Estimating functions and inhomogeneous point processes

Rasmus Waagepetersen
Department of Mathematics
Aalborg University
Denmark

February 7, 2017

Outline

Introduction to:

- estimating equations and quasi-likelihood
- inhomogeneous point processes
- estimating functions for inhomogeneous point processes

Examples of estimating equations

Least squares (non-linear) : suppose Y_i has mean $\mu_i(\beta)$.

Minimizing

$$\sum_{i=1}^{n}[Y_i - \mu_i(\beta)]^2$$

leads to estimating equation (first derivative)

$$D^{\mathsf{T}}[Y - \mu(\beta)] = 0$$

where (sensitivity)

$$D = \frac{\mathrm{d}\mu}{\mathrm{d}\beta^T} = \left[\mathrm{d}\mu_i/\mathrm{d}\beta_j\right]_{ij}$$

Moment estimation: suppose we know $\mathbb{E}_\theta g(Y)$ for some function g.

Then we estimate θ by solving

$$g(y) = \mathbb{E}_\theta g(Y) \Leftrightarrow \mathbb{E}_\theta g(Y) - g(y) = 0$$

I.e. choose θ so that empirical value of g matches its expected value.

Example:

$$\mathbb{E}SSE = \mathbb{E}\sum_{i=1}^{n}(Y_i - \bar{Y}_\cdot)^2 = (n-1)\sigma^2$$

Maximum likelihood estimation: suppose $f(y;\theta)$ is likelihood of observation y. Then maximum likelihood estimate is

$$\hat{\theta} = \operatorname*{argmax}_{\theta} f(y;\theta) = \operatorname*{argmax}_{\theta} \log f(y;\theta)$$

Typically we find $\hat{\theta}$ by differentiation and equating to zero:

$$s(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta} \log f(y;\theta) = 0$$

Exponential family:

$$f(y;\theta) = c(\theta)h(y)\exp[t(y)\cdot\theta]$$

Then score is

$$s(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta} \log f(y;\theta) = t(y) - \mathbb{E}_\theta t(Y)$$

Thus (moment estimation)

$$s(\theta) = 0 \Leftrightarrow t(y) = \mathbb{E}_\theta t(Y)$$

In general: estimating function e is function of data Y and unknown parameter θ. Estimate $\hat{\theta}$ is given as solution of estimating equation

$$e(\theta) = 0$$

(typically we suppress data Y from the notation).

Hopefully unique solution !

Optimality (one-dimensional case)

Let θ^* denote true value of θ. We want:

1. $e(\theta^*)$ close to zero
2. $e(\theta)$ differs much from zero when θ differs from θ^*

1. OK if $e(\theta)$ *unbiased* estimating function

$$\mathbb{E}_{\theta^*} e(\theta^*) = 0$$

and $\mathbb{V}\mathrm{ar}_{\theta^*} e(\theta^*)$ small.

2. OK if large sensitivity $e'(\theta^*)$ large

This leads to criteria $\left(\mathbb{E}_{\theta^*} e'(\theta^*)\right)^2 / \mathbb{V}\mathrm{ar}_{\theta^*} e(\theta^*)$ which should be as big as possible. Equivalently, $\mathbb{V}\mathrm{ar}_{\theta^*} e(\theta^*) / \left(\mathbb{E}_{\theta^*} e'(\theta^*)\right)^2$ should be as small as possible.

In the multidimensional case we consider

$$I = S(\theta^*)^{\mathsf{T}} \mathbb{V}\mathrm{ar}_{\theta^*} e(\theta^*)^{-1} S(\theta^*)$$

where S is *sensitivity matrix*

$$S(\theta) = -\mathbb{E}[\frac{\mathrm{d}}{\mathrm{d}\theta^{\mathsf{T}}} e(\theta)]$$

We then say that e_1 is better than e_2 if

$$I_1 - I_2$$

is positive semi-definite.

e is *optimal within a class* of estimating functions if it is better than any other estimating function in the class.

I is called the *Godambe information*.

