Analysis of multi-species point patterns using multivariate log Gaussian Cox processes

Rasmus Waagepetersen
Department of Mathematical Sciences
Aalborg University

Based on completed work with Abdollah Jalilian, Yongtao Guan, Jorge Mateu
and
ongoing work with Jeff Coeurjolly and Achmad Choiruddin

Tropical rain forest data

Large Spatio-Temporal point pattern data:

- Locations of high number ($\approx$ 300.000) of trees
- Many ($\approx$ 300) different types of trees
- temporal data: trees observed each 5 years

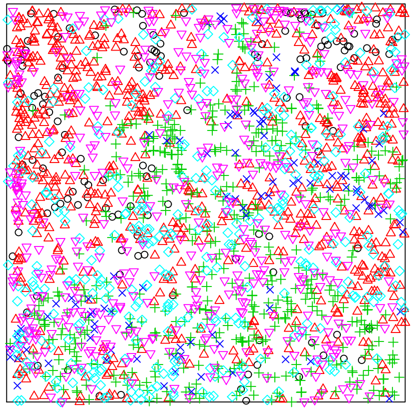
Aim: discuss selected approaches to statistical analysis of multivariate point patterns - and some plans for further development

Outline:

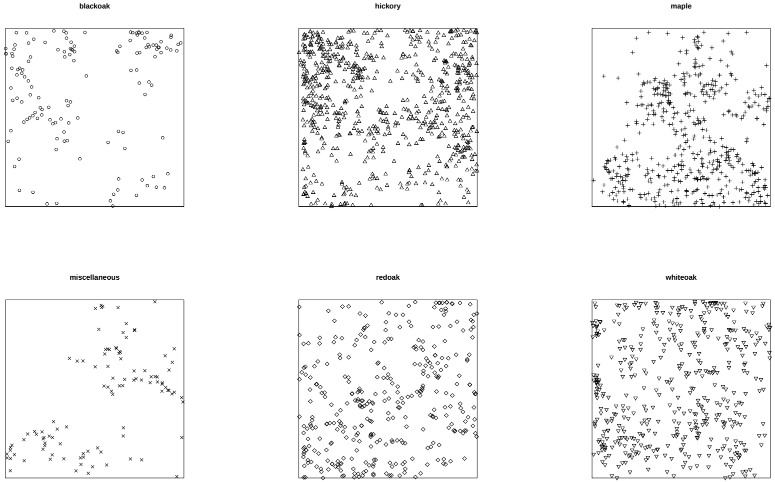
1. bivariate cross summary statistics
2. multivariate log Gaussian Cox process models
3. efficient algorithms and regularization

Example: Lansing Woods data (small)

Locations of 6 types of trees in Lansing Woods, Michigan.



Each type separately:



Objectives of statistical analysis

- Basic: study bivariate dependence for pairs of species
- Advanced: study underlying mechanisms that govern multivariate dependence structure

Both objectives can be addressed using statistics for spatial point processes.

Multivariate point process

Multivariate point process on $\mathbb{R}^2$:

$$\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_p)$$

collection of point processes $\mathbf{X}_i$.

Each $\mathbf{X}_i$ random set of points in $\mathbb{R}^2$ so that $\mathbf{X}_i \cap B$ is finite for any bounded $B \subseteq \mathbb{R}^2$.

Intensity function $\rho_i(\cdot)$:

$$\mathbb{E}\#\mathbf{X}_i \cap B = \int_B \rho_i(u)\mathrm{d}u$$

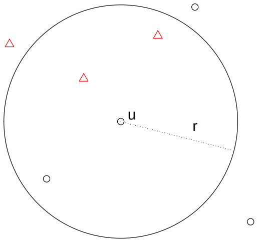
$$\rho_i(u)\mathrm{d}u \approx P(\mathbf{X}_i \text{ has a point at } u)$$

NB: u generic notation for location in $\mathbb{R}^2$.

Cross summary statistics (stationary case)

Stationary case: $\rho_j(u) = \rho_j$ constant.

Consider number of points in $\mathbf{X}_j$ within distance r from $u \in \mathbf{X}_i$.



