Analysis of multi-species point patterns using multivariate log Gaussian Cox processes

Rasmus Waagepetersen[1], Yongtao Guan[2], Abdollah Jalilian[3] and Jorge Mateu[4]

[1]Aalborg University, Aalborg, Denmark
[2]University of Miami, Coral Gables, USA
[3]Razi University, Kermanshah, Iran
[4]Universitat Jaume I, Castellón, Spain

Abstract

Multivariate log Gaussian Cox processes are flexible models for multivariate point patterns. However, they have so far only been applied in bivariate cases. In this paper we move beyond the bivariate case in order to model multi-species point patterns of tree locations. In particular we address the problems of identifying parsimonious models and of extracting biologically relevant information from the fitted models. The latent multivariate Gaussian field is decomposed into components given in terms of random fields common to all species and components which are species specific. This allows a decomposition of variance that can be used to quantify to which extent the spatial variation of a species is governed by common respectively species specific factors. Cross-validation is used to select the number of common latent fields in order to obtain a suitable trade-off between parsimony and fit of the data. The selected number of common latent fields provides an index of complexity of the multivariate covariance structure. Hierarchical clustering is used to identify groups of species with similar patterns of dependence on the common latent fields.

Keywords: cross correlation, cross-validation, hierarchical clustering, log Gaussian Cox process, multivariate point process, proportions of variance

1 Introduction

In tropical rain forest ecology, hypotheses regarding biodiversity are studied using large data sets containing locations of thousands of trees for each of hundreds of species. Statistical methodology based on spatial point processes is now well-established in such studies where the pattern of tree locations for each species is regarded as a realisation of a spatial point process (e.g. Seidler and Plotkin, 2006; Wiegand et al., 2007; Liu et al., 2007; Shen et al., 2009; Law et al., 2009; Lin et al.,

Corresponding author: Rasmus Waagepetersen, email: rw@math.aau.dk

2011). Often a first step is to fit regression models depending on habitat variables for the intensity function of each species. Second, variation not accounted for by the intensity function may be assessed using e.g. the inhomogeneous K-function or pair correlation function (Waagepetersen, 2007; Waagepetersen and Guan, 2009). A natural third step after such univariate (single species) analyses is to study possible interactions between species.

One approach to studying inter-species interactions is to consider cross K or cross pair correlation functions between all pairs of species possibly conducting simulation-based tests for the hypothesis of no interaction for each pair of species. As remarked in Perry et al. (2006) one problem is that it may be hard to grasp the information in the potentially high number of cross summary statistics and multiple testing becomes an issue. This essentially bivariate approach further does not provide insight in the multivariate dependence structure of several species. Finally, from the non-parametric estimates of K or pair correlation functions it is not possible to study biologically interesting questions regarding decomposition of variation according to sources common for several species (e.g. unobserved environmental covariates) and sources which are species specific (e.g. seed dispersal). To address such questions a suitable modelling framework is needed. The literature on multivariate spatial point process models is mainly restricted to the bivariate case, see for example Diggle and Milne (1983); Harkness and Isham (1983); Högmander and Särkkä (1999); Brix and Møller (2001); Allard et al. (2001); Diggle (2003); Picard et al. (2009); Liang et al. (2009); Funwi-Gabga and Mateu (2012). Two exceptions are Diggle et al. (2005) and Baddeley (2010) who modelled four- and six-variate point patterns using multivariate Poisson processes. The Poisson process assumption does, however, not seem appropriate for the clustered patterns of rain forest trees. A third is Illian et al. (2009) but the hierarchical model developed in this paper is quite specific for a case where so-called reseeders occur conditionally on locations of resprouters.

In this paper we consider a statistical analysis of two multivariate point pattern data sets containing locations of respectively six and nine species of trees. The first data set is the classical Lansing Woods data while the other contains species from the tropical rain forest plot at Barro Colorado Island. The data sets and the objectives of the analyses are described in more detail in Sections 1.1–1.2. For the analyses we develop an inferential framework for in principle an arbitrary number of species without a known hierarchy. We model multivariate point patterns using the well-known multivariate log Gaussian Cox processes (LGCPs) (Møller et al., 1998; Brix and Møller, 2001; Liang et al., 2009). The latent Gaussian fields are obtained as linear combinations of common Gaussian fields as well as Gaussian fields specific to each species as in Brix and Møller (2001) who considered the bivariate case. From a methodological point of view the first specific objective in this paper is to move beyond bivariate point processes and to address the challenges linked to the potentially quickly increasing number of parameters that occur for highly multivariate LGCPs. A second objective is to explore how biologically relevant information can be extracted from a fitted multivariate LGCP.

The number of parameters in our model depends strongly on the number q of common latent fields, and we introduce a cross-validation procedure to determine the number q. This leads to parsimonious models when the selected q is considerably

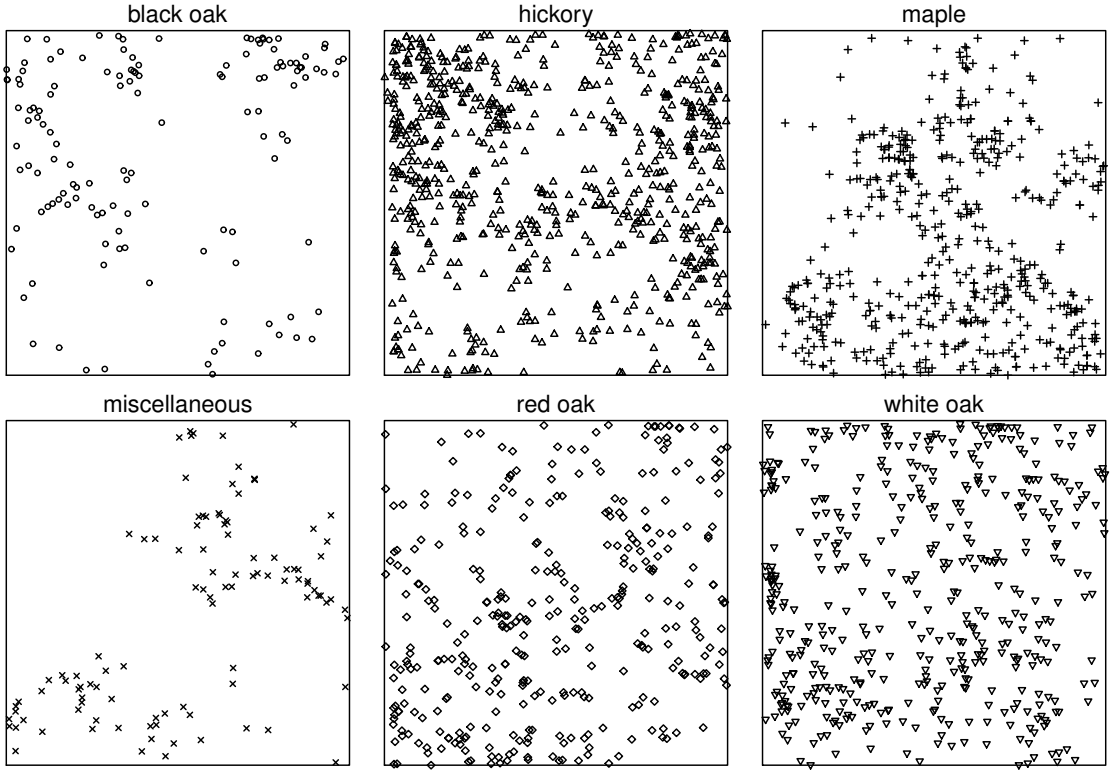

Figure 1: Plots of Lansing Woods data.

smaller than the number of species. The selected q provides an index of the complexity of the multivariate dependence structure and a test for $q = 0$ yields an overall test for the hypothesis of independence between all species. Using a decomposition of the fitted multivariate covariance structure we further quantify to which extent the spatial distribution of trees is controlled by respectively common and species specific factors. We finally identify clusters of species with similar patterns of dependence on the latent common factors.

1.1 Lansing Woods

The Lansing Woods data (Gerrard, 1969) contain locations of 2251 trees in a 19.6 acre square plot. The trees are grouped according to six species (abundances in parenthesis): black oak (135), hickory (703), maple (514), miscellaneous (105), red oak (346) and white oak (448). The point patterns of tree locations are shown in Figure 1. The original objective in Gerrard (1969) was to study a new type of competition index measuring the degree of competition among trees in a given region. Such indices are e.g. used to predict the growth of the trees. The Lansing Woods data were used in Diggle (2003) to demonstrate a range of statistical methods for spatial point patterns. However, Diggle (2003) did not consider multivariate analyses of the Lansing Woods data. Baddeley (2010) considered a multivariate Poisson point process model for the Lansing Woods data and rejected the null hypothesis of no segregation of species (i.e. the hypothesis of proportional intensity functions). Our focus is on the multivariate dependence structure of the tree species with the aim

Table 1: Family name, life form, seed dispersal mode and abundance for nine Barro Colorado Island species.

	Species	Family	Life form	Seed dispersal	Abundance
1	*Psychotria horizontalis*	*Rubiaceae*	shrub	bird	2640
2	*Protium tenuifolium*	*Burseraceae*	tree	big bird/mammal	3090
3	*Capparis frondosa*	*Capparaceae*	shrub	bird/mammal	3110
4	*Protium panamense*	*Burseraceae*	tree	big bird/mammal	3120
5	*Swartzia simplex*	*Fabaceae-papilionoideae*	understory	big bird/mammal	6370
6	*Hirtella triandra*	*Chrysobalanaceae*	midstory	big bird/mammal	4550
7	*Tetragastris panamensis*	*Burseraceae*	tree	big bird/mammal	4960
8	*Garcinia intermedia*	*Clusiaceae*	tree	big bird/mammal	5040
9	*Mouriri myrtilloides*	*Melastomataceae*	shrub	bird/mammal	7240

of obtaining a more incisive and parsimonious characterisation of the multivariate dependence than what is provided e.g. by consideration of the 15 distinct pair and cross pair correlation functions for the 6 species. We also wish to study whether the multivariate dependence structure can be related to the three major groups of species: oaks, hickories and maples.