Another view on optimality

By linear approximation (asymptotically) (assuming S^{-1} exists)

$$0 = e(\hat{\theta}) \approx e(\theta^*) - S(\hat{\theta} - \theta^*) \Leftrightarrow (\hat{\theta} - \theta^*) \approx S^{-1} e(\theta^*)$$

Thus

$$\mathbb{V}\mathrm{ar}\hat{\theta} \approx S^{-1}\Sigma(S^{-1})^{\mathsf{T}} = I^{-1} \quad \Sigma = \mathbb{V}\mathrm{ar}e(\theta)$$

Hence we say e_1 is better than e_2 if

$$\mathbb{V}\mathrm{ar}\hat{\theta}_2 - \mathbb{V}\mathrm{ar}\hat{\theta}_1 = S_2^{-1}\Sigma_2(S_2^{-1})^{\mathsf{T}} - S_1^{-1}\Sigma_1(S_1^{-1})^{\mathsf{T}}$$

is positive definite.

Same as before since

$$S_2^{-1}\Sigma_2(S_2^{-1})^{\mathsf{T}} - S_1^{-1}\Sigma_1(S_1^{-1})^{\mathsf{T}} = I_2^{-1} - I_1^{-1}$$

which is positive semi-definite if $I_1 - I_2$ is positive semi-definite (see useful matrix result on last slide).

Case of MLE

For likehood score (under suitable regularity conditions[1])

$$\mathbb{V}\mathrm{ar}_\theta s(\theta) = S$$

so that Godambe information

$$I = S$$

is equal to the Fisher information.

$$\mathbb{V}\mathrm{ar}\hat{\theta} \approx S^{-1}$$

[1] E.g. interchange of differentiation and integration allowed

Estimating functions and the likelihood score

The following result holds for an unbiased estimating function (under suitable regularity conditions) (one-dimensional case for ease of notation):

$$\mathbb{E}s(\theta)e(\theta) = \mathbb{C}\mathrm{ov}[s(\theta), e(\theta)] = S$$

This implies

$$\mathbb{C}\mathrm{orr}[s(\theta), e(\theta)]^2 = \frac{S^2}{\mathbb{V}\mathrm{ar}s(\theta)\mathbb{V}\mathrm{ar}e(\theta)} = \frac{I}{\mathbb{V}\mathrm{ar}s(\theta)}$$

That is the optimal estimating function has maximal correlation with the likelihood score.

Corollary: the likelihood score is optimal among all estimating functions.

Useful condition for optimality (Theorem 2.1, Heyde, 1997)

Consider a class $\mathcal{E}$ of estimating functions. e^o is optimal within $\mathcal{E}$ if for some constant invertible matrix K,

$$\Sigma_{ee^o} = \mathbb{C}\mathrm{ov}[e, e^o] = S_e K \tag{1}$$

for all $e \in \mathcal{E}$.

If $\mathcal{E}$ is convex then the converse is true too.

Note: if e^o is optimal then $(K^{-1})^{\mathsf{T}} e^o$ optimal too. Hence we can let $K = I$ without loss of generality. Then (1) implies $\mathbb{V}\mathrm{ar} e^0 = S_{e^o}$ and we obtain properties

$$I_{e^o} = S_{e^o} \quad \mathbb{V}\mathrm{ar}\hat{\theta}^o \approx S_{e^o}^{-1}$$

as for the likelihood score.

Proof of if part:

Define standardized estimating function $e_s = S_e^{\mathsf{T}} \Sigma_e^{-1} e$.

Then $\Sigma_{e_s} = \mathbb{V}\mathrm{ar} e_s = I_e$. Thus $I_{e^o} - I_e = \mathbb{V}\mathrm{ar} e_s^o - \mathbb{V}\mathrm{ar} e_s$.

Moreover (1) is equivalent to $\Sigma_{e_s e_s^o} = \Sigma_{e_s^o e_s} = \Sigma_{e_s}$. Then

$$\mathbb{V}\mathrm{ar}[e_s^o - e_s] = \Sigma_{e_s^o} - \Sigma_{e_s}$$

which proves the result since the LHS is positive semi-definite.

