{k}(\alpha_j-1)c_j}, & \text{if } y\notin\cup_{j=1}^{k}I_j,\end{cases}$$

## Model Assumptions

- When $k = 1$ ⟶ single step uniform distribution
  - Used for a single cluster case in Scan statistics literature;
    - See, e.g., Naus (1965), Nagarwalla (1996), and many others
- Alternative expression in terms of Poisson models
  - Current formulation to highlight the interpretation of the parameters $\alpha_j$'s
- Relation to Bayesian Method
  - Further assumption of α's being random + priors on parameters ⟶ a Bayesian hierarchical model

## Incorporation of Background

- Background value may not be the same across the time window (0,T).
  - Such as seasonal patterns or population sizes, etc.
- Known background function W(t), often assessed from separated sources, can be easily incorporated:

$$f_\theta(y|\mathbf{b},\mathbf{c},k)=\begin{cases}\frac{\alpha_1 W(y)}{\tilde{T}+\sum_{j=1}^{k}(\alpha_j-1)\tilde{c}_j}, & \text{if } y\in I_1\\ \cdots\cdots & \\ \frac{\alpha_k W(y)}{\tilde{T}+\sum_{j=1}^{k}(\alpha_j-1)\tilde{c}_j}, & \text{if } y\in I_k\\ \frac{W(y)}{\tilde{T}+\sum_{j=1}^{k}(\alpha_j-1)\tilde{c}_j}, & \text{if } y\notin\cup_{j=1}^{k}I_j\end{cases}$$

where $\tilde{T}=\int_0^T W(t)dt$ and $\tilde{c}_j=\int_{I_j}W(t)dt$

## Likelihood Based Approach

- Treat *k* as a model parameter, together with model parameters α's and λ's
  - Leads to an over-fitting problem
- Solution:
  - Use model selection techniques to determine the number of clusters *k*
  - For a fixed *k*, develop likelihood inference and an EM/MCMC algorithm
    - Estimate model parameters,
    - Detect/test significant clusters,
    - Identify cluster locations.

## Features of Proposed Methods

- Likelihood based approach
  - Large sample efficiency
  - More efficient than the Stepwise Regression method (under assumed model)
- Detecting multiple clusters of varying sizes
  - Multiple clusters all together
  - Most significant or each single cluster
- Others
  - Specificity in locating the unusual clusters
  - Estimation of parameters α's
  - No need to adjust for multiple comparisons
  - Infinity many possible choices of clusters
  - Flexible, with many possible extensions (regression, spatial, etc.)

## Likelihood Function

- Density functions:
  - Probability of having k clusters in (0,T):

$$P_\lambda(\{\delta=k\})=P_\lambda\{\sum_{j=1}^{k}(b_j+c_j)+b_{k+1}\geq T \text{ and } \sum_{j=1}^{k}(b_j+c_j)\leq T\}=\int_0^T\int_{T-s}^{\infty}\psi_b(t)\psi_{bc}^{[k]}(s)dtds.$$

  - Conditional density function of latent variables, give *k*:

$$f_\theta(\mathbf{b},\mathbf{c}|k)=\frac{\prod_{j=1}^{k}\{\psi_b(b_j)\psi_c(c_j)\}\psi_b(b_{k+1})\mathbf{1}_{\{\delta=k\}}}{P_\lambda(\delta=k)}.$$

  - Conditional density function of y's, given fixed *k* clusters:

$$f_\theta(\mathbf{y}|\mathbf{b},\mathbf{c},k)=\prod_{i=1}^{n}f(y_i|\mathbf{b},\mathbf{c},k)=e^{\sum_{j=1}^{k}(\log\alpha_j)Z_j-n\log\{T+\sum_{j=1}^{k}(\alpha_j-1)c_j\}},$$

where $Z_j=Z_j(\mathbf{y},\mathbf{b},\mathbf{c})=\sum_{i=1}^{n}\mathbf{1}_{\{y_i\in I_j\}}$

