Reconstructing ecological networks with hierarchical Bayesian regression and Mondrian processes

Andrej Aderhold
School of Biology,
University of St Andrews
St Andrews KY16 9TH, UK
aa796@st-andrews.ac.uk

Dirk Husmeier
School of Mathematics and Statistics,
University of Glasgow
Glasgow G12 8QW, UK
dirk.husmeier@glasgow.ac.uk

V. Anne Smith
School of Biology,
University of St Andrews
St Andrews KY16 9TH, UK
vas1@st-andrews.ac.uk

Abstract

Ecological systems consist of complex sets of interactions among species and their environment, the understanding of which has implications for predicting environmental response to perturbations such as invading species and climate change. However, the revelation of these interactions is not straightforward, nor are the interactions necessarily stable across space. Machine learning can enable the recovery of such complex, spatially varying interactions from relatively easily obtained species abundance data. Here, we describe a novel Bayesian regression and Mondrian process model (BRAMP) for reconstructing species interaction networks from observed field data. BRAMP enables robust inference of species interactions considering autocorrelation in species abundances and allowing for variation in the interactions across space. We evaluate the model on spatially explicit simulated data, produced using a trophic niche model combined with stochastic population dynamics. We compare the model's performance against L1-penalized sparse regression (LASSO) and non-linear Bayesian networks with the BDe scoring scheme. Finally, we apply BRAMP to real ecological data.

1 INTRODUCTION

Recent endeavours in systems biology aiming to elucidate the structure of complex interaction networks have sparked off a series of novel applications and methodological innovations in machine learning and computational statistics. This has become most evident in the field of molecular systems biology, where following up on the seminal paper by Friedman et al. (2000), which aimed to infer gene interaction networks from mRNA abundance profiles with static Bayesian networks, a large variety of more advanced methods have been developed. This includes, for instance, approximate Bayesian inference for pathway ranking (Vyshemirsky & Girolami, 2008), Gaussian process models for transcriptional regulation (Honkela et al., 2010), and non-stationary dynamic Bayesian networks for inferring time-varying gene interactions (Lèbre et al., 2010). The latter work in particular has motivated new machine learning research, related to the combination of dynamic Bayesian networks and multiple changepoint processes (Robinson & Hartemink, 2010; Grzegorczyk & Husmeier, 2012).

While interaction networks at the molecular level have been at the forefront of modern biology, due to the ever increasing amount of available postgenomic data, interaction networks at other scales are drawing growing attention from the global research community. This concerns, in particular, ecological networks, owing to their connection with climate change and biodiversity, which poses new challenges and opportunities for machine learning and computational statistics.

Ecosystems are complex dynamic systems, with interconnected networks of interactions among species and abiotic characteristics. This interconnectedness can lead to seemingly unpredictable behaviour: changing numbers of one species can influence unexpected others (Henneman & Memmott, 2001); the whole system can shift between dynamical states (Beisner, Haydon & Cuddington, 2003). Yet being able to predict such phenomena is of growing importance in the modern world, where perturbations from features such as climate change and invasive species can affect both natural biodiversity and human agriculture (Foley et al., 2005). Such prediction requires understanding the ecological networks underlying the system (Dunne,

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31. Copyright 2013 by the authors.

Williams & Martinez, 2002).

Unravelling these networks strains the limits of typical ecological studies, requiring intensive observation to determine trophic interactions (predator-prey interactions) in even simple ecosystems, e.g. in Memmott et al. (2000). And trophic interactions are not the whole story, as harder-to-observe interactions such as competition and mutualism (species interacting in a way that both partners benefit) also influence ecosystem dynamics (Werner & Peacor, 2003). Measures of such indirect interactions have been attempted (van Veen, Brandon & Godfray, 2009), but computational inference presents an alternative, and perhaps more comprehensive, route to revealing both direct and indirect interactions within ecosystems.

Ecosystem interactions will leave traces in species distribution across space, a measure relatively easily obtained and currently available for many ecosystems, e.g. by Hagemeijer & Blair (1997). Computational algorithms can make use of such observational data, as has also been done in other areas of biology (e.g. gene expression for transcriptional regulation (Friedman et al., 2000), neural activity for information flow in the brain (Smith et al., 2006)), to reverse engineer the ecological interactions which produced them (Milns et al., 2010; Faisal et al., 2010). Furthermore, as the algorithms recover interactions based on their influence on species distribution, they are not limited to any one particular type of interaction (e.g., trophic, competition), and instead are capable of revealing interactions of all types simultaneously.

The challenges for computational inference specific to ecological systems are that, first, the interactions take place in a spatially explicit environment which must be taken into account, and second, the interactions can vary across this environment depending on the make-up of the elements (species and abiotic factors) present. Here, we meet these challenges by modifying the method from Lèbre et al. (2010) for temporally explicit (1-dimensional) gene expression data to infer ecological interactions from spatially explicit species abundance data on a 2-dimensional grid. The authors describe a non-homogeneous dynamic Bayesian network based on the Bayesian hierarchical regression model of Andrieu & Doucet (1999), using a multiple global change-point process. We replace the latter by a Mondrian process following Roy & Teh (2008), allowing a more precise partitioning of 2-dimensional space. We make further use of the spatial explicit nature of ecological data by correcting for spatial auto-correlation with a parent node (in Bayesian network terminology) that explicitly represents the spatial neighbourhood of a node.

We evaluate our model's performance on data generated from a realistic simulation, which combines a trophic niche model of Lotka-Volterra type predator-prey interactions with a stochastic population model on a 2-dimensional lattice. We compare the model's performance on this simulated data with both L1-penalized sparse regression (LASSO) and non-linear Bayesian networks (BDe score). We then apply our model to species counts of ground cover flora and associated abiotic variables from a strip of land across an environmental gradient on the western shore of the Outer Hebrides, to assess our model's applicability and utility for real ecological data.

2 MODEL

2.1 Overview

This section describes briefly our modelling approach, which combines the Bayesian hierarchical regression model of Andrieu & Doucet (1999) and Punskaya et al. (2002) with a spatial Mondrian process partitioning model (Roy & Teh, 2008; Wang et al., 2011) and pursues Bayesian inference with reversible jump Markov chain Monte Carlo (RJMCMC) (Green, 1995). The value that the kth node in the graph takes on at a given location represents the abundance of the kth species in the population. This abundance is determined by various biotic and abiotic determinants, i.e. factors that influence the abundance of species k. Abiotic factors are related to the environment and include e.g. temperature, humidity, soil type etc. Biotic factors represent the abundance of other species. Their influences are indicative of how species interact, which is the primary interest of the present work. The strengths of these influences are allowed to vary geographically, based on a stochastic process of spatial variation. More specifically, the conditional probability of a species abundance at a given location is a conditional Gaussian distribution, where the conditional mean is a linear weighted sum of the abundance levels of the biotic and abiotic determinants. The weight parameters can vary between different segments of a spatial partition, which adds extra flexibility to the model and allows for unobserved or latent factors. The interaction weights, the variance parameters, and the number of potential determinants are given (conjugate) prior distributions in a hierarchical Bayesian model, and the spatial partition is modelled non-parametrically with a Mondrian process prior. For inference, all quantities are sampled from the posterior distribution with RJMCMC. Note that a complete specification of all species-determinant configurations determines the structure of a regulatory network: each node receives incoming directed edges from each node in its set of determinants (the so-called parent set).

