Quantifying the overall effect of biotic interactions on species distributions along environmental gradients

Marc Ohlmann, Catherine Matias, Giovanni Poggiato, Stéphane Dray, Wilfried Thuiller, Vincent Miele

Quantifying the overall effect of biotic interactions on species distributions along environmental gradients

Marc Ohlmann[1,*], Catherine Matias[2], Giovanni Poggiato[1,3], Stéphane Dray[4], Wilfried Thuiller[1], and Vincent Miele[4]

[1]Univ. Grenoble Alpes, CNRS, Univ. Savoie Mont Blanc, LECA, Laboratoire d'Ecologie Alpine, F-38000 Grenoble, France
[2]Sorbonne Université, Université Paris Cité, Centre National de la Recherche Scientifique, Laboratoire de Probabilités, Statistique et Modélisation, F-75005 Paris, France
[3]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, F-38000 Grenoble, France
[4]Université de Lyon, F-69000 Lyon; Université Lyon 1; CNRS, UMR5558, Laboratoire de Biométrie et Biologie Évolutive, F-69622 Villeurbanne, France
[*]Corresponding author: marcohlmann@live.fr

Open Research statement: ELGRIN is implemented in the function `elgrin` of the `R` package `econetwork` available on the code repository `https://plmlab.math.cnrs.fr/econetproject/econetwork` and at `CRAN` (`https://cran.r-project.org/`). The simulation procedure can be reproduced with the `R` code available along with this manuscript. The vertebrate data from O'Connor *et al.* (2020) can be found at `https://datadryad.org/stash/dataset/doi:10.5061/dryad.bcc2fqz79`. Climatic data were downloaded from the Worldclim v2 database (`http://www.worldclim.org/bioclim`) as described in the Methods, section 2.4. Land cover data were downloaded from Global Land cover v2.2 (`http://due.esrin.esa.int/page_globcover.php`), net primary productivity was downloaded from (`https://sedac.ciesin.columbia.edu/data/set/hanpp-net-primary-productivity/data-download`) and the human footprint index was downloaded from `http://sedac.ciesin.columbia.edu/data/set/wildareas-v2-human-footprint-geographic`, searching for the latest version (V2 the time of the article).

Key-words: biodiversity patterns, C-score, environmental niche, Markov random fields, metanetwork, species co-occurrence.

Abstract

Separating environmental effects from those of interspecific interactions on species distributions has always been a central objective of community ecology. Despite years of effort in analysing patterns of species co-occurrences and the developments of sophisticated tools, we are still unable to address this major objective. A key reason is that the wealth of ecological knowledge is not sufficiently harnessed in current statistical models, notably the knowledge on interspecific interactions.

Here, we develop ELGRIN, a statistical model that simultaneously combines knowledge on interspecific interactions (*i.e.*, the metanetwork), environmental data and species occurrences to tease apart their relative effects on species distributions. Instead of focusing on single effects of pairwise species interactions, which have little sense in complex communities, ELGRIN contrasts the overall effect of species interactions to that of the environment.

Using various simulated and empirical data, we demonstrate the suitability of ELGRIN to address the objectives for various types of interspecific interactions like mutualism, competition and trophic interactions. We then apply the model on vertebrate trophic networks in the European Alps to map the effect of biotic interactions on species distributions.

Data on ecological networks are everyday increasing and we believe the time is ripe to mobilize these data to better understand biodiversity patterns. ELGRIN provides this opportunity to unravel how interspecific interactions actually influence species distributions.

Introduction

Ecologists have always strived to understand the drivers of biodiversity patterns with the particular interest to tease apart the effects of environment and biotic interactions on species distributions and communities (Ricklefs, 2008; Thuiller *et al.*, 2015; Chase & Leibold, 2003; de Candolle, 1855). Species distributions are influenced by the abiotic environment (e.g. climate or soil properties) because of their own physiological constraints that allow them or not to sustain viable populations in specific environmental configurations (Austin, 2002; Pulliam, 2000). However, the occurrence of a species in a given site is also influenced by other species through all sort of interactions that can be trophic (e.g. a predator needs preys), non-trophic (e.g. plant species need to be pollinated by insects) or competitive (two species with the same requirements might exclude each other) (Guisan *et al.*, 2017; Gravel *et al.*, 2019; Lortie *et al.*, 2004; Soberón & Nakamura, 2009).

Teasing apart the effects of environmental variations and interspecific interactions on species distributions and communities from observed co-occurrence patterns has always been a hot topic in ecology since the earlier debate between Diamond (1975) and Connor & Simberloff (1979), to the recent syntheses on the subject (Blanchet *et al.*, 2020). More than anything, with a few exceptions, and despite recent advances like joint species distribution models (Ovaskainen *et al.*, 2017) or null model developments (Peres-Neto *et al.*, 2001; Chalmandrier *et al.*, 2013), the conclusion has been that it is almost impossible to retrieve and estimate interspecific interactions from observed spatial patterns of species communities (Zurell *et al.*, 2018; Blanchet *et al.*, 2020). This conclusion should thus preclude any attempt to disentangle the relative effects of environment and interspecific interactions. A major difficulty of this long-standing issue is that interspecific interactions could be of any type (i.e. positive, negative, asymmetric) and that observed patterns average out all these interactions. Observed communities indeed reflect the overall outcome of interspecific interactions that is difficult to dissect, especially when analysing pairwise species spatial associations as it is commonly done (e.g., Tikhonov *et al.*, 2017) . Yet, this overall outcome might be worth analysing on its own, for instance to measure the overall strength of interspecific interactions in a given community and between communities, how it depends on the co-existing species, and how it varies in space.

Interestingly, so far there have been few attempts to integrate the wealth of existing knowledge to address this fundamental ecological issue (Blanchet *et al.*, 2020; Holt, 2020). Indeed, the spatial analysis of biotic interactions is gaining an increased interest with novel technologies to measure interactions in the field (e.g. camera-traps, gut-content), open databases (e.g. GLOBI, Mangal) and the developments of new statistical tools to analyse them (Tylianakis & Morris, 2017; Pellissier *et al.*, 2018; Ohlmann *et al.*, 2019; Botella *et al.*, 2022). The combination of expert knowledge, literature, available databases, and phylogenetic hypotheses has also given rise to large metanetworks that generalise the regional species-pool of community ecology by incorporating the potential interactions between species from different trophic levels along with their functional and phylogenetic characteristics (Maiorano *et al.*, 2020; Morales-Castilla *et al.*, 2015). Despite a few attempts (e.g., Staniczenko *et al.*, 2017), information on interaction networks has been poorly integrated to understand and model biodiversity patterns. We believe that the time is ripe to incorporate network information into the process of modelling species distributions and communities. It implies to integrate both biotic and abiotic information (and their spatial variations) as explanatory factors in statistical models to weight their relative strength.

In this article, we propose a novel statistical model, called ELGRIN (in reference to Charles

Elton and Joseph Grinnell) that can handle the effects of both environmental factors and known interspecific interactions (aka a metanetwork) on species distributions. We rely on Markov random fields (MRF, also called Gibbs distribution, e.g., Brémaud, 1999), a family of flexible models that can handle dependencies between variables using a graph. More specifically, ELGRIN jointly models the presence and absence of all species in a given area in function of environmental covariates and the topological structure of the known metanetwork (Figure 1 left). It separates the interspecific interaction effects (Figure 1 top-right) from those of the environment (Figure 1 bottom-right) on species distributions. To our knowledge, ELGRIN is the first model whose outputs are the relative strengths of biotic factors needed on top of abiotic environmental variables to shape the species distributions and their spatial variation (see Latitude/Longitude in Figure 1 top-right). It thus provides a convenient way to integrate network ecology in joint species community modelling.

In this article, we first present the overall modelling framework and then assess its performances under different scenarios implying data simulated using three different dynamic models. In other words, although ELGRIN considers only static observational data (metaweb and community data), we evaluated the model using simulated data generated using different dynamic models that involve various underlying processes, including intraspecific competition. We test the ability of ELGRIN to decipher the relative importance of abiotic and interspecific interactions in these difficult cases so as to better understand what kind of signal ELGRIN can or cannot retrieve from the data. Finally, we apply the model on vertebrate trophic networks in the European Alps as an empirical study.

Material and methods

Species data and potential interactions

We consider a set of sites or locations indexed by $l \in \{1, \dots, L\}$, where the occurrence (presence/absence) of N species and a set of environmental variables (vector W_l) are observed.

For the same set of N species, we assume that we know all the pairwise interactions between them (e.g. who eats whom), an information summarised with a graph $G^\star = (V^\star, E^\star)$ over the set of nodes $V^\star = \{1, \dots, N\}$ and edges $E^\star$. This graph, usually called a metanetwork that represents a regional pool of both species and interactions, can be obtained, for instance, by aggregating local networks at different locations or from expert knowledge and literature review (e.g., Cirtwill *et al.*, 2019; Maiorano *et al.*, 2020). Note that various types of interactions can be considered here (e.g., trophic, mutualism, competition). However, while considering a mixture of interaction types is technically possible, the interpretation of results would be difficult because in our framework, $G^\star$ records the presence of an interaction and not its type. An additional note is that our model, like most species community models (e.g. Joint species distribution models, ordination techniques) relying on occurrence data, makes some assumptions about the ecological processes structuring species assemblages. In our current implementation of ELGRIN, we consider that only unimodal responses of species to environmental gradients and interspecific interactions shape communities, ignoring other processes such as dispersal limitation or mass effect for instance. Lastly, note also that our model supposes that the graph associated to the metanework is undirected with no self-loops (see model specifications below) and thus ignores intraspecific interactions. Hereafter, we

refer to co-present (or co-absent) species, pairs of species that are connected in the metanetwork and jointly present (or absent, respectively) at a given location.

The statistical model of ELGRIN

Model description The aim of ELGRIN model is to factorise the joint species presence distribution between a Grinnellian part, that consists in a regression on environmental covariables, and an Eltonian part that quantifies association strengths between species distribution according to the metanetwork. More formally, we consider a set of random variables $\{X_i^l\}_{i\in V^\star}$ taking values in $\{0,1\}$ and that represent the presence/absence of species $i \in V^\star$ at location $l \in \{1,\dots,L\}$. We rely on a *Markov random field* (see for instance Brémaud, 1999) to model the dependencies between species occurrences at location l. This is a multivariate model that encodes statistical dependencies between species distribution using a network. In our ELGRIN model, these dependencies are encoded through the metanetwork $G^\star$. For each location $l \in \{1,\dots,L\}$, we thus assume that these random variables are distributed according to a Gibbs distribution specifying the joint associations between the species occurrence variables $\{X_i^l\}_{i\in V^\star}$, as follows:

$$\mathbb{P}(\{X_i^l\}_{i\in V^\star}) = \frac{1}{Z}\exp\Big(\sum_{i\in V^\star}[a_l + a_i + W_l^\intercal b_i + (W_l^2)^\intercal c_i]X_i^l \tag{1a}$$

$$+ \beta_{l,co-pres}\sum_{(i,j)\in E^\star}\mathbf{1}\{X_j^l = X_i^l = 1\} \tag{1b}$$

$$+ \beta_{l,co-abs}\sum_{(i,j)\in E^\star}\mathbf{1}\{X_j^l = X_i^l = 0\}\Big), \tag{1c}$$

where $\mathbf{1}\{A\}$ is the indicator function of event A (either co-absence $X_j^l = X_i^l = 0$ or co-presence $X_j^l = X_i^l = 1$), notation $U^\intercal$ stands for the transpose of vector U and Z a normalising constant discussed below. Some model parameters have an ecological interpretation (Table 1). The use of W_l and W_l^2 (the vector of coordinate-wise squared values of W_l) allows modelling a quadratic species response to environmental gradient, following then a bell-shaped relationship as expected under classical niche theory (Chase & Leibold, 2003).

Sub-equation (1a) is the Grinnellian part of ELGRIN, as it represents some prior probability of species occurrences independently of their interactions. Parameters a_i, b_i, c_i capture the response of species i to environment, seen through a vector of environmental covariates W_l. The intercepts a_i and a_l are estimated up to a constant only (see Appendix S1: Section S.2.1) and may not be interpreted, whereas the vectors b_i, c_i deal with the species environmental niche, like in a standard species distribution model (Guisan *et al.*, 2017).

Sub-equations (1b) and (1c) form the Eltonian part of ELGRIN. It considers only interactions $(i,j) \in E^\star$, i.e. the edges of the metanetwork. The β_l represent the overall influence of the interactions (as encoded through $G^\star$) on all species presence/absence at location l. However, this influence may be different for co-presence and co-absence, with parameters $\beta_{l,co-pres}$ and $\beta_{l,co-abs}$ respectively (see Table 2). When a $\beta_{l,co-pres}$ is positive, it represents a positive driving force of co-presence on species distributions. By contrast, a negative value indicates that species co-presences are avoided. The same reasoning holds with $\beta_{l,co-abs}$ for co-absences. Since the interaction parameter $\beta_{l,co-abs}$ can also be influenced by co-absences between species that are both absent at

Variables	Ecological interpretation
$G^\star$	Metanetwork of known interactions (undirected)
$V^\star$	Species (node set) of the metanetwork
$E^\star$	Interactions (edge set) of the metanetwork
X_i^l	Presence/absence of species i at location l
W_l	Environmental covariates at location l
Parameters	
a_i	Species i intercept
a_l	Location l intercept
b_i, c_i	Environmental (abiotic) parameters of species i
$\beta_{l,co-pres}$	Co-presence strength (or avoidance when < 0) at location l
$\beta_{l,co-abs}$	Co-absence strength (or avoidance when < 0) at location l

Table 1: Definition of variables and parameters of the Markov random field model ELGRIN.

location l only because of unsuitable environmental conditions, we introduced a compatibility matrix so that the effect of interactions is only estimated in the environmental conditions where interacting species could co-occur (details are given in Appendix S1: Section S.2.2). Importantly, this compatibility matrix is estimated during the inference procedure and is not a required input by the user.

Note that we chose the parameters β_l to be specific to location $l \in \{1, \dots, L\}$ such that the effect of species interactions can vary across space. Finally, Z is a normalising constant that cannot be computed for combinatorial reasons, although the statistical inference procedure takes care of it. Full details of the estimation procedure and parameter identifiability are available in Appendix S1: Section S.3 and Appendix S1: Section S.2.1, respectively.

Lastly, it is important to note two specificities of the metanetwork $G^\star$ in our modelling procedure: it cannot be directed nor contain self-loops. Indeed, Markov random fields specify conditional dependencies between random variables $\{X_i^l\}$ in an undirected way, and self-loops have no meaning in this framework. Our model assumes that these dependencies are given by the interaction network without considering the direction of edges. Consequently, this statistical model of interaction cannot be read in the light of causality. In case of trophic interactions, it consists in assuming that presence/absence of a predator and its prey are intertwined, without specifying top-down or bottom-up control. Moreover, the absence of self-loops prevents from taking into account intraspecific effects. These effects are simply ignored by ELGRIN, as they are in any joint species distribution model or ordination technique (see Appendix S1: Section S.6).

ELGRIN is implemented in `C++` for efficiency and is available in the function `elgrin` of the `R` package `econetwork` available on the code repository `https://plmlab.math.cnrs.fr/econetproject/econetwork` and at `CRAN` (`https://cran.r-project.org/`). We assessed the performance of the method in inferring parameters from data sampled and resampled under the model (see Appendix S1: Section S.4).

Model interpretation In the hypothetical example where $G^\star$ is an empty graph (no edges, none of the species interact), the random variables $\{X_i^l\}_{i \in V^\star}$ are independent and each species is present with probability $e^{\alpha_{i,l}}/(1 + e^{\alpha_{i,l}}) \in (0, 1)$, where $\alpha_{i,l} = a_l + a_i + W_l^\intercal b_i + (W_l^2)^\intercal c_i$. In other

words, $\alpha_{i,l}$ is the logit of the probability of presence of species i at location l in the absence of interactions. Assuming that we have included all important environmental covariates, that there is no other ecological processes involved, and no model mis-specifications, $\alpha_{i,l}$ is analogous to the fundamental niche parameters of the species (sensu Hutchinson, 1959). It gives the probability of presence of species i at location l when only environmental filtering occurs.

In the case of species interactions, $G^\star$ is a non empty graph and the presence/absence information is smoothed across neighbouring nodes in $G^\star$. In Table 2, we detailed the ways both $\beta_{l,co-pres}$ and $\beta_{l,co-abs}$ parameters capture how the metanetwork influences species co-occurrences in a given location, notably the co-presence or co-absence of pairs of interacting species. This table describes expected patterns of species distribution according to the combination of positive, negative and zero values for the β parameters. More precisely, when species are known to interact positively (e.g. $G^\star$ encodes mutualism) and that these interactions, averaged over all species with suitable environmental conditions at location l, influence their co-occurrences at that location, $\beta_{l,co-pres}$ and/or $\beta_{l,co-abs}$ will be estimated as positive. On the other hand, in case of negative interactions (e.g. $G^\star$ encodes competition) that influence the co-occurrences at location l of species with favorable environmental conditions, the parameters $\beta_{l,co-pres}$ and/or $\beta_{l,co-abs}$ will be negative, co-presence configurations (or co-absence, respectively) tend to be avoided, meaning that only one of the two species tends to be present. Given a location with fixed total number of interacting co-present (resp. interacting co-absent) species, the larger the absolute value of $\beta_{l,co-pres}$ (resp. $\beta_{l,co-abs}$), the stronger the strength of the interactions.

Exploration on simulated data from complex dynamic processes

To test the ability of ELGRIN to infer the overall biotic and abiotic controls on species distributions, we used three theoretical models, different from the one underlying ELGRIN, to dynamically simulate spatial community data with 50 species and 400 sites along a single environmental gradient and combined them with multiple different interactions scenarios (competition, mutualism, and no interaction). To do that, we chose species niche optima evenly distributed along a single environmental gradient. The metanetworks were built so that interacting species have close niche optima (otherwise they would never co-occur). In the mutualistic scenario, we also considered a case where species that facilitate each other tend to have an abiotic niche that is also not too close (otherwise they would compete). Along this single environmental gradient, niche optima and associated metanetworks according the interaction scenarios, we used three theoretical dynamic models (Lotka-Volterra, colonisation-extinction, and co-existence model aka VirtualCom) to simulate the resulting species distribution data. These models have different underlying assumptions and processes, which allowed testing ELGRIN under a total of 9 different configurations.

Lotka-Volterra model The Lotka-Volterra model is one of the foundational models in community ecology (Takeuchi, 1996). This model simulates communities under both intra- and interspecific interactions, while ELGRIN is not able to handle intraspecific interactions (its metanetwork does not allow for self-loops). Thus we parameterized the Lotka-Volterra simulation with intraspecific interactions being negligible in regards to interspecific interactions. That way we generated species community data that meets the type of data and ecological questions ELGRIN is designed to tackle (for details, see Appendix S1: Section S.5.1). Nonetheless, we also explored the converse

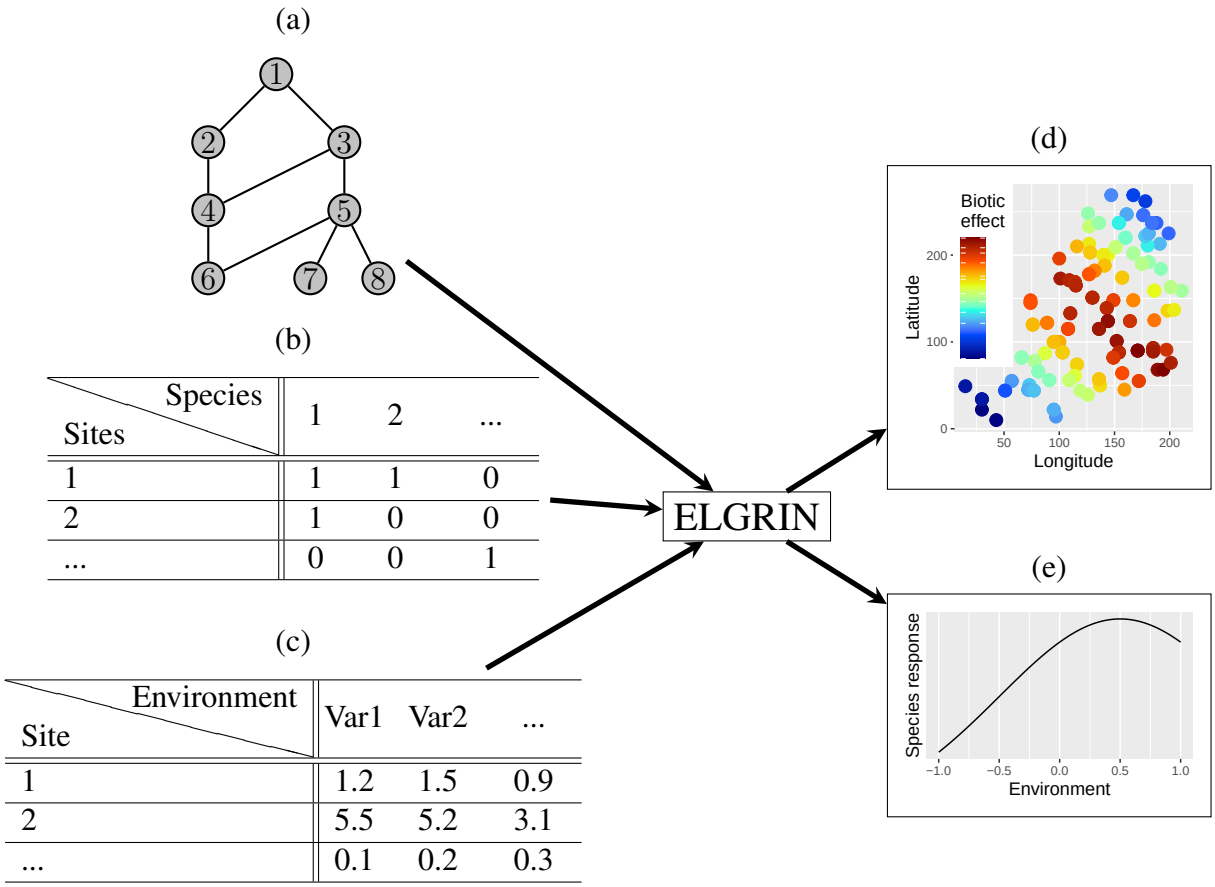


Figure 1: Schematic view of ELGRIN statistical model. Given (a) an interaction metanetwork that summarises known interactions (edges) between species (nodes), (b) species occurrences data and (c) environmental covariates for a set of sites, ELGRIN model estimates (d) the overall effect of known biotic interactions on species distributions in each site using two association parameters, and (e) the environmental response of each species along all sites using regression parameters on environmental covariates.

