Uncovering bidirectional ecological associations from co-occurrence and environmental data

Learning bidirectional ecological associations [running headline]

KEYWORDS: Biotic associations, ecological networks, habitat suitability models, network inference, plant interactions, representation learning, simulated communities.

AUTHORS:

Sara Si-moussi sara.si-moussi@inrae.fr
INRAE, IRD, CIRAD, Montpellier SupAgro, Université Montpellier, UMR Eco&Sols

Esther Galbrun esther.galbrun@uef.fi
School of Computing, University of Eastern Finland

Mickaël Hedde mickael.hedde@inrae.fr
INRAE, IRD, CIRAD, Montpellier SupAgro, Université Montpellier, UMR Eco&Sols

Wilfried Thuiller wilfried.thuiller@univ-grenoble-alpes.fr
Univ. Grenoble Alpes, CNRS, Univ. Savoie Mont Blanc, CNRS, LECA, Laboratoire d'Écologie Alpine

Correspondence to **Sara Si-Moussi**
Mailing adress: 2 Place Pierre Viala 34060 Montpellier, France
E-mail: sara.si-moussi@inrae.fr *Tel/Fax:* +33 (0)4.99.61.30.86

Article type Methods

Number of words abstract: 150, main text: 5000

Number of references 61

Number of tables or figures 6

Acknowledgements. We thank Laura Pollock (McGill university) and Tamara Münkenmüller (LECA) for guidance on and access to source code for simulating virtual communities; we thank Philippe Choler (LECA) for discussion and crucial explanations on the ecology of Alpine plant communities. We also thank Li Ping Liu for access to source code on exponential family embeddings.

SS is supported by a joint PhD fellowship between the French National Institute of Agricultural and Environmental Research (INRAE) and the French Research Institute for digital sciences (Inria). The research was also supported by the Agence Nationale pour la Recherche (ANR) through the GlobNet (ANR-16-CE02-0009), Gambas (ANR-18-CE02-0025) and Forbic (ANR-18-MPGA-0004) projects. Most of the computations presented in this paper were performed using the GRICAD infrastructure.[1]

Authorship. SS conceptualized the inference framework and reviewed it with EG. WT and SS designed the evaluation methodology. SS, WT and EG analyzed the results. MH gave additional perspectives to the paper. SS wrote the first version of the paper. All authors contributed critically to editing the manuscript and gave final approval for publication.

Data Availability. Data on Alpine plants [13], obtained from the *ade4* package. Simulated datasets and source codes are available in the online GitHub repository.

Source code. The source code for running the model is released as a Python package available at DOI:10.5281/zenodo.3611437. A tutorial for running it from R is provided in the Supplementary Materials.

[1] https://gricad.univ-grenoble-alpes.fr

Abstract

The interplay between environmental suitability, dispersal and biotic interactions induces spatial patterns of species' co-abundance. Existing statistical frameworks that infer the underlying interactions from these patterns either ignore the species response to the environment or they fail to account for the asymmetric nature of interactions.

Here, we propose a framework that (a) models pair-wise associations as directed influences from a source to a target species, parameterized with two species-specific latent variables: the response of the target species to the community, and the effect of the source species on the community; and (b) jointly fits these associations with a habitat suitability model through a conditional abundance model. Using both simulated and empirical data, we demonstrate the ability of the framework to recover known associations and highlight the properties of the learned association networks. Our framework should now pave the way for getting more accurate pictures of interspecific dependencies from empirical data.

1 Introduction

Understanding the drivers of species distribution and their abundance is a long-lasting goal of biogeography [32]. Niche theory explains the spatial distribution of species by a set of physiological and adaptive properties allowing them to thrive in specific environmental conditions and decline in others [12, 49]. The range of environmental (abiotic) variables, such as climate and soil characteristics, that matches the eco-physiological requirements of a species delimits its potential niche (Grinnellian niche, [25]) or (Fundamental niche, [33]). Habitat suitability models (HSM) or species distribution models (SDMs) [27] aim to infer and model this niche by establishing statistical relationships between observed occurrences or abundances of species and the environmental characteristics of the corresponding locations.

HSMs have proven useful to predict species ranges in space, or their shift in response to climate change, providing operational tools to conservation biologists [19,26]. However, as they model multiple species distributions separately, they fail to detect, or to account for, possible dependencies between species that can restrict or extend their ranges beyond what is expected when considering only abiotic factors. Indeed, species may exclude one another locally (*competitive exclusion*, [28]) or be different enough in terms of space and resource needs to co-exist (*niche partitioning*, [52]). Conversely, some species facilitate others by modifying the environment in a way that creates habitats or enables access to resources for other species (*engineering and facilitation*, [17]). Although these interactions take place at a local scale, some of them may alter the range of the species on a wider, macroscopic scale [22]. Consequently, they induce consistent patterns of co-location and dislocation that are unexplained by the abiotic environment. Such interactions are referred here as *associations*. The inability to take into account the presence or absence of other species is therefore an important source of errors for statistical models of species distributions ([61]).

Over the last decade, several approaches have been proposed to infer interspecific dependencies from the observations of many species. Probabilistic Graphical Models (PGM) [34] have been used to infer either directed (Bayesian Networks, BN) or undirected (Markov Random Fields, MRF) networks [20] involving plants [1], parasites and potential hosts [46], predators and preys [37, 56] or multi-trophic communities turnover [43]. More recently, Joint Species Distribution Models (JSDM) were introduced to address the same question while jointly predicting co-occurrences of multiple species [44, 48]. The gist is that once abiotic factors are accounted for, the unexplained variance, typically captured by the correlation matrix of the residuals, is attributed to the effect of species on one another or to unknown environmental variables [44]. As they rely purely on correlations, JSDMs and MRFs are limited to estimating symmetric associations where the involved parties influence one another with the same polarity and strength. Lany et al. [35] proposed a JSDM that allows to capture asymmetric associations but requires longitudinal data (see also [2, 30]). On the other hand, BNs support directed relationships but they impose an acyclic structure that does not allow modeling of bidirectional influences.

Inferring associations from co-occurrence data is a common task in text mining. Supposedly, the probability of a word occurring in a particular sentence of a text depends on the semantic compatibility (association) of this word with the list of words surrounding it, forming its context. The common approach is to use *word embedding* algorithms (e.g *word2vec* [38]) to learn multidimensional representations (embeddings) of words that encode this

contextual semantic compatibility. By analogy, in community ecology, the probability of the presence of a species in a given environmentally suitable site depends on its compatibility with other species occurring at that site, i.e. other species in the observed community.

