Statistical analysis of point patterns on the sphere

Jesper Møller, Aalborg University

Data example

Sky positions of 10546 galaxies outside the Milky Way band.

(cont.)

From the Revised New General Catalogue and Index Catalogue (RNGC/IC) and considered by Steinicke (2015) and Lawrence et al. (2016).

Dependence between the points such as spatial clustering between galaxies?

The k-dimensional unit sphere

$$\mathbb{S}^k = \{u \in \mathbb{R}^{k+1} \mid \|u\| = 1\}, \quad k = 1, 2, \ldots$$

Main interest: $k = 1$ (the circle) and $k = 2$ (the sphere).

Point pattern analysis

Models and tools in $\mathbb{R}^k$ and $\mathbb{S}^k$ are basically the same as long as the appropriate analogue is used:

- ▶ Homogeneity in $\mathbb{R}^k$ means that the point process distribution is invariant under translation. On the sphere the corresponding action is a rotation.

- ▶ Lebesgue measure on $\mathbb{R}^k$ should be replace by surface measure on $\mathbb{S}^k$.

- ▶ Eucledian distance should be replaced by great circle distance.

This has probably been well-known to people in the field 'forever', but never published in a mainstream understandable way until recently (apart from a short remark in Ripley, 1977).

History

▶ Robeson, Li and Huang (2014): Def. of $\hat{K}$ in the homogeneous case but without theoretical details.

▶ Lawrence, Baddeley, Milne and Nair (2016): 'Corrected' def. of $\hat{K}$ and taking care of inhomogeneity and boundaries + inhomogeneous Thomas process.

▶ Møller and Rubak (2016): Details on Palm distributions + functional summaries + DPPs.

▶ Cuevas-Pacheco and Møller (2018): LGCPs and GPs.

▶ Møller, Nielsen, Porcu and Rubak (2018): DPPs, in particular spectral representations.

Natural (transitive) group action is given by rotations

A rotation is a real $(k+1) \times (k+1)$ matrix R with $RR^T = I$ and $\det R = 1$. It acts on $\mathbb{S}^k$ as a linear map.

If $k = 1$: For $0 \le \theta < 2\pi$,

$$R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

If $k = 2$: See en.wikipedia.org/wiki/Rotation_matrix

Natural reference measure: Surface measure $\nu = \nu_k$

For $k = 1$ and $u = (\cos\theta, \sin\theta)$ with $0 \leq \theta < 2\pi$,

$$d\nu_1(u) = d\theta$$

is the usual Lebesgue measure on $[0, 2\pi)$.

For $k = 2$ and $u = (\sin\vartheta\cos\varphi, \sin\vartheta\sin\varphi, \cos\vartheta)$ where $\vartheta \in [0, \pi]$ is the polar latitude and $\varphi \in [0, 2\pi)$ is the polar longitude,

$$d\nu_2(u) = \sin\vartheta \, d\varphi \, d\vartheta.$$

For $k \geq 2$ and $u = (v\sin\vartheta, \cos\vartheta)$ with $v \in \mathbb{S}^{k-1}$ and $\vartheta \in [0, \pi]$,

$$d\nu_k(u) = \sin^{k-1}\vartheta \, d\nu_{k-1}(v) \, d\vartheta.$$

(cont.)

Let $\sigma_k = \nu_k(\mathbb{S}^k)$ be the surface area of the unit ball in $\mathbb{R}^{k+1}$, then

$$\sigma_k = \frac{2\pi^{(k+1)/2}}{\Gamma((k+1)/2)}.$$

Exercise: Verify the formula above when $k = 2$, using the definition of ν_2.

Further definitions

Natural metric d: <u>Geodesic metric</u> (shortest path distance or great circle distance)

$$d(u, v) = \arccos(u \cdot v), \quad u, v \in \mathbb{S}^k,$$

where $\cdot$ is the usual inner product, so $0 \leq d(u, v) \leq \pi$.

<u>Point process</u> on $\mathbb{S}^k$: a random finite subset $X \subset \mathbb{S}^k$; <u>isotropic</u> if $RX \sim X$ for every rotation R.

Its <u>intensity</u> (function) ρ, pair correlation function g (<u>pcf</u>), etc. are defined as in the Euclidean case but w.r.t. ν_k.
If g is <u>isotropic</u>, i.e., $g(u, v) = g_0(d(u, v))$, then X is said to be <u>SOIRI</u>)(second-order intensity reweighted isotropic) and we define the (inhomogeneous) K-function (for an arbitrary $u \in \mathbb{S}^k$) by

$$K(t) = \int_{d(u,v) \leq t} g_0(d(u, v)) \, d\nu_k(v) \quad \text{for } 0 \leq t \leq \pi.$$

Exercise:

1. Discuss what is meant by 'random finite subset $X \subset \mathbb{S}^k$'.

2. Verify that SOIRI is equivalent to that g is invariant under rotations, i.e., $g(Ru, Rv) = g(u, v)$ for all rotations R and all distinct $u, v \in \mathbb{S}^k$ if and only if g is of the form $g(u, v) = g_0(d(u, v))$ for all distinct $u, v \in \mathbb{S}^k$.

