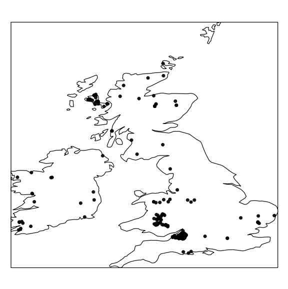
Non-parametric intensity estimation of spatial point processes by random forests

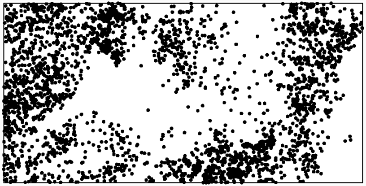
Christophe A. N. Biscio
joint work with Frédéric Lavancier (CREST, ENSAI, Rennes)

Motivation I

Let X a spatial point process observed on $W \subset \mathbb{R}^d$.



Brown trouts in the UK



Trees in a tropical rain forest

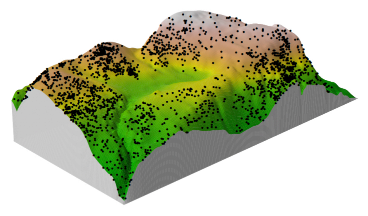
Aim: Estimate the intensity $\lambda(x)$, $x \in \mathbb{R}^d$, where

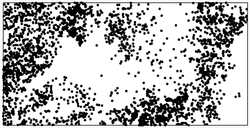
$$\lambda(x) \approx \mathbb{P}(X \text{ has a point at } x).$$

Formally: $\forall A \subset \mathbb{R}^d,\ \mathbb{E}(X(A)) = \int_A \lambda(x)dx.$

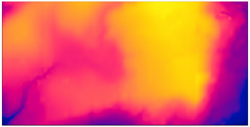
Motivation II

Sometimes we observe several covariates $z : \mathbb{R}^d \to \mathbb{R}^p$ on W.

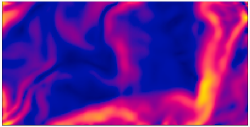




Trees



Elevation



Slope

In which case, we assume $\lambda(x) = f(z(x))$.

Usual methods

Usual methods to estimate $\lambda(x) = f(z(x))$:

Without covariates ($z(x) = x$): kernel smoothing, i.e.

$$\widehat{\lambda}(x) = \sum_{u \in X \cap W} k_h(\|x - u\|).$$

With covariates:

- parametric approach: assume $\log \lambda(x) = \theta' z(x)$ and get $\hat{\theta}$.
- non-parametric approach : assume $\lambda(x) = f(z(x))$ and

$$\widehat{\lambda}(x) = \sum_{u \in X \cap W} k_h(\|z(x) - z(u)\|).$$

Standard regression random forest in a nutshell

Aim: Predict an output y given covariates $x \in \mathbb{R}^p$.

Data: input/output, (x_i, y_i), $i = 1, \ldots n$.

Standard regression random forest in a nutshell

Aim: Predict an output y given covariates $x \in \mathbb{R}^p$.

Data: input/output, (x_i, y_i), $i = 1, \dots n$.

Regression tree:

- Build a partition $\pi = \{I_j\}$ of the covariates' space,
- Prediction for a new $\tilde{x} \in I_{j_0}$: average all y_i's such that $x_i \in I_{j_0}$.

Standard regression random forest in a nutshell

Aim: Predict an output y given covariates $x \in \mathbb{R}^p$.

Data: input/output, (x_i, y_i), $i = 1, \dots n$.

Regression tree:

- Build a partition $\pi = \{I_j\}$ of the covariates' space,
- Prediction for a new $\tilde{x} \in I_{j_0}$: average all y_i's such that $x_i \in I_{j_0}$.

Random Forest: Build M "diverse" trees :

- bootstrap the data before building each tree,
- build the partition with randomly selected covariates.

The random forest predictor is an average of the M tree predictors.

Standard regression random Forest in a nutshell

Advantages:

- Applies to a wide range of prediction problems
- Several ”success stories”
- Built-in selection of hyperparameters by ”Out-Of-Bag” (OOB).
- Assess importance of covariates: ”Variable Importance” (VIP).

But: Challenging theory (and other flaws not covered here)

Standard regression random Forest in a nutshell

Advantages:

- Applies to a wide range of prediction problems
- Several ”success stories”
- Built-in selection of hyperparameters by ”Out-Of-Bag” (OOB).
- Assess importance of covariates: ”Variable Importance” (VIP).

But: Challenging theory (and other flaws not covered here)

One exception: if the partitions are built independently of the data.