2.2 Species Interaction Network

We represent the N interacting species as nodes $n \in \{1, \ldots, N\}$ in a directed graph or network $\mathcal{G} = \{\pi_1, \ldots, \pi_N\}$, where π_n denotes the parents of node n, that is the set of nodes with a directed edge pointing to n. $\mathcal{G}_n$ is the subnetwork associated with target species n, which is determined by its parent set π_n. A node cannot be contained in its own parent set, $n \notin \pi_n$, i.e. we rule out self-interactions related to e.g. cannibalism. The species are observed or surveyed at $T_1 \times T_2$ locations defined by their (orthogonal) coordinates (x_1, x_2), at which their abundance levels $y = \{y_n(x_1, x_2)\}_{1 \leq n \leq N, 1 \leq x_1 \leq T_1, 1 \leq x_2 \leq T_2}\}$ are determined.

2.3 Nonparametric Spatial Partition with the Mondrian Process

Interactions among species are influenced by latent effects, which we assume to be similar in spatially adjacent locations, and we therefore introduce into our model a process of partitioning a 2-dimensional domain $\Theta_1 \times \Theta_2$ (longitude times latitude) inhabited by the species of interest. The Mondrian process, introduced by Roy & Teh (2008), is a generative recursive process for self-consistently partitioning the 2-dimensional domain in the following way. A hyperparameter λ (the so-called "budget") determines the average number of cuts in the partition. At each stage of the recursion, a Mondrian sample can either define a trivial partition $\Theta_1 \times \Theta_2$, i.e. a segment, or a cut that creates two subprocesses $m_<$ and $m_>$: $m = \langle d, \chi, \lambda', m_<, m_> \rangle$, where d is the horizontal or vertical direction and χ the position of the cut. The direction d and position χ are drawn from a binomial and uniform distribution, respectively, both depending on Θ_1 and Θ_2, as shown in line 5 of Algorithm 1. The process of cutting a segment is limited by the budget λ associated to each segment and the cost E of a cut. Conditional on halfperimeter $\tau = |\Theta_1| + |\Theta_2|$, a cut is introduced yielding $m_<$ and $m_>$ if the cost $E \sim \exp(\tau)$ does not exceed the budget λ, i.e. satisfies $\lambda' = \lambda - E > 0$. The process is recursively repeated on $m_<$ and $m_>$ until the budgets are exhausted, as shown in Algorithm 1. This creates a binary tree with the initial Mondrian sample $m_{k=1}$ as the root node spanning the unit square $[0; 1]^2$ and sub-nodes representing Mondrian samples $m_{1<k \leq K}$, $k \in \{1, \ldots, K\}$ where K is the total number of nodes in the tree, e.g. $K = 15$ in Fig. 1. The leaf nodes present non-overlapping segments and are associated each with a latent variable $h(k)$ labeled with $m^{h(k)}$ (Fig. 1). These latent variables determine the interactions among species, as described in Subsection 2.4. We denote by Z the number of uncut segments, e.g. $Z = 8$ in Fig. 1, and $h(k) \in \{1, \ldots, Z\}$.

The Mondrian process can be regarded as a 2-dimensional generalization of the Poisson process, and it has the same self-consistency property. We have chosen this approach over a global changepoint process in order to provide varying levels of fineness of the segments and thereby account for spatial alterations of the regulatory relationships among species on a local scale.

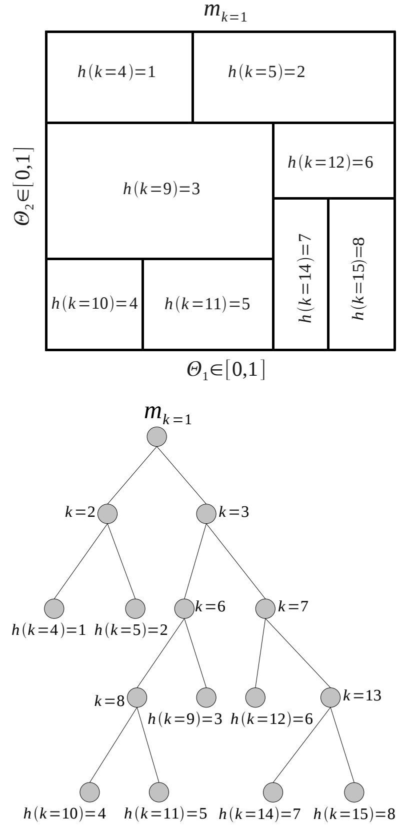

Figure 1: **Mondrian Process Example.** The upper panel shows an example partition with a Mondrian process. The lower panel dislays the associated tree with labels of the latent variable $h(k)$ identifying each non-overlapping segment in the partition.

2.4 Modelling Species Interactions with a Regression Model

For all species n, the random variable $Y_n(x_1, x_2)$ refers to the abundance of species n at location (x_1, x_2). Within any segment $h(k)$, this abundance depends on the abundance levels of the species in the parent set of species n, π_n, which we model with a segment specific linear regression model. Define the set of parameters $\{(a_{np}^{h(k)})_{p \in 0..N}, \sigma_n^{h(k)}\}$, $a_{np}^{h(k)} \in \mathbb{R}, \sigma_n^{h(k)} > 0$. For all $p \neq 0$, $a_{np}^{h(k)} = 0$ if $p \notin \pi_n$. For all species n, for all locations (x_1, x_2) in segment $h(k)$, $Y_n(x_1, x_2)$ depends on the N variables $\{Y_p(x_1, x_2)\}_{1 \leq p \leq N, p \neq n}$ according to

$$\begin{aligned} Y_n(x_1, x_2) \;=\; & a_{n0}^{h(k)} + \sum\nolimits_{p \in \pi_n} a_{np}^{h(k)}\, Y_p(x_1, x_2) \\ & + \varepsilon_n(x_1, x_2) \;+\; a_{nA}^{h(k)} A_n(x_1, x_2) \quad (1) \end{aligned}$$