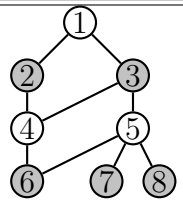
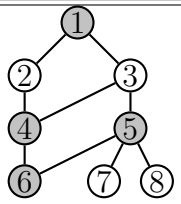
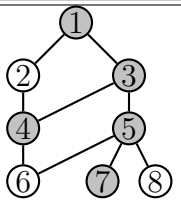
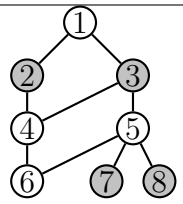
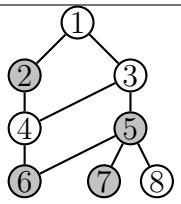
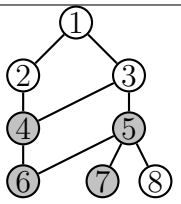
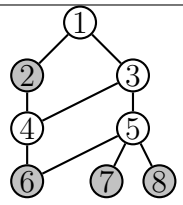
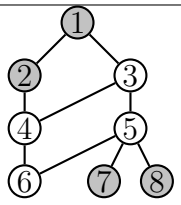
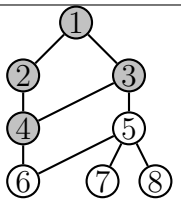
	$\beta_{l,co-pres} \ll 0$ (avoided co-presence)	$\beta_{l,co-pres} = 0$ (random presence)	$\beta_{l,co-pres} \gg 0$ (favored co-presence)
$\beta_{l,co-abs} \ll 0$ (avoided co-absence)			
$\beta_{l,co-abs} = 0$ (random absence)			
$\beta_{l,co-abs} \gg 0$ (favored co-absence)			

Table 2: Simplified view of the different behaviours of the model in function of the parameters $\beta_{l,co-pres}$ and $\beta_{l,co-abs}$. The graph represents the metanetwork containing all potential interactions where species can be either present (gray node) or absent (white node) in a given location l leading to different estimated $\beta_{l,co-pres}$ and $\beta_{l,co-abs}$. When $\beta_{l,co-pres} \ll 0$ or $\beta_{l,co-abs} \ll 0$, interacting species in the metanetwork tend to avoid each other: whenever one is absent, the other tend to be present and reversely. This situation favors a checkerboard pattern on the metanetwork. Reversely, whenever $\beta_{l,co-pres} \gg 0$ (resp. $\beta_{l,co-abs} \gg 0$), there are groups of interacting species that tend to be all present (resp. all absent), inducing sets of gray (resp. white) neighbour nodes in the metanetwork. Whenever $\beta_{l,co-pres} = 0$ or $\beta_{l,co-abs} = 0$, there are sets of interacting species whose states are independent from one another and thus purely random (the proportions of gray and white nodes are governed by the values of the parameters in the Grinnellian part of the model).

case to fully understand the limits of ELGRIN (see Appendix S1: Section S.6).

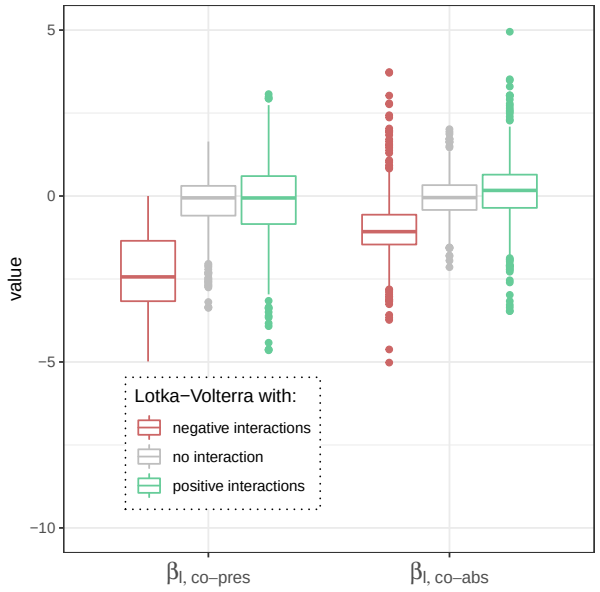


Figure 2: Distribution of co-presence ($\beta_{l,co-pres}$) and co-absence ($\beta_{l,co-abs}$) strengths inferred using ELGRIN on simulated ecological communities using a Lotka-Volterra model with competition (negative interactions), mutualism (positive interactions) or no interactions.

Colonisation-extinction model We used an updated version of the stochastic colonisation-extinction model developed in Ohlmann et al. (2022) to simulate the species community dataset for the three interaction scenarios (for details see Appendix S1: Section S.5.2). The model consists in a multivariate Markov chain that converges towards a stationary distribution from which we sampled the species community dataset.

VirtualCom model We used an updated version of the model developed by Münkemüller & Gallien (2015) to simulate communities whose composition is driven simultaneously by biotic and abiotic environmental effects, for the three interaction scenarios (for details see Appendix S1: Section S.5.3). In this model, each community has the same carrying capacity (i.e. the exact number of individuals in each location).

Application: a case study

We analyse the newly available Tetra-EU 1.0 database, a species-level trophic network of European tetrapods (Maiorano *et al.*, 2020) that combines all known potential interactions between terrestrial mammals, birds, reptiles and amphibians occurring in Europe. This metanetwork is based on data extracted from known interactions, scientific literature, including published articles, books, and grey literature (see Maiorano *et al.*, 2020, for a complete description of the data

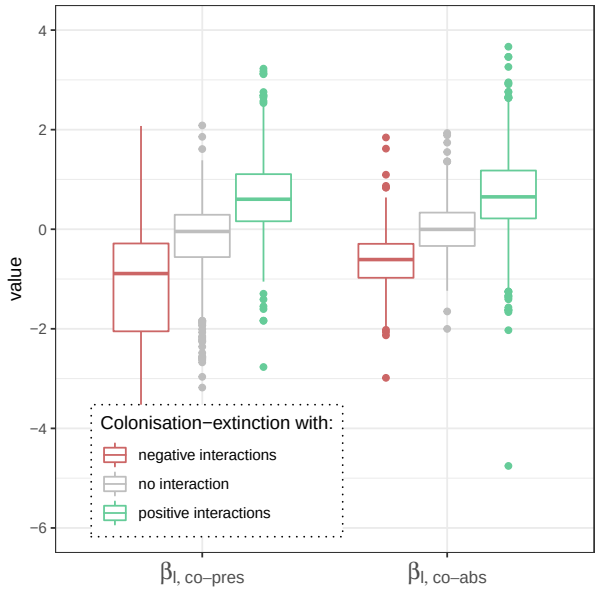


Figure 3: Distribution of co-presence ($\beta_{l,co-pres}$) and co-absence ($\beta_{l,co-abs}$) strengths inferred using ELGRIN on simulated ecological communities using a colonisation-extinction model with competition (negative interactions), mutualism (positive interactions) or no interactions.

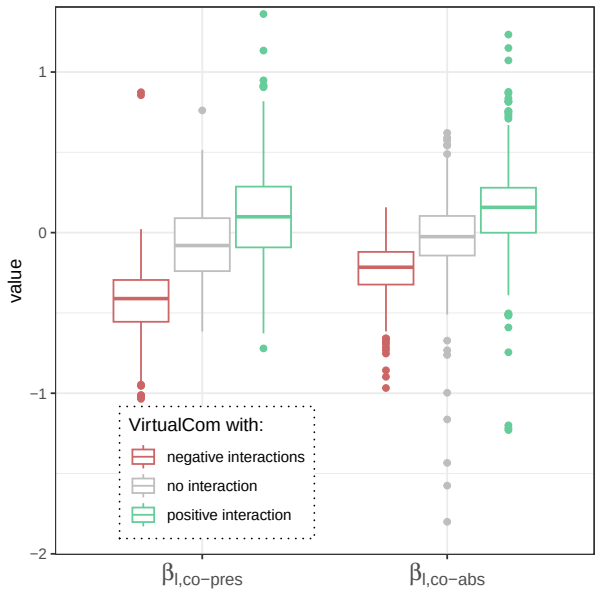


Figure 4: Distribution of co-presence ($\beta_{l,co-pres}$) and co-absence ($\beta_{l,co-abs}$) strengths inferred using ELGRIN on simulated ecological communities with VirtualCom model, with competition (negative interactions), mutualism (positive interactions) or no interactions.

and the reference list used to build the metanetwork). As usual with such data, this metanetwork does not provide information on interaction plasticity or intraspecific interactions. We re-

stricted our analyses on the European Alps that show sharp environmental gradients and varying trophic web distributions (O'Connor *et al.*, 2020). We extracted the species distribution data from Maiorano *et al.* (2013) at a 300 m resolution. We upscaled all species ranges maps to a 10x10 km equal-size area grid and cropped the distribution data to the European Alps. Species were considered present on a given 10x10 km cell if they were present in at least one of the 300 x 300 m cells within it. This yielded species distributions maps for 257 breeding birds, 99 mammals, 36 reptiles, and 30 amphibians over 2138 locations. Environmental covariates were extracted at the same resolution and were selected following previous work on those data (Braga *et al.*, 2019). For climate, we used mean annual temperature, temperature seasonality, temperature annual range, total annual precipitation and coefficient of variation of precipitation that were all extracted from the Worldclim v2 database (`http://www.worldclim.org/bioclim`). Using GlobCover (GlobCover V2.2; `http://due.esrin.esa.int/page_globcover.php`), we extracted the number of habitats present in a given pixel, habitat diversity in a given pixel based on Simpson index and habitat evenness as a measure of habitat complexity. Finally, we added an index of annual net primary productivity (Global Patterns in Net Primary Productivity, v1 (1995), `http://sedac.ciesin.columbia.edu/data/set/hanpp-net-primary-productivity`) and the human footprint index (`http://sedac.ciesin.columbia.edu/data/set/wildareas-v2-human-footprint-geographic`). Since these data were highly correlated, we used a PCA to retain the three leading vectors as environmental covariates (W_l) in ELGRIN.

Results

Tests on simulated species community data

Let us first recall that we assessed the performance of the method in inferring parameters from data sampled and re-sampled under ELGRIN model (see Appendix S1: Section S.4). We now turn to more involved dynamical theoretical models.

For the three theoretical models (Lotka-Volterra, colonisation-extinction and VirtualCom), ELGRIN was correct in identifying the no interaction scenario, with estimated interaction strengths close to 0 (Figures 2, 3 and 4). Similarly, ELGRIN was able to retrieve the negative effects of interactions in the case of competition as simulated by the three theoretical models. The $\beta_{l,co-pres}$ and $\beta_{l,co-abs}$ parameters were mostly negative (with much higher absolute values for $\beta_{l,co-pres}$), capturing the backbone of the competitive interactions. They indicated that co-presence and co-absence were avoided (as presented in Table 2 top-left), leading to some level of competitive exclusion. In the VirtualCom co-existence model, this phenomenon was clearly the by-product of the competitive interactions and the carrying capacity in terms of number of individuals (that explicitly induced exclusion). When positive interactions come into play (i.e. mutualism), the results should be contrasted between those obtained for the Lotka-Volterra model, where ELGRIN does not qualitatively identify the processes at stake and the two other models (colonisation-extinction and VirtualCom) where ELGRIN succeeds in identifying them. The Lotka-Volterra simulation with positive interactions scenario produced species that are essentially distributed along their respective niches (see Appendix S1: Figure S.11). As a consequence, this distribution can be simply fitted with the Grinellian part of the model and ELGRIN estimates the βs close to zero (Figure 2). That means

that the same dataset could have been produced by only abiotic environmental conditions and the actual species distribution does not contain anymore a pattern that ELGRIN would identify as the trace of the positive interspecific interactions. On the contrary, in the positive interactions scenario, with both competition-colonisation and VirtualCom co-existence models, ELGRIN correctly identified the process at play. The parameters $\beta_{l,co-pres}$ and $\beta_{l,co-abs}$ were mostly positive. During the simulation steps, the presence of one species was then favored by the presence of another species it interacted with, leading to a co-presence phenomena captured by the positive $\beta_{l,co-pres}$. Conversely, the inverse mechanism emerged for co-absence, implying that the $\beta_{l,co-abs}$ tended to be positive as revealed by ELGRIN (Figures 3, 4). To quantitatively investigate the difference between $\beta_{l,co-pres}$ and $\beta_{l,co-abs}$ distributions in the three simualations, we performed Kolmogorov-Smirnov (KS) tests. For each simulation, we tested whether $\beta_{l,co-pres}$ and $\beta_{l,co-abs}$ distributions were significantly different in the scenarios with interactions (either positive or negative) from the scenario without interaction. In the three simulations, the tests correctly identify significant differences between interactions and no interaction scenarios (see Appendix S1: Table 3).

Empirical case study

When fitted to the European vertebrate dataset, ELGRIN's parameters $\beta_{l,co-pres}$ and $\beta_{l,co-abs}$ were highly correlated (Pearson correlation of 0.84, see Appendix S1: Section S.7.1) suggesting that trophic interactions impact both predator/prey co-presence and co-absence. In what follows, we therefore mainly dealt with $\beta_{l,co-pres}$.

We first observed a structured spatial pattern of the effects of interactions, with regions of negative or positive $\beta_{l,co-pres}$ (bluish or reddish colors respectively in Figure 5). The largest $\beta_{l,co-pres}$ values were found mainly in the french Alps and in the Eastern zone.

In Figure 6, we present the values of different variables at each location, according to groups of estimated $\beta_{l,co-pres}$ parameters, where the width of each boxplot is proportional to the number of points in each class. Almost all the highest $\beta_{l,co-pres}$ (> 0.05) were revealed in locations below 1600 m of altitude (Figure 6a, p-value of the KS test inferior to $2.2\mathrm{e}{-16}$, details given in Appendix S1: Section S.7.2). In these regions, species richness was generally high (Figure 6b, p-value inferior to $2.2\mathrm{e}{-16}$). In the opposite, the higher up, the more likely $\beta_{l,co-pres}$ was negative (Figure 6a). This was particularly true above 1600 m in the central Alps, where almost all the negative $\beta_{l,co-pres}$ were estimated (bluish colors in Figure 5). Locations with negative $\beta_{l,co-pres}$ have a lower species richness (Figure 6b). Interestingly, locations with low connectance have lower absolute $\beta_{l,co-pres}$ values (Figure 6c, p-value inferior to $2.2\mathrm{e}{-16}$) indicating a lower effect of biotic interactions compared to abiotic effects in these locations. Here, connectance is the density of the graph induced by the metanetwork at location l, namely its nodes are species occurring at location l and edges are those from the metanetwork between those present species.

Discussion

Deciphering the mechanisms driving spatial patterns of species distributions and communities is likely one of the most active fields of ecological research since the early days of biogeography and community ecology. Still, there was so far no comprehensive statistical approach able to make the best of existing knowledge on interspecific interactions, species occurrence and environmental

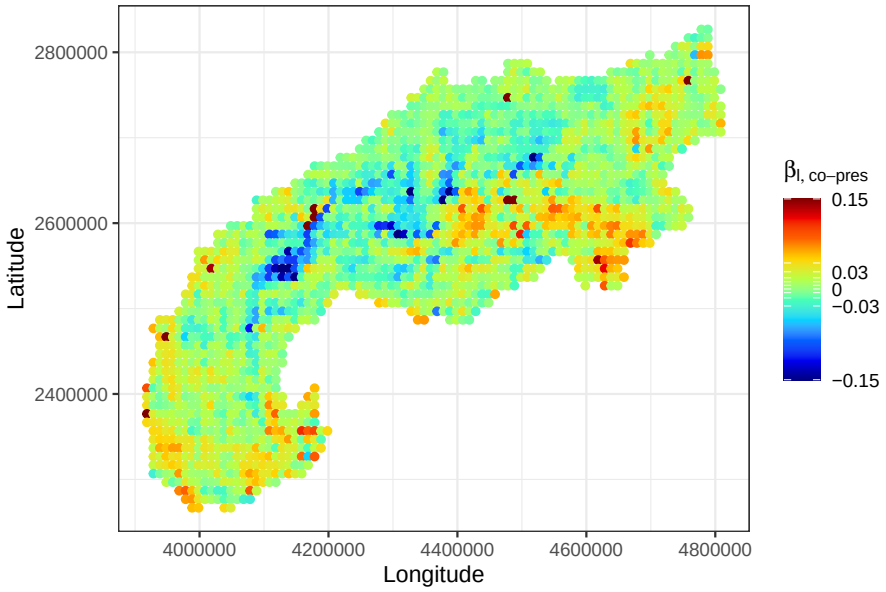


Figure 5: Results of ELGRIN on the European tetrapods case study. Map of estimated $\beta_{l,co-pres}$ (one dot per location). The color scale indicates the $\beta_{l,co-pres}$ values. For the sake of representation, $\beta_{l,co-pres}$ values above 0.15 in absolute value were set to 0.15.

data to measure and quantify the dual effects of environment and biotic interactions on species distributions. Our proposed model that relies on Markov random fields builds on the ability of graphical models to encode and analyse species distribution dependencies using the known species interactions. This formalism allows, within the same model, to account for both the effects of the environment and the interspecific interactions, which reconciles the Grinnellian vision of species niches (i.e. how species respond to the abiotic environment) with its Eltonian counterpart (i.e. how species respond to the biotic environment). The mathematical foundations of ELGRIN are strong and its framework is flexible allowing for useful extensions to handle interaction strength, sampling effects and plasticity of interactions (see Appendix S1: Section S.1).

A key element of ELGRIN is its ability to measure the overall relative effects of interspecific interactions on species distributions with respect to abiotic environmental conditions, which allows to summarise all local pairwise interactions in a single measure (i.e. $\beta_{l,co-abs}$ or $\beta_{l,co-pres}$). This measure can then be mapped, related to spatial layers to understand how the overall relative effect of interspecific interactions vary in space and in function of the environment or the ecosystem types. Importantly, this measure can also be carefully investigated at a given location in function of the constituent species, trophic groups, specialists vs generalists, connectance and so on. Interestingly, we can thus see our β_l estimates as an extended and more meaningful version of the famous checkerboard score or C-score (Stone & Roberts, 1990), which has been used to quantify local interspecific interactions from co-occurrence pattern (e.g., Boulangeat *et al.*, 2012). The main advantage of ELGRIN over the C-score is that instead of trying to infer biotic interactions only from co-occurrences (which we know to be notoriously difficult, nearly impossible), it quantifies, in a conditional way, the effects of the known interspecific interactions on species communities, while accounting for the environmental responses of the species. Our approach is

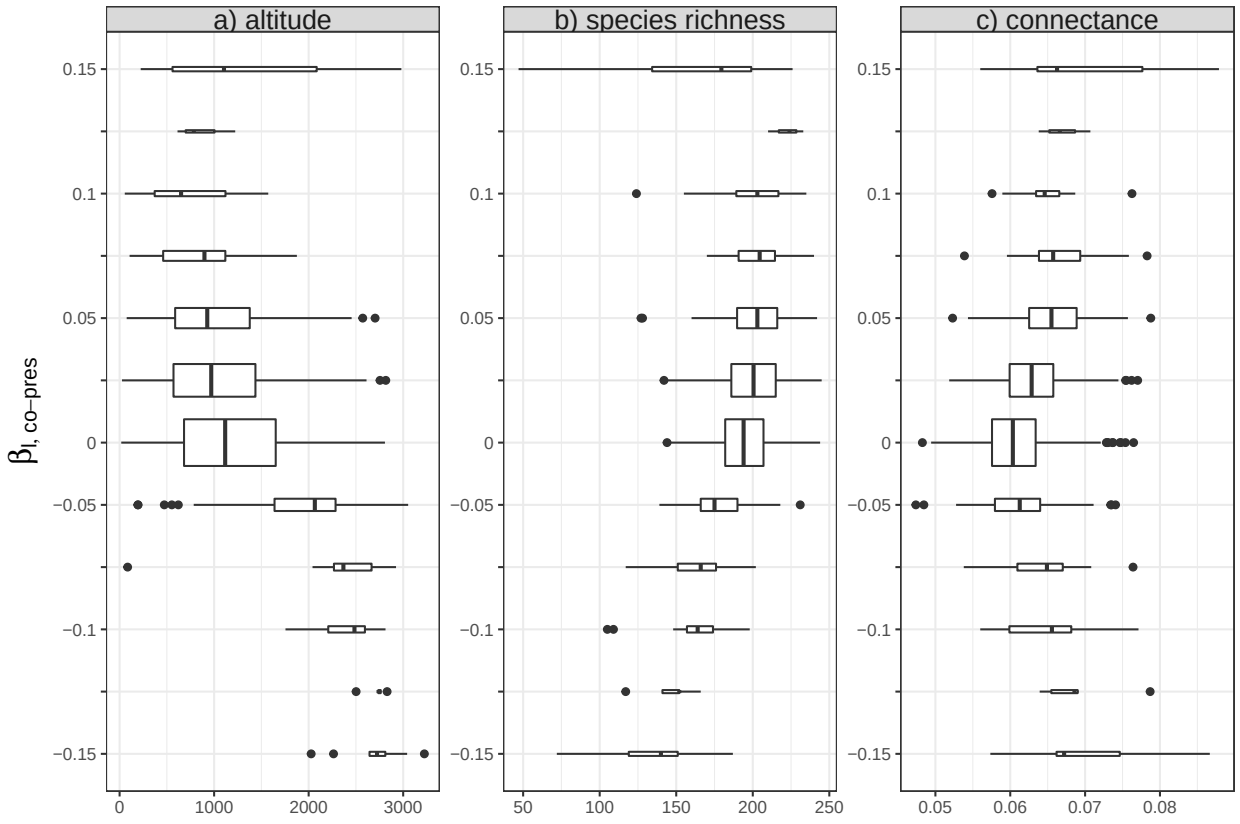


Figure 6: Results of ELGRIN on the European tetrapods case study. Boxplots representing the values of different variables at each location, according to the estimated $\beta_{l,co-pres}$ values (x axis). (a) altitude, (b) species richness, and (c) connectance (density of the graph induced by the metanetwork at location l) For the sake of representation, $\beta_{l,co-pres}$ values above 0.15 in absolute value were set to 0.15. Width of the boxplots is proportional to the number of points in each class.

thus not comparable with recent developments on joint species distribution models (JSDMs) that relate species occurrences to environmental conditions, and provides a residual covariance matrix that could be interpreted on the light of missing predictors, mis-specifications and biotic interactions (Ovaskainen *et al.*, 2017; Zurell *et al.*, 2018). This matrix represents covariances between model residuals (the left-over from the environmental effects) and actually provides little information about biotic interactions (Zurell *et al.*, 2018; Poggiato *et al.*, 2021). ELGRIN does not infer any residual covariance and directly accounts for the known interactions through the metanetwork. In JSDMs, missing covariates will inevitably lead to spurious estimates of biotic interactions. In ELGRIN, the parameter a_l is supposed to capture most of the unexplained information that is independent of the interspecific interactions. This parameter acts as a site random effect in mixed models and is expected to filter out the effects of missing covariates, although some remaining species-specific effects might still percolate into the β_l estimates.