1.2 Barro Colorado Island

The Barro Colorado Island (BCI) data (Hubbell and Foster, 1983; Condit et al., 1996; Condit, 1998) contain locations of hundreds of thousands of trees observed in a $1000\,\mathrm{m} \times 500\,\mathrm{m}$ plot. For computational reasons we are not able to handle a joint analysis of all the hundreds of species found in BCI plot. We therefore restrict attention to 9 species of intermediate abundance in the range 2500 to 7500. In addition to the point patterns of trees, a number of covariates are available regarding topography and soil properties. For each spatial location, the covariate vector is 11-dimensional and in addition to the constant 1 contains soil potassium content, pH, elevation, elevation gradient, multi-resolution index of valley bottom flatness, incoming mean solar radiation, topographic wetness index as well as soil contents of copper, mineralised nitrogen and phosphorus (plots of tree locations and selected covariates are provided in Section 1 in the supplementary material). Finally information of two types of functional traits are available: life form and mode of seed dispersal (Muller-Landau and Hardesty, 2005; Wright et al., 2007). Table 1 lists the species and their family names, life forms, modes of seed dispersal and abundances. Regarding the modes of seed dispersal, big birds are birds of biomass larger than 300 g. However, the distinction between the classes bird/mammal and big bird/mammal is not completely clear cut (Dr. Joseph Wright, personal communication).

As for the Lansing Woods data we wish to analyse in detail the multivariate dependence structure of the nine species. A further aim is to connect this analysis to the information regarding species families, life forms and modes of seed dispersal. For instance, we want to study whether species of the same family or life form tend to be positively correlated or share similar properties regarding the relative influence of common and species specific factors on their spatial pattern.

2 Multivariate log Gaussian Cox processes

We consider a multivariate Cox point process (Møller et al., 1998) $\mathbf{X} = (X_1, \ldots, X_p)$, $p > 1$, where each component X_i is a Cox process driven by a random intensity function Λ_i. That is, conditional on the Λ_i, the X_i are independent Poisson point processes each with intensity function Λ_i. The random intensity functions are of the form $\Lambda_i(\mathbf{u}) = \exp[Z_i(\mathbf{u})]$ with

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

For each i, μ_i is a deterministic function typically depending on covariates and Y_i and U_i are zero-mean Gaussian random fields. The U_i are assumed to be independent of each other and of the Y_i while the Y_i may be correlated across species. The idea is that the Y_i represent effects of e.g. unobserved environmental variables while the U_i serve to model clustering due to species specific factors such as seed dispersal. The U_i are assumed to be stationary and isotropic with variance σ_i^2 and correlation function $c_i(\,\cdot\,)$ so that $\mathbb{C}\mathrm{ov}[U_i(\mathbf{u}), U_i(\mathbf{u}+\mathbf{h})] = \sigma_i^2 c_i(\|\mathbf{h}\|)$, $\mathbf{h} \in \mathbb{R}^2$.

2.1 Model for correlated latent fields

Regarding the Y_i we assume the following factor-type model

$$\mathbf{Y}(\mathbf{u}) = [Y_1(\mathbf{u}), \ldots, Y_p(\mathbf{u})]^\mathsf{T} = \alpha \mathbf{E}(\mathbf{u})$$

where $\alpha = [\alpha_{ij}]_{ij}$ is a $p \times q$ coefficient matrix, and

$$\{\mathbf{E}(\mathbf{u})\}_{\mathbf{u}\in\mathbb{R}^2} = \{[E_1(\mathbf{u}), \ldots, E_q(\mathbf{u})]^\mathsf{T}\}_{\mathbf{u}\in\mathbb{R}^2}$$

is a q-dimensional stationary and isotropic zero-mean Gaussian process with independent components E_l. Without loss of generality we assume that $\mathbb{V}\mathrm{ar}[E_l(\mathbf{u})] = 1$ and we denote by $r_l(\,\cdot\,)$ the correlation function of E_l. Thus the multivariate covariance function for $\mathbf{E}$ is

$$R(t) = \mathbb{C}\mathrm{ov}[\mathbf{E}(\mathbf{u}), \mathbf{E}(\mathbf{u}+\mathbf{h})] = \mathrm{Diag}\big[r_1(t), \ldots, r_q(t)\big], \quad \|\mathbf{h}\| = t \geq 0,$$

where $\mathrm{Diag}[a_1, \ldots, a_n]$ means diagonal matrix with diagonal entries $a_1, \ldots, a_n$. It follows that the multivariate covariance function of $\mathbf{Y}$ is $C(t) = \alpha R(t) \alpha^\mathsf{T} = \sum_{l=1}^q \boldsymbol{\alpha}_{\cdot l} \boldsymbol{\alpha}_{\cdot l}^\mathsf{T} r_l(t)$ where $\boldsymbol{\alpha}_{\cdot l}$ is the lth column in α. Figure 2 shows the structure of the model in the case $p = 3$ and $q = 2$.

Our model for $\mathbf{Y}$ is well-known in the signal processing literature where the problem of estimating α and $\mathbf{E}$ from observations of $\mathbf{Y}$ is known as blind source separation (see e.g. Belouchrani et al., 1997). In spatial statistics the model was first proposed in Gelfand et al. (2004) as a generalisation of the so-called intrinsic or proportional correlation model (e.g. Section 5.6.4 in Chilès and Delfiner, 1999) which is obtained when $r_l = r_k$ for all l, k. Moreover, the model is a special case of the so-called linear model of coregionalisation (e.g. Section 5.6.5 in Chilès and Delfiner, 1999; Genton and Kleiber, 2014). In case of the proportional correlation model, the multivariate covariance function C only depends on α through $\alpha\alpha^\mathsf{T}$. Hence in this case one can without loss of generality take $\alpha = OD^{1/2}$ where ODO^T is the spectral factorisation of $\alpha\alpha^\mathsf{T}$. The latent processes E_l are then known as empirical orthogonal functions (Wackernagel, 2003).

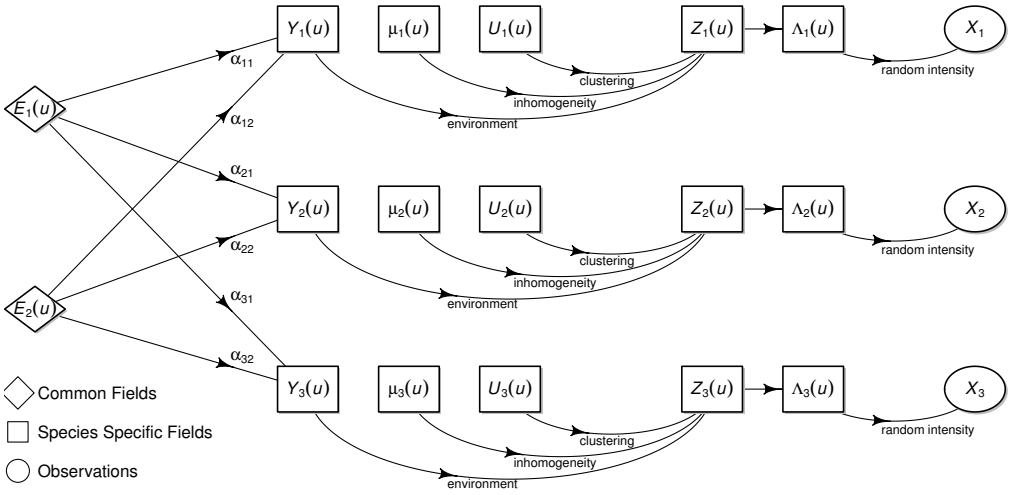

Figure 2: Structure of the multivariate log Gaussian Cox process model in the case $p = 3$ and $q = 2$.

2.2 Intensity function and multivariate pair correlation function

The intensity function of X_i is $\rho_i(\mathbf{u}) = \exp\left[\mu_i(\mathbf{u}) + \boldsymbol{\alpha}_{i\cdot}\boldsymbol{\alpha}_{i\cdot}^{\mathsf{T}}/2 + \sigma_i^2/2\right]$ where $\boldsymbol{\alpha}_{i\cdot}$ denotes the ith row of α. The matrix $g(t)$ of cross pair correlations of $\mathbf{X}$ at lag t has entries (Møller et al., 1998)

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

Large values of $\sum_{l=1}^{q} \alpha_{il}^2 r_l(t) + \sigma_i^2 c_i(t)$ lead to strong intra species correlation for X_i at lag t. Regarding between species interaction, $\sum_{l=1}^{q} \alpha_{il}\alpha_{jl} r_l(t) < 0$ (> 0) implies repulsion (attraction) between points of X_i and X_j at lag t. The cross pair correlation function g_{ij} and the intensity functions ρ_i and ρ_j determine the covariances of counts $N_i(A)$ and $N_j(B)$ of the points from X_i and X_j falling in subsets $A, B \subseteq \mathbb{R}^2$:

$$\begin{aligned}&\mathbb{C}\mathrm{ov}[N_i(A), N_j(B)]\\ &\quad = 1(i=j)\int_{A\cap B} \rho_i(\mathbf{u})\mathrm{d}\mathbf{u} + \int_A\int_B \rho_i(\mathbf{u})\rho_j(\mathbf{v})[g_{ij}(\|\mathbf{u}-\mathbf{v}\|) - 1]\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v}.\end{aligned} \tag{2.1}$$

Thus in the case $i \neq j$, g_{ij} equal to one implies that counts from respectively X_i and X_j are uncorrelated.

