CENTRE FOR **STOCHASTIC GEOMETRY**
AND ADVANCED **BIOIMAGING**

Achmad Choiruddin, Francisco Cuevas-Pacheco, Jean-François Coeurjolly and Rasmus Waagepetersen

Regularized estimation for highly multivariate log Gaussian Cox processes

No. 03, February 2020

Regularized estimation for highly multivariate log Gaussian Cox processes

Achmad Choiruddin[1], Francisco Cuevas-Pacheco[2], Jean-François Coeurjolly[2] and Rasmus Waagepetersen[3]

[1]Department of Statistics, Institut Teknologi Sepuluh Nopember (ITS), Indonesia
[2]Department of Mathematics, Université du Québec à Montréal (UQAM), Canada
[3]Department of Mathematical Sciences, Aalborg University, Denmark

Abstract

Statistical inference for highly multivariate point pattern data is challenging due to complex models with large numbers of parameters. In this paper we develop numerically stable and efficient parameter estimation and model selection algorithms for a class of multivariate log Gaussian Cox processes. The methodology is applied to a highly multivariate point pattern data set from tropical rain forest ecology.

Keywords: cross pair correlation, elastic net, LASSO, log Gaussian Cox process, multivariate point process, proximal Newton method.

1 Introduction

Highly multivariate point pattern data are becoming increasingly common. Tropical rain forest ecologists, for example, collect data on locations of thousands of trees belonging to hundreds of species. Likewise, huge space-time data sets regarding scene, time and type of crimes are recorded and made publicly available for many major cities across the world. The basic tools for inferring dependencies from multivariate point pattern data are bivariate summary statistics like cross K and cross pair correlation functions (e.g. Møller and Waagepetersen, 2003; Lan *et al.*, 2012). However, for highly multivariate point patterns it becomes very difficult to grasp the joint information in the resulting very large number of bivariate summary statistics. To better understand the multivariate dependence structure, parametric modeling strategies are needed since such strategies enable researchers to formulate and assess qualitative assumptions regarding the dependence structure. Moreover, concise quantitative conclusions can be obtained from the estimated model parameters. See also Section 2.4 in this paper.

Research on statistical modeling of multivariate point patterns has mainly considered bivariate or trivariate point patterns. Some exceptions are Diggle *et al.* (2005) and Baddeley *et al.* (2014) who considered four- and six-variate multivariate

Poisson processes and more recently, Jalilian *et al.* (2015) who considered five-variate product shot-noise Cox processes and Waagepetersen *et al.* (2016) who considered nine-variate log Gaussian Cox processes (LGCPs). To the best of our knowledge, the only truly high-dimensional analysis was conducted by Rajala *et al.* (2018) who introduced a multivariate Gibbs point process and applied it to a point pattern data set containing locations of 83 species of rain forest trees.

Gibbs and Cox point processes are very different and complementary model classes which each have their virtues, see e.g. the review in Møller and Waagepetersen (2007). Gibbs models have clear interpretations in terms of their Papangelou conditional intensities while their intensity functions are not tractable. In contrast, shot-noise Cox processes and LGCPs have tractable intensity and second order joint intensity functions which is advantageous in terms of marginal interpretations. Similar to generalized linear mixed models, the specific type of multivariate LGCP considered by Waagepetersen *et al.* (2016) further enables decomposition of the log random intensity function into different sources of variation with natural interpretations.

A particular challenge regarding modeling of highly multivariate point patterns is that models easily become very complex with large numbers of parameters. This leads to considerable computational challenges. Waagepetersen *et al.* (2016) employed a standard quasi-Newton optimization algorithm which is not very fast nor computationally stable and this is the main reason why they did not analyse more than nine species jointly. Their discussion section mentioned the possibility of using regularization to enhance interpretability of fitted models and numerical stability of estimation. This is the approach followed by Rajala *et al.* (2018) who used the group LASSO in the context of Gibbs processes.

The objective of this paper is to expand the applicability of the multivariate LGCP models defined by Waagepetersen *et al.* (2016) by developing a numerically stable and efficient parameter estimation methodology. This significally adds to the toolbox of spatial statistics since users are then free to choose among highly multivariate Gibbs and Cox processes according to their preferences. We achieve this by introducing regularization for certain parts of the multivariate LGCP and by constructing efficient convex optimization algorithms exploiting the particular structure of the estimation objective function.

The rest of the article is organized as follows. In Section 2, we detail the multivariate LGCP model considered in this study. The new estimation methodology is developed in Section 3 and in Appendices A-B, and we test it in a simulation study in Section 4 and on tropical rain forest data in Section 5. In particular we demonstrate the potential for highly multivariate point patterns by analyzing a point pattern of locations of 86 species of rain forest trees. Section 6 contains some concluding remarks.

2 Multivariate log Gaussian Cox processes

A multivariate LGCP (see Møller *et al.*, 1998) is a multivariate point process $\mathbf{X} = (X_1, \ldots, X_p)$, $p > 1$, where each component X_i, $i = 1, \ldots, p$, is a Cox process driven by a log Gaussian random intensity function Λ_i. Conditionally on the Λ_i, the X_i

are independent Poisson point processes each with intensity function Λ_i. As in Waagepetersen *et al.* (2016), we assume that the random intensity functions are of the form $\Lambda_i(\mathbf{u}) = \exp[Z_i(\mathbf{u})]$ with

$$Z_i(\mathbf{u}) = \mu_i(\mathbf{u}) + Y_i(\mathbf{u}) + U_i(\mathbf{u}),\ \mathbf{u} \in \mathbb{R}^2. \tag{2.1}$$

The terms μ_i are deterministic and typically given in terms of regressions on observed covariates. The terms Y_i and U_i are zero-mean Gaussian fields. The Y_i can be mutually correlated while the U_i are assumed to be independent. The U_i are assumed to be stationary with variances $\sigma_i^2 > 0$ and correlation functions c_i, $i = 1, \ldots, p$. For the Y_i we assume that

$$Y_i(u) = \sum_{l=1}^{q} \alpha_{il} E_l(u)$$

where $q \geq 1$, $\boldsymbol{\alpha} = [\alpha_{ij}]_{ij}$ is a $p \times q$ real valued coefficient matrix, and the E_l, $l = 1, \ldots, q$, are independent zero-mean stationary Gaussian fields with variance one. In our applications we also consider the case $q = 0$ meaning that the Y_i are omitted in (2.1). The Y_i can be interpreted as effects of unobserved spatial covariates while the U_i represent sources of clustering which are specific to each type of points. We denote by r_l the correlation function of E_l. For the correlation functions r_l and c_i we introduce isotropic parametric models $r_l(\cdot;\phi_l) = r(\|\cdot\|/\phi_l)$ and $c_i(\cdot;\psi_i) = c(\|\cdot\|/\psi_i)$, where ϕ_l and ψ_i are correlation scale parameters. Specifically, we consider in this paper exponential correlation functions $r(t) = c(t) = \exp(-t)$, $t \geq 0$, although many other choices are available (Chilès and Delfiner, 1999).

2.1 Intensity function and pair correlation function

Let $\boldsymbol{\alpha}_{i\cdot}$ denote the ith row of $\boldsymbol{\alpha}$. Following Møller *et al.* (1998), the intensity function of X_i is $\rho_i(\mathbf{u}) = \exp\left[\mu_i(\mathbf{u}) + \boldsymbol{\alpha}_{i\cdot}\boldsymbol{\alpha}_{i\cdot}^{\mathsf{T}}/2 + \sigma_i^2/2\right]$ while the cross pair correlation function for the pair X_i and X_j is

$$g_{ij}(t) = \exp\Big[\sum_{l=1}^{q} \alpha_{il}\alpha_{jl} r_l(t;\phi_l) + 1(i=j)\sigma_i^2 c_i(t;\psi_i)\Big] \tag{2.2}$$

for $t \geq 0$. Consider two spatial locations $\mathbf{u}$ and $\mathbf{v}$. Then $\rho_j(\mathbf{v}) g_{ij}(\|\mathbf{v}-\mathbf{u}\|)$ represents the cross-Palm intensity function (Coeurjolly *et al.*, 2017) and can be interpreted as the intensity function of X_j conditional on that $\mathbf{u} \in X_i$. Hence $g_{ij}(\|\mathbf{v}-\mathbf{u}\|) > 1$ (< 1) implies that presence of a point from X_i at $\mathbf{u}$ increases (decreases) the intensity of X_j at $\mathbf{v}$. Thus $\sum_{l=1}^{q} \alpha_{il}\alpha_{jl} r_l(t) < 0$ (> 0) implies repulsion (attraction) between points of X_i and X_j at lag t. Similarly, a large value of $\sum_{l=1}^{q} \alpha_{il}^2 r_l(t) + \sigma_i^2 c_i(t)$ leads to strong attraction among points of X_i separated by a lag t.

2.2 Kernel estimation of cross pair correlation functions

Non-parametric kernel estimates of the g_{ij} are given by

$$\hat{g}_{ij}(t) = \frac{1}{2\pi t} \sum_{\substack{\mathbf{u}\in X_i\cap W,\\ \mathbf{v}\in X_j\cap W,\\ \mathbf{u}\neq\mathbf{v}}} \frac{k_b(t - \|\mathbf{u}-\mathbf{v}\|)}{\hat{\rho}_i(\mathbf{u})\hat{\rho}_j(\mathbf{v})|W \cap W_{\mathbf{u}-\mathbf{v}}|}, \qquad t > 0, \tag{2.3}$$

where W is the observation window, k_b is a kernel function depending on a bandwidth $b > 0$, $|\cdot|$ denotes area and $W_{\mathbf{h}}$ denotes the translate of W by the vector $\mathbf{h} \in \mathbb{R}^2$ (Møller and Waagepetersen, 2003). The quantities $\hat{\rho}_i$ and $\hat{\rho}_j$ are estimates of the intensity functions of X_i and X_j, typically obtained from regression models depending on observed covariates through maximizing the composite likelihood (see e.g. Waagepetersen, 2007; Møller and Waagepetersen, 2007) or its regularized versions (e.g. Thurman *et al.*, 2015; Choiruddin *et al.*, 2018).

We use the non-parametric kernel estimates as response variables in a least squares estimation object function (Section 2.3) and in this context we believe that bias is more of a concern than variance. Hence bandwidths should be chosen in order to avoid oversmoothing. For the simulation studies in Section 4 the bandwidth was chosen after visual inspection of kernel estimates confirmed that the resulting estimates were not oversmoothed. The same was done for the estimates of the g_{ij}, $i \neq j$, in the data examples in Section 5. In case of $i = j$ the data-driven method in Jalilian and Waagepetersen (2018) was used for choosing the bandwidths. Unfortunately a similar method is not yet available in case of the cross pair correlation functions with $i \neq j$.

2.3 Least squares estimation

Let $\boldsymbol{\theta}$ be the parameter vector consisting of the components of $\boldsymbol{\alpha}$, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_p^2)^\mathsf{T}$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_q)^\mathsf{T}$, and $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_p)^\mathsf{T}$. Let further

$$\begin{aligned} \boldsymbol{\beta}_{ij}(\boldsymbol{\alpha}, \boldsymbol{\sigma}^2) &= (\alpha_{i1}\alpha_{j1}, \ldots, \alpha_{iq}\alpha_{jq})^\mathsf{T}, \qquad i \neq j, \\ \boldsymbol{\beta}_{ii}(\boldsymbol{\alpha}, \boldsymbol{\sigma}^2) &= (\alpha_{i1}^2, \ldots, \alpha_{iq}^2, \sigma_i^2)^\mathsf{T}. \end{aligned} \tag{2.4}$$

The objective function used by Waagepetersen *et al.* (2016) for parameter estimation is of the form

$$Q(\boldsymbol{\theta}) = \sum_{i,j=1}^{p} \|Y_{ij} - X_{ij}(\boldsymbol{\phi}, \boldsymbol{\psi})\boldsymbol{\beta}_{ij}(\boldsymbol{\alpha}, \boldsymbol{\sigma}^2)\|^2, \tag{2.5}$$

where

$$Y_{ij} = (\sqrt{w_{ij1}} \log \hat{g}_{ij}(t_1), \ldots, \sqrt{w_{ijL}} \log \hat{g}_{ij}(t_L))^\mathsf{T},$$

$\hat{g}_{ij}(t_k)$, $k = 1, \ldots, L$, are obtained using (2.3) for lags $0 < t_1 < t_2 < \ldots < t_L$ and the $w_{ij} \geq 0$ are non-negative weights. The matrix $X_{ij}(\boldsymbol{\phi}, \boldsymbol{\psi})$ is $L \times q$ $(i \neq j)$ or $L \times (q+1)$ $(i = j)$ with rows $\sqrt{w_{ijk}}\mathbf{r}(t_k; \boldsymbol{\phi})$ $(i \neq j)$ or $\sqrt{w_{iik}}[\mathbf{r}(t_k; \boldsymbol{\phi}), c_i(t_k; \psi_i)]$ $(i = j)$, $k = 1, \ldots, L$, where

$$\mathbf{r}(t_k; \boldsymbol{\phi}) = (r_1(t_k; \phi_1), \ldots, r_q(t_k; \phi_q)).$$

Waagepetersen *et al.* (2016) minimized $Q(\boldsymbol{\theta})$ using a standard quasi-Newton method.

