Check for updates

Received: 28 October 2019 | Accepted: 5 February 2020

DOI: 10.1111/2041-210X.13380

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

Tree-based inference of species interaction networks from abundance data

Raphaëlle Momal[1] | Stéphane Robin[1] | Christophe Ambroise[2]

[1]UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, Paris, France

[2]Laboratoire de Mathématiques et Modélisation d'Évry, Évry, France

Correspondence
Raphaëlle Momal
Email: raphaelle.momal@agroparistech.fr

Funding information
Next Generation Biomonitoring, Grant/Award Number: ANR-17-CE32-0011 NGB; Fondation Mathématique Jacques Hadamard, Grant/Award Number: ANR-11-LABX-0056-LMH, LabEx LMH

Handling Editor: David Warton

Abstract

1. The behaviour of ecological systems mainly relies on the interactions between the species it involves. We consider the problem of inferring the species interaction network from abundance data.
2. To be relevant, any network inference methodology needs to handle count data and to account for possible environmental effects. It also needs to distinguish between direct interactions and indirect associations and graphical models provide a convenient framework for this purpose.
3. A simulation study shows that the proposed methodology compares well with state-of-the-art approaches, even when the underlying graph strongly differs from a tree. The analysis of two datasets highlights the influence of covariates on the inferred network.
4. Accounting for covariates is critical to avoid spurious edges. The proposed approach could be extended to perform network comparison or to look for missing species.

KEYWORDS

abundance data, covariates adjustment, EM algorithm, graphical models, matrix tree theorem, Poisson log-normal model, species interaction network

1 | INTRODUCTION

There is a growing awareness of biotic interactions being crucial components of biodiversity and relevant descriptors of ecosystems (Jordano, 2016; Valiente-Banuet et al., 2015). Such interactions can be conveniently represented by networks, which have been increasingly studied and used in recent years for describing and understanding living systems in ecology (Poisot, Stouffer, & Kéfi, 2016), microbiology (Faust & Raes, 2012) or genomics (Evans, Kitson, Lunt, Straw, & Pocock, 2016). Observing species interactions is a laborious task which restricts them to certain categories (e.g. pollination, predation, parasitism), while many other types of interactions may be hard to observe and key in the system organization (e.g. communication, shelter sharing). Many efforts have been devoted in the last decade to get a more complete picture of the biotic interactions existing between species living in the same niche: all these interactions can be gathered in a so-called *species interaction network*.

Network reconstruction. A first attempt consists in using observed interactions to predict other possible links based on species traits matching (see e.g. Bartomeus et al., 2016; Graham & Weinstein, 2018; Olito & Fox, 2015; Weinstein & Graham, 2017). The interaction strength can also be predicted (Wells & O'Hara, 2013). This can be viewed as a prediction task, and modern approaches arising from signal processing and machine learning have been also proposed (Dallas, Park, & Drake, 2017; Desjardins-Proulx, Laigle, Poisot, & Gravel, 2017; Stock, Poisot, Waegeman, & Baets, 2017). We name these approaches *network reconstruction* to distinguish them from *network inference*, which is the problem we consider in this article.

Network inference. Network inference approaches also aim at retrieving the interactions among species, but do not rely on observed

 © 2020 British Ecological Society

interactions and therefore, remain agnostic as for their type. Such approaches have been developed in many domains ranging from cell biology (Friedman, 2004, to infer gene regulatory networks) to neurosciences (Zhu & Cribben, 2018, to decipher brain connectivity structures). In ecology, it will typically aim at inferring the set of biotic interactions linking species from the same guild. As summarized in Figure 1, network inference takes input measures on species (here abundances) at similar sites, and returns a network of direct interactions between species. The importance of distinguishing between direct interaction and indirect association between species is explained in Popovic, Warton, Thomson, Hui, and Moles (2019).

Species not engaged in biotic interactions can appear linked if they respond similarly or oppositely to an abiotic effect (spurious interaction). Therefore network inference must account for environmental covariates. Figure 2 illustrates this phenomenon: in (c) species (1 and 4) are not in direct interaction, but are affected by the variations of the same environmental covariate *x*. (d) displays the network when *x* is not accounted for: a spurious edge appears between species.

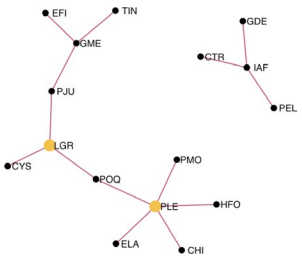
Joint species distribution models. The rationale behind network inference is that interactions between species must affect their joint distribution in a series of similar sites. Such approaches necessarily rely on a *joint* species distribution model (JSDM), as opposed to species distribution models (Elith & Leathwick, 2009) where species are traditionally considered as disconnected entities. A JSDM is a probabilistic model describing the species' simultaneous presence/absence (Harris, 2015; Ovaskainen et al., 2017) or joint abundances (Popovic, Hui, & Warton, 2018; Popovic et al., 2019). An important feature of JSDMs is to include environmental covariates to account for abiotic interactions.

Recently, latent variable models have received attention in community ecology as they provide a convenient way to model the dependence structure between species (Warton et al., 2015). The JSDM proposed by Popovic et al. (2018, 2019) involves a latent layer. So does the Poisson log-normal model (PLN; Aitchison & Ho, 1989), which combines generalized linear models to account for covariates and offsets, and a Gaussian latent structure to describe the species interactions. It can be seen as a multivariate mixed model, in which correlated random effects encode the dependency between the species abundances.

In (b), the network is disconnected: species 4 is independent from all others. This illustrates that graphical models enjoy all the desirable properties to represent interactions between species in an interpretable manner, so that they can be used as the mathematical counterpart of species interaction networks. Graphical models: a generic framework for network inference. Although they describe the dependence structure between the distributions of all the species from a same niche, JSDM are not sufficient to perform network inference as they do not distinguish indirect associations from direct interactions (Dormann et al., 2018). Graphical models (Lauritzen, 1996) provide a probabilistic framework to do so and, in the same time, a formal definition of the network to be inferred. This formalism is therefore especially appealing for the inference of species interaction networks (Popovic et al., 2019). In a undirected graphical model (which is the same as a Markov random field: Clark, Wells, & Lindberg, 2018), two species are connected if they are *dependent* conditional on all other species, that is if the variations of their respective abundances would still be correlated if ever both the environmental conditions and the abundances of all other species were kept fixed. Two species are unconnected if they are *independent* conditional on all other species: the observed correlation between them only results from a series of links with other species (Morueta-Holme et al., 2016) or environmental effects. Figure 2 illustrates the concept of conditional dependence/independence with toy graphical models. In (a), the network is connected so all species are interdependent: an association exists between any two of them. However, 1 is only directly interacting with 2 which

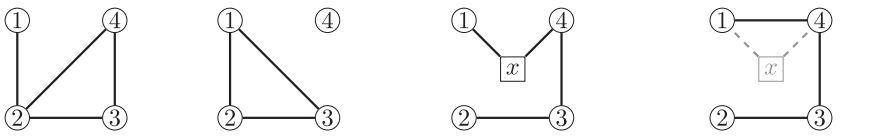
date	site	EFI	ELA	GDE	GME	HFA
apr93	km03	71	1	5	6	0
apr93	km03	118	2	3	0	0
apr93	km03	69	0	6	2	0
apr93	km03	56	0	0	0	0
apr93	km17	0	1	1	0	0
apr93	km17	0	0	2	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮

(a) Covariates **(b)** Species abundances



(c) Inferred network

FIGURE 1 Aim of species interaction network inference from abundance data. Data sample from the Fatala river dataset (see Section 2.3.4)



(a) Connected **(b)** Disconnected **(c)** With covariate **(d)** Missing covariate

FIGURE 2 Examples of graphical models. (a) All species are dependent, (b) 4 is independent from all others, (c) 1 and 4 are independent conditional on *x*, (d) not accounting for *x* induces a spurious dependence between 1 and 4

mediates its association with 3 and 4:1 is independent from them conditional on 2.

In (b), the network is disconnected: species 4 is independent from all others. This illustrates that graphical models enjoy all the desirable properties to represent interactions between species in an interpretable manner, so that they can be used as the mathematical counterpart of species interaction networks.