Cross K_{ij}-function:

$$\rho_j K_{ij}(r) = \mathbb{E}\left[\text{ number of points in } \mathbf{X}_j \text{ within distance } r \text{ from } u \,|u \in \mathbf{X}_i\right]$$

Can be generalized to the case of non-constant intensity $\rho_j(\cdot)$.

Cross pair correlation function

K_{ij} is a cumulative quantity.

Pair correlation function is derivative:

$$g_{ij}(r) = \frac{K'_{ij}(r)}{2\pi r}$$

Infinitesimal interpretation:

$$g_{ij}(\|u - v\|) \approx \frac{P(\mathbf{X}_j \text{ has point at } v \,|\, \mathbf{X}_i \text{ has point at } u)}{P(\mathbf{X}_j \text{ has point at } v)}$$

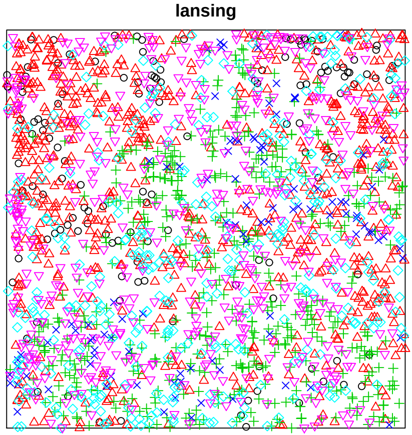
If $\mathbf{X}_i$ and $\mathbf{X}_j$ independent then

$$P(\mathbf{X}_j \text{ has point at } v \,|\, \mathbf{X}_i \text{ has point at } u) = P(\mathbf{X}_j \text{ has point at } v)$$
$$\Rightarrow g_{ij}(\cdot) = 1$$

$g_{ij}(\cdot) = 1 \Rightarrow \mathbf{X}_i$ and $\mathbf{X}_j$ uncorrelated

Example: Lansing woods

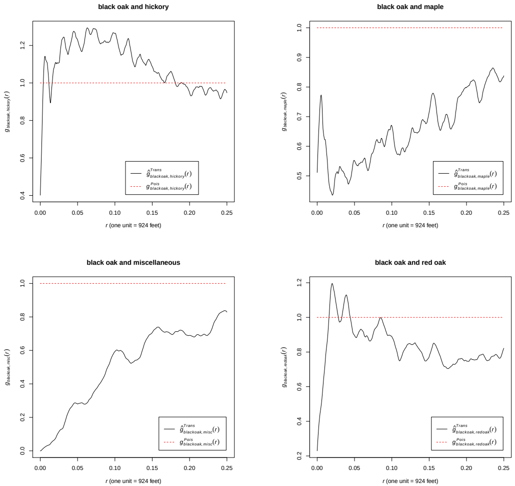
6 species ($\Rightarrow$ 15 pairs of species):



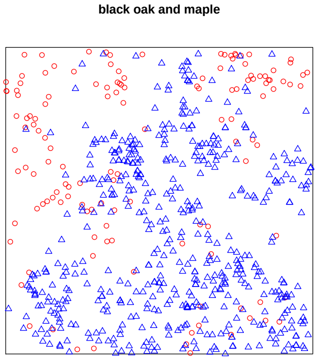
Empirical cross pair correlation functions

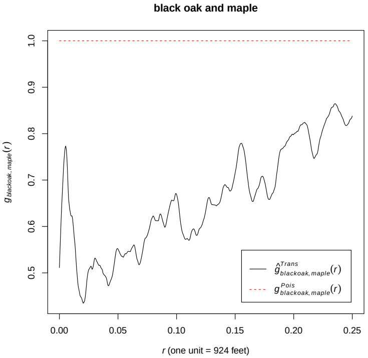
Pair correlation function can be estimated using kernel density estimate.

4 out of 15 cross pair correlation functions:



E.g. black oak and maple:





Seems that these species are segregated.

Perhaps species adapted to different environmental conditions ?

Issues with non-parametric analyses

1. given p species we have many - $O(p^2)$ - cross summary statistics.
   - hard to grasp information in $O(p^2)$ plots.
   - multiple testing.
2. pairwise/bivariate analyses only. Hard to get the big picture.