We model $\mu_i(\cdot)$ by a linear regression depending on the available covariate vectors of environmental variables. For the correlation functions r_l and c_i we introduce parametric models $r(\cdot\,;\phi_l)$ and $c_i(\cdot) = r(\cdot\,;\psi_i)$. The geostatistical literature offers a wide range of correlation function models, see e.g. Chilès and Delfiner (1999).

3 Least squares estimation and cross-validation

In this section we first consider a least squares approach to estimate the model parameters for a fixed q. In the least squares criterion, the dependent variables are given by log transformed non-parametric estimates of cross pair and pair correlation functions. Based on the least squares criterion we next introduce a cross-validation procedure for selecting q. In the case of no species specific latent variation $\sigma_i^2 = 0$, $i = 1, \ldots, p$, Section 3 in the supplementary material describes an alternative method of estimation for α based on spectral decomposition. The least squares estimation and cross-validation methods are assessed in a simulation study in Appendix B.

3.1 Non-parametric estimation

Fitted regression models $\hat{\rho}_i(\,\cdot\,)$ are obtained in a standard way using composite likelihood, see e.g. Waagepetersen (2007) and Waagepetersen and Guan (2009). In a second step we obtain non-parametric estimates (Baddeley et al., 2000; Møller and Waagepetersen, 2003) $\hat{g}_{ij}$ of the cross pair correlation functions:

$$\hat{g}_{ij}(t) = \frac{1}{2\pi t} \sum_{\substack{\mathbf{u}\in X_i\cap W, \mathbf{v}\in X_j\cap W \\ \mathbf{u}\neq\mathbf{v}}} \frac{k_b(t - \|\mathbf{u}-\mathbf{v}\|)}{\hat{\rho}_i(\mathbf{u})\hat{\rho}_j(\mathbf{v})|W \cap W_{\mathbf{u}-\mathbf{v}}|} \tag{3.1}$$

where W is the observation window, k_b is a kernel function depending on a smoothing parameter (bandwidth) $b > 0$, $|\,\cdot\,|$ is area and $W_{\mathbf{h}}$ denotes the translate of W by the vector $\mathbf{h} \in \mathbb{R}^2$ (Møller and Waagepetersen, 2003).

3.2 Least squares estimation

For each ij we obtain a parametric model $\log g_{ij}(\,\cdot\,;\boldsymbol{\theta}) = \sum_{l=1}^{q} \beta_{ijl} r(\,\cdot\,;\phi_l) + 1(i=j)\sigma_i^2 r(\,\cdot\,;\psi_i)$ where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{\sigma}^2)$ with $\boldsymbol{\phi} = (\phi_l)_l$, $\boldsymbol{\psi} = (\psi_i)_i$, $\boldsymbol{\sigma}^2 = (\sigma_i^2)_i$, $\beta_{ijl} = \alpha_{il}\alpha_{jl}$, and $\boldsymbol{\alpha} = (\alpha_{ij})_{ij} = (\alpha_{11}, \alpha_{12}, \ldots, \alpha_{pq})$ is the vector of lexicographically ordered entries in α. To estimate $\boldsymbol{\theta}$ we generalise the approaches in Møller et al. (1998) and Brix and Møller (2001). Consider distinct lags $0 < t_1 < \cdots < t_L$ where $L \geq q$. Then for each ij we let $\mathbf{y}_{ij} = [\log \hat{g}_{ij}(t_k)]_k$ and

$$\hat{\mathbf{y}}_{ij}(\boldsymbol{\theta}) = [\log g_{ij}(t_k;\boldsymbol{\theta})]_k = R(\boldsymbol{\phi})\boldsymbol{\beta}_{ij} + 1(i=j)\sigma_i^2 \mathbf{R}_U(\psi_i)$$

where $\boldsymbol{\beta}_{ij} = (\beta_{ij1}, \ldots \beta_{ijq})^{\mathsf{T}}$, $R(\boldsymbol{\phi})$ is $L \times q$ with klth entry $r(t_k;\phi_l)$ and $\mathbf{R}_U(\psi_i) = [r(t_k;\psi_i)]_k$ is $L \times 1$.

We then minimise

$$Q = \sum_{ij} [\mathbf{y}_{ij} - \hat{\mathbf{y}}_{ij}(\boldsymbol{\theta})]^{\mathsf{T}} W_{ij} [\mathbf{y}_{ij} - \hat{\mathbf{y}}_{ij}(\boldsymbol{\theta})] \tag{3.2}$$

with respect to $\boldsymbol{\theta}$, where $W_{ij} = \mathrm{Diag}(w_{ijk}, k = 1, \ldots, L)$ is a user-defined weight matrix of positive weights $w_{ijk} > 0$. One can show (Heinrich and Liebscher, 1997) that $\mathbb{V}\mathrm{ar} \log \hat{g}_{ij}(t_k) \approx 1/g_{ij}(t_k)$. Moreover, contributions from indices ij equal those for ji so that off-diagonal elements count twice. We hence choose weights $w_{ijk} = \hat{g}_{ij}(t_k)/2$ when $i \neq j$ and $w_{iik} = \hat{g}_{ii}(t_k)$.

3.3 Identifiability

Each cross pair correlation function g_{ij} is invariant to a) simultaneous permutation of the columns in α and the diagonal entries in $R(t)$ and b) multiplication with -1 of a column in α. Hence if one local minimum is found for Q given by (3.2), there will be $q!2^q - 1$ other local minima with the same value of Q. Rather than imposing constraints to resolve this identifiability issue our strategy is to restrict attention to estimates of functions of α which are invariant to the mentioned transformations of α. However, we need a notion of 'local identifiability' stating that the aforementioned local minima actually exist. In particular one question is how large a q can be used.

To address this question we consider the Hessian matrix (A.1) of Q with respect to $\boldsymbol{\alpha}$ which is given in Appendix A. Under an appropriate asymptotic setting with the non-parametric estimates $\hat{g}_{ij}$ tending to their true values, the Hessian matrix converges to $2(\mathrm{d}\boldsymbol{\beta}^\mathsf{T}/\mathrm{d}\boldsymbol{\alpha})\underline{R}^\mathsf{T}\underline{R}(\mathrm{d}\boldsymbol{\beta}/\mathrm{d}\boldsymbol{\alpha})$ where $\underline{R}$ is a block diagonal matrix with p^2 diagonal blocks $\mathrm{Diag}(\sqrt{w_{ijk}}, k = 1, \ldots, L)R(\boldsymbol{\phi})$ $i, j = 1, \ldots, p$. Hence, $\underline{R}$ has full rank if and only if $R(\boldsymbol{\phi})$ has full rank. This will in general be the case for any $q \leq L$ if all $\phi_l, l = 1, \ldots, q$, are distinct. Appendix A also provides the entries in the $pq \times p^2 q$ matrix $\mathrm{d}\boldsymbol{\beta}/\mathrm{d}\boldsymbol{\alpha}$. For this matrix to be of full rank it is sufficient (although not necessary) that all $\alpha_{ij} \neq 0$. Hence if the true ϕ_l are all distinct and the true α_{ij} are all non-zero, the object function will at least asymptotically have a local minimum with respect to α at the true parameter value for any $q \leq L$. These theoretical considerations thus do not rule out consideration of large q. However in practice the optimisation of the object function becomes increasingly cumbersome for increasing q and we have restricted attention to $q \leq p$.

We minimise the object function Q using a combination of a quasi-Newton algorithm and a spectral projected gradient method. Specifically we use the *R*-procedure `optimx` with "method" equal to `BFGS` or `spg` and supply the analytical expressions for the gradient and Hessian (the latter for the purpose of evaluating criteria for local minima), see Appendix A in this paper and Section 2 in the supplementary material. The choice of the starting point for the minimisation is crucial. In particular, $\boldsymbol{\alpha} = \mathbf{0}$ is a stationary point for Q with respect to $\boldsymbol{\alpha}$ since the derivative of Q with respect to $\boldsymbol{\alpha}$ is always zero when $\boldsymbol{\alpha}$ is the zero vector $\mathbf{0}$. We used as starting point a crude estimate of $\boldsymbol{\alpha}$ obtained using a spectral method (Section 3 in the supplementary material) or picked a random starting point centred around $\mathbf{0}$ if the spectral method failed.

4 Estimation of number of latent factors

To determine q we apply a variant of K-fold cross-validation (e.g. Hastie et al., 2013). That is, we split the indices ijk, $i \neq j$, into K sets $S_1, \ldots, S_K$. For each q and $c = 1, \ldots, K$ we then obtain an estimate $\hat{\boldsymbol{\theta}}_c$ by minimising (3.2) with w_{ijk} replaced by 0 for $ijk \in S_c$. A cross-validation score is then obtained by

$$CV(q) = \sum_{c=1}^{K} \sum_{ijk \in S_c} w_{ijk}[y_{ijk} - \hat{y}_{ijk}(\hat{\boldsymbol{\theta}}_c)]^2. \tag{4.1}$$

We do not include diagonal indices iik in the sets S_c. This is because the within species log pair correlation functions $\log g_{ii}$ have both species specific components and components due to the common random factors. They therefore do not provide much information about q. Including such indices in the S_c further makes the estimation of the species specific parameters less stable. For a given ij, y_{ijk} and $y_{ijk'}$ are strongly correlated when k and k' are close. Hence to obtain a sufficient sensitivity of the cross-validation score we need to leave out blocks of consecutive indices.