Network inference: The general problem. Network inference methods attempt to retrieve the graphical model underlying the distribution of abundance data. In every domain, network inference is impeded by the huge number of possible graphs for a given set of nodes, which increases super-exponentially with the latter (more than 10^{13} undirected graphs can be drawn between 10 nodes, and more than 10^{57} between 20). The exploration of the graph space is therefore intractable from a combinatorial point of view. To reduce the search space, a common and reasonable assumption is that a relatively small fraction of species pairs is in direct interaction: the network is sparse. In the case of continuous observations, one of the most popular approaches is the graphical lasso (glasso: Friedman, Hastie, & Tibshirani, 2008), which takes advantage of the properties of Gaussian graphical models (GGM) to efficiently infer a sparse network. Alternatively, tree-based approaches have been proposed: Chow and Liu (1968) first made the too stringent assumption that the network is made of a single spanning tree (that is connecting all nodes without any loop, as in Figure 3). More recent approaches introduced by Meilă and Jaakkola (2006) and Kirshner (2008) rely on efficient algebraic tools to average over all possible tree-structured graphical models. The inferred network resulting from such an averaging procedure is not restricted to be a tree: species or groups of species can be isolated (e.g. Figure 1), and loops can appear (e.g. Figure 3).

Network inference from species abundance data. This work focuses on network inference based on abundance data, and not only their presence/absence (as considered in Clark et al., 2018; Ovaskainen, Hottola, & Siitonen, 2010). Network inference from species abundance measures is a notoriously difficult problem (Ulrich & Gotelli, 2010), not only because network inference is complex, but also because it has to account for the data specificities. Abundance data may spread over a wide range of values and often result from sampling efforts (sample and/or species-specific), making them difficult to compare. Obviously, count data do not directly fit the Gaussian framework but many network inference methods dedicated to abundance data actually rely on a latent Gaussian structure (see Section 2.3.1).

Contribution. In the present work, we adopt a model-based approach to perform network inference from abundance data. To accommodate the data specificities, we use a PLN model, which includes the over-dispersion of the observed counts as well as the sampling effort. Importantly, the PLN model allows us to account for abiotic effects and avoid the detection of spurious interactions between species.

As for the network inference, we adopt a tree-based approach (as opposed to Biswas, McDonald, Lundberg, Dangl, & Jojic, 2016, which also uses a PLN model but resort to glasso), which provides a probability for each edge to be actually part of the underlying graphical model.

Outline. We introduce the method EMtree, which combines two (variational) expectation-maximization (EM) algorithms to estimate the model parameters. Importantly, our approach provides the probability for each possible edge to be part of the interaction network. We evaluate our approach on both synthetic and ecological datasets. An R package implementing EMtree is available on GitHub https://github.com/Rmomal/EMtree (https://doi.org/10.5281/zenodo.3660627).

2 | MATERIALS AND METHODS

2.1 | Model

Let us first describe the typical type of data we consider. We assume that p species have been observed in n sites. The abundances are gathered in the $n \times p$ matrix $\mathbf{Y}$. Y_{ij} is the abundance of species j in site i, and $\mathbf{Y}_i$ the abundance vector collected in site i (ith row of $\mathbf{Y}$). We further assume that a vector of covariates $\mathbf{x}_i$ of size d has been measured in each site i and that all covariates are gathered in the $n \times d$ matrix $\mathbf{X}$. The sites are supposed to be independent. Our aim is to decipher the dependency structure between the p species, accounting for the effect of the environmental covariates encoded in $\mathbf{X}$. As explained above, ignoring environmental covariates is more than likely to result in spurious edges.

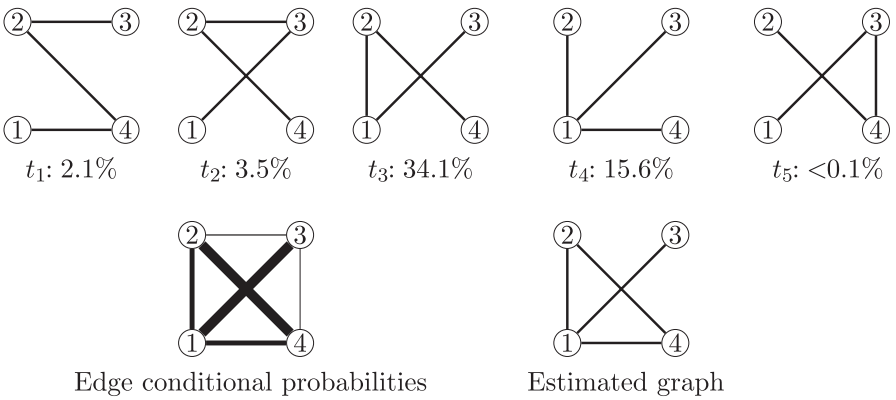


FIGURE 3 Tree averaging principle. *Top*: 5 spanning trees with 4 nodes (t_1, ..., t_5), with their respective conditional probability given the data $P(T = t|Y)$. *Bottom left*: Weighted graph resulting from tree averaging. Each edge has width proportional to its conditional probability. *Bottom right*: Estimated graph (obtained by thresholding edge probabilities) is not a tree

Mixed model. To distinguish between covariates effects and species interactions, we consider a mixed model which states that each abundance Y_{ij} has a (conditional) Poisson distribution

$$Y_{ij} \sim \mathcal{P}\left(\exp\left(\mathbf{x}_i^\intercal \boldsymbol{\theta}_j + o_{ij} + Z_{ij}\right)\right). \quad (1)$$

In model (1), o_{ij} is the sample- and species-specific offset which accounts for the sampling effort. $\boldsymbol{\theta}_j$ is the vector of fixed regression coefficients measuring the effect of each covariate on species j abundance. The regression part is similar to a general linear model as used in niche modelling (see e.g. Austin, 2007). Z_{ij} is the random effect associated with species j in site i. Importantly, the coordinates of the site-specific random vector $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})$ are not independent: the multivariate random term $\mathbf{Z}_i$ precisely accounts for the interactions that are not due to environmental fluctuations. For each site i, a vector $\mathbf{Z}_i$ is associated with the corresponding abundance vector $\mathbf{Y}_i$. The distribution given in Equation 1 is over-dispersed as the Poisson parameter is itself random, which suits ecological modelling of abundance data (Richards, 2008).

We now describe the distribution of the latent vector $\mathbf{Z}_i$. To this aim, we adopt a version of Kirshner's model (Kirshner, 2008), which states that a spanning tree T is first drawn with probability

$$p(T) = \prod_{(j,k)\in T} \beta_{jk}/B, \quad (2)$$

where $(j,k) \in T$ means that the edge connecting species j and k is part of the tree T and where B is a normalizing constant. Each edge weight β_{jk} controls the probability for the edge (j, k) to be in the interaction network.

Then for each site i, a vector $\mathbf{Z}_i$ is drawn independently with conditional Gaussian distribution $(\mathbf{Z}_i|T) \sim \mathcal{N}(0, \Sigma_T)$, where the subscript T means that the distribution of $\mathbf{Z}_i$ is faithful to T. When T is a spanning tree, this faithfulness simply means this distribution can be factorized on the nodes and edges of T as follows (see Kirshner, 2008):

$$p\left(\mathbf{Z}_i|T\right) = \prod_{j=1}^{p} p\left(Z_{ij}|T\right) \prod_{(j,k)\in T} \psi_{jk}(\mathbf{Z}_i), \quad (3)$$

where $\psi_{jk}(\mathbf{Z}_i)$ does not depend on T. This factorization means that each edge of T corresponds to a species pair in direct interaction; all other pairs are conditionally independent. Experiments are independent, and in the sequel we consider the product of all $p(\mathbf{Z}_i)$ and use the simpler notation $\psi_{jk} = \prod_i \psi_{jk}(\mathbf{Z}_i)$ instead.

According to Equation 2, each $\mathbf{Z}_i$ has a Gaussian distribution conditional on the tree T, so its marginal distribution is a mixture of Gaussians: $\mathbf{Z}_i \sim \sum_{T\in\mathcal{T}} p(T)\,\mathcal{N}(0, \Sigma_T)$, where $\mathcal{T}$ is the set of all spanning trees. As a consequence, the joint distribution of the $\mathbf{Z}_i$ is modelled by a mixture of distributions with tree-shaped dependency structure.

