CBMS Lecture 8

Alan E. Gelfand
Duke University

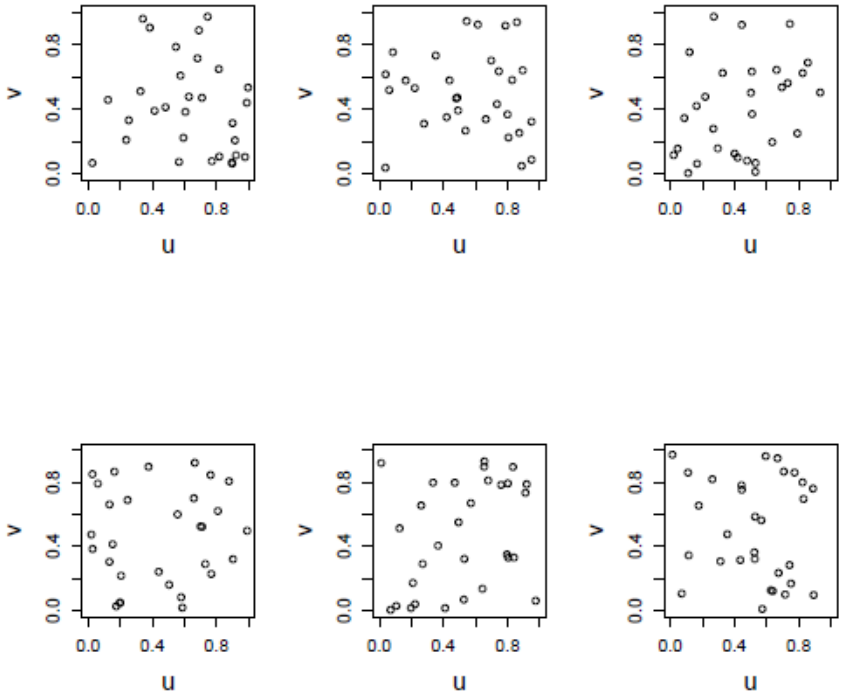
Outline

- Basics of spatial point patterns
- Diagnostic tools
- Models
- Model fitting within a Bayesian framework
- Posterior inference using simulated point patterns
- GNZ formula and variants
- Residual analysis, model adequacy, model comparison
- Examples

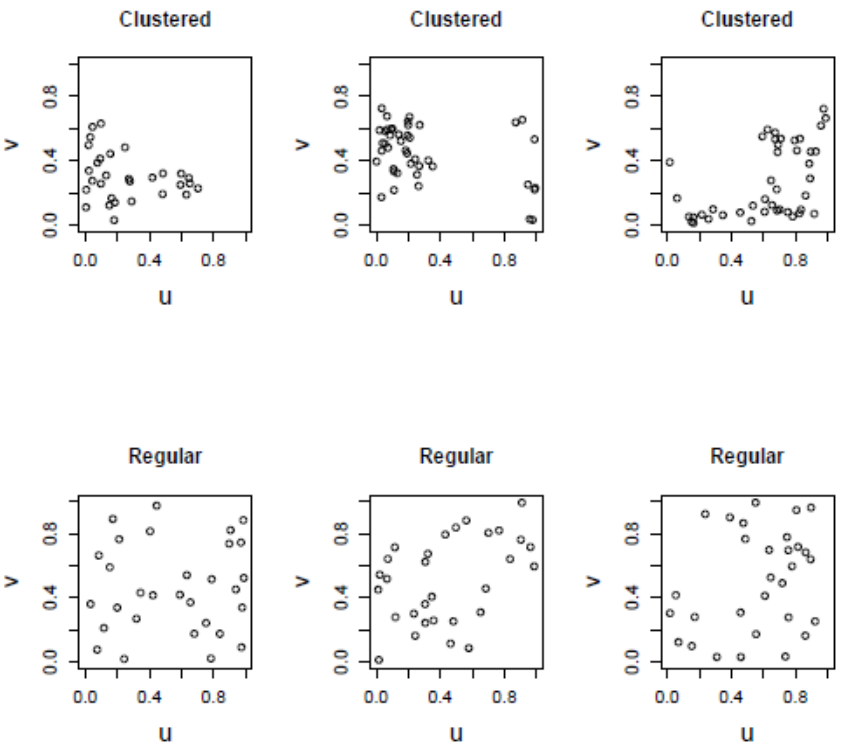
What is a point pattern?

- For a specified, bounded region D, a set of locations $s_i, i = 1, 2..., n$
- The locations are viewed as "random"
- Need not have variables at locations, just the pattern of points
- Crude features of patterns, e.g., complete randomness, clustering/attraction, inhibition/repulsion, regular/systematic
- Can add "marks", i.e., labels. Then, a point pattern for each mark; comparison of patterns

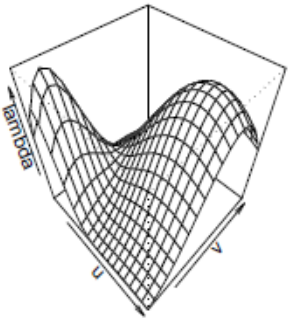
spatial homogeneity

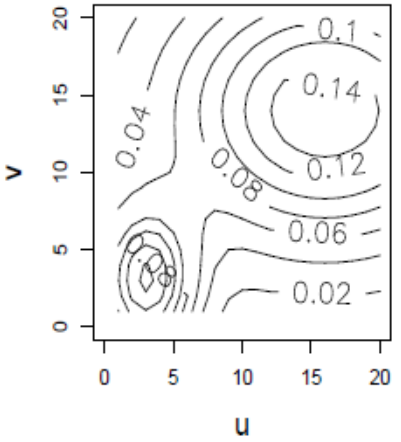


cluster pattern; systematic pattern

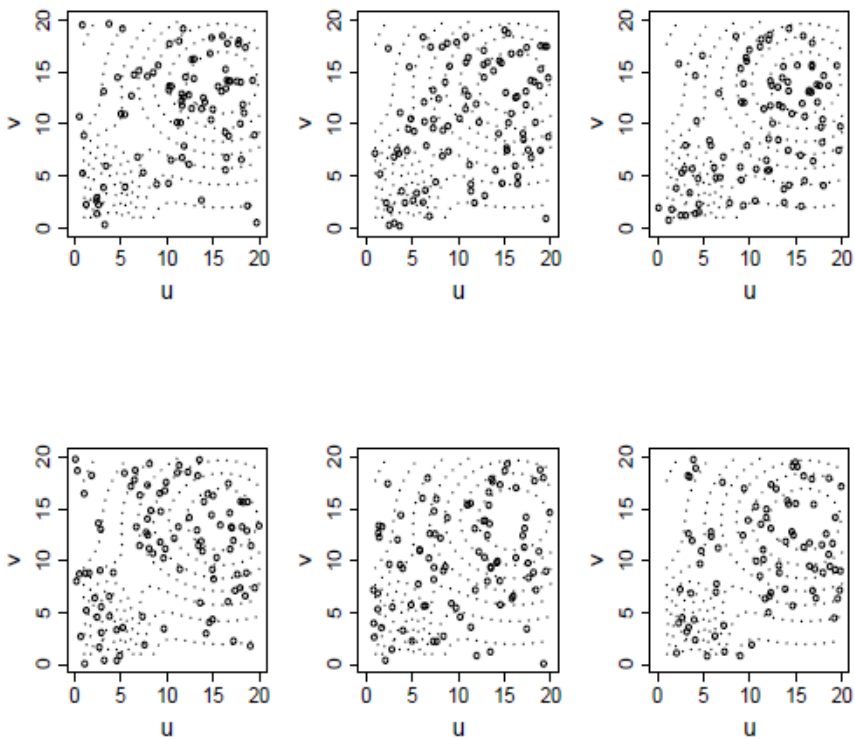


spatial heterogeneity





spatial heterogeneity



Examples

- In looking at ecological processes, interest in the pattern of occurrences of species, e.g., the pattern of trees in a forest, say junipers and pinions.
- In spatial epidemiology, we seek to find pattern in disease cases, perhaps different patterns for cases vs. controls; breast cancer cases: treatment option - mastectomy or radiation
- In *syndromic surveillance* we seek to identify disease outbreaks, looking for clustering of cases, over time.
- Evolution/growth of a city, i.e., urban development, pattern of development of single family homes or of commercial property over time.