2.4 Inference regarding multivariate dependence structure

The model (2.1) enables us to decompose the covariances of the latent Gaussian fields Z_i into contributions from the common fields E_l and the type-specific fields

U_i. Specifically, Waagepetersen *et al.* (2016) considered for each type i and lag t the proportion of variance (PV) due to the common fields:

$$\begin{aligned}\mathrm{PV}_i(t) &= \frac{\mathrm{cov}\{Y_i(\mathbf{u}), Y_i(\mathbf{u}+\mathbf{h})\}}{\mathrm{cov}\{Z_i(\mathbf{u}), Z_i(\mathbf{u}+\mathbf{h})\}} \\ &= \frac{\sum_{l=1}^q \alpha_{il}^2 r_l(t;\phi_l)}{\sum_{l=1}^q \alpha_{il}^2 r_l(t;\phi_l) + \sigma_i^2 c_i(t;\psi_i)}, \qquad \|\mathbf{h}\| = t.\end{aligned}$$

These are useful e.g. for grouping species based on how much of the variation is due to common factors respectively type-specific factors. Furthermore, from $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}^2$ we can compute the matrix of lag zero inter-type covariances $\boldsymbol{\alpha}\boldsymbol{\alpha}^\mathsf{T}$ due to the common latent fields with ijth entry

$$\mathrm{cov}\{Y_i(\mathbf{u}), Y_j(\mathbf{u})\} = \boldsymbol{\alpha}_{i\cdot}\boldsymbol{\alpha}_{j\cdot}^\mathsf{T}$$

as well as the lag zero covariances between the fields including both common and type-specific effects,

$$\mathrm{cov}\{Z_i(\mathbf{u}), Z_j(\mathbf{u})\} = \boldsymbol{\alpha}_{i\cdot}\boldsymbol{\alpha}_{j\cdot}^\mathsf{T} + 1[i=j]\sigma_i^2. \tag{2.6}$$

A row $\boldsymbol{\alpha}_{i\cdot}$ informs on the dependence of X_i on the common latent fields. Considering the norms of differences $\|\boldsymbol{\alpha}_{i\cdot} - \boldsymbol{\alpha}_{j\cdot}\|$, we are able to group the different types of point patterns according to their dependence on the latent factors E_l.

As discussed in Waagepetersen *et al.* (2016), the distribution of our multivariate LGCP is invariant to 1) simultaneous permutation of columns in $\boldsymbol{\alpha}$ and corresponding ϕ_i's and 2) multiplication of a column in $\boldsymbol{\alpha}$ by -1. Thus we can not identify individual parameters α_{il} and ϕ_l without imposing constraints on the parameter space.

In our simulation studies in Section 4, we therefore follow Waagepetersen *et al.* (2016) by restricting attention to identifiable functions of $\boldsymbol{\alpha}$ and $\boldsymbol{\psi}$ such as the aforementioned proportions of variances and covariances and norms of differences between rows of $\boldsymbol{\alpha}$. In the application, we also consider the percentage of zero entries when $\boldsymbol{\alpha}$ is estimated using elastic net regularization with $\xi > 0$, see next section. The more zeros, the less complex is the dependence structure of the multivariate LGCP.

3 Regularized least squares estimation

The parameter vector $\boldsymbol{\theta}$ is of potentially very high dimension, especially due to the many components of the $p \times q$ parameter matrix $\boldsymbol{\alpha}$. To enhance interpretability and numerical stability of estimation we suggest to introduce regularization and thus consider the regularized least squares criterion

$$Q_\lambda(\boldsymbol{\theta}) = \; Q(\boldsymbol{\theta}) + \lambda \sum_{i=1}^{p}\sum_{l=1}^{q} p(\alpha_{il}) \tag{3.1}$$

where $Q(\boldsymbol{\theta})$ is given by (2.5), λ is a nonnegative tuning parameter and $p(\cdot)$ is a convex penalty function. We consider in the following the elastic net penalization (Zou and

Hastie, 2005) $p(\alpha_{il}) = (1-\xi)\alpha_{il}^2/2 + \xi|\alpha_{il}|$, $0 \le \xi \le 1$, which embraces LASSO (Tibshirani, 1996) and ridge regression (Hoerl and Kennard, 1988) techniques by setting $\xi = 1$ or $\xi = 0$ respectively.

Using regularization in a related factor analysis was previously suggested by Choi *et al.* (2010). Their simpler setting corresponds to directly observing vectors $(Z_i(u_k))_{i=1}^p$, $k = 1, \ldots, n$, where $Z_i(u_k)$ is modeled as in (2.1) but with zero spatial correlation. In contrast, our Z_i are unobserved with spatial correlation modeled via the correlation functions r_l and c_i. Thus the computational methodology suggested by Choi *et al.* (2010) is not applicable in our situation.

To minimize (3.1) with respect to $\boldsymbol{\theta}$, we employ a cyclical block descent algorithm where $\boldsymbol{\sigma}^2$, $\boldsymbol{\alpha}$, $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are updated in turn. The updating is iterated until relative function convergence of the criterion (3.1). The details of the block updates are given in the following two sections and Appendices A–B. Pseudo-code for the full algorithm is given in Appendix B.3.

3.1 Update for $\boldsymbol{\sigma}^2$ and $\boldsymbol{\alpha}$

Our strategy for updating $\boldsymbol{\sigma}^2$ and $\boldsymbol{\alpha}$ is to use a least squares update of σ_i^2 for $i = 1, \ldots, p$ followed by an update of $\boldsymbol{\alpha}_{i\cdot}$ using a cyclical coordinate descent algorithm. The motivation for updating rows $\boldsymbol{\alpha}_{i\cdot}$ instead of other subsets of $\boldsymbol{\alpha}$ is that the update of $\boldsymbol{\alpha}_{i\cdot}$, keeping all other parameters fixed, is quite close to a standard least squares problem, as will be evident in the following.

The relevant part of the objective function for the updates of σ_i^2 and $\boldsymbol{\alpha}_{i\cdot}$ given all other parameters is

$$Q_{\lambda,i}(\boldsymbol{\alpha}_{i\cdot}, \sigma_i^2) = 2\sum_{\substack{j=1 \\ j\neq i}}^{p} \|Y_{ij} - \tilde{X}_{ij}\boldsymbol{\alpha}_{i\cdot}\|^2 + \|Y_{ii} - X_{ii}\boldsymbol{\beta}_{ii}(\boldsymbol{\alpha}, \boldsymbol{\sigma}^2)\|^2 + \lambda\sum_{l=1}^{q} p(\alpha_{il}) \tag{3.2}$$

where the lth column of $\tilde{X}_{ij}$ is the lth column of X_{ij} multiplied by α_{jl}. In other words, for $i \neq j$, $\tilde{X}_{ij} = X_{ij}\mathrm{Diag}(\alpha_{j1}, \ldots, \alpha_{jq})$ where $\mathrm{Diag}(\alpha_{j1}, \ldots, \alpha_{jq})$ is the diagonal matrix with diagonal entries $\alpha_{j1}, \ldots, \alpha_{jq}$. For ease of notation we here omit the dependence of $\tilde{X}_{ij}$ and X_{ii} on the fixed parameters $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$. Note that (3.2) is equivalent to a standard least squares objective function for $\boldsymbol{\alpha}_{i\cdot}$ except for the middle term that depends on α_{il}^2, $l = 1, \ldots, q$, cf. (2.4).

The minimization of $Q_{\lambda,i}$ with respect to σ_i^2 only involves the middle term in (3.2). This is a standard least squares problem except that we require σ_i^2 to be non-negative. Thus,

$$\hat{\sigma}_i^2 = \max\{0, \arg\min_{\sigma_i^2} Q_{\lambda,i}(\boldsymbol{\alpha}_{i\cdot}, \sigma_i^2)\}.$$

An explicit formula for this update is given in Appendix B.1.

To update $\boldsymbol{\alpha}_{i\cdot}$ (given σ_i^2 and all other parameters), we use a so-called proximal Newton update (Lee *et al.*, 2014, and Appendix A) where the middle term in (3.2) is replaced by a quadratic approximation around the current value $\boldsymbol{\alpha}_{i\cdot}^{(k)}$. We denote by

$\hat{Q}_{\lambda,i}(\boldsymbol{\alpha}_{i\cdot}, \sigma_i^2 | \boldsymbol{\alpha}_{i\cdot}^{(k)})$ the resulting approximate objective function (to be detailed in the next paragraph). Since $\hat{Q}_{\lambda,i}(\boldsymbol{\alpha}_{i\cdot}, \sigma_i^2 | \boldsymbol{\alpha}_{i\cdot}^{(k)})$ is a regularized linear least squares objective function, minimization can be performed using a standard coordinate descent algorithm (see e.g. Hastie *et al.*, 2015).

A very simple quadratic approximation of the middle term of (3.2) is

$$\|Y_{ii} - X_{ii}\boldsymbol{\beta}_{ii}(\boldsymbol{\alpha}, \boldsymbol{\sigma}^2)\|^2 \approx \|Y_{ii} - \tilde{X}_{ii}^k[\boldsymbol{\alpha}_{i\cdot}^\mathsf{T}, \sigma_i^2]^\mathsf{T}\|^2,$$

where $\tilde{X}_{ii}^k = X_{ii}\mathrm{Diag}\{\alpha_{i1}^{(k)}, \ldots, \alpha_{iq}^{(k)}, 1\}$. Nevertheless, the curvature of this quadratic approximation does not match the curvature of the original term at $\boldsymbol{\alpha}_{i\cdot}^{(k)}$. Instead we use a second-order Taylor approximation as detailed in the Appendix A.1 which results in the explicit expression for $\hat{Q}_{\lambda,i}(\boldsymbol{\alpha}_{i\cdot}, \sigma_i^2 | \boldsymbol{\alpha}_{i\cdot}^{(k)})$ given by

$$\begin{aligned} Q_{\lambda,i}(\boldsymbol{\alpha}_{i\cdot}, \sigma_i^2) &\approx \hat{Q}_{\lambda,i}(\boldsymbol{\alpha}_{i\cdot} | \boldsymbol{\alpha}_{i\cdot}^{(k)}) \\ &= \sum_{j=1}^{p} \|Y_{ij}^* - X_{ij}^*\boldsymbol{\alpha}_{i\cdot}\|^2 + \lambda \sum_{l=1}^{q} p(\alpha_{il}), \end{aligned} \tag{3.3}$$

where

$$\begin{aligned} Y_{ij}^* &= \sqrt{2}Y_{ij}, \text{ for } i \neq j, \\ X_{ij}^* &= \sqrt{2}X_{ij}D(\alpha_{j\cdot}^{(k)}), \text{ for } i \neq j, \\ Y_{ii}^* &= Y_{ii} + X_{ii,\cdot(1:q)}\boldsymbol{\alpha}_{i\cdot}^{2,(k)} - X_{ii,\cdot(q+1)}\sigma_i^2, \\ X_{ii}^* &= 2X_{ii,\cdot(1:q)}D(\boldsymbol{\alpha}_{i\cdot}^{(k)}) \end{aligned} \tag{3.4}$$

and $X_{ii,\cdot(1:q)}$ denotes the first q columns in X_{ii}.

We obtain

$$\hat{\boldsymbol{\alpha}}_{i\cdot} = \arg\min_{\boldsymbol{\alpha}_{i\cdot}} \hat{Q}_{\lambda,i}(\boldsymbol{\alpha}_{i\cdot} | \boldsymbol{\alpha}_{i\cdot}^{(k)})$$

using coordinate descent with an explicit formula for the updates given in Appendix B.2. Further, define for some $t > 0$,

$$\boldsymbol{\alpha}_{i\cdot}^{(k+1)} = \boldsymbol{\alpha}_{i\cdot}^{(k)} + t(\hat{\boldsymbol{\alpha}}_{i\cdot} - \boldsymbol{\alpha}_{i\cdot}^{(k)}). \tag{3.5}$$

Thus, $\boldsymbol{\alpha}_{i\cdot}^{(k+1)}$ is obtained using $(\hat{\boldsymbol{\alpha}}_{i\cdot} - \boldsymbol{\alpha}_{i\cdot}^{(k)})$ as a search direction with step size controlled by t. Following Lee *et al.* (2014, Proposition 2.3), one can show (see Appendix A.2) that $Q_{i,\lambda}(\boldsymbol{\alpha}_{i\cdot}^{(k+1)}) < Q_{i,\lambda}(\boldsymbol{\alpha}_{i\cdot}^{(k)})$ if t is small enough. That is, if the minimization of $\hat{Q}_{i,\lambda}$ is combined with a line search the resulting update is guaranteed to decrease the objective function $Q_{i,\lambda}$ written in (3.2).