In the presentation of ELGRIN and in our case studies, we focused on a single interaction type at a time (e.g. competition, mutualism or trophic interaction). When dealing with a single type

of interaction, competition for instance, the modelling is explicit since we clearly understand the effect that one species can have on another species. Although it is technically possible to manage a metanetwork composed of different types of interactions, the interpretation would become problematic. Different interaction types can have opposite effects, such as competition (a species excludes other species) and mutualism (a species facilitates other species) and, since ELGRIN captures an overall impact of these interactions on the distributions at each location, interpreting ELGRIN's results can be misleading in that case. Additionally, it is worth noting that since ELGRIN relies on a Markov random field, $G^\star$ is undirected. In other words, when the original metanetwork encodes asymmetric interactions (e.g. predator-prey), they are then converted in undirected edges that only represent the presence of interactions (whatever their direction). It is thus critical to keep that in mind when interpreting the results of ELGRIN, and when merging different types of interactions together. The same issue happens when hoping to interpret the residual covariance matrix of JSDM through the lens of biotic interactions, since the values of the covariance matrix could reflect any type of interactions between species, that could be asymmetric or symmetric, or both. Note that we explicitly used a bell-shaped relationship for modelling species response to environmental gradients. While it would be possible to modify ELGRIN to incorporate any other parametric relationship, the actual version of ELGRIN would lead to erroneous conclusions whenever used on data where this assumption is not satisfied.

More generally, it is important to underline that ELGRIN finds the most likely scenario under a model associated to underlying assumptions. This model represents up to date the most reasonable and simple model that integrates both interspecific interactions and abiotic factors in modelling the species distribution. In that sense, it goes beyond (joint) species distribution models or ordination models by including explicitly the effect of interspecific interactions. However, the most likely scenario under this model is not necessarily the real one that lead to observed data. For instance, ELGRIN was not able to identify the positive interspecific interactions present in the dynamics of a Lotka-Volterra model (even when restricting to negligible intraspecific interactions). Despite being a most widely studied model, the Lotka-Volterra model still raises important challenges. Indeed, whether the system reaches a single globally stable equilibrium point is known only in specific cases (Takeuchi, 1996). Since ELGRIN infers model interspecific interactions relative effects from the species distributions, existence of multiple equilibria in the Lotka-Voltera dynamics (depending on the initial conditions that are unknown) could pose serious identifiability problems. Even in presence of a unique and globally stable equilibrium point, several parameters or different interaction types could lead to the same equilibrium and thus same observed species distributions. This also raises tough identifiability issues. We hope that the recent developments around Lotka-Volterra model will help to circumvent those issues (Biroli *et al.*, 2018; Remien *et al.*, 2021). We could easily simulate species distributions, using models that include other ecological processes, on which ELGRIN would fail in recovering the true underlying generation processes. Indeed we present simulations scenarios beyond the assumptions of the model (i.e., a Lotka-Volterra model with intraspecific interactions stronger than interspecific ones, see Appendix S1: Section S.6), where ELGRIN again uncovered a completely different explanation of the data at hand. If the data contain the signature of different ecological processes (including ones not considered by ELGRIN), ELGRIN will not be able to infer properly the relative effects of interspecific interactions and abiotic factors. The question of knowing which ecological processes could indeed be recovered from species distribution patterns remains thus debated (e.g. Blanchet *et al.*, 2020). A last note

is that ELGRIN only deals with binary occurrence data rather than abundance or frequency data. In our simulation design, both the Lotka-Volterra and the VirtualCom models produced abundance data that we had to sample to obtain binary signals, loosing information during the process. On the contrary, ELGRIN performs better on colonisation-extinction simulations, where the dynamics directly generates binary data. Extending ELGRIN from the binary setup to the continuous one could improve the inference by considering more information in the species distribution data but it remains an important methodological challenge.

In terms of further perspectives, we might wonder whether this model could be extended for prediction purposes. In principle, it is possible to draw presence/absence data from the model for different values of the environment variables. These different values could allow for predictions in space but also in time. However, something to keep in mind is that metanetwork will not change in the model and will thus be considered as static and thus representative in space (or in time). If the metanetwork has not been built with that prediction perspective in mind, this might be an issue as we will miss interaction rewiring effects on species distributions. Instead, if the metanetwork is truly a potential metanetwork that tries to incorporate these potential interactions that have been observed yet (i.e. Maiorano *et al.*, 2020), it might be interesting to investigate how biotic interactions might further influence future species distributions in response to environmental changes.

Acknowledgements

The authors would like to thanks the anonymous reviewers that carefully reviewed a previous version of our model and manuscript, and notably proposed the Lotka-Volterra simulation scheme. Funding was provided by the French National Center for Scientific Research (CNRS) and the French National Research Agency (ANR) grant ANR-18-CE02-0010-01 EcoNet. VM would like to thank the LECA laboratory for hosting him in Chambéry.

Conflict of Interest Statement

The authors declare no conflict of interest.

References

Austin, M. (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, **157**, 101–118.

Biroli, G., Bunin, G. & Cammarota, C. (2018) Marginally stable equilibria in critical ecosystems. *New Journal of Physics*, **20**, 083051.

Blanchet, F.G., Cazelles, K. & Gravel, D. (2020) Co-occurrence is not evidence of ecological interactions. *Ecology Letters*, **23**, 1050–1063.

Botella, C., Dray, S., Matias, C., Miele, V. & Thuiller, W. (2022) An appraisal of graph embeddings for comparing trophic network architectures. *Methods in Ecology and Evolution*, **13**, 203–216.

Boulangeat, I., Gravel, D. & Thuiller, W. (2012) Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. *Ecology letters*, **15**, 584–593.

Braga, J., Pollock, L.J., Barros, C., Galiana, N., Montoya, J.M., Gravel, D., Maiorano, L., Montemaggiori, A., Ficetola, G.F., Dray, S. *et al.* (2019) Spatial analyses of multi-trophic terrestrial vertebrate assemblages in Europe. *Global Ecology and Biogeography*, **28**, 1636–1648.

Brémaud, P. (1999) *Markov chains: Gibbs fields, Monte Carlo simulation, and Queues*, volume 31. Springer.

Chalmandrier, L., Münkemüller, T., Gallien, L., De Bello, F., Mazel, F., Lavergne, S. & Thuiller, W. (2013) A family of null models to distinguish between environmental filtering and biotic interactions in functional diversity patterns. *Journal of Vegetation Science*, **24**, 853–864.

Chase, J.M. & Leibold, M.A. (2003) *Ecological niches: linking classical and contemporary approaches*. University of Chicago Press.

Cirtwill, A.R., Eklöf, A., Roslin, T., Wootton, K. & Gravel, D. (2019) A quantitative framework for investigating the reliability of empirical network construction. *Methods in Ecology and Evolution*, **10**, 902–911.

Connor, E.F. & Simberloff, D. (1979) The assembly of species communities: chance or competition? *Ecology*, **60**, 1132–1140.

de Candolle, A. (1855) *Géoographie botanique raisonnée ou, Exposition des faits principaux et des lois concernant la distribution géographique des plantes de l'époque actuelle*. Masson.

Diamond, J.M. (1975) Assembly of species communities. J. Diamond & M. Cody, eds., *Ecology and evolution of communities*, pp. 342–444. Harvard University Press.

Gravel, D., Baiser, B., Dunne, J.A., Kopelke, J.P., Martinez, N.D., Nyman, T., Poisot, T., Stouffer, D.B., Tylianakis, J.M., Wood, S.A. & Roslin, T. (2019) Bringing Elton and Grinnell together: a quantitative framework to represent the biogeography of ecological interaction networks. *Ecography*, **42**, 401–415.

Guisan, A., Thuiller, W. & Zimmermann, N.E. (2017) *Habitat Suitability and Distribution Models: With Applications in R*. Ecology, Biodiversity and Conservation. Cambridge University Press.

Holt, R.D. (2020) Some thoughts about the challenge of inferring ecological interactions from spatial data. *Biodiversity Informatics*, **15**, 61–66.

Hutchinson, G.E. (1959) Homage to Santa Rosalia or why are there so many kinds of animals? *The American Naturalist*, **93**, 145–159.

Lortie, C.J., Brooker, R.W., Choler, P., Kikvidze, Z., Michalet, R., Pugnaire, F.I. & Callaway, R.M. (2004) Rethinking plant community theory. *Oikos*, **107**, 433–438.

Maiorano, L., Montemaggiori, A., O'Connor, L., Ficetola, G. & W., T. (2020) TETRA-EU 1.0: A species-level trophic meta-web of European tetrapods. *Global Ecology & Biogeography*, **29**, 1452–1457.

Maiorano, L., Amori, G., Capula, M., Falcucci, A., Masi, M., Montemaggiori, A., Pottier, J., Psomas, A., Rondinini, C., Russo, D. *et al.* (2013) Threats from climate change to terrestrial vertebrate hotspots in Europe. *PLoS One*, **8**.

Morales-Castilla, I., Matias, M.G., Gravel, D. & Araújo, M.B. (2015) Inferring biotic interactions from proxies. *Trends in ecology & evolution*, **30**, 347–356.

Münkemüller, T. & Gallien, L. (2015) VirtualCom: a simulation model for eco-evolutionary community assembly and invasion. *Methods in Ecology and Evolution*, **6**, 735–743.

Ohlmann, M., Miele, V., Dray, S., Chalmandrier, L., O'Connor, L. & Thuiller, W. (2019) Diversity indices for ecological networks: a unifying framework using Hill numbers. *Ecology letters*, **22**, 737–747.

Ohlmann, M., Munoz, F., Massol, F. & Thuiller, W. (2022) Assessing mutualistic metacommunity capacity by integrating spatial and interaction networks. Technical report, arXiv:2206.11029.

Ovaskainen, O., Tikhonov, G., Norberg, A., Blanchet, F.G., Duan, L., Dunson, D., Roslin, T. & 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.

O'Connor, L.M.J., Pollock, L.J., Braga, J., Ficetola, G.F., Maiorano, L., Martinez-Almoyna, C., Montemaggiori, A., Ohlmann, M. & Thuiller, W. (2020) Unveiling the food webs of tetrapods across Europe through the prism of the Eltonian niche. *Journal of Biogeography*, **47**, 181–192.

Pellissier, L., Albouy, C., Bascompte, J., Farwig, N., Graham, C., Loreau, M., Maglianesi, M.A., Melián, C.J., Pitteloud, C., Roslin, T. *et al.* (2018) Comparing species interaction networks along environmental gradients. *Biological Reviews*, **93**, 785–800.

Peres-Neto, P.R., Olden, J.D. & Jackson, D.A. (2001) Environmentally constrained null models: site suitability as occupancy criterion. *Oikos*, **93**, 110–120.

Poggiato, G., Münkemüller, T., Bystrova, D., Arbel, J., Clark, J.S. & Thuiller, W. (2021) On the interpretations of joint modeling in community ecology. *Trends in Ecology and Evolution*, **36**, 391–401.

Pulliam, H. (2000) On the relationship between niche and distribution. *Ecology Letters*, **3**, 349–361.

Remien, C.H., Eckwright, M.J. & Ridenhour, B.J. (2021) Structural identifiability of the generalized Lotka–Volterra model for microbiome studies. *Royal Society Open Science*, **8**, 201378.

Ricklefs, R.E. (2008) Disintegration of the ecological community: American society of naturalists Sewall Wright Award Winner Address. *The American Naturalist*, **172**, 741–750.

Soberón, J. & Nakamura, M. (2009) Niches and distributional areas: Concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences*, **106**, 19644–19650.

Staniczenko, P.P., Sivasubramaniam, P., Suttle, K.B. & Pearson, R.G. (2017) Linking macroecology and community ecology: refining predictions of species distributions using biotic interaction networks. *Ecology letters*, **20**, 693–707.

Stone, L. & Roberts, A. (1990) The checkerboard score and species distributions. *Oecologia*, **85**, 74–79.

Takeuchi, Y. (1996) *Global Dynamical Properties of Lotka-Volterra Systems*. World Scientific, Singapore.

Thuiller, W., Pollock, L.J., Gueguen, M. & Münkemüller, T. (2015) From species distributions to meta-communities. *Ecology Letters*, **18**, 1321–1328.

Tikhonov, G., Abrego, N., Dunson, D. & Ovaskainen, O. (2017) Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context. *Methods in Ecology and Evolution*, **8**, 443–452.

Tylianakis, J.M. & Morris, R.J. (2017) Ecological networks across environmental gradients. *Annual Review of Ecology, Evolution, and Systematics*, **48**, 25–48.

Zurell, D., Pollock, L.J. & Thuiller, W. (2018) Do joint species distribution models reliably detect interspecific interactions from co-occurrence data in homogenous environments? *Ecography*, **41**, 1812–1819.

Appendix S1 for manuscript 'Quantifying the overall effect of biotic interactions on species distributions along environmental gradients', by M. Ohlmann, C. Matias, G. Poggiato, S. Dray, W. Thuiller & V. Miele.

S.1 Model extensions

S.1.1 Interaction strength

Besides the binary case, it is also possible to handle interaction strengths. An interaction strength can represent a frequency (e.g., the number of visits of a pollinator to a plant), an intensity (e.g., rate of predation, Berlow et al., 2004) or a preference (e.g. modulating trophic links with known affinities of a predator to its preys).

We write $A^\star = (A^\star_{ij})_{i,j\in V^\star}$ the adjacency matrix of the graph $G^\star$. Now, each edge $(i,j) \in E^\star$ is modulated through the weight $A^\star_{ij}$ of the interaction. In this case, sub-equations (1b) and (1c) are replaced by

$$\beta_{l,co-pres} \sum_{(i,j)\in E^\star} A^\star_{ij} \mathbf{1}\{X^l_j = X^l_i = 1\} = \beta_{l,co-pres} \sum_{(i,j)\in E^\star} A^\star_{ij} X^l_j X^l_i$$
$$\text{and } \beta_{l,co-abs} \sum_{(i,j)\in E^\star} A^\star_{ij} \mathbf{1}\{X^l_j = X^l_i = 0\} = \beta_{l,co-abs} \sum_{(i,j)\in E^\star} A^\star_{ij} (1 - X^l_j)(1 - X^l_i),$$

respectively.

S.1.2 Sampling effects

The random variables X^l_i that indicate the presence of species i at location l might not be exactly observed due to sampling effects. Here, we propose to account for these effects by assuming that each species $i \in V^\star$ is sampled with probability $p_{i,l} \in (0,1)$ at location $l \in \{1,\dots,L\}$. We therefore introduce a new set of random variables $Y^l_i, i \in V^\star, l \in \{1,\dots,L\}$ such that each Y^l_i only depends on X^l_i and is distributed as

$$\begin{aligned}\mathbb{P}(Y^l_i | X^l_i) &= p_{i,l}^{Y^l_i}(1-p_{i,l})^{1-Y^l_i} X^l_i + (1 - X^l_i)(1 - Y^l_i) \\ &= p_{i,l}^{X^l_i Y^l_i}(1-p_{i,l})^{X^l_i(1-Y^l_i)} \mathbf{1}\{(1 - X^l_i)Y^l_i \neq 1\}.\end{aligned}$$

Specifically, whenever $X^l_i = 0$ (species i is absent from location l), species i cannot be observed at location l and $Y^l_i = 0$. Now, when $X^l_i = 1$ (species i is present at location l), it is observed ($Y^l_i = 1$) with sampling probability $p_{i,l}$ and unobserved ($Y^l_i = 0$) with probability $1 - p_{i,l}$. The parameter $p_{i,l}$ must be given by the user considering three possible cases: species dependent sampling ($p_{i,l} := p_i; i \in V^\star$), location dependent sampling ($p_{i,l} := p_l; 1 \leq l \leq L$) or constant sampling ($p_{i,l} := p$). In this case, the X^l_i become latent variables as we only observe the Y^l_i's. The model turns out to be a hidden Markov random field (HMRF).

S.1.3 Plasticity of interactions

Our model is able to assume that interactions are not necessarily induced by the presence/absence variables (we can assume that two species interact in a given location but not in another location). In this case, we consider a sample of observed graphs $G^1, \dots, G^L$ where each $G^l = (V^l, E^l)$ is such that $V^l \subset V^\star$. These graphs represent local interactions that are observed at the different locations $l \in \{1, \dots, L\}$. The main point here is that we assume that these interactions are sampled from the pool of potential interactions encoded in the metanetwork $G^\star$. Let $A^l = (A^l_{i,j})_{i,j \in V^l}$ denote the adjacency matrix of the graph G^l. We assume that any two species that are observed and that can potentially interact (i.e., are linked in the metanetwork $G^\star$) do effectively interact at location l with a probability that depends only on these two species. Namely for any $(i,j) \in E^\star$, conditional on the fact that two species $i, j \in V^\star$ were observed at location l (namely $Y^l_i Y^l_j = 1$), we set

$$A^l_{i,j} | Y^l_i Y^l_j = 1 \sim \mathcal{B}(\epsilon_{ij}),$$

and $A^l_{i,j} \equiv 0$ whenever $(i,j) \notin E^\star$ or $Y^l_i = 0$ or $Y^l_j = 0$. This additional parameter $\epsilon = \{\epsilon_{i,j}\}_{i,j \in V^\star}$ allows us to handle interaction plasticity directly in the model.

S.2 Mathematical details on the model

S.2.1 Identifying the parameters of the Gibbs distribution

We first address the issue of the identifiability of the parameters from the Gibbs distribution. In what follows, we focus on the case of a binary metanetwork $G^\star$. However, our results remain valid in the weighted case, where degrees are replaced by weighted degrees and the cardinality $|E^\star|$ (total number of edges in $G^\star$) becomes the total sum of the weights.

Let us focus on the model with no covariates ($W_l = 0$) and consider for each location $l \in \{1, \dots, L\}$ the maps $\psi_l = (\{a_i\}_i, a_l, \beta_{l,co-pres}, \beta_{l,co-abs}) \mapsto \mathbb{P}_{\psi_l}$, where

$$\mathbb{P}_{\psi_l}(\{X^l_i\}_{i \in V^\star}) = \frac{1}{Z_{\psi_l}} \exp\Big(\sum_{i \in V^\star} (a_i + a_l) X^l_i + \beta_{l,co-pres} \sum_{(i,j) \in E^\star} X^l_j X^l_i + \beta_{l,co-abs} \sum_{(i,j) \in E^\star} (1 - X^l_j)(1 - X^l_i) \Big).$$

For any $\psi = (\{a_i\}_{i,l}, \{a_l, \beta_{l,co-pres}, \beta_{l,co-abs}\}_l)$ we also define the global probability distribution $\mathbb{P}_\psi$ as follows

$$\mathbb{P}_\psi(\{X^l_i\}_{i \in V^\star; 1 \le l \le L}) = \prod_{l=1}^{L} \mathbb{P}_{\psi_l}(\{X^l_i\}_{i \in V^\star}).$$

Proposition 1 (Identifying linear combinations of the parameter)**.** *In the model without covariate ($W_l = 0$, for any l), the probability distribution $\mathbb{P}_\psi$ uniquely defines the quantities*

$$\beta_{l,co-pres} + \beta_{l,co-abs}, \tag{S.2}$$

$$\textit{and } a_i + a_l + \beta_{l,co-pres} \deg_{G^\star}(i) \textit{ or equivalently } a_i + a_l - \beta_{l,co-abs} \deg_{G^\star}(i), \tag{S.3}$$

for any $i \in V^\star, l \in \{1,\dots,L\}$*, where* $\deg_{G^\star}(i)$ *is the degree of species* i *in the metanetwork* $G^\star$*. Moreover, if there exist 2 species* $1 \leq i,j \leq N$ *such that* $\deg_{G^\star}(i) \neq \deg_{G^\star}(j)$ *in* $G^\star$*, then the probability distribution* $\mathbb{P}_\psi$ *uniquely defines the additional quantities*

$$\beta_{l,co-abs} - \beta_{l',co-abs} \text{ or equivalently } \beta_{l,co-pres} - \beta_{l',co-pres}, \tag{S.4}$$

$$\text{and } a_l - a_{l'}, \tag{S.5}$$

for any $l, l' \in \{1,\dots,L\}$.

