A Spatial Attraction-Repulsion Point Process for Viral Infections

Murali Haran

Department of Statistics, Pennsylvania State University

ISBA 2014, Cancun, Mexico.
July 2014

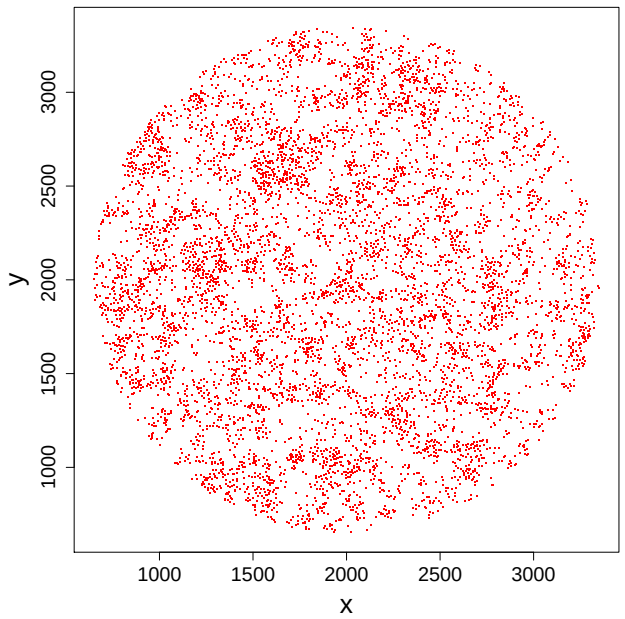
(joint work with **Josh Goldstein**, Ivan Simeonov, John Fricks, and Francesca Chiaromonte)

Modeling Virus Infections

- Of interest: investigating the progression of viral infections
- Our goal: use data from imaging of cell cultures to study the spatial structure of an infection under different conditions, e.g. individual strains, both together, different time lags
- An *in vitro* cell culture study identifies and locates cells infected with two strains of the human respiratory syncytial virus (RSV-A and RSV-B)

Question:

How does the presence of an infected cell impact infections in neighboring cells?



Points represent locations of cells infected with RSV.

Outline

- Spatial point processes in the plane provide a natural framework here
- Each point represents the 2D coordinates of an infected cell
- Goal: Infer spatial interaction among cells

Contributions of this work:

- A new spatial attraction-repulsion point process model
- Inferential methods for this computationally challenging problem
- Draw scientific conclusions from fitting this model to the RSV data

Spatial Point Processes

A spatial point process is a stochastic process, a realization of which consists of a set of points $X = (x_1, ..., x_n)$ in a bounded region $W \subseteq \mathbb{R}^d$.

Some SPPs can be used to model interactions:

$$f(X|\Theta) = \lambda^n \prod_{i \neq j} \phi(x_i, x_j)$$

where $\phi(x_i, x_j)$ is the *interaction function* between points i and j.
In the homogeneous case, $\phi(x_i, x_j) = \phi(\|x_i - x_j\|) = \phi(r)$

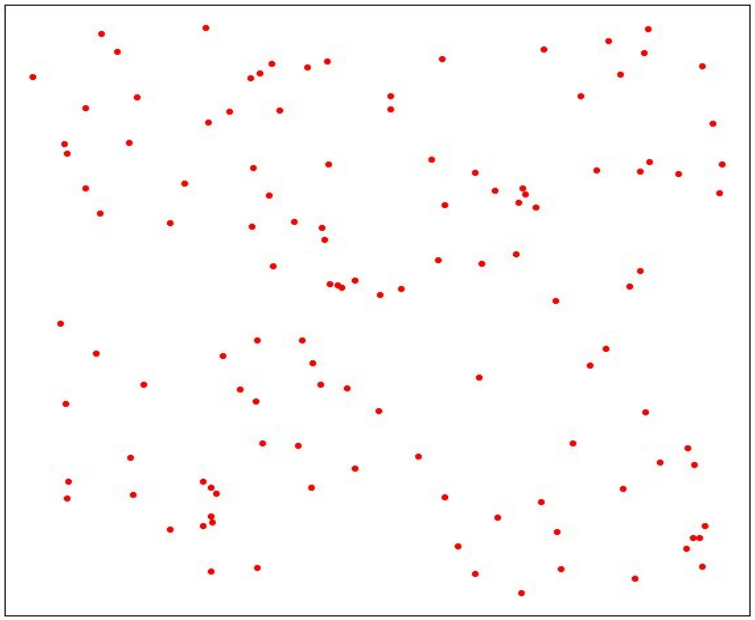
The Strauss process (Strauss, 1975; Geyer, 1999) is a simple example,

