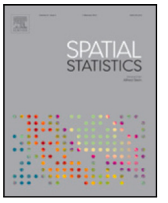
Spatial Statistics 74 (2026) 101005

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta



Robust variable selection for spatial point processes observed with noise

Dominik Sturm [a,b,c,d], Ivo F. Sbalzarini [a,b,c,d,e,*]

[a] Dresden University of Technology, Faculty of Computer Science, Dresden, Germany
[b] Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany
[c] Center for Systems Biology Dresden, Dresden, Germany
[d] Center for Scalable Data Analytics and Artificial Intelligence ScaDS.AI, Dresden/Leipzig, Germany
[e] Cluster of Excellence Physics of Life, TU Dresden, Dresden, Germany

ARTICLE INFO

Keywords:
Spatial point processes
Variable selection
Noise robustness
Lasso
Best-subset selection
Stability selection

ABSTRACT

We propose a method for variable selection in the intensity function of spatial point processes that combines sparsity-promoting estimation with noise-robust model selection. As high-resolution spatial data becomes increasingly available through remote sensing and automated image analysis, identifying spatial covariates that influence the localization of events is crucial to understand the underlying mechanism. However, results from automated acquisition techniques are often noisy, for example due to measurement uncertainties and detection errors. We study the impact of such noise on sparse point-process estimation across different models. To improve noise robustness without requiring additional knowledge about the true process, we propose to use stability selection based on point-process subsampling and to incorporate a non-convex best-subset penalty to enhance sparsity. In extensive simulations, we demonstrate that this approach reliably recovers true covariates under diverse noise scenarios and improves both selection accuracy and stability. We then apply the proposed method to a forestry data set, analyzing the distribution of trees in a tropical rain forest. This shows the practical utility of the method for robust variable selection in spatial point-process models under noise, without requiring additional knowledge of the process.

1. Introduction

Spatial data, described through the locations of points or events, are ubiquitous in applications. They are found in ecology (Renner and Warton, 2013; Renner et al., 2015), forestry (Waagepetersen and Guan, 2009), epidemiology (Zimmerman, 2008; Diggle, 2013), cell biology (Helmuth et al., 2010; Parra, 2021; Summers et al., 2022), and telecommunications (Li et al., 2015). Such data can be statistically modeled by spatial point processes. A spatial point-process model represents probabilities of observations as a random subset $X \subseteq W$, where W is an observation window of interest, i.e., the domain of the data. A fundamental quantity in spatial point processes is the expected number of events over W, which is described by the *intensity* of the process as the first moment of the distribution over X.

In practical applications, the intensity is often unknown, as the point process is observed through data, i.e., realizations of a point process. A common task then is to estimate the intensity at a location $u \in W$ from an observed point pattern with respect to some covariates $z(u) \in \mathbb{R}^p$, $p \geq 1$. While estimation of the intensity function is well studied (Møller and Waagepetersen, 2007),

* Corresponding author at: Dresden University of Technology, Faculty of Computer Science, Dresden, Germany.
E-mail address: sbalzarini@mpi-cbg.de (I.F. Sbalzarini).

https://doi.org/10.1016/j.spasta.2026.101005
Received 28 October 2025; Received in revised form 11 May 2026; Accepted 2 June 2026
Available online 8 June 2026

most methods do not explicitly account for uncertainty in the data, and the robustness of point-process estimation under noise has received little attention so far (Guttorp et al., 2023).

In spatial point processes, noise can occur by design, for example, in presence-only analysis in ecology (Renner et al., 2015), where events are modeled only if they are detected and not necessarily everywhere they are present. But it can also be introduced by automated data analysis, for example, in remote sensing (Gillespie et al., 2024), biomedical image analysis (Kuronen et al., 2021; Parra, 2021), and automated geocoding (Zimmerman, 2008; Briz-Redón, 2024). Noise comes in two flavors: point localization errors due to measurement uncertainties and misdetected events due to detection errors. Both types of noise can influence the performance of estimators in nontrivial ways.

Previous works on estimation under noise either focused on explicitly modeling noise and correcting the estimates (Assunção and Guttorp, 1999; Lund and Rudemo, 2000; Kuronen et al., 2021; Zimmerman, 2008; Briz-Redón, 2024) or on cross-validation (Yue and Loh, 2015; Rajala et al., 2018; Choiruddin et al., 2020). While this improves the accuracy of the estimates, it is inconsequential for the robustness of the estimator. Moreover, cross-validation becomes difficult if repeated experiments are not available or the number of samples is small, as correlations can lead to overfitting. Noise correction, on the other hand, requires a statistical noise model, which might not be available in applications where the data-acquisition process is not well understood. We therefore aim to devise an approach for estimating the intensity function without such assumptions and with high robustness, especially at small sample sizes.

Estimation from small sample sizes becomes particularly difficult when the number of covariates is large. Then, it becomes necessary — both for interpretability and computational efficiency — to use automatic variable-selection procedures (Hastie et al., 2015). The goal of variable selection is to identify a small set of covariates that are collectively sufficient to explain the observed spatial distribution of points. It is easy to see how changes in the observed point patterns due to noise can influence the set of selected variables.

The problem of variable selection has been addressed from different viewpoints. Fitting regularized maximum-likelihood estimators has been proposed for Poisson and clustering processes (Thurman and Zhu, 2014; Thurman et al., 2015; Choiruddin et al., 2018), as well as for Gibbs point processes (Yue and Loh, 2015; Ba and Coeurjolly, 2023). It has also been shown how such regularization techniques can be used to detect the correlation structure of highly multivariate point processes (Rajala et al., 2018; Choiruddin et al., 2020). Thurman et al. (2015), Choiruddin et al. (2018), Ba and Coeurjolly (2023) have established asymptotic results for the resulting estimators in an increasing domain for certain regularizations and processes. An alternative approach solved auxiliary tasks for variable-importance measures (Spychala et al., 2024). Collectively, these works have established the feasibility of variable selection in point-process modeling.

Variable selection in spatial point processes is not only feasible, but practically efficient. Experimental results have demonstrated that adaptive regularization methods, such as the adaptive Lasso (Zou, 2006), are particularly effective at recovering the sparse support of a model (Choiruddin et al., 2018; Ba and Coeurjolly, 2023; Coeurjolly et al., 2023). The problem then reduces to finding the optimal regularization parameter, which is commonly done by information criteria (Choiruddin et al., 2018, 2021; Ba and Coeurjolly, 2023). Information criteria, however, do not account for uncertainty in the data, and they require additional knowledge about the ground-truth point process to model the degrees of freedom. Specifically, variable selection using information criteria requires modeling or knowing the pair-correlation function of the process, which is often unknown in applications, and it does not improve noise robustness.

Here, we propose a method to improve the noise robustness of variable selection in point-process intensity estimation that does not require additional knowledge about the true process or its pair-correlation function. Specifically, we use stability selection (Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013) for model selection. Stability selection yields a robust statistic for model selection based on a consensus across multiple repetitions. This has been shown to increase variable selection performance under noise (Maddu et al., 2022), as well as selection stability compared to the Lasso (Nogueira et al., 2018). Stability selection is based on subsampling the data. For point processes, subsampling can be done by p-thinning. It has been shown that subsampling-based estimators are effective at finding the optimal bandwidth of kernel density estimates of point processes (Cronie et al., 2024). Inspired by these results, we derive estimating equations for stability selection over point processes. This allows controlling the per-family error rate (PFER) of the estimator under noise (Meinshausen and Bühlmann, 2010) and is easily integrated into existing algorithms and software.

The error control and noise robustness of stability selection can be modularly combined with different underlying sparsity-promoting inference losses. The most frequently used sparsity-promoting penalty for variable selection in point processes is the L_1 Lasso penalty. Lasso-type penalties are convex relaxations of the true variable-selection problem. Non-convex relaxations have also been considered in the literature, such as SCAD and MC+ (Choiruddin et al., 2018), but they remain relaxations. Variable selection amounts to best-subset selection, which is most directly modeled by a L_0 penalty. This, however, leads to a hard combinatorial optimization problem, since algorithms for L_0 optimization with guarantees have exponential time complexity. Motivated by results from compressed sensing (Blumensath and Davies, 2009) and PDE learning (Maddu et al., 2022), and to showcase the flexibility of stability selection to handle hard edge cases, we here consider the L_0 penalty in addition to the (adaptive) Lasso. Although it renders the resulting optimization problem non-convex, the L_0 penalty promises (in theory) faster convergence in the risk than the Lasso, without requiring assumptions on the design matrix (Bach, 2024). We leverage proximal operators to compute approximate, locally optimal solutions to the L_0 problem without theoretical guarantees on global optimality. We confirm in numerical experiments that the sparse local minima thus identified are sufficient to recover the true support of the model in many cases, while outperforming the Lasso due to stronger thresholding and reduced shrinkage bias. In combination with stability selection, L_0 penalization can provide additional improvements in selection performance under noise.

We validate the proposed combination of stability selection and L_0 penalization on synthetic data of simulated point processes with varying noise levels and different noise types at small sample sizes. The experiments suggest that the proposed algorithm is able to recover the true covariates under noise while being robust to overfitting, and the control over the PFER allows for principled model selection. We compare the model-selection performance of stability selection with that of various information criteria with and without oracle knowledge of the pair-correlation function. While stability selection introduces a computational overhead that scales linearly with the number of subsamples, we find that it improves variable-selection performance and robustness of the estimator, in particular when combined with a L_0 penalty.

To show the practical utility of the proposed method, we apply it to a real-world forestry data set, where we analyze tree occurrences in relation to elevation and soil nutrients. The identified models are consistent with results from the literature, but they tend to contain fewer variables and do not require knowledge of the underlying process or its pair-correlation function. We therefore believe that the proposed intensity estimation method is a useful addition to the toolbox of spatial point-process modeling, in particular in the presence of noise and at small sample sizes.

2. Methods

We review the formulation of spatial point processes considered here and introduce the notation. Then, we present the methods for sparse intensity estimation and model selection.

2.1. Spatial point processes and estimating function inference

A spatial point process X is a random process on $W \subseteq \mathbb{R}^d$, $d \geq 1$, whose realizations $X = \{u_i\}_{i=1}^n$ are a locally finite subset of W. We denote $N(B) = |X \cap B|$ as the number of points in some region $B \subseteq W$ and say X is locally finite if $N(B) < \infty$ for all $B \subseteq W$. As the distribution of X is in many cases difficult to express, either through counting measures or void probabilities, statistical inference usually revolves around the characterization of point processes through joint intensity functions (Møller and Waagepetersen, 2007; Baddeley, 2007). Given a set of points $u_1, \dots, u_m$ the mth order joint intensity $\rho^{(m)}(u_1, \dots, u_m)\mathrm{d}u_1, \dots, \mathrm{d}u_m$ can be interpreted as the probability of finding one point in each of the infinitesimal regions $\mathrm{d}u$. More formally, we can define $\rho^{(m)}$ through

$$\begin{aligned}&\mathbb{E}\left\{\sum_{u_1,\dots,u_m \in X}^{\neq} \mathbf{1}(u_1 \in B_1, \dots, u_m \in B_m)\right\} \\ &= \int_{B_1} \cdots \int_{B_m} \mathbf{1}(u_1 \in B_1, \dots, u_m \in B_m)\rho^{(m)}(u_1, \dots, u_m)\,\mathrm{d}u_1 \dots \mathrm{d}u_m. \end{aligned} \tag{1}$$

As such, $\rho^{(1)}$ and $\rho^{(2)}$ characterize the first and second moments of the distribution. We abbreviate $\rho(u)$ for the first-order intensity $\rho^{(1)}(u)$ and define the pair correlation

$$g(u,v) = \begin{cases} 0 & \text{if } \rho(u)\rho(v) = 0\,, \\ \dfrac{\rho^{(2)}(u,v)}{\rho(u)\rho(v)} & \text{otherwise}\,. \end{cases} \tag{2}$$

Using Palm conditioning, the product $\rho(u)g(u,v)$ can be interpreted as the intensity of observing a point at u given that $v \in X$ (Coeurjolly et al., 2017). A value of $g(u,v) > 1$ increases the probability of observing an additional point at u and therefore indicates *spatial attraction*. Conversely, $g(u,v) < 1$ decreases the probability of finding an additional point in some neighborhood and thus indicates *spatial inhibition*.

In the following, we consider processes that are second-order intensity-reweighted stationary, i.e., they only vary in their first moment, and their pair-correlation function is stationary (Diggle, 2013). Further restricting to isotropic interactions, this implies that the pair correlation simplifies to only depend on the distance r between two points, i.e., $g(u,v) = g(r)$. Many such processes can be modeled using a separable set of parameters $\theta = (\beta^\top, \alpha^\top) \in \mathbb{R}^{p_l+q}$. Here, $\beta = (\beta_1, \dots, \beta_p, \omega)^\top \in \mathbb{R}^{p_l}$ specifies the parameters of the intensity, which is given as a log-linear model $\rho(u;\beta) = \omega \exp\left\{\beta_{1:p}^\top z(u)\right\}$ depending on a set of spatial covariates $z(u)$ and $\omega > 0$ setting the scale. The parameters $\alpha \in \mathbb{R}^q$ model spatial interactions via the pair-correlation function $g(r;\alpha)$.

Since the likelihood is intractable for processes other than the Poisson process, the use of composite likelihood approaches has become popular for parametric estimation (Møller and Waagepetersen, 2007). Campbell's formula (Baddeley, 2007) can be used to specify a system of unbiased estimating equations (Lavancier et al., 2021). Given the above assumptions, we can estimate the mean using the Poisson score function

$$e(\beta) := \sum_{u \in X} \frac{\nabla_\beta \rho(u;\beta)}{\rho(u;\beta)} - \int_W \nabla_\beta \rho(u;\beta)\,\mathrm{d}u = 0\,. \tag{3}$$

This can be seen as the limit of composite likelihoods of Bernoulli random trials over partitions $c_i \in W$ as $\prod_{c_i} (\rho(c_i)|c_i|)^{N_i} (1 - \rho(c_i)|c_i|)^{1-N_i}$, for $N_i = \mathbf{1}(N(c_i) > 0)$, and it provides an unbiased estimating equation even if the underlying model is not Poisson (Møller and Waagepetersen, 2007, 2017).[1]

[1] This can be seen from Campbell's formula, choosing $f(u) = \nabla_\beta \log \rho(u;\beta)$ as the test function.

To estimate the parameters of the second-order properties of the process, one can use estimators based on the radially symmetric pair-correlation function $g(r)$, or the K-function, for the normalized number of events observed at distance r. If $g(r)$ or $K(r)$ are known, non-parametric models of $\hat{g}(r)$ or $\hat{K}(r)$ can be used to estimate α using minimum-contrast estimation (Møller and Waagepetersen, 2007; Waagepetersen and Guan, 2009). The non-parametric models $\hat{g}(r)$ and $\hat{K}(r)$ also depend on an initial estimate of the intensity $\rho(u;\beta)$, since they describe the "normalized" second-order properties of the process (Baddeley et al., 2000). Using the K-function, the objective function for the estimation of α can be written as

$$\hat{\alpha} = \arg\min_{\alpha\in\mathbb{R}^q} \int_{r_{\min}}^{r_{\max}} \left[K(r;\alpha)^b - \hat{K}(r)^b\right]^2 \mathrm{d}r\,. \tag{4}$$

This minimizes the distance between the estimated and observed K-functions over some range $[r_{\min}, r_{\max}]$, which has to be chosen small enough to capture local variations. The exponent b controls the variance of the estimator; it is chosen empirically (Diggle, 2013).

In this paper, we consider two different point processes to assess how the underlying correlation structure affects variable-selection performance. Specifically, we consider a Poisson point process with intensity $\rho(u;\beta)$, in which events are uncorrelated, and a Thomas process, which generates spatial clustering. To simulate a Thomas process with intensity $\rho(u;\beta)$ of the specified form, we first sample *parent points* Y from a homogeneous Poisson point process with intensity $\kappa > 0$. Conditional on each $v \in Y$, we sample a second Poisson point processes X_v with intensity

$$\rho_v(u) = \omega \exp\{\beta_{1:p}^\top z(u)\}\mathbf{Normal}(u; v, \sigma^2\mathbb{I}_d)/\kappa\,, \tag{5}$$

where $\mathbf{Normal}(\cdot)$ is the d-dimensional normal distribution, and $\omega > 0$ controls the number of *daughter points*. The Thomas point process is then given by the superposition of all daughter-point patterns $\cup_{v\in Y} X_v$. The pair correlation function in $\mathbb{R}^2$ is given by

$$g(u,v) = 1 + \frac{\exp\{-\left\|u-v\right\|_2^2/(4\sigma^2)\}}{4\pi\kappa\sigma^2}\,, \tag{6}$$

which is attractive ($g(r) \geq 1$) for all $r > 0$ (Møller and Waagepetersen, 2007). Event clusters can thus either arise due to an inhomogeneous intensity function or from attractive interactions between points. This makes variable selection more challenging.

2.2. Sparse intensity estimation

Parametric estimation of point-process models often revolves around identifying how spatial covariates $z(u)$ influence the expected number of points in some region of interest. If the dimensionality of $z(u)$ rapidly increases with $p \gg 1$, the question of sparse variable selection naturally arises. For point processes, we can adapt the Poisson likelihood in Eq. (3) to regularize the amount of selected variables similar to the Lasso:

$$\log \ell(\beta) = \sum_{u\in X} \log \rho(u;\beta) - \int_W \rho(u;\beta)\,\mathrm{d}u - \sum_{i=1}^{p} \lambda_i h(\beta_i)\,. \tag{7}$$

Here, h denotes a penalty function to be chosen. Since previous studies argued for the use of adaptive penalties, we here use the adaptive Lasso, which has been found to perform best in for both intensity (Choiruddin et al., 2018) and conditional intensity (Ba and Coeurjolly, 2023) estimation tasks. Following Zou (2006), we set the adaptive penalty to $\lambda_i = \lambda/|\hat{\beta}_i|$, where $\hat{\beta}$ is the unpenalized maximizer of Eq. (7). This provides a convex relaxation of the actual variable-selection problem.