Recently, word embeddings were generalized to any type of data that follow an exponential family distribution [36, 50], including binary and ordinal data, in so-called exponential family embeddings. Building on this work, we propose here a conditional probabilistic model of species co-distributions that can be trained jointly with any habitat suitability model on presence/absence or count data to infer interspecific associations. In this paper, we detail the methodology, the mathematical formulation and the underlying assumptions. We evaluate the capacity of the model to accurately recover associations using both simulated data for which we know the true interactions between species, and empirical data for which reliable expert knowledge is available [13, 60]. We show that our model critically infers meaningful associations. Finally, we demonstrate how the learnt parameters can be harnessed, to analyze the structure of biotic association networks.

2 The inference framework

Three main conditions should be satisfied for a species to be present at a given site. First, the site must be accessible. This relates to the species' intrinsic *dispersal* capacity and the presence of migration opportunities or barriers. Second, the abiotic conditions should allow the species' population to maintain a positive growth rate. This condition is referred to here as *habitat suitability* and is the target of Habitat Suitability Models. Third, the species should sustain the interactions with the other species of the community, since those interactions can also impact the species' survival chances and its abundance [27]. Although we recognize the importance of spatial dispersal processes, in this study we focus on the latter two factors, namely *habitat suitability* and *species interactions*.

Notation. We consider a dataset consisting of the abundances of a collection $\mathcal{S}$ of m species observed at a collection $\mathcal{K}$ of n sites, as well as abiotic variables measured at these same sites or in their vicinity. The abundance of species i at site k is denoted y_{ki}, while the vector x_k represents the *abiotic variables* at site k.

In what follows, we introduce the key concepts used in the inference model. In particular, we explain how we model the associations between a pair of species by decomposing them into effects and responses, represented as multi-dimensional embedding vectors, and how we use these embeddings to recover biotic interactions.

2.1 Spatial associations and biotic context

2.1.1 Representing species associations using embeddings

For a given pair of species, a *spatial association* describes the relative influence that they have on each other's abundance. The two directions of this influence can be of different types (positive, negative or neutral) and have different intensities Fig 1b. Several mechanisms can lead to such association: a direct interaction between these two species (e.g. competition, predator-prey), an indirect interaction through the environment or a shared correlation

to an unmeasured environmental variable or an unobserved group of organisms.

Here, we represent the association between species i and j with a pair of scalars a_{ij} and a_{ji}, representing the strength of the influence of species j on species i and vice-versa, respectively. More specifically, a_{ij} represents the change (excess if positive, deficit if negative, none otherwise) in *target* species i's abundance induced by the *source* species j. These values across all pairs of species, and in both directions, can be collected into an $m \times m$ asymmetric association matrix A.

The association strength depends on two parameters: the *effect* applied by the source on the target species, and the *response* of the target species. We assume these parameters are controlled by intrinsic traits or properties of the species, which we encode in two separate d-dimensional real-valued vectors referred to as embeddings. In practice, d is a user defined hyperparameter which is typically significantly smaller than half the number of species.

The *effect embedding* of species i, α_i, captures the type of organisms the species allows when it is present. The *response embedding* of species i, ρ_i, measures the type of biotic context the species would strive in. For instance, trees with spreading canopy create shade (effect) that selects only shade-tolerant (response) species and exclude others. The response and effect embeddings of the different species can be collected into two $m \times d$ matrices, respectively denoted as P and Q.

The association matrix Fig 1a is then written as $A = PQ^T$.

2.1.2 Biotic context

The biotic context encodes our assumptions about the potential biotic effects a target species is exposed to at a given site. In the simplest case, without any prior knowledge, it consists of individuals from other species observed at the same site. Formally, the biotic context of species i at site k, denoted C_{ki}, is defined as follows:

$$C_{ki} = \{j \in \mathcal{S}, j \neq i \text{ and } y_{kj} > 0\}\,.$$

We obtain the aggregated effect of the biotic context by averaging the effect embeddings of its elements weighted by their respective abundances:

$$z_{ki} = \frac{1}{|C_{ki}|} \sum_{j \in C_{ki}} y_{kj}\alpha_j\,.$$

This formulation allows the presence of facilitators and competitors to balance one another. By weighting with abundance, we implicitly consider that individuals from the same species are similar and contribute equally to the community structure. Rare species have a noticeable impact only if their per capita effect is stronger than the aggregated effect of dominant groups.

The biotic context carries implicit constraints on the structure of species association networks by restricting the set of potential associations a priori. For instance, it can be customized for each species according to its known interactions. Moreover, it can include species from neighboring locations (**spatially-explicit**) up to a chosen radius where their influence is relevant (e.g. species with high mobility). Similarly, we can construct the biotic context from previous observations (**temporally-explicit**) to perform a causal analysis. We detail the

mathematical adjustments of these alternative definitions along with the associated data requirements and relevant effect aggregation functions in the Appendix.

2.2 A conditional generative model of abundance

2.2.1 Formalization

The habitat suitability for species i at site k, denoted s_{ki}, is a binary variable that follows a Bernoulli distribution whose parameter (success rate) is estimated using a habitat suitability model (HSM), h, fitted on the target species's occurrences, i.e.

$$s_{ki} \sim \mathcal{B}\big(h_i(x_k)\big)\,.$$

At sites where the abiotic environment is unsuitable (i.e. where $s_{ki} = 0$), the probability mass of the species abundance is concentrated on zero. Otherwise (i.e. where $s_{ki} = 1$), the abundance of the species is a function of its biotic context. In other words, we assume that the abiotic environment conditions the presence or the absence of a given species, while the biotic context controls its abundance.

Following Rudolph et al. [50], we model the abundance of species i at site k using the canonical form of the exponential family $\mathcal{E}$, whose probability density function (pdf) $f_{\mathcal{E}}$ is parameterized by η_{ki}. Formally,

$$y_{ki} \sim \begin{cases} \mathcal{E}\big(\eta_{ki}, \tau(y_{ki})\big) & \text{if } s_{ki} = 1, \\ \delta_0 & \text{otherwise,} \end{cases}$$

such that

- δ_0 denotes Dirac (point-mass) distribution, whose density is equal to one at zero, and to zero elsewhere.
- τ denotes the sufficient statistic from the canonical form of the exponential family distribution. It depends only on the data point y_{ki}.

We let the canonical parameter η_{ki} depend on the response ρ_i of the target species and on the biotic context effect z_{ki}. An offset o_i is used to represent the baseline abundance of each species in the event of an empty biotic context. The link function f scales the outcome to the domain of the target variable. The canonical parameter η_{ki} is defined as

$$\eta_{ki} = f(\rho_i z_{ki} + o_i)\,,$$

which can be rewritten as an aggregate of pairwise association strengths:

$$\eta_{ki} = f\big(\sum_{j \in C_{ki}} y_{kj} a_{ij} + o_i\big)\,.$$

The type of data considered (presence/absence vs. abundance) might lead to different choices of probability distributions, which in turn require resorting to different variants of the generic model (cf. Table 1).

