Estimating functions and inhomogeneous point processes

Rasmus Waagepetersen
Department of Mathematics
Aalborg University
Denmark

April 17, 2024

Outline

Estimating equations and quasi-likelihood

Estimating functions for inhomogeneous spatial point processes

Composite information criteria 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)

$$D^{\mathsf{T}}[Y - \mu(\beta)] = 0 \tag{1}$$

where

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

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:

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

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

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

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 $e'(\theta^*)$

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 $S^{-1}(\theta^*)$ exists)

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

Thus

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

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 next slide).

Useful matrix result

Assume A and B invertible.

$$\begin{aligned} B^{-1}-A^{-1} = A^{-1}(A-B)B^{-1}AA^{-1} &= A^{-1}[(A-B)B^{-1}(B+A-B)]A^{-1} \\ &= A^{-1}[A-B+(A-B)B^{-1}(A-B)]A^{-1} \end{aligned}$$

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

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

Consider a class $\mathcal{E}$ of estimating functions. e^o is optimal within $\mathcal{E}$ if

$$\Sigma_{ee^o} = \mathbb{C}\mathrm{ov}[e, e^o] = S_e \tag{2}$$

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

The property (2) implies $\mathbb{V}\mathrm{ar} e^0 = S_{e^o} = S_{e^o}^{\mathsf{T}}$ and we obtain

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

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 (2) 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. calculate S and Σ and I for the non-linear least squares estimating function (1). Is the estimating function unbiased ?
2. Show that $\frac{\mathrm{d}}{\mathrm{d}\theta}\log(c(\theta)^{-1}) = \mathbb{E}_\theta t(Y)$ for the exponential family model on slide 5.
3. show results on slide 'Estimating functions and the likelihood score' (hint: use the rule for differentiation of a product to show the first result)

Exercises cntd.

4. (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

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

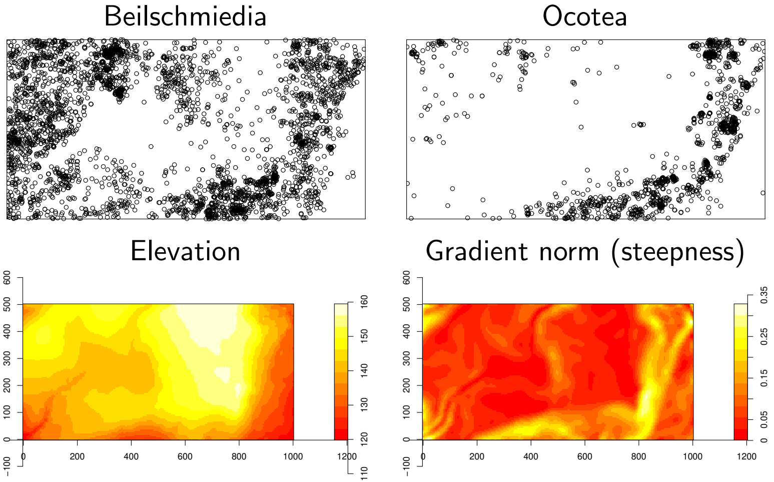
   where A $q \times n$ (all linear combinations of residual vector). Show that the optimal choice is $A = D^{\mathsf{T}} V^{-1}$.

   What is the Godambe information matrix ?
5. Check the proof on slide 14.

Now: inhomogeneous point processes.

Data example: tropical rain forest trees

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

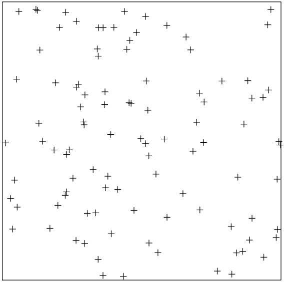


Sources of variation: elevation and gradient covariates *and* possible clustering/aggregation due to unobserved covariates and/or seed dispersal.

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 $N(A) = \#(\mathrm{X} \cap A)$.

Intensity of a spatial point process

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

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 $N(A) = \#(\mathrm{X} \cap A)$.