- We then say that the RF is a **purely random forest**.
- (Rarely the case in practice)

☞ J. Mourtada, S. Gaïffas and E. Scornet. *Minimax optimal rates for Mondrian trees and forests.* AOS (2020)

☞ E. O’Reilly and N. Mai Tran. *Minimax Rates for High-Dimensional Random Tessellation Forests.* JMLR (2024).

Random forest approach

Setting: we observe the point process X on W and $z(x)$ for all $x \in W$

$\longrightarrow$ We want to estimate $\lambda(x) = f(z(x))$.

Random forest approach

Setting: we observe the point process X on W and $z(x)$ for all $x \in W$

$\longrightarrow$ We want to estimate $\lambda(x) = f(z(x))$.

We first need an “intensity tree” estimator.

- Let $\pi = \{I_j\}$ be a finite partition of $z(W)$.
- Let $A_j = z^{-1}(I_j) \cap W$.

Thus

$$z(W) = \bigsqcup I_j \qquad \text{and} \qquad W = \bigsqcup A_j.$$

Random forest approach

Setting: we observe the point process X on W and $z(x)$ for all $x \in W$

$\longrightarrow$ We want to estimate $\lambda(x) = f(z(x))$.

We first need an "intensity tree" estimator.

- Let $\pi = \{I_j\}$ be a finite partition of $z(W)$.
- Let $A_j = z^{-1}(I_j) \cap W$.

Thus

$$z(W) = \bigsqcup I_j \qquad \text{and} \qquad W = \bigsqcup A_j.$$

Let $x \in W$ and denote $A(x)$: the cell A_j that contains x.

Random forest approach

Setting: we observe the point process X on W and $z(x)$ for all $x \in W$

$\longrightarrow$ We want to estimate $\lambda(x) = f(z(x))$.

We first need an "intensity tree" estimator.

- Let $\pi = \{I_j\}$ be a finite partition of $z(W)$.
- Let $A_j = z^{-1}(I_j) \cap W$.

Thus

$$z(W) = \bigsqcup I_j \qquad \text{and} \qquad W = \bigsqcup A_j.$$

Let $x \in W$ and denote $A(x)$: the cell A_j that contains x.

Then we define an intensity tree estimate by

$$\widehat{\lambda}^{(1)}(x) = \frac{X\left(A(x)\right)}{|A(x)|} = \frac{\text{number of points in the cell}}{\text{volume of the cell}}.$$

Random forest approach

Consider M different partition of $z(W)$.

Denote the corresponding intensity tree estimators by $\widehat{\lambda}^{(1)}, \dots, \widehat{\lambda}^{(M)}$.

We define the **random forest intensity estimator** by

$$\widehat{\lambda}^{(RF)}(x) = \frac{1}{M} \sum_{i=1}^{M} \widehat{\lambda}^{(i)}(x).$$

How can we generate partitions of $z(W)$?

We split the presentation in two cases:

1. **No covariate** : only the spatial coordinates are available
   Equivalently $z(x) = x$, so that $z(W) = W$

2. **With covariates**.

1st Case – No covariate

$z(x) = x$, $z(W) = W$

Tessellations

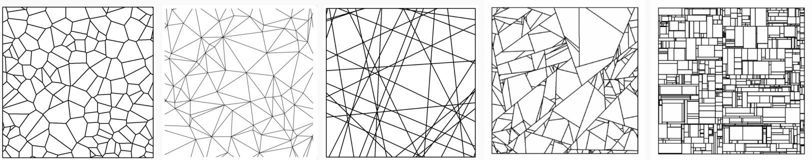
A partition of $W \iff$ A tessellation on W.

Tessellations

A partition of $W \iff$ A tessellation on W.

We consider independent random tessellations, that can be:

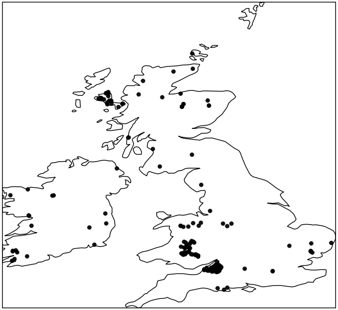
- Poisson Voronoï
- Poisson Delaunay
- Poisson hyperplane
- STIT tessellations (including the Mondrian process)

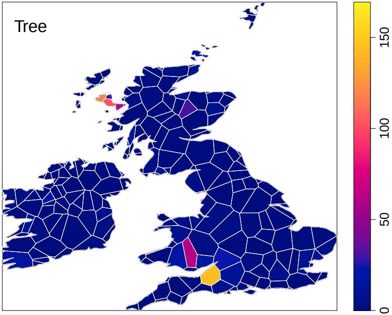


These tessellations depend on an intensity parameter h^{-d}.