2.2.2 Inference

We gather as positive and negative observations (instances), the species that are present and absent at each site. Absent species are typically over-represented in an ecological dataset as compared to present species, leading to a high imbalance between positive and negative observations. To address this issue, while all positive instances are included into the training set, negative instances are sub-sampled randomly (at rate $r\%$) at each training iteration. Then, we use Stochastic Gradient Descent [6] to learn the parameters (esp. response and effect embeddings matrices and HSM parameters) that minimize the negative loglikelihood (cf. Appendix) of the selected observations with the addition of lasso penalties on the embeddings to promote the sparsity of the resulting associations. Finally, we use cross-validation to select the hyper-parameters (including hyper-parameters for the abiotic suitability model, embedding dimension, vector of species offsets, negative examples subsample rate and regularization coefficient).

2.3 Unraveling inter-specific association networks

To identify meaningful associations, we apply two filtering steps to the estimated matrix A. First, the *statistical filtering* step consists in setting to zero all associations with a confidence interval containing zero and keeping the mean value for the rest. Second, the *biogeographic filtering* step aims to further eliminate associations that are predicted to potentially exist from the latent representations of the species, but are not realized by the species occurrences because they break some biogeographic constraints:

1. Mutualism or attraction between two species require co-existence. Thus, we set to zero any inferred positive effect involving two species never observed together in the same site [51], i.e. non-co-occurring.

2. Repulsive relationships do not require co-occurrence and may even explain the geographic separation. Hence, the involved species do not have to co-occur but should live in similar environments to be considered as a potential negative association. Specifically, we compute the ranges of the environmental values corresponding to the occurrences of each species and retain negative associations only if these ranges overlap or if the species are otherwise sufficiently similar (above a user-defined similarity threshold) in terms of their habitat suitability parameters.

Furthermore, we focus on the polarity of the associations, rather than their strength, hence we consider a discrete version of the association matrix, which we call the *adjacency matrix* defined as follows:

$$I_{ij} = \begin{cases} \text{positive} & \text{if } a_{ij} > \epsilon^+, \\ \text{negative} & \text{if } a_{ij} < \epsilon^-, \\ \text{neutral} & \text{otherwise.} \end{cases}$$

such that ϵ^+ and ϵ^- represent a user-defined threshold on the strength of the positive and negative associations, respectively.

The resulting matrix defines a network, where each species is represented by a vertex and a directed edge labelled as positive (resp. negative) from vertex i to vertex j represents a positive (resp. negative) influence of species i on species j.

By design (Fig. 1a), species with similar response embeddings constitute clusters of rows in the adjacency matrix, called *response groups*. Conversely, species with similar effect embeddings constitute clusters of columns called *effect groups*. These groups can be computed simultaneously using a *bi-clustering* algorithm [23]. The product of both types of groups results in the emergence of clusters of exchangeable or redundant species in the resulting network, called *structural roles* [21].

3 Theoretical validation of the framework

Before applying our model to infer associations from empirical data, we evaluated its ability to recover interspecific associations from simulated datasets with known associations.

3.1 Data generation

We used a process-based stochastic model adapted from `Virtualcom` [40] to simulate the assembly of individuals from a regional species pool into communities, on different locations sampled along an environmental gradient. The assembly process is controlled by three filtering mechanisms: the response to the abiotic environment, the outcome of biotic interactions, and reproduction.

We set up an experiment where multiple simulations were run on random points on a single abiotic gradient ranging from 0 to 100 with different hand-crafted configurations of the prior association matrix: absence of association, positive associations only, negative associations only and both positive and negative associations. Generally, few species were set to interact. We investigated two settings where species are involved in one association (sparse) or more than on associations (dense). Associated pairs were chosen such that their abiotic niches overlap.

In each configuration mode, we varied the number of species (5, 10 or 20), the density setting (sparse or dense) and whether the association matrix included asymmetric effects (semi-attraction or semi-repulsion). Positive (resp. negative) effects were all set to $+1$ (resp. -1) as we are interested in the polarity of the associations rather than their magnitude. The factorial design of this experiment produced 29 simulation datasets (Appendix).

3.2 Evaluation

We describe the inference procedure and model selection in detail in Appendix.

3.2.1 Observed relative abundances vs. inferred associations strengths

For each association type and simulation configuration, we compare the inferred association strengths to the observed relative abundance effects. To quantify these effects, we define the *relative abundance index* (RAI), an asymmetric metric that measures the change in abundance of the target species when the source species is present as compared to its mean abundance irrespective of whether the source species is present.

Formally, we define

$$\bar{y}_t = \text{avg}\left(\{\, y_{kt}, \text{ for all } k \in \mathcal{K} \text{ such that } y_{kt} > 0 \,\}\right), \quad \text{and}$$

$$\Delta_{st} = \{\, y_{kt} - \bar{y}_t, \text{ for all } k \in \mathcal{K} \text{ such that } y_{kt} > 0 \text{ and } y_{ks} > 0 \,\}.$$

Then $\text{RAI}_{st} = \text{avg}(\Delta_{st})$. The larger the standard deviation $\text{std}(\Delta_{st})$, the more ambiguous the strength of the effect of species s on species t. If the confidence interval $\text{avg}(\Delta_{st}) \pm 1.96\,\text{std}(\Delta_{st})$ does not contain zero, then the simulated dependencies unambiguously translate a polarized effect of species s on species t. Otherwise, the polarity of the effect is ambiguous, due to either confounding effects of other species or a neutral association if the mean is close to zero. We also compute the Jaccard coefficient between the binary presence/absence vectors of species s and t, a.k.a. Jaccard co-occurrence index, denoted as J_{st}.

3.2.2 Association classification evaluation

We discretized the learnt associations using the threshold values $(\epsilon^+, \epsilon^-) = (0, 0)$ to obtain the corresponding classes (positive, negative, neutral). Subsequently, we evaluated the latter against the simulated association classes as the ground-truth using standard multi-class performance metrics (recall, precision, F1-score). Recall measures the percentage of associations of a specific class correctly recovered by the model, whereas precision quantifies the percentage of true associations amongst those classified as the specified class. The F1-score is computed as the harmonic mean of recall and precision. A higher recall indicates lower false negatives whilst higher precision indicates lower false positives.

3.3 Results

Overall, we found a better fit for positively associated communities than those in competition, while mixing both types resulted in intermediate performances. In the case of simulations with competition, most sparse and asymmetric configurations induced better performances than their (dense and symmetric) counterparts (see Appendix).

The average relative abundance index reflected well the simulated associations with negative (resp. positive) effects below or around (resp. above) zero, while neutral associations were centered around zero. However, most positive effects yielded small relative abundance effects as compared to the negative effects. Although more clearly marked, the latter approached neutrality on larger and more densely connected communities (Fig. 2).