Besides, for all trees including the edge (j, k), the estimate of the covariance term between the coordinates j and k is the same (see Lauritzen, 1996; Schwaller, Robin, & Stumpf, 2019). Hence, we may define a global covariance matrix $\boldsymbol{\Sigma}$, filled with covariances that are each common to spanning trees containing a same edge. Each $\boldsymbol{\Sigma}_T$ is then built by extracting from $\boldsymbol{\Sigma}$ the covariances corresponding to the edges of T.

2.2 | Inference with EMtree

We now describe how to infer the model parameters. We gather the edge weights into the $p \times p$ matrix $\boldsymbol{\beta}$ and the vectors of regression coefficients into a $d \times p$ matrix $\boldsymbol{\theta}$. The $p \times p$ matrix $\boldsymbol{\Sigma}$ contains the variances and covariances between the coordinates of each latent vector $\mathbf{Z}_i$. Hence, the set of parameters to be inferred is $(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\theta})$.

Likelihood. The model described above is an incomplete data model, as it involves two hidden layers: the random tree T and the latent Gaussian vectors $\mathbf{Z}_i$. The most classical approach to achieve maximum likelihood inference in this context is to use the EM algorithm (Dempster, Laird, & Rubin, 1977). Rather than the likelihood of the observed data $p(\mathbf{Y})$, the EM algorithm deals with the often more tractable likelihood $p(T, \mathbf{Z}, \mathbf{Y})$ of the complete data (which consists of both the observed and the latent variables). It can be decomposed as

$$p_{\boldsymbol{\beta},\boldsymbol{\Sigma},\boldsymbol{\theta}}(T, \mathbf{Z}, \mathbf{Y}) = p_{\boldsymbol{\beta}}(T) \times p_{\boldsymbol{\Sigma}}(\mathbf{Z}|T) \times p_{\boldsymbol{\theta}}(\mathbf{Y}|\mathbf{Z}), \quad (4)$$

where the subscripts indicate on which parameter each distribution depends.

Observe that the dependency structure between the species is only involved in the first two terms, whereas the third term only depends on the regression coefficients $\boldsymbol{\theta}$. We take advantage of this decomposition to propose a two-stage estimation algorithm. The first stage deals with the observed layer $p_{\boldsymbol{\theta}}(\mathbf{Y}|\mathbf{Z})$, the second with the two hidden layers $p_{\boldsymbol{\beta}}(T)$ and $p_{\boldsymbol{\Sigma}}(\mathbf{Z}|T)$. The network inference itself takes place in the second step.

Inference in the observed layer. The variational EM (VEM, Blei, Kucukelbir, & McAuliffe, 2017; Ormerod & Wand, 2010) algorithm that provides an estimate of the regression coefficients matrix $\boldsymbol{\theta}$ is described in Appendix A.1 (along with a reminder on EM and VEM). It also provides the (approximate) conditional means $\mathbb{E}(Z_{ij}|\,\mathbf{Y}_i)$, variances $\mathbb{V}(Z_{ij}|\,\mathbf{Y}_i)$ and covariances $\mathbb{C}\mathrm{ov}(Z_{ij}, Z_{ik}|\,\mathbf{Y}_i)$ required for the inference in the hidden layer (Chiquet, Mariadassou, & Robin, 2018). As a consequence, this first step provides the estimates $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\Sigma}}$.

Inference in the hidden layer. The second step is dedicated to the estimation of $\boldsymbol{\beta}$. The EM algorithm actually deals with the conditional expectation of the complete log-likelihood, namely $\mathbb{E}\left(\log p_{\boldsymbol{\beta},\boldsymbol{\Sigma},\boldsymbol{\theta}}(T, \mathbf{Z}, \mathbf{Y})|\mathbf{Y}\right)$. As shown in Appendix A.2, this reduces to

$$\mathbb{E}\left(\log p_{\boldsymbol{\beta},\boldsymbol{\Sigma},\boldsymbol{\theta}}(T, \mathbf{Z}, \mathbf{Y})|\mathbf{Y}\right) \simeq \sum_{1\le j<k\le p} P_{jk} \log\left(\beta_{jk}\hat{\psi}_{jk}\right) - \log B + \mathrm{cst}, \quad (5)$$

where $\hat{\psi}_{jk}$ is the estimate of ψ_{jk} defined in Equation 3, and the 'cst' term depends on $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ but not on $\boldsymbol{\beta}$. P_{jk} is the approximate conditional probability (given the data) for the edge (j, k) to be part of the network: $P_{jk} \simeq \mathbb{P}\{(j,k) \in T|Y\}$. It is also shown in Appendix A.2 that $\hat{\psi}_{jk} = (1 - \hat{\rho}_{jk}^2)^{-n/2}$, where the estimated correlation $\hat{\rho}_{jk}$ depends on the conditional mean, variance and covariances of the Z_{ij}'s provided by the

first step. Equation 5 is maximized via an EM algorithm iterating the calculation of the P_{jk} and the maximization with respect to the β_{jk}:

Expectation step: Computing the P_{jk} with tree averaging The conditional probability of an edge is simply the sum of the conditional probabilities of the trees that contain this edge. Hence, computing P_{jk} amounts to averaging over all spanning trees. Figure 3 illustrates the principle of tree averaging for a toy network with $p = 4$ nodes. Here, five arbitrary spanning trees t_1 to t_5 (among the $p^{p-2} = 16$ spanning trees) are displayed, with their respective conditional probability $p(T|Y)$. The edge (1, 3) has a high conditional probability P_{13} because it is part of likely trees such as t_3 and t_4, whereas P_{23} is small because the edge (2, 3) is only part of unlikely trees (e.g. t_1, t_2).

Averaging over all spanning at the cost of a determinant calculus (i.e. with complexity $O(p^3)$) is possible using the Matrix Tree theorem (Chaiken & Kleitman, 1978, recalled as Theorem 1 in Appendix A.3). Kirshner (2008) further shows that all the P_{jk}'s can be computed at once with the same complexity $O(p^3)$, although the calculation may lead to numerical instabilities for large n and p.

Maximization step: Estimating the β_{jk} This step is not straightforward, as the normalizing constant $B = \sum_T \prod_{(j,k)\in T} \beta_{jk}$ involves all β_{jk}'s. We propose an exact maximization built upon the Matrix Tree theorem (see Appendix A.2).

Algorithm output: Edge scoring and network inference. EMtree provides the (approximate) conditional probability P_{jk} for each edge (j, k) to be part of the network. Whenever an actual inferred network $\hat{G}$ is needed (e.g. for a graphical purpose), it can be obtained by thresholding the P_{jk} (see Figure 3, bottom right). Because we are dealing with trees, a natural threshold is the density of a spanning tree, which is $2/p$. More robust results can be obtained using a resampling procedure similar to the stability selection proposed by Liu, Roeder, and Wasserman (2010). It simply consists in sampling a series of subsamples $s = 1, \ldots, S$, to get an estimate $\hat{G}^s$ from each of them and to collect the selection frequency for each edge. Again, these edge selection frequencies can be thresholded if needed.

2.3 | Simulation and illustrations

Because network inference is an unsupervised problem (as opposed to network reconstruction), we compare the accuracy of the methods described above on synthetic abundance datasets, for which the true underlying network is known.

2.3.1 | Alternative inference methods

We consider network inference methods dedicated to both metagenomics (SPIEC-EASI, gCoda and MInt) and ecology (MRFcov, ecoCopula). All methods can handle count data and rely on some (implicit) Gaussian setting. SPIEC-EASI (Kurtz et al., 2015), gCoda (Fang, Huang, Zhao, & Deng, 2017) and MRFcov (Clark et al., 2018) resort to data transformation to fit a Gaussian framework. MInt (Biswas et al., 2016) considers a Poisson mixed model similar to the one we consider and ecoCopula (Popovic et al., 2019) defines a multivariate count distribution, the dependency structure of which is encoded in a Gaussian copula. These methods rely on a GGM or a Gaussian copula, so that the network inference problem amounts to estimating a sparse version of the inverse covariance matrix (also named *precision* matrix).

Edge scoring. These methods build upon glasso penalization (Friedman et al., 2008). For each edge, there exists a minimal penalty value above which it is eliminated from the network. The higher this minimal penalty, the more reliable the edge in the network, so it can be used as a score reflecting the importance of an edge. Only SpiecEasi and gCoda provide unthresholded quantities (namely the glasso regularization path) that can be used for edge scoring; the other methods only return their optimal graph.