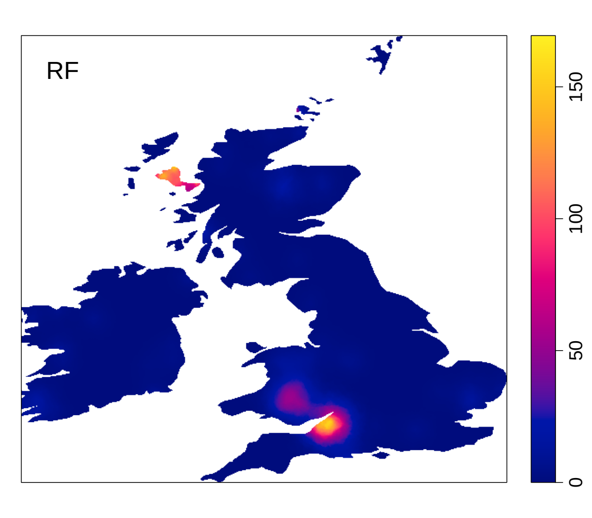
Remark: The RF is a genuine *pure* RF.

Example – One tree

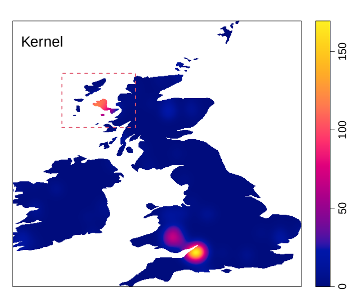


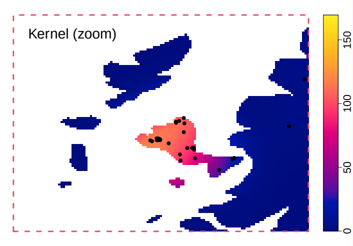


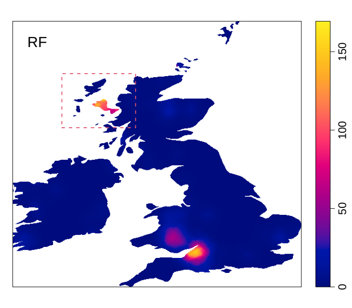
Example – RF (100 trees)

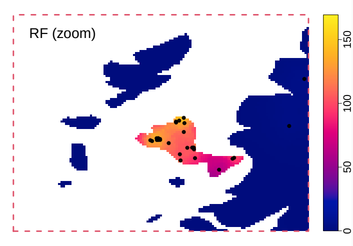


Example – Kernel smoothing versus RF









2st Case – With covariates

Tree

We need a partition of $z(W) = z_1(W) \times \cdots \times z_p(W)$ where $z_i(W) \subset \mathbb{R}$.

Tree

We need a partition of $z(W) = z_1(W) \times \cdots \times z_p(W)$ where $z_i(W) \subset \mathbb{R}$.

- We can generate a Voronoï tessellation of $z(W)$, as above.
  Then the RF will be a purely RF.

Tree

We need a partition of $z(W) = z_1(W) \times \cdots \times z_p(W)$ where $z_i(W) \subset \mathbb{R}$.

- We can generate a Voronoï tessellation of $z(W)$, as above. Then the RF will be a purely RF.

- Or, in the spirit of standard RF, we can construct an “optimal” tessellation, in relation with the output (here, the intensity).

Tree, in the spirit of standard RF

First step: for $i = 1, \dots, p$,

- Let $m_i = \mathrm{Median}(z_i(W))$
- Consider the possible split:

$$L_i = \{z_i(x) < m_i\} \text{ and } R_i = \{z_i(x) \geq m_i\}.$$

Tree, in the spirit of standard RF

First step: for $i = 1, \dots, p$,

- Let $m_i = \mathrm{Median}(z_i(W))$
- Consider the possible split:

$$L_i = \{z_i(x) < m_i\} \text{ and } R_i = \{z_i(x) \geq m_i\}.$$

Choose the best split out of these p possible splits.

$\longrightarrow$ The score of each split $L \cup R$ is based on the Poisson likelihood:

$$n_L \log\left(\frac{n_L - 1}{|L|}\right) + n_R \log\left(\frac{n_R - 1}{|R|}\right).$$

Tree, in the spirit of standard RF

First step: for $i = 1, \dots, p$,

- Let $m_i = \mathrm{Median}(z_i(W))$
- Consider the possible split:

$$L_i = \{z_i(x) < m_i\} \text{ and } R_i = \{z_i(x) \geq m_i\}.$$

Choose the best split out of these p possible splits.