Sparse variable selection penalizes the number of nonzero coefficients in β. This defines an L_0 problem with penalty function $\|\beta\|_0 := \sum_{i=1}^p \mathbf{1}(\beta_i \neq 0)$. This, however, constitutes a best-subset selection problem, which is NP-hard to solve exactly (Bach, 2024) and breaks the assumptions on the derivatives for oracle properties as provided by Choiruddin et al. (2018), such that theoretical results are difficult to obtain. Moreover, the L_0 penalty function is not differentiable, hampering the use of gradient-based optimizers. Nevertheless, it can be approximately solved using proximal gradient descent (PGD), where the best-subset penalty amounts to hard thresholding. This favors sparser solutions than the shrinkage induced by the Lasso. Proximal operators provide a general framework for non-differentiable penalties (Yang and Yu, 2020). The proximal operator of a penalty function h is defined as $\mathrm{prox}_h(x) := \arg\min_{v\in\mathbb{R}^p}\left[h(v) + \frac{1}{2}\left\|v-x\right\|_2^2\right]$. This amounts to a generalized projection operator (Parikh and Boyd, 2014). PGD computes a series of parameter updates by gradient descent over the negative Poisson likelihood followed by applying the proximal operator of the penalty. For the L_0-constrained problem with penalty weight λ, this becomes

$$\begin{aligned}
\beta^{t+1} &= \mathrm{prox}_{\lambda\gamma\|\cdot\|_0}\left(\beta^t + \gamma e(\beta)\right) && (8)\\
&= \arg\min_{v\in\mathbb{R}^p}\left[\lambda\left\|v\right\|_0 + \frac{1}{2\gamma}\left\|v - \left(\beta^t + \gamma e(\beta)\right)\right\|_2^2\right] && (9)\\
&= T_{\sqrt{2\lambda\gamma}}\left(\beta^t + \gamma e(\beta)\right)\,. && (10)
\end{aligned}$$

Here, γ is the step size of the gradient descent, and T_ξ is the hard-thresholding operator

$$T_\xi(\beta) = \begin{cases} \beta & |\beta| > \xi \\ 0 & \text{else}\,. \end{cases} \tag{11}$$

Table 1
Penalty functions h and their corresponding proximal operators, where λ is the penalty weight and γ the gradient-descent step size (see Algorithm 1). All proximal operators are evaluated element-wise.

Penalty h	prox_h	Name
$\|\beta\|_0$	$T_{\gamma\lambda}(\beta) = \begin{cases} \beta & \beta^2 > 2\gamma\lambda \\ 0 & \text{else} \end{cases}$	best-subset selection — hard thresholding
$\|\beta\|_1$	$S_{\gamma\lambda}(\beta) = \begin{cases} \beta - \gamma\lambda & \beta > \gamma\lambda \\ \beta + \gamma\lambda & \beta < -\gamma\lambda \\ 0 & \text{else} \end{cases}$	Lasso — soft thresholding

The same algorithm, with the correspondingly changed proximal operator, also applies to Lasso and elastic net penalties and their adaptive variants. Then, T_ξ becomes the soft-thresholding or scaled soft-thresholding operator, respectively, resulting in the classic ISTA algorithm (Hastie et al., 2015). This is summarized in Table 1. The proximal operators for other non-convex penalties, such as SCAD and MC+, can be found in Gong et al. (2013).

We obtain a fast algorithm for first-order PGD by computing an adaptive step size $\gamma \in \mathbb{R}_+$ using the Barzilai–Borwein (BB) method (Barzilai and Borwein, 1988). This computes the step size from a scalar approximation of the Hessian by solving the least-squares problem

$$\hat{\gamma} = \arg\min_{\gamma\in\mathbb{R}_+} \left\| \Delta\beta - \gamma\Delta g \right\|_2^2 = |\Delta\beta^\top \Delta g| / \left\| \Delta g \right\|_2^2 , \tag{12}$$

where $\Delta\beta = \beta^t - \beta^{t-1}$ and $\Delta g = e(\beta^t) - e(\beta^{t-1})$ for two subsequent iterations $t-1$ and t. We observed empirically for the L_1 problem that PGD with such an adaptive step size outperforms PGD with fixed step size and accelerated versions both in speed and stability (Beck and Teboulle, 2009; Hastie et al., 2015). In some of the cases presented below, however, we observed that BB step sizes can induce oscillations when using the L_0 penalty, where a model term oscillates between being in the support and being set to zero. This is due to the dependence of the hard thresholding on the step size γ close to the inclusion boundary. Therefore, we use BB-adaptive steps only for the L_1 penalty. For the L_0 penalty, we use a fixed step size of $\gamma = 10^{-3}$. We find that this performs comparably to accelerated PGD (Yang and Yu, 2020).

The algorithm stops upon convergence. We detect convergence if the likelihood does not increase over 1000 subsequent iterations or if the relative change in β is below $\epsilon = 10^{-4}$ between two subsequent iterations.

2.3. Model selection

The penalization weight λ controls the number of selected variables and hence the complexity of the resulting model. Choosing λ is crucial. This is typically done using information criteria, such as the Akaike (AIC) or Bayesian information criteria (BIC) (Choiruddin et al., 2018, 2021, 2023). However, information criteria underperform for misspecified likelihoods, e.g., when estimating non-Poisson models using Eq. (3), because they ignore the correlation structure (Choiruddin et al., 2021). The model may then include additional covariates to compensate for unmodeled clustering in the data.

Correcting for unmodeled correlations, composite information criteria for the estimating equation in Eq. (3) have been derived (Choiruddin et al., 2021; Ba and Coeurjolly, 2023). This includes the composite BIC (Choiruddin et al., 2021)

$$\mathbf{cBIC}(\hat{\beta}) = -2\log\ell(\beta) + \mathbf{df}(\rho_\beta)\log N(W) , \tag{13}$$

where $\mathbf{df}(\rho_\beta) = \text{tr}(S^{-1}\Sigma)$ are the effective degrees of freedom with sensitivity matrix S and asymptotic variance–covariance matrix Σ. While S can be estimated from the Hessian of $\log\ell(\beta)$, estimating Σ requires additional knowledge about the pair-correlation function $g(r)$ of the process. Choiruddin et al. (2021) proposed the simplified estimator

$$\mathbf{df}(\rho_\beta) = k + \text{tr}(S^{-1}T_2) , \tag{14}$$

with $k > 0$ the number of non-zero entries in $\hat{\beta}$ and

$$T_2 = \int_W \int_W z(u)z(v)^\top \rho(u;\hat{\beta})\rho(v;\hat{\beta})(g(u,v) - 1)\,\mathrm{d}u\mathrm{d}v . \tag{15}$$

This requires estimating $g(u,v)$ using a valid parametric model (Choiruddin et al., 2021), which implies additional assumptions on the underlying process. If the underlying process is Poisson, i.e., if $g(u,v) = 1$, the term T_2 vanishes and $\mathbf{df}(\rho_\beta) = k$ reduces to the standard BIC (Choiruddin et al., 2021).

Another generalization of the BIC to penalized likelihoods has been proposed with the extended regularized information criterion (ERIC) (Hui et al., 2015). The composite ERIC (cERIC) has been used to estimate Lasso-regularized conditional-intensity models of Gibbs point processes (Ba and Coeurjolly, 2023). It directly includes knowledge of the regularization parameter λ to weight the degrees of freedom:

$$\mathbf{cERIC}(\hat{\beta}) = -2\log\ell(\beta) + \mathbf{df}(\rho_\beta)\log\left(\frac{N(W)}{\lambda}\right) . \tag{16}$$

Composite information criteria like cBIC and cERIC can outperform the standard BIC by favoring models with higher penalization (Choiruddin et al., 2021; Ba and Coeurjolly, 2023).

Using composite information criteria on real data, however, poses two challenges: First, neither the pair-correlation function nor the intensity function of the underlying point process are usually known. Second, while composite information criteria correct for correlations in the data, they do not account for noise. We therefore propose an alternative model selection procedure based on stability selection. This does not require modeling the second-order moments, and it improves noise robustness by combining knowledge over several subsamplings of the data (Meinshausen and Bühlmann, 2010). In the context of sparse regression, stability selection can be understood as a bootstrap estimate of the inclusion probability of a covariate β_j in the support of the model S. As the support depends on the choice of $\lambda \geq 0$, we estimate the support $S^\lambda(Z_i)$ over K subsamples Z_i, $i = 1, \dots, K$, of the observed process X over the regularization path $\lambda \in [\lambda_{\min}, \lambda_{\max}] =: \Lambda$. The stability measure is then defined as (Meinshausen and Bühlmann, 2010):

$$\Pi_j^\lambda := \mathbb{P}\{j \in S^\lambda\} \approx \frac{1}{K} \sum_{i=1}^{K} \mathbf{1}(j \in S^\lambda(Z_i)), \tag{17}$$

which converges by the law of large numbers given a sufficient bootstrap size K. In our experiments, we find that $K = 50$ was sufficient. Larger K did not significantly improve the results. The decision to select a covariate β_j is taken by thresholding the inclusion probability $\mathbb{P}\{j \in S^\lambda\}$, i.e., β_j is selected if and only if $\Pi_j^\lambda \geq \pi_{\text{th}}$. The threshold π_{th} defines the required stability for a term to be included in the model. It is usually chosen between 0.7 and 0.9 as suggested by Meinshausen and Bühlmann (2010).

Stability selection estimates selection probabilities Π_j^λ and returns a set of selected variables based on π_{th}. This is different from information criteria, where model support selection and coefficient estimation are done simultaneously. In stability selection, coefficient estimation is performed after the selection step using the full point pattern X with fixed model support. Consequently, the method's primary objective is the accurate identification of the model support; once the support is identified, the Poisson likelihood yields consistent coefficient estimates.

A distinctive advantage of stability selection over (composite) information criteria and cross-validation is the possibility for error control during model selection. Meinshausen and Bühlmann (2010) derived an upper bound on the per-family error rate (PFER), which is the expected number of falsely selected variables, as a function of the selection threshold π_{th} and the expected number of selected terms over the regularization path Λ, $q_\Lambda = \mathbb{E}\{|\cup_{\lambda\in\Lambda} S^\lambda|\}$. Under an exchangeability assumption on the inclusion probabilities of noise variables, and for $\pi_{\text{th}} \in (0.5, 1)$, this bound is

$$\text{PFER} \leq \frac{1}{2\pi_{\text{th}} - 1} \frac{q_\Lambda^2}{p}. \tag{18}$$

Following Meinshausen and Bühlmann (2010), and unless otherwise stated, we choose $\lambda_{\max}$ such that $|S^{\lambda_{\max}}(Z_i)| = 0$ and the $\lambda_{\min}$ with $q_\Lambda = \sqrt{0.8p}$. At $\pi_{\text{th}} = 0.9$ this guarantees that PFER ≤ 1. While the exchangeability assumption might not be satisfied in practice, for example if predictors are correlated, the empirical PFER in our experiments always remains close to or below the bound. The tightness of the bound could potentially be improved using complimentary-pairs stability selection (CPSS) (Shah and Samworth, 2013). For the experiments below, this was not necessary, but we nevertheless describe how the approach presented here can be extended to CPSS. Bodinier et al. (2023) proposed model-based automatic calibration of the stability-selection hyperparameters for selecting the final model from a feasible set obtained by error control. We find that this procedure sometimes improves the F_1 score, albeit at the cost of decreasing the True Positive Rate (TPR). Since this means that we might miss important covariates, we do not use automatic hyperparameter calibration here, although it can naturally be combined with our approach. Instead, and to maintain methodological simplicity, we use Eq. (18) for error control, which provides a principled procedure for model selection.

We use error control according to Eq. (18) only to identify the support of the model S with stability selection over $K = 50$ subsamples Z_i. The coefficient values are then estimated on the entire data X with fixed model support. When creating the bootstrap samples Z_i, we distinguish between replicated experiments and single observations. For replicated experiments $\mathcal{X} = \{Z_i\}_{i=1}^N$, we draw subsets from $\mathcal{X}$ with replacement (Hastie et al., 2009). If there is only a single observation, the bootstrap samples are obtained by subsampling the point pattern X. Similar to Cronie et al. (2024), who used repeated subsampling of a point process to derive a point-process learning objective, we argue that subsampling X should be done by independent thinning with retention probability $p_{\text{thin}} : W \to [0, 1]$, resulting in the thinned point pattern Z_i. The intensity of the thinned point process is given by $\rho_Z(u) = p_{\text{thin}}(u)\rho(u)$, where $\rho(u)$ is the intensity function of the original point process (Møller and Waagepetersen, 2003). Motivated by classic stability selection and bootstrapping, we choose $p_{\text{thin}} = 0.5$ unless mentioned otherwise.

This also readily extends to CPSS to allow for additional error control. There, one would estimate the parameters over both Z_i and $X \setminus Z_i$ and evaluate the stability for all $2K$ subsamples using Eq. (17). This could be further extended to loss-guided stability selection (Werner, 2025). In that case, the set of coefficients would be chosen to minimize a given loss function, for which loss functions based on innovation measures, as used in point-process learning (Cronie et al., 2024), might be well suited.

Due to the known distributional properties of the thinned process, we can define estimating equations for the subsamplings Z_i in stability selection:

$$e_i(\beta) := \sum_{z\in Z_i} \nabla_\beta \log \rho(z;\beta) - p_{\text{thin}} \int_W \nabla_\beta \rho(u;\beta)\, \mathrm{d}u. \tag{19}$$

This follows from Poisson-likelihood inference over ρ_Z. Therefore, each Z_i can be used to obtain a bootstrap estimate of the underlying intensity function $\rho(u;\beta)$. The likelihood can be approximated by a weighted Poisson regression using the Berman–Turner device (Berman and Turner, 1992) as implemented in standard generalized linear model software, such as the `spatstat`

Algorithm 1 Compute the stability path $\{\Pi^{\lambda_i}\}_{i=1}^{M}$ for a penalty function h from Table 1 using proximal gradient descent (PGD) with warm starts.

for $k \in \{1, \dots, K\}$ **do**
  $X_k \leftarrow \text{subsample}(X, p_{\text{thin}})$ ▷ multinomial or p-thinning
  $\hat{\beta}_{\lambda_0} \leftarrow 0$ ▷ initialize with 0 for $\lambda_0 = \lambda_{\max}$
  $\gamma^0 \leftarrow 10^{-4}$ **if** $h = L_1$ **else** 10^{-3}
  for $\lambda_i \in \{\lambda_1, \dots, \lambda_M : \lambda_i > \lambda_{i+1}\}$ **do** ▷ solve Eq. (7)
    $\beta^0_{\lambda_i} \leftarrow \hat{\beta}_{\lambda_{i-1}}$
    for $t \in \{1, \dots, \text{max\_iter}\}$ **or** convergence **do** ▷ perform PGD
      if $t > 1$ **and** adaptive_step **then**
        $\Delta\beta \leftarrow \beta^{t-1}_{\lambda_i} - \beta^{t-2}_{\lambda_i}$
        $\Delta g \leftarrow \nabla(-\log \ell(\beta^{t-1}_{\lambda_i}; X_k)) - \nabla(-\log \ell(\beta^{t-2}_{\lambda_i}; X_k))$
        $\gamma^t \leftarrow |\Delta\beta^\top \Delta g| / \left\|\Delta g\right\|_2^2$
      end if
      $\beta^t_{\lambda_i} \leftarrow \text{prox}_{\gamma^t \lambda_i h}(\beta^{t-1}_{\lambda_i} - \gamma^t \nabla(-\log \ell(\beta^{t-1}_{\lambda_i}; X_k)))$
    end for
    $\hat{\beta}_{\lambda_i} \leftarrow \beta^{\text{max\_iter}}_{\lambda_i}$
  end for
end for

R package (Baddeley et al., 2016). The `spatstat` implementation has also been extended to Lasso and elastic net regularizations in the `ppmlasso` (Renner et al., 2015) and `kppmenet` (Choiruddin et al., 2024) R packages. Here, we numerically solve the estimating equations using `pytorch` (Paszke et al., 2019) by implementing the PGD algorithm and discretizing the integral in the estimating equations using midpoint quadrature. This provides the flexibility for using different penalties, as well as the possibility for tensorization and GPU acceleration. The pseudo-code for our implementation is given in Algorithm 1. It uses warm starts to accelerate the computation of the regularization path (Hastie et al., 2015). We empirically find this algorithm to be particularly effective at converging to *sparse* local minima for the non-convex L_0 penalty.

3. Results

We empirically evaluate the proposed estimation procedure and compare it with model selection based on information criteria for data containing different types and levels of noise. We use simulated data with known ground truth to quantify the performance of the methods. As a baseline, we use a Poisson process with uncorrelated events. Then, we consider a Thomas process with correlated events as an example of a process with spatial attraction. We compare the variable-selection accuracy of the adaptive L_0 and L_1 penalties in conjunction with selection strategies based on information criteria and stability selection. Since estimating the intensity function of a Thomas process using the Poisson likelihood constitutes a misspecified model, we also evaluate the composite cBIC and cERIC in comparison with stability selection. Following the simulation benchmarks, we apply the proposed method to estimating the distribution of trees in a tropical rainforest. There, we aim to identify a sparse set of covariates, such as soil nutrients or topography, that influence tree distribution. This shall demonstrate the practical applicability of the proposed method.