Proof. Let us denote $\alpha_{i,l} = a_i + a_l$. As $\mathbb{P}_{\psi_l}$ is a marginal of $\mathbb{P}_\psi$, we start by fixing the location $l \in \{1,\dots,L\}$ and consider the probabilities of specific configurations at this location. We let X^l_{-i} denote the set $\{X^l_j; j \in V^\star, j \neq i\}$. From the knowledge of $\mathbb{P}_\psi$, we obtain for $l \in \{1,...,L\}$ and $i \in V^\star$ the quantities

$$\begin{aligned}
s^l_0 &:= \log \mathbb{P}_{\psi_l}(\{0,...,0\}) = -\log(Z_{\psi_l}) + |E^\star|\beta_{l,co-abs} \\
s^l_1 &:= \log \mathbb{P}_{\psi_l}(\{1,...,1\}) = -\log(Z_{\psi_l}) + \sum_i \alpha_{i,l} + |E^\star|\beta_{l,co-pres} \\
s^{i,l}_{10} &:= \log \mathbb{P}_{\psi_l}(\{X^l_i = 1, X^l_{-i} = 0\}) = -\log(Z_{\psi_l}) + \alpha_{i,l} + \beta_{l,co-abs}(|E^\star| - \deg_{G^\star}(i)) \\
s^{i,l}_{01} &:= \log \mathbb{P}_{\psi_l}(\{X^l_i = 0, X^l_{-i} = 1\}) = -\log(Z_{\psi_l}) + \sum_{j\neq i} \alpha_{j,l} + \beta_{l,co-pres}(|E^\star| - \deg_{G^\star}(i)),
\end{aligned}$$

where $|E^\star|$ is the cardinality of the set $E^\star$. It follows

$$\begin{aligned}
r^l_1 &:= s^l_1 - s^l_0 = \sum_i \alpha_{i,l} + |E^\star|(\beta_{l,co-pres} - \beta_{l,co-abs}) \\
r^{i,l}_2 &:= s^{i,l}_{10} - s^l_0 = \alpha_{i,l} - \beta_{l,co-abs}\deg_{G^\star}(i) \\
r^{i,l}_3 &:= s^{i,l}_{01} - s^l_0 = \sum_{j\neq i} \alpha_{j,l} + (\beta_{l,co-pres} - \beta_{l,co-abs})|E^\star| - \beta_{l,co-pres}\deg_{G^\star}(i).
\end{aligned}$$

From these equations, we uniquely obtain

$$\begin{aligned}
t^{i,l}_1 &:= r^l_1 - r^{i,l}_3 = \alpha_{il} + \beta_{l,co-pres}\deg_{G^\star}(i) \\
t^{i,l}_2 &:= r^l_1 - r^{i,l}_2 - r^{i,l}_3 = (\beta_{l,co-abs} + \beta_{l,co-pres})\deg_{G^\star}(i).
\end{aligned}$$

As a consequence, as soon as there is at least one edge in the metanetwork $G^\star$ (inducing at least one species i with $\deg_{G^\star}(i) \neq 0$) we can obtain the quantities $\beta_{l,co-abs} + \beta_{l,co-pres}$ (recall that $\deg_{G^\star}(i)$ is known) as well as $\alpha_{i,l} + \beta_{l,co-pres}\deg_{G^\star}(i)$ uniquely from the distribution $\mathbb{P}_\psi$. Note also that combining the knowledge of these two quantities, the second is equivalent to knowing $\alpha_{i,l} - \beta_{l,co-abs}\deg_{G^\star}(i)$.

Now, let us recall that $\alpha_{i,l} = a_i + a_l$. For two different locations $l \neq l'$, we have access to

$$t^{i,l}_1 - t^{i,l'}_1 = a_l - a_{l'} + (\beta_{l,co-pres} - \beta_{l',co-pres})\deg_{G^\star}(i).$$

We now assume that there exist two species $1 \leq i,j \leq N$ such that $\deg_{G^\star}(i) \neq \deg_{G^\star}(j)$ in $G^\star$ and obtain (S.4) as follows

$$\beta_{l,co-pres} - \beta_{l',co-pres} = (t^{i,l}_1 - t^{i,l'}_1 - t^{j,l}_1 + t^{j,l'}_1)[\deg_{G^\star}(i) - \deg_{G^\star}(j)]^{-1}.$$

Combining this with (S.2), it is equivalent to the unique identification of $\beta_{l,co-abs} - \beta_{l',co-abs}$. Finally, going back to $t_1^{i,l} - t_1^{i,l'}$ we uniquely obtain $a_l - a_{l'}$. □

Definition 1 (Equivalence class)**.** *For any parameter* $\psi = (\{a_i\}_i, \{a_l, \beta_{l,co-pres}, \beta_{l,co-abs}\}_l)$*, its equivalence class* $[\psi]$ *is defined as*

$$[\psi] := \{(\{a_i + \gamma \deg_{G^\star}(i) - \delta\}_i, \{a_l + \delta, \beta_{l,co-pres} - \gamma, \beta_{l,co-abs} + \gamma\}_l); \gamma \in \mathbb{R}, \delta \in \mathbb{R}\}.$$

Corollary 1 (Parameter identifiability up to the equivalence class)**.** *In the model without covariate* ($W_l = 0$*, for any* l*) and assuming that there exist 2 species* $1 \leq i, j \leq N$ *such that* $\deg_{G^\star}(i) \neq \deg_{G^\star}(j)$ *in* $G^\star$*, we have that whenever there are two parameter values* $\psi, \tilde{\psi}$ *such that* $\mathbb{P}_\psi = \mathbb{P}_{\tilde{\psi}}$*, then* $\tilde{\psi} \in [\psi]$*. In other words, the equality* $\mathbb{P}_\psi = \mathbb{P}_{\tilde{\psi}}$ *implies that there exist real values* $\gamma, \delta \in \mathbb{R}$ *such that for any* $i \in V^\star$ *and* $l \in \{1, \ldots, L\}$*, we have*

$$\begin{aligned}
\tilde{a}_i &= a_i + \gamma \deg_{G^\star}(i) + \delta \\
\tilde{a}_l &= a_l - \delta \\
\tilde{\beta}_{l,co-pres} &= \beta_{l,co-pres} - \gamma \\
\tilde{\beta}_{l,co-abs} &= \beta_{l,co-abs} + \gamma.
\end{aligned}$$

Proof. Assume that $\mathbb{P}_\psi = \mathbb{P}_{\tilde{\psi}}$ and define for any location $l \in \{1, \ldots, L\}$ the quantity $\gamma_l := \beta_{l,co-pres} - \tilde{\beta}_{l,co-pres}$. We know from Proposition 1 that

$$\begin{aligned}
\beta_{l,co-abs} + \beta_{l,co-pres} &= \tilde{\beta}_{l,co-abs} + \tilde{\beta}_{l,co-pres} \\
\tilde{a}_i + \tilde{a}_l + \tilde{\beta}_{l,co-pres} \deg_{G^\star}(i) &= a_i + a_l + \beta_{l,co-pres} \deg_{G^\star}(i).
\end{aligned}$$

This induces that

$$\begin{aligned}
\gamma_l &= \tilde{\beta}_{l,co-abs} - \beta_{l,co-abs} \\
\text{and} \quad \tilde{a}_i + \tilde{a}_l &= a_i + a_l + \gamma_l \deg_{G^\star}(i).
\end{aligned}$$

Let us further prove that γ_l does not depend on l. From Proposition 1 and the additional assumption that at least two species have different degrees in the metanetwork, we have for any locations $l, l' \in \{1, \ldots, L\}$,

$$\beta_{l,co-pres} - \beta_{l',co-pres} = \tilde{\beta}_{l,co-pres} - \tilde{\beta}_{l',co-pres} = \beta_{l,co-pres} - \beta_{l',co-pres} - \gamma_l + \gamma_{l'},$$

which implies that $\gamma_l = \gamma_{l'}$ for any pair of locations. Finally, let us define for any location and any species

$$\delta_l = a_l - \tilde{a}_l \quad \text{and} \quad \delta_i = a_i - \tilde{a}_i.$$

We have established that $\delta_l + \delta_i = -\gamma \deg_{G^\star}(i)$. This implies that δ_l is constant through locations and equal to some δ. This concludes the proof. □

Corollary 1 tells us that the model parameter is identifiable up to the equivalence class in Definition 1. Note that it is possible to choose one specific representative parameter in this class.

Proposition 2 (Choosing a representative)**.** *In the model without covariate ($W_l = 0$, for any l) and assuming that there exist 2 species $1 \leq i, j \leq N$ such that $\deg_{G^\star}(i) \neq \deg_{G^\star}(j)$ in $G^\star$, for any parameter value $\tilde{\psi}$, it is possible to choose a unique representative $\psi \in [\tilde{\psi}]$ such that the estimated linear regression coefficients of the set of parameters $\{a_i\}_i$ over the degrees $\{\deg_{G^\star}(i)\}_i$ are equal to 0, namely*

$$(\hat{\gamma}, \hat{\delta}) := \underset{(\gamma,\delta)\in\mathbb{R}^2}{\text{Argmin}} \sum_{i\in V^\star} (a_i - \gamma \deg_{G^\star}(i) - \delta)^2$$

satisfies $(\hat{\gamma}, \hat{\delta}) = (0, 0)$.

Proof. Fix a parameter value $\tilde{\psi}$ and consider the linear regression of the set of parameters $\{\tilde{a}_i\}_i$ over the degrees $\{\deg_{G^\star}(i)\}_i$, namely

$$(\tilde{\gamma}, \tilde{\delta}) := \underset{(\gamma,\delta)\in\mathbb{R}^2}{\text{Argmin}} \sum_{i\in V^\star} (\tilde{a}_i - \gamma \deg_{G^\star}(i) - \delta)^2.$$

Then by setting the parameter $\psi = (\{a_i\}_{i,l}, \{a_l, \beta_{l,co-pres}, \beta_{l,co-abs}\}_l)$ as

$$\begin{aligned} a_i &:= \tilde{a}_i - \tilde{\gamma} \deg_{G^\star}(i) - \tilde{\delta}; \\ a_l &:= \tilde{a}_l + \tilde{\delta}; \\ \beta_{l,co-pres} &:= \tilde{\beta}_{l,co-pres} + \tilde{\gamma} \\ \beta_{l,co-abs} &:= \tilde{\beta}_{l,co-abs} - \tilde{\gamma} \end{aligned}$$

(for any i, l), we know from Definition 1 that $\psi \in [\tilde{\psi}]$ and also by definition, the estimated values

$$(\hat{\gamma}, \hat{\delta}) := \underset{(\gamma,\delta)\in\mathbb{R}^2}{\text{Argmin}} \sum_{i\in V^\star} (a_i - \gamma \deg_{G^\star}(i) - \delta)^2$$

will now satisfy $(\hat{\gamma}, \hat{\delta}) = (0, 0)$. □

Remark 1. *The choice of the representative parameter given by Proposition 2 is such that the response of species i to the environment does not depend on its degree in the metanetwork and thus on its number of interactions. This is a natural choice to separate the Grinellian part from the Eltonian one in our model. Note that this representative parameter is the one we rely on when interpreting the model. Thus, when we comment the behaviour of the model with respect to different values of its parameter, we always rely on this specific representative.*

Note however that whatever the choice of the representative, the intercept values a_i and a_l are inferred up to an additive constant.

S.2.2 A compatibility matrix to robustify the model

In this section, we slightly modify the model to handle cases where either there are species with tight environmental niches or where the metanetwork $G^\star$ contains edges between species with incompatible environmental niches (which would be a nonsense). Indeed, we aim at estimating Eltonian effects only when species are in their Grinnelian niche.

We introduce a binary matrix $C = (C_{il})_{i\in V^\star, 1\leq l\leq L}$ that encodes the possibility for species i to be present at location l given its niche properties. The matrix C is called a *compatibility matrix*. For

the model's presentation, it is supposed to be fixed and known. In practice, it is either obtained from expert knowledge, otherwise built from the realized niche of each species (our implementation in the function `elgrin` will pre-estimate the compatibility matrix from realized niche before fitting the model). In the latter case, for any species i, at each location l and for each covariate d, relying on the observation set $\{x_i^l\}_{i,l}$, we set

$$\omega_{id} = \inf_{1\leq l\leq L}\{W_{ld}; x_i^l = 1\}, \tag{S.6}$$

$$\Omega_{id} = \sup_{1\leq l\leq L}\{W_{ld}; x_i^l = 1\} \tag{S.7}$$

$$\text{and } C_{il} = 1\{\forall 1 \leq d \leq D, W_{ld} \in [\omega_{id}; \Omega_{id}]\}.$$

where location l is characterized by an environmental covariate vector $W_l = (W_{l1}, \ldots, W_{lD})$. Naturally, if $X_i^l = 1$ then $C_{il} = 1$.

Relying on the compatibility matrix, at each location l we restrict our attention to species compatible with the environment at this location. In particular, we now impose that $X_i^l = 0$ whenever $C_{il} = 0$. Thus the probability distribution of the species in ELGRIN is modified as follows

$$\mathbb{P}_{\psi_l}(\{X_i^l\}_{i\in V^\star}) = \left(\prod_{i\in V^\star; C_{il}=0}(1-X_i^l)\right) \times \frac{1}{Z_{\psi_l}} \exp\Big\{ \sum_{i\in V^\star; C_{il}=1} \Big[(a_l + a_i + W_l^\intercal b_i + (W_l^2)^\intercal c_i)X_i^l + \beta_{l,co-pres} \sum_{j;(i,j)\in E^\star} X_j^l X_i^l + \beta_{l,co-abs} \sum_{j;(i,j)\in E^\star} C_{jl}(1-X_j^l)(1-X_i^l)\Big]\Big\}.$$

Note that if the compatibility matrix is full of 1 (i.e. all the species may occur at all locations), we are back to our initial model. Otherwise, we now avoid mistaking co-absence of two interacting species with the event of two independent absences due to incompatible niches.

From a modeling point of view, the modified version of the model helps in robustifying our results. This is the case for instance when considering interacting species with tight niches. Indeed, at locations l where two interacting species i, j are absent due to incompatible environmental conditions (i.e. $C_{il} = C_{jl} = 0$), we observe that $X_i^l = X_j^l = 0$. In that case in our original model, this double absence would wrongly be interpreted as a co-absence and blur the inference of $\beta_{l,co-abs}$. Note also that whenever two species i, j are potentially interacting (i.e. $(i, j) \in E^\star$), we consider that their respective niches should overlap ($C_{il} = C_{jl} = 1$ for at least one location l). If this rule is not satisfied, it could happen that, without the additional factor $C_{il}C_{jl}$ regulating the co-absence term, an absence of species i would be interpreted as a co-absence due to its interaction with species j.

Note that at locations l where the environment covariates W_l prevent from the occurrence of a species i (i.e. $C_{il} = 0$), it is useless to try to fit the Grinellian part of the model, i.e. the non-informative intercepts a_i, a_l and the parameters b_i, c_i. So that when appropriate, we only consider the estimated maps $W \mapsto W^\intercal b_i + (W^2)^\intercal c_i$ on the environment values compatible with species i.

S.2.3 Hidden Markov random field and its interpretation

We discuss here the model in its full generality, including possible weights on the metanetwork, sampling effects, plasticity of interactions and the robust version relying on a compatibility matrix.

We thus have $\mathbf{X} := \{\mathbf{X}^l\}_{1\leq l\leq L} = \{X_i^l\}_{i\in V^\star, 1\leq l\leq L}$ (resp. $\mathbf{Y} := \{\mathbf{Y}^l\}_{1\leq l\leq L} = \{Y_i^l\}_{i\in V^\star, 1\leq l\leq L}$ and $\mathbf{A} := \{A^l\}_{1\leq l\leq L} = \{A_{i,j}^l\}_{i,j\in V^l, 1\leq l\leq L}$) denoting the set of true occurrence variables (resp. observed occurrences and observed interactions). We assume that we observe $(\mathbf{Y}, \mathbf{A})$, while $\mathbf{X}$ are latent random variables.

A Gibbs distribution specifies the joint associations between the species occurrence variables $\{X_i^l\}_{i\in V^\star}$, as follows

$$\mathbb{P}_{\psi_l}(\{X_i^l\}_{i\in V^\star}) = \left(\prod_{i\in V^\star; C_{il}=0}(1-X_i^l)\right) \times \frac{1}{Z_{\psi_l}} \exp\Big\{ \sum_{i\in V^\star; C_{il}=1} \Big[(a_l + a_i + W_l^\intercal b_i + (W_l^2)^\intercal c_i)X_i^l + \beta_{l,co-pres} \sum_{j;(i,j)\in E^\star} X_j^l X_i^l + \beta_{l,co-abs} \sum_{j;(i,j)\in E^\star} C_{jl}(1-X_j^l)(1-X_i^l)\Big]\Big\}. \quad \text{(S.8)}$$

First note that the normalizing constant Z_{ψ_l} is given by

$$Z_{\psi_l} = \sum_{i\in V^\star; C_{il}=1} \sum_{x_i\in\{0,1\}} \exp\Big(\sum_{i\in V^\star; C_{il}=1} [a_l + a_i + W_l^\intercal b_i + (W_l^2)^\intercal c_i]x_i + \beta_{l,co-pres} \sum_{j;(i,j)\in E^\star} A_{ij}^\star x_i x_j + \beta_{l,co-abs} \sum_{j;(i,j)\in E^\star} A_{ij}^\star C_{jl}(1-x_i)(1-x_j)\Big).$$

In general, this normalising constant Z_{ψ_l} cannot be computed due to the large number of possible configurations appearing in the sum. The statistical inference procedure needs to deal with that.

The model interpretation strongly builds on the *Markov property*, a fundamental characteristic of Markov random fields. In the following we focus on the species compatible with one location ($C_{il} = 1$); otherwise recall that its occurrence is set to zero with probability 1. Let us denote $\mathcal{N}_i^\star$ the set of species $j \in V^\star$ that are connected to i in the graph $G^\star$ (namely $\{j \in V^\star; A_{ij}^\star \neq 0\}$) and $X_{\mathcal{N}_i^\star}^l$, the set of corresponding random variables X_j^l for $j \in \mathcal{N}_i^\star$. We also recall that X_{-i}^l denotes the set $\{X_j^l; j \in V^\star, j \neq i\}$. Then, under the *Markov property* we have

$$\mathbb{P}_{\psi_l}(X_i^l | X_{-i}^l, C_{il}=1) = \mathbb{P}_{\psi_l}(X_i^l | X_{\mathcal{N}_i^\star}^l, C_{il}=1) \propto \exp\Big([a_l + a_i + W_l^\intercal b_i + (W_l^2)^\intercal c_i]X_i^l + \beta_{l,co-pres} \sum_{j\in\mathcal{N}_i^\star} A_{ij}^\star X_j^l X_i^l + \beta_{l,co-abs} \sum_{j\in\mathcal{N}_i^\star} A_{ij}^\star C_{jl}(1-X_j^l)(1-X_i^l)\Big), \quad \text{(S.9)}$$

where $\propto$ means proportional (equals up to a normalising constant). More specifically, it means that the conditional occurrence probability of a species i is modulated by the occurrences of the species interacting with i in $G^\star$. In other words, a species presence only depends on abiotic environment and on the species it interacts with. Moreover, the presence/absence variables of any two species are not statistically independent of each other if $G^\star$ is connected (namely, if there exists a path between any two species in $G^\star$). Meanwhile, if $G^\star$ has more than one connected component (i.e. disconnected compartments, Krause et al., 2003), then the presence/absence of species in different components are independent. The Markov property is the cornerstone idea of our model. Indeed,

the conditional probabilities of each random variable is specified through (S.9) and is rooted on the idea that the occurrence of a species i at location l depends both on a suitability term, specific to that species and the local environment, and on the presence/absence of other species with whom it interacts (as encoded in the metanetwork). From this set of conditional probabilities, the Hammersley-Clifford theorem (Besag, 1974) ensures that there exists a proper joint distribution on the random variables $\{X_i^l\}_{i,l}$ and that it is given by Equation (S.8).

Now, the observed species occurrence variables $Y_i^l, i \in V^\star, l \in \{1, \dots, L\}$ are distributed such that each Y_i^l only depends on X_i^l (the true occurrence variable) with

$$\begin{aligned}\mathbb{P}(Y_i^l|X_i^l) &= p_{i,l}^{Y_i^l}(1-p_{i,l})^{1-Y_i^l}X_i^l + (1-X_i^l)(1-Y_i^l) \\ &= p_{i,l}^{X_i^l Y_i^l}(1-p_{i,l})^{X_i^l(1-Y_i^l)}\mathbf{1}\{(1-X_i^l)Y_i^l \neq 1\}. \end{aligned} \tag{S.10}$$

In what follows, we choose to impose that the sampling parameters $p_{i,l}$ are set by the user. A consequence of this is that the quantity (S.10) will play no role in the inference procedure. Indeed, it is a constant quantity with respect to the parameter. Finally we set

$$A_{i,j}^l | Y_i^l Y_j^l = 1 \sim \mathcal{B}(\epsilon_{ij}), \tag{S.11}$$

and $A_{i,j}^l \equiv 0$ whenever $(i,j) \notin E^\star$ or $Y_i^l = 0$ or $Y_j^l = 0$.

Building on Equations (S.10) and (S.11), we first obtain the conditional distribution of all observations $(\mathbf{Y}, \mathbf{A})$ given the latent variables $\mathbf{X}$

$$\begin{aligned}\mathbb{P}_\phi(\mathbf{Y}, \mathbf{A}|\mathbf{X}) &= \prod_{l=1}^{L} \mathbb{P}_\phi(A^l|\mathbf{Y}^l)\mathbb{P}(\mathbf{Y}^l|\mathbf{X}^l) \\ &= \prod_{l=1}^{L}\prod_{i\in V^\star}\left[p_{i,l}^{X_i^l Y_i^l}(1-p_{i,l})^{X_i^l(1-Y_i^l)}1\{(1-X_i^l)Y_i^l \neq 1\}\right] \times \prod_{(i,j)\in E^\star} \epsilon_{ij}^{Y_i^l Y_j^l A_{i,j}^l}(1-\epsilon_{ij})^{Y_i^l Y_j^l (1-A_{i,j}^l)}.\end{aligned}$$

Here, the parameter $\epsilon = \{\epsilon_{ij}\}_{i,j\in V^\star}$ drives the distribution of the observation process from the latent one.

Finally, our model is obtained by combining this with Equation (S.8) for the distribution of the latent variables $\mathbf{X}$. Thus the global model is parameterised by $\theta = \{\theta_l\}_{1\leq l\leq L}$ where each $\theta_l = (\psi_l, \epsilon)$. This amounts to the following sets of parameters

$$(\{a_i, b_i, c_i\}_{i\in V^\star}, \{a_l, \beta_{l,co-abs}, \beta_{l,co-pres}\}_{1\leq l\leq L}, \{\epsilon_{ij}\}_{i,j\in V^\star})$$

so there are $3N + 3L + N(N-1)$ parameters when the observed graphs A^l are directed (and $3N + 3L + N(N-1)/2$ when the observed graphs A^l are undirected) compared with $N(N-1)L$ observations. However note that in the model inference (see next section), the parameters ϵ_{ij} are pre-estimated (see Equation (S.12)) and do not appear in the main inference algorithm (see Algorithm 1). In what follows, we often use the notation

$$\alpha_{i,l} = a_i + a_l + W_l^\intercal b_i + (W_l^2)^\intercal c_i.$$

A chain graph (Lauritzen, 1996) describing the dependencies among the random variables in this model is given in Fig. S.7.