$\longrightarrow$ The score of each split $L \cup R$ is based on the Poisson likelihood:

$$n_L \log\left(\frac{n_L - 1}{|L|}\right) + n_R \log\left(\frac{n_R - 1}{|R|}\right).$$

And so on, until a stopping criterion.

$\longrightarrow$ We choose a minimal number of points per cell ($minpts$).

Tree to Forest

To build the forest, consider M “diverse” trees, by

Tree to Forest

To build the forest, consider M “diverse” trees, by

Resampling: Each tree is based on a bootstrapped version of X

Tree to Forest

To build the forest, consider M “diverse” trees, by

Resampling: Each tree is based on a bootstrapped version of X

Pick variables: at each node, $mtry$ covariates are used, at random.

Tree to Forest

To build the forest, consider M “diverse” trees, by

Resampling: Each tree is based on a bootstrapped version of X

Pick variables: at each node, $mtry$ covariates are used, at random.

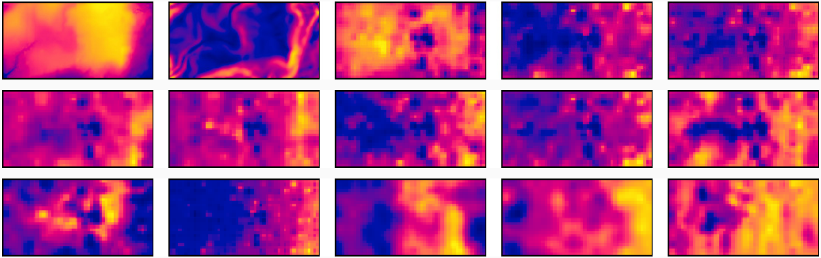
Like for standard RF:

- **Out-of-Bags cross-validation** (based on the Poisson likelihood score) is available.
- We can also compute the **VIP (variable importance)** of each variable.

Simulation Study

Simulation Study

$p = 15$ covariates: $z = (z_1, \ldots, z_{15})$.

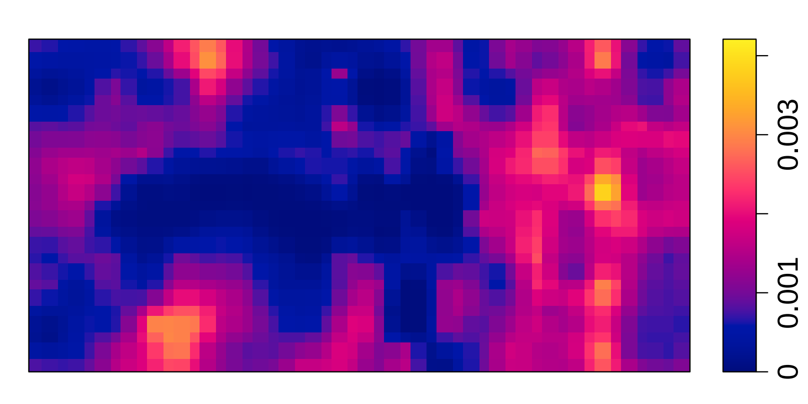


We simulate an inhomogeneous Poisson point process with intensity:

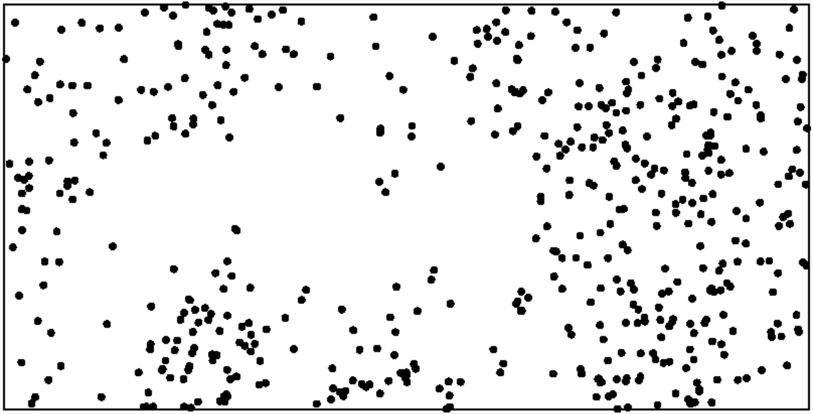
$$\lambda(x) = f(z_{10}(x))$$

with 500 points in average.

