Point Process Learning: A new statistical approach for point processes

Ottmar Cronie

Department of Mathematical Sciences, Chalmers University of Technology & University of Gothenburg, Sweden

Joint work with Christophe A.N. Biscio (Aalborg University) and Mehdi Moradi (Umeå University).
Let $X = \{x_i\}_{i=1}^N$, $0 \leq N \leq \infty$, be a (simple) point process in a general (Polish) space $S$, with Borel sets $A \subseteq S$ and reference measure $A \mapsto |A| = \int_A du$, $A \subseteq S$ ("size/volume").

Heuristically, $X$ is a **generalised random sample** allowing dependencies and a random sample size, $N$. Formally, $X$ is a random element in the measurable space $(\mathcal{X}, \mathcal{N})$, where

i) $\mathcal{X} = \{x = \{x_1, \ldots, x_n\} \subseteq S : A \subseteq S$ bounded, $\#(x \cap A) < \infty\}$ is the space of locally finite point patterns/configurations,

ii) $\mathcal{N}$ is generated by the mappings $x \mapsto \#(x \cap A) \in \{0, 1, \ldots\}$, $A \subseteq S$, $x \in \mathcal{X}$.

Typically, we observe only one realisation $x$ of a point process $X \cap W$, $W \subseteq S$. 
Point process characteristics

Important distributional characteristics of $X = \{x_i\}_{i=1}^N$:

- The (Papangelou) conditional intensity $\lambda$ has the infinitesimal interpretation that
  $$\mathbb{P}(X(du) = 1|X \cap du^c = x \cap du^c) = \lambda(u; x)du, \quad u \in du \subseteq S, x \in X.$$  
  If $X$ finite ($N < \infty$ a.s.) then $\lambda(u; x) = j((x \setminus \{u\}) \cup \{u\})/j(x \setminus \{u\})$, where the Janossy density/likelihood $j(\cdot)$ governs the joint distribution of all points of $X$ in $S$.

  When $\lambda(\cdot; y) \leq \lambda(\cdot; x)$ if $y \subseteq x$ we have attractiveness

  When $\lambda(\cdot; y) \geq \lambda(\cdot; x)$ if $y \subseteq x$ we have repulsiveness

- The intensity function satisfies
  $$\rho(u) = \mathbb{E}[\lambda(u; X)],$$
  which coincides with $\lambda$ for a Poisson process (so both attractive/repulsive).

Objective: Estimate some unknown characteristic of $X \cap W$ based on $x$.
Example: true intensity (marginal model) or conditional intensity (full model).
Estimation

- A model is typically characterised as a **general parametrised estimator family**

\[ \Xi_\Theta = \{ \xi_\theta(u; y) : u \in S, y \in \mathcal{X}, \theta \in \Theta \}, \quad \Theta \subseteq \mathbb{R}^l, \ l \geq 1. \]

* Example: a non-parametric intensity estimator, \( \xi_\theta = \hat{\rho}_\theta \), or a parametric family of conditional intensities, \( \xi_\theta = \lambda_\theta \).

- In some cases they are constant on \( \mathcal{X} \):

\[ \xi_\theta(\cdot; y) = \xi_\theta(\cdot), \quad y \in \mathcal{X}. \]

* Example: a parametric intensity estimator, \( \xi_\theta = \rho_\theta \).

- **Estimation**: Find \( \hat{\theta} = \hat{\theta}(x) \in \Theta \) by minimising some loss \( \mathcal{L}(\theta) = \mathcal{L}(\xi_\theta(\cdot; x)) \geq 0. \)

Our new approach, **point process learning**, combines two new concepts:

1. **Point process cross-validation (CV)** – enables conditional repeated sampling of the underlying point process.

2. **Point process prediction errors** – measure how well a given model predicts a validation set.
Point process cross-validation

A marked point process may be obtained by attaching some (random/dependent) element $m(x)$ to each point $x \in X$.

**Definition (General thinning)**

Attach marks $m(x) \in \{0, 1\}$ to all $x \in X$ and define the associated **thinning** as

$$\{x \in X : m(x) = 1\}.$$ 

**Independent thinning**: indep. marking; retention prob. $p(x) = \mathbb{P}(m(x) = 1), x \in S.$

$p$-thinning: independent thinning with $p(\cdot) \equiv p \in (0, 1)$.