$$\phi(r) = \begin{cases} \gamma, & 0 < r \leq R \\ 1, & r > R \end{cases}$$

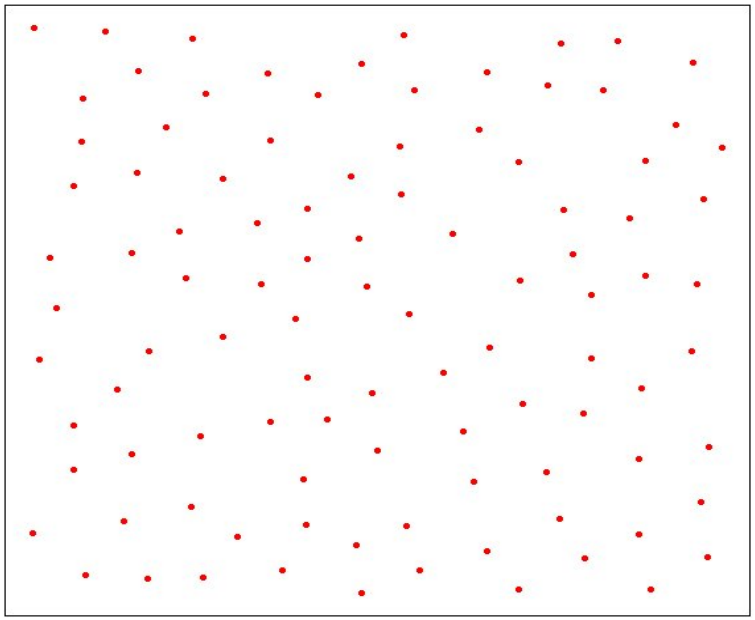
for $0 \leq \gamma \leq 1$. Since $\phi(r) \leq 1$ this is a repulsion point process.

Poisson Process vs. Strauss Process

Realization of Poisson Process



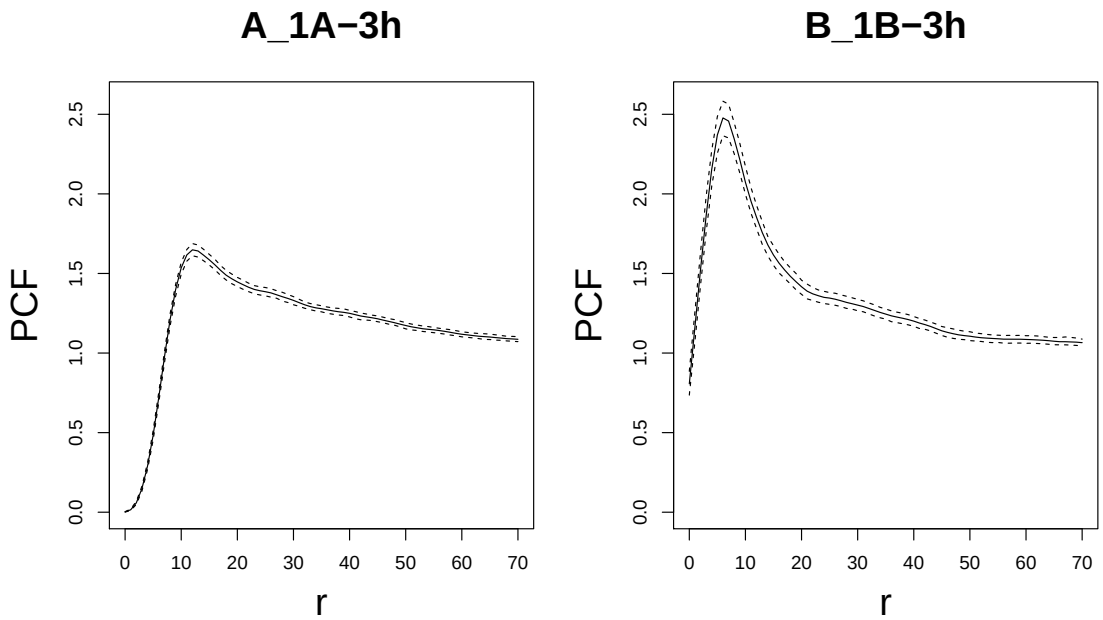
Realization of Strauss Process



Exploratory Analysis of RSV Data: Need for a New Model

The pair correlation function (PCF) $g(r)$ is an exploratory summary statistic that tells us the attraction-repulsion behavior of points separated by distance r in a spatial point process.