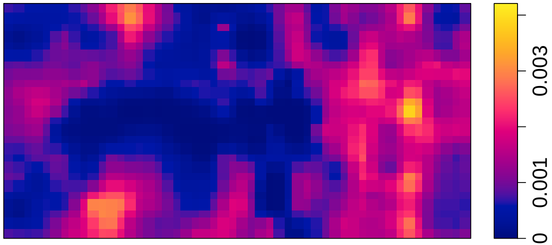
True intensity

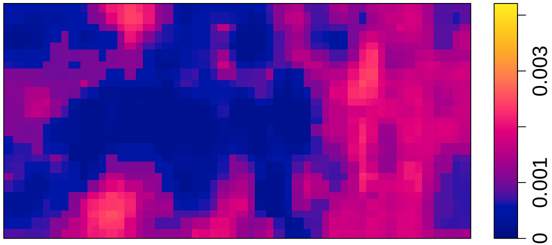


Realisation

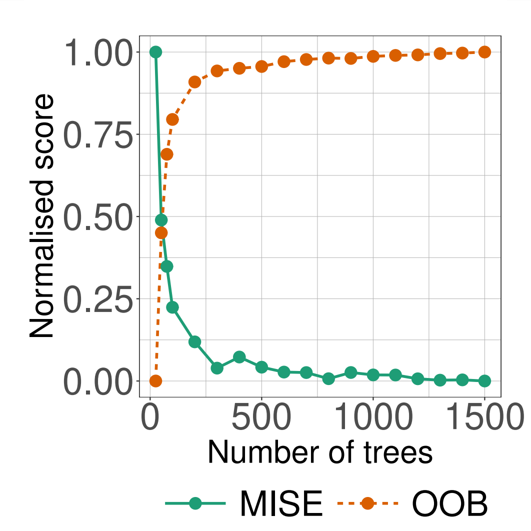


True intensity vs Random Forest estimate

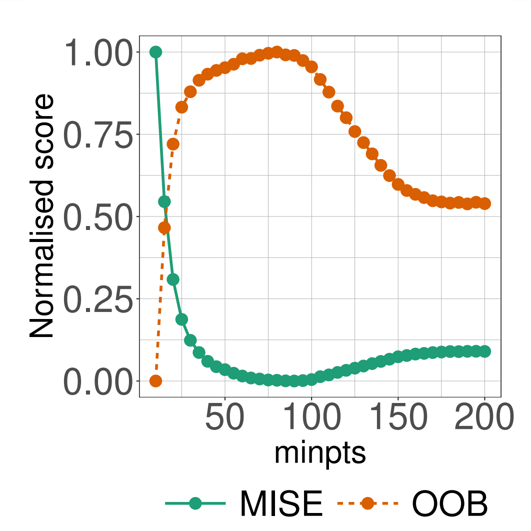




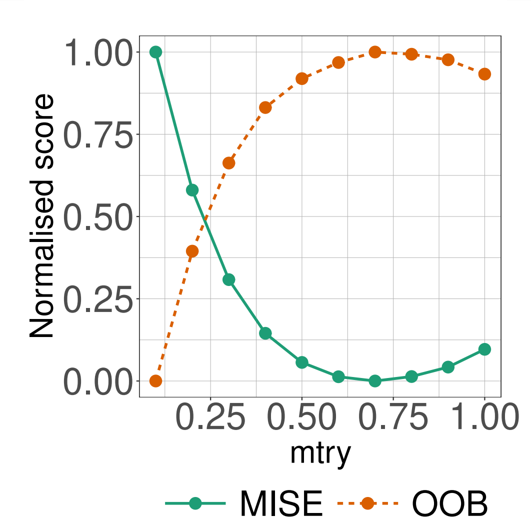
Choosing the number of trees



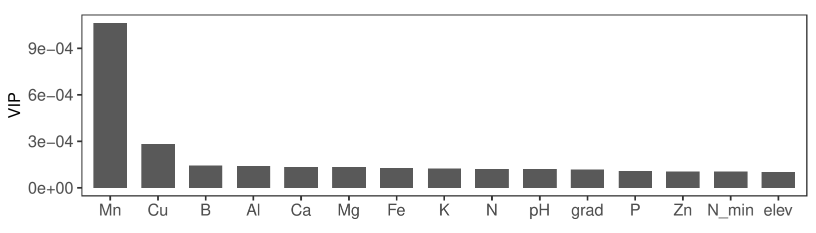
Choosing $minpts$



Choosing $mtry$



VIP



Mn is clearly detected as the most important one.