The inference model was able to discriminate positive from negative effects while maintaining an average value for non interacting pairs centered on zero with a small variance. On simulations with a dense mix of positive and negative associations, both observed effects and inferred associations were close to zero, possibly due to opposing effects canceling each other. The absence of associations led to the systematic prediction of the offset hence the constant deviance on mixed types simulations.

On average, recall did not vary significantly between positive and negative associations, whereas precision was higher for negative than for positive associations (Table 2), indicating the detection of spurious positive

associations. Much higher precision was achieved for neutral associations (absence of association). The prediction performance was better for small species pool, with higher recall on the dense configurations but higher precision on sparse ones. The sparse asymmetric positive simulation resulted in the worst predictive performance with low recall.

4 Empirical case study

We applied our model to empirical observations (also analyzed by [60]) of species abundances to show how it can unravel meaningful associations. We used the plant dataset from [13] that consists of 75 vegetation plots, of size 5×5m sampled around July 2000. Across the vegetation plots, the abundance of the 82 occurring plant species were registered. In addition, a set of environmental and topographic variables was recorded on each plot. We describe the data preprocessing, framework adaptation and model selection procedures in Appendix.

4.1 Network analysis

We performed a hierarchical bi-clustering on the inferred association matrix to obtain effect and response groups (Fig. 3a). In parallel, we applied the modularity maximization algorithm [42] on the association network to identify densely connected modules, referred to as communities [21]. After that, we mapped the structural roles within the modules to create the group-level network. Finally, we analyzed the resulting patterns in light of existing literature on alpine plants interactions [14].

4.2 Results

We identified four densely connected modules of different sizes, within which species occupied various structural roles in the plant association network. Modules were structured along the snow melting date gradient (Fig. 3a).

Species from early-melting sites were classified into the same module. We found a prominence of positive associations, specifically an unselective mostly asymmetric attraction of forbs and grasses to tall dominant graminoids (*Carex, Kobresia*). Forbs and grasses formed two distinct groups linked by negative associations. Besides, some of them acted as hubs connecting the high elevation sites to the adjacent sites where they also occurred. The second module encompassed two groups of grasses: (i) Tall herb grasslands occuring in favorables conditions, mostly structured by negative associations (ammensalism and competition); (ii) Short herb meadows, exposed to zoogenic disturbances. They presented higher abundances when co-occuring with tall herbs. The third module consisted of chinopholous (cold-resistant) vegetation appearing on late-melting sites. The last module included north-facing isolated communities dominated by *Salix Herbacea* positively associated with high-altitude communities and characterized by high eccentricity (Fig. 4).

In general, positive associations were prominent on stressful conditions. For instance, on early melting sites, species are exposed to wind and erosion due to snow melting [13]. The positive associations could be explained by the facilitative effect of graminoids through multiple hypothetical mechanisms. Graminoids have the ability to maintain the soil stability [9, 29], they can also prevent dessication and frost heaving on stones in favor of seedling

survival [14]. They also sustain a suitable microclimate for small forbs and grasses, while offering protection from the wind [59]. On the other hand, negative associations were found on the richest sites, hypothetically reflecting a competition for resources: water and Nitrogen [14].

As reported in the literature, the abiotic conditions strongly structured the predicted response [13] of the plant species and the dominant interaction types [10]. Specifically, network modules were distributed along the gradient following their composition's response to the average snow duration. Negative associations inflicted by competitive tall grasses on mid slope communities connected early-melting communities to the chinopholous vegetation from late-melting sites in the resulting association network (Fig. 3b). Response and effect groups were included in one or at most two close (in terms of position in the gradient) modules (Fig. 3a). Non neutral associations had consistent types (either negative or positive) within effect groups regardless of the responding species, suggesting that species roles (effects) in their community might be predictable from their own characteristics and the surrounding abiotic conditions. At last, learnt associations were symmetric within groups but asymmetric (mostly semi-attraction or semi-repulsion) between them.

5 Discussion

In this article, we tackled the challenge of inferring interspecific associations from multiple species co-abundances along environmental gradients. To do so, we formalized pairwise associations as a function of two sets of latent variables representing the response and the effect of each species in respect to the others. We incorporated these associations into a conditional probabilistic model of abundance that controls for habitat suitability. The evaluation of the model's ability to recover known associations from simulated data showed that it is able to discriminate the association types (positive, negative or neutral), but the inferred strength depended on the species pool size, niche overlap, network density and the presence of multiple confounding associations. When we applied the model to the co-abundance data of plants along a mesotopographic gradient in the French Alps, the model identified most of the important relationships expected in these plant communities [14].

5.1 Disentangling confounding effects

Inferring species associations from co-occurrence patterns is a very challenging task [11]. Even in our simulated dataset for which we have pre-defined interactions between species, the resulting co-occurrence levels could be high even for known competing pairs of species (case of a small species pool with large carrying capacity). Instead, the co-abundance structure better reflects the nature of associations. Indeed, species' abundances are lower than in average in presence of negative associations and higher with positive associates. Nonetheless, pairwise abundance effects may turn out to be neutral in presence of multiple confounding effects [7]. Our proposed framework is able to tease apart the different opposing influences (Fig. 2) by estimating the pairwise associations conditioned on all other species.

5.2 Interaction of habitat suitability and biotic associations

Following a hierarchical filtering scheme of community assembly [7, 44, 45], we assumed the species occurred only on suitable habitats. However, species may occur outside their abiotic niche if the local conditions are ameliorated by their peers (facilitation) [8, 16, 53].

In the absence of micro habitat descriptors, two scenarios are possible depending on the location's distance in environmental space from the species' true fundamental niche. If the distance is small, any HSM would probably overestimate the species ranges [5] with no residuals on the occurrence probabilities. Consequently, the associations would go undetected by joint presence/absence models (e.g JSDMs), whereas the proposed model will detect them as long as there is an observed increase in abundance. Conversely, if the distance is large then a robust HSM would correctly learn the true abiotic niche, leaving the unexplained presence to the biotic effect. But since the habitat was found unsuitable by the HSM component, our model would fail to detect the underlying biotic association.

Indeed, the model constraints considered co-occurrence as a prerequisite for positive associations (including facilitation) while in the described scenario co-occurrence is a consequence of facilitation. A compromise solution would be to consider that species do not respond separately to each factor : abiotic captured by the HSM parameters θ_h and biotic through the response embeddings ρ. Instead, they respond to the outcome of the aggregation f of the abiotic features x and the biotic context effects z. By choosing f to be a universal function approximator (e.g a feed-forward neural network [31]), we can learn which of the hierarchical, additive or interactive models of the abiotic and biotic filters best fit the observed abundances/distributions without any *a priori* assumption.