To learn more we need joint model-based approach.

Waagepetersen, Jalilian, Guan, Mateu (2016): multivariate log Gaussian Cox processes ($p = 9$)

Rajala, Olhede, Murrell (2017): multivariate Gibbs point processes ($p = 83$) - penalized pseudo-likelihood estimation.

I prefer Cox due to easier interpretation - but want higher p ☺

Multivariate Cox processes

Consider a multivariate non-negative random process

$$\Lambda(u) = [\Lambda_1(u), \ldots, \Lambda_p(u)], \quad u \in \mathbb{R}^2$$

A multivariate point process

$$\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$$

is a multivariate Cox process if $\mathbf{X}|\Lambda$ is a multivariate Poisson process with intensity function Λ.

Within- and between-species dependence originates from dependencies within and between the Λ_i.

Note Λ is *unobserved* latent process

Multivariate log Gaussian Cox process

$$\log \Lambda_i(u) = z(u)^\mathsf{T}\beta_i + Y_i(u) + U_i(u)$$

where

$$Y_i(u) = \sum_{l=1}^{q} \alpha_{il} E_l(u)$$

and $E_1, \ldots, E_q$, $U_1, \ldots, U_p$ independent Gaussian random fields.

- $z(u)$ observed spatial covariate
- E_l common latent factors (e.g. unobserved environmental covariates).
- U_i species-specific factors (within-species clustering - e.g. seed dispersal)
- known as linear model of coregionalization in geostatistics

Recall:

$$\Lambda_i(u) = \exp[z(u)^\mathsf{T}\beta_i + Y_i(u) + U_i(u)]$$

Intensity function:

$$\rho_i(u) = \mathbb{E}\Lambda_i(u) = \exp[z(u)^\mathsf{T}\beta_i + \sum_{l=1}^{q}\alpha_{il}^2/2 + \sigma_i^2/2] = \exp[\mu + z(u)^\mathsf{T}\beta_i]$$

Cross pair correlation function:

$$g_{ij}(h) = \begin{cases} \exp\left[\sum_{l=1}^{q}\beta_{ijl}c_l(h)\right] & i \neq j \\ \exp\left[\sum_{l=1}^{q}\beta_{ijl}c_l(h) + 1[i=j]\sigma_i^2 c_i(h)\right] & i = j \end{cases} \quad \beta_{ijl} = \alpha_{il}\alpha_{jl}$$

where $c_l(\cdot)$ and $c_i(\cdot)$ correlation functions of the E_l and U_i.

Estimation I

Intensity function can be estimated using composite likelihood approach:

$$\hat{\rho}_i(u) = \exp[\hat{\mu} + z(u)\hat{\beta}_i^\mathsf{T}]$$

Non-parametric kernel density estimates $\hat{g}_{ij}(r)$ of cross pair correlation functions.

Estimation II

Use exponential correlation models for E_l and U_i:
$c(r;\phi) = \exp(-r/\phi)$.

For fixed q minimize weighted least squares criterion to estimate θ (α and covariance parameters)

$$Q(\theta) = \sum_{k,i,j} w_{ijk} [\log \hat{g}_{ij}(t_k) - \log g_{ij}(t_k; \theta, q)]^2$$

Determination of q: K-fold cross-validation based on least squares criterion ($1/K$ of $\log \hat{g}_{ij}(t_k)$ left out)

E.g. $K = 8$ on a multicore machine with 8 CPUs $\rightarrow$ parallel computation

What can be inferred from fitted multivariate LGCP ?

- How many common latent fields $E_1, \ldots, E_q$? q measure of 'complexity'
- Decomposition of covariance into covariance due to common fields $E_1, \ldots, E_q$ and species specific fields U_i.
- Group species according to their pattern of dependence $\alpha_{i1}, \ldots, \alpha_{iq}$ on common fields:

$$Y_i(u) = \alpha_{i1} E_1(u) + \cdots + \alpha_{iq} E_q(u)$$

Decomposition of covariance

$$\log \Lambda_i(u) = z(u)^\mathsf{T}\beta_i + Y_i(u) + U_i(u)$$

Proportions of covariance due to common factors:

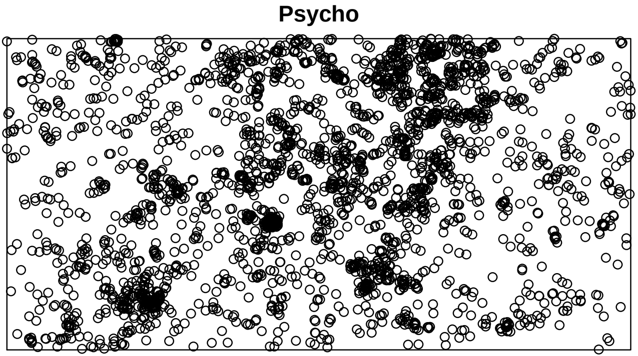
$$PV_i(h) = \frac{\mathbb{C}\mathrm{ov}[Y_i(u), Y_i(u+h)]}{\mathbb{C}\mathrm{ov}[\log \Lambda_i(u), \log \Lambda_i(u+h)]} = \frac{\sum_{l=1}^{q} \alpha_{il}^2 c_l(h)}{\sum_{l=1}^{q} \alpha_{il}^2 c_l(h) + \sigma_i^2 c_i(h)}$$

Application

9 abundant species from Barro Colorado Island plot.

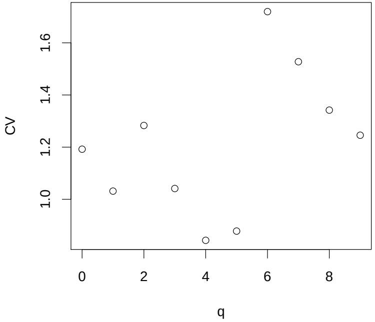
Covariates regarding topography, soil nutrients,...

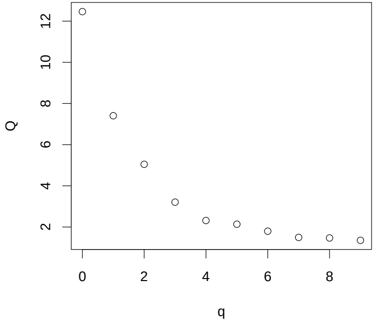
One species *Psychotria* (2640 trees):



Cross-validation

$CV(q)$ $Q(q)$





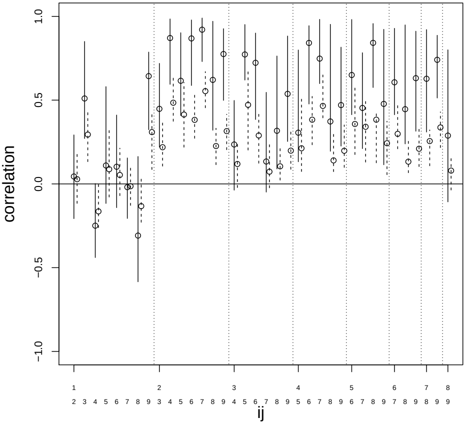
Smallest CV score for $q = 4$.

36 α_{il} and 4 correlation scale parameters for fields E_l. Total 40 parameters for 36 cross g_{ij} functions, $i < j$.

1.1 parameter for each cross pair correlation function.

Estimated correlations at zero lag (with bootstrap confidence intervals)

$$\log \Lambda_i(u) = z(u)^{\mathsf{T}}\beta_i + Y_i(u) + U_i(u)$$

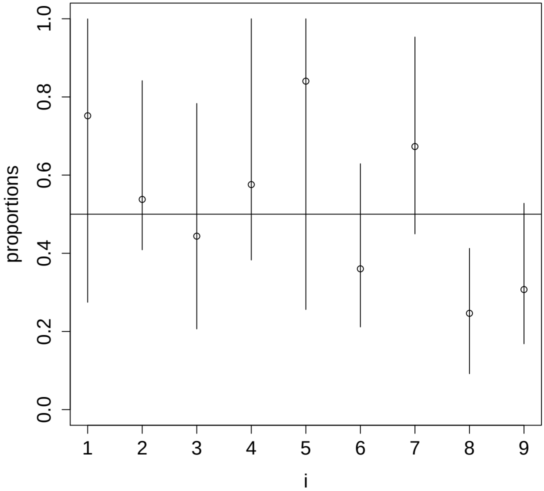


Solid: between $Y_i(u) = \sum_l \alpha_{il} E_l(u)$ and $Y_j(u) = \sum_l \alpha_{jl} E_l(u)$