To obtain the subsets 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. Moreover, if ijk, $i < j$ is assigned to S_c so is jik. That is, the S_c are symmetric in the sense that $ijk \in S_c$ implies $jik \in S_c$. In a simulation study a value of b equal to 50 % of the number of lags L worked well. This choice of b is also used in the applications in Section 6. Often K between 5 and 10 are used (Hastie et al., 2013). We chose $K = 8$ to use efficient parallel computing on a server with 8 CPUs. Following the discussion in Section 3.3 we consider in practice q in the range $0, \ldots, p$.

5 Inferences regarding multivariate dependence structure

The first pertinent question is whether species are at all correlated. To assess this we use the least squares criterion Q with $q = 0$ as a test statistic and compare the observed Q with its distribution obtained using a parametric bootstrap (Davison and Hinkley, 1997) under the model fitted with $q = 0$.

The cross pair correlation functions or equivalently $\sum_{l=1}^{q} \alpha_{il}\alpha_{jl} r(t; \phi_l) = \mathbb{C}\mathrm{ov}[Y_i(\mathbf{u}), Y_j(\mathbf{u} + \mathbf{h})]$, $\|\mathbf{h}\| = t$, determine the sizes of cross covariances of count variables associated with the point processes X_i and X_j, $i \neq j$, cf. (2.1). Our approach provides parametric estimates of the cross pair correlation functions but this is not a key contribution since essentially the same information is obtained from the non-parametric estimates $\hat{g}_{ij}$. Our parametric model on the other hand allows us to decompose the covariances of the latent Gaussian fields into contributions from respectively the common fields and the species specific fields. For a given spatial lag t and species i we can consider the proportion of covariance due to the common fields of the log random intensity function Z_i,

$$PV_i(t) = \frac{\mathbb{C}\mathrm{ov}[Y_i(\mathbf{u}), Y_i(\mathbf{u} + \mathbf{h})]}{\mathbb{C}\mathrm{ov}[Z_i(\mathbf{u}), Z_i(\mathbf{u} + \mathbf{h})]} = \frac{\sum_{l=1}^{q} \alpha_{il}^2 r(t; \phi_l)}{\sum_{l=1}^{q} \alpha_{il}^2 r(t; \phi_l) + \sigma_i^2 r(t; \psi_i)}, \quad \|\mathbf{h}\| = t. \quad (5.1)$$

The proportions of variances $PV_i(t)$, $i = 1, \ldots, p$ can thus be used to group species according to how much of the variation in the log random intensity function is due to common factors as opposed to species specific factors. In analogy with Jalilian et al. (2013) the proportions of variances are further related to a certain R^2-type statistic measuring how big proportion of the variance in Λ_i is due to the common latent fields.

Considering the between species correlation structure for a given spatial lag we

have for $i \neq j$,

$$\mathbb{C}\text{orr}[Z_i(\mathbf{u}), Z_j(\mathbf{u}+\mathbf{h})] = \mathbb{C}\text{orr}[Y_i(\mathbf{u}), Y_j(\mathbf{u}+\mathbf{h})]\sqrt{PV_i(0)}\sqrt{PV_j(0)}.$$

Thus the correlation between two different log random intensity functions is factored into the correlation due to the common factors and the square roots of the proportions of variances. A small $PV_i(0)$ thus immediately implies that the latent field Z_i has a small correlation with any other species. To study the implications at the scale of counts of X_i and X_j (Section 2.2) note that $\hat{N}_i(A) = \mathbb{E}[N_i(A)|\Lambda_i] = \int_A \Lambda_i(\mathbf{u})\mathrm{d}\mathbf{u}$ can be viewed as the spatially structured part of the count $N_i(A)$ (Jalilian et al., 2013). For small A and B containing locations $\mathbf{u}$ and $\mathbf{v}$, $\hat{N}_i(A) \approx |A|\Lambda_i(\mathbf{u})$ and $\hat{N}_j(B) \approx |B|\Lambda_j(\mathbf{v})$. Thus the correlation $\mathbb{C}\text{orr}[\hat{N}_i(A), \hat{N}_j(B)]$ can be approximated by $\mathbb{C}\text{orr}[\Lambda_i(\mathbf{u}), \Lambda_j(\mathbf{v})]$. Employing further $\exp(x) \approx 1 + x$, we obtain $\mathbb{C}\text{orr}[\hat{N}_i(A), \hat{N}_j(B)] \approx \mathbb{C}\text{orr}[Z_i(\mathbf{u}), Z_j(\mathbf{v})]$.

Suppose we want to group species according to their pattern of dependence on the latent factors E_l. A simple distance measure between species i and j would be $\|\boldsymbol{\alpha}_{i\cdot} - \boldsymbol{\alpha}_{j\cdot}\|$ which is invariant to the kind of transformations of α mentioned in Section 3.3. The covariance $\mathbb{C}\text{ov}[Y_i(\mathbf{u}), Y_j(\mathbf{u})] = \boldsymbol{\alpha}_{i\cdot}^{\mathsf{T}}\boldsymbol{\alpha}_{j\cdot}$ or correlation $\mathbb{C}\text{orr}[Y_i(\mathbf{u}), Y_j(\mathbf{u})] = \boldsymbol{\alpha}_{i\cdot}^{\mathsf{T}}\boldsymbol{\alpha}_{j\cdot}/\sqrt{\|\boldsymbol{\alpha}_{i\cdot}\|^2\|\boldsymbol{\alpha}_{j\cdot}\|^2}$ seem less useful in this context. For two species which are similar in both having small values of $\|\boldsymbol{\alpha}_{i\cdot}\|^2$ and $\|\boldsymbol{\alpha}_{j\cdot}\|^2$ for example, the covariance will nevertheless be small. On the other hand, if two species have the same relative patterns of dependence on the E_l in the sense $\boldsymbol{\alpha}_{i\cdot} = k\boldsymbol{\alpha}_{j\cdot}$, $k \neq 0$, then $|\mathbb{C}\text{orr}[Y_i(\mathbf{u}), Y_j(\mathbf{u})]| = 1$ regardless of k.

In the following applications we will focus on estimation of $PV_i(0)$ and the zero lag cross correlations $\mathbb{C}\text{orr}[Y_i(\mathbf{u}), Y_j(\mathbf{u})]$ among the common fields. We also look at the mean cross correlation over a range $[0, T]$ of lags, i.e. $\int_0^T \mathbb{C}\text{orr}[Y_i(\mathbf{u}), Y_j(\mathbf{u} + (0,t)^{\mathsf{T}})]\mathrm{d}t/T$. We further perform clustering of species using the distances $\|\boldsymbol{\alpha}_{i\cdot} - \boldsymbol{\alpha}_{j\cdot}\|$. To obtain confidence intervals for correlations and proportions of variances we use a parametric bootstrap based on simulations from the fitted model. In the bootstrap we consider q as known and given by the q selected by cross-validation. This will lead to some underestimation of variances of parameter estimates but doing a full bootstrap including selection of q can be very time consuming when the number of species is large.

6 Multi-species dependence structures in temperate and tropical forests

In the following sections we return to the applications presented in Sections 1.1 and 1.2.

6.1 A joint analysis of the Lansing Woods data

Covariates are not available for the Lansing Woods data so for the intensity functions we just fit an intercept for each species. For the correlation of the latent fields we use the exponential correlation model $r(\,\cdot\,;\psi) = \exp(-\|\,\cdot\,\|/\psi)$ where $\psi > 0$ is the correlation scale parameter. We fit seven stationary multivariate log Gaussian Cox

processes with numbers of latent processes q ranging from 0 to 6. Initially we test the hypothesis of independent species ($q = 0$) using a parametric bootstrap. That is, we simulate 400 datasets under the model fitted with $q = 0$ and fit the model with $q = 0$ to all the simulated datasets. Only 0.25 % of the simulated Q lie above the observed value 65.3 of Q. Hence we reject the hypothesis of independent species. For each q the left plot in Figure 3 shows the minimised object function (3.2) while the right plot shows the cross-validation score (4.1).

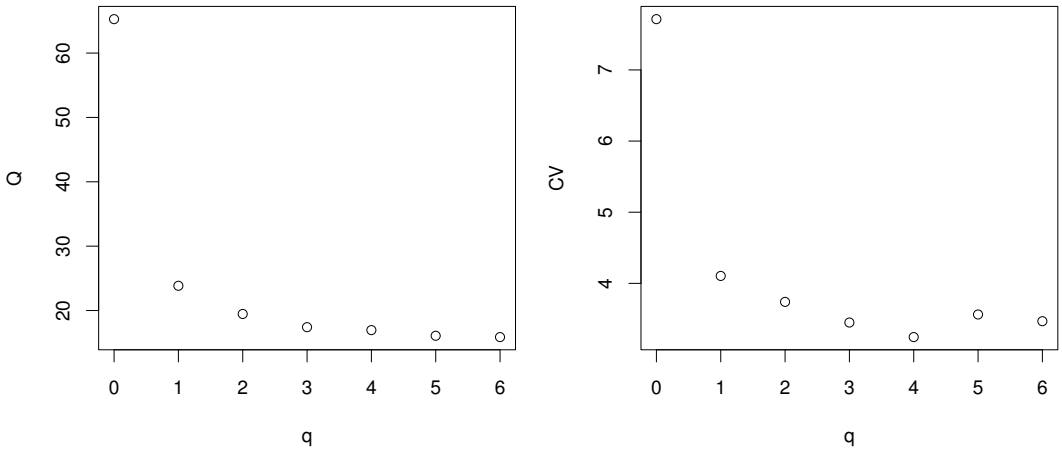

Figure 3: (Lansing data) Left: minimised object function Q. Right: cross-validation scores CV. In both plots, $q = 0, \ldots, 6$.