The key players (my view)

- Adrian Baddeley - impressive theoretical contributions; recently, more applied effort - likelihood methods, exploratory tools, residual analysis, spatstat package
- Peter Diggle - ahead of his time; lovely early theory; broad spatial interests, always a strong practical bent, accessible (classic) books and useful website
- Jesper Møller - outstanding theoretician; rich classes of models and model fitting; simulation and fitting algorithms for Markov and Cox processes; a book
- Recent book of Illian, Pentinen, Stoyan, and Stoyan - a broad, richly exemplified, accessible volume
- Handbook of Spatial Statistics (Gelfand et al., 2010); Hierarchical Modeling and Analysis for Spatial Data, 2nd Edition (Banerjee et al., 2014)

The contribution

- At the heart is modeling and distribution theory for spatial point patterns
- Given model fitting, focus on inference within a Bayesian framework
- From an inferential perspective, spatial point pattern work is least developed and even more the case within the Bayesian framework
- Use simulation as the tool, enables full inference, with uncertainty
- Ideas for residual analysis, model adequacy, model comparison
- Lots of preliminaries

The basics

- Point patterns consider the randomness associated with the locations of the points.
- "No spatial pattern?" A *uniform* distribution of points? Complete spatial randomness (CSR)?
- For a bounded region D, denote the realization as $\mathbf{s}_i, i = 1, 2..., n$ with both n and the $\mathbf{s}_i$ random.
  - Are we seeing a finite realization of an infinite point pattern as a result of imposing D (edge effects and the shape of D might matter)?
  - Are we seeing a finite point pattern associated with a specified D (e.g., an island, a forest, a city)?
- Modeling depends upon setting. Second case better suited to application, more flexible modeling

cont.

- Need not have variables at locations, just the pattern of points provided by the locations.
- Crude features of the patterns. CSR is a place to start, hope to criticize. Why? In applications, it would not be operating.
- We seek to shed light on where there is departure from randomness and what its nature might be.
- Departure can result from environmental features, regression models to explain pattern we observe
- Instead, clustering or attraction, possibly inhibition or repulsion, perhaps regular or systematic behavior which we seek to explain.

Modeling

- We focus on point patterns over $D \subset R^2$
- We consider a bounded, connected subset D. We denote a random realization of a point pattern by $\mathbf{S}$ with elements $\mathbf{s}_1, ..., \mathbf{s}_n$.
- $\mathbf{S}$ is random and so are any features calculated from it.
- A probabilistic model for $\mathbf{S} \in D$ must place a distribution over all possible realizations in D.
- In practice, often easier to examine features/functionals of this distribution than to specify the distribution.
- Generative specification: (i) distribution over $\{0, 1, 2, ...\}$ to provide number of points then, (ii) distribution to *jointly* locate these points over D.

More explicitly:

- Two ingredients to specify a generative probabilistic model for $\mathbf{S}$
- Distribution for $N(D)$, the number of points in D, a distribution over the set $n \in \{0, 1, .., \infty\}$.
- A multivariate *location density* over D^n, for any n, say $f(\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n)$. Since points are unordered/unlabeled, f must be symmetric in its arguments.
- With $\partial\mathbf{s}$ denoting a small circular neighborhood around $\mathbf{s}$, $P(N(\partial\mathbf{s}_1) = 1, N(\partial\mathbf{s}_2) = 1, ..., N(\partial\mathbf{s}_n) = 1) \approx f(\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n)\Pi_i|\partial\mathbf{s}_i|$, with $|\partial\mathbf{s}|$ the area of $\partial\mathbf{s}$.
- We need to specify f consistently over all $\mathbf{S}$.
- Joint dist has marginal-conditional form $P(N(D) = n)n!f(\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n)$.

Stationarity

- A stationary point pattern model:
  $f(\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n) = f(\mathbf{s}_1 + \mathbf{h}, \mathbf{s}_2 + \mathbf{h}, \ldots, \mathbf{s}_n + \mathbf{h})$ for all n, $\mathbf{s}_i$, and $\mathbf{h}$.
- This condition would naturally be proposed over R^2 and applied, suitably, over D.
- Stationarity is a model property, not a model specification.
- Will return to below

Counting measure

- Analogous to $N(D)$, introduce count variables, $N(B)$, i.e., $N(B) = \sum_{\mathbf{s}_i \in \mathbf{S}} 1(\mathbf{s}_i \in B)$.
- $N(B)$ is computed by looking at the points in $\mathbf{S}$ individually, a first order property.
- Pairs of points, a second order property (below)
- Random counting measure over a σ-algebra through finite dimensional distributions, i.e., the joint distribution for a finite collection of count variables.
- A realization of a point pattern is equivalent to a realization of a counting measure (*void sets*).

Poisson process

- Recall Poisson process over a set D, intensity $\lambda(\mathbf{s})$.
  $N(B) \sim \text{Po}(\lambda(B))$ where $\lambda(B) = \int_B \lambda(\mathbf{s})d\mathbf{s}$.
- In addition, if B_1 and B_2 are disjoint, then $N(B_1)$ and $N(B_2)$ are independent
- The random Poisson measure induced by $\lambda(\mathbf{s})$:
  $\lim_{\partial\mathbf{s}\to 0} \frac{N(\partial\mathbf{s})}{|\partial\mathbf{s}|} = N(\mathbf{s})$ or equivalently, $N(B) = \int_B N(\mathbf{s})d\mathbf{s}$
- Independence of disjoint sets implies
  $f(\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n) = \Pi_i f(\mathbf{s}_i) = \Pi_i \lambda(\mathbf{s}_i)/\lambda(D)$ where
  $\lambda(D) = \int_D \lambda(\mathbf{s})d\mathbf{s}$.
- $P(N(\partial\mathbf{s}) = 1) \approx E(N(\partial\mathbf{s})) = \lambda(\partial\mathbf{s}) \approx \lambda(\mathbf{s})|\partial\mathbf{s}| \equiv \lambda(\mathbf{s})d\mathbf{s}$.

Moment measures

- First order properties, i.e., the first moment measure, $\{E(N(B)) : B \in \mathcal{B}\}$. Given $\lambda(\mathbf{s})$, we can compute $E(N(B)) = \int_B \lambda(\mathbf{s}) d\mathbf{s}$.
- However, given that the collection, $\{E(N(B)) : B \in \mathcal{B}\}$, is a measure, we can extract the *first-order* intensity: $\lambda(\mathbf{s}) = \lim_{|\partial \mathbf{s}| \to 0} \frac{E(N(\partial \mathbf{s}))}{|\partial \mathbf{s}|}$.
- If $f(\mathbf{s}_1, ..., \mathbf{s}_n) = \Pi_i f(\mathbf{s}_i)$, then $\lambda(\mathbf{s}) = f(\mathbf{s})\lambda(D)$.