3.2 Update for $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$

To update $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ given all other parameters, we first reparameterize the objective function in terms of $\mathbf{f} = (\log\phi_1, \ldots, \log\phi_q)^\mathsf{T}$ and $\mathbf{s} = (\log\psi_1, \ldots, \log\psi_p)^\mathsf{T}$. We then update $\mathbf{f}$ and $\mathbf{s}$ in turn using a standard quasi-Newton update as implemented in the `optim` routine in the `R` language with method `bfgs` (Broyden-Fletcher-Goldfarb-Shanno update). Finally, we transform back using the exponential to get updates of $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$.

We also tried other options: joint update of $(\boldsymbol{\phi}, \boldsymbol{\psi})$ without log-transformation but introducing box constraints to avoid negative values and joint quasi-Newton update of the log-transformed parameters $(\mathbf{f}, \mathbf{s})$. For simulated data examples, the option with separate updates of $\mathbf{f}$ and $\mathbf{s}$ performed best.

3.3 Initialization

We initialize the components $\boldsymbol{\alpha}$ by a sample of independent random normals with mean zero and standard deviation 0.05 while we choose 1 for the initial values of the components in $\boldsymbol{\sigma}^2$. For $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ we choose initial values that depend on the scale of the observation window to avoid that the corresponding covariance functions become essentially constant equal to zero (too small initial values) or to one (too large initial values). For the unit square observation window, for example, the initial values for $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ were chosen randomly from the uniform distribution on $[0.01, 0.05]$. Regarding the choice of weights w_{ijk} introduced in Section 2.3, we follow arguments by Waagepetersen *et al.* (2016) and fix $w_{ijk} = \hat{g}_{ij}(t_k)/2$ for $i \neq j$ and $w_{iik} = \hat{g}_{ii}(t_k)$ for $i, j = 1, \ldots, p$ and $k = 1, \ldots, L$.

3.4 Strategy to determine q and regularization parameters λ and ξ

In our applications we consider just a few values $\xi = 0$ (ridge), $\xi = 0.5$ (mix of ridge and LASSO, i.e. elastic net) and $\xi = 1$ (LASSO). For each of the values of ξ we use a two-dimensional K-fold cross validation (CV) approach to select optimal values λ_{opt} and q_{opt} among prespecified values $\lambda_1, \ldots, \lambda_M$ and $q_1, \ldots, q_N$ (e.g. Hastie *et al.*, 2013, Chapter 7). The procedure is as follows.

1. We split indices ijk ($i, j = 1, \ldots, p$ and $k = 1, \ldots, L$) into K sets $S_1, \ldots, S_K$ (see details below).
2. For each $\lambda \in \{\lambda_1, \ldots, \lambda_M\}$ and $q \in \{q_1, \ldots, q_N\}$, we obtain an estimate $\hat{\boldsymbol{\theta}}_c$ by minimizing equation (3.1) with w_{ijk} replaced by 0 for $ijk \in S_c, c = 1, \ldots, K$. The CV score for λ and q is then obtained by

$$\text{CV}(\lambda, q) = \frac{1}{K} \sum_{c=1}^{K} \text{CV}_c, \tag{3.6}$$

   where $\text{CV}_c = \sum_{ijk \in S_c} (Y_{ijk} - \hat{Y}_{ijk}(\hat{\boldsymbol{\theta}}_c))^2$ and $\hat{Y}_{ij}(\hat{\boldsymbol{\theta}}_c) = X_{ij}(\hat{\boldsymbol{\phi}}_c, , \hat{\boldsymbol{\psi}}_c)\boldsymbol{\beta}_{ij}(\hat{\boldsymbol{\alpha}}_c, \hat{\boldsymbol{\sigma}}_c^2)$.

3. To obtain λ_{opt} and q_{opt}, we minimize $\text{CV}(\lambda, q)$ w.r.t λ and q, i.e.,

$$(\lambda_{\text{opt}}, q_{\text{opt}}) = \underset{\substack{m=1,\ldots,M \\ n=1,\ldots,N}}{\arg\min}\, \text{CV}(\lambda_m, q_n). \tag{3.7}$$

The sets S_c in Step 1 need to be chosen carefully. First, since $\log(\hat{g}_{ijk})$ and $\log(\hat{g}_{ijk'})$ are strongly correlated when k and k' are close, we leave out blocks of consecutive indices. Second, we do not include diagonal indices iik in the sets S_c since values

Y_{iik} include contributions from the type-specific random fields. The diagonal values thus do not provide so much information about q and omission of these values further makes the estimation procedure less stable regarding $\boldsymbol{\sigma}^2$ and $\boldsymbol{\psi}$. So, to determine each subset S_c, we arrange the ijk with $i < j$ lexicographically in a vector $(121, 122, \ldots)$ and split this vector into consecutive blocks of length b. These blocks are then assigned to the different S_c at random.

The one standard error (1-SE) rule is an alternative way to select λ and q based on the CV scores obtained from (3.6) (e.g. Hastie *et al.*, 2013). In case of q fixed, the 1-SE rule chooses the largest λ for which the CV score is less than the smallest CV score plus one standard deviation. In the case where both λ and q is to be selected, we adapt the 1-SE rule by starting with $(\lambda_{\text{opt}}, q_{\text{opt}})$ given by (3.7) and then choosing (λ, q) to be the smallest q and largest λ possible such that the following condition holds:

$$\text{CV}(\lambda, q) \leq \text{CV}(\lambda_{\text{opt}}, q_{\text{opt}}) + \text{SE}(\lambda_{\text{opt}}, q_{\text{opt}}),$$

where

$$\text{SE}(\lambda_{\text{opt}}, q_{\text{opt}}) = \sqrt{\frac{\sum_{c=1}^{K} (\text{CV}_c - \text{CV}(\lambda, q))^2}{(K-1)K}}.$$

Hence, the 1-SE rule attempts to select the most simple model whose CV score is within one standard error of the minimal CV score.

Finally, note that when $\xi = 0.5$ or $\xi = 1$ and $\lambda > 0$ is chosen, the resulting estimate of $\boldsymbol{\alpha}$ may contain columns that consist entirely of zeros. The effective number q_{eff} of columns in $\boldsymbol{\alpha}$ then becomes smaller than q_{opt}.

4 Simulation study

We conduct simulations under two settings of varying complexity with p either 5 or 10 as detailed in the following subsections. The aim is to evaluate the regularized least squares technique for parameter estimation and the CV method to select q and λ. The first setting is identical to the one used for the simulation study in Waagepetersen *et al.* (2016). Under this setting, our objective in Section 4.1 is to compare the estimates obtained using the new cyclical block descent (CBD) algorithm developed in Section 3 with the method proposed by Waagepetersen *et al.* (2016). In this regard, we consider values of $q = 1, \ldots, 5$ and fix $\lambda = 0$ since regularization was not used in Waagepetersen *et al.* (2016). Next, we consider in Section 4.2 only the new algorithm with the objective of comparing different CV options for selecting q and λ, cf. Section 3.4, and to study the effect of regularization. Section 4.3 has the same objective as Section 4.2 but now under the more complex second setting. In both simulation studies we use $K = 8$ for the CV and we only consider the LASSO option ($\xi = 1$) for regularization.

To assess the parameter estimates, we consider the root mean squared errors (RMSEs) of the estimates. For a real parameter ω and estimate $\hat{\omega}$, the RMSE is

$$\text{RMSE}(\hat{\omega}) = \sqrt{\mathbb{E}\big((\hat{\omega} - \omega)^2\big)}.$$

For each of the parameter matrices/vectors $\boldsymbol{\alpha}\boldsymbol{\alpha}^\mathsf{T}$, $\boldsymbol{\sigma}^2$, $\boldsymbol{\psi}$, or the vector of proportions of variances at lag 0 (PV), we evaluate the average of RMSEs for the components in these quantities. For example, we compute the average of RMSEs for each entry in the $p \times p$ matrix $\boldsymbol{\alpha}\boldsymbol{\alpha}^\mathsf{T}$.

4.1 Comparison of methods for least squares estimation

The first study follows the one in Waagepetersen *et al.* (2016) for which 200 point patterns in $W = [0,1]^2$ are generated from a multivariate LGCP as defined in Section 2, with $p = 5$ and $q = 2$. The true parameters are: $\boldsymbol{\sigma}^2 = (1,1,1,1,1)$, $\boldsymbol{\phi} = (0.02, 0.1)$, $\boldsymbol{\psi} = (0.01, 0.02, 0.02, 0.03, 0.04)$ and

$$\boldsymbol{\alpha}^\mathsf{T} = \begin{bmatrix} \sqrt{0.5} & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0.5 \end{bmatrix}.$$

The trend models $\mu_i(u) = m_i$ are set such that the expected number of points is 1000 for each $i = 1, \ldots, 5$. A uniform kernel with bandwidth 0.005 is used for the non-parametric estimation of the cross pair correlation function at $L = 25$ equispaced lags between 0.025 and 0.25.

For each simulation we compare two methods for minimizing (3.1) with $\lambda = 0$ and $q \in \{1, \ldots, 5\}$:

1. The standard quasi-newton (SQN) optimization algorithm considered by Waagepetersen *et al.* (2016) and implemented in the `R` package `optimx`. This algorithm updates all parameters jointly.
2. The new CBD algorithm described in Section 3.

The comparison is in terms of minimization of the objective function, computing time and RMSEs.

Table 1 reports the averages of the values of the minimized objective functions and the computational times over the 200 simulations. All timings are carried out on a Dell R740 2 x 14 cores (Intel(R) Xeon(R) Gold 6132 CPU @ 2.60GHz) 768 GB

Table 1: Averages of the minimized objective function $Q(\boldsymbol{\theta})$ given by (2.5) and the computing time (in seconds) based on 200 simulations from a multivariate LGCP ($p = 5, q = 2$), modeled with $q \in \{1,2,3,4,5\}$, for two optimization methods.

Method	q				
	1	2	3	4	5
	Minimized objective function				
SQN	6.61	4.76	5.39	6.32	4.51
CBD	3.55	1.96	1.73	1.62	1.57
	Timings (seconds)				
SQN	0.96	1.98	3.97	6.45	8.99
CBD	1.99	3.11	4.26	5.30	5.92

RAM 2x200gb SSD 960 GB NVME. CBD performs considerably better in terms of minimizing the objective function than SQN. SQN is somewhat faster than CBD for small q but slower for larger q. The computing times for SQN grow quite quickly with increasing q while the computing times seems more stable for CBD.

The RMSE results are shown in Table 2. For the calculation of the RMSEs, we exclude small percentages of very extreme parameter estimates. These percentages are reported in the last column of Table 2. CBD performs better than SQN since smaller RMSEs are obtained and there are no outlying parameter estimates. For SQN quite large percentages of extreme parameter estimates are observed.

Table 2: Average RMSEs for $\hat{\boldsymbol{\alpha}}\hat{\boldsymbol{\alpha}}^\mathsf{T}$, $\hat{\boldsymbol{\sigma}}^2$, and $\hat{\boldsymbol{\psi}}$ (see explanation in text) obtained from 200 simulations from a multivariate LGCP ($p = 5, q = 2$), modeled with $q \in \{1, 2, 3, 4, 5\}$. The estimates are obtained by minimizing (2.5) with two optimization methods. Last column shows the percentages of outlying parameter estimates removed in the RMSE calculation.

Method	q					Outliers (%)
	1	2	3	4	5	
			$\hat{\boldsymbol{\alpha}}\hat{\boldsymbol{\alpha}}^\mathsf{T}$			
SQN	0.41	0.93	1.10	1.17	1.09	10.3
CBD	0.41	0.25	0.29	0.32	0.39	0
			$\hat{\boldsymbol{\sigma}}^2$			
SQN	0.58	0.54	0.44	0.89	0.98	1.1
CBD	0.34	0.18	0.28	0.39	0.50	0
			$\hat{\boldsymbol{\psi}}$			
SQN	0.0791	0.1752	0.1337	0.4091	0.4566	11.5
CBD	0.0050	0.0091	0.0110	0.0005	0.0004	0

4.2 Assessment of cross validation and regularization with $p = 5$

In this section we continue with the simulations from the previous setting but restrict attention to CV selection of q and λ using CBD for optimization with the LASSO regularization ($\xi = 1$). We select values of q in $\boldsymbol{q} = \{1, 2, 3, 4, 5\}$ and values of λ in $\boldsymbol{\lambda} = \{0, 10^{-3}, \ldots, 5\}$ which has 20 elements and where the non-zero values of $\boldsymbol{\lambda}$ grow log-linearly from $\log 10^{-3}$ to $\log 5$. We consider three situations: (1) we select q from $\boldsymbol{q}$ with $\lambda = 0$ fixed, thus least squares estimation (LSE) is performed; (2) we search for the jointly optimal (q, λ); (3) we fix $q = 5$ and select λ from $\boldsymbol{\lambda}$. Recall that the selection of a relatively big λ may lead to zero columns in the $\boldsymbol{\alpha}$ estimate. We therefore consider the effective q_{eff} as defined in Section 3.4. Thereby we can also evaluate the selection of q in situation (3). In case of (2) we both consider the minimum CV (Min) and the 1-standard error (1-SE) rules to select q and λ.