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 $N(A) = \#(\mathsf{X} \cap A)$.

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

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

Covariance of counts and pair correlation function

Pair correlation function

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

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:

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

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

The Poisson process

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

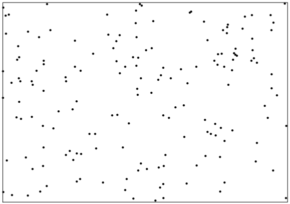
The Poisson process

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

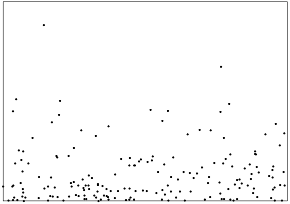
X is a Poisson process with intensity measure μ if for any bounded region B with $\mu(B) > 0$:

1. $N(B) \sim \text{Poisson}(\mu(B))$
2. Given $N(B)$, points in $X \cap B$ 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

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

Inhomogeneous Poisson process with covariates

Log linear intensity function

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

Consider indicators $N_i = 1[\mathrm{X} \cap C_i \neq \emptyset]$ 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

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

Consider indicators $N_i = 1[\mathrm{X} \cap C_i \neq \emptyset]$ 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

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

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):

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

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

By Campbell formulae

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

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

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

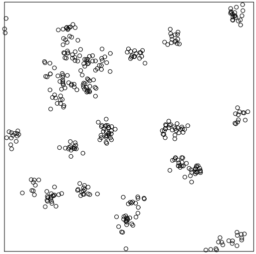
'n' (number of observations) tends to infinity ?

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



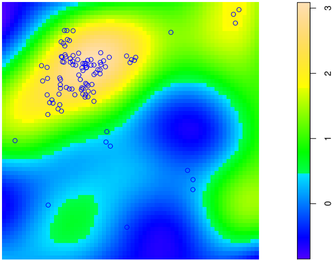
Parents stationary Poisson point process intensity κ

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

Inhomogeneity: offspring survive according to probability

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

depending on covariates (independent thinning).



Intensity and pair correlation function for Thomas

We can write Thomas process X as

$$X = \cup_{c \in C} X_c$$

where C stationary Poisson process of intensity κ and given C, the X_c are independent Poisson processes with intensity functions $p(u)\alpha k(u-c)$ where $k(\cdot)$ density of $N_2(0, \omega^2 I)$.

With $p(u) = \exp(z(u)^\mathsf{T}\beta)/M$ the intensity becomes

$$\rho(u) = \alpha\kappa \exp[z(u)^\mathsf{T}\beta]/M = \exp[\beta_0 + z(u)^\mathsf{T}\beta]$$

where $\exp(\beta_0) = \alpha\kappa/M$.

The pair correlation function becomes (for Thomas process in $\mathbb{R}^d$)

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

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 $\mathbb{V}\mathrm{ar} s(\beta) = \mathbb{V}\mathrm{ar} \sum_{u \in \mathrm{X} \cap W} z(u) = I(\beta)$.
3. Compute the inverse Godambe information for the estimating function $s(\beta)$ when 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 $(g = 1)$.
4. Verify the expressions for the intensity and pair correlation function of a Thomas process (slide 35).

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

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

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

Residual measure

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

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

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

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

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

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

$$\mathbb{C}\mathrm{ov}[e_\phi, e_f] = S_f \tag{3}$$

for all f.

Remember: ϕ is optimal if

$$\mathbb{C}\mathrm{ov}[e_\phi, e_f] = S_f \tag{3}$$

for all f.

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

$$\phi(u;\beta) + \int_W t(u,v)\phi(v;\beta)\mathrm{d}v = \frac{\mathrm{d}}{\mathrm{d}\beta}\log\rho(u;\beta) \quad u \in W \tag{4}$$

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

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

which coincides with score of Poisson process log likelihood.

