A Spatial Point Process Model for Viral Infections

Murali Haran

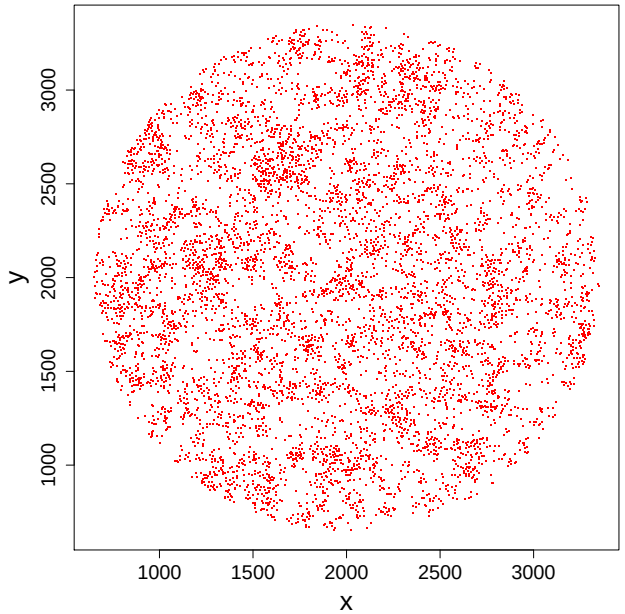
Department of Statistics, Pennsylvania State University

TIES, Al Ain, United Arab Emirates.
November, 2015

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

Spatial point process modeling of virus infections

- Biologists are often interested in investigating the progression of viral infections
- Our goal: use data from imaging of cell cultures to study the spatial structure of an infection
- 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)



Points represent locations of cells infected with RSV.

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

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 point process model for the RSV data
- Inferential methods for this computationally challenging problem
- Draw scientific conclusions from fitted model

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, $\phi(x_i, x_j) = \phi(\|x_i - x_j\|) = \phi(r)$

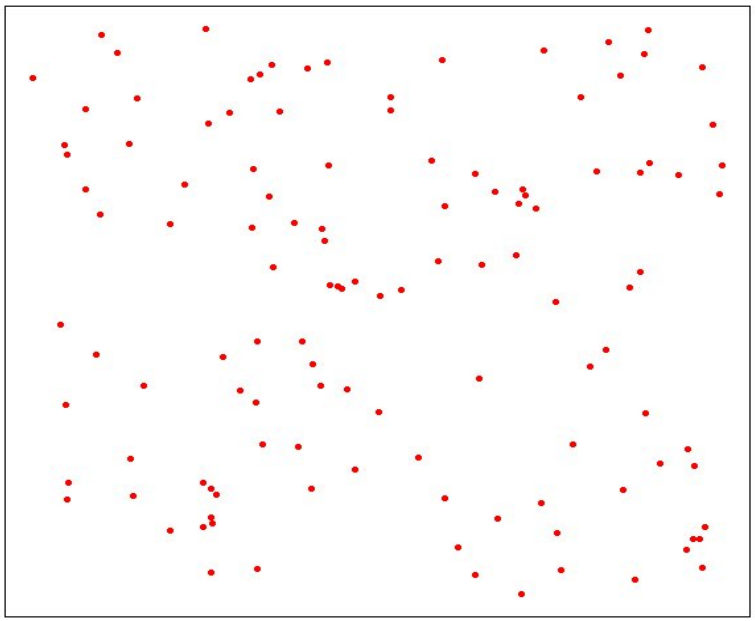
The Strauss process is a simple example,