Exercises

1. show results on slide 'Estimating functions and the likelihood score' (hint: use the rule for differentiation of a product to show the first result)
2. (Quasi-likelihood) Suppose $Y = (Y_1, \ldots, Y_n)$ has mean vector $\mu(\beta)$ and (known) covariance matrix V.

   Consider the class of estimating functions

   $$A[Y - \mu(\beta)]$$

   where A $q \times n$ (all linear combinations of residual vector). Show that the optimal choice is $A = D^{\mathsf{T}} V^{-1}$.

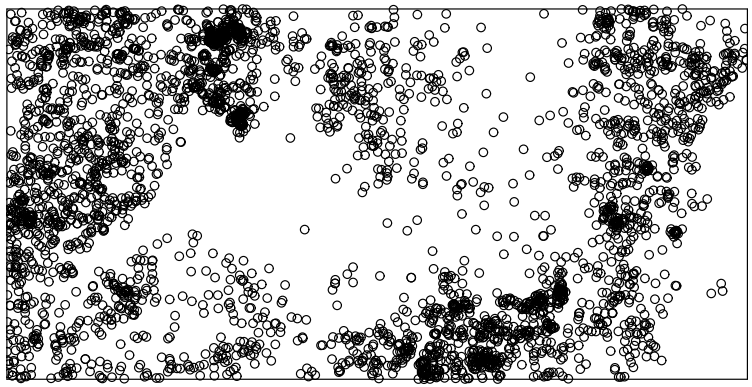
   What is the Godambe information matrix ?

Now: inhomogeneous point processes.

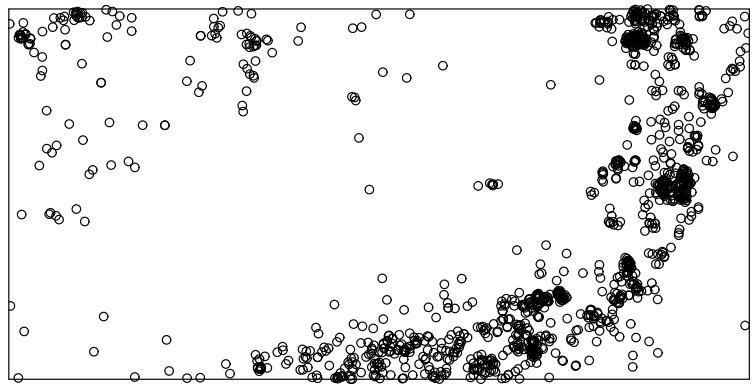
Data example: tropical rain forest trees