3. Why is X SOIRI if X is isotropic?

4. Show that

$$K(t) = \sigma_{k-1} \int_0^t g_0(\varphi) \sin^{k-1} \varphi \, \mathrm{d}\varphi.$$

The Poisson process

'Usual definition' of a Poisson process with intensity (function) ρ...

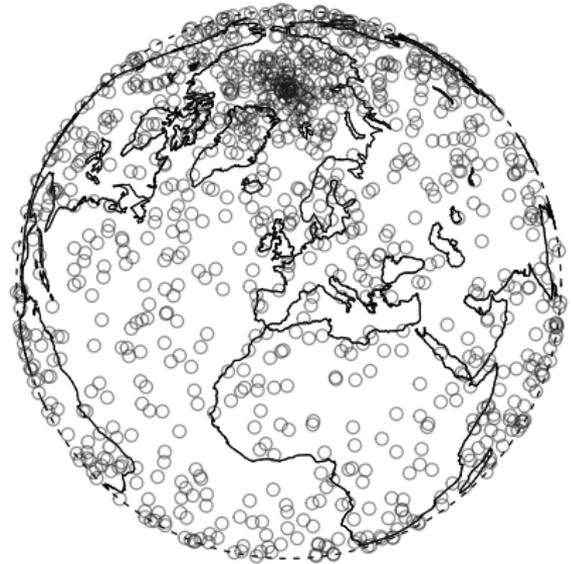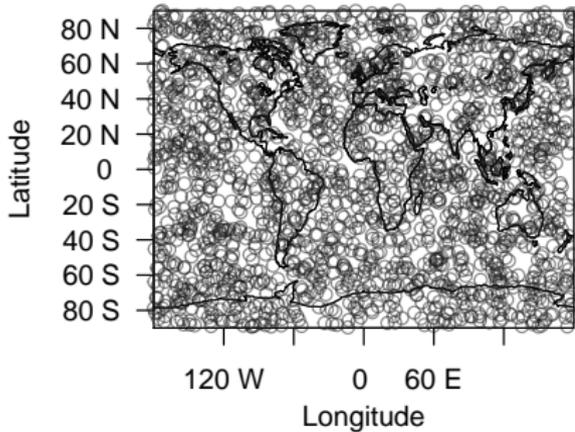For a Poisson process, $g = 1$, so SOIRI and

$$K(t) = \nu_k(\text{cap with polar longitude } t) = \sigma_{k-1} \int_0^t \sin^{k-1} \varphi \, d\varphi.$$

Exercise: Show that for a Poisson process,

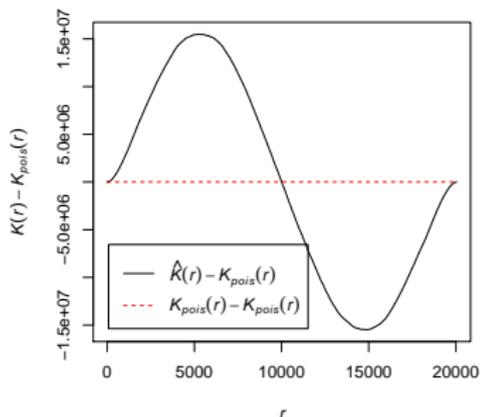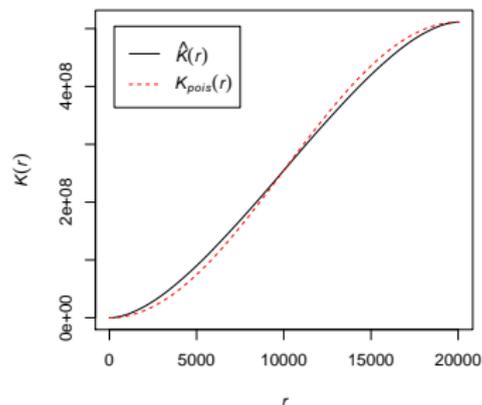$$K(t) = 2\pi(1 - \cos t) \quad \text{if } k = 2.$$

Not so uniformly distributed points on the sphere

▶ As you might know uniformly sampled points on a flat map doesn't correspond to uniform points on the sphere:

K-function for not so uniform points

▶ A plot of the K-function clearly shows departure
 (clustering) from what is expected under complete spatial
  randomness:

Proper uniform points on the sphere

▶ Instead of the previous naive approach the latitude should be the arccosine of a uniform point in $[-1, 1]$ which gives a nice homogeneous distribution:
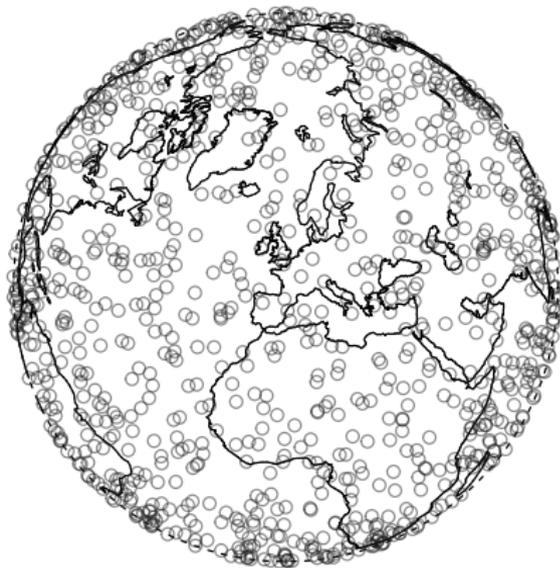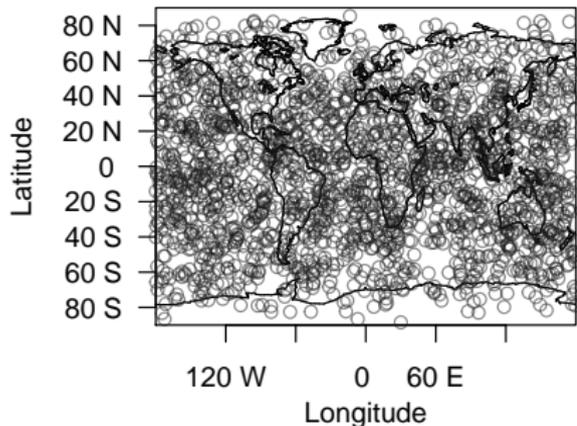
K-function for proper uniform points on the sphere

▶ Then the K-function agrees with complete spatial randomness (notice the completely different scale on the plot of the difference):
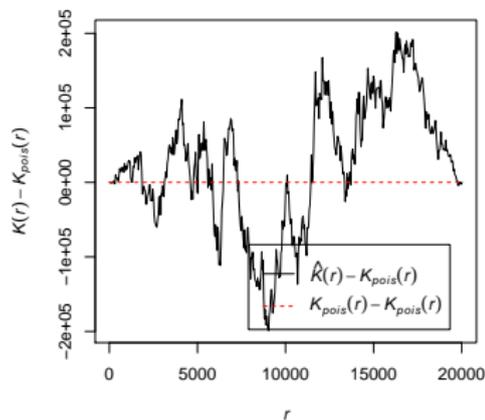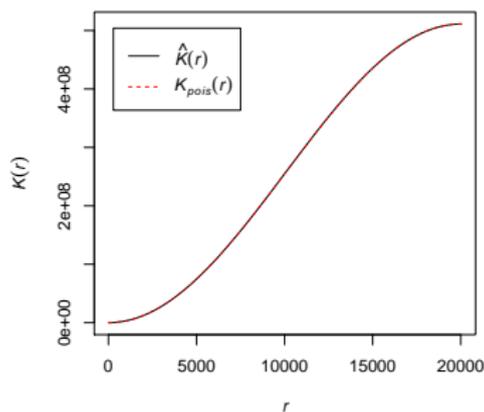
Reduced Palm distribution

For $u \in \mathbb{S}^k$, X_u follows the reduced Palm distribution at u if

$$\mathbb{E} \sum_{u \in X} h(u, X \setminus \{u\}) = \int \rho(u) \mathbb{E} h(u, X_u) \, d\nu_k(u)$$

for any non-negative (measurable) function h
(X_u can be arbitrary if $\rho(u) = 0$).

If $\mathcal{D}(X)$ is absolutely continuous w.r.t. Poisson (int.=1):

$$f_{\mathbf{x}}^!(\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}) = f(\{\mathbf{x}, \mathbf{x}_1, \ldots, \mathbf{x}_n\})/\rho(\mathbf{x}).$$

For a Poisson process X, we have $X_u \sim X$.

If X is isotropic, $\rho > 0$, and R_u rotates u to $e = (0, \ldots, 0, 1)$:

$$X_u \sim R_u X_e, \quad \mathbb{P}(X_e \in F) = \frac{1}{\rho \sigma_k} \mathbb{E} \sum_{u \in X} 1 \left[R_u^\top (X \setminus \{u\}) \in F \right].$$

Exercise

Suppose X is SOIRI. Let $A \subseteq \mathbb{S}^k$ (Borel) with $\nu_k(A) > 0$. It can be shown that

$$K(t) = \mathrm{E} \sum_{u \in X \cap A} \sum_{v \in X \setminus \{u\}} \frac{1(d(u,v) \leq t)}{\rho(u)\rho(v)\nu_k(A)}$$

$$= \int_{u \in A:\, \rho(u)>0} \mathbb{E} \sum_{v \in X_u} \frac{1(d(u,v) \leq t)}{\rho(v)\nu_k(A)} \, \mathrm{d}\nu_k(u).$$

1. Using the first formula above, discuss how a non-parametric estimator $\hat{K}(t)$ may be obtained if X has been observed within a window W and ρ is known (or has been estimated).

2. Assuming X is isotropic and $\rho > 0$, show that

$$\rho K(t) = \mathbb{E} \sum_{v \in X_e} 1(d(e,v) \leq t)$$

and interpret $\rho K(t)$.

Exercise

Assuming X is isotropic and $\rho > 0$, discuss how to define F, G, and J-functions.

Models

- ▶ Gibbs/Markov point processes, LGCPs, cluster point processes, and DPPs are defined in a similar way as in the Euclidean case.