5.3 Associations validation

With no prior or guidance on the expected associations network, validating inferred associations is challenging.

It is now agreed upon that associations are not equivalent to biological interactions [18]. They represent significant spatial co-location patterns, that are informative in a predictive rather than in a causal way [39]. The specific mechanism that led to these patterns may vary from pair to pair, ranging from direct interactions (e.g. trophic), to indirect interactions (e.g. engineering, shared habitat). Consequently, the validation of inferred associations should consider other explanations than biotic interactions.

In general, because many processes influence community assembly, multiple scenarios could lead to the same communities making this problem unidentifiable [41]. In this case, rather than a single expected list of associations, we need all the possible combinations of associations, or a goodness-of-fit measure that accounts for equivalence between different combinations. A possible way to prevent this issue is to include prior knowledge of ecological interactions in the model [11]. For instance, [54] used a Bayesian network with a predefined structure, and trained its parameters using a HSM to predict species occurrence probabilities. In our case, such constraints can be defined by altering the biotic context definition. One direct way to do it is to consider a customized biotic context for each species composed of the set of its potential interaction partners in a pre-built regional metaweb.

There is now growing evidence that ecological interactions are context-dependent [47, 55], we showed in the Appendix how to adapt the framework to infer associations whose strength is modulated by other covariates (e.g. stress, presence of predator, etc.). Recently developed models account for association variability as a function of

the environmental context [15, 55]. Despite these new possibilities, the question of how to validate their results still arises itself.

5.4 Species roles and association's asymmetry

In the plant network, modules were strongly structured by the abiotic gradient. Because of the HSM conditioning, this pattern would be expected in gradients with strong taxonomic turnover. Modules and structural roles provided two complementary information. The modules brought insight into the connectivity therefore the stability [24] of the meta-network while structural roles, including response and/or effect groups, were useful to evaluate the functional redundancy within locally projected networks. In the future, identifying characteristic traits within structural roles would allow the elicitation of the functional drivers of network structure.

Many studies [57] of interaction networks reported that interactions tend to be asymmetric, both in terms of "type" in binary networks [3, 58] and in terms of "strength" in quantitative networks [4]. By analyzing various types of observed ecological networks, [57] suggested that asymmetric interaction strengths arise from mismatch in species relative abundances. Since the proposed model learns per capita effects, the abundance is already controlled for. In the case of Alpine plants, the response and effect groups encompassed species occurring in similar habitats. Knowing that, the predominance of symmetrical (resp. asymmetric) associations within (resp. between) groups suggests the degree of asymmetry might be inversely related to habitat sharing or niche overlap.

Nevertheless, the ability to discern this asymmetry sheds light on the imbalance and direction of interspecific dependencies, drawing a more accurate picture for biodiversity forecasting models.

6 Conclusion

Biological interactions and other processes induce spatial patterns of co-occurrence and co-abundance. We presented and validated a model of species co-abundances as a function of the habitat and biotic associations. We proposed an asymmetric scheme for modeling associations that is based on learning latent representations of species' responses and effects. Future efforts should be directed towards a combination of prior knowledge on the complete or partial topology of the association networks to guide the inference process. Along with that, a strong theory of how known ecological interactions influence the co-distribution of species is needed to support all these models.

References

[1] Aderhold, A., Husmeier, D., Lennon, J. J., Beale, C. M., and Smith, V. A. Hierarchical bayesian models in ecology: reconstructing species interaction networks from non-homogeneous species abundance data. *Ecological Informatics 11* (2012), 55–64.

[2] Barraquand, F., Picoche, C., Detto, M., and Hartig, F. Inferring species interactions using granger causality and convergent cross mapping. *arXiv preprint arXiv:1909.00731* (2019).

[3] Bascompte, J., Jordano, P., Melián, C. J., and Olesen, J. M. The nested assembly of plant–animal mutualistic networks. *Proceedings of the National Academy of Sciences 100*, 16 (2003), 9383–9387.

[4] Bascompte, J., Jordano, P., and Olesen, J. M. Asymmetric coevolutionary networks facilitate biodiversity maintenance. *Science 312*, 5772 (2006), 431–433.

[5] Beale, C. M., and Lennon, J. J. Incorporating uncertainty in predictive species distribution modelling. *Philosophical Transactions of the Royal Society B: Biological Sciences 367*, 1586 (2012), 247–258.

[6] Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[7] Boulangeat, I., Gravel, D., and Thuiller, W. Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. *Ecology letters 15*, 6 (2012), 584–593.

[8] Bulleri, F., Bruno, J. F., Silliman, B. R., and Stachowicz, J. J. Facilitation and the niche: implications for coexistence, range shifts and ecosystem functioning. *Functional Ecology 30*, 1 (2016), 70–78.

[9] Callaway, R. M. *Positive interactions and interdependence in plant communities*. Springer, 2007.

[10] Callaway, R. M., Brooker, R., Choler, P., Kikvidze, Z., Lortie, C. J., Michalet, R., Paolini, L., Pugnaire, F. I., Newingham, B., Aschehoug, E. T., et al. Positive interactions among alpine plants increase with stress. *Nature 417*, 6891 (2002), 844.

[11] Cazelles, K., Araújo, M. B., Mouquet, N., and Gravel, D. A theory for species co-occurrence in interaction networks. *Theoretical Ecology 9*, 1 (2016), 39–48.

[12] Chase, J. M., and Leibold, M. A. *Ecological niches: linking classical and contemporary approaches*. University of Chicago Press, 2003.

[13] Choler, P. Consistent shifts in alpine plant traits along a mesotopographical gradient. *Arctic, Antarctic, and Alpine Research 37*, 4 (2005), 444–453.

[14] Choler, P., Michalet, R., and Callaway, R. M. Facilitation and competition on gradients in alpine plant communities. *Ecology 82*, 12 (2001), 3295–3308.

[15] Clark, N. J., Wells, K., and Lindberg, O. Unravelling changing interspecific interactions across environmental gradients using markov random fields. *Ecology 99*, 6 (2018), 1277–1283.

[16] Crotty, S. M., and Bertness, M. D. Positive interactions expand habitat use and the realized niches of sympatric species. *Ecology 96*, 10 (2015), 2575–2582.

[17] Cuddington, K., Byers, J. E., Wilson, W. G., and Hastings, A. *Ecosystem engineers: plants to protists*, vol. 4. Academic Press, 2011.

[18] Dormann, C. F., Bobrowski, M., Dehling, D. M., Harris, D. J., Hartig, F., Lischke, H., Moretti, M. D., Pagel, J., Pinkert, S., Schleuning, M., et al. Biotic interactions in species distribution modelling: 10 questions to guide interpretation and avoid false conclusions. *Global ecology and biogeography 27*, 9 (2018), 1004–1016.