$\varepsilon_n(x_1, x_2)$ is assumed to be white Gaussian noise with mean 0 and variance $(\sigma_n^{h(k)})^2$, $\varepsilon_n(x_1, x_2) \sim N(0, (\sigma_n^{h(k)})^2)$. We define $a_n^{h(k)} = (a_{np}^{h(k)})_{n \in 0..N}$ to denote the vector of all regression parameters of species n. This includes the parameters defining the strength of interactions with other species p, $a_{np}^{h(k)}$, as well as a species-specific offset term, $a_{n0}^{h(k)}$. Spatial autocorrelation effects are represented with $A_n(x_1, x_2)$ weighted by an additional edge $a_{nA}^{h(k)}$. They reflect the influence of neighboring cells that can have a strong effect on statistical inference (Lennon, 2000). $A_n(x_1, x_2)$ denotes the average densities in the vicinity of (x_1, x_2), weighted inversely proportional to the distance of the neighbors:

$$A_n(x_1, x_2) = \frac{\sum_{(\tilde{x}_1,\tilde{x}_2)\in\mathcal{N}(x_1,x_2)} d^{-1}[(x_1,x_2),(\tilde{x}_1,\tilde{x}_2)]Y_n(\tilde{x}_1,\tilde{x}_2)}{\sum_{(\tilde{x}_1,\tilde{x}_2)\in\mathcal{N}(x_1,x_2)} d^{-1}[(x_1,x_2),(\tilde{x}_1,\tilde{x}_2)]} \tag{2}$$

where $\mathcal{N}(x_1, x_2)$ is the spatial neighborhood of location (x_1, x_2) (e.g. the four nearest neighbors), and $d[(x_1, x_2), (\tilde{x}_1, \tilde{x}_2)]$ is the Euclidean distance between (x_1, x_2) and $(\tilde{x}_1, \tilde{x}_2)$.

2.5 Prior Distribution

Interaction network: To encourage sparse network structures, we impose a truncated Poisson prior with mean κ and maximum $\overline{p} = 5$ on the number p_n of parents for node n: $P(p_n|\kappa) \propto \frac{\kappa^{p_n}}{p_n!}\mathbb{1}_{\{p_n \leq \overline{p}\}}$. Conditional on p_n, the prior for the parent set π_n is a uniform distribution over all parent sets with cardinality p_n: $P(\pi_n \mid |\pi_n| = p_n) = 1/\binom{N-1}{p_n}$. The overall prior on the network structure $\mathcal{G}$ is given by factorization and marginalization in equation (3):

$$\begin{aligned} P(\mathcal{G}|\kappa) &= \prod\nolimits_{n=1}^{N} P(\pi_n|\kappa); \\ P(\pi_n|\kappa) &= \sum\nolimits_{p_n=1}^{\overline{p}} P(\pi_n|p_n)P(p_n|\kappa) \end{aligned} \tag{3}$$

Conditional on the parent set π_n of size p_n, the $p_n + 2$ regression coefficients, denoted by $a_n^{h(k)} = (a_{n0}^{h(k)}, a_{nA}^{h(k)}, (a_{np}^{h(k)})_{p\in\pi_n})$, are assumed zero-mean multivariate Gaussian distributed with covariance matrix $(\sigma_n^{h(k)})^2\Sigma_{n,h(k)}$,

$$P(a_n^{h(k)}|\pi_n, \sigma_n^{h(k)}, \delta^2) = |2\pi(\sigma_n^{h(k)})^2\Sigma_{n,h(k)}|^{-\frac{1}{2}}\exp\left(-\frac{[a_n^{h(k)}]^\dagger\Sigma_{n,h}^{-1}a_n^{h(k)}}{2(\sigma_n^{h(k)})^2}\right) \tag{4}$$

where the symbol † denotes matrix transposition, $\Sigma_{n,h(k)} = \delta^{-2}D^\dagger_{n,h(k)}(y)D_{n,h(k)}(y)$ and $D_{n,h(k)}(y)$ is the $s_{n,h(k)} = |\widehat{\Theta}_1^{h(k)}||\widehat{\Theta}_2^{h(k)}| \times (p_n + 2)$ matrix whose first column is a vector of 1s, for the constant in (1), the second column is a vector of autocorrelation variables, defined in (2), and the remaining columns contain the observed abundance values $y_n(x_1, x_2)$ for all species n and all locations (x_1, x_2) that map into segment $h(k)$. This so-called g-prior is widely used in Bayesian statistics; see e.g. by Andrieu & Doucet (1999). Finally, the conjugate prior for the variance $(\sigma_n^{h(k)})^2$ is the inverse gamma distribution, $P((\sigma_n^{h(k)})^2) = \mathcal{IG}(\upsilon_0, \gamma_0)$. Following Lèbre et al. (2010), we set the hyper-hyperparameters for shape, $\upsilon_0 = 0.5$, and scale, $\gamma_0 = 0.05$, to fixed values that give a non-informative prior distribution. The term κ can be interpreted as the expected number of parents and δ^2 is the expected signal-to-noise ratio. Following Lèbre et al. (2010), these hyperparameters are drawn from vague conjugate hyperpriors, which are in the (inverse) gamma distribution family: $P(\kappa) = \mathcal{G}a(0.5, 1)$ and $P(\delta^2) = \mathcal{IG}(2, 0.2)$.

Mondrian process: The prior disrtribution of the Mondrian process depends on the hyperparameter λ and is defined via the generative process described in Algorithm 1 and Section 2.3. However, for the RJMCMC scheme described below all that is needed is the prior ratio, which is given by (8).

Algorithm 1 MCMC Mondrian cut: Note, the Mondrian generative process corresponds to lines 1-4 and 7, i.e. the MCMC move extends it by considering the acceptance probability in lines 5-6.

```
 1: Input: $m$, $\lambda$
 2: $h(k) \leftarrow \mathcal{U}(1, Z)$        ▷ uniformly select uncut segment $h(k)$
 3: $\lambda^{'} \leftarrow \lambda - E$ with $E \sim \exp(|\Theta_1^{h(k)}| + |\Theta_2^{h(k)}|)$
 4: if $\lambda^{'} \geq 0$ then                              ▷ if budget sufficient
 5:  ▷ draw direction $d \in \{1, 2\}$, 1 is vertical and 2 is horizontal
 6:     $d \sim \mathcal{B}(|\Theta_1^{h(k)}|/(|\Theta_1^{h(k)}| + |\Theta_2^{h(k)}|))$
 7:     $\chi|d \sim \mathcal{U}(\Theta_d^{h(k)})$                        ▷ draw cut position $\chi$
 8:     $\alpha \leftarrow min\{1, r\}$      ▷ acceptance probability, equation 7
 9:     if $\alpha > u \sim \mathcal{U}(0, 1)$ then ▷ accept with subtrees $m_< \ m_>$
10:         $m^{h(k)} \leftarrow \langle d, \chi, \lambda^{'}, m_<, m_>\rangle$
11:     end if
12: end if
```