## Likelihood Function

- Likelihood function of observing $y$'s and having $k$ clusters:

$$\ell_k(\theta|\mathbf{y}) = \log\{f_\theta(\mathbf{y},k)\} = \log\{\int\int f_\theta(\mathbf{y}|\mathbf{b},\mathbf{c},k) f_\theta(\mathbf{b},\mathbf{c}|k) P_\lambda(\delta=k) d\mathbf{b} d\mathbf{c}\}.$$

  - Involve multiple integrations
  - Complicated to directly compute this likelihood function and its first and second derivatives
- Explicit joint density function of ($y$, $b$, $c$, $\delta = k$):

$$\begin{aligned} f_\theta(\mathbf{y},\mathbf{b},\mathbf{c},k) &= f_\theta(\mathbf{y}|\mathbf{b},\mathbf{c},k) f_\theta(\mathbf{b},\mathbf{c}|k) P_\lambda(\{\delta=k\}) \\ &= e^{\sum_{j=1}^k (\log\alpha_j) Z_j - n\log\{T+\sum_{j=1}^k(\alpha_j-1)c_j\}} \prod_{j=1}^k \{\psi_b(b_j)\psi_c(c_j)\}\psi_b(b_{k+1}) 1_{\{\delta=k\}} \end{aligned}$$

→ Use EM algorithm

## EM/MCMC algorithm (EM Steps)

*Step 1.* Select a starting point $\theta^{(0)} = (\alpha^{(0)}, \lambda^{(0)})'$ of $\theta = (\alpha, \lambda)'$.

*Step 2.* (*E*-step) For a given $\theta^{(s)}$ at the sth iteration, $s = 0, 1, 2, \ldots$, calculate

$$Q(\theta|\theta^{(s)}) = Q_1(\alpha|\theta^{(s)}) + Q_2(\lambda|\theta^{(s)}),$$

where

$$\begin{aligned} Q_1(\alpha|\theta^{(s)}) &= \sum_{j=1}^k \mathrm{E}(Z_j|\mathbf{y},k,\theta^{(s)}) \log\alpha_j - n\mathrm{E}[\log\{T+\sum_{j=1}^k(\alpha_j-1)c_j\}|\mathbf{y},k,\theta^{(s)}], \\ Q_2(\lambda|\theta^{(s)}) &= \sum_{j=1}^{k+1} \mathrm{E}\{\log\psi_b(b_j)|\mathbf{y},k,\theta^{(s)}\} + \sum_{j=1}^k \mathrm{E}\{\log\psi_c(c_j)|\mathbf{y},k,\theta^{(s)}\}. \end{aligned}$$

*Step 3.* (*M*-step) For each $s = 0, 1, 2, \ldots$, update the parameter estimates, $\theta^{(s+1)} = (\alpha^{(s+1)}, \lambda^{(s+1)})'$, by maximizing the following functions,

$$\alpha^{(s+1)} = \mathrm{argmax}\, Q_1(\alpha|\theta^{(s)}), \quad \text{and} \quad \lambda^{(s+1)} = \mathrm{argmax}\, Q_2(\lambda|\theta^{(s)}).$$

In the case with $\psi_b$ and $\psi_c$ being density functions of exponential distributions $\mathrm{Exp}(\lambda_b)$ and $\mathrm{Exp}(\lambda_c)$, the updating formula of $\lambda^{(s+1)}$ is simply $\lambda_b^{(s+1)} = (k+1)/\sum_{j=1}^{k+1}\mathrm{E}(b_j|\mathbf{y},k,\theta^{(s)})$ and $\lambda_c^{(s+1)} = k/\sum_{j=1}^k \mathrm{E}(c_j|\mathbf{y},k,\theta^{(s)})$.

*Step 4.* Repeat steps 2 and 3 until $\|\theta^{(s+1)} - \theta^{(s)}\|$ is very small; that is, until the algorithm numerically converges.