Second order properties

- For second-order properties, consider $\gamma(B_1 \times B_2) \equiv E_{\mathbf{S}} \sum_{\mathbf{s},\mathbf{s}' \in \mathbf{S}} 1(\mathbf{s} \in B_1, \mathbf{s}' \in B_2)$. Define $\gamma(\mathbf{s}, \mathbf{s}')$, *second order intensity* through $\gamma(B_1 \times B_2) = \int_{B_1} \int_{B_2} \gamma(\mathbf{s}, \mathbf{s}') d\mathbf{s}' d\mathbf{s}$.
- So, if B_1, B_2 disjoint, $E_{\mathbf{S}}(N(B_1)N(B_2)) = \int_{B_1} \int_{B_2} \gamma(\mathbf{s}, \mathbf{s}') d\mathbf{s}' d\mathbf{s}$.
- Hence, wwith sufficiently small sets, $\gamma(\mathbf{s}, \mathbf{s}') = \lim_{|\partial \mathbf{s}| \to 0, |\partial \mathbf{s}'| \to 0} \frac{E(N(\partial \mathbf{s})N(\partial \mathbf{s}'))}{|\partial \mathbf{s}||\partial \mathbf{s}'|}$.
- The *pair correlation function*, $\gamma(\mathbf{s}, \mathbf{s}')/\lambda(\mathbf{s})\lambda(\mathbf{s}')$. When $\lambda(\mathbf{s}) = \lambda$ simplifies to $\gamma(\mathbf{s}, \mathbf{s}')/\lambda^2$ and, in fact, equals 1 under CSR. > 1 implies attraction, < 1 implies repulsion.
- Under stationarity, $\gamma(\mathbf{s}, \mathbf{s}') = \gamma(\mathbf{s} - \mathbf{s}')$. Isotropic means $\gamma(\mathbf{s}, \mathbf{s}') = \gamma(||\mathbf{s} - \mathbf{s}'||)$.

Papangelou conditional intensity

- Consider $\lambda(\mathbf{s}|\mathbf{S})$ for a given location $\mathbf{s}$ and a given realization $\mathbf{S}$?
- $\lambda(\partial\mathbf{s})|\mathbf{S}) \approx \lambda(\mathbf{s}|\mathbf{S})d\mathbf{s}$ is interpreted as the conditional probability that there is a point of the process in $\partial\mathbf{s}$ and the rest of the process coincides with $\mathbf{S}$.
- Roughly, $\lambda(\partial\mathbf{s}|\mathbf{S})$ is the probability that there is a point of $\mathbf{S}$ in $\partial\mathbf{s}$ and the rest of $\mathbf{S}$ lies outside of $\partial\mathbf{s}$.
- $\lambda(\mathbf{s}|\mathbf{S}) = \lambda(\mathbf{s}|\mathbf{S}/\mathbf{s}), \mathbf{s} \in \mathbf{S}; = \lambda(\mathbf{s}|\mathbf{S}), \mathbf{s}$ not $\in \mathbf{S}$
- $\lambda(\mathbf{s}|\mathbf{S})$ is random since $\mathbf{S}$ is and $E_{\mathbf{S}}(\lambda(\mathbf{s}|\mathbf{S})) = \lambda(\mathbf{s})$.
- $\lambda(\mathbf{s}|\mathbf{S}) = \frac{f(\mathbf{s},\mathbf{S})}{f(\mathbf{S})}$ where $f(\mathbf{S})$ is the *density* of the spatial point process (with respect to an HPP(1))
- $f(\mathbf{S})$ is not fixed dimension; usually specified up to normalizing constant which cancels from ratio for $\lambda(\mathbf{s}|\mathbf{S})$
- For conditionally independent locations $\lambda(\mathbf{s}|\mathbf{S}) = \lambda(\mathbf{s})$.

Homogeneous Poisson Process (HPP)

- CSR: $\lambda(\mathbf{s}) = \lambda$ (HPP), $\lambda(B) = \lambda|B|$
- Stationarity implies that $\lambda(\mathbf{s}) = \lambda$ for all $\mathbf{s}$ and thus, $\lambda(B) = \lambda|B|$ for all $B \subseteq D$.
- $f(\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n) = 1/|D|^n$.
- The HPP is only one stationary process specification. It specifies a constant intensity with conditionally independent locations.
- More general models include interactions between points, e.g., the stationary Gibbs processes.
- Can be a *null model* for certain types of data, e.g., physical processes in a homogeneous environment, for example, interacting particle models.

Exploratory tools, the G function

- Again, complete spatial randomness (CSR) $\equiv$ HPP(λ). Want to criticize CSR
- Distance based approaches; G, F, and K functions
- $G(d)$, the "nearest neighbor" distribution, i.e., the c.d.f. of the nearest neighbor distance, event to event.
- $G(d) = Pr(\texttt{nearest event} \leq d)$.
- $F(d)$ is the "empty space" distribution, i.e., for an arbitrary location, the c.d.f. of the nearest neighbor distance, point to event
- $F(d) = Pr(\texttt{nearest event} \leq d)$.
- Under CSR, $G(d) = F(d) = 1 - \exp(-\lambda\pi d^2)$.
- G places a lot of mass on small distances. We expect to see some clustering under CSR.

cont.

- Empirical c.d.f., $\hat{G}(d)$, arises from the n nearest neighbor distances (for $\mathbf{s}_1$, for $\mathbf{s}_2$, etc.). Denote this set by $\{d_1, d_2, ..., d_n\}$.
- With bounded D, we will need an edge correction, e.g., if, for $\mathbf{s}_i$, $d > b_i$, where b_i is the *distance* from $\mathbf{s}_i$ to edge of D.
- $\hat{G}(d) = \frac{\sum_i I(d_i \leq d < b_i)}{\sum_i I(d < b_i)}$
- So, if $d > b_i$, then the event $\{d_i < d\}$ is not observed.
- Comparison of $\hat{G}$ with G under CSR is usually through a theoretical Q-Q plot.
- Shorter tails suggest clustering/attraction; longer tails suggest inhibition/repulsion.

The K function

- The K function considers the *expected number* of points within distance d of an arbitrary point.
- Under stationarity, this expectation is the same for any point.
- $K(d) = (\lambda)^{-1}\mathrm{E}(\#$ of points within d of an arbitrary point) The scaling by $1/\lambda$, along with stationarity, scales $K(d)$ to be free of λ.
- For example, under CSR, $K(d) = \lambda\pi d^2/\lambda = \pi d^2$, i.e., the area of a circle of radius d.

cont.

- $\hat{K}(d) = (\hat{\lambda})^{-1} \sum_i \sum_{j \neq i} 1(d_{ij} \equiv ||\mathbf{s}_i - \mathbf{s}_j|| \leq d)/n$
  $= (n\hat{\lambda})^{-1} \sum_i r_i$ where $\hat{\lambda} = n/|D|$ and r_i is number of $\mathbf{s}_j$ within d of $\mathbf{s}_i$
- Edge correction, w_{ij} for $\mathbf{s}_i$ too near boundary of D.
- w_{ij} is the conditional probability that an event is in D given that it is exactly distance d_{ij} from $\mathbf{s}_i$
- Approximated as the proportion of the circumference of a circle centered at $\mathbf{s}_i$ with radius $\|\mathbf{s}_i - \mathbf{s}_j\|$ that lies within D.
- In fact, for a stationary process, can define $K(d) = \int_{||\mathbf{u}|| \leq d} g(\mathbf{u}) d\mathbf{u}$, with g the pair correlation function
- As a result, the second moment measure $\gamma(d) = \frac{\lambda^2 K'(d)}{2\pi d}$.
- Suggests the possibility of $\hat{\gamma}(d)$.

Finite point pattern models (restriction to D)

- Nonhomogeneous Poisson process (NHPP) - $\lambda(\mathbf{s})$, conditionally independent locations with location density, $f(\mathbf{s}) = \lambda(\mathbf{s})/\lambda(D)$
- Scaling form: $\lambda(\mathbf{s};\theta) = \lambda f(\mathbf{s};\theta)$, f a bivariate density function truncated to D.
- Sufficiently rich choices for f? Mixture models, e.g., $f(\mathbf{s}) = \sum_{k=1}^{K} p_k f_k(\mathbf{s})$. But fitting challenges.
- Nonnegativity challenges for trend surfaces.
- Most common: $\log \lambda(\mathbf{s}) = \mathbf{X}^T(\mathbf{s})\gamma$; spatial covariates drive the point pattern
- Need to calculate $\int_D e^{\mathbf{X}^T(\mathbf{s})\gamma} d\mathbf{s}$ to obtain the likelihood. $\mathbf{X}(\mathbf{s})$? Discrete approximation, ecological fallacy, tiled surface. But, without finer covariate resolution, can't do better.