- ▶ For LGCPs and DPPs we deal with covariance functions: $c : \mathbb{S}^k \times \mathbb{S}^k \mapsto \mathbb{C}$ is a complex covariance function if it is Hermitian and positive (semi-)definite. It is isotropic if $c(u, v) = c_0(d(u, v))$ for all $u, v \in \mathbb{S}^k$; then c is real. So $c(u, v) = c_0(d(u, v))$ is a covariance function if and only if c_0 is a real function so that $\sum_{i,j=1}^{n} a_i a_j c_0(d(u_i, u_j)) \geq 0$ for all $a_1, \ldots, a_n \in \mathbb{R}$ and all pairwise distinct $u_1, \ldots, u_n \in \mathbb{S}^k$.

- ▶ The theory for isotropic covariance functions on $\mathbb{S}^k$ (mainly due to classic papers by Schoenberg) differs from that of stationary covariance functions on $\mathbb{R}^k$ – similar parametric models have different ranges (see 'linear networks').

Parametric models for $c_0(r)$, $0 \leq r \leq \pi$

Model	Correlation function $c_0(r)$	Parameter range
Powered exp.	$\exp\left(-r^\alpha/\phi\right)$	$\alpha \in (0,1]$, $\phi > 0$
Matérn	$\frac{2}{\Gamma(\nu)}\left(\frac{r}{2\phi}\right)^\nu K_\nu\left(\frac{r}{\phi}\right)$	$0 < \nu \leq \frac{1}{2}$, $\phi > 0$
Gen. Cauchy	$(1 + (\frac{r}{\phi})^\alpha)^{-\tau/\alpha}$	$\phi, \tau > 0$, $\alpha \in (0,1]$
Dagum	$1 - ((\frac{r}{\phi})^\tau/(1 + (\frac{r}{\phi})^\tau))^{\frac{\alpha}{\tau}}$	$\phi > 0$, $\tau \in (0,1]$
multiquadric	$\left(\frac{(1-\delta)^2}{1+\delta^2-2\delta\cos r}\right)^\tau$	$\delta \in (0,1)$, $\tau > 0$
Sine power	$1 - \sin(r/2)^\alpha$	$\alpha \in (0,2)$
Spherical	$(1 + \frac{1}{2}\frac{r}{\phi})(1 - \frac{r}{\phi})_+^2$	$\phi > 0$
Askey	$(1 - \frac{r}{\phi})_+^\tau$	$\phi > 0$, $\tau \geq 2$
C^2-Wendland	$(1 + \tau\frac{r}{\phi})(1 - \frac{r}{\phi})_+^\tau$	$\phi \in (0,\pi]$, $\tau \geq 4$
C^4-Wendland	$(1 + \tau\frac{r}{\phi} + \frac{\tau^2-1}{3}\frac{r^2}{\phi^2})(1 - \frac{r}{\phi})_+^\tau$	$\phi \in (0,\pi]$, $\tau \geq 6$

Table 1: Here Γ is the gamma function, K_ν is the modified Bessel function of the second kind, and $t_+ := \max\{t, 0\}$ for $t \in \mathbb{R}$.

Comments to Table 1

For the powered exponential, Matérn, generalized Cauchy, Dagum, multiquadric, and sine power models, $k \in \{1, 2, \ldots\}$, whilst for the spherical, Askey, C^2-Wendland, and C^4-Wendland models, $k \in \{1, 2, 3\}$.

For each model, the specified parameter range ensures that $c_0(r)$ is well-defined, cf. Gneiting (2013), and hence for any $\sigma^2 > 0$, $c(r) := \sigma^2 c_0(r)$ is an isotropic covariance function.

Analysing the sky positions of galaxies

Sky positions of 10546 galaxies outside the Milky Way band.

Cox process

Conditional on a non-negative process $\Lambda = \{\Lambda(u) \mid u \in \mathbb{S}^k\}$, X is assumed to be a Poisson process with intensity function Λ.

This Cox process is well-defined if $\int_{\mathbb{S}^k} \mathbb{E}\Lambda(u) \, d\nu_k(u) < \infty$; then the intensity function is $\rho(u) = \mathbb{E}\Lambda(u)$.

Pair correlation function:

$$g(u, v) = \frac{\mathbb{E}[\Lambda(u)\Lambda(v)]}{\mathbb{E}\Lambda(u)\mathbb{E}\Lambda(v)}$$

taking $0/0 = 0$ and provided
$\int_{\mathbb{S}^k} \int_{\mathbb{S}^k} \mathbb{E}[\Lambda(u)\Lambda(v)] \, d\nu_k(u) \, d\nu_k(v) < \infty$.

Model ρ and the residual process $\Lambda_0(u) = \Lambda(u)/\rho(u)$ (so $g(u, v) = \mathbb{E}[\Lambda_0(u)\Lambda_0(v)]$).

Data analysis

Lawrence et al. (2016) fitted the intensity function

$$\rho(u) = 6.06 - 0.112 \sin\theta \cos\phi - 0.149 \sin\theta \sin\phi + 0.320 \cos\theta + 1.971 \cos^2\theta,$$

where $u = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$, $\theta \in [0, \pi]$ is the colatitude, and $\phi \in [0, 2\pi)$ is the longitude.