Table 3 shows the distribution of absolute distance between q_{eff} and the true $q = 2$. For LSE, using the Min rule, q_{eff} coincides with the true q for 47% of the

Table 3: Distribution of $|q_{\text{eff}} - 2|$ (in %) over 200 simulations from a multivariate LGCP ($p = 5, q = 2$) using CBD for minimization.

	LSE $q \in \boldsymbol{q},\ \lambda = 0$				LASSO $q \in \boldsymbol{q},\ \lambda \in \boldsymbol{\lambda}$				LASSO $q = 5\,; \lambda \in \boldsymbol{\lambda}$			
$\lvert q_{\text{eff}} - 2\rvert$	0	1	2	3	0	1	2	3	0	1	2	3
Min	47	28	13	12	42	32	21	5	16	37	30	17
1-SE	46	32	22	0	15	20	65	0	10	22	65	3

Table 4: Average RMSEs obtained from 200 simulations from a multivariate LGCP ($p = 5$, $q = 2$) for different methods of selecting q and λ.

	$q = 2$		LSE		LASSO		LASSO	
	$\lambda = 0$	$\lambda \in \boldsymbol{\lambda}$	$q \in \boldsymbol{q}, \lambda = 0$		$q \in \boldsymbol{q}, \lambda \in \boldsymbol{\lambda}$		$q = 5, \lambda \in \boldsymbol{\lambda}$	
	Min	Min	Min	1-SE	Min	1-SE	Min	1-SE
$\hat{\boldsymbol{\alpha}}\hat{\boldsymbol{\alpha}}^{\mathsf{T}}$	0.26	0.33	0.33	0.40	0.36	0.54	0.40	0.54
$\hat{\boldsymbol{\sigma}}^2$	0.42	0.54	0.54	0.58	0.56	0.75	0.63	0.76
$\hat{\boldsymbol{\psi}}$	0.04	0.05	0.05	0.02	0.03	0.01	0.04	0.01
$\hat{\text{PV}}$	0.28	0.31	0.32	0.35	0.33	0.41	0.37	0.42

simulations and differs at most by 1 from the true q in 75% of the simulations. The results with the 1-SE rule are similar with percentages 46 and 78. LASSO with Min rule for joint selection of (q, λ) performs similarly to LSE with the corresponding percentages 42 and 74 %. With fixed $q = 5$ the percentages are reduced to 16% and 53 %. Using 1-SE rule, the LASSO forces many columns to be zero leading to quite small percentages where $|q_{\text{eff}} - 2| \leq 1$.

RMSEs are reported in Table 4 for all three situations. In addition, in the first columns, we consider the case fixed $q = 2$ assuming the true q is known. We first note that LASSO gives worse results than LSE when $q = 2$ is fixed. In general, for unknown q, LSE and LASSO perform quite similarly when the Min rule is used. The results are worse when 1-SE is used and in particular for LASSO. When q is fixed to 5 and only λ is selected the results are worse than for LASSO with q selected by the Min rule while the results with $q = 5$ are similar to LASSO with q selected by the 1-SE rule.

The overall impression is that LSE performs slightly better than LASSO, especially in estimating $\boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathsf{T}}$. This may indicate that when p is relatively small, selection of q with $\lambda = 0$ (LSE) already gives sparse results. Another reason that LASSO does not improve RMSE may be that the true $\boldsymbol{\alpha}$ is not that sparse having only 40% zero components. Thus the bias introduced by regularization is not counterbalanced by a reduction in variance. On the other hand, the performance of LSE and LASSO are quite similar showing that the CV does a good job in selecting a small λ for the LASSO. Following a comment by a referee, we also tried out a more sparse scenario with

$$\boldsymbol{\alpha}^{\mathsf{T}} = \begin{bmatrix} 0.1 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & -0.1 \end{bmatrix}.$$

In this case, as expected, LASSO outperforms LSE both in terms of selecting q and in terms of RMSE for $\boldsymbol{\alpha}\boldsymbol{\alpha}^\mathsf{T}$ (see supplementary material, page 26) when the Min rule is used. In the next section we further explore LASSO and LSE in a more complex setting with $p = 10$.

4.3 Assessment of cross validation and regularization with $p = 10$

In this experiment, we study a more complex situation with a higher p and more variation in the parameters. We simulate 200 point patterns from a multivariate LGCP with $p = 10$, $q = 4$, $W = [0, 1]^2$, and parameters

$$\boldsymbol{\alpha} = \begin{bmatrix} \sqrt{0.5} & 0.10 & -1 & 0 \\ 0 & 0 & -0.70 & 1 \\ 0 & -0.15 & \sqrt{0.5} & 0.10 \\ -1 & 0 & 0 & 0 \\ -0.70 & 1 & 0 & -0.15 \\ \sqrt{0.5} & 0.10 & -1 & 0 \\ 0 & 0 & -0.70 & 1 \\ 0 & -0.15 & \sqrt{0.5} & 0.10 \\ -1 & 0 & 0 & 0 \\ -0.70 & 1 & 0 & -0.15 \end{bmatrix},$$

$$\boldsymbol{\phi} = (0.02, 0.03, 0.03, 0.05),$$

$$\boldsymbol{\sigma}^2 = (1, 1, 1.5, 1, 0.2, 0.2, 1, 1.5, 1.5, 1.5)^\mathsf{T},$$

and

$$\boldsymbol{\psi} = (0.01, 0.02, 0.02, 0.03, 0.04, 0.04, 0.05, 0.06, 0.06, 0.07)^\mathsf{T}.$$

The settings for the trend models, the kernel estimation and the CV are as in the previous simulation study except that $\boldsymbol{q} = \{0, \ldots, 8\}$. In $\boldsymbol{\alpha}$, 40% of the components are zeros and 20% are of absolute value less than 0.15. The remaining components have absolute value greater than 0.7.

Table 5 shows the distribution of the absolute distance $|q_{\text{eff}} - 4|$ between q_{eff} and the true $q = 4$. Considering first the Min rule, with LSE, q_{eff} concurs with the true q in 19% of the simulations and differs at most by 2 from the true q in 58% of the simulations. The corresponding percentages are 14% and 65 % for LASSO, and 6% and 41 % for LASSO with $q = 8$ fixed. In this situation, the 1-SE rule seems advantageous for selecting q. For example, the percentage of q_{eff}'s which differ from the true q by at most 2 improves from 58% to 83 % for LSE, from 65% to 80 % for LASSO, and from 41% to 68 % for LASSO with fixed $q = 8$.

Table 6 details the RMSE results. The superiority of the 1-SE rule when selecting q does not translate into better results in terms of RMSE except for LASSO with fixed $q = 8$ where better results are obtained with 1-SE than with Min. The best results are obtained with LASSO using the Min rule for selecting q and λ. This indicates that regularization is indeed helpful in complex settings with relatively large p.

Table 5: Distribution of $|q_{\text{eff}} - 4|$ from 200 simulations of a multivariate LGCP ($p = 10$ and $q = 4$).

	LSE $q \in \boldsymbol{q},\ \lambda = 0$					LASSO $q \in \boldsymbol{q},\ \lambda \in \boldsymbol{\lambda}$					LASSO $q = 8\,; \lambda = \boldsymbol{\lambda}$				
$\lvert q_{\text{eff}} - 4\rvert$	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
Min	19	21	18	19	23	14	31	20	19	16	6	15	20	21	38
1-SE	27	36	20	12	5	22	37	21	8	12	21	22	25	11	21

Table 6: Average of RMSEs obtained from 200 simulations from a multivariate LGCP ($p = 10, q = 4$) for different methods of selecting q and λ.

	LSE		LASSO		$q = 8$ (LASSO)	
	Min	1-SE	Min	1-SE	Min	1-SE
$\hat{\boldsymbol{\alpha}}\hat{\boldsymbol{\alpha}}^{\mathsf{T}}$	0.50	0.67	0.44	0.48	0.78	0.51
$\hat{\boldsymbol{\sigma}}^2$	0.58	0.89	0.54	0.70	0.88	0.76
$\hat{\boldsymbol{\psi}}$	0.02	0.02	0.01	0.02	0.02	0.02
$\hat{\text{PV}}$	0.35	0.35	0.34	0.39	0.35	0.40

Based on the simulation studies, for analyzing highly multivariate point pattern data, we recommend to use regularization with the Min rule for selecting q and λ.

5 Application

In a 50-hectare 1000 m × 500 m region of the tropical moist forest of Barro Colorado Island (BCI) in central Panama, censuses have been carried out where all free-standing woody stems with at least 10 mm diameter at breast height were identified, tagged, and mapped, resulting in maps of over 350 000 individual trees with around 300 species (see e.g. Hubbell and Foster, 1983; Condit *et al.*, 1996; Condit, 1998). In addition, 13 spatial covariates are also available containing topological attributes and soil nutrients (see supplementary material, page 26). Our main objective is to study the impact of regularization and the computational feasibility of our method. We first consider 9 tree species, *Psychotria, Protium t., Capparis, Protium p., Swartzia, Hirtella, Tetragastris, Garcinia, Mourmiri*, with intermediate abundances ranging from 2500 to 7500 and previously analyzed by Waagepetersen *et al.* (2016). The plots of locations of each species are shown in Figure 1 in the supplementary material. The main aim of this analysis is to compare the results with our new algorithm to those obtained by Waagepetersen *et al.* (2016). Secondly, to test our algorithm in a more challenging situation, we analyze a highly multivariate point pattern involving species of trees with at least 400 individuals, resulting in 86 species.

For each species, we use maximum composite likelihood to fit log-linear regression models involving the spatial covariates for the μ_i-terms in (2.1). We then estimate the cross pair correlation function using (2.3). Therefore, the variation due to observed covariates are filtered out and the non-parametric estimates of cross

pair correlation function hence capture the residual correlation due to unobserved covariates, species-specific factors, and any other sources. The bandwidths for the g_{ii} were chosen using the data-driven method in Jalilian and Waagepetersen (2018). The bandwidth 2 was chosen for the g_{ij}, $i \neq j$, after visual inspection of kernel estimates to avoid oversmoothing of the resulting estimates (see discussion in Section 2.2).

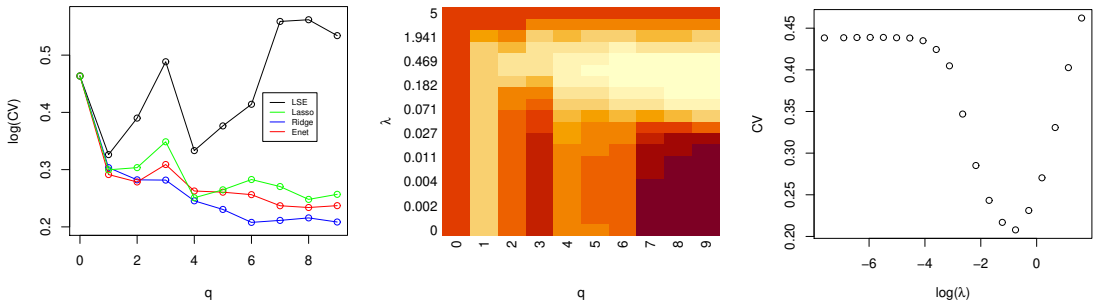

Figure 1: CV scores for 9-species data analysis. Left: $\min_{\lambda \in \boldsymbol{\lambda}} \log \mathrm{CV}(q, \lambda)$ against q for ridge, elastic net and LASSO and $\log \mathrm{CV}(q, 0)$ against q for LSE. Middle: image plot of $\mathrm{CV}(q, \lambda)$ in case of ridge (lighter color corresponds to smaller CV score). Right: $\mathrm{CV}(6, \lambda)$ plotted against $\log \lambda$ in case of ridge.

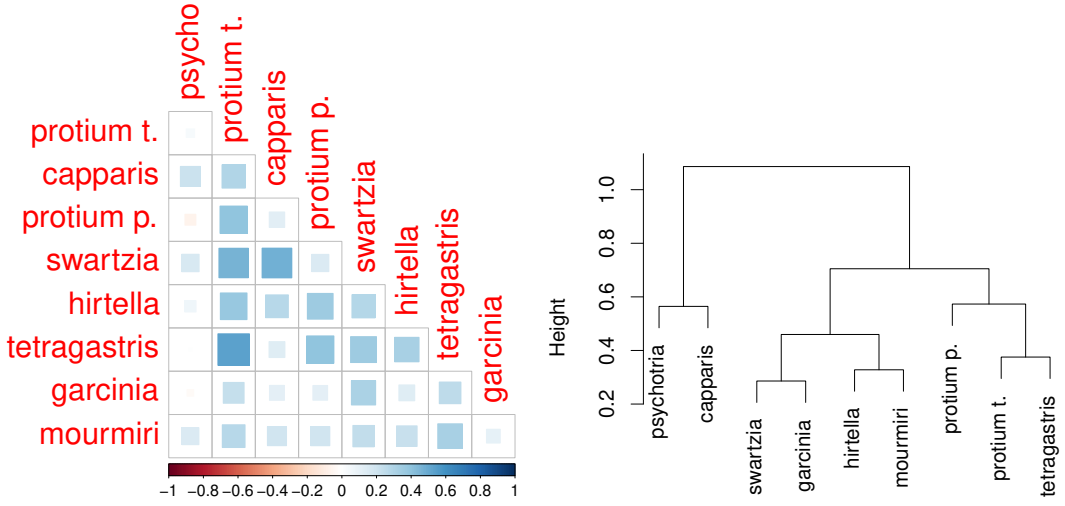

Figure 2: Left: Estimated inter-species correlations $\mathrm{corr}\{Z_i(\mathbf{u}), Z_j(\mathbf{u})\}$ at lag zero. Right: 9-species clustering based on $\|\hat{\boldsymbol{\alpha}}_{i\cdot} - \hat{\boldsymbol{\alpha}}_{j\cdot}\|$.