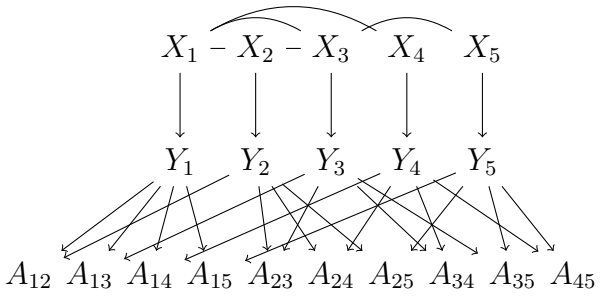


Figure S.7: Example of a metanetwork $G^\star$ (relations among the random variables $\{X_i\}_{i\in V^\star}$ with $V^\star = \{1,\dots,5\}$, on the top row) and induced dependency chain graph of all the variables in the model for one observed undirected graph $A = (A_{ij})_{i<j}$ with no self-loops.

S.3 Model inference

We present the inference procedure in the most general case, namely with weighted metanetwork, sampling effects and plasticity of interactions. This means that our inference procedure takes place in the context of a hidden Markov random field model.

S.3.1 Likelihood

The log-likelihood for observing independent interaction graphs $G^1,\dots,G^L$ at the different locations (and thus species occurrences variables ; indeed it is equivalent to observe $G^1,\dots,G^L$ or $(\mathbf{Y}^1,A^1,\dots,\mathbf{Y}^L,A^L)$) in this model is given by

$$\ell_{n,L}(\theta) = \sum_{l=1}^{L} \log \mathbb{P}_{\theta_l}(G^l),$$

where

$$\mathbb{P}_{\theta_l}(G^l) = \sum_{\{x_i^l\}_{i\in V^\star}\in\{0,1\}^N} \mathbb{P}_{\theta_l}(G^l, \{X_i^l = x_i^l; i\in V^\star\}).$$

As usual in latent variables models, this sum over all possible configurations $\{x_i^l\}_{i\in V^\star}\in\{0,1\}^N$ cannot be computed (unless N is really small). The inference procedure in latent variable models generally relies on the Expectation-Maximisation (EM) algorithm (Dempster et al., 1977). In the context of hidden Markov random fields, many difficulties arise that prevent from using this simple strategy.

The complete log-likelihood $\ell^c_{n,L}(\theta)$ contribution of all observations and all latent configurations is given by

$$\begin{aligned}\ell^c_{n,L}(\theta) &:= \log \mathbb{P}_\theta(\mathbf{X}, G^l,\dots,G^L) = \sum_{l=1}^{L} \log \mathbb{P}_{\theta_l}(\mathbf{X}^l,\mathbf{Y}^l,A^l)\\ &= \sum_{l=1}^{L} \log \mathbb{P}_{\psi_l}(\mathbf{X}^l) + \sum_{l=1}^{L}\sum_{i\in V^\star} \log \mathbb{P}(Y_i^l|X_i^l) + \sum_{l=1}^{L}\sum_{i,j\in V^l} \log \mathbb{P}_\phi(A_{i,j}^l|Y_i^l,Y_j^l).\end{aligned}$$

This can be written as

$$\begin{aligned}\ell^c_{n,L}(\theta) =& \sum_{l=1}^{L}\sum_{i\in V^\star} C_{il}\log(1-\alpha_{i,l}) + \sum_{l=1}^{L}\sum_{i\in V^\star} C_{il}X_i^l \log\left(\frac{\alpha_{i,l}}{1-\alpha_{i,l}}\right) + \sum_{l=1}^{L}\sum_{(i,j)\in E^\star} C_{il}C_{jl}A^\star_{ij}X_j^l X_i^l \\ &+ \sum_{l=1}^{L}\beta_{l,co-abs}\sum_{(i,j)\in E^\star} A^\star_{ij}C_{il}C_{jl}(1-X_j^l)(1-X_i^l) - \sum_{l=1}^{L}\log(Z_{\psi_l}) \\ &+ \sum_{i\in V^\star}\sum_{l=1}^{L} X_i^l\Big\{Y_i^l\log(p_{i,l}) + (1-Y_i^l)\log(1-p_{i,l})\Big\} \\ &+ \sum_{i,j\in V^\star}\sum_{l=1}^{L} Y_i^l Y_j^l\Big\{A_{i,j}^l\log\epsilon_{ij} + (1-A_{i,j}^l)\log(1-\epsilon_{ij})\Big\} + \text{cst.}\end{aligned}$$

Here, we restrict our attention to complete datasets $(\mathbf{X}^l, G^l)$ which are compatible, in the sense that whenever $X_i^l = 0$ we also have $Y_i^l = 0$. Otherwise the probability above is 0 and its log is $-\infty$.

S.3.2 Estimating the frequency of interactions

First, it is important to note that a consequence of the dependence among the $\{X_i^l\}_{i\in V^\star}$ is that the random variables $A_{i,j}^l$ and $A_{i'j'}^l$ are dependent. However, this dependency is entirely carried by the species observations Y_i^l's (which themselves are dependent through the species latent presences X_i^l's). In other words, we have $\mathbb{P}_\phi(A^l|\mathbf{Y}^l, \mathbf{X}^l) = \mathbb{P}_\phi(A^l|\mathbf{Y}^l)$. A consequence is that the parameters ϵ that describe the graph distribution are directly estimated from the data. While the sampling parameters and the random field ones ($\beta_{l,co-abs}$, $\beta_{l,co-pres}$ and $\alpha_{i,l}$'s) require a sophisticate inference procedure, the ϵ_{ij} parameters are directly estimated by the frequencies

$$\hat{\epsilon}_{ij} = \frac{\sum_{l=1}^{L} A_{ij}^l}{\sum_{l=1}^{L} Y_i^l Y_j^l}. \tag{S.12}$$

Here, the normalising term $\sum_{l=1}^{L} Y_i^l Y_j^l$ is simply the number of simultaneous observations of species i and j across the L different locations, while the numerator counts the number of observed interactions between those species across locations.

S.3.3 Inference of the random field parameters with simulated field algorithm

Now, we focus on the estimation of random field parameters $\beta_{l,co-abs}, \beta_{l,co-pres}$ and $\alpha_{i,l}$'s. A classical EM algorithm would consist in (iteratively) optimising with respect to $\psi = \{\psi_l\}_{1\le l\le L}$ the quantity

$$Q(\psi) = \sum_{l=1}^{L}\mathbb{E}\big(\log\mathbb{P}_{\psi_l}(\mathbf{X}^l, \mathbf{Y}^l)|\psi_l^{(t)}, \mathbf{Y}^l\big) = \sum_{l=1}^{L}\mathbb{E}\big[\log\mathbb{P}_{\psi_l}(\mathbf{X}^l)\big|\psi_l^{(t)}, \mathbf{Y}^l\big] + \text{cst}, \tag{S.13}$$

computed with the current value of the parameter $\psi^{(t)} = \{\psi_l^{(t)}\}_{1\leq l\leq L}$. (Recall that in our setup, the observations $\mathbf{Y}$ are obtained from $\mathbf{X}$ through a random function with fixed and known parameters). The above quantity has many drawbacks: first it contains the partition functions Z_{ψ_l} that are unknown and cannot be computed. Second, the conditional distribution of $\mathbf{X}^l$ given $\mathbf{Y}^l$ has an intricate dependency structure and thus may not be computed (in fact it is also a Markov random field).

We thus follow the *simulated field algorithm* proposed in Celeux et al. (2003). It is based on two different approximations of probability distributions plus a simulation step, as follows. First, the distribution $\mathbb{P}_\psi(\mathbf{X})$ appearing in the complete likelihood is replaced by a mean-field approximation, namely the product distribution

$$\mathbb{P}^1(\mathbf{X}|\psi,\tilde{\mathbf{x}}) = \prod_{l=1}^{L}\prod_{i\in V^\star} \mathbb{P}_{\psi_l}(X_i^l|X_{\mathcal{N}_i^\star}^l = \tilde{x}_{\mathcal{N}_i^\star}^l), \tag{S.14}$$

for some well chosen fixed configuration $\tilde{\mathbf{x}} = (\tilde{x}_i^l)_{1\leq l\leq L, i\in V^\star}$. Second, the conditional distribution $\mathbb{P}_\psi(\mathbf{X}|\mathbf{Y})$ used for integrating the complete log-likelihood in (S.13) is also replaced by a mean-field approximation, that is

$$\mathbb{P}^2(\mathbf{X}|\psi,\tilde{\mathbf{x}},\mathbf{Y}) = \prod_{l=1}^{L}\prod_{i\in V^\star} \mathbb{P}_{\psi_l}(X_i^l|X_{\mathcal{N}_i^\star}^l = \tilde{x}_{\mathcal{N}_i^\star}^l, Y_i^l). \tag{S.15}$$

Note that both distributions (S.14) and (S.15) are probability distributions, contrarily to what happens when relying on pseudo-likelihoods. Third, the choice of the fixed configuration $\tilde{\mathbf{x}}$ relies on a sequential Gibbs sampling from the approximate distribution (S.15). With these three tools at hand, the algorithm consists in iteratively optimising with respect to $\psi = \{\psi_l\}_{1\leq l\leq L}$ the quantity

$$\mathbb{E}^2\big[\log \mathbb{P}^1(\mathbf{X}|\psi,\tilde{\mathbf{x}})\big|\psi^{(t)},\tilde{\mathbf{x}},\mathbf{Y}\big],$$

computed with the current value of the parameter $\psi^{(t)}$ and current simulated field $\tilde{\mathbf{x}}$. Here, $\mathbb{E}^2$ denotes expectation under the probability distribution $\mathbb{P}^2$. This quantity should be compared to the original criterion (S.13).

Let us now fully describe the procedure. For any current parameter value $\psi^{(t)}$ and fixed state value $\tilde{\mathbf{x}}$, we let

$$\tilde{Q}(\psi|\psi^{(t)},\tilde{\mathbf{x}}) = \sum_{l=1}^{L}\sum_{i\in V^\star}\sum_{x\in\{0,1\}} \mathbb{P}_{\psi^{(t)}}(X_i^l = x|X_{\mathcal{N}_i^\star}^l = \tilde{x}_{\mathcal{N}_i^\star}^l, Y_i^l)\log \mathbb{P}_\psi(X_i^l = x|X_{\mathcal{N}_i^\star}^l = \tilde{x}_{\mathcal{N}_i^\star}^l).$$

The algorithm consists in iterating the following two steps at time t,

- SE-step: sequentially sample a configuration $\tilde{\mathbf{x}}^{(t)}$ as follows for $1\leq l\leq L$ and $1\leq i\leq n$, sample $(X_i^l)^{(t)}$ according to the conditional distribution

$$x \mapsto \mathbb{P}_{\psi^{(t-1)}}\big(X_i^l = x\big|\{X_j^l = (\tilde{x}_j^l)^{(t)}, j\in\mathcal{N}_i^\star, j<i\}, \{X_j^l = (\tilde{x}_j^l)^{(t-1)}, j\in\mathcal{N}_i^\star, j>i\}, Y_i^l\big).$$

Thus, if $C_{il} = 0$ we set $X_i^l = 0$ and whenever $C_{il} = 1$, we sample the value 0 with probability

$$c \exp\Big(\beta_{l,co-abs}^{(t-1)} \sum_{j \in \mathcal{N}_i^\star} A_{ij}^\star C_{jl} \big[1\{(\tilde{x}_j^l)^{(t)} = 0, j < i\} + 1\{(\tilde{x}_j^l)^{(t-1)} = 0, j > i\}\big]\Big) 1\{Y_i^l = 0\} \tag{S.16}$$

and we sample the value 1 with probability

$$c \exp\Big(\alpha_{i,l}^{(t-1)} + \beta_{l,co-pres}^{(t-1)} \sum_{j \in \mathcal{N}_i^\star} A_{ij}^\star \big[1\{(\tilde{x}_j^l)^{(t)} = 1, j < i\} + 1\{(\tilde{x}_j^l)^{(t-1)} = 1, j > i\}\big] + Y_i^l \log(p_{i,l}^{(t-1)}) + (1 - Y_i^l) \log(1 - p_{i,l}^{(t-1)})\Big), \tag{S.17}$$

where c is a normalising constant (set such that the 2 probabilities sum to 1).

- M-step: Optimize $\tilde{Q}(\psi|\psi^{(t)}, \tilde{\mathbf{x}}^{(t)})$ with respect to $\psi = \{\alpha_{i,l}, \beta_{l,co-abs}, \beta_{l,co-pres}\}_{i,l}$.

We now express the quantity $\tilde{Q}$ in our model and derive update formulas in our model. First we set

$$\tilde{p}_{i,l,t}(0) = c \exp\Big(\beta_{l,co-abs}^{(t)} \sum_{j \in \mathcal{N}_i^\star} A_{ij}^\star C_{jl} 1\{(\tilde{x}_j^l)^{(t)} = 0\}\Big) 1\{Y_i^l = 0\}$$

$$\tilde{p}_{i,l,t}(1) = c \exp\Big(\alpha_{i,l}^{(t)} + \beta_{l,co-pres}^{(t)} \sum_{j \in \mathcal{N}_i^\star} A_{ij}^\star 1\{(\tilde{x}_j^l)^{(t)} = 1\} + Y_i^l \log(p_{i,l}^{(t)}) + (1 - Y_i^l) \log(1 - p_{i,l}^{(t)})\Big),$$

with the normalising constant c such that $\tilde{p}_{i,l,t}(0) + \tilde{p}_{i,l,t}(1) = 1$. Then the vector $(\tilde{p}_{i,l,t}(0), \tilde{p}_{i,l,t}(1))$ is nothing else than the probability distribution $\mathbb{P}_{\psi_l^{(t)}}(X_i^l = \cdot | X_{\mathcal{N}_i^\star}^l = \tilde{x}_{\mathcal{N}_i^\star}^l, Y_i^l)$. From this quantity, we obtain

$$\begin{aligned}
&\tilde{Q}(\psi|\psi^{(t)}, \tilde{\mathbf{x}}) \\
&= \sum_{l=1}^{L} \sum_{i \in V^\star} C_{il} \Big\{ \tilde{p}_{i,l,t}(0) \Big[\beta_{l,co-abs} \sum_{j \in \mathcal{N}_i^\star} A_{ij}^\star C_{jl} (1 - \tilde{x}_j^l)\Big] + \tilde{p}_{i,l,t}(1) \Big[\alpha_{i,l} + \beta_{l,co-pres} \sum_{j \in \mathcal{N}_i^\star} A_{ij}^\star \tilde{x}_j^l\Big] \\
&\qquad - \log\Big[\exp\Big(\beta_{l,co-abs} \sum_{j \in \mathcal{N}_i^\star} A_{ij}^\star C_{jl} (1 - \tilde{x}_j^l)\Big) + \exp\Big(\alpha_{i,l} + \beta_{l,co-pres} \sum_{j \in \mathcal{N}_i^\star} A_{ij}^\star \tilde{x}_j^l\Big)\Big]\Big\}.
\end{aligned} \tag{S.18}$$

Optimising this quantity with respect to ψ is done numerically. To this aim, we provide below the derivatives of $\tilde{Q}$ wrt ψ.

Let us introduce the following quantities

$$w_i^\star = \sum_{j \in \mathcal{N}_i^\star} A_{ij}^\star C_{jl},$$

$$w_{i,l}^\star = \sum_{j \in \mathcal{N}_i^\star} A_{ij}^\star \tilde{x}_j^l$$

which are the sum of weights of the neighbours of i in $G^\star$ compatible with the location l and the sum of weights of the neighbours of i in $G^\star$ that are present at location l, respectively. Remembering that $C_{jl}\tilde{x}_j^l = \tilde{x}_j^l$, we have that

$$\sum_{j \in \mathcal{N}_i^\star} A_{ij}^\star C_{jl}(1 - \tilde{x}_j^l) = w_i^\star - w_{i,l}^\star$$

is the sum of weights of the neighbours of i in $G^\star$ that are absent at location l while compatible with that location. We also use

$$\text{den}_{i,l}(\beta_{l,co-abs}, \beta_{l,co-pres}, \alpha_{i,l}) = \exp[\beta_{l,co-abs}(w_i^\star - w_{i,l}^\star)] + \exp(\alpha_{i,l} + \beta_{l,co-pres} w_{i,l}^\star).$$

With these quantities at hand and relying on (S.18), we obtain

$$\begin{aligned}\tilde{Q}(\psi|\psi^{(t)}, \tilde{\mathbf{x}}) = \sum_{l=1}^{L} \sum_{i \in V^\star} C_{il} \Big\{ & \tilde{p}_{i,l,t}(0)\beta_{l,co-abs}(w_i^\star - w_{i,l}^\star) + \tilde{p}_{i,l,t}(1)[\alpha_{i,l} + \beta_{l,co-pres} w_{i,l}^\star] \\ & - \log \text{den}_{i,l}(\beta_{l,co-abs}, \beta_{l,co-pres}, \alpha_{i,l}) \Big\}.\end{aligned}$$

Let us recall that $\alpha_{i,l}$ is a shorthand for the quantity $a_i + a_l + W_l^\intercal b_i + (W_l^2)^\intercal c_i$, so that we finally get, for each $1 \leq l \leq L$ and each $1 \leq i \leq n$, the derivatives

$$\frac{\partial \tilde{Q}}{\partial a_i} = \sum_{l=1}^{L} C_{il} \Big[\tilde{p}_{i,l,t}(1) - \frac{\exp(\alpha_{i,l} + \beta_{l,co-pres} w_{i,l}^\star)}{\text{den}_{i,l}(\beta_{l,co-abs}, \beta_{l,co-pres}, \alpha_{i,l})} \Big] \tag{S.19}$$

$$\frac{\partial \tilde{Q}}{\partial a_l} = \sum_{i \in V^\star} C_{il} \Big[\tilde{p}_{i,l,t}(1) - \frac{\exp(\alpha_{i,l} + \beta_{l,co-pres} w_{i,l}^\star)}{\text{den}_{i,l}(\beta_{l,co-abs}, \beta_{l,co-pres}, \alpha_{i,l})} \Big] \tag{S.20}$$

$$\frac{\partial \tilde{Q}}{\partial b_i} = \sum_{l=1}^{L} C_{il} W_l^\intercal \Big[\tilde{p}_{i,l,t}(1) - \frac{\exp(\alpha_{i,l} + \beta_{l,co-pres} w_{i,l}^\star)}{\text{den}_{i,l}(\beta_{l,co-abs}, \beta_{l,co-pres}, \alpha_{i,l})} \Big] \tag{S.21}$$

$$\frac{\partial \tilde{Q}}{\partial c_i} = \sum_{l=1}^{L} C_{il} (W_l^2)^\intercal \Big[\tilde{p}_{i,l,t}(1) - \frac{\exp(\alpha_{i,l} + \beta_{l,co-pres} w_{i,l}^\star)}{\text{den}_{i,l}(\beta_{l,co-abs}, \beta_{l,co-pres}, \alpha_{i,l})} \Big] \tag{S.22}$$

$$\frac{\partial \tilde{Q}}{\partial \beta_{l,co-abs}} = \sum_{i \in V^\star} C_{il} \left[\tilde{p}_{i,l,t}(0)(w_i^\star - w_{i,l}^\star) - \frac{(w_i^\star - w_{i,l}^\star) \exp[\beta_{l,co-abs}(w_i^\star - w_{i,l}^\star)]}{\text{den}_{i,l}(\beta_{l,co-abs}, \beta_{l,co-pres}, \alpha_{i,l})} \right] \tag{S.23}$$

$$\frac{\partial \tilde{Q}}{\partial \beta_{l,co-pres}} = \sum_{i \in V^\star} C_{il} \left[\tilde{p}_{i,l,t}(1) w_{i,l}^\star - \frac{w_{i,l}^\star \exp[\alpha_{i,l} + \beta_{l,co-pres} w_{i,l}^\star]}{\text{den}_{i,l}(\beta_{l,co-abs}, \beta_{l,co-pres}, \alpha_{i,l})} \right]. \tag{S.24}$$

The simulated field algorithm is described in Algorithm 1.

Remark 2. *In the case with no sampling effects (namely $p_{i,l} = 1$), the simulation step is skipped (since $\mathbf{X} = \mathbf{Y}$), the quantities $\tilde{p}_{i,l,t}$ become $\tilde{p}_{i,l,t}(x) = 1\{X_i^l = x\}$ and the criteria to optimize reduces to the quantity*

$$\tilde{Q}_{direct}(\psi) = \sum_{l=1}^{L} \sum_{i \in V^\star} \sum_{x \in \{0,1\}} \log \mathbb{P}_\psi(X_i^l = x | X_{\mathcal{N}_i^\star}^l).$$

Algorithm 1: Simulated field algorithm

Input: Observed presence/absence data $\mathbf{Y}$, adjacency matrix of metanetwork $A^\star$.
Initialization: Choose initial values $\tilde{\mathbf{x}}^{(0)}, \psi^{(0)}$.
Set $t = 1$.
while not converged **do**
  Simulation step:
  for $1 \leq l \leq L$ **do**
    for $1 \leq i \leq n$ **do**
      Sample $(\tilde{x}_i^l)^{(t)}$ from $\{0, 1\}$ relying on the vector of probabilities (S.16) and (S.17).
    end for
  end for
  Compute $\tilde{Q}(\psi|\psi^{(t)}; \tilde{\mathbf{x}})$ from (S.18).
  Maximization step:
  Compute the value $\hat{\psi}$ zeroing the derivatives (S.19)–(S.24).
  Update parameter $\psi^{(t)} = \hat{\psi}$.
  Increment t.
end while

This means that in this specific case, our method consists exactly in a pseudo-likelihood estimation, which is known to be consistent as the number of observations increases (Besag, 1975). Therefore, the estimation algorithm is more computationally affordable in this case since it consists in a simple iteration of the `M`*-step (i.e. the 'maximization step' in Algorithm 1).*

S.3.4 Additional details on the implementation

The 'maximization step' in Algorithm 1 is performed using the vector Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm implemented in the GNU Scientific Library (`https://www.gnu.org/software/gsl/`). We observed that this algorithm was sensitive to the initial value of the parameters. After analyzing synthetic datasets simulated from the model and estimating the model with various initial values, we validated the following combination of initial parameters:

$$\begin{aligned} a_i = a_l &= \frac{a_0}{2} \\ b_i = c_i &= 0 \\ \beta_{l,co-abs} = \beta_{l,co-pres} &= 0 \end{aligned}$$

with $a_0 = \log(\frac{\bar{Y}}{1-\bar{Y}})$ and $\bar{Y} = \sum_{il} Y_{il}/(nL)$.