3.1. Simulation benchmarks

We quantify the accuracy of the estimator using synthetic data. For this, we use the covariate data available from the Barro Colorado Island (BCI) research plot in Panama (Condit, 1998; Hubbell et al., 1999), which contains measurements of elevation, elevation gradients, and soil nutrients for a total of 15 covariates. We standardize all covariates and interpolate them to a common grid of size 201×101, which is the standard grid size for this dataset in `spatstat`. We consider the observation window $W = [0, 250] \times [0, 125]$ for the Poisson process and an erosion by 4σ for the Thomas process to avoid edge effects. This is the same simulation setup as used in previous works (Choiruddin et al., 2018, 2021; Ba and Coeurjolly, 2023), and it allows assessing variable selection performance under realistic covariates without needing to model the covariates by Gaussian processes. We consider both Poisson (in the following indicated by **P**) and Thomas (in the following indicated by **T**) point processes, where elevation and elevation gradients are used as the two true covariates. The intercept ω of the log-linear model is chosen to achieve a desired number of points in the observation window W. Specifically, we simulate $\mathbb{E}N(W) = 50, 100, \dots, 250$ to investigate performance over varying sample sizes. In line with previous studies (Waagepetersen and Guan, 2009; Choiruddin et al., 2023), these sample sizes are deliberately chosen small, as variable selection is particularly challenging in the data-limited regime. For the Poisson process, we choose moderate effect sizes $\beta_{1:2} = (1, 0.5)^\top$, as larger effect sizes generally improve estimator performance. For the Thomas process, we consider $\kappa = 4 \times 10^{-3}$ and scale parameter $\sigma = 1.5$, which results in a strong positive correlation between points and is adjusted for size (Choiruddin et al., 2021). The coefficients of the covariates are set to $\beta_{1:2} = (2, 0.75)^\top$ following previous works (Choiruddin et al., 2018, 2021). Therefore, one covariate has a strong effect, the other a moderate effect. In all cases we use the 201×101 grid

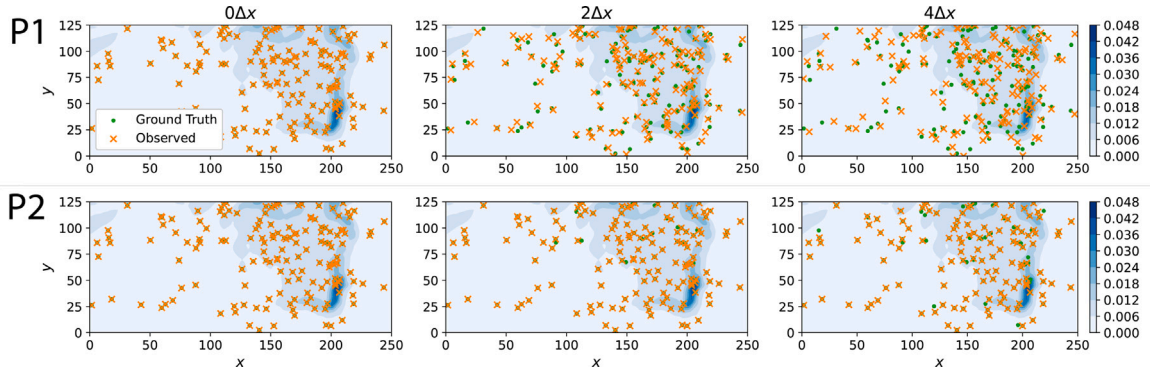


Fig. 1. Illustration of the simulation setup and the noise types considered here for a Poisson point process with $\mathbb{E}N(W) = 150$ and parameters $\beta = (1, 0.5)^\top$. The top row shows samples with different levels (from left to right: $c = 0, 2, 4$) of localization uncertainty (scenario **P1**). The bottom row shows the same with detection uncertainty (scenario **P2**). Green points indicate the true simulated point locations; orange crosses show the locations observed with noise. The blue shades visualize the intensity function of the underlying point process (color bar).

nodes as quadrature points and compute regularization paths $\Lambda = [10^{-4}, 5 \times 10^2]$ for the Poisson process and $\Lambda = [10^{-3}, 10^3]$ for the Thomas process, each with 35 log-equidistant points to ensure sufficient coverage of the parameter space.

Since we are particularly interested in the robustness of the variable-selection schemes against noise in the data, we artificially corrupt the data with two different types of noise common in practical applications:

Localization Uncertainty (P1/T1): Given $\mathbb{E}N(W)$, the process (Poisson or Thomas) is simulated to obtain a sample X. Localization uncertainty is modeled by adding random displacements to the sampled points in X. We use Gaussian displacements to model measurement errors and change each point's location u_i to $\tilde{u}_i = u_i + \epsilon$, $\epsilon \sim \mathbf{Normal}(0, \delta^2 \mathbb{I}_d)$. The standard deviation $\delta = c\Delta x$ depends on the grid spacing Δx with $c = 0, 1, \ldots, 4$ setting the dimensionless noise magnitude.

Detection Uncertainty (P2/T2): Given $\mathbb{E}N(W)$, the process (Poisson or Thomas) is simulated to obtain a sample X. Detection uncertainty is modeled by missing events using distance-dependent thinning. For each point $u_i \in X$ we draw a random cutoff distance $r_i \sim |\mathbf{Normal}(0, \delta^2 \mathbb{I}_d)|$ with $\delta = c\Delta x$. The point u_i is retained if and only if no previously retained point lies inside the ball $B(u_i, r_i)$. The noise magnitude is set by $c = 0, 1, \ldots, 4$.

Both types of noise can lead to biased estimates and affect the performance of variable-selection methods. This is particularly relevant in automated data acquisition, where location uncertainty can stem from inaccuracies in the detection process and points might be missed, e.g., due to occlusion in dense regions. Examples of noise realizations for $c = 0, 2, 4$ are shown in Fig. 1 for a simulated Poisson process.

Fig. 2 shows the regularization paths when using adaptive L_0 and L_1 penalties for a Poisson process with localization uncertainty (scenario **P1**) and $\mathbb{E}N(W) = 200$, $c = 4$. Here, the adaptive L_0 penalty achieves better separation between true covariates (solid colored lines with symbols) and noise covariates (dashed black lines) than the adaptive L_1 penalty. Notably, as seen from the coefficient paths in the left panels, the L_1 penalty assigns a nonzero coefficient to a noise covariate together with β_2, whereas for the L_0 penalty noise covariates only show up for low λ. The right panels show the stability paths Π_j^λ for both penalties with the selection threshold $\pi_{\text{th}} = 0.9$ indicated by a dotted line. The adaptive L_0 penalty yields lower stability scores for noise covariates than the adaptive L_1 penalty. As a result, noise variables are selected only at low λ, leaving a wider range of λ in which the true covariates are correctly identified. This means that the L_0-regularized estimator is more robust to noise in the data than the L_1 estimator.

To systematically quantify the influence of different penalties and selection criteria on the variable-selection performance under noise, we repeat the experiment for different sample sizes $\mathbb{E}N(W)$ and noise levels c with 100 independent repetitions each.[2] We measure the performance of variable selection using the True Positive Rate (TPR), False Positive Rate (FPR), Positive Predictive Value (PPV), the F_1 score, and the feature-selection stability Φ_S (Nogueira et al., 2018). The TPR is the fraction of correctly selected covariates over all true covariates. The FPR is the fraction of selected noise covariates over all noise covariates. The PPV is the fraction of correctly selected covariates over all selected covariates. The F_1 score is the harmonic mean of the TPR and PPV and is a standard metric in machine learning. The feature-selection stability Φ_S was proposed by Nogueira et al. (2018) and quantifies the

[2] We also ran case T1 with 250 independent repetitions of each experiment, which consumed 400 CPU hours for the additional repetitions without changing any of the conclusions, thus confirming that 100 repetitions are sufficient. This is also in line with previous simulation studies on variable selection for point processes (Rajala et al., 2018) and has been shown sufficient for estimating selection stability (Nogueira et al., 2018).

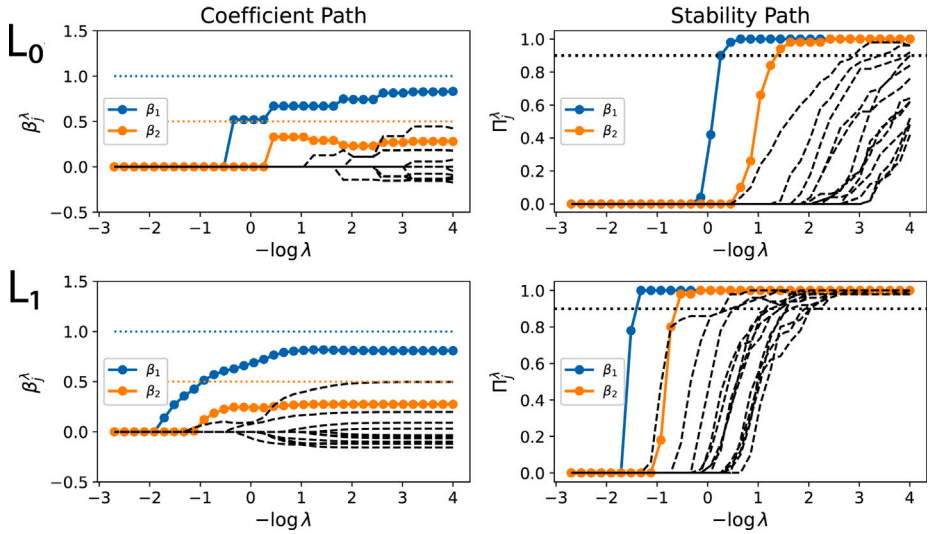


Fig. 2. Regularization paths ($\lambda \in [10^{-4}, 5 \times 10^2]$) for a Poisson process with parameters $\beta = (1, 0.5)^\top$ observed with localization uncertainty (scenario **P1**, $\mathbb{E}N(W) = 200$, $c = 4$). The penalty (L_0, L_1) is indicated by the row labels on the left. The left panels show the coefficient paths β_j^λ with the coefficients of the true covariates as symbol lines (ground truth values indicated by dotted lines) and noise covariates as dashed lines. The right panels show the corresponding stability paths Π_j^λ with the dotted horizontal line indicating the threshold $\pi_{\text{th}} = 0.9$ corresponding to PFER ≤ 1.

variance of the selection indicator random variable, i.e., how reliably a feature is selected. It decreases with increasing indicator variance as:

$$\Phi_S = 1 - \frac{\frac{1}{p}\sum_{f=1}^{p} s_f^2}{\frac{\bar{k}}{p}\left(1 - \frac{\bar{k}}{p}\right)}. \tag{20}$$

Here, $s_f^2 = \frac{M}{M-1}\hat{p}_f(1 - \hat{p}_f)$ is the empirical sample variance for selecting the fth feature with selection probability $\hat{p}_f$ over M repetitions (here $M = 100$). The denominator is the expected sample variance under random selection, with $\bar{k}$ the average number of selected features over all selections. Like all other performance metrics considered here, Φ_S is between 0 and 1 (for large M), where 0 indicates no stability and 1 indicates perfect stability.

Fig. 3 shows the performance of the considered algorithms for scenario **P1** (Poisson point process with localization uncertainty). It compares model selection based on the BIC (A), ERIC (B), and stability selection with PFER ≤ 1 (C) for the adaptive L_0 (top row of each subfigure) and adaptive L_1 (bottom row of each subfigure) penalties. Throughout all cases, performance decreases for lower sample sizes and higher noise magnitudes, as sampling bias increases or the true covariates are masked. The adaptive L_0 penalty outperforms the adaptive L_1 penalty in all metrics except in the TPR, where the adaptive L_1 penalty achieves better results in particular for small sample sizes. This is because the adaptive L_1 penalty selects more covariates than the adaptive L_0 penalty. Choosing the penalty therefore allows tuning the tradeoff between TPR and FPR depending on the incurred costs of type I and type II errors, respectively. The F_1 score and the stability Φ_S, however, are always better when using the adaptive L_0 penalty, despite its non-convex nature, as it reflects the true variable-selection objective.

Comparing selection strategies in Figs. 3A–C, we note that the BIC achieves the lowest performance in all metrics except the TPR, for the same reasons as discussed above. As already reported by Ba and Coeurjolly (2023), ERIC performs better than BIC, since it includes information about the regularization weight λ. ERIC achieves particularly good performance in combination with the adaptive L_0 penalty. For small sample sizes, however, this combination sometimes selects empty models (low TPR) because the likelihoods are attenuated by down-weighting the degrees of freedom for large λ. Stability selection achieves the best F_1 scores with L_0 performance comparable to the ERIC. This highlights that stronger penalization alone can already improve variable-selection performance. Also for stability selection, some empty models were selected for the smallest sample size and the highest noise level, where no covariate reached the threshold $\pi_{\text{th}} = 0.9$. However, this behavior is correct in light of the imposed error bound.

Since stability selection only yields the support of the model, the final estimates of the coefficients are obtained by fitting the model with fixed support to the entire data using a Poisson likelihood. The RMSE of the final parameter estimates for the selected models are reported in Fig. 4. The RMSE is generally lower for the adaptive L_0 penalty, which is partly explained by its better variable-selection performance, partly by the L_0 penalty not shrinking the coefficient as the L_1 penalty does.

The greatest difference between stability selection and ERIC is observed for the adaptive L_1 penalty. There, stability selection reduces the FPR and increases the F_1 score. This is because the PFER bound leads to fewer selected covariates. In Fig. 5 we show the empirically achieved PFER in all cases for both the adaptive L_0 and L_1 penalties. It is always well below the imposed bound of 1, confirming that stability selection is able to bound the number of false positives.

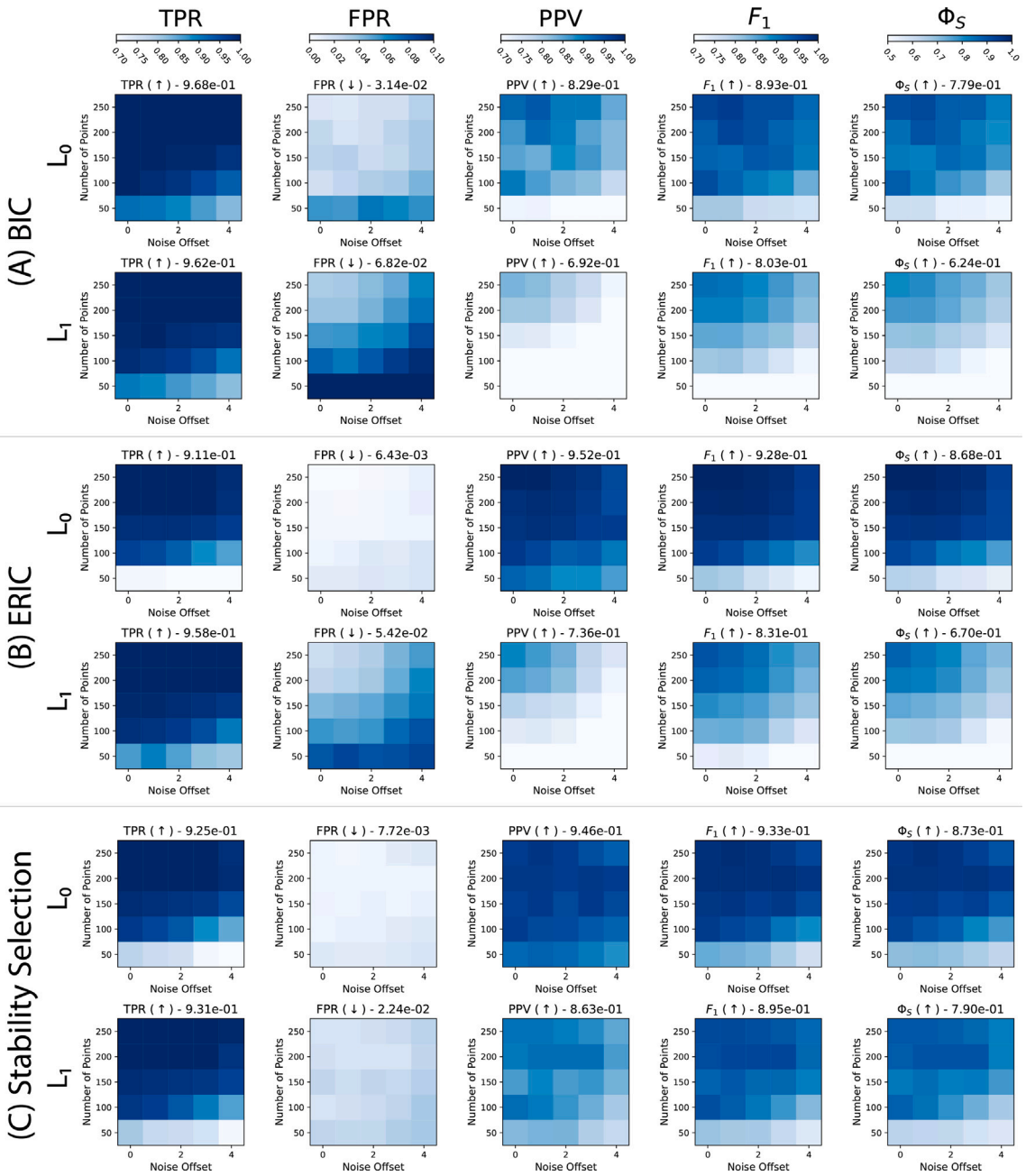


Fig. 3. Variable-selection performance for a Poisson point process with localization uncertainty (scenario **P1**). We show the mean (over 100 independent repetitions of each experiment) True Positive Rate (TPR), False Positive Rate (FPR), Positive Predictive Value (PPV), F_1 score, and feature-selection stability Φ_S for model selection using the BIC **(A)**, ERIC **(B)**, and stability selection with PFER ≤ 1 **(C)** with adaptive L_0 (top row of each subfigure) and adaptive L_1 (bottom row of each subfigure) penalties. Each panel shows a performance metric (top titles, color bars) for different noise magnitudes c (x-axis) and sample sizes $\mathbb{E}N(W)$ (y-axis). The average metrics over all 25 experiments are given in the panel titles with arrows (↑ / ↓) indicating the direction of improvement.

In order to derive scientific conclusions from the selected models, it is key that variable selection be stable under noise, i.e., that the same covariates are reproducibly selected for different realizations of the process. We quantify this by the feature-selection stability Φ_S as defined in Eq. (20). The results are shown in the last column of Fig. 3. The adaptive L_0 penalty achieves better stability than the adaptive L_1 penalty. This is expected, as L_0 selects fewer covariates and thus has a lower variance in the selection indicator. Using stability selection instead of information criteria generally improves the stability for both penalties, especially at low sample sizes and high noise levels. For high sample sizes ERIC with L_0 penalization performs best, indicating that in this case

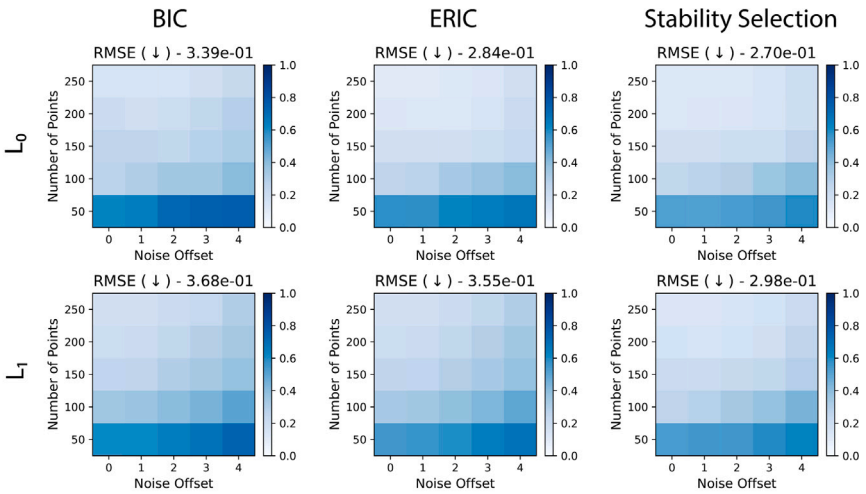


Fig. 4. Parameter estimation performance quantified by the RMSE of the selected models in scenario **P1** (Fig. 3) when using BIC, ERIC, and stability selection with either the L_1 or L_0 penalty. The averages over all 25 experiments are given in the panel titles.