Log Gaussian Cox process (LGCP)

- A particular Cox process; we write $\lambda(\mathbf{s}) = g(\mathbf{X}(\mathbf{s})^T\gamma)\lambda_0(\mathbf{s})$
- Require $g(\cdot) \geq 0$ and think of $\lambda_0(\mathbf{s})$ as the *error* process, a realization of a positive stochastic process
- Natural center is mean 1
- Conditional on $\{\lambda_0(\mathbf{s}), \mathbf{s} \in D\}$ (and γ), we have a NHPP
- Log Gaussian Cox process (LGCP) iff $\lambda(\mathbf{s}) = \exp(Z(\mathbf{s}))$, $Z(\mathbf{s})$ from a spatial Gaussian process with mean say $\mathbf{X}^T(\mathbf{s})\gamma$ and covariance function $\sigma^2\rho(\cdot)$
- Two stage process: $[\mathbf{S}|\lambda(\mathbf{s})][\lambda(\mathbf{s})]$

The likelihood

- For an NHPP or a LGCP, what is the likelihood?
- As a function of $\lambda(\mathbf{s})$, $L(\{\lambda(\mathbf{s}), \mathbf{s} \in D\}; \mathbf{S}_{obs}) = e^{-\lambda(D)} \Pi_i \lambda(\mathbf{s}_i)$
- A function of an entire surface. For NHPP, a parametric function, for LGCP, a process realization
- So, $\lambda(D) = \int_D \lambda(\mathbf{s}) d\mathbf{s}$ is a regular or a stochastic integral.
- Discrete approximation for $\int_D e^{\mathbf{X}^T(\mathbf{s})\boldsymbol{\gamma} + Z(\mathbf{s})} d\mathbf{s}$ using *representative* points
- Challenges: For NHPP, ecological fallacy, for LGCP, convergence

More general Cox processes

- Neyman Scott process, Matérn process, Thomas process; shot noise, e.g., Poisson Gamma process
- Suppose we generate *parent* events from a NHPP with $\lambda(\mathbf{s})$ say K, and their locations say $\boldsymbol{\mu}_k, k = 1, 2, ..., K$.
- Next, suppose each parent produces a random (but i.i.d.) number of offspring, N_k, where the N_k are i.i.d. according say, $g = \text{Po}(\delta)$.
- Next, locate the offspring relative to the parent.
- For kth parent, locate offspring according to i.i.d. draws from a bivariate density, $f(\mathbf{s}; \boldsymbol{\mu}_k)$.
- Only the offspring are retained to yield the point pattern.

cont.

- If bivariate density is $N(\boldsymbol{\mu}_k, \sigma^2 I)$, a (modified) Thomas process
- Compound Poisson process: degenerate offspring density at $\boldsymbol{\mu}_k$. Count at $\boldsymbol{\mu}_k$ is a 'mark' at that location.
- The Matérn process: offspring at $\boldsymbol{\mu}_k$ uniform in a circle of radius R (a parameter) around $\boldsymbol{\mu}_k$
- More generally, combine the steps of generating the number of children and their locations. That is, generate N i.i.d $\sim g_K$ and generate $\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_N$ i.i.d $\sim \sum_{k=1}^{K} \frac{1}{K} f(\mathbf{s}; \boldsymbol{\mu}_k, \Sigma)$
- For example, with above, say, $g_K = Po(K\lambda)$.

Shot noise processes

- A Cox process that is also conditionally a NHPP; an alternative to a LGCP.
- Again, $\lambda(\mathbf{s}) = e^{X^T(\mathbf{s})\beta}\lambda_0(\mathbf{s})$
- Now, $\lambda_0(\mathbf{s})$ is a mean 1 shot noise process so that $\lambda(\mathbf{s})$ is *centered* around the deterministic component.
- Usual form: $\lambda_0(\mathbf{s}) = \sum_{\mathbf{s}_i \in \mathbf{S}} f(\mathbf{s} - \mathbf{s}_i)m(\mathbf{s}_i)$, with $\mathbf{S}$ drawn from a HPP(λ) and $m(\mathbf{s}_i)$ a constant, m
- f is a density over D and $m(\mathbf{s}_i) \geq 0$.
- $m(\mathbf{s}_i)$ perhaps m (or from a regression on say $X(\mathbf{s})$ over D or a process realization over D)
- $m(\mathbf{s}_i)$ denotes contribution to $\lambda_0(\mathbf{s})$ from $\mathbf{s}_i$ and $\lambda_0(\mathbf{s})$ accumulates the "shots" arising from $\mathbf{S}$

Poisson-Gamma process

- Poisson Gamma process is an example of a shot noise process. Allows both over and under-dispersion relative to an HPP.
- General gamma process provides a random positive spatial surface, i.e., $\Gamma(d\mathbf{u}) \sim Ga(\alpha(d\mathbf{u}), \beta^{-1})$ (i.e., $\int_A \Gamma(d\mathbf{u}) = \Gamma(A)$)
- We kernel mix to obtain the random intensity $\lambda(\partial\mathbf{s}) \approx \lambda(\mathbf{s}|\partial\mathbf{s}| = \int_D f(\mathbf{s}-\mathbf{u})\Gamma(d\mathbf{u})|\partial\mathbf{s}|$
- We draw a realization of an HPP over D to obtain $\mathbf{S}^* = \{s_j^*, j = 1, 2, ..., m\}$
- We simplify the intensity by discretizing D to $\mathbf{S}^*$ yielding $\lambda(\mathbf{s}) = \sum_{s_j^* \in \mathbf{S}^*} f(\mathbf{s}-\mathbf{s}_j^*)w(\mathbf{s}_j^*)$, $w(\mathbf{s}_j^*)$ a Gamma variable.

Markov, Gibbs processes

- Markov processes, for us Gibbs processes. Examples here: Strauss process, hardcore process
- A finite Gibbs process if location density is $f(\mathbf{S}) = \exp(-Q(\mathbf{S}))$ with regard to an HPP with unit intensity
- $Q(\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n) =$ $c_0 + \sum_{i=1}^{n} h_1(\mathbf{s}_i) + \sum_{i \neq j} h_2(\mathbf{s}_i, \mathbf{s}_j) + ... + h_n(\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n)$
- h's have parameters, c_0 is a *normalizing* constant over $\times D^n$, a function of the parameters in the h's
- c_0 is almost always intractable, making $E(N(D))$ intractable

cont.

- The h's are potentials of order 1, 2,...n, respectively, each symmetric in its arguments.
- With potentials only of order 1, NHPP with $\lambda(\mathbf{s}) = e^{-h_1(\mathbf{s})}$.
- Higher order potentials capture/control interaction.
- Pairwise interactions: only include h_1 and h_2. To guarantee integrability, we must take $h_2 \geq 0$.
- This implies we can only capture inhibition.
- If $h_1(\mathbf{s})$ is constant, homogeneous Gibbs process.

cont.

- Most common form for h_2 is $\phi(||\mathbf{s} - \mathbf{s}'||)$, e.g., $||\mathbf{s} - \mathbf{s}'||^2/\tau^2$
- Papangelou conditional intensity has a simple form in this case, $\lambda(\mathbf{s}|\mathbf{S}) = \exp(-(h_1(\mathbf{s}) + \sum_{i=1}^{n} \phi(\|\mathbf{s} - \mathbf{s}_i\|)))$.
- Unknown normalizing constant cancels from the conditional intensity
- Examples through $\phi(d)$, d is an interpoint distance.
- Strauss process sets $\phi(d) = \beta, d \leq d_0, = 0, d > d_0$. $\beta > 0$, $e^{-\phi(d)} \leq 1$ for all d, interaction term downweights patterns with more points close to each other.
- Hardcore process sets $\phi(d) = \infty, d \leq d_0, = 0, d > d_0$. Now, density is 0 for all $\mathbf{S}$ with a pair of points less than d_0 apart.