S.4 Simulation under ELGRIN model and inference

In order to test the statistical performance of ELGRIN model, we simulated under ELGRIN model and tried to recover the sample parameters. Once the parameters inferred, we simulated new data using ELGRIN again while relying on these inferred values. We then inferred the parameters of this new dataset to test the stability of parameters inference under resampling. HTLM vignette

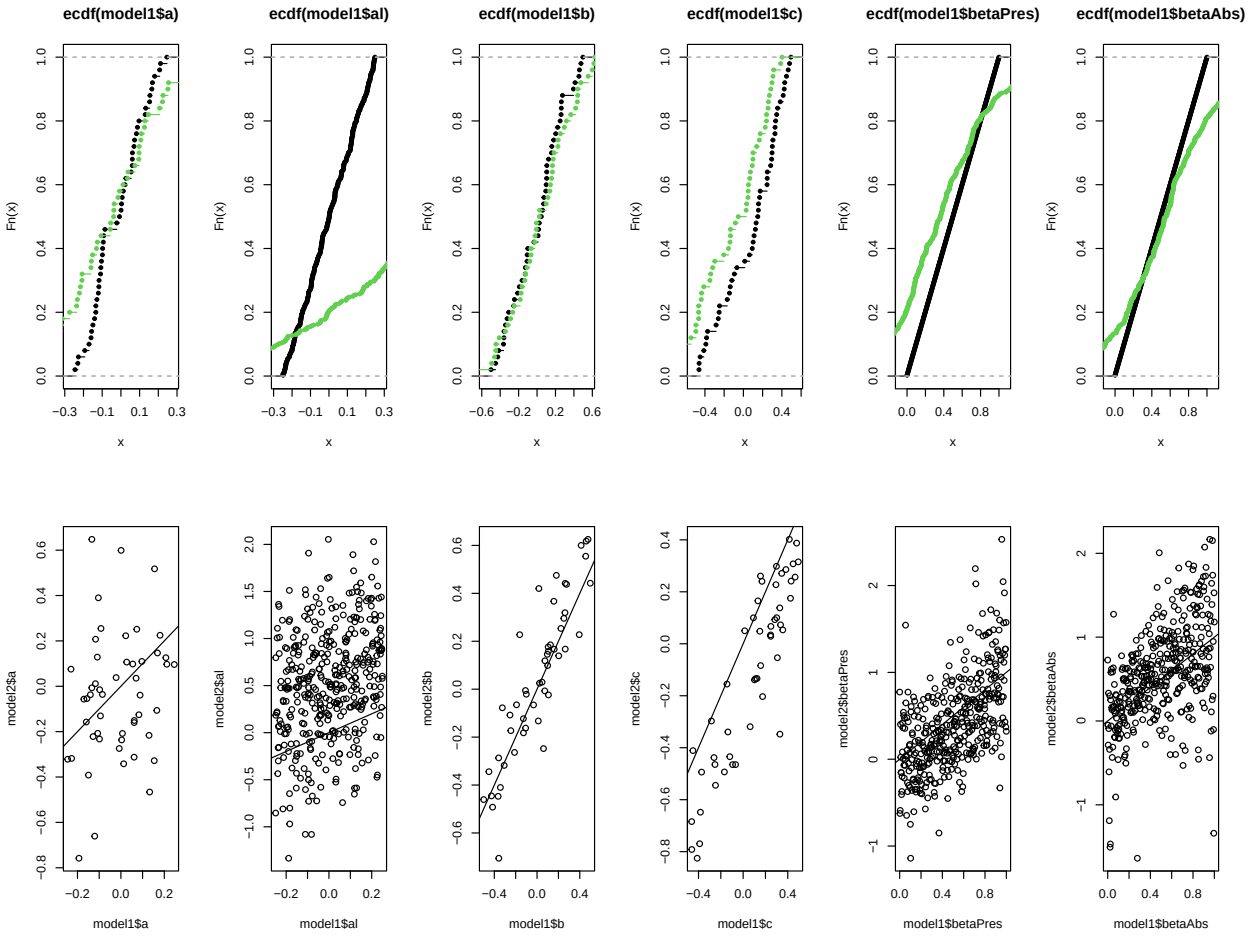


Figure S.8: Comparison between the sample parameters and the inferred parameters using ELGRIN inference algorithm. Top: Cumulative distribution of the sample parameters (model 1, black) and the inferred ones (model 2, green). Bottom: Scatter plot of pairs of parameters (sample and inferred) and the diagonal axis.

is available at `https://plmlab.math.cnrs.fr/econetproject/econetwork/-/blob/master/vignettes/simul_under_elgrin.html`.
We chose the same metaweb and environmental gradient as in the colonisation-extinction simulation (see section S.5.2), with 50 species and used 400 sites. We draw a_i and a_l uniformly in $[-0.25, 0.25]$, b_i and b_l uniformly in $[-0.5, 0.5]$. We chose a gradient of $\beta_{l,co-abs}$ and $\beta_{l,co-pres}$ ranging from 0 to 1.
We represent the comparison of the original model and the inferred model in Figs. S.8 and S.9. The parameters sampled with ELGRIN are reasonably recovered by the inference algorithm and stable under another sampling and inference step.

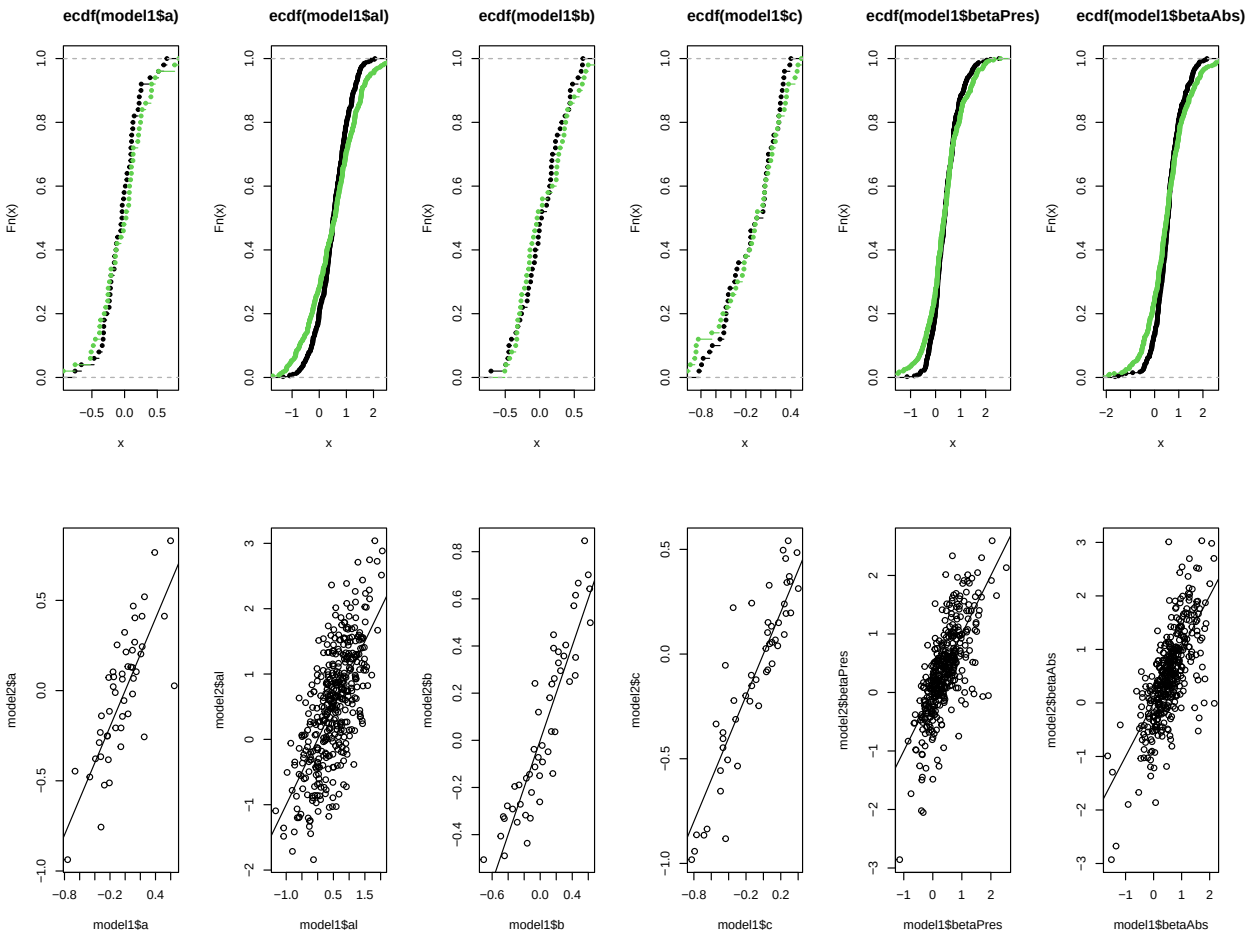


Figure S.9: Comparison between the re-sample parameters (used to sample under ELGRIN model) and the inferred parameters using ELGRIN inference algorithm. Top: Cumulative distribution of the re-sample parameters (model 1, black) and the inferred ones (model 2, green). Bottom: Scatter plot of pairs of parameters (re-sample and inferred) and the diagonal axis.

S.5 Simulations with three different theoretical models

We provide here models and details on the three simulations of the paper. HTLM vignettes for the three simulations are available at `https://plmlab.math.cnrs.fr/econetproject/econetwork`. For each model, we simulated three different scenarios: positive (i.e. mutualism), negative (i.e. competition) or no interactions. The scenario with no interactions uses an empty metanetwork to generate the data. However, inference with ELGRIN in this case relied on a metanetwork with interactions (to be specified below).

In the following, when ELGRIN is fitted on a dataset, the inference procedure outputs estimated parameters values. For any species i, its niche optima was estimated from these values, relying on the optimum of the estimated function $w \mapsto \hat{a}_i + \hat{b}_i w + \hat{c}_i w^2$ within the interval $[\omega_i, \Omega_i]$ defined in (S.6) (here dimension $d = 1$).

S.5.1 Lotka-Volterra model: details and simulation set-up

We sampled species communities from the equilibrium of a deterministic Lotka-Volterra model (Takeuchi, 1996). We defined the environmental niche of each species as a Gaussian distribution centered on a given optimum. The environmental niches optima were evenly taken on a grid whereas the standard deviations were all equal to a given value σ for simplicity.

Building the network from niche values

We constructed the metanetwork $G^\star$ used for generating the data in scenarios with interactions (positive and negative) and later used for inference with ELGRIN in the three scenarios (i.e. including when there are no interactions). Let μ_i and μ_j be the niche optima of two distinct species. We sampled symmetric interaction between species i and j according to a Bernoulli law of parameter $\lambda m |\mu_i - \mu_j|^{-1}$, where $m = \max_{i,j}(|\mu_i - \mu_j|^{-1})$ and λ is a parameter modulating the overall edge number. We obtained a metanetwork $G^\star$ symmetric with no self-loops.

Modelling the dynamics

We assume, for species i, a per-capita growth rate $r_i(w)$ depending on the environment value w and following a Gaussian function of mean μ_i. We then model $N_{iw}(t)$, the abundance of species i at environment value w and time t, using a generalised Lotka-Volterra dynamical model with intraspecific competition. In the negative interactions scenario, we used

$$\frac{1}{N_{iw}} \frac{dN_{iw}}{dt} = r_i(w) - \sum_j C_{ij} N_{jw}, \tag{S.25}$$

where $C_{ij} = A^\star_{ij} + c\mathbf{1}(i = j)$ with $A^\star$ the adjacency matrix of $G^\star$ and c the intraspecific competition coefficient. For the positive interaction scenario, we used

$$\frac{1}{N_{iw}} \frac{dN_{iw}}{dt} = r_i(w) + \sum_j M_{ij} N_{jw}, \tag{S.26}$$

where $M_{ij} = A^\star_{ij}/M_0 - c\mathbf{1}(i = j)$ where M_0 is a constant ($M_0 > 1$) that reduces the strength of positive interactions in order to get convergence towards a finite abundance value. In the no

interactions scenario, we use $C_{ij} = M_{ij} = 0$ for all i, j in the above equations. Note though that we used the simulated metanetwork $G^\star$ for inference with ELGRIN in the three scenarios. From the equilibrium point $\mathbf{N}_w^\star = (N_{iw}^\star)_i$ (with $N_{iw}^\star = \lim_{t\to+\infty} N_{iw}(t)$, limit that is assumed here to be unique and independent of initial conditions), we sample presence or absence X_i^l of each species i at location l using a Bernoulli law of parameter $\min(1, N_{iw_l}^\star/5)$.

Parameter values

We performed simulations with $N = 50$ species and $L = 400$ locations. The environmental niches optima were evenly taken on a grid between -2 and 2 whereas the environmental gradient ranged from -3 to 3. We set the standard deviations of niche distributions to $\sigma = 1$ and we set the intraspecific competition term to $c = 1/10$ for all species. The constant M_0 is set to 50. We ran the simulation of the Lotka-Volterra dynamics for $10,000$ time steps. Fig. S.10 shows growth rates in function of the environment and metanetwork. We also represented the distribution of species presence-absences and species richness under the three interaction scenarios in Fig. S.11 and Fig. S.12. Niche optima inferred from ELGRIN on this dataset are shown in Fig. S.13. Association parameters $\beta_{l,co-abs}$ and $\beta_{l,co-pres}$ are represented in the main text, Fig. 2.

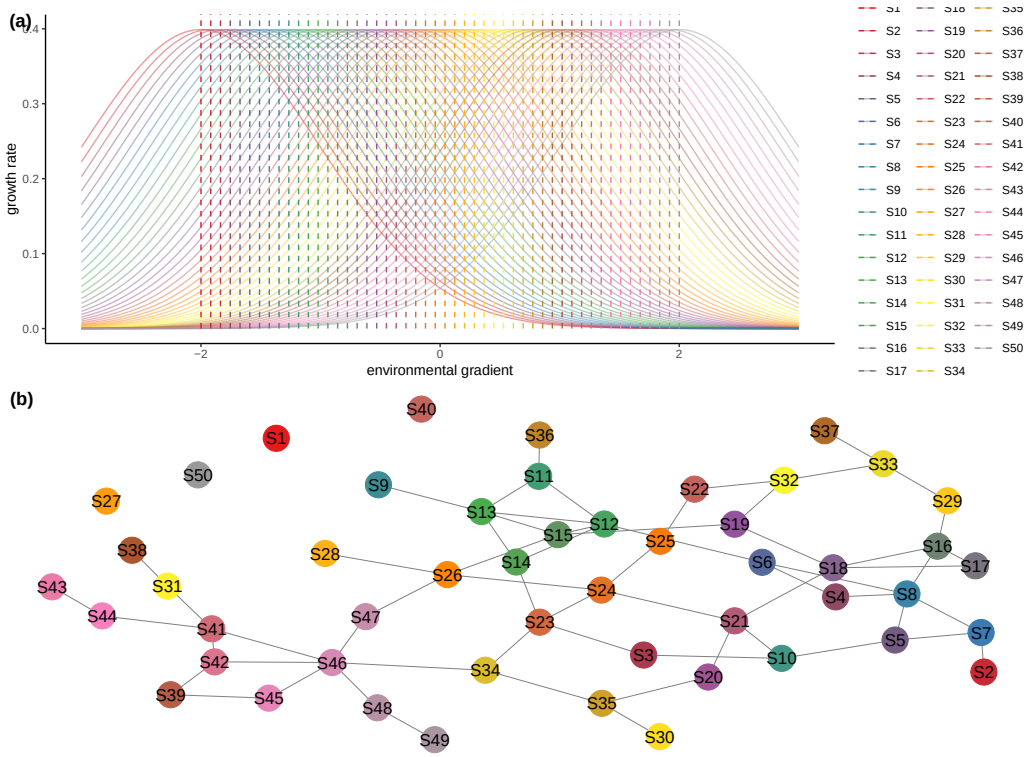


Figure S.10: Simulations under Lotka-Volterra and colonisation-extinction models. (a) Growth rates in function of the environment for the 50 considered species. (b) Representation of the metanetwork used for simulations in the two scenarios with interactions and for estimation with ELGRIN in the three scenarios. Nodes are colored according to the value of niche optima along the environmental gradient.

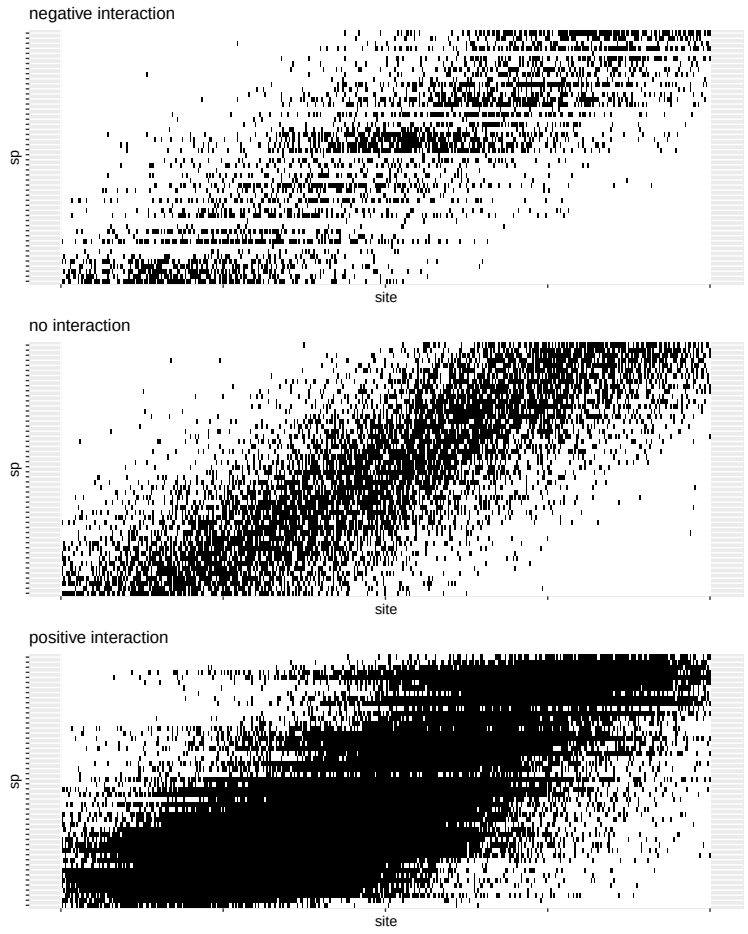


Figure S.11: Presence-absence of species (y-axis) along the environmental gradient (x-axis) for the Lotka-volterra simulations across the three interaction scenarios.

Results and discussion

We notice that species richness increases from the competition to the mutualistic scenarios, as positive interactions enhance the possibility of species to be present (vice-versa for competition). We see that except for the positive interactions scenario, ELGRIN reasonably infers niche optima and association parameters ($\beta_{l,co-abs}$ and $\beta_{l,co-pres}$, as shown in Fig. 2 in the main text) on this community data built from Lotka-Volterra model (see the discussion in the main text for further insights). We however acknowledge a large variance on association parameters for the negative interaction model and as already underlined in the main text, the inability of ELGRIN to identify the positive interactions scenario. We remark that this positive interaction scenario of Lotka-Volterra model is a particularly harsh test for ELGRIN. Indeed, positive interactions increase the effective growth rate, leading to the risk of explosion of the system. For this reason, we were obliged to reduce the overall interaction strength when simulating the data (parameter M_0 in Equation (S.26)), thus reducing their signal in resulting data. Moreover, positive interactions cause species to be present everywhere along their fundamental niche (e.g. Fig. S.11), so that their distribution can be completely explained by the Grinellian part of the model. In other words, the data look exactly as if they

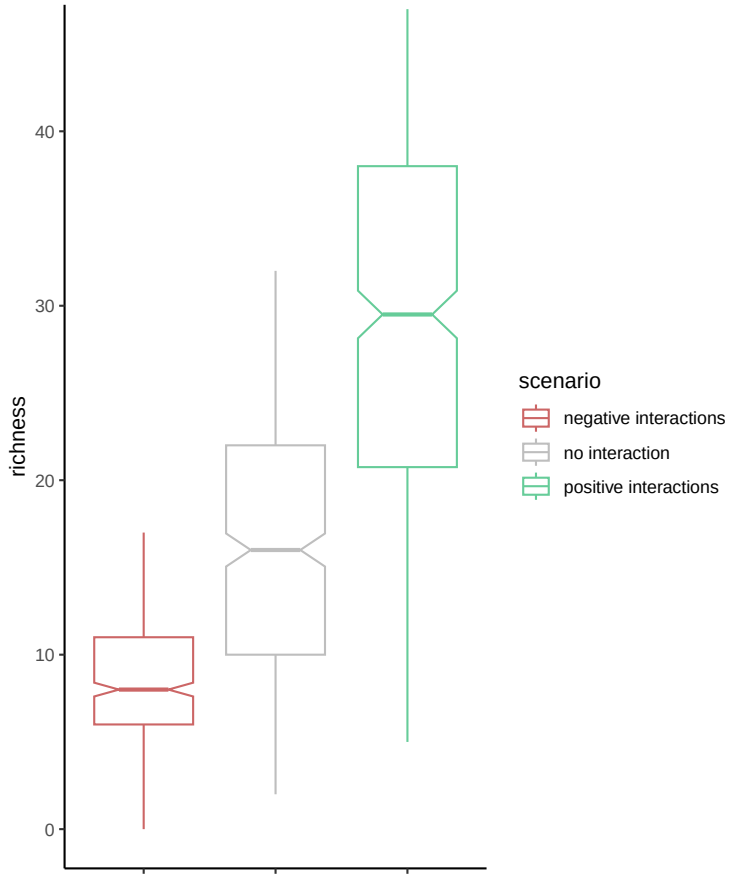


Figure S.12: Distribution of species richness (observed number of present species) for the Lotka-Volterra simulation under the three interaction scenarios.

were obtained from a Grinellian model, where only the environmental variables shape the species distribution. Therefore, no signal is left for the Eltonian part, and the association parameters are inferred to be close to zero.