## EM/MCMC algorithm (Gibbs Sampling)

- Fully conditional distributions

$$f(b_j|b_l, l=1,2,\ldots,k+1, l\neq j, \mathbf{c},\mathbf{y},k) \propto f(\mathbf{b},\mathbf{c},\mathbf{y}|k) \propto e^{\sum_{s=1}^k Z_s(\log\alpha_s)}\psi_b(b_j)1_{\{\delta=k\}},$$

$$f(c_j|c_l, l=1,2,\ldots,k, l\neq j, \mathbf{b},\mathbf{y},k) \propto f(\mathbf{b},\mathbf{c},\mathbf{y}|k) \propto \frac{e^{\sum_{s=1}^k Z_s(\log\alpha_s)}}{\{T+\sum_{s=1}^k(\alpha_s-1)c_s\}^n}\psi_c(c_j)1_{\{\delta=k\}}.$$

- Gibbs sampling (repeat M times) ➔ M sets of Gibbs Samples

$\mathbf{b}^* = (b_1^*, b_2^*, \ldots, b_{k+1}^*)'$ and $\mathbf{c}^* = (c_1^*, c_2^*, \ldots, c_k^*)'$ from $f(\mathbf{b},\mathbf{c}|\mathbf{y},k,\hat{\theta}^{(\hat{s})})$

- Four expectations in the E-step estimated by

$\sum_* Z_j^*/M$, $\sum_* \log\{T+\sum_{j=1}^k(\alpha_j-1)c_j^*\}/M$, $\sum_* \log\{\psi(b_j^*)\}/M$, and $\sum_* \log\{\psi(c_j^*)\}/M$

## EM/MCMC algorithm (Importance Sampling)

- Simulate from $f(b_j|b_l, l=1,2,\ldots,k+1, l\neq j, \mathbf{c},\mathbf{y},k)$

*Step A.* Simulate a large number, say $N$, random deviates $e_1, e_2, \ldots, e_N$ from a candidate distribution $\tilde{\psi}_b(b_j)$. Then, compute weight $w_l = \{\psi_b(e_l)/\tilde{\psi}_b(e_l)\}e^{\sum_{s=1}^k Z_s^{[l]}(\log\alpha_s)}1_{\{\delta^{[l]}=k\}}$, for $l = 1, \ldots, N$, where $Z_s^{[l]}$ is the total number of incidences in sth cluster and $\{\delta^{[l]} = k\}$ is the constraint of having $k$ clusters but with the $b_j$ being replaced by $e_l$ and the rest of $b$'s and $c$'s kept the same. In the case of simulating $b_{k+1}$ given the rest $b$'s and $c$'s, the weight can be simplified to $w_l = \{\psi_b(e_l)/\tilde{\psi}_b(e_l)\}1_{\{\delta^{[l]}=k\}}$.

*Step B.* Simulate $b_j$ from one of the $N$ values $\{e_1, e_2, \ldots, e_N\}$ with respective probabilities $(p_1, p_2, \ldots, p_N)$; Here, $p_l = w_l/\sum_{s=1}^N w_s$.

- Similar to simulate from $f(c_j|c_l, l=1,2,\ldots,k, l\neq j, \mathbf{b},\mathbf{y},k)$

## EM/MCMC algorithm (Information/variance Estimation)

- Missing Information Principle & Louis Method

$$\begin{aligned} H_n &\overset{d}{=} -\left\{\frac{\partial^2}{\partial\theta^2}\ell_k(\theta|\mathbf{y})\right\} \\ &= -E\left\{\frac{\partial^2}{\partial\theta^2}\ell_k(\theta|\mathbf{b},\mathbf{c},\mathbf{y})|\mathbf{y},\delta=k\right\} - \mathrm{Var}\left\{\frac{\partial}{\partial\theta}\ell_k(\theta|\mathbf{b},\mathbf{c},\mathbf{y})|\mathbf{y},\delta=k\right\}, \end{aligned}$$

where $\ell_k(\theta|\mathbf{b},\mathbf{c},\mathbf{y}) = \log\{f_\theta(\mathbf{y},\mathbf{b},\mathbf{c},k)\}$ is the log-likelihood function of the complete data.