Thinning leads to a natural definition of **cross-validation (CV)**.

**Definition (General cross-validation)**

Generate training-validation set pairs $(x^T_i, x^V_i), i = 1, \ldots, k$, by letting $x^V_i$ be a thinning of $x$ and $x^T_i = x \setminus x^V_i$. Analogous for a point process $X$. 
Argument: **CV is most naturally achieved through independent thinning.**

**Lemma**

Let \( X^V \) be an independent thinning of a point process \( X \). Then \( \rho_{X^V}(u) = p(x)\rho_X(u) \) and 
\[
\lambda_{X^V}(u; X^V) = p(x)\mathbb{E}[\lambda_X(u; X)|X^V], \ u \in S.
\]

Examples of independent thinning-based CV:

- **Monte-Carlo CV** (\( p \)-thinning): Independently retain each \( x \in x \) \( (x \in X) \) with prob. \( p \in (0, 1) \) to generate \( x^V \); here \( p(\cdot) = p \in (0, 1) \).

- **Multinomial \( k \)-fold CV** \( (p \)-thinning): Give \( x \) iid multinomial marks with parameters \( k \) and \( p_1 = \ldots = p_k = 1/k \), and set \( x^V = \{x \in x : m(x) = i\} \); here \( p(\cdot) = p = 1/k \).

- **block-CV**: \( x^V = x \cap W_i \) for \( p(x) = 1\{x \in W_i\} \) and a partition \( \{W_i\}_{i=1}^k \) of \( W \).
Point process cross-validation

We will focus on Monte-Carlo CV and multinomial $k$-fold CV.

Figure: Monte-Carlo CV and block CV
Want to measure how well $\xi_\theta(\cdot; X_i^T)$ predicts the points of $X_i^V$.

**Definition**

Consider two general parametrised estimator families, $\Xi_\Theta = \{\xi_\theta : \theta \in \Theta\}$ and $\mathcal{H}_\Theta = \{h_\theta : \theta \in \Theta\}$ (test/weight functions). For an arbitrary pair $(X_i^T, X_i^V)$ and $A \subseteq S$, the ($\mathcal{H}$-weighted) prediction errors:

$$I_{\xi_\theta}^{h_\theta}(A; X_i^V, X_i^T) = \sum_{x \in X_i^V \cap A} h_\theta(x; X_i^T \setminus \{x\}) - \int_A h_\theta(u; X_i^T)\xi_\theta(u; X_i^T)du, \quad \theta \in \Theta.$$ 

- Note that we allow $X_i^T \cap X_i^V \neq \emptyset$, e.g. **auto-prediction**: $X_i^T = X_i^V = X$.
- **Sum**: collects $h_\theta$-generated ($X_i^T$-based) "predictions" of the points of $X_i^V \cap A$.
- **Integral**: represents a $\xi_\theta$-governed "expected counterpart" /"compensator".
Special cases (auto-prediction):

- When $\xi_\theta = \lambda_\theta$ is a parametric conditional intensity estimator, (the norm of)

$$
\mathcal{L}(\theta) = \mathcal{I}_{\lambda_\theta}^{h_\theta}(W; X, X) = \sum_{x \in X \cap W} h_\theta(x; X \setminus \{x\}) - \int_W h_\theta(u; X)\lambda_\theta(u; X)du,
$$
called an innovation (Baddeley et al., 2005), forms the basis of Takacs-Fiksel estimation.

Here, $h_\theta(u; X) = \nabla_\theta \log \lambda_\theta(u; X)$ yields pseudolikelihood estimation.

- When $\xi_\theta = \hat{\rho}_\theta$ is a non-parametric intensity estimator, (the norm of)

$$
\mathcal{L}(\theta) = \mathcal{I}_{\hat{\rho}_\theta}^{h_\theta}(W; X, X) = \sum_{x \in X \cap W} h_\theta(x; X \setminus \{x\}) - \int_W h_\theta(u; X)\hat{\rho}_\theta(u; X)du
$$
The intuition behind the prediction errors is made precise by the following result:

**Theorem**

For an arbitrary pair \((X^T_i, X^V_i)\), let \(\tilde{\lambda}\) be the conditional intensity of the marked point process \(\tilde{X} = \{(x, 0) : x \in X^T_i\} \cup \{(x, 1) : x \in X^V_i\} \subseteq S \times \{0, 1\}\).