Covariates. Only MInt, MRFcov and ecoCopula may include covariates. In order to draw a fair comparison, we give SPIEC-EASI and gCoda access to the covariate information by feeding them with residuals of the linear regression of the transformed data onto the covariates.

2.3.2 | Comparison criteria

False discovery rate (FDR) and density ratio criteria. Inferred networks are mostly useful to detect potential interactions between species, which then need to be studied by experts to determine their exact nature. Falsely including an edge leads to meaningless interpretation or useless validation experiments.

A network with a few reliable edges will be preferred to a one having more edges with a larger risk of possible false discoveries. Therefore we choose the FDR as an evaluation criterion, which should be close to 0. Comparing FDR's only makes sense for networks with similar densities. We then compute the ratio between the densities of the inferred and the true network (*density ratio*).

Area under the curve (AUC) criterion. The AUC criterion allows to evaluate the inference quality without resorting to any threshold. It evaluates the probability for a method to score the presence of a present edge higher than that of an absent one; it should be close to 1. Note that this criterion cannot be computed for MRFcov, ecoCopula and MInt as they provide a unique network.

2.3.3 | Simulation design

Simulated graphs. We consider three typical graph structures: Scale-free, Erdös (short for Erdös-Reyni) and Cluster. Scale-free structure bears the closest similarity to the tree one, with almost the same density and no loops; it is popular in social networks and in genomics as it corresponds to a preferential-attachment behaviour. It is simulated following the Barbási-Albert model as implemented in the HUGE R package (Zhao, Liu, Roeder, Lafferty, & Wasserman, 2012). The degree distribution of Scale-free structure follows a power law, which constrains the edges probabilities such that the

network density cannot be controlled. Erdös structure is the most even as the edges all have the same existence probability. It is a step away from the tree as it may contain loops and its density can be increased arbitrarily. Cluster structure spreads edges into highly connected clusters, with few connections between the clusters; the *ratio* parameter controls the intra/inter connection probability ratio.

Simulated counts. The datasets are simulated under the Poisson mixed model described in Equation 1. We first build the covariance matrix Σ_G associated with a graph *G* following Zhao et al. (2012) and randomly choosing the sign of the link, so that in our simulations we consider both positive and negative interactions. For each site *i*, we simulate $\mathbf{Z}_i \sim \mathcal{N}(0, \Sigma_G)$, then use these parameters together with a set of covariates to generate count data **Y**. We use three covariates (one continuous, one ordinal and one categorical), with their regression coefficients θ drawn from a standard uniform distribution to create heterogeneity in environmental response across species.

Experiments. For each set of parameters and type of structure, we generate 100 graphs, simulate a dataset under a heterogeneous environment and infer the dependency structure using EMtree, gCoda, SpiecEasi MInt, ecoCopula and MRFcov (the three latter only for Experiment 1). The settings of all methods are set to default, except for ecoCopula for which we use the 'AIC' selection criterion ('BIC' gives too many empty results). All computation times are obtained with a 2.5 GH Intel Core 17 processor and 8G of RAM.

Experiment 1 **Effect of the data dimensions on the inferred network.** We compare performances in terms of FDR and density ratios on two scenarios: *easy* (n = 100, p = 20), and *hard* (n = 50, p = 30). The network density for Erdös and Cluster structures is set to $\log(p)/p$.

Experiment 2 **Effect of the network structure on edge rankings.** AUC measures are collected for alternate variations of n and p to get a general idea of each performance. For comparison's sake, the same density is fixed for all structures in this case, so that only n and p vary in turn; the Scale-free structure imposes a common density of $2/p$. The default values are n = 100, p = 20.

Experiment 3 **Effect of the graph density on edge rankings.** AUC measures are collected for variations of n and p with a density of $5/p$ (5 neighbours per node on average), and for variations of density parameters. The default values are n = 100, p = 20.

2.3.4 | Illustrations

The first application deals with fish population measurements in the estuary of the Fatala River, Guinea (Baran, 1995, available in the R package ADE4). The data consist of 95 count samples of 33 fish species, and two covariates *date* and *site*. We infer the network using four models including no covariates, either one or both covariates (i.e. respectively the *null*, *site*, *date* and *site + date* models).

The second example is a metabarcoding experiment designed to study oak powdery mildew (Jakuschkin et al., 2016), caused by the fungal pathogen *Erysiphe alphitoides* (Ea). To study the pathobiome of oak leaves, measurements were done on three trees with different infection status. The resulting dataset is composed of 116 count samples of 114 fungal and bacterial operational taxonomic units (OTUs) of oak leaves, including the Ea agent. The original raw data are available at https://www.ebi.ac.uk/ena/data/view/PRJEB7319. Several covariates are available, among which the tree status, the orientation of the branch, and three covariates measuring the distances of oak leaves to the ground (D1), to the base of the branch (D2) and to the tree trunk (D3). The experiment used different depths of coverage for bacteria and fungi, which we account for via the offset term. We fitted three Poisson mixed models including none, the tree status or all of the covariates (i.e. respectively *null*, *tree* and *tree* + D1 + D2 + D3 models).

To further analyse the inferred networks, we use the betweenness centrality (Freeman, 1978), a centrality measure popular in social network analysis. It measures a node's ability to act as a bridge in the network. High betweenness scores identify sensitive nodes that can efficiently describe a network structure. We compute these using the R package IGRAPH.

3 | RESULTS

3.1 | On simulated data

3.1.1 | Effect of dataset dimensions

Behaviours are compared on an easy setting (n = 100, p = 20) and a hard setting (n = 50, p = 30). Figure 4 displays FDR and density ratio measures for all methods on the different cases. Detailed values of medians and standard-deviations are given in Tables S3 and S4. The behaviour of methods remains virtually the same across Erdös and Cluster structures. Scale-free structure appears to entail a greater difficulty for all methods except ecoCopula: the FDR increases in easy cases of about 15% for SpeacEasi, MRFcov and EMtree, and about 35% for MInt.

The greater difficulty affects all methods. gCoda standard deviation increases by 10%. MRFcov, EMtree and MInt show an increase in FDR of about 5%, 20% and 30% respectively. Density ratios overall decrease, especially for ecoCopula which ratio is close to 0 and yields a proportion of empty networks of 15%–25% (Table S5).

Considering FDRs and density ratios combined, EMtree appears to be the method with the lower FDR which maintains a density ratio reasonably close to 1. As a consequence, the proposed methodology compares well to existing tools on problems with varying difficulties. EMtree is also comparable on running times. Table 1 shows that for Erdös and Cluster it is the third quicker method in easy cases and the second in hard ones. Table S6 shows that on scale-free problems, EMtree is the second quicker method in hard cases, and is curiously slow on easy ones.

Interestingly, in easy cases when the network density is well estimated, methods yield FDR of 10%–30% in median. This reminds that network inference from abundance data is a difficult task, and that perfect inference of the network remains an out-of-reach goal.