Fitting spatial point process models

- HPP: MLE is straightforward. Closed form likelihood.
- Minimum contrast method (Diggle), essentially a method of moments idea, e.g., $\int(\hat{K}(d) - K(d))^2 dd$, or with pair correlation, g, or introduce powers
- Likelihood-based methods more attractive, common
- NHPP - Berman-Turner device: connects NHPP log likelihood to a weighted Poisson regression log likelihood using quadrature to do a numerical integration
- LGCP - Numerical integration; "representative points"
- Markov and Gibbs processes - pseudo-maximum likelihood using the Papangelou conditional intensity; i.e., pseudo-likelihood through $\Pi_i \lambda(\mathbf{s}_i|\mathbf{S}/\mathbf{s}_i)$.

Bayesian model fitting of spatial point processes

- HPP - Bayes is straightforward. With Gamma prior, posterior for λ is again a Gamma
- NHPP - Berman-Turner provides the likelihood. Adding a prior enables routine MCMC, again integral approximation
- LGCP - Elliptical slice sampling from Murray and Adams (2010), MALA (Møller et al.(1998), Hamiltonian MC (Girolami and Calderhead, 2010); recently INLA (Simpson et al. 2011)
- Cox, Shot noise processes (Møller and Waagepetersen, 2004, 2007)
- Markov and Gibbs processes - Auxiliary variables in Metropolis Hastings (Berthelsen and Møller papers)

Bayesian inference in the literature

- Aspects of spatial point process modelling and Bayesian inference, J. Møller (http://conferences.inf.ed.ac.uk/bayeslectures/moeller.pdf)
- J. Møller and R.P. Waagepetersen (2007). Modern statistics for spatial point processes (with discussion). Scandinavian Journal of Statistics, 34, 643-711
- K.K. Berthelsen and J. Møller (2008). Non-parametric Bayesian inference for inhomogeneous Markov point processes. Australian and New Zealand Journal of Statistics, 50, 627-649.
- J.B. Illian, J. Møller and R.P. Waagepetersen (2009). Hierarchical spatial point process analysis for a plant community with high biodiversity. EES, 16, 389-405.
- P. Guttorp and T.L. Thorarinsdottir (2012) Bayesian inference for non-Markovian point processes (in Advances and Challenges in Space-time Modelling...).

A general inference approach

- Model - generic form $[\mathbf{S}|\boldsymbol{\theta}][\boldsymbol{\theta}]$
- Observe $\mathbf{S}_{obs}$
- Fit, obtain posterior samples $\boldsymbol{\theta}_b^*$ from $[\boldsymbol{\theta}|\mathbf{S}_{obs}]$
- Sample - using composition, create samples $\mathbf{S}_b^*$ from $[\mathbf{S}_{new}|\mathbf{S}_{obs}]$ by drawing $\mathbf{S}_b^*$ from $[\mathbf{S}|\boldsymbol{\theta}_b^*]$
- Infer - create posterior samples of any function say h of $\mathbf{S}$ as $\{h(\mathbf{S}_b^*), b = 1, 2, ..., B\}$ from $[h(\mathbf{S})|\mathbf{S}_{obs}]$
- **So, if we can fit and if we can sample, arbitrary inference**
- Also, if we can sample, prior-posterior comparison

Generating samples

- For HPP, trivial
- For NHPP, usual thinning of an HPP(λ_{max})
- For LGCP, two stage - tiled realization of the GP, followed by NHPP thinning given the GP surface
- For cluster process, usually directly generative
- For Gibbs process, perfect simulation (CFTP); birth-death MCMC algorithm
- General thinning for generation - p-thinning, $p(\mathbf{s})$ thinning
- Other mechanisms: displacement, censoring, superposition

The broad challenge

- A primary feature we are trying to infer about is a (random) surface, i.e., an intensity. But we never observe a point on this surface.
- Analogue with density estimation. In fact, we have empirical kernel intensity estimates
- But also, number of points is random
- For example, consider an HPP setting. Any observed point pattern will give an empirical intensity estimate which is not close to flat
- In fact, null hypotheses, $H_o : \lambda(\mathbf{s}) = \lambda$ seems *silly*
- Instead, compare inference under HPP model with that from other models.
- In general, hard to criticize models, hard to choose between models. Not much literature, no Bayesian work

Posterior study of features

- For example, posterior distributions: $[N(A)|\mathbf{S}_{obs}]$, $[N(A), N(B)|\mathbf{S}_{obs}]$, $[N(A)|N(B), \mathbf{S}_{obs}]$, $[N(A)/N(D)|\mathbf{S}_{obs}]$; posterior for G and K functions; prior comparison.
- Posterior distribution of *realized* residuals, e.g., in NHPP, $[N(A)_{obs} - \int_A \lambda(\mathbf{s})d\mathbf{s}|\mathbf{S}_{obs}]$.
- Posterior distribution of *predicted* residuals, $[N(A)_{obs} - N(A)|\mathbf{S}_{obs}]$.
- Predictive residuals better for model checking
- As in linear regression: $Y_{i,obs} - X_i^T\hat{\beta}$ vs. $Y_{i,obs} - \hat{Y}_i$
- Important point: $\lambda(\mathbf{s})$ informs about observed data points, also about unobserved points.
- In a point pattern, more information than just the locations of the observed points. *Absence* at other locations is informative (Baddeley et al., 2005).

More explicitly

- Under the model, interest in $b(\boldsymbol{\theta})$ using $[b(\boldsymbol{\theta})|\mathbf{S}_{obs}]$.
- With posterior samples $\{\boldsymbol{\theta}_l^*\}$, we obtain $\{b(\boldsymbol{\theta}_l^*)\}$
- If interest is in $[h(\mathbf{S})|\mathbf{S}_{obs}]$, then for each $\boldsymbol{\theta}_l^*$, we generate $\mathbf{S}_l^*$ obtaining $\{\mathbf{S}_l^*\}$ and thus $\{h(\mathbf{S}_l^*)\}$
- Back to $b(\boldsymbol{\theta})$, often not available explicitly. So, find $h(\mathbf{S})$ such that $E(h(\mathbf{S})|\boldsymbol{\theta}) = b(\boldsymbol{\theta})$.
- Then, to obtain $b(\boldsymbol{\theta}_l^*)$, for each $\boldsymbol{\theta}_l^*$, generate $\mathbf{S}_{lb}^*$'s obtaining the set $\{\mathbf{S}_{lb}^*\}$ so a Monte Carlo integration for $b(\boldsymbol{\theta}_l^*)$ is $\frac{1}{B}\sum_b h(\mathbf{S}_{lb}^*)$.
- Most generally, $[f(\mathbf{S}, \boldsymbol{\theta})|\mathbf{S}_{obs}]$ with f available explicitly, can use $\{\boldsymbol{\theta}_l^*, \mathbf{S}_l^*\}$.

cont.

- Examples of $b(\boldsymbol{\theta})$'s include: $\lambda(\mathbf{s};\boldsymbol{\theta}), \gamma(d;\boldsymbol{\theta}), \lambda(A;\boldsymbol{\theta}), E(N(A)N(B)|\boldsymbol{\theta}), g(d;\boldsymbol{\theta}), G(d;\boldsymbol{\theta})$.
- Examples of $h(\mathbf{S})$'s include: $N(A), (N(A), N(B)), N(A)/N(D)$, predictive residuals $([N_{obs}(A) - N(A)|\mathbf{S}_{obs}])$ and conditional events with distribution $[N(A)|N(B) = m; \mathbf{S}_{obs}]$.
- Examples of $f(\mathbf{S}, \boldsymbol{\theta})$ include: realized residuals $([N(A) - \lambda(A;\boldsymbol{\theta})|\mathbf{S}_{obs}])$, $K(d:\boldsymbol{\theta}), K_{inhom}(d;\boldsymbol{\theta})$ (here $f(\mathbf{S};\boldsymbol{\theta})$ takes forms like $\sum_{\mathbf{s}_i \in \mathbf{S} \cap D} \frac{1}{\lambda(D)} g(\mathbf{s}_i; \mathbf{S} \backslash \mathbf{s}_i)$ or $\sum_i \sum_{j \neq i} \frac{g(\mathbf{s}_i, \mathbf{s}_j)}{\lambda(\mathbf{s}_i)\lambda(\mathbf{s}_j)}$), $\hat{\lambda}(\mathbf{s}) - \lambda(\mathbf{s};\boldsymbol{\theta})$ where $\hat{\lambda}(\mathbf{s})$ is a kernel intensity estimate of $\lambda(\mathbf{s})$.
- So, full inference is clear.