- Estimate of observed information matrix

$$\begin{aligned} \hat{H}_n &= -\frac{1}{M}\sum\frac{\partial^2}{\partial\theta^2}\ell_k(\theta|\mathbf{b}^*,\mathbf{c}^*,\mathbf{y}) \\ &\quad -\left[\frac{1}{M}\sum\left\{\frac{\partial}{\partial\theta}\ell_k(\theta|\mathbf{b}^*,\mathbf{c}^*,\mathbf{y})\right\}^2 - \left\{\frac{1}{M}\sum\frac{\partial}{\partial\theta}\ell_k(\theta|\mathbf{b}^*,\mathbf{c}^*,\mathbf{y})\right\}^2\right] \end{aligned}$$

where the summations are over the set of *M* Gibbs samples b* and c* in the final round of the EM algorithm.

## Likelihood Inference/Cluster Detection

- Wald test for a single cluster
  - Hypotheses:

$H_0 : \alpha_j = 1$ versus $H_1 : \alpha_j \neq 1$ Or $H_0 : \alpha_j = 1$ versus $H_1 : \alpha_j > 1$

  - Test Statistic:

$t = \hat{\alpha}_j / \hat{se}(\hat{\alpha}_j)$ where $\hat{\alpha}_j$ is the estimator of the parameter $\alpha_j$; $\hat{se}(\hat{\alpha}_j)$ is an estimator of the standard error of $\hat{\alpha}_j$

## Likelihood inference/Cluster Detection

- Likelihood ratio test for multiple clusters
  - Hypothesis:

$$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_k = 1 \text{ versus } H_1 : \text{at least one } \alpha_j \neq 1$$

  - Test Statistic:

$$\begin{aligned} R &= 2\log\{\frac{\max_{H_1} f_\theta(\mathbf{y},k)}{\max_{H_0} f_\theta(\mathbf{y},k)}\} = 2\ell_k(\hat{\theta}|\mathbf{y})|_{\theta=\hat{\theta}} + 2n\log(T) - 2\max_\lambda \log P_\lambda(\delta=k) \\ &= 2\{\log\int\int f_\theta(\mathbf{y}|\mathbf{b},\mathbf{c},k) f_\theta(\mathbf{b},\mathbf{c}|k)d\mathbf{b}d\mathbf{c} + \log P_\lambda(\delta=k) + n\log(T) - \max_\lambda \log P_\lambda(\delta=k)\} \end{aligned}$$

Calculation

$$R^{**} = 2\left[\log\{\frac{1}{M}\sum_{**} f(\mathbf{y}|\mathbf{b}^{**},\mathbf{c}^{**},k)\} + \log P_\lambda(\delta=k) + n\log(T) - \max_\lambda \log P_\lambda(\delta=k)\right]$$

where $\sum_{**}$ the summation over the $M$ sets of $\mathbf{b}^{**}$ and $\mathbf{c}^{**}$ samples from $f_\theta(\mathbf{b},\mathbf{c}|k)$.

## Cluster Location Estimation

- The lower and upper bounds of the *j*th cluster:

$$L_j = \sum_{l=1}^{j-1}(b_l + c_l) + b_j \;\&\; U_j = \sum_{l=1}^{j}(b_l + c_l)$$

  - M Gibbs copies L and U (from the last iteration of EM)

$$L_j^* = \sum_{l=1}^{j-1}(b_l^* + c_l^*) + b_j^* \;\&\; U_j^* = \sum_{l=1}^{j}(b_l^* + c_l^*)$$

- Estimation methods
  - Means: Mean ($L_j^*$) & Mean($U_j^*$)
  - Medians: Median ($L_j^*$) & Median($U_j^*$)

## Cluster Location Estimation

- Four measurements for accuracy (in simulation)

Sensitivity = $\dfrac{\text{Number of A} \cap \text{B}}{\text{Number of A}}$

PPV = (Positive Predictive Value) $\dfrac{\text{Number of A} \cap \text{B}}{\text{Number of B}}$

Specificity = $\dfrac{\text{Number of A}^c \cap \text{B}^c}{\text{Number of A}^c}$