5.1 Application with 9 species

For each value of $\xi = 0, 0.5, 1$ we apply 8-fold CV to select q and λ where $\lambda \in \boldsymbol{\lambda} = \{0, 10^{-3}, \ldots, 5\}$ as in the simulation studies and $q \in \boldsymbol{q} = \{0, \ldots, 9\}$. The upper left plot in Figure 1 shows for each ξ, $\min_{\lambda \in \boldsymbol{\lambda}} \log \mathrm{CV}(q, \lambda)$ as a function of q. For comparison with Waagepetersen *et al.* (2016) we also show in this plot $\log \mathrm{CV}(q, 0)$ against q (LSE). A general pattern for ridge, elastic net and LASSO is that the CV scores decrease quite quickly as a function of q until around $q = 4$ and after

that the CV scores stabilize or decrease slowly. The CV scores for ridge ($\xi = 0$) are consistently smaller than those for elastic net ($\xi = 0.5$) and LASSO ($\xi = 1$). Hence we select $\xi = 0$. The minimal CV score for $\xi = 0$ is obtained with $q = 6$ and $\lambda = 0.4693$.

For comparison, the minimal CV score with LASSO is obtained with $q = 8$ and $\lambda = 0.11$. However, in this case, the resulting effectively selected q_{eff} is three since the resulting estimate of $\boldsymbol{\alpha}$ has 5 zero columns. In case of LSE ($\lambda = 0$), the CV procedure chooses $q = 1$. The second-smallest CV score with LSE is obtained with $q = 4$ which was the value chosen in Waagepetersen *et al.* (2016). The difference in CV results for LSE compared with Waagepetersen *et al.* (2016) may be due to our new more efficient optimization algorithm, cf. the comparison in Section 4.1.

The middle plot in Figure 1 is an image plot of the CV scores for ridge ($\xi = 0$) where darker color corresponds to smaller CV score. The development of the CV scores across values of q for fixed λ appears quite erratic with several local minima. In contrast, for each q there appears to be a well-defined minimum for λ. As an example, the right plot in Figure 1 shows $\text{CV}(6, \lambda)$ plotted against $\log \lambda$ (where we replace the undefined $\log 0$ by $\log 5 \times 10^{-4}$). The computing time required to run the CV method with $\xi = 0$ is 2.4 hours with the same processor as used in the simulation study. Approximately 16 seconds is required to estimate the parameters for the 9-species application using ridge with $q = 6$ and $\lambda = 0.4693$.

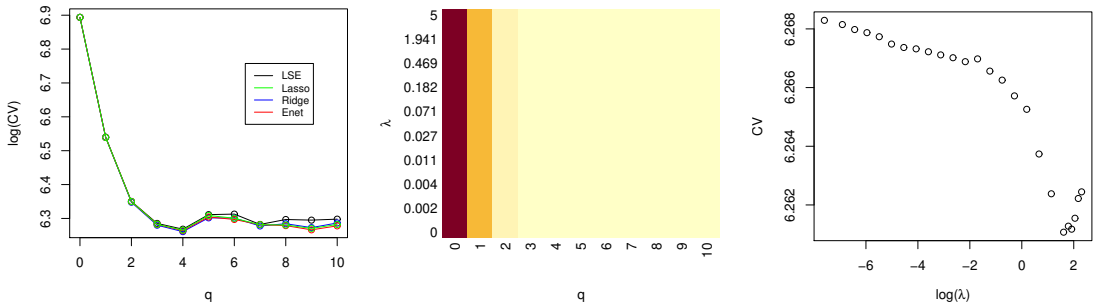

Figure 3: CV scores for 86-species data analysis. Left: $\min_{\lambda \in \boldsymbol{\lambda}} \text{CV}(q, \lambda)$ against q for ridge, elastic net and LASSO and $\text{CV}(q, 0)$ against q for LSE. Middle: image plot of $\text{CV}(q, \lambda)$ in case of ridge (lighter color corresponds to smaller CV score). Right: $\text{CV}(4, \lambda)$ plotted against $\log \lambda$ in case of ridge.

The results regarding the multivariate dependence structure of the 9 species are qualitatively rather similar to those obtained by Waagepetersen *et al.* (2016). The estimated inter-species correlations $\text{corr}\{Z_i(u), Z_j(u)\}$, cf. (2.6), are shown in the left plot of Figure 2. Most of the pairs of species have a positive correlation. However, the correlations between *Psychotria* and the other species are mainly close to zero. The right plot in Figure 2 shows a hierarchical clustering of the species based on the estimated coefficient rows $\boldsymbol{\alpha}_{i\cdot}$. Compared with Waagepetersen *et al.* (2016) where *Psychotria* was isolated, it forms a cluster with *Capparis* in the current analysis. The clustering pattern may have some relation to the families of species as shown by the cluster of *Protium p.*, *Protium t.* and *Tetragastris* which come from the same family (see Table 9 in the supplementary material).

5.2 Application with 86 tree species

For the 86-species application, we apply the 8-fold CV procedure with $\xi = 0, 0.5, 1$ and $\lambda \in \{0, 10^{-3}, \ldots, 5\}$ as in the previous section and $q \in \{0, \ldots, 10\}$. Figure 3 is similar to Figure 1. The left plot shows that the CV scores for LASSO, ridge and elastic net are very similar for all q while LSE tends to have worse CV scores for large q. For all types of regularization, the smallest CV score is obtained for $q = 4$ with slightly smallest CV score for ridge. The remaining plots are obtained with $\xi = 0$. The image plot of CV scores in the middle plot looks much smoother than in the 9 species case. The right plot shows a minimum for $\lambda = 5$ given $q = 4$. Note that to check that $\lambda = 5$ is indeed a (local) minimum we extended the CV for ridge and $q = 4$ to additional values of λ ranging from 6 to 10. For LASSO and elastic net the optimal λ values are 1.21 and 1.94, respectively, given $q = 4$.

The computing time for the CV is 6.2 hours for $\xi = 0$ and the computing time to estimate the parameters for the chosen $q = 4$ and $\lambda = 5$ is 3.3 minutes. Considering the ridge results, we model $86 \times 87/2 = 3741$ distinct pair and cross pair correlation functions using only $6 \times 86 + 4 = 520$ parameters. For LASSO or elastic net respectively 507 or 510 parameters are used since 13 or 10 parameters were set to zero in the estimated $\boldsymbol{\alpha}$ with LASSO or elastic net. Thus we can indeed obtain a sparse model for the given data.

Table 7 shows the distribution of estimated inter-species correlations due to common latent fields and the combination of common and species-specific fields (see Section 2.4) across 6 intervals. Most estimated correlations are positive. However, the correlations decrease a lot in absolute value when the species-specific fields are included (last row of Table 7).

The distribution of estimated PVs is shown in Table 8. Most species (48%) have estimated proportions of variances due to common factors less than 0.25.

Table 7: Distribution (in %) of estimated inter-species correlations $\text{corr}[Y_i(u), Y_j(u)]$ and $\text{corr}[Z_i(u), Z_j(u)]$, $i \neq j$, over different intervals [Lower, Upper] for the 86 species application using ridge ($\xi = 0$) with $q = 4$ and $\lambda = 5$.

Lower	-1	-0.5	-0.2	0	0.2	0.5
Upper	-0.5	-0.2	0	0.2	0.5	1
$\text{corr}[Y_i(u), Y_j(u)]$	2	8	10	13	22	45
$\text{corr}[Z_i(u), Z_j(u)]$	0	2	19	58	19	3

Table 8: Distribution of estimated $\text{PV}_i(0)$ for 86 species application using ridge ($\xi = 0$) with $q = 4$ and $\lambda = 5$.

Interval	0–0.25	0.25–0.5	0.5–0.75	0.75–1
Species (%)	48	30	9	13

6 Conclusion

We developed in this study a regularized estimation method for highly multivariate point patterns modeled by multivariate LGCPs. The procedure is numerically stable and performs well both in the considered simulations and applications. In our truly highly multivariate second application, we were able to fit a sparse model for a multivariate point pattern with 86 types of points.

Our method requires selection of tuning parameters q, λ and ξ as well as the bandwidths for the kernel estimates of cross pair correlation functions. Our CV procedure provides a useful solution for the choice of tuning parameters. Regarding the choice of bandwidths, it would be desirable and should be quite feasible to generalize the methods in Guan (2007) and Jalilian and Waagepetersen (2018) to cover also data-driven choice of bandwidth in case of cross pair correlation functions. One limitation of our method is that we choose a fixed class of correlation models for the latent Gaussian fields. We believe that the exact shape of the correlation models is not very crucial but choosing among different correlation models could be a topic for further research.

The results of the CV are somewhat sensitive to Monte Carlo error due to the random allocation of observations into K folds. This is especially the case where the CV score curve is quite flat as for large values of q in the left plots in Figures 1-3. At the expense of a higher computational load, the sensitivity can be reduced by averaging CV scores over several replications of K-fold CV.

An interesting application of obtained estimates is to group types of points according to their estimated dependence on common latent fields as expressed by the rows $\boldsymbol{\alpha}_{i\cdot}$. Hence a further development could be to consider an extension of the so-called fused LASSO (Tibshirani *et al.*, 2005) by introducing regularization for differences $\boldsymbol{\alpha}_{i\cdot} - \boldsymbol{\alpha}_{j\cdot}$. A further possibility would be to consider a sparse group LASSO (Simon *et al.*, 2013) to obtain estimates of $\boldsymbol{\alpha}$ with some zeros of α_{il} as developed in this paper and, in addition, with entire rows of zeros implying independence of corresponding types of points and all other types of points.

Acknowledgements

We thank the editor and the two reviewers for their constructive and helpful comments. The research by A. Choiruddin, F. Cuevas-Pacheco, and R. Waagepetersen is supported by The Danish Council for Independent Research | Natural Sciences, grant DFF – 7014-00074 "Statistics for point processes in space and beyond", and by the "Centre for Stochastic Geometry and Advanced Bioimaging", funded by grant 8721 from the Villum Foundation.

The BCI forest dynamics research project was made possible by National Science Foundation grants to Stephen P. Hubbell:

DEB-0640386, DEB-0425651, DEB-0346488, DEB-0129874, DEB-00753102, DEB-9909347, DEB-9615226, DEB-9615226, DEB-9405933, DEB-9221033, DEB-9100058, DEB-8906869, DEB-8605042, DEB-8206992, DEB-7922197,

support from the Center for Tropical Forest Science, the Smithsonian Tropical Research k+1 Institute, the John D. and Catherine T. MacArthur Foundation, the

Mellon Foundation, the Celera Foundation, and numerous private individuals, and through the hard work of over 100 people from 10 countries over the past two decades. The plot project is part of the Center for Tropical Forest Science, a global network of large-scale demographic tree plots.

The BCI soils data set were collected and analyzed by J. Dalling, R. John, K. Harms, R. Stallard and J. Yavitt with support from NSF DEB021104, 021115, 0212284, 0212818 and OISE 0314581, STRI and CTFS. Paolo Segre and Juan Di Trani provided assistance in the field. The covariates `dem`, `grad`, `mrvbf`, `solar` and `twi` were computed in SAGA GIS by Tomislav Hengl (`spatial-analyst.net`). We thank Dr. Joseph Wright for sharing data on dispersal modes and life forms for the BCI tree species.

References

Baddeley, A., Jammalamadaka, A. & Nair, G. (2014). Multitype point process analysis of spines on the dendrite network of a neuron. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **63**(5), 673–694.

Chilès, J.-P. & Delfiner, P. (1999). *Geostatistics: modeling spatial uncertainty*. Probability and Statistics, Wiley, New York.

Choi, J., Oehlert, G. & Zou, H. (2010). A penalized maximum likelihood approach to sparse factor analysis. *Statistics and its Interface* **3**(4), 429–436.

Choiruddin, A., Coeurjolly, J.-F. & Letué, F. (2018). Convex and non-convex regularization methods for spatial point processes intensity estimation. *Electronic Journal of Statistics* **12**(1), 1210–1255.