Different plots and tests in the accompanying supporting information to Lawrence et al. (2016): the galaxies are aggregated and not well-described by an inhomogeneous Poisson process model.

Thomas process

Lawrence et al. (2016) proposed the 'inhomogeneous Thomas process', that is,

$$\Lambda_0(u) = \sum_{y \in Y} f_{y,\xi}(u)/\kappa$$

where Y is a homogeneous Poisson point process with intensity $\kappa > 0$ and

$$f_{y,\xi}(u) = \frac{\xi}{4\pi \sinh \xi} \exp(\xi(u \cdot y)), \quad u \in \mathbb{S}^2,$$

is the density of the von Mises-Fisher distribution on $\mathbb{S}^2$ ($\xi > 0$).

Then

$$K(r) = K_{(\kappa,\xi)}(r) = K_{\mathrm{Pois}}(r) + \frac{\cosh(2\xi) - \cosh\left(\sqrt{2\xi^2(1 + \cos r)}\right)}{4\kappa \sin^2 \xi}.$$

LGCP

Cuevas-Pacheco and Møller (2018): LGCP with the same fitted
intensity function as above and $\log \Lambda_0$ a Gaussian process with
an isotropic covariance function given by the multiquadric
model

$$c_0(r) = \sigma^2 \left(\frac{(1-\delta)^2}{1 + \delta^2 - 2\delta \cos r} \right)^\tau, \quad \delta \in (0,1), \tau, \sigma > 0,$$

and s.t. $\log \Lambda_0$ has mean $-\sigma^2$ ($\Leftrightarrow \mathbb{E}\Lambda_0 = 1$).

Model fitting

To fit both point process models a minimum contrast procedure was used, that is,

$$\int_a^b \left(\widehat{K}(t)^{0.25} - K_\theta(t)^{0.25} \right)^2 \, \mathrm{d}t$$

where $b > a \geq 0$ are user-specified parameters and θ is the parameter vector for the residual process.

Thomas: The expression for K above.
LGCP: Numerical approximation.

Model checking

Because the K-function has been used for the estimation procedure, we need other functional summaries for model checking.

The F, G and J-functions are not well defined....

However, an independent thinning with retention probabilities $\rho_{min}/\rho(u)$ ensures that the resultant point process X_{thin} is isotropic.

Exercise

Is X_{thin} a Cox process and then what is the residual process?

Model checking: details

Under each fitted Thomas and LGCP model based on the data:

1. Simulate 7499 realizations of X_{thin} and compute $\widehat{F}$, $\widehat{G}$, $\widehat{J}$-functions for each simulation.

2. Compute 95%-global envelopes and calculate the p-value of the global rank envelope test.

A small detail: Cuevas-Pacheco and Møller (2018) used the global rank envelope test and considered intervals of p-value given by liberal (lower) and conservative (upper) values, but today I would had used the extreme rank length test and a single p-value, see Myllymäki et al. (2017), Mrkvička et al. (2018), Myllymäki and Mrkvička (2019). However, the conclusions in the following would be the same.

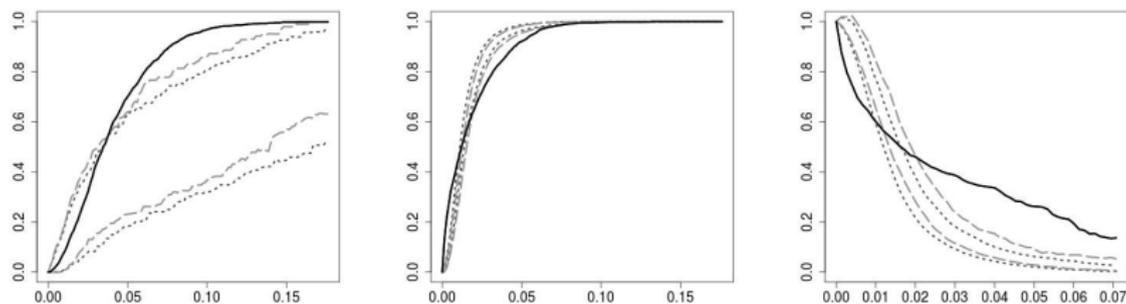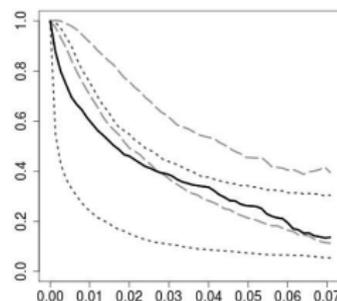# Model checking for fitted Thomas process

Figure 1: F-function   Figure 2: G-function   Figure 3: J-function

Full lines: Functional summary statistics for each simulation.
Dashed lines: Envelopes using integration interval $[0, 1.396]$ (as in Lawrence et al., 2016); $0.01\% \leq p \leq 1.28\%$.
Dotted lines: Envelopes using integration interval $[0, 0.175]$ (corresponding to 0–10 degrees); $0.01\% \leq p \leq 1.05\%$.
In both cases, p is less than about 1%.
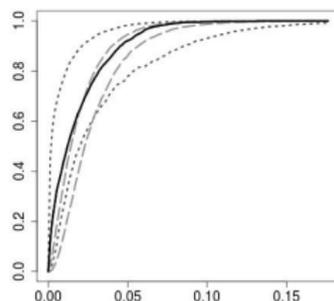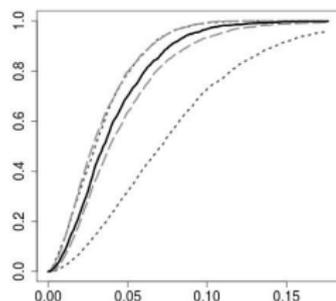
Model checking for fitted LGCP process