Summary of the methodology

Benefits:

- Works with any window shape (possibly not connected)
- Works with high number of covariates
- OOB cross-validation available
- VIP available

Flaws:

- Hyperparameters to choose (M, $minpts$, $mtry$)
- VIP sensitive to correlation between covariates
- Can be computationally involved
- Theory more involved than for purely RF

Some theory

Some theory

We want to study the performance of $\hat{\lambda}^{(RF)}(x)$ based on a single realisation of X in W.

Questions:

Some theory

We want to study the performance of $\hat{\lambda}^{(RF)}(x)$ based on a single realisation of X in W.

Questions:

1. What asymptotic framework do we consider ?

Some theory

We want to study the performance of $\hat{\lambda}^{(RF)}(x)$ based on a single realisation of X in W.

Questions:

1. What asymptotic framework do we consider ?
2. What point process models do we consider ?

Some theory

We want to study the performance of $\hat{\lambda}^{(RF)}(x)$ based on a single realisation of X in W.

Questions:

1. What asymptotic framework do we consider ?
2. What point process models do we consider ?
3. Is the procedure consistent ? minimax ?

Some theory

We want to study the performance of $\hat{\lambda}^{(RF)}(x)$ based on a single realisation of X in W.

Questions:

1. What asymptotic framework do we consider ?
2. What point process models do we consider ?
3. Is the procedure consistent ? minimax ?
4. What is the interest to leverage on covariates ?

Some theory

We want to study the performance of $\hat{\lambda}^{(RF)}(x)$ based on a single realisation of X in W.

Questions:

1. What asymptotic framework do we consider ?
2. What point process models do we consider ?
3. Is the procedure consistent ? minimax ?
4. What is the interest to leverage on covariates ?
5. What is the advantage of an RF over a single tree ?

Some theory

We want to study the performance of $\hat{\lambda}^{(RF)}(x)$ based on a single realisation of X in W.

Questions:

1. What asymptotic framework do we consider ?
2. What point process models do we consider ?
3. Is the procedure consistent ? minimax ?
4. What is the interest to leverage on covariates ?
5. What is the advantage of an RF over a single tree ?

We will assume that our RF are purely random forests.

1. The asymptotic regime

Setting:

$X = X_n$ is observed on W_n with intensity $\lambda_n = a_n \lambda$ with $a_n > 0$.

$\longrightarrow$ We want to estimate λ.

1. The asymptotic regime

Setting:

$X = X_n$ is observed on W_n with intensity $\lambda_n = a_n\lambda$ with $a_n > 0$.

$\longrightarrow$ We want to estimate λ.

Remark:

$$\mathbb{E}(X_n(W_n)) = \int_{W_n} \lambda_n(x)dx = a_n \int_{W_n} \lambda(x)dx \asymp a_n|W_n|.$$

Increasing the number of observations means $a_n|W_n| \to \infty$.

1. The asymptotic regime

Setting:

$X = X_n$ is observed on W_n with intensity $\lambda_n = a_n \lambda$ with $a_n > 0$.

$\longrightarrow$ We want to estimate λ.

Remark:

$$\mathbb{E}(X_n(W_n)) = \int_{W_n} \lambda_n(x)dx = a_n \int_{W_n} \lambda(x)dx \asymp a_n|W_n|.$$

Increasing the number of observations means $a_n|W_n| \to \infty$.

Different possible asymptotic regimes:

- *Infill*: $W_n = W$ is fixed but $a_n \to \infty$
- *Increasing domain*: $a_n = 1$ but $|W_n| \to \infty$
- *Intermediate regimes*: $a_n \to \infty$ and $|W_n| \to \infty$.

2. Point process models

Concerning the dependence structure of X_n, we assume that

$$\forall n, \forall A \subset W_n, \quad a_n \int_{A^2} |g_n(x,y) - 1| dxdy \leq c|A|, \tag{1}$$

where g_n is the pair correlation function of X_n.

2. Point process models

Concerning the dependence structure of X_n, we assume that

$$\forall n, \forall A \subset W_n, \quad a_n \int_{A^2} |g_n(x,y) - 1| dx dy \leq c|A|, \tag{1}$$

where g_n is the pair correlation function of X_n.

Typically, if for a certain underlying pcf g,

$$g_n(x,y) = g(a_n x, a_n y) \quad \text{or} \quad g_n(x,y) - 1 = \frac{1}{a_n}(g(x,y) - 1),$$

then (1) is ok whenever $\sup_y \int_{\mathbb{R}^d} |g(x,y) - 1| dx < \infty$.