Observation window $W = [0, 1000] \times [0, 500]$

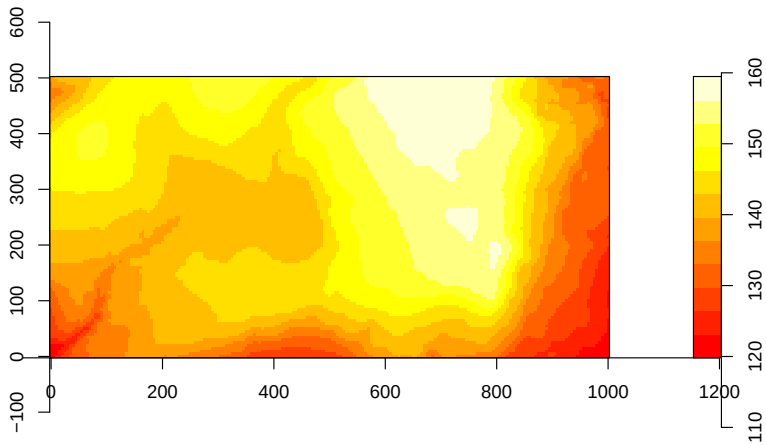
Beilschmiedia



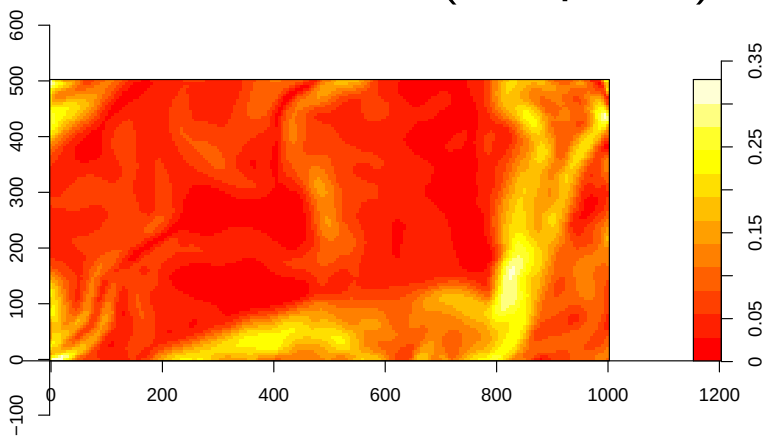
Ocotea



Elevation



Gradient norm (steepness)

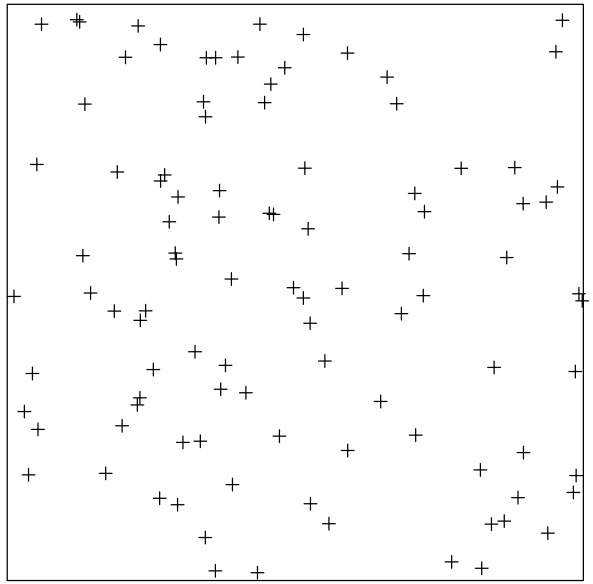


Sources of variation: elevation and gradient covariates *and* possible clustering/aggregation due to unobserved covariates and/or seed dispersal.

Spatial point process

Spatial point process: random collection of points

(finite number of points in bounded sets)



Intensity of a spatial point process

Fundamental characteristic of point process: mean of counts $N(A) = \#(\mathbf{X} \cap A)$.

Intensity of a spatial point process

Fundamental characteristic of point process: mean of counts $N(A) = \#(\mathbf{X} \cap A)$.

Intensity measure μ:

$$\mu(A) = \mathbb{E}N(A), \quad A \subseteq \mathbb{R}^2$$

Intensity of a spatial point process

Fundamental characteristic of point process: mean of counts $N(A) = \#(\mathbf{X} \cap A)$.

Intensity measure μ:

$$\mu(A) = \mathbb{E}N(A), \quad A \subseteq \mathbb{R}^2$$

In practice often given in terms of *intensity function*

$$\mu(A) = \int_A \rho(u)\mathrm{d}u$$

Intensity of a spatial point process

Fundamental characteristic of point process: mean of counts $N(A) = \#(\mathbf{X} \cap A)$.

Intensity measure μ:

$$\mu(A) = \mathbb{E}N(A), \quad A \subseteq \mathbb{R}^2$$

In practice often given in terms of *intensity function*

$$\mu(A) = \int_A \rho(u)\mathrm{d}u$$

Infinitesimal interpretation: $N(A)$ binary variable (presence or absence of point in A) when A very small. Hence

$$\rho(u)|A| \approx \mathbb{E}N(A) \approx P(\mathbf{X} \text{ has a point in A})$$

Covariance of counts and pair correlation function

Pair correlation function

$$\mathbb{E} \sum_{u,v \in \mathbf{X}}^{\neq} \mathbf{1}[u \in A,\, v \in B] = \int_A \int_B \rho(u)\rho(v) g(u,v)\, \mathrm{d}u\, \mathrm{d}v$$

Covariance between counts:

$$\mathbb{C}\mathrm{ov}[N(A), N(B)] = \int_{A \cap B} \rho(u) \mathrm{d}u + \int_A \int_B \rho(u)\rho(v)(g(u,v) - 1) \mathrm{d}u \mathrm{d}v$$

Pair correlation $g(u,v) > 1$ implies positive correlation.