Figure 4: F-function   Figure 5: G-function   Figure 6: J-function

Full lines: Functional summary statistics for each simulation.
Dashed lines: Envelopes using integration interval $[0, 1.396]$;
$0.01\% \leq p \leq 1.23\%$.
Dotted lines: Envelopes using integration interval $[0, 0.175]$;
$24.02\% \leq p \leq 24.09\%$.

Model checking sensitivity

To study the sensitivity of the model checking procedure based on thinning and a global rank envelope test, Cuevas-Pacheco and Møller (2018) generated 1000 times 'a new point pattern dataset' and repeated the model checking procedure each time.

Thereby, for each combination of the two fitted models and each of the two integration intervals, we obtained 1000 estimates of $(\hat{F}, \hat{G}, \hat{K})$ and 1000 p-values.

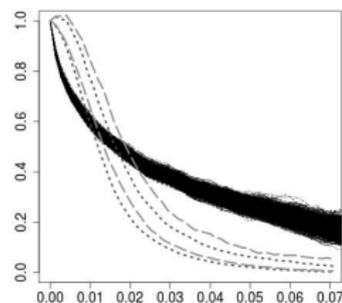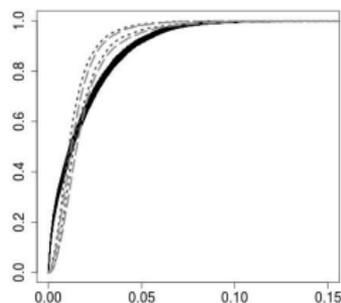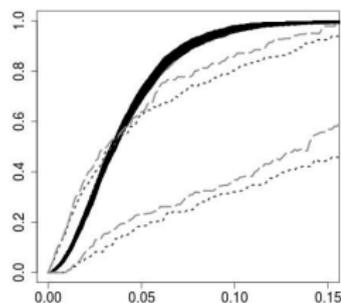# Model checking sensitivity for Thomas process

Figure 7: F-function   Figure 8: G-function   Figure 9: J-function

Full lines: Functional summary statistics for each simulation.
Dashed lines: Envelopes using integration interval $[0, 1.396]$.
Dotted lines: Envelopes using integration interval $[0, 0.175]$.

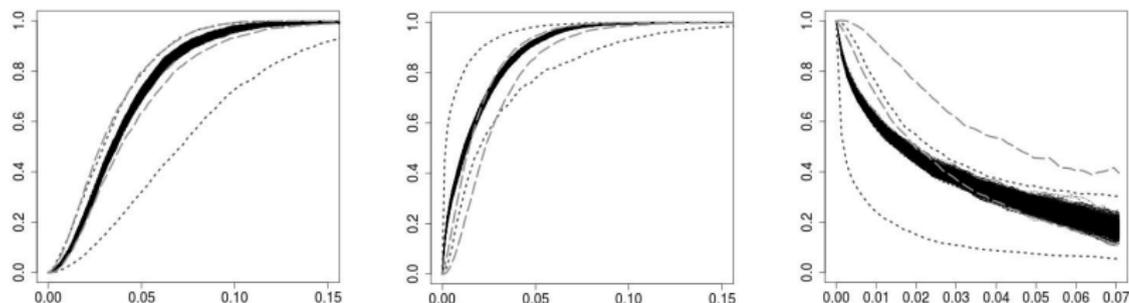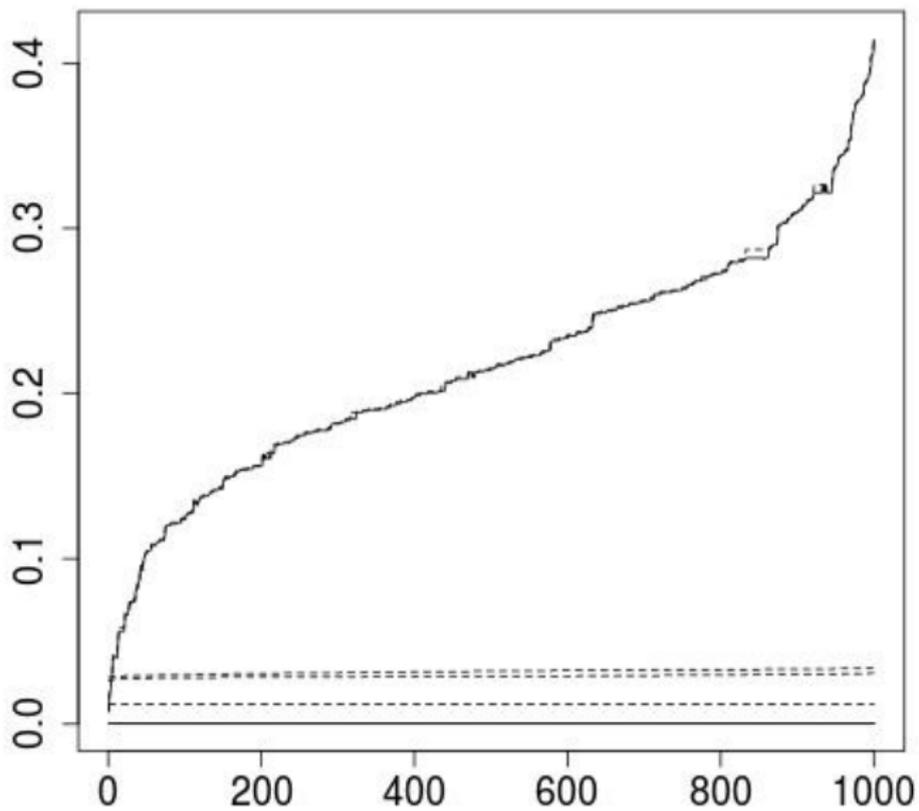# Model checking sensitivity for LGCP process