Details about Nyström method

Use Riemann sum dividing W into cells C_i with representative points u_i, $i = 1, \ldots, n$. Then we obtain linear equations

$$\phi(u_i;\beta) + \sum_{j=1}^{n} t(u_i, u_j)|C_j|\phi(u_j;\beta) = \frac{\mathrm{d}}{\mathrm{d}\beta}\log\rho(u_i;\beta) \quad i = 1, \ldots, n \tag{5}$$

which in matrix form become

$$(I + T)\bar{\phi} = [\frac{\mathrm{d}}{\mathrm{d}\beta}\log\rho(u_i;\beta)]_i$$

where $\bar{\phi} = (\phi(u_i))_i$ and $T_{ij} = t(u_i, u_j)|C_j|$.

Defining $\mu_i = \rho(u_i;\beta)|C_i|$, $M = \mathrm{diag}(\mu_1, \ldots, \mu_n)$, and $G = [G_{ij}]_{ij}$ with $G_{ij} = \mu_i\mu_j[g(u_i, v_j) - 1]$, this is equivalent to

$$(M + G)\bar{\phi} = M[\frac{\mathrm{d}}{\mathrm{d}\beta}\log\rho(u_i;\beta)]_i = D$$

where D is matrix of partial derivatives $\mathrm{d}\mu_i/\mathrm{d}\beta_j$.

Quasi-likelihood

Using solution

$$\bar{\phi} = (M + G)^{-1} = V^{-1} D$$

with $V = M + G$ the resulting approximated optimal estimating function becomes the *quasi-likelihood* score

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

where

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

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

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

Exercise

1. Show that (5) implies (3).

   Hint: start by evaluating (3) using the Campbell formulae

All models are wrong...

“All models are wrong but some are useful”

If any model we propose/select/estimate is wrong how can we talk of a ‘true’ parameter value, true model, optimal estimation method... ?

Approach:

- consider ‘least false’ model - i.e. model among a set of candidate models which is closest to the unknown true model
- consider ‘least false’ parameter value - i.e. parameter value that makes a given model closest to unknown true model

Kullback-Leibler divergence

Consider two densities f and g with same support and $X \sim f$. Then Kullback-Leibler divergence of g from f is

$$D_{KL}(f,g) = \int f(x) \log \frac{f(x)}{g(x)} \mathrm{d}x = -\mathbb{E} \log \frac{g(X)}{f(X)} = -\mathbb{E}[\log g(X) - \log f(X)]$$

By Jensen's inequality or just $\log(x) \leq x - 1$,

$$D_{KL}(f,g) \geq 0 \tag{6}$$

and "=" only if $f = g$ f-almost surely (Gibbs' inequality).

Suppose f represents true distribution of data and $g_1, \ldots, g_K$ are candidate models.

We may then declare g_l to be the least false model if

$$l = \operatorname*{argmin}_{k=1,\ldots,K} D_{KL}(f, g_k)$$

Similar, if the g_k are parametrized by some unknown parameter $\theta_k \in \Theta_k$ we may declare θ_k^* to be the least false parameter value for g_k if

$$\theta_k^* = \underset{\theta_k \in \Theta_k}{\operatorname{argmin}} D_{KL}(f, g(\cdot; \theta_k))$$

Case of composite likelihood for point process

Suppose X is a point process with true intensity function λ and ρ is some other intensity function.

Also let $l(\cdot;\lambda)$ and $l(\cdot;\rho)$ denote corresponding Poisson log density functions (first order composite likelihood functions)

Then we may define composite Kullback-Leibler divergence as

$$CD_{KL}(\lambda,\rho) = -\mathbb{E}[l(\mathsf{X};\rho) - l(\mathsf{X};\lambda)]$$

Again

$$CD_{KL}(\lambda,\rho) \geq 0 \tag{7}$$

and "=" only if $\lambda = \rho$ almost surely with respect to distribution of X (exercise).

Least false intensity function among $\rho_1,\ldots,\rho_K$ minimizes $CD_{KL}(\lambda,\rho_l)$.