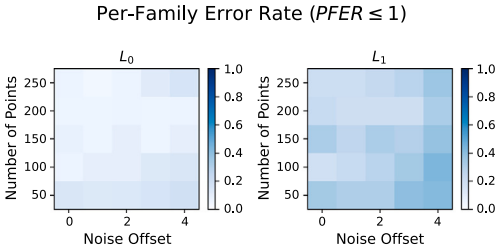


Fig. 5. Empirical confirmation that stability selection achieves the desired error bound PFER ≤ 1 for all experiments in scenario **P1**.

strong penalization helps to stabilize the selection. The bootstrap procedure of stability selection implicitly improves the stability of the selected variables as previously reported (Nogueira et al., 2018). This improved variable-selection stability is particularly important in the presence of noise.

We next consider scenario **P2**, a Poisson point process with detection uncertainty. The variable-selection performance is reported in Fig. 6 and the RMSE in Fig. 7 in the same format as before. They are qualitatively similar to those for scenario **P1** (Fig. 3) but with better average performance. Again, the adaptive L_0 penalty outperforms the adaptive L_1 penalty in all metrics except the TPR, as it favors higher sparsity. Also for this noise type, stability selection improves the performance for both penalties, but particularly for the adaptive L_1 penalty. The PFER bound is met in all cases also here (not shown), and stability selection again has better Φ_S than the information criteria, in particular for high noise magnitudes. We therefore conclude that the adaptive L_0 penalty in combination with stability selection achieves the best variable-selection performance for uncorrelated Poisson point processes under noise, for both noise types.

In scenario **P1**, performance monotonically decreases with increasing noise magnitude. Curiously, in scenario **P2**, we sometimes observe better performance for higher noise magnitudes. While this seems counter-intuitive at first, it can be explained by thinning noise removing points from the process according to their local hardcore neighborhoods. This introduces an apparent repulsive interaction between points, significantly influencing the second-order properties of the process (Kuronen et al., 2021). This is specific to the chosen noise model, and missing points by independent thinning would not have this effect (Møller and Waagepetersen, 2003). The resulting regular process has fewer effective degrees of freedom (see Eq. (15) for $g(r) \leq 1$), allowing more stable estimation of the parameters. This is also reflected in the stability scores being higher for thinning noise (scenario **P2**, Fig. 6) than for displacement noise (scenario **P1**, Fig. 3). We expect that similar observations might hold also for other repulsive processes. This exemplifies that noise can influence variable selection in non-trivial ways that can be qualitatively different for different noise processes.

After having established the baseline for uncorrelated Poisson processes, we consider a Thomas point process with spatial attraction leading to clustering of points. We use this as a representative example of a clustering point process, which also include shot-noise Cox processes and log-Gaussian Cox processes. We again consider both localization uncertainty (scenario **T1**) and detection uncertainty (scenario **T2**).

The results are shown in Fig. 8 for scenario **T1**. The performance is generally lower than for a Poisson process, which is expected for the more complex Thomas process. Again, the adaptive L_0 penalty achieves a better performance than the adaptive L_1 penalty

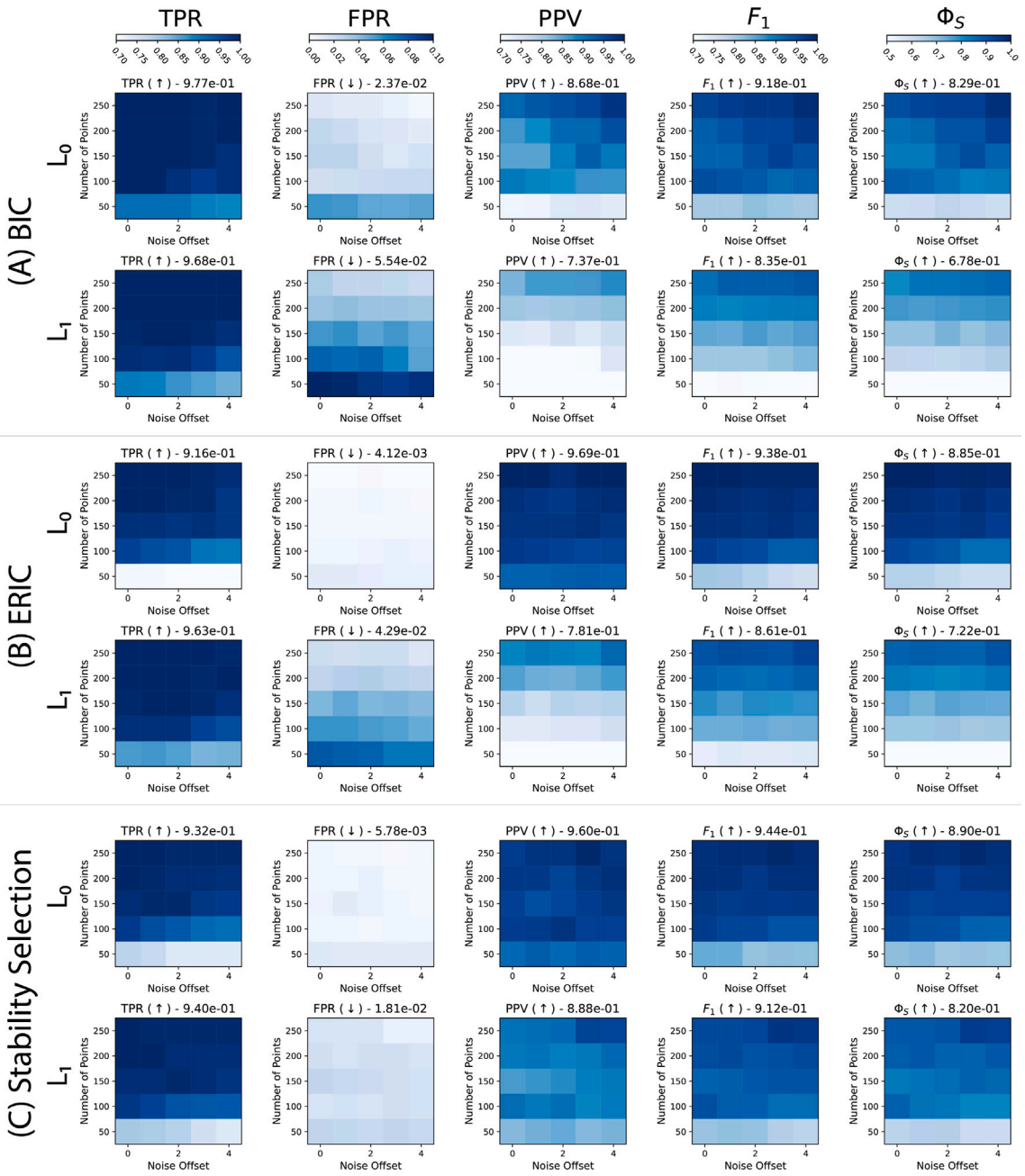


Fig. 6. Variable-selection performance for a Poisson point process with detection uncertainty (scenario **P2**). We show the mean (over 100 independent repetitions of each experiment) True Positive Rate (TPR), False Positive Rate (FPR), Positive Predictive Value (PPV), F_1 score, and feature-selection stability Φ_S for model selection using the BIC **(A)**, ERIC **(B)**, and stability selection with PFER ≤ 1 **(C)** with adaptive L_0 (top row of each subfigure) and adaptive L_1 (bottom row of each subfigure) penalties. Each panel shows a performance metric (top titles, color bars) for different noise magnitudes c (x-axis) and sample sizes $\mathbb{E}N(W)$ (y-axis). The average metrics over all 25 experiments are given in the panel titles with arrows (↑ / ↓) indicating the direction of improvement.

in all metrics except the TPR. Also as in the Poisson scenarios, the BIC achieves the overall lowest performance, followed by ERIC, which again works particularly well in conjunction with the adaptive L_0 penalty (37% improvement in the average F_1 score over all experiments). Both BIC and ERIC achieve high TPR but low PPV, suggesting under-penalization of the clustering Thomas process. Stability selection achieves the best F_1 scores and feature-selection stability Φ_S for both L_0 and L_1 penalties. The PFER is again below the threshold in all cases (not shown), albeit much closer to 1 in the case of the L_1 penalty and even reaching the bound for $\mathbb{E}N(W) = 250$ and $c = 1$. This becomes apparent in the reduced FPR compared to BIC/ERIC, especially for the adaptive L_1 penalty.

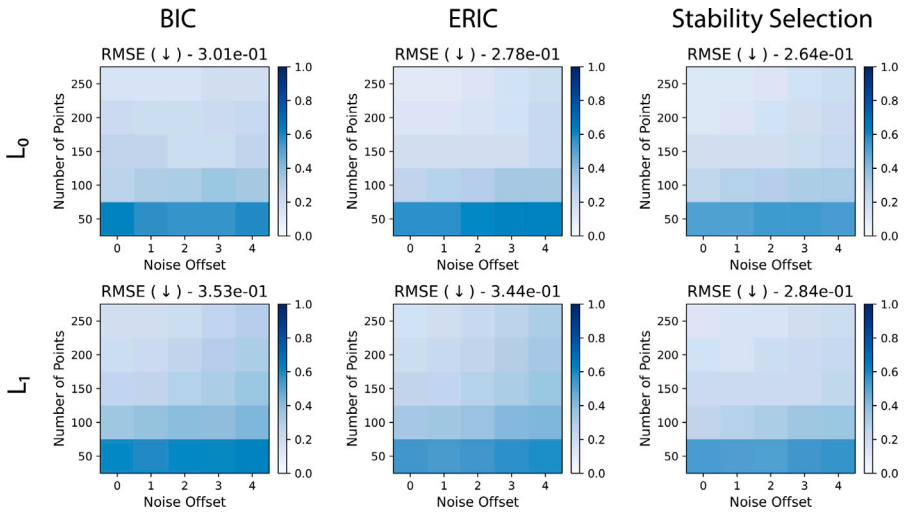


Fig. 7. Parameter estimation performance quantified by the RMSE of the selected models in scenario **P2** (Fig. 6) when using BIC, ERIC, and stability selection with either the L_1 or L_0 penalty. The averages over all 25 experiments are given in the panel titles.

The adaptive L_0 penalty further improves the FPR over the adaptive Lasso. For both penalties, the FPR achieved by stability selection for the Thomas process is comparable to scenarios **P1** and **P2**. We also report results using $p = 28$ covariates in the Appendix, where we observe similar trends. This suggests that the adaptive L_0 penalty in combination with stability selection is able to achieve high variable-selection performance without requiring knowledge about the second-order structure of the point process.

The performance of all methods tends to improve for smaller sample sizes or higher noise magnitudes, except for the TPR. This counter-intuitive behavior is because the clustering in the Thomas process decreases with increasing noise, or it is not observed at low sample sizes. At high noise or for small samples, the Thomas process therefore appears more Poisson-like. The effect is particularly strong for the information criteria, which become increasingly appropriate the more Poisson-like the process appears. This provides another example where the effect of noise (this time simple displacement noise) on variable selection is not straightforward. Stability selection is the least sensitive to this effect, showing a more uniform distribution of performance metrics versus noise and sample size (Fig. 8C). This is because stability selection does not model the noise process but generally reduces its influence.

The RMSE of the parameter estimates for the selected models are shown in Fig. 9. The RMSE is generally lower for the adaptive L_0 penalty and when using stability selection, consistent with the improved variable-selection performance. Also here, we observe the counter-intuitive result that the RMSE decreases with increasing noise when using information criteria. This can again be attributed to the process becoming more Poisson-like under noise, leading to improved estimation performance. For stability selection, however, the RMSE improves with larger sample sizes and lower noise levels, as expected. This is because model parameters are estimated using a Poisson likelihood only after variable selection, which — even for a correctly specified model — is expected to deteriorate with increasing noise and decreasing sample size.

Since the Thomas process introduces correlations between events, estimating its intensity using the Poisson likelihood constitutes a composite likelihood estimate (see Section 2.3). We therefore also compare stability selection with the cBIC and cERIC, which have been shown to outperform the standard BIC and ERIC for clustering point processes (Choiruddin et al., 2021). We compute the effective degrees of freedom using Eq. (14) in two different ways: (1) Using the true pair-correlation function of the Thomas process with known parameters. While this is unrealistic in practical applications, it constitutes the best case for composite information criteria. This is because the pair-correlation estimates generally contain errors, which are amplified by model misspecification, small sample size, and noise. This case hence provides a best-case baseline. (2) By estimating the parameters of the K-function, the structure of which is analytically known, using minimum-contrast estimation. For this estimation, we use a two-step procedure (Waagepetersen and Guan, 2009) with $r_{\min} = 0$, $r_{\max} = 25$, and $b = 0.25$ to reduce the variance for clustering point processes as recommended by Diggle (2013). We use K-function estimation because it is less sensitive to noise than directly estimating the pair-correlation function. The latter would additionally require choosing a bandwidth parameter and would exhibit high variance. Moreover, we find that likelihood-based approaches, such as those proposed by Waagepetersen (2007), can become unstable in the presence of noise or for small sample sizes when clustering is weak. The minimum-contrast estimate is computed using the L-BFGS-B optimizer as implemented in the `scipy` Python package (Virtanen et al., 2020). The resulting parameter estimates are then used to compute the pair-correlation function required by composite information criteria with the variance–covariance matrix computed on a two-fold coarsened grid.

Fig. 10 shows the results for cBIC and cERIC when using minimum-contrast estimation. Both composite information criteria perform better than their uncorrected versions in all cases (Fig. 8). Adaptive L_0 penalization achieves better performance and stability than the Lasso. However, neither cBIC nor cERIC achieve the performance of stability selection (Fig. 8C), except for cERIC at low noise and large sample sizes. In these cases the estimation of the K-function is expected to work best and the stronger

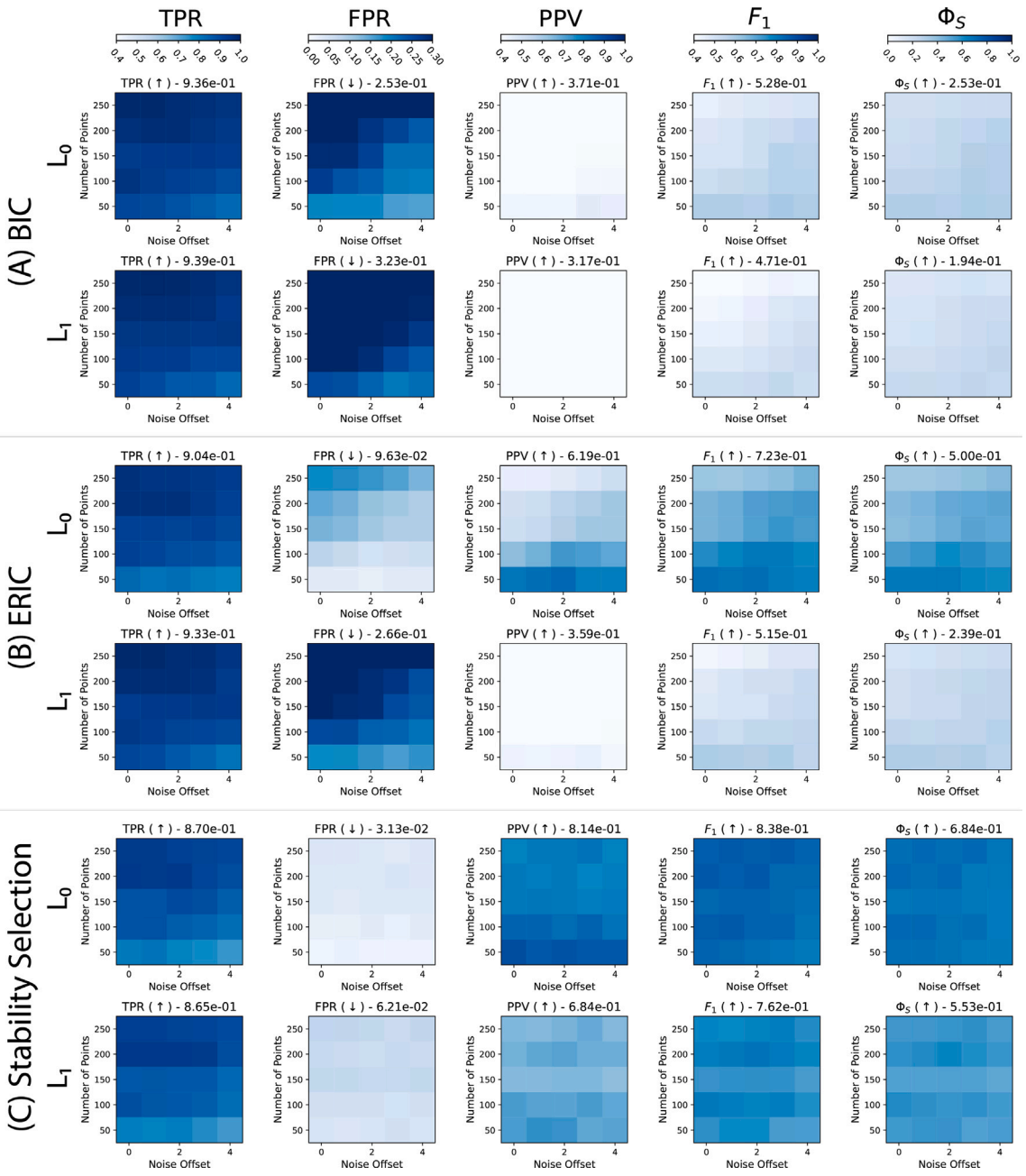


Fig. 8. Variable-selection performance for a Thomas point process with localization uncertainty (scenario **T1**). We show the mean (over 100 independent repetitions of each experiment) True Positive Rate (TPR), False Positive Rate (FPR), Positive Predictive Value (PPV), F_1 score, and feature-selection stability Φ_S for model selection using the BIC **(A)**, ERIC **(B)**, and stability selection with PFER ≤ 1 **(C)** with adaptive L_0 (top row of each subfigure) and adaptive L_1 (bottom row of each subfigure) penalties. Each panel shows a performance metric (top titles, color bars) for different noise magnitudes c (x-axis) and sample sizes $\mathbb{E}N(W)$ (y-axis). The average metrics over all 25 experiments are given in the panel titles with arrows (↑ / ↓) indicating the direction of improvement.

penalization leads to improved performance. Overall, the average F_1 score across all experiments is 6% higher for stability selection than for cERIC when using the L_0 penalty and 26% higher for the L_1 penalty. For cBIC, the differences are even larger. The same trends are observed for the RMSE of the parameter estimates (Fig. 12). We also observe that when using minimum-contrast estimation, composite information criteria increasingly select empty models for small sample sizes when using the L_0 penalty. We believe this is because the K-function overestimates the clustering due to the higher variance in the data and the renormalization by the estimated intensity function. In combination with the stronger L_0 penalty, this can lead to an over-penalization, favoring the

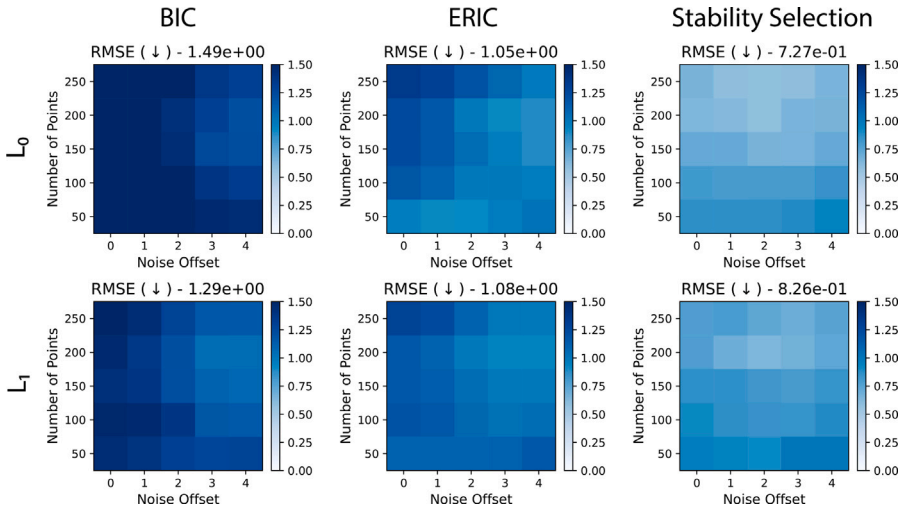


Fig. 9. Parameter estimation performance quantified by the RMSE of the selected models in scenario **T1** (Fig. 8) when using BIC, ERIC, and stability selection with either the L_1 or L_0 penalty. The averages over all 25 experiments are given in the panel titles.