- A value of $g(r) > 1$ indicates attraction, a tendency for points to cluster at distance r. Similarly, $g(r) < 1$ indicates repulsion at distance r. For our data:



Observed attraction-repulsion in RSV data varies smoothly in r.

New Attraction-repulsion Model: Interaction Function

Goal: Allow attraction-repulsion to vary smoothly with distance to model observed RSV behavior.

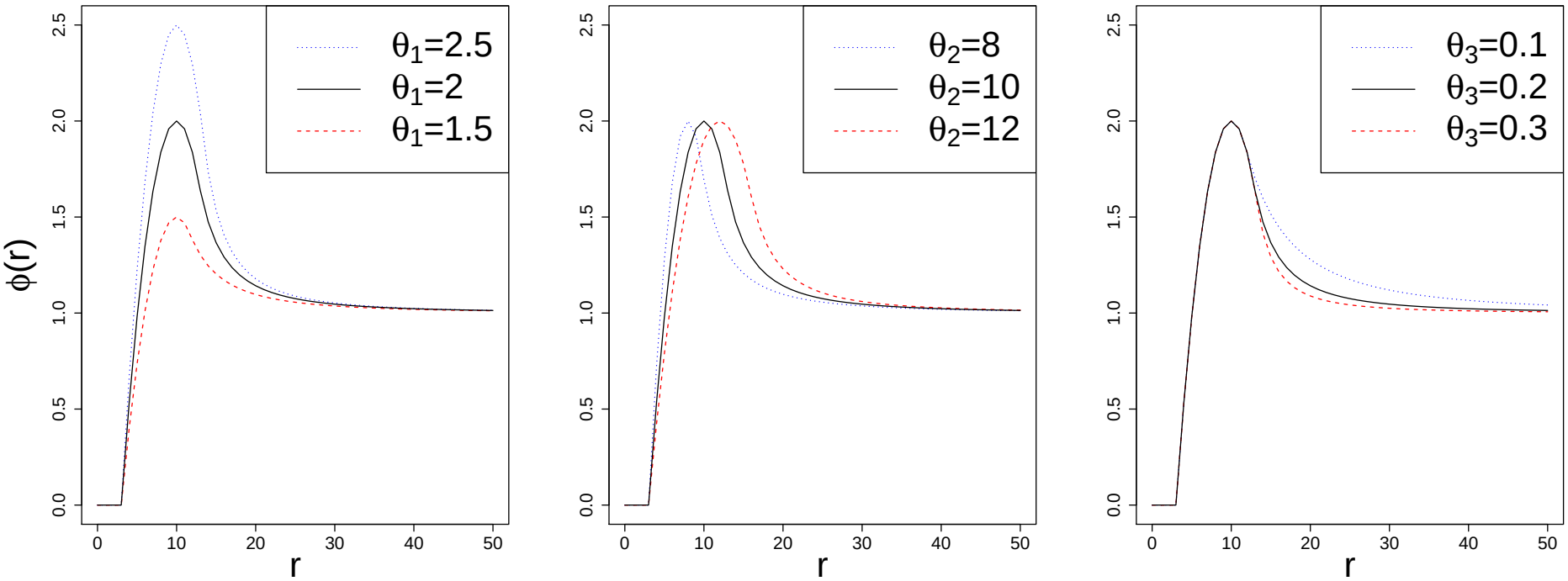
The interaction function $\phi(r)$ is defined piecewise,

$$\phi(r) = \begin{cases} 0, & 0 \leq r \leq R \\ \theta_1 - \left(\dfrac{\sqrt{\theta_1}}{\theta_2 - R}(r - \theta_2) \right)^2, & R < r \leq r_1 \\ 1 + \dfrac{1}{(\theta_3(r - r_2))^2}, & r > r_1 \end{cases}$$

where

- θ_1: value of $\phi(\cdot)$ at the peak (height)
- θ_2: the value of r at the peak (location of peak)
- θ_3: rate of descent after the peak
- R: minimum allowable distance between points
- r_1, r_2: chosen to ensure $\phi(\cdot)$ is continuously differentiable.