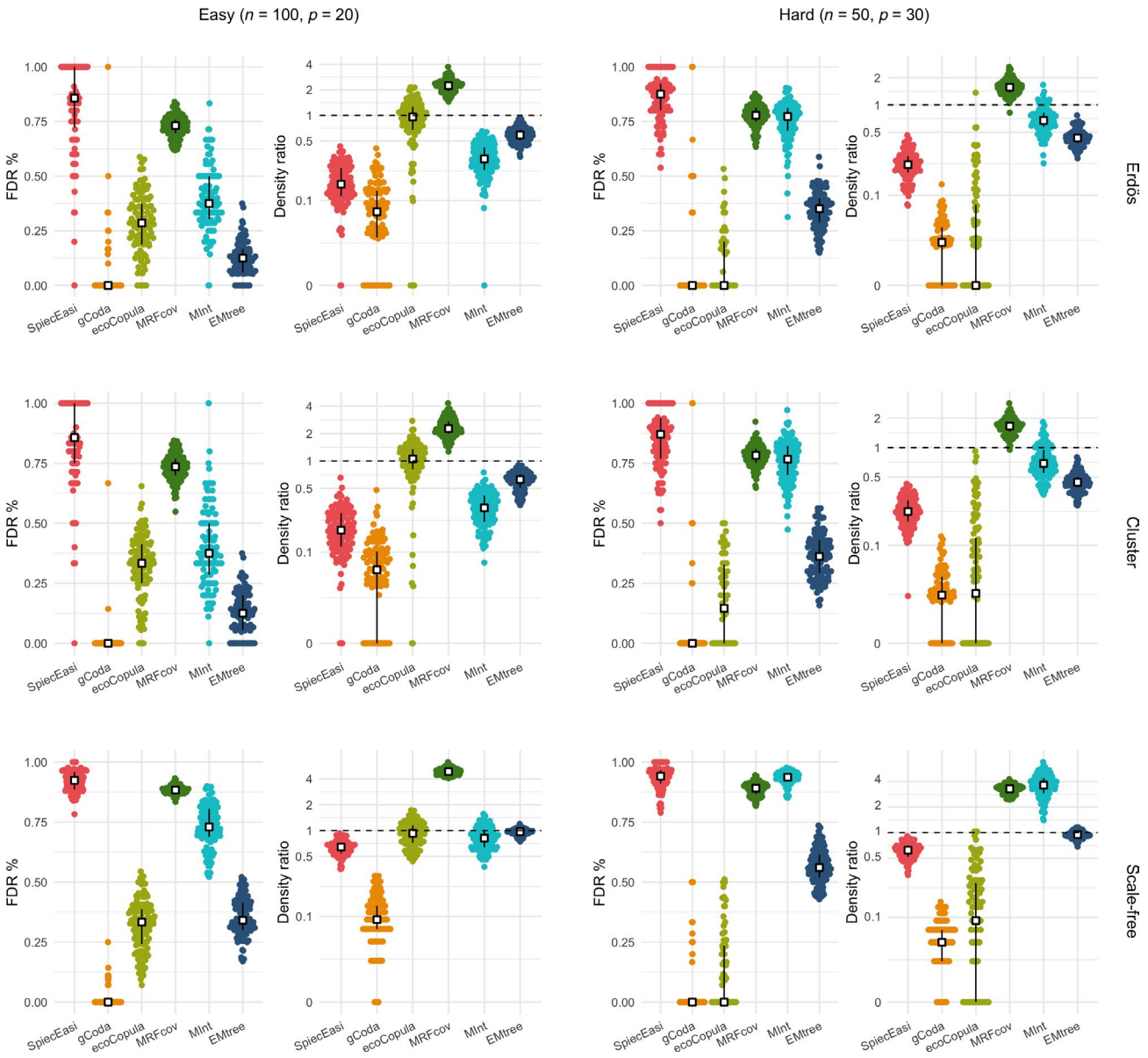


FIGURE 4 False discovery rate and density ratio measures for all methods at two different difficulty levels and 100 networks of each type. White squares and black plain lines represent medians and quartiles respectively. *ecoCopula selection method: AIC. Number of subsamples for SpiecEasi and EMtree: S* = 20. *SpiecEasi and gCoda: lambda.min.ratio* = 0.001, *nlambda* = 100

TABLE 1 Median and standard-deviation running-time values (in seconds) for Cluster and Erdös structures, including resampling with *S* = 20 for SpiecEasi and EMtree

	SpiecEasi	gCoda	ecoCopula	MRFcov	MInt	EMtree
Easy	25.45 (1.87)	0.11 (0.06)	5.55 (0.64)	34.51 (3.68)	43.04 (19.76)	11.72 (1.89)
Hard	28.43 (1.30)	0.53 (0.25)	9.6 (0.65)	8.29 (0.36)	33.77 (18.20)	8.17 (0.50)

3.1.2 | Effect of network structure

As expected for a fixed *p*, the higher the number of observations *n*, the better the performance for all methods and structures. Interestingly, the same happens when *p* increases for a fixed *n* = 100 (except for SpiecEasi). EMtree performs well on Scale-free structures (Figure 5) which was also expected; the other methods performance worsen compared to other structures. When lowering *n* to 30, EMtree performance deteriorates along with *p*, yet remaining above 70% in median in the extreme case where *p* = *n* (Figure 5, right). The structure being Erdös or Cluster, each method is affected in the same way by an increase of *n* or *p* (Figure 6). Besides, increasing the difference between the two structures by tuning up the *ratio* parameter has no effect. Overall EMtree performs better

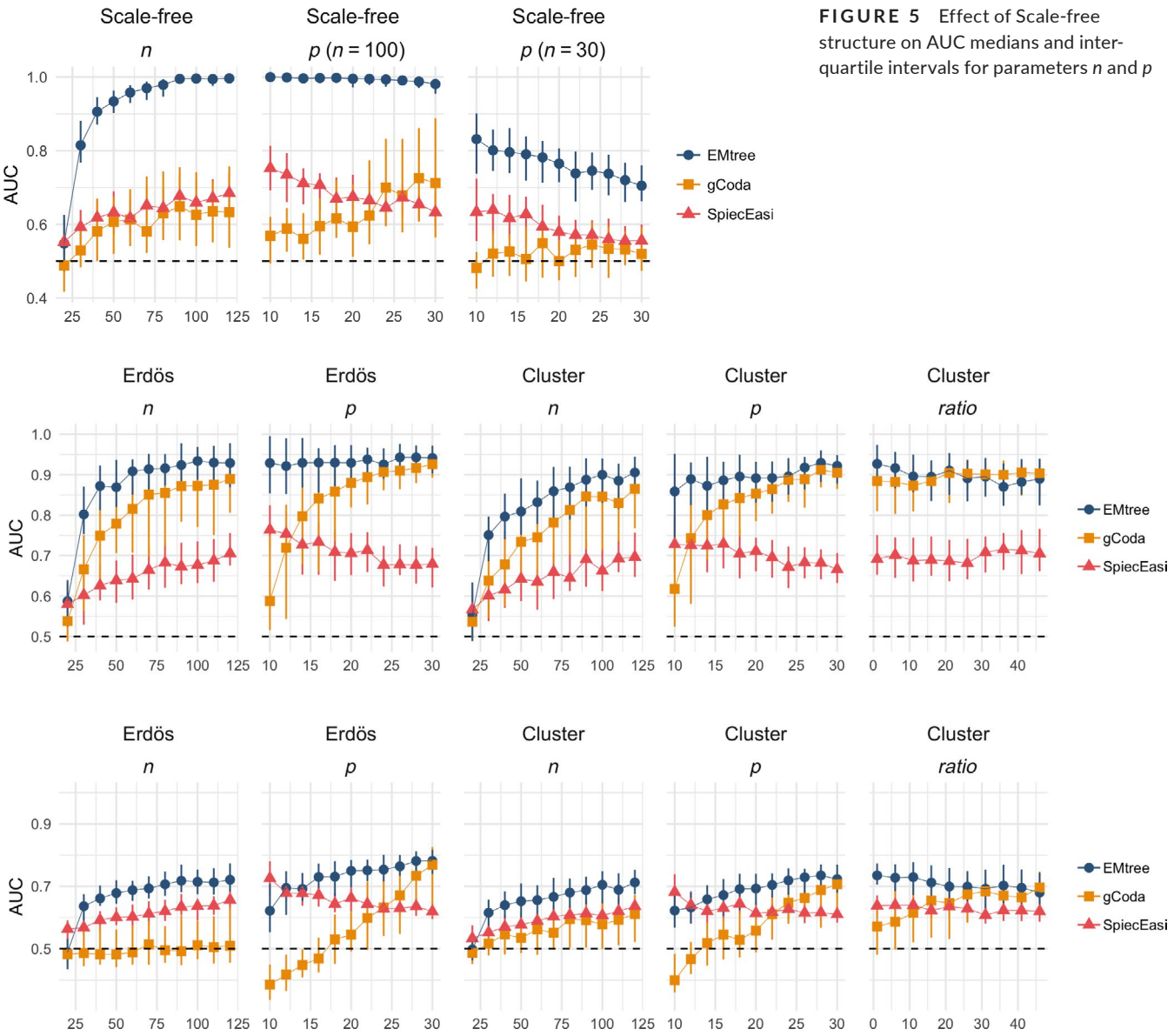


FIGURE 5 Effect of Scale-free structure on AUC medians and inter-quartile intervals for parameters *n* and *p*

FIGURE 6 Effect of Erdös and Cluster structures on AUC medians and inter-quartile intervals for parameters *n*, *p* and *ratio*. *Top*: densities set to 2/*p*, *bottom*: densities set to 5/*p*

than gCoda and SpiecEasi on all the studied configurations. Running times are summarized in Table 2. EMtree is about 10 times slower than gCoda (4 for small *n*), and four times faster than SpiecEasi. The high standard deviation for small *n* seems to be due to gCoda struggling with Scale-free structures.