NPV = (Negative Predictive Value) $\dfrac{\text{Number of A}^c \cap \text{B}^c}{\text{Number of B}^c}$

A = { $y$ | $y$ is truly inside a cluster (the truth) }
B = { $y$ | $y$ is estimated inside a cluster (estimation)}

## Determine the Number of Clusters *k*

- AIC and BIC criteria:

$$\begin{aligned} \text{AIC}(k) &= -2\log f_\theta(\mathbf{y},k) + 2k \\ &= -2\log\{\int\int f_\theta(\mathbf{y}|\mathbf{b},\mathbf{c},k) f_\theta(\mathbf{b},\mathbf{c}|k)d\mathbf{b}d\mathbf{c}\} - 2\log P_\lambda(\delta=k) + 2k \\ \text{BIC}(k) &= -2\log f_\theta(\mathbf{y},k) + k\log(n) \\ &= -2\log\{\int\int f_\theta(\mathbf{y}|\mathbf{b},\mathbf{c},k) f_\theta(\mathbf{b},\mathbf{c}|k)d\mathbf{b}d\mathbf{c}\} - 2\log P_\lambda(\delta=k) + k\log(n) \end{aligned}$$

Calculation

$$\widehat{\text{AIC}}(k) = -2\log\{\frac{1}{M}\sum_{**} f(\mathbf{y}|\mathbf{b}^{**},\mathbf{c}^{**},k)\} - 2\log P_\lambda(\delta=k) + 2k,$$

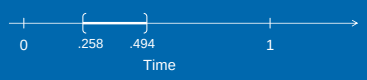
$$\widehat{\text{BIC}}(k) = -2\log\{\frac{1}{M}\sum_{**} f(\mathbf{y}|\mathbf{b}^{**},\mathbf{c}^{**},k)\} - 2\log P_\lambda(\delta=k) + k\log(n),$$

where $\sum_{**}$ the summation over the $M$ sets of $\mathbf{b}^{**}$ and $\mathbf{c}^{**}$ samples from $f_\theta(\mathbf{b},\mathbf{c}|k)$.

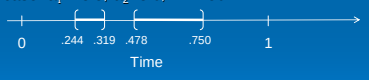
- Pick *k* that minimizes $\widehat{\text{AIC}}(k)$ or $\widehat{\text{BIC}}(k)$

## Simulation Study I (setting one: fixed clusters)

- *k* = 1 case: α = 3.0, n = 100

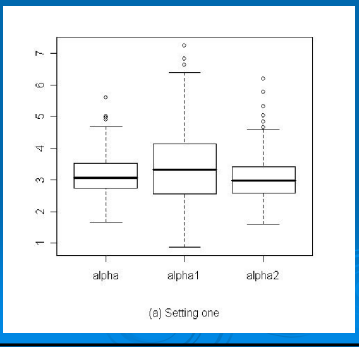



*k* = 2 case: $\alpha_1$ = 3.5, $\alpha_2$=3.0, n = 150




- In each case
  - Simulate a data set and apply the proposed methodology
  - Repeat 300 times

## Parameter Estimation




(a) Setting one

Table 1. Simulated Powers and Sizes for Wald and LRT Tests

| | $k=1$ | | | | | |
|---|---|---|---|---|---|---|
| | Power | | | Size | | |
| Method | Wald | LRT | (WLS) | Wald | LRT | (WLS) |
| Reject | 299 | 298 | (180) | 28 | 10 | (50) |
| Accept | 1 | 2 | (120) | 272 | 290 | (250) |
| Total | 300 | 300 | (300) | 300 | 300 | (300) |

$k = 2$
$c_1$=.075, $c_2$=.272
$\alpha_1$= 3.5, $\alpha_2$ =3.0
$n$ = 150

$k = 1$
$c_1$= .236
$\alpha$ = 3.0
$n$ = 100

| | $k=2$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Power | | | | Size | | | |
| Method | Wald | | LRT | (SR) | Wald | | LRT | (SR) |
| | 1 | 2 | | | 1 | 2 | | |
| Reject | 258 | 297 | 297 | (193) | 9 | 16 | 3 | (78) |
| Accept | 42 | 3 | 3 | (107) | 291 | 284 | 297 | (222) |
| Total | 300 | 300 | 300 | (300) | 300 | 300 | 300 | (300) |