The smallest cross-validation score is obtained with $q = 4$. The object function drops markedly from $q = 0$ to $q = 3$ and then starts to level off. Hence $q = 4$ seems to be a good choice for the number of latent processes.

Continuing with the model fitted with $q = 4$, the left plot in Figure 4 shows the estimated cross correlations at lag zero for pairs Y_i, Y_j and Z_i, Z_j as well as 95 % parametric bootstrap confidence intervals obtained from 400 simulations of the fitted model. The indices $i, j = 1, \ldots, 6$ correspond to black oak, hickory, maple, miscellaneous, red oak, white oak. Due to the species specific random fields U_i, the cross correlations are smaller for the Z_i than for the Y_i. The estimated cross correlations for species pairs (black oak, maple), (black oak, miscellaneous) and (hickory, maple) are pretty small. However, due to large sample variation all bootstrap confidence intervals contain zero. The right plot shows estimated mean cross correlations over the range $[0, 0.25]$. Overall the patterns of cross correlations are similar in the two plots in Figure 4 but some of the confidence intervals in the right plot are narrower than in the left plot. In particular, zero is not contained in the bootstrap intervals for the mean cross correlations for the species pairs (black oak, maple), (black oak, miscellaneous) and (hickory, maple).

The left plot in Figure 5 shows estimated proportions of variances $PV_i(0)$ at lag zero with 95 % parametric bootstrap intervals. According to the estimates, the main proportion of the variance of Z_i is due to the common latent factors for black oak, hickory, maple and miscellaneous while the common latent factors and the species specific factors have roughly equal contributions for red and white oak. Thus black oak seems to be distinct from red and white oak regarding the relative influence of

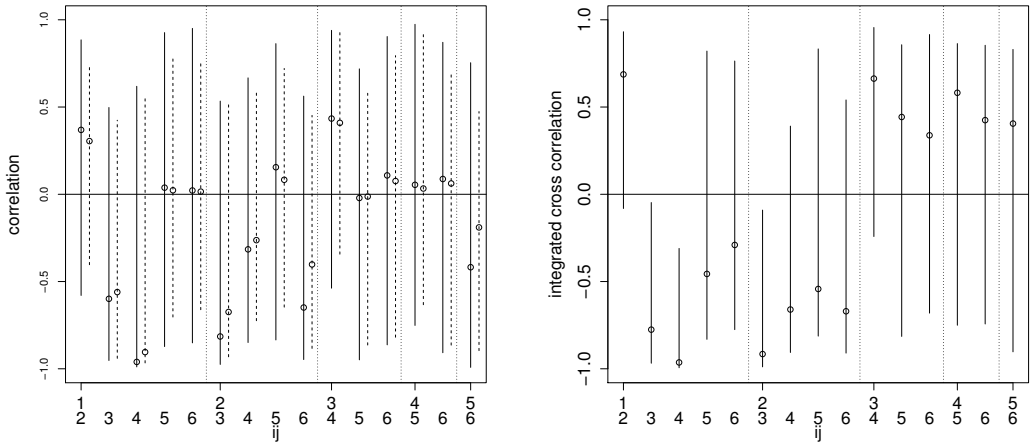

Figure 4: (Lansing data) Left plot: for each (i, j) circles show estimated cross correlations $\mathbb{C}\text{orr}[(Y_i(u), Y_j(u)]$, i, j (first circle) and $\mathbb{C}\text{orr}[Z_i(u), Z_j(u)]$ (second circle). Index $i, j = 1, \ldots, 6$ corresponds to black oak, hickory, maple, miscellaneous, red oak, white oak. Vertical lines show parametric bootstrap confidence intervals. Right plot: estimates and parametric bootstrap confidence intervals for mean cross correlations of pairs Y_i, Y_j over lag range $[0, 0.25]$.

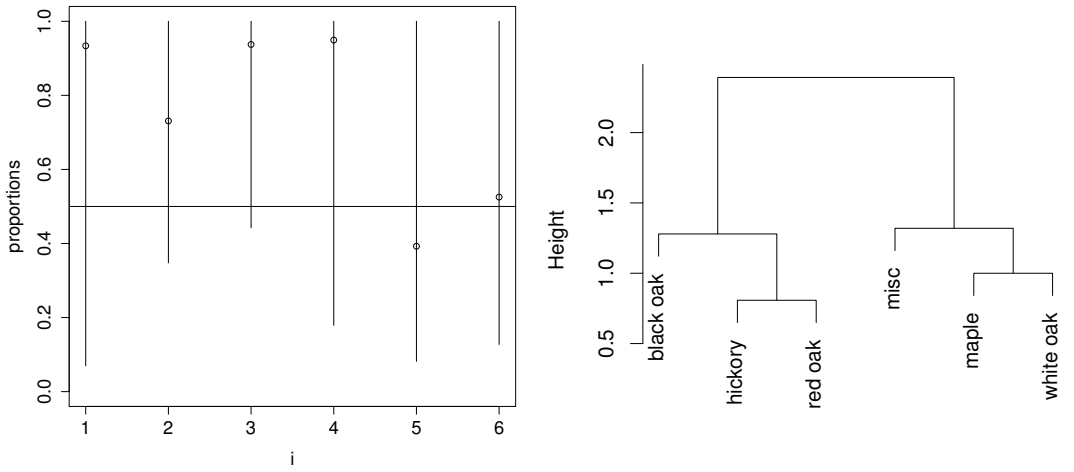

Figure 5: (Lansing data) Left: estimated proportions of variance $PV_i(0)$ due to common latent factors. Index $i = 1, \ldots, 6$ corresponds to black oak, hickory, maple, miscellaneous, red oak, white oak. Right: hierarchical clustering based on fitted $\boldsymbol{\alpha}_i$.'s.

common factors on the spatial pattern. However, as for the cross correlations, the rather wide confidence intervals show that the estimates are quite uncertain.

The right plot in Figure 5 shows a hierarchical clustering of the species based on the fitted coefficient rows $\boldsymbol{\alpha}_{i\cdot}$. In agreement with the fitted correlations, black oak and maple belong to separate clusters. The same holds for the species pairs (black oak, miscellaneous) and (hickory, maple). The clustering does not support a grouping into the coarser categories oak, hickory, maple.

The model with $q = 4$ has 28 parameters (in α and $\boldsymbol{\phi}$) used to fit 15 unique cross

pair correlation functions g_{ij}, $i < j$. Thus on average 1.9 parameters are used for each g_{ij}, $i < j$. As an assessment of model fit, Figures 2 and 3 in the supplementary material show non-parametric estimates of the L- and cross L-functions (e.g. Chapter 4 in Møller and Waagepetersen, 2003) together with 95 % pointwise envelopes obtained from simulations of the fitted model. None of these plots disclose any severe deficiencies of the fitted model.

6.1.1 Analysis without miscellaneous

The miscellaneous category corresponds not to a single species, but to a residual group of trees belonging to a mixture of less abundant species. For this reason, as pointed out by a referee, omitting this group in the analysis could potentially lead to a simpler model and hence smaller uncertainty in the inference. We therefore repeated the analysis without miscellaneous. In this case the cross-validation identified $q = 1$ as a suitable number of latent processes (left plot in Figure 6).

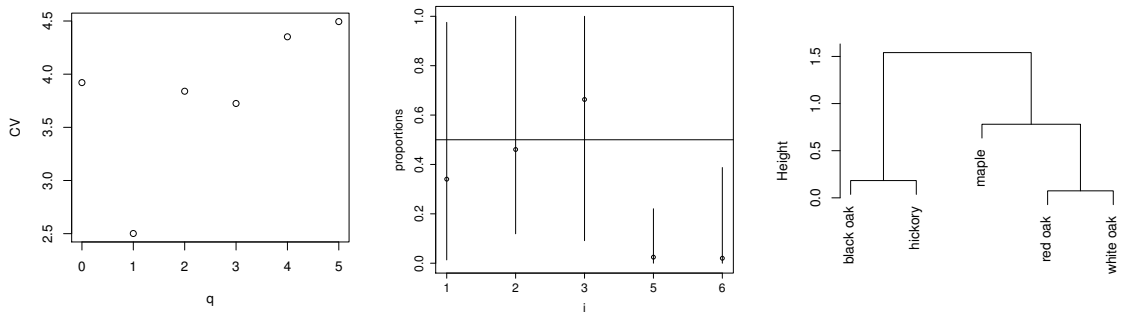

Figure 6: (Lansing data - analysis without miscellaneous) Left: cross-validation scores for $q = 0, \ldots, 5$. Middle: estimated proportions of variance $PV_i(0)$ due to common latent factors. Index $i = 1, 2, 3, 5, 6$ corresponds to black oak, hickory, maple, red oak, white oak. Right: hierarchical clustering based on fitted α_{i1}'s.