2. Point process models

Concerning the dependence structure of X_n, we assume that

$$\forall n, \forall A \subset W_n, \quad a_n \int_{A^2} |g_n(x,y) - 1| dxdy \leq c|A|, \tag{1}$$

where g_n is the pair correlation function of X_n.

Typically, if for a certain underlying pcf g,

$$g_n(x,y) = g(a_n x, a_n y) \quad \text{or} \quad g_n(x,y) - 1 = \frac{1}{a_n}(g(x,y) - 1),$$

then (1) is ok whenever $\sup_y \int_{\mathbb{R}^d} |g(x,y) - 1| dx < \infty$.

This is a mild assumption satisfied for most usual models:

- Inhomogeneous Poisson point process,
- Neyman-Scott point process,
- LGCP with suitable mean and covariance functions,
- Matern hardcore point process (type I and II),
- Standard DPPs (Gaussian, Ginibre,...).

3. Consistency

Assume $\lambda(x) = f(z(x))$ where f is continuous at $z(x)$ and let

- $z(W_n) = \bigsqcup I_{n,j}$
- $I_n(x)$ = the cell $I_{n,j}$ that contains $z(x)$
- $A_n(x) = z^{-1}(I_n(x)) \cap W_n$

Theorem

For a **purely** *RF intensity estimator, if*

(1) $\operatorname{diam}(I_n(x)) \to 0$ *in probability,*

(2) $\mathbb{E}\left(1/(a_n|A_n(x)|)\right) \to 0$,

Then $\mathbb{E}\left[\left(\hat{\lambda}^{(RF)}(x) - \lambda(x)\right)^2\right] \to 0.$

(1) : $I_n(x)$ must concentrate around $z(x)$ $\;(bias \to 0)$

(2) : number of points in $A_n(x)$ must tend to infinity $\;(variance \to 0)$

3. Consistency: the case without covariate

When are the assumptions satisfied ?

$$(1)\ \mathrm{diam}(I_n(x)) \to 0 \quad \text{and} \quad (2)\ \mathbb{E}\left(1/(a_n|A_n(x)|)\right) \to 0.$$

3. Consistency: the case without covariate

When are the assumptions satisfied ?

$$(1)\ \mathrm{diam}(I_n(x)) \to 0 \quad \text{and} \quad (2)\ \mathbb{E}\left(1/(a_n|A_n(x)|)\right) \to 0.$$

Without covariate: $z(x) = x$ and $I_n(x) = A_n(x)$

For a regular tessellation of W_n (say Voronoï) with intensity h_n^{-d},

$A_n(x) = I_n(x)$ is the *zero cell* of the tessellation and we have:

$$\mathrm{diam}(I_n(x)) = O(h_n) \quad \text{and} \quad \mathbb{E}\left(1/|A_n(x)|\right) = h_n^{-d}.$$

3. Consistency: the case without covariate

When are the assumptions satisfied ?

$$(1)\ \mathrm{diam}(I_n(x)) \to 0 \quad \text{and} \quad (2)\ \mathbb{E}\left(1/(a_n|A_n(x)|)\right) \to 0.$$

Without covariate: $z(x) = x$ and $I_n(x) = A_n(x)$

For a regular tessellation of W_n (say Voronoï) with intensity h_n^{-d},

$A_n(x) = I_n(x)$ is the *zero cell* of the tessellation and we have:

$$\mathrm{diam}(I_n(x)) = O(h_n) \quad \text{and} \quad \mathbb{E}\left(1/|A_n(x)|\right) = h_n^{-d}.$$

Therefore:

(1) is ok whenever $h_n \to 0$

(2) depends on the asymptotic regime:

- if $a_n \to \infty$ (infill or intermediate), then ok whenever $a_n h_n^d \to \infty$
- if $a_n = 1$ (increasing domain): no consistency

3. Consistency: the case with covariates

When are the assumptions satisfied ?

$$(1)\ \mathrm{diam}(I_n(x)) \to 0 \quad \text{and} \quad (2)\ \mathbb{E}\left(1/(a_n|A_n(x)|)\right) \to 0.$$

With covariates:

For a regular tessellation of $z(W_n)$ with intensity h_n^{-p},

(1) ok if $h_n \to 0$ since $\mathrm{diam}(I_n(x)) = O(h_n)$.