[19] Elith, J., and Leathwick, J. R. Species distribution models: ecological explanation and prediction across space and time. *Annual review of ecology, evolution, and systematics 40* (2009), 677–697.

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

[21] Gauzens, B., Thébault, E., Lacroix, G., and Legendre, S. Trophic groups and modules: two levels of group detection in food webs. *Journal of The Royal Society Interface 12*, 106 (2015), 20141176.

[22] Gotelli, N. J. Ecology: Biodiversity in the scales. *Nature 419*, 6907 (2002), 575.

[23] Govaert, G., and Nadif, M. *Co-clustering: models, algorithms and applications*. John Wiley & Sons, 2013.

[24] Grilli, J., Rogers, T., and Allesina, S. Modularity and stability in ecological communities. *Nature communications 7* (2016), 12031.

[25] Grinnell, J. The niche-relationships of the california thrasher. *Auk 34*, 4 (1917), 427–433.

[26] Guisan, A., and Thuiller, W. Predicting species distribution: offering more than simple habitat models. *Ecology letters 8*, 9 (2005), 993–1009.

[27] Guisan, A., Thuiller, W., and Zimmermann, N. E. *Habitat suitability and distribution models: with applications in R*. Cambridge University Press, 2017.

[28] Hardin, G. The competitive exclusion principle. *science 131*, 3409 (1960), 1292–1297.

[29] Heilbronn, T., and Walton, D. W. Plant colonization of actively sorted stone stripes in the subantarctic. *Arctic and Alpine Research 16*, 2 (1984), 161–172.

[30] Hollingdale, E., Pérez-Barbería, F. J., and Walker, D. M. Inferring symmetric and asymmetric interactions between animals and groups from positional data. *PloS one 13*, 12 (2018).

[31] Hornik, K., Stinchcombe, M., and White, H. Multilayer feedforward networks are universal approximators. *Neural networks 2*, 5 (1989), 359–366.

[32] Humboldt, A. v., Bonpland, A., et al. *Essai sur la géographie des plantes.* Chez Levrault, Schoell et compagnie, libraires, 1805.

[33] Hutchinson, G. The multivariate niche. In *Cold Spring Harbor Symposia on Quantitative Biology* (1957), vol. 22, pp. 415–421.

[34] Koller, D., and Friedman, N. *Probabilistic graphical models: principles and techniques.* MIT press, 2009.

[35] Lany, N. K., Zarnetske, P. L., Schliep, E. M., Schaeffer, R. N., Orians, C. M., Orwig, D. A., and Preisser, E. L. Asymmetric biotic interactions and abiotic niche differences revealed by a dynamic joint species distribution model. *Ecology 99*, 5 (2018), 1018–1023.

[36] Liu, L.-P., and Blei, D. M. Zero-inflated exponential family embeddings. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (2017), JMLR. org, pp. 2140–2148.

[37] Majdi, N., Hette-Tronquart, N., Auclair, E., Bec, A., Chouvelon, T., Cognie, B., Danger, M., Decottignies, P., Dessier, A., Desvilettes, C., et al. There's no harm in having too much: A comprehensive toolbox of methods in trophic ecology. *Food webs 17* (2018), e00100.

[38] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (2013), pp. 3111–3119.

[39] Milns, I., Beale, C. M., and Smith, V. A. Revealing ecological networks using bayesian network inference algorithms. *Ecology 91*, 7 (2010), 1892–1899.

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

[41] Muñoz-Tamayo, R., Puillet, L., Daniel, J.-B., Sauvant, D., Martin, O., Taghipoor, M., and Blavy, P. To be or not to be an identifiable model. is this a relevant question in animal science modelling? *animal 12*, 4 (2018), 701–712.

[42] Newman, M. E. Modularity and community structure in networks. *Proceedings of the national academy of sciences 103*, 23 (2006), 8577–8582.

[43] Ohlmann, M., Mazel, F., Chalmandrier, L., Bec, S., Coissac, E., Gielly, L., Pansu, J., Schilling, V., Taberlet, P., Zinger, L., et al. Mapping the imprint of biotic interactions on β-diversity. *Ecology letters 21*, 11 (2018), 1660–1669.

[44] Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T., and Abrego, N. How to make more out of community data? a conceptual framework and its implementation as models and software. *Ecology Letters 20*, 5 (2017), 561–576.

[45] Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M., and Araújo, M. B. *Ecological niches and geographic distributions (MPB-49)*, vol. 56. Princeton University Press, 2011.

[46] Pilosof, S., Fortuna, M. A., Cosson, J.-F., Galan, M., Kittipong, C., Ribas, A., Segal, E., Krasnov, B. R., Morand, S., and Bascompte, J. Host–parasite network structure is associated with community-level immunogenetic diversity. *Nature communications 5* (2014), 5172.

[47] Poisot, T., Stouffer, D. B., and Gravel, D. Beyond species: why ecological interaction networks vary through space and time. *Oikos 124*, 3 (2015), 243–251.

[48] Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., Vesk, P. A., and McCarthy, M. A. Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (jsdm). *Methods in Ecology and Evolution 5*, 5 (2014), 397–406.

[49] Pulliam, H. R. On the relationship between niche and distribution. *Ecology letters 3*, 4 (2000), 349–361.

[50] Rudolph, M., Ruiz, F., Mandt, S., and Blei, D. Exponential family embeddings. In *Advances in Neural Information Processing Systems* (2016), pp. 478–486.

[51] Sanderson, J. G., and Pimm, S. L. *Patterns in Nature: The analysis of species co-occurrences*. University of Chicago Press, 2015.

[52] Schoener, T. W. Resource partitioning in ecological communities. *Science 185*, 4145 (1974), 27–39.

[53] Stachowicz, J. Niche expansion by positive interactions: realizing the fundamentals. a comment on rodriguez-cabal et al. *Ideas in Ecology and Evolution 5* (2012).

[54] Staniczenko, P. P., Sivasubramaniam, P., Suttle, K. B., and Pearson, R. G. Linking macroecology and community ecology: refining predictions of species distributions using biotic interaction networks. *Ecology letters 20*, 6 (2017), 693–707.

[55] Tikhonov, G., Abrego, N., Dunson, D., and Ovaskainen, O. Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context. *Methods in Ecology and Evolution 8*, 4 (2017), 443–452.

[56] Trifonova, N., Kenny, A., Maxwell, D., Duplisea, D., Fernandes, J., and Tucker, A. Spatio-temporal bayesian network models with latent variables for revealing trophic dynamics and functional networks in fisheries ecology. *Ecological informatics 30* (2015), 142–158.