Coeurjolly, J.-F., Møller, J. & Waagepetersen, R. (2017). A tutorial on Palm distributions for spatial point processes. *International Statistical Review* **85**(3), 404–420.

Condit, R. (1998). *Tropical Forest Census Plots*. Springer-Verlag and R. G. Landes Company, Berlin, Germany and Georgetown, Texas.

Condit, R., Hubbell, S. P. & Foster, R. B. (1996). Changes in tree species abundance in a neotropical forest: impact of climate change. *Journal of tropical ecology* **12**(2), 231–256.

Diggle, P., Zheng, P. & Durr, P. (2005). Nonparametric estimation of spatial segregation in a multivariate point process: bovine tuberculosis in Cornwall, UK. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54**(3), 645–658.

Friedman, J., Hastie, T. & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**(1), 1–22.

Guan, Y. (2007). A least-squares cross-validation bandwidth selection approach in pair correlation function estimations. *Statistics and Probability Letters* **77**(18), 1722–1729.

Hastie, T., Tibshirani, R. & Friedman, J. (2013). *The elements of statistical learning: data mining, inference, and prediction*. Springer Series in Statistics, Springer New York Inc., New York, 2nd edition.

Hastie, T., Tibshirani, R. & Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. Monographs on Statistics and Applied Probability, Chapman & Hall/CRC Press, Boca Raton.

Hoerl, A. E. & Kennard, R. W. (1988). Ridge regression. *Encyclopedia of Statistical Sciences* **8**.

Hubbell, S. P. & Foster, R. B. (1983). Diversity of canopy trees in a neotropical forest and implications for conservation. In: *Tropical Rain Forest: Ecology and Management* (eds. S. L. Sutton, T. C. Whitmore and A. C. Chadwick), Blackwell Scientific Publications, Oxford, 25–41.

Jalilian, A. & Waagepetersen, R. (2018). Fast bandwidth selection for estimation of the pair correlation function. *Journal of Statistical Computation and Simulation* **88**(10), 2001–2011.

Jalilian, A., Guan, Y., Mateu, J. & Waagepetersen, R. (2015). Multivariate product-shot-noise Cox models. *Biometrics* **71**(4), 1022–1033.

Lan, G., Getzin, S., Wiegand, T., Hu, Y., Xie, G., Zhu, H. & Cao, M. (2012). Spatial distribution and interspecific associations of tree species in a tropical seasonal rain forest of China. *PloS one* **7**(9), e46074.

Lee, J. D., Sun, Y. & Saunders, M. A. (2014). Proximal Newton-type methods for minimizing composite functions. *SIAM Journal on Optimization* **24**(3), 1420–1443.

Møller, J. & Waagepetersen, R. (2003). *Statistical inference and simulation for spatial point processes*. Chapman and Hall/CRC, Boca Raton.

Møller, J. & Waagepetersen, R. (2007). Modern statistics for spatial point processes. *Scandinavian Journal of Statistics* **34**(4), 643–684.

Møller, J., Syversveen, A. R. & Waagepetersen, R. (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics* **25**(3), 451–482.

Rajala, T., Murrell, D. J. & Olhede, S. C. (2018). Detecting multivariate interactions in spatial point patterns with Gibbs models and variable selection. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **67**(5), 1237–1273.

Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics* **22**(2), 231–245.

Thurman, A. L., Fu, R., Guan, Y. & Zhu, J. (2015). Regularized estimating equations for model selection of clustered spatial point processes. *Statistica Sinica* **25**(1), 173–188.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **58**(1), 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(1), 91–108.

Waagepetersen, R. (2007). An estimating function approach to inference for inhomogeneous Neyman-Scott processes. *Biometrics* **63**(1), 252–258.

Waagepetersen, R., Guan, Y., Jalilian, A. & Mateu, J. (2016). Analysis of multi-species point patterns using multivariate log Gaussian Cox processes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **65**(1), 77–96.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2), 301–320.

A Proximal Newton Method

Suppose we want to find the solution of

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^n} f(\boldsymbol{\theta}) := a(\boldsymbol{\theta}) + c(\boldsymbol{\theta}), \tag{A.1}$$

where the function $f(\cdot)$ can be separated into two parts: the function $a(\cdot)$ which is a convex and twice continuously differentiable loss function and the function $c(\cdot)$ which is a convex but not necessarily differentiable penalty function. The proximal-Newton method is an iterative optimization algorithm that uses a quadratic approximation of the differentiable part $a(\cdot)$:

$$\begin{aligned} f(\boldsymbol{\theta}) &\approx \hat{f}(\boldsymbol{\theta}) \\ &= \hat{a}(\boldsymbol{\theta}) + c(\boldsymbol{\theta}) \\ &= a(\boldsymbol{\theta^{(k)}}) + \nabla a(\boldsymbol{\theta^{(k)}})^\mathsf{T}(\boldsymbol{\theta} - \boldsymbol{\theta^{(k)}}) \\ &\quad + (\boldsymbol{\theta} - \boldsymbol{\theta^{(k)}})^\mathsf{T} H(\boldsymbol{\theta^{(k)}})(\boldsymbol{\theta} - \boldsymbol{\theta^{(k)}}) + c(\boldsymbol{\theta}), \end{aligned} \tag{A.2}$$

where $\boldsymbol{\theta}^{(k)}$ is the current value of $\boldsymbol{\theta}$, $\nabla a(\cdot)$ is the first derivative of $a(\cdot)$ and $H(\cdot)$ is an approximation to the Hessian matrix $\nabla^2 a(\cdot)$. Letting $\tilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \hat{f}(\boldsymbol{\theta})$, the next value of $\boldsymbol{\theta}$ is obtained as

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + t(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^{(k)})$$

for some $t > 0$. That is, $\tilde{\boldsymbol{\theta}}$ is used to construct a search direction for the $k+1$th value of $\boldsymbol{\theta}$. Theoretical results in Lee *et al.* (2014) show that t can be chosen so that $f(\boldsymbol{\theta}^{(k+1)}) < f(\boldsymbol{\theta}^{(k)})$. The matrix $H(\cdot)$ can be chosen in various ways, see Lee *et al.* (2014) and Hastie *et al.* (2015) for more details.

In the following sections, we adapt the proximal Newton method to minimization of our objective function.

A.1 Quadratic approximation for updating $\boldsymbol{\alpha}_{i\cdot}$

Let us first regard (3.2) as a function of $\boldsymbol{\alpha}_{i\cdot}$,

$$\begin{aligned} Q_{\lambda,i}(\boldsymbol{\alpha}_{i\cdot}, \sigma_i^2) &= 2\sum_{\substack{j=1 \\ j\neq i}}^{p} \|Y_{ij} - \tilde{X}_{ij}\boldsymbol{\alpha}_{i\cdot}\|^2 \\ &\quad + \|Y_{ii} - X_{ii}\boldsymbol{\beta}_{ii}(\boldsymbol{\alpha}, \boldsymbol{\sigma}^2)\|^2 + \lambda\sum_{l=1}^{q} p(\alpha_{il}) \\ &= a(\boldsymbol{\alpha}_{i\cdot}) + b(\boldsymbol{\alpha}_{i\cdot}) + c(\boldsymbol{\alpha}_{i\cdot}). \end{aligned} \tag{A.3}$$

To minimize (3.2), we consider the proximal Newton method stated in (A.2). In particular, we approximate $b(\boldsymbol{\alpha}_{i\cdot})$ by a quadratic approximation around the current value $\boldsymbol{\alpha}_{i\cdot}^{(k)}$:

$$\begin{aligned} b(\boldsymbol{\alpha}_{i\cdot}) &\approx \hat{b}(\boldsymbol{\alpha}_{i\cdot}) \\ &= b(\boldsymbol{\alpha}_{i\cdot}^{(k)}) + \nabla b(\boldsymbol{\alpha}_{i\cdot}^{(k)})^{\mathsf{T}}(\boldsymbol{\alpha}_{i\cdot} - \boldsymbol{\alpha}_{i\cdot}^{(k)}) \\ &\quad + \tfrac{1}{2}(\boldsymbol{\alpha}_{i\cdot} - \boldsymbol{\alpha}_{i\cdot}^{(k)})^{\mathsf{T}} H(\boldsymbol{\alpha}_{i\cdot}^{(k)})(\boldsymbol{\alpha}_{i\cdot} - \boldsymbol{\alpha}_{i\cdot}^{(k)}). \end{aligned} \tag{A.4}$$

Here, the first derivative is

$$\nabla b(\boldsymbol{\alpha}_{i\cdot}^{(k)}) = -4D(\boldsymbol{\alpha}_{i\cdot}^{(k)})X_{ii,\cdot(1:q)}^{\mathsf{T}}\big(Y_{ii} - X_{ii}\boldsymbol{\beta}_{ii}(\boldsymbol{\alpha}^{(k)}, \boldsymbol{\sigma}^2)\big)$$

while $H(\boldsymbol{\alpha}_{i\cdot}^{(k)})$ is an approximation of the second derivative,

$$\nabla^2 b(\boldsymbol{\alpha}_{i\cdot}^{(k)}) = 8D(\boldsymbol{\alpha}_{i\cdot}^{(k)})X_{ii,\cdot(1:q)}^{\mathsf{T}}X_{ii,\cdot(1:q)}D(\boldsymbol{\alpha}_{i\cdot}^{(k)}) - C(\boldsymbol{\alpha}_{i\cdot}^{(k)}),$$

where $D(\boldsymbol{\alpha}_{i\cdot}^{(k)}) = \mathrm{Diag}(\alpha_{i1}^{(k)}, \ldots, \alpha_{iq}^{(k)})$, $X_{ii,\cdot(1:q)}$ denotes the first q columns in X_{ii}, and $C(\boldsymbol{\alpha}_{i\cdot}^{(k)}) = 4\mathrm{Diag}\big(X_{ii,\cdot(1:q)}^{\mathsf{T}}\big(Y_{ii} - X_{ii}\boldsymbol{\beta}_{ii}(\boldsymbol{\alpha}^{(k)}, \boldsymbol{\sigma}^2)\big)\big)$. Specifically,

$$\begin{aligned} H(\boldsymbol{\alpha}_{i\cdot}^{(k)}) &= 8D(\alpha_{i\cdot}^{(k)})X_{ii,\cdot(1:q)}^{\mathsf{T}}X_{ii,\cdot(1:q)}D(\alpha_{i\cdot}^{(k)}) \\ &\approx \nabla^2 b(\boldsymbol{\alpha}_{i\cdot}^{(k)}). \end{aligned}$$

To ease the presentation and computation, we write $\hat{b}(\boldsymbol{\alpha}_{i\cdot})$ from (A.4) in the form of a least squares problem

$$\begin{aligned} \hat{b}(\boldsymbol{\alpha}_{i\cdot}) &= \|Y_{ii} - X_{ii}\boldsymbol{\beta}_{ii}(\boldsymbol{\alpha}^{(k)}, \boldsymbol{\sigma}^2)\|^2 \\ &\quad - 2\Big(Y_{ii} - X_{ii}\boldsymbol{\beta}_{ii}(\boldsymbol{\alpha}^{(k)}, \boldsymbol{\sigma}^2))\Big)^{\mathsf{T}} \\ &\quad \times [2X_{ii,\cdot(1:q)}D(\boldsymbol{\alpha}_{i\cdot}^{(k)})](\boldsymbol{\alpha}_{i\cdot} - \boldsymbol{\alpha}_{i\cdot}^{(k)}) \\ &\quad + \tfrac{1}{2}(2)(\boldsymbol{\alpha}_{i\cdot} - \boldsymbol{\alpha}_{i\cdot}^{(k)})^{\mathsf{T}}[2D(\boldsymbol{\alpha}_{i\cdot}^{(k)})X_{ii,\cdot(1:q)}^{\mathsf{T}}] \\ &\quad \times [2X_{ii,\cdot(1:q)}D(\boldsymbol{\alpha}_{i\cdot}^{(k)})](\boldsymbol{\alpha}_{i\cdot} - \boldsymbol{\alpha}_{i\cdot}^{(k)}) \\ &= \mathbf{v}^{\mathsf{T}}\mathbf{v} - 2\mathbf{v}^{\mathsf{T}}X_{ii}^{*}\boldsymbol{\gamma} + \boldsymbol{\gamma}^{\mathsf{T}}(X_{ii}^{*})^{\mathsf{T}}X_{ii}^{*}\boldsymbol{\gamma} \\ &= \|\mathbf{v} - X_{ii}^{*}\boldsymbol{\gamma}\|^2 \\ &= \|Y_{ii}^{*} - X_{ii}^{*}\boldsymbol{\alpha}_{i\cdot}\|^2 \end{aligned}$$

where

$$\begin{aligned}\mathbf{v} &= Y_{ii} - X_{ii}\boldsymbol{\beta}_{ii}(\boldsymbol{\alpha}^{(k)}, \boldsymbol{\sigma}^2),\\ X^*_{ii} &= 2X_{ii,\cdot(1:q)}D(\boldsymbol{\alpha}^{(k)}_{i\cdot}),\\ \boldsymbol{\gamma} &= \boldsymbol{\alpha}_{i\cdot} - \boldsymbol{\alpha}^{(k)}_{i\cdot},\\ Y^*_{ii} &= Y_{ii} + X_{ii,\cdot(1:q)}\boldsymbol{\alpha}^{2,(k)}_{i\cdot} - X_{ii,\cdot(q+1)}\sigma_i^2.\end{aligned}$$

Replacing b in (A.3) with $\hat{b}$ we obtain the approximate objective function $\hat{Q}_{\lambda,i}(\boldsymbol{\alpha}_{i\cdot}|\boldsymbol{\alpha}^{(k)}_{i\cdot})$ given in (3.3). Since (3.3) is a standard regularized least squares problem, we minimize (3.3) using a coordinate descent algorithm to obtain $\hat{\boldsymbol{\alpha}}_{i\cdot}$ as detailed in Section B.2.