Figure 10: F-function   Figure 11: G-function   Figure 12: J-function

Full lines: Functional summary statistics for each simulation.
Dashed lines: Envelopes using integration interval $[0, 1.396]$.
Dotted lines: Envelopes using integration interval $[0, 0.175]$.

Sensitivity for the *p*-values (intervals)

Comments to the plot

Intervals for *p*-values obtained from the global envelope test based on combining the *F*, *G*, *J*-functions when repeating the independent thinning procedure 1000 times.

The three lower solid lines are very close and therefore appear as one thick solid line in the plot.

From below to the top: Each pair of the dashed and solid lines corresponds to conservative and liberal p-values for the fitted Thomas process, using first the long and second the short integration interval, and for the fitted LGCP, using first the long and second the short integration interval.

Determinantal point processes (DPPs)

▶ Defined by a covariance function $c(u, v)$ s.t.

$$\rho^{(n)}(u_1, \ldots, u_n) = \det\{c(u_i, u_j)\}_{i,j=1,\ldots,n}.$$

▶ Well defined if and only if $\mathrm{spectrum}(c) \leq 1$.

▶ DPPs have beautiful mathematical structure.

▶ They model inhibition between points.

▶ They have been studied extensively in probability theory and physics the past 40 years.

▶ Statistical methodology was developed in Lavancier, Møller and Rubak (2015).

▶ Extensions to the sphere were made in Møller and Rubak (2016) and Møller, Nielsen, Porcu and Rubak (2018).

Parametric DPP models

- ▶ For a fixed intensity (i.e., fixed mean number of points on the sphere) there is a limit to how strong inhibition a DPP can have.

- ▶ We have a 'most repulsive DPP' which in many cases can be the most interesting case since we have strongest interaction here.

- ▶ Ideally we would like to have parametric classes than span from Poisson to most repulsive.

- ▶ There are two promising parametric model classes:

  - ▶ Multi-quadric: Has two parameters and has nice closed form expressions for the moments etc., but does not quite cover the most repulsive case.

  - ▶ A certain spectral model: Has three parameters (is more flexible) but the parameters are less easy to interpret.

Simulations
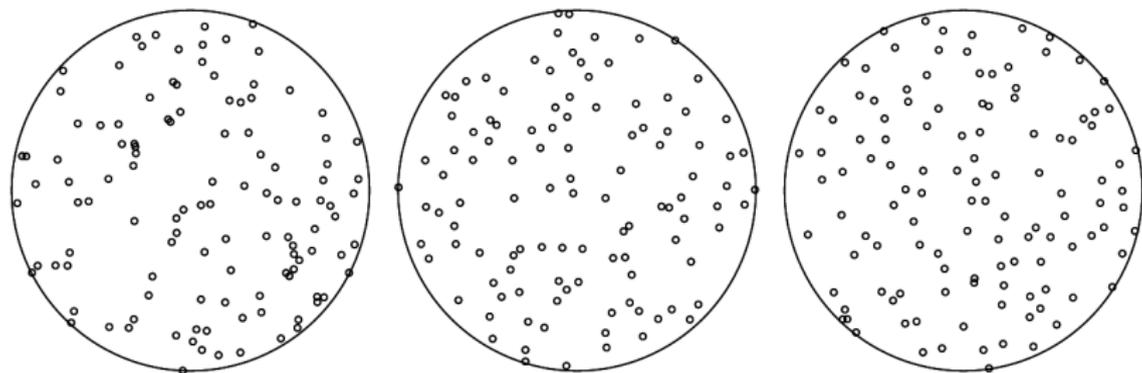
Simulated realizations from 3 DPP models on $\mathbb{S}^2$ with mean number of points 225 when the Northern Hemisphere spherical point patterns have been projected to the unit disc (with an equal-area azimuthal projection). Left: Poisson; middle: multi-quadric; right: most repulsive DPP.