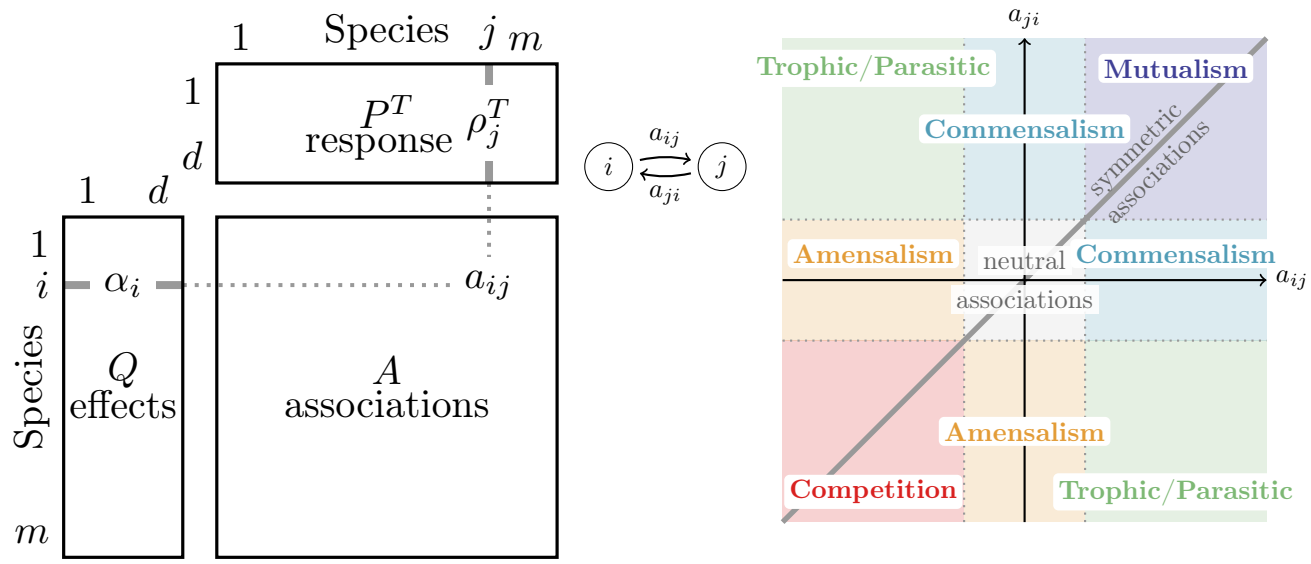
[57] Vázquez, D. P., Melián, C. J., Williams, N. M., Blüthgen, N., Krasnov, B. R., and Poulin, R. Species abundance and asymmetric interaction strength in ecological networks. *Oikos 116*, 7 (2007), 1120–1127.

[58] Vázquez, D. P., Poulin, R., Krasnov, B. R., and Shenbrot, G. I. Species abundance and the distribution of specialization in host–parasite interaction networks. *Journal of Animal Ecology 74*, 5 (2005), 946–955.

[59] Wardle, D., Barker, G., Bonner, K., and Nicholson, K. Can comparative approaches based on plant ecophysiological traits predict the nature of biotic interactions and individual plant species effects in ecosystems? *Journal of ecology 86*, 3 (1998), 405–420.

[60] Warton, D. I., Blanchet, F. G., O'Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C., and Hui, F. K. So many variables: joint modeling in community ecology. *Trends in Ecology & Evolution 30*, 12 (2015), 766–779.

[61] Wisz, M. S., Pottier, J., Kissling, W. D., Pellissier, L., Lenoir, J., Damgaard, C. F., Dormann, C. F., Forchhammer, M. C., Grytnes, J.-A., Guisan, A., et al. The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biological reviews 88*, 1 (2013), 15–30.



(a) Association matrix factorization.

(b) Pairwise associations classification

Fig. 1: Association strengths are computed from species response and effects (a). Pairwise association strengths are mapped to potential interaction classes (b). The different quarters of the bi-plot represent the various types of associations between species. The first bissector represents the association domain covered by correlation-based approaches (JSDM) and undirected graphical models (MRF).

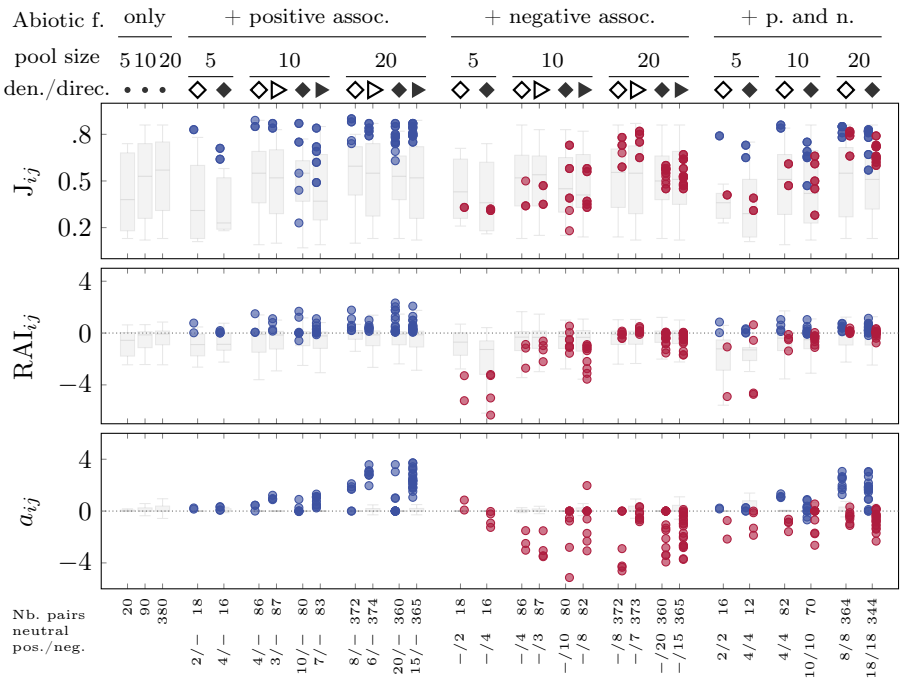
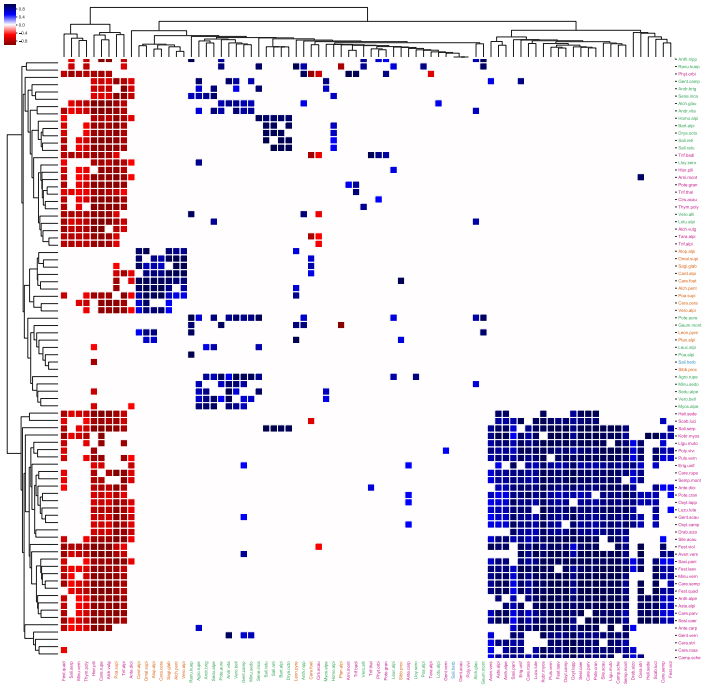
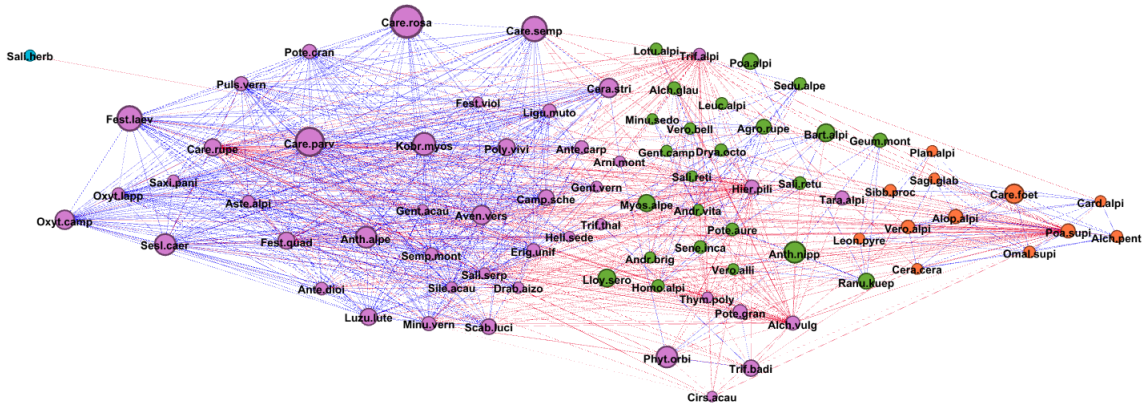