TABLE 2 Median and standard-deviation of running times for each method in seconds, for *n* and *p* parameters

	$n < 50$	$n \geq 50$	$p < 20$	$p \geq 20$
EMtree	0.44 (0.14)	0.60 (0.17)	0.41 (0.13)	0.76 (0.21)
gCoda	0.11 (26.8)	0.05 (0.05)	0.05 (0.04)	0.09 (0.54)
SpiecEasi	2.09 (0.26)	2.37 (0.28)	2.42 (0.27)	2.42 (0.26)

3.1.3 | Effect of network density

The comparison of top and bottom panels of Figure 6 shows that network inference gets harder as the network gets denser, whatever the method and the structure of the true graph. Running times are not affected (Table S8). Figure 7 shows that EMtree performance does not deteriorate faster than that of other methods, demonstrating that the tree hypothesis is not a limitation.

3.2 | Illustrations

In this section we emphasize the importance of covariates for network inference. Accounting for environmental effects changes the structure of all inferred networks we present; nodes with the highest betweenness scores are highlighted to spot these

changes. Most frequently, it results in reducing the number of edges (i.e. making the network sparser). However new edges can appear as well, as adjusting for a covariate also reduces the variability, which improves the detection power. In all examples, we used the resampling method described in Section 2.2, which provides edge selection frequencies. Eventually, we have to threshold these frequencies to draw actual networks; the value of the threshold obviously affects the density of the plotted networks (see Figure S12).

3.2.1 | Fish populations in the Fatala River estuary

Networks on Figure 8 suggest a predominant role of the *site* covariate compared to the *date*. Indeed, adjusting for the *site* results in much sparser networks (Figure S12). It deeply modifies the network structure: the *site* network has 12 new edges and only six in common with the *null* network. Besides, the highlighted nodes only change when introducing the *site* covariate. This suggests that the environmental heterogeneity between the sites has a major effect on the variations of species abundances, while the effect of the date of sampling is moderate.

3.2.2 | Oak powdery mildew

When providing the inference with more information (tree status, distances), the structure of the resulting network is significantly modified. Nodes with high betweenness scores differ from one model to another. There is an important gap in density between the *null* model and the others, starting from a 25% selection threshold (Figure S12). From a more biological point of view, the features of the pathogen node are greatly modified too: its betweenness score is among the smallest in the *null* network (quantile 16%), and among the highest in the two other networks (quantiles 93% and 96%). Its connections to the other nodes vary as well. Accounting for covariates results in less interactions with the pathogen but a greater role of the latter in the pathobiome organization (Figure 9).

Using the dataset restricted to infected samples (39 observations for 114 OTUs) and correcting for the leaves position in the tree (proxy for their abiotic environment), Jakuschkin et al. (2016) identifies a list of 26 OTUs likely to be directly interacting with the pathogen. Running EMtree on the same restricted dataset with the same correction yields a good concordance with edge selection frequencies, as shown in Figure 10.

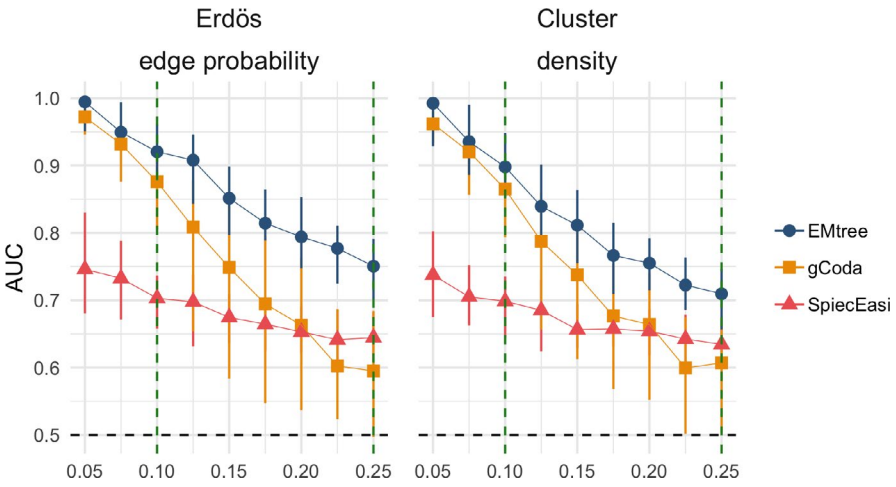


FIGURE 7 AUC median and interquartile intervals for parameters controlling the number of edges in Erdös (*edge probability*) and in Cluster (*density*) structures, $p = 20$, $n = 100$

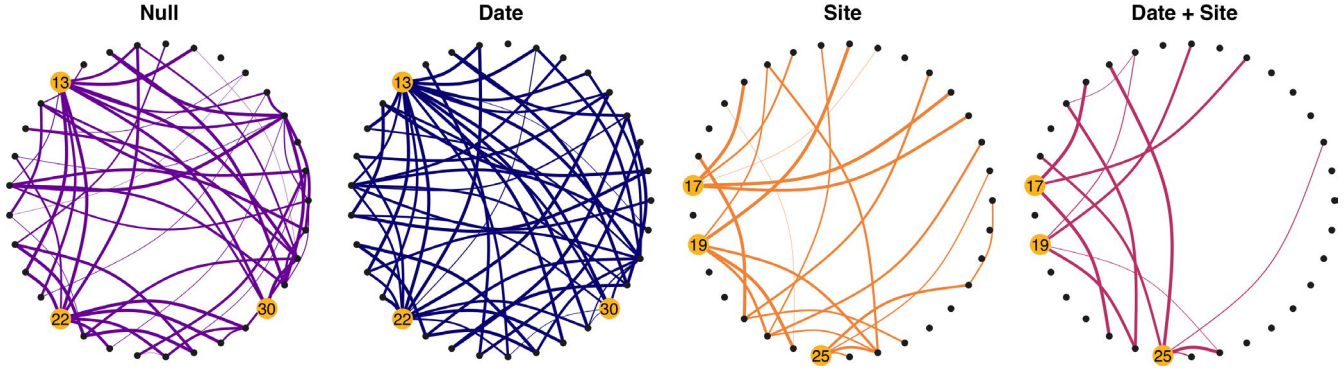


FIGURE 8 Interaction networks of Fatala River fishes inferred when adjusting for none, both or either one of the covariates among *site* and *date*. Highlighted nodes spot the highest betweenness centrality scores. Widths are proportional to selection frequencies. $S = 100$, $f' = 90\%$

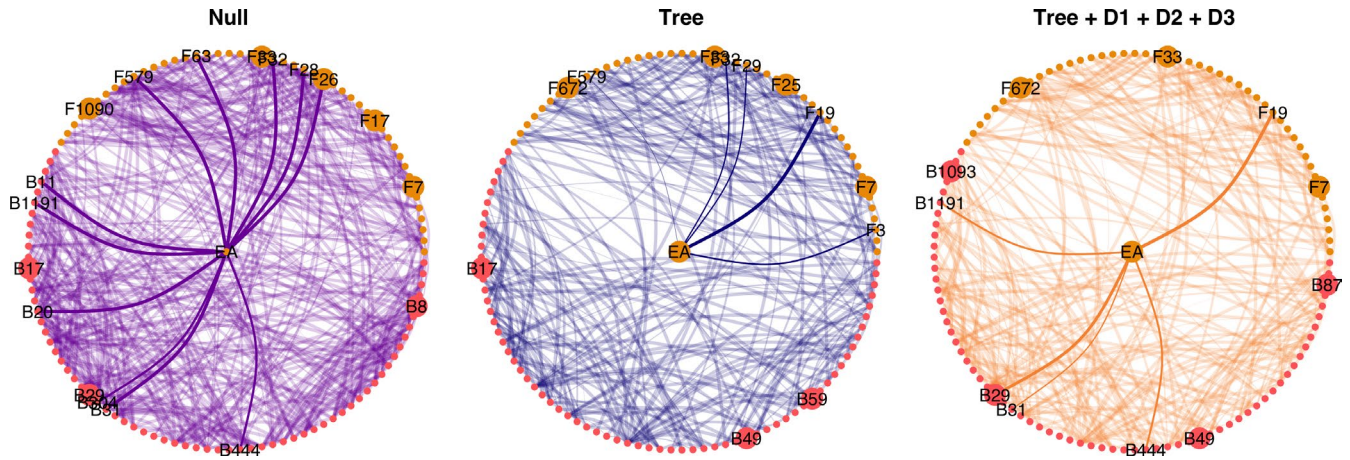


FIGURE 9 Pathogen interaction networks on oak leaves inferred with EMtree when adjusting for none, the *tree* covariate or *tree* and distances. Bigger nodes represent OTUs with highest betweenness values, colors differentiate fungal and bacterial OTUs. Widths are proportional to selection frequencies. $S = 100$, $f' = 90\%$