Table 2. Sensitivity, Specificity, PPV and NPV Summary

| | Method | Statistics | Min | 1st QT | Median | 3rd QT | Max | Mean |
|---|---|---|---|---|---|---|---|---|
| $k=1$ | Mean based | Sensitivity | 58.33 | 92.94 | 98.11 | 100 | 100 | 95.60 |
| | | Specificity | 43.40 | 94.12 | 98.00 | 100 | 100 | 95.35 |
| | | PPV | 61.04 | 94.12 | 98.00 | 100 | 100 | 95.49 |
| | | NPV | 79.10 | 94.06 | 98.08 | 100 | 100 | 96.44 |
| | Median based | Sensitivity | 69.77 | 94.12 | 98.15 | 100 | 100 | 96.19 |
| | | Specificity | 43.40 | 94.06 | 98.04 | 100 | 100 | 95.21 |
| | | PPV | 61.04 | 94.29 | 98.11 | 100 | 100 | 95.44 |
| | | NPV | 79.69 | 95.06 | 98.15 | 100 | 100 | 96.93 |
| | (SR) | Sensitivity | 0 | 2.15 | 100 | 100 | 100 | 69.00 |
| | | Specificity | 23.53 | 84.81 | 96.49 | 99.22 | 100 | 89.26 |
| | | PPV | 0 | 59.32 | 90.91 | 99.23 | 100 | 69.64 |
| | | NPV | 36.11 | 59.09 | 100 | 100 | 100 | 84.82 |
| $k=2$ | Mean based | Sensitivity | 39.77 | 93.62 | 97.14 | 99.02 | 100 | 95.13 |
| | | Specificity | 39.06 | 85.00 | 93.10 | 98.02 | 100 | 89.57 |
| | | PPV | 62.16 | 91.11 | 95.79 | 98.87 | 100 | 93.83 |
| | | NPV | 49.52 | 89.47 | 95.00 | 98.33 | 100 | 93.11 |
| | Median based | Sensitivity | 34.09 | 94.38 | 97.78 | 100 | 100 | 95.18 |
| | | Specificity | 35.94 | 86.00 | 92.73 | 98.25 | 100 | 89.33 |
| | | PPV | 67.46 | 91.37 | 95.83 | 98.96 | 100 | 93.71 |
| | | NPV | 48.67 | 90.65 | 95.83 | 100 | 100 | 93.36 |
| | (SR) | Sensitivity | 0 | 59.89 | 97.78 | 100 | 100 | 74.71 |
| | | Specificity | 9.23 | 55.65 | 75.00 | 90.81 | 100 | 70.79 |
| | | PPV | 0 | 70.37 | 84.00 | 94.15 | 100 | 73.06 |
| | | NPV | 6.82 | 59.22 | 96.00 | 100 | 100 | 76.38 |

Table 3. AIC and BIC Model Selection

| | AIC | | | | BIC | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Estimated | | | | Estimated | | | | |
| | $k=1$ | $k=2$ | $k=3$ | $k=4$ | $k=1$ | $k=2$ | $k=3$ | $k=4$ | Total |
| True $k=1$ | 282 | 9 | 4 | 5 | 299 | 1 | 0 | 0 | 300 |
| $k=2$ | 46 | 203 | 38 | 13 | 110 | 184 | 5 | 1 | 300 |

## Simulation Study II
## (setting two: non-repeating clusters)

- $k$ = 1 case: $\alpha$ = 3.0, n = 100
  latent: $b_1$, $b_2$ ~ exp($\lambda_b$= 1.2) &
  $c_1$ ~ exp($\lambda_c$ = 5.2)