Campbell formulae

From definitions of intensity and pair correlation function we obtain the Campbell formulae:

$$\mathbb{E} \sum_{u \in \mathbf{X}} h(u) = \int h(u) \rho(u) \mathrm{d}u$$

$$\mathbb{E} \sum_{u,v \in \mathbf{X}}^{\neq} h(u,v) = \iint h(u,v) \rho(u) \rho(v) g(u,v) \mathrm{d}u \mathrm{d}v$$

The Poisson process

Assume μ locally finite measure on $\mathbb{R}^2$ with density ρ.

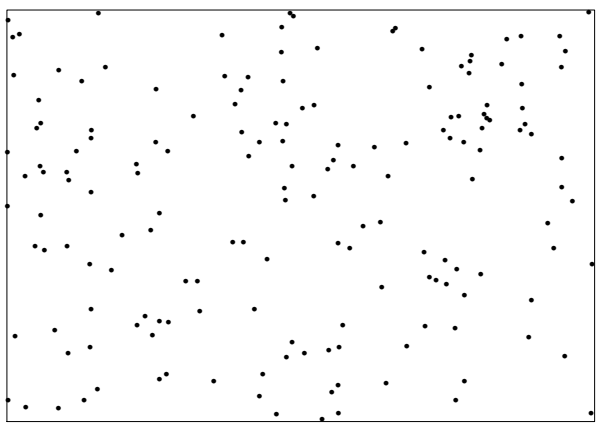
The Poisson process

Assume μ locally finite measure on $\mathbb{R}^2$ with density ρ.

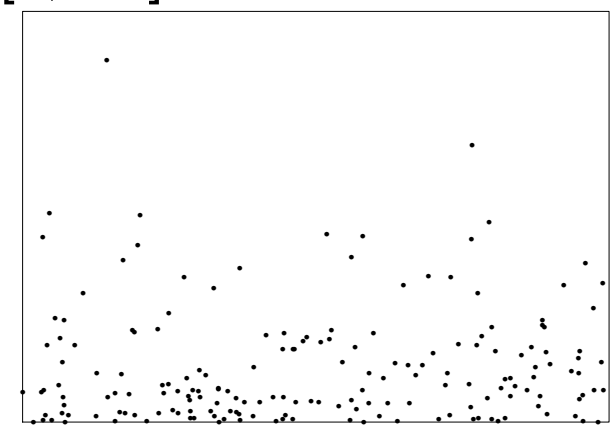
X is a Poisson process with intensity measure μ if for any bounded region B with $\mu(B) > 0$:

1. $N(B) \sim \text{Poisson}(\mu(B))$
2. Given $N(B)$, points in $\mathbf{X} \cap B$ i.i.d. with density $\propto \rho(u)$, $u \in B$

$B = [0, 1] \times [0, 0.7]$:



Homogeneous: $\rho = 150/0.7$



Inhomogeneous: $\rho(x, y) \propto \mathrm{e}^{-10.6y}$

Independence properties of Poisson process

1. if A and B are disjoint then $N(A)$ and $N(B)$ independent
2. - this implies $\mathbb{C}\mathrm{ov}[N(A), N(B)] = 0$ if $A \cap B = \emptyset$
3. - which in turn implies $g(u, v) = 1$ for a Poisson process

Inhomogeneous Poisson process with covariates

Log linear intensity function

$$\rho_\beta(u) = \exp(z(u)\beta^\mathsf{T}), \quad z(u) = (1, z_{\mathsf{elev}}(u), z_{\mathsf{grad}}(u))$$