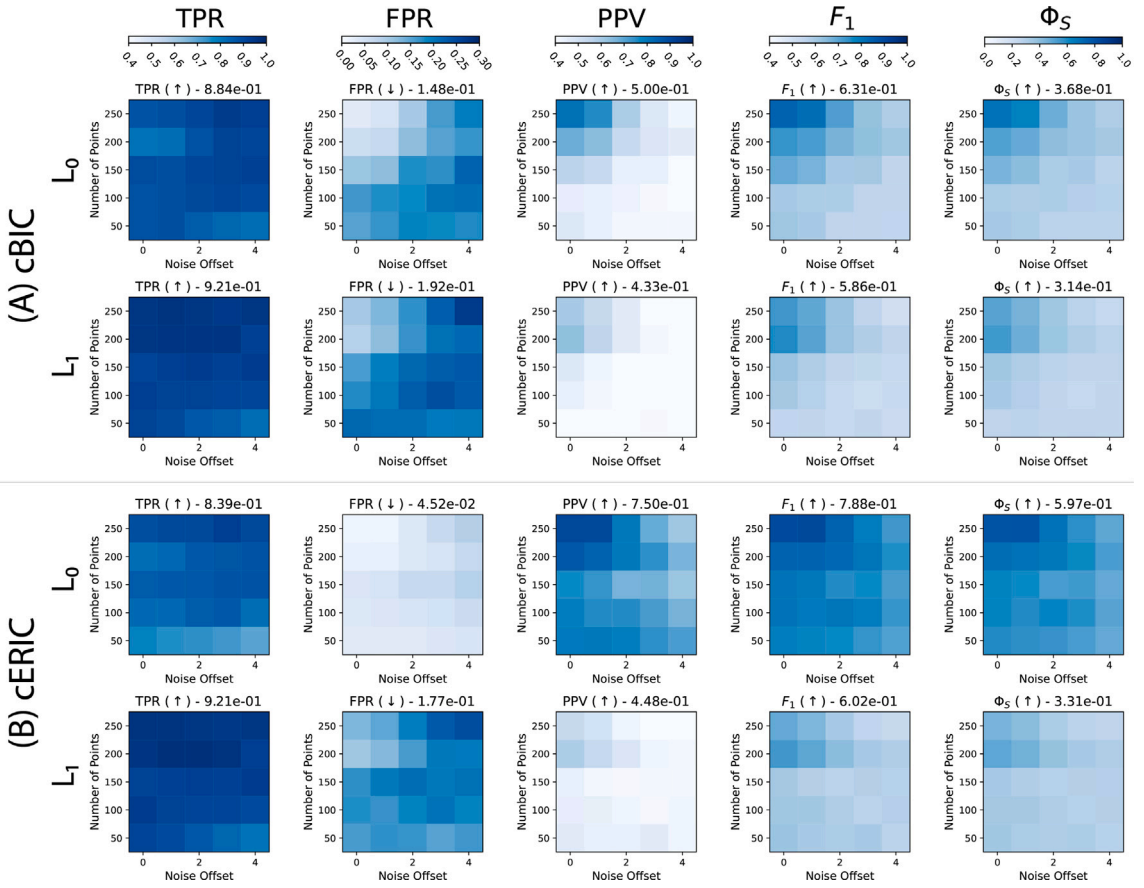


Fig. 10. Variable-selection performance for a Thomas point process with localization uncertainty (scenario **T1**) using composite information criteria with parameter estimation. We show the mean (over 100 independent repetitions of each experiment) True Positive Rate (TPR), False Positive Rate (FPR), Positive Predictive Value (PPV), F_1 score, and feature-selection stability Φ_S for model selection using the cBIC **(A)** and cERIC **(B)** with adaptive L_0 (top row of each subfigure) and adaptive L_1 (bottom row of each subfigure) penalties. The parameters of the pair-correlation function are estimated using minimum-contrast estimation of the K-function. Each panel shows a performance metric (top titles, color bars) for different noise magnitudes c (x-axis) and sample sizes $\mathbb{E}N(W)$ (y-axis). The average metrics over all 25 experiments are given in the panel titles with arrows (↑ / ↓) indicating the direction of improvement.

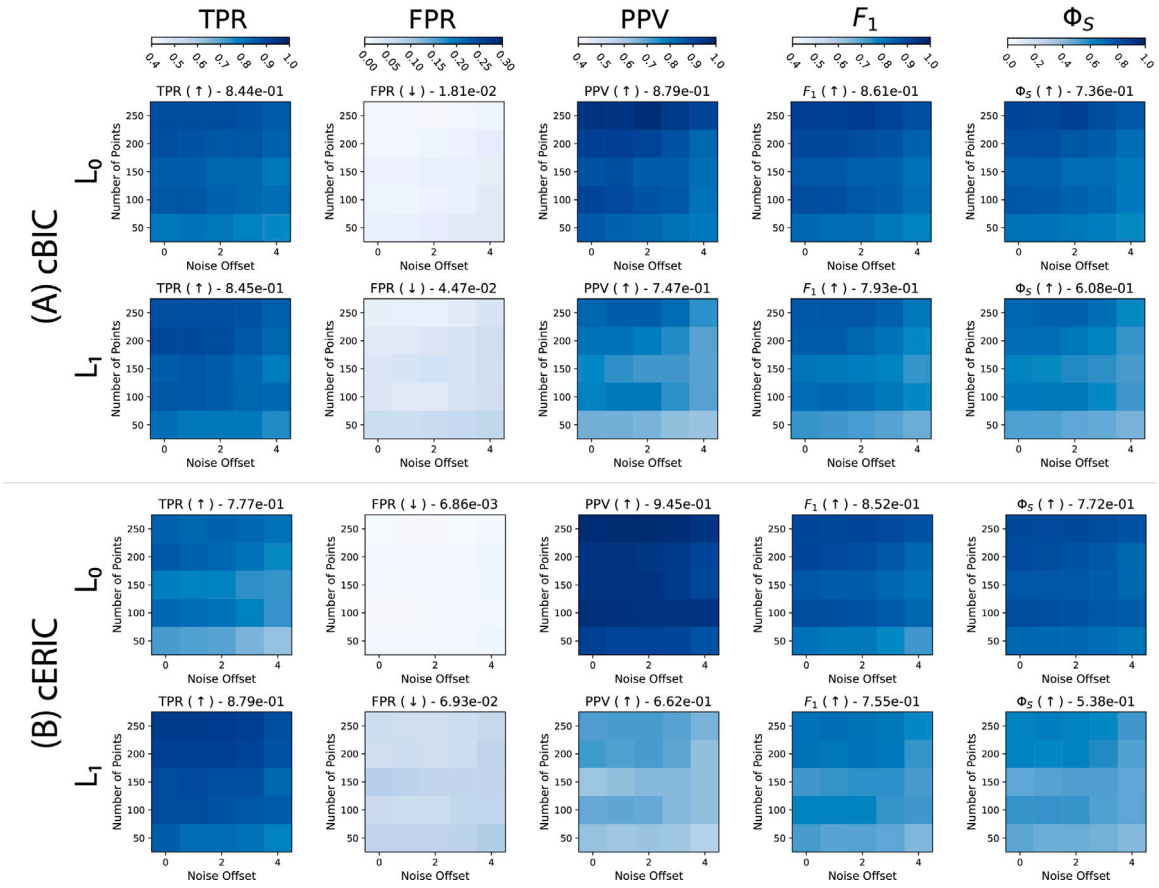


Fig. 11. Variable-selection performance for a Thomas point process with localization uncertainty (scenario **T1**) using composite information criteria with exact knowledge. We show the mean (over 100 independent repetitions of each experiment) True Positive Rate (TPR), False Positive Rate (FPR), Positive Predictive Value (PPV), F_1 score, and feature-selection stability Φ_S for model selection using the cBIC **(A)** and cERIC **(B)** with adaptive L_0 (top row of each subfigure) and adaptive L_1 (bottom row of each subfigure) penalties. The parameters of the pair-correlation function are assumed to be known exactly. Each panel shows a performance metric (top titles, color bars) for different noise magnitudes c (x-axis) and sample sizes $\mathbb{E}N(W)$ (y-axis). The average metrics over all 25 experiments are given in the panel titles with arrows (↑ / ↓) indicating the direction of improvement.

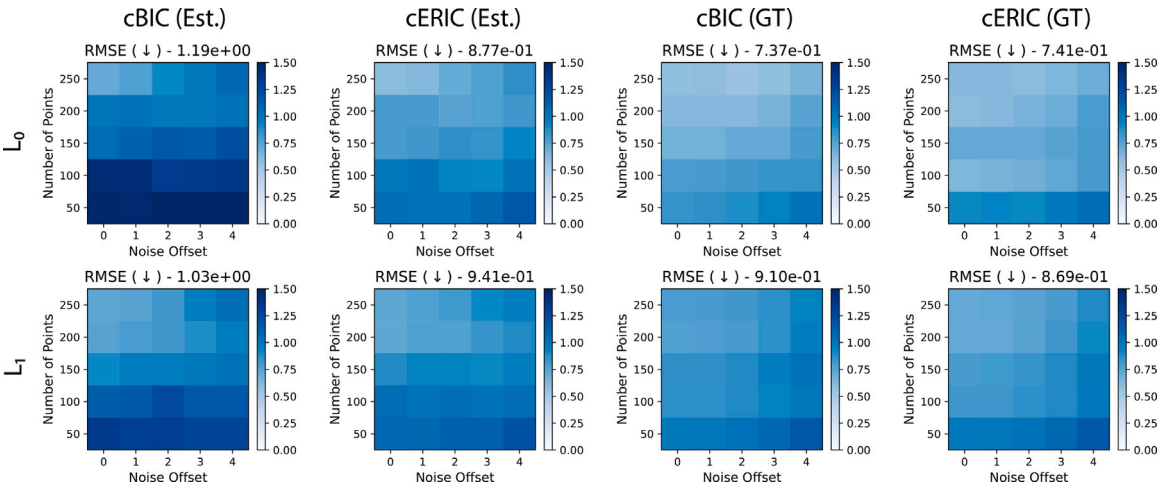


Fig. 12. Parameter estimation performance quantified by the RMSE of the selected models in scenario **T1** (Fig. 8) when using composite information criteria as shown in Figs. 10 (Est.) and 11 (GT). The averages over all 25 experiments are given in the panel titles.

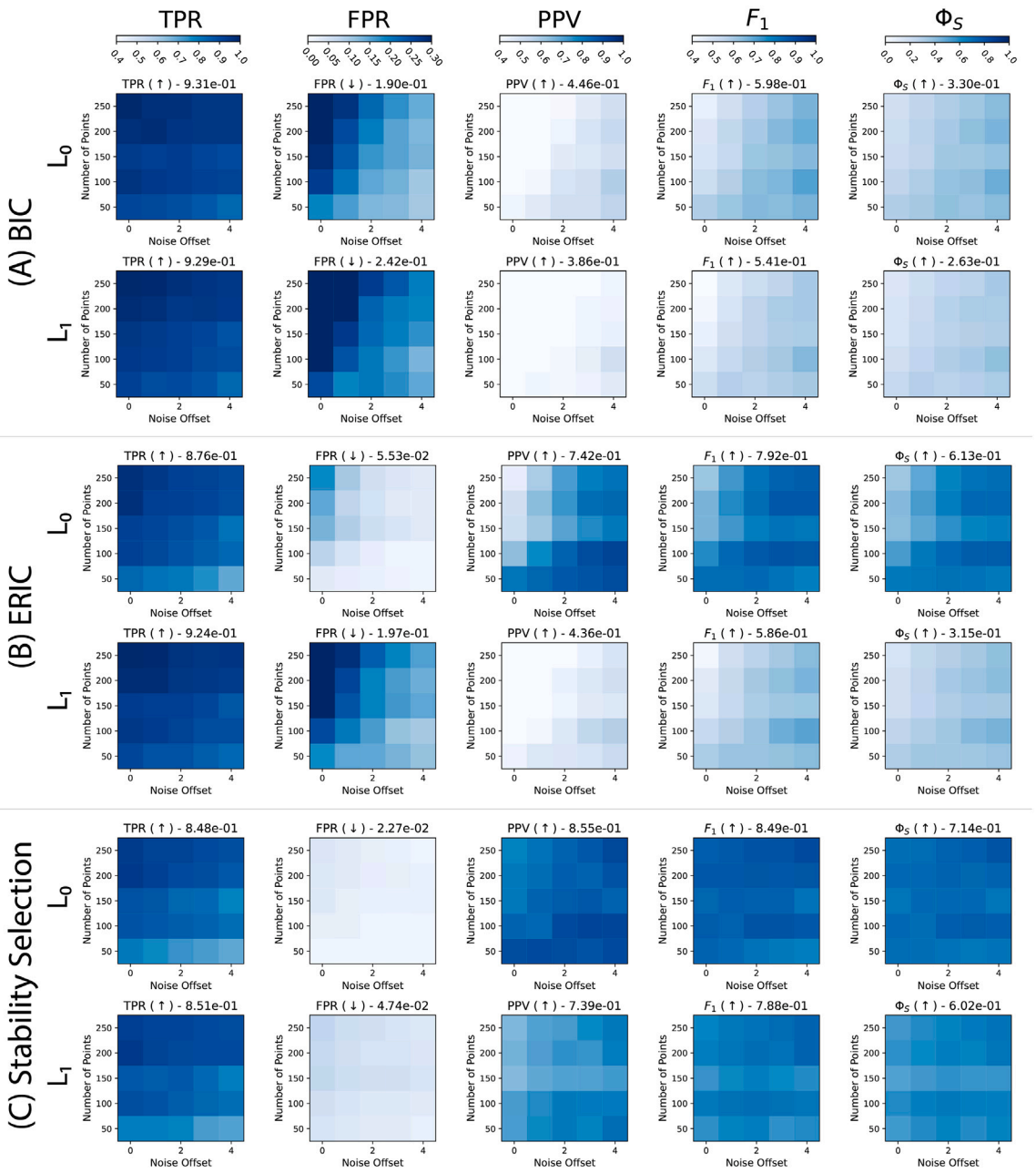


Fig. 13. Variable-selection performance for a Thomas point process with detection uncertainty (scenario **T2**). We show the mean (over 100 independent repetitions of each experiment) True Positive Rate (TPR), False Positive Rate (FPR), Positive Predictive Value (PPV), F_1 score, and feature-selection stability Φ_S for model selection using BIC **(A)**, ERIC **(B)**, and stability selection with PFER ≤ 1 **(C)** with adaptive L_0 (top row of each subfigure) and adaptive L_1 (bottom row of each subfigure) penalties. Each panel shows a performance metric (top titles, color bars) for different noise magnitudes c (x-axis) and sample sizes $\mathbb{E}N(W)$ (y-axis). The average metrics over all 25 experiments are given in the panel titles with arrows (↑ / ↓) indicating the direction of improvement.

homogeneous model where clustering is explained by the estimated K-function. This illustrates the potentially detrimental effect of feedback between the first- and second-order models in the estimation procedure, which can lead to suboptimal variable selection in the presence of noise. Such feedback is not present in stability selection, as it directly estimates the parameters of the intensity function directly from the data without requiring a second-order model.

Fig. 11 shows the performance of cBIC and cERIC when the pair-correlation function is assumed to be known exactly. The performance is always better than when estimating the second-order parameters from data. This is expected, as the estimation

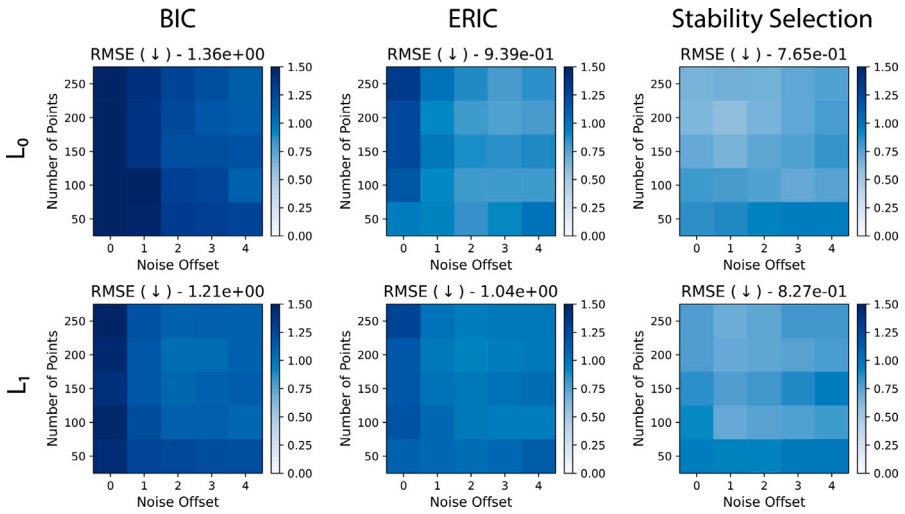


Fig. 14. Parameter estimation performance quantified by the RMSE of the selected models in scenario **T2** (Fig. 13) when using BIC, ERIC, and stability selection with either the L_1 or L_0 penalty. The averages over all 25 experiments are given in the panel titles.

procedure introduces additional errors, especially for small samples or high noise where the clustering is less apparent in the data. We also observe that with perfect knowledge, cBIC performs better than cERIC in the average F_1 score but achieves lower selection stability Φ_S, indicating higher variance in the selected models. This is a consequence of the additional regularization in cERIC, through the penalization weight λ, leading to more conservative selection. Indeed, the higher PPV of cERIC indicates that it selects only the most relevant covariates, ignoring smaller effects. This could also explain the higher feature-selection stability. Overall, cBIC achieves a better trade-off in this setting, which explains its higher average F_1 score. When using the Lasso, PPV drops for cERIC. In this case, cERIC selects lower λ values than cBIC, which leads to more complex models due to the soft thresholding of the Lasso. Similar observations have been made when using the adaptive Lasso and cERIC for Gibbs point processes (Ba and Coeurjolly, 2023). However, this effect seems to be mitigated by the L_0 penalty, which leads to less smooth coefficient paths than the Lasso (see Fig. 2). In the λ-ranges over which the estimated model does not change due to the hard thresholding of the L_0 penalty, cERIC can better prioritize models with higher penalization due to the λ-weighting.