Going further

- Campbell's Theorem (a feature with one argument): $E_{\mathbf{S}}(\sum_{\mathbf{s}_i \in \mathbf{S}} h(\mathbf{s}_i)) = \int h(\mathbf{s})\lambda(\mathbf{s})d\mathbf{s}$
- Why? Let $h(\mathbf{s}_i) = 1(\mathbf{s}_i \in A)$, then left side is $E_{\mathbf{S}}N(A)$ and right side is $\int_A \lambda(\mathbf{s})d\mathbf{s} = \lambda(A)$
- If $h(\mathbf{S}) = \sum_{\mathbf{s}_i \in \mathbf{S}} h(\mathbf{s}_i)$, $E_{\mathbf{S}|\boldsymbol{\theta}} h(\mathbf{S}) = \int h(\mathbf{s})\lambda(\mathbf{s})d\mathbf{s} \equiv b_h(\boldsymbol{\theta})$
- So, with posterior samples, $\{\boldsymbol{\theta}_l^*\}$, $b_h(\boldsymbol{\theta}_l^*)$ are posterior samples from $[b_h(\boldsymbol{\theta})|\mathbf{S}_{obs}]$ and $\frac{1}{L}\sum_l b_h(\boldsymbol{\theta}_l^*)$ is a MC integration for $E(b_h(\boldsymbol{\theta})|\mathbf{S}_{obs})$
- If we can't calculate $b_h(\boldsymbol{\theta})$, then, with $\mathbf{S}_l^* \sim [\mathbf{S}|\mathbf{S}_{obs}]$, $\frac{1}{L}\sum_l h(\mathbf{S}_l^*)$ is a MC integration for $E(h(\mathbf{S})|\mathbf{S}_{obs})$
- And, $E_{\mathbf{S}|\mathbf{s}_{obs}}(h(\mathbf{S})) = E_{\boldsymbol{\theta}|\mathbf{s}_{obs}} E_{\mathbf{S}|\boldsymbol{\theta}}(h(\mathbf{S})) = E_{\boldsymbol{\theta}|\mathbf{s}_{obs}}(b_h(\boldsymbol{\theta})$
- So, $\frac{1}{L}\sum_l h(\mathbf{S}_l^*)$ is a MC integration for $E(b_h(\boldsymbol{\theta})|\mathbf{S}_{obs})$

Going further, cont

- If we want posterior samples of $b_h(\boldsymbol{\theta})$, they are $b_h(\boldsymbol{\theta}_l^*)$. If we can not calculate $b_h(\boldsymbol{\theta})$, we need a MC integration.
- Now, we need, for each $\boldsymbol{\theta}_l^*$, $\{\mathbf{S}_{lb}^*, b = 1, 2, ..., B\} \sim [\mathbf{S}|\boldsymbol{\theta}_l^*]$ so $\frac{1}{B} h(\mathbf{S}_{lb}^*)$ is a MC integration for $E_{\mathbf{S}|\boldsymbol{\theta}_l^*}(h(\mathbf{S}) = b_h(\boldsymbol{\theta}_l^*)$
- Campbell's Theorem (a feature with two arguments): $E_{\mathbf{S}}\left(\sum_{\mathbf{s}_i,\mathbf{s}_j \in \mathbf{S}, i \neq j} h(\mathbf{s}_i, \mathbf{s}_j)\right) = \int\int h(\mathbf{s}, \mathbf{s}')\gamma(\mathbf{s}, \mathbf{s}')d\mathbf{s}d\mathbf{s}'$ (e.g., $h(\mathbf{s}, \mathbf{s}') = 1(\mathbf{s} \in A, \mathbf{s}' \in B)$ yields $E_{\mathbf{S}}(N(A)N(B)))$
- Existence of expectations: Countable point pattern vs. restriction to $D \Rightarrow$ a finite point pattern.

Parametric-nonparametric

- Explicitly, $E[N(A)|\mathbf{S}_{\text{obs}}] \approx \frac{1}{L}\sum_{l=1}^{L}\sum_{\mathbf{s}_{li}^*\in\mathbf{S}_l^*}\mathbf{1}(\mathbf{s}_{li}^*\in A)$
- Could create model-based Bayesian intensity estimates. Taking $A = \partial\mathbf{s}$ yields Bayes estimate for $\lambda(\partial\mathbf{s}) \approx \lambda(\mathbf{s})|\partial\mathbf{s}|$, hence for $\lambda(\mathbf{s})$.
- With a fine grid of $\mathbf{s}$, an estimated intensity surface. Size of $\partial\mathbf{s} \Leftrightarrow$ bandwidth for a kernel intensity estimate.
- Usual *kernel* smoothing yields kernel intensity estimate, $\lambda_\tau(\mathbf{s}0 = \frac{1}{\tau^2}\sum_{\mathbf{s}_i\in\mathbf{S}} h(||\mathbf{s}-\mathbf{s}_i||/\tau)$
- If we can write λ as a parametric function, $\lambda(\mathbf{s};\boldsymbol{\theta})$ (say for an NHPP but not for a LGCP), posterior samples of the $\boldsymbol{\theta}$ yield an estimate of $\lambda(\mathbf{s};\boldsymbol{\theta})$.
- Rao-Blackwellized vs. non-Rao-Blackwellized estimation

A bit deeper

- Features which depend upon entire $\mathbf{S}$, e.g., $h(\mathbf{s}_i; \mathbf{S}/\mathbf{s}_i)$
- Need Georgii - Nguyen - Zessin (GNZ) result: $E_{\mathbf{S}}(\sum_{\mathbf{s}_i \in \mathbf{S}} h(\mathbf{s}_i; \mathbf{S}/\mathbf{s}_i)) = E_{\mathbf{S}} \int \lambda(\mathbf{s}|\mathbf{S}) h(\mathbf{s}; \mathbf{S}) d\mathbf{s}$
- Now, we have Papangelou conditional intensity
- Will expectation exist? Restrict to $\mathbf{s}_i \in \mathbf{S} \cap D$ with $\int_D$. Can bring expectation under integral
- Examples: $h(\mathbf{u}; \mathbf{S}/\mathbf{u}) = 1(\mathbf{u} \in B)$ yields $E_{\mathbf{S}} N(\mathbf{S} \cap B) = \int_B E_{\mathbf{S}} \lambda(\mathbf{u}|\mathbf{S}) d\mathbf{u}$
- Suggests $N(\mathbf{S} \cap B) - \int_B \lambda(\mathbf{s}|\mathbf{S}) d\mathbf{s}$, realized *innovation* residuals, which have mean 0 (Baddeley et al., 2005)
- $h(\mathbf{u}; \mathbf{S}/\mathbf{u}) = 1(\mathbf{u} \in B)/\lambda(\mathbf{u}|\mathbf{S})$ yields $E_{\mathbf{S}}(\sum_{\mathbf{s}_i \in \mathbf{S}} 1(\mathbf{s}_i \in B)/\lambda(\mathbf{s}_i|\mathbf{S}/\mathbf{s}_i) = |B|$ (Stoyan and Grabarnik, 1991; "inverse" residuals, cute but ...)
- Other *scaled* residuals

cont.