S.5.2 Colonisation-extinction model: details and simulation set-up

We sampled species communities from the stationary distribution of a stochastic colonisation-extinction model (see Ohlmann et al. 2022). We kept the same environmental gradient, niches and metanetwork as in the Lotka-Volterra simulation. We also combined this model with the three interaction scenarios: negative interactions (i.e. competition), positive interactions (i.e. mutualism) or no interactions.

Modelling the dynamics

We note X_i^t the binary random variable associated to presence of species i at discrete time t and w the value of the environmental gradient. We model niche and interaction effects trough conditional probabilities.

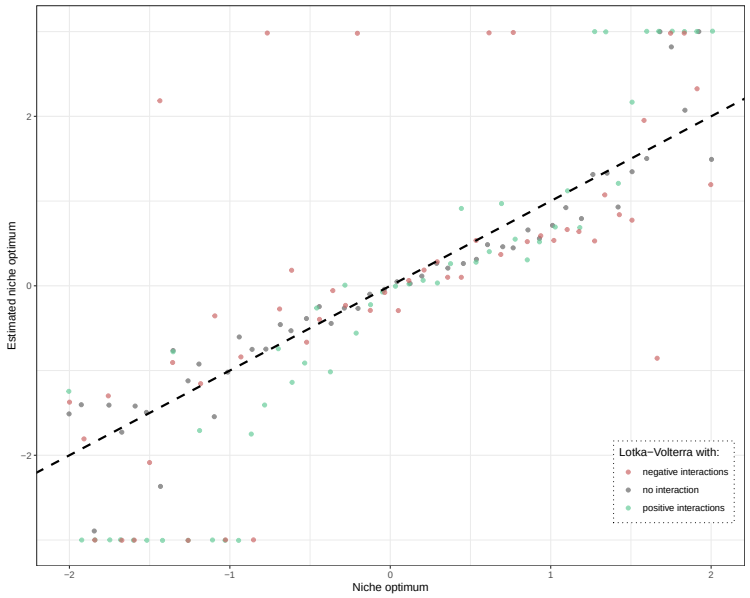


Figure S.13: Estimated niche optima versus true niche optima for the Lotka-Volterra simulation under the three interaction scenarios.

Colonisation-extinction model without interaction In this scenario, we assume that interaction effects do not impact colonisation or extinction. Extinction probability p_e is constant whereas colonisation probability $c_i(w)$ for species i depends on the value of the environmental gradient w only

$$P(X_i^{t+1} = 1 | \{X_j^t\}_j, w) = P(X_i^{t+1} = 1 | X_i^t, w) \propto c_i(w)(1 - X_i^t) + (1 - p_e)X_i^t,$$

where $c_i(w)$ is a Gaussian function with mean μ_i and variance η^2 and $\propto$ means up to a normalizing constant. We simulated this Markov chain and sampled from the stationary distribution to generate a joint species distribution.

Colonisation-extinction model with positive and negative interactions In these scenarios, we assume that both abiotic niche effects and interspecific interactions do impact colonisation-extinction processes. Environmental gradient modulates colonisation probability whereas interactions modulate both colonisation and extinction probabilities.

For the positive interactions scenario, we have:

$$\begin{aligned} P(X_i^{t+1} = 1 | \{X_j^t\}_j, w) &= P(X_i^{t+1} = 1 | X_i^t, X_{N(i)}^t, w) \\ &\propto c_i(w) \exp\Big(\frac{\sum_{k \in N(i)} X_k^t}{|N(i)|}\Big)(1 - X_i^t) + \Big[1 - p_e \exp\Big(-\frac{\sum_{k \in N(i)} X_k^t}{|N(i)|}\Big)\Big] X_i^t, \end{aligned}$$

and for the negative interactions scenario

$$\begin{aligned} P(X_i^{t+1} = 1 | \{X_j^t\}_j, w) &= P(X_i^{t+1} = 1 | X_i^t, X_{N(i)}^t, w) \\ &\propto c_i(w) \exp\Big(-\frac{\sum_{k \in N(i)} X_k^t}{|N(i)|}\Big)(1 - X_i^t) + \Big[1 - p_e \exp\Big(\frac{\sum_{k \in N(i)} X_k^t}{|N(i)|}\Big)\Big] X_i^t, \end{aligned}$$

where $N(i)$ is the set of neighbour species of species i in the metanetwork. Similarly as for the no interaction scenario, we sampled the species co-occurrences in the stationary distributions of each of these scenarios.

Parameter values

We performed simulations with $N = 50$ species and $L = 400$ locations. Extinction probability was set to $p_e = 2\%$ and colonisation probability $c_i(w)$ is Gaussian with mean μ_i and standard deviation $\eta = 1$. We ran each simulation dynamics for $3,000$ time steps. We represented the distribution of species presence-absences and species richness under the three interaction scenarios in Fig. S.14 and Fig. S.15. Niche optima inferred from ELGRIN on this dataset are shown in Fig. S.16. Association parameters $\beta_{l,co-abs}$ and $\beta_{l,co-pres}$ are represented in the main text, Fig. 3.

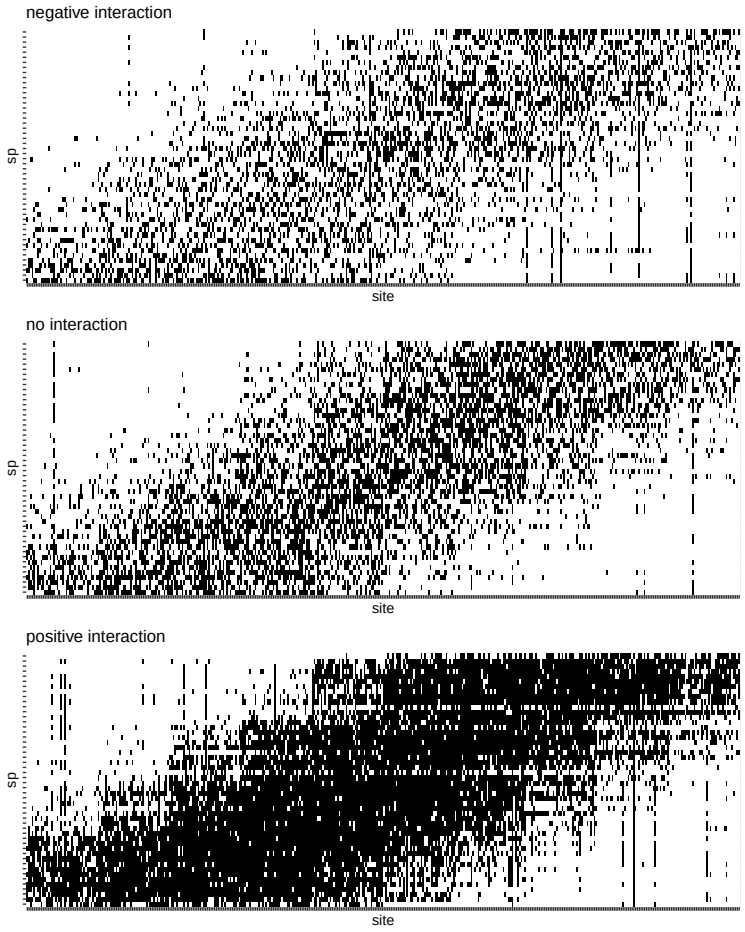


Figure S.14: Presence-absence of species (y-axis) along the environmental gradient (x-axis) for the colonisation-extinction simulations, across the three interaction scenarios.

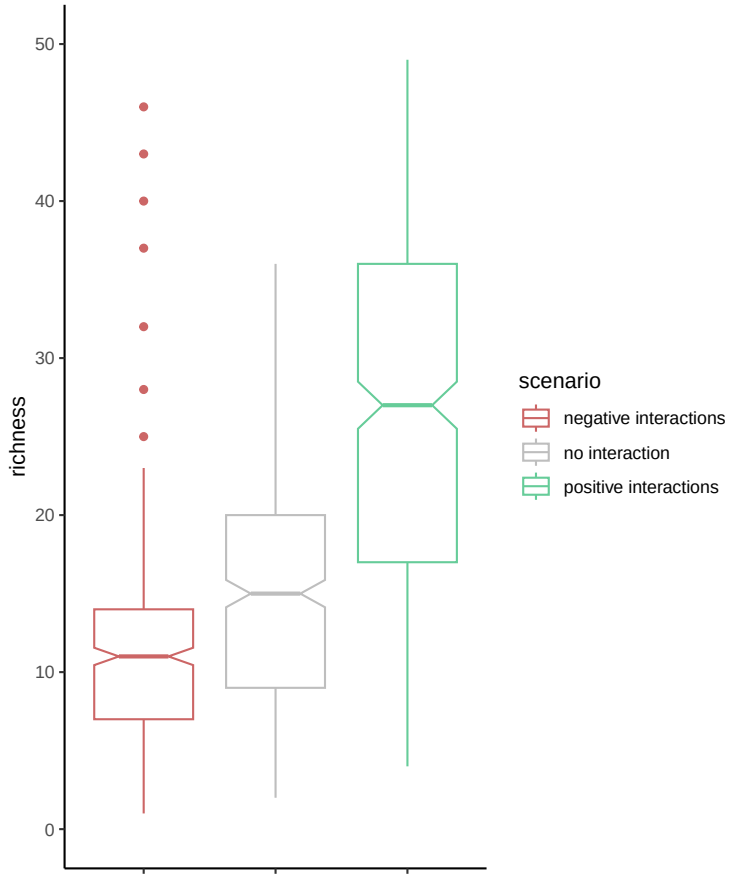


Figure S.15: Distribution of species richness (observed number of present species) for the colonisation-extinction simulation under the three interaction scenarios.

Results and discussion

We notice that species richness increases from the competition to the mutualistic scenarios, as positive interactions enhance the possibility of species to be present (vice-versa for competition, Fig. S.15). For each scenario, the distribution of association parameters ($\beta_{l,co-pres}$ and $\beta_{l,co-abs}$, see Fig.3 in the main text) have a negative median for negative interactions, a median close to zero for the case without interaction and a positive median for positive interactions. The sign of inferred (static) association parameters is the same as the sign of dynamic interaction parameters.

Moreover, ELGRIN correctly infers niche optima in the three interaction scenarios (Fig S.16). Consequently, on these simulations, ELGRIN separates environmental effects for biotic interactions (see the discussion in the main text for further insights).

S.5.3 VirtualCom model: details and simulation set-up

We considered N species in the species pool and L communities to simulate (i.e. the number of locations). We defined the environmental niche (or preference) of each species as a Gaussian distribution centered on a given optimum. The environmental niches optima were regularly taken

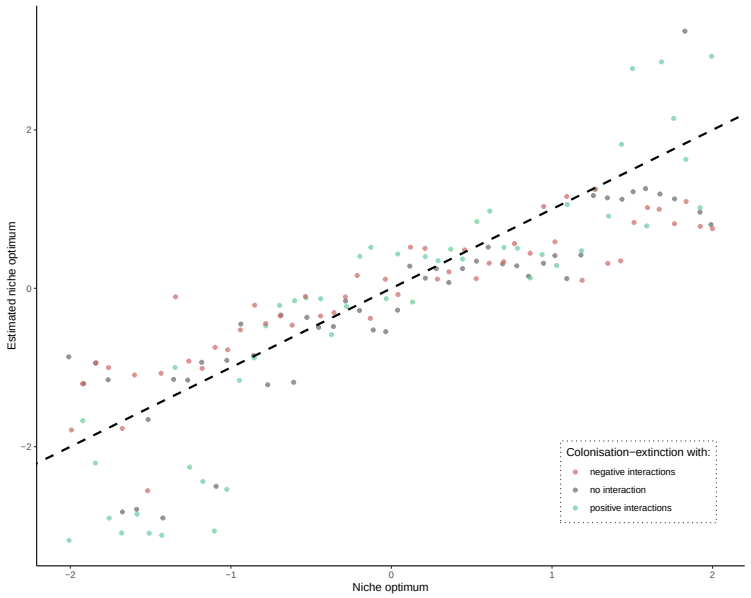


Figure S.16: Estimated niche optima versus true niche optima for the colonisation-extinction simulation under the three interaction scenarios.

on a grid between -2 and 2, whereas the standard deviations were all equal to a given value σ for simplicity. Each community or location l has the same carrying capacity K (i.e. the exact number of individuals in each location).

Building the interaction networks from niche values

Here we constructed two different metanetworks $G^\star$ used to simulate data in the two interaction scenarios. Let μ_i and μ_j be the niche optima of two species and σ the standard deviation of their niche. We considered that the two considered species potentially interact in the mutualistic metanetwork if $\sigma < |\mu_i - \mu_j| < 2\sigma$. Regarding competition, we considered that the two species potentially compete if they share the same environmental niche, and thus if $|\mu_i - \mu_j| < \sigma$. Among all potential species interactions, we randomly sampled 50% of them for both competition and mutualism. Inference with ELGRIN in the scenarios with interactions relied on the respective metanetworks used for simulation. In the no interaction scenario, ELGRIN inference relied on the positive interactions metanetwork (corresponding to mutualism).

Modelling the dynamics

The community assembly process was randomly initialized with a set of individuals that were randomly selected in the species pool until the carrying capacity K was reached. At each time step, the probability of an individual from species i to replace a random individual of the community l is R_{il}. This probability depends on how the environmental conditions at location l are suitable for species i (environmental filter) and on the number of individuals present in community l that interact with species i (competition or mutualism filter). More precisely, we consider the following

equation defining the relative importance of environmental and biotic filters respectively:

$$R_{il} \propto \exp\left[\gamma_{env} \log(p_{il}^{env}) \ + \gamma_{inter} \log(p_{il}^{inter})\right],$$

where γ_{env} and γ_{inter} are tuning parameters giving weights to abiotic and biotic components, and p_{il}^{env} and p_{il}^{inter} are probabilities of species replacement with different filters. The probability p_{il}^{env} accounts for the environmental filtering and is a rescaled density of the Gaussian niche of species i at the environmental value of location l (the scaling ensures this value ranges in $[0, 1]$). When the environment in community l is suitable to species i, the probability that this species enters this community becomes high.

We then have a term dealing with species interactions. In the no interaction scenario, the constant γ_{il} is set to 0. Otherwise, the interaction term is set as

$$p_{il}^{inter} = \begin{cases} K^{-1} \sum_{j;(i,j)\in E^\star} K_{jl} & \text{for mutualism,} \\ 1 - K^{-1} \sum_{j;(i,j)\in E^\star} K_{jl} & \text{for competition,} \end{cases}$$

where K_{jl} is the number of individuals of species j in community l, such that the total carrying capacity $K = \sum_j K_{jl}$. In case of mutualism, the larger number of individuals of species connected with i in the metanetwork are present in location l, the higher is the probability of an individual of species i to enter the community. For competition, the opposite effect is induced. The tuning parameters γ_{env} and γ_{inter} weight the relative importance of the different filters. This algorithm updates the communities until an equilibrium is reached. To assess the equilibrium state, we calculated the Shannon diversity for each location over time, and checked for convergence. Lastly, we deduced species presence/absence by examining species composition in each location.

Parameter values

We performed simulations with $N = 50$ species and $L = 400$ locations, with a carrying capacity of $K = 40$ individuals. The standard deviations of the Gaussian niche distributions were set to $\sigma = 1$ for all species. We chose $\gamma_{env} = 1$ and $\gamma_{metanetwork} = 10$ in case of competition and 5 in case of mutualism. Fig. S.17 shows growth rates in function of the environment and the two metanetworks (for positive and negative interactions). We simulated 100 time steps such that the algorithm convergence was achieved in practice. We repeated the whole procedure 10 times and verified that we obtained equivalent qualitative results. Simulations were implemented with R version 3.6.2 and a modified version of the VirtualCom package. We represented the distribution of species presence-absences and species richness under the three interaction scenarios in Fig. S.18 and Fig. S.19. Niche optima inferred from ELGRIN on this dataset are shown in Fig. S.20. Association parameters $\beta_{l,co-abs}$ and $\beta_{l,co-pres}$ are represented in the main text, Fig. 4.

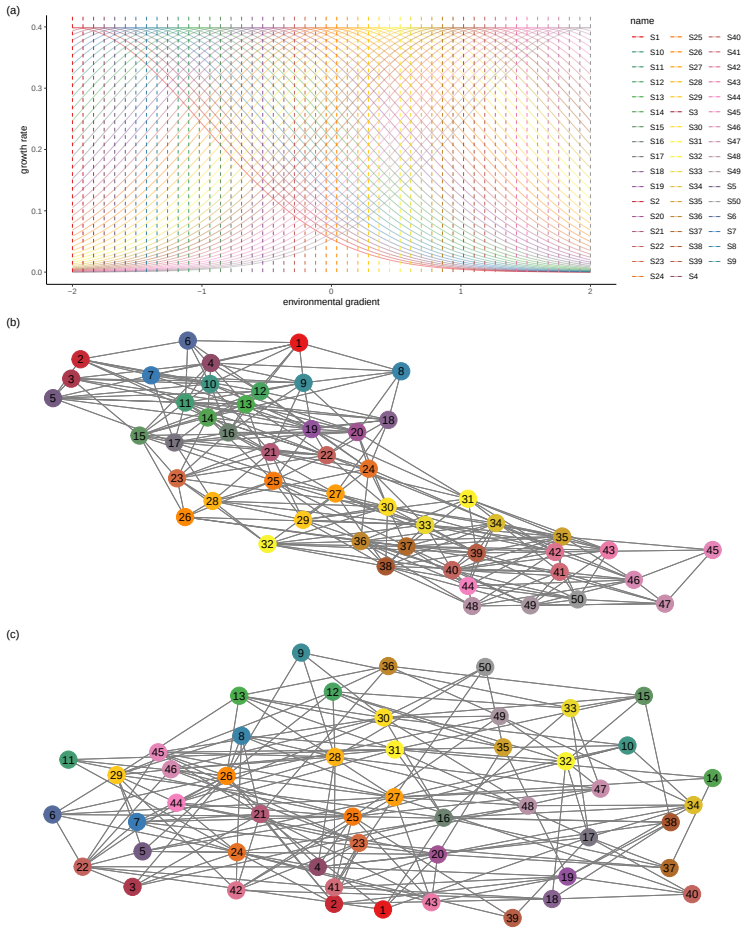


Figure S.17: Simulations under VirtualCom model. (a) Growth rates in function of the environment for the 50 considered species. Representation of the metanetworks used for simulations of VirtualCom in the facilitation case (b) and in the competition case (c). Nodes are colored according to the value of niche optima along the environmental gradient.

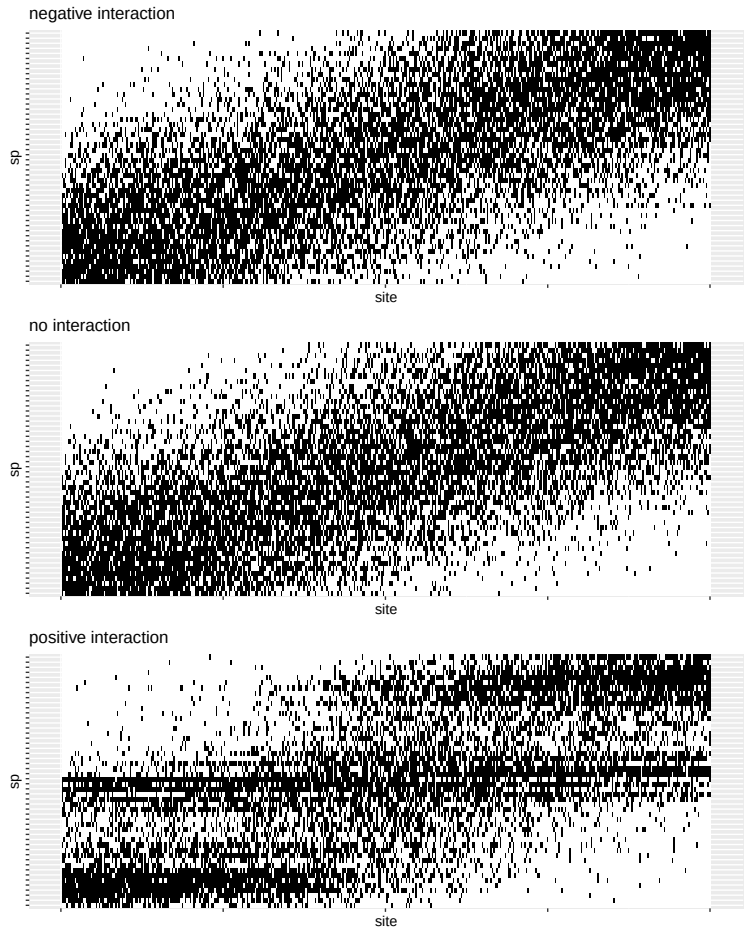


Figure S.18: Presence-absence of species (y-axis) along the environmental gradient (x-axis) for the VirtualCom simulations across the three interaction scenarios.

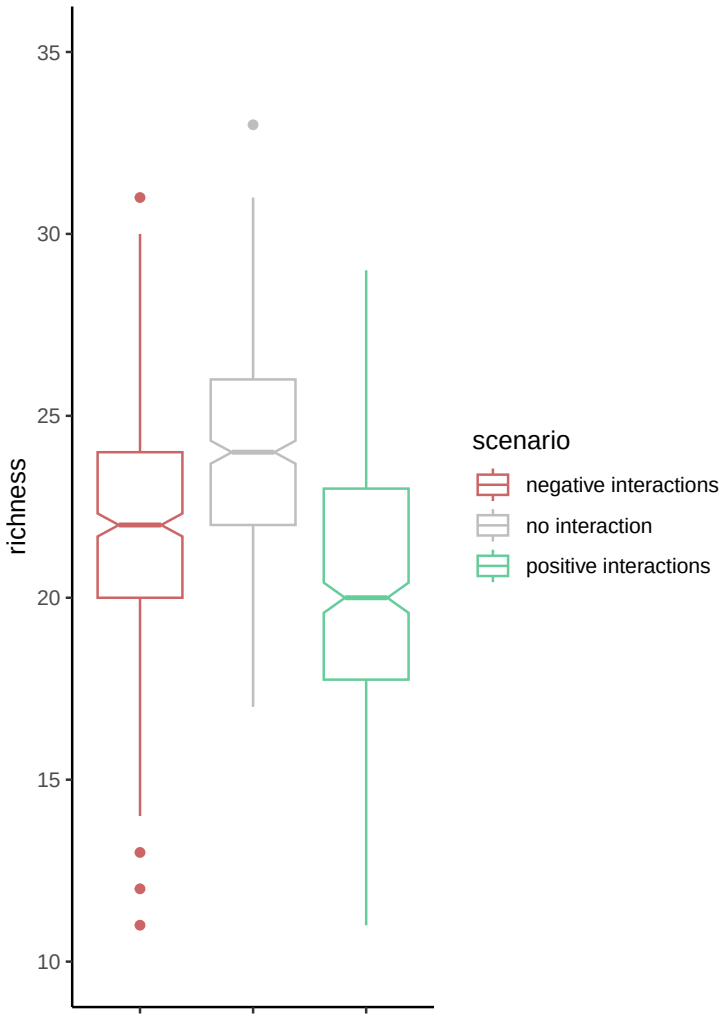


Figure S.19: Distribution of species richness (observed number of present species) for the VirtualCom simulation under the three interaction scenarios.