For parametric model $\rho_k(\cdot;\theta_k)$, least false θ_k is

$$\theta_k^* = \underset{\theta_k \in \Theta_k}{\operatorname{argmin}} \, CD_{KL}(\lambda, \rho(\cdot;\theta_k))$$

Regression model for the intensity function

X spatial point process observed in window $W \subset \mathbb{R}^d$.

Popular log-linear model for the intensity function:

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

where $\mathrm{z}(u) = (z_1(u), \ldots, z_p(u))^\mathsf{T}$ covariate vector associated to spatial location u.

Model selection problem: which subset of covariates should be used ?

One approach is to use information criteria (AIC, BIC,....)

How to do this in case of a spatial point process ?

I got this question back in 2008 while I was in Spar Nord Bank :)

Notation: l index for collection of models M_l characterized by varying subsets $\mathsf{z}_l(u)$ of covariates and with parameter vectors β_l. I.e. $\mathsf{z}_l(u) = (z_j(u))_{u \in I_l}$, $I_l \subseteq \{1, \ldots, p\}$.

The log-likelihood for model M_l in case of a Poisson process is

$$l(\beta_l; \mathrm{X}) = \sum_{u \in \mathrm{X}} \mathsf{z}_l(u)^{\mathsf{T}} \beta_l - \int_W \rho(u; \beta_l) \mathrm{d}u$$

AIC:

$$-2l(\hat{\beta}_; \mathrm{X}) + 2p_l$$

Is this theoretically justified for a Poisson process ?

Moreover, we often use $l(\beta_l; \mathrm{X})$ as a kind of composite likelihood in case X is not a Poisson process.

Can we still use AIC or do we need to consider composite information criterion (CIC) ?

Bayesian information criterion

What about BIC:

$$-2l(\hat{\beta}_I; \mathrm{X}) + \log(n)p_I$$

What is n ? ("number of observations") ?

- 1 ?
- Number N of points in $\mathrm{X} \cap W$?
- Size of observation window $|W|$?
- Number of points used in quadrature scheme for approximation of likelihood ? (analogy to logistic regression)

Asymptotic results for misspecified model

'Least false β_I', β_I^*, minimizes Kullback-Leibler distance:

$$\beta_I^* = \underset{\beta_I}{\operatorname{argmin}}\, CD_{KL}(\rho(\cdot;\beta_I),\lambda) = \underset{\beta_I}{\operatorname{argmin}}\, \mathbb{E}[-l(\beta_I;\mathrm{X})]$$

Given (wrong) model M_I we can under reasonable conditions show that

$$\hat{\beta}_I - \beta_I^* \approx N(0, V)$$

That is, composite likelihood estimate will asympotically make the fitted model M_I least false.

The covariance matrix has the following expression:

$$S_I(\beta_I^*)^{-1}\Sigma_I S_I(\beta_I^*)^{-1}$$

where unfortunately Σ_I is not known...

Under reasonable conditions, $S_I(\beta_I^*)^{-1}\Sigma_I S_I(\beta_I^*)^{-1}$ is of the order $|W|^{-1}$!

Model selection

Choose model so that

$$C(\rho(\cdot;\beta_I^*)) = \mathbb{E}[-l(\beta_I^*;\mathrm{X})]$$

is minimal.

Issue: β_I^* unknown in practice since it depends on unknown $\lambda(\cdot)$.

Suggestion: given data X and resulting estimates $\hat{\beta}_I$, minimize

$$\mathbb{E}C(\rho(\cdot;\hat{\beta}_I))$$

over models M_I.

Note: $\mathbb{E}C(\rho(\cdot;\hat{\beta}_I)) = \mathbb{E}\mathbb{E}[-l(\hat{\beta}_I;\tilde{\mathrm{X}})|\mathrm{X}]$

Problem: both expectations unknown.

Estimation of $\mathbb{E}C(\rho(\cdot;\hat{\beta}_l))$

Suppose we have two independent copies of the point process X and $\tilde{\text{X}}$ and we obtain $\hat{\beta}_l$ from X.