2.6 Likelihood and Marginal Likelihood

Equation (1) implies that the likelihood is given by

$$\mathcal{L}(y_n^{h(k)}|\mathcal{G}_n, a_n^{h(k)}, \sigma_n^{h(k)}) = \left(\sqrt{2\pi}\sigma_n^{h(k)}\right)^{-s_{n,h(k)}} \times \exp\left(-\frac{(y_n^{h(k)} - D_{n,h(k)}(y)a_n^{h(k)})^\dagger\ (y_n^{h(k)} - D_{n,h(k)}(y)a_n^{h(k)})}{2(\sigma_n^{h(k)})^2}\right)$$

An attractive feature of the chosen model is that the marginalization over the parameters $a = \{a_n^{h(k)}, 1 \leq n \leq N, 1 \leq h(k) \leq Z\}$ and $\sigma^2 = \{(\sigma_n^{h(k)})^2, 1 \leq n \leq N, 1 \leq h(k) \leq Z\}$ is analytically tractable (Lèbre et al., 2010; Andrieu & Doucet, 1999), and we obtain a closed-form expression for the marginal likelihood:

$$\mathcal{L}(y_n^{h(k)}|\mathcal{G}_n, \delta^2) = \int \mathcal{L}(y_n^{h(k)}|\mathcal{G}_n, a_n^{h(k)}, \sigma_n^{h(k)}) P([\sigma_n^{h(k)}]^2)P(a_n^{h(k)}|\pi_n, [\sigma_n^{h(k)}]^2, \delta^2)da_n^{h(k)}d[\sigma_n^{h(k)}]^2 \tag{5}$$

For space restrictions, the reader is referred to Lèbre et al. (2010) for an explicit expression.

2.7 Posterior Distribution and Bayesian Inference

The objective of Bayesian inference is to sample from the posterior distribution

$$P(m, \mathcal{G}, \kappa, \delta^2|y) \propto \mathcal{L}(y_n^{h(k)}|\mathcal{G}_n, \delta^2)P(\delta^2)P(E) P(\mathcal{G}|\kappa)P(\kappa)P(m|\lambda) \tag{6}$$

where $\mathcal{L}(y_n^{h(k)}|\mathcal{G}_n, \delta^2)$ is the marginal likelihood, from (5), and all prior distributions have been defined above. To this end, we pursue a Gibbs sampling like strategy, where we iteratively sample new hyperparameters from $P(\kappa,\delta^2|\mathcal{G}, m, y)$, a new network structure from $P(\mathcal{G}|\kappa,\delta^2, m, y)$, and a new Mondrian process partition of

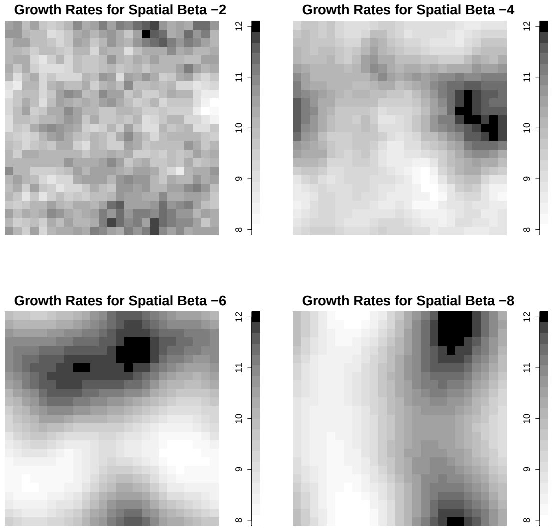

Figure 2: **Spatial Distribution.** The figure shows the spatial distribution of growth rates r_n entering equation (9) as the spatial β parameter, defined in Section 3.2, decreases from -2 to -8. A value of 0 would correspond to uniformly random noise, and -2 is Brownian noise.

the spatial domain from $P(m|\mathcal{G},y,\lambda)$. The first distribution is of standard form due to conjugacy of the prior, and the hyperparameters can be sampled directly. However, direct sampling from the other two distributions is intractable, and we therefore apply RJMCMC (Green, 1995). To sample new network structures $\mathcal{G}$, we follow the scheme described by Lèbre et al. (2010), which is based on edge birth and death moves. To sample from the Mondrian process, we adopt the method proposed by Wang et al. (2011), which we will briefly outline in the next subsection. The scheme could be extended to infer λ, but that has not been done yet, and we assume this hyperparameter to be fixed (and hence have not made it explicit on the left-hand side of equation (6)). We are primarily interested in a sample of network structures from the posterior distribution $P(\mathcal{G}|y)$, which we obtain by marginalizing over the hyperparameters and Mondrian process partition. By further marginalization, we get the posterior probabilities of all species interactions $P(n \to \tilde{n}|y)$, which defines a ranking of the interactions in terms of posterior confidence. If the true network structure is known, this ranking allows the computation of the areas under the ROC (AUROC) and precision-recall (AUPRC) curves (Davis & Goadrich, 2006), which are two measures widely used in the systems biology literature to quantify the overall network reconstruction accuracy (Prill et al., 2010), with larger values indicating a better prediction performance.

2.8 Mondrian Process RJMCMC

As described above, an essential step of the inference procedure is to sample a new Mondrian process segment m from $P(m|\mathcal{G},y,\lambda)$. Following Wang et al. (2011), and as described in Section 2.3, the current state of the Mondrian process m is represented by a structure tree and a model parameter vector $\vec{\zeta}$, which contains all previous costs E_k and cut locations χ_k. Note that all budgets and domains can be computed from that recursively. When a cut move is proposed (marked with $+$), the current parameter values are augmented by supplementary random variates u_1 and u_2 in such a way that the dimensions in the higher and lower dimensional parameter spaces are matched. We uniformly sample a spatial segment $h(k)$ draw u_1 and u_2 from the density $q(u_1,u_2)$ and set $\vec{\zeta} \to \vec{\zeta}^+ = \langle \vec{\zeta}, E^{h(k)} = u_1, \chi^{h(k)} = u_2 \rangle$. If $E^{h(k)}$ does not exceed the budget $\lambda^{h(k)}$, as described in Section 2.3, the cut move proceeds as shown in Algorithm 1, where $\chi^{h(k)}$ defines the position proportional to the sample domain size, which follows a Bernoulli distribution $\mathcal{B}$. The proposed new Mondrian process state m^+ is accepted with probability $\alpha = min\{1,r\}$,

$$r = \frac{P(m^+|\lambda)}{P(m|\lambda)} \times \frac{q(m|m^+)}{q(m^+|m)} \times \frac{\mathcal{L}(y_<^{h(k)}|\mathcal{G},\delta^2)\mathcal{L}(y_>^{h(k)}|\mathcal{G},\delta^2)}{\mathcal{L}(y^{h(k)}|\mathcal{G},\delta^2)} \times J \tag{7}$$

$$\frac{q(m|m^+)}{q(m^+|m)} = \frac{Z}{\phi(m^+)q(E^{h(k)},\chi^{h(k)})},$$
$$\frac{P(m^+|\lambda)}{P(m|\lambda)} = \frac{\omega_<^{h(k)}\omega_>^{h(k)}P(E^{h(k)})P(\chi^{h(k)})}{\omega^{h(k)}} \tag{8}$$