- $k$ = 2 case: $\alpha_1$ = 3.5, $\alpha_2$=3.0, n = 150
  latent: $b_1$, $b_2$, $b_3$ ~ exp($\lambda_b$ = 4) &
  $c_1$, $c_2$ ~ exp($\lambda_c$ = 3)

## Parameter Estimation

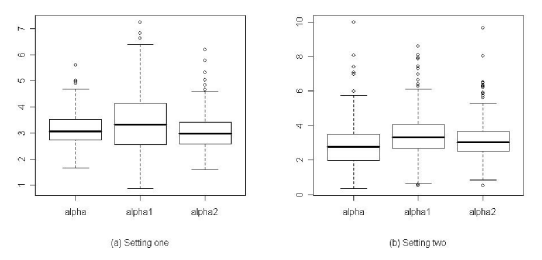



Table 4. Part I: Simulated Powers and Sizes for Wald and LRT Tests

| | $k=1$ | | | | $k=2$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Power | | Size | | Power | | | Size | | |
| Method | Wald | LRT | Wald | LRT | Wald | | LRT | Wald | | LRT |
| | | | | | 1 | 2 | | 1 | 2 | |
| Reject $H_0$ | 223 | 203 | 30 | 5 | 258 | 269 | 287 | 19 | 30 | 5 |
| Accept $H_0$ | 77 | 97 | 270 | 295 | 42 | 31 | 13 | 281 | 270 | 295 |
| Total | 300 | 300 | 300 | 300 | 300 | 300 | 300 | 300 | 300 | 300 |

Part II. Sensitivity, Specificity, PVP and PVN Summary

| | Method | Statistics | Min | 1st QT | Median | 3rd QT | Max | Mean |
|---|---|---|---|---|---|---|---|---|
| $k=1$ | Mean | Sensitivity | 0 | 59.41 | 94.87 | 100 | 100 | 73.21 |
| | | Specificity | 0 | 88.71 | 95.96 | 100 | 100 | 88.14 |
| | | PVP | 0 | 21.62 | 94.00 | 100 | 100 | 67.17 |
| | | PVN | 0 | 90.53 | 96.24 | 100 | 100 | 90.16 |
| | Median | Sensitivity | 0 | 80.00 | 96.15 | 100 | 100 | 78.25 |
| | | Specificity | 0 | 88.88 | 96.94 | 100 | 100 | 87.95 |
| | | PVP | 0 | 42.71 | 94.12 | 100 | 100 | 70.85 |
| | | PVN | 0 | 92.51 | 98.04 | 100 | 100 | 91.12 |
| $k=2$ | Mean | Sensitivity | 0 | 88.21 | 95.16 | 98.76 | 100 | 90.65 |
| | | Specificity | 14.29 | 73.75 | 87.10 | 96.49 | 100 | 82.35 |
| | | PVP | 0 | 89.69 | 95.50 | 98.37 | 100 | 91.33 |
| | | PVN | 8.33 | 69.39 | 88.00 | 96.95 | 100 | 80.82 |
| | Median | Sensitivity | 0 | 89.66 | 96.19 | 99.08 | 100 | 91.53 |
| | | Specificity | 29.27 | 73.53 | 90.11 | 98.15 | 100 | 84.10 |
| | | PVP | 0 | 90.30 | 95.83 | 99.12 | 100 | 91.12 |
| | | PVN | 8.60 | 76.00 | 89.29 | 97.50 | 100 | 84.28 |

Table 5. AIC and BIC Model Selection

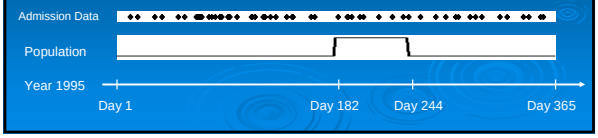
| | AIC | | | | BIC | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Estimated | | | | Estimated | | | | |
| | $k=1$ | $k=2$ | $k=3$ | $k=4$ | $k=1$ | $k=2$ | $k=3$ | $k=4$ | Total |
| True $k=1$ | 272 | 21 | 5 | 2 | 294 | 6 | 0 | 0 | 300 |
| $k=2$ | 136 | 113 | 38 | 13 | 194 | 95 | 6 | 5 | 300 |