Even in this best-case scenario, stability selection performs comparably with cBIC and cERIC. The average F_1 score of stability selection is 3% lower than for cBIC when using the L_0 penalty and around 4% lower with the L_1 penalty. Feature-selection stability is between 8% and 12% lower than for cBIC and cERIC. For the L_1 penalty, stability selection performs better than cERIC (by 1%) but worse than cBIC (by 4%). However, we observe that the RMSE slightly improves when using stability selection compared to cBIC and cERIC (Fig. 12). This, in combination with the higher TPR and smaller PPV of stability selection, suggests that stability selection produces more false positives than composite information criteria with perfect knowledge of the second-order structure of the process. In practical applications, where perfect second-order knowledge is not available, the performance of composite information criteria rapidly deteriorates. Even for high-quality estimates from data (Fig. 10, parameter estimate for ground-truth analytical form), cBIC and cERIC fall behind. Stability selection therefore presents a good choice in practice. It remains robust to noise without requiring additional knowledge of the true process.

We repeat the same experiments for a Thomas point process with detection uncertainty (scenario **T2**). The results are shown in Fig. 13. Like for scenario **T1**, the overall performance is again lower than in the Poisson case. In contrast to **T1**, but similar to **P2**, variable-selection performance is better than for localization uncertainty, and it generally improves with increasing detection noise magnitude. We hypothesize that this is again because thinning creates a more regular process. This again particularly impacts BIC and ERIC, with ERIC achieving better performance than BIC. The best performance overall is achieved by stability selection with L_0 penalty. Stability selection is also the least sensitive to noise and sample size. Across methods, the adaptive L_0 penalty performs better in all metrics than the adaptive L_1 penalty, except for the TPR. This is because the adaptive Lasso includes more covariates overall, also leading to higher FPRs. Using stability selection instead of information criteria, however, reduces the FPR of the Lasso by an order of magnitude while also reducing the variance in the estimated models.

For the RMSE of the parameter estimates (Fig. 14), we again observe similar trends as for the variable-selection performance. The RMSE is generally lower for the adaptive L_0 penalty and when using stability selection. Interestingly, when using stability selection, the RMSE for scenario **T2** is higher than for **T1**, unlike the variable-selection performance. This suggests that, while thinning noise can lead to more accurate support recovery, the resulting Poisson likelihood estimates still have higher errors.

Fig. 15 shows the variable-selection performance and Fig. 17 the RMSE of cBIC and cERIC in scenario **T2** when the second-order parameters are estimated using minimum-contrast estimation. While thinning noise increases the local regularity of the point pattern, which tends to underestimate pair correlations, we still observe an improvement in performance when using composite information criteria over standard information criteria. This is especially visible for the cBIC, where the average F_1 score is around 12% or

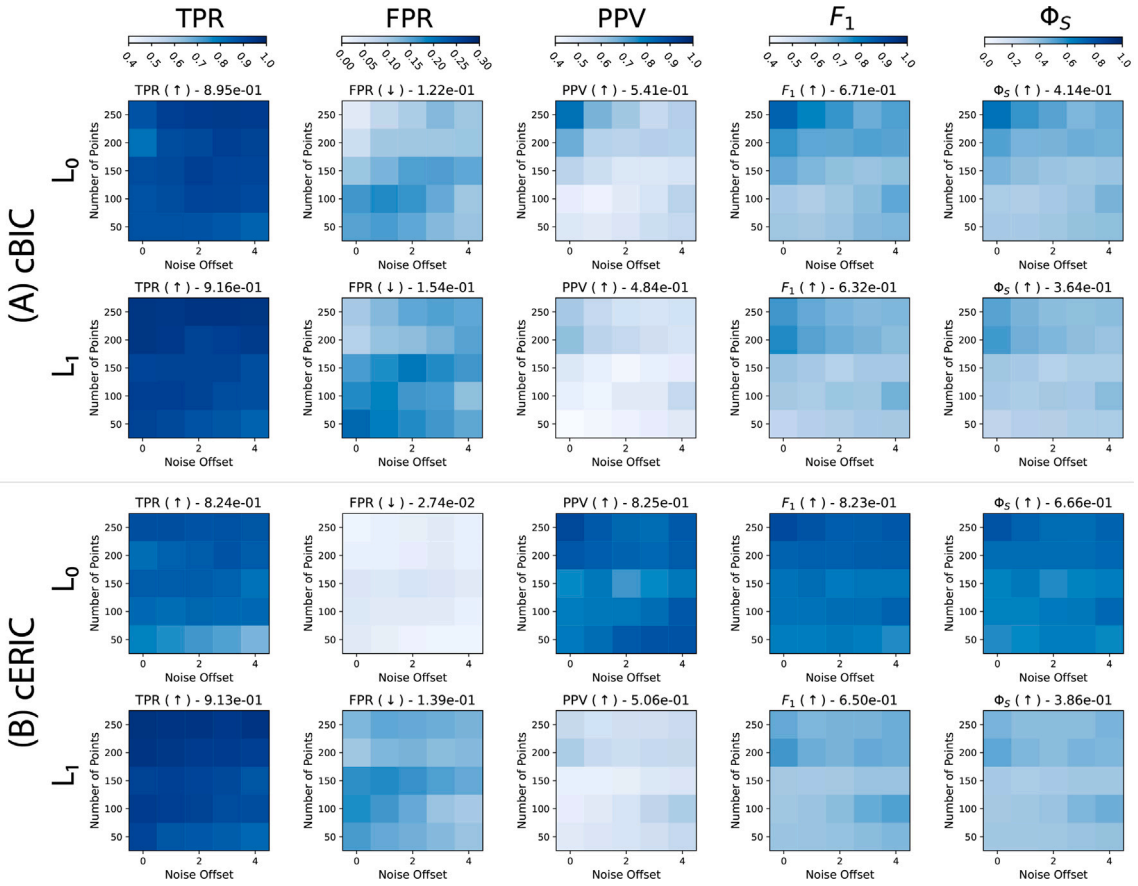


Fig. 15. Variable-selection performance for a Thomas point process with detection uncertainty (scenario **T2**) using composite information criteria with parameter estimation. We show the mean (over 100 independent repetitions of each experiment) True Positive Rate (TPR), False Positive Rate (FPR), Positive Predictive Value (PPV), F_1 score, and feature-selection stability Φ_S for model selection using the cBIC **(A)** and cERIC **(B)** with adaptive L_0 (top row of each subfigure) and adaptive L_1 (bottom row of each subfigure) penalties. The parameters of the pair-correlation function are estimated using minimum-contrast estimation of the K-function. Each panel shows a performance metric (top titles, color bars) for different noise magnitudes c (x-axis) and sample sizes $\mathbb{E}N(W)$ (y-axis). The average metrics over all 25 experiments are given in the panel titles with arrows (↑ / ↓) indicating the direction of improvement.

17% higher than for BIC with L_0 or L_1 penalties, respectively. The cERIC achieves an improvement over the ERIC of 4% and 11%, respectively, for the L_0 and L_1 penalties. However, no composite information criterion achieves the performance or robustness of stability selection. While the F_1 score of cERIC with L_0 penalty is comparable to that of stability selection, stability selection achieves higher feature-selection stability Φ_S.

Like for the previous scenario **T1**, we repeat the analysis with perfect knowledge of the pair-correlation function of the Thomas process under detection uncertainty. This constitutes the best case for cBIC and cERIC. The results are shown in Fig. 16. Again, like in scenario **T1**, performance is better than when using a data-estimated pair-correlation function. The cBIC again outperforms cERIC in terms of the average F_1 score for both penalties. Like for localization uncertainty, the λ-weighting in cERIC leads to a more conservative selection of covariates under the L_0 penalty and less penalization under the L_1 penalty. Also in this case, stability selection achieves near-best performance without requiring knowledge of the pair correlation function.

3.2. Application to a forestry data set

To illustrate the practical applicability of the proposed method, we consider the well-known data from the Barro Colorado Island (BCI) research plot in Panama. Over a 50 ha site (1000 m × 500 m), the locations of tree stems with at least 1 cm diameter at breast height have been recorded. This results in a data set of over 350,000 trees from 300 species (Condit, 1998; Hubbell et al., 1999). A central question is how so many species are able to coexist, and how they carve environmental niches. Therefore, one aims to identify environmental covariates — such as elevation or soil nutrients — that explain the distribution of tree species. This data set

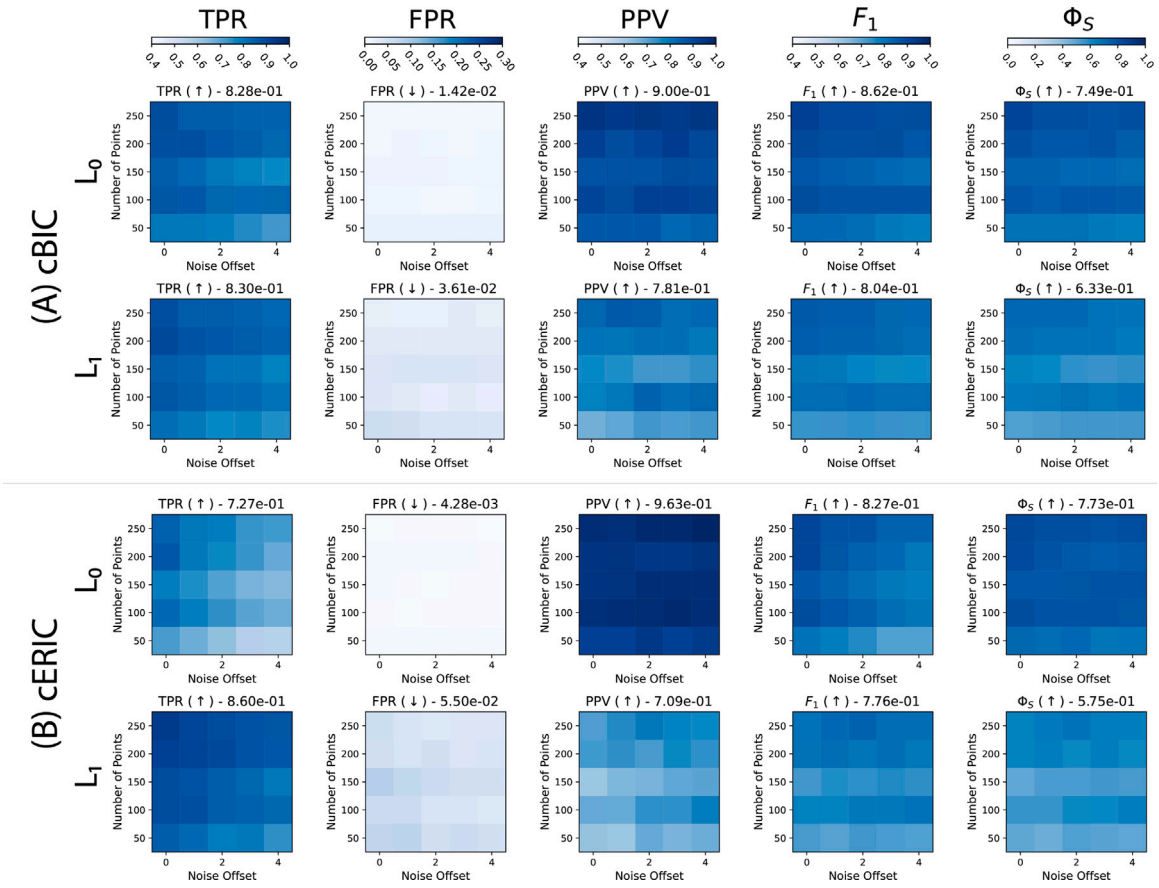


Fig. 16. Variable-selection performance for a Thomas point process with detection uncertainty (scenario **T2**) using composite information criteria with exact knowledge. We show the mean (over 100 independent repetitions of each experiment) True Positive Rate (TPR), False Positive Rate (FPR), Positive Predictive Value (PPV), F_1 score, and feature-selection stability Φ_S for model selection using the cBIC **(A)** and cERIC **(B)** with adaptive L_0 (top row of each subfigure) and adaptive L_1 (bottom row of each subfigure) penalties. The parameters of the pair-correlation function are assumed to be known exactly. Each panel shows a performance metric (top titles, color bars) for different noise magnitudes c (x-axis) and sample sizes $\mathbb{E}N(W)$ (y-axis). The average metrics over all 25 experiments are given in the panel titles with arrows (↑ / ↓) indicating the direction of improvement.

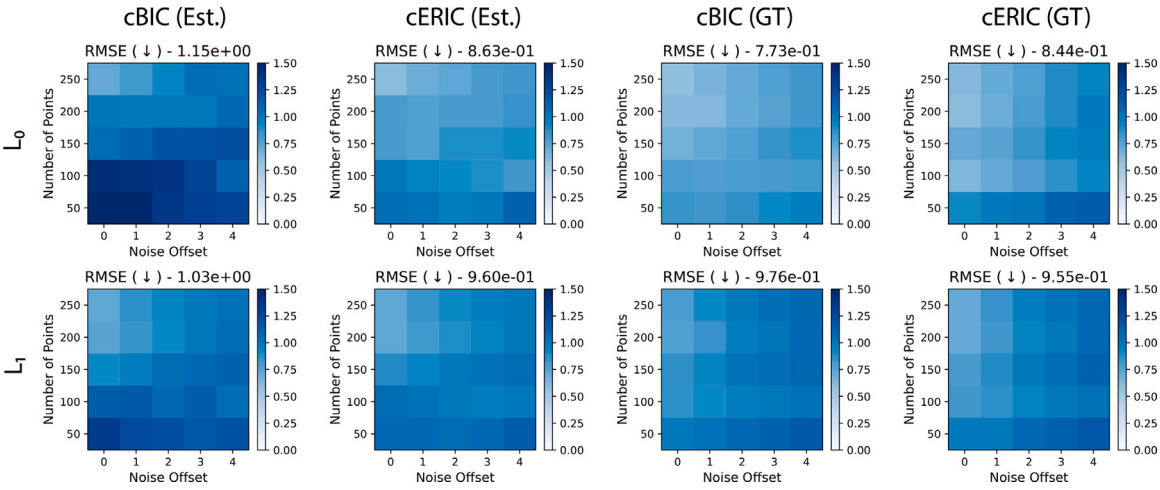


Fig. 17. Parameter estimation performance quantified by the RMSE of the selected models in scenario **T2** (Fig. 8) when using composite information criteria as shown in Figs. 15 (Est.) and 16 (GT). The averages over all 25 experiments are given in the panel titles.

Table 2
Stability-selection results for the BCI data set using adaptive Lasso (L_1) and adaptive Best-Subset (L_0) penalties. The table shows the estimated effect sizes (standardized data) for PFER $\leq 1, 2, 3$. The effect sizes are estimated over the support identified by stability selection by solving the unpenalized composite likelihood problem on the whole data set. The considered covariates are: elevation (Elev.), slope (Slope), aluminum (Al), boron (B), calcium (Ca), copper (Cu), iron (Fe), potassium (K), magnesium (Mg), manganese (Mn), phosphorus (P), zinc (Zn), nitrogen (N), mineralized nitrogen (N(min)), and soil pH (pH). The last row shows the number of selected covariates $\left\|\hat{\beta}\right\|_0 = \sum_{j=1}^{p} \mathbf{1}(\hat{\beta}_j \neq 0)$ for the respective penalty and error bound. For each PFER, the first sub-column uses the full data set, while the second sub-column uses a thinned point pattern with $p_{\text{thin}} = 0.1$.

	Adaptive Lasso (L_1)						Adaptive Best Subset (L_0)					
	PFER ≤ 1		PFER ≤ 2		PFER ≤ 3		PFER ≤ 1		PFER ≤ 2		PFER ≤ 3	
	Full	Thin	Full	Thin	Full	Thin	Full	Thin	Full	Thin	Full	Thin
Elev.	0.35	0.37	0.39	0.37	0.38	0.36	0.35	0	0.35	0.37	0.36	0.37
Slope	0.33	0.25	0.27	0.25	0.30	0.27	0.33	0	0.33	0.25	0.33	0.25
Al	0	0	0	0	0	0	0	0	0	0	0	0
B	0	0	0.40	0	0.22	0.21	0	0	0	0	0	0
Ca	0	0	0	0	0	0	0	0	0	0	0	0
Cu	0	0	0	0	0	0	0	0	0	0	0	0
Fe	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0	0	0
Mg	0	0	0	0	0	0	0	0	0	0	0	0
Mn	0	0	0	0	0.26	0	0	0	0	0	0.33	0
P	−0.59	−0.58	−0.68	−0.58	−0.61	−0.65	−0.59	−0.59	−0.59	−0.58	−0.54	−0.58
Zn	−0.28	0	−0.58	0	−0.57	0	−0.28	0	−0.28	0	−0.45	0
N	0	0	0	0	0	0	0	0	0	0	0	0
N(min)	0	0	0	0	0	−0.36	0	0	0	0	0	0
pH	0	0	0	0	0	0	0	0	0	0	0	0
$\left\|\hat{\beta}\right\|_0$	4	3	5	3	6	5	4	1	4	3	5	3

has already been used in previous studies to identify sparse sets of predictors (Choiruddin et al., 2018; Ba and Coeurjolly, 2023), which allows direct comparison of results.