Here, the subscripts $>$ and $<$ refer to the two new spatial segments associated with the cut, $y_>^{h(k)}$ and $y_<^{h(k)}$ are the corresponding subsets of $y^{h(k)}$, and $y^{h(k)}$ denotes the species abundance data associated with segment/leaf node $h(k)$. Following the standard RJMCMC scheme (Green, 1995), the four terms in (7) are the prior ratio, inverse proposal ratio, marginal likelihood ratio and Jacobian. The latter is one, $J = 1$, the marginal likelihood is given by (5), and the prior and proposal ratios are given by (8), where ϕ denotes the number of Mondrian leaf siblings, i.e. adjacent segments that can be merged in order to restore m, and $\omega^{h(k)} = \int_\lambda^\infty \tau^{h(k)} \exp(-\tau^{h(k)}e)de = \exp(-\tau^{h(k)}\lambda^{h(k)})$ denotes the probability of no further cut. By setting $q(E_k,\chi_k) = P(E_k)P(\chi_k)$, the expression naturally simplifies. The state m is replaced by the proposal m^+ in the case the move is accepted. The probability of removing a cut is given by the inverse of (7). A shift move replaces the direction d and position χ of a cut, which separates the adjacent segments $h(k_1)$ and $h(k_2)$ yielding the proposal segments $h(k_1)^+$ and $h(k_2)^+$. The acceptance probability is $\alpha = min\{1, \mathcal{L}(y^{h(k_1)^+})\mathcal{L}(y^{h(k_2)^+})/\mathcal{L}(y^{h(k_1)})\mathcal{L}(y^{h(k_2)})\}$ after canceling the proposal and prior ratios because budget, cost and number of Mondrian samples remain invariant. Whenever a segment is cut or merged, the affected regression coefficients are sampled from the posterior. A more detailed discussion of inference in Mondrian processes with RJMCMC can be found in Wang et al. (2011).

3 DATA

3.1 Synthetic Data

For an objective measure of network recovery, we tested the model's ability to recover the true network structure from test data generated from a piecewise linear regression model following equation (1). The data grid was partitioned according to a single Mondrian process generated from Algorithm 1 and the number of grid cells was selected to be 15 in each direction. The number of nodes n in the

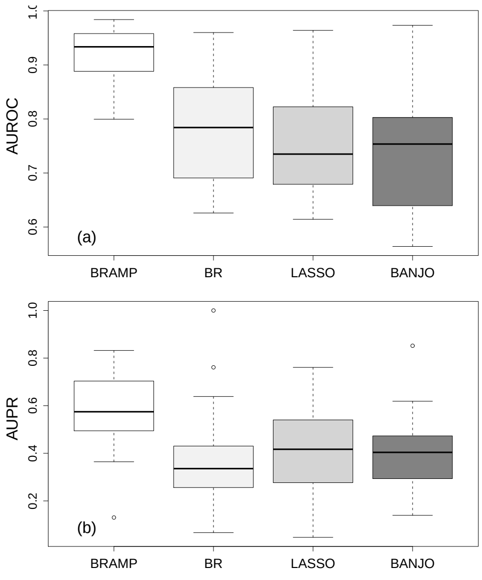

Figure 3: **Comparison on Synthetic Data.** Boxplots of AUROC (upper panel) and AUPRC (lower panel) scores obtained with three methods on the synthetic data described in Section 3.1: the proposed model (BRAMP), a Bayesian linear regression model without changepoints (BR), L1-penalized sparse regression (LASSO), and a homogeneous Bayesian network with the BDe score (BANJO). Each boxplot shows the distribution of scores of 30 independent data sets.

network $\mathcal{G}$ was set to 10 and the number of parents for each node was sampled from a Poisson distribution. The regression coefficients $a_n^{h(k)}$ together with the bias $a_0^{h(k)}$ of each segment $h(k)$ were sampled from a uniform distribution in the interval of $[-1; -0.5]$ and $[0.5, 1.0]$. The noise ε_n was sampled from a normal distribution. Nodes without incoming edge were initialized to a Gaussian random number with a variance of 1. The values of the remaining nodes were calculated at each grid cell following equation (1).

3.2 Realistic Simulation of Trophic Interactions

For a realistic evaluation, we followed Faisal et al. (2010) and generated data from an ecological simulation that combines a niche model (Williams & Martinez, 2000) with a stochastic population model (Lande, Engen & Saether, 2003) in a 2-dimensional lattice.

Niche model and species interactions. The niche model defines the structure of the trophic network and has two parameters: the number of species N and the connectance (or network density) defined as L/N^2 where L is the number of interactions (edges) in the network. Each species n is assigned a niche value x_n, drawn uniformly from $[0, 1]$. This gives an ordering of the species, where higher values mean that species are higher up in the food chain. For each species a niche range R_n is drawn from a beta distribution with expected value $2C$ (where C is the desired connectance), and species n consumes all species falling in a range R_n that is placed by uniformly drawing the centre of the range from $[R_n/2, x_n]$. An illustration is given in Figure 1 of Williams & Martinez (2000). Despite its simplicity, it was shown there that the resulting networks share many characteristics with real food webs.

Stochastic population dynamics. The population model is defined by a stochastic differential equation where the dynamics of the log abundance $X_n(t)$ of species n at time t can be expressed as:

$$\begin{aligned}\frac{dX_n(t)}{dt} &= r_n + \frac{\sigma_d}{\sqrt{e^{X_n(t)}}}\frac{dA_n(t)}{dt} + \sigma_e\frac{dB_n(t)}{dt} - \\ &\qquad \gamma X_n(t) - \Omega(X) + \sigma_E\frac{dE(t)}{dt} \qquad (9)\end{aligned}$$

where X is the set of all $X_N(t)$, r_n is the growth rate of species n, σ_d is the standard deviation of the demographic effect, $A_n(t)$ is the species-specific demographic effect, σ_e is the standard deviation of the species-specific environmental effect, $B_n(t)$ is the species-specific environmental effect, γ is the intra-specific density dependence, Ω is the effect of competition for common resources, σ_E is the standard deviation of the general environmental effect and $E(t)$ is the general community environment. The growth rates r_n are location dependent (depending on the cell of a rectangular grid), with a spatial pattern that is generated by noise with spectral density f^β (with $\beta < 0$, and f denoting the spatial frequency at which the noise is measured). An illustration is given in Figure 2. To model species migration, we included an exponential dispersal model, where the probability of a species moving from one location to another is determined by the Euclidean distance between the locations. To incorporate the niche model, we modified the term Ω in (9) to include predator-prey interactions in the Lotka-Volterra form. A detailed description is available by Faisal et al. (2010).