## Hospital Hemoptysis Admission Data

Days of hemoptysis admission at Nice University Hospital from January 1 to December 31, 1995 (Molinari, et al. 2001):

2 8 23 29 43 48 58 60 61 63
69 71 74 74 78 80 85 86 86 87
93 105 106 108 115 117 121 126 135 140
141 156 159 179 187 188 188 191 191 198
201 214 225 225 235 235 239 249 262 271
279 279 282 292 296 302 317 323 337 342
352 354

Nice, France (south)




## Hospital Hemoptysis Admission Data Example

Cluster Detection:

| | K =1 | | K = 2 | | |
|---|---|---|---|---|---|
| | Cluster Interval | p-value | Cluster Interval 1 | Cluster Interval 2 | p-values |
| (WLS) | [58,87] (Feb. 27-Mar.28) | .02* .13* | [58.87] (Feb.27-Mar.28) | [187,201] (Jul.6-Jul.20) | .30 |
| Proposed | [58, 108] (Feb.27-Apr.18) | .141(Wald) .262(LRT) | [58, 108] (Feb.27-Apr.18) | [198,235] (Jul.17 – Aug.23) | .107(Wald) .837(Wald) .330 (LRT) |

Demattei and Molinari's (2006)

Parameter Estimation:

k = 1 case: $\hat{\alpha} = 1.961$

k = 2 case: $\hat{\alpha}_1 = 1.935$, $\hat{\alpha}_2 = 1.117$

Conclusion: No significant cluster!

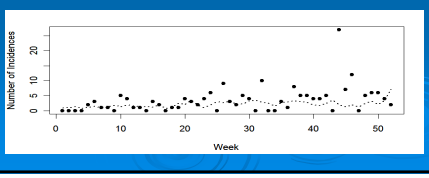
## 2004 Weekly Brucellosis Incidence Data (Collected by CDC)

2004 Data (number of incidences by week):

0 0 0 0 2 3 1 1 0 5 4 1 1 0 3 2 0 1 1 4 3 2 4 6 0 9 3
2 5 4 0 10 0 0 3 1 8 5 5 4 4 5 0 27 7 12 0 5 6 6 4 2

1997-2003 Average:

0.86 1.00 1.29 0.72 1.15 1.43 0.86 1.29 1.86 1.29 2.00
1.58 1.29 1.29 1.15 2.00 0.72 1.29 2.43 2.15 3.58 1.86
1.00 2.00 3.00 2.58 3.29 2.15 2.29 3.29 3.43 2.72 2.43
1.58 2.58 2.86 3.29 3.00 2.72 1.86 1.72 2.43 3.58 2.00
1.29 2.00 1.29 2.43 3.15 2.15 3.43 7.15




## CDC Brucellosis Data Example

Cluster Detection:

| | K =1 | | K = 2 | | |
|---|---|---|---|---|---|
| | Cluster Interval | p-value | Cluster Interval 1 | Cluster Interval 2 | p-values |
| Proposed | [44, 46] (Oct.6-20) | <.001(Wald) <.001(LRT) | [20, 24] (Apr.28-May.19) | [44,46] (Oct.6–20) | .12(Wald) <.001(Wald) <.001 (LRT) |

Parameter Estimation:

k = 1 case: $\hat{\alpha} = 6.574$

k = 2 case: $\hat{\alpha}_1 = 1.088$, $\hat{\alpha}_2 = 6.715$

Conclusion: One significant cluster week 44 to week 46.

## Summary

- Mimic a typical procedure of cluster generation
  - Develop a latent model to model clusters
  - Given clusters, data come from multiple step uniform function
- Likelihood approach applicable
  - EM/MCMC for estimation and likelihood inference for testing
  - Model selection criteria to determine the number of clusters
- Efficient approach to detect multiple clusters of varying sizes
  - Extension for spatial data (Sun, 2008 – Thesis)

## Thank You!