Inhomogeneous Poisson process with covariates

Log linear intensity function

$$\rho_\beta(u) = \exp(z(u)\beta^\mathsf{T}), \quad z(u) = (1, z_{\text{elev}}(u), z_{\text{grad}}(u))$$

Consider indicators $N_i = \mathbf{1}[\mathbf{X} \cap C_i \neq \emptyset]$ of occurrence of points in disjoint C_i ($W = \cup C_i$) where $P(N_i = 1) \approx \rho_\beta(u_i)|C_i|$, $u_i \in C_i$

Inhomogeneous Poisson process with covariates

Log linear intensity function

$$\rho_\beta(u) = \exp(z(u)\beta^{\mathsf{T}}), \quad z(u) = (1, z_{\text{elev}}(u), z_{\text{grad}}(u))$$

Consider indicators $N_i = \mathbf{1}[\mathbf{X} \cap C_i \neq \emptyset]$ of occurrence of points in disjoint C_i ($W = \cup C_i$) where $P(N_i = 1) \approx \rho_\beta(u_i)|C_i|$, $u_i \in C_i$

Limit ($|C_i| \to 0$) of likelihood ratios

$$\prod_{i=1}^{n} \frac{(\rho_\beta(u_i)|C_i|)^{N_i}(1 - \rho_\beta(u_i)|C_i|)^{1-N_i}}{(1|C_i|)^{N_i}(1 - 1|C_i|)^{1-N_i}} \equiv \prod_{i=1}^{n} \frac{\rho_\beta(u_i)^{N_i}(1 - \rho_\beta(u_i)|C_i|)^{1-N_i}}{(1 - 1|C_i|)^{1-N_i}}$$

is

$$L(\beta) = \Big[\prod_{u \in \mathbf{X} \cap W} \rho_\beta(u) \Big] \exp(|W| - \int_W \rho_\beta(u) \mathrm{d}u)$$

This is the Poisson likelihood function.

Maximum likelihood parameter estimate

Score function:

$$s(\beta) = \frac{\mathrm{d}}{\mathrm{d}\beta} \log L(\beta) = \sum_{u \in \mathbf{X} \cap W} z(u) - \int_W z(u) \rho_\beta(u) \mathrm{d}u$$

Maximum likelihood estimate $\hat{\beta}$ maximizes $L(\beta)$. I.e. solution of

$$s(\beta) = 0.$$

Note by Campbell $s(\beta)$ unbiased:

$$\mathbb{E} s(\beta) = 0.$$

Observed information ($p \times p$ matrix):

$$I(\beta) = -\frac{\mathrm{d}}{\mathrm{d}\beta^\mathsf{T}} s(\beta) = \int_W z(u)^\mathsf{T} z(u) \rho_\beta(u) \mathrm{d}u$$

Unique maximum/root if $I(\beta)$ positive definite.

By Campbell formulae

$$\mathbb{V}\mathrm{ar}\, u(\beta) = I(\beta)$$

and according to standard asymptotic results for MLE (β^* 'true' value)

$$\hat{\beta} \approx N(\beta^*, I(\beta^*)^{-1})$$

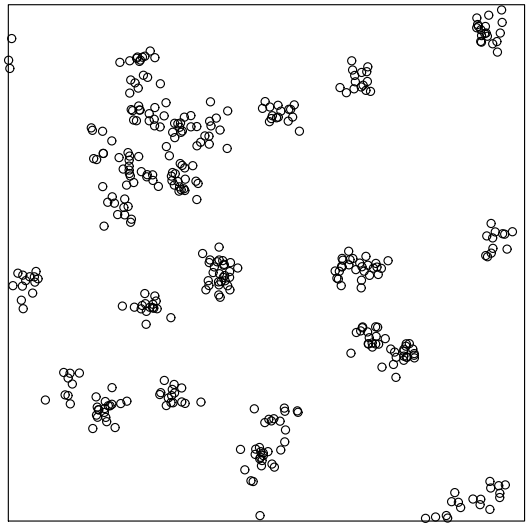
'n' (number of observations) tends to infinity ?