Simulation. We applied this model to 10 species living in a 25-by-25 rectangular grid. We simulated the dynamics of this model for 3000 steps and then recorded species abundance levels in all grid cells at the final step; this corresponds to an ecological survey carried out at a fixed moment in time. For each grid cell we counted the number of species that went extinct. These counts were added up over all cells, yielding a total number of extinctions. A simulation was rejected if these extinctions exceeded the value 50. For each of the spatial β parameters displayed in Figure 4, 30 surveys were collected by running the simulation repeatedly with different networks and parameter initializations.

3.3 Real World Plant Data

We have applied the method to real-world data from Lennon et al. (2011), including 106 vascular plants and 12 physical variables collected from a 200m x 2162m land stripe at the western shore of the Outer Hebrides. Samples were taken at 217 locations, each 1m x 1m in size, equally distributed with a 50m spacing. Plant samples were measured as ground coverage in percentage and physical samples as absolute values (such as moisture, pH value, organic matter and slope). The data was log-normal transformed after observing substantial skewness in the distributions. Each sample point was mapped into a 2D grid ignoring locations with no sample data available. The single spatial autocorrelation value for each plant and location was calculated with equation 2 using neighbors inside a radius of 70m. Since we are interested only in plant interactions, we defined each plant to have all 12 physical soil variables as fixed input, i.e., permanent predictor variables.

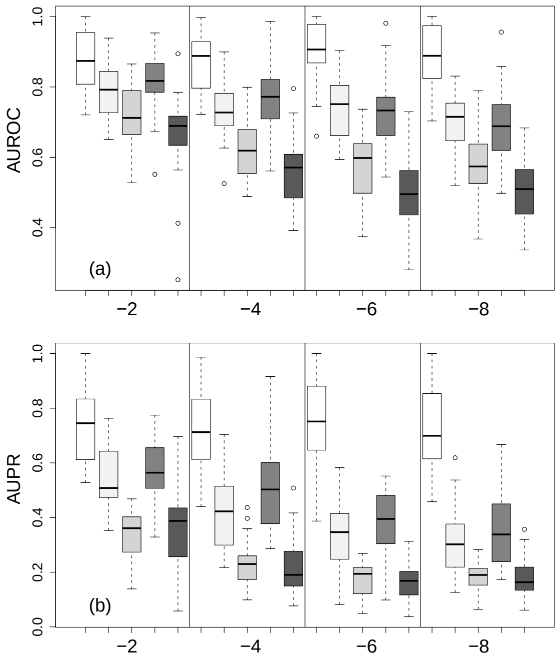

Figure 4: **Comparative Evaluation of Five Network Reconstruction Methods for the Parasitism Data.** AUROC (upper panel) and AUPRC (lower panel) scores obtained on the realistic simulated data described in Section 3.2. Box color scheme: BRAMP (white), BR (light gray), BR-0 (gray), LASSO (dark gray), Banjo (darkest gray).

4 COMPARATIVE EVALUATION

We followed Wang et al. (2011) and set the hyperparameter of the Mondrian process to the fixed value $\lambda = 1$ for all our simulations. We compared the performance of BRAMP with two alternative Bayesian regression models: a Bayesian regression without changepoints (BR) allowing for spatial autocorrelation and Bayesian regression without changepoints and without allowing for spatial autocorrelation (BR-0). We included a comparison with L1-penalized sparse regression (LASSO: Tibshirani (1996)), using the optimization algorithm proposed by Grandvalet (1998). This method is widely applied in molecular systems biology van Someren et al. (2006), has been recommended to be used more widely in ecology (Dahlgren, 2010), and was found to outperform all competing methods in Faisal et al. (2010). The L1 regularization parameter, which controls the network sparsity, was inferred with 10-fold cross-validation, which led to better results than optimizing the BIC score. The method produces edge weights indicating the strength and sign of interactions among species. For obtaining the ROC and precision-recall curves, we ranked the potential interactions based on the absolute values of the non-zero interaction parameters. We further included a comparison with a non-linear Bayesian network, as implemented in the software package BANJO. We discretized the data with Hartemink's pairwise mutual information method (Hartemink, 2001) (implemented in *R* package *bnlearn*)[1]. Search was done using simulated annealing with random walk proposals. Simulated annealing was run on each dataset until convergence. Using the top 100 high-scoring (BDe score) networks we computed edge probabilities for ranking. Application of both LASSO and BANJO included taking spatial autocorrelation into account. Finally, we applied BRAMP to real world data, revealing putative plant interactions.

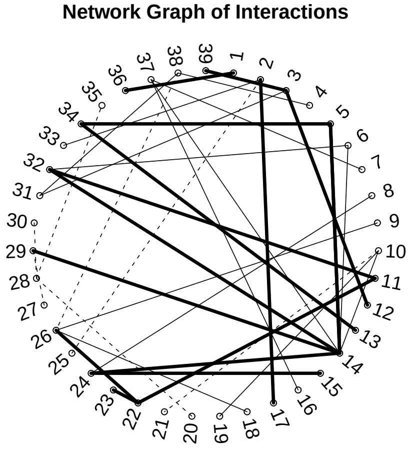

Figure 5: **Species Interaction Network.** Species interactions as inferred with BRAMP, described in Section 3.3, with an inferred marginal posterior probability of 0.5 (thick lines) and 0.1 (thin lines). Solid lines are positive (e.g. mutualism, facilitation) and dashed are negative interactions (e.g. resource competition). Species are represented by numbers and have been ordered phylogenetically (see Supplementary Material).

5 RESULTS

On the synthetic data of Section 3.1, BRAMP outperforms all competing schemes (Figure 3). This is not surprising, in that the data have been generated from a process that is consistent with the modeling assumptions of BRAMP. However, it is reassuring both that the MCMC inference scheme can successfully deal with the increased model complexity, and that it leads to an improvement over the competing models in terms of actual network reconstruction accuracy. For the data simulated from the realistic niche model, described in Section 3.2 we found that BRAMP consistently outperforms the other methods (Figure 4). Table 1 shows the corresponding p-values of paired Wilcoxon tests for the AUROC and AUPREC values comparing BRAMP against the other methods. The low p-values indicate significant performance gain of BRAMP and suggest that the Mondrian process better captures spatial heterogeneity. The improvement over BANJO underlines the detrimental effect of the information loss inherent in data discretization.