A.2 Theoretical result regarding proximal Newton update

Let $\Delta(\boldsymbol{\alpha}^{(k)}_{i\cdot}) = \hat{\boldsymbol{\alpha}}_{i\cdot} - \boldsymbol{\alpha}^{(k)}_{i\cdot}$ where $\hat{\boldsymbol{\alpha}}_{i\cdot}$ is the minimizer of (3.3) and according to a line search strategy let

$$\boldsymbol{\alpha}^{(k+1)}_{i\cdot} = \boldsymbol{\alpha}^{(k)}_{i\cdot} + t\Delta(\boldsymbol{\alpha}^{(k)}_{i\cdot})$$

for some $t > 0$. Following the proof of Proposition 2.3 in Lee *et al.* (2014), we can verify the following theorem.

Theorem 1 *Let* $H(\boldsymbol{\alpha}^{(k)}_{i\cdot}) = 8D(\boldsymbol{\alpha}^{(k)}_{i\cdot})X_{ii}^{\mathsf{T}}X_{ii}D(\boldsymbol{\alpha}^{(k)}_{i\cdot})$. *Then*

$$Q_{i,\lambda}(\boldsymbol{\alpha}^{(k+1)}_{i\cdot}, \sigma_i^2) \le Q_{i,\lambda}(\boldsymbol{\alpha}^{(k)}_{i\cdot}, \sigma_i^2) - t\Delta(\boldsymbol{\alpha}^{(k)}_{i\cdot})^{\mathsf{T}}H(\boldsymbol{\alpha}^{(k)}_{i\cdot})\Delta(\boldsymbol{\alpha}^{(k)}_{i\cdot}) + O(t^2).$$

Thus, by Theorem 1, if $H(\boldsymbol{\alpha}^{(k)}_{i\cdot})$ is positive definite, we can choose $t > 0$ so that $Q_{i,\lambda}(\boldsymbol{\alpha}^{(k+1)}_{i\cdot}, \sigma_i^2) < Q_{i,\lambda}(\boldsymbol{\alpha}^{(k)}_{i\cdot}, \sigma_i^2)$. That is, the update of $\boldsymbol{\alpha}_{i\cdot}$ results in a decrease of the objective function (3.2).

B Algorithm

In our block descent algorithm, we minimize (3.1) with respect to $\boldsymbol{\sigma}^2, \boldsymbol{\alpha}, \boldsymbol{\phi}$, and $\boldsymbol{\psi}$ in turn. For $i = 1, \ldots, p$, we first update σ_i^2 by minimizing (3.2) using least squares estimation followed by an update of $\boldsymbol{\alpha}_{i\cdot}$ by minimizing (3.3) using a coordinate descent method. We denote by $X_{ij,\cdot k}$ the kth column of X_{ij} for $k = 1, \ldots, q$ $(i \neq j)$ or $k = 1, \ldots, q+1$ $(i = j)$. We detail, respectively in Appendices B.1 and B.2, the updates of σ_i^2 and the coordinate descent updates of α_{il} for $l = 1, \ldots, q$. A summary of the final algorithm is given by Appendix B.3.

B.1 Update of σ_i^2

The parameter $\hat{\sigma}_i^2$ is updated using least squares methods. More precisely, the gradient of (3.2) with respect to σ_i^2 is

$$\frac{\partial Q_{\lambda,i}(\boldsymbol{\alpha}_{i\cdot}, \sigma_i^2)}{\partial \sigma_i^2} = -2X_{ii,\cdot(q+1)}^{\mathsf{T}}(Y_{ii} - X_{ii}\boldsymbol{\beta}_{ii}(\boldsymbol{\alpha}, \boldsymbol{\sigma}^2)).$$

By solving $\frac{\partial Q_{\lambda,i}(\boldsymbol{\alpha}_{i\cdot},\sigma_i^2)}{\partial \sigma_i^2} = 0$, we obtain the update

$$\sigma_i^2 \leftarrow \max\left\{ \frac{X_{ii,\cdot(q+1)}^{\mathsf{T}} \left(Y_{ii} - \sum_{l=1}^{q} X_{ii,\cdot l}\alpha_{il}^2\right)}{X_{ii,\cdot(q+1)}^{\mathsf{T}} X_{ii,\cdot(q+1)}}, 0 \right\} \tag{B.1}$$

where $\max\{a, 0\}$ is used to avoid negative results of the update.

B.2 Update of α_{il}

Let $r_{ij} = Y_{ij}^* - \sum_{\substack{k=1 \\ k\neq l}}^{q} X_{ij,\cdot k}^* \alpha_{ik}$, where Y_{ij}^* and X_{ij}^* are specified in (3.4). Then we rewrite (3.3) as

$$\begin{aligned}
\hat{Q}_{\lambda,i}(\boldsymbol{\alpha}_{i\cdot}) = & \sum_{j=1}^{p} \|r_{ij} - X_{ij,\cdot l}^* \alpha_{il}\|^2 \\
& + \lambda \sum_{\substack{k=1 \\ k\neq l}}^{q} \big((1-\xi)\tfrac{1}{2}\alpha_{ik}^2 + \xi|\alpha_{ik}| \big) \\
& + \lambda \big((1-\xi)\tfrac{1}{2}\alpha_{il}^2 + \xi|\alpha_{il}| \big).
\end{aligned}$$

The gradient with respect to α_{il} is

$$\begin{aligned}
\frac{\partial \hat{Q}_{\lambda,i}(\alpha_{il})}{\partial \alpha_{il}} = & -2 \sum_{j=1}^{p} (X_{ij,\cdot l}^*)^{\mathsf{T}} (r_{ij} - X_{ij,\cdot l}^* \alpha_{il}) \\
& + \lambda \big((1-\xi)\alpha_{il} + \xi \operatorname{sign}(\alpha_{il}) \big).
\end{aligned}$$

Following the main argument by Friedman *et al.* (2010), the coordinate-wise update for α_{il} is of the form

$$\alpha_{il} \leftarrow \frac{S\left(2\sum_{j=1}^{p} (X_{ij,\cdot l}^*)^{\mathsf{T}} r_{ij}, \lambda\xi\right)}{2\sum_{j=1}^{p} (X_{ij,\cdot l}^*)^{\mathsf{T}} X_{ij,\cdot l}^* + \lambda(1-\xi)}, \tag{B.2}$$

where $S(A, \lambda\xi) = \operatorname{sign}(A)(|A| - \lambda\xi)_+$.

B.3 Algorithm to update $\boldsymbol{\alpha}, \boldsymbol{\sigma}^2, \boldsymbol{\phi}, \boldsymbol{\psi}$

For a given q and sequence of λ values $0 \leq \lambda_1, \ldots, \lambda_M$, the overall procedure to estimate the parameters: $\boldsymbol{\alpha}, \boldsymbol{\sigma}^2, \boldsymbol{\phi}, \boldsymbol{\psi}$ is described by Algorithm 1. Note that estimates obtained with λ_{s-1} are used as initial values for the estimation with λ_s, $s = 2, \ldots, M$.

Algorithm 1 Cyclical block descent method for minimization of regularized least squares objective function (3.1).

Set initial values $\hat{\boldsymbol{\alpha}}^{(0)}, \hat{\boldsymbol{\sigma}}^{2,(0)}, \hat{\boldsymbol{\phi}}^{(0)}$ and $\hat{\boldsymbol{\psi}}^{(0)}$
for $s = 1$ to M **do**
  $\boldsymbol{\sigma}^2 := \hat{\boldsymbol{\sigma}}^{2,(s-1)}$
  $\boldsymbol{\alpha} := \hat{\boldsymbol{\alpha}}^{(s-1)}$
  $\boldsymbol{\phi} := \hat{\boldsymbol{\phi}}^{(s-1)}$
  $\boldsymbol{\psi} := \hat{\boldsymbol{\psi}}^{(s-1)}$
  while Relative function convergence not achieved **do**
    for $i = 1$ to p **do**
      Update σ_i^2 using (B.1)
      Update $\boldsymbol{\alpha}_{i\cdot}$ using cyclical descent over α_{il}, $l = 1, \ldots, q$ using (B.2)
      Apply line search for $\boldsymbol{\alpha}_{i\cdot}$.
    end for
    update $\boldsymbol{\phi}$ using quasi-Newton
    update $\boldsymbol{\psi}$ using quasi-Newton
  end while
  $\hat{\boldsymbol{\sigma}}^{2,(s)} := \boldsymbol{\sigma}^2$
  $\hat{\boldsymbol{\alpha}}^{(s)} := \boldsymbol{\alpha}$
  $\hat{\boldsymbol{\phi}}^{(s)} := \boldsymbol{\phi}$
  $\hat{\boldsymbol{\psi}}^{(s)} := \boldsymbol{\psi}^2$
end for

Supplementary material

S.1 Plots and detail information of BCI data used in the analysis

Figure 4 shows locations of the 9 selected tree species. Plots of 13 spatial covariates used for analysis are depicted in Figure 5. Finally, information regarding family, seed dispersal mode, life form, as well as the abundance of each of the 86 selected tree species is reported in Table 9.

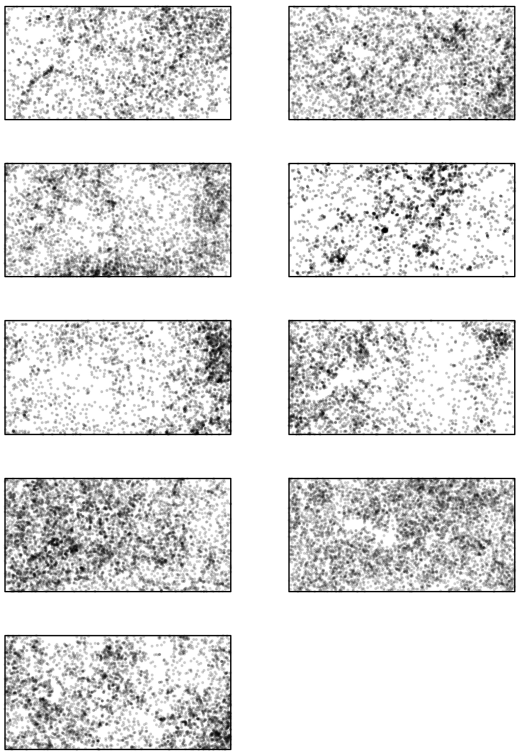

Figure 4: Locations of 9 selected tree species (from top to bottom): 1st column: Capparis frondosa, Hirtella triandra, Protium tenuifolium, Mouriri myrtilloides, and Tetragastris panamensis; 2nd column: Garcinia intermedia, Psychotria horizontalis, Protium panamense, and Swartzia simplex.