Something on estimation of the K-function

Assume X is SOIRI, $\rho(u) = \rho$ is constant, and X is fully observed. Natural estimator:

$$\widehat{K}(t) = \frac{\sigma_k}{N(N-1)} \sum_{u,v \in X}^{\neq} 1 \left[d(u,v) \leq t \right]$$

corresponding to estimating ρ^2 by $N(N-1)/\sigma_k^2$ (with $N = \#X$), which is unbiased for the Poisson process.
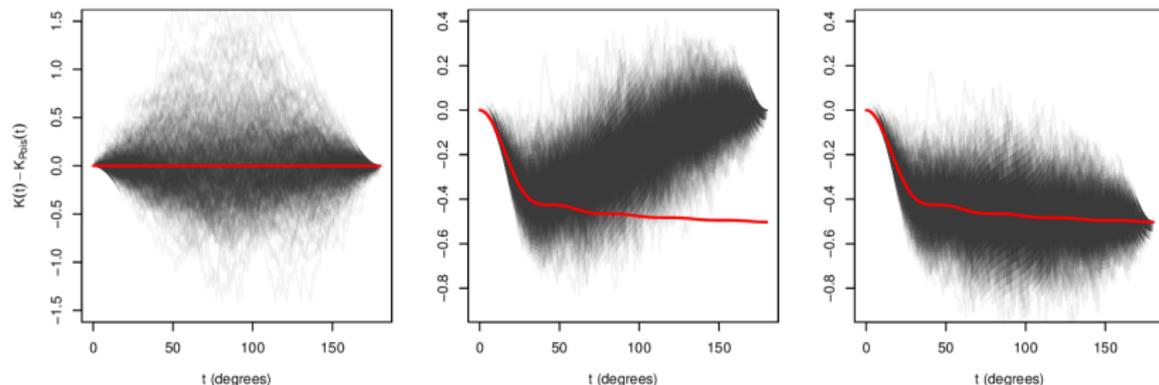
At a first glance the normalization seems unproblematic since

$$\widehat{K}(\pi) = K_{\mathrm{Pois}}(\pi) = \sigma_k.$$

A bias problem

However, in general $K(\pi) = \widehat{K}(\pi) + \frac{1}{\rho}\left(\frac{\mathrm{Var}(N)}{\mathrm{E}(N)} - 1\right)$.

For 3 DPP models and $k = 2$: $\hat{K} - K_{\mathrm{Pois}}$ for 500 simulated point patterns

- ▶ Left: Poisson model and usual $\widehat{K}$.
- ▶ Middle: Most repulsive DPP and usual $\widehat{K}$.
- ▶ Right: Most repulsive DPP and modified estimator $(\widehat{\rho^2} = \frac{N^2}{\sigma_k^2})$.

Exercise

1. Verify that

$$K(\pi) = \sigma_k + \frac{1}{\rho}\left(\frac{\mathbb{V}\mathrm{ar}(N)}{\mathbb{E}(N)} - 1\right).$$

2. Argue why it helps to use $\widehat{\rho}^2 = \frac{N^2}{\sigma_k^2}$ instead of $\widehat{\rho}^2 = \frac{N(N-1)}{\sigma_k^2}$ when estimating K for a DPP.

3. Would this work for a point process with clustering?

Software

Use

- `spatstat` package in R for simulation and estimation,
- `GET` (Myllymäki and Mrkvička, 2019) to make global envelope tests,
- `ggplot2` (Wickham, 2016) for visualisation.

References

Cuevas-Pacheco, F. and Møller, J. (2018). Log Gaussian Cox processes on the sphere. *Spatial Statistics*, **26**, 69-82.

Gneiting, T. (2013). Strictly and non-strictly positive definite functions on spheres. *Bernoulli*, **19**, 1327–1349.

Lavancier, F., Møller, J. and Rubak, E. (2015). Determinantal point process models and statistical inference. *Journal of Royal Statistical Society: Series B (Statistical Methodology)*, **77**, 853-877.

Lawrence, T., Baddeley, A., Milne, R.K. and Nair, G. (2016). Point pattern analysis on a region of a sphere. *Stat*, **5**, 144–157.

(cont.)

Møller, J., Nielsen, M., Porcu, E. and Rubak, E. (2018). Determinantal point process models on the sphere. *Bernoulli*, **24**, 1171-1201.

Møller, J. and Rubak, E. (2016). Functional summary statistics on the sphere with an application to determinantal point processes. *Spatial Statistics*, **18**, 4-23.

Mrkvička, T., Myllymäki, M., Jílek, M. and Hahn, U. (2018). A one-way ANOVA test for functional data with graphical interpretation. Available at arXiv:1612.03608.

Myllymäki, M. and Mrkvička, T. (2019). GET: Global envelopes in R. Available at arXiv:1911.06583.

Myllymäki, M. and Mrkvička, T., Grabarnik, P., Seijo, H. and Hahn, U. (2017). Global envelope tests for spatial processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **79**, 381–404.

Ripley, B.D. (1977). Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **39**, 172–212.

Robeson, S.M., Li, A. and Huang, C. (2014). Point-pattern analysis on the sphere. *Spatial Statistics*, **10**, 76–86.

Steinicke, W.: Revised new general catalogue and index catalogue, 2015. (Available from http://www.klima-luft.de/steinicke/index.htm). [Accessed February 1, 2018].

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.