Then

$$-l(\hat{\beta}_l, \tilde{\text{X}}) = -\sum_{u \in \tilde{\text{X}}} z_l(u)^{\mathsf{T}} \hat{\beta}_l + \int_W \rho(u; \hat{\beta}_l) \mathrm{d}u$$

would be an unbiased estimate of

$$\mathbb{E}C(\rho(\cdot;\hat{\beta}_l)) = \mathbb{E}\mathbb{E}[-l(\hat{\beta};\tilde{\text{X}})|\text{X}]$$

(similar to cross validation)

However, we only have the single realization X.

The observed likelihood

$$-\sum_{u \in \text{X}} z_l(u)^{\mathsf{T}} \hat{\beta}_l + \int_W \rho(u; \hat{\beta}_l) \mathrm{d}u$$

is a biased (too small) estimate due to overfitting.

Estimation of bias

We can approximate log likelihood using second-order Taylor expansion:

$$l(\hat{\beta}_I; \tilde{X}) \approx l(\beta_I^*; \tilde{X}) + \nabla l(\beta_I^*; \tilde{X})^{\mathsf{T}}(\hat{\beta}_I - \beta_I^*) - \frac{1}{2}(\hat{\beta}_I - \beta_I^*)^{\mathsf{T}} S(\beta_I^*)(\hat{\beta}_I - \beta_I^*))$$

and (observed likelihood)

$$l(\hat{\beta}_I; X) \approx l(\beta_I^*; X) + \nabla l(\beta_I^*; X)^{\mathsf{T}}(\hat{\beta}_I - \beta_I^*) - \frac{1}{2}(\hat{\beta}_I - \beta_I^*)^{\mathsf{T}} S(\beta_I^*)(\hat{\beta}_I - \beta_I^*)$$

Here $S(\beta)$ is sensitivity

$$S(\beta) = \int_W z_I(u)^{\mathsf{T}} z_I(u) \rho(u; \beta_I) \mathrm{d}u$$

Bias (recall first Bartlett identity $\mathbb{E}\nabla l(\beta_I^*; \tilde{X})^{\mathsf{T}} = 0$):

$$\mathbb{E}l(\hat{\beta}_I; \tilde{X}) - \mathbb{E}l(\hat{\beta}_I; X) = -\mathbb{E}\nabla l(\beta_I^*; X)^{\mathsf{T}}(\hat{\beta}_I - \beta_I^*) + \mathbb{E}[o_P(1)]$$

Using first order Taylor

$$\nabla l(\beta_l^*; \mathsf{X}) \approx S(\hat{\beta}_l)(\hat{\beta}_l - \beta_l^*) \Rightarrow (\hat{\beta}_l - \beta_l^*) \approx S(\beta_l^*)^{-1} \nabla l(\beta_l^*; \mathsf{X})$$

we get

$$\mathbb{E}\nabla l(\beta_l^*; \mathsf{X})^\mathsf{T}(\hat{\beta}_l - \beta_l^*) = \mathbb{E}\nabla l(\beta_l^*; \mathsf{X})^\mathsf{T} S(\beta_l^*)^{-1} \nabla l(\beta_l^*; \mathsf{X}) + \mathbb{E} o_P(1)$$
$$= \text{trace}\left[S(\beta_l^*)^{-1}\Sigma_l\right] + \mathbb{E} o_P(1)$$

where

$$\Sigma_l = \mathbb{V}\mathrm{ar}\nabla l(\beta_l^*; \mathsf{X})$$

The previous expansions work when we have

$$\hat{\beta}_l - \beta_l^* = O_P(|W|^{-1/2})$$

'consistency wrt least false parameter value under M_l'

As mentioned before we can obtain this consistency for wide class of point processes (including Cox and Cluster)

To obtain $\mathbb{E} o_P(1) = o(1)$ we need technical condition of uniform integrability. Often ignored in literature.

What about AIC ?

Suppose X is a Poisson process and M_l is the true model. Then by standard Bartlett identity

$$\Sigma_l = S_n(\beta_l^*)$$

and

$$\text{trace}\Sigma_l S_n(\beta_l^*)^{-1} = \text{trace} I_{p_l} = p_l = \text{length}\beta_l$$

This gives AIC criterion for model M_l !