(2) $A_n(x) \approx$ level set of z at $z(x)$.
If z takes often the value $z(x)$, then $|A_n(x)|$ can be "large"

Example : z is binary, $z(W_n) = \{0, 1\}$ for n large. Say $z(x) = 0$.
Then $A_n(x) = z^{-1}(0) \cap W_n$ and typically $|A_n(x)| \to \infty$
$\Longrightarrow$ consistency in all asymptotics regimes

Other examples: z periodic or z realisation of an ergodic process

3. Minimax rates

3. Minimax rates

In $\lambda(x) = f(z(x))$, assume that z is α-Hölder continuous and that f is β-Hölder continuous, so that λ is $\alpha\beta$-Hölder continuous. Then

(i) for a pure RF based on a "regular tessellation" of $z(W_n)$ with intensity h_n^{-p},

$$\mathbb{E}\left[\left(\hat{\lambda}^{(RF)}(x) - \lambda(x)\right)^2\right] \leq c\left(\frac{1}{a_n h_n^{d/\alpha}} + h_n^{2\beta}\right).$$

(ii) pure RF based on a "regular tessellation" of W_n with intensity h_n^{-d},

$$\mathbb{E}\left[\left(\hat{\lambda}^{(RF)}(x) - \lambda(x)\right)^2\right] \leq c\left(\frac{1}{a_n h_n^{d}} + h_n^{2\alpha\beta}\right).$$

In both cases the minimax rate $a_n^{-2\alpha\beta/(2\alpha\beta+d)}$ is achieved when $a_n \to \infty$ for a proper choice of $h_n \to 0$.

4. What is the interest to leverage on covariates?

Conclusion : for Hölder-continuous functions, the optimal rate is minimax when $a_n \to \infty$ whether or not we use the covariates.

4. What is the interest to leverage on covariates?

Conclusion : for Hölder-continuous functions, the optimal rate is minimax when $a_n \to \infty$ whether or not we use the covariates.

What is the interest to leverage on covariates?

4. What is the interest to leverage on covariates?

Conclusion : for Hölder-continuous functions, the optimal rate is minimax when $a_n \to \infty$ whether or not we use the covariates.

What is the interest to leverage on covariates?

- If $a_n = 1$ (increasing domain):
  - $\hat{\lambda}(x)$ is not consistent if we do not use covariates
  - $\hat{\lambda}(x)$ is consistent if we use the covariates z and z is “ergodic”.

4. What is the interest to leverage on covariates?

Conclusion : for Hölder-continuous functions, the optimal rate is minimax when $a_n \to \infty$ whether or not we use the covariates.

What is the interest to leverage on covariates?

- If $a_n = 1$ (increasing domain):
  - $\hat{\lambda}(x)$ is not consistent if we do not use covariates
  - $\hat{\lambda}(x)$ is consistent if we use the covariates z and z is "ergodic".
- If $a_n \to \infty$ (infill or intermediate regime): the rate when using covariates can be faster in some cases.

Example: If z is binary and continuous at x then

- with covariates: $\mathbb{E}\left[\left(\hat{\lambda}^{(RF)}(x) - \lambda(x)\right)^2\right] \leq c/(a_n|W_n|)$,
- without covariates: $\mathbb{E}\left[\left(\hat{\lambda}^{(RF)}(x) - \lambda(x)\right)^2\right] \leq c/(a_n h_n^d)$.

5. Benefits of a RF over a single tree

5. Benefits of a RF over a single tree

We may prove that for a pure RF

$$\mathbb{E}\left[\left(\hat{\lambda}^{(RF)}(x) - \lambda(x)\right)^2\right] \leq \mathbb{E}\left[\mathbb{V}(\hat{\lambda}^{(1)}(x)|\pi_n^{(1)})\right] + \frac{1}{M}\mathbb{V}(B_n) + \mathbb{E}(B_n)^2,$$

where $B_n = \mathbb{E}\left(\hat{\lambda}^{(1)}(x)|\pi_n^{(1)}\right) - \lambda(x)$: conditional bias of a single tree.

For a single tree, the bias can be large, i.e. $\mathbb{V}(B_n)$ may be large.

Consequently,

- For a single tree ($M = 1$), the rate can be sub-optimal when $a_n \to \infty$ (this happens for instance if λ is $\mathcal{C}_1$ and λ' is β-Hölder)
- For a pure RF with M large enough, we recover the minimax rate.

Conclusion

RF approach adapts nicely to point process intensity estimation

Without covariate:

- Based on i.i.d. tessellations
- Works with any window shape
- Pure RF $\longrightarrow$ Theory pretty exhaustive

With covariates:

- Similar as standard RF: same benefits, same flaws
- Our theory is restricted to pure RF
- It is generally beneficial to leverage on covariates

Thank you