This yields a much simpler model than with the previously selected $q = 4$. In particular on average only 0.6 parameters are used for each of the 10 unique cross pair correlation function g_{ij}, $i < j$. The fitted parameters are $(\alpha_{11}, \alpha_{21}, \alpha_{31}, \alpha_{51}, \alpha_{61}) = (-0.68, -0.49, 0.87, 0.16, 0.08)$. Thus both black oak and hickory have a negative dependence on the latent field, while maple has a positive dependence on the latent field. Red and white oak both have a relatively weak dependence on the latent field. In case of $q = 1$, the correlations $\mathbb{C}\mathrm{orr}[(Y_i(u), Y_j(u)]$ are either precisely one or minus one depending on whether the corresponding parameters α_{i1} and α_{j1} are of the same or different sign. The fitted correlations $\mathbb{C}\mathrm{orr}[Z_i(u), Z_j(u)]$ are quite similar to the fitted correlations for the same pairs of species obtained in the previous analysis. However, the bootstrap confidence intervals are much narrower (plots omitted). Regarding proportions of variance (middle plot in Figure 6) all the fitted proportions of variance are smaller than those obtained with $q = 4$. This is consistent with the much more sparse representation of the correlated latent fields Y_i which implies that more variation is explained by the species specific fields. In particular the proportions of variances for red oak and white oak are close to zero with much narrower confidence intervals than in the previous analysis. The grouping from the hierarchical clustering

is a bit more consistent with the coarser groups oaks, hickory and maple than before, placing red and white oak in one cluster and with maple forming a single species cluster. However, there is still a heterogeneous cluster consisting of black oak and hickory.

6.2 Multivariate dependence and functional traits for species in the Barro Colorado Island plot

For the Barro Colorado Island data, following Section 3.1, we fit regression models for the μ_i terms using composite likelihood for each species separately. In the subsequent non-parametric estimation of the cross pair correlation functions using (3.1) the variations due to the observed covariates are filtered out. The non-parametric estimates thus capture residual correlation due e.g. to unobserved covariates, seed dispersal and other sources of correlation. As for the Lansing data we use the exponential correlation model for the latent random fields. In the following we consider similar analyses as those for the Lansing data. For ease of presentation we refer to the species by the first part of their genus, see Table 1, adding a *t.* for *Protium tenuifolium* and a *p.* for *Protium panamense*.

6.2.1 Statistical analyses

The hypothesis of independent species ($q = 0$) is rejected with a parametric bootstrap p-value of 0.5 %. For each $q = 0, \ldots, 9$ the left plot in Figure 7 shows the minimised object function Q while the right plot shows the cross-validation scores. The smallest

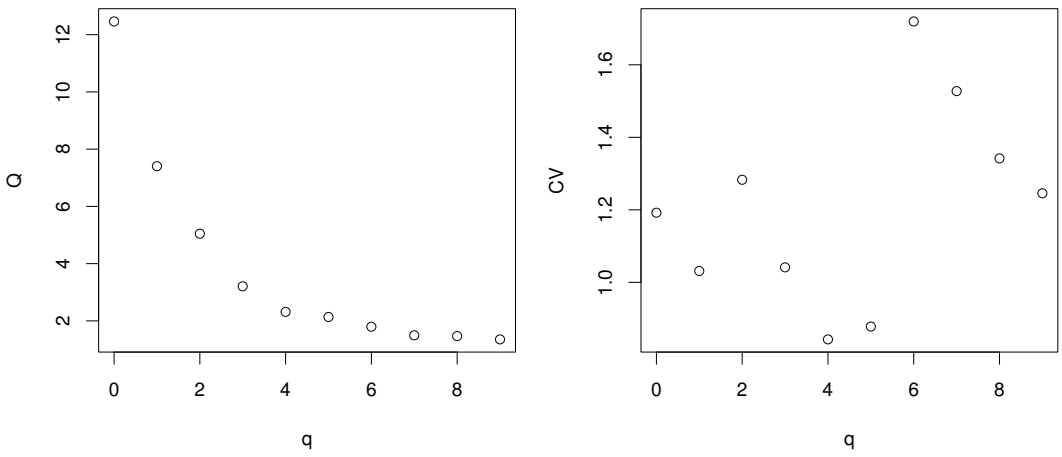

Figure 7: (BCI data) Left: minimised object function Q. Right: cross-validation scores CV. In both plots, $q = 0, \ldots, 9$.

cross-validation score is obtained with $q = 4$. This choice of q is also supported by the left plot where the decrease in the object function is relatively modest for $q > 4$.

Proceeding with the model fitted with $q = 4$, Figure 8 shows estimated cross correlations as well as 95 % parametric bootstrap confidence intervals obtained from 400 simulations of the fitted model. Most of the cross correlations (whether for Y or Z) appear to be significantly larger than zero. There is some evidence of negative correlation

between *Psychotria* and *Protium p.* while there is no evidence of positive or negative correlation for the pairs (*Psychotria,Protium t.*), (*Psychotria,Swartzia*), (*Psychotria,Hirtella*), (*Psychotria,Tetragastris*) and (*Psychotria,Garcinia*). Integrated cross correlations show a similar pattern (plot omitted) but with wider confidence intervals for some species.

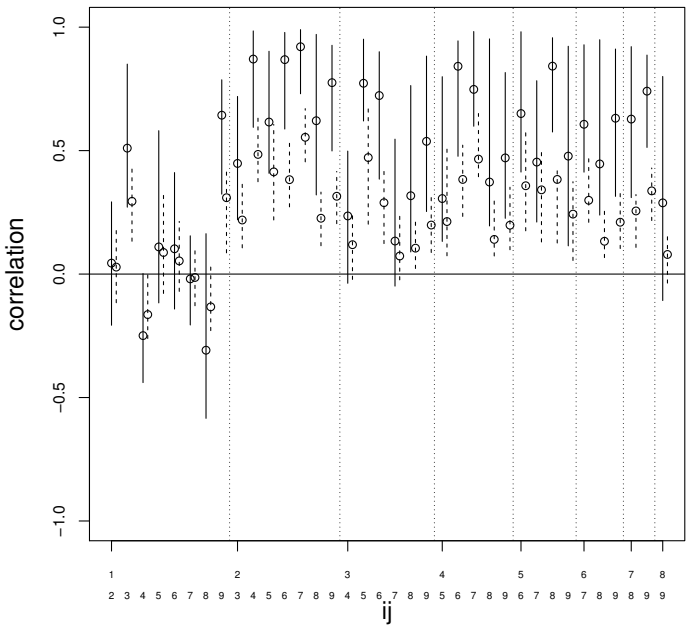

Figure 8: (BCI data) For each (i, j) circles show estimated cross correlations $\mathbb{C}\text{orr}[(Y_i(u), Y_j(u)]$, i, j (first circle) and $\mathbb{C}\text{orr}[Z_i(u), Z_j(u)]$ (second circle). Indices $i, j = 1, \ldots, 9$ correspond to *Psychotria, Protium t., Capparis, Protium p., Swartzia, Hirtella, Tetragastris, Garcinia, Mouriri*. Vertical lines show parametric bootstrap confidence intervals.

The left plot in Figure 9 shows estimated proportions of variances $PV_i(0)$ at lag zero with 95 % parametric bootstrap intervals. *Psychotria* and *Swartzia* have quite high estimated proportions of variances while *Garcinia* and *Mouriri* have the smallest proportions of variances. As for the Lansing data the estimates are quite uncertain as indicated by the width of the confidence intervals. However, the proportion of variance due to the common factors for *Garcinia* is significantly smaller than the benchmark value of 0.5. The right plot in Figure 9 shows a hierarchical clustering of the species based on the fitted coefficient rows $\boldsymbol{\alpha}_{i\cdot}$. In agreement with the fitted correlations, *Psychotria* forms its own cluster.

For the model with $q = 4$ on average 1.1 parameters are used for each unique g_{ij}, $i < j$. Figures 4-7 in the supplementary material provide model assessment using L-functions as for the Lansing data. Out of 45 unique L or cross L-functions there only appears to be issues with the two intra-species L-functions for *Swartzia* and *Garcinia*.

6.2.2 Relation to species families, life form and mode of seed dispersal

Protium t., Protium p. and *Tetragastris* all belong to the family *Burseraceae* while the other species belong to distinct families. It is interesting to see that the family related

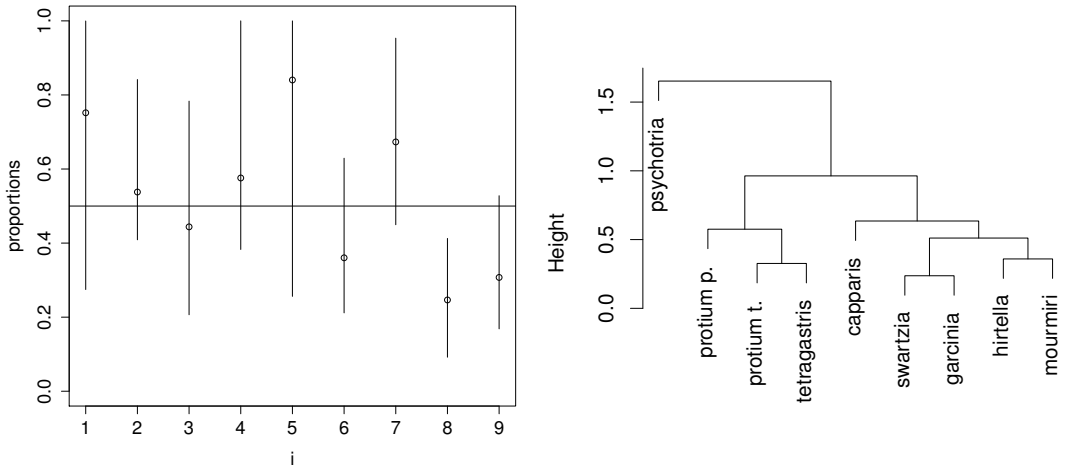

Figure 9: (BCI data) Left: estimated proportions of variance due to common latent factors. Index $i = 1, \ldots, 9$ corresponds to *Psychotria, Protium t., Capparis, Protium p., Swartzia, Hirtella, Tetragastris, Garcinia, Mouriri.* Right: hierarchical clustering based on fitted $\boldsymbol{\alpha}_i$.'s.

species *Protium t., Protium p.* and *Tetragastris* have fairly similar fitted proportions of variance and that the hierarchical clustering creates a cluster consisting of precisely these three species.