In general we need to estimate (Takeuchi) bias correction

$$\text{trace} S(\beta_l^*)^{-1}\Sigma_l$$

Suggestion so far: estimate $S(\beta_l^*)$ by $S(\hat{\beta}_l)$

Regarding Σ_l:

$$\Sigma_l = \mathbb{V}\mathrm{ar}\nabla l(\beta_l^*)$$
$$= \int_W \mathsf{z}_l(u)^\mathsf{T}\mathsf{z}_l(u)\lambda(u)\mathrm{d}u + \int_{W^2} \mathsf{z}_l(u)^\mathsf{T}\mathsf{z}_l(v)\lambda(u)\lambda(v)[g(u,v)-1]\mathrm{d}u\mathrm{d}v$$

We approximate $\lambda(u) \approx \rho(u;\hat{\beta}_l)$ and obtain

$$\mathrm{trace}\Sigma_l S(\beta_l^*)^{-1} \approx p_l + \mathrm{trace}[T(\hat{\beta}_l)S(\beta_l^*)^{-1}]$$

where

$$T(\hat{\beta}_l) = \int_{W^2} \mathsf{z}_l(u)^\mathsf{T}\mathsf{z}_l(v)\rho(u;\hat{\beta}_l)\rho(v;\hat{\beta}_l)[\hat{g}(u-v)-1]\mathrm{d}u\mathrm{d}v$$

These quantities and estimate $\hat{g}$ can be obtained from output of `spatstat` procedure `kppm`.

Bayesian information Criterion

Very different type of reasoning compared to AIC.

Impose prior $P(M = M_l)$ for model M and prior $p(\beta_l|M_l)$ for β_l given $M = M_l$.

Given M_l and β_l assume X Poisson process with density $f(\mathrm{x}|\beta_l, M_l)$.

Suppose uniform prior on models M_l. Then posterior of M is

$$P(M = M_l|\mathrm{X}) \propto P(\mathrm{X}|M_l)P(M_l) \propto P(\mathrm{X}|M_l)$$
$$= \int_{\mathbb{R}^{p_l}} f(\mathrm{X}|\beta_l, M_l)p(\beta_l|M_l)\mathrm{d}\beta_l$$

Using a Laplace approximation of the integral one obtains

$$\log P(\mathrm{X}|M_l) = l(\hat{\beta}_l; \mathrm{X}) - \frac{p_l}{2}\log(\mu) + O(1)$$

where μ is marginal mean of number of points in X.

Neglecting $O(1)$ terms and estimating $\mu \approx N$ where N is number of points in X we obtain

$$\mathrm{BIC}(M_l) = -2l(\hat{\beta}_l; \mathrm{X}) + \log(N)p_l$$

I.e. ‘number of observations’ is number of points !

Comparison with AIC/CIC:

- In Bayesian setting, we by assumption use the true model. No mention of ‘least false parameter value’.
- $\hat{\beta}_l$ convenient starting point for second order Taylor expansion underlying Laplace approximation.
- For technical reasons need *almost sure convergence* of $\hat{\beta}_l$ to fixed value β_l^*
- Asymptotics underlying Laplace approximation deterministic since conditioning on X.

Simulation studies

BIC: use of window size $|W|$ or number of points in quadrature approximation of likelihood useless.

AIC vs BIC (Poisson process): AIC tends to choose too complex models

CIC (cluster process): for cluster point processes CIC works better than AIC and BIC that both choose too complex models

Exercises

1. show (6) and (7).
2. Show that if sensitivity $S(\beta_I)$ is positive definite then least false parameter value β_I^* is well-defined (exists and is unique)
3. Show $\mathbb{E}\nabla I(\beta_I^*;\mathsf{X})^\mathsf{T} S(\beta_I^*)^{-1}\nabla I(\beta_I^*;\mathsf{X}) = \mathsf{trace}\left[S(\beta_I^*)^{-1}\Sigma_I\right]$