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

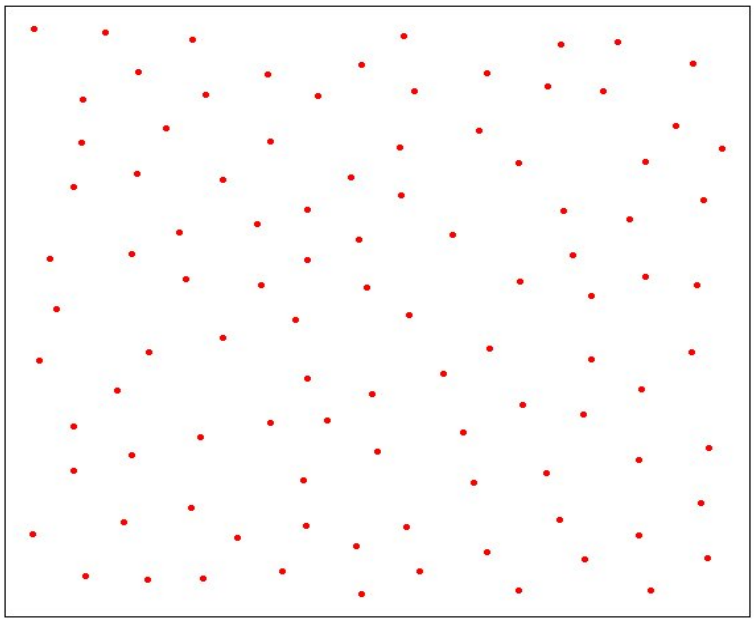
for $0 \leq \gamma \leq 1$. Since $\phi(r) \leq 1$ this is a repulsion point process.

Poisson process vs. Strauss process

Realization of Poisson Process



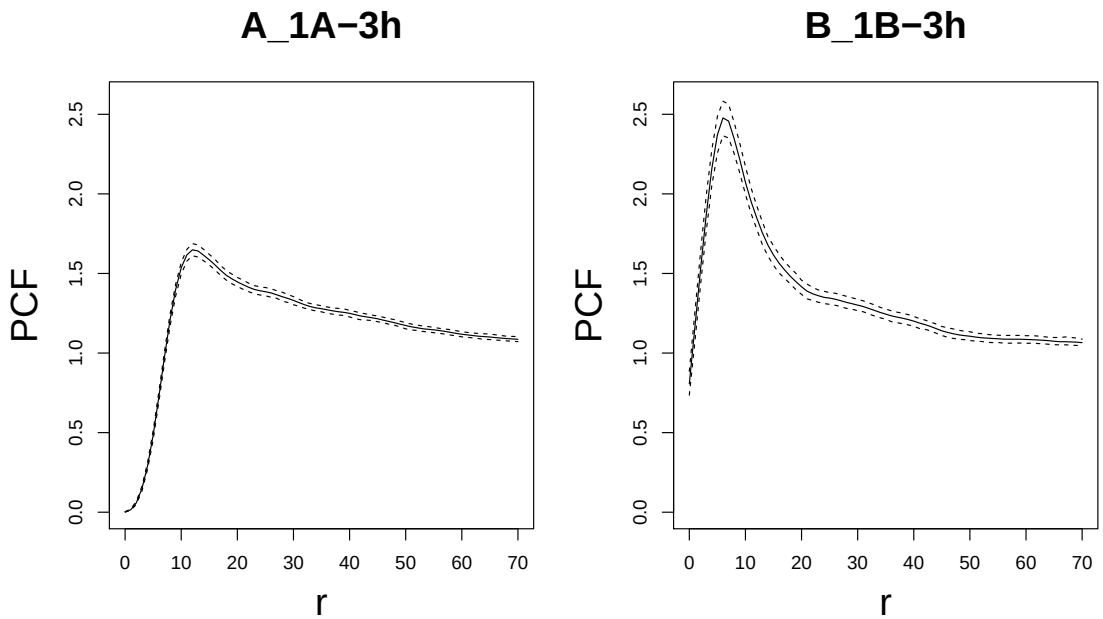
Realization of Strauss Process



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.