Regarding life form there is not a clear pattern in relation to the previous results as each of the categories trees (*Protium t., Protium p., Tetragastris, Garcinia*) and shrubs (*Psychotria, Capparis, Mouriri*) both display high and low proportions of variances and do not correspond to groups in the hierarchical clustering.

The spatial pattern of a species is influenced by the mode of seed dispersal (e.g. Muller-Landau and Hardesty, 2005; Seidler and Plotkin, 2006). The mode of seed dispersal for *Psychotria* is bird while it is bird/mammal or big bird/mammal for the remaining species. This could explain why *Psychotria* seems to be distinct from the other species both in terms of correlations and the results of the hierarchical clustering. Regarding the distinction between bird/mammal and big bird/mammal there does not seem to be a clear pattern in relation to the fitted proportions of variances, the fitted correlations or the hierarchical clustering.

7 Discussion

A basic problem with multivariate log Gaussian Cox processes is to model the cross covariance structure of the latent multivariate Gaussian field. Genton and Kleiber (2014) is a nice review of approaches to cross covariance modelling. In practice we need a flexible, interpretable and parsimonious model. The linear model of coregionalisation has some deficiencies in terms of flexibility. It e.g. enforces symmetric cross covariances $\mathbb{C}\mathrm{ov}[Y_i(u), Y_j(v)] = \mathbb{C}\mathrm{ov}[Y_i(v), Y_j(u)]$ but this seems a minor problem in the considered practical context of modelling point patterns of tree species. The model components have a reasonable interpretation as explained in the beginning of Section 2. Parsimony is sought by selection of a hopefully small q by cross-validation. In both practical examples this results in fairly parsimonious

models as measured by the number of parameters per unique cross pair correlation function. Another way to obtain parsimony would be to consider a lasso approach (Tibshirani, 1996). In this case one would fix q a priori and then use cross-validation to select a suitable L_1 regularisation which would typically result in a number of entries α_{il} being set to zero. The problem of identifying a suitable q however remains.

Regarding the biological implications of our work, the number of latent processes q selected by cross-validation gives an index of the complexity of the multivariate dependence structure. The plots of cross correlations in Section 6 provide compact presentations of the correlation structure of the species while fitted proportions of variances quantify to which extent the spatially structured random variation of the species is governed by common or species specific factors. As shown for the BCI data, there is further scope for linking proportions of variances and results of hierarchical clustering to families of species and functional traits such as life forms, reproductive strategies and growth/mortality patterns.

The hierarchical clustering results did not come with a measure of uncertainty. Inspired by Kerr and Churchill (2001) one may study the stability of the clustering by applying the hierarchical clustering to parametric bootstrap simulations from the fitted models. This is considered in Section 5 of the supplementary material. It appears that the clustering results are very stable for the abundant BCI species but less so for the Lansing data where the species are less abundant.

Acknowledgments

Abdollah Jalilian and Rasmus Waagepetersen's research was supported by the Danish Natural Science Research Council, grant 09-072331 'Point process modelling and statistical inference', Danish Council for Independent Research | Natural Sciences, Grant 12-124675, 'Mathematical and Statistical Analysis of Spatial Data', and by Centre for Stochastic Geometry and Advanced Bioimaging, funded by a grant from the Villum Foundation. Yongtao Guan's research was supported by NSF grant DMS-0845368, by NIH grant NIH grant 1R01CA169043 and by the VELUX Visiting Professor Programme. Jorge Mateu's research was supported by grants P1-1B2012-52 and MTM2013-43917-P.

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 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 (`http://spatial-analyst.net/`).

We thank Dr. Joseph Wright for sharing data on dispersal modes and life forms for the BCI tree species.

References

Allard, D., A. Brix, and J. Chadoeuf (2001). Testing local independence between two point processes. *Biometrics 57*(2), pp. 508–517.

Baddeley, A. J. (2010). Multivariate and marked point processes. In A. Gelfand, P. Diggle, P. Guttorp, and M. Fuentes (Eds.), *Handbook of Spatial Statistics*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, pp. 371–402. Taylor & Francis.

Baddeley, A. J., J. Møller, and R. Waagepetersen (2000). Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica 54*, 329–350.

Belouchrani, A., K. Abed-Meraim, J. F. Cardoso, and E. Moulines (1997). A blind source separation technique using second-order statistics. *IEEE Transactions on Signal Processing 45*(2), 434–444.

Brix, A. and J. Møller (2001). Space-time multi type log Gaussian Cox processes with a view to modelling weeds. *Scandinavian Journal of Statistics 28*(3), 471–488.

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

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

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

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap methods and their application*. Cambridge University Press.

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

Diggle, P. J. (2003). *Statistical analysis of spatial point patterns* (second ed.). London: Arnold.

Diggle, P. J. and R. K. Milne (1983). Bivariate Cox processes: some models for bivariate spatial point patterns. *Journal of the Royal Statistical Society, Series B 45*, 11–21.

Funwi-Gabga, N. and J. Mateu (2012). Understanding the nesting spatial behaviour of gorillas in the Kagwene sanctuary, Cameroon. *Stochastic Environmental Research and Risk Assessment 26*(6), 793–811.

Gelfand, A., A. Schmidt, S. Banerjee, and C. Sirmans (2004). Nonstationary multivariate process modeling through spatially varying coregionalization. *Test 13*, 263–312.

Genton, M. and W. Kleiber (2014). Cross-covariance functions for multivariate geostatistics. *Statistical Science*. To appear.

Gerrard, D. J. (1969). Competition quotient: a new measure of the competition affecting individual forest trees. Research Bulletin 20, Agricultural Experiment Station, Michigan State University.

Harkness, R. D. and V. Isham (1983). A bivariate spatial point pattern of ants' nests. *Journal of the Royal Statistical Society, Series C 32*(3), 293–303.

Hastie, T., R. Tibshirani, and J. Friedman (2013). *The Elements of Statistical Learning* (2 ed.). Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Heinrich, L. and E. Liebscher (1997). Strong convergence of kernel estimators for product densities of absolutely regular point processes. *Journal of Nonparametrical Statistics 8*, 65–96.

Högmander, H. and A. Särkkä (1999). Multitype spatial point patterns with hierarchical interactions. *Biometrics 55*, 1051–1058.

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

Illian, J., J. Møller, and R. Waagepetersen (2009). Hierarchical spatial point process analysis for a plant community with high biodiversity. *Environmental and Ecological Statistics 16*(3), 389–405.

Jalilian, A., Y. Guan, and R. Waagepetersen (2013). Decomposition of variance for spatial Cox processes. *Scandinavian Journal of Statistics 40*, 119–137.

Kerr, M. K. and G. A. Churchill (2001). Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Sciences 98*(16), 8961–8965.

Law, R., J. Illian, D. F. R. P. Burslem, G. Gratzer, C. V. S. Gunatilleke, and I. A. U. N. Gunatilleke (2009). Ecological information from spatial patterns of plants: insight from point process theory. *Journal of Ecology 97*, 616–628.

Liang, S., B. P. Carlin, and A. E. Gelfand (2009). Analysis of Minnesota colon and rectum cancer point patterns with spatial and nonspatial covariate information. *Annals of Applied Statistics 3*, 943–962.

Lin, Y.-C., L.-W. Chang, K.-C. Yang, H.-H. Wang, and I.-F. Sun (2011). Point patterns of tree distribution determined by habitat heterogeneity and dispersal limitation. *Oecologia 165*, 175–184.

Liu, D., M. Kelly, P. Gong, and Q. Guo (2007). Characterizing spatial-temporal tree mortality patterns associated with a new forest disease. *Forest Ecology and Management 253*, 220–231.

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

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

Muller-Landau, H. C. and B. D. Hardesty (2005). Seed dispersal of woody plants in tropical forests: concepts, examples, and future directions. In D. Burslem, M. Pinard, and S. Hartley (Eds.), *Biotic Interactions in the Tropics*, Chapter 11, pp. 267–309. Cambridge: Cambridge University Press.

Perry, G. L. W., B. P. Miller, and N. J. Enright (2006). A comparison of methods for the statistical analysis of spatial point patterns in plant ecology. *Plant Ecology 187*, 59–82.

Picard, N., A. Bar-Hen, F. Mortier, and J. Chadoeuf (2009). The multi-scale marked area-interaction point process: A model for the spatial pattern of trees. *Scandinavian Journal of Statistics 36*, 23–41.

Seidler, T. G. and J. B. Plotkin (2006). Seed dispersal and spatial pattern in tropical trees. *PLoS Biology 4*, 2132–2137.

Shen, G., M. Yu, X.-S. Hu, X. Mi, H. Ren, I.-F. Sun, and K. Ma (2009). Species-area relationships explained by the joint effects of dispersal limitation and habitat heterogeneity. *Ecology 90*, 3033–3041.

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

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

Waagepetersen, R. and Y. Guan (2009). Two-step estimation for inhomogeneous spatial point processes. *Journal of the Royal Statistical Society, Series B 71*, 685–702.

Wackernagel, H. (2003). *Multivariate geostatistics*. Springer.

Wiegand, T., S. Gunatilleke, N. Gunatilleke, and T. Okuda (2007). Analyzing the spatial structure of a Sri Lankan tree species with multiple scales of clustering. *Ecology 88*, 3088–3102.

Wright, S. J., A. Hernandéz, and R. Condit (2007). The bushmeat harvest alters seedling banks by favoring lianas, large seeds, and seeds dispersed by bats, birds, and wind. *Biotropica 39*, 363–371.