New attraction-repulsion model: Interaction function



- $\phi(r) > 1$: attraction, points tend to cluster at distance r
- $\phi(r) < 1$: repulsion

Attraction-repulsion Model

The likelihood can be written as

$$\mathcal{L}(X|\Theta) = \frac{f(X|\Theta)}{c(\Theta)}, f(X|\Theta) = \lambda^n \Big[\prod_{i=1}^{n} e^{\min\left[\sum_{i \neq j} log(\phi(x_i, x_j)), k\right]} \Big]$$

Model parameters:

- λ is the intensity of the process
- $\theta_1, \theta_2, \theta_3$ control the shape of $\phi(r)$.
- R is the minimum distance allowed between points
- k: truncation constant to prevent degenerate "clumping" behavior

Important: $c(\Theta)$ is intractable. This makes computing very challenging.

Inference

- Let $\Theta = (\lambda, k, \theta_1, \theta_2, \theta_3)$. Likelihood $\mathcal{L}(\Theta; X)$.
- *(Pretend we are not at an ISBA conference)*
  First investigate maximum likelihood-based inference:
  - Unknown normalizing function poses a major challenge.
  - MCMC-maximum likelihood (e.g. Geyer and Thompson, 1992), should work in principle: essentially importance sampling combined with MCMC.
    - Problem 1: Good initial values are crucial *and* difficult to obtain.
    - Problem 2: Standard errors are difficult because we need analytical gradients of unnormalized likelihood. Difficult/intractable for our model.
- Bayesian inference (surprisingly?) offers computational tractability and convenience

Bayesian Inference

- Bayesian inference to the rescue: turns out that, although it is challenging, computing for Bayesian inference is feasible even though MCMC-MLE is not.
- Bayesian inference for Θ is based on the posterior distribution

$$\pi(\Theta|X) \propto \mathcal{L}(X|\Theta)p(\Theta) = \frac{f(X|\Theta)p(\Theta)}{c(\Theta)}$$

- Markov chain Monte Carlo (MCMC) is a convenient approach to learning about $\pi(\Theta|X)$.
- Choose a gamma prior on k, prior for remaining parameters are uniform over a scientifically plausible range.

MCMC

- Construct a Markov chain with stationary distribution $\pi(\Theta|X)$.
- In MCMC, propose Θ' from $q(\Theta, \Theta')$ and calculate the following acceptance probablity:

$$\alpha = \min\left(1, \frac{p(\Theta')q(\Theta', \Theta)f(X|\Theta')}{p(\Theta)q(\Theta, \Theta')f(X|\Theta)} \frac{c(\Theta)}{c(\Theta')}\right)$$

- The intractable normalizing constant $c(\Theta)$ does not cancel. Traditional MCMC methods cannot be applied.

Solution: Introduce an auxiliary variable.

- Perfect sampling for $f(X \mid \Theta)$ unavailable so cannot use Møller et al. (2006) and Murray et al. (2007).
- Double Metropolis-Hastings algorithm of Liang (2010).
- Two nested MCMC samplers; the "inner sampler" generates an auxiliary point pattern at each step of the "outer" sampler. Normalizing constants cancel so do not need their ratio.

Computational Challenges

The largest datasets consist of 13,000-14,000 spatial locations. For data this large, the nested samplers are expensive; the inner sampler must be run for thousands of iterations at each step of the outer sampler.

- Inner sampler updates fast in practice since we only propose to add or remove a single point (birth-death sampler)
- R too slow, code in C and optimize.
- Greatly reduce computing by truncating the interaction function for large values of r (evaluate $\phi(r)$ to 1 when $r > R_{max}$).
- Inference for three replicates of the largest dataset takes a few days on our linux clusters.

Conclusions

- Our model captures the complex scale-varying attraction and repulsion behavior observed in the RSV dataset; this flexibility is not available in existing models
- Inference works well for simulated examples – we can recover the truth. Also verified goodness-of-fit.
- Parametric specification of the interaction function lets us draw meaningful conclusions based on parameter inference
- Draw meaningful scientific conclusions as a result of inference on model parameters across multiple RSV experiments. Answer questions about infection status of a cell as a function of proximity. For e.g. RSV-B infected cells have a higher propensity to lump together than RSV-A; suggests RSV-B induces stronger increase in susceptibility to infection.
- Computational challenges are still considerable.

Goldstein, Haran, Simeonov, Fricks, Chiaromonte (2014) (on `arxiv.org`)