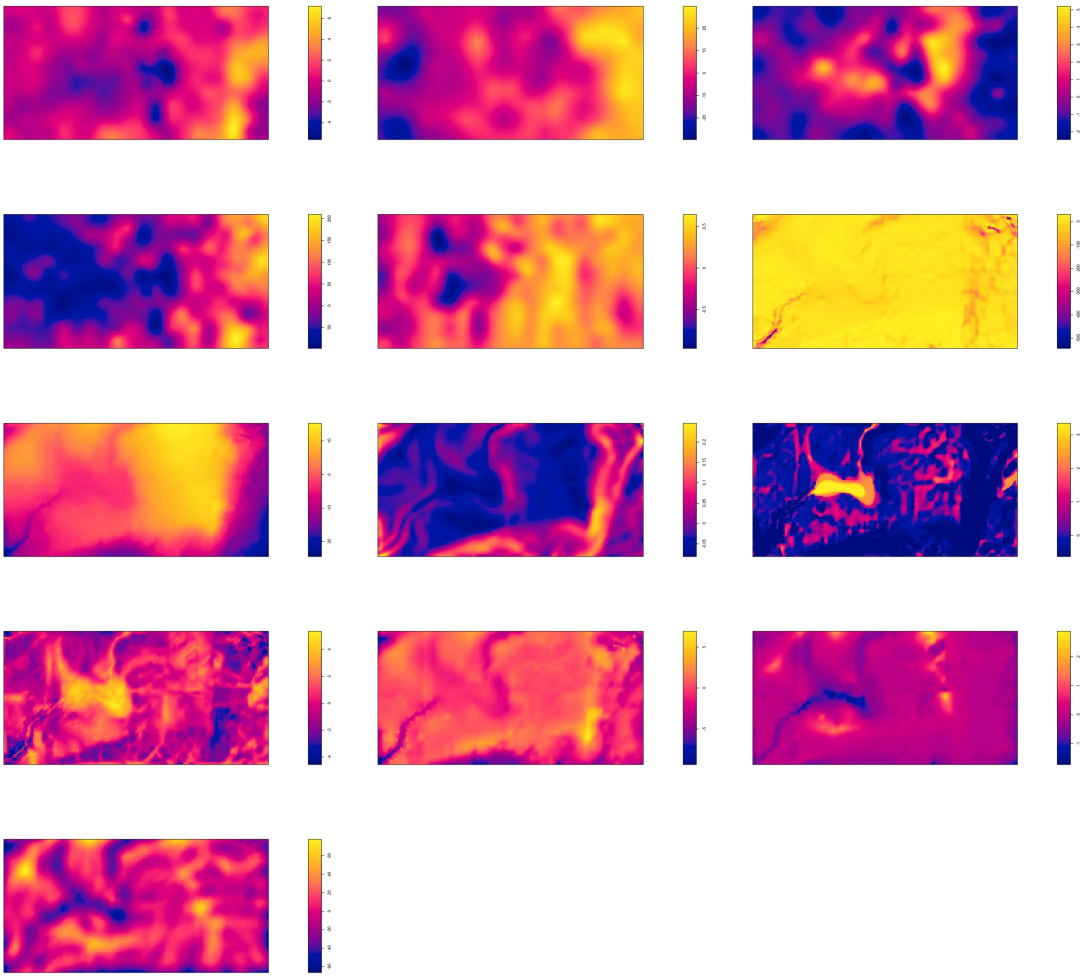

Figure 5: Covariates involved in the analysis (from left to right): 1st row: Copper content (mg/kg of soil) in the surface soil, mineralization needs for Nitrogen (mg/kg of soil) after a 30-day incubation period and Phosphorus content (mg/kg of soil) in the surface soil; 2nd row: Potassium content (mg/kg of soil) in the surface soil, pH content in the surface soil, and incoming mean annual solar radiation; 3rd row: elevation, slope, and multiresolution index of valley bottom flatness; 4th row: topographic wetness index, difference from the mean value in 15 pixels search radius, and deviation from mean value in 15 pixels search radius; 5th row: convergence index (search radius) with direction to the center cell.

Table 9: Family, mode of seed dispersal, life form and abundance of each of the selected 86 species of tree in BCI data.

No.	Mnemonic	Species	Family	Seed dispersal	Life form	Abundance (Per plot)
1	acaldi	Acalypha diversifolia	Euphorbiaceae	Explosion	Shrub	1146
2	alibed	Alibertia edulis	Rubiaceae	Big Bird-Mammals	Shrub	450
3	alsebl	Alseis blackiana	Rubiaceae	Wind	Understory	8983
4	annoac	Annona acuminata	Annonaceae	Bird-Mammals	Understory	556
5	aspicr	Aspidosperma spruceanum	Apocynaceae	Wind	Tree	524
6	beilpe	Beilschmiedia pendula	Lauraceae	Big Bird-Mammals	Tree	2252
7	brosal	Brosimum alicastrum	Moraceae	Bat-Mammals	Tree	878
8	calolo	Calophyllum longifolium	Clusiaceae	Bat-Mammals	Tree	1878
9	cappfr	Capparis frondosa	Capparaceae	Bird-Mammals	Shrub	3112
10	caseac	Casearia aculeata	Salicaceae	Bird-Mammals	Understory	546
11	cassel	Cassipourea elliptica	Rhizophoraceae	Bird-Mammals	Understory	1149
12	cecrin	Cecropia insignis	Urticaceae	Bat-Bird-Mammals	Tree	894
13	chr1ec	Chrysochlamys eclipes	Clusiaceae	Bird-Mammals	Understory	423
14	chr2ar	Chrysophyllum argenteum	Sapotaceae	Big Bird-Mammals	Tree	775
15	coccma	Coccoloba manzinellensis	Polygonaceae	Bird-Mammals	Midstory	479
16	cordbi	Cordia bicolor	Boraginaceae	Bird-Mammals	Midstory	693
17	cordla	Cordia lasiocalyx	Boraginaceae	Bird-Mammals	Understory	1188
18	cou2cu	Coussarea curvigemmia	Rubiaceae	Bird-Mammals	Understory	2111
19	crotbi	Croton billbergianus	Euphorbiaceae	Explosion	Understory	635
20	cupasy	Cupania seemannii	Sapindaceae	Bird-Mammals	Understory	1609
21	des2pa	Desmopsis panamensis	Annonaceae	Big Bird-Mammals	Understory	11654
22	drypst	Drypetes standleyi	Putranjivaceae	Bat-Mammals	Tree	2210
23	eugeco	Eugenia coloradoensis	Myrtaceae	Big Bird-Mammals	Understory	653
24	eugega	Eugenia galalonensis	Myrtaceae	Bird-Mammals	Midstory	2095
25	eugene	Eugenia nesiotica	Myrtaceae	Big Bird-Mammals	Understory	634
26	eugeoe	Eugenia oerstediana	Myrtaceae	Bird-Mammals	Understory	1956
27	faraoc	Faramea occidentalis	Rubiaceae	Big Bird-Mammals	Understory	25739
28	gar2in	Garcinia intermedia	Clusiaceae	Big Bird-Mammals	Tree	5036
29	gar2ma	Garcinia madruno	Clusiaceae	Mammals	Tree	420
30	guargu	Guarea guidonia	Meliaceae	Big Bird-Mammals	Midstory	1993
31	guarsp	Guarea bullata	Meliaceae	Bird-Mammals	Tree	793
32	guatdu	Guatteria dumetorum	Annonaceae	Bird-Mammals	Tree	941
33	gustsu	Gustavia superba	Lecythidaceae	Big Bird-Mammals	Understory	745
34	hassfl	Hasseltia floribunda	Salicaceae	Bird-Mammals	Understory	424
35	heisco	Heisteria concinna	Erythropalaceae	Big Bird-Mammals	Midstory	900
36	herrpu	Herrania purpurea	Malvaceae	Big Bird-Mammals	Shrub	601
37	hirttr	Hirtella triandra	Chrysobalanaceae	Big Bird-Mammals	Midstory	4552
38	hybapr	Hybanthus prunifolius	Violaceae	Explosion	Shrub	30130
39	ingama	Inga marginata	Fabaceae-mimosoideae	Big Bird-Mammals	Understory	804
40	ingaqu	Inga nobilis	Fabaceae-mimosoideae	Big Bird-Mammals	Understory	598
41	ingas1	Inga acuminata	Fabaceae-mimosoideae	unknown	Understory	619
42	ingaum	Inga umbellifera	Fabaceae-mimosoideae	Big Bird-Mammals	Understory	823
43	laciag	Lacistema aggregatum	Lacistemataceae	Bird-Mammals	Shrub	1395
44	laetth	Laetia thamnia	Salicaceae	Bird-Mammals	Understory	432
45	loncla	Lonchocarpus heptaphyllus	Fabaceae-papilionoideae	Wind	Tree	712
46	malmsp	Mosannona garwoodii	Annonaceae	Bird-Mammals	Midstory	530
47	maquco	Maquira guianensis	Moraceae	Bird-Mammals	Midstory	1352
48	micoaf	Miconia affinis	Melastomataceae	Bird-Mammals	Shrub	469
49	micoar	Miconia argentea	Melastomataceae	Bird-Mammals	Understory	688
50	micone	Miconia nervosa	Melastomataceae	Bird-Mammals	Shrub	412
51	mourmy	Mouriri myrtilloides	Melastomataceae	Bird-Mammals	Shrub	7241
52	ocotce	Ocotea cernua	Lauraceae	Bird-Mammals	Midstory	477
53	ocotwh	Ocotea whitei	Lauraceae	Big Bird-Mammals	Tree	406
54	oenoma	Oenocarpus mapora	Arecaceae	Big Bird-Mammals	Midstory	2049
55	ouralu	Ouratea lucens	Ochnaceae	Bird-Mammals	Shrub	1401
56	paligu	Palicourea guianensis	Rubiaceae	Bird-Mammals	Shrub	1119
57	picrla	Picramnia latifolia	Picramniaceae	Bird-Mammals	Understory	1131
58	poular	Poulsenia armata	Moraceae	Bat-Mammals	Tree	996
59	poutre	Pouteria reticulata	Sapotaceae	Big Bird-Mammals	Tree	1327
60	pri2co	Prioria copaifera	Fabaceae-caesalpinioideae	Mammals-Water	Tree	1353
61	protco	Protium costaricense	Burseraceae	Big Bird-Mammals	Tree	748
62	protpa	Protium panamense	Burseraceae	Big Bird-Mammals	Tree	3119
63	protte	Protium tenuifolium	Burseraceae	Big Bird-Mammals	Tree	3091
64	psycho	Psychotria horizontalis	Rubiaceae	Bird	Shrub	2639
65	psycma	Psychotria marginata	Rubiaceae	Bird	Shrub	834
66	pterro	Pterocarpus rohrii	Fabaceae-papilionoideae	Wind	Tree	1406
67	quaras	Quararibea asterolepis	Malvaceae	Bat-Mammals	Tree	2227
68	randar	Randia armata	Rubiaceae	Big Bird-Mammals	Understory	951
69	simaam	Simarouba amara	Simaroubaceae	Bird-Mammals	Understory	1600
70	sipapa	Siparuna pauciflora	Siparunaceae	Bird-Mammals	Shrub	481
71	sloate	Sloanea terniflora	Elaeocarpaceae	Bird-Mammals	Tree	469
72	socrex	Socratea exorrhiza	Arecaceae	Mammals	Midstory	500
73	soroaf	Sorocea affinis	Moraceae	Bird-Mammals	Shrub	2404
74	stylst	Stylogyne turbacensis	Myrsinaceae	Bird-Mammals	Understory	833
75	swars1	Swartzia simplex	Fabaceae-papilionoideae	Big Bird-Mammals	Understory	3189
76	swars2	Swartzia simplex	Fabaceae-papilionoideae	Big Bird-Mammals	Understory	3185
77	tab2ar	Tabernaemontana arborea	Apocynaceae	Bird-Mammals	Tree	1818
78	tachve	Tachigali versicolor	Fabaceae-caesalpinioideae	Wind	Tree	2135
79	taline	Talisia nervosa	Sapindaceae	Big Bird-Mammals	Understory	746
80	talipr	Talisia croatii	Sapindaceae	Big Bird-Mammals	Understory	881
81	tet2pa	Tetragastris panamensis	Burseraceae	Big Bird-Mammals	Tree	4961
82	tri2pa	Trichilia pallida	Meliaceae	Big Bird-Mammals	Midstory	499
83	tri2tu	Trichilia tuberculata	Meliaceae	Big Bird-Mammals	Tree	11293
84	unonpi	Unonopsis pittieri	Annonaceae	Big Bird-Mammals	Midstory	667
85	virose	Virola sebifera	Myristicaceae	Big Bird-Mammals	Tree	1289
86	xyl1ma	Xylopia macrantha	Annonaceae	Bird-Mammals	Midstory	1698

S.2 Additional simulation study for $p = 5$

This section repeats the study in Section 4.2 in the main article except that a more sparse $\boldsymbol{\alpha}^\mathsf{T}$ is used, that is

$$\boldsymbol{\alpha}^\mathsf{T} = \begin{bmatrix} 0.1 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & -0.1 \end{bmatrix}.$$

Table 10: Distribution of $|q_{\text{eff}} - 2|$ (in %) over 200 simulations from a multivariate log Gaussian Cox process ($p = 5, q = 2$) using CBD for minimization.

	LSE $q \in \boldsymbol{q},\ \lambda = 0$				LASSO $q \in \boldsymbol{q},\ \lambda \in \boldsymbol{\lambda}$			
$\lvert q_{\text{eff}} - 2\rvert$	0	1	2	3	0	1	2	3
Min	10	17	69	5	12	33	53	2
1-SE	1	6	93	0	1	3	96	0

Table 11: Average RMSEs obtained from 200 simulations from a multivariate log Gaussian Cox process ($p = 5, q = 2$) for different methods of selecting q and λ.

	LSE $q \in \boldsymbol{q}, \lambda = 0$		LASSO $q \in \boldsymbol{q}, \lambda \in \boldsymbol{\lambda}$	
	Min	1-SE	Min	1-SE
$\hat{\boldsymbol{\alpha}}\hat{\boldsymbol{\alpha}}^\mathsf{T}$	0.48	0.31	0.43	0.30
$\hat{\boldsymbol{\sigma}}^2$	0.62	0.62	0.59	0.63
$\hat{\boldsymbol{\psi}}$	0.01	0.01	0.01	0.01
$\hat{\text{PV}}$	0.37	0.37	0.37	0.37