We have applied BRAMP to the plant abundance data from the ecological survey described in Section 3.3. We sampled interaction network structures from the posterior distribution with MCMC and computed the marginal posterior probabilities of the individual potential species interactions, as described in Section 2.7. We kept all species interactions with a marginal posterior probability greater

[1] There are 3 discretization levels following Yu et al. (2004).

than 0.1 , resulting in 39 out of 106 species with relevant interaction in the reconstructed network shown in Figure 5. Since we had defined the 12 soil attributes as fixed predictors to each plant, the interactions in this network represent plant-plant interactions not mediated by similar soil preferences. This network can lead to the formation of new ecological hypotheses. For instance, *Ranunculus bulbosus* (species 14) is densely connected with four interspecific links above the posterior threshold of 0.5. A possible explanation for this observation might be a relation to its tolerance for nutrient-poor soil and preferred occurrence in species-rich patches. There is a noticeable imbalance between positive and negative interactions. An initial consultation with ecologists indicates that the fact that our analysis tends to find more positive than negative links is interesting in that it points to a dominance of facilitation over competition. The importance of facilitation was emphasised by Bruno, Stachowicz & Bertness (2003). Ecologists also suggest that positive interactions may be more characteristic for harsh environments (e.g. by Brooker & Callaghan (1998)) as it is found in the Marchair-vegetation. These results demonstrate that the proposed method provides a useful tool for exploratory data analysis in ecology with respect to both species interactions and spatial heterogeneity.

6 CONCLUSION

We have addressed the problem of reconstructing species interaction networks from species abundance data. To this end, we have proposed a Bayesian model combining Bayesian piecewise linear regression with a Mondrian process. The work is motivated by a model recently proposed in the molecular systems biology literature (Lèbre et al., 2010), but has been adapted from the temporal domain (gene expression time series) to the spatial one (snapshot of species distributions in space, typical of ecological surveys). We have introduced and tested two essential modifications: Firstly, we have expanded the data space into 2-dimensions and applied a Mondrian process following (Roy & Teh, 2008), which corresponds to a richer latent variable structure that allows modeling unobserved effects with smooth geographical variation. Secondly, we have explicitly introduced an additional enforced parent node for each species, which represents the average species abundance from the spatial neighborhood of the current location and thereby allows a correction for spatial autocorrelation. We have tested our model on data from a realistic simulation, which combines spatial species dispersal with demographic and environmental effects and predator-prey interactions of the Lotka-Volterra form defined by a trophic network obtained from a realistic niche model. Our results show that the proposed model consistently outperforms competing models (Figure 4 and Table 1). An application to plant species abundance data from a recent ecological survey has demonstrated how the proposed method can be used as a tool for hypothesis generation with respect to species interactions and spatial distribution patterns.

Table 1: **Improvement of BRAMP on Lotka-Volterra Data.** P-values for paired one-sided Wilcoxon test of AUROC and AUPREC values for several spatial β values. The alternative hypothesis states that BRAMP scores are greater than the competing methods with low p-values indicating significant performance gain of BRAMP.

SPATIAL β:	**-2**	**-4**	**-6**	**-8**
AUROCS				
BR	1.1e-06	2.8e-07	1.0e-07	1.8e-09
LASSO	6.1e-04	7.2e-04	1.3e-08	9.3e-10
BANJO	9.3e-10	9.3e-10	9.3e-10	9.3e-10
AUPRECS				
BR	1.7e-08	9.3e-10	1.8e-09	9.3e-10
LASSO	2.8e-07	5.3e-06	4.6e-09	9.3e-10
BANJO	9.3e-10	9.3e-10	9.3e-10	9.3e-10

7 FUTURE WORK

The Mondrian process is intrinsically based on two distinguished perpendicular directions. This may be more appropriate for some applications than for others. For the application in our study, the plant ecosystem on the island of Uist, these two distinguished perpendicular directions exist. The island's ecogeography, with the open sea in the west, and abutting land in the other directions, implies that the east-west soil profile (longitudinal coordinate) differs systematically from the north-south profile (latitudinal coordinate); see Lennon et al. (2011). Similar patterns can be found on many other coastal islands, where for principal directions that do not coincide with latitude and longitudinal, the Mondrian process can be formulated in terms of a local, rotated coordinate system. However, the Mondrian process will not always be the most appropriate model. For instance, for applications with rotational invariance other models, e.g. based on a Voronoi tesselation, might be more appropriate. While this provides a direction for future research, we expect that small model inadequacies, e.g. related to a violation of rotational invariance, have comparatively little effect compared to the clear advantage of the Mondrian process over a global changepoint model: namely, that it adapts the number of segments locally and therefore can deal with ecosystems that change rapidly in some areas, but slowly in others.

Further future work will explore different priors on the interaction parameters. The present prior, expressed in (4), is independent among segments and symmetric around zero. This is the most cautious approach, which allows mutualistic interactions to become neutral or antagonistic over parts of the range of the interacting species, as occasionally observed (Brooker & Callaghan, 1998). Such drastic changes in the interactions appear to be rather infrequent, though, and one may therefore want to assume that, a priori, interactions in adjacent spatial segments are, in general, more likely to be similar than different. This idea can be implemented with some mechanism of information sharing, as recently proposed in the context of time series segmentation (Grzegorczyk & Husmeier, 2012), and generalizing this method to the spatial domain provides an interesting avenue for future research.

References

Andrieu, C. & Doucet, A. (1999). Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC. *IEEE Transactions on Signal Processing*, *47*(10), 2667–2676.

Beisner, B., Haydon, D., & Cuddington, K. (2003). Alternative stable states in ecology. *Front. Ecol. Environ.*, *1*(7), 376–382.

Brooker, R. & Callaghan, T. (1998). The balance between positive and negative plant interactions and its relationship to environmental gradients: a model. *Oikos*, 196–207.

Bruno, J., Stachowicz, J., & Bertness, M. (2003). Inclusion of facilitation into ecological theory. *Evolution*, *18*(3), 119–125.

Dahlgren, J. (2010). Alternative regression methods are not considered in Murtaugh (2009) or by ecologists in general. *Ecology letters*, *13*(5), 7–9.

Davis, J. & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proc. of Int. Conf. on Machine Learning*, (pp. 233–240).

Dunne, J., Williams, R., & Martinez, N. (2002). Network structure and biodiversity loss in food webs: robustness increases with connectance. *Ecology Letters*, *5*, 558–567.