Given \(\xi(u; X^T_i) = W_{\tilde{X}}(u) \lambda_X(u; X^T_i)\), under suitable second moment conditions, for any sufficiently nice \(h\), we have

\[
E[I_h(\xi; X^V_i, X^T_i)] = E \left[ \sum_{x \in X^V_i \cap A} h(x; X^T_i \setminus \{x\}) - \int_A h(u; X^T_i) \xi(u; X^T_i) du \right] = 0
\]

for any \(A \subseteq S\) if and only if \(W_{\tilde{X}}(u) = E[\tilde{\lambda}((u, 1); \tilde{X})/\lambda_X(u; X^T_i)|X^T_i]\).

In particular, when \((X^T_i, X^V_i)\) is independent thinning-based, \(W_{\tilde{X}}(u) = p(u)E[\lambda_X(u; X)/\lambda_X(u; X^T_i)|X^T_i]\). Moreover, \(W_{\tilde{X}}(u) \leq p(u)\) if \(X\) is repulsive, \(W_{\tilde{X}}(u) \geq p(u)\) if \(X\) is attractive and \(W_{\tilde{X}}(u) = p(u)\) if \(X\) is a Poisson process.
For training-validation set pairs \((X^V_i, X^T_i), \ i = 1, \ldots, k\), let

\[
\mathcal{I}_i(\theta) = \mathcal{I}^{h_\theta}_\xi(W; X^V_i, X^T_i) = \sum_{x \in X^V_i \cap W} h_\theta(x; X^T_i) - \int_W h_\theta(u; X^T_i) \hat{W}_\hat{X} \lambda_\theta(u; X^T_i) \, du,
\]

where \(\hat{W}_\hat{X}(u)\) is an estimator of the conditional expectation weight \(W_\hat{X}(u)\).

Assuming that \(\lambda_X = \lambda_{\theta_0}\) for some \(\theta_0 \in \Theta\), since we would like to minimise

\[
\mathbb{E}[|\mathcal{I}_i(\theta) - \mathcal{I}_i(\theta_0)|^j] \leq \mathbb{E}[|\mathcal{I}_i(\theta)|^j] + \mathbb{E}[|\mathcal{I}_i(\theta_0)|^j], \quad j = 1, 2,
\]

we treat \(\theta \mapsto \mathbb{E}[|\mathcal{I}_i(\theta)|^j]\) as a risk function and let its empirical counterpart

\[
\mathcal{L}_j(\theta) = \frac{1}{k} \sum_{i=1}^k |\mathcal{I}_i(\theta)|^j, \quad j = 1, 2,
\]

be a **loss function** to minimise. Another option: \(\mathcal{L}_3(\theta) = \left(\frac{1}{k} \sum_{i=1}^k \mathcal{I}_i(\theta)\right)^2\).

Things to specify: CV parameters \((k \text{ and } p)\) and test functions \(\mathcal{H}_\Theta = \{h_\theta : \theta \in \Theta\}\).
Kernel intensity estimation

The point pattern $x$ comes from a point process $X \subseteq W \subseteq S = \mathbb{R}^d$, $d \geq 1$, with unknown intensity $\rho(u)$, $u \in W$. **Kernel intensity estimation** of $\rho$:

$$\hat{\rho}_\theta(u; x) = \sum_{x \in x \cap W} \frac{\kappa_\theta(u-x)}{e_\theta(u, x)} = \sum_{x \in x \cap W} \frac{\theta^{-d}\kappa(((u-x)/\theta))}{e_\theta(u, x)}, \quad u \in W,$$

where

- the kernel $\kappa$ is a symmetric density function on $\mathbb{R}^d$,
- $e_\theta(u, x)$ is an edge correction term: compensates for possible interaction between points inside and outside $W$; e.g. $e_\theta(u, x) = \int_W \kappa(v-x)dv$.
- the smoothing parameter $\theta \in \Theta = (0, \infty)$ is called the **bandwidth**.