Possibilities: increasing observation window or increasing intensity

Problem: Poisson process does not fit rain forest data due to excess clustering (e.g. seed dispersal) !

Hence variance of $\hat{\beta}$ is underestimated by $I(\beta^*)^{-1}$ when a Poisson process is assumed.

Cluster process: Inhomogeneous Thomas process



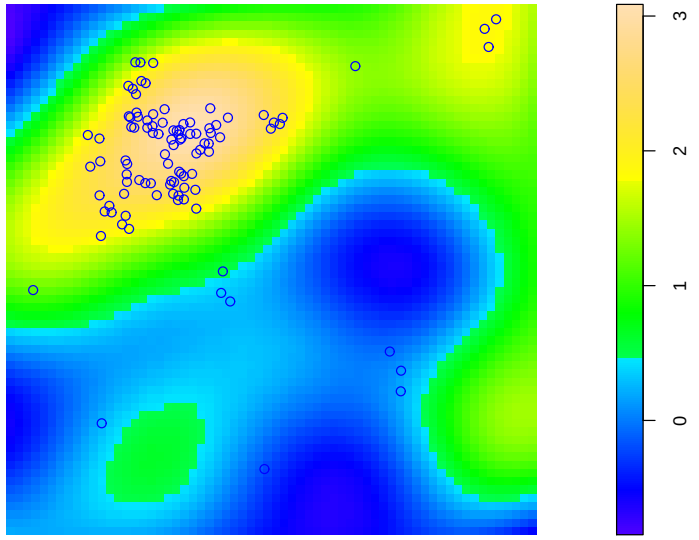
Parents stationary Poisson point process intensity κ

Poisson(α) number of offspring distributed around parents according to bivariate Gaussian density with std. dev. ω

Inhomogeneity: offspring survive according to probability

$$p(u) \propto \exp(z(u)\beta^{\mathsf{T}})$$

depending on covariates (independent thinning).



Intensity and pair correlation function for Thomas

$$\rho_\beta(u) = \exp[z(u)\beta^{\mathsf{T}}]$$

and

$$g(u,v) = 1 + (4\pi\omega^2)^{-d/2} \exp[-\{r/(2\omega)\}^2]/\kappa$$

Note $g(u,v) > 1$!

Parameter estimation: regression parameters

Likelihood function for inhomogeneous Thomas process is complicated.

Can instead use Poisson score $s(\beta)$ as an *estimating function* (Poisson likelihood now *composite likelihood*).

I.e. estimate $\hat{\beta}$ again solution of

$$s(\beta) = 0$$

But now larger variance of $s(\beta)$ due to positive correlation !

Exercises

1. Show that $s(\beta)$ is an unbiased estimating function (both in the Poisson case and for the inhom. Thomas).
2. For a Poisson process, show that $\mathbb{V}\mathrm{ar}\, s(\beta) = \mathbb{V}\mathrm{ar} \sum_{u \in \mathbf{X} \cap W} z(u) = I(\beta)$.
3. Compute the Godambe information for the estimating function $s(\beta)$ when $\mathbf{X}$ is a general point process with pair correlation function $g \neq 1$ (hint: use second-order Campbell formula). Compare with the case of a Poisson process ($g = 1$).

Quasi-likelihood for spatial point processes

Quasi-likelihood based on data vector Y was optimal linear transformation

$$D^{\mathsf{T}} V^{-1} R$$

of residual vector

$$R = Y - \mu(\beta)$$

Can we adapt quasi-likelihood to spatial point processes ?

What is residual in this case ?

Residual measure

For point process $\mathbf{X}$ and $A \subset \mathbb{R}^2$ *residual measure* is

$$R(A) = N(A) - \mathbb{E}N(A) = \sum_{u \in \mathbf{X}} 1[u \in A] - \int 1[u \in A]\rho(u;\beta)\mathrm{d}u$$

($N(A)$ number of points in A).