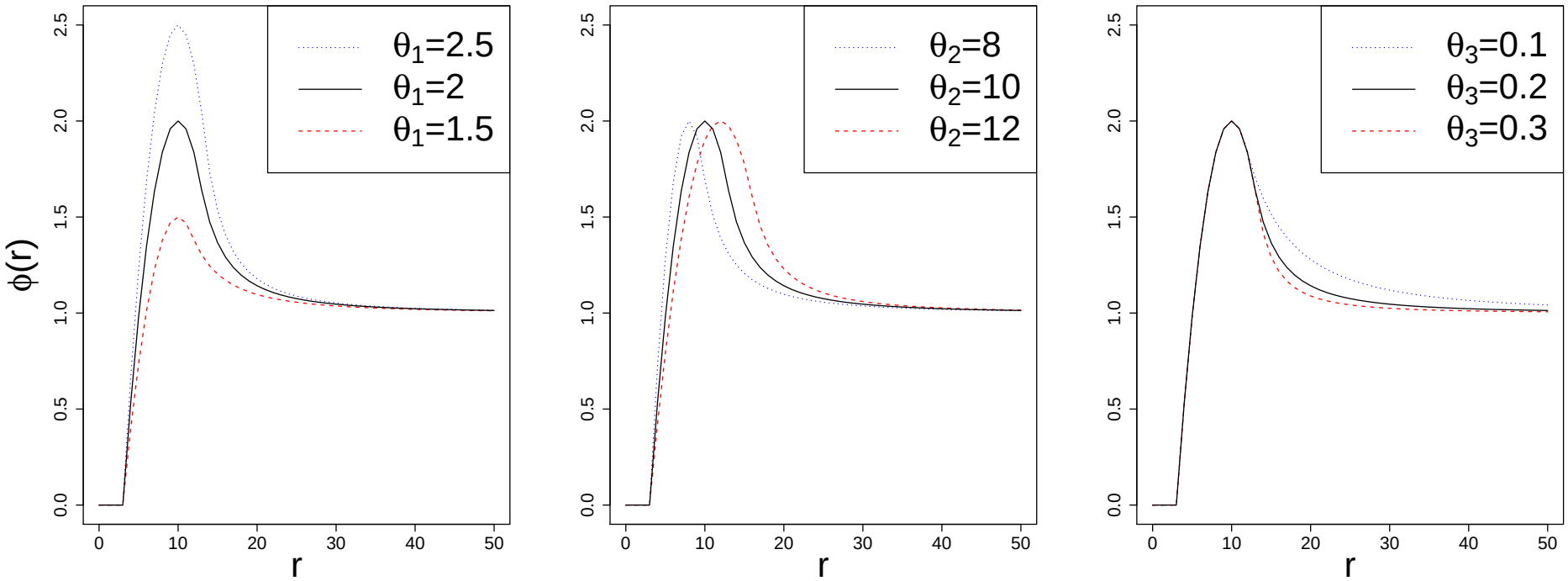
The interaction function $\phi(r)$ is defined piecewise,

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

where

- θ_1: value of $\phi(\cdot)$ at the peak
- θ_2: the value of r at the 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 is a truncation constant necessary to prevent "clumping" behavior

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

Inference

Let $\Theta = (\lambda, k, \theta_1, \theta_2, \theta_3)$. Likelihood $\mathcal{L}(X|\Theta)$.
Is maximum likelihood an option?

- Unknown normalizing function poses a major challenge.
- Existing methods: MCMC-maximum likelihood (e.g. Geyer and Thompson, 1992), essentially importance sampling combined with MCMC, should work in principle.
- However, this algorithm is infeasible in practice:
  - Problem 1: Initial values are important *and* difficult to obtain.
  - Problem 2: Standard errors are difficult because we need an estimate of inverse Hessian, in turn need estimate of gradients of unnormalized likelihood. Difficult/intractable for our model.
- Bayesian inference (contrary to what some people might think?) is more tractable and convenient in this normalizing function problem.

Bayesian Inference

- Bayesian inference to the rescue: turns out that, although it is challenging, computing for Bayesian inference is feasible even though MCMCMLE 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.

Auxiliary Variable MCMC algorithm

Idea: Introduce an auxiliary variable. That is, augment the target distribution.

- 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.

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 the Lion-X cluster.

Conclusions

Can make meaningful scientific conclusions as a result of inference on model parameters across multiple RSV experiments (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).

Advantages of our method:

- Model captures the complex scale-varying attraction and repulsion behavior observed in the RSV dataset; this flexibility is not available in other models
- Parametric specification of the interaction function lets us draw meaningful conclusions based on parameter inference
- We use a simulation-based approach with PCFs to check the goodness of fit for our model to the data.
- Inference works well for simulated examples – we recover truth.

Goldstein, Haran, Simeonov, Fricks, Chiaromonte *Biometrics*, **2015.**