We focus on the locations of the 3604 *Beilschmiedia pendula* (BPL) trees from the *Lauraceae* family, which we interpret as a realization of a spatial point process (see Fig. 18, top-left). We aim to model the intensity function of the point process using the 15 covariates listed in Table 2. As in Section 3.1, we interpolate all covariates onto a common grid of size 201×101, which we also use a quadrature points. The observation window corresponds to the entire 50 ha region, $W = [0, 1000] \times [0, 500]$. All covariates are standardized as in previous analyses (Choiruddin et al., 2018; Ba and Coeurjolly, 2023).

We compare the models identified by stability selection with adaptive L_1 and adaptive L_0 penalties. The PGD step size is $\gamma = 10^{-4}$ in both cases, which ensures numerical stability. For stability selection, we obtain 50 bootstrap samples by p-thinning with $p_{\text{thin}} = 0.5$. The λ-path is chosen so as to achieve a desired PFER bound according to Eq. (18) with $\pi_{\text{th}} = 0.9$. The resulting candidate interval for λ spans six orders of magnitude and is discretized with 40 log-equidistant points. The $\lambda_{\max}$ is set to obtain an empty model under the respective penalty. Using these settings, the stability path for the L_1 penalty is computed in 20s and for the L_0 penalty in 38s on a personal laptop (Apple MacBook Pro 2023, Apple M3 Pro CPU, 36 GB LPDDR5 RAM). The higher computational cost of the L_0 penalty is due to the fixed step sizes in contrast to the BB-PGD used for the L_1 penalty. Table 2 reports the resulting models for PFER $\leq 1, 2, 3$ and compares the two penalization methods. The final parameter estimates are obtained by solving the unpenalized composite likelihood problem for the identified set of stable predictors using all observations.

With either penalty, stability selection yields similar models. As expected, the adaptive Lasso (L_1) selects slightly more covariates than adaptive Best-Subset selection (L_0). For PFER ≤ 1, both penalties find the same covariates elevation (Elev.), slope (Slope), phosphorus (P), and zinc (Zn). This agrees with previous results on this data set (Choiruddin et al., 2018). These covariates are visualized in Fig. 18 together with the tree location pattern, visually supporting the conclusion that BPL trees prefer higher elevation and slopes with low concentration of phosphorus and zinc.

The influence of error control can nicely be seen in the number of selected covariates monotonically increasing for looser bounds and does not require recomputing the path. At higher PFER, the L_0 models only slowly include additional terms, with the first one being manganese (Mn) at PFER ≤ 3. The L_1 models first and additionally include boron (B). In this case, the effect sizes are similar as previously reported using weighted composite likelihood estimates, which rely on estimating the pair-correlation function (Choiruddin et al., 2018). Stability selection achieves the same level of sparsity without modeling assumptions about the underlying process, and without the need for estimating the pair-correlation function. The similarity of the obtained models hints at the practical utility of stability selection for variable selection in spatial point processes.

To better mimic the data-limited scenario that motivates our work, we thin the original point pattern with $p_{\text{thin}} = 0.1$ to simulate studying a rare species. The resulting pattern of 345 points is shown in Fig. 18 (bottom-left). Following the same procedure as for the full data set, we obtain the models shown in the second sub-column of each PFER bound in Table 2. Stability selection consistently selects fewer terms for the thinned data. This is due to higher uncertainty in the candidate models and error control thus enforcing

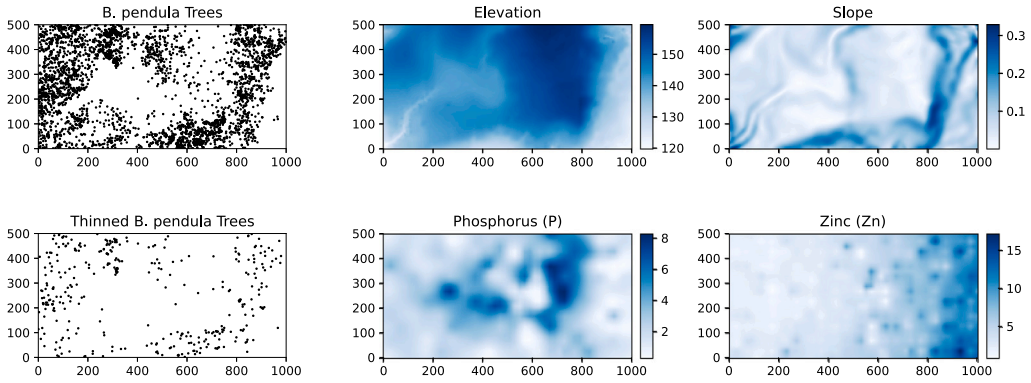


Fig. 18. Point pattern of *B. pendula* tree stem locations in the BCI research plot (black dots, top-left) and a thinned version with $p_{\text{thin}} = 0.1$ (black dots, bottom-left) with visualizations of the selected covariates from Table 2 for PFER ≤ 1. The covariates are: elevation (Elev.), slope (Slope), phosphorus (P), and zinc (Zn).

more regularization. For the L_0 penalty, stability selection always identifies a subset of the covariates identified for the full data set. The L_1 penalty chooses a not previously selected term (N(min)) for high PFER. For the L_0 penalty, we find only phosphorus (P) for PFER ≤ 1. For PFER ≤ 2 and PFER ≤ 3, elevation (Elev.) and slope (Slope) are additionally selected. This can be attributed to the stronger penalization leading to lower selection frequencies for smaller effects. The results confirm that stability selection remains useful in the data-limited regime, which is often the case when studying rare species.

4. Discussion and outlook

We presented a method for sparse intensity estimation of spatial point processes under noise. We showed that noise, in the form of localization or detection uncertainty, significantly influences variable-selection performance. We proposed stability selection as a noise-robust variable-selection method for spatial point processes, providing a flexible method that can be combined with different sparsity-inducing penalties without requiring additional knowledge about the true process or its correlation structure. We benchmarked stability selection with both the familiar adaptive Lasso (L_1 penalty) and the more difficult-to-solve adaptive best-subset selection (L_0 penalty) to illustrate the flexibility of the approach. Despite the non-convex and combinatorial nature of the L_0 penalty, we observed that the L_0 penalty generally achieves better variable-selection performance than the L_1 penalty and avoids shrinkage bias. Simulation studies showed that the proposed method provides stable estimates and allows for error control in the variable-selection procedure. This improves variable-selection performance under noise, for small samples, and under model misspecification.

The presented method uses p-thinning for generating the bootstrap subsamples required for stability selection. The estimating equation in Eq. (19) allows for straightforward integration of the presented method into existing frameworks for spatial point processes that use generalized linear model software (Baddeley et al., 2016). The computational overhead incurred by our method scales linearly with the number of bootstrap samples K, where $K = 50$ was sufficient in our benchmarks. While the compute time depends on many factors, including the number of covariates, the discretization grid size, the step size, the convergence criteria, and the quality of the warm starts when computing the λ-path, the times for the forestry dataset were below one minute per case.

We compared the proposed method to existing model-selection procedures based on (composite) information criteria. The present method consistently outperformed information criteria both in terms of stability and accuracy. Typically, model selection based on information criteria produced more false positives. For the correlated Thomas point process, stability selection performed almost as well as composite information criteria with perfect knowledge of the pair-correlation function of the process, albeit without requiring such knowledge. This suggests that stability selection is able to cope with simultaneous clustering and noise in the data without having to explicitly model pair correlations or noise statistics. This is of practical importance, since such knowledge is not usually available in applications, and it is difficult to estimate for small samples. But even when second-order information could be estimated from the data, the performance of composite information criteria markedly dropped. However, it was still better than using the standard BIC or ERIC, which we deem not advisable for correlated processes or noisy observations.

Our results also suggested that the adaptive L_0 penalty, despite its non-convexity, generally achieves better performance than the adaptive L_1 penalty at comparable computational cost, especially in terms of the false-positive rate and F_1 score. This indicates that proximal gradient descent is able to effectively find sparse local minima that represent good-enough estimates of the true process. Choosing between L_0 and L_1 penalties allows trading off between type I and type II errors, since L_0 generally selects fewer variables. Stability selection provides control over the PFER for either penalization. The empirical PFER was either below or close to the imposed threshold in all presented simulations. This error control filters out bad local minima, leading to greatly reduced

false-positive rates, particularly when using the adaptive Lasso. Existing estimators based on the Lasso can therefore directly benefit from stability selection.

Finally, we illustrated the applicability of the proposed method on a real-world forestry data set, identifying the relevant covariates for the spatial distribution of *Beilschmiedia pendula* trees in a tropical rain forest area of 50 ha. The identified models were in line with previous results, validating the method. Comparing different error bounds and penalties also confirmed the statistical consistency of the proposed method, with looser error bounds leading to monotonically larger models. This indicates that stability selection can be used to identify a sparse set of relevant covariates in spatial point processes without any assumptions on the underlying model.

While these results are encouraging, they also highlight the need for future work. An obvious limitation of our work is that we only considered uncorrelated noise. Real-world data, however, often contains structured noise. Examples include presence-only analysis of plants and animals in ecological studies, where they are more likely to be spotted near roads, or digital imagery recorded with CMOS or CCD sensors, which generate correlated readout noise. The effect of structured noise on stability selection remains to be studied.

In order to establish a baseline, we here only considered the classic formulation of stability selection. It would be interesting in the future to extend the presented approach to more recent stability-selection variants, such as complimentary-pairs stability selections (Shah and Samworth, 2013). While we have empirically shown the effectiveness of L_0 penalization for variable selection in spatial point processes, further study of its theoretical properties would be interesting. It would also be worthwhile to explore the use of stability selection for estimating interactions across multiple spatial scales in multivariate point process models, particularly when combined with group sparsity as suggested by Rajala et al. (2018). Group-sparse multivariate models are relevant in biology and ecology, where many (molecular) species interact with each other and with their environment across scales. Stability selection may offer a systematic approach to identifying interaction structures in such settings, while being robust to noise in the data.

Finally, future work could extend the methodology proposed here to other types of point processes. Since similar discretizations of composite likelihoods are, for example, also used for Gibbs point processes (Baddeley and Turner, 2000; Ba and Coeurjolly, 2023). We therefore think that the presented method could be adapted to estimating the Papangelou conditional intensity in those models as well.

Despite these open questions, we believe that the idea of applying stability selection to spatial point-process modeling opens several doors. We hope that the present work laid the foundations by systematically benchmarking the method, deriving the estimating equations, and establishing the effectiveness of proximal gradient descent for non-convex L_0 penalties.

Acknowledgments

The authors thank Dr. Nandu Gopan and Dr. Abhishek Behera (both Sbalzarini group) for helpful discussions. The authors acknowledge financial support by the German Federal Ministry of Research, Technology and Space and by the Saxon State Ministry of Science, Culture, and Tourism in the program “Center of Excellence for AI-research” — “Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig”, project identification number: ScaDS.AI. This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) under Germany’s Excellence Strategy — Cluster of Excellence EXC2068 “Physics of Life” of TU Dresden.

Appendix. Extended simulation study results ($p = 28$)

We extend the simulation study of the Thomas point process with location uncertainty (scenario **T1**) by considering a larger set of covariates. We follow the same simulation setup as described in Section 3.1 but consider additional interactions between the 13 soil nutrients, resulting in a total of 28 covariates. The additional interactions are constructed as described by Ba and Coeurjolly (2023).

The results are shown in Fig. A.19. The overall trends are similar to the case with $p = 15$ covariates, with stability selection outperforming information criteria in all metrics except the TPR. However, the performance of all methods is generally lower than for $p = 15$, which is expected due to the higher problem dimensionality. Nevertheless, we find that also in this case the error control of stability selection limits the FPR, leading to better overall performance. The adaptive L_0 penalty again achieves better performance than the adaptive L_1 penalty. The reduced selection stability Φ_S for $p = 28$ compared to $p = 15$ indicates that the higher dimensionality leads to more variability in the selected models across repetitions. Still, stability selection achieves higher stability than information criteria in all cases.

However, while the error control of stability selection effectively limits the FPR, it does not achieve the imposed bound of PFER ≤ 1 in this case. This is best seen in Fig. A.20, comparing the empirical PFER for all experiments. For the L_1 penalty, the empirical PFER is mostly larger than the bound (indicated by red crosses in the figure) and reaches around 1.7 in the worst case, which is, however, still low compared to the total number of covariates ($p = 28$). The higher empirical PFER compared to the imposed bound can be attributed to the higher problem dimensionality, where it becomes more difficult to ensure small correlations between true and noise covariates, such that the assumption of the PFER bound — exchangeability of the selection of noise variables — is likely not met (Meinshausen and Bühlmann, 2010). For the L_0 penalty, we find that the empirical PFER is generally below the imposed bound, only reaching 1.01 in the worst case. This highlights the importance of the underlying selection method.

Overall, these results further support the conclusion that stability selection with an adaptive L_0 penalty is a robust and effective method for variable selection in spatial point processes under noise. Even if error control is not perfect, it still provides a principled way for model selection, which is crucial for interpretability.

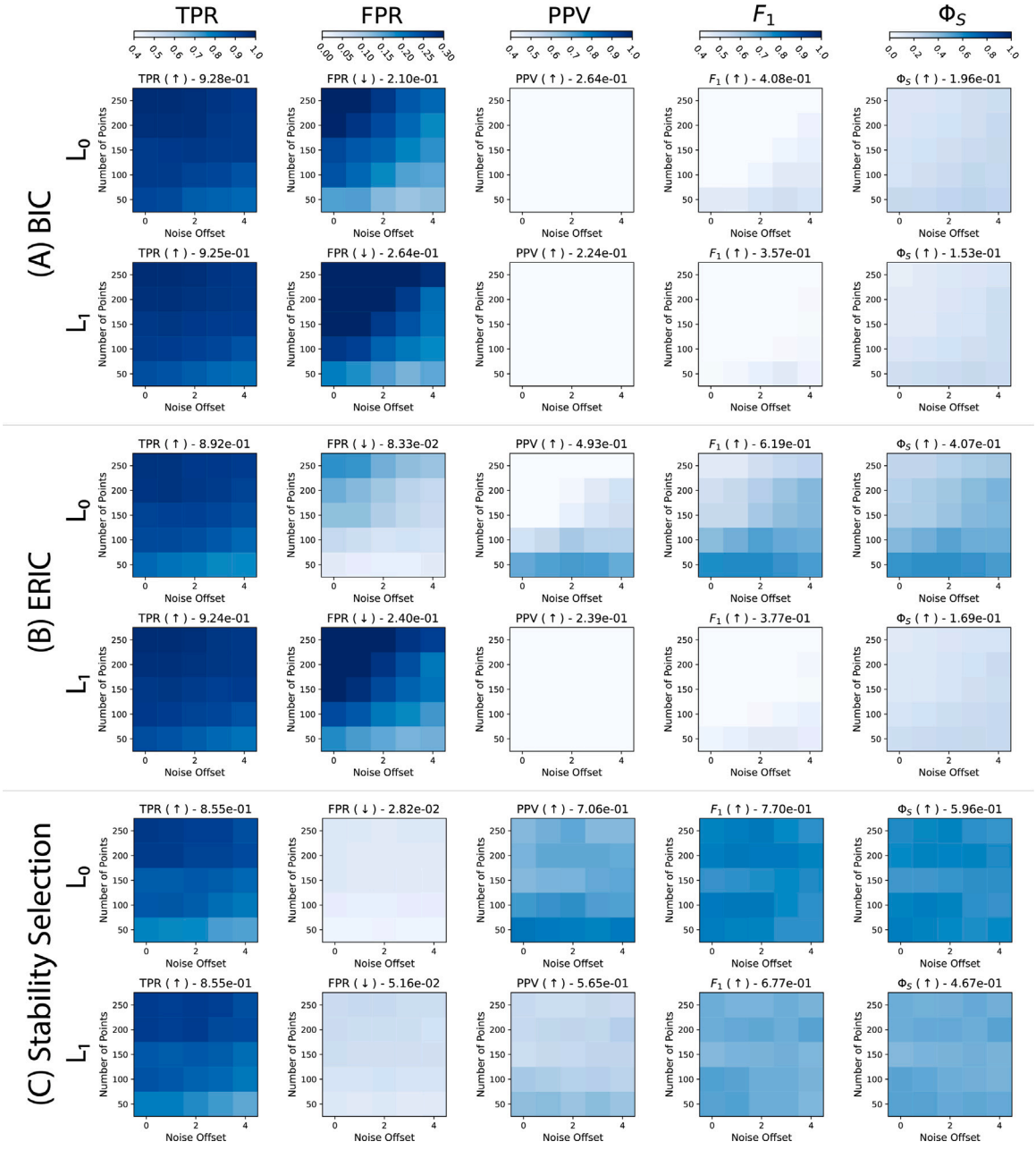


Fig. A.19. Variable-selection performance for a Thomas point process with localization uncertainty (scenario **T1**) using $p = 28$ covariates. We show the mean (over 100 independent repetitions of each experiment) True Positive Rate (TPR), False Positive Rate (FPR), Positive Predictive Value (PPV), F_1 score, and feature-selection stability Φ_S for model selection using the BIC **(A)**, ERIC **(B)**, and stability selection with PFER ≤ 1 **(C)** with adaptive L_0 (top row of each subfigure) and adaptive L_1 (bottom row of each subfigure) penalties. Each panel shows a performance metric (top titles, color bars) for different noise magnitudes c (x-axis) and sample sizes $\mathbb{E}N(W)$ (y-axis). The average metrics over all 25 experiments are given in the panel titles with arrows (↑ / ↓) indicating the direction of improvement.

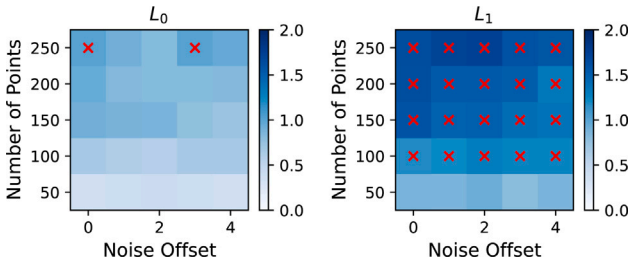


Fig. A.20. Empirical PFER compared with the desired bound PFER ≤ 1 when using stability selection for all experiments in scenario **T1** with $p = 28$ covariates. The red crosses indicate cases where the empirical PFER exceeds the desired bound.

Data availability

The data, models, and source codes that support the findings of this study are openly available at the URL/DOI given in Sturm and Sbalzarini (2025).

References

Assunção, R., Guttorp, P., 1999. Robustness for Inhomogeneous Poisson Point Processes. Ann. Inst. Statist. Math. 51 (4), 657–678. http://dx.doi.org/10.1023/A:1004079013014.