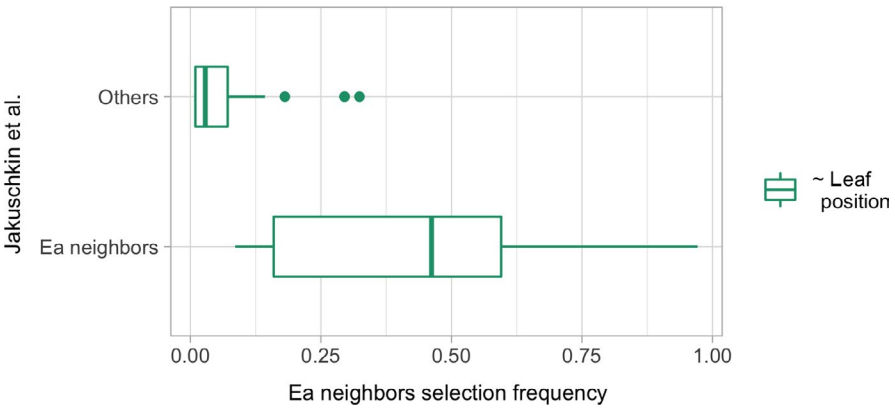


FIGURE 10 EMtree selection frequencies of pathogen neighbors compared to Jakuschkin et al. (2016) results, computed on infected samples and adjusting for the leaf position (100 subs-samples)

4 | DISCUSSION

The inference of species interaction network is a challenging task, for which a series of methods have been proposed in the past years. Abundance data seem to be a promising source of information for this purpose. Here we adopt the formalism of graphical models to define a probabilistic model-based framework for the inference of such networks from abundance data. Using a model-based approach offers several important advantages. First, it enables easy and explicit integration of environmental and experimental effects. These could be modelled in a more flexible way using generalized additive models, which include non-linear effects (Hastie, 2017) generalized. Then, as it also relies on a formal statistical definition of a *species interaction network* in the context of graphical models, accounting for abiotic effects and modelling species interactions are two clearly defined and distinguished goals. Finally, all the underlying assumptions are explicitly stated in the model definition itself, and can therefore be discussed and criticized.

We developed an efficient method to infer sparse networks, which combines a multivariate Poisson mixed model for the joint distribution of abundances, with an averaging over all spanning trees to efficiently infer direct species interactions. As we do consider a mixture over all spanning trees, our approach remains flexible and can infer most types of statistical dependencies. An EM algorithm (EMtree) maximizes the likelihood of the result and returns each edge probability to be part of the network. An optional resampling step increases network robustness.

A simulation study in a heterogeneous environment demonstrates that EMtree compares very well to alternative approaches. The proposed model can take all kind of covariates into account, which when ignored can have dramatic effects on the inferred network structure, as showed here on empirical datasets. Experiments on simulated data and illustrations also demonstrate that EMtree computational cost remains very reasonable.

Alternative methods used in this work all rely on an optimized threshold to tell an edge presence. This particular threshold is obtained after testing a grid of possible values which all yield a different network, and altogether build a path. Making this path available to the user is useful, as the final threshold might need modification and it gives the possibility to build edges scores and get more than a binary result. We found few recent approaches doing this, which

prevented us to study their performance in a way that did not impose a threshold.

The proposed methodology could be extended in several ways. Species abundances and interactions indeed vary across space, and depend on local conditions (Poisot, Canard, Mouillot, Mouquet, & Gravel, 2012; Poisot, Stouffer, & Gravel, 2015). This can either be considered as nuisance parameter or as feature of interest. In the first case, the method could be extended to account for the spatial autocorrelation of sampling sites, to obtain a 'regional' interaction network corrected for this effect, that is, assuming the network is the same in all sites. If of interest, variation across space and local conditions could be studied by comparing networks inferred from the different sampling locations. Networks comparison is a wide and interesting question and tools lack to check which edges are shared by a set of networks. The approach introduced by Schwaller and Robin (2017) could be adapted to EMtree framework. Lastly, it is also very likely that not all covariates nor even all species have been measured or observed. Another extension may therefore be to detect ignored covariates or missing species. To this purpose EMtree could probably be combined with the approach developed by Robin, Ambroise, and Robin (2019) to identify missing actors.

ACKNOWLEDGEMENTS

The authors thank P. Gloaguen and M. Authier for helpful discussions and C. Vacher for providing the oak dataset. This work is partly funded by ANR-17-CE32-0011 NGB and by a public grant as part of the Investissement d'avenir project, reference ANR-11-LABX-0056-LMH, LabEx LMH.

AUTHORS' CONTRIBUTIONS

All authors conceived the ideas and designed methodology; R.M. developed and tested the algorithm. All authors led the writing of the manuscript, contributed critically to the drafts and gave final approval for publication.

DATA AVAILABILITY STATEMENT

The method developed in this paper is implemented in the R package EMtree available on GitHub: https://github.com/Rmomal/EMtree (https://doi.org/10.5281/zenodo.3660627 (Rmomal, 2020a)). Simulation code is available at https://github.com/Rmomal/MEE_supplement_code (https://doi.org/10.5281/zenodo.3662705 (Rmomal, 2020b)).

ORCID

Raphaëlle Momal https://orcid.org/0000-0002-1550-4530
Stéphane Robin https://orcid.org/0000-0003-1045-069X

REFERENCES

Aitchison, J., & Ho, C. (1989). The multivariate Poisson-log normal distribution. *Biometrika*, *76*, 643–653. https://doi.org/10.1093/biomet/76.4.643

Austin, M. (2007). Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling*, *200*, 1–19. https://doi.org/10.1016/j.ecolmodel.2006.07.005

Baran, E. (1995). *Dynamique spatio-temporelle des peuplements de poissons estuariens en Guinée*. PhD thesis, Universite de Bretagne Occidentale.

Bartomeus, I., Gravel, D., Tylianakis, J. M., Aizen, M. A., Dickie, I. A., & Bernard-Verdier, M. (2016). A common framework for identifying linkage rules across different types of interactions. *Functional Ecology*, *30*, 1894–1903. https://doi.org/10.1111/1365-2435.12666

Biswas, S., McDonald, M., Lundberg, D. S., Dangl, J. L., & Jojic, V. (2016). Learning microbial interaction networks from metagenomic count data. *Journal of Computational Biology*, *23*, 526–535. https://doi.org/10.1089/cmb.2016.0061

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*, 859–877. https://doi.org/10.1080/01621459.2017.1285773

Chaiken, S., & Kleitman, D. J. (1978). Matrix tree theorems. *Journal of Combinatorial Theory, Series A*, *24*, 377–381. https://doi.org/10.1016/0097-3165(78)90067-5

Chiquet, J., Mariadassou, M., & Robin, S. (2018). Variational inference for probabilistic Poisson PCA. *The Annals of Applied Statistics*, *12*, 2674–2698. https://doi.org/10.1214/18-aoas1177

Chow, C., & Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, *14*, 462–467. https://doi.org/10.1109/tit.1968.1054142

Clark, N. J., Wells, K., & Lindberg, O. (2018). Unravelling changing interspecific interactions across environmental gradients using Markov random fields. *Ecology*, *99*, 1277–1283. https://doi.org/10.1002/ecy.2221

Dallas, T., Park, A. W., & Drake, J. M. (2017). Predicting cryptic links in host-parasite networks. *PLoS Computational Biology*, *13*, e1005557. https://doi.org/10.1371/journal.pcbi.1005557

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*, 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Desjardins-Proulx, P., Laigle, I., Poisot, T., & Gravel, D. (2017). Ecological interactions and the Netflix problem. *PeerJ*, *5*, e3644. https://doi.org/10.7717/peerj.3644