**Main challenge**: choose the bandwidth ”optimally” (given some kernel $\kappa$).

**Important note**: Mathematically, $\hat{\rho}_\theta$ is an attractive conditional intensity.
Bandwidth selection

Cronie & van Lieshout (2018)’s (implicit) heuristics:
1) As $x$ is ”central” in the distribution of $X \cap W$, $\rho(u) = E[\lambda(u; X)] \approx \lambda(u; x)$, $u \in W$.
2) If we model $\lambda$ well with $\hat{\rho}_\theta$, $\theta \in \Theta$, then additionally $\lambda(u; x) \approx \hat{\rho}_\theta(u; x)$.
3) Hence, use $\hat{\rho}(u) = \hat{\rho}_\theta(u, x)$, $u \in W$, as the final intensity estimate.

About Step 2):
- Conditional intensity estimation under model misspecification, since we employ $\hat{\rho}_\theta$ regardless of the distribution of $X$.
- We do it by means of point process learning.
- $\hat{\rho}_\theta$ attractive model $\Rightarrow$ main theorem tells us that we should have $\hat{W}_\hat{X} \geq \rho$; heuristics suggest $\hat{W}_\hat{X} = \rho/(1 - \rho)$. 

Bandwidth selection

Numerical comparison of point process learning and the Cronie-van Lieshout approach: In both we use test function $h_\theta(x; \mathbf{x}) = 1/\hat{\rho}_\theta(x, \mathbf{x})$, setting prediction error integrals to $|W|$.

Performance measures (average-estimates):
- integrated absolute bias, $IAB = \int_W |\mathbb{E}[\hat{\rho}_\theta(u, X)] - \rho(u)|\,du$,
- integrated squared bias, $ISB = \int_W (\mathbb{E}[\hat{\rho}_\theta(u, X)] - \rho(u))^2\,du$,
- integrated variance, $IV = \int_W \mathbb{V}ar(\hat{\rho}_\theta(u, X))\,du$,
- mean integrated squared error, $MISE = ISB + IV$.

Using Monte-Carlo CV in combination with $\mathcal{L}_3$, when $k \to \infty$ and $p \to 0$, we have that Point Process Learning tends to the method of Cronie & van Lieshout (2018).
100 realisations from a homogeneous determinantal point process with kernel 

\[(u, v) \mapsto \sigma^2 \exp\{-r\|u - v\|_2\}, \quad (\sigma^2, \beta) = (250, 50), \]

independently thinned with retention probability 

\[u \mapsto (10 + 80u_1)/90; \quad \rho(u) = \sigma^2(10 + 80u_1)/90 \]

and \(\mathbb{E}\{X(W)\} \approx 138.9\).

IAB, IV, ISB and MISE. Loss function: \(L_2\). \(k=100\) (black), \(k=200\) (red), \(k=300\) (blue), \(k=400\) (gold).

Numerics: Poisson process (no interaction)

100 realisations from a Poisson process with intensity $\rho(u) = 10 + 480u_1$, $u = (u_1, u_2) \in W = [0, 1]^2$.

IAB, IV, ISB and MISE. Loss function: $\mathcal{L}_2$. $k=100$ (black), $k=200$ (red), $k=300$ (blue), $k=400$ (gold).

Cronie & van Lieshout (2018): $IAB = 15.80$, $ISB = 921.82$, $IV = 4408.21$, $MISE = 5330.04$. 
100 realisations from a log-Gaussian Cox process with random intensity $u \mapsto \eta(u) e^{Z(u)}$, 
$\eta(u_1, u_2) = (10 + 80u_1)$, $u = (u_1, u_2) \in W = [0, 1]^2$; $Z$ is a Gaussian r.f. with $\mathbb{E}[Z(u)] = 0$ and $\text{Cov}(Z(u), Z(v)) = \sigma^2 e^{-\beta \|u-v\|}$, $(\sigma^2, \beta) = (2 \log 5, 50)$. $\rho(u) = \eta(u) e^{\sigma^2/2}$.

IAB, IV, ISB and MISE. Loss function: $L_2$. k=100 (black), k=200 (red), k=300 (blue), k=400 (gold).

Cronie & van Lieshout (2018): $\text{IAB} = 19$, $\text{ISB} = 964$, $\text{IV} = 17598$, $\text{MISE} = 18561$. 
Kernel intensity estimation

**Figure:** Locations of tropical rain forest trees on Barro Colorado Island, Panama, with a kernel intensity estimate overlaid.
Kernel intensity estimation

Figure: A point pattern of locations of spines on one branch of the dendritic tree of a rat neuron, with an obtained kernel intensity estimate overlaid.
References


Thanks for listening