Dashed: between $\log \Lambda_i(u)$ and $\log \Lambda_j(u)$

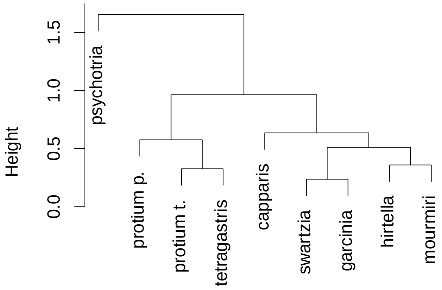
Most species positively correlated. Species 1 (*Psychotria*) is exception.

Proportions of variances at lag zero due to common fields



Clustering of species

Based on similarity of vectors $(\alpha_{i1}, \ldots, \alpha_{iq})$ and $(\alpha_{j1}, \ldots, \alpha_{jq})$.



Psychotria: distinct mode of seed dispersal (bird)

Protium p., Protium t., Tetragastris: the members of the Burseraceae family.

Challenges:

- only considered 9 species
- stability of estimation (numerical minimization)
- interpretability of model

Wishes:

- fast and stable computation.
- encourage sparse results.

Estimation of α_{il}

Focus on parameters α_{il}, $i = 1, \ldots, p$, $l = 1, \ldots, q$.

Fixing all other parameters, object function is of the form

$$\sum_{i,j} \|y_{ij} - x_{ij}\beta_{ij}\|^2$$

where y_{ij} $L \times 1$ 'response vector' and x_{ij} $L \times q$ 'design matrix'.

$$y_{ijk} = \log \hat{g}_{ij}(t_k) \quad (x_{ij})_{kl} = c(t_k; \phi_l) \quad \beta_{ijl} = \alpha_{il}\alpha_{jl}$$

One challenge: non-linear least-squares problem.

Another challenge: high-dimensional α - would be nice to use regularization to promote sparsity and stability of least squares solution.

Regularized least squares

Introduce elastic net penalty

$$\sum_{i,j} \|y_{ij} - x_{ij}\beta_{ij}\|^2 + \lambda p_\xi(\alpha)$$

where

$$p_\xi(\alpha) = \sum_{il} \left[(1-\xi)\alpha_{il}^2 + \xi|\alpha_{il}|\right]$$

Efficient algorithms available for regularized linear models but our problem is non-linear.

Block updates

Consider iterative procedure where α^m is value after m iterations.

Now update ith row $\alpha_{i\cdot} = (\alpha_{i1}, \ldots, \alpha_{iq})$ keeping other rows fixed: minimize

$$Q_i(\alpha_{i\cdot}) = 2 \sum_{\substack{j=1 \\ j \neq i}}^{p} \|y_{ij} - \tilde{x}_{ij}^m \alpha_{i\cdot}\|^2 + \|y_{ii} - x_{ii}\alpha_{i\cdot}^2\|^2 + \lambda p_\xi(\alpha_{i\cdot})$$

Here we rewrote

$$\|y_{ij} - x_{ij}\beta_{ij}^m\|^2 = \|y_{ij} - x_{ij}\mathrm{Diag}(\alpha_{j\cdot}^m)\alpha_{i\cdot}^m\|^2 = \|y_{ij} - \tilde{x}_{ij}^m \alpha_{i\cdot}^m\|^2$$

Note except for 'ii' term $Q_i(\alpha_{i\cdot})$ looks exactly like regularized least squares !

Approximate block update

Consider modified criterion

$$\tilde{Q}_i(\alpha_{i\cdot}) = 2\sum_{\substack{j=1\\ j\neq i}}^{p} \|y_{ij} - \tilde{x}_{ij}^m \alpha_{i\cdot}\|^2 + 2\|y_{ii} - \tilde{x}_{ii}^m \alpha_{i\cdot}\|^2 + \lambda p_\xi(\alpha_{i\cdot})$$

where

$$\tilde{x}_{ii}^m = x_{ii}\mathrm{Diag}(\alpha_{i\cdot}^m)$$

Minimizing $\tilde{Q}_i(\alpha_{i\cdot})$ standard regularized least squares problem (e.g. `glmnet`).