Faisal, A., Dondelinger, F., Husmeier, D., & Beale, C. (2010). Inferring species interaction networks from species abundance data: A comparative evaluation of various statistical and machine learning methods. *Ecological Informatics*, *5*(6), 451–464.

Foley, J. A., DeFries, R., Asner, G. P., Barford, C., Bonan, G., Carpenter, S. R., Chapin, F. S., Coe, M. T., Daily, G. C., Gibbs, H. K., Helkowski, J. H., Holloway, T., Howard, E. A., Kucharik, C. J., Monfreda, C., Patz, J. A., & Prentice, I. C. (2005). Global consequences of land use. *Science*, *309*, 570–574.

Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, *7*, 601–620.

Grandvalet, Y. (1998). Least absolute shrinkage is equivalent to quadratic penalization. In *ICANN 98*, volume 1 of *Perspectives in Neural Computing*, (pp. 201–206). Springer.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*, 711–732.

Grzegorczyk, M. & Husmeier, D. (2012). Bayesian regularization of non-homogeneous dynamic Bayesian networks by coupling interaction parameters. In *Fifth Int. Conf. on Art. Int. and Stat. (AISTATS)*.

Hagemeijer, W. & Blair, M. (1997). *The EBCC atlas of European breeding birds: their distribution and abundance*. Poyser London.

Hartemink, A. (2001). *Principled Computational Methods for the Validation and Discovery of Genetic Regulatory Networks*. PhD thesis, MIT.

Henneman, M. L. & Memmott, J. (2001). Infiltration of a Hawaiian Community by Introduced Biological Control Agents. *Science*, *293*(5533), 1314–1316.

Honkela et al., A. (2010). Model-based method for transcription factor target identification with limited data. *PNAS*, *107*(17), 7793–7798.

Lande, R., Engen, S., & Saether, B. (2003). *Stochastic population dynamics in ecology and conservation*. Oxford University Press.

Lèbre, S., Becq, J., Devaux, F., Lelandais, G., & Stumpf, M. (2010). Statistical inference of the time-varying structure of gene-regulation networks. *BMC Systems Biology*, *4*(130).

Lennon, J. (2000). Red-shifts and red herrings in geographical ecology. *Ecography*, *23*, 101–113.

Lennon, J., Beale, C., Reid, C., Kent, M., & Pakeman, R. (2011). Are richness patterns of common and rare species equally well explained by environmental variables. *Ecography*, *34*, 529–539.

Memmott, J., Fowler, S., Paynter, Q., Sheppard, A., & Syrett, P. (2000). The invertebrate fauna on broom, *Cytisus scoparius*, in two native and two exotic habitats. *Acta Oecol.*, *21*(3), 213–222.

Milns, I., Beale, C., & Smith, V. (2010). Revealing ecological networks using Bayesian network inference algorithms. *Ecology*, *91*, 1892–1899.

Prill, R. J., Marbach, D., Saez-Rodriguez, J., Sorger, P. K., Alexopoulos, L. G., Xue, X., Clarke, N. D., Altan-Bonnet, G., & Stolovitzky, G. (2010). Towards a Rigorous Assessment of Systems Biology Models: The DREAM3 Challenges. *PLoS ONE*, *5*(2), e9202.

Punskaya, E., Andrieu, C., Doucet, A., & Fitzgerald, W. (2002). Bayesian curve fitting using MCMC with applications to signal segmentation. *IEEE Transactions on Signal Processing*, *50*(3), 747–758.

Robinson, J. & Hartemink, A. (2010). Learning non-stationary dynamic Bayesian networks. *Journal of Machine Learning Research*, *11*, 3647–3680.

Roy, D. & Teh, Y. (2008). The Mondrian Process. In *Proc. of Advances in Neural Information Processing Systems (NIPS)*, volume 21.

Smith, V., Yu, J., Smulders, T., Hartemink, A., & Jarvis, E. (2006). Computational inference of neural information flow networks. *PLoS Comput. Biol.*, *2*(11), 1436–1449.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. Roy. Stat. Soc. B*, *58*, 267–288.

van Someren, E. P., Vaes, B. L. T., Steegenga, W. T., Sijbers, A. M., Dechering, K. J., & Reinders, M. J. T. (2006). Least absolute regression network analysis of the murine osterblast differentiation network. *Bioinformatics*, *22*(4), 477–484.

van Veen, F., Brandon, C., & Godfray, H. (2009). A positive trait-mediated indirect effect involving the natural enemies of competing herbivores. *Oecologia*, *160*(1), 195–205.

Vyshemirsky, V. & Girolami, M. A. (2008). Bayesian ranking of biochemical system models. *Bioinformatics*, *24*(6), 833–839.

Wang, P., Laskey, K., Domeniconi, C., & Jordan, M. (2011). Nonparametric Bayesian Co-clustering Ensembles. In *Proceedings of the SIAM International Conference on Data Mining*, (pp. 331–342).

Werner, E. & Peacor, S. (2003). A review of trait-mediated indirect interactions in ecological communities. *Ecology*, *84*(5), 1083–1100.

Williams, R. & Martinez, N. (2000). Simple rules yield complex food webs. *Nature*, *404*, 180–183.

Yu, J., Smith, V., Wang, P., Hartemink, A., & Jarvis, E. (2004). Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, *20*, 3594–3603.

A APPENDIX-SUPPLEMENTARY MATERIAL

Table 2: Indices with full scientific names as appearing in Figure 5. These plants can be assigned to four taxonomies of forbs (1-19), grasses (20-29), rushes (30-33) and sedges (34-39).

ID	Name
1	*Anagallis tenella*
2	*Calluna vulgaris*
3	*Drosera rotundifolia*
4	*Epilobium palustre*
5	*Galium verum*
6	*Hypochaeris radicata*
7	*Leontodon autumnalis*
8	*Lychnis flos-cuculi*
9	*Odontites verna*
10	*Plantago lanceolata*
11	*Potentilla erecta*
12	*Potentilla palustris*
13	*Prunella vulgaris*
14	*Ranunculus bulbosus*
15	*Ranunculus repens*
16	*Sagina procumbens*
17	*Succia pratensis*
18	*Trifolum repens*
19	*Viola riviniana*
20	*Agrostis capillaris*
21	*Aira praecox*
22	*Anthoxanthum odoratum*
23	*Cynosurus cristatus*
24	*Festuca rubra*
25	*Festuca vivipara*
26	*Holcus lanatus*
27	*Koeleria macrantha*
28	*Molinia caerulea*
29	*Poa pratensis*
30	*Juncus effusus*
31	*Juncus kochii*
32	*Luzula campestris*
33	*Luzula pilosa*
34	*Carex arenaria*
35	*Carex demissa*
36	*Carex dioica*
37	*Carex flacca*
38	*Carex nigra*
39	*Eriophorum angustifolum*