Dormann, C. F., Bobrowski, M., Dehling, D. M., Harris, D. J., Hartig, F., Lischke, H., … Kraan, C. (2018). Biotic interactions in species distribution modelling: 10 questions to guide interpretation and avoid false conclusions. *Global Ecology and Biogeography*, *27*, 1004–1016. https://doi.org/10.1111/geb.12759

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Evans, D. M., Kitson, J. J., Lunt, D. H., Straw, N. A., & Pocock, M. J. (2016). Merging DNA metabarcoding and ecological network analysis to understand and build resilient terrestrial ecosystems. *Functional Ecology*, *30*, 1904–1916. https://doi.org/10.1111/1365-2435.12659

Fang, H., Huang, C., Zhao, H., & Deng, M. (2017). gCoda: Conditional dependence network inference for compositional data. *Journal of Computational Biology*, *24*, 699–708. https://doi.org/10.1089/cmb.2017.0054

Faust, K., & Raes, J. (2012). Microbial interactions: From networks to models. *Nature Reviews Microbiology*, *10*, 538–550. https://doi.org/10.1038/nrmicro2832

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, *1*, 215–239. https://doi.org/10.1016/0378-8733(78)90021-7

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*, 432–441. https://doi.org/10.1093/biostatistics/kxm045

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*, *303*, 799–805. https://doi.org/10.1126/science.1094068

Graham, C. H., & Weinstein, B. G. (2018). Towards a predictive model of species interaction beta diversity. *Ecology Letters*, *21*, 1299–1310. https://doi.org/10.1111/ele.13084

Harris, D. J. (2015). Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution*, *6*, 465–473. https://doi.org/10.1111/2041-210X.12332

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S* (pp. 249–307). Abingdon-on-Thames, UK: Routledge.

Jakuschkin, B., Fievet, V., Schwaller, L., Fort, T., Robin, C., & Vacher, C. (2016). Deciphering the pathobiome: Intra- and interkingdom interactions involving the pathogen *Erysiphe alphitoides*. *Microbial Ecology*, *72*, 870–880. https://doi.org/10.1007/s00248-016-0777-x

Jordano, P. (2016). Sampling networks of ecological interactions. *Functional Ecology*, *30*(1881–1893), 1883. https://doi.org/10.1111/1365-2435.12763

Kirshner, S. (2008). Learning with tree-averaged densities and distributions. *Advances in Neural Information Processing Systems*, *20*, 761–768.

Kurtz, Z. D., Muller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., & Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Computational Biology*, *11*, e1004226. https://doi.org/10.1371/journal.pcbi.1004226

Lauritzen, S. L. (1996). *Graphical models*. Oxford Statistical Science Series. Oxford, UK: Clarendon Press.

Liu, H., Roeder, K., & Wasserman, L. (2010). Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models. *Advances in Neural Information Processing Systems*, *24*(2), 1432.

Meilă, M., & Jaakkola, T. (2006). Tractable Bayesian learning of tree belief networks. *Statistics and Computing*, *16*, 77–92. https://doi.org/10.1023/a:1008932416310

Morueta-Holme, N., Blonder, B., Sandel, B., McGill, B. J., Peet, R. K., Ott, J. E., … Svenning, J.-C. (2016). A network approach for inferring species associations from co-occurrence data. *Ecography*, *39*, 1139–1150. https://doi.org/10.1111/ecog.01892

Olito, C., & Fox, J. W. (2015). Species traits and abundances predict metrics of plant-pollinator network structure, but not pairwise interactions. *Oikos*, *124*, 428–436. https://doi.org/10.1111/oik.01439

Ormerod, J. T., & Wand, M. P. (2010). Explaining variational approximations. *The American Statistician*, *64*, 140–153. https://doi.org/10.1198/tast.2010.09058

Ovaskainen, O., Hottola, J., & Siitonen, J. (2010). Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology*, *91*, 2514–2521. https://doi.org/10.1890/10-0173.1

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., … Abrego, N. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters*, *20*, 561–576. https://doi.org/10.1111/ele.12757

Poisot, T., Canard, E., Mouillot, D., Mouquet, N., & Gravel, D. (2012). The dissimilarity of species interaction networks. *Ecology Letters*, *15*, 1353–1361. https://doi.org/10.1111/ele.12002

Poisot, T., Stouffer, D. B., & Gravel, D. (2015). Beyond species: Why ecological interaction networks vary through space and time. *Oikos*, *124*, 243–251. https://doi.org/10.1111/oik.01719

Poisot, T., Stouffer, D. B., & Kéfi, S. (2016). Describe, understand and predict: Why do we need networks in ecology? *Functional Ecology*, *30*, 1878. https://doi.org/10.1111/1365-2435.12799

Popovic, G. C., Hui, F. K., & Warton, D. I. (2018). A general algorithm for covariance modeling of discrete data. *Journal of Multivariate Analysis*, *165*, 86–100. https://doi.org/10.1016/j.jmva.2017.12.002

Popovic, G. C., Warton, D. I., Thomson, F. J., Hui, F. K. C., & Moles, A. T. (2019). Untangling direct species associations from indirect mediator species effects with graphical models. *Methods in Ecology and Evolution*, *10*, 1571–1583. https://doi.org/10.1111/2041-210x.13247

Richards, S. A. (2008). Dealing with overdispersed count data in applied ecology. *Journal of Applied Ecology*, *45*, 218–227. https://doi.org/10.1111/j.1365-2664.2007.01377.x

Rmomal. (2020a). Data from: Rmomal/EMtree: EMtree (Version v1.0). *Zenodo*, https://doi.org/10.5281/zenodo.3660627

Rmomal. (2020b). Data from: Rmomal/MEE_supplement_code: Supplementary Code Information (Version v1.0). *Zenodo*, https://doi.org/10.5281/zenodo.3662705

Robin, G., Ambroise, C., & Robin, S. (2019). Incomplete graphical model inference via latent tree aggregation. *Statistical Modelling*, *19*, 545–568. https://doi.org/10.1177/1471082x18786289

Schwaller, L., & Robin, S. (2017). Exact Bayesian inference for offline change-point detection in tree-structured graphical models. *Statistics and Computing*, *27*, 1331–1345. https://doi.org/10.1007/s11222-016-9689-3

Schwaller, L., Robin, S., & Stumpf, M. (2019). Closed-form Bayesian inference of graphical model structures by averaging over trees. *Journal of the French Statistical Society*, *160*, 1.

Stock, M., Poisot, T., Waegeman, W., & De Baets, B. (2017). Linear filtering reveals false negatives in species interaction data. *Scientific Reports*, *7*, 45908. https://doi.org/10.1038/srep45908

Ulrich, W., & Gotelli, N. J. (2010). Null model analysis of species associations using abundance data. *Ecology*, *91*, 3384–3397. https://doi.org/10.1890/09-2157.1

Valiente-Banuet, A., Aizen, M. A., Alcántara, J. M., Arroyo, J., Cocucci, A., Galetti, M., … Zamora, R. (2015). Beyond species loss: The extinction of ecological interactions in a changing world. *Functional Ecology*, *29*, 299. https://doi.org/10.1111/1365-2435.12356

Warton, D. I., Blanchet, F. G., O'Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C., & Hui, F. K. (2015). So many variables: Joint modeling in community ecology. *Trends in Ecology & Evolution*, *30*, 766–779. https://doi.org/10.1016/j.tree.2015.09.007

Weinstein, B. G., & Graham, C. H. (2017). Persistent bill and corolla matching despite shifting temporal resources in tropical hummingbird–plant interactions. *Ecology Letters*, *20*, 326–335. https://doi.org/10.1111/ele.12730

Wells, K., & O'Hara, R. B. (2013). Species interactions: Estimating per-individual interaction strength and covariates before simplifying data into per-species ecological networks. *Methods in Ecology and Evolution*, *4*, 1–8. https://doi.org/10.1111/j.2041-210x.2012.00249.x

Zhao, T., Liu, H., Roeder, K., Lafferty, J., & Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in R. *Journal of Machine Learning Research*, *13*, 1059–1062.

Zhu, Y., & Cribben, I. (2018). Sparse graphical models for functional connectivity networks: Best methods and the autocorrelation issue. *Brain Connectivity*, *8*(3), 139–165. https://doi.org/10.1089/brain.2017.0511

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Momal R, Robin S, Ambroise C. Tree-based inference of species interaction networks from abundance data. *Methods Ecol Evol*. 2020;11:621–632. https://doi.org/10.1111/2041-210X.13380