Does it work ?

Gradients of $Q_i(\alpha_{i\cdot})$ and $\tilde{Q}_i(\alpha_{i\cdot})$ coincide.

In simulation studies method works well - although increase in least criterion may be observed in first few iterations.

Wish: more convincing argument that approximate block updates are doing the right thing.

Issue: $\log \hat{g}_{ij}$ is biased estimate of $\log g_{ij}$ (due to kernel smoothing and log transformation)

One more wish: ‘unbiased’ response and design matrix:
$\mathbb{E} Y_{ij} = x_{ij}\beta_{ij}$

Variational approach

Variational point process identity (Jeff) specialized to the isotropic case:

$$\mathbb{E}\left\{\sum_{u\in\mathbf{X}_i,v\in\mathbf{X}_j}^{\neq} e(u,v)h(\|v-u\|)(\log g_{ij})'(\|v-u\|)\right\} =$$
$$-\mathbb{E}\left\{\sum_{u\in\mathbf{X}_i,v\in\mathbf{X}_j}^{\neq} e(u,v)h'(\|v-u\|)\right\},$$

where

$$e(u,v)=\frac{1[u\in W, v\in W]}{\rho_i(u)\rho_j(v)|W\cap W_{v-u}|}$$

and h is continously differentiable with compact support.

In our case,

$$\log g_{ij}(t) = \mathbf{c}(t)\beta_{ij}^{\mathsf{T}} \quad \mathbf{c}(t) = [c(t;\phi_1), \ldots, c(t;\phi_q)] \quad \beta_{ijl} = \alpha_{il}\alpha_{jl}$$

Let

$$h(t) = h_0(t)\mathbf{c}'(t)$$

where h_0 compact support and let

$$\mathbf{A} = \sum_{u,v\in\mathbf{X}\cap W}^{\neq} e(u,v)h_0(\|v-u\|)\mathbf{c}'(\|v-u\|)\{\mathbf{c}'(\|v-u\|)\}^{\top}$$

$$\mathbf{b} = -\sum_{u,v\in\mathbf{X}\cap W}^{\neq} e(u,v)\{h_0'(\|v-u\|)\mathbf{c}'(\|v-u\|) + h_0(\|v-u\|)\mathbf{c}''(\|v-u\|)$$

$\mathbf{A}$ is $q\times q$ and $\mathbf{b}$ is $q\times 1$.

Then from variational equation we obtain unbiased estimating function

$$\mathbf{A}\beta_{ij} - \mathbf{b}.$$

That is,

$$\mathbb{E}[\mathbf{A}\beta_{ij}] = \mathbb{E}\mathbf{b}$$

In terms of α_{il}, procedure can be recast as a least squares problem

$$\|\mathbf{b} - \mathbf{A}\beta_{ij}\|^2 \quad \beta_{ijl} = \alpha_{il}\alpha_{jl}$$

and we can introduce regularization as before.

Further open problems

- validity of using approximate block updates
- choice of λ and ξ (cross validation, BIC,...)
- Inference for proportions of variances, correlations... (bootstrap ?)
- choice of function h_0 in variational equation

References

- Møller, Syversveen, Waagepetersen (1998) Log Gaussian Cox processes, SJS.
- Brix and Møller (2001) Space-time multi type log Gaussian Cox processes, SJS.
- Guan, Jalilian, Mateu, Waagepetersen (2016) Analysis of multi-species point patterns using multivariate log Gaussian Cox processes, *Journal of the Royal Statistical Society, Series C*, **65**, 77-96.
- Jalilian, Guan, Mateu, Waagepetersen, R. (2015) Multivariate product-shot-noise Cox models, *Biometrics*, **71**, 1022-1033.
- Rajala, Olhede, Murrell (2017) Detecting multivariate interactions in spatial point patterns with Gibbs models and variable selection, *Journal of the Royal Statistical Society, Series C*, to appear.