Residual measure

For point process $\mathbf{X}$ and $A \subset \mathbb{R}^2$ *residual measure* is

$$R(A) = N(A) - \mathbb{E}N(A) = \sum_{u \in \mathbf{X}} 1[u \in A] - \int 1[u \in A]\rho(u;\beta)\mathrm{d}u$$

($N(A)$ number of points in A).

In analogy with quasi-likelihood look for optimal linear transformation of the residual measure

$$e_f(\beta) = \int f(u;\beta)R(\mathrm{d}u) = \sum_{u \in \mathbf{X}} f(u;\beta) - \int f(u;\beta)\rho(u;\beta)\mathrm{d}u$$

where $f : \mathbb{R}^2 \to \mathbb{R}^p$ real vector-valued "weight" function.

Estimate $\hat{\beta}_f$ solves estimating equation

$$e_f(\beta) = 0$$

Remember: ϕ is optimal if

$$\mathbb{C}\mathrm{ov}[e_\phi, e_f] = S_f$$

for all f.

Remember: ϕ is optimal if

$$\mathbb{C}\mathrm{ov}[e_\phi, e_f] = S_f$$

for all f.

Using the Campbell formulae one can show that this is satisfied if ϕ solves following integral equation:

$$\phi(u;\beta) + \int_W t(u,v)\phi(v;\beta)\mathrm{d}v = \frac{\mathrm{d}}{\mathrm{d}\beta}\log\rho(u;\beta) \quad u \in W$$

where integral operator kernel is

$$t(u,v) = \rho(v;\beta)[g(u,v) - 1]$$

Poisson process case

Poisson process case: $g(u, v) = 1$ so integral equation simplifies:

$$\phi(u) + \int_W \rho(v; \beta)[g(u, v) - 1]\phi(v)\mathrm{d}v = \frac{\mathrm{d}}{\mathrm{d}\beta} \log \rho(u; \beta) \Rightarrow$$

$$\phi(u) = \frac{\mathrm{d}}{\mathrm{d}\beta} \log \rho(u; \beta) = \frac{\rho'(u; \beta)}{\rho(u; \beta)}$$

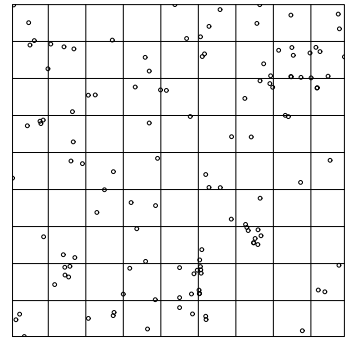
Hence resulting estimating function is

$$\sum_{u \in \mathbf{X} \cap W} \frac{\rho'(u; \beta)}{\rho(u; \beta)} - \int_W \rho'(u; \beta)\mathrm{d}u$$

which coincides with score of Poisson process log likelihood.

Quasi-likelihood

Integral equation approximated using Riemann sum dividing W into cells C_i with representative points u_i.



Resulting estimating function is *quasi-likelihood* score

$$D^{\mathsf{T}} V^{-1}[Y - \mu]$$

based on

$$Y = (Y_1, \ldots, Y_m)^{\mathsf{T}}, \quad Y_i = 1[\mathbf{X} \text{ has point in } C_i].$$

μ mean of Y:

$$\mu_i = \mathbb{E} Y_i = \rho(u_i; \beta)|C_i| \text{ and } D = \left[\mathrm{d}\mu(u_i)/\mathrm{d}\beta_j\right]_{ij}$$

V covariance of Y

$$V_{ij} = \mathbb{C}\mathrm{ov}[Y_i, Y_j] = \mu_i 1[i = j] + \mu_i \mu_j [g(u_i, u_j) - 1]$$

Useful matrix result

Assume A and B invertible.

$$B^{-1}-A^{-1} = A^{-1}(A-B)B^{-1}AA^{-1} = A^{-1}[(A-B)B^{-1}(B+A-B)]A^{-1}$$
$$= A^{-1}[A - B + (A - B)B^{-1}(A - B)]A^{-1}$$

Hence if $A - B$ is positive definite so is $B^{-1} - A^{-1}$.