A Gradient and Hessian matrix for least squares object function

Let $\tilde{\mathbf{y}}$ denote the p^2L vector consisting of the $\tilde{y}_{ijk} = \sqrt{w_{ijk}}y_{ijk}$ concatenated in lexicographic order $(\tilde{y}_{111}, \tilde{y}_{112}, \ldots)$ and let $\underline{R}$ be the $p^2L \times p^2q$ block matrix with diagonal $L \times q$ blocks $\mathrm{Diag}[\sqrt{w_{ijk}}, k = 1, \ldots, L]R(\boldsymbol{\phi})$ and zeros outside the diagonal blocks. Let further $\underline{R}_U(\boldsymbol{\psi})$ denote the $p^2L \times p$ matrix with (iik, i)th entry $\exp(-t_k/\psi_i)$, $i = 1, \ldots, p$, $k = 1, \ldots, L$, and zeros elsewhere. In the following we derive derivatives and second derivatives of the object function (3.2) with respect to $\boldsymbol{\alpha}$.

By the multivariate chain rule, the derivative of Q with respect to $\boldsymbol{\alpha}$ is

$$g = \frac{\mathrm{d}Q}{\mathrm{d}\boldsymbol{\alpha}} = \frac{\mathrm{d}\boldsymbol{\beta}^\mathsf{T}}{\mathrm{d}\boldsymbol{\alpha}}\frac{\mathrm{d}Q}{\mathrm{d}\boldsymbol{\beta}} = -2\frac{\mathrm{d}\boldsymbol{\beta}^\mathsf{T}}{\mathrm{d}\boldsymbol{\alpha}}\underline{R}^\mathsf{T}(\tilde{\mathbf{y}} - \underline{R}\boldsymbol{\beta} - \underline{R}_U(\boldsymbol{\psi})\boldsymbol{\sigma}^2)$$

and the Hessian matrix is

$$H = \frac{\mathrm{d}^2Q}{\mathrm{d}\boldsymbol{\alpha}^\mathsf{T}\mathrm{d}\boldsymbol{\alpha}} = 2\frac{\mathrm{d}\boldsymbol{\beta}^\mathsf{T}}{\mathrm{d}\boldsymbol{\alpha}}\underline{R}^\mathsf{T}\underline{R}\frac{\mathrm{d}\boldsymbol{\beta}}{\mathrm{d}\boldsymbol{\alpha}} + \left[\frac{\partial\boldsymbol{\beta}^\mathsf{T}}{\partial\alpha_{k_1l_1}\partial\alpha_{k_2l_2}}\frac{\mathrm{d}Q}{\mathrm{d}\boldsymbol{\beta}}\right]_{k_1l_1,k_2l_2} \tag{A.1}$$

The $pq \times p^2q$ matrix $\mathrm{d}\boldsymbol{\beta}^\mathsf{T}/\mathrm{d}\boldsymbol{\alpha}$ has entries

$$\frac{\partial\beta_{jlk}}{\partial\alpha_{ik'}} = \begin{cases} 2\alpha_{ik} & i = j = l,\ k = k' \\ \alpha_{lk} & i = j,\ i \neq l,\ k = k' \\ \alpha_{jl} & i = l,\ i \neq j,\ k = k' \\ 0 & \text{otherwise} \end{cases}$$

and the vector $\partial\boldsymbol{\beta}^\mathsf{T}/(\partial\alpha_{i_1k_1}\partial\alpha_{i_2k_2})$ has entries

$$\frac{\partial\beta_{jlk}}{\partial\alpha_{i_1k_1}\partial\alpha_{i_2k_2}} = \begin{cases} 2 & i_1 = i_2 = j = l,\ k = k_1 = k_2 \\ 1 & i_1 = j,\ i_2 = l,\ k = k_1 = k_2 \\ 1 & i_1 = l,\ i_2 = j,\ k = k_1 = k_2 \\ 0 & \text{otherwise.} \end{cases}$$

The remaining derivatives of Q are given in Section 2 of the supplementary material.

B Simulation study

To asses the least squares method for parameter estimation and the cross-validation method for choosing q we conducted a simulation study on the unit square with $p = 5$ and q either zero or two. Regarding parameter estimation, we focused on the estimation of the proportions of variance $PV_i(0)$ and the off-diagonal correlations $\mathbb{C}\mathrm{orr}[Y_i(\mathbf{u}), Y_j(\mathbf{u})]$ at lag 0. For both $q = 0$ and $q = 2$ we considered $\boldsymbol{\sigma}^2 = (1, 1, 1, 1, 1)$, $\boldsymbol{\psi} = (0.01, 0.02, 0.02, 0.03, 0.04)$. In the case $q = 2$,

$$\alpha^\mathsf{T} = \begin{bmatrix} \sqrt{1/2} & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & -0.5 \end{bmatrix}$$

and $\boldsymbol{\phi} = (0.02, 0.1)$. This produced off-diagonal correlations $\mathbb{C}\mathrm{orr}[Y_i(\mathbf{u}), Y_j(\mathbf{u})]$ ranging between -0.41 and 0.41 and proportions of variances between $1/5$ and $2/3$. For the trend models we used $\mu_i(\mathbf{u}) = m_i$ where m_i was adjusted for each $i = 1, \ldots, 5$ to produce an expected number of 1000 points. For the least squares estimation a uniform kernel with bandwidth 0.005 was used for the non-parametric estimation of the cross pair correlation functions at 100 equispaced lags between 0.025 and 0.25.

We first considered parameter estimation in the case $q = 2$ using the least squares method assuming q known and equal to the true value. Tables 2 and 3 show quantiles of estimates of the off-diagonal correlations $\mathbb{C}\mathrm{orr}[Y_i(\mathbf{u}), Y_j(\mathbf{u})]$ and the proportions

Table 2: Quantiles of estimates of off-diagonal correlations $\mathbb{C}\mathrm{orr}[Y_i(\mathbf{u}), Y_j(\mathbf{u})]$ for known $q = 2$.

2.5 %	0.93	−0.91	−0.36	−0.58	−0.92	−0.36	−0.61	−0.98	0.15	−1
true	1	−0.71	0	0	−0.71	0	0	−0.71	0.71	−1
50 %	0.99	−0.7	0.02	0.06	−0.7	0.01	0.07	−0.72	0.67	−0.97
97.5 %	1	−0.4	0.57	0.58	−0.41	0.58	0.56	−0.39	0.98	−0.67

Table 3: Quantiles of estimates of proportions of variances $CV_i(0)$ for known $q = 2$.

2.5 %	0.11	0.17	0.34	0.18	0.06
true	0.33	0.5	0.67	0.5	0.2
50 %	0.29	0.47	0.69	0.62	0.24
97.5 %	0.61	0.95	1	1	0.64

of variances $PV_i(0)$ obtained from 1000 simulations of the multivariate model. In general there is good agreement between the true values and the medians of the estimates.

We next applied the cross-validation method to 200 simulations of the multivariate model both with $q = 2$ and $q = 0$ and using block lengths 10, 20 or 50, see Section 4. For each simulation and block size we identified the value q_s of q with the smallest cross-validation score. Table 4 shows for $q = 2$ and $q = 0$ the empirical distributions of the differences between q_s and the true q. Both for $q = 2$ and $q = 0$, the cross-validation method works best with $b = 50$. In the case $q = 2$ and $b = 50$, q_s coincides with the true q for 41 % of the simulations and differs at most by one from the true q in 67 % of the cases. For $q = 0$ and $b = 50$ the corresponding percentages are 0.76 and 0.90.

We also considered the distribution of the estimates of the off-diagonal $\mathbb{C}\mathrm{orr}[Y_i(\mathbf{u}), Y_j(\mathbf{u})]$ and the $PV_i(0)$ in the case of unknown $q = 2$. For each simulation the least squares method was applied with the selected q_s using $b = 50$ which as shown in Table 4 sometimes differs markedly from the true q. Quantiles of the simulated parameter estimates are shown in Tables 5 and 6. For most parameters there is reasonable agreement between medians of estimates and true values but the estimates are more variable than for the case of known q.

Table 4: Distributions of differences between selected $q = q_s$ and true q.

q	b	−2	−1	0	1	2	3
2	10	0	0	0.24	0.22	0.23	0.31
2	20	0	0.01	0.36	0.24	0.23	0.16
2	50	0.04	0.04	0.41	0.22	0.16	0.13
q	b	0	1	2	3	4	5
0	10	0.32	0.24	0.10	0.12	0.08	0.14
0	20	0.69	0.14	0.06	0.05	0.04	0.01
0	50	0.76	0.14	0.05	0.02	0.03	0

Table 5: Quantiles of estimates of off-diagonal correlations $\mathbb{C}\mathrm{orr}[Y_i(\mathbf{u}), Y_j(\mathbf{u})]$ in the case of unknown $q = 2$.

2.5 %	0.46	−1	−0.61	−1	−1	−0.52	−1	−1	−0.19	−1
50 %	0.98	−0.66	−0.02	0.01	−0.65	−0.02	0.01	−0.65	0.58	−0.87
True	1	−0.71	0	0	−0.71	0	0	−0.71	0.71	−1
97.5 %	1	−0.19	1	0.61	0.22	1	0.61	−0.22	1	−0.01

Table 6: Quantiles of estimates of proportions of variances $PV_i(0)$ in the case of unknown $q = 2$.

2.5 %	0	0	0	0	0
true	0.33	0.5	0.67	0.5	0.2
50 %	0.32	0.51	0.77	0.72	0.32
97.5 %	0.96	1	1	1	1