Ba, I., Coeurjolly, J.-F., 2023. Inference for low- and high-dimensional inhomogeneous Gibbs point processes. Scand. J. Stat. 50 (3), 993–1021. http://dx.doi.org/10.1111/sjos.12616, URL https://onlinelibrary.wiley.com/doi/10.1111/sjos.12616.

Bach, F., 2024. Learning Theory from First Principles. In: Adaptive Computation and Machine Learning series, MIT Press, URL https://books.google.de/books?id=R_T8EAAAQBAJ.

Baddeley, A., 2007. Spatial Point Processes and their Applications. In: Weil, W. (Ed.), Stochastic Geometry. In: Lecture Notes in Mathematics, vol. 1892, Springer Berlin Heidelberg, pp. 1–75. http://dx.doi.org/10.1007/978-3-540-38175-4_1, URL http://link.springer.com/10.1007/978-3-540-38175-4_1. Series Title: Lecture Notes in Mathematics.

Baddeley, A.J., Møller, J., Waagepetersen, R., 2000. Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. Stat. Neerl. 54 (3), 329–350. http://dx.doi.org/10.1111/1467-9574.00144, URL https://onlinelibrary.wiley.com/doi/10.1111/1467-9574.00144.

Baddeley, A., Rubak, E., Turner, R., 2016. Spatial point patterns: methodology and applications with r. Chapman & Hall / CRC Interdisciplinary Statistics, CRC Press, Taylor & Francis Group, Boka Raton London New York, http://dx.doi.org/10.1201/b19708.

Baddeley, A., Turner, R., 2000. Practical Maximum Pseudolikelihood for Spatial Point Patterns: (with Discussion). Aust. N. Z. J. Stat. 42 (3), 283–322. http://dx.doi.org/10.1111/1467-842X.00128, URL https://onlinelibrary.wiley.com/doi/10.1111/1467-842X.00128.

Barzilai, J., Borwein, J.M., 1988. Two-Point Step Size Gradient Methods. IMA J. Numer. Anal. 8 (1), 141–148. http://dx.doi.org/10.1093/imanum/8.1.141.

Beck, A., Teboulle, M., 2009. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. SIAM J. Imaging Sci. 2 (1), 183–202. http://dx.doi.org/10.1137/080716542, URL http://epubs.siam.org/doi/10.1137/080716542.

Berman, M., Turner, T.R., 1992. Approximating point process likelihoods with glim. J. R. Stat. Soc. Ser. C. Appl. Stat. 41 (1), 31–38. http://dx.doi.org/10.2307/2347614.

Blumensath, T., Davies, M.E., 2009. Iterative hard thresholding for compressed sensing. Appl. Comput. Harmon. Anal. 27 (3), 265–274. http://dx.doi.org/10.1016/j.acha.2009.04.002, URL https://www.sciencedirect.com/science/article/pii/S1063520309000384.

Bodinier, B., Filippi, S., Nøst, T.H., Chiquet, J., Chadeau-Hyam, M., 2023. Automated calibration for stability selection in penalised regression and graphical models. J. R. Stat. Soc. Ser. C. Appl. Stat. 72 (5), 1375–1393. http://dx.doi.org/10.1093/jrsssc/qlad058, URL https://academic.oup.com/jrsssc/article/72/5/1375/7223787.

Briz-Redón, A., 2024. Dealing with location uncertainty for modeling network-constrained lattice data. Spat. Stat. 59, 100807. http://dx.doi.org/10.1016/j.spasta.2023.100807, URL https://linkinghub.elsevier.com/retrieve/pii/S2211675323000829.

Choiruddin, A., Coeurjolly, J.-F., Letué, F., 2018. Convex and non-convex regularization methods for spatial point processes intensity estimation. Electron. J. Stat. 12 (1), http://dx.doi.org/10.1214/18-EJS1408, URL https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-12/issue-1/Convex-and-non-convex-regularization-methods-for-spatial-point-processes/10.1214/18-EJS1408.full.

Choiruddin, A., Coeurjolly, J.-F., Letué, F., 2023. Adaptive lasso and Dantzig selector for spatial point processes intensity estimation. Bernoulli 29 (3), 1849–1876. http://dx.doi.org/10.3150/22-BEJ1523, URL https://projecteuclid.org/journals/bernoulli/volume-29/issue-3/Adaptive-lasso-and-Dantzig-selector-for-spatial-point-processes-intensity/10.3150/22-BEJ1523.full.

Choiruddin, A., Coeurjolly, J.-F., Waagepetersen, R., 2021. Information criteria for inhomogeneous spatial point processes. Aust. N. Z. J. Stat. 63 (1), 119–143. http://dx.doi.org/10.1111/anzs.12327, URL https://onlinelibrary.wiley.com/doi/10.1111/anzs.12327.

Choiruddin, A., Cuevas-Pacheco, F., Coeurjolly, J.-F., Waagepetersen, R., 2020. Regularized estimation for highly multivariate log Gaussian Cox processes. Stat. Comput. 30 (3), 649–662. http://dx.doi.org/10.1007/s11222-019-09911-y, URL http://link.springer.com/10.1007/s11222-019-09911-y.

Choiruddin, A., Susanto, T.Y., Husain, A., Kartikasari, Y.M., 2024. Kppmenet: combining the kppm and elastic net regularization for inhomogeneous cox point process with correlated covariates. J. Appl. Stat. 51 (5), 993–1006. http://dx.doi.org/10.1080/02664763.2023.2207786, arXiv:https://doi.org/10.1080/02664763.2023.2207786.

Coeurjolly, J.-F., Ba, I., Choiruddin, A., 2023. Regularization techniques for inhomogeneous (spatial) point processes intensity and conditional intensity estimation. URL http://arxiv.org/abs/2305.13470. arXiv:2305.13470 [math, stat].

Coeurjolly, J.-F., Møller, J., Waagepetersen, R., 2017. A Tutorial on Palm Distributions for Spatial Point Processes. Int. Stat. Rev. 85 (3), 404–420. http://dx.doi.org/10.1111/insr.12205, URL https://onlinelibrary.wiley.com/doi/10.1111/insr.12205.

Condit, R., 1998. Tropical Forest Census Plots. Springer Berlin Heidelberg, Berlin, Heidelberg, http://dx.doi.org/10.1007/978-3-662-03664-8, URL http://link.springer.com/10.1007/978-3-662-03664-8.

Cronie, O., Moradi, M., Biscio, C.A.N., 2024. A cross-validation-based statistical theory for point processes. Biometrika 111 (2), 625–641. http://dx.doi.org/10.1093/biomet/asad041, URL https://academic.oup.com/biomet/article/111/2/625/7208865.

Diggle, P.J., 2013. Statistical Analysis of Spatial and Spatio-Temporal Point Patterns, third ed. Monographs on Statistics and Applied Probability 128, Chapman and Hall/CRC, New York, http://dx.doi.org/10.1201/b15326, URL https://www.taylorfrancis.com/books/9781466560246.

Gillespie, L.E., Ruffley, M., Exposito-Alonso, M., 2024. Deep learning models map rapid plant species changes from citizen science and remote sensing data. Proc. Natl. Acad. Sci. 121 (37), e2318296121. http://dx.doi.org/10.1073/pnas.2318296121, URL https://pnas.org/doi/10.1073/pnas.2318296121.

Gong, P., Zhang, C., Lu, Z., Huang, J., Ye, J., 2013. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In: Dasgupta, S., McAllester, D. (Eds.), Proceedings of the 30th International Conference on Machine Learning. PMLR, Atlanta, Georgia, USA, pp. 37–45, URL https://proceedings.mlr.press/v28/gong13a.html.

Guttorp, P., Illian, J., Kostensalo, J., Kuronen, M., Myllymäki, M., Särkkä, A., Thorarinsdottir, T.L., 2023. What you see is not what is there: Mechanisms, models, and methods for point pattern deviations. URL http://arxiv.org/abs/2310.02292. arXiv:2310.02292 [stat].

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. In: Springer Series in Statistics, Springer New York, New York, NY, http://dx.doi.org/10.1007/978-0-387-84858-7, URL http://link.springer.com/10.1007/978-0-387-84858-7.

Hastie, T., Tibshirani, R., Wainwright, M., 2015. Statistical learning with sparsity: the lasso and generalizations. In: Monographs on statistics and applied probability, (143), CRC Press, Taylor & Francis Group, Boca Raton, http://dx.doi.org/10.1201/b18401.

Helmuth, J.A., Paul, G., Sbalzarini, I.F., 2010. Beyond co-localization: inferring spatial interactions between sub-cellular structures from microscopy images. BMC Bioinformatics 11 (1), 372. http://dx.doi.org/10.1186/1471-2105-11-372, URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-372.

Hubbell, S.P., Foster, R.B., O'Brien, S.T., Harms, K.E., Condit, R., Wechsler, B., Wright, S.J., De Lao, S.L., 1999. Light-Gap Disturbances, Recruitment Limitation, and Tree Diversity in a Neotropical Forest. Science 283 (5401), 554–557. http://dx.doi.org/10.1126/science.283.5401.554, URL https://www.science.org/doi/10.1126/science.283.5401.554. Publisher: American Association for the Advancement of Science (AAAS).

Hui, F.K.C., Warton, D.I., Foster, S.D., 2015. Tuning Parameter Selection for the Adaptive Lasso Using ERIC. J. Amer. Statist. Assoc. 110 (509), 262–269. http://dx.doi.org/10.1080/01621459.2014.951444, URL http://www.tandfonline.com/doi/full/10.1080/01621459.2014.951444.

Kuronen, M., Myllymäki, M., Loavenbruck, A., Särkkä, A., 2021. Point process models for sweat gland activation observed with noise. Stat. Med. 40 (8), 2055–2072. http://dx.doi.org/10.1002/sim.8891, URL https://onlinelibrary.wiley.com/doi/10.1002/sim.8891.

Lavancier, F., Poinas, A., Waagepetersen, R., 2021. Adaptive estimating function inference for nonstationary determinantal point processes. Scand. J. Stat. 48 (1), 87–107. http://dx.doi.org/10.1111/sjos.12440, URL https://onlinelibrary.wiley.com/doi/10.1111/sjos.12440.

Li, Y., Baccelli, F., Dhillon, H.S., Andrews, J.G., 2015. Statistical Modeling and Probabilistic Analysis of Cellular Networks With Determinantal Point Processes. IEEE Trans. Commun. 63 (9), 3405–3422. http://dx.doi.org/10.1109/TCOMM.2015.2456016, URL https://ieeexplore.ieee.org/document/7155510/.

Lund, J., Rudemo, M., 2000. Models for point processes observed with noise. Biometrika 87 (2), 235–249. http://dx.doi.org/10.1093/biomet/87.2.235, URL https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/87.2.235.

Maddu, S., Cheeseman, B.L., Sbalzarini, I.F., Müller, C.L., 2022. Stability selection enables robust learning of differential equations from limited noisy data. Proc. R. Soc. A: Math. Phys. Eng. Sci. 478 (2262), 20210916. http://dx.doi.org/10.1098/rspa.2021.0916, URL https://royalsocietypublishing.org/doi/10.1098/rspa.2021.0916.

Meinshausen, N., Bühlmann, P., 2010. Stability Selection. J. R. Stat. Soc. Ser. B Stat. Methodol. 72 (4), 417–473. http://dx.doi.org/10.1111/j.1467-9868.2010.00740.x, URL https://academic.oup.com/jrsssb/article/72/4/417/7076513.

Møller, J., Waagepetersen, R.P., 2003. Statistical Inference and Simulation for Spatial Point Processes, first ed. Monographs on Statistics and Applied Probability, vol. 100, Chapman and Hall/CRC, http://dx.doi.org/10.1201/9780203496930, URL https://www.taylorfrancis.com/books/9781135442286.

Møller, J., Waagepetersen, R.P., 2007. Modern Statistics for Spatial Point Processes. Scand. J. Stat. 34 (4), 643–684. http://dx.doi.org/10.1111/j.1467-9469.2007.00569.x, URL https://onlinelibrary.wiley.com/doi/10.1111/j.1467-9469.2007.00569.x.

Møller, J., Waagepetersen, R., 2017. Some Recent Developments in Statistics for Spatial Point Patterns. Annu. Rev. Stat. Appl. 4 (1), 317–342. http://dx.doi.org/10.1146/annurev-statistics-060116-054055, URL https://www.annualreviews.org/doi/10.1146/annurev-statistics-060116-054055.

Nogueira, S., Sechidis, K., Brown, G., 2018. On the Stability of Feature Selection Algorithms. J. Mach. Learn. Res. 18 (174), 1–54, URL http://jmlr.org/papers/v18/17-514.html.

Parikh, N., Boyd, S., 2014. Proximal Algorithms. Found. Trends® Optim. 1 (3), 127–239. http://dx.doi.org/10.1561/2400000003, URL http://www.nowpublishers.com/articles/foundations-and-trends-in-optimization/OPT-003.

Parra, E.R., 2021. Methods to Determine and Analyze the Cellular Spatial Distribution Extracted From Multiplex Immunofluorescence Data to Understand the Tumor Microenvironment. Front. Mol. Biosci. 8, 668340. http://dx.doi.org/10.3389/fmolb.2021.668340, URL https://www.frontiersin.org/articles/10.3389/fmolb.2021.668340/full.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. Pytorch: an imperative style, high-performance deep learning library. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, URL https://dl.acm.org/doi/10.5555/3454287.3455008.

Rajala, T., Murrell, D.J., Olhede, S.C., 2018. Detecting Multivariate Interactions in Spatial Point Patterns with Gibbs Models and Variable Selection. J. R. Stat. Soc. Ser. C. Appl. Stat. 67 (5), 1237–1273. http://dx.doi.org/10.1111/rssc.12281, URL https://academic.oup.com/jrsssc/article/67/5/1237/7058395.

Renner, I.W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S.J., Popovic, G., Warton, D.I., 2015. Point process models for presence-only analysis. In: O'Hara, R.B. (Ed.), Methods Ecol. Evol. 6 (4), 366–379. http://dx.doi.org/10.1111/2041-210X.12352, URL https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.12352.

Renner, I.W., Warton, D.I., 2013. Equivalence of MAXENT and Poisson Point Process Models for Species Distribution Modeling in Ecology. Biometrics 69 (1), 274–281. http://dx.doi.org/10.1111/j.1541-0420.2012.01824.x, URL https://academic.oup.com/biometrics/article/69/1/274-281/7490909.

Shah, R.D., Samworth, R.J., 2013. Variable Selection with Error Control: Another Look at Stability Selection. J. R. Stat. Soc. Ser. B Stat. Methodol. 75 (1), 55–80. http://dx.doi.org/10.1111/j.1467-9868.2011.01034.x, URL https://academic.oup.com/jrsssb/article/75/1/55/7075434.

Spychala, C., Dombry, C., Goga, C., 2024. Variable selection methods for Log-Gaussian Cox processes: A case-study on accident data. Spat. Stat. 61, 100831. http://dx.doi.org/10.1016/j.spasta.2024.100831, URL https://linkinghub.elsevier.com/retrieve/pii/S2211675324000228.

Sturm, D., Sbalzarini, I.F., 2025. Git repository for the code of the paper. URL https://git.mpi-cbg.de/mosaic/point_process_stability_selection.

Summers, H.D., Wills, J.W., Rees, P., 2022. Spatial statistics is a comprehensive tool for quantifying cell neighbor relationships and biological processes via tissue image analysis. Cell Rep. Methods 2 (11), 100348. http://dx.doi.org/10.1016/j.crmeth.2022.100348, URL https://linkinghub.elsevier.com/retrieve/pii/S2667237522002454.

Thurman, A., Fu, R., Guan, Y., Zhu, J., 2015. Regularized Estimating Equations for Model Selection of Clustered Spatial Point Processes. Statist. Sinica 25 (1), 173–188. http://dx.doi.org/10.5705/ss.2013.208w, URL http://www3.stat.sinica.edu.tw/statistica/J25N1/J25N110/J25N110.html.

Thurman, A.L., Zhu, J., 2014. Variable selection for spatial Poisson point processes via a regularization method. Stat. Methodol. 17, 113–125. http://dx.doi.org/10.1016/j.stamet.2013.08.001, URL https://linkinghub.elsevier.com/retrieve/pii/S1572312713000622.

Virtanen, P., Gommers, R., Oliphant, T.E., et al., 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods 17 (3), 261–272. http://dx.doi.org/10.1038/s41592-019-0686-2, URL https://www.nature.com/articles/s41592-019-0686-2. Publisher: Springer Science and Business Media LLC.

Waagepetersen, R.P., 2007. An Estimating Function Approach to Inference for Inhomogeneous Neyman–Scott Processes. Biometrics 63 (1), 252–258. http://dx.doi.org/10.1111/j.1541-0420.2006.00667.x, URL https://academic.oup.com/biometrics/article/63/1/252-258/7321690. Publisher: Oxford University Press (OUP).

Waagepetersen, R., Guan, Y., 2009. Two-Step Estimation for Inhomogeneous Spatial Point Processes. J. R. Stat. Soc. Ser. B Stat. Methodol. 71 (3), 685–702. http://dx.doi.org/10.1111/j.1467-9868.2008.00702.x, URL https://academic.oup.com/jrsssb/article/71/3/685/7092929.

Werner, T., 2025. Loss-guided stability selection. Adv. Data Anal. Classif. 19 (1), 5–30. http://dx.doi.org/10.1007/s11634-023-00573-3, URL https://link.springer.com/10.1007/s11634-023-00573-3.

Yang, Y., Yu, J., 2020. Fast Proximal Gradient Descent for a Class of Non-convex and Non-smooth Sparse Learning Problems. In: Adams, R.P., Gogate, V. (Eds.), Proceedings of the 35th Uncertainty in Artificial Intelligence Conference. PMLR, pp. 1253–1262, URL https://proceedings.mlr.press/v115/yang20b.html.

Yue, Y.R., Loh, J.M., 2015. Variable selection for inhomogeneous spatial point process models. Can. J. Stat. / Rev. Can. Stat. 43 (2), 288–305, URL http://www.jstor.org/stable/44708472. Publisher: [Statistical Society of Canada, Wiley].

Zimmerman, D.L., 2008. Estimating the Intensity of a Spatial Point Process from Locations Coarsened by Incomplete Geocoding. Biometrics 64 (1), 262–270. http://dx.doi.org/10.1111/j.1541-0420.2007.00870.x, URL https://academic.oup.com/biometrics/article/64/1/262-270/7331608.

Zou, H., 2006. The Adaptive Lasso and Its Oracle Properties. J. Amer. Statist. Assoc. 101 (476), 1418–1429. http://dx.doi.org/10.1198/016214506000000735, URL https://www.tandfonline.com/doi/full/10.1198/016214506000000735.