Fig. 2: Distribution of Jaccard co-occurrence indices J_{ij}, relative abundance indices RAI_{ij}, and inferred association strengths a_{ij} per simulation. Each data point represents a directed association (positive in red, negative in blue and neutral in gray) involving two species from the corresponding simulation. For clarity, we represent neutral associations with boxplots. The symbols represent combinations of density (sparse:hollow shape/dense:filled shape) and directionality (asymmetric:triangle/symmetric:diamond)



(a) Inferred plant association matrix. Species in the association matrix are grouped based on a hierarchical bi-clustering performed row-wise (yielding response groups) and column-wise (yielding effect groups).



(b) Network of plant associations. Blue (resp. red) edges indicate negative (resp. positive) edge weights. Node colors on the graph represent communities identified by the modularity maximization algorithm [42] whilst node sizes are scaled according to the plant height. Nodes (except *Salix Herbacea*, which represents the vegetation on the northern face of the gradient) are placed from left to right following an ascending order of their response to Snow duration (regression coefficient from the Generalized Linear Model used as a Habitat Suitability Model).

Fig. 3: Plant associations on an Alpine mesotopographic gradient. We highlight the communities (node colors) in figure (b) using colored labels on the matrix (a).

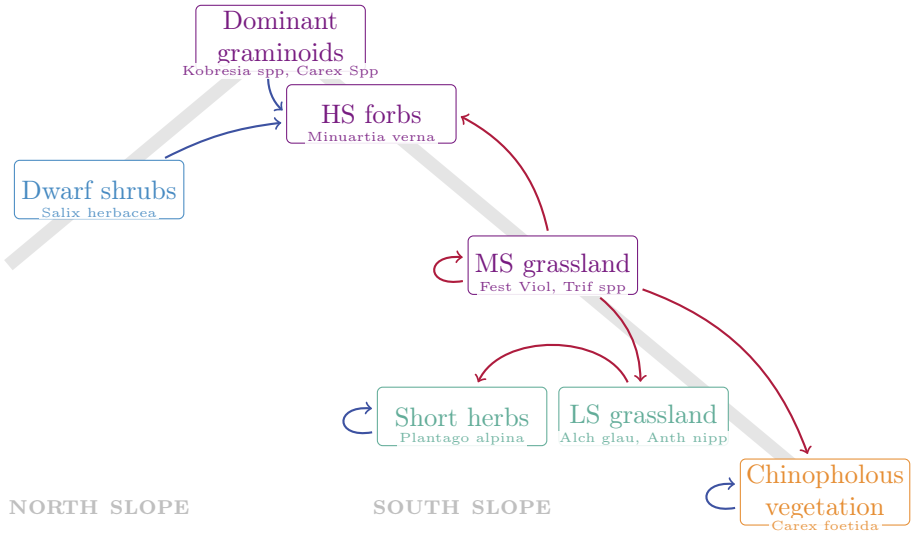


Fig. 4: The summary association network. Structural roles (nodes) are mapped to position in the gradient (Higher-slope HS, mid-slope MS, lower-slope LS) and plant classes (graminoids, grasses/herbs, forbs) and network modules (node colors). Edges go from a source (effect group) to a target (response group). Blue (resp. red) edges represent positive (resp. negative) associations.

Data type	Distribution	Link function	Natural parameter mapping
Presence/ Absence	Binomial	identity	Probability of occurrence $p_{kj} = \sigma(\sum_{i \in C_{ki}} y_{kj} a_{ij} + o_i)$ σ : logistic function
Count	Poisson	identity	Mean count $\lambda_{ki} = \exp(\sum_{j \in C_{ki}} y_{kj} a_{ij} + o_i)$
Count	Poisson	logarithm	Mean count $\lambda_{ki} = (\sum_{j \in C_{ki}} y_{kj} a_{ij} + o_i)$
Count	Negative binomial	identity	Mean count $p_{ki} = \exp(\sum_{j \in C_{ki}} y_{kj} a_{ij} + o_i)$

Table 1: Natural parameter mapping to the expression of the mean for common distributions used for presence/absence or count data, for different choices of the link function.

Association type	Support	Recall (%)	Precision (%)	F1-score (%)
Neutral	[12, 380]	60.75	98.64	74.50
Negative	[2, 20]	72.00	34.02	41.23
Positive	[2, 20]	77.60	17.60	26.72
Averages	-	62.45	94.71	73.09

Table 2: Association classification performances and class supports (number of true associations of each class). The averages are obtained by weighting the score of each association type by its support.