- So, now if $h(\mathbf{S}) = \sum_{\mathbf{s}_i \in \mathbf{S} \cap D} h(\mathbf{s}_i; (\mathbf{S} \cap D \backslash \mathbf{s}_i)$, $E_{\mathbf{S} \cap D | \boldsymbol{\theta}}(h(\mathbf{S})) = E_{\mathbf{S} \cap D | \boldsymbol{\theta}} \int_D h(\mathbf{s}; (\mathbf{S} \cap D \backslash \mathbf{s}) \lambda(\mathbf{s}|\mathbf{S}) d\mathbf{s} \equiv b_h(\boldsymbol{\theta})$
- Instead, we will work with $\bar{h}(\mathbf{S}) \equiv h(\mathbf{S})/N(\mathbf{S} \cap D)$
- If $N(\mathbf{S} \cap D) = 0$, then $h(\mathbf{S}) = 0$ and we define $\frac{0}{0} = 1$
- So, consider $E_{\mathbf{S} \cap D | \boldsymbol{\theta}}(\bar{h}(\mathbf{S})) \equiv b_{\bar{h}}(\boldsymbol{\theta})$
- We need a different version of the GNZ result

An iterated expectation version

- We can view $\mathbf{S}$ over R^2 which induces $\mathbf{S} \cap D$ over D with $N(\mathbf{S} \cap D)$. Alternatively, suppose, given D, first generate $N(\mathbf{S} \cap D) = n$, then locate $\mathbf{S}$ over D given $N(\mathbf{S} \cap D) = n$, assuming the $\mathbf{s}_i$ are exchangeable
- A *generative* view (e.g., a NHPP or a cluster process) vs. a *modeling* view (e.g., a Gibbs process)
- Either way, there is a joint distribution $[\mathbf{S} \cap D, N(\mathbf{S} \cap D)]$, hence $[\mathbf{S} \cap D | N(\mathbf{S} \cap D)][N(\mathbf{S} \cap D)]$. So, we can calculate the expectation iteratively
- $E_{\mathbf{S} \cap D}(\sum_{\mathbf{s}_i \in \mathbf{S} \cap D} h(\mathbf{s}_i; (\mathbf{S} \cap D)/\mathbf{s}_i)) =$
  $E_{N(\mathbf{S} \cap D)} E_{\mathbf{S} \cap D | N(\mathbf{S} \cap D)} \sum_{\mathbf{s}_i \in \mathbf{S} \cap D} h(\mathbf{s}_i; (\mathbf{S} \cap D)/\mathbf{s}_i) =$
  $E_{N(\mathbf{S} \cap D)}(N(\mathbf{s} \cap D) E_{\mathbf{S} \cap D | N(\mathbf{s} \cap D}(h(\mathbf{s}, (\mathbf{S} \cap D)/\mathbf{s}))$
- And, $E_{\mathbf{S} \cap D}(\sum_{\mathbf{s}_i \in \mathbf{S} \cap D} h(\mathbf{s}_i; (\mathbf{S} \cap D)/\mathbf{s}_i))/N(\mathbf{S} \cap D) =$
  $E_{\mathbf{S} \cap D} h(\mathbf{s}; (\mathbf{S} \cap D)/\mathbf{s})$ (defining $0/0 = 1$)

so

- $E_{\mathbf{S}\cap D|\boldsymbol{\theta}}\bar{h}(\mathbf{S}) = b_{\bar{h}}(\boldsymbol{\theta})$ where
  $b_{\bar{h}}(\boldsymbol{\theta}) = E_{\mathbf{S}\cap D|\boldsymbol{\theta}}\mathbf{S}\cap D h(\mathbf{s};(\mathbf{S}\cap D)/\mathbf{s})$
- A usual Bayes estimate for $b_{\bar{h}}(\boldsymbol{\theta})$ is $E(b_{\bar{h}}(\boldsymbol{\theta})|\mathbf{S}_{obs})$
- With posterior samples, $\{\boldsymbol{\theta}_l^*\}$, a Monte Carlo integration for the posterior mean is $\frac{1}{L}\sum_l b_h(\boldsymbol{\theta}_l^*)$
- Typically, we can not calculate $b_{\bar{h}}(\boldsymbol{\theta})$ explicitly
- However, $E_{\mathbf{S}\cap D|\mathbf{S}_{obs}}\bar{h}(\mathbf{S}) = E_{\boldsymbol{\theta}|\mathbf{S}_{obs}}E_{\mathbf{S}\cap D|\boldsymbol{\theta}}\bar{h}(\mathbf{S}) = E_{\boldsymbol{\theta}|\mathbf{S}_{obs}}b_{\bar{h}}(\boldsymbol{\theta})$
- So, a direct MC integration becomes $\frac{1}{L}\bar{h}(\mathbf{S}_l^*)$

Two examples

- We now examine two features under a model with stationarity: G and K function
- Again, view process over all of R^2, i.e., an infinite point pattern which becomes finite under restriction to D.
- Suppose $\mathbf{s} \in \mathbf{S}$, $\partial_d \mathbf{s}$ is a circle of radius d centered at $\mathbf{s}$, and $N(\mathbf{s}, d; \mathbf{S})$ counts the number of points in $\partial_d \mathbf{s}$ from $\mathbf{S}$, excluding $\mathbf{s}$.
- Under stationarity, $\mathbf{S} \sim \mathbf{S} - \mathbf{s}$ so $N(\mathbf{s}, d; \mathbf{S}) \sim N(\mathbf{0}, d; \mathbf{S} - \mathbf{s})$, where $\mathbf{S} - \mathbf{s}$ is the translation of $\mathbf{S}$ by $\mathbf{s}$
- Every point in $\mathbf{S}$ is a *typical* point, i.e., equivalent to $\mathbf{0}$ under translation.

Back to G function

- Recall, $N_D(\mathbf{s}_i, d, \mathbf{S}) \equiv N(\mathbf{s}_i, d, \mathbf{S} \cap D)$; we only observe $N_D(\mathbf{s}_i, d, \mathbf{S})$
- Recall, $G(d) = Pr[N(\mathbf{s}, d, \mathbf{S}) > 0]$; consider

$$\bar{h}_{G,d}(\mathbf{S}) = \sum_{\mathbf{s}_i \in \mathbf{S} \cap D} \frac{1(N_D(\mathbf{s}_i, d, \mathbf{S}) > 0)}{N(\mathbf{S} \cap D)}$$

  which has expected value $Pr[N_D(\mathbf{s}, d, \mathbf{S}) > 0]$
- So, we are estimating $Pr[N_D(\mathbf{s}, d, \mathbf{S}) > 0]$; we want $Pr[N(\mathbf{s}, d, \mathbf{S}) > 0]$
- Of course $N_D(\mathbf{s}, d, \mathbf{S}) \leq N(\mathbf{s}, d, \mathbf{S})$ for any $\mathbf{s}$ and any $\mathbf{S}$, so $G(d) = Pr[N(\mathbf{s}, d, \mathbf{S}) > 0] \geq Pr[N_D(\mathbf{s}, d, \mathbf{S}) > 0]$.
- We need edge correction
- Bayesian edge correction available, details omitted

Back to the K function

- For a model with constant first order intensity λ, $E_{\mathbf{S}\cap D}\sum_{\mathbf{s}_i\in\mathbf{S}\cap D}\frac{N_D(\mathbf{s}_i,d,\mathcal{S}\backslash\mathbf{s}_i)}{N(\mathbf{S}\cap D)} = E_{\mathbf{S}\cap D}N_D(\mathbf{s},d,\mathcal{S}\backslash\mathbf{s})$
- $K(d)\equiv EN(s,d,\mathcal{S}\backslash\mathbf{s})/\lambda$ while what we can create is $K_D(d)\equiv E_{\mathbf{S}\cap D}N_D(s,d,\mathcal{S}\backslash\mathbf{s})/\lambda$.
- So the uncorrected estimator is based on $\bar{h}_{K,d}(\mathbf{S}) = \sum_{\mathbf{s}_i\in\mathbf{S}\cap D}\frac{N_D(\mathbf{s}_i,d,\mathcal{S}\backslash\mathbf{s}_i)}{N(\mathbf{S}\cap D)\lambda}$ whose expectation is $K_D(d)$.
- Again, we see the need for edge correction. We are estimating $K_D(d)$ rather than $K(d)$.
- In fact, since $N_D(\mathbf{s},d,\mathbf{S})\leq N(\mathbf{s},d,\mathbf{S})$, $K_D(d)\leq K(d)$.