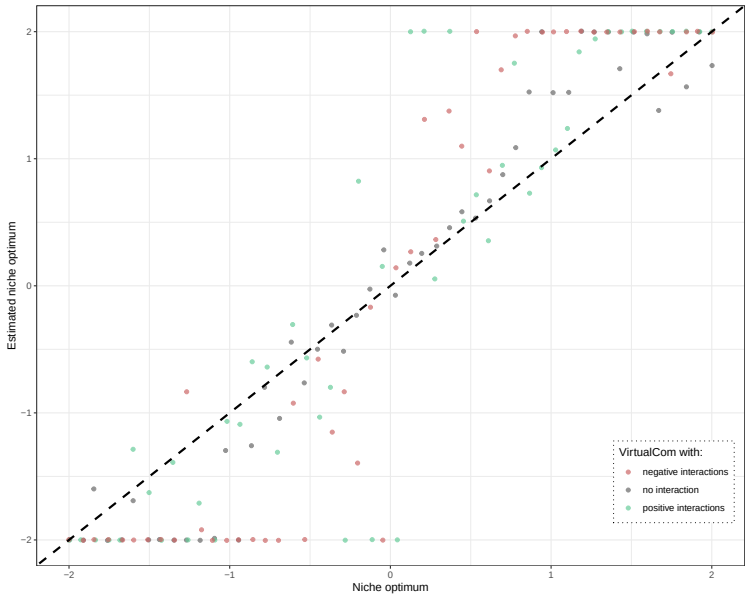


Figure S.20: Estimated niche optima versus true niche optima for the VirtualCom simulation under the three interaction scenarios.

Results and discussion

We notice here that species richness is lower in the facilitation case than in the other cases. This might seem counter intuitive, but comes from the constraint of VirtualCom of keeping fixed the number of individual at each site. Therefore, positive interactions tend to produce communities with a lower number of species, since the few species that facilitate each other and that can survive at the given environmental conditions keep enhancing their probability of presence and cannot be replaced by other species. Instead, the cases with negative interactions, or without interactions, reduces the probability of competitive species, thus favoring all the other species to replace them, leading to an overall higher richness. We see that ELGRIN reasonably infers the niche parameters and association parameters ($\beta_{l,co-pres}$ and $\beta_{l,co-abs}$, see Fig.4 in the main text) on this community data built from VirtualCom model (see the discussion in the main text for further insights). We however acknowledge a large variance on association parameters for the negative interactions model, which could be due to the fact that the VirtualCom model does not express as a ELGRIN one. In the no interactions scenario, we correctly infer that the association parameters under ELGRIN model are estimated around zero.

S.5.4 Kolmogorov-Smirnov tests on association parameters

To quantitatively investigate the difference between $\beta_{l,co-pres}$ and $\beta_{l,co-abs}$ distributions in the three simulations, we performed Kolmogorov-Smirnov tests. For each simulation, we tested whether $\beta_{l,co-pres}$ and $\beta_{l,co-abs}$ distributions were significantly greater (resp. lower) in the scenarios with positive interactions (resp. negative) from the scenario without interactions. Namely, denoting β^{pos} (resp. $\beta^{\text{no int}}$ and β^{neg}) the β values inferred under the positive interaction scenario (resp. no interactions and negative interactions) and $F_{\beta^{\text{pos}}}$ (resp. $F_{\beta^{\text{no int}}}$ and $F_{\beta^{\text{neg}}}$) the corresponding cdfs,

Results of Kolmogorov-Smirnov tests		
Simulation and parameter	p-value of the test $H_0 : F_{\beta^{\text{pos}}} = F_{\beta^{\text{no int}}}$ against $H_1 : F_{\beta^{\text{pos}}} \geq F_{\beta^{\text{no int}}}$	p-value of the test $H_0 : F_{\beta^{\text{neg}}} = F_{\beta^{\text{no int}}}$ against $H_1 : F_{\beta^{\text{neg}}} \leq F_{\beta^{\text{no int}}}$
LV, $\beta_{l,co-pres}$	0.0020	$< 2.2\text{e}{-16}$
LV, $\beta_{l,co-abs}$	0.0040	$< 2.2\text{e}{-16}$
CE, $\beta_{l,co-pres}$	$< 2.2\text{e}{-16}$	$< 2.2\text{e}{-16}$
CE $\beta_{l,co-abs}$	$< 2.2\text{e}{-16}$	$< 2.2\text{e}{-16}$
VC, $\beta_{l,co-pres}$	$7.8\text{e}{-16}$	$< 2.2\text{e}{-16}$
VC, $\beta_{l,co-abs}$	$< 2.2\text{e}{-16}$	$< 2.2\text{e}{-16}$

Table 3: Results of Kolmogorov-Smirnov tests on association parameters ($\beta_{l,co-abs}$ and $\beta_{l,co-pres}$) in the three simulations settings: Lotka-Volterra (LV), Colonisation-extinction (CE) and VirtualCom (VC). The tests compare the distribution of association parameters between the positive interactions and the no interaction scenario (second column) and also between the negative interactions scenario and the no interaction scenario (third column).

we tested in the positive scenario the null hypothesis $H_0 : F_{\beta^{\text{pos}}} = F_{\beta^{\text{no int}}}$ against the alternative $H_1 : F_{\beta^{\text{pos}}} \geq F_{\beta^{\text{no int}}}$ (this corresponds to stochastic ordering). In the same way, for the negative interaction scenario, we tested $H_0 : F_{\beta^{\text{neg}}} = F_{\beta^{\text{no int}}}$ against the alternative $H_1 : F_{\beta^{\text{neg}}} \leq F_{\beta^{\text{no int}}}$. We recall the concept of stochastic ordering for 2 random variables U, V: we have that 'U is stochastically larger than V', denoted $U \succeq V$ when the corresponding cdfs satisfy $F_U \geq F_V$ which in practice corresponds to the fact that a random observation from U 'tends to be larger' than one from V.

In the three simulations, the tests correctly identify significant differences between interactions and no interaction scenarios (Table 3). For the Lotka-Volterra simulation, the p-values for the comparison between the positive interaction scenario and the no-interaction scenario were slightly greater than the p-values of the other tests, in accordance to the qualitative assessment of the simulation results.

S.6 Simulation beyond model assumptions

We provide here a test of our model when species communities are simulated based on processes that are not accounted for by ELGRIN. In particular, we take the example of Lotka-Volterra models (see Section S.5.1) where intraspecific interactions are higher than interspecific ones. ELGRIN does not account for these intraspecific interactions (i.e., it does not model self-loops), and we might thus expect that it will struggle in correctly retrieving model parameters. An HTLM vignette for this simulation is available at `https://plmlab.math.cnrs.fr/econetproject/econetwork`.

S.6.1 Simulation set-up

We set-up simulations accordingly to Section S.5.1. We simulate three different scenarios: positive (i.e. mutualism), negative (i.e. competition) or no interactions, using the same interaction network, niche optima (Fig. S.10) and simulation parameters. However, we increase the intraspecific com-

petition term (i.e., parameter c in Equation (S.25)), from $1/10$ to 2, in order to make it stronger than interspecific interactions. The distribution of species presence-absences and species richness under the three interaction scenarios is represented in Fig. S.21 and Fig. S.22. Niche optima inferred from ELGRIN on this dataset are shown in Fig. S.23. Inferred association parameters $\beta_{l,co-abs}$ and $\beta_{l,co-pres}$ are represented in Fig. S.24.

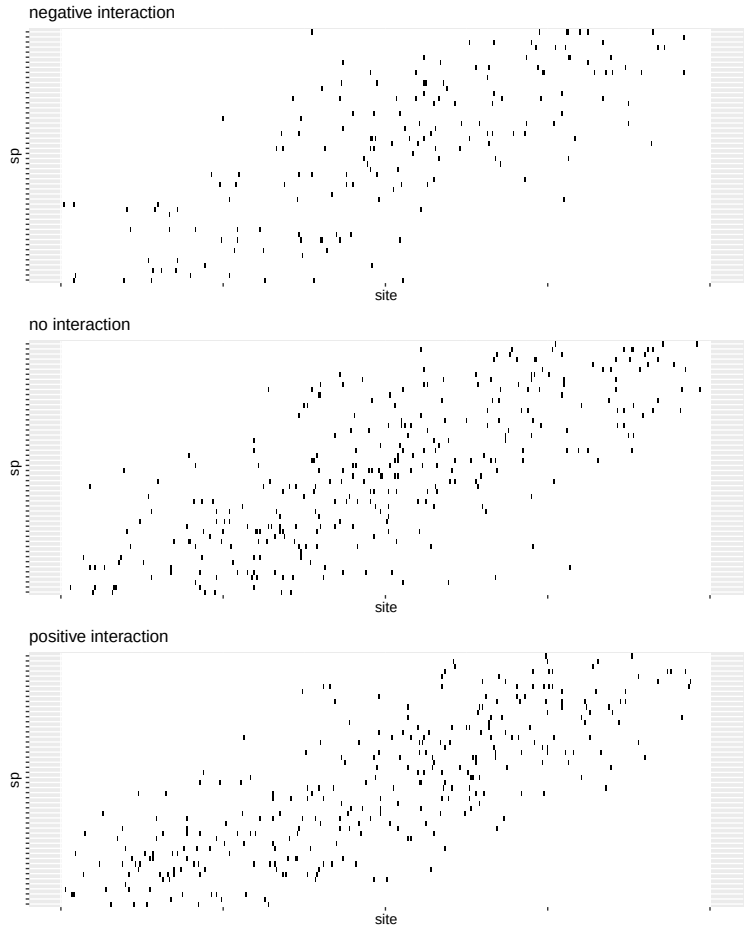


Figure S.21: Presence-absence of species (y-axis) along the environmental gradient (x-axis) for the Lotka-Volterra simulations with intraspecific interactions larger than interspecific ones, across the three interaction scenarios.

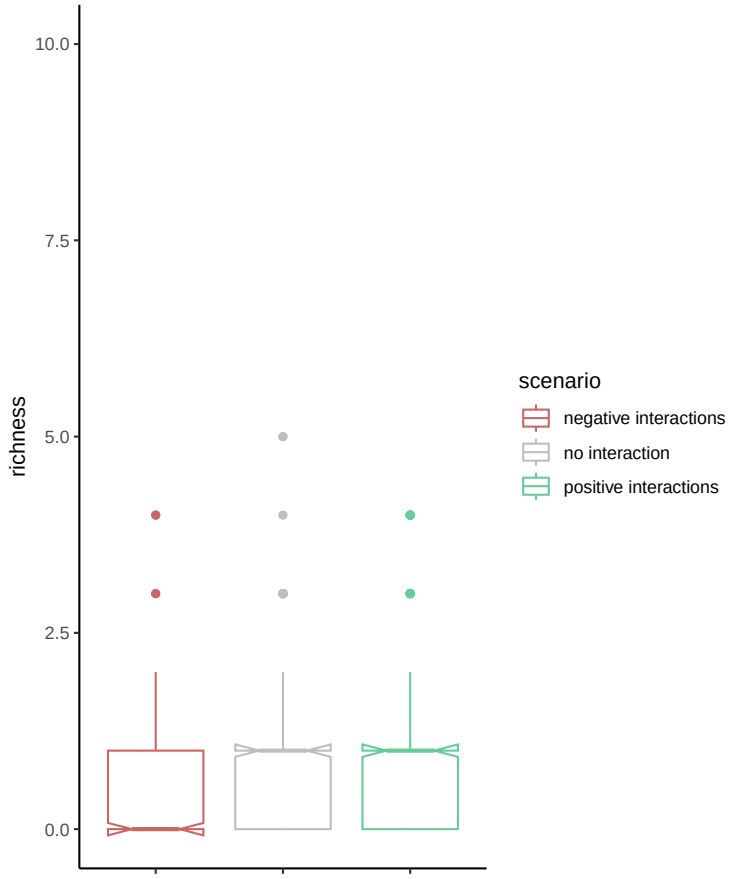


Figure S.22: Distribution of species richness (observed number of present species) for the Lotka-Volterra simulation with intraspecific interactions larger than interspecific ones, under the three interaction scenarios.

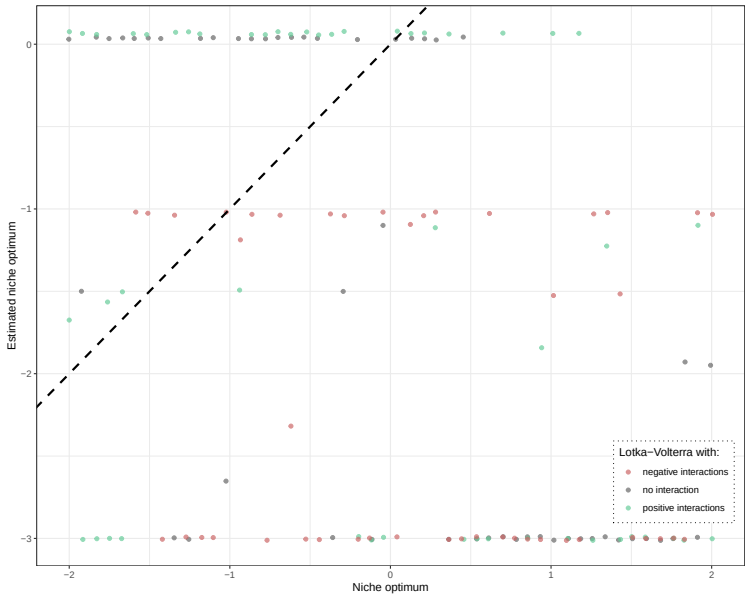


Figure S.23: Estimated niche optima versus true niche optima for the Lotka-Volterra simulation with intraspecific interactions larger than interspecific ones, under the three interaction scenarios.

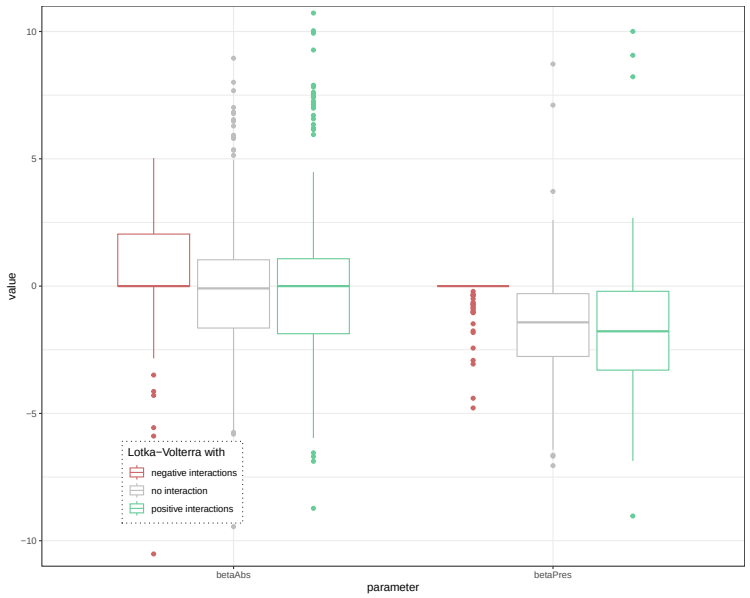


Figure S.24: Distribution of co-absence $\beta_{l,co-abs}$ and co-presence $\beta_{l,co-pres}$ strengths inferred using ELGRIN on simulated ecological communities using a Lotka-Volterra model with intraspecific interactions larger than interspecific ones, under the three interaction scenarios.

Results and discussion

We notice that mean species richness increases from the competition to the mutualistic scenarios, as positive interactions enhance the possibility of species to be present (vice-versa for competition). However, the community matrices are very sparse and species richness is overall very low for the three different scenarios (Fig. S.22). Overall, ELGRIN does not correctly infer model parameters on this community data built from Lotka-Volterra model with large intraspecific interactions. Niche optima are badly estimated (Fig S.23) and the inferred association parameters do not show the expected patterns ($\beta_{l,co-abs}$ and $\beta_{l,co-pres}$, as shown in Fig. S.24). We see that ELGRIN cannot correctly disentangle between the three different simulated scenarios. Indeed, we might expect $\beta_{l,co-abs}$ and $\beta_{l,co-pres}$ parameters to be negative in the negative interaction case, positive in the positive interaction one, and around zero in the no-interaction case. This is generally not the case here, where the inferred β parameters are generally close to zero and are lower for the positive interactions scenario than for the negative one. The poor performance of ELGRIN when intraspecific interactions are higher than interspecific ones is not surprising. As discussed in the main text it is possible to simulate species distributions on which ELGRIN will fail in recovering the true underlying generation process, because these datasets simply do not show anymore enough information about the process that generated them, and could be the result of a completely different scenario, in particular the one inferred by ELGRIN. As such, when using ELGRIN - or any other statistical model - we must bear in mind its model assumptions, knowing that inference might be blurred when other processes are at play.

S.7 Empirical case study

S.7.1 Relation between $\beta_{l,co-pres}$ and $\beta_{l,co-abs}$

Fig. S.25 shows the correlation between the values $\beta_{l,co-pres}$ and $\beta_{l,co-abs}$ estimated through ELGRIN on the European tetrapods case study.

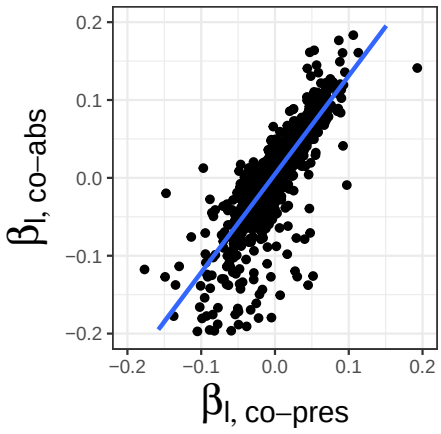


Figure S.25: Results of ELGRIN on the European tetrapods case study. The parameters $\beta_{l,co-pres}$ and $\beta_{l,co-abs}$ were highly correlated.

S.7.2 Kolmogorov-Smirnov tests on the distributions

We performed the following tests on the $\beta_{l,co-pres}$ parameters estimated from the data:

- Denoting $F_{\beta^{\text{high alt}}}$ (resp. $F_{\beta^{\text{low alt}}}$) the cdf of the $\beta_{l,co-pres}$ values inferred at locations with altitude above 1600m (resp. below 1600m), we tested the null hypothesis $H_0 : F_{\beta^{\text{high alt}}} = F_{\beta^{\text{low alt}}}$ against the alternative $H_1 : F_{\beta^{\text{high alt}}} \leq F_{\beta^{\text{low alt}}}$. The resulting p-value is inferior to $2.2\text{e}{-16}$.

- Denoting $F_{\beta^{\text{high richness}}}$ (resp. $F_{\beta^{\text{low richness}}}$) the cdf of the $\beta_{l,co-pres}$ values inferred at locations with richness larger than 200, we tested the null hypothesis $H_0 : F_{\beta^{\text{high richness}}} = F_{\beta^{\text{low richness}}}$ against the alternative $H_1 : F_{\beta^{\text{high richness}}} \geq F_{\beta^{\text{low richness}}}$. The resulting p-value is inferior to $2.2\text{e}{-16}$.

- Denoting $F_{|\beta^{\text{high connect}}|}$ (resp. $F_{|\beta^{\text{low connect}}|}$) the cdf of the $|\beta_{l,co-pres}|$ values inferred at locations with connectance larger than its median value (0.062), we tested the null hypothesis $H_0 : F_{|\beta^{\text{high connect}}|} = F_{|\beta^{\text{low connect}}|}$ against the alternative $H_1 : F_{|\beta^{\text{high connect}}|} \geq F_{|\beta^{\text{low connect}}|}$. The resulting p-value is inferior to $2.2\text{e}{-16}$.

References for Appendix

Berlow, E. L., A.-M. Neutel, J. E. Cohen, P. C. De Ruiter, B. Ebenman, M. Emmerson, J. W. Fox, V. A. A. Jansen, J. Iwan Jones, G. D. Kokkoris, D. O. Logofet, A. J. McKane, J. M. Montoya, and O. Petchey (2004). Interaction strengths in food webs: issues and opportunities. *Journal of Animal Ecology 73*(3), 585–598.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological) 36*(2), 192–225.

Besag, J. (1975). Statistical analysis of non-lattice data. *The statistician 24*(3), 179–195.

Celeux, G., F. Forbes, and N. Peyrard (2003). EM procedures using mean field-like approximations for Markov model-based image segmentation. *Pattern Recognition 36*(1), 131–144.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B 39*, 1–38.

Krause, A. E., K. A. Frank, D. M. Mason, R. E. Ulanowicz, and W. W. Taylor (2003). Compartments revealed in food-web structure. *Nature 426*(6964), 282–285.

Lauritzen, S. L. (1996). *Graphical models*, Volume 17 of *Oxford Statistical Science Series*. New York: The Clarendon Press Oxford University Press. Oxford Science Publications.

Ohlmann, M., F. Munoz, F. Massol, and W. Thuiller (2022). Assessing mutualistic metacommunity capacity by integrating spatial and interaction networks. Technical report, arXiv:2206.11029.

Takeuchi, Y. (1996). *Global Dynamical Properties of Lotka-Volterra Systems*. Singapore: World Scientific.