Cross-validation for Point Patterns

- Cross-validation can provide model assessment without encouraging overfitting.
- Limited discussion of cross-validation methods for point processes. Leave-one-out cross-validation from Diggle for bandwidth-selection for kernel intensity estimate
- With a model having dependence between locations of the points, is leave-one-out sensible? (e.g., for Gibbs processes can't remove points without altering the interpoint distances)
- For models with conditionally independent locations given the intensity, leave-one-out does make sense and, more efficient, fitting data and validation data.
- To choose the fitting data, can't remove say 10% of the data? This will *fix* the size of the point pattern.

cont.

- Rather, the p-thinning approach.
- p-thinning independently deletes each point $\mathbf{s}_i \in \mathbf{S}$ with probability $1 - p$. Yields $\mathbf{S}^{fit}$ and $\mathbf{S}^{val}$. They are independent conditional on $\lambda(\mathbf{s})$.
- $\mathbf{S}^{fit}$ has intensity $p\lambda(\mathbf{s})$, $\mathbf{S}^{val}$ has intensity $(1 - p)\lambda(\mathbf{s})$
- To use fitted model for cross-validation purposes, we thin the posterior draws from fitted model to predictive draws $(\frac{p}{1-p})$ to compare with the held-out data for model adequacy and model selection
- For implementation: Partition domain D into subregions $B_1, B_2, \ldots B_K$ (any shape but equal area) and evaluate a residual measure on each

Model adequacy

- Here, the situation is a bit less clear. There is no single criterion for model adequacy
- Posterior predictive model checking (Gelman, Meng, Stern) or prior predictive model checking (Dey, Gelfand, Swartz, Vlachos)
- GMS is more common, easier to do, but doesn't criticize the model well enough, uses the data twice (once to fit, once to check).
- DGSV is more computationally demanding but is formally cleaner, uses the data only once.

Predictive model checking

- Both GMS and DGSV look at discrepancy measures, $D(\mathbf{S}; \theta)$, for example, $N(A) - \lambda(A; \theta)$.
- GMS compares $[D(\mathbf{S}; \theta)|\mathbf{S}_{obs}]$ with $[D(\mathbf{S}_{obs}; \theta)|\mathbf{S}_{obs}]$.
- The problem: draws of $\mathbf{S}$ from $\mathbf{S}; \theta|\mathbf{S}_{obs}]$ will be too much like $\mathbf{S}_{obs}$, discrepancies will be too much like $D(\mathbf{S}_{obs}; \theta)$, the model checking won't be critical enough.
- DGSV create $\mathbf{S}_l^*$'s from the marginal distribution of $\mathbf{S}$ by drawing θ_l^* from $[\theta]$ and then $\mathbf{S}_l^*$ from $[\mathbf{S}|\theta_l^*]$.
- Then, they obtain $[\mathbf{S}, \theta|\theta_l^*]$ and then compare $[D(\mathbf{S}_l^*; \theta)|\mathbf{S}_l^*]$ with $[D(\mathbf{S}_{obs}; \theta)|\mathbf{S}_{obs}]$.
- Apples with apples comparison, uses the data once
- DGSV compare the observed discrepancy with discrepancies you expect under the model; GMS compare the observed discrepancies with what you expect under the model **and** the data.

cont.

- Empirical coverage for model adequacy checking suffers the GMS problem; it will not be critical enough.
- For a collection of B_k's, consider $\{[N_{obs}(B_k) - N(B_k)|\mathbf{S}_{obs}]\}$. Check empirical coverage vs. nominal coverage.
- The $\mathbf{S}_l^*$'s will be too similar to $\mathbf{S}_{obs}$ (with weak priors) so the $N(B_k)$ that we generate given $\mathbf{S}_{obs}$ will be too much like $N_{obs}(B_k)$ (a function of $\mathbf{S}_{obs}$)
- Better to generate $N(B)$ through $\mathbf{S}_l^*$'s from the marginal distribution rather than from the posterior distribution.
- Now, a Monte Carlo test comparison between $[N_{obs}(B_k) - N(B_k)|\mathbf{S}_{obs}]$ and $\{[N_l^*(B_k) - N(B_k)|\mathbf{S}_l^*]\}$.
- A lot of comparison - for each B_k, compare an "observed" vs. say 99 generated posterior distributions, say using quantiles. Lots of simultaneous inference!
- No role for empirical coverage here unless out-of-sample. In-sample will be inadequate to criticize the model.

Model comparison

- Lack of useful model selection tools, especially for Bayesian models. Ad hoc tests of the homogeneity and independence assumptions of CSR, but not much for comparing models.
- Lack of likelihood precludes customary tools - AIC, BIC, DIC, Bayes factors
- In some cases, there may be a *natural* process, behavioral/mechanistic, to guide the choice of model prior to the analysis.
- First discussions of Bayesian model selection in Akman and Raftery - computing Bayes factors for NHPPs and change point Poisson processes.
- Guttorp and Thorarinsdottir (2012) perform model choice via a reversible jump algorithm to move between a nested pair of models.

cont.

- Model comparison should be done in predictive space since parameters don't mean anything across models
- For a collection of choices of $A \subset D$, focus on $[N(A)|\mathbf{S}_{obs}]$. In particular, compare $N_{obs}(A)$ with $[N(A)|\mathbf{S}_{obs}; M_j]$ for each model, $j = 1, 2, ..., J$.
- For model j with parameters $\boldsymbol{\theta}_j$, we obtain posterior samples, $\boldsymbol{\theta}^*_{j,l}$ and then $\mathbf{S}^*_{j,l}$.
- We want to do this out-of-sample, through $p-$thinning.
- We can do this for NHPP's, LGCP's and for cluster processes (superpositions of NHPP's)
- Criteria: PMSE, perhaps normalized by the expected number, empirical coverage, RPS
- If our only option, can do it in-sample

Ranked Probability Scores

- We propose ranked probability score (RPS) for general use, applied to predictive distributions for set counts.
- Specifically, we propose choosing subregions B_k uniformly over D, with each B_k having the same size and potentially overlapping other $B_{k'}$.
- In fact, we use the same B_k as in the Monte Carlo assessment above. Obtain $N(B_k)$ from the hold-out dataset, compare with $[N(B_k|\mathbf{S}_{fitted})]$ using posterior predictive point patterns
- For any B_k, we can write the *RPS* as $\text{RPS}(B_k) = \sum_{n=0}^{\infty}[F_{N(B_k)|\mathbf{S}_{fitted}}(n) - \mathbf{1}[n \geq N_{obs}(B_k)]]^2$. Can average over k to compare models
- Can also calculate in-sample *RPS* and compare with out of sample to see if model choice differs.

Finally!

- Can't use G, F, K, K_{inhom} to compare models.
- Model features. For example, can't say that G for one model is "better" than G for another model?
- Posterior distributions, e.g., $[G(d : \boldsymbol{\theta}_j)|\mathbf{S}_{obs}; M_j]$, can criticize say CSR which has known distance functions when CSR is nested within the fitted model, M_j.
- Compare, e.g., $[G(d : \boldsymbol{\theta}_j)|\mathbf{S}_{obs}; M_j]$ with empirical estimate $\hat{G}(d)$? Since latter is a *nonparametric* estimate, such comparison could be used to criticize M_j.
- Since K functions involve parameters, the empirical estimate will be semiparametric with parameter estimates based upon some model.
- Analogy with theoretical Q-Q plots